

University of Crete
Department of Computer Science

**Improving Generative Adversarial Networks and its
Applications in Speech Synthesis**

Ph.D. Thesis

Dipjyoti Paul

Heraklion

February 2024

Είμαι ο αποκλειστικός συγγραφέας της υποβληθείσας Διδακτορικής Διατριβής με τίτλο **Improving Generative Adversarial Networks and its Applications in Speech Synthesis**. Η συγκεκριμένη Διδακτορική Διατριβή είναι πρωτότυπη και εκπονήθηκε αποκλειστικά για την απόκτηση του Διδακτορικού διπλώματος του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης. Κάθε βοήθεια, την οποία είχα για την προετοιμασία της, αναγνωρίζεται πλήρως και αναφέρεται επακριβώς στην εργασία. Επίσης, επακριβώς αναφέρω στην εργασία τις πηγές, τις οποίες χρησιμοποίησα, και μνημονεύω επώνυμα τα δεδομένα ή τις ιδέες που αποτελούν προϊόν πνευματικής ιδιοκτησίας άλλων, ακόμη κι εάν η συμπερίληψή τους στην παρούσα εργασία υπήρξε έμμεση ή παραφρασμένη. Γενικότερα, βεβαιώνω ότι κατά την εκπόνηση της Διδακτορικής Διατριβής έχω τηρήσει απαρέγκλιτα όσα ο νόμος ορίζει περί πνευματικής ιδιοκτησίας και έχω συμμορφωθεί πλήρως με τα προβλεπόμενα στο νόμο περί προστασίας προσωπικών δεδομένων, με τις αρχές της ηθικής και δεοντολογίας της έρευνας και της εν γένει ακαδημαϊκής δεοντολογίας.

Improving Generative Adversarial Networks and its Applications in Speech Synthesis

Submitted by

Dipjyoti Paul

in partial fulfilment of the requirements for the Doctor of Philosophy degree in Computer Science

Author



Dipjyoti Paul

Department of Computer Science

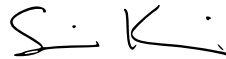
Examination Committee:

Supervisor



Yannis Stylianou, Professor, University of Crete, Greece

Member



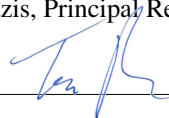
Simon King, Professor, University of Edinburgh, United Kingdom

Member



Yannis Pantazis, Principal Researcher, FORTH, Greece

Member



Panagiotis Tsakalidis, Professor, University of Crete, Greece

Member



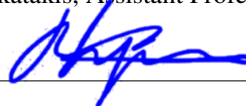
Nikolaos Komodakis, Assistant Professor, University of Crete, Greece

Member



Grigoris Tsagkatakis, Assistant Professor, University of Crete, Greece

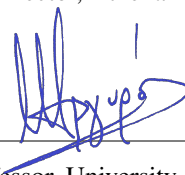
Member



Vasilis Katsouros, Research Director, Athena Research Center, Greece

Departmental Approval:

Chairman
of the Department



Antonis Argyros, Professor, University of Crete, Greece

Heraklion, February 2024

Acknowledgements

I want to express my profound and heartfelt gratitude to my esteemed mentor, Professor Yannis Stylianou, and co-supervisor, Yannis Pantazis, for his continuous support and encouragement, which provided me with insight to conduct research in a focused and organized way. More importantly, their guidance and valuable advice encouraged me throughout my PhD tenure.

I must also thank my secondary supervisor, Simon King, a Professor at the University of Edinburgh, United Kingdom, for his continuous support and timely monitoring of my work, which helped me keep the research on track. I should not forget to thank Professor Panagiotis Tsakalides, Assist. Professor Nikolaos Komontakis and Assist. Professor Grigoris Tsagkatakis from the University of Crete Greece, and Vasilis Katsouros, Research Director, Athena Research Center, for spending their valuable time reading and commenting on the manuscript.

While pursuing my Ph.D., I had the opportunity to collaborate with colleagues from the Speech Signal Processing Laboratory (SSPL) at the University of Crete, which proved to be an excellent working environment that greatly contributed to my personal and professional growth. I would extend my heartfelt thanks to Eirini Sisamaki, Dr Nagraj Adiga, Dr. Shifas Muhammed PV, and Dr. Anna Sfakianaki for making the time together most memorable. I extend a special thanks to Dr. George Kafentzis (Researcher, SSPL) for his continuous support and collaboration during this duration. I like to say a big ‘thank you’ to all my colleagues Dr. Petko Petkov, Dr. Langzhou Chen, Dr. Javier Latorre Martinez, Rafael Tsirmpas and Dr. Vasileios Tsiaras for all the scientific discussions. Furthermore, I am sincerely grateful to Eleftheria Lydaki for her invaluable assistance in helping me write my thesis.

While being a PhD researcher at the University, I was also associated with the European research network called ENRICH. I want to express my appreciation to my fellow colleagues in the ENRICH circle, whose timely exchange of research ideas during our peer-to-peer meetings has greatly contributed to the development and progress of my research.

This dissertation is incomplete without mentioning Soumi, my partner, for her utmost care, consideration, love, laughter, and kindness. Last but not least, I extend the greatest “thank you” to my parents for their prayers and support in reaching this moment in life.

Finally, I dedicate this work to the voices in my head that compelled me to write this thesis.

Thank you all!

Ευχαριστίες

Θα ήθελα να εκφράσω την βαθιά και εγκάρδια ευγνωμοσύνη μου στον αξιότιμο μέντορα μου, καθηγητή Yannis Stylianiou, και τον συνεπιβλέποντα μου Dr. Yannis Pantazis για τη συνεχή υποστήριξή τους και την ενθάρρυνση που μου παρείχαν κίνητρο για να διεξάγω την έρευνα μου με εστιασμένο και οργανωμένο τρόπο. Η καθοδήγησή τους και οι πολύτιμες συμβουλές τους με ενθάρρυναν στην διάρκεια των σπουδών μου.

Πρέπει ακόμα να ευχαριστήσω τον δευτερεύοντα επόπτη μου Simon King, Καθηγητή του Πανεπιστημίου του Εδιμβούργου, Ηνωμένο Βασίλειο, για τη συνεχή υποστήριξή του και τη συνεχή παρακολούθηση της εργασίας μου, που βοήθησε στη διατήρηση της έρευνάς μου σε καλό δρόμο. Δεν πρέπει να ξεχάσω να ευχαριστήσω τον καθηγητή Panagiotis Tsakalides, τον Επίκ. Καθηγητή Nikolaos Komontakis τον Επίκ. Καθηγητή Grigoris Tsagkatakis από το Πανεπιστήμιο Κρήτης, Ελλάδα, και τον Vassilis Katsouros, Κύριο Ερευνητή στο Ερευνητικό Κέντρο Αθήνας, που πέρα από την συγκαταθεσή τους να γίνουν μέλη του πάνελ της διατριβής μου, αφιέρωσαν τον πολύτιμο χρόνο τους για να διαβάσουν και να σχολιάσουν το κείμενο.

Κατά τη διάρκεια των σπουδών μου, είχα την ευκαιρία να συνεργαστώ με συναδέλφους από το Εργαστήριο Επεξεργασίας Σήματος Ομιλίας (SSPL) στο Πανεπιστήμιο της Κρήτης, το οποίο αποδείχθηκε ένα εξαιρετικό περιβάλλον εργασίας που συνέβαλε σημαντικά στην προσωπική μου και επαγγελματική μου ανάπτυξη. Θα ήθελα να ευχαριστήσω από καρδιάς την Eirini Sisamaki, τον Dr. Nagraj Adiga, τον Dr. Shifas Muhammed PV και την Dr. Anna Sfakianaki, που έκαναν τον χρόνο που περάσαμε μαζί αξέχαστο. Εκφράζω ιδιαίτερες ευχαριστίες στον Dr. George Kafentzis, Ερευνητής, SSPL, για τη συνεχή υποστήριξη και τη συνεργασία του. Θα ήθελα να πω ένα μεγάλο “ευχαριστώ” σε όλους τους συνεργάτες μου, τους Dr. Petko Petkov, Dr. Langzhou Chen, Dr. Javier Latorre Martinez, Rafael Tsirmpas και Dr. Vasileios Tsiaras, για όλες τις επιστημονικές συζητήσεις. Επιπροσθέτως, είμαι ειλικρινά ευγνώμων στην Eleftheria Lydaki για την πολύτιμη βοήθεια στην συγγραφή της διατριβής μου.

Κατά τη διάρκεια της διδακτορικής μου έρευνας στο Πανεπιστήμιο, ήμουν επίσης μέλος του Ευρωπαϊκού Ερευνητικού Δικτύου που ονομάζεται ENRICH. Θα ήθελα να εκφράσω την εκτίμησή μου προς τους συναδέλφους μου στον κύκλο του ENRICH, οι οποίοι με την έγκαιρη ανταλλαγή ερευνητικών ιδεών κατά τη διάρκεια των συναντήσεών μας συνέβαλαν σημαντικά στην ανάπτυξη και πρόοδο της έρευνάς μου.

Αυτή η διατριβή θα είναι ελλιπής αν δεν αναφέρω την Soumi, τη σύντροφό μου, για την απόλυτη φροντίδα, συνειδητοποίηση, αγάπη, γέλιο, και ευγένεια της. Τέλος, επεκτείνω το μεγαλύτερο “ευχαριστώ” στους γονείς μου για τις προσευχές τους και τη στήριξή τους για να φτάσω σε αυτή τη στιγμή στη ζωή μου.

Τέλος, αφιερώνω αυτό το έργο στις φωνές στο κεφάλι μου που με ώθησαν να γράψω αυτήν τη διατριβή.

Σας ευχαριστώ όλους!

Abstract

In this thesis, we explore significant advancements in machine learning. We focus on improving algorithms for Generative Adversarial Networks (GANs) and using them to improve image generation and computer speech generation.

Given the recent strides in GAN training, it is imperative to address and enhance the stability of the training process. Consequently, the first part of this thesis places a distinct emphasis on exploring algorithmic advancements tailored to improved GAN training. The objective is to delve into strategies that mitigate challenges and instabilities encountered during the training of GANs, thereby contributing to the overall refinement of the training process. We propose a novel weight-based algorithm aimed at strengthening the Generator. The theoretical underpinnings of this approach suggest that it outperforms the baseline algorithm by creating a more potent Generator at each iteration. Empirical results show substantial accuracy improvements and faster convergence rates across synthetic and image datasets. The improvements range between 5% and a remarkable 50%.

In the realm of GAN loss functions, we introduce a novel approach based on cumulant generating functions. This technique offers a fresh perspective on GAN loss functions by encompassing various divergences and distances based on cumulant generating functions and relies on a recently derived variational formula. We show that the corresponding optimization is equivalent to Rényi divergence minimization, thus offering a (partially) unified perspective of GAN losses: the Rényi family encompasses Kullback-Leibler divergence (KLD), reverse KLD, Hellinger distance, and χ^2 -divergence. Besides, it enhances training stability, particularly when weaker discriminators are employed, and demonstrates substantial improvements in synthetic image generation on datasets like CIFAR-10 and Imagenet.

Disentangled representations are crucial for capturing probability distributions and measuring divergences effectively. Mutual Information (MI) estimation, specifically through Kullback-Leibler Divergence (KLD), is commonly used to enforce disentanglement. We explore using variational representations, particularly based on minimizing Rényi divergences, as an alternative to KLD. Rényi divergences offer advantages in comparing different types of distributions. The text emphasizes using scalable neural network estimators for efficient MI estimation. Despite the potential for large statistical estimation, incorporating a variational representation based on Rényi divergences proves feasible and effective. The method is particularly successful in enhancing stability in real biological data, enabling the detection of rare sub-

populations even with limited samples. Moreover, the difficulty of precisely estimating divergences poses a significant challenge in many machine learning tasks, especially when dealing with high-dimensional datasets that can lead to increased variance. In addressing this challenge, we suggest a solution: incorporating an explicit variance penalty (VP) into the objective function of the divergence estimator. This added penalty aims to decrease the variance associated with the estimator, providing a potential way to enhance the accuracy of divergence estimations.

In this part of the thesis, our attention shifts to practical uses in speech synthesis, such as transforming one voice into another (voice conversion) and turning written text into spoken words (text-to-speech synthesis). We introduce innovative techniques for voice conversion that focus on many-to-many voice conversion. Leveraging concepts from the previous weight-based algorithm, we propose a weight multiplication approach to enhance the Generator’s gradients, making it more adept at fooling the Discriminator. This results in a robust Weighted StarGAN (WeStarGAN) system. Notably, WeStarGAN achieves significantly superior performance compared to conventional methods. It garners preference scores of 75% and 65% in terms of speech subjective quality and speaker similarity, respectively.

Neural vocoders often struggle with generalization, especially to unseen speakers and conditions. Here, we introduce the Speaker Conditional WaveRNN (SC-WaveRNN), which leverages speaker embeddings to improve speech quality and performance. This variant significantly outperforms baseline WaveRNN, achieving impressive improvements of up to 95% in terms of Mean Opinion Score (MOS) for unseen speakers and conditions. We extend this work further by implementing a multi-speaker text-to-speech (TTS) synthesis approach, effectively tackling zero-shot speaker adaptation.

In the realm of Universal TTS, we present a system capable of generating speech with various speaking styles and speaker characteristics, all without the need for explicit style annotation or speaker labels. We propose a novel approach based on Rényi Divergence and Disentangled Representation. This innovative method effectively reduces content and style leakage, resulting in substantial improvements in word error rate and speech quality. Our proposed algorithm achieves improvements of approximately 16-20% in MOS speech quality, alongside a 15% boost in MOS-style similarity.

Lastly, the growing use of digital assistants highlights the importance of TTS synthesis systems on modern devices. Ensuring clear speech generation in noisy environments is crucial. Our innovative transfer learning approach in TTS harnesses the power of amalgamating two effective strategies: Lombard speaking style data and Spectral Shaping and Dynamic Range Compression (SSDRC). This extended system, Lombard-SSDRC TTS, significantly improves intelligibility, with relative enhancements ranging from 110% to 130% in speech-shaped noise (SSN) and 47% to 140% in competing-speaker noise (CSN) compared to state-of-the-art TTS methods. Subjective evaluations further confirm substantial improvements, with a median keyword correction rate increase of 455% for SSN and 104% for CSN compared to the baseline TTS method.

Περίληψη

Σε αυτή τη διατριβή, εξετάζουμε σημαντικές προόδους στον τομέα της μηχανικής μάθησης. Generative Adversarial Networks (GANs) και στη χρήση τους για τη βελτίωση της δημιουργίας εικόνων και του τρόπου που οι υπολογιστές παράγουν ομιλία.

Δεδομένων των πρόσφατων αλλαγών στην εκπαίδευση των GANs, είναι επιτακτική η ενασχόληση και η βελτίωση της σταθερότητας της διαδικασίας εκπαίδευσης. Επομένως, το πρώτο μέρος αυτής της διατριβής δίνει ξεχωριστή έμφαση στην διερεύνηση αλγοριθμικών βελτιώσεων με σκοπό την καλύτερη εκπαίδευση GANs. Στόχος είναι η διεύθυνση σε στρατηγικές που αντιμετωπίζουν δυσκολίες και αστάθειες κατά την εκπαίδευση των GANs, και επομένως συνεισφέρουν στην συνολική αναβάθμιση της διαδικασίας εκπαίδευσης. Προτείνουμε έναν καινοτόμο βαρο-κεντρικό αλγόριθμο που στοχεύει στην ενίσχυση της Γεννήτριας. Τα θεωρητικά θεμέλια αυτής της προσέγγισης υποδεικνύουν καλύτερες επιδόσεις σε σχέση με τον κατεστημένο αλγόριθμο, με την δημιουργία μιας πιο ικανής Γεννήτριας σε κάθε επανάληψη. Εμπειρικά αποτελέσματα στηρίζουν αυτή την υπόθεση, αναδεικνύοντας σημαντική βελτίωση στην ακρίβεια και ταχύτερους ρυθμούς σύγκλισης μεταξύ συνθετικών συλλογών δεδομένων και συλλογών δεδομένων με εικόνες. Το ποσοστό βελτίωσης κυμαίνεται ανάμεσα σε ένα 5% και ένα εντυπωσιακό 50%.

Αναφορικά με τις συναρτήσεις κόστους, εισάγουμε μια νέα προσέγγιση βασισμένη σε ανθρωπιστικές γεννήτριες συναρτήσεις. Αυτή η μέθοδος προσφέρει μία νέα οπτική στις συναρτήσεις κόστους στα GANs, με την χρήση ενός μεγάλου εύρους αποκλίσεων και αποστάσεων, βασισμένων σε ανθρωπιστικές γεννήτριες συναρτήσεις, και στηρίζεται σε μία πρόσφατη σχέση διακυμάνσεων. Δείχνουμε ότι η αντίστοιχη βελτιστοποίηση είναι ισοδύναμη με την μέθοδο ελαχιστοποίησης της απόκλισης του Renyi, και άρα προσφέρει μια (μερικώς) καθολική οπτική στα κόστη GANs: η οικογένεια Renyi χρησιμοποιεί Kullback-Leibler απόκλιση KLD, αντίστροφο KLD, απόσταση Hellinger απόκλιση χ^2 . Συγχρόνως, βελτιώνει την σταθερότητα εκπαίδευσης, ιδίως όταν χρησιμοποιούνται πιο αδύναμοι διακριτές, και αναδεικνύει σημαντική βελτίωση στην παραγωγή συνθετικών εικόνων σε συλλογές δεδομένων όπως CIFAR-10 και Imagenet .

Οι αποσυνδεδεμένες αναπαραστάσεις είναι απαραίτητες για την αποτύπωση των κατανομών πιθανοτήτων και την μέτρηση της απόκλισης. Η εκτίμηση της Αμοιβαίας Πληροφορίας, συγκεκριμένα μέσω

του KLD, είναι μία συνήθης προσέγγιση για την ενίσχυση της αποσύνδεσης. Μελετάμε την χρήση μεταβαλλόμενων αναπαραστάσεων, βασισμένων ιδίως στην ελαχιστοποίηση των αποκλίσεων Renyi, ως εναλλακτική του KLD. Οι αποκλίσεις Renyi προσφέρουν πλεονεκτήματα στην σύγκριση διαφορετικών τύπων κατανομών. Το κείμενο δίνει έμφαση στην χρήση κλιμακούμενων νευρωνικών δικτύων εκτιμητών για την αποτελεσματική εκτίμηση της Αμοιβαίας Πληροφορίας. Παρά τη δυνατότητα για μια μεγάλη στατιστική εκτίμηση, η χρήση μίας μεταβαλλόμενης αναπαράστασης, βασισμένης στις αποκλίσεις Renyi, αποδεικνύεται εφικτή και αποτελεσματική. Η μέθοδος είναι ιδιαίτερα επιτυχής στην βελτίωση της σταθερότητας σε πραγματικά βιολογικά δεδομένα, επιτρέποντας την ανίχνευση σπάνιων υποπληθυσμών ακόμη και με περιορισμένα δείγματα. Ακόμη, η δυσκολία στην ακριβή εκτίμηση των αποκλίσεων αποτελεί μία σημαντική πρόκληση σε πολλά προβλήματα μηχανικής μάθησης, ειδικά σε μεγάλης διάστασης δεδομένα που οδηγούν σε αυξημένη διακύμανση. Για την αντιμετώπιση αυτής της πρόκλησης προτείνουμε μία λύση: την χρήση μίας ποινής διακύμανσης στην αντικειμενική συνάρτηση της εκτίμησης της απόκλισης. Αυτή η πρόσθετη ποινή στοχεύει στην μείωση της διακύμανσης που σχετίζεται με τον εκτιμητή, παρέχοντας ένα πιθανό τρόπο βελτίωσης της ακρίβειας της εκτίμησης των αποκλίσεων.

Σε αυτό το μέρος της διατριβής, η προσοχή μας στρέφεται στις πρακτικές χρήσεις της σύνθεσης φωνής, όπως η μετατροπή μίας φωνής σε άλλη (μετασχηματισμός φωνής) και η παραγωγή λόγου από κείμενο (κείμενο-σε-φωνή-σύνθεση, TTS). Εισάγουμε καινοτόμες τεχνικές για μετασχηματισμό φωνής που στοχεύουν κυρίως στον πολλές-σε-πολλές μετασχηματισμό φωνής. Χρησιμοποιώντας έννοιες από τον προηγούμενο βαρο-κεντρικό αλγόριθμο, προτείνουμε μια προσέγγιση πολλαπλασιασμού βαρών για την βελτίωση των παραγώγων της Γεννήτριας, καθιστώντας την πιο ικανή στο να 'ξεελαεί' τον Διακριτή. Αυτό οδηγεί σε ένα εύρωστο σύστημα Weighted StarGAN (WeStarGAN). Είναι αξιοσημείωτο ότι το WeStarGAN επιτυγχάνει σημαντικά ανώτερη επίδοση σε σχέση με συμβατικές μεθόδους. Σημειώνει σκορ επίδοσης της τάξης του 75% και 65% σε ότι αφορά την υποκειμενική ποιότητα φωνής και την ομοιότητα ομιλητή αντίστοιχα.

Οι νευρωνικοί vocoders συχνά αντιμετωπίζουν δυσκολίες στην γενίκευση, ειδικά σε άγνωστους ομιλητές και συνθήκες. Εδώ, εισάγουμε το Speaker Conditional WaveRNN (SC-WaveRNN), που χρησιμοποιεί ενσωματώσεις ομιλητών για την βελτίωση της ποιότητας της φωνής και της επίδοσης. Αυτή η εναλλακτική ξεπερνά σημαντικά το βασικό WaveRNN, επιτυγχάνοντας εντυπωσιακή βελτίωση της τάξης έως και 95% σε ότι αφορά το Σκορ Μέσης Άποψης (MOS) για άγνωστους ομιλητές και συνθήκες. Ως επιπλέον επέκταση υλοποιούμε μία προσέγγιση πολλαπλών-ομιλητών κείμενο-σε-φωνή σύνθεσης, αντιμετωπίζοντας την προσαρμογή σε άγνωστους κατά την εκπαίδευση ομιλητές.

Αναφορικά με το Universal TTS, παρουσιάζουμε ένα σύστημα, ικανό να παράγει φωνή με ποικίλα στυλ ομιλίας και χαρακτηριστικά ομιλητή, χωρίς την ανάγκη επισημείωσης του στυλ ή του ομιλητή. Παρουσιάζουμε μία νέα προσέγγιση βασισμένη στην Απόκλιση Renyi και την αποσυνδεδεμένη αναπαράσταση. Αυτή η καινοτόμα μέθοδος μειώνει αποτελεσματικά την διαρροή περιεχομένου και στυλ,

επιφέροντας ουσιώδη βελτίωση στον ρυθμό λάθος λέξεων και στην ποιότητα φωνής. Ο προτεινόμενος αλγόριθμος μας επιτυγχάνει βελτίωση περίπου 16%- 20% στην ποιότητα φωνής MOS, μαζί με μία αναβάθμιση της τάξης του 15% στην ομοιότητα κατά MOS

Τέλος, η αυξανόμενη χρήση ψηφιακών βοηθών τονίζει την σημασία των συστημάτων TTS στις σύγχρονες συσκευές. Η εξασφάλιση της παραγωγής καθαρού λόγου σε θορυβώδη περιβάλλοντα είναι επιτακτική. Η καινοτόμα προσέγγιση μας μεταφοράς μάθησης στο TTS αξιοποιεί τη δύναμη του συνδυασμού δύο αποτελεσματικών στρατηγικών: δεδομένα στυλ ομιλίας Lombard και SSDRC. Αυτό το επεκταμένο σύστημα, Lombard-SSDRC TTS, βελτιώνει σημαντικά την κατανοησιμότητα, με σχετικές αναβαθμίσεις που κυμαίνονται από 110% έως 130% στο θόρυβο με μορφή φωνής (SSN) και από 47% έως 140% στο θόρυβο από ανταγωνιστές-ομιλητές (CSN), συγκριτικά με σύγχρονες μεθόδους TTS. Υποκειμενικές αξιολογήσεις επιβεβαιώνουν περαιτέρω σημαντική βελτίωση, με μια αύξηση στο μέσο ρυθμό διόρθωσης λέξεων κλειδιών της τάξης του 455% στο (SSN) και 104% στο (CSN) σε σχέση με την βασική μέθοδο TTS.

Abbreviations and symbols

STFT	Short-time Fourier transform
ISTFT	Inverse STFT
SNR	Signal to noise ratio
SE	Speech enhancement
LE	Listening enhancement
FFT	Fast Fourier transform
F0	Fundamental frequency
SS	Spectral shaping
DRC	Dynamic range compression
2D	Two-dimensional
Bi	Bi-directional
GAN	Generative adversarial network
WGAN	Wasserstein GAN (WGAN)
GP	Gradient penalty
WeGAN	Weighted Generative adversarial network
KLD	Kullback–Leibler divergence
VP	Variance penalty
MI	Mutual information
DNE	Divergence neural estimation
CNN	Convolutional neural networks
LSTM	Long short-term memory
RNN	Recurrent neural network
GRU	Gated recurrent unit
SGD	Stochastic gradient descent
TTS	Text-to-speech synthesis
VC	Voice Conversion
SV	Speaker verification
SC-WaveRNN	Speaker Conditional WaveRNN
RDDR	Rényi Divergence based Disentangled Representation

Table 1: Abbreviations.

Σ	Summation
Π	Multiplication
$p(\cdot)$	Probability of
$f(\cdot)$	Function of
$ \cdot $	Magnitude of
t	Time dimation

Table 2: Symbols.

Contents

Title	1
Acknowledgements	3
Acknowledgements	5
Ευχαριστίες	7
Abstract	9
Περίληψη	11
Abstract	15
List of Tables	23
List of Figures	25
1 Introduction	25
1.1 Speech Production Mechanism	25
1.2 Thesis Contribution	26
I Generative Adversarial Networks	31
2 Introduction	33
2.1 GAN Preliminaries	33

2.2	Training Issues	34
2.3	Variants of GANs	36
3	Training Generative Adversarial Networks with Weights	43
3.1	Introduction	43
3.2	GAN formulation	44
3.3	Weighted GAN algorithm	45
3.3.1	Theoretical properties of WeGAN algorithm	45
3.4	Results	46
3.4.1	An illustrative example	46
3.4.2	MNIST	47
3.4.3	CIFAR	48
3.5	Conclusions	49
4	Cumulant GAN	51
4.1	Introduction	51
4.2	Background	52
4.2.1	Wasserstein GAN	52
4.2.2	Cumulant Generating Functions	53
4.3	Cumulant GAN	53
4.3.1	Definition	53
4.3.2	A Variational Formula for Rényi Divergence	54
4.3.3	Concavity Property of Cumulant GAN	55
4.3.4	KLD, Reverse KLD and Rényi Divergence as Special Cases	55
4.3.5	Cumulant GAN as a Weighted Version of the SGD Algorithm	58
4.3.6	Convergence Guarantees for Linear Discriminator	60
4.4	Cumulant GAN Implementation	62

4.5	Demonstrations	63
4.5.1	Traversing the (β, γ) -plane: from Mode Covering to Mode Selection	63
4.5.2	Learning the Covariance Matrix of a Multivariate Gaussian	64
4.5.3	Image Generation	65
4.6	Conclusions and Future Directions	70
 II Disentanglement Learning		77
 5 Advancements in Neural-Based Divergence Estimation		79
5.1	Introduction	79
5.1.1	Related Work	80
5.2	Variational Formulas for Rényi and f -Divergences.	81
5.3	Statistical Estimators and Variance Reduction	82
5.3.1	Variance Penalty	82
5.3.2	Variance-Reduced Divergence Estimation Algorithm	84
5.4	Proofs of Transformed Variational Formula Identities	84
5.5	Bias Bounds	89
5.6	Results on Synthetic Datasets	92
5.7	Real Data Applications	93
5.7.1	Detecting Rare Biological Sub-Populations	93
 III Speech Synthesis		95
 6 Introduction		97
6.1	Voice Conversion	98
6.2	Text to Speech Synthesis	99
6.2.1	Deep Learning based Speech Synthesis	99
6.3	Intelligible Speech Synthesis	103

7	Non-parallel Voice Conversion using Weighted Generative Adversarial Networks	105
7.1	Introduction	105
7.2	GAN Architectures	108
7.2.1	Star Generative Adversarial Networks	108
7.2.2	Training StarGAN with Weights (WeStarGAN)	109
7.3	Experimental Setup	110
7.3.1	Experimental conditions	110
7.3.2	Network architectures	111
7.4	Results and Discussion	111
7.5	Conclusion	113
8	Speaker Conditional WaveRNN: Towards Universal Neural Vocoder for Unseen Speaker and Recording Conditions	115
8.1	Introduction	115
8.2	Neural Speaker Encoder	117
8.2.1	Training Encoder Network	117
8.2.2	Generalized End-to-End Loss	118
8.3	Speaker-conditional WaveRNN	119
8.3.1	Preliminaries	119
8.3.2	Training WaveRNN with Speaker Embeddings	120
8.4	Zero-shot Text-to-Speech	121
8.5	Experimental Setup	122
8.6	Results and Discussion	122
8.6.1	Universal vocoder	122
8.6.2	Zero-shot TTS Synthesis	123
8.7	Conclusions	124
9	Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learn-	

ing based on Rényi Divergence Minimization	125
9.1 Introduction	125
9.2 Universal TTS (UTTS)	127
9.2.1 Speaker Encoder	127
9.2.2 Style Encoder	127
9.2.3 TTS Module	128
9.3 Proposed Disentangled Representation	128
9.3.1 Preliminaries	128
9.3.2 Rényi Divergence based Disentangled Representation	129
9.4 Results and Discussion	130
9.4.1 Objective Evaluation	131
9.4.2 Subjective Evaluation	132
9.4.3 Disentangled Representation Learning using variance reduction method	132
9.5 Conclusions	133
10 Enhancing Speech Intelligibility in TTS using Speaking Style Conversion	135
10.1 Introduction	135
10.2 Factors defining speech intelligibility	136
10.3 Spectral shaping and dynamic range compression (SSDRC)	139
10.3.1 Spectral shaping (SS):	139
10.3.2 Dynamic range compression (DRC):	141
10.4 Neural TTS architecture	143
10.4.1 Tacotron	143
10.4.2 WaveRNN	144
10.5 Transfer learning	145
10.6 Database and Hyperparameters Selection	145
10.7 Observations and discussion	146

10.8 Conclusions and perspective	148
Conclusions and Future Work	149
11 Conclusions and Future Work	149
11.1 Overview	149
11.2 Future research directions	151
A Publications	153
Bibliography	157

List of Tables

1	Abbreviations.	15
2	Symbols.	16
4.1	Inception scores on CIFAR-10 dataset.	67
4.2	Inception scores on Imagenet dataset.	67
5.1	Mean values and standard deviation for the histograms shown in Figure 5.3.	94
8.1	Objective evaluation tests.	122
9.1	Objective evaluation tests. Lower scores indicate better performance.	131
9.2	Average cosine-similarity evaluation.	131
9.3	MOS scores (95% confidence interval) of audio quality and speaking style similarity for different TTS modules.	132
9.4	Objective evaluation tests. Lower scores indicate better performance.	133
10.1	$SII B^{Gauss}$ intelligibility measure at different SNR levels under speech-shaped and competing-speaker noise.	146

List of Figures

1.1	Human speech production system.	26
1.2	(a) Time domain waveform, and (b) spectrum and corresponding spectral envelope of a voiced sound [i/] uttered by a male speaker.	27
1.3	(a) Time domain waveform, and (b) spectrum and corresponding spectral envelope of an unvoiced sound [s/] uttered by a male speaker.	27
3.1	Stochastic gradient ascent/descent training of WeGAN. For a direct comparison with the original GAN, we follow the formulation of [GPM ⁺ 14].	44
3.2	Upper & Middle plot: Relative improvement as a function of the epochs in terms of mean MMD with respect to vanilla GAN for a mixture of 8 Gaussians. Lower values for η resulted in improved convergence of WeGAN (lines with circles, squares and stars). Lower plot: Similar to the other plots for Wasserstein GAN. Higher values for η gave faster convergence while IWGAN is not applicable.	47
3.3	Relative improvement as a function of the epochs in terms of IS (upper plot) and FID (lower plot) with respect to vanilla GAN for the MNIST digit dataset. As in the benchmark example, lower values for η result in improved convergence of the WeGAN.	48
3.4	Similar to Fig. 3.3 but for the CIFAR-10 dataset. Improvements still happen but they are less prominent while the performance metrics unfortunately produce inconsistent results.	49
4.1	Special cases of <i>cumulant GAN</i> . Line defined by $\beta + \gamma = 1$ has a point symmetry. The central point, $(\frac{1}{2}, \frac{1}{2})$, corresponds to the Hellinger distance. For each point, $(\alpha, 1 - \alpha)$, there is a symmetric one, i.e., $(1 - \alpha, \alpha)$, which has the same distance from the symmetry point. The respective divergences have reciprocal probability ratios (e.g., KLD & reverse KLD, χ^2 -divergence & reverse χ^2 -divergence, etc.).	58

4.2	Interpretation of <i>cumulant GAN</i> as a weighted variation of SGD for $\beta, \gamma > 0$. Both real and generated samples for which the discriminator outputs a value closer to the decision boundary are assigned with larger weights because these are the samples which most probably confuse the discriminator.	60
4.3	Generated samples using the Wasserstein distance using clipping (1st row), KL divergence (2nd row), reverse KLD (3rd row) and Hellinger distance (last row). The boundedness condition is not enforced on this example but it is necessary to be satisfied when the hyper-parameters take negative values.	62
4.4	Covariance estimation error for the exact cumulant loss function (upper plot) and for the statistically-approximated cumulant loss function (lower plot).	64
4.5	Inception score for CIFAR-10 using various hyper-parameters of cumulant GAN and various architectures. In all cases, WGAN has a lower inception score relative to the cumulant GAN with the hyper-parameter corresponding to Hellinger minimization achieving the best overall performance.	68
4.6	Same as Fig. 4.5 but for ImageNet. Cumulant GAN achieves higher inception score relative to WGAN for both weak (left panel) and strong (right panel) discriminator.	69
4.7	WGAN: Samples of CIFAR-10 from generator and discriminator trained with convolutional networks.	70
4.8	KLD: Samples of CIFAR-10 from generator and discriminator trained with convolutional networks.	71
4.9	Reverse KLD: Samples of CIFAR-10 from generator and discriminator trained with convolutional networks.	71
4.10	Hellinger: Samples of CIFAR-10 from generator and discriminator trained with convolutional networks.	72
4.11	WGAN: Samples of CIFAR-10 from generator and discriminator trained with residual networks.	72
4.12	KLD: Samples of CIFAR-10 from generator and discriminator trained with residual networks.	73
4.13	Reverse KLD: Samples of CIFAR-10 from generator and discriminator trained with residual networks.	73
4.14	Hellinger: Samples of CIFAR-10 from generator and discriminator trained with residual networks.	74

4.15	WGAN: Samples of ImageNet from generator and discriminator trained with residual networks.	74
4.16	KLD: Samples of ImageNet from generator and discriminator trained with residual networks.	75
4.17	Reverse KLD: Samples of ImageNet from generator and discriminator trained with residual networks.	75
4.18	Hellinger: Samples of ImageNet from generator and discriminator trained with residual networks.	76
5.1	Comparison between the estimator without VP (DNE) and with VP (DNE- VP_λ) for Rényi divergence between two one-dimensional Gaussians with $Q = \mathcal{N}(0, 1.1)$ and $P = \mathcal{N}(0, 1)$. We use $N = 5K$ sample size, 512 as batch size and results are averaged over 100 i.i.d. runs. Left column: DNE and DNE- VP_λ estimators for increasing values of α . The variance of DNE becomes uncontrollably high for $\alpha > 3$. Middle column: Relative MedAE (the lower, the better) for varying penalty coefficient λ and two values of α . The relative MedAE for large values of λ is close to one which implies that the estimated value of DNE- VP_λ approaches zero. Right column: Relative MedAE for increasing sample size N . We additionally present a penalty coefficient that varies with sample size, shown in blue ($\lambda_N = \frac{500}{N}$ and $\lambda_N = \frac{2000}{N}$ for $\alpha = 0.5$ and $\alpha = 10$, respectively).	91
5.2	Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. The number of samples is set to $512K$ and batch size to 64. Panels with $R_{\alpha=0.5}$ in their titles present the Rényi-based MI with $\alpha = 0.5$ whereas the rest of the methods estimate the standard MI (i.e., the KL divergence). In each panel, the true values are shown as a step function (black line). The correlation coefficient of the Gaussian, ρ , for each step is: 0.3084, 0.4257, 0.5091, 0.5741, 0.6273 and 0.6717. The running estimates per minibatch are displayed as shadow blue curves. The dark blue curves shows the moving average of the estimated MI, with a bandwidth equal to 200 steps.	93

5.3	Comparison of DNE and DNE-VP $_{\lambda}$ estimators for Rényi divergence on biological data. The histograms of the estimated divergence value are constructed from 100 i.i.d. runs between datasets of $N = 20K$ samples each. Healthy dataset's distribution is denoted by P whereas healthy + 1% diseased dataset's by Q . Left column: Rényi divergence with $\alpha = 0.5$. Neither DNE nor DNE-VP $_{\lambda}$ are able to discriminate between the healthy and the 1% contaminated dataset. Right column: Rényi divergence with $\alpha = 1.1$. For this α value, VP is compulsory for a stable estimation of Rényi divergence. Furthermore, we are able to discriminate between healthy and 1% contaminated distributions with high accuracy (87.5%).	94
7.1	Overview of StarGAN (in black), consisting of two modules, a Discriminator D (identical neural network architecture is used for Classifier except for the last convolutional layer) and a Generator G . The weights (in red) are introduced during the training optimization process in our proposed algorithm.	109
7.2	Training algorithm of WeStarGAN. For a direct comparison with the StarGAN, we follow the formulation of [KKTH18].	110
7.3	Overview of StarGAN [KKTH18], consisting of two modules, a discriminator D and a generator G . In the input and output layers, h , w , and ch represent height, width, and number of channels, respectively. In each convolutional layer, k , c , and s denote kernel size, number of output channels and stride size, respectively. "Conv", "IN", "ReLU", "LReLU", and "Deconv" denote convolution, instance normalization, rectified linear unit, leaky rectified linear unit and transposed convolution, respectively. D_{cls} provides a probability distribution over domain labels where the domain corresponds to the number of speakers used to train VC.	111
7.4	Subjective preference test in (%) for speaker similarity and speech quality.	112
8.1	System overview of speaker encoder [WWPM18]. Features, speaker embeddings and similarity scores from different speakers are represented by different color codes. 'spk' denotes speakers and 'emb' represents embedding vectors.	117
8.2	UMAP projection of 10 utterances for each of the 10 random speakers. Different colors represent different speakers.	118
8.3	Block diagram of WaveRNN architecture.	119
8.4	Block diagram of proposed SC-WaveRNN training.	120
8.5	Block diagram of the proposed zero-shot TTS system.	121

8.6	Vocoder Subjective listening test (MOS) for speech quality and preference test in (%) for speaker similarity.	123
8.7	Zero-shot TTS Subjective listening test (MOS) for speech quality and preference test for (%) for speaker similarity.	124
9.1	System overview of our universal TTS framework. (a) Universal TTS conditioned on speaker (E_{sp}) and style (E_{st}) encoders that can synthesise well-controllable speech. TransformerTTS is employed as a backbone TTS infrastructure. (b) The proposed training protocol considers a novel adversarial RDDR approach combined with minimising the TTS reconstruction loss. A reference utterance is used to extract speaker and style factors, whereas, during inference (c), the system may take any arbitrary speaker or style as input.	127
9.2	Training algorithm of RDDR.	130
10.1	Average relative spectra for all frames (from Godoy et.al [GKS14]).	138
10.2	Spectral shaping fixed filter	140
10.3	Input-Output Envelope Characteristic (IOEC) Curve	141
10.4	Speech waveform modified for intelligibility with SSDRC algorithm.	143
10.5	Block diagram of Tacotron architecture.	144
10.6	A functional block diagram of the proposed adaptation techniques used in this study. Each block represents a TTS system (Tacotron + WaveRNN), which takes text as input and generates speech samples.	145
10.7	Box plot results for listeners' keyword scores across of methods for SSN and CSN.	147

Chapter 1

Introduction

Human speech is a vital part of how we communicate. It is not just about conveying information; it is also how we express our feelings and connect with others. Speech is not just words; it includes how we say them—our tone, pitch, speed, and rhythm—all of which make our conversations rich and meaningful. Speech helps us share knowledge, explain complex ideas, and learn from one another, whether we're in school, at work, or just chatting. It's not just about facts; it's also about emotions. By changing how we speak, we can convey a wide range of feelings, from happiness to sadness or anger. This emotional aspect makes our conversations deeper and more meaningful. Speech also brings people together. When we talk, share stories, and listen to each other, we build connections and understand each other better. It shows why we need technology that can understand and replicate human speech, including its information, emotions, and social aspects. [Sta80, KM82, RLL89]

Over the past decades, scientists and engineers have made remarkable advancements, uncovering tools that have had a lasting impact on society. Advances in speech coding algorithms and technologies have revolutionized voice communication and storage, making them more effective and efficient. The integration of natural voice in interactive systems, made possible by sophisticated speech synthesis algorithms, has elevated the understanding of human-machine communication to new heights, and the communication process has become smoother, more seamless, and potentially more productive due to these technological advancements. In the medical domain, speech analysis models and principles have provided profound insights into the complexities of the human speech production system. This, in turn, has assisted medical professionals in the swift and reliable detection of speech-related pathologies and anomalies. Algorithms designed to enhance speech in noisy conditions have significantly bolstered the resilience of terrestrial and satellite communications, ensuring clear and reliable transmission even in challenging environments. The entertainment industry has harnessed advanced speech transformation techniques to introduce life-like artificial voices in toys, films, and video games, enriching the overall user experience. As the field of speech processing continues to evolve, it is poised for even more applications in the future, especially with the strengthening convergence of computers, communications, and the Internet.

1.1 Speech Production Mechanism

Given that this thesis is focused on speech, it is essential to include a brief review of the human speech production mechanism from an acoustic standpoint [TM06, Lev92]. The illustrative diagram in Fig. 1.1 provides a comprehensive depiction of the intricate human vocal mechanism and the associated speech organs instrumental in producing speech. Voiced sound production is a nuanced orchestration

involving the vibration of the vocal folds induced by airflow, generating a quasi-periodic sound replete with harmonics. In contrast, unvoiced sounds, akin to random noise, result from turbulent airflow when they do not involve the use of the vocal cords. The conceptual framework of the source-filter model aptly captures the essence of speech production, portraying the vocal folds as the source and the vocal tract as the filter, dynamically altering the spectral attributes of the source signal. The vocal tract, a pivotal component encompassing the pharynx, oral cavity, and nasal cavity, influences the production of speech sounds.

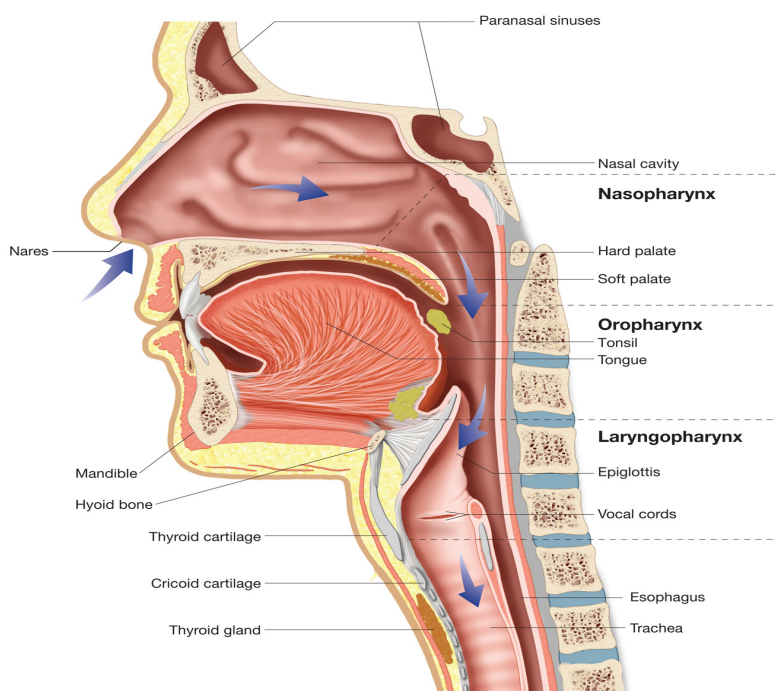


Figure 1.1: Human speech production system.

The temporal and spectral aspects of a voiced sound are portrayed in Fig.1.2, explaining the periodic nature of the waveform and the distinctive formants within the frequency domain spectrum. The pitch period, linked to the glottal cycle, and the fundamental frequency (F_0) are influenced by the dynamic characteristics of the vocal folds, with variations observed between male and female speakers. The spectral envelope shows the presence of formants, representing the resonant frequencies of the vocal tract.

Fig. 1.3 shows the unvoiced speech scenario, characterized by the absence of periodicity and harmonic structure, as evident in both the time and frequency domain plots. The speech perception resides in the amalgamation of the glottal flow wave with the vocal tract, the shape of which is inherently determined by the positioning of articulators. In summary, what makes each person's voice unique comes from a mix of different factors in how we speak. Two important factors are the pitch of our voice (fundamental frequency) and the distinctive patterns in our speech sounds (formants). Techniques that simplify these traits into a few key details, known as voice characterization, prove that these features are not only important but also practical for identifying people by their voices.

1.2 Thesis Contribution

This thesis is structured into three interrelated segments, each serving a distinct purpose in advancing our research objectives. The initial two sections are dedicated to comprehensively exploring the the-

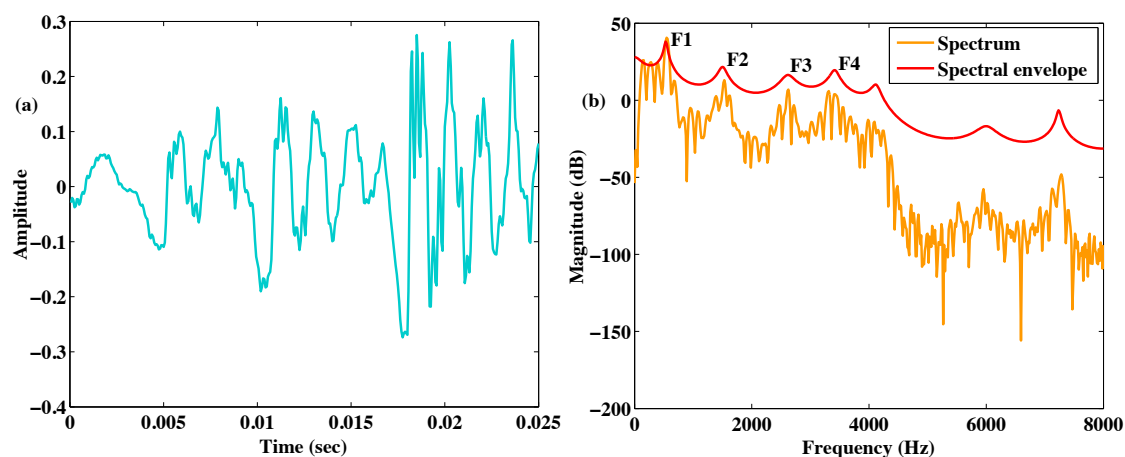


Figure 1.2: (a) Time domain waveform, and (b) spectrum and corresponding spectral envelope of a voiced sound [i/] uttered by a male speaker.

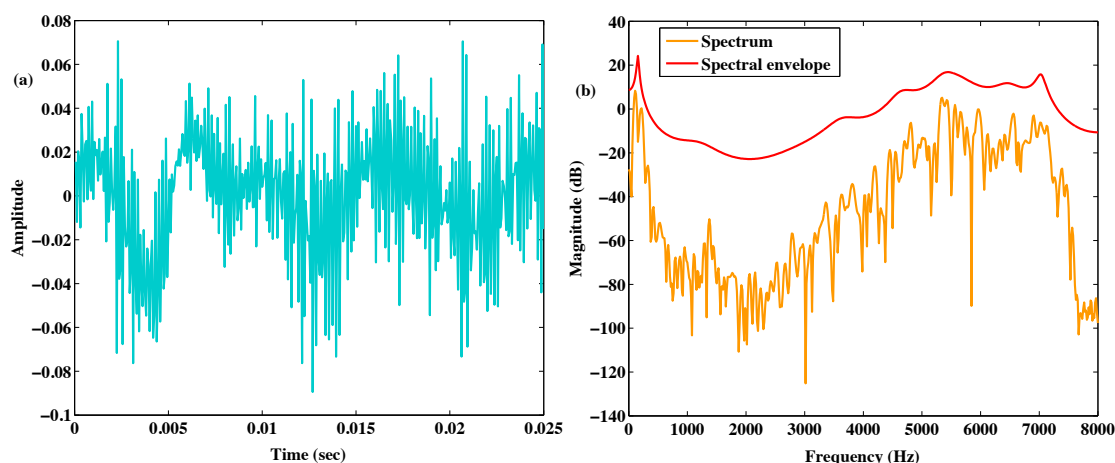


Figure 1.3: (a) Time domain waveform, and (b) spectrum and corresponding spectral envelope of an unvoiced sound [s/] uttered by a male speaker.

oretical underpinnings. Here, we engage in a detailed examination of the conceptual groundwork and methodological enhancements introduced, providing a robust theoretical framework for our research.

The subsequent part of the thesis seamlessly transitions from theory to practical implementation, where we apply the established theoretical framework and algorithmic modifications to the domain of speech synthesis. Within this phase, we demonstrate the real-world application of the theoretical advancements established in the first segment. This deliberate division into theoretical exploration and practical implementation facilitates a comprehensive understanding of both dimensions of our research endeavour. By adding these segments, we aim to contribute not only to the theoretical advancements in our field but also to their meaningful and practical implications in the domain of speech synthesis.

The key contributions of this thesis are outlined below:

1. Contribution to Generative Adversarial Models:

- (a) GANs showcased their capability to generate novel, pseudo-real, and high-quality data that closely mirrors their training set. It is crucial to acknowledge that training GANs poses challenges due to the inherent min-max game, and they are susceptible to a phenomenon known as

mode collapse. These challenges underscore the need for careful consideration and optimization when employing GANs in various applications. To mitigate these issues, we propose a novel GAN training algorithm, Weighted GAN (WeGAN), inspired by the multiplicative weight update method from Game Theory. We enhance Generator training by assigning higher weights to fake samples likely to deceive the Discriminator while reducing weights for confidently identified fake samples. Our contributions include improved training performance with minimal computational cost and rigorous arguments demonstrating that WeGAN’s weights reduce the loss function as effectively as equally weighted stochastic gradient descent for the Generator.

- (b) Despite the remarkable achievements of WeGANs, its training processes often exhibit instability, necessitating extensive experimentation to fine-tune parameters such as loss functions, optimization algorithms, and architectures. This thesis introduces a novel loss function grounded in cumulant generating functions, leading to the development of the Cumulant GAN. The key advantage of utilizing cumulants over expectations lies in their ability to capture higher-order information about underlying distributions, contributing to more effective learning processes without any mode collapses or training instabilities.

2. Contribution to Disentanglement Learning:

Model performance deteriorates if data representations are not invariant and disentangled. Model performance refers to how well a machine learning model accomplishes the task it was designed for and can vary depending on the specific task. High performance means the model achieves the desired outcome effectively and accurately, while low performance indicates that the model struggles to achieve the desired outcome. In the context provided, disentangled refers to the separation or isolation of different factors or features within the data representation. When data representations are disentangled, it means that the model can distinguish and manipulate individual factors independently of each other. Here, a good representation often involves capturing the probability distribution and measuring the divergences as a metric. One popular way to enforce disentangled representations is via MI estimation, which can be defined as the KLD between the joint distribution and the product of the marginals. Indeed, when MI equals zero, then the two random variables/vectors are independent.

- (a) Recent works [GPM⁺14, NCT16, CGK⁺02] have shown that variational representations of KLD can be utilized to estimate MI via scalable, flexible, and completely trainable neural network estimators. However, the statistical estimation of mutual information becomes exponentially large. To this end, we analyzed the feasibility of incorporating our proposed variational representation based on the minimization of Rényi divergences. Rényi divergences have several advantages over the commonly used KLD, including comparing heavy-tailed distributions and certain non-absolutely continuous distributions. Moreover, conventional divergence estimators exhibit robust performance in low-dimensional scenarios but encounter difficulties when confronted with large, high-dimensional datasets typical in contemporary machine learning applications. The challenge of accurately estimating divergences is a critical aspect of various machine learning tasks that often face heightened variance, particularly in high-dimensional datasets. In response to this challenge, we propose an innovative approach known as Variance Penalty (VP) to address and mitigate the variance associated with divergence estimators.

3. Contribution to Speech Synthesis

This thesis section addresses challenges within the realm of speech synthesis, specifically focusing on voice conversion and multi-speaker multi-style scenarios, collectively referred to as universal

Text-to-Speech (TTS) scenarios. In these contexts, accommodating all potential speaker variations and styles during training becomes impractical, prompting the need for better TTS systems.

- (a) WeGAN’s flexibility extends across diverse GAN types. We expanded the application of our weighting approach into voice conversion, introducing WeStarGAN, a variation of StarGAN tailored for non-parallel multi-domain voice conversion tasks. Despite incurring minor additional computational costs, this approach significantly enhanced the training process by reinforcing the generator at each minibatch iteration. Subjective evaluations revealed substantial improvements in sound quality and speaker similarity compared to baseline methods.
- (b) Neural vocoder techniques, driven by data-centric learning, often exhibit limitations in generalization due to the specialization of the training data [PPS20]. Additionally, in multi-speaker scenarios, it becomes impractical to encompass all potential in-domain (seen) and out-of-domain (unseen) cases within the training database. Seen refers to the speakers that are already present in the training, and unseen speakers are the new speakers during testing. To address these challenges, the proposed universal vocoder, Speaker Conditional WaveRNN (SC-WaveRNN), investigates the efficacy of incorporating explicit speaker information, specifically speaker embeddings, as a conditioning factor. This innovative approach aims to enhance the quality of generated speech across the widest possible range of speakers without necessitating adaptation or retraining. We extend our innovative approach to establish an efficient zero-shot TTS system. This advancement demonstrates that our proposed zero-shot TTS, coupled with a universal vocoder, can enhance both speaker similarity and the naturalness of synthetic speech, offering improvements for both seen and unseen speakers.
- (c) In the pursuit of creating a universal TTS synthesis system capable of replicating the characteristics and speaking style of a reference speaker, we are faced with some significant challenges: the potential occurrence of “content leakage” and “style leakage”. During training, content information is leaked into the style embeddings (“content leakage”) and speaker information into style embeddings (“style leakage”). We put forth a novel disentangled representation approach to address these issues, leveraging cumulant-generating functions in speech synthesis. Our system approximates and minimizes the Rényi divergence between content-style and style-speaker pairs. Finally, we integrate the Variance Penalty into speech representation learning, disentangling text, speaker, and style components, resulting in a marked improvement in training performance compared to baseline systems.
- (d) The growing prevalence of digital assistants underscores the vital role of TTS synthesis systems in modern devices. Ensuring clear speech generation in noisy environments is paramount. Our innovative TTS approach, Lombard-SSDRC, combines Lombard speaking style data with Spectral Shaping and Dynamic Range Compression (SSDRC). This extended system exhibits significant improvements, boasting relative enhancements ranging from 47% to 140% in speech-shaped noise (SSN) and competing-speaker noise (CSN) when compared to state-of-the-art TTS methods. Subjective evaluations further affirm substantial progress, revealing a median keyword correction rate increase of 455% for SSN and 104% for CSN in comparison to the baseline TTS method.

Part I

Generative Adversarial Networks

Chapter 2

Introduction

Machine learning has advanced impressively in recent years, achieving performance on par with or even surpassing human beings in many challenging tasks. This progress has been driven by developing new machine-learning models, such as generative models. Generative models are a type of machine learning model that can create new data similar to existing data. This makes them useful for a variety of tasks, such as generating realistic images and speech applications.

Generative adversarial networks (GANs) are generative models that learn to generate samples from a target distribution by training a generator network to produce synthetic data that appears similar to real data [GPAM⁺14]. The generator takes random noise as input and transforms it into synthetic samples, while the discriminator aims to distinguish between real and fake samples. The generator and discriminator improve iteratively through an adversarial training process, resulting in increasingly realistic synthetic data.

2.1 GAN Preliminaries

1. Adversarial Training:

The adversarial training process in GANs involves two main steps: training the discriminator and training the generator. Let's denote the target data as x , the generator network as G , and the discriminator network as D .

The discriminator's goal is to classify real samples from synthetic samples correctly. It is trained using a binary cross-entropy loss function defined as:

$$\mathcal{L}_D = -\frac{1}{m} \sum_{i=1}^m \left[\log D(x^{(i)}) + \log(1 - D(G(z^{(i)}))) \right]$$

Where m is the mini-batch size, $x^{(i)}$ represents real data samples, and $z^{(i)}$ represents random noise samples used as target data for the generator.

The generator aims to fool the discriminator by generating synthetic samples that are classified as real. Its objective is to minimize the following loss function:

$$\mathcal{L}_G = -\frac{1}{m} \sum_{i=1}^m \log D(G(z^{(i)}))$$

The generator seeks to maximize the probability of the discriminator misclassifying the generated samples as real.

2. Minimax Game and Optimization:

The training of GANs can be formulated as a minimax game between the generator and the discriminator. The objective is to find a Nash equilibrium [Kre89]. A Nash equilibrium refers to a stable state where both the generator and the discriminator have reached an optimal strategy, resulting in a balance between generating realistic samples and distinguishing between real and fake samples.

The minimax objective function for GANs is given by:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

where $p_{\text{data}}(x)$ represents the real data distribution, and $p_z(z)$ is the distribution of the input noise.

To solve the minimax game, both the generator and discriminator are optimized alternately. Popular optimization algorithms include stochastic gradient descent (SGD), Adam, and RMSprop. The discriminator is updated by ascending its gradient concerning its parameters, while the generator is updated by descending its gradient.

3. Convergence Analysis:

Ensuring convergence of the GAN training process is a fundamental challenge. GANs often suffer from mode collapse, where the generator produces a limited variety of samples, and the discriminator fails to provide meaningful feedback.

2.2 Training Issues

Indeed, despite their impressive accomplishments, GANs can be used for several practical challenges. One of the most prominent issues is the instability of the training process, which can lead to problems like mode collapse or oscillation. Additionally, evaluating the quality of generated data poses a challenge since conventional assessment criteria may not fully capture the diversity and realism of the synthesized samples. Moreover, GANs have been found to exhibit biases that can potentially mirror the biases inherent in the training data. These challenges underscore the ongoing efforts required to harness the full potential of GANs while addressing their limitations and ethical considerations [SC21, SMAMG21]. Here are some of the challenges that GANs face during training:

- Mode Collapse:

Issues like mode collapse [SGZ⁺16, ACB17] refer to a phenomenon in which the generator of a GAN produces a limited variety of outputs despite having a diverse range of input noise samples. In other words, the generator fails to capture the full complexity of the target distribution and instead converges to a single mode or a small subset of modes. As a result, the generated samples lack diversity and do not accurately represent the entire dataset.

Mode collapse can occur for various reasons, including the discriminator becoming too strong or dominant. If the discriminator is too effective at distinguishing between real and fake samples, it provides strong and consistent feedback to the generator. This can lead to the generator overfitting to a specific subset of the data, neglecting other modes.

- Instability:

GAN training is inherently unstable, meaning that the training process can be challenging to control and prone to fluctuations [AB17, MNG18, ZXL⁺19]. Several factors contribute to this instability:

Sensitivity to network architecture and hyperparameters: The performance of GANs is highly sensitive to the choice of network architecture and hyperparameters, such as learning rate, batch size, and regularization methods. The learning rate is a hyperparameter that determines the step size during optimization, affecting the convergence speed and stability of the training process; batch size refers to the number of training examples processed simultaneously in each iteration, influencing computational efficiency and generalization performance and regularization methods are techniques used to prevent overfitting by introducing constraints or penalties on the model's parameters, promoting simpler and more generalizable solutions. These factors are crucial in training GANs, where small changes in architecture and hyperparameters can significantly impact training dynamics and the quality of generated samples, highlighting the importance of careful selection and tuning for optimal GAN performance.

Balancing the generator and discriminator: GANs involve a delicate balance between the generator and discriminator networks. If the discriminator is too weak, it fails to provide meaningful feedback to the generator, hindering its learning. On the other hand, if the discriminator is too strong, it can dominate the training process and lead to mode collapse. Finding the right equilibrium between the two networks is crucial for stable GAN training.

Oscillations and convergence issues: GAN training can suffer from oscillations, where the performance of the generator and discriminator fluctuate, hindering convergence. These oscillations can make it challenging to determine when the training process has converged to an optimal solution.

- Data scarcity:

One of the challenges in GAN training is the requirement for a large amount of training data. GANs are data-hungry models that rely on a diverse and representative dataset to learn the underlying distribution and generate realistic samples. However, in many domains, obtaining a large amount of labeled or high-quality data can be difficult [LWY19, BCNM06, WC18].

Overfitting to limited data: When the training dataset is limited, GANs may struggle to capture the true distribution adequately. The generator might memorize the few available samples, leading to poor generalization and limited diversity in the generated samples.

Domain shift and dataset bias: GANs are sensitive to the distribution of the training data. If the available data does not sufficiently cover the target data distribution, the generated samples may exhibit biases or fail to represent the desired characteristics of the target domain accurately.

Preprocessing and data augmentation challenges: In some domains, preprocessing and augmenting the data to increase its diversity and quality can be challenging. For instance, in medical imaging, obtaining large annotated datasets can be time-consuming and expensive, limiting the potential for GAN training.

Addressing data scarcity in GAN training often involves techniques such as transfer learning, data augmentation, and utilizing auxiliary data sources to enhance the diversity and quantity of available training samples.

To mitigate mode collapse and stabilize the training process, various regularization techniques have been proposed [KAHK17, MLX⁺17, GSW⁺21]. Feature matching aims to match the statistics of the real data's intermediate features to those of the generated samples. It involves minimizing the discrepancy between the expected feature values of the real and fake samples. The generator is trained to generate samples that match the statistics of the real data's intermediate representations.

2.3 Variants of GANs

Since their introduction [GPAM⁺14], Generative Adversarial Networks (GANs) have undergone remarkable advancements, resulting in various specialized variants that excel in data generation across diverse domains. Conditional GANs, for instance, enable the generation of data based on specific conditions or desired attributes, facilitating tasks like synthesizing images of a particular class. CycleGANs have proven effective in image-to-image translation, even when unavailable paired data. StyleGAN, known for its versatility, can generate images with various styles and distinctive features. GANs have also expanded beyond visual domains, showing potential in generating textual, musical, and 3D modelling, future cities, time series data, and many other areas.

1. Deep Convolutional GANs (DCGANs):

Deep Convolutional GANs (DCGANs) improve upon the traditional GAN architecture by incorporating deep convolutional neural networks (CNNs) in both the generator and discriminator [RMC15, OOS16]. This architectural enhancement enables DCGANs to capture spatial information and generate high-quality images.

In DCGANs, the generator and discriminator architectures are designed with convolutional layers, allowing them to handle image data effectively. The generator takes random noise as input and progressively upsamples it using transposed convolutions to generate synthetic images. The discriminator, on the other hand, processes the input image through convolutional layers to classify it as real or fake.

DCGANs have demonstrated superior performance in generating visually appealing images with sharp details and realistic textures. Using convolutional layers enables the model to exploit the spatial relationships present in images, capturing local features and generating coherent and visually consistent samples.

2. Wasserstein GAN (WGAN):

Wasserstein GAN (WGAN) introduces the Wasserstein distance, also known as Earth Mover’s Distance (EMD), as a measure of discrepancy between the real and generated distributions [ACB17]. By using the Wasserstein distance instead of traditional divergence measures, WGAN improves the stability of GAN training and mitigates mode collapse.

The key idea behind WGAN is using a critic network instead of a traditional discriminator. The critic network assigns a real-valued score to each sample, estimating the Wasserstein distance between the real and generated distributions. The generator’s objective in WGAN is to minimize this estimated Wasserstein distance, while the critic aims to maximize it.

The use of the Wasserstein distance offers several advantages over traditional GAN training. It provides a more informative and stable gradient signal to the generator, enabling better convergence and preventing mode collapse. Additionally, WGANs exhibit smoother training dynamics, making it easier to monitor and assess the progress of the training process.

The WGAN objective function can be expressed as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [D(x)] - \mathbb{E}_{z \sim p_z(z)} [D(G(z))]$$

where $p_{\text{data}}(x)$ represents the real data distribution and $p_z(z)$ represents the distribution of the input noise.

3. Wasserstein GAN with Gradient Penalty (WGAN-GP):

In addition to Wasserstein GAN (WGAN), there is a variant called WGAN with Gradient Penalty (WGAN-GP), which addresses some limitations of the original WGAN and improves the stability of training further [GAA⁺17b].

WGAN-GP introduces a gradient penalty term to the WGAN objective function to enforce the Lipschitz continuity constraint on the discriminator. The Lipschitz constraint ensures that the discriminator's gradients do not grow too large, leading to more stable training dynamics and improved convergence.

The objective function of WGAN-GP can be expressed as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [D(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [D(G(\mathbf{z}))] + \lambda \mathbb{E}_{\hat{\mathbf{x}} \sim p_{\hat{\mathbf{x}}}} \left[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2 \right]$$

where \mathbf{x} represents real data samples, \mathbf{z} represents the input noise, G is the generator, and D is the discriminator. $p_{\text{data}}(\mathbf{x})$ is the data distribution, and $p_z(\mathbf{z})$ is the noise distribution. $\hat{\mathbf{x}}$ is a randomly sampled point between real and generated samples defined as $\hat{\mathbf{x}} = \epsilon \mathbf{x} + (1 - \epsilon)G(\mathbf{z})$, where ϵ is sampled uniformly between 0 and 1. The term λ is a hyperparameter that controls the weight of the gradient penalty.

The objective function of WGAN with Gradient Penalty (WGAN-GP) consists of three terms: the Wasserstein distance between the discriminator's output on real samples and generated samples, the negative Wasserstein distance between the discriminator's output on generated samples, and the gradient penalty term that enforces the Lipschitz constraint on the discriminator. By minimizing this objective function, the generator aims to generate samples that can fool the discriminator, while the discriminator aims to distinguish between real and generated samples. The gradient penalty term encourages the discriminator to have gradients with a norm of 1, penalizing deviations from this value. WGAN-GP has demonstrated improved stability and convergence properties compared to traditional GANs and even the original WGAN. It alleviates the need for careful weight clipping in the discriminator, which was a requirement in the original WGAN. By imposing the Lipschitz constraint through the gradient penalty, WGAN-GP provides a more principled and effective way to enforce a stable training process while achieving high-quality generated samples. Overall, WGAN-GP is a powerful variant of GAN that combines the benefits of Wasserstein GAN with the gradient penalty technique to enhance training stability and generate high-quality samples.

4. Least Squares GAN (LSGAN):

Least Squares GAN (LSGAN) introduced in [MLX⁺17] presents an innovative approach to address a common issue in traditional GAN models. Traditional GANs use a discriminator modelled as a classifier with the sigmoid cross entropy loss function. However, this choice of loss function can lead to vanishing gradients during training, which hinders the deep representation learning process. LSGAN tackles this problem by introducing the least squares loss function for the discriminator.

Mathematically, in the LSGAN model, there are two loss functions: the Generator loss function (L_G) and the Discriminator loss function (L_D). L_G is defined as two times the expected value of the discriminator's response to the generator's output ($D(G(z))$) minus a specific constant (c) for fake data sampled from a noise distribution (z).

(L_D) comprises two terms: the first term is two times the expected value of the discriminator's response to real data ($D(x)$) minus another constant (b) for real data sampled from the actual data distribution, and the second term is similar to (L_G), involving fake data and another constant (a).

$$L_G = \frac{1}{2} \mathbb{E}_{z \sim p_z} (D(G(z)) - c)^2$$

$$L_D = \frac{1}{2} \mathbb{E}_{x \sim p_{data}} (D(x) - b)^2 + \frac{1}{2} \mathbb{E}_{z \sim p_z} (D(G(z)) - a)^2$$

This encoding scheme of a , b , and c helps define the labels for fake and real data and guides the discriminator's learning process. The LSGAN framework represents a significant improvement over traditional GANs, offering a solution to the vanishing gradient issue and enhancing the stability of GAN training.

5. Conditional GANs (cGANs):

Conditional GANs (cGANs) extend the GAN framework to enable conditional generation, where the generator generates samples conditioned on additional information. This additional information can be class labels, input images, or any other relevant conditioning variables [MO14, IZZE17].

The architecture of cGANs includes both a generator and a discriminator, similar to traditional GANs. However, in cGANs, the generator takes an additional input, known as the conditioning variable, alongside the random noise. This conditioning variable provides information to guide the generation process, allowing control over specific attributes or characteristics of the generated samples.

The discriminator in cGANs also considers the conditioning variable when assessing the authenticity of the generated samples. It aims to classify whether the pair of the conditioning variable and the generated sample is real or fake.

cGANs have greatly succeeded in various applications, including image synthesis, style transfer, and image-to-image translation tasks. By conditioning the generation process on specific attributes, cGANs enable targeted and controlled synthesis, allowing users to specify desired characteristics or transform input samples based on given conditioning variables.

The cGAN objective function can be formulated as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x, y \sim p_{data}(x, y)} [\log D(x, y)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(x, G(z)))]$$

where x represents the input data, y represents the conditioning variable, G is the generator, and D is the discriminator. The discriminator aims to maximize the probability of correctly classifying the pair (x, y) , while the generator aims to minimize it.

6. CycleGAN:

CycleGAN is a variant of GANs designed for unpaired image-to-image translation tasks [KXRS19, KXRS19]. Unlike cGANs that require paired data (input and corresponding output) for training, CycleGAN can learn mappings between two domains using only unpaired datasets. It aims to learn a mapping from one domain to another while preserving the underlying structure and content of the input images.

CycleGAN incorporates two generators and two discriminators. The generators map images from one domain to the other and vice versa. The discriminators assess the authenticity of the generated images and provide feedback to the generators. In addition to the adversarial loss, CycleGAN introduces cycle-consistency loss, which encourages the reconstructed images to be similar to the original input images when translated back and forth between the domains.

The objective function of CycleGAN can be expressed as follows:

$$\min_{G_{A \rightarrow B}, G_{B \rightarrow A}} \max_{D_A, D_B} \mathcal{L}_{GAN}(G_{A \rightarrow B}, D_B, A, B) + \mathcal{L}_{GAN}(G_{B \rightarrow A}, D_A, B, A) + \lambda \mathcal{L}_{cycle}(G_{A \rightarrow B}, G_{B \rightarrow A})$$

where A and B represent images from domains A and B, respectively. $G_{A \rightarrow B}$ and $G_{B \rightarrow A}$ are the generators, and D_A and D_B are the discriminators. \mathcal{L}_{GAN} denotes the adversarial loss, and \mathcal{L}_{cycle} represents the cycle-consistency loss. λ is a hyperparameter that controls the importance of the cycle-consistency loss.

CycleGAN has also demonstrated successful applications in speech-related tasks [KKH⁺18, KKH⁺20]. It has been utilized for non-parallel voice conversion, allowing transformation between different speakers without needing paired training data. Additionally, CycleGAN has been employed for mel-spectrogram conversion, enabling the conversion of speech representations across different acoustic characteristics. These applications showcase the versatility of CycleGAN in the field of speech processing, providing solutions for tasks like voice transformation and speech representation modification without relying on aligned training data.

7. Self-Attention GAN (SAGAN):

Self-Attention GAN (SAGAN) introduces a self-attention mechanism to the generator and discriminator architectures [ZXL⁺19]. This mechanism allows the model to focus on important spatial relationships across different image regions, improving the generation quality and reducing artefacts. SAGANs generate images with better global coherence and sharpness by capturing long-range dependencies.

The self-attention mechanism in SAGAN utilizes query, key, and value operations to compute attention weights for each spatial position in the image. These weights determine the importance of each position and are used to create an attention map. This attention map is then applied to the feature maps to compute the attended feature representation, enhancing the model's ability to capture global dependencies. SAGANs have demonstrated superior performance in image generation tasks, especially when long-range dependencies and fine details are crucial, such as generating high-resolution images or complex scenes.

8. Vector Quantized GAN (VQGAN):

Vector Quantized GAN (VQGAN) is an innovative approach that combines the power of GANs with vector quantization techniques to create really good images [ERO21]. First, it trains a special computer program to understand images and their hidden meanings. This program can take an image and turn it into a secret code that represents what's in the picture. This code is like a special language that only computers understand. Then, it uses this secret code to make better pictures. But here's the cool part: you can change this secret code to make the pictures look different. It's like having a magic wand to change the colors or shapes in a picture. However, there are some challenges. VQGAN needs lots of pictures to learn from and really powerful computers to work its magic. This means it's not something you can use quickly in everyday situations. Also, sometimes the pictures it makes can look kind of similar because of the secret code.

9. StyleGAN and StyleGAN2:

StyleGAN [KLA19] and its subsequent improvement, StyleGAN2 [KALL20], introduce a novel architecture that allows for fine-grained control over the style and appearance of generated images. These models enable the manipulation of various attributes such as facial expressions, hair color, and other visual characteristics while maintaining high-quality image synthesis.

StyleGAN and StyleGAN2 achieve this control by disentangling the latent space into style and content components. The style component captures the high-level properties, such as lighting conditions and global styles, while the content component focuses on specific object details. This disentanglement enables independent manipulation of the style and content, providing greater flexibility in generating diverse and realistic images. Additionally, StyleGAN2 introduces a series

of architectural improvements, including adaptive instance normalization (AdaIN) and progressive growing techniques, further enhancing the quality of generated images.

StyleGAN and StyleGAN2, known for their success in creative image generation, have also found applications in the field of speech processing. These models have been employed in tasks such as speech synthesis, voice conversion, and speech modification, expanding their versatility beyond visual domains. Their wide usage in generating photorealistic faces, creating novel artworks, and enabling image-to-image translation tasks has paved the way for exploring their potential in speech-related applications [KKH⁺19, WBK⁺20].

10. MelGAN:

MelGAN is designed for mel-spectrogram synthesis [KKdB⁺19, WBK⁺20]. The objective function of MelGAN can be written as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{m \sim p_{\text{data}}(m)} [\log D(m)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

where m represents real mel-spectrograms, G is the generator network, D is the discriminator network, $p_{\text{data}}(m)$ is the distribution of real mel-spectrograms, and $p_z(z)$ is the distribution of input noise. These equations capture the adversarial training process in the respective GAN variants for speech applications, guiding the generator to produce realistic speech waveforms, converted acoustic features, or mel-spectrograms.

These advanced GAN variants, along with the techniques mentioned earlier, demonstrate the continuous efforts to enhance the training process, stability, and capabilities of GAN models for various applications in image synthesis, translation, and beyond.

GANs have emerged as a groundbreaking technology with applications across various fields. One of their most prominent applications is in computer vision, where they excel at generating highly realistic images of objects, animals, and characters that don't actually exist. This ability has significant implications for creating visual content, from art and design to video game development. In addition to images, GANs have shown promise in generating synthetic videos, which is a more complex task due to the need for coherence and continuity. GANs can create videos that look remarkably close to real-life footage by combining generators and discriminators. GANs also play a crucial role in addressing data scarcity issues in machine learning. They can generate synthetic data that complements real datasets, improving the performance of deep learning models. This is particularly valuable when collecting large amounts of real data is challenging. Another fascinating application is style transfer, where GANs can take the artistic style of one image and apply it to another, resulting in entirely new and visually appealing artwork. This has the potential to revolutionize creative industries and art forms. In the realm of Natural Language Processing (NLP), GANs have been adapted to generate coherent and contextually relevant text. They can synthesize textual content that reads like it was written by humans, opening up possibilities for content generation, storytelling, and more. Musicians also benefit from GAN technology, which can assist in music composition by analyzing existing musical patterns and structures to create original compositions. GANs help musicians explore new styles and ideas, making them valuable tools in the creative process. In the medical domain, GANs are helping improve disease diagnosis by generating synthetic medical images, overcoming limitations caused by limited real-world data. In geoscience and remote sensing, GANs can generate synthetic data that retains the statistical characteristics of real data. [MO14, OOS16, BAC⁺18, PBS17, FGD18]. In summary, GANs are incredibly versatile and have the potential to transform various sectors by creating, enhancing, and safeguarding data. As this technology continues to advance, we can expect even more innovative applications in real-world problem-solving.

Despite significant advancements in the field of GANs, training instability remains a prominent issue. The process of training GANs often requires extensive experimentation and fine-tuning, involving the selection of suitable loss functions, optimization algorithms, and architectural choices. The existing trial-and-error approach to finding the optimal combination of these elements can be time-consuming and inefficient. To tackle this challenge, this thesis explores different variants of GANs that propose novel loss functions. By investigating alternative approaches to formulating the loss function, we aim to enhance the stability and effectiveness of GAN training. This research seeks to provide a deeper understanding of the relationship between loss functions and the overall performance of GANs. Through systematic experimentation and analysis, we aim to identify and evaluate the effectiveness of these novel loss functions in GAN training. By comparing their performance against traditional loss functions, we can assess their potential to overcome the limitations of current approaches. The thesis will delve into a comprehensive examination of these innovative loss functions and their effects on critical aspects of GAN training, including convergence speed, output quality, and the prevalent issue of mode collapse. The objective is to discern how these novel loss functions can significantly enhance the training process of GANs. Furthermore, the study aims to leverage these improvements to enhance speech synthesis applications, ultimately contributing to the advancement of this field.

Chapter 3

Training Generative Adversarial Networks with Weights

3.1 Introduction

A fully data-driven paradigm has emerged during the last years with the advent of GANs [GPM⁺14]. A GAN offers a new methodology for drawing samples from an unknown distribution where only samples from this distribution are available making them one of the strong areas in machine learning/artificial intelligence research. Indicatively, GANs have been successfully utilized in (conditional) image creation [MO14, RMC15, OOS16], generating very realistic samples [KALL17, BAC⁺18], speech signal processing [PBS17, STS18], natural language processing [CLZ⁺17] and astronomy [SZZ⁺17], to name a few.

A GAN is a two-player *zero-sum game* [GPM⁺14, OR94] between a Discriminator and a Generator, both being powerful neural networks. They are simultaneously trained to achieve a *Nash equilibrium*, where the Discriminator cannot distinguish the real and the fake samples while the Generator has learned the unknown distribution. It is well-known that the training procedure of GANs often fails and several specific heuristics and hacks have been devised [SGZ⁺16] along with general-purpose acceleration techniques such as batch normalization [IS15]. Extensions and generalizations stemming from the utilization of a different loss function have been proposed to alleviate the difficulties of training. For instance, f-GAN [NCT16] is a generalization where the f -divergence is used instead of the *Shannon-Jensen divergence* of the original GAN. Another widely-applied extension is *Wasserstein GAN* [ACB17] which has been further improved in [GAA⁺17b]. On the other hand, there are relatively few studies that aim directly to improve the convergence speed of training of an existing GAN.

In this chapter, instead of proposing a new GAN architecture or a new GAN loss function we propose a new training algorithm inspired by the *multiplicative weight update method* (MWUM) [AHK12]. Our goal is to improve the training of the Generator by transferring ideas from Game Theory. Intuitively, the new algorithm puts more weight on fake samples that are more likely to fool the Discriminator and simultaneously reduces the weight of samples that are confidently discriminated as fake. Our contributions are summarized as follows: (i) By adding weights to the training of GANs, we manage to improve the training performance with minor additional computational costs. The new approach is called *Weighted GAN* (WeGAN). (ii) We provide rigorous arguments that the weights of WeGAN locally reduce the loss function more or at least as much as the equally weighted stochastic gradient descent for the Generator. (iii)

The proposed algorithm is not specific to vanilla GAN [GPM⁺14], but it is directly transferable to other extensions such as *conditional GANs*, *Wasserstein GAN* and *f-GAN*. This is an important generalization property of WeGAN.

Before proceeding, it is worth noting that training methods utilizing weights for the Generator have been recently proposed [HJC⁺17, CLZ⁺17, HYSX17]. These methods are essentially equivalent to each other since they assign importance weights to the generated samples in order to obtain a tighter lower bound for their variational formula. However, the importance weights of GAN (IWGAN) cannot be applied to any type of objective function, and additionally, these variational GANs might diverge due to their unboundedness. We implemented IWGAN and presented its performance in the Results section comparing it to our algorithm.

3.2 GAN formulation

MWUM basics. The multiplicative weight update method (MWUM) is a classic algorithmic technique with numerous applications. The main idea behind this method is the existence of a number of “experts” that give some kind of advice to a decision maker. To any “expert” a specific weight is assigned and the initial weights are equal for any “expert”. Then, the decision maker takes the decision according to the advice of the “experts” taking into account the weight of any of them. After this the weights are multiplicatively updated according to the performance of the advice of any individual “expert”; increasing the weights of the “experts” with good performance and decreasing them otherwise and so on. We continue with the description of our algorithm and the connection to this method.

Algorithm

number of iterations k steps Sample $\{x_1, \dots, x_m\}$ from the data distribution $p_{data}(x)$.
 Sample $\{z_1, \dots, z_m\}$ from the input distribution $p_z(z)$.
 Update the Discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m [\log D(x_i) + \log(1 - D(G(z_i)))]. \quad (3.1)$$

Sample $\{z_1, \dots, z_m\}$ from the input distribution $p_z(z)$.
 Compute the unnormalized weights:

$$w_i = \eta^{(1 - D(G(z_i)))}, \quad i = 1, \dots, m.$$

Normalize:

$$w_i = \frac{w_i}{\sum_{j=1}^m w_j}, \quad i = 1, \dots, m.$$

Update the Generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \sum_{i=1}^m w_i \log(1 - D(G(z_i))). \quad (3.2)$$

Figure 3.1: Stochastic gradient ascent/descent training of WeGAN. For a direct comparison with the original GAN, we follow the formulation of [GPM⁺14].

3.3 Weighted GAN algorithm

The proposed algorithm presented in Fig. 3.1 is a modification of the original GAN training algorithm. Inspired by the MWUM, instead of equally-weighted 'fake' samples, we assign a weight to each sample (the "expert" in MWUM) which multiplies the respective gradient term of the Generator. The weighting aims to put more strength to samples that fool the Discriminator and thus are closer to the real data. Indeed, when $D(G(z)) = 0$ and the Discriminator understands that the sample is fake the weight decreases by a factor $\eta \in (0, 1]$. On the other hand, when $D(G(z)) = 1$ the weight remains the same and after the normalization step it has a value greater or equal than the previous one. Notice also that the weights in Algorithm Fig. 3.1 depend only on the current value of the Discriminator while in the standard MWUM the weights are updated cumulatively. This modification was necessary because the input samples are different at each iteration. Indeed, new samples are generated and there is no obvious map between the current samples and the samples from the previous iteration.

3.3.1 Theoretical properties of WeGAN algorithm

A key assumption of our algorithm as well as in other weighting algorithms is that the Discriminator is faithful in the sense that it produces sound decisions for both real and fake samples. Quantitatively, it means that the Discriminator should return on average values above 0.5 when the sample comes from the real distribution and below 0.5 when fake samples are fed to the Discriminator. Next, we show that for a fixed Discriminator, the optimal Generator with weights as in Algorithm 1 achieves a lower or equal loss value than the optimal Generator with equally-weighted samples. Hence, we expect that the inferred Generator is stronger favourably affecting the convergence properties.

Theorem 1. Fix Discriminator D and let $G_{D;w}^*$ and $G_{D;\frac{1}{m}}^*$ be the respective optimum Generator under-weighted and equally-weighted loss function defined by

$$L(G, D; w) = \frac{1}{m} \sum_{i=1}^m \log(D(x_i)) + \sum_{i=1}^m w_i \log(1 - D(G(z_i))). \quad (3.3)$$

Let the weight vector, w , be defined according to Algorithm 1 then

$$L(G_{D;w}^*, D; w) \leq L(G_{D;\frac{1}{m}}^*, D; \frac{1}{m}). \quad (3.4)$$

Proof. By definition, it holds for the optimum Generator that

$$L(G_{D;w}^*, D; w) \leq L(G_{D;\frac{1}{m}}^*, D; w).$$

If we prove that for any G , it holds that $L(G, D; w) \leq L(G, D; \frac{1}{m})$ when w is defined as in Algorithm 1, we get the desired result for $G = G_{D;\frac{1}{m}}^*$. Without loss of generality, we prove the case with $m = 2$ samples. Using a more elaborate but similar argument we can prove it for the general case.

Assuming that $D(G(z_1)) > D(G(z_2))$, it is easy to show that $w_1 > w_2$ and $\log(1 - D(G(z_1))) < \log(1 - D(G(z_2)))$. Next, let n, k be positive integers such that $w_1 = \frac{k}{2n} + \varepsilon_1$ and $w_2 = \frac{2n-k}{2n} + \varepsilon_2$, with ε_i be arbitrarily small constants for $i \in \{1, 2\}$. This is possible due to the fact that the set of rational

numbers is a dense subset of real numbers. Since $w_1 > w_2$ implies $k > n$, then, it holds

$$\begin{aligned}
& w_1 \log(1 - D(G(z_1))) + w_2 \log(1 - D(G(z_2))) + \varepsilon \\
&= \frac{1}{2n} [k \log(1 - D(G(z_1))) + (2n - k) \log(1 - D(G(z_2)))] + \varepsilon \\
&= \frac{1}{2n} [n \log(1 - D(G(z_1))) + n \log(1 - D(G(z_2)))] \\
&\quad + (k - n)(\log(1 - D(G(z_1))) - \log(1 - D(G(z_2))))] + \varepsilon \\
&\leq \frac{1}{2n} [n \log(1 - D(G(z_1))) + n \log(1 - D(G(z_2)))] + \varepsilon \\
&= \frac{1}{2} \log(1 - D(G(z_1))) + \frac{1}{2} \log(1 - D(G(z_2))) + \varepsilon,
\end{aligned} \tag{3.5}$$

for arbitrarily small positive ε . Thus, we prove for $m = 2$ that

$$L(G, D; w) \leq L(G, D; \frac{1}{2}). \tag{3.6}$$

At equilibrium. It is straightforward to show that at the Nash equilibrium, the weights of WeGAN are uniform. Indeed, it holds that $D(x) = 0.5$ for all x and thus

$$w_i = \frac{\eta^{1-D(G(z_i))}}{\sum_{j=1}^m \eta^{1-D(G(z_j))}} = \frac{\eta^{0.5}}{\sum_{j=1}^m \eta^{0.5}} = \frac{1}{m}. \tag{3.7}$$

This observation can serve either as a criterion to stop the training process or as an evaluation metric to assess whether or not the training process converged to an optimum. Monitoring the variance of the weights is the simplest statistic for both tasks.

WeGAN generalization. The proposed algorithm is not exclusive to vanilla GAN and it can be easily extended and applied to any variation of GANs that incorporates a Discriminator mechanism. Therefore, we do not propose just an extension of vanilla GAN but rather a novel training algorithm for general GANs. For instance, we could assign the same formula as in vanilla GAN for the weights for Wasserstein GAN. The presented theoretical analysis still holds for this case.

3.4 Results

For a fair comparison, we evaluate the performance of the various training algorithms without changing the architecture of the networks.

3.4.1 An illustrative example

We present a benchmark example where the new algorithm converges to the data distribution faster than vanilla GAN. The ‘real’ data are drawn from a mixture of 8 normal distributions with each of the 8 components being equally-probable. The mean values are equally-distributed on a circle with radius 3 and covariance matrix I_d . Moreover, both the Generator and Discriminator are fully-connected neural networks with 2 hidden layers and 32 units per layer. The input random variable has a 2-dimensional standard normal while the output of the Discriminator is the sigmoid function.

The upper and middle plots of Fig. 3.2 show the relative improvement of WeGAN with respect to vanilla GAN for various values of η (circle, square & star lines) as a function of the number of epochs. The chosen performance metric is the maximum mean discrepancy (MMD) [GBR⁺12] which measures the closeness between the real data and the generated ones. The relative improvement is higher at the early stage when only $k = 1$ iteration in the training of the Discriminator is performed (upper plot of Fig. 3.2). In contrast, the highest relative improvement occurs closer to the convergence regime when

$k = 5$ iterations in Discriminator's training are performed (middle plot). For comparison purposes, we added IWGAN (dashed line) which also outperforms vanilla GAN but it is slightly worse than WeGAN with $\eta = 0.01$. Moreover, there were cases where IWGAN diverged because it produced a weight with infinite value. In the lower plot of Fig. 3.2, we present the relative performance improvement between the baseline training algorithm for the Wasserstein GAN and the respective weighted variation. We observe that improvements happen but they are less prominent. Additionally, higher values of η result in better performance which is the opposite situation when compared with the vanilla GAN.

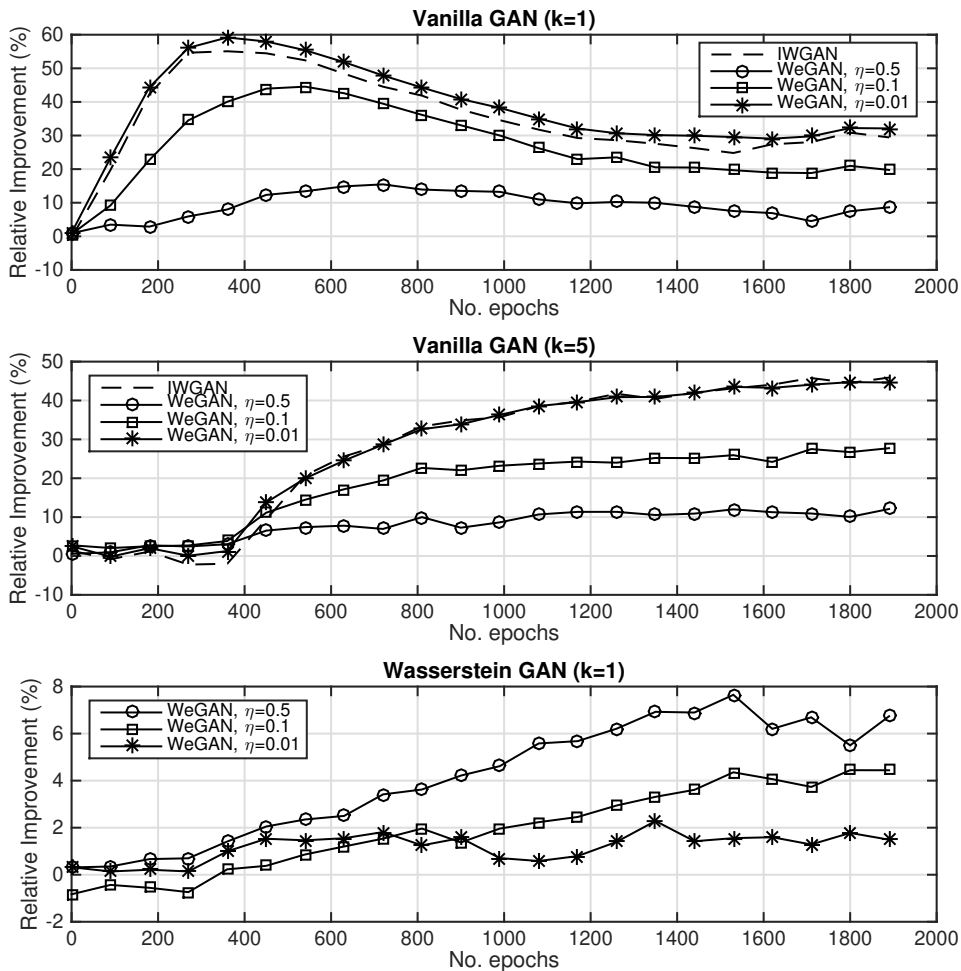


Figure 3.2: Upper & Middle plot: Relative improvement as a function of the epochs in terms of mean MMD with respect to vanilla GAN for a mixture of 8 Gaussians. Lower values for η resulted in improved convergence of WeGAN (lines with circles, squares and stars). Lower plot: Similar to the other plots for Wasserstein GAN. Higher values for η gave faster convergence while IWGAN is not applicable.

3.4.2 MNIST

We extend our experiments on a common benchmark MNIST image database of handwritten digits [LeC98, LBBH98]. In this experiment, a single hidden layer-based fully connected neural network has been used for both Generator and Discriminator with 128 hidden units. Whereas, the input to the Generator is set to 100 dimensional standard normal random variables. Two popular evaluation metrics i.e., Inception Score (IS) [SGZ⁺16] and Fréchet Inception Distance (FID) [HRU⁺17] are used to quan-

tatively assess the performance of GANs. Both metrics assume access to a pre-trained classifier and provide an objective score based on the distribution of the sample that is to be evaluated. Overall relative performance, for IWGAN and various versions of WeGAN with respect to vanilla GAN in terms of IS (upper plot) and FID (lower plot) metrics, are presented in Fig. 3.3. Evidently, WeGAN algorithm outperforms standard vanilla GAN with relative improvement of almost 10% in IS and 30% in FID metrics. Results reveal that WeGAN with $\eta = 0.01$ has the best improvement when compared to other variations of η values which is consistent with the earlier reported results. By examining Fig. 3.3, we also observe that IWGAN achieves higher relative improvement in the early epochs, however, fails to maintain the performance as opposed to WeGAN at $\eta = 0.01$ which procures the best performance.

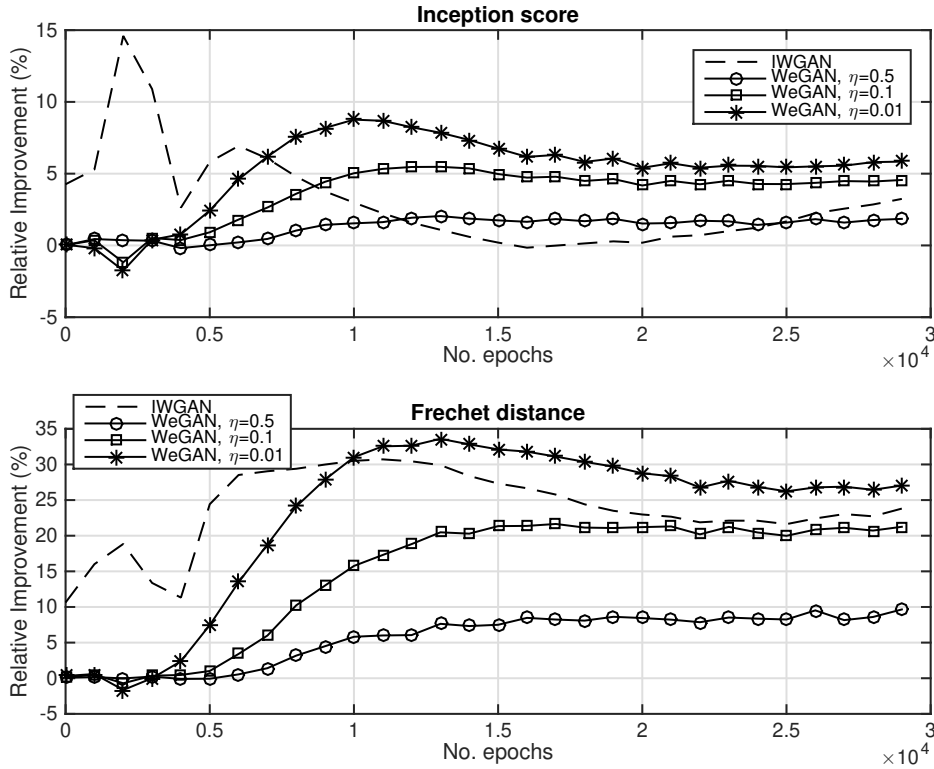


Figure 3.3: Relative improvement as a function of the epochs in terms of IS (upper plot) and FID (lower plot) with respect to vanilla GAN for the MNIST digit dataset. As in the benchmark example, lower values for η result in improved convergence of the WeGAN.

3.4.3 CIFAR

CIFAR-10 is a well-studied dataset of natural images [KH09]. We use this dataset to examine the performance of GANs in generating images. For the Generator, we use a deep convolutional network with a single linear layer followed by 3 convolutional layers. Whereas, the Discriminator has 4 convolutional layers and 1 linear layer at the end. Batch normalization is applied to both networks. The input noise with a dimensionality of 100 is drawn from a uniform distribution. Fig. 3.4 shows IS (upper plot) and FID (lower plot) scores for the CIFAR-10 dataset in terms of relative improvement with reference to vanilla GAN. It can be observed that the proposed WeGAN with $\eta = 0.01$ is preferred over all respective weighted variations in IS score with 5–10% improvement. Whereas, WeGAN with $\eta = 0.5$ & 0.1 both perform comparatively well in FID score. Unfortunately, the performance metrics produce conflicting outcomes making it hard to draw a clear conclusion for this dataset. We also evaluate IWGAN, however,

its performance remains approximately the same against the baseline vanilla GAN.

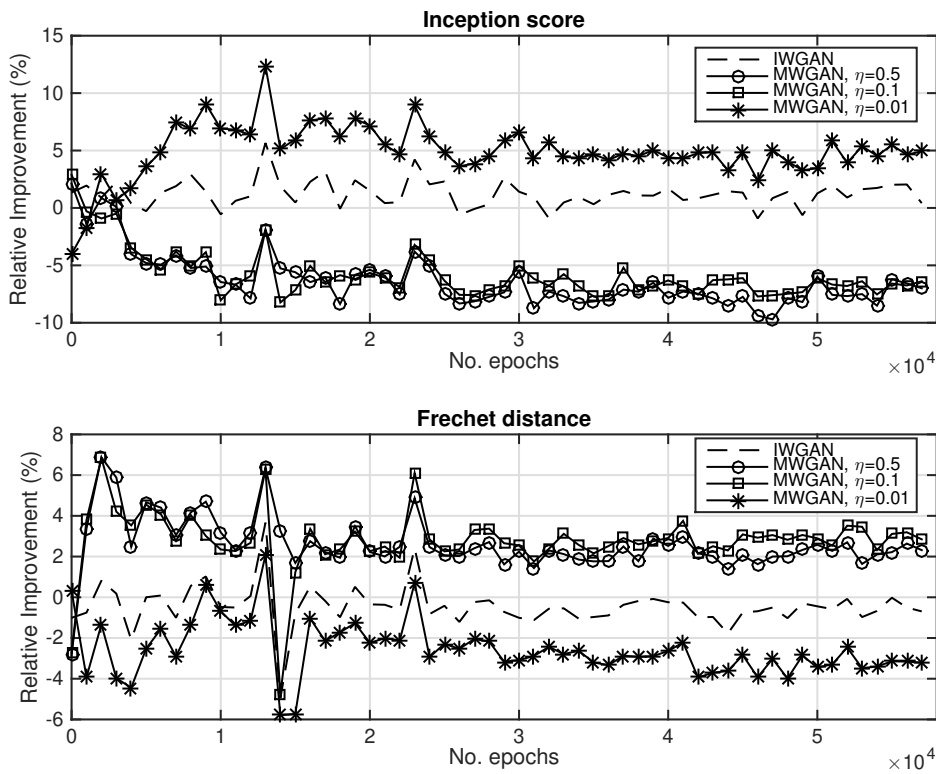


Figure 3.4: Similar to Fig. 3.3 but for the CIFAR-10 dataset. Improvements still happen but they are less prominent while the performance metrics unfortunately produce inconsistent results.

Results indicated that the performance is improved when compared to the baseline training procedure. Moreover, WeGAN is not restricted to a particular type of GAN but it can be easily applied to any type.

3.5 Conclusions

Building upon the principles of the multiplicative weight update method, our research introduces a novel training algorithm for GANs. Our experimentation reveals promising outcomes, with notable improvements in performance compared to the conventional baseline training procedures for GANs. What sets our approach, which we term WeGAN (Weighted GAN), apart is its adaptability, as it is not confined to a specific GAN variant but can be readily applied to various GAN architectures.

Chapter 4

Cumulant GAN

4.1 Introduction

It is well-documented that the training procedure of GANs often fails, and several heuristics have been devised [SGZ⁺16] to alleviate the training train stance, a recurring impediment with GAN training is the oscillatory behaviour of the optimization algorithms due to the fact that the optimal solution is a saddle point of the loss function. Standard optimization algorithms such as stochastic gradient descent (SGD) may fail even for simple loss functions [MPP18, DISZ18].

Since their introduction, GANs have also been described as a tractable approach to minimize a divergence or a distance between the real data distribution and the model distribution. Indeed, the original formulation of GAN [GPM⁺14] can be seen as the minimization of the *Shannon-Jensen divergence*, *f*-GAN [NCT16] is a generalization of vanilla GAN where a variational lower bound for the *f*-divergence is minimized, *Wasserstein GAN* (WGAN) [ACB17] which has been further improved in [GAA⁺17b] aims to minimize the Wasserstein distance showing increased training stability and similarly *Least-Squares GAN* [MLX⁺17] which minimizes a softened version of the Pearson χ^2 -divergence.

However, training might still be unstable and searching for the proper loss function, optimization algorithm, and architecture can involve tedious trial and error. In this chapter, we concentrate on the loss function selection. We propose a novel loss function based on cumulant generating functions with the resulting model referred to as *Cumulant GAN*. A key advantage of cumulants over expectations is that cumulants capture *higher-order* information about the underlying distributions, which often results in more effective learning. Using this property, we rigorously prove that cumulant GAN converges exponentially fast when the gradient descent algorithm is used for the special case with linear generator, linear discriminator and Gaussian distributions. Despite being a simple case, this theoretical result offers a rigorous and valuable differentiation between WGAN, which fails to converge, and the proposed cumulant GAN which demonstrates exponential convergence to the Nash equilibrium, when the same gradient descent algorithm is used on both.

Interestingly, the optimization of cumulant GAN can be described as a *weighting* extension of the standard stochastic gradient descent where the samples that confuse the discriminator the most receive a higher weight, thus, contributing more to the update of the neural network's parameters. Furthermore, by applying a recent variational representation formula [BDK⁺20b], we show that cumulant GAN is capable of interpolating between several GAN formulations, thus, offering a partially-unified mathematical framework. Indeed, the optimization of the proposed loss function is equivalent to the minimization

of divergence for a wide set of cumulant GAN’s hyper-parameter values. It is also worth noting that despite f -GAN’s (partial) unification property [NCT16], cumulant GAN and f -GAN formulations are not equivalent even when they minimize the same divergence and there is a subtle but important difference: the underlying variational representation which is eventually optimized is different. Ours is based on the Donsker-Varadhan representation formula while f -GAN is based on the Legendre transform of f divergence. For KLD, the Donsker-Varadhan formula is tighter than the Legendre duality formula¹. Additionally, our formulation is computationally more manageable because the hyper-parameters of cumulant GAN are of continuous nature while f -GAN requires different f ’s for different divergences.

Our numerical demonstrations aim to provide insights into cumulant GAN’s representational ability and learnability advantages. Experiments on synthetic multi-modal data revealed the differences in the dynamics of learning for different hyper-parameter values of cumulant GAN. Even though the optimal solution is the same, the SGD sequence of the training parameters driven by the chosen hyper-parameters’ values resulted in very different distributional realizations with the two extremes being mode covering and mode selection. Moreover, using cumulant GAN, we were able to recover higher-order statistics even when the discriminator is linear. Finally, we demonstrated increased robustness and improved performance on image generation for both CIFAR10 and ImageNet datasets. Indeed, we performed relative comparisons with WGAN under the standard as well as distressed settings which is a primary reason for training instabilities in GANs and demonstrated that cumulant GAN not only is more stable but also it is better up to 58% in terms of averaged inception score.

The chapter is organized as follows. Section 4.2 introduces the necessary background theory, while Section 4.3 defines cumulant GAN and highlights the derivation of several of its theoretical properties. In Section 4.5, numerical simulations on both synthetic and real datasets are presented, while Section 4.6 concludes the chapter.

4.2 Background

The proposed GAN is a fundamental generalization of WGAN by means of cumulant generating functions. These concepts are briefly discussed in this section.

4.2.1 Wasserstein GAN

WGAN [ACB17, GAA⁺17b] minimizes the Earth-Mover (Wasserstein-1) distance and primarily aims to stabilize the training procedure of GANs. Based on the Kantorovich-Rubinstein duality formula for Wasserstein distance, the loss function of WGAN can be written as

$$\min_G \max_{D \in \mathcal{D}} \mathbb{E}_{p_r}[D(x)] - \mathbb{E}_{p_g}[D(x)], \quad (4.1)$$

where p_g $D(\cdot)$ is the discriminator (called critic in the WGAN setup) while \mathcal{D} is the function space of all 1-Lipschitz continuous functions. In WGAN, Lipschitz continuity is imposed by adding a (soft) regularization term on gradient values called Gradient Penalty (GP). It has been shown that GP regularization produces superior performance relative to weight clipping [GAA⁺17b].

¹Simply by the fact that $x \geq e \log x$; see also [BBR⁺18a].

4.2.2 Cumulant Generating Functions

The cumulant generating function (CGF), also known as the log-moment generating function, is defined for a random variable with pdf $p(x)$ as

$$\Lambda_{f,p}(\beta) = \log \mathbb{E}_p[e^{\beta f(x)}], \quad (4.2)$$

where f is a measurable function with respect to p . The standard CGF is obtained when $f(x) = x$. CGF is a convex function with respect to β and it contains information for all moments of p . CGF also encodes the tail behaviour of distributions and plays a key role in the theory of Large Deviations for the estimation of rare events [DE11]. A power series expansion of the CGF reveals that the lower order statistics dominate when $|\beta| \ll 1$ while all statistics contribute to the CGF when $|\beta| \gg 1$. In statistical mechanics, CGF is the logarithm of the partition function while $-\beta^{-1}\Lambda_{f,p}(-\beta)$ is called the Helmholtz free energy while β is interpreted as the inverse temperature and f as the Hamiltonian [LRS10]. Furthermore, it is straightforward to show that $\Lambda_{f,p}(0) = 0$ as well as $\Lambda'_{f,p}(0) = \mathbb{E}_p[f(x)]$, hence, the following limit for CGF holds

$$\lim_{\beta \rightarrow 0} \beta^{-1} \Lambda_{f,p}(\beta) = \mathbb{E}_p[f(x)]. \quad (4.3)$$

We are now ready to introduce the new GAN.

4.3 Cumulant GAN

4.3.1 Definition

We define a novel GAN training by substituting the expectations in the loss function of WGAN with the respective CGFs. Thus, we propose to optimize the following minimax problem:

$$\begin{aligned} & \min_G \max_{D \in \mathcal{D}} \{(-\beta)^{-1} \Lambda_{D,p_r}(-\beta) - \gamma^{-1} \Lambda_{D,p_g}(\gamma)\} \equiv \\ & \min_G \max_{D \in \mathcal{D}} \underbrace{-\beta^{-1} \log \mathbb{E}_{p_r}[e^{-\beta D(x)}] - \gamma^{-1} \log \mathbb{E}_{p_g}[e^{\gamma D(x)}]}_{=L(\beta,\gamma)}, \end{aligned} \quad (4.4)$$

where the hyper-parameters β and γ are two non-zero real numbers which control the learning dynamics as well as the optimal solution. Since the loss function is the difference of two CGFs, we call $L(\beta, \gamma)$ in (4.4) the *cumulant loss function* and the respective generative model as *Cumulant GAN*. Throughout this thesis, we assume the mild condition that both CGFs are finite for a neighbourhood of $(0, 0)$, therefore, the cumulant loss is well-defined for $|\beta| + |\gamma| < \epsilon$, for some $\epsilon > 0$.

The definition of the loss function is extended on the axes and the origin of the (β, γ) -plane using the limit in (4.3). Hence, the cumulant loss function is defined for all values of β and γ for which the new loss function is finite. It is straightforward to show that WGAN is a special case of cumulant GAN.

Let \mathcal{D} be the set of all 1-Lipschitz continuous functions. Then, cumulant GAN with $(\beta, \gamma) = (0, 0)$ is equivalent to WGAN.

Proof. The proposition is a consequence of the fact that

$$\lim_{\beta, \gamma \rightarrow 0} L(\beta, \gamma) = L(0, 0) = \mathbb{E}_{p_r}[D(x)] - \mathbb{E}_{p_g}[D(x)]. \quad (4.5)$$

□

Next, we rigorously demonstrate that cumulant GAN can be seen as a unified and smooth interpolation between several well-known divergence minimization problems.

4.3.2 A Variational Formula for Rényi Divergence

Similarly to the Donsker-Varadhan variational formula for the Kullback-Leibler divergence that can be obtained from the convex duality formula, we prove a variational formula for the Rényi divergence using the variational representation of exponential integrals also known as risk-sensitive functionals/observables.

Theorem 2. (Variational Representation of Rényi Divergences) *Let p and q be probability distributions. Then, the following formula holds:*

$$\mathcal{R}_\alpha(p||q) = \sup_{f \in C_b} \left\{ \frac{1}{\alpha - 1} \log \mathbb{E}_p[e^{(\alpha-1)f}] - \frac{1}{\alpha} \log \mathbb{E}_q[e^{\alpha f}] \right\}, \quad (4.6)$$

where C_b is the space of all bounded and measurable functions.

Proof. The authors in [ACD15] proved that for all bounded and measurable functions f we have:

$$\frac{1}{\alpha - 1} \log \mathbb{E}_p[e^{(\alpha-1)f}] = \inf_q \left\{ \frac{1}{\alpha} \log \mathbb{E}_q[e^{\alpha f}] + \mathcal{R}_\alpha(p||q) \right\}.$$

Therefore, for any q ,

$$\begin{aligned} \frac{1}{\alpha - 1} \log \mathbb{E}_p[e^{(\alpha-1)f}] &\leq \frac{1}{\alpha} \log \mathbb{E}_q[e^{\alpha f}] + \mathcal{R}_\alpha(p||q) \\ \mathcal{R}_\alpha(p||q) &\geq \frac{1}{\alpha - 1} \log \mathbb{E}_p[e^{(\alpha-1)f}] - \frac{1}{\alpha} \log \mathbb{E}_q[e^{\alpha f}] \end{aligned}$$

For simplicity in the presentation, here we provide the proof based on the assumption that the function $f = \log \frac{dp}{dq}$ is bounded and measurable. Based on the aforementioned assumption we have:

$$\begin{aligned} &\frac{1}{\alpha - 1} \log \mathbb{E}_p[e^{(\alpha-1) \log \frac{dp}{dq}}] - \frac{1}{\alpha} \log \mathbb{E}_q[e^{\alpha \log \frac{dp}{dq}}] \\ &= \frac{1}{\alpha - 1} \log \mathbb{E}_q \left[\left(\frac{dp}{dq} \right)^\alpha \right] - \frac{1}{\alpha} \log \mathbb{E}_q \left[\left(\frac{dp}{dq} \right)^\alpha \right] \\ &= \frac{1}{(\alpha - 1)\alpha} \log \mathbb{E}_q \left[\left(\frac{dp}{dq} \right)^\alpha \right] \\ &= \mathcal{R}_\alpha(p||q) \end{aligned}$$

Therefore, the supremum is attained; hence, we proved (4.6). We refer to [BDK⁺20b] for the complete and general proof. \square

The variational formula for Rényi divergence reduces to the well-known Donsker-Varadhan variational formula for the Kullback-Leibler divergence, when $\alpha \rightarrow 1$, [BDK⁺20b].

4.3.3 Concavity Property of Cumulant GAN

The concavity of the logarithmic function implies that

$$\beta^{-1}\Lambda_{f,p}(\beta) \geq \mathbb{E}_p[f(x)],$$

which is nothing else but Jensen's inequality. If additionally f is bounded, i.e., there is $M > 0$ such that $|f(x)| \leq M$ for all x then a stronger inequality is obtained due to the fact that the domain of the logarithm is also bounded. Indeed, the logarithm is strongly concave with a modulus equal to the infimum value of the domain. In our case, strongly Jensen's inequality deduces that

$$\beta^{-1}\Lambda_{f,p}(\beta) \geq \mathbb{E}_p[f(x)] - \beta e_p^{-\beta M}(f(x))$$

From Jensen's inequality (4.3.3), it is easy to show that for all $\beta, \gamma \neq 0$

$$L(\beta, \gamma) \geq L(0, 0) = \mathbb{E}_{p_r}[D(x)] - \mathbb{E}_{p_g}[D(x)]$$

A stricter inequality called Jensen's inequality for strongly convex/concave functions can be obtained if the function D is bounded. Indeed, if $|D(x)| < M$ for all x then the domain of the logarithmic function is also bounded leading to the stronger inequality

$$L(\beta, \gamma) \geq \mathbb{E}_{p_r}[D(x)] - \mathbb{E}_{p_g}[D(x)] - \beta e_{p_r}^{-\beta M}(D(x)) - \gamma e_{p_g}^{-\gamma M}(D(x)).$$

Generally speaking, strong concavity/convexity is a strengthening of the notion of concavity/convexity, and some properties of strongly concave/convex functions are just "stronger versions" of analogous properties of concave/convex functions.

4.3.4 KLD, Reverse KLD and Rényi Divergence as Special Cases

A major inconvenience of many GAN formulations is their inability to interpret the loss function value and understand the properties of the obtained solution. Even when the stated goal is to minimize a divergence as in the original GAN and the f -GAN, the utilization of training tricks such as non-saturating generators may result in the minimization of something completely different as it was recently observed [Sha20]. In contrast, the proposed cumulant loss function can be interpreted for several choices of its hyper-parameters. Below there is a list of values for β and γ that result in interpretable loss functions. Indeed, several well-known divergences are recovered when the function space for the discriminator is the set of all measurable and bounded functions. In the context that follows, we adopt the convention that a forward divergence, or simply divergence, refers to the utilization of the probability ratio, $\frac{p_r}{p_g}$, whereas a reverse divergence involves the reciprocal ratio.

Let \mathcal{D} be the set of all bounded and measurable functions. Then, the optimization of cumulant loss in (4.4) is equivalent to the minimization of

- a. Kullback-Leibler divergence for $(\beta, \gamma) = (0, 1)$:

$$\min_G \max_{D \in \mathcal{D}} L(0, 1) \equiv \min_G D_{KL}(p_r || p_g).$$

b. Reverse KLD for $(\beta, \gamma) = (1, 0)$:

$$\min_G \max_{D \in \mathcal{D}} L(1, 0) \equiv \min_G D_{KL}(p_g || p_r).$$

c. Rényi divergence for $(\beta, \gamma) = (\alpha, 1 - \alpha)$ with $\alpha \neq 0$ and $\alpha \neq 1$:

$$\min_G \max_{D \in \mathcal{D}} L(\alpha, 1 - \alpha) \equiv \min_G \mathcal{R}_\alpha(p_g || p_r),$$

as well as for $(\beta, \gamma) = (1 - \alpha, \alpha)$ with $\alpha \neq 0$ and $\alpha \neq 1$:

$$\min_G \max_{D \in \mathcal{D}} L(1 - \alpha, \alpha) \equiv \min_G \mathcal{R}_\alpha(p_r || p_g),$$

where $\mathcal{R}_\alpha(p || q)$ is the Rényi divergence defined by

$$\mathcal{R}_\alpha(p || q) = \frac{1}{\alpha(1 - \alpha)} \log \mathbb{E}_q \left[\left(\frac{p}{q} \right)^\alpha \right],$$

when p and q are absolutely continuous with respect to each other and $\alpha > 0^2$.

Proof. a. Using the definition of $L(\beta, \gamma)$, we have:

$$\begin{aligned} \max_{D \in \mathcal{D}} L(0, 1) &= \max_{D \in \mathcal{D}} \left\{ \mathbb{E}_{p_r}[D(x)] - \log \mathbb{E}_{p_g}[e^{D(x)}] \right\} \\ &= D_{KL}(p_r || p_g), \end{aligned} \quad (4.7)$$

where the last equation is the Donsker-Varadhan variational formula [DV83, DE11].

b. Similarly,

$$\begin{aligned} \max_{D \in \mathcal{D}} L(1, 0) &= \max_{D \in \mathcal{D}} \left\{ -\log \mathbb{E}_{p_r}[e^{-D(x)}] - \mathbb{E}_{p_g}[D(x)] \right\} \\ &= \max_{D' = -D \in \mathcal{D}} \left\{ \mathbb{E}_{p_g}[D'(x)] - \log \mathbb{E}_{p_r}[e^{D'(x)}] \right\} \\ &= D_{KL}(p_g || p_r), \end{aligned} \quad (4.8)$$

where we applied again the Donsker-Varadhan variational formula.

c. Generalizing a. and b. we now have:

$$\begin{aligned} &\max_{D \in \mathcal{D}} L(\alpha, 1 - \alpha) \\ &= \max_{D \in \mathcal{D}} \left\{ -\frac{1}{\alpha} \log \mathbb{E}_{p_r}[e^{-\alpha D(x)}] - \frac{1}{1 - \alpha} \log \mathbb{E}_{p_g}[e^{(1 - \alpha)D(x)}] \right\} \\ &= \max_{D' = -D \in \mathcal{D}} \left\{ \frac{1}{\alpha - 1} \log \mathbb{E}_{p_g}[e^{(\alpha - 1)D'(x)}] - \frac{1}{\alpha} \log \mathbb{E}_{p_r}[e^{\alpha D'(x)}] \right\} \\ &= \mathcal{R}_\alpha(p_g || p_r), \end{aligned} \quad (4.9)$$

where the last equation is an extension of the Donsker-Varadhan variational formula to Rényi divergence and was recently proved in ([BDK⁺20b, Theorem 5.4]). For completeness, we provide proof of the Rényi divergence variational representation in Appendix A of Supplementary Materials.

The proof for the case $L(1 - \alpha, \alpha)$ is similar and agrees with the symmetry identity for the Rényi

²The definition is extended for $\alpha < 0$ using the symmetry identity $\mathcal{R}_\alpha(p || q) = \mathcal{R}_{1 - \alpha}(q || p)$.

divergence, $\mathcal{R}_\alpha(p||q) = \mathcal{R}_{1-\alpha}(q||p)$. \square

The Rényi divergence, \mathcal{R}_α , interpolates between KLD ($\alpha \rightarrow 0$) and reverse KLD ($\alpha \rightarrow 1$). Interestingly, there are additional special cases that belong to the family of Rényi divergences. The following corollary states some of them, while Fig. 4.1 depicts schematically the obtained divergences and distances on the (β, γ) -plane.

Under the same assumption as in Theorem 1, the optimization of (4.4) is equivalent to the minimization of

a. Hellinger distance for $(\beta, \gamma) = (\frac{1}{2}, \frac{1}{2})$:

$$\min_G \max_{D \in \mathcal{D}} L\left(\frac{1}{2}, \frac{1}{2}\right) \equiv \min_G -4 \log(1 - D_H^2(p_g, p_r)),$$

where $D_H^2(p, q) = \frac{1}{2} \mathbb{E}_q \left[\left(\left(\frac{p}{q} \right)^{1/2} - 1 \right)^2 \right]$ is the square of the Hellinger distance [Tsy08].

b. χ^2 -divergence for $(\beta, \gamma) = (-1, 2)$:

$$\min_G \max_{D \in \mathcal{D}} L(-1, 2) \equiv \min_G \frac{1}{2} \log(1 + \chi^2(p_r||p_g)),$$

and reverse χ^2 -divergence for $(\beta, \gamma) = (2, -1)$:

$$\min_G \max_{D \in \mathcal{D}} L(2, -1) \equiv \min_G \frac{1}{2} \log(1 + \chi^2(p_g||p_r)),$$

where $\chi^2(p||q) = \mathbb{E}_q \left[\left(\frac{p}{q} - 1 \right)^2 \right]$ is the χ^2 -divergence³ [Tsy08].

c. All-mode covering or worst-case regret in minimum description length principle [G⁺07] for $(\beta, \gamma) = (\infty, -\infty)$:

$$\min_G \lim_{\alpha \rightarrow \infty} \alpha \max_{D \in \mathcal{D}} L(\alpha, 1 - \alpha) \equiv \min_G \log \left(\text{ess sup}_{x \in (p_g)} \frac{p_g(x)}{p_r(x)} \right) \quad (4.10)$$

where ess sup is the essential supremum of a function.

d. Largest-mode selector for $(\beta, \gamma) = (-\infty, \infty)$:

$$\min_G \lim_{\alpha \rightarrow \infty} \alpha \max_{D \in \mathcal{D}} L(1 - \alpha, \alpha) \equiv \min_G \log \left(\text{ess sup}_{x \in (p_r)} \frac{p_r(x)}{p_g(x)} \right). \quad (4.11)$$

All cases a.to d. follow from Theorem 1.c as special instances of Rényi divergence:

$$\begin{aligned} R_{1/2}(p||q) &= -4 \log(1 - D_H^2(p, q)), \\ R_2(p||q) &= \frac{1}{2} \log(1 + \chi^2(p||q)), \\ R_{-1}(p||q) &= R_2(q||p), \\ \lim_{\alpha \rightarrow \infty} \alpha R_\alpha(p||q) &= \log \left(\text{ess sup}_{x \in (q)} \frac{p(x)}{q(x)} \right). \end{aligned} \quad (4.12)$$

³Forward χ^2 -divergence is often called Pearson χ^2 -divergence while the reverse χ^2 -divergence is often called Neyman χ^2 -divergence.

We refer to [M⁺05, Bis06] and the references therein for detailed proofs.

The flexibility of the two hyper-parameters is significant since they offer a simple recipe to remedy some of the most frequent issues of GAN training. For instance, KLD tends to cover all the modes of the real distribution while reverse KLD tends to select a subset of them [M⁺05, Bis06, HLLR⁺16, LT16, Sha20] (see also Fig. 4.3 for a benchmark). Therefore, if mode collapse is observed during training, then, increasing γ with $\beta = 1 - \gamma$ will push the generator towards generating a wider variety of samples. In the other limit, more realistic samples (e.g. less blurry images) with less variability will be generated when β is increased while $\gamma = 1 - \beta$.

Remark. From a practical perspective, the boundedness condition required in the above theoretical formulation can be easily enforced by considering a clipped discriminator with clipping factor M , i.e., $D_M(x) = M \tanh(\frac{D(x)}{M})$. On the other hand, the set of all measurable functions is a very large class of functions and it might be difficult to be represented by a neural network. However, one can approximate measurable functions with continuous functions via Lusin's theorem [Fol99] which states that every finite Lebesgue measurable function is approximated arbitrarily well by a continuous function except on a set of arbitrarily small Lebesgue measure. Therefore, a sufficiently-large neural network can accurately approximate any measurable function.

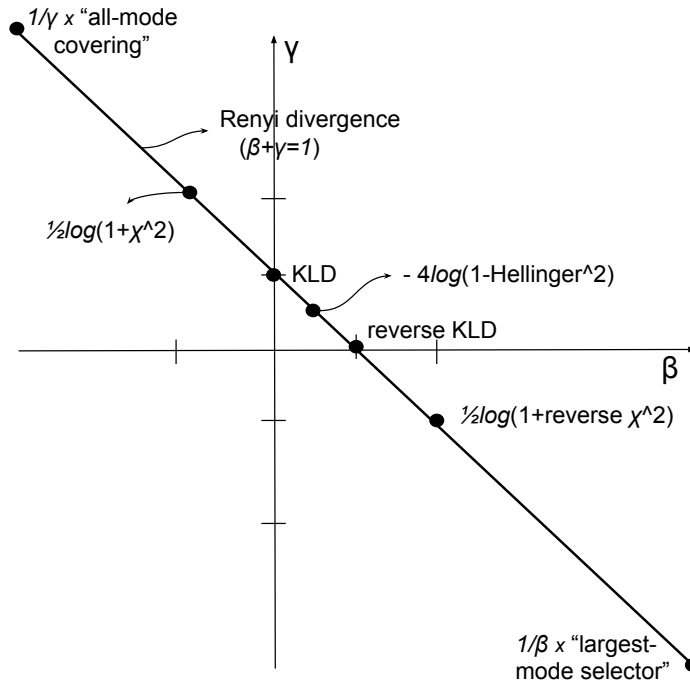


Figure 4.1: Special cases of *cumulant GAN*. Line defined by $\beta + \gamma = 1$ has a point symmetry. The central point, $(\frac{1}{2}, \frac{1}{2})$, corresponds to the Hellinger distance. For each point, $(\alpha, 1 - \alpha)$, there is a symmetric one, i.e., $(1 - \alpha, \alpha)$, which has the same distance from the symmetry point. The respective divergences have reciprocal probability ratios (e.g., KLD & reverse KLD, χ^2 -divergence & reverse χ^2 -divergence, etc.).

4.3.5 Cumulant GAN as a Weighted Version of the SGD Algorithm

The parameter estimation for the cumulant GAN is performed using the SGD algorithm. Algorithm 1 presents the core part of SGD's update steps where we exclude any regularization terms for clarity purposes. Namely, λ is the learning rate. The proposed loss function is not the difference between two

expected values, therefore, the order between differentiation and expectation approximation does matter. We choose to first approximate the expected values with the respective statistical averages as

$$\hat{L}_m(\beta, \gamma) = -\frac{1}{\beta} \log \sum_{i=1}^m e^{-\beta D(x_i)} - \frac{1}{\gamma} \log \sum_{i=1}^m e^{\gamma D(G(z_i))}. \quad (4.13)$$

Then, we apply the differentiation operator which results in a weighted version of SGD as shown in Algorithm 1. Interestingly, several recent papers [BGS16, LT16, HYSX18, HJC⁺18, PPFS19] included a weighting perspective in their optimization approach.

Algorithm 1 Core of SGD Iteration

Input: data batch: $\{x_i\}$, noise batch: $\{z_i\}$
for k steps **do**

$$\eta \leftarrow \eta + \lambda \left(\sum_{i=1}^m w_i^\beta \nabla_\eta D(x_i) - \sum_{i=1}^m w_i^\gamma \nabla_\eta D(G(z_i)) \right) \quad (4.14)$$

end for

$$\theta \leftarrow \theta + \lambda \left(\sum_{i=1}^m w_i^\gamma \nabla_\theta D(G(z_i)) \right) \quad (4.15)$$

The difference between WGAN and cumulant GAN for the update steps is the weights w_i^β and w_i^γ . In WGAN, the weights are constant and equal to $\frac{1}{m}$ while in cumulant GAN they are defined for any $i = 1, \dots, m$ by

$$w_i^\beta = \frac{e^{-\beta D(x_i)}}{\sum_{j=1}^m e^{-\beta D(x_j)}}, \quad \text{and}, \quad w_i^\gamma = \frac{e^{\gamma D(G(z_i))}}{\sum_{j=1}^m e^{\gamma D(G(z_j))}}.$$

The weights redistribute the sample distributions based on the assessment of the current discriminator. Fig. 4.2 demonstrates the change of the weight relative to uniform weights for $\beta, \gamma > 0$. The weights emphasise the real samples associated with the smallest $D(x_i)$ values. Similarly, they place more emphasis on the synthetic samples that give the highest $D(G(z_i))$ values.

The intuition behind the weighting mechanism is that samples that confuse the discriminator, i.e., the samples around the “fuzzy” decision boundary, are more valuable for the training process than samples that are easily distinguished, thus, they should weigh more. Essentially, the discriminator is updated with samples produced by a better generator than the current one, as well as with more challenging real samples. Similarly, the generator is also updated using samples from a generator which is better than the current one. Overall, due to the use of the weights w_i^β, w_i^γ in Algorithm 1, both generator and discriminator updates will be more affected by synthetic samples that are more indistinguishable from the real ones.

Additionally, the update of the discriminator is performed k times more than the generator’s update offering two important advantages. First, more iterations for the discriminator implies that it better distinguishes the real data from the generated ones, making the weighting perspective more valid. Second, it better approximates the optimal discriminator, thus, the theory presented in the previous section becomes more credible in practice.

The Monte Carlo approximation in (4.13) is biased. However, it has been shown that it is consistent [LT16], hence, the error due to the statistical approximation decreases as the size of minibatch increases. Bias correction gradients using moving averages have been utilized in [BBR⁺18a] for the estimation of CGF. However, the modification of the loss function and the lack of an interpretation analogous to the weights w_i^β, w_i^γ are two key reasons to avoid adding any correction terms.

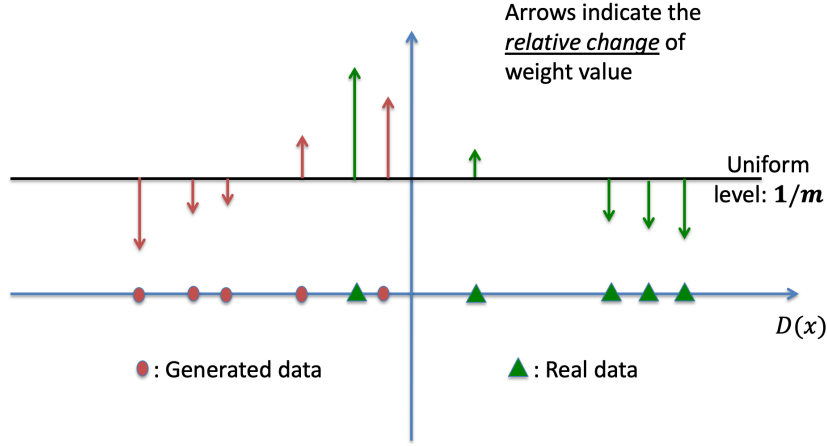


Figure 4.2: Interpretation of *cumulant GAN* as a weighted variation of SGD for $\beta, \gamma > 0$. Both real and generated samples for which the discriminator outputs a value closer to the decision boundary are assigned with larger weights because these are the samples which most probably confuse the discriminator.

4.3.6 Convergence Guarantees for Linear Discriminator

Let \mathcal{D} be the set of all linear functions (i.e., $D(x) = \eta^T x$ with $\eta, x \in \mathbb{R}^d$) and assume that the real data follow a Gaussian distribution with mean value $\mu \in \mathbb{R}^d$ and covariance matrix, I_d . The generator is defined by $G(z) = z + \theta$, where z is a standard d -dimensional Gaussian. The exact loss function for WGAN is

$$\min_{\theta} \max_{\eta} \eta^T (\mu - \theta), \quad (4.16)$$

while the respective exact cumulant loss function from (4.4) is

$$\min_{\theta} \max_{\eta} \eta^T (\mu - \theta) - \frac{\beta + \gamma}{2} \eta^T \eta. \quad (4.17)$$

It is known that the above WGAN loss function oscillates without converging to the optimum if gradient descent is used [MPP18] and more sophisticated algorithms are required to guarantee convergence [DISZ18]. In contrast, the following theorem demonstrates that the proposed cumulant loss function converges if gradient descent is used. Evidently, the use of the cumulant generating function transforms the optimization problem from a concave to a strongly concave problem for η . Next, without loss of generality, we assume $\gamma = 0$.

Theorem 3. *The gradient descent method with learning rate λ converges exponentially fast to the (unique) Nash equilibrium with rate $1 - \lambda\beta$ if $\beta \in (0, \lambda^{-1})$. Mathematically, for the t -th iteration of the gradient descent we have*

$$\|(\theta_t, \eta_t) - (\mu, 0)\|_2^2 \leq c(1 - \lambda\beta)^t, \quad (4.18)$$

where $(\theta^*, \eta^*) = (\mu, 0)$ is the Nash equilibrium while c is a computable positive constant.

Proof. The update step of gradient descent for the cumulant loss is given by

$$\begin{aligned}\eta_{t+1} &= \eta_t + \lambda(\mu - \theta_t - \beta\eta_t), \\ \theta_{t+1} &= \theta_t + \lambda\eta_t.\end{aligned}\tag{4.19}$$

The proof is separated into two sub-cases depending on the value of β . We will consider first the case where $0 < \beta \leq 1$ and then the reciprocal case where $1 \leq \beta < \lambda^{-1}$. This separation is needed because different auxiliary functionals are defined.

Case $0 < \beta \leq 1$: Define the energy function

$$E(\eta, \theta) = \eta^T \eta - \beta \eta^T (\mu - \theta) + (\mu - \theta)^T (\mu - \theta).$$

$E(\eta, \theta)$ is a second order polynomial for η ; it is straightforward to show that if $0 < \beta \leq 1$ then $E(\eta, \theta) \geq 0$ for all η and θ and it is equal to 0 iff $\eta = \eta^* = 0$ and $\theta = \theta^* = \mu$. Additionally, it generally holds that

$$\|(\theta, \eta) - (\mu, 0)\|_2^2 \leq 2E(\eta, \theta),$$

since $2E(\eta, \theta) - \|(\theta, \eta) - (\mu, 0)\|_2^2 = \eta^T \eta - 2\beta \eta^T (\mu - \theta) + (\mu - \theta)^T (\mu - \theta) \geq 0$ for all $0 < \beta \leq 1$.

Next, we show that $E(\eta_t, \theta_t)$ converges exponentially fast to 0. Since, $E(\eta, \theta) = \sum_{i=1}^d \eta_i^2 - \beta \eta_i (\mu_i - \theta_i) + (\mu_i - \theta_i)^2$, we can proceed with $d = 1$ without sacrificing the generality of the proof. After some calculations, we obtain

$$\begin{aligned}E(\eta_{t+1}, \theta_{t+1}) &= (1 - \lambda\beta)E(\eta_t, \theta_t) \\ &\quad - \lambda^2[\eta_t^2 + \beta\eta_t(\mu - \theta_t) + (\mu - \theta_t)^2] \\ &\leq (1 - \lambda\beta)E(\eta_t, \theta_t),\end{aligned}\tag{4.20}$$

since $\eta_t^2 + \beta\eta_t(\mu - \theta_t) + (\mu - \theta_t)^2 \geq 0$ for $\beta \leq 1$. The iterative application of this inequality yields

$$E(\eta_{t+1}, \theta_{t+1}) \leq (1 - \lambda\beta)^{t+1} E(\eta_0, \theta_0).\tag{4.21}$$

Combining the above inequalities we prove (4.18) with $c = 2E(\eta_0, \theta_0)$.

Case $1 \leq \beta < \lambda^{-1}$: Repeat the steps of the first case but this time for the modified energy function

$$\bar{E}(\eta, \theta) = \eta^T \eta - \beta^{-1} \eta^T (\mu - \theta) + (\mu - \theta)^T (\mu - \theta).$$

Here the positive constant is given by $c = 2\bar{E}(\eta_0, \theta_0)$. \square

It is worth noting that the above theorem suggests a learning rate below but close to $\frac{1}{\beta}$. However, a statistical approximation of the exact cumulant loss is used in practice and the optimal learning rate is affected by the minibatch size, thus, it is safer to assign a smaller value. Moreover, the proof is quite broad in the sense that it uses the concept of energy functions (a.k.a. Lyapunov functionals), a tool from the mathematical theory of Dynamical Systems that can be transferred to more general/complex settings as the following remark reveals.

For the same discriminator and generator, the above theorem can be generalized to the case where $x \sim \mathcal{N}(\mu, \Sigma)$ and $z \sim \mathcal{N}(0, \Sigma)$ with Σ being a positive-definite covariance matrix. The proof follows the same steps for the modified energy function $E(\eta, \theta) = \eta^T \Sigma \eta - \beta \eta^T L(\mu - \theta) + (\mu - \theta)^T (\mu - \theta)$, where L is the Cholesky decomposition of Σ (i.e., $\Sigma = LL^T$).

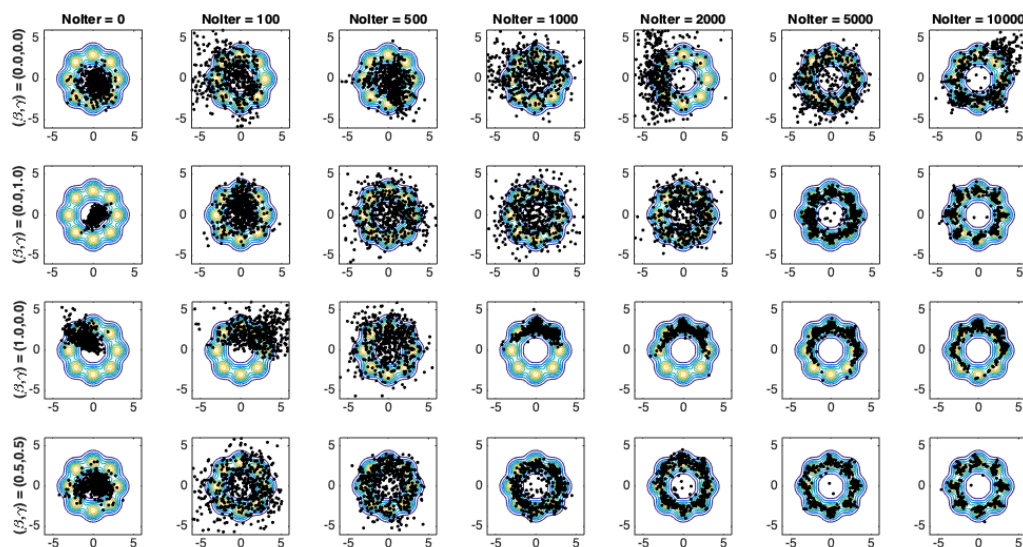


Figure 4.3: Generated samples using the Wasserstein distance using clipping (1st row), KL divergence (2nd row), reverse KLD (3rd row) and Hellinger distance (last row). The boundedness condition is not enforced on this example but it is necessary to be satisfied when the hyper-parameters take negative values.

4.4 Cumulant GAN Implementation

Here, we present the core part of the implementation of *cumulant GAN*.

```
fake_data = Generator()
disc_real = Discriminator(real_data)
disc_fake = Discriminator(fake_data)

def loss_function(disc_real, disc_fake, beta, gamma):

    max_val = tf.reduce_max((-beta) * disc_real)
    disc_cost_real =
        -(1.0/beta)*(tf.log(tf.reduce_mean(tf.exp((-beta)*disc_real-max_val)))+max_val)

    max_val = tf.reduce_max((gamma) * disc_fake)
    disc_cost_fake =
        (1.0/gamma)*(tf.log(tf.reduce_mean(tf.exp(gamma*disc_fake-max_val)))+max_val)
    gen_cost =
        -(1.0/gamma)*(tf.log(tf.reduce_mean(tf.exp(gamma*disc_fake-max_val)))+max_val)

    disc_cost = disc_cost_fake - disc_cost_real

    alpha = tf.random_uniform(
        shape=[64,1],
        minval=0.,maxval=1.)

    differences = fake_data - real_data
    interpolates = real_data + (alpha*differences)
    gradients = tf.gradients(Discriminator(interpolates), [interpolates])[0]
    slopes = tf.sqrt(tf.reduce_sum(tf.square(gradients), reduction_indices=[1]))
    gradient_penalty = tf.reduce_mean((slopes-1.)*2)
    disc_cost += 10*gradient_penalty
```



```

gen_train_op = tf.train.AdamOptimizer(learning_rate=1e-4, beta1=0.,
    beta2=0.9).minimize(gen_cost,
    var_list=lib.params_with_name('Generator'),
    colocate_gradients_with_ops=True)

disc_train_op = tf.train.AdamOptimizer(learning_rate=1e-4, beta1=0.,
    beta2=0.9).minimize(disc_cost,
    var_list=lib.params_with_name('Discriminator.'),
    colocate_gradients_with_ops=True)

return gen_train_op, disc_train_op

```

4.5 Demonstrations

4.5.1 Traversing the (β, γ) -plane: from Mode Covering to Mode Selection

As demonstrated in Section III.B and Fig. 4.1, the optimization of cumulant GAN for the set of bounded and measurable functions and various hyper-parameter values is equivalent to the minimization of a divergence. It is well-known that different divergences result in fundamentally different behaviour of the solution. For instance, KLD minimization tends to produce a distribution that covers all the modes while the reverse KLD tends to produce a distribution that is focused on a subset of the modes [M⁺05, Bis06, HLLR⁺16]. Taking the extreme cases, an all-mode covering is obtained as $\beta \rightarrow -\infty$ while the largest mode selection is observed at the other limit direction.

Our first example aims at highlighting the above characteristics of divergences and additionally to verify that the sub-optimal approximation of the function space of all bounded functions by a family of neural networks does not significantly affect the expected outcomes. Fig. 4.3 presents generated samples for various values of the (β, γ) pair at different stages of the training process as quantified by the number of iterations (denoted by ‘NoIter’). The target distribution is a mixture of 8 equiprobable and equidistant Gaussian random variables. Both discriminator and generator are neural networks with 2 hidden layers with 32 units each and ReLU as the activation function. Input noise for the generator is an 8-dimensional standard Gaussian. In all cases, the discriminator is updated $k = 5$ times followed by an update for the generator.

KLD minimization that corresponds to $(\beta, \gamma) = (0, 1)$ (second row) tends to cover all modes while reverse KLD that corresponds to $(\beta, \gamma) = (1, 0)$ (third row) tends to select a subset of them. This is particularly evident when the number of iterations is between 500 and 2000. Hellinger distance minimization (last row) produces samples with statistics that lie between KLD and reverse KLD minimization while Wasserstein distance minimization (first row) has a less controlled behavior. It is also noteworthy that reverse KLD was not able to fully cover all the modes after 10K iterations. This is not necessarily a drawback since the divergence of choice is primarily an application-specific decision. For instance, the lack of diversity might be sacrificed in image generation for the sake of sharpness of the synthetic images.

Finally, we remark that despite demonstrating a single run, the plots in Fig. 4.3 are not cherry-picked. We have tested several architectures with more or fewer layers, as well as more or fewer units per layer, repeating each run several times, with qualitatively similar results.

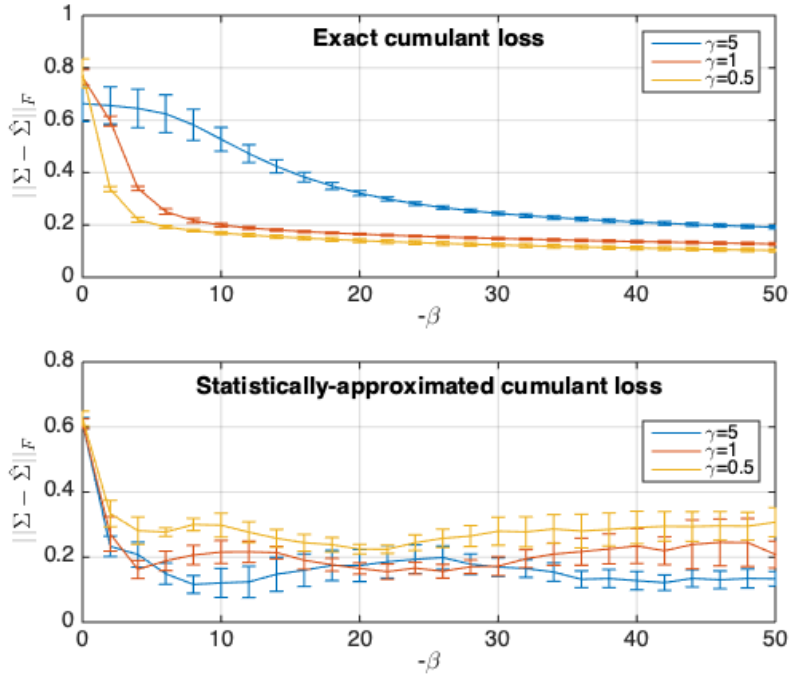


Figure 4.4: Covariance estimation error for the exact cumulant loss function (upper plot) and for the statistically-approximated cumulant loss function (lower plot).

4.5.2 Learning the Covariance Matrix of a Multivariate Gaussian

A CGF can uniquely determine a distribution and contains information on all moments. Therefore, the use of simple discriminators which may fail under the WGAN loss might be sufficient under the cumulant loss in order to successfully train the generator. In this section, we provide an example where the discriminator is a linear function and the target is to learn the second order statistic of a multivariate Gaussian distribution. Thus, the real data, $x \in \mathbb{R}^d$, follow a zero-mean Gaussian with covariance matrix Σ , the discriminator is given by $D(x) = \eta^T x$ while the generator is given by $G(z) = Az$ where A is a $d \times k$ matrix and z is a standard k -dimensional Gaussian. The aim is to obtain a solution, $\hat{\Sigma} = \hat{A}\hat{A}^T$, close to the true covariance matrix.

The loss function of WGAN is $L(0, 0) = \eta^T \mathbb{E}_{p_r}[x] - \eta^T A \mathbb{E}_{p_z}[z] = 0$, therefore it is impossible here to learn the covariance matrix. On the other hand, the cumulant loss reads

$$L(\beta, \gamma) = -\frac{1}{2} \eta^T (\beta \Sigma + \gamma A A^T) \eta$$

allowing the possibility of a (β, γ) pair that makes the Nash equilibrium non-trivially informative regarding the covariance matrix. Indeed, we calculated the best response diagrams for $d = 1$ with fixed positive values of γ and inferred that suitable values are $\beta \ll -1$. Fig. 4.4 presents the average error of the covariance matrix evaluated using the Frobenius norm as a function of β . The covariance is computed using either the above exact loss function (upper plot) or the statistical approximation of the cumulant loss along with stochastic gradient descent (lower plot) for three values of γ . We use 10K samples for the latter case, average over 10 iterations and a different covariance matrix is used at each iteration. The true covariance matrix is rescaled so that its Frobenius norm equals to 1. We observe that the covariance ma-

trix is learned satisfactorily when the exact loss function is used for large negative values of β . When the approximated, yet realistic, loss is used, the error between the true and the estimated covariance matrices increases after a certain value of $-\beta$ because tail statistics (requiring a large amount of samples) start to take control. Overall, the direct conclusion is that cumulant GAN is able to learn higher-order statistics and produce samples with the correct covariance structure despite the fact that a very simple discriminator was deployed.

4.5.3 Image Generation

A series of experiments have been conducted demonstrating the effectiveness of cumulant GAN on standard CIFAR-10 [KH09] and ImageNet [DDS⁺09] datasets. In the experiments, we select pairs of (β, γ) that correspond to well-known divergences in order to highlight their effect on the training process as well as to facilitate connections with existing literature.

Experimental Details

Here, we describe the experimental setup and architectural details for all the experiments presented in the chapter. Three architectures have been used to compare the performance of four GAN losses: Wasserstein, Kullback-Leibler divergence (KLD), reverse KLD and Hellinger distance. The architectures whose successful training we demonstrate are described as follows: (i) convolutional layer for CIFAR-10 data, (ii) residual blocks for CIFAR-10 data (iii) residual blocks for ImageNet data. In the convolutional architecture, batch normalization is applied only for the generator but not for the discriminator. Whereas, we implemented layer normalization in both generator and discriminator. We used Adam as the optimizer with a learning rate of 0.0001. We trained the model for a total of 100,000 iterations on CIFAR-10 and 50,000 iterations on ImageNet, with a mini-batch of 128 and 64, respectively.

CIFAR-10 Convolutional Architecture

Generator				
Layer	Kernel	Output shape	Stride	Activation function
Input z	-	128	-	-
Linear	-	$512 \times 4 \times 4$	-	-
Transposed convolution 1	5×5	$256 \times 8 \times 8$	1	ReLU
Transposed convolution 2	5×5	$128 \times 16 \times 16$	1	ReLU
Transposed convolution 3	5×5	$3 \times 32 \times 32$	1	tanh
Discriminator				
Convolution	5×5	$64 \times 32 \times 32$	4	Leaky ReLU
Linear	-	1	-	-

CIFAR-10 Residual Architecture

Generator				
Layer	Kernel	Output shape	Stride	Activation function
Input z	-	128	-	-
Linear	-	$512 \times 2 \times 2$	-	-
Residual block 1	3×3	$512 \times 4 \times 4$	1	ReLU
Residual block 2	3×3	$256 \times 8 \times 8$	1	ReLU
Residual block 3	3×3	$128 \times 16 \times 16$	1	ReLU
Residual block 4	3×3	$64 \times 32 \times 32$	1	ReLU
Convolution	3×3	$3 \times 32 \times 32$	1	tanh
Discriminator				
Convolution	3×3	$64 \times 32 \times 32$	1	-
Residual block 1	3×3	$128 \times 16 \times 16$	1	ReLU
Residual block 2	3×3	$128 \times 8 \times 8$	1	ReLU
Linear	-	1	-	-

ImageNet Residual Architecture

Generator				
Layer	Kernel	Output shape	Stride	Activation function
Input z	-	128	-	-
Linear	-	$512 \times 4 \times 4$	-	-
Residual block 1	3×3	$512 \times 8 \times 8$	1	ReLU
Residual block 2	3×3	$256 \times 16 \times 16$	1	ReLU
Residual block 3	3×3	$128 \times 32 \times 32$	1	ReLU
Residual block 4	3×3	$64 \times 64 \times 64$	1	ReLU
Convolution	3×3	$3 \times 64 \times 64$	1	tanh
Discriminator				
Convolution	3×3	$64 \times 64 \times 64$	1	-
Residual block 1	3×3	$64 \times 32 \times 32$	1	ReLU
Residual block 2	3×3	$128 \times 16 \times 16$	1	ReLU
Linear	-	1	-	-

CIFAR-10 Dataset

CIFAR-10 is a well-studied dataset of $32 \times 32 \times 3$ RGB color images with 10 classes. We evaluate the quality of the generated images using four different architectures: one with convolutional layers (CNN) and three with residual blocks (resnet). The generator for the CNN consists of one linear layer followed by three convolutional layers while the discriminator is a single convolutional layer followed by one linear layer. The generator for the three resnets consists of four residual blocks while the discriminator consists of two or three residual blocks. We train two versions with three residual blocks for the discriminator but with different channel dimensions and learning rates. In all cases, we deliberately choose a weaker discriminator to challenge the training procedure.

To show that the proposed algorithm is extendable to different architectures, the performance of trained GANs is tested on four different architectures. The first corresponds to convolutional layers while the latter

Table 4.1: Inception scores on CIFAR-10 dataset.

		CIFAR-10			
		Conv layers	Residual blocks	Residual blocks (V1)	Residual blocks (V2)
Loss function	Architecture	Gen: 3 & Dis: 1	Gen: 4 & Dis: 2	Gen: 4 & Dis: 3	Gen: 4 & Dis: 3
	Wasserstein	3.95 \pm 0.21	4.24 \pm 0.19	4.63 \pm 0.22	5.97 \pm 0.03
	KLD	4.39 \pm 0.06	6.70 \pm 0.07	6.53 \pm 0.05	6.44 \pm 0.09
	Reverse KLD	4.01 \pm 0.17	6.67 \pm 0.06	6.60 \pm 0.08	6.39 \pm 0.04
	Hellinger	4.53 \pm 0.04	6.74 \pm 0.06	6.58 \pm 0.08	6.59 \pm 0.10

Table 4.2: Inception scores on Imagenet dataset.

		ImageNet	
		Residual blocks	Residual blocks
Loss function	Architecture	Gen: 4 & Dis: 2	Gen: 4 & Dis: 4
	Wasserstein	6.77 \pm 0.24	7.53 \pm 0.11
	KLD	7.21 \pm 0.22	7.48 \pm 0.09
	Reverse KLD	7.43 \pm 0.18	7.73 \pm 0.11
	Hellinger	7.22 \pm 0.16	7.79 \pm 0.13

two utilize residual networks of different capacities. For the generator of the first architecture, we use one linear layer followed by three convolutional layers, while the discriminator is a single convolutional layer followed by one linear layer. The generator for the second architecture has four residual blocks, while the discriminator consists of two residual blocks. The last two architectures i.e., residual network - 1 (strong discriminator) and residual network - 2 (strong discriminator) comprise of four residual blocks in the generator and three residual blocks in the discriminator but differs in terms of channel dimensions and learning rates.

The plots in Fig. 4.5 present the inception score as a function of the number of iterations for the four architectures. The inception score is a standard metric to evaluate the visual quality of generated image samples [SGZ⁺16]. It assumes access to a pre-trained classifier and provides an objective score based on the distribution of the multiple randomly generated samples that are to be evaluated. We test four different hyper-parameter values that correspond to the minimization of Wasserstein distance (blue line), KLD (red starred line), reverse KLD (green line) and Hellinger distance⁴ (black dashed line). In all cases, the same gradient penalty term is added resulting in optimization over Lipschitz continuous function space. The implementation of cumulant GAN is based on available open-source code⁵. Following the reference code, we train the models with the Adam optimizer and the discriminator’s parameters are updated $k = 5$ times more often than the parameters of the generator.

The averaged inception score results with standard deviation over five runs are reported in Table 4.1 and 4.2. We observe that all hyper-parameter choices for cumulant GAN outperform the baseline WGAN. The relative improvement ranged from 1.5% (reverse KLD) up to 14.9% (Hellinger distance) for the CNN architecture while the relative improvement for the resnet with the weaker discriminator ranged from 57.2% (reverse KLD) up to 58.8% (Hellinger distance) revealing that cumulant GAN takes into

⁴Actually, we minimize $-4 \log(1 - Hel^2)$, see Corollary 1.

⁵https://github.com/igul222/improved_wgan_training

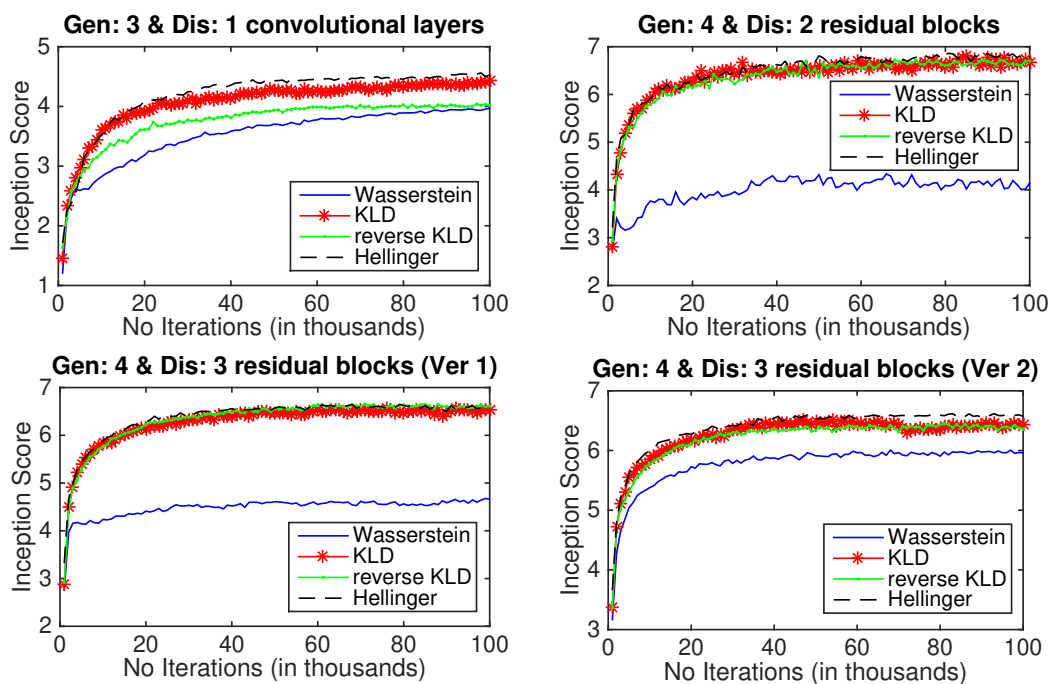


Figure 4.5: Inception score for CIFAR-10 using various hyper-parameters of cumulant GAN and various architectures. In all cases, WGAN has a lower inception score relative to the cumulant GAN with the hyper-parameter corresponding to Hellinger minimization achieving the best overall performance.

consideration all the moments of the discriminator, i.e., all higher-order statistics and not just mean values, leading to better realization of the target data distribution. Cumulant GAN achieves higher inception scores than WGAN for the two versions of resnets with three residual blocks for the discriminator (lower panels in Fig. 4.5), too. All cumulant GAN variations (KLD, reverse KLD and Hellinger) obtain similar results for both versions while the performance of WGAN is significantly affected by the choice of the hyper-parameter values, e.g., learning rate and channel dimension. This discrepancy in the performance highlights the enhanced robustness of cumulant GAN relative to WGAN implying that cumulant GAN might require less tuning in order to enjoy excellent performance. Finally, the samples generated by cumulant GAN also exhibit larger diversity and are visually better (we refer to Appendix E in Supplementary Materials).

Results reveal that KLD minimization is preferred with a relative improvement of 11.1% for convolutional architecture and 57.8% for residual network - weak discriminator over the baseline WGAN-GP. Reverse KLD has 1.5% & 57.2% relative improvement for convolutional architecture and residual network with weaker discriminator, respectively. The superior performance can be seen for the Hellinger distance where convolutional layer has 14.9% and residual - weak discriminator has 58.8% relative improvements. Moreover, significant improvements were found for strong residual discriminator networks. Version 1 of Residual network - strong discriminator achieves 41.03%, 42.5% and 42.1% relative improvements for KLD, reverse KLD and Hellinger minimizations respectively. In order to experiment with the performance when channel size and learning rate are changed, Residual network - strong discriminator version 2 is included where both KLD and reverse KLD approach have around 7% relative improvements. The best performance is attained for Hellinger loss with an inception score of 6.59 having an relative improvement of 10.4% against Wasserstein loss.

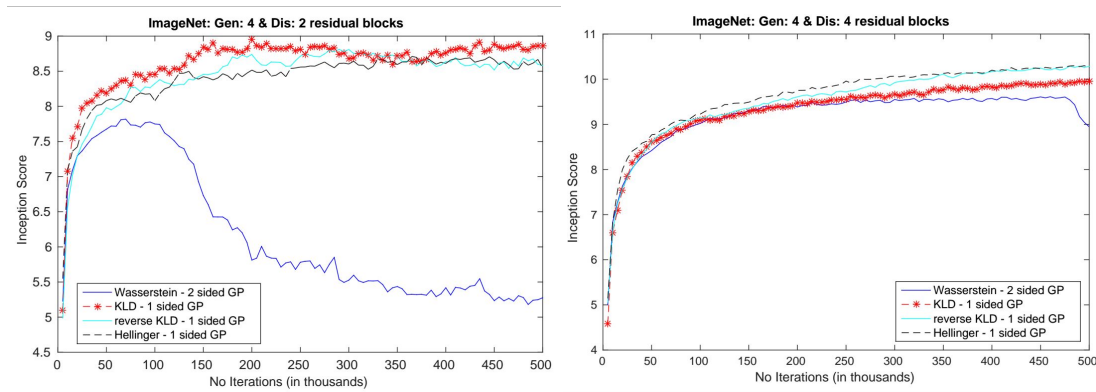


Figure 4.6: Same as Fig. 4.5 but for ImageNet. Cumulant GAN achieves higher inception score relative to WGAN for both weak (left panel) and strong (right panel) discriminator.

ImageNet Dataset

This large scale dataset consists of $64 \times 64 \times 3$ color images with 1000 object classes. The large number of classes is challenging for GAN training due to the tendency to underestimate the entropy in the distribution [SGZ⁺16]. We evaluate the performance on two different architectures which both have a generator with four residual blocks. The difference is in the number of residual blocks for the discriminator where we employ a weak discriminator with two residual blocks and a strong discriminator with four residual blocks. Even though improved performance can be potentially achieved by exploring a wider range of architectures, we choose to test the two architectures, one with a stronger discriminator than the other. Residual network with weak discriminator comprises of four residual blocks for the generator and two residual blocks for the discriminator. As for the other architecture (residual network - strong discriminator) has four residual blocks for both generator and discriminator. Fig. 4.6 presents the performance in terms of inception score both for baseline WGAN and for the variants of cumulant GAN when a weak discriminator (left panel) or a strong discriminator (right panel) is utilized. Improved inception scores are obtained with cumulant GAN for both architectures. The mean inception scores along with the standard deviation over five repetitions are reported in Table 4.1. In terms of relative improvement, cumulant GAN is between 6.5% (KLD) to 9.5% (reverse KLD) better than WGAN for the weak discriminator and a similar trend is observed when the strong discriminator is used. By visual inspection of the generated images (Appendix E in Supplementary Materials), we conclude that all generators learn some basic and contiguous shapes with natural color and texture. Nevertheless, cumulant GAN provides better images with object specifications that are clearly more realistic.

Despite not being exhaustive, the presented examples demonstrated a preference of cumulant GAN over WGAN. In general, GAN optimization has essentially two critical components: the first being the function space where the discriminator lives while the other is the loss function to be optimized. WGAN's breakthrough was the restriction of the function space to Lipschitz continuous functions that resulted in increased stability. However, there is no evidence that the best-performing loss function is the difference of two expectations as in WGAN. The presented examples revealed that there are better and more flexible options for the loss function and the proposed cumulant loss is one of them.

4.6 Conclusions and Future Directions

We proposed the cumulant GAN by establishing a novel loss function based on the CGF of the real and generated distributions. The use of CGFs allows for an inclusive characterization of the distributions' statistics, making it possible to partially remove complexity from the discriminator. The net result is improved and more stable GAN training. Furthermore, cumulant GAN has the capacity to interpolate between a wide range of divergences and distances by simply changing the two hyper-parameter values (β, γ) , and thus offers a flexible and comprehensive mechanism to choose –possibly adaptively– which objective to minimize. Yet, most of the (β, γ) cumulant GAN plane remains *terra incognita* and we plan to promptly explore its properties. Additional research directions include the use of Rényi variational representations for other estimation and inference applications and the application of the proposed cumulant loss function beyond image generation applications.

Generated Images

In this section, generated samples from all the trained models are presented. We remark that all models are trained with GP regularization.

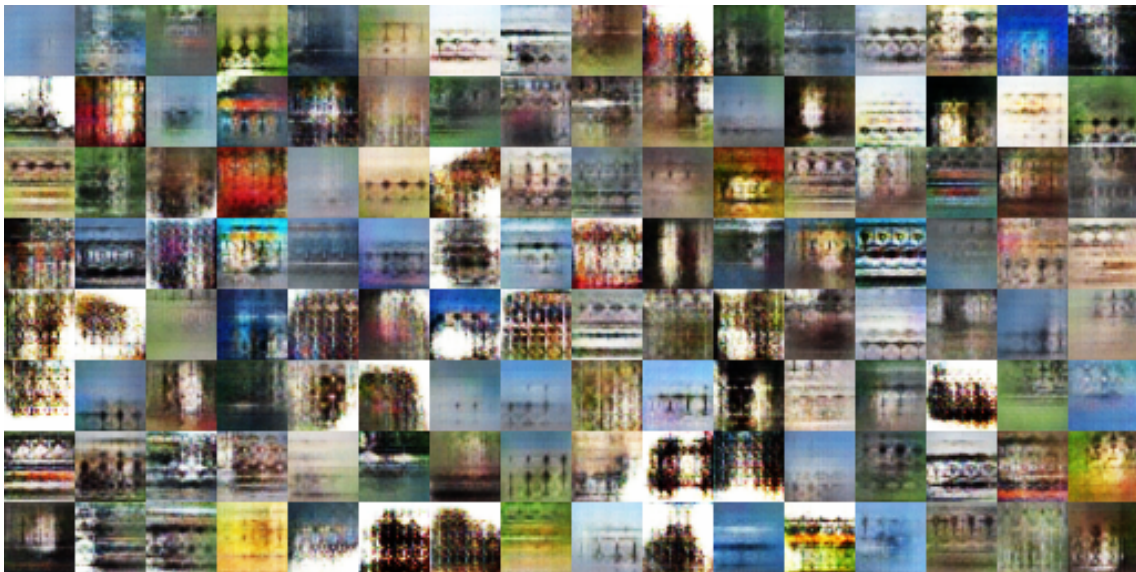


Figure 4.7: WGAN: Samples of CIFAR-10 from generator and discriminator trained with convolutional networks.

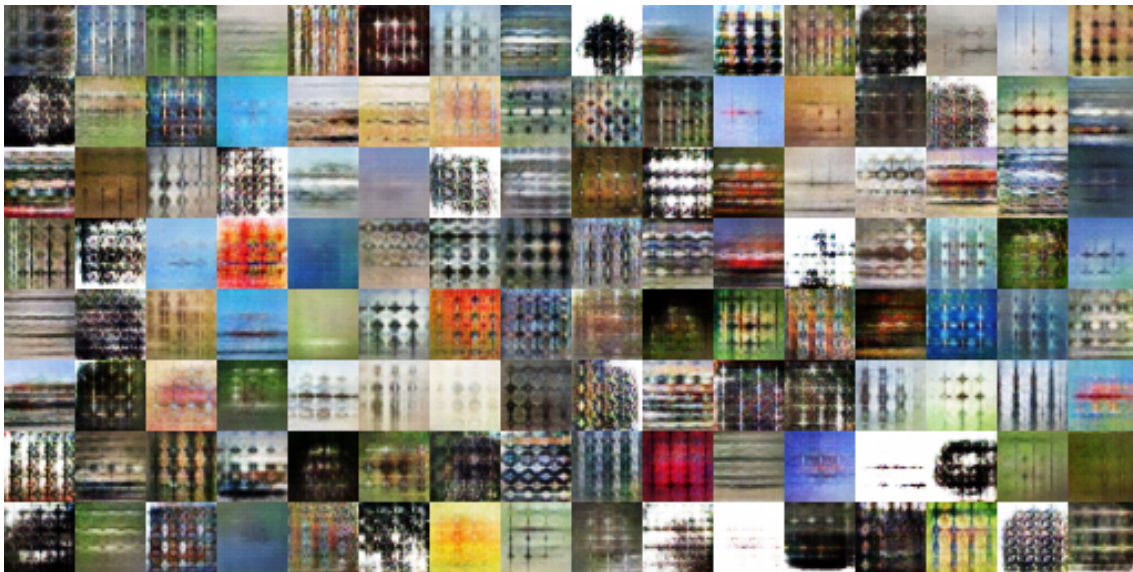


Figure 4.8: KLD: Samples of CIFAR-10 from generator and discriminator trained with convolutional networks.

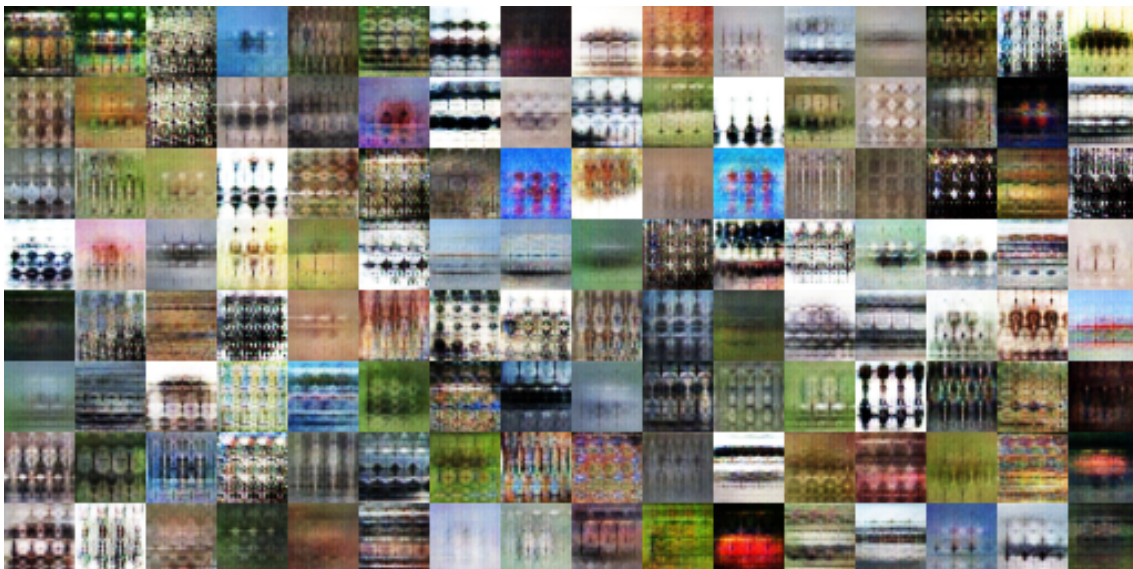


Figure 4.9: Reverse KLD: Samples of CIFAR-10 from generator and discriminator trained with convolutional networks.



Figure 4.10: Hellinger: Samples of CIFAR-10 from generator and discriminator trained with convolutional networks.



Figure 4.11: WGAN: Samples of CIFAR-10 from generator and discriminator trained with residual networks.

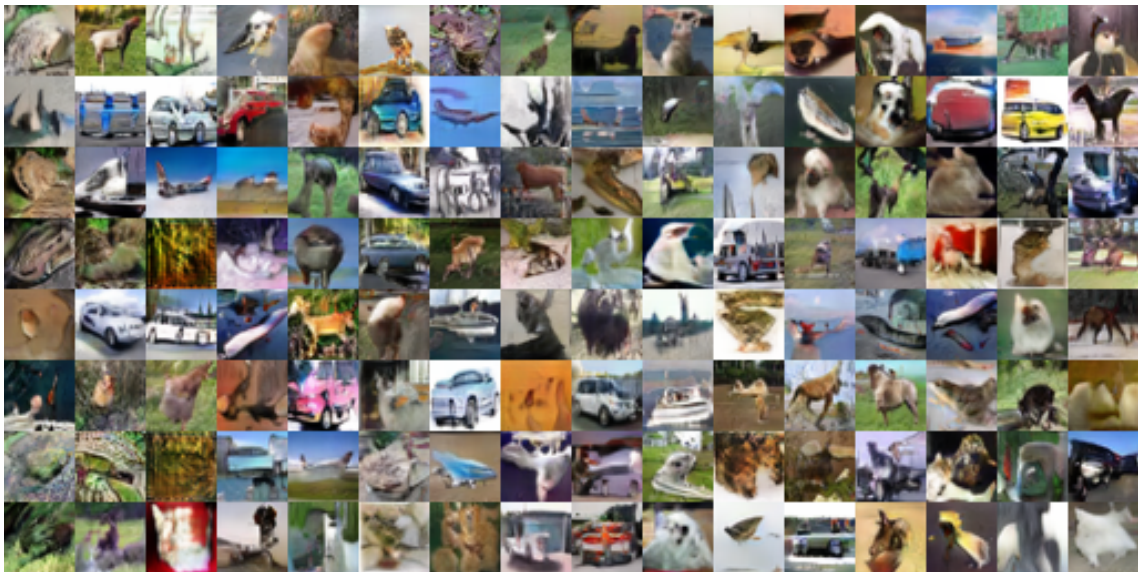


Figure 4.12: KLD: Samples of CIFAR-10 from generator and discriminator trained with residual networks.



Figure 4.13: Reverse KLD: Samples of CIFAR-10 from generator and discriminator trained with residual networks.



Figure 4.14: Hellinger: Samples of CIFAR-10 from generator and discriminator trained with residual networks.

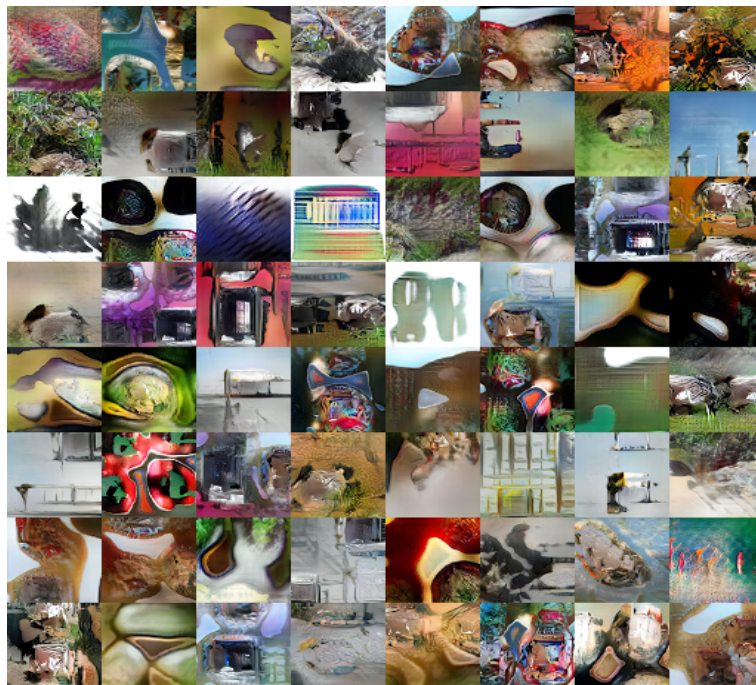


Figure 4.15: WGAN: Samples of ImageNet from generator and discriminator trained with residual networks.

Part II

Disentanglement Learning

Chapter 5

Advancements in Neural-Based Divergence Estimation

5.1 Introduction

Divergences such as Kullback-Leibler (KL) divergence, f -divergences, Hellinger divergence, α divergences and Rényi divergences, which were initially developed in the fields of information theory and statistical physics, are indispensable tools in a growing number of machine learning applications. They have been used in adversarial training of generative models [GPM⁺14, NCT16], in the estimation of generalization errors [EGI21] and in hypothesis testing [BK09], to name a few. Mutual information (MI), in particular, which is defined as the KL divergence between the joint distribution of a pair of variables and their marginals (and can be generalized to divergences other than KL), plays a crucial role in Bayesian networks and (conditional) independence [CGK⁺02], self-supervised learning via contrastive losses [vdOLV18, LKHS20] as well as in representation learning [HFLM⁺19, CDH⁺16].

Classical divergence estimators perform reasonably well for low dimensional cases, however they scale poorly to large, high dimensional datasets which are typically encountered in modern machine learning. The most compelling estimation approach of divergence is optimising a lower variational bound parametrized by neural networks. These lower bounds, which are likelihood-free approximations, are maximized in order to compute the divergence value at the optimizer. Well-known variational representations are the Legendre transformation of an f -divergence [BK06, NWJ10] as well as the Donsker-Varadhan (DV) variational formula [DV83] for KL divergence and its extension to Rényi divergence [BDK⁺20c]. Their tractability stems from their objective functionals, which are computed from expected values and approximated using statistical averages from the available or generated samples.

Despite their scalability and tractability, the estimation of divergence based on variational formulae is a notoriously difficult problem. One challenge stems from the potentially high bias since any approximation for the worst-case scenario requires an exponential number of samples in order to attain the true divergence value [MS20]. Additionally, the statistical variance, which scales exponentially with respect to the divergence's value for certain variational estimators [SE19], is often prohibitively high. Focusing on the elevated MI, there are several further lower bounds [BA03, BBR⁺18b, vdOLV18, POVDO⁺19, GCW⁺21] and a few upper bounds [CHD⁺20, POVDO⁺19] which aim to provide more reliable estimates of MI in the low sample size regime. However, most of these MI estimators are not transferable to the general estimation of divergences and frequently produce instabilities during training which are further magnified

by the small batch and/or sample size.

In this chapter, we propose to reduce a divergence estimator’s variance via an explicit variance penalty (VP) which is added to the objective functional. Our contributions are summarized as follows:

- We present a novel variance reduction penalty for f -divergence and expand it via the delta method to the nonlinear setting, including the DV formula for KL divergence and the variational formula for the Rényi divergences. The proposed VP is able to flexibly trade-off bias and variance.
- We present numerical evidence on synthetic datasets that the proposed approach improves both mean squared error (MSE) and median absolute error (MedAE) in a range of sample sizes and types of divergences. Furthermore, we implemented the proposed VP in several other lower and upper bounds of MI, showing that our variance reduction approach is not restricted to particular variational formulas. Still, it is generic and applicable to most existing variational representations.
- When applied to real datasets, we demonstrate the ability of the proposed approach to reduce the variance of the estimated Rényi divergence, thus enabling the detection of rare biological sub-populations which are otherwise difficult to identify. Interestingly, the baseline estimator is unstable when the order value is above one, but it becomes stable when the VP is added.

5.1.1 Related Work

There are several general-purpose variance reduction techniques in Monte Carlo stochastic sampling, with the most popular approaches being antithetic sampling or more broadly coupling methods, control of variates and importance sampling [RC05, Gla04, Sri13]. These methods have not been explicitly applied to the variational divergence estimation problem. We speculate that either they are not applicable due to the unavailability of analytical probability density formulas or they are inefficient (e.g., the control of the variates approach requires a second estimator and potentially a second parametric model in order to be applied).

Another way to reduce the variance is to restrict the function space to smoother and/or controlled test (or critic) functions, balancing again between bias and variance. For instance, the restriction to Lipschitz continuous functions has the potential to reduce the variance since there exist favourable concentration inequality results for the Lipschitz space [Wai19]. In the GAN literature, Wasserstein GAN [GAA⁺17a], and spectral normalization [MKKY18] impose Lipschitz continuity which resulted in significant gains in terms of training stability. Similarly, the restriction of test functions to an appropriately designed reproducing kernel Hilbert space could reduce the variance [STN20]. Such approaches can be combined with our proposed variance penalties, as our formulation allows for general test-function spaces. However, we do not focus on this point here.

Given the importance of MI, several estimators aim towards improved statistical properties. Lower bounds such as MINE [BBR⁺18b], which uses the DV variational formula with an exponential moving average, NWJ estimator [NWJ10] and BA estimator [BA03] as well as upper bounds such as CLUB [CHD⁺20] still have high variance. InfoNCE [vdOLV18] is one of the few MI estimators that has low variance, but at the cost of either high bias or high computational cost due to the need for many negative samples and thus large batch size. [POVDO⁺19] and [GCW⁺21] aims to clarify the relationships and trade-offs between those variational bounds. A different approach to reducing variance is by appropriately working on the gradients of the objective function [WZH⁺20, WBH⁺21].

Finally, we discuss the approach of truncating the test function inside a bounded region as proposed in [SE19]. The determination of the truncation threshold is quite difficult since it requires an a priori under-

standing of the log-likelihood ratio. Moreover, a high truncation threshold will not affect the estimation since a high threshold implies no real benefit in terms of variance reduction. On the other hand, a low threshold will result in a large bias. Overall, using a high truncation threshold in order to avoid extreme values is a good practice, even though it will have a limited impact on variance reduction.

5.2 Variational Formulas for Rényi and f -Divergences.

While our variance reduction method can be applied to any divergence that possesses a variational formula, here, our focus will be on the Rényi and f -divergences, including the KL divergence. For Rényi divergences, an appropriate objective functional can be constructed from a difference of cumulant generating functions [BDK⁺20c]

$$R_\alpha(Q|P) = \sup_{g \in \mathcal{M}_b(\Omega)} \left\{ \frac{1}{\alpha - 1} \log \mathbb{E}_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha} \log \mathbb{E}_P[e^{\alpha g}] \right\}, \quad \alpha \neq 0, 1. \quad (5.1)$$

Here Q and P are probability distributions on the set Ω , \mathbb{E}_Q and \mathbb{E}_P denote the expectations with respect to Q and P respectively, and $\mathcal{M}_b(\Omega)$ is the space of bounded measurable real-valued functions on Ω . For f divergences, f being a lower semicontinuous convex function with $f(1) = 0$, one has the well-known Legendre transform variational formula [BK06, NWJ10]

$$D_f(Q|P) = \sup_{g \in \mathcal{M}_b(\Omega)} \{ \mathbb{E}_Q[g] - \mathbb{E}_P[f^*(g)] \} \quad (5.2)$$

where $f^*(y) = \sup_{x \in \mathbb{R}} \{ yx - f(x) \}$ is the Legendre transform of f . Here and in the following, the function of g that is being optimized will be called the objective function. 5.2 can be generalized to the (f, Γ) -divergences [BDK⁺20a], where $\Gamma \subset \mathcal{M}_b(\Omega)$ is a restricted test-function space

$$D_f^\Gamma(Q|P) = \sup_{g \in \Gamma} \{ \mathbb{E}_Q[g] - \Lambda_f^P[g] \} \quad (5.3)$$

$$\Lambda_f^P[g] = \inf_{\nu \in \mathbb{R}} \{ \nu + \mathbb{E}_P[f^*(g - \nu)] \} \quad (5.4)$$

In particular, if $f_{\text{KL}}(x) = x \log(x)$ corresponds to the KL divergence then

$$\Lambda_{f_{\text{KL}}}^P[g] = \log(\mathbb{E}_P[\exp(g)]) \equiv \Lambda^P[g] \quad (5.5)$$

is the classical cumulant generating function and (5.3) (with $\Gamma = \mathcal{M}_b(\Omega)$) becomes the Donsker-Varadhan variational formula [DE97, Appendix C.2]

$$D_{\text{KL}}(Q|P) = \sup_{g \in \mathcal{M}_b(\Omega)} \{ \mathbb{E}_Q[g] - \log \mathbb{E}_P[e^g] \} \quad (5.6)$$

For general f , we will often write (5.3) as

$$D_f^\Gamma(Q|P) = \sup_{g \in \Gamma, \nu \in \mathbb{R}} \{\mathbb{E}_Q[g - \nu] - \mathbb{E}_P[f^*(g - \nu)]\} \quad (5.7)$$

and if Γ is closed under the shifts $g \mapsto g - \nu$, $\nu \in \mathbb{R}$ then we can write it simply as

$$D_f^\Gamma(Q|P) = \sup_{g \in \Gamma} \{\mathbb{E}_Q[g] - \mathbb{E}_P[f^*(g)]\}. \quad (5.8)$$

In particular, if $\Gamma = \mathcal{M}_b(\Omega)$ then $D_f^\Gamma = D_f$. The generalizations of Rényi and KL divergence obtained by using a restricted space Γ in place of $\mathcal{M}_b(\Omega)$ in (5.1) or (5.2) will be denoted by R_α^Γ and D_{KL}^Γ , respectively.

5.3 Statistical Estimators and Variance Reduction

Variational representations of divergences are especially useful for creating statistical estimators in a data-driven setting; a naive estimator is obtained by simply replacing expectations with the corresponding statistical averages in any of the equations (5.1), (5.2), (5.3), etc. More formally, the naive estimators can be written as $D_f^\Gamma(Q_n|P_n)$, $R_\alpha^\Gamma(Q_n|P_n)$, etc., where Γ is some parameterized space of functions (e.g., a neural network), Q_n and P_n are the n -sample empirical measures from Q and P respectively (i.e., $\mathbb{E}_{P_n}[g] = \frac{1}{n} \sum_{j=1}^n g(X_j)$ where X_j are i.i.d. samples from P and similarly for \mathbb{E}_{Q_n} ; we also assume that the samples from Q and P are independent of one another), and the divergences are expressed in terms of the variational formulas from Section 5.2. However, in practice, these naive methods often suffer from high variance [SE19, BDK+20c]. We address this via variance-penalized divergences, which are constructed by introducing a variance penalty into the objective functional of the variational representation, e.g.,

$$D_f^\lambda(Q|P) \equiv \sup_{g \in \mathcal{M}_b(\Omega)} \{\mathbb{E}_Q[g] - \mathbb{E}_P[f^*(g)] - \lambda V[g; Q, P]\}, \quad (5.9)$$

where the variance penalty, λV , is proportional to the variance of $\mathbb{E}_{Q_n}[g] - \mathbb{E}_{P_n}[f^*(g)]$ with strength $\lambda > 0$. Using this, we construct the following divergence estimator

$$\sup_{\eta} \{\mathbb{E}_{Q_n}[g_\eta] - \mathbb{E}_{P_n}[f^*(g_\eta)] - \lambda V[g_\eta; Q_n, P_n]\}, \quad (5.10)$$

where g_η is a neural network with parameters η . Similar variance penalties can be derived to other divergences with variational representations.

5.3.1 Variance Penalty

In this subsection, we provide details on the variance penalty for (f, Γ) -divergences, the KL-divergence, and Rényi divergences. The same framework can be repeated to other divergences with a variational representation.

To introduce the variance penalty, first consider the (f, Γ) -divergence representation (5.8). Our goal is to penalize g 's for which $\mathbb{E}_{Q_n}[g]$ or $\mathbb{E}_{P_n}[f^*(g)]$ have large variance. Hence we introduce a penalty term proportional to (Var_Q denotes variance with respect to Q)

$$\text{Var}[\mathbb{E}_{Q_n}[g] + \mathbb{E}_{P_n}[f^*(g)]] = \frac{1}{n} (\text{Var}_Q[g] + \text{Var}_P[f^*(g)]). \quad (5.11)$$

Specifically, for $\lambda > 0$ we define the variance-penalized (f, Γ) -divergence

$$D_f^{\Gamma, \lambda}(Q|P) \equiv \sup_{g \in \Gamma, \nu \in \mathbb{R}} \{ \mathbb{E}_Q[g - \nu] - \mathbb{E}_P[f^*(g - \nu)] - \lambda(\text{Var}_Q[g - \nu] + \text{Var}_P[f^*(g - \nu)]) \}. \quad (5.12)$$

As noted above, if Γ is invariant under constant shifts, then the optimization over ν can be omitted. A similar result to (5.12) can be derived for any objective functional that is a linear combination of expectations, e.g., integral probability metrics [M97, SFG⁺12] such as the Wasserstein metric.

For nonlinear objective functional terms of the generic form $G(\mathbb{E}_P[h(g)])$, such as appear in (5.1) and (5.2), we cannot compute the variance of the corresponding statistical estimator at finite n but we can use the delta method to obtain the asymptotic variance

$$\lim_{n \rightarrow \infty} n \text{Var} [G(\mathbb{E}_{P_n}[h(g)])] = (G'(\mathbb{E}_P[h(g)]))^2 \text{Var}_P[h(g)]. \quad (5.13)$$

Thus, we propose for the nonlinear case to use the above asymptotic variance as a penalty and obtain the following variance-penalized KL and Rényi divergence variational formulas:

$$D_{\text{KL}}^{\Gamma, \lambda}(Q|P) \equiv \sup_{g \in \Gamma} \left\{ \mathbb{E}_Q[g] - \log \mathbb{E}_P[e^g] - \lambda (\text{Var}_Q[g] + \text{Var}_P[e^g]/(\mathbb{E}_P[e^g])^2) \right\}, \quad (5.14)$$

$$R_\alpha^{\Gamma, \lambda}(Q|P) \equiv \sup_{g \in \Gamma} \left\{ \frac{1}{\alpha - 1} \log \mathbb{E}_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha} \log \mathbb{E}_P[e^{\alpha g}] - \lambda \left(\frac{1}{(\alpha - 1)^2} \frac{\text{Var}_Q[e^{(\alpha-1)g}]}{(\mathbb{E}_Q[e^{(\alpha-1)g}])^2} + \frac{1}{\alpha^2} \frac{\text{Var}_P[e^{\alpha g}]}{(\mathbb{E}_P[e^{\alpha g}])^2} \right) \right\}. \quad (5.15)$$

Remark. Both (5.11) and (5.13) suggest that the statistical estimators for the above-penalized divergences should use a variance penalty strength that decays with the sample size $\lambda = \lambda_0/n$, though other forms of n -dependence may be useful in practice.

Though the variance penalty introduces bias, as $\lambda \rightarrow 0$, the penalized divergence converges to the corresponding non-penalized divergence, as made precise by the following theorem.

Theorem 4. Let $\Gamma \subset \mathcal{M}_b(\Omega)$. We have the following convergence results:

$$\lim_{\lambda \rightarrow 0^+} D_{\text{KL}}^{\Gamma, \lambda}(Q|P) = D_{\text{KL}}^\Gamma(Q|P), \quad (5.16)$$

$$\lim_{\lambda \rightarrow 0^+} R_\alpha^{\Gamma, \lambda}(Q|P) = R_\alpha^\Gamma(Q|P), \quad (5.17)$$

and if $f^*(y) < \infty$ for all $y \in \mathbb{R}$ then

$$\lim_{\lambda \rightarrow 0^+} D_f^{\Gamma, \lambda}(Q|P) = D_f^\Gamma(Q|P). \quad (5.18)$$

Moreover, under fairly general assumptions, it holds that

$$\lim_{\lambda \rightarrow \infty} D_{\text{KL}}^{\Gamma, \lambda}(Q|P) = \lim_{\lambda \rightarrow \infty} R_\alpha^{\Gamma, \lambda}(Q|P) = \lim_{\lambda \rightarrow \infty} D_f^{\Gamma, \lambda}(Q|P) = 0. \quad (5.19)$$

Remark. Note that the corresponding statistical estimators, $D_f^{\Gamma, \lambda}(Q_n|P_n)$, etc., have additional bias due to the supremum over g . We present partial results on bias bounds in Appendix D.

The proof of Theorem 4 is given in Appendix B for the zero limit and Theorem 9 for the infinity limit. The same proof techniques can be applied to other divergences with a variational characterization.

Finally, for non-zero λ the penalized divergences (5.12), (5.14), (5.15) retain the divergence property

and are therefore appropriate for quantifying the “distance” between probability distributions:

Theorem 5. *Under fairly general assumptions on f and Γ (see Appendix B for details) and letting $D^{\Gamma,\lambda}$ denote any of $D_f^{\Gamma,\lambda}$, $D_{KL}^{\Gamma,\lambda}$, or $R_\alpha^{\Gamma,\lambda}$ we have $D^{\Gamma,\lambda}(Q|P) \geq 0$ and $D^{\Gamma,\lambda}(Q|P) = 0$ if and only if $Q = P$.*

The proof of Theorem 5 can be found in Appendix B.

5.3.2 Variance-Reduced Divergence Estimation Algorithm

We now propose the following divergence neural estimation (DNE) methods with variance penalty, generalizing equations (5.9)-(5.10).

$$\text{(DNE-VP}_\lambda) \quad \sup_{\eta} \{H[g_\eta; Q_n, P_n] - \lambda V[g_\eta; Q_n, P_n]\}. \quad (5.20)$$

We compare the above method to the non-penalized estimator (i.e., with $\lambda = 0$)

$$\text{(DNE)} \quad \sup_{\eta} H[g_\eta; Q_n, P_n]. \quad (5.21)$$

In the above, the test function space is a neural network $\Gamma = \{g_\eta, \eta \in E\}$ with parameters η and H denotes the objective functional of the divergence, e.g., for the Rényi divergences (5.1)

$$H_\alpha[g; Q, P] = \frac{1}{\alpha - 1} \log \mathbb{E}_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha} \log \mathbb{E}_P[e^{\alpha g}], \quad \alpha \neq 0, 1 \quad (5.22)$$

and for f divergences (5.2)

$$H_f[g; Q, P] = \mathbb{E}_Q[g] - \mathbb{E}_P[f^*(g)]. \quad (5.23)$$

Finally, V is the variance penalty corresponding to the chosen divergence (see Section 5.3.1), e.g., for Rényi divergences

$$V_\alpha[g; Q_n, P_n] = \frac{1}{(\alpha - 1)^2} \frac{\text{Var}_{Q_n}[e^{(\alpha-1)g}]}{(\mathbb{E}_{Q_n}[e^{(\alpha-1)g}])^2} + \frac{1}{\alpha^2} \frac{\text{Var}_{P_n}[e^{\alpha g}]}{(\mathbb{E}_{P_n}[e^{\alpha g}])^2}, \quad \alpha \neq 0, 1 \quad (5.24)$$

and for f divergences

$$V_f[g; Q_n, P_n] = \text{Var}_{Q_n}[g] + \text{Var}_{P_n}[f^*(g)]. \quad (5.25)$$

We solve (5.20) via Adam algorithm [KB14]; a stochastic gradient descent method.

5.4 Proofs of Transformed Variational Formula Identities

Given $-\infty \leq a < 1 < b \leq \infty$ we define $\mathcal{F}_1(a, b)$ to be the set of convex functions $f : (a, b) \rightarrow \mathbb{R}$ with $f(1) = 0$. The convex lower semicontinuous extension of $f \in \mathcal{F}_1(a, b)$ will also be denoted by $f : \mathbb{R} \rightarrow (-\infty, \infty]$. The Legendre transform of f will be denoted $f^*(y) = \sup_{z \in \mathbb{R}} \{yz - f(z)\}$; recall that f^* is continuous on $\{\overline{f^*} < \infty\}$ [Roc70, Theorem 10.1], where \overline{A} denotes the closure of the set A . We let (Ω, \mathcal{M}) be a measurable space and $\mathcal{P}(\Omega)$ be the set of probability measures on (Ω, \mathcal{M}) . For $P \in \mathcal{P}(\Omega)$ we let E_P denote the expectation with respect to P and V_P denote the variance with respect to P . Finally, for $k \in \mathbb{Z}^+$ we let $\mathcal{M}_b(\Omega, \mathbb{R}^k)$ denote the space of bounded measurable functions $g : \Omega \rightarrow \mathbb{R}^k$ (if $k = 1$ we simply write $\mathcal{M}_b(\Omega)$).

Here, we will derive variational characterizations of f -divergences and Rényi divergences that incorporate an additional transformation family.

Lemma 6. *Let $f \in \mathcal{F}_1(a, b)$, $Q, P \in \mathcal{P}(\Omega)$ with $Q \ll P$, and $\Psi : \mathbb{R}^k \rightarrow \mathbb{R}$ be continuous. Then*

$$\sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k)} \{E_Q[\Psi(g)] - E_P[f^*(\Psi(g))]\} = E_P[\sup_{x \in \mathbb{R}^k} \{\Psi(x)dQ/dP - f^*(\Psi(x))\}]. \quad (5.26)$$

Proof. Define $I = \{y : f^*(y) < \infty\}$. If $(\Psi) \subset I^c$ then both sides of (5.26) equal $-\infty$ and we are done. Hence, for the remainder of the proof we suppose there exists x_0 with $\Psi(x_0) \in I$. First compute

$$\begin{aligned} & \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k)} \{E_Q[\Psi(g)] - E_P[f^*(\Psi(g))]\} & (5.27) \\ &= \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k)} E_P[\Psi(g)dQ/dP - f^*(\Psi(g))] \\ &\leq E_P[\sup_{x \in \mathbb{R}^k} \{\Psi(x)dQ/dP - f^*(\Psi(x))\}]. \end{aligned}$$

Now we prove the reverse inequality. Define $A_m = \Psi^{-1}(\bar{I}) \cap [-m, m]^k$ and restrict to m large enough such that $x_0 \in A_m$. We have the bounds

$$-\infty < \Psi(x_0)y - f^*(\Psi(x_0)) \leq \sup_{x \in A_m} \{\Psi(x)y - f^*(\Psi(x))\} \leq \sup_{x \in [-m, m]^k} \{\Psi(x)y - \Psi(x)\} < \infty, \quad (5.28)$$

where here we used the inequality

$$f^*(y) = \sup_{z \in \mathbb{R}} \{yz - f(z)\} \geq y - f(1) = y. \quad (5.29)$$

By continuity of $x \in A_m \mapsto \Psi(x)y - f^*(\Psi(x))$, for every $\epsilon > 0$ there exists a measurable function $x_{m,\epsilon} : \mathbb{R} \rightarrow A_m$ such that

$$|\Psi(x_{m,\epsilon}(y))y - f^*(\Psi(x_{m,\epsilon}(y)))) - \sup_{x \in A_m} \{\Psi(x)y - f^*(\Psi(x))\}| < \epsilon. \quad (5.30)$$

The functions $x_{m,\epsilon}$ are valued in $[-m, m]^k$, hence $x_{m,\epsilon}(dQ/dP) \in \mathcal{M}_b(\Omega, \mathbb{R}^k)$ and we can use (5.30) to compute

$$\begin{aligned} & \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k)} \{E_Q[\Psi(g)] - E_P[f^*(\Psi(g))]\} & (5.31) \\ &\geq E_Q[\Psi(x_{m,\epsilon}(dQ/dP))] - E_P[f^*(\Psi(x_{m,\epsilon}(dQ/dP)))] \\ &= E_P[\Psi(x_{m,\epsilon}(dQ/dP))dQ/dP - f^*(\Psi(x_{m,\epsilon}(dQ/dP)))] \\ &\geq E_P[\sup_{x \in A_m} \{\Psi(x)dQ/dP - f^*(\Psi(x))\}] - \epsilon. \end{aligned}$$

This holds for all $\epsilon > 0$ and all m and so

$$\begin{aligned} & \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k)} \{E_Q[\Psi(g)] - E_P[f^*(\Psi(g))]\} & (5.32) \\ &\geq \liminf_{m \rightarrow \infty} E_P[\sup_{x \in A_m} \{\Psi(x)dQ/dP - f^*(\Psi(x))\}]. \end{aligned}$$

5.28 implies

$$\sup_{x \in A_m} \{\Psi(x)dQ/dP - f^*(\Psi(x))\} \geq \Psi(x_0)dQ/dP - f^*(\Psi(x_0)) \in L^1(P) \quad (5.33)$$

for all m , therefore we can apply Fatou's lemma to (5.32) to compute

$$\begin{aligned} & \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k)} \{E_Q[\Psi(g)] - E_P[f^*(\Psi(g))]\} \\ & \geq \liminf_{m \rightarrow \infty} E_P[\sup_{x \in A_m} \{\Psi(x)dQ/dP - f^*(\Psi(x))\}] \\ & \geq E_P[\liminf_{m \rightarrow \infty} \sup_{x \in A_m} \{\Psi(x)dQ/dP - f^*(\Psi(x))\}] \\ & = E_P[\sup_m \sup_{x \in A_m} \{\Psi(x)dQ/dP - f^*(\Psi(x))\}] \\ & = E_P[\sup_{x \in \Psi^{-1}(\bar{I})} \{\Psi(x)dQ/dP - f^*(\Psi(x))\}] \\ & = E_P[\sup_{x \in \mathbb{R}^k} \{\Psi(x)dQ/dP - f^*(\Psi(x))\}]. \end{aligned} \quad (5.34)$$

This completes the proof. \square

Theorem 7. Let $f \in \mathcal{F}_1(a, b)$, $Q, P \in \mathcal{P}(\Omega)$, and for every $\eta \in E$ suppose we have a continuous map $\Psi_\eta : \mathbb{R}^k \rightarrow \mathbb{R}$. If there exists $\eta_0 \in E$ with $(\Psi_{\eta_0}) = \mathbb{R}$ then

$$D_f(Q|P) = \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \{E_Q[\Psi_\eta(g)] - E_P[f^*(\Psi_\eta(g))]\}. \quad (5.35)$$

Proof. First suppose $Q \ll P$: We have $\Psi_\eta(g) \in \mathcal{M}_b(\Omega)$ for all η and g , hence (5.2) implies

$$\sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \{E_Q[\Psi_\eta(g)] - E_P[f^*(\Psi_\eta(g))]\} \leq D_f(Q|P). \quad (5.36)$$

On the other hand, using Lemma 6 and the assumption $(\Psi_{\eta_0}) = \mathbb{R}$ we can compute

$$\begin{aligned} & \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \{E_Q[\Psi_\eta(g)] - E_P[f^*(\Psi_\eta(g))]\} \\ & \geq \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k)} \{E_Q[\Psi_{\eta_0}(g)] - E_P[f^*(\Psi_{\eta_0}(g))]\} \\ & = E_P[\sup_{x \in \mathbb{R}^k} \{\Psi_{\eta_0}(x)dQ/dP - f^*(\Psi_{\eta_0}(x))\}] \\ & = E_P[\sup_{z \in (\Psi_{\eta_0})} \{zdQ/dP - f^*(z)\}] \\ & = E_P[\sup_{z \in \mathbb{R}} \{zdQ/dP - f^*(z)\}] \\ & = E_P[f(dQ/dP)] = D_f(Q|P). \end{aligned} \quad (5.38)$$

This proves the claim when $Q \ll P$.

Now suppose $Q \not\ll P$: In this case there exists a measurable set A with $P(A) = 0$ and $Q(A) > 0$. Take a sequence $x_n \in \mathbb{R}^k$ with $\Psi_{\eta_0}(x_n) \rightarrow \infty$, $y_0 \in \mathbb{R}$ with $f^*(y_0) < \infty$, and $x_0 \in \mathbb{R}^k$ with $\Psi_{\eta_0}(x_0) =$

y_0 . Define $g_n = x_n 1_A + x_0 1_{A^c} \in \mathcal{M}_b(\Omega, \mathbb{R}^k)$. Therefore

$$\sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \{E_Q[\Psi_\eta(g)] - E_P[f^*(\Psi_\eta(g))]\} \quad (5.39)$$

$$\begin{aligned} &\geq E_Q[\Psi_{\eta_0}(g_n)] - E_P[f^*(\Psi_{\eta_0}(g_n))] \\ &= \Psi_{\eta_0}(x_n)Q(A) + \Psi_{\eta_0}(x_0)Q(A^c) - f^*(y_0)P(A^c) \rightarrow \infty \end{aligned} \quad (5.40)$$

as $n \rightarrow \infty$. Hence we can conclude

$$\sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \{E_Q[\Psi_\eta(g)] - E_P[f^*(\Psi_\eta(g))]\} = \infty = D_f(Q|P). \quad (5.41)$$

This completes the proof. \square

Under stronger assumptions on f we can weaken the assumption that $(\Psi) = \mathbb{R}$ and still prove the transformed variational formula.

Theorem 8. Let $f \in \mathcal{F}_1(a, b)$ be C^1 with f' strictly increasing, $Q, P \in \mathcal{P}(\Omega)$ with $Q \ll P$, and $\Psi_\eta : \mathbb{R}^k \rightarrow \mathbb{R}$ be continuous for all $\eta \in E$. Suppose $a \leq dQ/dP \leq b$ and if the value a (resp. b) is achieved then $f'(a) \equiv \lim_{x \searrow a} f'(x)$ (resp. $f'(b) \equiv \lim_{x \nearrow b} f'(x)$) exists and is finite. Finally, define $I = \{f^* < \infty\}$ and suppose there exists $\eta_0 \in E$ with $(f'(dQ/dP)) \subset \overline{(\Psi_{\eta_0}) \cap I}$.

Then

$$D_f(Q|P) = \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \{E_Q[\Psi_\eta(g)] - E_P[f^*(\Psi_\eta(g))]\}. \quad (5.42)$$

Proof. Using the definition of the Legendre transform it is a straightforward calculus exercise to show that

$$f(dQ/dP) = \frac{dQ}{dP} f'(dQ/dP) - f^*(f'(dQ/dP)). \quad (5.43)$$

Using the continuity of f^* on \bar{I} we have

$$\begin{aligned} \sup_{x \in \mathbb{R}^k} \{\Psi_{\eta_0}(x)dQ/dP - f^*(\Psi_{\eta_0}(x))\} &= \sup_{y \in (\Psi_{\eta_0}) \cap I} \{ydQ/dP - f^*(y)\} \\ &= \sup_{y \in \overline{(\Psi_{\eta_0}) \cap I}} \{ydQ/dP - f^*(y)\} \\ &\leq (f^*)^*(dQ/dP) = f(dQ/dP). \end{aligned} \quad (5.44)$$

However, (5.43) together with the assumption $(f'(dQ/dP)) \subset \overline{(\Psi_{\eta_0}) \cap I}$ implies

$$\sup_{y \in \overline{(\Psi_{\eta_0}) \cap I}} \{ydQ/dP - f^*(y)\} \geq f(dQ/dP). \quad (5.45)$$

Therefore we have equality, with the maximum occurring at $f'(dQ/dP)$, and

$$f(dQ/dP) = \sup_{x \in \mathbb{R}^k} \{\Psi_{\eta_0}(x)dQ/dP - f^*(\Psi_{\eta_0}(x))\}. \quad (5.46)$$

Combining this with Lemma 6 we see that

$$\begin{aligned} D_f(Q|P) &= \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k)} \{E_Q[\Psi_{\eta_0}(g)] - E_P[f^*(\Psi_{\eta_0}(g))]\} \\ &\leq \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \{E_Q[\Psi_\eta(g)] - E_P[f^*(\Psi_\eta(g))]\}. \end{aligned} \quad (5.47)$$

The reverse inequality follows from (5.2) and the fact that $\Psi_\eta(g) \in \mathcal{M}_b(\Omega)$ for all η and g . \square

Remark. In particular, if one has the a priori bounds $a < m \leq dQ/dP \leq M < b$ then Theorem 8 justifies the use of a truncated function space, i.e., using Ψ_{η_0} with $[f'(m), f'(M)] \subset \overline{(\Psi_{\eta_0})} \cap I$. Similar truncated discriminators were previously used for variance reduction in 2019arXiv191006222S. In the examples shown in Figure 5.1 above, we use transformations of the form $\Psi_\theta(x) = h(x + T_\theta(x))$ where h is a truncation and T has the growth bound $\limsup_{x \rightarrow \infty} |T_\theta(x)/x| < 1$, thus ensuring $(x + T_\theta(x)) = \mathbb{R}$.

In the next corollary, we apply Theorem 8 to the α -divergences, i.e., the f -divergences obtained by using

$$f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha - 1)}, \quad \alpha > 0, \alpha \neq 1, \quad (5.48)$$

which have the Legendre transforms

$$f_\alpha^*(y) = \begin{cases} y^{\alpha/(\alpha-1)} \alpha^{-1} (\alpha-1)^{\alpha/(\alpha-1)} \mathbf{1}_{y \geq 0} + \frac{1}{\alpha(\alpha-1)}, & \alpha > 1, \\ \infty \mathbf{1}_{y \geq 0} + \left(|y|^{-\alpha/(1-\alpha)} \alpha^{-1} (1-\alpha)^{-\alpha/(1-\alpha)} - \frac{1}{\alpha(1-\alpha)} \right) \mathbf{1}_{y < 0}, & \alpha \in (0, 1). \end{cases} \quad (5.49)$$

Corollary 9. Let $Q, P \in \mathcal{P}(\Omega)$ with $Q \ll P$ and $\Psi_\eta : \mathbb{R}^k \rightarrow \mathbb{R}$, $\eta \in E$ be continuous. Let $\alpha > 0$, $\alpha \neq 1$.

1. If $\alpha > 1$ suppose that there exists $\eta_0 \in E$ with $(0, \infty) \subset (\Psi_{\eta_0})$.
2. If $\alpha \in (0, 1)$ suppose that $dQ/dP > 0$ and there exists $\eta_0 \in E$ with $(-\infty, 0) \subset (\Psi_{\eta_0})$.

Then

$$D_{f_\alpha}(Q|P) = \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \{E_Q[\Psi_\eta(g)] - E_P[f_\alpha^*(\Psi_\eta(g))]\}. \quad (5.50)$$

We can use Corollary 9 to derive a variational formula for the Rényi divergences that includes a transformation family. Rényi divergences for $\alpha < 0$ can be expressed in terms of R_α for $\alpha > 1$ van2014renyi and so we focus on the cases $\alpha > 1$ and $\alpha \in (0, 1)$.

Theorem 10. Let $\alpha > 0$, $\alpha \neq 1$, $Q, P \in \mathcal{P}(\Omega)$, and $\Psi_\eta : \mathbb{R}^k \rightarrow \mathbb{R}$, $\eta \in E$ be continuous. Suppose there exists $\eta_0 \in E$ with $(\Psi_{\eta_0}) = \mathbb{R}$. If $\alpha \in (0, 1)$ suppose also that $Q \ll P$ and $dQ/dP > 0$. Then

$$R_\alpha(Q|P) = \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \left\{ \frac{1}{\alpha - 1} \log(E_Q[e^{(\alpha-1)\Psi_\eta(g)}]) - \frac{1}{\alpha} \log(E_P[e^{\alpha\Psi_\eta(g)}]) \right\}. \quad (5.51)$$

Proof. First suppose $Q \ll P$. Applying Corollary 9 to $\Psi_{\eta,c}(x) = \pm c \exp((\alpha - 1)\Psi_\eta(x))$ (positive sign for $\alpha > 1$, negative sign for $\alpha \in (0, 1)$) with $c > 0$ and then evaluating the supremum over $c > 0$ (as was

done in the Appendix to [BKP20]) we obtain

$$\begin{aligned}
& D_{f_\alpha}(Q|P) \tag{5.52} \\
&= \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \sup_{c > 0} \{E_Q[\pm c e^{(\alpha-1)\Psi_\eta(g)}] - E_P[f_\alpha^*(\pm c e^{(\alpha-1)\Psi_\eta(g)})]\} \\
&= \begin{cases} \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \left\{ \frac{1}{\alpha(\alpha-1)} E_Q[e^{(\alpha-1)\Psi_\eta(g)}]^\alpha E_P[e^{\alpha\Psi_\eta(g)}]^{-(\alpha-1)} - \frac{1}{\alpha(\alpha-1)} \right\}, & \alpha > 1, \\ \sup_{g \in \mathcal{M}_b(\Omega, \mathbb{R}^k), \eta \in E} \left\{ \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha(1-\alpha)} E_Q[e^{(\alpha-1)\Psi_\eta(g)}]^\alpha E_P[e^{\alpha\Psi_\eta(g)}]^{1-\alpha} \right\}, & \alpha \in (0, 1) \end{cases}.
\end{aligned}$$

Using the relationship between α -divergences and Rényi divergences, we arrive at the claimed result.

Finally, (5.51) for $\alpha > 1$ and $Q \not\ll P$ is proven using the same technique as in Theorem 7. \square

5.5 Bias Bounds

In this section we derive bounds on the bias of Rényi and f -divergence variational formula estimators. We again let Q_n, P_n be n -sample empirical measures from Q and P respectively. The key lemma is the following simple results regarding the expectation of a supremum or infimum.

Lemma 11. *Given an objective functional, $H : \mathcal{M}_b(\Omega) \times \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \rightarrow \overline{\mathbb{R}}$, and a test function space $\Gamma \subset \mathcal{M}_b(\Omega)$ we have*

$$\begin{aligned}
\mathbb{E}[\sup_{g \in \Gamma} H[g; Q_n, P_n]] &\geq \sup_{g \in \Gamma} \mathbb{E}[H[g; Q_n, P_n]], \tag{5.53} \\
\mathbb{E}[\inf_{g \in \Gamma} H[g; Q_n, P_n]] &\leq \inf_{g \in \Gamma} \mathbb{E}[H[g; Q_n, P_n]].
\end{aligned}$$

The next lemma provides a bound on the bias of statistical estimators of Λ_f^P from (5.4).

Lemma 12. *Let f be convex with $f(1) = 0$, $P \in \mathcal{P}(\Omega)$, and P_n be n -sample empirical measures from Q and P respectively. Then for all $g \in \mathcal{M}_b(\Omega)$ the generalized cumulant generating function satisfies*

$$\mathbb{E}[\Lambda_f^{P_n}[g]] \leq \Lambda_f^P[g]. \tag{5.54}$$

Proof. Using (5.4) we can compute

$$\begin{aligned}
\mathbb{E}[\Lambda_f^{P_n}[g]] &= \mathbb{E}\left[\inf_{\nu \in \mathbb{R}} \{\nu + E_{P_n}[f^*(g - \nu)]\}\right] \tag{5.55} \\
&\leq \inf_{\nu \in \mathbb{R}} \mathbb{E}[\nu + E_{P_n}[f^*(g - \nu)]] \\
&= \inf_{\nu \in \mathbb{R}} \{\nu + E_P[f^*(g - \nu)]\} = \Lambda_f^P[g].
\end{aligned}$$

\square

Using similar reasoning, one can bound the bias of divergence estimators that are constructed from variational formulas.

Lemma 13. *Given an objective functional, $H : \mathcal{M}_b(\Omega) \times \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \rightarrow \overline{\mathbb{R}}$, and a test function space $\Gamma \subset \mathcal{M}_b(\Omega)$ we have*

$$\mathbb{E}[\sup_{g \in \Gamma} H[g; Q_n, P_n]] \geq \sup_{g \in \Gamma} \mathbb{E}[H[g; Q_n, P_n]]. \tag{5.56}$$

Lemmas 12 and 13 allow us to bound the bias of both f -divergences and Rényi divergences.

Corollary 14 (Rényi Divergence Bias Bound). *For $\alpha \in (0, 1)$ and $g \in \mathcal{M}_b(\Omega)$ we have*

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{\alpha - 1} \log E_{Q_n} [e^{(\alpha-1)g}] - \frac{1}{\alpha} \log E_{P_n} [e^{\alpha g}] \right] \\ & \geq \frac{1}{\alpha - 1} \log E_Q [e^{(\alpha-1)g}] - \frac{1}{\alpha} \log E_P [e^{\alpha g}] \end{aligned} \quad (5.57)$$

and

$$\mathbb{E}[R_\alpha^\Gamma(Q_n|P_n)] \geq R_\alpha^\Gamma(Q|P). \quad (5.58)$$

Proof. To prove (5.57) we compute

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{\alpha - 1} \log E_{Q_n} [e^{(\alpha-1)g}] - \frac{1}{\alpha} \log E_{P_n} [e^{\alpha g}] \right] \\ & = -\frac{1}{1 - \alpha} \mathbb{E} [\Lambda^{Q_n} [(\alpha - 1)g]] - \frac{1}{\alpha} \mathbb{E} [\Lambda^{P_n} [\alpha g]] \\ & \geq -\frac{1}{1 - \alpha} \Lambda^Q [(\alpha - 1)g] - \frac{1}{\alpha} \Lambda^P [\alpha g] \\ & = \frac{1}{\alpha - 1} \log E_Q [e^{(\alpha-1)g}] - \frac{1}{\alpha} \log E_P [e^{\alpha g}]. \end{aligned} \quad (5.59)$$

(5.58) then follows from Lemma 13. \square

Remark. When $\alpha > 1$ the biases of the two terms in (5.59) compete and so we can not obtain a bias bound via the above method.

Similarly, we have:

Corollary 15 (f -Divergence Bias Bound).

$$\mathbb{E}[E_{Q_n} [g] - \Lambda_f^{P_n} [g]] \geq E_Q [g] - \Lambda_f^P [g]$$

for all $g \in \mathcal{M}_b(\Omega)$ and

$$\mathbb{E}[D_f^\Gamma(Q_n|P_n)] \geq D_f^\Gamma(Q|P). \quad (5.60)$$

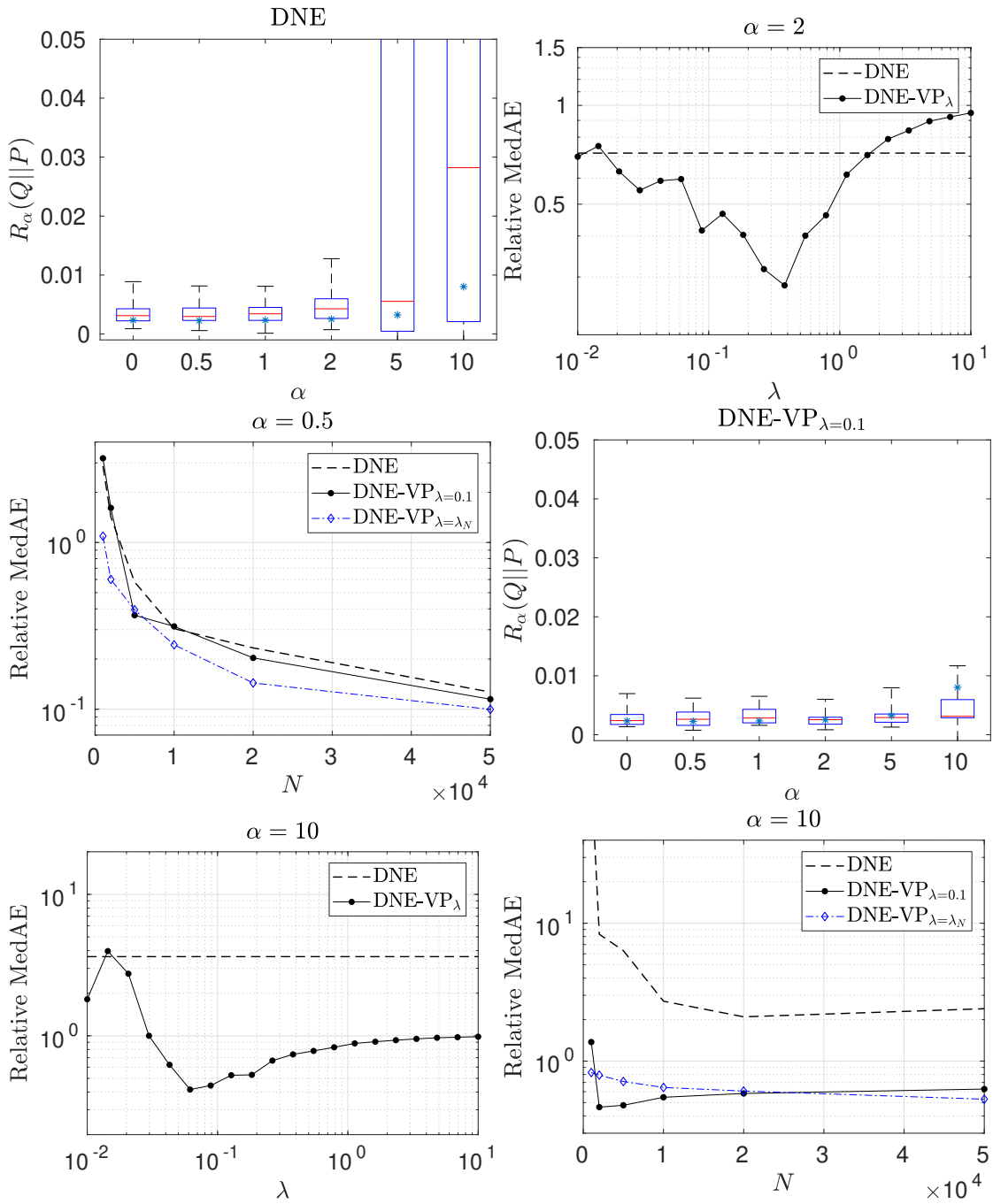


Figure 5.1: Comparison between the estimator without VP (DNE) and with VP (DNE-VP $_\lambda$) for Rényi divergence between two one-dimensional Gaussians with $Q = \mathcal{N}(0, 1.1)$ and $P = \mathcal{N}(0, 1)$. We use $N = 5K$ sample size, 512 as batch size and results are averaged over 100 i.i.d. runs. **Left column:** DNE and DNE-VP $_\lambda$ estimators for increasing values of α . The variance of DNE becomes uncontrollably high for $\alpha > 3$. **Middle column:** Relative MedAE (the lower, the better) for varying penalty coefficient λ and two values of α . The relative MedAE for large values of λ is close to one which implies that the estimated value of DNE-VP $_\lambda$ approaches zero. **Right column:** Relative MedAE for increasing sample size N . We additionally present a penalty coefficient that varies with sample size, shown in blue ($\lambda_N = \frac{500}{N}$ and $\lambda_N = \frac{2000}{N}$ for $\alpha = 0.5$ and $\alpha = 10$, respectively).

5.6 Results on Synthetic Datasets

Figure 5.1 presents the statistical estimation of Rényi divergence between two one-dimensional Gaussians which both have zero mean but different variance values. The order of Rényi divergence, α , controls how much weight to put on the tails of the distributions, thus it can become very sensitive to the few samples from the tails. The same conclusion can be deduced from the variational formula (i.e., (5.1) where α multiplies the exponentials' argument). Therefore, a larger α value implies larger statistical variance. Indeed, high estimation variance is observed with DNE (upper leftmost panel of Figure 5.1) despite the fact that we applied truncation as proposed by [SE19] with truncation threshold set to 1. In contrast, the DNE-VP $_{\lambda}$ estimator with $\lambda = 0.1$ greatly reduces the statistical variance even when α is large (lower leftmost panel). For fairness, we imposed the same truncation operation in the output of DNE-VP $_{\lambda}$. We report a 80% reduction of variance for $\alpha = 2$ which becomes 99% for $\alpha = 10$.

We demonstrate the large variance in the numerical estimation of Rényi divergence with respect to hyper-parameter α . First, we consider two zero-centered univariate Gaussian distributions with different standard deviations, $Q = \mathcal{N}(0, 1.1\sigma_0^2)$ and $P = \mathcal{N}(0, \sigma_0^2)$.

In our simulations, we set $\sigma_0 = 1$. Figure 5.1 shows Rényi divergence estimations as parameter α increases, for the case of $T_{\theta}(\phi) = 0$, $\lambda = 0$ and for the case of variance reduction using T_{θ} and $\lambda = 0.1$.

The proposed approach introduces an additional hyper-parameter, λ , which controls the strength of the VP. Our theory suggests that λ should depend on the sample size (and perhaps also on the other parameters), therefore we perform two sets of experiments. In the first experiment, we explore the range of optimal values for λ in terms of MedAE¹. As is evident from the middle panels of Figure 5.1, λ -values in the vicinity of 0.1 are a reasonable compromise between variance and bias. In the second experiment, we demonstrate the performance in terms of MedAE as a function of the sample size, N . As suggested in Remark 5.3.1, monotone performance is obtained when λ is inversely proportional to N (blue dashed line in rightmost upper panel of Figure 5.1).

Our second synthetic example constitutes the estimation of MI using various approaches with and without VP. Here, we let Q be a zero-mean multivariate correlated Gaussian random vector of dimension d . We impose element-wise correlation, i.e., $\text{corr}(x_i, x_{\frac{d}{2}+j}) = \delta_{i,j}\rho$ to the samples $x \sim Q$ where $i, j = 1, \dots, \frac{d}{2}$ and $\delta_{i,j}$ is Kronecker's delta. With P we denote the product of the marginals, which in this case is simply a zero-mean standardized multivariate Gaussian. Figure 5.2 presents the estimated MI per training step. We consider the Renyi-based MI with $\alpha = 0.5$ as well as the standard MI using the DV variational formula. Notice that these two variants result in different true values (black lines in Figure 5.2). The plotted results demonstrate the successful reduction of variance when VP is added to the objective functional. Interestingly, the extension of VP to InfoNCE and CLUB estimators (second row of panels in Figure 5.2) implies that our approach can be applied to any MI estimators, thus offering a general variance reduction framework. Bias, variance and MSE plots as well as several more experiments can be found in Appendix F.

¹Recall that MedAE stands for median absolute error and it is a more robust-to-outliers metric.

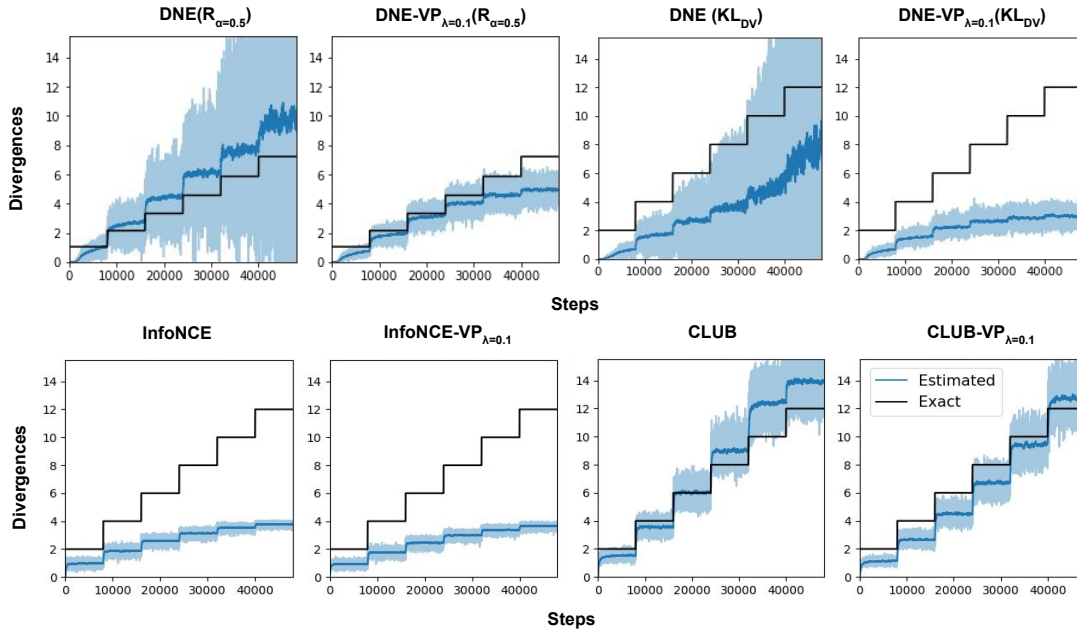


Figure 5.2: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. The number of samples is set to $512K$ and batch size to 64. Panels with $R_{\alpha=0.5}$ in their titles present the Rényi-based MI with $\alpha = 0.5$ whereas the rest of the methods estimate the standard MI (i.e., the KL divergence). In each panel, the true values are shown as a step function (black line). The correlation coefficient of the Gaussian, ρ , for each step is: 0.3084, 0.4257, 0.5091, 0.5741, 0.6273 and 0.6717. The running estimates per minibatch are displayed as shadow blue curves. The dark blue curves shows the moving average of the estimated MI, with a bandwidth equal to 200 steps.

5.7 Real Data Applications

5.7.1 Detecting Rare Biological Sub-Populations

Using the dataset from [LSB⁺15], we test the efficacy of $DNE-VP_{\lambda}$ in discriminating cell populations which are contaminated with a rare sub-population with distinguishable statistical properties. Specifically, we consider single-cell mass cytometry measurements on 16 bone marrow protein markers² (i.e., $d = 16$) coming from healthy and diseased individuals with acute myeloid leukemia. For each run we created three subsets of healthy samples with sample size $N = 20K$ which we denote by P and one dataset as a mixture of 99% healthy and 1% diseased samples which is denoted by Q . Notice that the actual number of diseased samples is only 200 thus it is considered as a rare sub-population.

For Rényi divergence with $\alpha = 0.5$ (left panels in Figure 5.3), both DNE and $DNE-VP_{\lambda}$ are stable. Despite the improvement in the separation of the two histograms, the observed variance reduction of $DNE-VP_{\lambda}$ is minimal and not enough to discriminate between the healthy and the contaminated with 1% diseased samples distributions. When considering Rényi divergence with $\alpha = 1.1$, we observe that DNE fails to produce stable estimates. In contrast, $DNE-VP_{\lambda}$ always computes stable estimates. Additionally, the two histograms are satisfactorily separated, implying that larger values of α are crucial, provided there is a way to handle the statistical variance. For completeness, Table 5.1 reports the first and second order statistics of the histograms shown in Figure 5.3.

²Data was accessed from <https://community.cytobank.org/cytobank/experiments/46098/illustrations/121588>

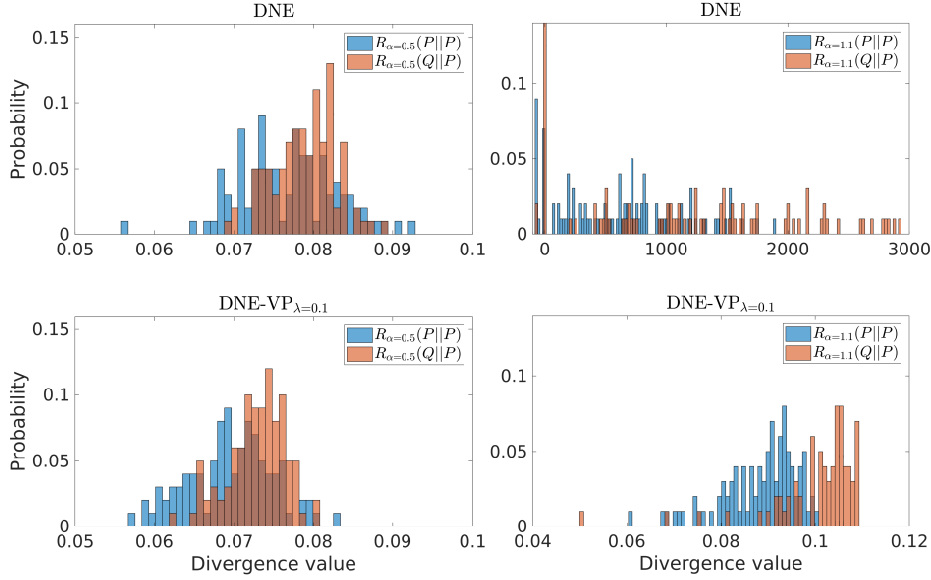


Figure 5.3: Comparison of DNE and DNE-VP $_{\lambda}$ estimators for Rényi divergence on biological data. The histograms of the estimated divergence value are constructed from 100 i.i.d. runs between datasets of $N = 20K$ samples each. Healthy dataset's distribution is denoted by P whereas healthy + 1% diseased dataset's by Q . **Left column:** Rényi divergence with $\alpha = 0.5$. Neither DNE nor DNE-VP $_{\lambda}$ are able to discriminate between the healthy and the 1% contaminated dataset. **Right column:** Rényi divergence with $\alpha = 1.1$. For this α value, VP is compulsory for a stable estimation of Rényi divergence. Furthermore, we are able to discriminate between healthy and 1% contaminated distributions with high accuracy (87.5%).

Table 5.1: Mean values and standard deviation for the histograms shown in Figure 5.3.

	DNE		DNE-VP $_{\lambda=0.1}$	
	mean	std	mean	std
$R_{\alpha=0.5}(P P)$	0.0765	0.0066	0.0695	0.0053
$R_{\alpha=0.5}(Q P)$	0.0789	0.0039	0.0720	0.0036
$R_{\alpha=1.1}(P P)$	676	515	0.0890	0.0089
$R_{\alpha=1.1}(Q P)$	1445	1165	0.1000	0.0120

Part III

Speech Synthesis

Chapter 6

Introduction

Human speech is a vital aspect of communication, encompassing more than conveying information—it's a way we express emotions, connect with others, and make conversations rich and meaningful. Speech involves not just words but also tone, pitch, speed, and rhythm, facilitating the sharing of knowledge and emotions. Speech synthesis applications take on an increasingly crucial role. By harnessing the nuances of human speech, these technologies can offer more natural and empathetic interactions. Speech synthesis has a diverse range of applications that are continually evolving. Voice conversion (VC) is a transformative technology that transcends the boundaries of speech synthesis and speech recognition. It focuses on altering the characteristics of a speaker's speech while retaining the linguistic content and message. Essentially, voice conversion allows one person's speech to be transformed into the speech of another. On the other hand, Text-to-speech (TTS) systems facilitate the conversion of written text into spoken words, and in the realm of virtual assistants and chatbots, TTS empowers these digital entities to comprehend text inputs and respond with spoken outputs, making interactions more natural and intuitive.

All such speech synthesis contributes to accessibility by enabling screen readers to deliver information in a manner that preserves emotional depth and interpersonal connections. This empowers individuals with visual impairments to engage with digital content on a more profound level. This technology bridges the gap between human communication and machines, ensuring that information is accessible and interactions are seamless for a wide array of users and contexts. In entertainment and gaming, speech synthesis brings characters to life, infusing them with distinct voices and emotions, thereby enhancing the authenticity of the experience for players. Furthermore, in language learning applications, it assists learners in achieving mastery of pronunciation and verbal expression, contributing significantly to effective language acquisition and communication skills development. Language learners find such speech synthesis systems invaluable for honing accurate pronunciation and enabling consistent language practice. In essence, speech synthesis stands as a testament to human creativity, forging a connection between people and machines by allowing technology to communicate in a way that closely resembles the natural and nuanced aspects of human speech. This innovation goes beyond mere information transfer; it replicates the emotional richness and interpersonal bonds intrinsic to human language.

In the ever-advancing landscape of technology, it's crucial to recognize that speech synthesis goes beyond merely relaying information; it plays a pivotal role in preserving the core elements of human communication. It acts as a bridge, seamlessly connecting technology with the nuances of human expression and connection. This underscores the importance of embracing these technological strides with empathy, understanding, and inclusivity as guiding principles. In doing so, we ensure that as we harness these innovations, they remain true to the essence of human interaction and contribute to a more harmonious and interconnected world.

6.1 Voice Conversion

Voice conversion (VC) is a subset of speech synthesis, focused on altering one's voice to mimic another's characteristics without changing the content. It plays a crucial role in applications involving voice transformation, emotion, and accent modulation.

Most of the traditional voice conversion techniques assume the availability of parallel training data, meaning paired utterances of the same linguistic content spoken by both the source and target speakers. Voice conversion research began in the late 1980s and has since been categorized into parametric and non-parametric mapping techniques. Parametric techniques, such as Gaussian mixture models (GMM) [SCM98b] and Dynamic Kernel Partial Least Square Regression [GSA⁺15], involve assumptions about the statistical distributions of speech features and their mapping. Non-parametric approaches, like vector quantization [ANSK90] and fuzzy vector quantization [SNA91], are less reliant on assumptions and strive to find the best mapping function while retaining the capacity to generalize to unseen data.

Conventional voice conversion faces challenges due to limited training data, but deep learning leverages large datasets effectively. Neural vocoders can handle low-level detail, while general-purpose acoustic models handle phonetic systems. Deep learning shifts the analysis-mapping-reconstruction pipeline, introducing embeddings for speech content and speaker identity. This aids disentanglement and addresses issues in parallel and non-parallel data voice conversion, pushing voice conversion research forward. Early DNN-based voice conversion focused on spectral transformation, offering non-linear mapping and feature dimension flexibility. Conversion of other features like fundamental frequency and energy contour was also explored [DRY⁺09, MK14]. DNNs mapped spectral representations between source and target speakers, and DBNs extracted latent features [NTTA13, CLLD14b]. Deep autoencoders and layer-wise generative training extended these ideas. LSTM networks improved temporal correlations, especially with bidirectional LSTM (BLSTM) networks. Deep BLSTM networks outperformed DNNs, even without dynamic features, leading to high-quality synthesized voice [NTA14a, HS97].

The attention mechanism [BCB14, VSP⁺17] revolutionized neural networks. Initially applied in machine translation [BCB14], speech recognition [CJLV16], and sequence-to-sequence speech synthesis [WSRS⁺17, PPG⁺17, TUA18], it spurred research in voice conversion. This mechanism enables networks to learn feature mapping and alignment simultaneously during training, eliminating the need for a frame-aligner at runtime. Variations like sequence-to-sequence conversion network (SCENT) [ZLL⁺19] and AttS2S-VC [TKKH19] are based on recurrent neural networks, employing the encoder-decoder with attention architecture [CVMG⁺14, LPM15]. CycleGAN, founded on adversarial learning, employs a generative model in a min-max game with two neural networks: generator and discriminator. It excels in tasks with unpaired training data, such as image manipulation and synthesis [ZPIE17, ZZZ⁺17], speech enhancement [MLG⁺18] and speech recognition [MSK17]. Adversarial training effectively addresses the over-smoothing problem, a major contributor to speech-quality degradation. Recently, CycleGAN-VC2, an enhanced version, was introduced [MSK17], featuring improved objectives (two-step adversarial losses), a better generator (2-1-2D CNN), and an enhanced discriminator (PatchGAN). CycleGAN finds applications in mono-lingual [TWH⁺19], cross-lingual voice conversion [SZDL19], emotional voice conversion [ZSL20], and rhythm-flexible voice conversion [YHC⁺18].

The VAE-based voice conversion framework [HHW⁺16], operates with a decoder that reconstructs utterances based on a latent code from the encoder and a separate speaker code. This speaker code can be an one-hot vector [HHW⁺16] for a closed set of speakers, or alternative representations like i-vectors [DKD⁺10], bottleneck speaker representations [LLY⁺18], or d-vectors [SINT18] for open sets of speakers. Conditioning the decoder on speaker identity compels the encoder to capture speaker-independent information from a multi-speaker database using the latent code. However, VAE decoders tend to produce

overly smoothed speech, potentially resulting in low-quality, buzzy-sounding output. To address this, GANs [GPAM⁺14] were introduced as a solution to the over-smoothing issue [KKHK17]. GANs offer a general framework where a data generator is trained to deceive a discriminator that distinguishes real from fake data produced by the generator. By integrating the GAN concept into VAE, VAE-GAN has been employed in voice conversion with non-parallel training data [HHW⁺17] and cross-lingual voice conversion [SZDL19], yielding more natural-sounding speech compared to standard VAE methods. Recent research on non-parallel voice conversion using sequence-to-sequence models [LCK⁺20] has demonstrated the possibility of explicitly modelling the transfer of other speech aspects, including source rhythm, speaking style, and emotion, to the target speech.

6.2 Text to Speech Synthesis

6.2.1 Deep Learning based Speech Synthesis

In response to these challenges, modern deep learning-based approaches have surged to prominence. These methods leverage extensive datasets and advanced neural networks, showcasing their effectiveness in capturing the intricate nuances of human speech. This has led to the production of synthetic speech that is more natural and fluent. Despite these advancements, some elusive aspects of human speech remain challenging to replicate using traditional synthesis methods. In recent years, the application of deep learning to speech synthesis has witnessed a remarkable surge in interest and development, representing a significant paradigm shift in the field. Deep learning, a subset of machine learning, employs artificial neural networks inspired by the human brain to learn patterns and representations from data automatically. These neural networks, consisting of multiple interconnected layers, have showcased exceptional capabilities in various domains, including image recognition, natural language processing, and, notably, speech synthesis.

Deep learning-based speech synthesis systems have proven to be exceptionally effective and have gained substantial traction in commercial products and applications. The reason behind their success lies in their remarkable ability to capture the subtleties and complexities of human speech. Here's how deep learning achieves this:

Learning from Large Datasets: One of the key strengths of deep learning is its capability to learn from massive datasets. In the context of speech synthesis, this means training neural networks on extensive collections of recorded human speech. The vast amount of data enables these systems to grasp the intricate patterns that define human speech, such as intonation (the rise and fall of voice that conveys meaning), rhythm (the pattern of stressed and unstressed syllables), and timbre (the unique quality of a person's voice).

Hierarchical Representation: Deep neural networks employ a hierarchical approach to feature extraction, where each layer learns progressively more abstract representations of the input data. In speech synthesis, this hierarchical representation allows the model to capture both the fundamental phonetic elements and the higher-level characteristics that make speech sound natural and expressive.

End-to-End Learning: Deep learning systems can be designed for end-to-end learning, where they directly map text input to speech output. This contrasts with traditional speech synthesis methods that involve multiple intermediate steps, such as text-to-phoneme conversion and phoneme-to-speech wave generation. End-to-end models can learn complex mappings directly from data, simplifying the synthesis process.

Adaptive and Context-Aware: Deep learning models are highly adaptive and context-aware. They can adjust their output based on the input text and context, capturing the fine-grained variations in speech that convey emotions, emphasis, and meaning.

Advanced deep learning techniques have created a transformative era for speech synthesis, pushing the boundaries of what's possible and delivering remarkably natural and fluent synthetic speech. These cutting-edge technologies are having a profound impact across a spectrum of applications.

Restricted Boltzmann machines (RBMs): In recent years, restricted Boltzmann machines (RBMs) [LDY13] have found extensive use in modelling speech signals for various applications, including speech recognition, spectrogram coding, and acoustic-articulatory inversion mapping. RBMs are often employed for pre-training deep auto-encoders (DAEs) [DSY+10] or deep neural networks (DNNs) in these contexts. In speech synthesis, RBMs serve as density models for generating spectral envelopes of acoustic parameters, addressing issues like over smoothing in HMM-based synthesis. After training HMMs, state alignments are performed, and RBMs estimate parameters through maximum likelihood estimation (MLE). RBM-HMMs are then constructed to model spectral envelopes. In synthesis, the optimal spectral envelope sequence is estimated based on input sentences and trained RBM-HMMs. While this method improves subjective evaluation results and spectral envelope accuracy compared to traditional HMM-GMM systems, it still struggles with data fragmentation issues inherent to the traditional HMM-based method.

Deep Recurrent Neural Networks (RNNs): In [GS05], the authors presented a modelling approach based on recurrent neural networks (RNNs), leveraging the advantage of RNNs in utilizing context information for input-output mapping. However, traditional RNNs have limitations in accessing extensive context due to issues like vanishing or exploding gradients and difficulty in learning long-term dependencies. To overcome these challenges, [GFGS06] introduced a memory cell and introduced the long short-term memory (LSTM) model, which has become a popular choice. For effective utilization of contextual information, bidirectional LSTM is commonly employed to map input linguistic features to acoustic features.

Convolutional Neural Networks (CNNs): Originally developed for image processing, CNNs have found a second home in the field of speech synthesis. Their adaptability extends to various purposes within this domain, such as feature extraction from spectrograms, which are graphical representations of the acoustic properties of speech. CNNs excel at capturing local patterns, which proves invaluable in improving the quality of synthesized speech by identifying and extracting pertinent acoustic features. They are often deployed in conjunction with other neural network architectures, contributing to the robustness and naturalness of the generated speech. Unlike the WaveNet model, which serves as a vocoder or back-end, [TUA18] primarily functions as a front-end (along with much of the back-end processing) capable of synthesizing spectrograms. Moreover, in [PPG+17], a novel fully-convolutional character-to-spectrogram architecture called Deep Voice 3 was introduced for speech synthesis, enabling fully parallel computation and faster training compared to models relying on recurrent units.

Transformer-based speech Synthesis: Transformer-based models have significantly advanced the field of speech synthesis in recent years. The introduction of the Transformer architecture, originally developed for machine translation tasks, has brought about notable improvements in the quality and efficiency of speech synthesis. Transformer TTS [LLL+19] incorporates a multi-head self-attention mechanism into both the encoder and decoder components of the speech synthesis model. By doing so, they enable the simultaneous construction of hidden states in a parallel fashion, which leads to significant enhancements in training efficiency. FastSpeech [RRT+19] is a feed-forward network for parallel mel-spectrogram generation that extracts attention alignments from a teacher model for phoneme duration prediction and uses a length regulator to align the source phoneme sequence with the mel-spectrogram sequence. Whereas, FastSpeech 2 [RHQ+20] is designed to overcome the challenges encountered in Fast-

Speech and provide an improved solution for the one-to-many mapping problem in TTS. It achieves this by training the model directly using ground-truth targets instead of simplified teacher-generated outputs and incorporating additional variation information such as pitch, energy, and more precise duration as conditional inputs.

Sequence-to-Sequence Speech Synthesis: Sequence-to-sequence (seq2seq) neural networks have proven highly versatile, capable of transducing input sequences into output sequences of varying lengths. These networks have found applications in diverse fields, including machine translation, speech recognition, and image caption generation, consistently delivering promising results. Given that speech synthesis is essentially the reverse process of speech recognition, seq2seq modelling techniques have gained traction in this domain as well. For instance, researchers have utilized content-based attention structures to model acoustic features for speech synthesis, as seen in [WXX⁺16]. Another example is Char2Wav [SMK⁺17], which employs location-based attention to construct an encoder-decoder acoustic model. However, current seq2seq models still grapple with stability issues related to missing or repeating phones. To address this, a forward attention approach for seq2seq acoustic modelling in speech synthesis has been proposed, as discussed in [ZLD18]. The highly acclaimed Tacotron model, also based on seq2seq architecture with an attention mechanism, has been introduced for mapping input text to mel-spectrograms, representing a significant advancement in speech synthesis.

End-to-End Speech Synthesis: TTS systems traditionally comprise a text analysis front-end, an acoustic model, and a speech synthesizer, each trained separately, which can introduce errors that accumulate across these components. To tackle this issue, the field of speech synthesis has seen the rise of end-to-end methods that unify these components into a single framework. End-to-end TTS systems offer several advantages: (1) they can be trained on a large dataset of $\{text, speech\}_i$ pairs with minimal human annotation; (2) they eliminate the need for phoneme-level alignment; and (3) errors do not compound since they rely on a single model.

Notable architectures like Tacotron, WaveNet and WaveRNN have fundamentally altered the landscape. WaveNet [vdODZ⁺16], an evolution of the PixelCNN [vdOKE⁺16] and PixelRNN [VOKK16] models initially employed in image generation, represents a significant breakthrough in the realm of raw audio waveform generation. Introduced by Deepmind in 2016, this model has paved the way for end-to-end speech synthesis. WaveNet stands out for its capacity to produce relatively realistic and human-like voices directly from waveform data trained on real speech recordings. It operates as a complete probabilistic autoregressive model, predicting the probability distribution of the current audio sample based on all preceding samples. A crucial element of WaveNet is the use of dilated causal convolutions, which ensure that the model can only consider the sampling points from 0 to $t - 1$ when generating the t th sample. This innovation has played a pivotal role in WaveNet's ability to generate high-quality audio waveforms. The original WaveNet model operates using autoregressive connections, enabling it to synthesize audio waveforms sequentially, one sample at a time. This process conditions the generation of each new sample on the preceding samples. The joint probability of a waveform X , denoted as x_1, x_2, \dots, x_T , can be factorized as follows:

$$p(X) = \prod p(x_{i+1} | x_1, x_2, \dots, x_i)$$

Where T represents the total number of samples in the waveform. This factorization breaks down the probability of the entire waveform into a product of conditional probabilities for each sample in the sequence, with each sample being dependent on the previous ones. While the original WaveNet paper primarily focuses on its use in TTS systems, there is also the possibility of employing the WaveNet architecture as a statistical vocoder. In this scenario, the generation of speech waveforms is locally conditioned solely by acoustic features [ATS18].

The WaveNet model, while capable of producing high-quality audio, face inherent challenges: first, its slow processing, as each sampling point's prediction relies on preceding ones; and second, its dependence on linguistic features from a text-to-speech (TTS) front-end, rendering it vulnerable to errors in text analysis. To tackle these issues, the parallel WaveNet was introduced, significantly boosting sampling efficiency and generating high-fidelity speech samples over 20 times faster [OLB⁺18a]. Additionally, the neural model Deep Voice [ACC⁺17] emerged as an alternative, replacing each TTS component with a corresponding neural network. However, it falls short of true end-to-end synthesis as its components are trained independently.

Tacotron [WSRS⁺17] is a fully end-to-end speech synthesis model that has revolutionized the field by enabling the training of speech synthesis models directly from $\langle \text{text}, \text{audio} \rangle$ pairs, eliminating the need for labor-intensive feature engineering. One of its remarkable features is its applicability to a wide range of languages, including Chinese Mandarin, as it operates at the character level. Tacotron utilizes a sequence-to-sequence (seq2seq) model with an attention mechanism to convert text into a mel-spectrogram, a robust representation of speech. While mel-spectrograms lack phase information crucial for audio reconstruction, Tacotron employs the Griffin–Lim algorithm [GL84] to iteratively estimate this phase information from the spectrogram during audio reconstruction. Tacotron's end-to-end nature has garnered significant research attention, leading to several improved versions and open-source clones that reproduce speech quality akin to the original work. Some researchers have incorporated deep generative models like Variational Auto-encoders (VAE) [KW13] into Tacotron to explicitly model speaker states and control speaking styles. Additionally, there are hybrid systems that combine Tacotron and WaveNet for speech synthesis, such as Deep Voice 2 [GAD⁺17], which employs Tacotron to transform text into a linear scale spectrogram and then uses WaveNet to generate speech. Tacotron2 [SPW⁺18], another notable system, has achieved high mean opinion scores (MOS) comparable to human speech by unifying a seq2seq Tacotron-style model for mel-spectrogram generation with a WaveNet vocoder for speech synthesis.

WaveRNN [KES⁺18] vocoder represents a groundbreaking advancement in the domain of text-to-speech (TTS) synthesis, characterized by its intricate neural network architecture and impressive waveform generation capabilities from mel spectrograms. At its core, WaveRNN employs a combination of recurrent neural networks (RNNs), particularly long short-term memory (LSTM) or gated recurrent unit (GRU) cells, with autoregressive generative models. This fusion of techniques allows it to tackle the challenging task of generating high-quality speech waveforms directly from textual input. What distinguishes WaveRNN from conventional TTS models is its ability to model the conditional probability distribution of audio waveforms, given the input text and context. By adopting a fully autoregressive approach, WaveRNN generates audio samples one at a time, conditioning each sample on the previously generated ones. This autoregressive process inherently captures the temporal dependencies and fine-grained details essential for natural speech, including pitch variations, phonetic nuances, and prosodic features.

GAN-based vocoders represent a significant advancement in speech synthesis, often surpassing autoregressive models in terms of both speed and speech quality. These vocoders leverage the principles of Generative Adversarial Networks (GANs) [GPM⁺14] to generate speech waveforms. GAN-based vocoders typically consist of a generator network responsible for modelling the waveform signal in the time domain and a discriminator network to evaluate and enhance the quality of the generated speech. Two notable models in this category are MelGAN and Parallel WaveGAN. MelGAN [KKdB⁺19] adopts the standard GAN architecture for rapid waveform generation. It utilizes a fully convolutional model for high-quality Mel-Spectrogram inversion. With fewer parameters compared to autoregressive models, MelGAN achieves a higher real-time factor on both GPU and CPU, all without requiring hardware-specific optimization. On the other hand, Parallel WaveGAN [YSK20] is a distillation-free, fast, and compact model designed for waveform synthesis. This architecture optimizes both the waveform-domain adversarial loss and the multi-resolution short-time Fourier transform (STFT) loss, offering a compelling balance between

synthesis speed and speech quality. These GAN-based vocoders represent cutting-edge innovations in the field of speech synthesis, promising improved efficiency and performance.

Diffusion probabilistic models introduce a unique approach to generative models, consisting of two fundamental processes: the diffusion process and the reverse process [HJA20]. In the diffusion process, a Markov chain gradually introduces Gaussian noise to the original signal until it becomes degraded. Conversely, the reverse process is a denoising procedure that progressively eliminates the added Gaussian noise, ultimately restoring the original signal. Within our study, we explore two diffusion-based vocoders: WaveGrad and DiffWave. WaveGrad [CZZ+20] draws inspiration from prior work on score matching and diffusion probabilistic models. This model takes white Gaussian noise as input and, conditioned on the Mel-Spectrogram, iteratively refines the signal using a gradient-based sampling technique. DiffWave [KPH+20] is a versatile diffusion probabilistic model designed for waveform synthesis, showcasing robust performance in both conditional and unconditional scenarios. It operates by employing white Gaussian noise as input, initiating a Markov chain process with a fixed number of steps to generate a structured waveform progressively. The model's training objective focuses on optimizing a variation of the variational bound on the data likelihood. These diffusion-based vocoders represent a novel direction in speech synthesis, offering intriguing possibilities for generating high-quality speech waveforms.

These state-of-the-art techniques represent a seismic shift in speech synthesis, offering an unprecedented level of naturalness and fluency in synthetic speech. They are revolutionizing not only TTS systems but also applications across diverse domains where lifelike speech generation is crucial. As technology continues to advance, we can anticipate even more refined and natural synthetic speech, further blurring the line between machine-generated and human-generated voices. As technology continues to evolve, deep learning-based speech synthesis is expected to advance further, with models becoming even more natural-sounding and capable of accommodating a broader range of linguistic nuances and languages. This transformative shift is reshaping the landscape of human-computer interaction and communication, with profound implications for accessibility, entertainment, education, and beyond.

6.3 Intelligible Speech Synthesis

The significant progress in speech synthesis over the past decade has created opportunities to enhance real-world speech communication, but the challenge of background noise remains critical. Speech intelligibility, denoting the extent to which spoken content is understandable, is vital in various contexts, from emergency alerts to human-machine interactions. Conversely, speech quality gauges how natural and engaging speech sounds, impacting user satisfaction in applications like virtual assistants and entertainment systems. The presence of background noise poses substantial challenges, distorting speech signals and diminishing both intelligibility and quality. Novel human assistive devices, such as hearing aids and voice assistants, depend on clear and natural speech, making it imperative to ensure their effectiveness in noisy environments. Thus, harnessing neural models to address noise-related issues in preserving speech quality and intelligibility is a key focus, holding promise for improved real-world communication experiences for both humans and machines.

In a typical speech communication scenario, both the speaker (far-end) and the listener (near-end) contend with the challenges posed by background noise. Speech processing models are designed with the primary objective of effectively conveying the speaker's message to the listener, even in the presence of these disruptive noises. The noise that affects the speech acquisition at the speaker's side is termed "far-end noise," given its spatial relation to the listener. Conversely, the noise at the listener's end, which impairs the listener's perception of the speech, is referred to as "near-end noise." Although distortions may arise during speech transmission between the two ends, this discussion predominantly focuses on the

impact of ambient noise on speech quality and intelligibility, assuming an ideal transmission channel.

Efforts to enhance listeners' speech intelligibility have historically involved the modification of speech spectral and/or temporal structures before its presentation in challenging listening environments. This approach parallels observations in human speech, where individuals adjust their articulation when speaking in noisy surroundings to mitigate the masking effect caused by background noise—an adaptation known as the Lombard reflex [Jun96]. Lombard speech, characterized by a higher fundamental frequency, reduced speaking rate, and flatter spectral tilt compared to normal speech in quiet conditions, has proven to be more intelligible to listeners in noisy environments, even after accounting for loudness variations [CMV14, Jun96, BC20, CL12]. Consequently, it becomes evident that speech needs to adapt to the listening context to ensure its intelligibility to the interlocutor. Likewise, artificial modifications of speech are essential in speech output devices to guarantee intelligibility across diverse operating conditions.

Chapter 7

Non-parallel Voice Conversion using Weighted Generative Adversarial Networks

7.1 Introduction

Speech, being the most convenient and effective mode of communication, has prompted significant interest in man-machine interface research due to recent advancements in computer technology. Interacting with computers through speech has been a longstanding goal, necessitating research in various domains. In particular, speech recognition and speech synthesis are crucial techniques for emulating human communication with accuracy and naturalness.

The process of producing speech begins with the expulsion of air from the lungs, followed by its passage through the trachea, larynx, and ultimately into the vocal tract. Voiced sounds are generated when the airflow reaches the glottis, causing the vocal folds to vibrate and produce a quasi-periodic puff-like sound source. This source signal consists of harmonics, which are different frequencies present in the periodic waveform. Conversely, unvoiced sounds are created when the vocal folds do not vibrate, and the airflow through the glottis becomes turbulent, resembling random noise. By considering the source-filter model of speech production, we can perceive the source signal as the airflow waveform originating from the vocal folds. Simultaneously, the physical vocal tract can be viewed as a filter that modifies the spectral characteristics of the source signal. Comprising the pharynx, oral cavity, and nasal cavity, the vocal tract possesses varying cross-sectional areas corresponding to the articulators' positions.

A speech signal contains both linguistic and para-linguistic information, each conveying distinct aspects. Linguistic information pertains to language-specific or dialect-related characteristics, while para-linguistic information encompasses speaker timbre and prosody. Thus, speech encompasses more than just the words spoken and carries additional details such as the emotion, attitude, and individuality of the speaker. Speaker identity, in particular, can be characterized by the following factors:

Linguistic factors: These factors are inherent in the spoken utterance of a speaker. Speaker identity is influenced by the speaker's language or dialect, specific terminology, and individual lexicon patterns. These linguistic characteristics are often shaped by factors such as place of birth, social status, family background, and community affiliations.

Supra-segmental factors: Supra-segmental features refer to prosodic characteristics that contribute to speaker identity. They include the duration of phonemes, syllables, and words, as well as aspects such as fundamental frequency (pitch contour), rhythm, duration and placement of pauses, tone, etc.

Segmental factors: Segmental acoustic descriptors play a significant role in speaker identity. These factors involve short-term features such as the spectrum, formants (resonant frequencies of the vocal tract), and the shape of the glottal excitation pulse.

Together, these linguistic, supra-segmental, and segmental factors contribute to the holistic understanding and recognition of speaker identity.

Speech synthesis is the process of artificially generating human-like speech. This thesis's specific area of focus is voice conversion (VC), which falls under the umbrella of speech synthesis. VC systems aim to transform an utterance spoken by a source speaker into a different target speaker's perceived voice while preserving the original utterance's linguistic content.

In VC, the desired conversion factors pertaining to speaker identity include supra-segmental and segmental acoustic features within a fixed linguistic context. However, due to their relatively easier extraction and modelling, as well as their rich speaker characteristics, this thesis primarily focuses on segmental-level conversion, specifically spectral mapping. The conversion of supra-segmental level factors, which encompass prosodic characteristics associated with speaker identity, falls outside the scope of this thesis and remains a challenging task to be addressed in future research.

Voiced sound exhibits periodicity, corresponding to the opening and closing of the vocal folds. The closed phase of the vocal folds is when they are shut, while the open phase is when they are apart. The duration of a complete glottal cycle determines the pitch period of the resulting speech signal, and its inverse is known as the fundamental frequency. The pitch frequency (F_0) represents the rate at which the vocal folds vibrate and is influenced by physical factors such as vocal fold elasticity and mass. Typically, male speakers have a lower pitch frequency range (60-150 Hz) compared to female speakers (200-400 Hz). The peaks in the spectral envelope, known as formants, represent vocal tract resonances. The human vocal tract is a tube excited at one end during voiced speech production. The resonant frequencies of this tube correspond to the formants in speech. Formant frequencies provide a concise representation of the time-varying speech signal. The vocal cords' fundamental frequency (F_0) and the first four formant frequencies (F_1 - F_4) of the vocal tract play a crucial role in characterizing a speaker's identity.

In contrast, unvoiced sounds occur when the vocal folds do not vibrate. The airflow becomes turbulent either through the glottis or constrictions formed by the articulators. Here, the speech waveform lacks periodicity and appears random. The spectrum does not exhibit harmonic structure, and the spectral envelope differs from that of voiced speech. As mentioned earlier, the speech signal listeners perceive results from filtering the glottal flow wave through the vocal tract. The positions of the articulators determine the shape of the vocal tract. Speakers control the articulator positions to produce specific phonemes. However, the vocal folds and oral and nasal cavities differ among speakers, leading to distinct speech waveforms even when the same phonetic content is uttered.

In summary, speaker individuality is characterized by a combination of factors, including segmental and supra-segmental aspects, with each factor exerting varying influence depending on the speaker. From the perspective of speech perception, fundamental frequency and formants carry essential information for speaker identification. Voice conversion, which aims to represent speaker individuality using a reduced number of parameters, confirms the suitability of these features.

Voice conversion technology finds applications in various domains. For instance, it can be utilized to customize text-to-speech (TTS) systems. TTS involves converting written text into speech sig-

nals. Corpus-based TTS, which produces high-quality synthetic speech, requires a substantial amount of recorded speech data from a specific speaker. The process of recording and processing such data to build a personalized TTS system is costly and time-consuming. In this context, VC provides an economical and efficient solution for creating new voices for TTS by utilizing a limited amount of data to convert the voice of the source speaker (a TTS system) to that of the target speaker. Additionally, VC technology is applicable in other domains such as automatic speech-to-speech translation, voice dubbing, education, speaking aids, and entertainment [KM98b, NTSS12, TNS12, EB08]. Presently, VC methods are also employed in investigating vulnerabilities associated with automatic speaker verification (ASV) systems.

Voice conversion can be formulated as a regression problem of estimating a mapping function from source to target speech. A large number of popular statistical approaches like *linear multivariate regression* (LMR) [VMT92], *Gaussian mixture model* (GMM) [SCM98a], *joint density GMM* (JD-GMM) [KM98a] were introduced more than two decades ago which proved quite successful. Over the time, several non-linear spectral mapping techniques based on *restricted Boltzmann machine* (RBM) [NTA14b], *feed-forward deep neural networks* (DNNs) [DBYP10, CLLD14a], *recurrent DNNs* [SKLM15] and *non-negative matrix factorization* (NMF) [WVCL14] have also been proposed. However, most of these conventional VC methods require aligned parallel source and target speech data for training. In many scenarios, it is troublesome to collect parallel utterances. Even when parallel data is accessible, the required alignment procedures introduce artefacts and lead to speech-quality degradation. Numerous attempts have been made to overcome these limitations to develop non-parallel VC methods. *Sequence-to-sequence* (Seq2Seq) learning has proved to be outstanding at various research tasks and was successfully adopted in VC [MSTS17, ZLL⁺19, TKKH18]. Seq2Seq VCs mainly use multiple modules such as Automatic speech recognition (ASR) and TTS, which are trainable with pairs of speech and its transcript rather than the source-target speech. These approaches convert both acoustic features and the duration of the source speech. Nonetheless, these techniques consist of several training procedures, and they are expensive in terms of both external data and computation.

Conditional variational autoencoders (CVAEs) approach were recently adopted for VC [HHW⁺16, SINT18]. CVAEs are an extended version of *variational autoencoders* where the encoder and decoder networks can take additional auxiliary input variables. The VC has experienced significant improvements following the introduction of *generative adversarial networks* (GANs). The VAE-GAN framework is an alternate approach for non-parallel VC that overcomes the weakness of VAEs [HHW⁺17]. Furthermore, a variation of GANs named *cycle-consistent GAN* (CycleGAN) was presented in [KK17]. CycleGAN utilizes a frame-by-frame approach, which is designed to learn forward and inverse mappings simultaneously using an *adversarial loss* and *cycle-consistency loss*. One of the drawbacks of CycleGAN-VC is the ability to learn only one-to-one mappings. To resolve this issue, *Star generative adversarial network* based VC (StarGAN-VC) was recently introduced [KKTH18], which was originally proposed as a method for simultaneously learning images among multiple domains [CCK⁺18]. It possesses a unified model architecture which allows simultaneous training of multiple domains, i.e., many-to-many mapping within a single network.

Even though a significant amount of research has been provided in the literature for non-parallel methods, generating high-quality audio quality is still very challenging and has room for improvement. This chapter extends the work of StarGAN-VC and proposes a novel training algorithm inspired by *Weighted GAN* (WeGAN) [PPFSed]. Furthermore, the existing StarGAN-VC utilizes three loss functions. However, it lacks stable training, which can be overcome by *Wasserstein GANs with gradient penalty* (WGAN-GP) [GAA⁺17b]. Our proposed approach introduces a new and effective weight factor for WGAN-GP. The proposed *Weighted StarGAN* (WeStarGAN) algorithm improves the training of the Generator by transferring ideas from Game Theory. The new algorithm puts more weight on generated samples whose data distribution is closer to the real samples and is more likely to fool the Discriminator. Simultaneously,

it reduces the weights of generated samples that are confidently discriminated against as fake. By doing so, WeStarGAN enhances the robustness of the weak Generator by adding weights to the training process, and we expect that the inferred Generator will be stronger, favourably affecting the convergence properties. Experimental results based on subjective performance evaluation confirm that our proposed method achieves better speaker similarity and perceptual speech quality than the baseline StarGAN-VC system.

7.2 GAN Architectures

7.2.1 Star Generative Adversarial Networks

Our proposed model is adapted from the StarGAN [CCK⁺18], which was proposed for multi-domain image-to-image translation and slightly differs from the StarGAN-VC [KKTH18] in terms of both cost functions and DNN architectures.

The objective is to train a single Generator G that learns mappings among multiple domains, i.e., many-to-many speaker conversion. To achieve this, we train G to convert the attribute of source \mathbf{x} speaker domain into target \mathbf{y} speaker domain conditioned on the target domain label c , $\mathbf{y}' = G(\mathbf{x}, c)$. The target domain label c is generated randomly so that G can learn the flexibility to transform the source speech. An auxiliary classifier is introduced that allows the Discriminator to control multiple domains. Fig. 7.1 illustrates the training process of StarGAN-VC approach.

We applied three losses in the objective function, Adversarial Loss, Domain Classification Loss and Reconstruction Loss.

Adversarial Loss: G generates fake data $G(\mathbf{x}, c)$ conditioned on both the source speaker's data \mathbf{x} and the target domain label c , while D tries to distinguish between real and fake data. While training, G tries to minimize this objective, while the Discriminator D tries to maximize it. Moreover, we implemented Wasserstein GAN with gradient penalty [GAA⁺17b], which uses a penalty term in the loss and provides strong performance and stability. The modified adversarial loss for D is defined as,

$$\mathcal{L}_{adv-gp}^D = \mathbb{E}_{\mathbf{x} \sim p_{src}}[-D(\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim p_{src}, c} D(G(\mathbf{x}, c)) + \lambda_{gp} \mathbb{E}_{\hat{\mathbf{x}}} [(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1)^2], \quad (7.1)$$

$$\mathcal{L}_{adv}^G = -\mathbb{E}_{\mathbf{x} \sim p_{src}, c} [D(G(\mathbf{x}, c))], \quad (7.2)$$

Where $\hat{\mathbf{x}}$ is sampled uniformly along a straight line between a pair of real and generated data samples, and λ_{gp} is a constant value.

Domain Classification Loss: An auxiliary classifier is implemented similar to D , which imposes the domain classification loss while optimizing the cost function. Two loss terms are incorporated here: domain classification loss of real speech data, which optimizes D , and a domain classification loss of fake speech data, which optimizes G . The losses are as follows,

$$\mathcal{L}_{cls}^{real} = \mathbb{E}_{\mathbf{x} \sim p_{src}, c'} [-\log D_{cls}(c'|\mathbf{x})], \quad (7.3)$$

$$\mathcal{L}_{cls}^{fake} = \mathbb{E}_{\mathbf{x} \sim p_{src}, c} [-\log D_{cls}(c|G(\mathbf{x}, c))], \quad (7.4)$$

where $D_{cls}(c'|\mathbf{x})$ represents a probability distribution of real data \mathbf{x} over domain labels computed by D . D learns to classify real data to its corresponding original domain c' . Whereas $D_{cls}(c|G(\mathbf{x}, c))$

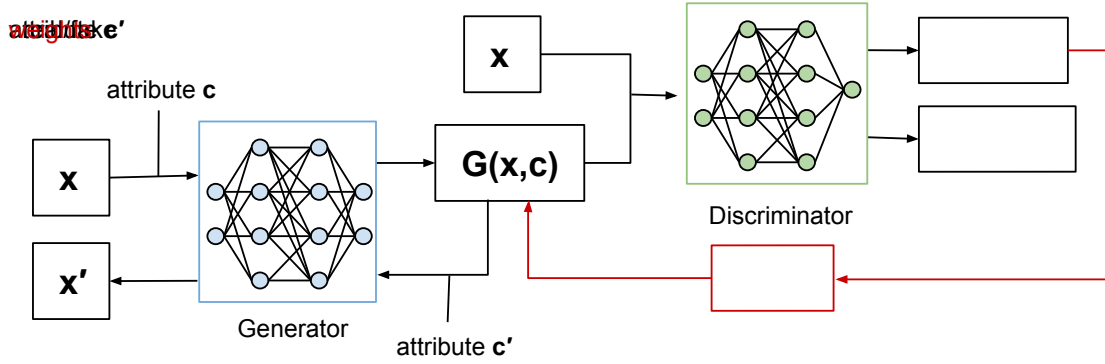


Figure 7.1: Overview of StarGAN (in black), consisting of two modules, a Discriminator D (identical neural network architecture is used for Classifier except for the last convolutional layer) and a Generator G . The weights (in red) are introduced during the training optimization process in our proposed algorithm.

represents the probability distribution of fake data $G(x, c)$ over domain labels computed by D . G tries to minimize this objective to generate data that will be classified as target domain c .

Reconstruction Loss: The adversarial and classification losses assist G to generate speech that is realistic and can be classified to its correct target domain. However, this does not guarantee preserving the content of the linguistic information while changing only the speaker domain-related information. To alleviate this problem, a reconstruction loss is introduced to the Generator, defined as,

$$\mathcal{L}_{rec} = \mathbb{E}_{x \sim p_{src, c, c'}} [\|x - G(G(x, c), c')\|_1], \quad (7.5)$$

where $G(x, c)$ is the generated data conditioned on x and the target domain label c and $G(G(x, c), c')$ is reconstruct the original speech x which is conditioned on $G(x, c)$ and the original domain label c' . We applied $L1$ norm as a reconstruction loss.

The overall objective functions to be minimized with respect to G and D can be written as

$$\mathcal{L}_D = \mathcal{L}_{adv-gp}^D + \lambda_{cls} \mathcal{L}_{cls}^{real}, \quad (7.6)$$

$$\mathcal{L}_G = \mathcal{L}_{adv}^G + \lambda_{rec} \mathcal{L}_{rec} + \lambda_{cls} \mathcal{L}_{cls}^{fake}, \quad (7.7)$$

where λ_{rec} and λ_{cls} are the hyper-parameters for domain classification loss and reconstruction loss, respectively.

7.2.2 Training StarGAN with Weights (WeStarGAN)

In [PPFSed], authors presented a training algorithm based on weights that improved the performance of vanilla GANs. Instead of equally-weighted 'fake' samples, a weight to each sample is assigned which multiplies to the respective gradient term of the Generator. The weights are designed to impose more strength on samples that fool the Discriminator and thus are closer to the real data. Intuitively, the weighted algorithm puts more weight on fake samples that are more likely to fool the Discriminator and simultaneously reduces the weight of samples that are confidently discriminated as fake. A theoretical argument reveals that the optimal Generator with weights achieves a lower or equal loss value than the optimal Generator with equally weighted samples for a fixed Discriminator. Hence, it is expected that the inferred generator will be stronger and favourably affect both the point and the speed of convergence with minor additional computational costs. The proposed algorithm is presented in Fig. 7.2.

Algorithm 1

number of iterations k steps Sample $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ from the source data distribution $p_{src}(\mathbf{x})$.

Update the Discriminator to minimize the objective function:

$$\begin{aligned} & \frac{1}{m} \sum_{i=1}^m [-D(\mathbf{x}_i) + D(G(\mathbf{x}_i, c))] \\ & - \frac{1}{\eta} \sum_{i=1}^m \lambda_{cls} \log D_{cls}(c' | \mathbf{x}_i) \\ & + \frac{1}{m} \sum_{i=1}^m \lambda_{gp} (\|\nabla_{\hat{\mathbf{x}}_i} D(\hat{\mathbf{x}}_i)\|_2 - 1)^2 \end{aligned}$$

Sample $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ from the source data distribution $p_{src}(\mathbf{x})$.

Normalize:

$$\bar{D}_i = D(G(\mathbf{x}_i, c)) - \frac{1}{2m} [\sum_{j=1}^m D(\mathbf{x}_j) + D(G(\mathbf{x}_j, c))].$$

Compute the unnormalized weights:

$$w_i = e^{\eta \min(0, \bar{D}_i)}, \quad i = 1, \dots, m.$$

Normalize:

$$w_i = \frac{w_i}{\sum_{j=1}^m w_j}, \quad i = 1, \dots, m.$$

Update the Generator to minimize the objective function:

$$\begin{aligned} & \sum_{i=1}^m -w_i D(G(\mathbf{x}_i, c)) \\ & + \frac{1}{\eta} \sum_{i=1}^m \lambda_{rec} \|\mathbf{x}_i - G(G(\mathbf{x}_i, c), c')\|_1 \\ & - \frac{1}{m} \sum_{i=1}^m \lambda_{cls} \log D_{cls}(c | G(\mathbf{x}_i, c)) \end{aligned}$$

Figure 7.2: Training algorithm of WeStarGAN. For a direct comparison with the StarGAN, we follow the formulation of [KKT18].

We extend the training algorithm to Wasserstein GANs with gradient penalty (WGAN-GP). In WGAN-GP, the discriminator does not return the probability of a real sample but a continuous regression-type value. Taking this fact into account, we uniformly scale the output of the Discriminator $D(G(\mathbf{x}, c))$ based on the output of the Discriminator conditioned on both real and fake data and translate the data around the axis 0. The normalized output is then employed in the weight function. The proper weights for WeStarGAN's Generator are defined by

$$w_i = e^{\eta \min(0, \bar{D}_i)} \quad (7.8)$$

where η corresponds to the hyper-parameter, which weighs the factor of the weight values. Note that the normalized \bar{D}_i is only used to estimate the weights. We empirically set $\eta = 0.1$ for our experiments.

The choice for the weights is dictated by the fact that we focus on improving the Generator training by putting more attention on the data that are closer to real distribution. Therefore, when the Discriminator output is $\bar{D}_i < 0$, the weight decreases by an exponential factor. On the other hand, when $\bar{D}_i > 0$, Our algorithm takes into account the samples which almost follow the real data distribution.

7.3 Experimental Setup

7.3.1 Experimental conditions

The experiments have been conducted with the CMU Arctic database [KB04a] that consists of speech spoken by two male speakers (rms and bdl) and two female speakers (clb and slt) and are divided into two subsets i.e., training and evaluation, without overlap. As there are four speakers involved in our experiments, the attribute c is represented as a four-dimensional one-hot vector depending upon the target speaker attribute. Although, the database contains parallel speech, we randomly select training data as our system operates on non-parallel data, The sampling rate of the speech signals is 16 kHz. For each utter-

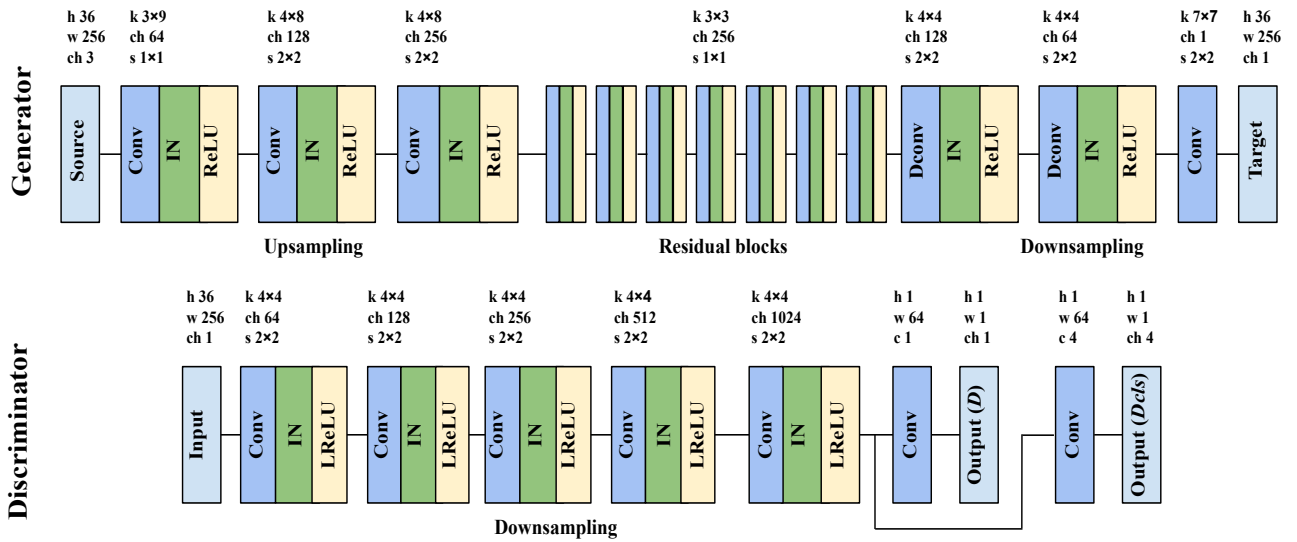


Figure 7.3: Overview of StarGAN [KKTH18], consisting of two modules, a discriminator D and a generator G . In the input and output layers, h , w , and ch represent height, width, and number of channels, respectively. In each convolutional layer, k , c , and s denote kernel size, number of output channels and stride size, respectively. “Conv”, “IN”, “ReLU”, “LReLU”, and “Deconv” denote convolution, instance normalization, rectified linear unit, leaky rectified linear unit and transposed convolution, respectively. D_{cls} provides a probability distribution over domain labels where the domain corresponds to the number of speakers used to train VC.

ance, 36 dimension mel-cepstral coefficients (MCCs), logarithmic fundamental frequency ($\log F0$), and aperiodicities (APs) were extracted for every 5 ms using the WORLD analyzer [MYO16]. The $\log F0$ is converted using the logarithm normalized transformation, and the aperiodicities are used directly without any modification. Once the training process is completed, we use WORLD vocoder to generate speech from converted features.

7.3.2 Network architectures

In the Generator, an acoustic feature sequence is inserted, and the output is an acoustic feature sequence of the same length. We normalize the source and target MCCs per dimension. The generator network comprises three convolutional layers (conv), six residual blocks, three transposed convolutional layers (Dconv), and seven conv layers, which are used for the discriminator. Whereas, in [KKTH18], five conv layers and five Dconv layers are considered in the Generator, and two separate five conv layers are used for Discriminator and Classifier networks. Instance normalization is used for the generator, but no normalization is used for the discriminator. All models are trained using Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batch size is set to 32. The overview of network architecture is depicted in Fig 7.3.

7.4 Results and Discussion

In this section, we present the experimental results to evaluate the performance of voice-converted speech samples. To assess the performance based on subjective evaluation experiments, we conducted listening tests for the speech quality (i.e., naturalness) and speaker similarity of the converted speech to the target speech. Our proposed WeStarGAN-VC architecture was compared against the recently proposed StarGAN-VC architecture. Two separate listening tests are reported, the ‘ABX’ and ‘AB’ tests. In the ‘ABX’ test, experimental subjects have to decide whether a given sentence ‘X’ is closer in vocal quality to

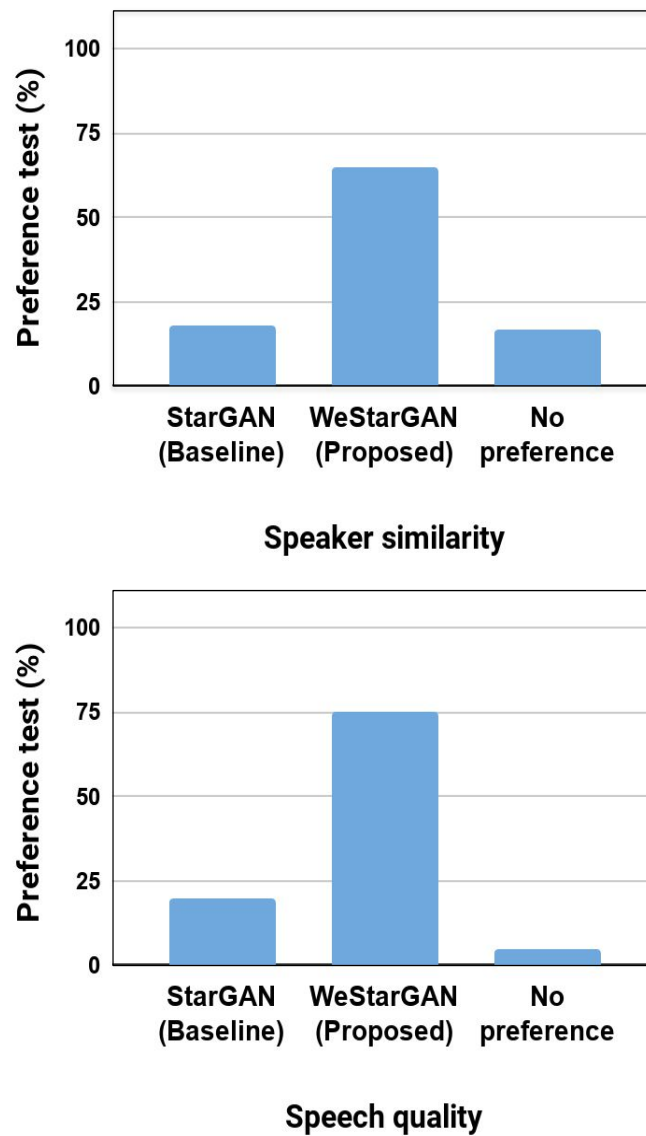


Figure 7.4: Subjective preference test in (%) for speaker similarity and speech quality.

one of a pair of sentences ‘A’ and ‘B’, which are converted speech samples obtained with the proposed and baseline methods, not necessarily in that order. Meanwhile, the ‘AB’ test compares the quality or naturalness of the converted speech. Fifteen native and non-native English listeners participated in our listening tests. All the converted speech samples were presented randomly from the evaluation set. Furthermore, the evaluation samples contain both intra-gender pairs and cross-gender pairs.

The evaluation results of the preference test are demonstrated in Fig. 7.4. The proposed WeStarGAN algorithm obtained the majority of preferences for best conversion in terms of sound quality and speaker similarity. For speaker similarity, the result shows that 17% preferences were given to the ‘No preference’ option, which indicates similar speaker characteristics in the speech samples generated using both approaches. Nevertheless, the proposed method performs better with 65% preference. Moreover, WeStarGAN significantly outperforms the baseline in generating good speech quality. The significant improvement in speech quality might be attributed to the fact that weights are only multiplied by the fake samples of the adversarial loss function, which generates real-like speech samples. On the contrary, no

weights are introduced to the domain classification loss that is responsible for speaker mapping. We finally remark that WeStarGAN has the potential to be used for the training of lighter Generators which are necessary in cases such as operating on mobile devices.

7.5 Conclusion

In this chapter, we proposed WeStarGAN, a novel algorithmic variation of StarGAN capable of performing non-parallel multi-domain voice conversion tasks. With minor additional computational costs, the suggested approach improved the training process by devising a stronger generator at each minibatch iteration. This development is crucial because our approach can overcome the limitation of using a weaker generator and still successfully train it to generate good-quality speech samples. In addition, we extended the weighting approach to the more stable WGAN-GP model. The subjective evaluation revealed that the proposed method obtained higher sound quality and speaker similarity than the baseline method.

Chapter 8

Speaker Conditional WaveRNN: Towards Universal Neural Vocoder for Unseen Speaker and Recording Conditions

8.1 Introduction

Speech synthesis has received attention in the research community as voice interaction systems have been implemented in various applications, such as personalized Text-to-Speech (TTS) systems, voice conversion, dialogue systems and navigations [Dut97, Tay09, SCM98b, PPS19]. In the past, conventional statistical parametric speech synthesis (SPSS) exhibited high naturalness under best-case conditions [ZTB09, Kin11]. Hybrid synthesis was also proposed as a way to take advantage of both SPSS and unit-selection approach [QSY12, MCW⁺16]. Most of these TTS systems consist of two modules: the first module converts textual information into acoustic features while the second one, i.e., the vocoder, generates speech samples from the previously generated acoustic features.

Traditional vocoder approaches mostly involved source-filter models for the generation of speech parameters [MQ86, MC90, KMKDC99, MYO16]. The parameters were defined by voicing decisions, fundamental frequency (F0), spectral envelope or band aperiodicities. Algorithms like Griffin-Lim utilized spectral representation to generate speech [GL84, PBS13]. However, the speech quality of such vocoders was restricted by the inaccuracies in parameter estimation. Recently, the naturalness of vocoders has been significantly improved by benefiting from the direct waveform modelling approach. Neural vocoders like WaveNet utilize an autoregressive generative model that can reconstruct waveform from intermediate acoustic features [ODZ⁺16, THK⁺17]. To overcome the time complexity at inference, a parallel wave generation approach was adopted to generate speech in real time [OLB⁺18b, PPC19]. Wave Recurrent Neural Networks (WaveRNN), which employs recurrent layers, increase sampling efficiency without compromising their quality [KES⁺18]. In particular, introducing a gated recurrent unit (GRU) can realise real-time high-quality synthesis. Although WaveRNN has been suggested to focus on text-to-speech synthesis, our work exercises it as a vocoder while changing the conditioning criteria from linguistic information to acoustic information. Other recent works have been also found in literature, notable among them are SampleRNN [MKG⁺17], WaveGlow [PVC19], LPCNet [VS19] and MelNet [VL19].

Techniques in neural vocoders involve data-driven learning and are prone to specialize to the training data which leads to poor generalization capabilities. Moreover, in multi-speaker scenarios, covering all possible in-domain (or seen) and out-of-domain (or unseen) cases in the training database is practically impossible. Previous studies also attempted to improve the adaptation capabilities of vocoders [SZL18], either with or without providing speaker information [LLJ⁺18, HTK⁺17]. However, these studies did not address the generalization capabilities for unseen out-of-domain data. In [LTDL⁺19], a potential universal vocoder was introduced, claiming that speaker encoding is not essential to train a high-quality neural vocoder.

Inspired by the performance and computational aspects of WaveRNN, we propose a novel approach for designing a universal WaveRNN vocoder. The proposed universal vocoder-speaker conditional WaveRNN (SC-WaveRNN) explores the effectiveness of explicit speaker information, i.e., speaker embeddings as a condition and improves the quality of generated speech across the broadest possible range of speakers without any adaptation or retraining. Even though conventional WaveRNN can model good temporal structure for a single speaker, it fails to capture the dynamics of multiple speakers. We have experimentally demonstrated that our proposed SC-WaveRNN overcomes such limitation by modelling temporal structure from a large data variability, making it possible to generate high-quality synthetic voices. Our work involves independent training of a speaker-discriminative neural encoder on a speaker verification (SV) task using a state-of-the-art generalized end-to-end loss [WWPM18]. The SV model, trained on a large amount of disjoint data, can attain robust speaker representations that are independent of channel conditions and capture a large space of speaker characteristics. Coupling such speaker information with speech synthesis training also reduces the need for ample high-quality multi-speaker training data. At the same time, it increases the model’s ability to generalize. Experimental results based on both objective and subjective evaluation confirm that the proposed method achieves better speaker similarity and perceptual speech quality than baseline WaveRNN in both seen and unseen speakers.

In parallel with the above-mentioned studies on the universal vocoder, there has been substantial development in multi-speaker TTS where the speaker encoder is jointly trained with TTS [CAS⁺18, PZPP19]. These jointly trained speaker encoders lead to poor inference performance when applied to data which are not included in the training dataset. Fine-tuning the pre-trained TTS model in combination with speaker embeddings was addressed in [DHS18, HMW⁺19, ACP⁺18]. Such approaches always require transcribed adaptation data and more computational time and resources to adapt to a new speaker. To overcome this, TTS models can be adapted from a few seconds of the target speaker’s voice in a zero-shot manner by solely using speaker embedding without retraining the entire model. [JZW⁺18, CCL⁺19, CLY⁺20].

Unfortunately, limitations still exist and human-level naturalness is not achieved yet. Additionally, prosody information was mismatched, especially for unseen speakers. To address those issues, we first train a multi-speaker Tacotron, which is conditioned on the speaker embeddings obtained from the independently trained speaker encoder. Tacotron [WSRS⁺17] is a sequence-to-sequence network which predicts mel-spectrograms from text. Next, we incorporate the proposed SC-WaveRNN as a vocoder using the same speaker encoder and synthesize the temporal waveform from the sequence of Tacotron’s mel-spectrograms. We compare our system with the baseline TTS method [CLY⁺20], which studies the effectiveness of several neural speaker embeddings in the context of zero-shot TTS. Our results demonstrate that the proposed zero-shot TTS system outperforms baseline zero-shot TTS in [CLY⁺20] in terms of both speech quality and speaker similarity on both seen and unseen conditions.

The chapter is organized as follows: In Section 2, the speaker encoder is explained. In Section 3, the details of Conditional WaveRNN are introduced. In Section 4, the experimental evaluations demonstrating the effectiveness of conditioning are presented. In Section 5, the implementation of zero-shot TTS is explained. Finally, the chapter is concluded in Section 6.

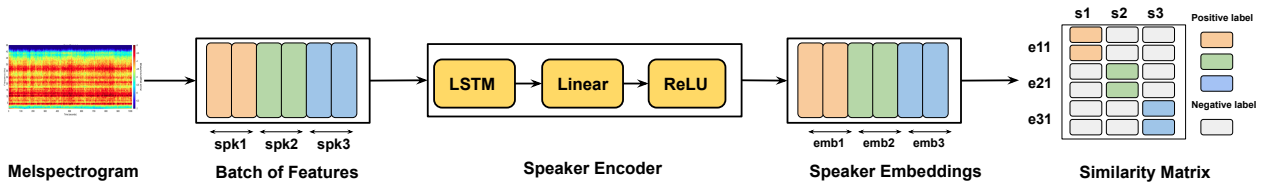


Figure 8.1: System overview of speaker encoder [WWPM18]. Features, speaker embeddings and similarity scores from different speakers are represented by different color codes. ‘spk’ denotes speakers and ‘emb’ represents embedding vectors.

8.2 Neural Speaker Encoder

Speaker verification (SV) refers to the process of determining whether an utterance belongs to a specific speaker by comparing it with that speaker’s known enrollment utterances. SV finds applications in Voice Match and similar systems. Speaker verification models typically fall into two main categories, depending on the constraints imposed on the enrollment and verification utterances: text-dependent speaker verification (TD-SV) and text-independent speaker verification (TI-SV). In TD-SV, both the enrollment and verification utterances have a phonetic constraint on their transcripts. This means that the spoken words or phrases in the utterances are predetermined and limited to a specific set of phonetic content. On the other hand, TI-SV does not impose any lexicon constraints on enrollment and verification utterance transcripts. As a result, there is a wider variability in terms of the phonemes used and the durations of the utterances.

In our work, we emphasize text-independent speaker verification (TI-SV) and its relevance in universal vocoders. We specifically focus on the significance of a speaker encoder within this context. We employ the generalized end-to-end (GE2E) approach for the speaker verification task, which has been trained on a large dataset containing thousands of speakers [WWPM18]. This trained model allows us to generate speaker embeddings using only a few seconds of reference speech from a target speaker without requiring any specific text or utterance constraints. By leveraging the GE2E model and the generated embeddings, we aim to enhance the capabilities of universal vocoders in capturing and reproducing the unique characteristics of individual speakers, thereby enabling more accurate and personalized speech synthesis.

The encoder network initially computes a frame-level feature representation and then summarizes these features to utterance-level fixed-dimensional speaker embeddings. Next, the classifier uses GE2E loss, where embeddings from the same speaker have high cosine similarity and embeddings from different speakers are far apart in the embedding space. As depicted in Fig. 8.2, Uniform Manifold Approximation and Projection (UMAP) [MHM18] shows that the speaker embeddings are perfectly separated with large inter-speaker distances and very small intra-speaker variance.

8.2.1 Training Encoder Network

The speaker encoder structure is depicted in Figure 8.1. The log mel-spectrograms are extracted from speech utterances of arbitrary window length. The feature vectors are then assembled in the form of a batch that contains S different speakers, and each speaker has U utterances. Each feature vector \mathbf{x}_{ij} ($1 \leq i \leq S$ and $1 \leq j \leq U$) represents the features extracted from speaker i utterance j . The features \mathbf{x}_{ij} are then passed to an encoder architecture. The final embedding vector \mathbf{e}_{ij} is L2 normalized and calculated by averaging each window separately.

A linear layer is attached to the last LSTM layer as an additional transformation to obtain network output. $f(\mathbf{x}_{ij}; \mathbf{w})$ where \mathbf{w} represents network parameters. The final embedding vector \mathbf{e}_{ij} is regularized by L2 normalization: $\mathbf{e}_{ij} = f(\mathbf{x}_{ij}; \mathbf{w}) / \|f(\mathbf{x}_{ij}; \mathbf{w})\|_2$. During inference, final embeddings are calculated by averaging each window separately.

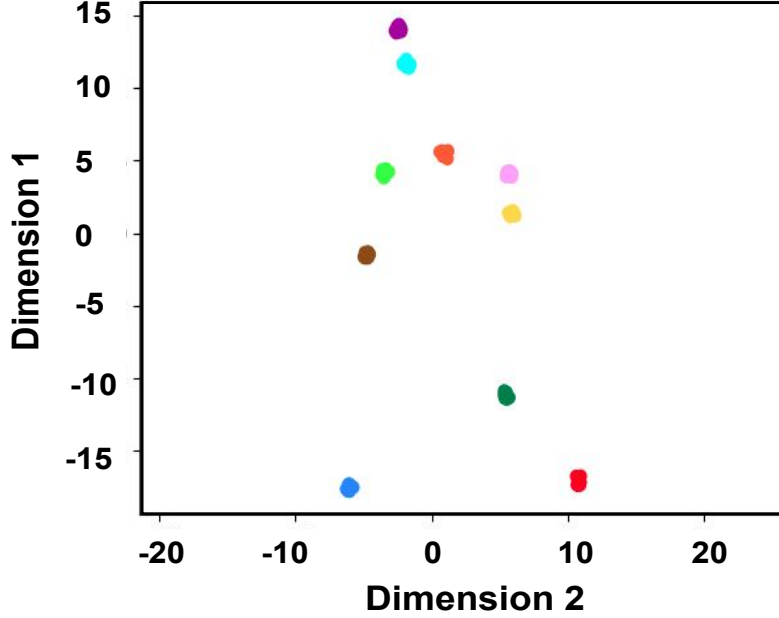


Figure 8.2: UMAP projection of 10 utterances for each of the 10 random speakers. Different colors represent different speakers.

8.2.2 Generalized End-to-End Loss

During training, the embedding of all utterances for a particular speaker should be close to the centroid of that particular speaker's embeddings while far from other speakers' centroids. The similarity matrix $\mathbf{SM}_{ij,k}$ is defined as the scaled cosine similarities between each embedding vector \mathbf{e}_{ij} to all speaker centroids \mathbf{c}_k ($1 \leq i, k \leq S$ and $1 \leq j \leq U$).

As depicted in Figure 8.1, the goal is to increase the similarity values of coloured areas and minimize the values of grey areas.

$$\mathbf{SM}_{ij,k} = \begin{cases} w \cdot \cos(\mathbf{e}_{ij}, \mathbf{c}_i^{-j}) + b & \text{if } k = i \\ w \cdot \cos(\mathbf{e}_{ij}, \mathbf{c}_k) + b & \text{otherwise} \end{cases}$$

$$\text{where } \mathbf{c}_i^{-j} = \frac{1}{U-1} \sum_{u=1; u \neq j}^U \mathbf{e}_{iu} \text{ and } \mathbf{c}_k = \frac{1}{U} \sum_{u=1}^U \mathbf{e}_{ku}$$

Here, w and b are trainable parameters. The ultimate GE2E loss L is the accumulative loss over similarity matrix ($1 \leq i \leq S$ and $1 \leq j \leq U$) on each embedding vector \mathbf{e}_{ij} :

$$L(\mathbf{x}; \mathbf{w}) = \sum_{i,j} L(\mathbf{e}_{ij}) = -\mathbf{SM}_{ij,i} + \log \sum_{k=1}^S \exp(\mathbf{SM}_{ij,k})$$

The use of the softmax function on the similarity matrix makes the output equal to 1 if $k = i$; otherwise,

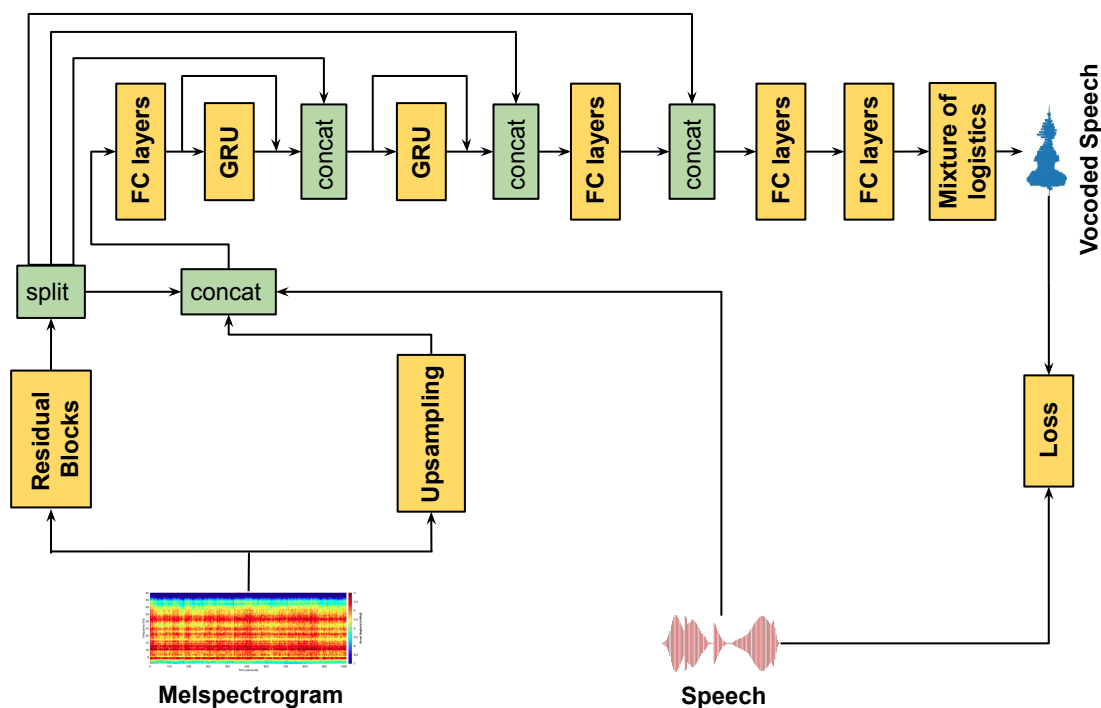


Figure 8.3: Block diagram of WaveRNN architecture.

it is 0.

8.3 Speaker-conditional WaveRNN

In literature, convolutional models have been thoroughly explored and achieved excellent performance in speech synthesis [ODZ⁺16, PPC19], yet they are prone to instabilities. A recurrent neural network (RNN) is expected to provide a more stable, high-quality speech due to the persistence of the hidden state.

8.3.1 Preliminaries

Our WaveRNN implementation is based on the repository¹ which is heavily inspired by WaveRNN training [KES⁺18]. This architecture combines residual blocks and an upsampling network, followed by GRU and FC layers, as depicted in Fig. 8.3. The architecture can be divided into two major networks: conditional and recurrent. The conditioning network consists of a pair of residual and upsampling networks with three scaling factors. At the input, we first map the acoustic features, i.e., mel-spectrograms, to a latent representation with the help of multiple residual blocks. The latent representation is then split into four parts, later fed as input to the recurrent network. The upsampling network is implemented to match the desired temporal size of the input signal. The outputs of these two convolutional networks, i.e., residual and upsampling networks, along with speech, are fed into the recurrent network. As part of the recurrent network, two uni-directional GRUs are employed with a few fully connected (FC) layers at the end. By design, the overhead complexity is reduced with fewer parameters, and the temporal context is taken advantage of for better prediction.

¹<https://github.com/fatchord/WaveRNN>

$$\text{waverrnn}(\mathbf{y}) = p(y_t|y_{t-1}; \mathbf{h}_t; \lambda)$$

where, p is probability function, \mathbf{h} is acoustic features from the conditioning network and λ is trainable network parameters.

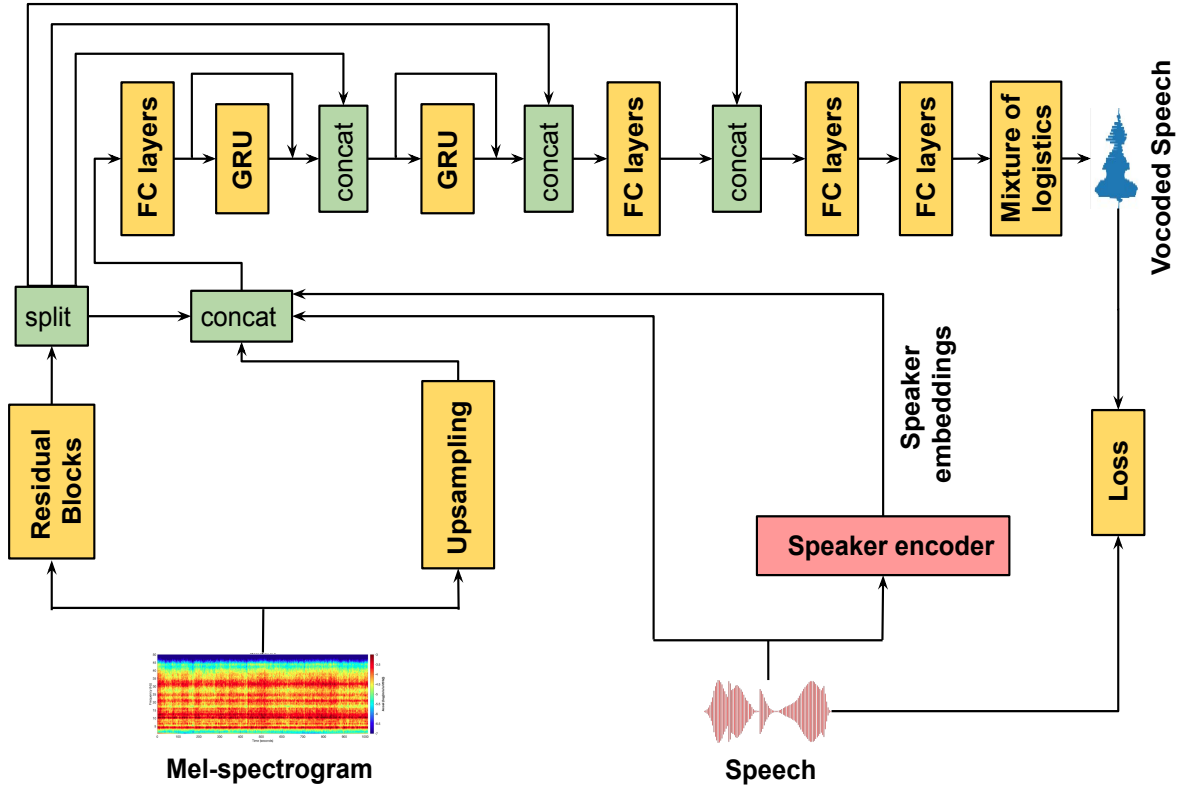


Figure 8.4: Block diagram of proposed SC-WaveRNN training.

8.3.2 Training WaveRNN with Speaker Embeddings

The above auto-regressive model can generate state-of-the-art, natural-sounding speech; however, it needs large amounts of training data to train a stable, high-quality model, and data scarcity remains a core issue. Moreover, a key challenge is its generalization ability. We observe degradation in speech quality and speaker similarity when the model generates waveforms from speakers that are not seen during training.

In order to assist the development of a stable universal vocoder and remove data dependency, we propose in this chapter an alternative module referred to as speaker conditional WaveRNN (SC-WaveRNN). In SC-WaveRNN, the output of the speaker encoder is used as additional information to control the speaker characteristics during both training and inference. The additional information is pivotal in generating more stable, high-quality speech across different speaker conditions. The direct estimation of raw audio waveform $\mathbf{y} = \{y_1, y_1, \dots, y_N\}$ is described by the conditional probability distribution:

$$\text{sc-waverrnn}(\mathbf{y}) = p(y_t|y_{t-1}; \mathbf{h}_t; \mathbf{e}; \lambda)$$

Where e is the 256 dimension speaker embeddings vector, the speaker encoder is independently trained using a large diversity of multi-speaker data that can generalize sufficiently to produce meaningful embeddings. The embedding vector e is computed utterance-wise. For each utterance, the final embedding vector is averaged over all frames, and hence, it is fixed for any utterance. The embedding vector is concatenated with the conditional network output and speech samples to form the conditional network. The details of the SC-WaveRNN algorithm are presented in Figure 8.4. In addition, we apply continuous univariate distribution constituting a mixture of logistic distributions [OLB⁺18b], which allows us to easily calculate the probability of the observed discretized value y . Finally, discretized mix logistic loss is applied to the discretized speech.

8.4 Zero-shot Text-to-Speech

Using the auxiliary speaker encoder enables us to propose a TTS system capable of generating high-fidelity synthetic voice for unseen speakers without retraining the Tacotron and vocoder. Such speaker adaptation to completely new speakers is called zero-shot. This speaker-aware TTS system mimics voice characteristics from a completely unseen speaker with only a few seconds of speech sample.

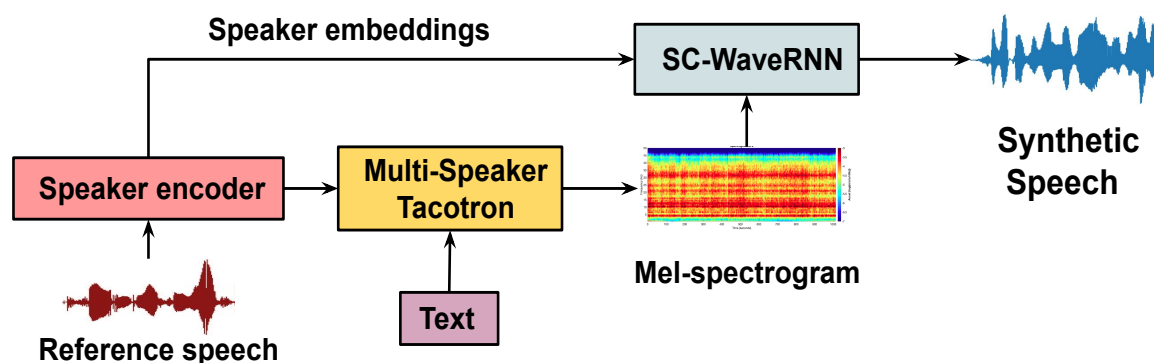


Figure 8.5: Block diagram of the proposed zero-shot TTS system.

Our proposed system is composed of three separately trained networks, illustrated in Figure 8.5: (a) a neural speaker encoder, based on GE2E training, (b) a multi-speaker Tacotron architecture [WSRS⁺17], which predicts a mel-spectrogram from text, conditioned on speaker embedding vector, and (c) the proposed speaker conditional WaveRNN, which converts the spectrogram into time domain waveforms. First, the speaker embeddings are extracted from each target speaker’s utterance using the speaker encoder. At each time step, the embedding vector for the target speaker is then concatenated with the embeddings of the characters before being fed into the encoder-decoder module. The final output is mel-spectrograms. To convert the predicted mel-spectrograms into audio, we use SC-WaveRNN, which is independently trained by conditioning on the additional speaker embeddings. Due to the generalization capabilities of the models, combining multi-speaker Tacotron with SC-WaveRNN can achieve efficient zero-shot adaptation for unseen speakers. We compare the proposed zero-shot system with a recently proposed zero-shot TTS [CLY⁺20] as the baseline system. There, the best-performing system uses a multi-speaker Tacotron with gender-dependent WaveNet vocoders as a TTS system and an x-vector with a learnable dictionary encoding as a speaker encoder network.

8.5 Experimental Setup

The speaker encoder training has been conducted on three public datasets: LibriSpeech, VoxCeleb1 and VoxCeleb2, containing utterances from over 8k speakers [JZW⁺18]. The log mel-spectrograms are first extracted from audio frames of width 25ms and step 10ms. Voice Activity Detection (VAD) and a sliding window approach is used. The GE2E model consists of 3 LSTM layers of 768 cells and a projection to 256 dimensions. While training, each batch contains $S = 64$ speakers and $U = 10$ utterances per speaker.

Tacotron and WaveRNN models are trained using VCTK English corpus [CJK16] from 109 different speakers. To evaluate generalization performance, we consider three scenarios: seen speakers-seen sound quality (SS-SSQ), unseen speakers-seen sound quality (UNS-SSQ) and unseen speakers-unseen sound quality (UNS-USQ). Seen speakers refer to the speakers that are already present in the training, and unseen speakers are the new speakers during testing. Sound quality refers to the recording conditions, such as recording equipment, reverberation, etc. We train the network using 100 speakers, leaving 9 speakers for UNS-SSQ scenarios that are chosen to be a mix of genders and have enough unique utterances per speaker. CMU-ARCTIC database [KB04a] is used for UNS-USQ scenario having 2 male and 2 female speakers. Moreover, to overcome the limited linguistic variability in VCTK data, we initially train the Tacotron model on the LJSpeech database as a “warm-start” training approach similar to [CLY⁺20]. Code and sound samples can be found in ².

8.6 Results and Discussion

8.6.1 Universal vocoder

In this section, we evaluate the performance of vocoded speech shown in Table 8.1. To assess the effectiveness of speaker embeddings in SC-WaveRNN, PESQ and STOI objective measures are computed from 50 random samples. We carry out evaluations on three conditions: SS-SSQ, UNS-SSQ and UNS-USQ. The purpose of each condition is to evaluate the proposed vocoder not only on seen or unseen speakers but also for the quality of the recordings. As expected, seen scenarios perform better with respect to unseen samples. However, we observe that SC-WaveRNN significantly improves both the objective scores when compared to baseline WaveRNN for all scenarios.

Table 8.1: Objective evaluation tests.

	SS-SSQ		UNS-SSQ		UNS-USQ	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
WaveRNN	2.2575	0.8173	2.1497	0.7586	1.4850	0.8620
SC-WaveRNN	2.7948	0.9049	2.8657	0.8984	1.8063	0.9195

Concerning the perceptual assessment of speech quality and speaker similarity, two separate listening tests are reported: mean opinion score (MOS) and ‘ABX’ preference test. The subjects are asked to rate the naturalness of generated utterances on a five-point scale (1:Bad, 2:Poor, 3:Fair, 4:Good, 5:Excellent). In the ‘ABX’ test, experimental subjects have to decide whether a given reference sentence ‘X’ is closer in speaker identity to one of ‘A’ and ‘B’ sentences, which are samples obtained either from the proposed

²<https://dipjyoti92.github.io/SC-WaveRNN/>

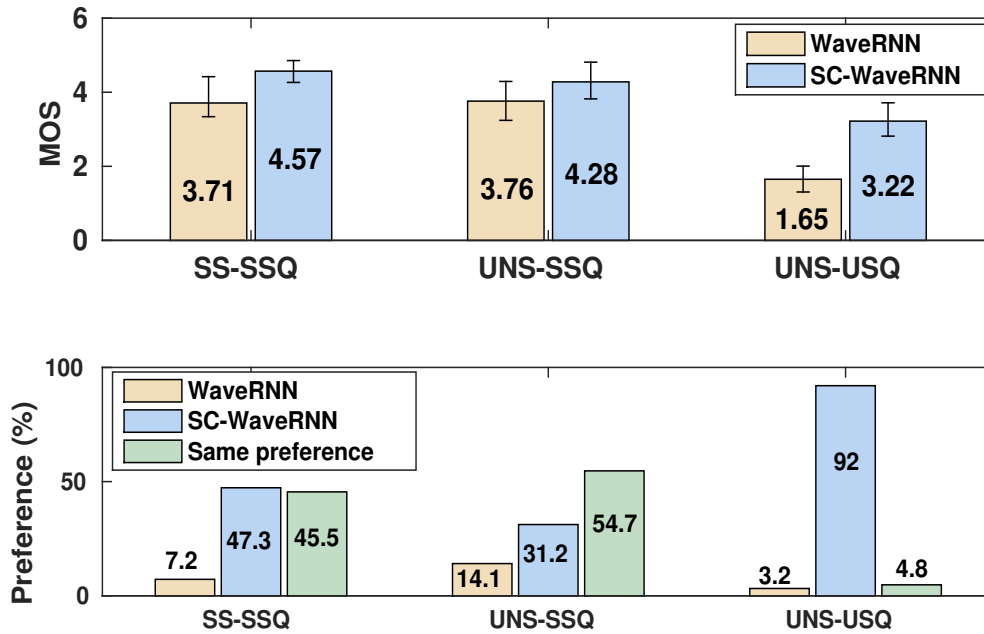


Figure 8.6: Vocoder Subjective listening test (MOS) for speech quality and preference test in (%) for speaker similarity.

or the baseline method, not necessarily in that order. Fifteen native and non-native English listeners participated in our listening tests. The evaluation results of both MOS and 'ABX' tests are demonstrated in Figure 8.6. Error bars represent 95% confidence intervals. For all seen and unseen scenarios, the MOS scores for the proposed SC-WaveRNN are much higher than the baseline WaveRNN (between 14% to 95% relative improvement). Under the same sound quality conditions (SS-SSQ and UNS-SSQ), although the proposed technique is preferred in terms of the speaker similarity preference test, most preference is given to the 'same preference' option, which indicates similar speaker characteristics for both methods. In contrast, experimental analysis shows a significant preference score (92%) in unseen sound quality for the proposed SC-WaveRNN. We conclude that additional speaker information in the form of embeddings is effective for improvements in naturalness and speaker similarity, especially for unseen data, and can achieve a truly universal vocoder. This is attributed to the fact that unseen scenarios are handled more efficiently by the model since additional embeddings are able to capture a broad spectrum of speaker characteristics. Moreover, SC-WaveRNN does not compromise the performance in seen conditions but also enhances generalization in unseen conditions.

8.6.2 Zero-shot TTS Synthesis

'MOS' and 'ABX' tests are employed to evaluate the proposed zero-shot TTS performance, as depicted in Figure 8.7. We subjectively evaluate both baseline [CLY+20] and our methods by synthesizing sample utterances from seen speakers and unseen speakers. Different sound qualities are not considered in the evaluation experiments of zero-shot TTS. As expected, a gap between seen and unseen speakers is visible: seen speakers' synthetic speech has a slightly higher quality than unseen speakers. MOS scores indicate that the proposed TTS is superior in quality, with 19.2% and 14.5% relative improvement for seen and unseen speakers, respectively. We also found that our proposed TTS mimic better speaker characteristics and significantly improves under both conditions. With regard to speaker similarity, the proposed TTS obtains the majority of preferences with 60% and 60.9% compared to 15.5% and 32.6% of the baseline

TTS for seen and unseen speakers, respectively.

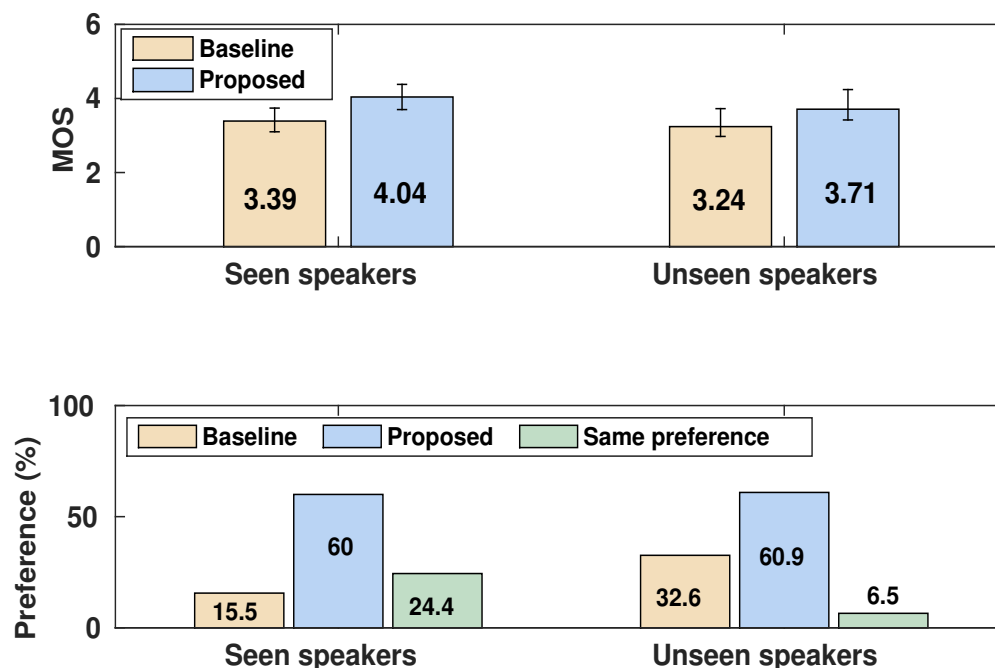


Figure 8.7: Zero-shot TTS Subjective listening test (MOS) for speech quality and preference test for (%) for speaker similarity.

8.7 Conclusions

In this chapter, we proposed a robust universal SC-WaveRNN vocoder that is capable of synthesizing high-quality speech. The system was conditioned on extracted speaker embeddings covering a diverse range of seen and unseen conditions. The main advantage of SC-WaveRNN is its high controllability since it improves multi-speaker vocoder training and better generalization ability by allowing reliable transfer to unseen speaker characteristics. Furthermore, speaker conditioning is typically more data efficient and computationally less expensive than training separate models for each speaker. The subjective and objective evaluation revealed that the proposed method generated higher sound quality and speaker similarity than the baseline method. In addition, we extended our approach to devising an efficient zero-shot TTS system. We demonstrated that the proposed zero-shot TTS with a universal vocoder could improve speaker similarity and the naturalness of synthetic speech for seen and unseen speakers.

Chapter 9

Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning based on Rényi Divergence Minimization

9.1 Introduction

Speech synthesis, which attracts a lot of attention in communication and voice interaction systems, aims to synthesize intelligible and high-quality speech signals which are indistinguishable from human recordings. The realization of a spoken utterance can be categorized into three principal components: the content, the speaker and the style component. The content component refers to the linguistic content of speech (what). The speaker characteristics are attributed to the speaker component (who). The definition of style component is associated with pitch variation and loudness (how). Style covers all aspects of speech that do not contribute to content information or the speaker’s identification.

Recently, the superiority of deep neural network (DNN) based speech synthesis surpassed the conventional speech synthesis models [WSRS⁺17, SPW⁺18, ACC⁺17, RHQ⁺20, vdODZ⁺16, KES⁺18]. Given a sufficient amount of training data, such TTS systems are capable of producing speech with superior quality, particularly for single-speaker synthesis. However, generated speech usually tends to be neutral and less expressive. Synthetic speech expressivity is also restricted because collecting labelled speech data, especially information describing prosody, is cumbersome due to concerns on cost, complexity and privacy, making unsupervised representation learning immensely popular.

Relating to expressive TTS, previous works aimed to transfer the style factor of a reference speech into the given text without prosody labels [HZW⁺18, SRBX⁺18]. In Global Style Token (GST), the reference speech is encoded into a fixed-length style embedding using a trainable style encoder that is conditioned along with content features and speaker embeddings in an unsupervised manner [WSZ⁺18]. Disentangling speech styles with a hierarchy of variational autoencoder (VAE) was introduced in [KW13, TPZK20]. Estimating mutual information using the Mutual-Information Neural Estimator (MINE) between the style and the content has been proposed in [HSTD20]. To improve the controllability in style modelling, fine-grained style transfer approaches were investigated in [LK19, KRRD19, DT20]. On the

other hand, to facilitate a semi-supervised approach, an auxiliary style classification task was proposed to capture style information from the reference utterances accurately [WLL⁺19, LYXX21]. The most recent studies in this direction also take into account the speaker identity [HZW⁺18, JZW⁺18, MMS19], which may be hard to extend for TTS models where only a few seconds of target voices are available. In [CLH⁺21, THZL21], the authors combined speaker and style embeddings to build an all-around TTS system.

A universal TTS synthesis system can generate speech from text with speaker characteristics and a speaking style similar to a given reference signal. The major challenge for universal TTS is speaker perturbation along with style transfer. The ultimate goal is to transplant prosody from arbitrary speakers, especially in the context of zero-shot, where only a few seconds of data is available. Towards this aim, we employ a universal TTS (UTTS) framework, which consists of four major components: content encoder, style encoder, speaker encoder and speech decoder. The content encoder generates a content embedding from the text. The style encoder represents the style factors in a style embedding, while the speaker encoder provides the speaker identity in the form of a speaker embedding. Finally, the speech decoder, conditioned on all the above embeddings, synthesizes the desired target speech.

When considering generalising the models with multiple speakers and multiple styles using just the reconstruction loss, performance unfortunately deteriorates. During training, content information is leaked into the style embeddings (“content leakage”) and speaker information into style embeddings (“style leakage”). Thus, at inference, when the reference speech has different content from the input text, the decoder expects the content from the style vector, ignoring some parts of the content text. Moreover, speaker information could be expected from the style encoder, leading to completely different speaker attributes.

To alleviate those issues, we suggest a novel Rényi Divergence based Disentangled Representation (RDDR) algorithm. The minimization of Rényi divergence becomes feasible via a variational representation formula that involves the cumulant generating function. We introduce two variations of this framework: Hellinger distance RDDR (H-RDDR) and sum of Rényi divergences RDDR (S-RDDR). Both variants are selected aiming to reduce the statistical variance of the adversarial component. Moreover, cumulants are preferred over expectations because they capture higher-order statistical information about the underlying distributions, which often leads to more stable training [PPF⁺20]. Similar to mutual information minimization where a lower bound of the Kullback-Leibler divergence is utilized, the proposed RDDR algorithm estimates, via neural network approximations, a lower bound of the Rényi divergence between the joint distribution and the product of the marginals of two pairs: content-style and speaker-style and then, minimize the estimated Rényi divergence in an adversarial manner. Rényi divergence minimization between those distributions pushes the various modalities to become independent. Our approach effectively disentangles content, style, and speaker information, not only alleviating leakage issues but also assisting the decoder in training on the proper data, leading to the synthesizing of high-quality speech. Our work involves independent training of a speaker-discriminative neural encoder to produce utterance-level speaker embeddings using a state-of-the-art generalized end-to-end loss [WWPM18]. Hence, the TTS system can synthesise speech from unseen speakers in a zero-shot manner. We train the style encoder using a set of trainable vectors, which are linearly combined using style factors generated from the input reference speech. Style tokens are trainable parameters that are optimized together with the TTS network parameters. Finally, due to low computational complexity, TransformerTTS is employed for the content encoder and speech decoder [LLL⁺19]. Experimental results based on both objective and subjective evaluation confirm that the proposed method achieves better style similarity and perceptual speech quality than the baseline TTS system, which is trained without a disentangled loss. Code and sound samples can be found in ¹.

¹<https://dipjyoti92.github.io/Universal-TTS/>

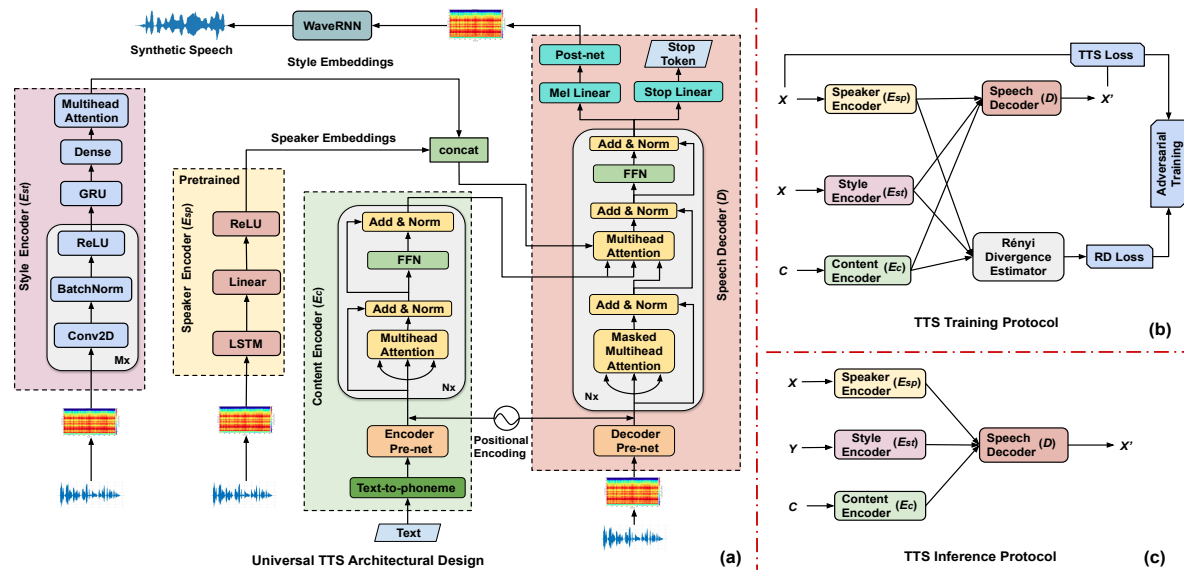


Figure 9.1: System overview of our universal TTS framework. (a) Universal TTS conditioned on speaker (E_{sp}) and style (E_{st}) encoders that can synthesise well-controllable speech. TransformerTTS is employed as a backbone TTS infrastructure. (b) The proposed training protocol considers a novel adversarial RDDR approach combined with minimising the TTS reconstruction loss. A reference utterance is used to extract speaker and style factors, whereas, during inference (c), the system may take any arbitrary speaker or style as input.

9.2 Universal TTS (UTTS)

Figure 1 shows an overview of the proposed universal TTS framework along with training and inference protocols. On the encoder side, we employ three different encoders with different architectures for content, style, and speaker, respectively. A speech decoder that is conditioned on all the above-mentioned latent embeddings produces the output speech features.

9.2.1 Speaker Encoder

The Speaker Encoder (E_{sp}) deploys generalized end-to-end (GE2E) loss that is trained on thousands of speakers [WWPM18]. The log mel-spectrograms are extracted from speech utterances of arbitrary length. The feature vectors are then assembled in the form of a batch that contains S different speakers, and each speaker has U utterances. Each feature vector f_{ij} ($1 \leq i \leq S$ and $1 \leq j \leq U$) represents the features extracted from speaker i and utterance j , respectively. The features are then passed to an encoder architecture. The final embedding vector is L^2 normalized, and they are calculated by averaging all window frames. The encoder consists of 3 LSTM layers of 768 cells and a projection to 256 dimensions, as depicted in the previous chapter. During training, the embedding of all utterances for a particular speaker should be closer to the centroid of that particular speaker’s embedding and, at the same time, far from other speakers’ centroid.

9.2.2 Style Encoder

The style encoder (E_{st}) is comprised of a convolutional stack, followed by a gated recurrent unit (GRU) similar to the GST-Tacotron paper [WSZ⁺18]. Mel spectrograms, which are extracted from the

reference speech, are passed to the stack of six 2D convolutional layers with kernel size 3×2 and 2×2 stride. The channel sizes of convolutional layers are (32, 32, 64, 64, 128, 128) followed by batch normalization and ReLU activation. The output from the last convolutional layer is summarized with a single-layer 128-unit unidirectional GRU. The style token layer is implemented with ten style token embeddings and a multi-head attention module [VSP⁺17]. As with speaker and content embeddings, the dimension of the style embeddings is 256. Finally, we apply tanh activation to GSTs before attention since it leads to greater token diversity. The overall style encoder is jointly trained with the entire TTS model without using prosodic labels.

9.2.3 TTS Module

Given its performance and computational gains, we implemented TransformerTTS as our backbone TTS [LLL⁺19]. Here, the multi-head attention mechanism constructs the hidden states for the encoder and the decoder in parallel, which improves training efficiency. In addition to this backbone TransformerTTS model, a style encoder E_{st} and a speaker encoder E_{sp} is introduced to construct a truly universal TTS, also depicted in 9.1(a).

The TTS module converts textual, style and speaker information into acoustic features with the help of the encoder-decoder paradigm. While the final module is a vocoder, WaveRNN [KES⁺18] in our case generates speech waveform from the previously generated acoustic information. The first stage of the TTS module is the conversion from text to phonemes. The text-to-phoneme converter not only assists the model in training on the vast majority of cases but also resolves cases where some letters can be pronounced differently under different contexts, leading to major performance degradation when data are insufficient. Given a set of speech and phoneme content pairs (\mathbf{x}, \mathbf{c}) , the baseline UTTS minimizes the feature-domain reconstruction loss between the predicted output of the speech decoder D and the original speech,

$$\mathcal{L}_{tts} = \min_{E_{st}, E_c, D} \| D(E_c(\mathbf{c}), E_{st}(\mathbf{x}), E_{sp}(\mathbf{x})) - \mathbf{x} \|_1 \quad (9.1)$$

where $\| \cdot \|_1$ is the L^1 norm. We are not optimizing E_{sp} weights due to the fact that it is already pre-trained on thousands of speakers. Therefore, the speaker embeddings should reflect a well-balanced speaker universe. TTS module is first pre-trained with LJSpeech, which has a broad range of linguistic variability, and then we freeze E_c for the remaining training process.

9.3 Proposed Disentangled Representation

Although baseline UTTS tries to synthesize speech using content, style and speaker factors, training just on an L^1 reconstruction loss is not enough. The style embeddings still manage to carry non-style information, leading to content leakage and style leakage. To efficiently decouple all representations without explicit labels, we estimate and minimize the Rényi divergence (RD) between their embedding representation pairs $(E_c(\mathbf{c}), E_{st}(\mathbf{x}))$ and $(E_{sp}(\mathbf{x}), E_{st}(\mathbf{x}))$. The overall training and inference protocols are demonstrated in 9.1 (c) & (d).

9.3.1 Preliminaries

The standard approach to disentangling two modalities is through Mutual Information (MI) minimization since zero MI implies Independence. MI is the Kullback-Leibler (KL) divergence, and it can be represented in the form of Donsker-Varadhan representation [DV83]. Given two random variable \mathbf{X} and

\mathbf{Y} , MI i.e., $\mathcal{I}(\mathbf{X}, \mathbf{Y})$ is equivalent to KL divergence between the joint distribution, $\mathbb{P}_{\mathbf{XY}}$ and the product of marginals, $\mathbb{P}_{\mathbf{X}} \otimes \mathbb{P}_{\mathbf{Y}}$. MINE [BBR⁺18a] estimate a lower bound of MI:

$$\mathcal{I}(\mathbf{X}, \mathbf{Y}) \geq \mathcal{I}_{\Theta}(\mathbf{X}, \mathbf{Y}) = \sup_{\theta \in \Theta} \{ \mathbb{E}_{\mathbb{P}_{\mathbf{XY}}} [T_{\theta}] - \log(\mathbb{E}_{\mathbb{P}_{\mathbf{X}} \otimes \mathbb{P}_{\mathbf{Y}}} [e^{T_{\theta}}]) \} \quad (9.2)$$

where T_{θ} is an NN-based parametrization of the space of all continuous and bounded functions. For the proposed UTTS, T_{θ} is parametrized by three fully connected layers, each layer followed by ReLU, with parameters $\theta \in \Theta$ while the optimization is performed through stochastic gradient descent.

We introduce this approach to disentangle content, style and speaker representations in UTTS, similar to [HSTD20]. We use two separate neural estimator T_{θ} and $T'_{\theta'}$ to approximate MI from the pairs $(E_c(\mathbf{c}), E_{st}(\mathbf{x}))$ and $(E_{st}(\mathbf{x}), E_{sp}(\mathbf{x}))$, respectively. By minimizing MI, we force the encoders to learn information that is independent of each other. Thus, the overall objective function is a min-max problem where we seek to maximize the lower-bound of MI w.r.t. T_{θ} and $T'_{\theta'}$ and minimize the MI and the reconstruction loss w.r.t. E_{st} and D .

$$\begin{aligned} \mathcal{L} = & \min_{E_{st}, D} \max_{T_{\theta}, T'_{\theta'}} \{ \| D(E_c(\mathbf{c}), E_{st}(\mathbf{x}), E_{sp}(\mathbf{x})) - \mathbf{x} \|_1 \\ & + \lambda \max(0, \mathcal{I}_{\Theta}(E_c(\mathbf{c}), E_{st}(\mathbf{x}))) + \lambda \max(0, \mathcal{I}_{\Theta'}(E_{st}(\mathbf{x}), E_{sp}(\mathbf{x}))) \} \end{aligned} \quad (9.3)$$

where λ is a hyperparameter, set to 0.1. We also bound from below the estimated MI to non-negative values.

9.3.2 Rényi Divergence based Disentangled Representation

Learning representative latent embeddings is a challenging problem. In order to decouple all the information factors properly, it is necessary to estimate the MI between the embedding pairs. However, it has been demonstrated that the statistical variance of the finite-sampling MI estimator can be exponentially high [SE20, MS20], often resulting in inferior estimation performance. In this chapter, we propose a different family of information-theoretic divergences to reduce estimator’s variance. We present two alternatives based on the Rényi divergence family for disentangled speech representation learning. The proposed algorithm is presented in 9.2.

The cumulant generating function (CGF), also known as the log-moment generating function, is defined for a random variable with pdf $p(X)$ as $\Lambda_{f,p}(\beta) = \log \mathbb{E}_p[e^{\beta f(X)}]$, where f is a measurable function with respect to p . RDDR employs the expectation of the cumulant loss function and substitutes it in the loss function of MINE in Equation (9.3.1).

Given two random variables \mathbf{X} and \mathbf{Y} , RDDR employs a DNN (i.e., T_{θ}) to approximate the maximum lower bound of the Rényi divergence (RD) variational formula which is defined via two cumulant generating functions (CGFs):

$$\mathcal{R}_{\beta, \gamma}(\mathbf{X}, \mathbf{Y}) \geq \sup_{\theta \in \Theta} -\frac{1}{\beta} \mathbb{E}_{\mathbb{P}_{\mathbf{XY}}} [e^{-\beta T_{\theta}}] - \frac{1}{\gamma} \log(\mathbb{E}_{\mathbb{P}_{\mathbf{X}} \otimes \mathbb{P}_{\mathbf{Y}}} [e^{\gamma T_{\theta}}]) \quad (9.4)$$

Where hyper-parameters β and γ are two non-zero real numbers which control the learning dynamics,

Algorithm: Pseudo-code for proposed RDDR training

Input: Speech and text pairs $\langle \mathbf{x}_i, \mathbf{c}_i \rangle$.
Pre-training: Optimize E_c, D on LJSpeech using
 $\min_{E_c, E_{st}, D} \sum_i \| D(E_c(\mathbf{c}_i), E_{st}(\mathbf{x}_i), E_{sp}(\mathbf{x}_i)) - \mathbf{x}_i \|_1$
 $E_{sp} \leftarrow$ GE2E training
 $E_{st}, T_\theta, T'_{\theta'} \leftarrow$ initialization with random weights
 $E_{st}, D, T_\theta, T'_{\theta'}$ not converged Sample mini-batch from $\langle \mathbf{x}_i, \mathbf{c}_i \rangle; i = \{1, 2, \dots, b\}$
 $\{\mathbf{p}_i\} \leftarrow \{E_c(\mathbf{c}_i) | i = 1, 2, \dots, b\}$
 $\{\mathbf{q}_i\} \leftarrow \{E_{st}(\mathbf{x}_i) | i = 1, 2, \dots, b\}$
 $\{\mathbf{r}_i\} \leftarrow \{E_{sp}(\mathbf{x}_i) | i = 1, 2, \dots, b\}$
 $\{\hat{\mathbf{p}}_i\}, \{\hat{\mathbf{r}}_i\} \leftarrow$ random permutation of $\{\mathbf{p}_i\}, \{\mathbf{r}_i\}$
 $\mathcal{L}_{RD^1} = \sum_k \left[-\frac{1}{\beta_k} \log \frac{1}{b} \sum_{i=1}^b e^{-\beta_k T_\theta(\mathbf{p}_i, \mathbf{q}_i)} \right.$
 $\left. - \frac{1}{\gamma_k} \log \frac{1}{b} \sum_{i=1}^b e^{\gamma_k T_\theta(\hat{\mathbf{p}}_i, \mathbf{q}_i)} \right]$
 $\mathcal{L}_{RD^2} = \sum_k \left[-\frac{1}{\beta_k} \log \frac{1}{b} \sum_{i=1}^b e^{-\beta_k T'_{\theta'}(\mathbf{r}_i, \mathbf{q}_i)} \right.$
 $\left. - \frac{1}{\gamma_k} \log \frac{1}{b} \sum_{i=1}^b e^{\gamma_k T'_{\theta'}(\hat{\mathbf{r}}_i, \mathbf{q}_i)} \right]$
The overall objective function:
 $\mathcal{L} = \frac{1}{b} \sum_{i=1}^b \| D(\mathbf{p}_i, \mathbf{q}_i, \mathbf{r}_i) - \mathbf{x}_i \|_1$
 $+ \lambda \max(0, \mathcal{L}_{RD^1}) + \lambda \max(0, \mathcal{L}_{RD^2})$
 $D = D - \epsilon \nabla_D \mathcal{L}; E_{st} = E_{st} - \epsilon \nabla_{E_{st}} \mathcal{L}$
 $T_\theta = T_\theta + \epsilon \nabla_D \mathcal{L}_{RD^1}; T'_{\theta'} = T'_{\theta'} + \epsilon \nabla_D \mathcal{L}_{RD^2}$

Figure 9.2: Training algorithm of RDDR.

the use of CGFs allows for an inclusive characterization of the distributions' statistics, making it possible for T_θ to enforce independence better. This, in turn, leverages improved disentanglement representation. The proposed algorithm can be interpreted for several choices of its hyper-parameters. Thus, the optimization of the proposed loss function is equivalent to the minimization of divergence for a wide set of hyperparameter values. For our experiments, we choose two variations. First, $(\beta, \gamma) = (0.5, 0.5)$ which is equivalent to the minimization of Hellinger distance, we refer to this as Hellinger RDDR (H-RDDR). Internal numerical simulations conducted by our group have shown that the variance of the estimated Hellinger distance is significantly smaller than the variance of the estimated KL divergence. Second, we choose a combination of $\beta = [0, 0.5, 1]$ and $\gamma = [1, 0.5, 0]$, which is equivalent to the minimization of the sum of Rényi divergences. This variant is called the sum of RDDR (S-RDDR). This particular choice for the hyper-parameters tries to optimize KL divergence and reverse KL and Hellinger distance simultaneously. By doing so, we anticipate enhanced independence between the speech factors. Similar to Equation (9.3), we can formulate the overall objective function through adversarial training.

9.4 Results and Discussion

The speaker encoder training has been conducted on LibriSpeech, VoxCeleb1 and VoxCeleb2 datasets containing utterances from over 8k speakers [PPS20]. TransformerTTS and WaveRNN models are trained using VCTK English corpus [VYM⁺16] from 109 different speakers. We initially trained the TransformerTTS model on the LJSpeech database [Kei17], which contains 13,100 audio clips from a single speaker, as a “warm-start” approach.

Table 9.1: Objective evaluation tests. Lower scores indicate better performance.

	No Shuffle			Shuffle		
	RMSE-F0	MCD	WER(%)	RMSE-F0	MCD	WER(%)
UTTS	29.80	5.28	22.7	47.02	6.56	32.1
UTTS MINE	30.33	5.38	25.4	47.26	6.59	31.9
UTTS S-RDDR	28.59	5.35	21.6	45.75	6.39	28.7
UTTS H-RDDR	28.59	5.26	18.3	47.26	6.59	26.6

9.4.1 Objective Evaluation

In this section, we evaluate the performance of disentanglement strategies shown in Table 1. To assess the effectiveness of the proposed methods, we calculated three performance scores from 100 random samples. Mel-cepstral distortion (MCD) measures the spectral distance between the synthesized and reference mel-spectrum features. Root mean squared error (RMSE) evaluates the similarity in F0 modelling between reference and synthesized speech. Lastly, word error rate (WER) evaluates the content preservation criterion. During inference, we evaluate the performance on two conditions: ‘no shuffle’ and ‘shuffle’. No shuffle feeds same reference speech \mathbf{x}_i into style and speaker encoders and its corresponding text \mathbf{c}_i to predict the speech features \mathbf{x}' with the decoder. Whereas, shuffle feeds speech \mathbf{x}_i into speaker, \mathbf{x}_j into style and \mathbf{c}_k into the content encoder given ($i \neq j \neq k$). We observe that the proposed S-RDDR and H-RDDR algorithms outperform both the baseline and MINE approaches regarding RMSE-F0 and MCD evaluation metrics with relative improvements. As expected, no shuffle scenarios perform better with respect to shuffled samples. Furthermore, one of the main objectives of the RDDR algorithm is to improve the content leakage of the generated speech, which we objectively measure using Google’s open-source automatic speech recognizer (ASR) [asr]. For the VCTK dataset, ASR achieves a WER of 14.6% on the held-out real data. Although both RDDR variants perform better, the performance of H-RDDR is the best so far, with WER of 18.3% and 26.6% for no shuffle and shuffle scenarios, respectively. We overall conclude that the disentanglement module during training assists the TTS in achieving a more accurate rendering of prosodic patterns as well as synthesizing proper speech content to its corresponding text without any significant leakage issues.

Table 9.2: Average cosine-similarity evaluation.

Methods	Baseline	MINE	S-RDDR	H-RDDR
No Shuffle	0.828	0.840	0.836	0.839
Shuffle	0.734	0.732	0.737	0.739

Next, we employ cosine similarity as a speaker similarity measure between the generated and reference speaker’s speech. As shown in Table 2, cosine similarity does not vary much across different systems. This is attributed to the fact that speaker embeddings are pre-trained and do not jointly train with the TTS module. Therefore, different TTS modules perform equally better, and speaker identities are much closer to reference speakers.

Table 9.3: MOS scores (95% confidence interval) of audio quality and speaking style similarity for different TTS modules.

Methods	MOS-Speech-Quality	MOS-Style-Similarity
UTTS	3.01 ± 0.05	2.98 ± 0.08
UTTS MINE	3.11 ± 0.06	2.92 ± 0.08
UTTS S-RDDR	3.62 ± 0.05	3.41 ± 0.07
UTTS H-RDDR	3.51 ± 0.06	3.36 ± 0.07

9.4.2 Subjective Evaluation

We conduct listening tests to evaluate different TTS modules and the choice of RDDR approach, as depicted in Table 3. Twenty native and non-native English listeners participated in our listening tests. We conducted two separate mean opinion score (MOS) listening tests, and subjects were asked to rate the synthesized speech on a scale of five (1:Bad, 2:Poor, 3:Fair, 4:Good, 5:Excellent). MOS-Speech-Quality (MSQ) assesses the perceptual speech quality, whereas MOS-Style-Similarity (MSS) evaluates the speaking style expressiveness w.r.t. the reference style. MSQ scores indicate that compared to UTTS, the proposed S-RDDR and H-RDDR UTTSs are superior in quality, with 20.3% and 16.6% relative improvement, respectively. We also found that our proposed TTS mimic better style characteristics than the baseline and shows significant relative improvement of 14.4% and 12.7% for S-RDDR and H-RDDR, respectively. Results indicate that disentanglement helps the system to properly learn all the information factors relating to content, speaker and style. Therefore, it enhances the speech decoder’s ability to synthesize high-quality speech and also preserves the style of the reference speech better. We did not conduct listening tests for speaker similarity as objective results clearly indicate equal performance for all TTS modules.

9.4.3 Disentangled Representation Learning using variance reduction method

An important application of MI is disentangled representation learning. In the context of representation disentanglement, the extraction of meaningful latent features for high-dimensional data is challenging, especially when explicit knowledge needs to be distilled into interpretable representations. One popular approach to enforce representation disentanglement is via MI minimization. Moreover, a superior disentanglement will allow a greater degree of interpretability and controllability, especially for generative models maintaining high production capacity. In this section, we employ the proposed DNE- VP_λ estimator for MI estimation in order to learn disentangled representation and, particularly, in the context of speech synthesis and analysis.

A universal text-to-speech synthesizer can generate speech from text with a speaker factor and speaking style similar to a reference signal. Previous works aimed to encode the information from reference speech into a fixed-length style and speaker embedding using trainable encoders [WSZ⁺18, TPZK20, CLH⁺21, THZL21]. The major challenges for such speech synthesizers are controllability and generalisability, especially when trying to generalize the models with multiple speakers and multiple styles. During training, content information is leaked into the style embeddings (“content leakage”) and speaker information into style embeddings (“style leakage”). Thus, at inference, when the reference speech has content different from the input text, the decoder expects the content to be from the style vector, ignoring some parts of the content text. Moreover, speaker information could be expected from the style

encoder, leading to completely different speaker attributes. To alleviate that, [PMPS21] suggested replacing the KL-based MI with Rényi-based MI and minimizing the Rényi divergence between the joint distribution and the product of marginals for the content-style and style-speaker pairs. However, reliable estimation of Rényi divergence was problematic due to high statistical variance. Taking advantage of the proposed variance reduction technique, we employ a VP term in the loss function, which is denoted as $\text{DNE-VP}_\lambda(R_\alpha)$. By doing so, content, style, and speaker spaces become representative and (ideally) independent of each other. We introduce two variations of this framework: sum of three Rényi divergences $\text{DNE}(R_{\alpha=0} + R_{\alpha=0.5} + R_{\alpha=1})$ (i.e., sum of the corresponding objective functionals) and $\text{DNE}(R_{\alpha=0.5})$. We tested several different λ values, aiming to reduce the statistical variance of the adversarial component. Notice that larger λ values were helpful in this application.

Table 9.4: Objective evaluation tests. Lower scores indicate better performance.

	No Shuffle			Shuffle		
	RMSE-F0	MCD	WER(%)	RMSE-F0	MCD	WER(%)
$\text{DNE}(R_{\alpha=0} + R_{\alpha=0.5} + R_{\alpha=1})$	28.59	5.35	21.6	45.75	6.39	28.7
$\text{DNE}(R_{\alpha=0.5})$	28.59	5.27	18.3	47.26	6.60	26.6
$\text{DNE-VP}_{\lambda=5}(R_{\alpha=0} + R_{\alpha=0.5} + R_{\alpha=1})$	30.29	5.23	21.2	48.15	6.39	27.3
$\text{DNE-VP}_{\lambda=10}(R_{\alpha=0} + R_{\alpha=0.5} + R_{\alpha=1})$	27.76	5.36	18.1	47.62	6.48	28.7
$\text{DNE-VP}_{\lambda=5}(R_{\alpha=0.5})$	28.69	5.87	17.3	46.53	6.72	25.4
$\text{DNE-VP}_{\lambda=10}(R_{\alpha=0.5})$	29.71	5.33	22.8	45.47	6.54	26.2

We evaluate the performance of disentanglement strategies using three performance scores from 100 random samples shown in Table 9.4. Mel-cepstral distortion (MCD) measures the spectral distance between the synthesized and reference mel-spectrum features. Root mean squared error (RMSE) evaluates the similarity in F0 modelling between reference and synthesized speech. Lastly, word error rate (WER) evaluates the content preservation criterion. During inference, we evaluate the performance on two conditions: ‘no shuffle’ and ‘shuffle’. During inference, ‘no shuffle’ feeds the same reference speech into style and speaker encoders and its corresponding text to predict the speech features, whereas ‘shuffle’ feeds random speech. We observe that the proposed DNE-VP_λ variants outperform baseline approaches without VP regarding all evaluation metrics. Our proposed systems greatly reduced content leakage by improving the word error rate by approximately 5-18% relative to the baseline systems. Furthermore, RMSE-F0 and MCD scores show that the disentanglement module during training assists the TTS in achieving a more accurate rendering of prosodic patterns and synthesizing proper speech content to its corresponding text without any significant leakage issues.

9.5 Conclusions

We proposed a novel disentangled representation by exploiting cumulant-generating functions in speech synthesis. Our system approximates and then minimizes the Rényi divergence between content-style and style-speaker pairs, and it is jointly trained with TTS reconstruction loss in an adversarial manner. The subjective and objective evaluation revealed that the proposed approach outperforms both the baseline and my algorithm and is able to eliminate the issues of content and style leakage, resulting in a truly universal TTS system. The main advantage of universal TTS is its high controllability since it improves multi-speaker multi-style training along with better generalization ability by allowing reliable transfer to speaker and style information.

Furthermore, reliable estimation of Rényi divergence faced challenges due to high statistical variance.

A VP term was introduced in the loss function to overcome this. The aim was to achieve independence between content, style, and speaker spaces. Various λ values were experimented with to reduce the statistical variance of the adversarial component, with larger λ values proving beneficial in this context.

Chapter 10

Enhancing Speech Intelligibility in TTS using Speaking Style Conversion

10.1 Introduction

Over the years, text-to-speech (TTS) systems have become more prevalent, and they have a substantial range of applications, including personal voice assistants, public address systems, and navigation devices. In a quiet environment, the intelligibility of synthetic speech corresponds to that of natural speech. However, the intelligibility typically falls below the natural speech level in noisy conditions [CMVB⁺13c]. Listeners in real-world scenarios often hear speech in noisy surroundings where the intelligibility of synthetic speech is also compromised. Therefore, highly efficient TTS systems which are able to simulate Lombard effect and make the speech more intelligible are essential for the end listeners. Such speaking style conversion retains the linguistic and speaker-specific information of the original speech.

Few studies have explicitly adapted Lombard speech onto speech synthesis models by focusing on articulatory effort changes [RSVA11, PDD14]. Previously, the majority of such studies were conducted using hidden Markov model (HMM)-based statistical parametric speech synthesis (SPSS) due to its superior adaptation abilities and flexibility. The HMM model trained on normal speech was then adapted using a small amount of Lombard speech and improvements were shown under different noisy conditions [CMVB⁺13c]. Yet, these approaches were limited to poor acoustic modelling and the inability to synthesize high-fidelity speech samples. To overcome this, deep neural network approaches were implemented where the robustness of acoustic modelling is improved by efficient mapping between linguistic and acoustic features. Inspired by the success of adversarial generative models, Cycle-consistent adversarial networks (CycleGANs) showed promising results in terms of speech quality and the magnitude of the perceptual change between speech styles [SJY⁺19, SJA⁺19]. An extension to recurrent neural networks and particularly long short-term memory networks (LSTMs) was proposed that it successfully adapted normal speaking style to Lombard style [BAA17]. In [BJA⁺19b], the authors demonstrated results with sequence-to-sequence (seq2seq) TTS models along with the recently proposed Wavenet vocoder, where the audio samples are generated through a non-linear autoregressive manner. Along with different adaptation approaches, various TTS vocoders are compared in the context of style transfer and assessment was performed in terms of speaking style similarity and speech intelligibility [SJRA19, BJA⁺19a].

A sizable amount of training data is required to train a TTS system with Lombard style. However, the collection of a large portion of Lombard speech is difficult. Such data sparsity limits the usage of typical

data-driven approaches similar to the recent end-to-end TTS systems. Our work considers the use of speaking style adaptation techniques leveraging on large quantities of widely available normal speech data referred to as transfer learning. It assumes the prior knowledge from a previous model trained with large variations in linguistic and acoustic information. It adapts to the target styles even with a limited amount of data. In the literature, most of the vocoders for style transfer in TTS systems are either source-filter based models or convolutional models [SJRA19, BJA⁺19b]. However, such techniques are limited by their inefficiency both in modelling proper acoustic parameters and in the computational complexity of sample generation. Inspired by the performance and computational aspects of recurrent neural networks, in this work, we employ WaveRNN as a vocoder [KES⁺18], which generates speech samples from acoustic features, i.e., mel-spectrograms. Experimental results indicate that WaveRNN is capable of adapting to appropriate target speech styles and able to provide more stable high-quality speech samples. To generate the mel-spectrograms from text, we utilize a popular architecture Tacotron, a seq2seq encoder–decoder neural network with attention mechanism [WSRS⁺17].

Improvement of speech intelligibility in noise can also be achieved by signal processing techniques such as amplitude compression [NG76], changes in spectral tilts [LC09], formant sharpening and dynamic range compression [ZS14]. The method Spectral Shaping and Dynamic Range Compression (SSDRC) has been shown to provide high intelligibility gains in various noisy conditions by redistributing signal energy on time-frequency information [ZKS12]. In [VBYKS13], the best performing method was achieved by applying additional processing, i.e., dynamic range compression after generating Lombard-style adapted TTS. The results, however, failed to increase the intelligibility under competing-speaker noise. In order to develop a highly intelligible communication system and restrict the latency imposed by additional processing after the TTS synthesis. Here, we implement Lombard-SSDRC TTS where the TTS model is trained with Lombard speech processed through the SSDRC algorithm. Hence, we combine the advantages of naturally-modified Lombardness with speech enhancement strategies in the frequency domain (spectral shaping) and in the time domain (dynamic range compression) into an intelligibility-enhanced TTS synthesis system.

10.2 Factors defining speech intelligibility

The understanding of speech is influenced by various factors, including the clarity of articulation, the sensitivity of the listener’s ear, the language proficiency of both the speaker and the listener, and the quality of the communication environment. When speakers perceive a loss of intelligibility, they adjust their articulation to facilitate communication.

Two main categories of speech production changes have been identified: (1) those based on the interlocutor, such as foreign-directed speech (FDS) or machine-directed speech (MDS), and (2) those based on the communication environment, such as Lombard speech (LS) or speech addressed to a distant listener. Each category exhibits unique features. For instance, LS is produced with increased vocal effort to maximize speech audibility in noisy conditions where clarity may be compromised, while FDS often demonstrates reduced lexical variability and clearer intonation patterns [UKB07, FAFZ12]. However, the boundaries between these categories are not absolute, as they share many acoustic and phonetic features.

Although there is intra-speaker variability within each speaking style, previous studies have reported clear acoustic and phonetic changes associated with individual production changes. This section provides a brief summary of the observed acoustic-phonetic changes in both interlocutor-induced and environment-induced modifications and their impact on listener intelligibility.

Lombard styles exhibit a decreased speaking rate, resulting in longer phone durations than casual or

normal speech [PDB86, Lu09]. The decrease in speaking rate is also accompanied by an increase in the number of pauses and the duration of various sound segments. Bradlow et al. [BKH00] observed an overall increase in sentence duration of 51% and 116% for male and female speakers, respectively when transitioning from casual to clear speech. Cooke et al. [CMV14] conducted experiments by artificially modifying speech durations and evaluated their impact on listener intelligibility, but no clear benefits were observed with durational modifications. Nevertheless, speaking slowly with appropriate pauses can give listeners sufficient time to process and comprehend the message effectively.

Analysis of the energy difference between consonants and vowels has shown an increased consonant-to-vowel energy ratio (CVR) in clear speech compared to casual speech [BKH03]. Hazan and Markham [HM04] investigated the correlation between CVR and intelligibility and found no significant correlation between word intelligibility and CVR for nasals, fricatives, and stop consonants in naturally produced speech. However, studies have demonstrated that enhancing consonant energy in words, consonant-vowel syllables (CV), and vowel-consonant-vowel (VCV) syllables improves consonant identification for both normal hearing listeners [HS98, GS86] and hearing-impaired listeners [GS87, ME88]. Building on these findings, Skowronski and Harris [SH06] performed energy redistribution from vowels to consonants, which was shown to improve speech intelligibility. Similarly, Godoy and Stylianou [GS12] evaluated the contributions of voiced and unvoiced regions to Lombard intelligibility and found that the increase in Lombard intelligibility is primarily attributed to vowel segments of the speech signal.

Short-term spectral analysis (STSA) allows the examination of the frequency information of speech at different time points. STSA has revealed that vowel sounds in clear speech exhibit higher spectral prominences than in casual speech.

Based on this observation, Krause [Kra01] amplified the magnitudes of the first and second formants in segments of casual speech to match the spectral characteristics of clear speech. This formant sharpening technique enhanced the intelligibility of casual speech for normal-hearing listeners but had limited benefits for the hearing impaired [Kra01].

Further analysis of the spectral differences between casual and clear speech was conducted in [KS14]. A mixed-filtering technique was developed to isolate the information from clear speech and incorporate it into casual speech to improve intelligibility, which proved beneficial for enhancing intelligibility in noisy environments.

The analysis of the long-term average spectrum (LTAS), which is the spectral information averaged over time, revealed an increase in energy within the frequency region spanning formants for Lombard speech compared to normal articulation. This increase in energy resulted in a reduction in spectral tilt. A similar, albeit less pronounced, tendency was observed in the case of clear speech compared to casual speech [HM04]. Godoy et al. [GKS14] further investigated the influence of relative spectral amplitude differences between different speech styles on speech intelligibility. Figure 2.5 depicts the LTAS difference between Lombard and normal speech, as well as clear and casual speech. The figure illustrates that Lombard speech exhibits a noticeable increase in average energy within the 500-4500 Hz frequency band compared to normal speech, while the exaggeration of spectral content is less pronounced in clear speech compared to casual speech. This migration of spectral energy from low and high frequency bands to the mid-frequency range is attributed to the increased intelligibility of these speech styles.

In line with this investigation, Lu and Cooke [LC09] artificially redistributed the spectral energy of normal speech to match that of Lombard speech, thereby reducing the spectral tilt of the overall spectrum. This reduction in spectral tilt was found to improve speech intelligibility in noisy environments.

Vowel sounds can be categorized based on the position of articulation. The positioning of articulators is reflected in the formants of each vowel sound, particularly the first (F1) and second (F2) formants,

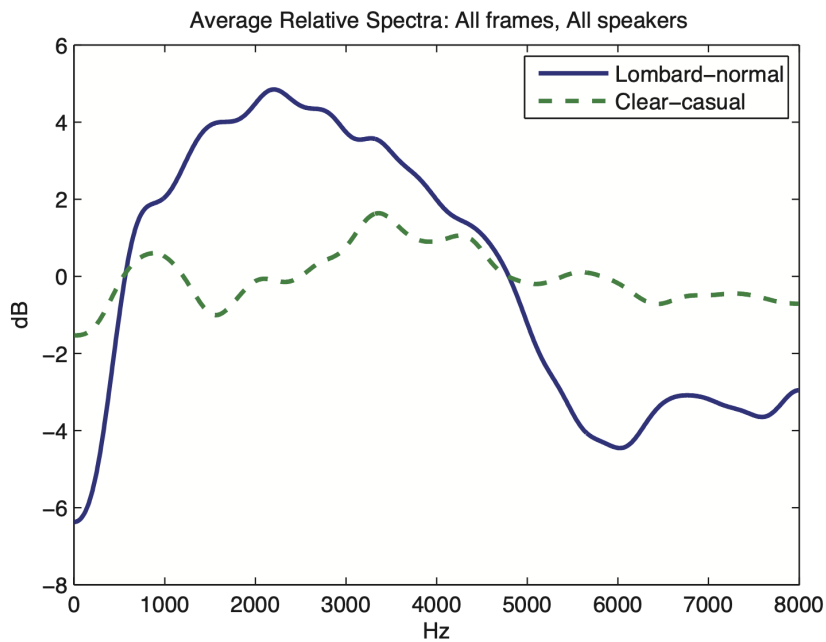


Figure 10.1: Average relative spectra for all frames (from Godoy et.al [GKS14]).

which together constitute the two-dimensional vowel space. Analysis of the correlation between vowel space and intelligibility has shown that speakers with larger vowel spaces tend to be more intelligible than those with reduced vowel spaces [HM04, BTP96]. Specifically, speakers with a wider range of F1 values appear to produce highly intelligible speech. Additionally, the F2 range has been found to be significantly correlated with sentence intelligibility [HM04] but less so with word intelligibility [BTP96]. Studies by Godoy et al. [GKS14] have also revealed an expansion of vowel space in clear speech compared to casual speech, while no such expansion is observed in Lombard speech. However, Lombard speech consistently exhibits an increase in F1 frequency, resulting in a shift in vowel space.

Motivated by the observation of vowel space expansion in clear speech, frequency wrapping techniques were tested in attempts to achieve vowel space expansion [MKS12, GKS14]. However, these techniques did not yield improvements in intelligibility. On the other hand, since formants and their transitions play a crucial role in perceiving and classifying different sound segments, sharpening the formants using statistical approaches has been found to be helpful in improving intelligibility in noise [ZKS12].

A variation in the speaker's fundamental frequency (F0) is also observed in Lombard speech [SB06]. This variation may contribute to the improved intelligibility associated with this style. However, modifying the F0 characteristics of normal speech to match those of Lombard speech did not enhance word recognition intelligibility in noise for normal listeners [LC09]. Furthermore, artificial flattening of F0 was found to degrade intelligibility [LB03, WS08], leaving the true impact of F0 on speech intelligibility uncertain.

Speech, as a real-valued signal, can be decomposed into a set of amplitude-modulated (AM) signals with carrier frequencies falling within the signal bandwidth [DFP94]. The temporal variation of these AM components is called the modulation of speech over time. Studies on clear speech have revealed higher modulation depths for temporal envelopes, which appear to correlate with the intelligibility advantage of clear speech [KB04b, LZ06]. The study by Drullman et al. [DFP94] demonstrated that smearing low-frequency modulations resulted in intelligibility degradation, with modulation frequencies in the range of 4 to 16 Hz being the most relevant for intelligibility. The modulation index metric is used to quantify the modulation depth of temporal envelopes [HS85], traditionally serving as a benchmark measure for

speech intelligibility in noisy and reverberant conditions. This argument is supported by studies in neuroscience, which show that speech is decomposed in the auditory cortex into spectro-temporal modulation content, and perception is driven by sounds that combine both temporal and spectral modulations effectively [MS05, SZ09, KDS96]. Consequently, modulation domain processing of speech has been proposed for applications such as noise reduction and echo cancellation, as it better isolates masking components [WL12, SJS18, JSS16]. Transplanting enhanced amplitude modulation from clear speech to casual speech has contributed to improved intelligibility in noise [KS16]. In a recent study, Bosker and Cooke [BC20] observed the same trend in Lombard speech, with enhanced modulations in the frequency range of 1-8 Hz. Subsequently, they demonstrated that transplanting Lombard amplitude modulation onto plain speech yielded additional intelligibility benefits, emphasizing the importance of amplitude modulation for speech intelligibility.

10.3 Spectral shaping and dynamic range compression (SSDRC)

Inspired by the intelligibility benefits of various speaking adaptations, artificial modification of speech to improve its intelligibility by altering the acoustic features has been recommended. Among the multitude of features contributing to intelligibility, spectral energy redistribution and increasing consonant-to-vowel ratio with dynamic time-domain energy reallocation were found to contribute largely to the intelligibility benefits in noise [RVD09, GKS14]. A combination of spectral shaping (SS) and dynamic range compression (DRC) was proposed in the work of Zorila et. al [ZKS12] as the SSDRC algorithm. SSDRC was tested in various listening settings on different languages since its introduction and has been found to produce the best intelligibility benefit in noise for normal and hearing-impaired listeners [CMVB13a, CMVB+13c, ZSFM17, SSCS20]. Therefore, we consider the SSDRC style over many natural styles as a reference for our research because it produces the highest intelligibility. Since the feature modifications elicited by SSDRC are used in the neural network architectures in the following chapters, a brief description of the SSDRC algorithm must be informative at this stage. SSDRC performs a two-stage speech processing to increase its intelligibility: 1) spectral shaping in the frequency domain and 2) dynamic range compression in the time domain.

10.3.1 Spectral shaping (SS):

In the initial phase of the enhancement framework, the SS module serves as an adaptive spectral shaper in the Fourier domain. Its primary objective is to impart a "crisp" and "clean" quality to speech by sharpening formants, as these formants play a crucial role in speech perception and contribute significantly to intelligibility, even under quiet listening conditions.

The entire process is carried out adaptively based on the voicing probability.

This module takes a plain speech signal, denoted as $x(t)$, as its input. The processing is performed frame-by-frame, where each frame has a fixed duration. The Discrete Fourier Transform (DFT) is applied to each frame, resulting in the magnitude spectral components denoted as $X(w, t)$. Adaptive shaping is applied considering the voicing probability, which helps avoid processing artefacts in regions with fewer sonorant characteristics, such as fricatives.

The voicing probability is determined using the following equation:

$$P_v(t) = \alpha \frac{rms(t)}{z(t)} \quad (10.1)$$

where $\alpha = 1/\max(P_v(t))$ is a normalization constant, and $rms(t)$ and $z(t)$ is the RMS value and zero crossings of the segment, respectively, for a window centred around the instant t with the length of 2.5 times the fundamental period (8.3 ms and 4.5 ms for male and female voices, respectively).

For every DFT frame $X(\omega, t)$, we employ the SEEVOC spectral envelope estimator [Pau81] on the magnitude spectrum to obtain the envelope estimate $E(\omega_k)$. Subsequently, we calculate the spectral envelope's tilt $T(\omega)$ as follows:

$$\log T(\omega) = c_0 + 2c_1 \cos(\omega), \quad (10.2)$$

where the variable c_m denotes the m^{th} cepstrum coefficient computed as

$$c_m = \frac{1}{N/2 + 1} \sum_{k=0}^{N/2} \log E(\omega_k) \cos(m\omega_k). \quad (10.3)$$

Therefore, the final adaptive spectral shaper has the transfer function function (over frame instance t)

$$H_s(\omega, t) = \left(\frac{E(\omega, t)}{T(\omega, t)} \right)^{\beta P_v(t)}. \quad (10.4)$$

In this way, the formant inclusive regions of voiced spectra are sharpened by selectively isolating the unvoiced segments with parameters $P_v(t)$. The variable β was set to 0.25 in most cases.

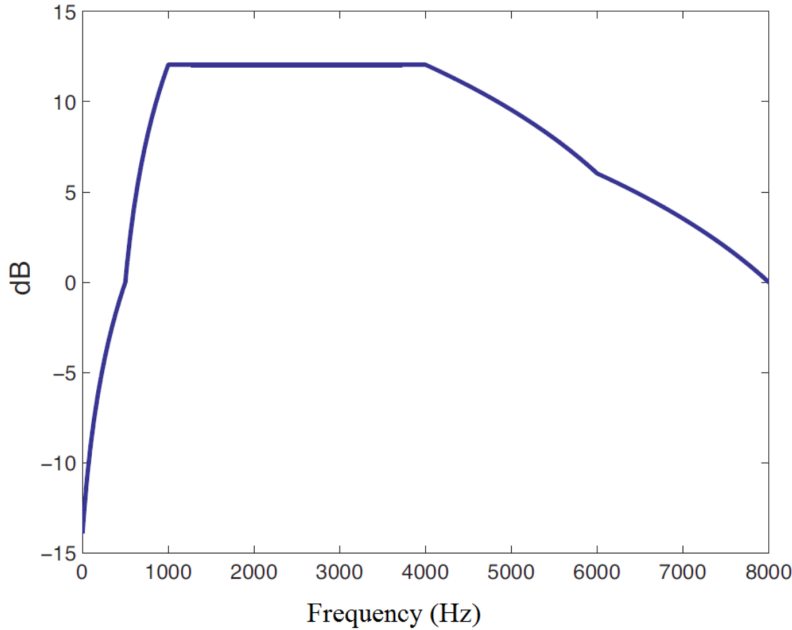


Figure 10.2: Spectral shaping fixed filter

Previous research has indicated that pre-emphasizing the spectrum above 1100 Hz enhances intelligibility in noisy environments [NG76]. Therefore, we employ an adaptive pre-emphasis filter as the second spectral shaping filter.

To avoid introducing a noisy quality to the speech during the filtering process, we utilize an adaptive pre-emphasis technique that is adapted based on the voicing probability. The following transfer function characterizes this adaptive pre-emphasis filter:

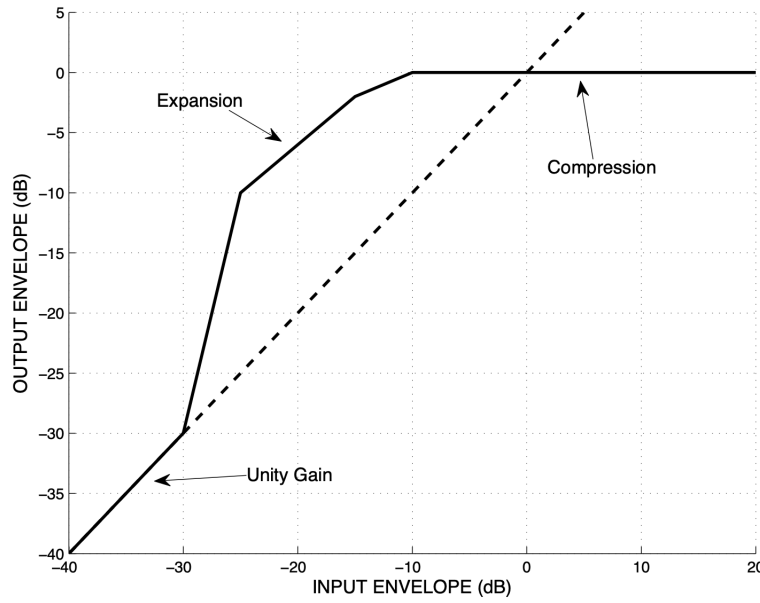


Figure 10.3: Input-Output Envelope Characteristic (IOEC) Curve

$$H_p(\omega, t) = \begin{cases} 1 & \omega \leq \omega_0 \\ 1 + \frac{\omega - \omega_0}{\pi - \omega_0} g P_v(t) & \omega > \omega_0 \end{cases} \quad (10.5)$$

where $\omega_0 = 0.125\pi$ for 16 kHz sampled speech, and the variable g is selected to 0.3.

Therefore, the adaptive cascaded spectral filtering can be expressed as

$$Y_{aSS}(\omega, t) = H_s(\omega, t) H_p(\omega, t) X(\omega, t). \quad (10.6)$$

Drawing inspiration from the spectral characteristics of Lombard Speech, we incorporate a fixed spectral gain filter as the final step to enhance energy in the mid-frequency range of the spectrum. This filter, denoted as $H_r(\omega)$, is non-adaptive or time-invariant. It increases the amplitudes of frequencies within the range of 1000 to 4000 Hz by 12 dB, while simultaneously attenuating components below 500 Hz with a slope of 6 dB per octave. This Lombard-inspired filter's transfer function is illustrated in Figure 10.2, and it aligns with the average spectral distribution observed in Lombard-style speech, as demonstrated in Figure 10.1.

Consequently, the output of the entire spectral shaping process is the final spectral-shaped signal, which is obtained by applying the Lombard-inspired spectral gain filter.

$$Y_{SS}(\omega, t) = H_r(\omega, t) Y_{aSS}(\omega, t). \quad (10.7)$$

Inverse Fourier transforms with overlap and add technique reconstructs the spectral-shaped waveform.

10.3.2 Dynamic range compression (DRC):

Following the spectral shaping module, the speech signal undergoes amplitude compression using a dynamic range compressor (DRC). The primary objective of the DRC is to reduce the variations in the

envelope of the signal. To achieve this, the gain applied by the DRC is derived from a desired input/output envelope characteristic (IOEC) curve. In the SSRC algorithm, the IOEC utilized is depicted in Figure 10.3, which consists of three distinct zones: unity gain, expansion, and compression.

Initially, the envelope of the speech signal is computed by employing the analytic signal technique with the assistance of the Hilbert transform. This results in the envelope estimation, denoted as $e(n)$. The estimated envelope, $e(n)$, is then subjected to dynamic compression with a release time constant of 2 ms and an almost instantaneous attack time constant. This compression process can be expressed using the following formulation:

$$\hat{e}(n) = \begin{cases} a_r \hat{e}(n-1) + (1 - a_r) e(n), & \text{if } e(n) < \hat{e}(n-1) \\ a_a \hat{e}(n-1) + (1 - a_a) e(n), & \text{if } e(n) \geq \hat{e}(n-1) \end{cases} \quad (10.8)$$

where the time constants are set to be $a_r = 0.15$ and $a_a = 0.0001$.

The 0 dB reference level of the envelope e_0 were set to the 30% of the maximum of the input signal envelope. With this reference value, the input envelope is computed as

$$e_{in}(n) = 20 \log_{10} (\hat{e}(n)/e_0). \quad (10.9)$$

The corresponding output level $e_{out}(n)$ is obtained by projecting $e_{in}(n)$ onto the IOEC curve in Figure 10.3 and the equivalent gain is computed as:

$$g(n) = 10^{(e_{out}(n) - e_{in}(n))/20}.$$

Therefore, the dynamic range compressed signal would be

$$s_g(n) = g(n)s(n).$$

To ensure that the loudness remains unchanged, the global energy of the output signal, denoted as $s_g(n)$, is rescaled to match that of the original unmodified speech. This rescaling process guarantees that the overall energy level of the modified speech aligns with that of the original signal, thereby maintaining consistent loudness.

Figure 10.4 illustrates the alteration induced by the spectral shaping and dynamic range compression (SSDRC) algorithm on a speech segment. The dynamic range of the SSDRC output is lower compared to the original signal. This reduction in dynamic range helps amplify low-intensity phonemes such as /p/ and /k/, contributing to improved intelligibility. However, it should be noted that this amplification comes at the cost of reducing the intensity of high-sonorant segments. Additionally, both the original signal and the SSDRC output have equal root mean square (RMS) energy, ensuring that the increase in intelligibility is not a result of direct signal amplification.

Controlled modifications of speech, such as those introduced by the SSDRC algorithm, can enhance intelligibility for listeners in noisy or distant environments. However, it is crucial to recognize that these modifications also alter the natural modulations of the speech signal, leading to a degradation in signal quality or naturalness. As a result, careful consideration must be given to parameters such as β in the spectral shaping and the attack (a_a) and release (a_r) time constants in the dynamic range compression of SSDRC when using the algorithm for different applications.

Furthermore, variations in pitch-period between male and female voices can impact quantities like

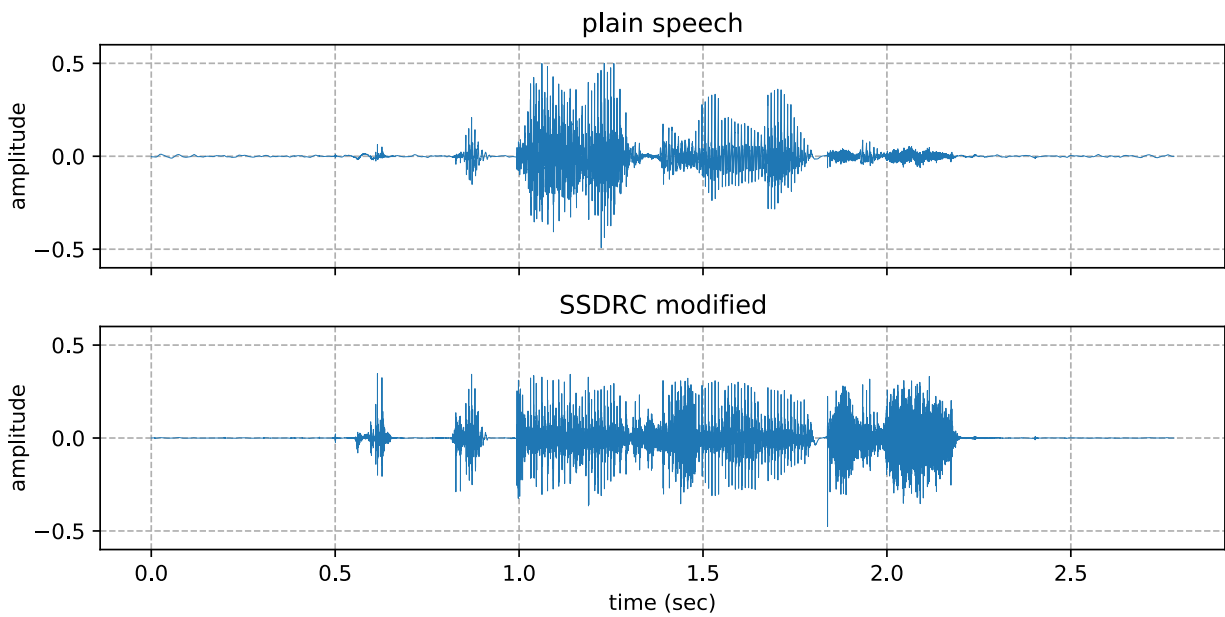


Figure 10.4: Speech waveform modified for intelligibility with SSDRC algorithm.

the window size used for computing the voice probability $P_v(t)$, which in turn can affect the quality and intelligibility of the output signal. Selecting appropriate trade-offs between quality and intelligibility becomes an application-specific task, considering the specific requirements and priorities of the intended use. For instance, in high-noise conditions, effortless message understanding is crucial, while in quiet environments, high quality and naturalness may be preferred. However, it is important to note that the scope of this thesis is limited to extremely low signal-to-noise ratio (SNR) listening scenarios, where maximizing intelligibility or message understanding is of paramount importance, with less emphasis on preserving naturalness. The debate surrounding the trade-off between quality and intelligibility extends beyond the scope of this thesis.

10.4 Neural TTS architecture

The proposed TTS system comprises two separately trained neural networks: (a) Tacotron, which predicts mel-spectrograms from text and (b) WaveRNN vocoder, which converts the mel-spectrograms into time-domain waveforms.

10.4.1 Tacotron

Tacotron [WSRS⁺17] (Figure 10.5) is a seq2seq architecture with an attention mechanism, and the encoder-decoder neural network framework heavily inspires it. The system has two main components: (a) an encoder and (b) an attention decoder. The encoder consists of 1-D convolutional filters, followed by fully-connected (FC) layers and a bidirectional gated recurrent unit (GRU). It takes text as input and extracts sequential representations of text. The attention decoder is a set of recurrent layers which produces the attention query at each decoder time step. The input to the decoder RNN can be produced by concatenating the context vector and output of the attention RNN. The decoder RNN is basically a 2-layer residual GRU, whereas the attention RNN has a single GRU layer. The output of the attention decoder is a sequence of mel-spectrograms, which is then passed to the vocoding stage.

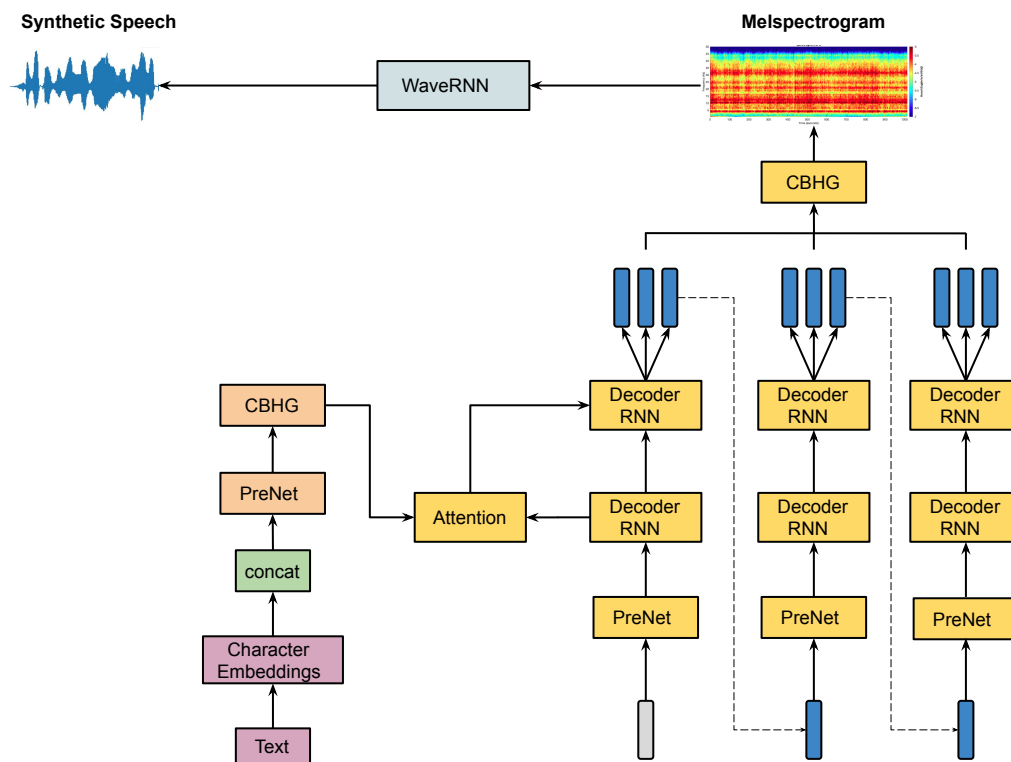


Figure 10.5: Block diagram of Tacotron architecture.

10.4.2 WaveRNN

The implemented WaveRNN vocoder is based on the repository¹ which in turn is heavily inspired by WaveRNN training [KES⁺18]. This architecture combines residual blocks and an upsampling network, followed by GRU and FC layers, as depicted in Figure 8.3.

The architecture can be divided into two major networks: the conditional network and the recurrent network. The conditional network consists of a pair of a residual network and an upsampling network with three scaling factors. At the input, we first map the acoustic features, i.e., the mel-spectrograms to a latent representation with the help of multiple residual blocks. The latent representation is then split into four parts which are later used as input to the subsequent recurrent network. The upsampling network is implemented to match the desired temporal size of the input signal. The outputs of these two convolutional networks, i.e., residual and upsampling networks, along with speech, are fed into the recurrent network. As part of the recurrent network, two uni-directional GRUs are employed with a few FC layers. By design, such a network not only reduces the overhead complexity with fewer parameters but also takes advantage of temporal context, resulting in better prediction.

In addition, we apply continuous univariate distribution to be a mixture of logistic distributions [OLB⁺18b], which allows us to calculate the probability of the observed discretized value easily. Finally, discretized mix logistic loss is applied to the discretized speech samples.

¹<https://github.com/fatchord/WaveRNN>

10.5 Transfer learning

Most deep learning methods perform well under the standard assumption that the training and inference data are drawn from similar feature space and data distribution. When the distribution changes, models must be trained from scratch using new training data. Under the condition of data scarcity, such as in our case for Lombard data, training a new model on such a limited sample size might lead to poor execution. In such cases, transfer learning (TL) offers a desirable and extremely important adaptation framework [PY09]. Assuming that there are two tasks, the source task and the target task, TL tries to boost the performance of the target task by utilizing knowledge learned from the source task via fine-tuning prior distributions of the hyper-parameters.

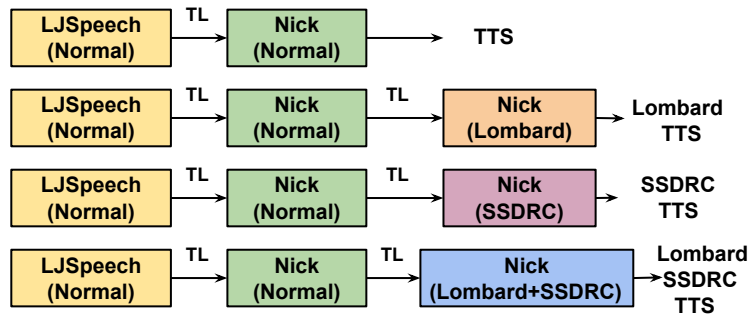


Figure 10.6: A functional block diagram of the proposed adaptation techniques used in this study. Each block represents a TTS system (Tacotron + WaveRNN), which takes text as input and generates speech samples.

We develop four TTS systems based on the speaking styles: normal TTS, Lombard TTS, SSDRC TTS and Lombard-SSDRC TTS. To effectively transfer the prior knowledge, we initially train the TTS system with normal speech (single female speaker from LJSpeech corpora), which has a large amount of linguistic variability. Then, we adapt the learned model with normal speech from a male speaker (Nick). This normal TTS serves as the baseline system for our experiments. The Lombard TTS system is then fine-tuned using the TL approach on the limited Lombard data from the same male speaker (Nick) again. Whereas SSDRC TTS uses training data processed with the SSDRC algorithm applied to Nick’s normal speech. The last TTS system is fine-tuned on data that is prepared by applying SSDRC algorithm on Nick’s Lombard speech, referred to as Lombard-SSDRC TTS. Please note that all proposed TTS systems comprise Tacotron and WaveRNN modules [PPSed], and each module is trained separately using data from the corresponding target speech style.

10.6 Database and Hyperparameters Selection

The proposed TTS systems are trained using two publicly available databases, i.e., LJSpeech corpus [Kei17] and Nick Hurricane Challenge speech data [CMVB⁺13b]. LJSpeech comprises 13,100 short audio clips of a single female professional speaker reading passages. The Nick data has both normal and Lombard styles of British male voice professional speech. The normal speech consists of 2592 utterances (~2 hours), whereas the Lombard speech data has 720 utterances (~30 minutes). During training, we always consider 2400 utterances for normal and 500 utterances for Lombard speech. We additionally compare with the baseline Lombard TTS system, which is built on Tacotron and WaveNet architecture [BJA⁺19b]. The WaveNet configuration used in their system consists of three repetitions of a 10-layer convolution stack with exponentially growing dilations, 64 residual channels and 128 skip channels, whereas the Tacotron architecture is similar to ours. The proposed Tacotron and WaveRNN models use 80-dimensional normalized mel-spectrograms extracted from audio frames of width of 50ms, hop

Table 10.1: $SIIB^{Gauss}$ intelligibility measure at different SNR levels under speech-shaped and competing-speaker noise.

Systems	SSN			CSN		
	-10 dB	-5 dB	0 dB	-21 dB	-14 dB	-7 dB
TTS	15.03	26.80	42.43	13.3	17.86	28.27
Lombard TTS [BJA ⁺ 19b]	17.89	33.89	54.53	9.91	18.1	36.21
Lombard TTS (ours)	20.02	37.43	58.65	13.52	22.51	41.65
SSDRC TTS	29.90	51.02	77.97	16.73	29.75	55.56
Lombard-SSDRC TTS	35.04	58.68	88.35	19.13	35.84	68.35

length of 12.5ms, and 2048-point Fourier transform. In Tacotron, character embeddings are set to 256, and a progressive training schedule is employed to reduce batch size from 32 to 8. WaveRNN architecture is based on a set of 10-layer convolution stacks inside residual blocks followed by 2 GRUs. Each GRU has 512 hidden units. Code and audio samples can be found in ².

10.7 Observations and discussion

Objective intelligibility scores are computed first for the five style adapted methods (TTS, Lombard TTS [BJA⁺19b], proposed Lombard TTS, also refer to as Lombard TTS (ours), SSDRC TTS and Lombard-SSDRC TTS) under two different noisy conditions. A recently developed intelligibility metric called ‘speech intelligibility in bits’ ($SIIB^{Gauss}$) [VKKH18] is implemented as an objective evaluation metric. It considers the information capacity of a Gaussian channel between clean and noisy signals. Higher values refer to better intelligibility. The scores are evaluated from 250 utterances, and each adaptation approach has 50 distinct utterances. Table 10.1 presents $SIIB^{Gauss}$ intelligibility scores. We consider three different Signal-to-Noise Ratio (SNR) levels masked with two types of noise: speech-shaped noise (0, -5 and -10 dB) and competing-speaker (-7, -14 and -21 dB). Since we are focusing on the context of TTS, we omitted the scores for natural speech in our experiments.

It can be observed that the standard synthesis system trained with normal speech, referred to here as the speech type ‘TTS’, is the worst performer when compared to the rest of the methods under any condition as expected. To enhance the intelligibility, TTS is re-trained with limited Lombard style data. We observe that the proposed Lombard TTS i.e., Lombard TTS (ours) is able to mimic the Lombardness successfully and outperforms baseline Lombard TTS from [BJA⁺19b] with a relative improvement between 8% and 12% in SSN and 15% to 36% in CSN conditions across different SNR levels: from low to high SNRs. The results also show high performance gain of 18% and 36% in Low SNR i.e., -10 dB for SSN and -21 dB in CSN conditions, respectively. The use of WaveRNN instead of WaveNet vocoder as in the baseline Lombard TTS, demonstrates how the choice of vocoder affects the intelligibility of synthesized speech. WaveRNN effectively adapts to the new style while trained with a limited amount of target style data. Furthermore, considering the SSDRC approach, we aim towards additional intelligibility gains under adverse noise conditions. Our results reveal that SSDRC TTS archives further improvement compared to the Lombard TTS. Motivated by the boosting effect of Lombard style, along with the enhancement by SSDRC data in terms of speech intelligibility, the proposed Lombard-SSDRC TTS shows significant intelligibility gains between 110% and 130% in SSN, and 47% to 140% in CSN against TTS. Those results can be attributed by the fact that the combined model exploits efficiently both Lombardness and spectral shaping with range compression by modifying time-frequency regions.

²<https://dipjyoti92.github.io/TTS-Style-Transfer/>

To assess the performance on subjective evaluation, metric scores were computed based on the number of keywords correctly identified in each sentence. The short common words ‘a’, ‘the’, ‘in’, ‘to’, ‘on’, ‘of’, and ‘for’ were excluded. The listening test was conducted via a web-based interface and ten native listeners participated in the test. No listener heard the same sentence twice, and each condition was heard by the same number of listeners. Since intelligibility level varies from one listener to another and large variability in scores can be possible when listeners use different hearing devices or backgrounds, intelligibility gains should be observed from a common reference point. This was achieved by designing an initial pilot study where subject-specific SNR levels are matched with the speech reception threshold (SRT) at which 40% of normal speech is intelligible for each individual listener. In the final listening test, we choose SNR levels based on the values obtained from the pilot study for each listener individually.

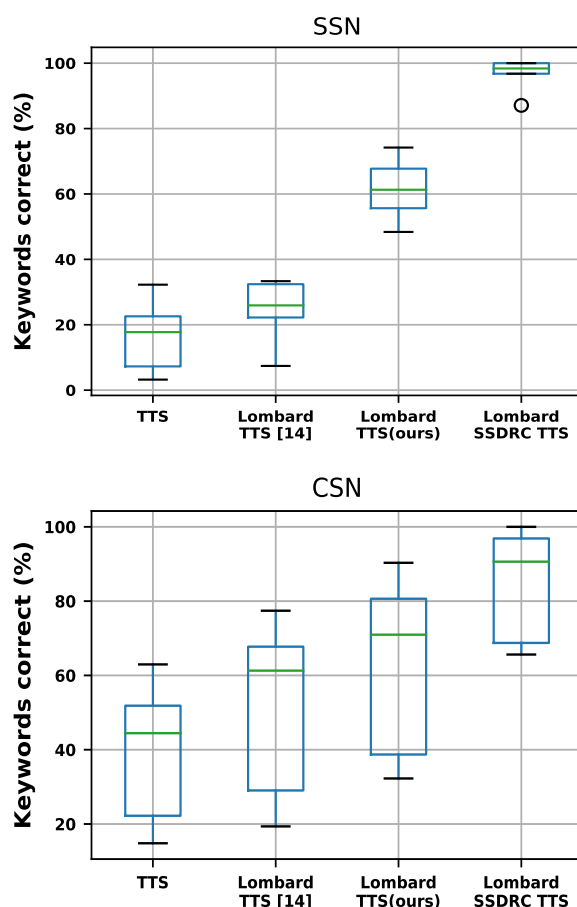


Figure 10.7: Box plot results for listeners' keyword scores across of methods for SSN and CSN.

Box plots reported in Figure 10.7 allow comparing different TTS modification algorithms. The subjective results reveal a similar pattern to the objective metrics. The proposed Lombard-SSDRC TTS outperforms all other methods by a remarkable margin under all noisy conditions. Lombard-SSDRC TTS shows superior performance by achieving a remarkable relative improvement of 455% for SSN and 104% for CSN in the median keyword correction rate compared to the TTS method. It is worth noting that the performance gains are immensely higher in SSN conditions, although we observe outstanding performance gains in both noisy conditions. Moreover, the comparison between Lombard TTS [BJA⁺19b] and Lombard TTS (ours) adaptation methods highlights that Lombard TTS (ours) method achieves significantly better performance in terms of keyword correction rate. This confirms the adaptability of WaveRNN for limited data scenarios and shows its effectiveness in the transfer learning approach. The results indicate a

relative improvement of 136% in SSN and 16% in CSN compared to Lombard TTS [BJA⁺19b] in terms of median keyword correction rate.

10.8 Conclusions and perspective

In summary, we built and evaluated a set of intelligible TTS systems for various speaking styles with the help of transfer learning in Tacotron + WaveRNN architecture. The synthesized voice was adapted to two strategies: Lombard-style recordings and SSDRC algorithm. First, we showed that the Lombard-adapted TTS system (ours) is able to learn the Lombard style under limited training data successfully and outperforms the baseline Lombard TTS system [BJA⁺19b] by a significant margin when masked either with SSN or CSN noise. This shows the advantage of applying a neural-based WaveRNN vocoder and its importance in achieving highly intelligible Lombard synthetic speech.

The SSDRC adaptation of TTS was found to improve further the intelligibility substantially compared to both the normal and Lombard-adapted TTS systems in objective metric. Furthermore, to enjoy larger intelligibility gains, we combined the benefits of Lombardness with the SSDRC modification strategy. Experiments on both objective and subjective intelligibility scores confirmed that the combined system contributed to significant gains under all noisy conditions.

In conclusion, these observations further underline the fact that neural networks can be optimized to learn various speaking styles and generate intelligible speech in adverse conditions, an observation that was reported in the previous chapter with speech input.

Chapter 11

Conclusions and Future Work

11.1 Overview

This thesis represents a comprehensive exploration of the potential of neural networks across various algorithmic levels and applications in the domains of image and speech synthesis. The findings underscore the critical importance of algorithmic advancements in the realm of artificial intelligence and demonstrate their versatility in a wide array of applications. This research serves as a testament to the adaptability and effectiveness of neural networks, showcasing their ability to drive innovation and address diverse challenges in today's technology-driven world.

Our thesis introduces the innovative WeGAN training algorithm, which has shown substantial improvements in GAN performance compared to traditional training methods. We demonstrate WeGAN's adaptability, highlighting its versatility, by applying it to various GAN architectures. We conducted benchmark experiments on synthetic data drawn from a mixture of 8 normal distributions and real-world datasets such as CIFAR and ImageNET. The results indicate that our new algorithm converges to the data distribution more rapidly than the vanilla GAN, resulting in enhanced performance compared to the baseline training procedures.

Furthermore, WeGAN's applicability extends beyond specific GAN types, making it a versatile tool for many applications. We extended our weighting approach to the domain of voice conversion, introducing WeStarGAN, an algorithmic variation of StarGAN capable of non-parallel multi-domain voice conversion tasks. Despite minor additional computational costs, this approach significantly improved the training process by strengthening the generator at each minibatch iteration. Subjective evaluations showed notable enhancements in speech quality and speaker similarity when compared to baseline methods. Our research thus demonstrates the potential of WeGAN and related techniques to enhance GAN performance across various applications, from data synthesis to voice conversion, offering promising avenues for future research and development in the field of generative adversarial networks.

Although GANs have achieved impressive results, their training process can often be unstable and require extensive experimentation to find the right loss function, optimization algorithm, and architecture. In this thesis, we focused on addressing the challenge of loss function selection and introduced a novel loss function based on cumulant generating functions, resulting in the Cumulant GAN. Using Cumulant GAN's loss function, grounded in cumulant generating functions, provides a comprehensive way to characterize distribution statistics, simplifying the discriminator's complexity. This leads to improved and more stable GAN training. Moreover, Cumulant GAN offers flexibility by allowing interpolation between

various divergences and distances through simple adjustments of two hyperparameters (β, γ), providing a versatile mechanism for choosing and potentially adapting the objective to minimize.

This thesis also addresses the challenge of accurately estimating divergences, which are crucial for various machine learning tasks but often suffer from high variance, especially in high-dimensional datasets. Common divergence estimators, like Kullback-Leibler (KL) divergence, f-divergences, Hellinger divergence, α -divergences, and Rényi divergence estimators, perform well in low-dimensional scenarios but struggle with large, high-dimensional data typical in modern machine learning.

To tackle this issue, we propose a novel approach called Variance Penalty (VP) to reduce the variance in divergence estimators. During optimisation, the VP is added to the objective function, effectively trading off bias and variance. This thesis presents significant contributions across four key areas: Firstly, we introduce a versatile VP that reduces variance in divergence estimators, designed specifically for f-divergences and extended to non-linear settings like KL divergence and Rényi divergences. Second, through extensive empirical validation on synthetic datasets, we demonstrate the VP's effectiveness in improving divergence estimation across diverse scenarios, enhancing both mean squared error and median absolute error. We also highlight its applicability to various mutual information bounds. Third, in real-world applications, we deploy the VP to actual datasets, enabling more accurate estimation of Rényi divergence and aiding in identifying rare biological subpopulations, with the added benefit of stabilizing estimators when order values exceed one.

In this thesis, we address the challenges faced by neural vocoders, particularly in multi-speaker scenarios, where it's impractical to cover all possible speaker variations during training. We propose a robust universal SC-WaveRNN vocoder designed to synthesize high-quality speech across a wide range of speakers without needing adaptation or retraining. The key innovation in our approach is the use of speaker embeddings that encompass a diverse set of seen and unseen conditions, enhancing both the vocoder's controllability and generalization capabilities. SC-WaveRNN offers several advantages, including improved multi-speaker vocoder training and better generalization to unseen speaker characteristics. This speaker conditioning technique is data-efficient and computationally less demanding than training separate models for each speaker. Subjective and objective evaluations confirm the effectiveness of our method, demonstrating higher speech quality and speaker similarity compared to baseline approaches.

Moreover, we extend our approach to create an efficient zero-shot TTS system. This innovation shows that our proposed zero-shot TTS, combined with a universal vocoder, can enhance both speaker similarity and the naturalness of synthetic speech, even for seen and unseen speakers. In future work, we plan to explore the construction of speaker embeddings and their potential applications with unseen data further, opening up exciting possibilities for improving speech synthesis across various domains and scenarios.

In pursuing a universal TTS synthesis system capable of generating speech that mimics a reference speaker's characteristics and speaking style, we encounter a substantial challenge: the potential distortion of speaker attributes and speaking style. This challenge becomes particularly pronounced in zero-shot learning scenarios where only limited reference data is available. To tackle this challenge, we introduce a Universal TTS framework comprising four key components: a content encoder, a style encoder, a speaker encoder, and a speech decoder. The content encoder focuses on generating content embedding from the input text. Meanwhile, the style encoder handles the representation of style factors, converting them into a style embedding. The speaker encoder identifies and encodes the speaker's identity as a speaker embedding. The speech decoder, informed by all the embeddings, synthesizes the target speech with the desired characteristics and style. However, a significant hurdle arises when attempting to generalize these models to handle multiple speakers and styles using only reconstruction loss. During training, information leaks across different embeddings, leading to issues known as "content leakage" and "style leakage." This means that the decoder, during inference, might expect content from the style vector or even misinterpret

speaker attributes.

We propose a novel disentangled representation approach to overcome these issues, leveraging cumulant-generating functions in speech synthesis. Our system approximates and minimizes the Rényi divergence between content-style and style-speaker pairs. This joint training process, incorporating an adversarial component, eliminates the problems of content and style leakage, resulting in a truly universal TTS system. The primary advantage of this universal TTS approach lies in its high controllability. It enhances multi-speaker and multi-style training while improving generalization capabilities by enabling reliable transfers of speaker and style information. The conditioning of speaker and style attributes can be achieved with minimal reference data in an unsupervised manner, making it a powerful tool for various TTS applications. Lastly, we integrate the VP into speech representation learning, disentangling text, speaker, and style components, leading to significantly improved training performance over baseline systems.

The remarkable advancements in speech synthesis have opened up new avenues for improving real-world speech communication, but the persistent challenge of background noise remains a critical issue. Speech intelligibility, which measures the degree to which spoken content is understandable, holds immense importance in diverse applications, ranging from emergency alerts to machine interactions. In this context, our research has yielded significant contributions. We developed and evaluated a series of intelligible TTS systems, utilizing transfer learning within the TTS architecture. These systems were adapted to different speaking styles, particularly Lombard style recordings and the Speech Separation and Dereverberation with deep Recurrent Neural Networks for Cochlear Implants (SSDRC) algorithm. Our findings demonstrate that the Lombard-adapted TTS system outperforms baseline Lombard TTS systems, especially in the presence of both Single-Sided Noise (SSN) and Colored Stationary Noise (CSN). Moreover, SSDRC adaptation further substantially enhances intelligibility, surpassing both normal and Lombard-adapted TTS systems in objective metrics. The combined approach, which combines Lombardness with SSDRC modification, yields significant gains in intelligibility under various noisy conditions. These results underscore the potential of neural networks to learn diverse speaking styles and generate intelligible speech in adverse environments, as highlighted in the previous chapter with speech input.

11.2 Future research directions

The research conducted in this thesis has opened up various avenues for future investigations and extensions. While the algorithms presented here represent significant advancements, they are not without limitations and open challenges. Future research could explore these areas to enhance further the work done in this thesis.

We introduced innovative loss functions derived from a novel approach based on the CGF; this CGF-based replacement of expected values isn't confined solely to the WGAN framework. It has the potential to extend to various other GAN loss functions, leading to the development of entirely new loss functions that can contribute to advancements in the field of generative adversarial networks. This underscores the versatility and broader applicability of our proposed approach beyond WGAN.

Variational representations for divergences translate the estimation of divergence into an optimization problem. It offers a valuable mathematical tool to analyze probabilistic models between multivariate probability distributions. However, learning proper representations from these high-dimensional data is challenging, especially when trying to distill that knowledge into useful representations. Even though our proposed neural-based divergence estimators provide a good representation, they often have different statistical variances that may result in unreliable divergence estimation. The future directions can be explored by two methods that have the potential to reduce the variance of the estimators: Weighted sum

of divergences and new families of transformations. Our superior disentanglement learning will allow a greater degree of interpretability and controllability, especially for generative models, maintaining high production value - be it audio or images. Therefore, the research outcome can be utilized in domains such as Augmented Reality Audio and AR/VR Human Understanding. The following domains could benefit from utilizing our disentanglement approach, which is able to extract useful features where high-dimensional observed data is disentangled into a low-dimensional representation comprising semantically meaningful factors of variation. Finally, this allows each factor to be extracted from a different distribution and then combined together for generation purposes.

Future research directions in Universal TTS entail several critical domains. The need for enhanced style control is evident. Users should be able to intricately define and manipulate speaking styles, allowing for the expression of emotions, formality, and other nuanced characteristics. This empowers the technology to cater to diverse user preferences and situational demands, making the synthesized speech more personalized and adaptable. Additionally, there is a pressing requirement to improve the data efficiency of Universal TTS systems. Achieving this involves investigating techniques like low-resource and zero-shot learning. These methodologies are essential because they address the challenge of training TTS models effectively even when limited data is available. They pave the way for systems to adapt to and generate unseen styles, expanding the utility of TTS technology across a broader spectrum of use cases. Furthermore, researchers should delve into the realm of unsupervised or weakly supervised style learning methods. These approaches hold the promise of allowing TTS systems to autonomously recognize and adapt to various speaking styles without relying on explicit style labels. This autonomy is pivotal for creating TTS systems that are highly versatile and capable of handling diverse styles and contexts seamlessly.

Another intriguing avenue for future research would involve extending the conditioning information with style and speaker attributes to the vocoder component in the context of Universal TTS systems. While significant progress has been made in conditioning the acoustic models to generate speech with specific speaking styles and speaker characteristics, extending this conditioning to the vocoder could yield promising results. This approach would involve aligning the vocoder's capabilities with those of the acoustic models, allowing for a more comprehensive and synchronized control over the entire TTS pipeline.

Appendix A

Publications

During this work, the following publications took place (in chronological order):

1. Conference and online publications

- (a) Pantazis, Y, **Paul, D.**, Fasoulakis, M., Stylianou, Y.,
Training Generative Adversarial Networks with Weights.,
in EUSIPCO 2019, pp. 1–5
- (b) **Paul, D.**, Pantazis, Y. and Stylianou, Y.,
Weighted generative adversarial network for many-to-many voice conversion.,
in Proc. International Congress on Acoustics (ICA) 2019, pp. 5742–5744.
- (c) **Paul, D.**, Pantazis, Y., Stylianou, Y.,
Non-parallel Voice Conversion using Weighted Generative Adversarial Networks.,
in Proc. Interspeech 2019, pp. 659–663.
- (d) **Paul, D.**, Pantazis, Y., Stylianou, Y.,
Speaker Conditional WaveRNN: Towards Universal Neural Vocoder for Unseen Speaker and Recording Conditions.,
in Proc. Interspeech 2020, pp. 235–239.
- (e) **Paul, D.**, Shifas, M. P., Pantazis, Y., Stylianou, Y.
Enhancing Speech Intelligibility in Text-To-Speech Synthesis Using Speaking Style Conversion.,
in Proc. Interspeech 2020, pp. 1361–1365.
- (f) **Paul, D.**, Chermaz, C, Shifas, M. P., Raman, S., Govender, A., Simantiraki, O.,
Enriched Speech for Effortless Listening.

In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Show & Tell, ST-P4.2., 2020.

- (g) **Paul, D.**, Mukherjee, S. P., Pantazis, Y., Stylianou, Y.,

A Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning based on Renyi Divergence Minimization,

in Proc. Interspeech 2021, pp. 3625–3629.

- (h) Birrell, J., Katsoulakis, M.A., Pantazis, Y., **Paul, D.**, Tsourtis, A.,

A Variance Reduction Method for Neural-based Divergence Estimation.

[Online]. Available: <https://openreview.net/forum?id=6g4VoBTaq6I>

2. Journals

- (a) Pantazis, Y., **Paul, D.**, Fasoulakis, M., Stylianou, Y., Katsoulakis, M.,

Cumulant GAN,

in IEEE Transactions on Neural Networks and Learning Systems 2022, vol. 34, no. 11, pp. 9439-9450, Nov. 2023, doi: 10.1109/TNNLS.2022.3161127.

- (b) **Paul, D.**, Pantazis, Y., Stylianou, Y.

Redefining Disentangled Representations in Universal Multi-Speaker Multi-Style Text-to-Speech Systems,

in IEEE/ACM Transactions on Audio, Speech, and Language Processing (To be submitted).

Bibliography

- [AB17] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.
- [ACB17] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.
- [ACC⁺17] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *International Conference on Machine Learning*, pages 195–204. PMLR, 2017.
- [ACD15] R. Atar, K. Chowdhary, and P. Dupuis. Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):18–33, 2015.
- [ACP⁺18] Sercan Arık, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*, pages 10019–10029, 2018.
- [AHK12] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.
- [ANSK90] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)*, 11(2):71–76, 1990.
- [asr] Google’s speech-to-text. <https://cloud.google.com/speech-to-text>.
- [ATS18] Nagaraj Adiga, Vassilis Tsiaras, and Yannis Stylianou. On the use of wavenet as a statistical vocoder. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678. IEEE, 2018.
- [BA03] David Barber and Felix V. Agakov. The IM algorithm: A variational approach to information maximization. In *NIPS*, pages 201–208, 2003.
- [BAA17] Bajibabu Bollepalli, Manu Airaksinen, and Paavo Alku. Lombard speech synthesis using long short-term memory recurrent neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5505–5509, 2017.
- [BAC⁺18] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.
- [BBR⁺18a] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Shertil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018.

- [BBR⁺18b] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.
- [BC20] Hans Rutger Bosker and Martin Cooke. Enhanced amplitude modulations contribute to the lombard intelligibility benefit: Evidence from the nijmegen corpus of lombard speech. *The Journal of the Acoustical Society of America*, 147(2):721–730, 2020.
- [BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [BCNM06] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.
- [BDK⁺20a] Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, Γ) -Divergences: Interpolating between f -Divergences and Integral Probability Metrics. *arXiv e-prints*, page arXiv:2011.05953, November 2020.
- [BDK⁺20b] Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational representations and neural network estimation of rényi divergences. *arXiv preprint arXiv:2007.03814*, 2020.
- [BDK⁺20c] Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational Representations and Neural Network Estimation of Rényi Divergences. *arXiv e-prints*, page arXiv:2007.03814, July 2020.
- [BGS16] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *Proceedings of the International Conference on Learning Representations*, 2016.
- [Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.
- [BJA⁺19a] Bajjibabu Bollepalli, Lauri Juvela, Manu Airaksinen, Cassia Valentini-Botinhao, and Paavo Alku. Normal-to-Lombard adaptation of speech synthesis using long short-term memory recurrent neural networks. *Speech Communication*, 110:64–75, 2019.
- [BJA⁺19b] Bajjibabu Bollepalli, Lauri Juvela, Paavo Alku, et al. Lombard speech synthesis using transfer learning in a Tacotron text-to-speech system. in *Proc. Interspeech*, pages 2833–2837, 2019.
- [BK06] Michel Broniatowski and Amor Keziou. Minimization of divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442, 2006.
- [BK09] Michel Broniatowski and Amor Keziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16–36, 2009.
- [BKH00] Ann R Bradlow, Nina Kraus, and Erin Hayes. Speaking clearly for learning-disabled children: Sentence perception in noise. *The Journal of the Acoustical Society of America*, 108(5):2603–2603, 2000.
- [BKH03] AR Bradlow, N Kraus, and E Hayes. Speaking clearly for learning-impaired children: Sentence perception in noise. *Journal of Speech, Language, and Hearing Research*, 46(1):80–97, 2003.
- [BKP20] Jeremiah Birrell, Markos A. Katsoulakis, and Yannis Pantazis. Optimizing Variational Representations of Divergences and Accelerating their Statistical Estimation. *arXiv e-prints*, page arXiv:2006.08781, June 2020.
- [BTP96] Ann R Bradlow, Gina M Torretta, and David B Pisoni. Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech communication*, 20(3):255, 1996.

- [CAS⁺18] Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C Cobo, Andrew Trask, Ben Laurie, et al. Sample efficient adaptive text-to-speech. In *International Conference on Learning Representations*, 2018.
- [CCK⁺18] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.
- [CCL⁺19] Mengnan Chen, Minchuan Chen, Shuang Liang, Jun Ma, Lei Chen, Shaojun Wang, and Jing Xiao. Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding. In *Proc. Interspeech*, pages 2105–2109, 2019.
- [CDH⁺16] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the International Conference on Neural Information Processing Systems*, page 2180–2188, 2016.
- [CGK⁺02] Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137:43–90, 2002.
- [CHD⁺20] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *ICML 2020*, July 2020.
- [CJK16] Veaux Christophe, Yamagishi Junichi, and MacDonald Kirsten. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. *The Centre for Speech Technology Research (CSTR)*, 2016.
- [CJLV16] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE, 2016.
- [CL12] Martin Cooke and Maria Luisa García Lecumberri. The intelligibility of lombard speech for non-native listeners. *The Journal of the Acoustical Society of America*, 132(2):1120–1129, 2012.
- [CLH⁺21] Chung Ming Chien, Jheng Hao Lin, Chien Yu Huang, Po Chun Hsu, and Hung Yi Lee. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. *arXiv preprint arXiv:2103.04088*, 2021.
- [CLLD14a] L.H. Chen, Z.H. Ling, L.J. Liu, and L.R. Dai. Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(12):1859–1872, 2014.
- [CLLD14b] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Li-Rong Dai. Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1859–1872, 2014.
- [CLY⁺20] Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *Proc. ICASSP*, pages 6184–6188, 2020.
- [CLZ⁺17] T. Che, Y. Li, R. Zhang, R. D. Hjelm, W. Li, Y. Song, and Y. Bengio. Maximum-likelihood augmented discrete generative adversarial networks. *arXiv preprint arXiv:1702.07983*, 2017.
- [CMV14] Martin Cooke, Catherine Mayo, and Julián Villegas. The contribution of durational and spectral changes to the lombard speech intelligibility benefit. *The Journal of the Acoustical Society of America*, 135(2):874–883, 2014.

- [CMVB13a] Martin Cooke, Catherine Mayo, and Cassia Valentini-Botinhao. Intelligibility-enhancing speech modifications: the hurricane challenge. In *Interspeech*, pages 3552–3556, 2013.
- [CMVB⁺13b] Martin Cooke, Catherine Mayo, Cassia Valentini-Botinhao, et al. Hurricane natural speech corpus. *LISTA Consortium, Language and Speech Laboratory, Universidad del Pais.*, 2013.
- [CMVB⁺13c] Martin Cooke, Catherine Mayo, Cassia Valentini-Botinhao, Yannis Stylianou, Bastian Sauert, and Yan Tang. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585, 2013.
- [CVMG⁺14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [CZZ⁺20] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.
- [DBYP10] S. Desai, A.W. Black, B. Yegnanarayana, and K. Prahallad. Spectral mapping using artificial neural networks for voice conversion. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(5):954–964, 2010.
- [DDS⁺09] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [DE97] P. Dupuis and R.S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley series in probability and statistics. John Wiley & Sons, New York, 1997. A Wiley-Interscience Publication.
- [DE11] Paul Dupuis and Richard S Ellis. A weak convergence approach to the theory of large deviations. *John Wiley & Sons*, 902, 2011.
- [DFP94] Rob Drullman, Joost M Festen, and Reinier Plomp. Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95(2):1053–1064, 1994.
- [DHS18] Yan Deng, Lei He, and Frank Soong. Modeling multi-speaker latent space to improve neural TTS: Quick enrolling new speaker and enhancing premium voice. *arXiv preprint arXiv:1812.05253*, 2018.
- [DISZ18] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *Proceedings of the International Conference on Learning Representations*, 2018.
- [DKD⁺10] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.
- [DRY⁺09] Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad. Voice conversion using artificial neural networks. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3893–3896. IEEE, 2009.
- [DSY⁺10] Li Deng, Michael L Seltzer, Dong Yu, Alex Acero, Abdel-rahman Mohamed, and Geoff Hinton. Binary coding of speech spectrograms using a deep auto-encoder. In *Eleventh annual conference of the international speech communication association*, 2010.
- [DT20] Tan Daxin and Lee Tan. Fine-grained style modelling and transfer in text-to-speech synthesis via content-style disentanglement. *arXiv preprint arXiv:2011.03943*, 2020.

- [Dut97] Thierry Dutoit. *An introduction to text-to-speech synthesis*, volume 3. Springer Science & Business Media, 1997.
- [DV83] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. IV. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.
- [EB08] Daniel Erro Eslava and Asunción Moreno Bilbao. Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models. *Barcelona, Spain: PhD Thesis, Universitat Politècnica de Catalunya*, 2008.
- [EGI21] Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67:4986–5004, 2021.
- [ERO21] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [FAFZ12] Thibaut Fux, Véronique Aubergé, Gang Feng, and Véronique Zimpfer. Speaker’s prosodic strategy for a large physical distance communication task. *Acoust. Soc. Am*, 45(1):47–53, 2012.
- [FGD18] William Fedus, Ian Goodfellow, and Andrew M Dai. MaskGAN: Better text generation via filling in the .. In *Proceedings of the International Conference on Learning Representations*, 2018.
- [Fol99] Gerald B. Folland. *Real analysis: Modern Techniques and Their Applications*. Wiley, New York, 1999.
- [G+07] Peter D Grünwald et al. The minimum description length principle. *MIT Press Books*, 1, 2007.
- [GAA+17a] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 5769–5779, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [GAA+17b] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.
- [GAD+17] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *Advances in neural information processing systems*, 30, 2017.
- [GBR+12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.
- [GCW+21] Qing Guo, Junya Chen, Dong Wang, Yuewei Yang, Xinwei Deng, Lawrence Carin, Fan Li, and Chenyang Tao. Tight mutual information estimation with contrastive fenchel-legendre optimization. *CoRR*, abs/2107.01131, 2021.
- [GFGS06] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.
- [GKS14] Elizabeth Godoy, Maria Koutsogiannaki, and Yannis Stylianou. Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles. *Computer Speech & Language*, 28(2):629–647, 2014.

- [GL84] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.
- [Gla04] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Applications of mathematics : stochastic modelling and applied probability. Springer, 2004.
- [GPAM⁺14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [GPM⁺14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative Adversarial Nets. In *Proceedings of Annual Conference on Neural Information Processing Systems (NIPS '14)*, pages 2672–2680, 2014.
- [GS86] Sandra Gordon-Salant. Recognition of natural and time/intensity altered cvs by young and elderly subjects with normal hearing. *The Journal of the Acoustical Society of America*, 80(6):1599–1607, 1986.
- [GS87] Sandra Gordon-Salant. Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects. *The Journal of the Acoustical Society of America*, 81(4):1199–1202, 1987.
- [GS05] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE, 2005.
- [GS12] Elizabeth Godoy and Yannis Stylianou. Unsupervised acoustic analyses of normal and lombard speech, with spectral envelope transformation to improve intelligibility. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [GSA⁺15] Mostafa Ghorbandoost, Abolghasem Sayadiyan, Mohsen Ahangar, Hamid Sheikhzadeh, Abdoreza Sabzi Shahrehabaki, and Jamal Amini. Voice conversion based on feature combination with limited training data. *Speech Communication*, 67:113–128, 2015.
- [GSW⁺21] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*, 35(4):3313–3332, 2021.
- [HFLM⁺19] Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR 2019*. ICLR, April 2019.
- [HHW⁺16] Chin Cheng Hsu, Hsin Te Hwang, Yi Chiao Wu, Yu Tsao, and Hsin Min Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6. IEEE, 2016.
- [HHW⁺17] Chin Cheng Hsu, Hsin Te Hwang, Yi Chiao Wu, Yu Tsao, and Hsin Min Wang. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint arXiv:1704.00849*, 2017.
- [HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [HJC⁺17] R. D. Hjelm, A. P. Jacob, T. Che, A. Trischler, K. Cho, and Y. Bengio. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*, 2017.
- [HJC⁺18] R Devon Hjelm, Athul Paul Jacob, Tong Che, Adam Trischler, Kyunghyun Cho, and Yoshua Bengio. Boundary-seeking generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

- [HLLR⁺16] JM Hernández-Lobato, Y Li, M Rowland, D Hernández-Lobato, TD Bui, and RE Turner. Black-box α -divergence minimization. In *Proceedings of the International Conference on Machine Learning*, volume 48, pages 1511–1520, 2016.
- [HM04] Valerie Hazan and Duncan Markham. Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, 116(5):3108–3118, 2004.
- [HMW⁺19] Qiong Hu, Erik Marchi, David Winarsky, Yannis Stylianou, Devang Naik, and Sachin Kajarekar. Neural text-to-speech adaptation from low quality public recordings. In *Speech Synthesis Workshop*, volume 10, 2019.
- [HRU⁺17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.
- [HS85] Tammo Houtgast and Herman JM Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77(3):1069–1077, 1985.
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [HS98] Valerie Hazan and Andrew Simpson. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24(3):211–226, 1998.
- [HSTD20] Ting Yao Hu, Ashish Shrivastava, Oncel Tuzel, and Chandra Dhir. Unsupervised style and content separation by minimizing mutual information for speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3267–3271. IEEE, 2020.
- [HTK⁺17] Tomoki Hayashi, Akira Tamamori, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda. An investigation of multi-speaker training for wavenet vocoder. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 712–718, 2017.
- [HYSX17] Z. Hu, Z. Yang, R. Salakhutdinov, and E. P. Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017.
- [HYSX18] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P. Xing. On unifying deep generative models. In *Proceedings of the International Conference on Learning Representations*, 2018.
- [HZW⁺18] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, et al. Hierarchical generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1810.07217*, 2018.
- [IS15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [IZZE17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [JSS16] EP Jayakumar, PV Muhammed Shifas, and PS Sathidevi. Integrated acoustic echo and noise suppression in modulation domain. *International Journal of Speech Technology*, 19(3):611–621, 2016.
- [Jun96] Jean-Claude Junqua. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex. *Speech communication*, 20(1-2):13–22, 1996.

- [JZW⁺18] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*, pages 4480–4490, 2018.
- [KAHK17] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.
- [KALL17] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [KALL20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Training generative adversarial networks with limited data. In *Proceedings of the International Conference on Learning Representations*, 2020.
- [KB04a] John Kominek and Alan W Black. The cmu arctic speech databases. In *Fifth ISCA workshop on speech synthesis*, 2004.
- [KB04b] Jean C Krause and Louis D Braid. Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, 115(1):362–378, 2004.
- [KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [KDS96] Nina Kowalski, Didier A Depireux, and Shihab A Shamma. Analysis of dynamic spectra in ferret primary auditory cortex. ii. prediction of unit responses to arbitrary dynamic spectra. *Journal of Neurophysiology*, 76(5):3524–3534, 1996.
- [Kei17] Keithito. The LJspeech dataset. <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [KES⁺18] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419, 2018.
- [KH09] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- [Kin11] Simon King. An introduction to statistical parametric speech synthesis. *Sadhana*, 36(5):837–852, 2011.
- [KK17] Takuhiro Kaneko and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293*, 2017.
- [KKdB⁺19] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. MELGAN: Generative adversarial networks for conditional waveform synthesis. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 14881–14892, 2019.
- [KKH⁺18] Takuhiro Kaneko, Hirokazu Kameoka, Kazuhiro Hiramatsu, Kunio Kashino, and Kazuyoshi Tanaka. Cyclegan-vc2: Improved cycle-consistent adversarial networks for non-parallel voice conversion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5274–5278. IEEE, 2018.
- [KKH⁺19] Hirokazu Kameoka, Takuhiro Kaneko, Kazuhiro Hiramatsu, Kunio Kashino, and Kazuyoshi Tanaka. Acvae-vc: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5279–5283, 2019.

- [KKH⁺20] Takuhiro Kaneko, Hirokazu Kameoka, Kazuhiro Hiramatsu, Kunio Kashino, and Kazuyoshi Tanaka. Cyclegan-vc3: Examining and improving the cyclegan-vc for mel-spectrogram conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:370–381, 2020.
- [KKHK17] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks. In *Interspeech*, volume 2017, pages 1283–1287, 2017.
- [KKTH18] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks. *arXiv preprint arXiv:1806.02169*, 2018.
- [KLA19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [KM82] Raymond D Kent and Ann D Murray. Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *The Journal of the Acoustical Society of America*, 72(2):353–365, 1982.
- [KM98a] A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 285–288. IEEE, 1998.
- [KM98b] Alexander Kain and Michael W Macon. Spectral voice conversion for text-to-speech synthesis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 1, pages 285–288. IEEE, 1998.
- [KMKDC99] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, 27(3):187–207, 1999.
- [KPH⁺20] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.
- [Kra01] Jean Christine Krause. Properties of naturally produced clear speech at normal rates and implications for intelligibility enhancement. 2001.
- [Kre89] David M Kreps. Nash equilibrium. In *Game Theory*, pages 167–177. Springer, 1989.
- [KRRD19] Viacheslav Klimkov, Srikanth Ronanki, Jonas Rohnke, and Thomas Drugman. Fine-grained robust prosody transfer for single-speaker neural text-to-speech. *arXiv preprint arXiv:1907.02479*, 2019.
- [KS14] M. Koutsogiannaki and Y. Stylianou. Simple and artefact-free spectral modifications for enhancing the intelligibility of casual speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4648–4652, 2014.
- [KS16] Maria Koutsogiannaki and Yannis Stylianou. Modulation enhancement of temporal envelopes for increasing speech intelligibility in noise. 2016.
- [KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [KXRS19] He Kaiming, Saining Xie, Shaoqing Ren, and Jian Sun. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.
- [LB03] Jacqueline S Laures and Kate Bunton. Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. *Journal of communication disorders*, 36(6):449–464, 2003.

- [LBBH98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [LC09] Youyi Lu and Martin Cooke. The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51(12):1253–1262, 2009.
- [LCK⁺20] Songxiang Liu, Yuewen Cao, Shiyin Kang, Na Hu, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. Transferring source style in non-parallel voice conversion. *arXiv preprint arXiv:2005.09178*, 2020.
- [LDY13] Zhen-Hua Ling, Li Deng, and Dong Yu. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE transactions on audio, speech, and language processing*, 21(10):2129–2139, 2013.
- [LeC98] Y. LeCun. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [Lev92] Willem JM Levelt. Accessing words in speech production: Stages, processes and representations. *Cognition*, 42(1-3):1–22, 1992.
- [LK19] Younggun Lee and Taesu Kim. Robust and fine-grained prosody control of end-to-end speech synthesis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5911–5915. IEEE, 2019.
- [LKHS20] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.
- [LLJ⁺18] Li Juan Liu, Zhen Hua Ling, Yuan Jiang, Ming Zhou, and Li-Rong Dai. WaveNet vocoder with limited training data for voice conversion. In *Proc. Interspeech*, pages 1983–1987, 2018.
- [LLL⁺19] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713, 2019.
- [LLY⁺18] Yanping Li, Kong Aik Lee, Yougen Yuan, Haizhou Li, and Zhen Yang. Many-to-many voice conversion based on bottleneck features with variational autoencoder for non-parallel training data. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 829–833. IEEE, 2018.
- [LPM15] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.
- [LRS10] Tony Lelièvre, Mathias Roussel, and Gabriel Stoltz. Free energy computations: A mathematical perspective. *World Scientific*, 2010.
- [LSB⁺15] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El Ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe’er, and Garry P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, 2015.
- [LT16] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.
- [LTDL⁺19] Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal. Towards achieving robust universal neural vocoding. In *Proc. Interspeech*, pages 4879–4883, 2019.
- [Lu09] Y Lu. Production and perceptual analysis of lombard effect. *Department of Computer Science, The University of Sheffield (Ph. D. thesis)*, 2009.

- [LWY19] Han Liu, Yiming Wu, and Yuxin Yang. Deep learning for extreme multi-label text classification. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 736–745, 2019.
- [LYXX21] Tao Li, Shan Yang, Liუმeng Xue, and Lei Xie. Controllable emotion transfer for end-to-end speech synthesis. In *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE, 2021.
- [LZ06] Sheng Liu and Fan-Gang Zeng. Temporal properties in clear speech perception. *The Journal of the Acoustical Society of America*, 120(1):424–432, 2006.
- [M97] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.
- [M⁺05] Tom Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.
- [MC90] Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467, 1990.
- [MCW⁺16] Thomas Merritt, Robert AJ Clark, Zhizheng Wu, Junichi Yamagishi, and Simon King. Deep neural network-guided unit selection synthesis. In *Proc. ICASSP*, pages 5145–5149, 2016.
- [ME88] Allen A Montgomery and Rodney A Edge. Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults. *Journal of Speech, Language, and Hearing Research*, 31(3):386–393, 1988.
- [MHM18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [MK14] Seyed Hamidreza Mohammadi and Alexander Kain. Voice conversion using deep neural networks with speaker-independent pre-training. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 19–23. IEEE, 2014.
- [MKG⁺17] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. SampleRNN: An unconditional end-to-end neural audio generation model. In *International Conference on Learning Representations*, 2017.
- [MKKY18] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.
- [MKS12] Seyed Hamidreza Mohammadi, Alexander Kain, and Jan PH van Santen. Making conversational vowels more clear. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [MLG⁺18] Zhong Meng, Jinyu Li, Yifan Gong, et al. Cycle-consistent speech enhancement. *arXiv preprint arXiv:1809.02253*, 2018.
- [MLX⁺17] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.
- [MMS19] Shuang Ma, Daniel Mcduff, and Yale Song. A generative adversarial network for style modeling in a text-to-speech system. In *International Conference on Learning Representations*, volume 2, 2019.
- [MNG18] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Which training methods for gans do actually converge? *International Conference on Machine Learning*, pages 3478–3487, 2018.

- [MO14] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [MPP18] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717, 2018.
- [MQ86] Robert McAulay and Thomas Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.
- [MS05] Nima Mesgarani and Shihab Shamma. Speech enhancement based on filtering the spectrotemporal modulations. In *Proceedings.(ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–1105. IEEE, 2005.
- [MS20] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *Proceedings of Machine Learning Research*, pages 875–884. PMLR, 26–28 Aug 2020.
- [MSK17] Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks. In *2017 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 134–140. IEEE, 2017.
- [MSTS17] Hiroyuki Miyoshi, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Voice conversion using sequence-to-sequence learning of context posterior probabilities. *arXiv preprint arXiv:1704.02360*, 2017.
- [MYO16] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.
- [NCT16] S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.
- [NG76] R Niederjohn and J Grotelueschen. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):277–282, 1976.
- [NTA14a] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion. In *Fifteenth annual conference of the international speech communication association*, 2014.
- [NTA14b] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. Voice conversion based on speaker-dependent restricted boltzmann machines. *IEICE TRANSACTIONS on Information and Systems*, 97(6):1403–1410, 2014.
- [NTSS12] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech. *Speech Communication*, 54(1):134–146, 2012.
- [NTTA13] Toru Nakashika, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Voice conversion in high-order eigen space using deep belief nets. In *Interspeech*, pages 369–372, 2013.
- [NWJ10] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

- [ODZ⁺16] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- [OLB⁺18a] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR, 2018.
- [OLB⁺18b] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel WaveNet: Fast high-fidelity speech synthesis. In *International Conference on Machine Learning*, pages 3918–3926, 2018.
- [OOS16] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. *arXiv preprint arXiv:1610.09585*, 2016.
- [OR94] M. J. Osborne and A. Rubinstein. *A course in game theory*. MIT press, 1994.
- [Pau81] D Paul. The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(4):786–794, 1981.
- [PBS13] Nathanaël Perraudin, Peter Balazs, and Peter L Søndergaard. A fast Griffin-Lim algorithm. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, 2013.
- [PBS17] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.
- [PDB86] Michael A Picheny, Nathaniel I Durlach, and Louis D Braida. Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4):434–446, 1986.
- [PDD14] Benjamin Picart, Thomas Drugman, and Thierry Dutoit. Analysis and HMM-based synthesis of hypo and hyperarticulated speech. *Computer Speech & Language*, 28(2):687–707, 2014.
- [PMPS21] Dipjyoti Paul, Sankar Mukherjee, Yannis Pantazis, and Yannis Stylianou. A Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning Based on Rényi Divergence Minimization. In *Proc. Interspeech 2021*, pages 3625–3629, 2021.
- [POVDO⁺19] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR, 09–15 Jun 2019.
- [PPC19] Wei Ping, Kainan Peng, and Jitong Chen. ClariNet: Parallel wave generation in end-to-end text-to-speech. In *International Conference on Learning Representations*, 2019.
- [PPF⁺20] Yannis Pantazis, Dipjyoti Paul, Michail Fasoulakis, Yannis Stylianou, and Markos Katsoulakis. Cumulant GAN. *arXiv preprint arXiv:2006.06625*, 2020.
- [PPFS19] Yannis Pantazis, Dipjyoti Paul, Michail Fasoulakis, and Yannis Stylianou. Training generative adversarial networks with weights. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2019.
- [PPFSed] Yannis Pantazis, Dipjyoti Paul, Michail Fasoulakis, and Yannis Stylianou. Training generative adversarial networks with weights. in *European Signal Processing Conference, EUSIPCO*, 2019 (accepted).

- [PPG⁺17] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*, 2017.
- [PPS19] Dipjyoti Paul, Yannis Pantazis, and Yannis Stylianou. Non-parallel voice conversion using weighted generative adversarial networks. In *Proc. Interspeech*, pages 659–663, 2019.
- [PPS20] Dipjyoti Paul, Yannis Pantazis, and Yannis Stylianou. Speaker conditional WaveRNN: Towards universal neural vocoder for unseen speaker and recording conditions. *Proc. Interspeech 2020*, pages 235–239, 2020.
- [PPSed] Dipjyoti Paul, Yannis Pantazis, and Yannis Stylianou. Speaker conditional WaveRNN: Towards universal neural vocoder for unseen speaker and recording conditions. In *Proc. Interspeech*, 2020 (accepted).
- [PVC19] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: A flow-based generative network for speech synthesis. In *Proc. ICASSP*, pages 3617–3621, 2019.
- [PY09] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.
- [PZPP19] Jihyun Park, Kexin Zhao, Kainan Peng, and Wei Ping. Multi-speaker end-to-end speech synthesis. *arXiv preprint arXiv:1907.04462*, 2019.
- [QSY12] Yao Qian, Frank K Soong, and Zhi-Jie Yan. A unified trajectory tiling approach to high quality speech rendering. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):280–290, 2012.
- [RC05] Christian P Robert and George Casella. *Monte Carlo statistical methods; 2nd ed.* Springer texts in statistics. Springer, Berlin, 2005.
- [RHQ⁺20] Yi Ren, Chenxu Hu, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text-to-speech. *arXiv preprint arXiv:2006.04558*, 2020.
- [RLL89] Liselotte Roug, Ingrid Landberg, and L-J Lundberg. Phonetic development in early infancy: A study of four swedish children during the first eighteen months of life. *Journal of child language*, 16(1):19–40, 1989.
- [RMC15] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [Roc70] R.T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970.
- [RRT⁺19] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.
- [RSVA11] Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku. Analysis of HMM-based Lombard speech synthesis. In *Proc. Interspeech*, 2011.
- [RVD09] Koenraad S Rhebergen, Niek J Versfeld, and Wouter A Dreschler. The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise. *The Journal of the Acoustical Society of America*, 126(6):3236–3245, 2009.
- [SB06] Mitchell S Sommers and Joe Barcroft. Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America*, 119(4):2406–2416, 2006.
- [SC21] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3):1–42, 2021.

- [SCM98a] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *Speech and Audio Processing, IEEE Transactions on*, 6(2):131–142, 1998.
- [SCM98b] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on speech and audio processing*, 6(2):131–142, 1998.
- [SE19] Jiaming Song and Stefano Ermon. Understanding the Limitations of Variational Mutual Information Estimators. *arXiv e-prints*, page arXiv:1910.06222, October 2019.
- [SE20] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020.
- [SFG⁺12] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electron. J. Statist.*, 6:1550–1599, 2012.
- [SGZ⁺16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.
- [SH06] Mark D Skowronski and John G Harris. Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, 48(5):549–558, 2006.
- [Sha20] Matt Shannon. The divergences minimized by non-saturating GAN training, 2020.
- [SINT18] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi. Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5274–5278. IEEE, 2018.
- [SJA⁺19] Shreyas Seshadri, Lauri Juvela, Paavo Alku, Okko Räsänen, et al. Augmented CycleGANs for continuous scale normal-to-lombard speaking style conversion. *Proc. Interspeech 2019*, pages 2838–2842, 2019.
- [SJRA19] Shreyas Seshadri, Lauri Juvela, Okko Räsänen, and Paavo Alku. Vocal effort based speaking style conversion using vocoder features and parallel learning. *IEEE Access*, 7:17230–17246, 2019.
- [SJS18] PV Muhammed Shifas, EP Jayakumar, and PS Sathidevi. Robust acoustic echo suppression in modulation domain. In *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, pages 527–537. Springer, 2018.
- [SJY⁺19] Shreyas Seshadri, Lauri Juvela, Junichi Yamagishi, Okko Räsänen, and Paavo Alku. Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6835–6839, 2019.
- [SKLM15] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4869–4873. IEEE, 2015.
- [SMAMG21] Vignesh Sampath, Iñaki Murtua, Juan Jose Aguilar Martin, and Aitor Gutierrez. A survey on generative adversarial networks for imbalance problems in computer vision tasks. *Journal of big Data*, 8:1–59, 2021.
- [SMK⁺17] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

- [SNA91] Kiyohiro Shikano, Satoshi Nakamura, and Masanobu Abe. Speaker adaptation and voice conversion by codebook mapping. In *1991 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 594–597. IEEE, 1991.
- [SPW⁺18] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.
- [SRBX⁺18] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In *Proc. international conference on machine learning*, pages 4693–4702. PMLR, 2018.
- [Sri13] R. Srinivasan. *Importance Sampling: Applications in Communications and Detection*. Springer Berlin Heidelberg, 2013.
- [SSCS20] Muhammed PV Shifas, Anna Sfakianaki, Theognosia Chimona, and Yannis Stylianou. Evaluating the intelligibility benefits of neural speech enrichment for listeners with normal hearing and hearing impairment using the greek harvard corpus. *arXiv preprint arXiv:2011.06548*, 2020.
- [Sta80] Rachel E Stark. Stages of speech development in the first year of life. In *Child phonology*, pages 73–92. Elsevier, 1980.
- [STN20] P. Aditya Sreekar, Ujjwal Tiwari, and Anoop M. Namboodiri. Reducing the variance of variational estimates of mutual information by limiting the critic’s hypothesis space to RKHS. In *ICPR*, pages 10666–10674. IEEE, 2020.
- [STS18] Y. Saito, S. Takamichi, and H. Saruwatari. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):84–96, 2018.
- [SZ09] Marc Schönwiesner and Robert J Zatorre. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fmri. *Proceedings of the National Academy of Sciences*, 106(34):14611–14616, 2009.
- [SZDL19] Berrak Sisman, Mingyang Zhang, Minghui Dong, and Haizhou Li. On the study of generative adversarial networks for cross-lingual voice conversion. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 144–151. IEEE, 2019.
- [SZL18] Berrak Sisman, Mingyang Zhang, and Haizhou Li. A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder. In *Interspeech*, pages 1978–1982, 2018.
- [SZZ⁺17] K. Schawinski, C. Zhang, H. Zhang, L. Fowler, and G. K. Santhanam. Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit. *Monthly Notices of the Royal Astronomical Society: Letters*, 467(1):L110–L114, 2017.
- [Tay09] Paul Taylor. *Text-to-speech synthesis*. Cambridge University Press, 2009.
- [THK⁺17] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda. Speaker-dependent WaveNet vocoder. In *Proc. Interspeech*, pages 1118–1122, 2017.
- [THZL21] Daxin Tan, Hingpang Huang, Guangyan Zhang, and Tan Lee. CUHK-EE voice cloning system for ICASSP 2021 M2VoC challenge. *arXiv preprint arXiv:2103.04699*, 2021.
- [TKKH18] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo. Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. *arXiv preprint arXiv:1811.04076*, 2018.

- [TKKH19] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo. Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6805–6809. IEEE, 2019.
- [TM06] Mark Tatham and Katherine Morton. *Speech production and perception*. Springer, 2006.
- [TNS12] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2505–2517, 2012.
- [TPZK20] Andros Tjandra, Ruoming Pang, Yu Zhang, and Shigeki Karita. Unsupervised learning of disentangled speech content and style representation. *arXiv preprint arXiv:2010.12973*, 2020.
- [Tsy08] Alexandre B Tsybakov. Introduction to nonparametric estimation. *Springer Science & Business Media*, 2008.
- [TUA18] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4784–4788. IEEE, 2018.
- [TWH⁺19] Patrick Lumban Tobing, Yi-Chiao Wu, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda. Voice conversion with cyclic recurrent neural network and fine-tuned wavenet vocoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6815–6819. IEEE, 2019.
- [UKB07] Maria Uther, Monja A Knoll, and Denis Burnham. Do you speak e-ng-li-sh? a comparison of foreigner-and infant-directed speech. *Speech communication*, 49(1):2–7, 2007.
- [VBYKS13] Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King, and Yannis Stylianou. Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of hmm-based synthetic speech in noise. In *Proc. Interspeech*, pages 3567–3571, 2013.
- [vdODZ⁺16] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *9th ISCA Speech Synthesis Workshop*, pages 125–125, 2016.
- [vdOKE⁺16] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.
- [vdOLV18] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.
- [VKKH18] Steven Van Kuyk, W Bastiaan Kleijn, and Richard Christian Hendriks. An evaluation of intrusive instrumental intelligibility metrics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2153–2166, 2018.
- [VL19] Sean Vasquez and Mike Lewis. MelNet: A generative model for audio in the frequency domain. *arXiv preprint arXiv:1906.01083*, 2019.
- [VMT92] H. Valbret, E. Moulines, and J.P. Tubach. Voice transformation using PSOLA technique. In *ICASSP*, volume 1, pages 145–148. IEEE, 1992.
- [VOKK16] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR, 2016.

- [VS19] Jean Marc Valin and Jan Skoglund. LPCnet: Improving neural speech synthesis through linear prediction. In *Proc. ICASSP*, pages 5891–5895, 2019.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.
- [VYM⁺16] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. 2016.
- [Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.
- [WBH⁺21] Liangjian Wen, Haoli Bai, Lirong He, Yiji Zhou, Mingyuan Zhou, and Zenglin Xu. Gradient estimation of information measures in deep learning. *Knowledge-Based Systems*, 224:107046, 2021.
- [WBK⁺20] Eric Weitz, Yizhak Belinkov, Zohaib Khan, Brian Baucom, Jiampeng Qin, Zhou Yu, Furu Wang, Li Dai, Amanda Stent, Najim Dehak, et al. Extending for multi-speaker text-to-speech synthesis. *arXiv preprint arXiv:2010.05288*, 2020.
- [WC18] Wei Wang and Kyunghyun Cho. Data augmentation for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 567–578, 2018.
- [WL12] Kamil K Wójcicki and Philipos C Loizou. Channel selection in the modulation domain for improved speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 131(4):2904–2913, 2012.
- [WLL⁺19] Pengfei Wu, Zhenhua Ling, Lijuan Liu, Yuan Jiang, Hongchuan Wu, and Lirong Dai. End-to-end emotional speech synthesis using style tokens and semi-supervised training. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 623–627. IEEE, 2019.
- [WS08] Peter J Watson and Robert S Schlauch. The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours. *American Journal of Speech-Language Pathology*, 2008.
- [WSRS⁺17] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint:1703.10135*, 2017.
- [WSZ⁺18] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. PMLR, 2018.
- [WVCL14] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li. Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1506–1521, 2014.
- [WWPM18] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *Proc. ICASSP*, pages 4879–4883, 2018.
- [WXX⁺16] Wenfu Wang, Shuang Xu, Bo Xu, et al. First step towards end-to-end parametric tts synthesis: Generating spectral parameters with neural attention. In *Interspeech*, pages 2243–2247, 2016.
- [WZH⁺20] Liangjian Wen, Yiji Zhou, Lirong He, Mingyuan Zhou, and Zenglin Xu. Mutual information gradient estimation for representation learning. In *International Conference on Learning Representations*, 2020.

- [YHC⁺18] Cheng-chieh Yeh, Po-chun Hsu, Ju-chieh Chou, Hung-yi Lee, and Lin-shan Lee. Rhythm-flexible voice conversion without parallel data using cycle-gan over phoneme posterior-gram sequences. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 274–281. IEEE, 2018.
- [YSK20] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE, 2020.
- [ZKS12] Tudor-Catalin Zorila, Varvara Kandia, and Yannis Stylianou. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [ZLD18] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai. Forward attention in sequence-to-sequence acoustic modeling for speech synthesis. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 4789–4793. IEEE, 2018.
- [ZLL⁺19] Jingxuan Zhang, Zhenhua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai. Sequence-to-sequence acoustic modeling for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.
- [ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.
- [ZS14] Tudor Cătălin Zorilă and Yannis Stylianou. On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement. In *Proc. Interspeech*, 2014.
- [ZSFM17] Tudor-Cătălin Zorilă, Yannis Stylianou, Sheila Flanagan, and Brian CJ Moore. Evaluation of near-end speech enhancement under equal-loudness constraint for listeners with normal-hearing and mild-to-moderate hearing loss. *The Journal of the Acoustical Society of America*, 141(1):189–196, 2017.
- [ZSL20] Kun Zhou, Berrak Sisman, and Haizhou Li. Transforming spectrum and prosody for emotional voice conversion with non-parallel training data. *arXiv preprint arXiv:2002.00198*, 2020.
- [ZTB09] Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.
- [ZXL⁺19] Han Zhang, Ian J Xu, Guanhua Li, Tongyuan Zhang, Qiuyue Wang, and Xiaowei Huang. Self-attention generative adversarial networks. *International Conference on Learning Representations*, 2019.
- [ZZP⁺17] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.