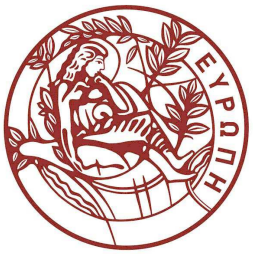


University of Crete
Department of Computer Science

University of Rennes 1

Adaptive Sinusoidal Models for Speech with Applications in Speech Modifications and Audio Analysis

Ph.D. Thesis

George P. Kafentzis

Heraklion
June 2014

UNIVERSITY OF CRETE
SCHOOL OF SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE

Adaptive Sinusoidal Models for Speech with Applications in Speech Modifications and Audio Analysis

Submitted by

George P. Kafentzis

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

Author:

George P. Kafentzis
Department of Computer Science

Examination Committee:

Supervisor

Yannis Stylianou, Professor, University of Crete

Member

Olivier Boeffard, Professor, University of Rennes 1

Member

Apostolos Traganitis, Professor, University of Crete

Member

Athanasios Mouchtaris, Assistant Professor, University of Crete

Member

Georgios Tziritas, Professor, University of Crete

Member

Olivier Rosec, Chief Researcher, France Telecom

Member

Régine Le Bouquin-Jeannes, Professor, University of Rennes 1

Departmental Approval:

Chairman
of the Department

Panagiotis Trahanias, Professor, University of Crete

Heraklion, June 2014

Acknowledgements

During this 4-year Ph.D. “quest”, I had the opportunity to meet and work with exceptional people and scientists who helped me evolve my work and my personality. First of all, I would like to thank my supervisor, Professor Yannis Stylianou, for his continuous support, encouragement, guidance, motivation, and above all, trust and patience that he showed during the time we worked together. His advice made me a better researcher, more persistent when things were not going well, and he taught me how to view things from different perspectives. However, the most important and emotive for me was his help and care on matters beyond academia, and for this I thank him twice.

At this point, I also have to thank my thesis supervisor on behalf of the University of Rennes 1, Professor Olivier Boeffard, and the member of my thesis jury, Dr. Olivier Rosec, former senior researcher in Orange Labs - France Telecom R & D, and currently a R & D director at Voxygen S.A., who gave me the opportunity to work with them for one year in Orange Labs - France Telecom R & D, Lannion, France. They both gave me important support and advice during the completion of this thesis.

In addition, I should not neglect to thank Assist. Professor Athanasios Mouchtaris, Professor Georgios Tziritas, and Professor Apostolos Traganitis, members of faculty of the Computer Science Department, University of Crete, and Professor Régine Le Bouquin-Jeannes, from the University of Rennes 1, for spending their valuable time on reading this thesis and for accepting to be members of my thesis jury. I also thank Professor Gernot Kubin, from the Technological University of Graz, Austria, and Professor W. Bastiaan Kleijn, from the University of Wellington, New Zealand, for accepting to be reporters of my thesis. I should not forget to thank Dr. Yannis Pantazis, for helping me in the beginning of this work.

For the past four years of the Ph.D. programme, I shared my days at the University of Crete with colleagues who made it a wonderful workplace and home, and helped me in so many ways. I would specially like to thank Pavlos Mattheakis, Maria Koutsogiannaki, Christos and George Tzagkarakis, Vangelis Vasilakis, Marcelo Caetano, and Gilles Degottex for standing by me and for being real friends.

Furthermore, I would like to thank the graduate students of the Multimedia Informatics Laboratory, Sofia Yanikaki, Olina Simantiraki, Veronica Morfi, and Dora Yakoumaki for sharing their working time with me during their studies. Moreover, I would like to say “thank you” to a few others, including Nikoletta Palivakou, Evgenios Kornaropoulos, Chris Margiolas, Petros Androvitsaneas, Kostas Sipsas, Pepi Katsiyanni, and Kostas Katsaros for their care and friendship, even if some of them are away from Crete.

Last but not least, the greatest “thank you” to my family: my parents, Panayotis Kafentzis and Diamanto Tsirolia, for everything: I owe them the man I am today. My sister Maria-Chrysanthi for showing incredible patient and care, and for bringing the love of our family all these years we were together in Crete. My little brother Stelios for giving me strength to begin every next day by listening his voice almost every night. Finally, my wife Anna, for her extreme patience and love, and for bringing in life our most valuable gift, our son, Panos.

Thank you all!

Remerciements

Au cours de cette 4 ans Ph.D. “quête”, j’ai eu l’occasion de rencontrer et de travailler avec des gens exceptionnels et des scientifiques qui m’a aidé à évoluer mon travail et ma personnalité. Tout d’abord, je tiens à remercier mon superviseur, le professeur Yannis Stylianou, pour son soutien continu, encouragements, des conseils, de la motivation, et surtout, la confiance et la patience qu’il a montré pendant le temps nous avons travaillé ensemble. Son conseil m’a fait quand une meilleure chercheur, plus persistante les choses n’allaient pas bien, et il m’a appris à voir les choses sous des angles différents. Cependant, la plupart importante et émotive pour moi était son aide et les soins qui dépassent le milieu universitaire, et pour cela je le remercie deux fois.

A ce point, je dois aussi remercier mon directeur de thèse, au nom de l’Université de Rennes 1, le professeur Olivier Boeffard, et le Dr. Olivier Rosec, ancien rechercheur principal de Orange Labs - France Télécom R & D, et a un directeur R & D chez Voxygen S.A., qui m’a donné l’occasion de travailler avec eux pendant un an dans les Orange Labs - France Télécom R & D, Lannion, France. Ils ont tous deux m’ont donné un soutien et des conseils importants lors de la réalisation de cette thèse.

En outre, je ne devrais pas négliger à remercier Assoc. Professeur Athanasios Mouchtaris, Professeur Georgios Tziritas, et professeur Apostolos Traganitis, les membres du corps professoral de Département d’Informatique, Université de Crète, et professeure Régine Le Bouquin-Jeannes, membre de Université de Rennes 1, pour dépenser leur temps précieux à la lecture de cette thèse et pour avoir acceptè d’ être membres de mon jury de thèse. Je remercie ègalement le Professeur Gernot Kubin, de l’Université Technologique de Graz, en Autriche, et le professeur W. Bastiaan Kleijn, de l’Université de Wellington, Nouvelle-Zélande, pour avoir accepté d’être des rapporteurs de ma thèse. Je ne devrais pas oublier de remercier Dr. Yannis Pantazis, pour m’aider dans le début de ce travail.

Pour les quatre dernières années du doctorat programme, j’ai partagé mes jours à l’Université de Crète avec des collègues qui a fait un merveilleux travail et la maison, et ils m’ont aidé à bien des égards. Je voudrais spécialement merci Pavlos Mattheakis, Maria Koutsogiannaki, Christos et George Tzagkarakis, Vangelis Vasilakis, Marcelo Caetano, et Gilles Degottex pour debout par moi et d’être de vrais amis.

En outre, je tiens à remercier les étudiants des cycles supérieurs de l’ informatique de laboratoire multimédia, Sofia Yanikaki, Olina Simantiraki, Veronica Morfi, et Dora Yakoumaki pour partager leur temps de travail avec moi pendant leurs études. En outre, je tiens à dire “merci” à quelques autres, y compris Nikoletta Palivakou, Evgenios Kornaropoulos, Chris Margiolas, Petros Androvitsaneas, Kostas Sipas, Pepi Katsiyanni, et Kostas Katsaros pour leurs soins et leur amitié, même si certains d’entre eux se trouvent de Crète.

Enfin et surtout, le plus grand “merci” à ma famille : mes parents, Panayotis Kafentzis et Diamanto Tsirolia, pour tout : je leur dois l’homme que je suis aujourd’hui. Ma soeur Maria - Chrysanthi pour montrer patient et soins incroyables, et pour apporter l’amour de notre famille pendant toutes ces années, nous étions ensemble en Crète. Mon petit frère Stelios pour me donner la force de commencer chaque lendemain en écoutant sa voix presque chaque nuit dernière. Enfin, mon èpouse Anna, pour son extrême patience et d’amour, et d’avoir dans la vie de notre cadeau le plus prècieux, notre fils, Panos.

Merci à tous !

Ευχαριστίες

Κατά τη διάρκεια αυτής της 4χρονης "περιπέτειας "της Διδακτορικής μου διατριβής, είχα την ευκαιρία να γνωρίσω και να συνεργαστώ με εξαιρετικούς ανθρώπους και επιστήμονες, που με βοήθησαν να εξελίξω τη δουλειά μου και την προσωπικότητά μου. Πρώτον απ΄ όλους, θα ήθελα να ευχαριστήσω τον επόπτη μου, Καθηγητή Ιωάννη Στυλιανού, για τη συνεχή υποστήριξή του κατά τη διάρκεια του προγράμματος μεταπτυχιακών σπουδών. Είμαι ειλικρινά ευγνώμων για τις συμβουλές, την ενθάρρυνση, την καθοδήγηση, το κίνητρο, και πάνω απ΄ όλα, την εμπιστοσύνη και την υπομονή που έδειξε κατά το διάστημα που δουλέψαμε μαζί. Επίσης, μου έμαθε πώς να βλέπω τα πράγματα από διαφορετικές οπτικές γωνίες, πώς να κάνω έρευνα, και πώς να είμαι επίμονος όταν τα πράγματα δεν πηγαίνουν καλά. Τέλος, η βοήθεια και η έγνοια του για θέματα εκτός του ακαδημαϊκού περιβάλλοντος ήταν πραγματικά συγκινητική, και γι΄ αυτό τον ευχαριστώ διπλά.

Επιπλέον, θα ήθελα να ευχαριστήσω τον καθηγητή Olivier Boeffard, που ήταν ο επόπτης της διατριβής μου εκ μέρους του Πανεπιστημίου της Rennes 1, καθώς και τον Δρ. Olivier Rosec, ερευνητή στα Orange Labs - France Telecom R & D - και τώρα Διευθυντή R & D στη Voxygen S.A. - με τους οποίους είχα την ευκαιρία να συνεργαστώ για ένα χρόνο, κατά τη διαμονή μου στις εγκαταστάσεις των Orange Labs - France Telecom R & D, στη Lannion της Γαλλίας. Αμφότεροι υπήρξαν σημαντικοί αρωγοί στην προσπάθειά μου για την ολοκλήρωση αυτής της διατριβής, με τις συμβουλές τους και την υποστήριξή τους.

Επιπροσθέτως, δε θα ήθελα να παραλείψω να ευχαριστήσω τους καθηγητές Αθανάσιο Μουχτάρη, Γιώργο Τζιρίτα, και Απόστολο Τραγανίτη, από το Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης, και την καθηγήτρια Régine Le Bouquin-Jeannes, από το Πανεπιστήμιο της Rennes 1, για το χρόνο που αφιέρωσαν στην ανάγνωση αυτής της διατριβής και για την αποδοχή τους να συμμετάσχουν στην εξεταστική επιτροπή. Θα ήθελα επίσης να ευχαριστήσω τον καθηγητή Gernot Kubin, από το Τεχνολογικό Πανεπιστήμιο του Graz της Αυστρίας, και τον καθηγητή W. Bastiaan Kleijn, από το Πανεπιστήμιο του Wellington της Νέας Ζηλανδίας, για το ότι δέχθηκαν να γράψουν αναφορά για τη διατριβή μου. Επίσης, θα ήθελα να ευχαριστησω το Δρ. Γιάννη Πανταζή, που μου προσέφερε πολύτιμη βοήθεια στην αρχή αυτής της διατριβής.

Επίσης, ευχαριστώ τους συναδέλφους μου, πρώην και νυν, με τους οποίους μοιράστηκα της μέρες μου στο Πανεπιστήμιο Κρήτης, και ειδικότερα στο Εργαστήριο Πολυμέσων του Τμήματος Επιστήμης Υπολογιστών. Η παρουσία τους το έκανε ένα υπέροχο εργασιακό περιβάλλον και ένα δεύτερο σπίτι για μένα, στα χρόνια που πέρασα εκεί. Θα ήθελα ιδιαίτερα να αναφέρω τους Παύλο Ματθαιάκη, Μαρία Κουτσογιαννάκη, Χρήστο και Γιώργο Τζαγκαράκη, Βαγγέλη Βασιλάκη, Marcelo Caetano και Gilles Deggotex, για τη βοήθειά τους με χίλιους δυο τρόπους, και για το ότι μου στάθηκαν πραγματικοί φίλοι.

Επίσης, θα ηθελα να ευχαριστήσω τις μεταπτυχιακές συναδέλφους - μέλη του Εργαστηρίου Πολυμέσων, Σοφία Γιαννικάκη, Ολίνα Σημαντηράκη, Βερόνικα Μόρφη, και Δώρα Γιακουμάκη, που μοιράστηκαν τον εργασιακό χρόνο τους μαζί μου κατά τη διάρκεια των σπουδών τους. Ακόμα, θα ήθελα να πω ένα "ευχαριστώ"σε πολλούς ακόμα, στους οποίους συμπεριλαμβάνονται οι Νικολέττα Παλιβάκου, Ευγένιος Κορναρόπουλος, Χρηστος Μαργιόλας, Κώστας Κατσάρος, Πέτρος Ανδροβιτσανέας, Κώστας Σιψάς, και Πέπη Κατσιγιάννη, για την έγνοια και τη φιλία τους, έστω κι αν κάποιοι απ΄ αυτούς είναι μακριά απ΄ την Κρήτη.

Τελευταίους αλλά όχι έσχατους, θα ήθελα να ευχαριστήσω θερμά την οικογένειά μου: τους γονείς μου, Παναγιώτη Καφεντζή και Διαμάντω Τσιρολιά, για τα πάντα. Τους χρωστάω αυτό που είμαι σήμερα. Την αδελφή μου, Μαρία-Χρυσάνθη, για την υπομονή, την αγάπη, την έγνοια, την ανεκτικότητα, και την υποστήριξή της, σε όλα αυτά τα χρόνια που είμαστε μαζί στην Κρήτη. Τον αδελφό μου, Στέλιο, για την αγάπη και την έγνοια του. Τέλος, τη σύζυγό μου Άννα, για την τεράστια υπομονή και αγάπη της, και για το ότι θα φέρει στη ζωή το πιο πολύτιμο δώρο μας, το γιο μας, Παναγιώτη.

Abstract

Sinusoidal Modeling is one of the most widely used parametric methods for speech and audio signal processing. The accurate estimation of sinusoidal parameters (amplitudes, frequencies, and phases) is a critical task for close representation of the analyzed signal. In this thesis, based on recent advances in sinusoidal analysis, we propose high resolution adaptive sinusoidal models for analysis, synthesis, and modifications systems of speech. Our goal is to provide systems that represent speech in a highly accurate and compact way.

Inspired by the recently introduced adaptive Quasi-Harmonic Model (aQHM) and adaptive Harmonic Model (aHM), we overview the theory of adaptive Sinusoidal Modeling and we propose a model named the extended adaptive Quasi-Harmonic Model (eaQHM), which is a non-parametric model able to adjust the instantaneous amplitudes and phases of its basis functions to the underlying time-varying characteristics of the speech signal, thus significantly alleviating the so-called local stationarity hypothesis. The eaQHM is shown to outperform aQHM in analysis and resynthesis of voiced speech. Based on the eaQHM, a hybrid analysis/synthesis system of speech is presented (eaQHNM), along with a hybrid version of the aHM (aHNM). Moreover, we present motivation for a full-band representation of speech using the eaQHM, that is, representing all parts of speech as high resolution AM-FM sinusoids. Experiments show that adaptation and quasi-harmonicity is sufficient to provide transparent quality in unvoiced speech resynthesis. The full-band eaQHM analysis and synthesis system is presented next, which outperforms state-of-the-art systems, hybrid or full-band, in speech reconstruction, providing transparent quality confirmed by objective and subjective evaluations.

Regarding applications, the eaQHM and the aHM are applied on speech modifications (time and pitch scaling). The resulting modifications are of high quality, and follow very simple rules, compared to other state-of-the-art modification systems. The concepts of relative phase and relative phase delays are crucial for the development of artefact-free, shape-invariant, high quality modifications. Results show that harmonicity is preferred over quasi-harmonicity in speech modifications due to the embedded simplicity of representation. Moreover, the full-band eaQHM is applied on the problem of modeling audio signals, and specifically of musical instrument sounds. The eaQHM is evaluated and compared to state-of-the-art systems, and is shown to outperform them in terms of resynthesis quality, successfully representing the attack, transient, and stationary part of a musical instrument sound. Finally, another application is suggested, namely the analysis and classification of emotional speech. The eaQHM is applied on the analysis of emotional speech, providing its instantaneous parameters as features that can be used in recognition and Vector-Quantization-based classification of the emotional content of speech. Although the sinusoidal models are not commonly used in such tasks, results are promising.

Résumé en français

La modélisation sinusoïdale est l'une des méthodes paramétriques les plus largement utilisées pour le traitement de la parole et du signal audio. L'estimation précise des paramètres sinusoïdaux (des amplitudes, des fréquences et des phases) est une tâche essentielle pour une représentation de haute qualité du signal analysé. Dans cette thèse, basée sur les avances récents de l'analyse sinusoïdale, nous proposons des modèles sinusoïdaux adaptatifs à haute résolution pour l'analyse, la synthèse et la transformation de la parole. Notre objectif est de fournir des systèmes qui représentent la parole d'une manière très précise et compacte.

Inspiré par le modèle adaptatif quasi-harmonique (aQHM) et le modèle adaptatif harmonique (aHM) récemment introduits, nous présentons la vue d'ensemble de la théorie de modéles sinusoïdaux adaptatifs et ensuite nous proposons le modèle étendu adaptif quasi-harmonique (eaQHM), qui est un modèle non-paramétrique capable d'ajuster les amplitudes et les phases instantanées de ses fonctions de base aux variations temporelles caractéristiques du signal de parole, ainsi qu'atténuer significativement l'hypothèse de stationnarité locale. On montre que la performance d'analyse et de resynthèse de la parole voisée de eaQHM est superieure à celle de aQHM. Un système hybride d'analyse et synthèse de la parole basé sur eaQHM est présenté (eaQHNM), ainsi qu'une version hybride de aHM (aHNM). Ensuite, nous présentons la motivation pour une représentation à bande pleine de la parole en utilisant eaQHM, c'est à dire, en représentant toutes les partiels de la parole avec des sinusoïdes AM-FM à haute résolution. Les expériences montrent que l'adaptation et la quasi-harmonicité sont suffisantes pour une représentation transparente de la parole synthétique non voisée. Le système eaQHM d'analyse et syntèse à bande pleine est présenté après. eaQHM surpasse l' etat de l'art des systèmes soit hybrides soit à bande pleine de reconstruction de la parole, offrant une qualité transparente confirmée par des évaluations objectives et subjectives.

En ce qui concerne les applications, le eaQHM et l'aHM sont appliqués sur les modifications de la parole (modification de durée ou de hauteur). Les modifications qui en résultent sont de haute qualité, et suivent des règles très simples, par rapport à d'autres systèmes de modification dans l'état de l'art. Les concepts de phase relative et les retards de phase relatifs sont cruciales pour le développement de modifications de haute qualité sans artefactes. Les résultats montrent que l'harmonicité est préféré à la quasi-harmonicité de modifications de la parole du fait de la simplicité de la représentation intégrée. En plus, eaQHM à bande pleine est appliqué à la modélisation des signaux audio, en particulier aux sons d'instruments de musique. La méthode eaQHM est évaluée et comparée avec l'état de l'art, avec une

performance supérieure en termes de qualité de resynthèse, représentant avec succès l'attaque, les transitoires, et la partie stationaire des sons d'instruments de musique. Enfin, une autre application est suggérée, l'analyse et la classification de la parole émotive. La eaQHM est appliquée à l'analyse de la parole émotive, offrant des paramètres instantanés qui peuvent être utilisés dans la reconnaissance et la classification à quantification vectorielle du contenu émotionnel de la parole. Bien que les modèles sinusoïdaux sont pas couramment utilisés dans ces tâches, les résultats sont prometteurs.

Περίληψη

Η Ημιτονοειδής Μοντελοποίηση είναι μια από τις πιο ευρέως χρησιμοποιούμενες παραμετρικές μεθόδους για την επεξεργασία σήματος φωνής και ήχου. Η ακριβής εκτίμηση των ημιτονοειδών παραμέτρων (πλάτη, συχνότητες, και φάσεις) είναι ένα κρίσιμο σημείο για τη ακριβή αναπαράσταση των σημάτων που αναλύονται. Στην παρούσα εργασία, με βάση τις πρόσφατες εξελίξεις στην ημιτονοειδή ανάλυση, προτείνουμε υψηλής ανάλυσης, προσαρμόσιμα ημιτονοειδή μοντέλα για συστήματα ανάλυσης, σύνθεσης, και τροποποίησης ομιλίας. Στόχος μας είναι να προσφέρουμε συστήματα που αναπαριστούν σήματα φωνής με εξαιρετικά ακριβή και συμπαγή τρόπο.

Εμπνευσμένοι από πρόσφατα προταθέντα μοντέλα, όπως το προσαρμόσιμο Σχεδον - Αρμονικό Μοντέλο (aQHM) και το προσαρμόσιμο Αρμονικό Μοντέλο (aHM), διατυπώνουμε τη θεωρία της προσαρμόσιμης Ημιτονοειδούς Μοντελοποίησης και προτείνουμε ένα μοντέλο που ονομάζεται εκτεταμένο προσαρμόσιμο Σχεδον - Αρμονικό Μοντέλο (eaQHM), το οποίο είναι ένα μη παραμετρικό μοντέλο, ικανό να προσαρμόσει τα στιγμιαία πλάτη και φάσεις των συναρτήσεων βάσης του στα τοπικά χρονικά μεταβαλλόμενα χαρακτηριστικά του σήματος της φωνής, αμβλύνοντας έτσι τη γνωστή υπόθεση της τοπικής στασιμότητας. Αποδεικνύεται ότι το eaQHM παρουσιάζει υψηλότερες επιδόσεις από το aQHM στην ανάλυση και ανασύνθεση των έμφωνων τμημάτων φωνής. Με βάση το eaQHM, ένα υβριδικό σύστημα ανάλυσης / σύνθεσης ομιλίας παρουσιάζεται (eaQHNM), μαζί με μια υβριδική έκδοση του του aHM (aHNM). Επιπλέον, παρουσιάζουμε κίνητρα για μια αναπαράσταση του σήματος της φωνής σε όλο το φάσμο και σε όλη τη διάρκεια του, χρησιμοποιώντας το eaQHM, αναπαριστώντας έτσι όλα τα μέρη του σήματος της φωνής, με υψηλής ανάλυσης AM-FM ημίτονα. Η αξιολόγηση δείχνει ότι η προσαρμοσιμότητα και η σχεδόν-αρμονικότητα είναι αρκετή για να παράξει πολύ υψηλή ποιότητα στην ανασύνθεση των άφωνων τμημάτων της φωνής. Στη συνέχεια, παρουσιάζεται το σύστημα πλήρους φάσματος ανάλυσης και σύνθεσης βασισμένο στο eaQHM, το οποίο υπερτερεί συστημάτων που θεωρούνται state-of-the-art, υβριδικά ή πλήρους ανάλυσης, στην ανάλυση και ανασύνθεση φωνής. Η υπεροχή του στην ποιότητα ανασύνθεσης επιβεβαιώθηκε με αντικειμενικές και υποκειμενικές αξιολογήσεις.

Όσον αφορά τις εφαρμογές, το eaQHM και το aHM εφαρμόζονται σε μετασχηματισμούς φωνής (κλιμάκωση χρόνου και κλιμάκωση θεμελιώδους συχνότητας). Οι μετασχηματισμοί που προκύπτουν είναι υψηλής ποιότητας, ακολουθώντας πολύ απλούς κανόνες, σε σύγκριση με άλλα συστήματα state-of-the-art. Οι έννοιες της σχετικής φάσης και της καθυστέρησης σχετικής φάσης είναι ζωτικής σημασίας για την ανάπτυξη μετασχηματισμένου σήματος

με χαρακτηριστικά αναλλοίωτου σχήματος, χωρίς τεχνικά ελαττώματα, και υψηλής ποιότητας. Τα αποτελέσματα δείχνουν ότι τα συστηματα βασισμένα στην αρμονικότητα προτιμούνται έναντι αυτών της σχεδόν-αρμονικότητας, λόγω της απλότητας της αναπραστάσης. Επιπλέον, το eaQHM εφαρμόζεται στο πρόβλημα της μοντελοποίησης σημάτων ήχου, και συγκεκριμένα ήχων μουσικών οργάνων. Το eaQHM αξιολογείται και σύγκρινεται με state-of-the-art συστήματα, και έχει υψηλές επιδόσεις όσον αφορά την ποιότητα επανασύνθεσης, αναπαριστωντας με επιτυχία τα στάδια της επίθεσης, της μετάβασης, και της στατικότητας ενός ήχου μουσικού οργάνου. Τέλος, μια άλλη προτεινόμενη εφαρμογή έγκειται στην ανάλυση και ταξινόμηση της εκφραστικής ομιλίας. Το eaQHM εφαρμόζεται στην ανάλυση της εκφραστικής ομιλίας, παρέχοντας τις στιγμιαίες παραμέτρους του ως χαρακτηριστικά που μπορούν να χρησιμοποιηθούν στην αναγνώριση και ταξινόμηση, βασισμένη σε διανυσματικούς χβαντιστές, εκφραστικής ομιλίας. Αν και τα ημιτονοειδή μοντέλα δεν χρησιμοποιούνται συνήθως σε τέτοιες εφαρμογές, τα αποτελέσματα είναι ελπιδοφόρα.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

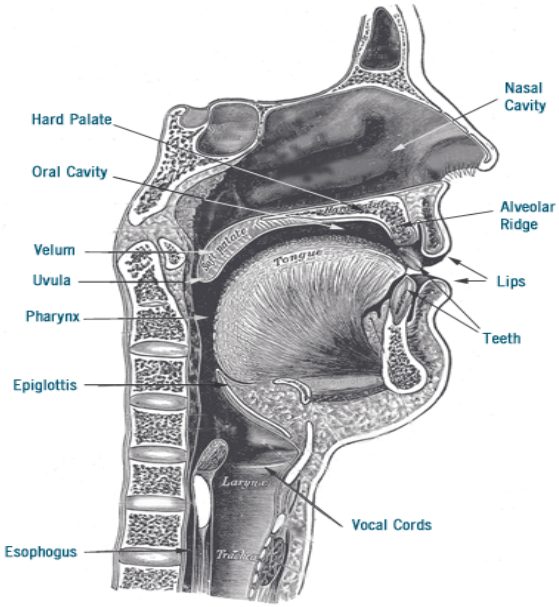
Without any doubt, speech is the most important and the most natural means of communication between humans. For this, digital processing of speech signals has been one of the most exciting areas of signal processing. In the last decades, speech research has led scientists and engineers in the discovery of several tools that still have tremendous impact on society. Voice communication and storage has been made effective and efficient due to advances in *speech coding* [RD01, HJA02] algorithms and technologies. Human-computer interaction became more convenient because of *speech recognition* [HD10] techniques that were able to make computers understand not only human speech but also human languages. Interactive systems increased their efficiency by corresponding to humans with natural voice: *speech synthesis* [ZTB09] algorithms are responsible for that. *Speech analysis* [Qua02] models and principles managed to provide deeper knowledge about human speech production system, helping medical doctors towards faster and reliable detection of pathologies and anomalies in speech. Algorithms for *enhancement of speech* [BMC05] under noisy conditions made terrestrial and satellite communications more robust. Entertainment industry got benefit from advanced *speech transformation* [Sty08] techniques to provide artificial voices in toys, films, and video-games. The list is endless and gets more and more populated day by day.

Apparently, speech processing has numerous applications and probably will have more in the future, when the convergence of computers, communications, and the Internet will grow stronger.

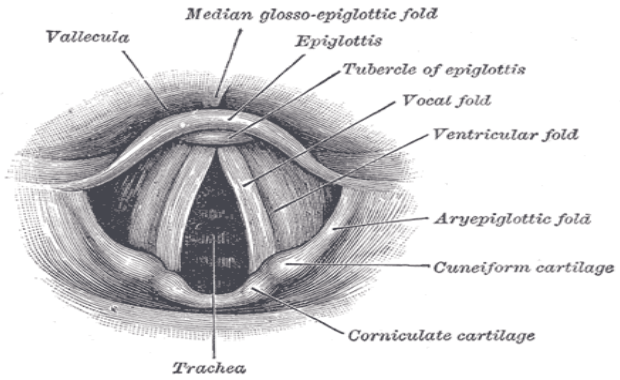
1.1 The Human Speech Production Mechanism

This thesis is related to speech, hence a brief review of the human speech production mechanism from an acoustic point of view should not be omitted. Figures 1.1a and 1.1b show the basic parts and organs that work together in speech production. These are roughly divided into three groups: the lungs, the larynx, and the vocal tract. First, the lungs provide the energy in the form of an airflow travelling into the trachea, where it is modulated by the glottis (see Figure 1.1b). This modulation results in either a quasi-periodic or a noisy source, which comes into the vocal tract and determines the phonation type. The vocal tract consists of the three cavities: the oral, the nasal, and the pharyngeal. Inside the vocal tract the source is shaped, hence giving sound its naturalness. Finally, the resulting waveform is radiated by the lips. More specifically, when *voiced sounds* are produced, the larynx is the source of the sound energy. First, the vocal folds are coming closer until fully attached to each other, temporarily blocking the airflow from the lungs and leading to increased subglottal pressure. When the resistance offered by the vocal folds becomes less than the subglottal pressure, the vocal folds re-open. Then, a combination of factors, including elasticity, laryngeal muscle tension, and the Bernoulli effect, lead to an immediate closure of the vocal folds. The vocal cords will continue to open and close in a quasi-periodic fashion as long as the process is maintained by a steady supply of pressurized air from the lungs. As they open and close, pulses of air flow through the glottal opening. The frequency of these pulses determines the fundamental frequency (f_0) of the source and contributes to the perceived pitch of the sound that is produced. This fundamental frequency varies over time, providing linguistic information, as in the different intonation patterns associated with questions and statements, and information about emotional content, such as differences in the emotional situation of the speaker. The vocal tract, consisting of both the oral and nasal cavities can serve as a time-varying acoustic filter that suppresses sound energy at certain frequencies while allowing it at others. The frequencies where local energy maxima are sustained by the vocal tract are called *formants* and are determined by the shape, length, and volume of the vocal tract, whereas the frequencies where local energy is suppressed are named *anti-formants*.

When *unvoiced sounds* are produced, the larynx is again the source of the sound energy; however, the vocal folds may be completely open, as in unvoiced consonants /s/ and /f/, whereas an intermediate position may also occur in phonemes like /h/. In *stop consonants*, such as /p/, /t/, or /k/, the vocal cords may act suddenly from a completely closed position in which they cut the air flow completely, to totally open position producing a glottal stop.



(a) The human vocal tract (from [GL18]).



(b) The vocal folds area (from [GL18]).

Figure 1.1: *Anatomy of the human speech production system.*

1.2 Modeling Speech

Over the years, scientists and engineers have invented numerous ways to represent the speech production mechanism in a mathematical context. Roughly speaking, there are two different, but not distinct, approaches: (a) a mathematical model that takes into account the actual speech production mechanism, considering it as a linear, time-varying system, excited by an input signal that differs according to the type of voicing (voiced or unvoiced speech), and (b) a mathematical model that represents the speech signal as a time-series, that is, a sum of amplitude and frequency modulated sinusoids. Approaches that follow the former model can be generally named as *source-filter models*, whereas for the latter they are said to follow a *sinusoidal model*.

1.2.1 The Source-Filter Model

The source-filter theory of speech finds its origins in the work of Fant [Fan70]. In brief, the acoustic speech output is commonly considered as a result from a combination of a *source* of sound energy, which is the larynx, modulated by a *filter* with its characteristics determined by the shape of the vocal tract. This combination results in a shaped spectrum with broadband energy peaks. This model is the so-called *source-filter model of speech production*.

From a signal processing point of view, the source-filter theory is implemented as follows. In voiced speech, the *source* of sound is modeled as a series of glottal pulses that represent the glottal volume velocity. The distance of the successive pulses determines the fundamental frequency of the signal. In unvoiced speech, there are no pulses, since the vocal folds do not quasi-periodically open and close. Thus, zero-mean white noise is used to model the characteristics of unvoiced speech. Thereafter, the source signal excites a time-varying filter that represents the vocal tract characteristics, as described earlier. Some models incorporate the lip radiation into the source, since the spectral characteristics of the lip radiation follow a high-pass-like structure. The source is then the derivative of the glottal flow [FLL85, PQR99], and is driven into the vocal tract filter. The accurate estimation of the source, especially in voiced speech, and the vocal tract signal is a notoriously challenging problem in digital speech processing [Mak75, Alk92, EJM91, LM95, YV98, PQR99, COCM01, BDDD05, FM06, DWBH06, VRR06, VRC07, Deg10, Alk11].

1.2.2 The Sinusoidal Models

The sinusoidal-based models for speech representation generalize the binary glottal excitation model (glottal pulses for voiced and random noise for unvoiced speech) used in the source-filter theory. In these models, the excitation waveform is assumed to be composed of sinusoidal components of arbitrary amplitudes, frequencies, and phases. The process of estimating the sinusoidal parameters, along with the assumptions on the nature of the model, is the main difference among sinusoidal-based representations [MQ86, Gri87, Ser89, SMFS89, Sty96, GS97, JJH99, HVK02, JH02].

Thus, the speech waveform is modeled as time and frequency modulations of these sinusoids, as they pass through the vocal tract and radiated by the lips. The basic assumption these models hold is that speech does not significantly change in short time intervals, or - as it is often stated in the literature - speech is *locally stationary*. In practical terms, this means that in a short time analysis window ($20 - 30$ ms), one can model speech as sinusoids that have constant amplitude and frequency values. This is a convenient assumption that is not entirely valid but has been proven useful in practical implementations of the models. An example of sinusoidal frequency decomposition is given in Figure 1.2.

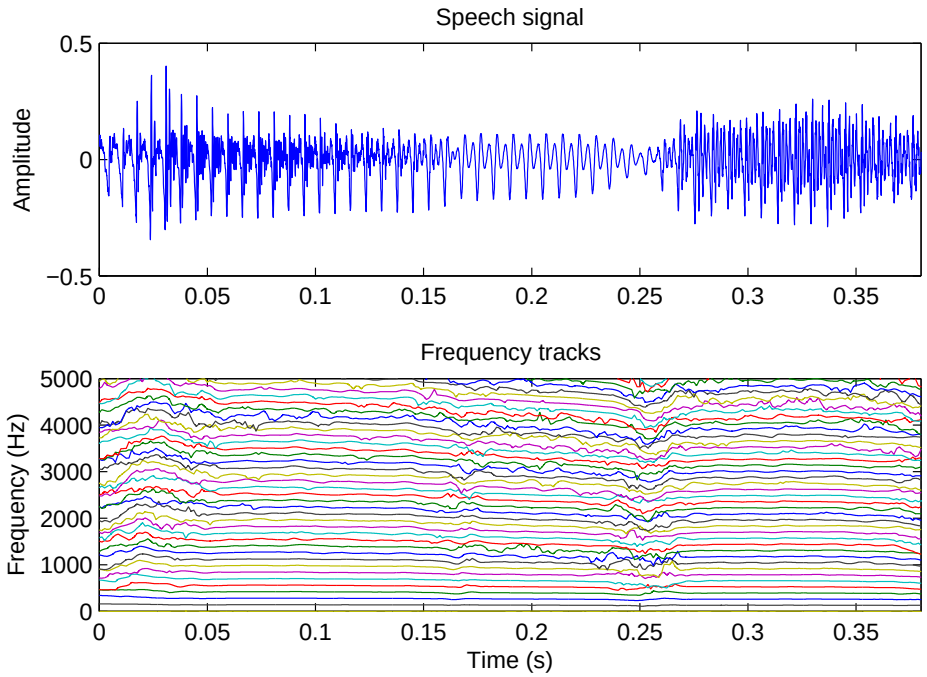


Figure 1.2: *An example of sinusoidal analysis. Upper panel: the speech signal. Lower panel: the decomposed frequencies present in the signal.*

1.3 Thesis Subject

In this thesis, the main focus is on **Adaptive Sinusoidal Models of Speech with Applications on Speech Modifications and Audio Analysis**.

Adaptive Sinusoidal Models (aSMs), although not all of them developed in this thesis, are presented under a common perspective, as sinusoidal models that are adaptive to the local characteristics of the underlying speech signal. In other words, they do *not* share the stationarity assumption the models discussed earlier do. The first aSM was developed in 2010 by Pantazis [Pan10], following a quasi-harmonic model suggested by Laroche [Lar89], and successively refining it [PRS08, PRS09b, PRS09a, PRS10] until reaching a model that is non-parametric and adaptive to the local phase characteristics of the analyzed signal. This model is called the *adaptive Quasi-Harmonic Model (aQHM)* [PRS11]. Two years later, Degottex and Stylianou proposed a full-band *adaptive Harmonic Model (aHM)* [DS13], where the analysis uses a quasi-harmonic scheme for an accurate f_0 estimation but the synthesis is strictly harmonic. In this thesis, we introduce a new model, the *extended adaptive Quasi-Harmonic Model* [KPRS12, KRS14], where adaptation includes both amplitude and phase. The main advantage of aSMs is the accuracy of the parameter estimation process, that is either a very accurate f_0 estimation (aHM) or very accurate sets of amplitudes, frequencies, and phases (aQHM, eaQHM). The process to obtain such accuracy is *adaptation*. Later, a detailed description of the models will be presented, the advantages of each model will be highlighted, and a comparison on speech analysis and resynthesis will take place.

In addition, Speech Transformations refer to the various modifications that may be applied to the speech signal. It covers a wide area of research from speech production modelling and understanding to speech perception, and from natural language processing, modelling and control of speaking style, to pattern recognition and statistical signal processing.

Speech Transformations (the terms ”transformations” and ”modifications” are used interchangeably in the rest of this text) have many potential applications in areas such as speech synthesis (e.g. for interactive Voice Response Systems, dialogue systems, text-to-speech systems), entertainment, film and music industry, toys, chat rooms and games, high-end

hearing aids, vocal pathology, and voice restoration. An objective of this research is to propose and implement algorithms for speech modifications based on the aSMs. More specifically, two main goals are to be addressed: time and pitch-scale modifications. Let us now give a formal definition on these two terms.

- ***In time-scale modification, the rate of articulation is changed without affecting the perceptual quality of the original speech.***
- ***In pitch-scale modification, the fundamental frequency is changed while preserving the short-time envelope (vocal tract) characteristics as well as the duration of the original speech.***.

Although other types of transformations do exist, such as voice conversion (where speech of a *source* speaker is transformed to match as closely as possible the speech of a *target* speaker), this work is focused only on the modifications with no specific target speaker. It is interesting to note that modifications include terms such as *rate of articulation* and *short-time envelope (vocal tract)*, implying that there is a combination of sinusoidal-model and source-filter theory in order to provide high-quality speech modifications [QM86, QM92, Sty96].

Moreover, a variety of applications of the aSMs are presented in this thesis. These include the **Modeling of Musical Instrument Sounds** and the **Analysis and Classification of Expressive Speech** using the adaptive Sinusoidal Models (and mostly, the eaQHM). For the former, the eaQHM is compared to the Exponentially Damped Sinusoidal Model (EDSM), which is considered as the state of the art in audio modeling, and the superiority of the eaQHM is demonstrated in terms of musical instrument sounds representation. Regarding the latter, the eaQHM is compared to the standard SM in representing expressive speech, and some preliminary results regarding recognition and classification are shown. The motivation behind applications is to examine if the parameters obtained from the models can yield higher performance in fields other than speech analysis and resynthesis.

1.4 Thesis Contribution

The contribution of this thesis with respect to the aforementioned speech processing fields of the introductory paragraph can be depicted in Figure 1.3.

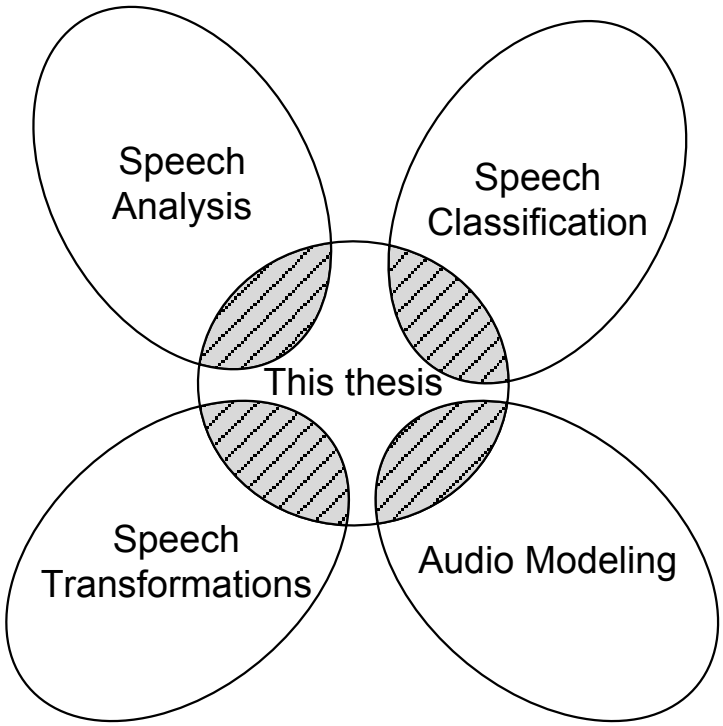


Figure 1.3: *Contribution of this thesis with respect to some well known speech processing research fields.*

Namely, this thesis contributes the following achievements:

- An extension of the adaptive Quasi-Harmonic Model, called the *extended adaptive Quasi-Harmonic Model - eaQHM*. The eaQHM performs full signal adaptation, e.g. not only the phase is adapted to the signal but also the amplitude as well. Thus, the signal can be modelled more accurately, and this can be demonstrated both subjectively and objectively. **This work has been published in ICASSP 2012 [KPRS12].**
- The family of *adaptive Sinusoidal Models* is introduced, which is a collection of models that can locally adapt their parameters on the analyzed signal. This set of models include the recently proposed *adaptive Quasi-Harmonic Model - (aQHM)*, the *extended adaptive Quasi-Harmonic Model - (eaQHM)*, which is proposed in this thesis, and the recently suggested *adaptive Harmonic Model - (aHM)*.
- The eaQHM is shown to accurately model unvoiced parts of speech, such as stop consonants and fricatives, as adaptive, high-resolution AM-FM components. Since the eaQHM models both voiced and unvoiced speech very accurately, the eaQHM is suggested as a model for a residual-free, full-band model of speech. **Part of this work has been published in Interspeech 2012 [KRS13].**
- In this context, hybrid and full-band systems of analysis and synthesis of speech based on the aSMs are revised, developed, and described. The eaQHM-based system is shown to stand among adaptive models and outperforms the state-of-the-art in terms of both accuracy in signal reconstruction and perceptual transparency compared to the original signal. **This work has been published in ICASSP 2014 [KRS14].**
- Speech modifications are proposed: pitch and time-scaling. The modifications are characterized by high quality, simplicity, flexibility, and freedom of common artefacts. Hybrid and full-band analysis, synthesis, and modifications systems based on aSMs are presented. The modifications are of high quality and are governed by very simple rules, making the transformations very attractive. Specifically, time-scaling is simply implemented by resampling and interpolating the instantaneous amplitude and frequency of the speech signal. For the phase, a very simple approach is suggested based on the notions of *relative phase* and *relative phase delays*. For pitch scaling, the estimation of a spectral magnitude envelope is necessary, but no phase envelope estimation is followed, thus significantly simplifying modification. **Part of this work has been published in ICASSP 2013 [KDRS13] and in ICASSP 2014 [KDRS14].**
- Applications of the eaQHM are discussed:
  1. The eaQHM is applied on modelling musical instrument sounds. It is shown that the eaQHM can handle *all parts* of the sounds, including silence, onset, transient, and stationary part. The accuracy of the reconstruction is shown to be high and the synthesized sound is indistinguishable from the original. Extensive comparisons to the state-of-the-art have been made and the superiority of the proposed model over the competition is demonstrated. **This work has been published in EUSIPCO [CKMS13], in WASPAA [CKD+13], and in IEEE Transactions on Audio, Speech, and Language Processing [CKMS] (under peer review).**
  2. The eaQHM is applied on expressive speech analysis and classification. It is shown that aSMs can analyze and resynthesize various styles of expressive speech with very high accuracy and quality, whereas the classification rates are higher than standard sinusoidal approaches in emotion classification. **This work has been accepted in EUSIPCO 2014 [KYMS14] and in Interspeech 2014 [YKS14].**

1.5 Thesis Organization

The rest of this thesis is organized in three parts, as follows:

1. In Part I: Adaptive Sinusoidal Models
   (a) Chapter 2 discusses the adaptive Sinusoidal Models, as standalone signal analysis tools for speech.
2. In Part II: Speech Analysis, Synthesis, and Modifications
   (a) Chapter 3 briefly presents the related work in the field of speech analysis, synthesis, and modifications. The most important systems are categorized and described.
   (b) Chapter 4 describes analytically the suggested analysis and synthesis systems and compares them with other, well-known, state-of-the-art systems. Objective and subjective results are presented.
   (c) Chapter 5 suggests speech modification methods using the systems presented in Chapter 4, and are compared to the state-of-the-art.
3. In Part III: Applications

   (a) Chapter 6 presents applications of the newly proposed systems in musical instrument sound analysis.
   (b) Chapter 7 discusses the application of the aSMs in expressive speech analysis and classification.

4. Chapter 8 concludes the thesis and suggests possible future research directions.
5. Appendix I presents a residual analysis of musical instrument sounds obtained from sinusoidal models.
6. Appendix II presents the publications made during this thesis.

Subjective evaluations are supported with on-line demonstration pages that verify the conclusions derived.

Part I

Adaptive Sinusoidal Models

Chapter 2

Adaptive Sinusoidal Modeling

The term *adaptive Sinusoidal Models - aSM* refers to the family of sinusoidal models that are able to *adapt* their parameters to the local characteristics (phase and/or amplitude) of the analyzed speech signal. It should be highlighted that non-adaptive (conventional) sinusoidal models do consider local stationarity of a signal in their representation.

Before going into details, the notions of *stationarity* and *adaptivity* should be precisely defined. It should be noted that stationarity is a general term that can be used to characterize any signal; however, adaptation will be defined in a context of representing a signal with a set of complex exponential basis functions. Now let us define the *stationary* and *adaptation* terms for sinusoidal analysis.

Consider a set of complex exponential basis functions that a signal is projected onto.

- *A complex exponential is called **stationary** in a well-defined time interval when its frequency and/or its amplitude are constant.*
- *A complex exponential is called **adaptive** in a well-defined time interval when its amplitude and/or frequency are computed by taking into account the local characteristics of the signal which is projected onto it.*

In general, the aSMs are founded on the principle of projecting a signal segment onto a set of non-parametric, time-varying, non-stationary set of complex exponential basis functions *inside an analysis window*, whereas conventional sinusoidal models consider that speech characteristics remain relatively unchanged in a local level, thus their basis functions are stationary in amplitude and frequency. The construction of this set of time-varying basis function depends on the adaptive model, as it will become apparent in this chapter.

First, the heart of all aSMs is presented, the so-called *Quasi-Harmonic Model - QHM*. The QHM is not an adaptive model itself, but it provides the mechanism for adaptation, that is a *frequency correction mechanism*, which yields an estimate of the mismatch between the actual and estimated frequencies. This frequency correction is added to the estimated frequencies to allow a closer representation of the underlying signal. An iterative procedure on the parameter estimation leads to the *iterative Quasi-Harmonic Model - iQHM*, which successively updates the frequencies until a convergence criterion is met. However, both QHM and iQHM hold the local stationarity assumption, that is, all parameters are obtained by projecting the signal on a *stationary* set of exponential basis functions.

To alleviate the local stationarity assumption, the *adaptive Quasi-Harmonic Model - aQHM* developed by Pantazis et al [PRS11] is presented and then, we propose an extension, the *extended adaptive Quasi-Harmonic Model - eaQHM* [KPRS12]. These models go one step further and exploit the frequency correction mechanism of QHM to refine their frequency estimations, along with the iterative construction of a time-varying, non-parametric, and non-stationary set of basis functions where the signal is projected onto. Hence, a definition for the *adaptation* term is given as follows:

Consider a set of complex exponential basis functions that a signal is projected onto.

***Adaptation** is an iterative construction of a set of complex exponential basis functions according to the local characteristics of the underlying signal, which are successively used to refine the instantaneous components of the signal, e.g. the instantaneous amplitude, frequency, and phase.*

In simpler words, the i^{th} adaptation cycle refines the instantaneous parameters of the signal, and uses them to form the non-stationary exponential basis functions of the $(i+1)^{th}$ adaptation cycle.

Moreover, QHM can work as an initializer for aQHM and eaQHM, providing a well-estimated set of frequency tracks. However, any AM-FM algorithm can be used as an initializer (such as a simple Harmonic Model (HM) or a peak

picking process via FFT). Finally, another recent approach, referred to as the *adaptive Harmonic Model - aHM* [DS13] is briefly presented. The aHM is a purely harmonic model but benefits from the idea of non-stationary basis functions to provide a closer representation of speech via an iterative refinement of the fundamental frequency.

2.1 Introduction

In general, an aSM can be described as

$$x(t) = \left(\sum_{k=-K}^{K} C_k(t)\psi_k(t) \right) w(t) \tag{2.1}$$

where $\psi_k(t)$ denotes the set of non-stationary basis functions, $C_k(t)$ denotes the amplitude term of the model, $2K+1$ is the number of exponentials (hence, $K+1$ sinusoids), and finally $w(t)$ is the analysis window with support in $[-T, T]$. In conventional sinusoidal models, including the Sinusoidal Model [MQ86], the Harmonic Model (HM) [Sty96], and others, the set of basis functions $\psi_k(t)$ in the analysis part is stationary in frequency and in amplitude. For example, the basis functions in the SM are in the form of

$$\psi_k^{SM}(t) = 1 \cdot e^{j2\pi f_k t}, \;\; C_k^{SM}(t) = a_k \tag{2.2}$$

where the amplitudes and frequencies of the basis functions are constant inside the analysis window (1 and f_k, respectively). However, in the aSMs, as will be described in the following sections, amplitudes and frequencies of the basis functions are non-parametric and depend on the actual characteristics of the analyzed signal:

$$\psi_k^{aSM}(t) = \alpha_k(t) \cdot e^{j\phi_k(t)}, \tag{2.3}$$

where $\alpha_k(t)$ is the time-varying instantaneous amplitude of the k^{th} basis function, $\phi_k(t)$ is the instantaneous phase of the k^{th} basis function, computed as the integral of the instantaneous frequency, $f_k(t)$. The amplitude term of the model, $C_k^{aSM}(t)$, is time-varying for all aSMs. Specifically,

$$\psi_k^{aQHM}(t) = 1 \cdot e^{j\phi_k(t)}, \;\; C_k^{aQHM}(t) = a_k + tb_k, \tag{2.4}$$

$$\psi_k^{eaQHM}(t) = \alpha_k(t) \cdot e^{j\phi_k(t)}, \;\; C_k^{eaQHM}(t) = a_k + tb_k \tag{2.5}$$

and

$$\psi_k^{aHM}(t) = 1 \cdot e^{jk\phi_0(t)}, \;\; C_k^{aHM}(t) = a_k + tb_k \tag{2.6}$$

where a_k, b_k are the complex amplitude and slope of the k^{th} component of the model, $\phi_0(t)$ is the instantaneous phase of the fundamental frequency, computed as the integral of the latter.

Based on this introductory analysis, the description of the aSMs follows next, starting from the QHM, which is fundamental for the discussion and understanding of aSMs.

2.2 The Quasi Harmonic Model (QHM)

As we mentioned in sinusoidal modeling, a signal can be represented as follows:

$$x(t) = \left(\sum_{k=-K}^{K} a_k e^{j2\pi f_k t} \right) w(t), \tag{2.7}$$

where $2K+1$ is the number of components with complex amplitudes a_k at frequencies f_k, and $w(t)$ is the analysis window. Let us assume that f_k denote the correct frequencies of the signal components. In sinusoidal modelling, frequencies are estimated first (e.g., by peak-picking, by considering harmonics of a fundamental frequency, etc.), before the estimation of the complex amplitudes. The estimated frequencies will be denoted here by $\hat{f}_k$. Then, we may write:

$$f_k = \hat{f}_k + \eta_k, \quad k = -K, ..., K \tag{2.8}$$

If the error, η_k, is high, then the estimation of the complex amplitudes, a_k, is severely biased.

To cope with this problem, in [PRS09a] and [PRS08] the use of the Quasi-Harmonic Model (QHM) for the represen-

tation of speech was suggested:

$$x(t) = \left(\sum_{k=-K}^{K} (a_k + tb_k) e^{j2\pi \hat{f}_k t} \right) w(t), \tag{2.9}$$

where b_k denotes the complex slope of the k^{th} component. In the frequency domain, the k^{th} component is written as:

$$X_k(f) = a_k W(f - \hat{f}_k) + j\frac{b_k}{2\pi} W'(f - \hat{f}_k) \tag{2.10}$$

where $W(f)$ is the Fourier transform of the analysis window and $W'(f)$ is the derivative of $W(f)$ over f. In [PRS08], it was shown that QHM is able to correct frequency mismatches using the projection of b_k onto a_k:

$$b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k \tag{2.11}$$

where ja_k denotes the perpendicular (vector) to a_k. The ρ_1, ρ_2 parameters are computed as:

$$\rho_1 = \frac{\Re\{a_k\}\Re\{b_k\} + \Im\{a_k\}\Im\{b_k\}}{|a_k|^2}, \tag{2.12}$$

and

$$\rho_2 = \frac{\Re\{a_k\}\Im\{b_k\} - \Im\{a_k\}\Re\{b_k\}}{|a_k|^2}, \tag{2.13}$$

where $\Re\{a_k\}$, $\Re\{b_k\}$ and $\Im\{a_k\}$, $\Im\{b_k\}$ are the real and imaginary parts of a_k and b_k, respectively. By substitution in (2.10) and if we consider the Taylor Series expansion of $W(f - \hat{f}_k - \frac{\rho_{2,k}}{2\pi})$, and the value of term $W''(f)$ at f_k as small, then for small values of $\rho_{2,k}$, it can be shown [PRS08] that the k^{th} component can be expressed in time domain as:

$$X_k(f) = a_k \left[W(f - \hat{f}_k) - \frac{\rho_{2,k}}{2\pi} W'(f - \hat{f}_k) + j\frac{\rho_{1,k}}{2\pi} W'(f - \hat{f}_k) \right] \tag{2.14}$$

If we consider the Taylor series expansion of $W(f - \hat{f}_k - \frac{\rho_{2,k}}{2\pi})$,

$$W(f - \hat{f}_k - \frac{\rho_{2,k}}{2\pi}) = W(f - \hat{f}_k) - \frac{\rho_{2,k}}{2\pi} W'(f - \hat{f}_k) + O(\rho_{2,k}^2 W''(f - \hat{f}_k)) \tag{2.15}$$

and the value of term $W''(f)$ at f_k as small, then for small values of $\rho_{2,k}$, it can be shown [PRS08] that the k^{th} component can be expressed in time domain as:

$$x_k(t) \approx a_k \left[e^{j(2\pi \hat{f}_k + \rho_{2,k})t} + \rho_{1,k} t e^{j2\pi \hat{f}_k t} \right] w(t) \tag{2.16}$$

Thus, $\rho_{2,k}/2\pi$ is an estimator of the frequency error η_k:

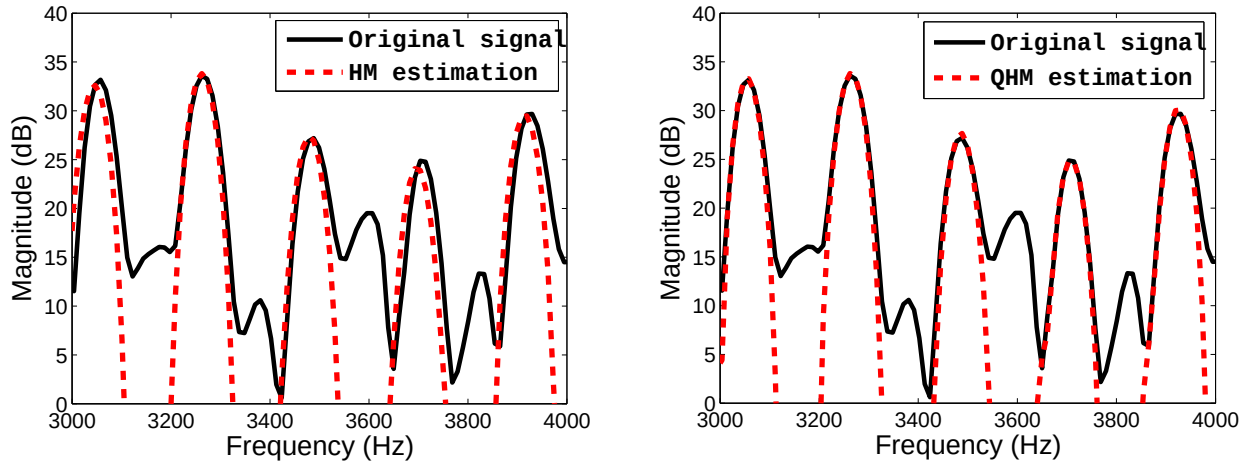
$$\hat{\eta}_k = \frac{\rho_{2,k}}{2\pi} = \frac{1}{2\pi} \frac{\Re\{a_k\}\Im\{b_k\} - \Im\{a_k\}\Re\{b_k\}}{|a_k|^2}, \tag{2.17}$$

where $\rho_{1,k}$ accounts for the normalized amplitude slope of the k^{th} component. In [PRS08], it was also shown that this correction depends on the magnitude of $\rho_{2,k}$ and the value of the term $W''(f)$ at f_k. The frequency mismatch correction mechanism can be depicted in Figure 2.1, where in Figure 2.1a, a harmonic template is estimated from the known f_0 value, and in Figure 2.1b, the frequency correction estimates $\hat{\eta}_k$ are applied.

The estimation of a_k, b_k is performed via Least Squares (LS) in the following way:

Let us define the parameter vector $\mathbf{x}$ as

$$\mathbf{x} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \tag{2.18}$$



(a) *Before applying frequency correction*

(b) *After applying frequency correction*

Figure 2.1: *Frequency correction with QHM. Left panel: Frequency estimation without frequency correction using a harmonic model. Right panel: Frequency estimation after frequency correction using QHM. Red dashed line denotes the estimated magnitude spectrum, and black solid line denotes the original magnitude spectrum of a speech frame.*

The error is defined in discrete time by

$$\epsilon(\mathbf{a},\mathbf{b}) = \sum_{n=-N}^{N} |s[n] - s_q[n]|^2 \tag{2.19}$$

$$= \sum_{n=-N}^{N} (s[n] - s_q[n])^*(s[n] - s_q[n]) \tag{2.20}$$

where $s[n]$ is the original windowed signal, $s_q[n]$ is the Quasi-Harmonic representation of Eq. (2.7), and $2N+1$ is the window size. In matrix notation, if we separate the window values from the samples, Eq. (2.20) becomes

$$\epsilon(\mathbf{a},\mathbf{b}) = (\mathbf{Ws} - \mathbf{Ws_q})^H(\mathbf{Ws} - \mathbf{Ws_q}) \tag{2.21}$$

$$= (\mathbf{W}(\mathbf{s} - \mathbf{s_q}))^H\mathbf{W}(\mathbf{s} - \mathbf{s_q}) \tag{2.22}$$

$$= (\mathbf{s} - \mathbf{s_q})^H\mathbf{W}^H\mathbf{W}(\mathbf{s} - \mathbf{s_q}) \tag{2.23}$$

where $\mathbf{W}$ is a square matrix having the analysis window values in its diagonal, $\mathbf{s}$ is the original signal samples in a vector, and ^{H} denotes the Hermitian operator.

Now, the QHM can be written in matrix notation as

$$s_q[n] = \sum_{n=-N}^{N} (a_k + nb_k)e^{j2\pi f_k n/f_s} \tag{2.24}$$

$$= \sum_{n=-N}^{N} a_k e^{j2\pi f_k n/f_s} + \sum_{n=-N}^{N} nb_k e^{j2\pi f_k n/f_s} \tag{2.25}$$

$$\mathbf{s_q} = \mathbf{E_0 a} + \mathbf{E_1 b} = \begin{bmatrix}\mathbf{E_0}|\mathbf{E_1}\end{bmatrix}\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix} = \mathbf{Ex} \tag{2.26}$$

where

$$\mathbf{E_0} = (E_0)_{n,k} = e^{j2\pi\frac{f_k n}{f_s}}\ ,\ \mathbf{E_1} = (E_1)_{n,k} = n(E_0)_{n,k} = ne^{j2\pi\frac{f_k n}{f_s}} \tag{2.27}$$

and

$$\mathbf{E} = \begin{bmatrix}\mathbf{E_0}|\mathbf{E_1}\end{bmatrix} \tag{2.28}$$

Hence, the minimization comes to

$$\frac{\partial \epsilon(\mathbf{x})}{\partial \mathbf{x}} = 0 \tag{2.29}$$

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{s}-\mathbf{E}\mathbf{x})^H\mathbf{W}^H\mathbf{W}(\mathbf{s}-\mathbf{E}\mathbf{x}) = 0 \tag{2.30}$$

The solution is given by

$$\mathbf{x} = \begin{bmatrix}\hat{\mathbf{a}}\\ \hat{\mathbf{b}}\end{bmatrix} = (\mathbf{E}^H\mathbf{W}^H\mathbf{W}\mathbf{E})^{-1}\mathbf{E}^H\mathbf{W}^H\mathbf{W}\mathbf{s} \tag{2.31}$$

Finally, the signal can be locally approximated as

$$x(t) = \sum_{k=-K}^{K} |\hat{a}_k| e^{j(2\pi(\hat{f}_k+\hat{\eta}_k)t+\hat{\phi}_k)} w(t) \tag{2.32}$$

where

$$\hat{\phi}_k = \angle \hat{a}_k \tag{2.33}$$

Although the QHM has been proved to perform better than standard Sinusoidal or Harmonic Models [PRS09a], it still assumes that the signal is stationary inside the analysis window.

2.3 The adaptive Quasi-Harmonic Model, aQHM

To alleviate the non-stationarity assumption, an adaptive Quasi-Harmonic Model (aQHM) has been suggested by Pantazis et al [PRS11].

$$x(t) = \left(\sum_{k=-K}^{K} (a_k + tb_k) e^{j(\hat{\phi}_k(t+t_i)-\hat{\phi}_k(t_i))}\right) w(t), \;\; t \in [-T, T] \tag{2.34}$$

where $\phi_k(t)$ denotes the instantaneous phase function of the k^{th} component and t_i is the center of the analysis window. The term b_k plays the same role as in QHM; it provides a means to update the frequency of the underlying sine wave at the center of the analysis window, t_i. Given the samples of the input signal in vector s, the model parameters are found via LS, as in QHM:

$$\begin{bmatrix}\hat{\mathbf{a}}\\ \hat{\mathbf{b}}\end{bmatrix} = (\mathbf{E}^H\mathbf{W}^H\mathbf{W}\mathbf{E})^{-1}\mathbf{E}^H\mathbf{W}^H\mathbf{W}\mathbf{s} \tag{2.35}$$

where $\mathbf{W}$ is the matrix containing the window values in the diagonal, $\mathbf{s}$ is the input signal vector, the matrix $\mathbf{E}$ is defined as $\mathbf{E} = [\mathbf{E_0}|\mathbf{E_1}]$, the submatrices $\mathbf{E_i}$, $i = 0, 1$ have elements given by

$$(E_0)_{n,k} = e^{j(\phi_k(t_n+t_i)-\phi_k(t_i))} \tag{2.36}$$

and

$$(E_1)_{n,k} = t_n e^{j(\phi_k(t_n+t_i)-\phi_k(t_i))} = t_n (E_0)_{n,k}, \tag{2.37}$$

and the instantaneous phase of the k^{th} component can be computed as

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \int_{t_i}^{t_i+t} 2\pi f_k(u)du, \;\; t \in [-T, T], \tag{2.38}$$

where $f_k(t)$ is the frequency trajectory of the k^{th} component.

Using the definition of phase, the instantaneous phase of a single component, $\phi(t)$, is computed as the integral of the instantaneous frequency, $f(t)$. The instantaneous frequency is obtained from an initial parameter estimation, such as in QHM. In order to interpolate phase values between two successive time instants, t_i, t_{i+1}, the following equation is proposed:

$$\phi(t) = \hat{\phi}(t_i) + \int_{t_i}^{t+t_i} 2\pi \hat{f}(u)du \tag{2.39}$$

where $\hat{\phi}(t_i)$ is the instantaneous phase estimate at time instant t_i. However, this solution does not take into account the

frame boundary conditions at time instant t_{i+1}. Hence, there is no guarantee that the phase value at time instant t_{i+1},

$$\phi(t)\big|_{t=t_{i+1}} = \hat{\phi}(t_{i+1}) + 2\pi M \tag{2.40}$$

where M is an integer appropriately selected to be as close as possible to

$$M = \mathbf{round}\Big(\frac{\phi(t_{i+1}) - \hat{\phi}(t_i)}{2\pi}\Big) \tag{2.41}$$

where $\mathbf{round}(\cdot)$ is the rounding to closest integer function. In order to ensure phase continuation over frame boundaries, it is suggested [PRS11] to modify Eq. (2.39) as:

$$\phi(t) = \hat{\phi}(t_i) + \int_{t_i}^{t+t_i} (2\pi\hat{f}(u) + c(u))du \tag{2.42}$$

where $c(u)$ is given by

$$c(u) = r(t_{i+1}) \sin\Big(\frac{\pi(u - t_i)}{t_{i+1} - t_i}\Big) \tag{2.43}$$

This way, Eq. (2.40) is satisfied if we choose $r(t_{i+1})$ as

$$r(t_{i+1}) = \frac{\pi(\phi(t_{i+1}) + 2\pi M - \hat{\phi}(t_{i+1})}{2(t_{i+1} - t_i)} \tag{2.44}$$

where M is computed as in Eq. (2.41).

In contrast to QHM, where the argument of the basis functions is parametric and stationary, in aQHM the argument of the basis functions is neither parametric nor necessarily stationary. Moreover, the aQHM basis functions use the instantaneous phases which have been estimated from the input signal. In that sense, these are also adaptive to the estimates of the current phase characteristics of the signal. The process of successive adaptations is shown in Figure 2.2. However, only phase adaptation is allowed in this model. Naturally, amplitude adaptation should be included as well,

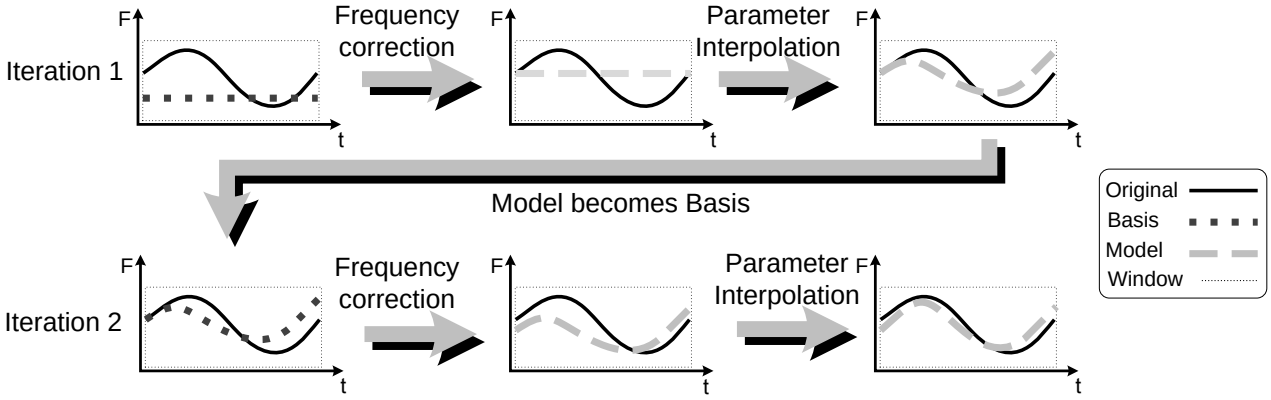


Figure 2.2: *Illustration of the adaptation of the frequency trajectory of a sinusoidal partial inside the analysis window in aQHM. The figure depicts the first and second iterations of aQHM, showing local adaptation as iterative projection of the original waveform onto the model. Horizontal axes represent time, vertical axes represent frequency.*

and this the concept of the next section.

2.4 The extended adaptive Quasi-Harmonic Model - eaQHM

In this thesis, we propose an extension of aQHM to include amplitude adaptation, called the extended adaptive Quasi-Harmonic Model (eaQHM):

$$x(t) = \left(\sum_{k=-K}^{K} (a_k + tb_k)\hat{\alpha}_k(t)e^{j(\hat{\phi}_k(t+t_i) - \hat{\phi}_k(t_i))}\right) w(t), \;\; t \in [-T, T] \tag{2.45}$$

where

$$\hat{\alpha}_k(t) = \frac{A_k(t+t_i)}{A_k(t_i)} \tag{2.46}$$

where t_i is still the center of the analysis window and $A_k(t)$ is the instantaneous amplitude of the k^{th} component obtained from a previous adaptation step. The estimation of the unknown parameters of eaQHM is similar to that of QHM:

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = (\mathbf{E_e}^H \mathbf{W}^H \mathbf{W} \mathbf{E_e})^{-1} \mathbf{E_e}^H \mathbf{W}^H \mathbf{W} \mathbf{s} \tag{2.47}$$

where $\mathbf{W}$ and s remain as previously defined, the matrix $\mathbf{E_e}$ is defined as $\mathbf{E_e} = [\mathbf{E_{e0}}|\mathbf{E_{e1}}]$, and the submatrices $\mathbf{E_{ei}}$, $i = 0, 1$ have elements given by

$$(E_{e0})_{n,k} = \alpha_k(t_n) e^{j(\phi_k(t_n+t_i)-\phi_k(t_i))} \tag{2.48}$$

and

$$(E_{e1})_{n,k} = t_n \alpha_k(t_n) e^{j(\phi_k(t_n+t_i)-\phi_k(t_i))} = t_n (E_{e0})_{n,k}, \tag{2.49}$$

It is clear that the basis functions are adapted to the local amplitude characteristics of the signal. Please note that the instantaneous amplitude $A_k(t)$ is divided by $A_k(t_i)$ before its use in the basis functions.

In picturing the amplitude and frequency modelling of the eaQHM within the analysis window, Figures 2.3 and 2.4 show how conventional sinusoidal models like HM, SM, or QHM behave inside their analysis window. Their exponential basis functions are stationary in frequency, thus being inefficient on the representation of highly non-stationary frequency curves. The same argument applies for amplitude curves, although frequency estimation is far more important than amplitude estimation.

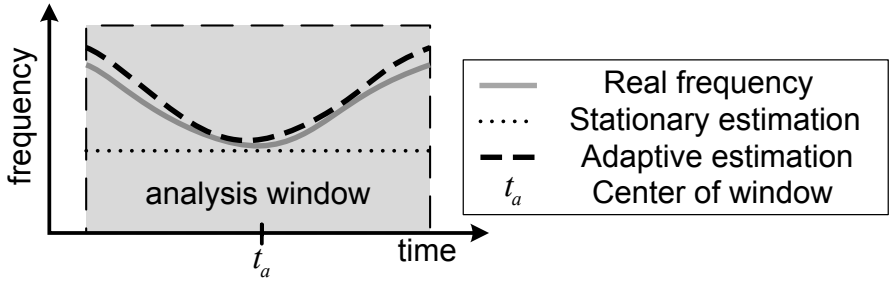


Figure 2.3: *Inside the analysis window, the frequency trajectory of a partial (solid grey line) is assumed to be constant for stationary sinusoidal models (dotted line), while eaQHM (dashed line) iteratively adapts to the shape of the instantaneous frequency.*

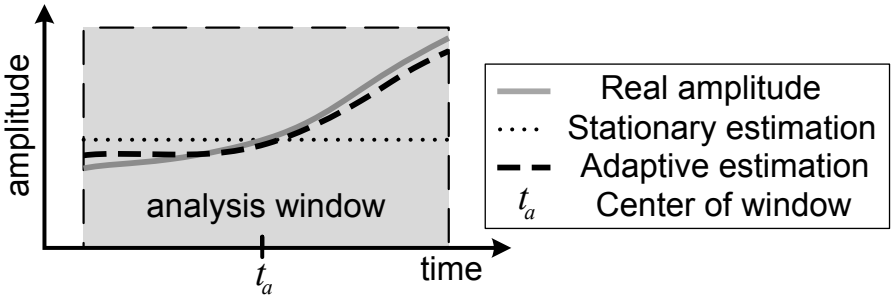


Figure 2.4: *Inside the analysis window, the amplitude trajectory of a partial (solid grey line) is assumed to be constant for stationary sinusoidal models (dotted line), while eaQHM (dashed line) iteratively adapts to the shape of the instantaneous amplitude.*

2.5 Algorithm for Adaptive Sinusoidal Analysis

Adaptive Sinusoidal models require an initialization step, so QHM will be used for this purpose, although any frequency estimation algorithm can be used. Thus, the initials steps consist of the following:

$$\hat{f}_k^0(t_i) = \hat{f}_k^0(t_{i-1}) + \hat{\eta}_k \tag{2.50}$$

$$\hat{A}_k^0(t_i) = |a_k^i|, \;\; \hat{\phi}_k^0(t_i) = \angle a_k^i \tag{2.51}$$

where t_i is the center of the i^{th} analysis frame. The AM-FM decomposition algorithm using eaQHM is provided in Algorithm 1.

Algorithm 1 *AM-FM decomposition using eaQHM*

Require: Provide initial frequency estimate $\mathbf{f_k^0(t_1)}$
  for frame $i = 1$ **to** L **do**
    Compute $\mathbf{a_k^i}, \mathbf{b_k^i}$ using LS
    Update $\mathbf{\hat{f}_k^0(t_i)}$ using Eq. (2.50)
    Compute $\mathbf{\hat{A}_k^0(t_i)}$ and $\mathbf{\hat{\phi}_k^0(t_i)}$ using Eq. (2.51)
    Set $\mathbf{f_k^0(t_{i+1}) = \hat{f}_k^0(t_i)}$
  end for
  Interpolation of the parameters $\{\mathbf{\hat{A}_k^0(t)}, \mathbf{\hat{f}_k^0(t)}, \mathbf{\hat{\phi}_k^0(t)}\}$
  Adaptation of amplitudes and phases:
  for adaptation $m = 1$ **to** $\cdots$ **do**
    for frame $i = 1$ **to** L **do**
      Compute $\mathbf{a_k^i}, \mathbf{b_k^i}$ using $\mathbf{\hat{\phi}_k^{m-1}(t)}$ of Eq. (2.42) and Eq. (2.47)
      Set $\mathbf{\hat{f}_k^m(t_i) = \hat{f}_k^m(t_{l-1}) + \hat{\eta}_k}$
      Set $\mathbf{\hat{A}_k^m(t_i) = |a_k^i|}$ and $\mathbf{\hat{\phi}_k^m(t_i) = \angle a_k^i}$
    end for
  end for
  Interpolation of the parameters $\{\mathbf{\hat{A}_k^m(t)}, \mathbf{\hat{f}_k^m(t)}, \mathbf{\hat{\phi}_k^m(t)}\}$

The convergence criterion for the algorithm was selected to be the following:

$$\frac{SRER^{m-1} - SRER^m}{SRER^{m-1}} < \epsilon \tag{2.52}$$

where $SRER^m$ is the Signal-to-Reconstruction-Error Ratio of the resynthesized signal in the m^{th} adaptation, defined as

$$SRER = 20 \log_{10} \frac{\sigma_{x(t)}}{\sigma_{x(t)-\hat{x}(t)}} \tag{2.53}$$

where σ_x denotes the standard deviation of $x(t)$, $x(t)$ is the actual signal and $\hat{x}(t)$ is the reconstructed signal, and where ϵ is a threshold for convergence, typically set to 0.02. As a last step of the algorithm, the signal can finally be approximated as the sum of its AM-FM components:

$$\hat{x}(t) = \sum_{k=-K}^{K} \hat{A}_k(t) e^{j\hat{\phi}_k(t)}$$

2.6 Evaluations

In this section, a performance comparison between the two adaptive quasi-harmonic models described so far on synthetic and real voiced speech is presented.

2.6.1 Validation on Synthetic Signals

For the purpose of demonstrating the performance of eaQHM, we consider a two-component signal with modulated amplitudes and frequencies, defined as:

$$x(t) = a_1(t) e^{j(2\pi f_1 t + \phi_1(t))} + a_2(t) e^{j(2\pi f_2 t + \phi_2(t))} \tag{2.54}$$

where the above parameters are given in Table 2.1, and the sampling frequency is $F_s = 8$ kHz, while the window length is

Sinusoid	1st	2nd
f_i	700	1000
$\phi_i(t)$	$\frac{\pi}{10} + \cos(2\pi 80t)$	$\frac{\pi}{3} + \cos(2\pi 50t)$
$a_i(t)$	$2 + 0.8\cos(2\pi 100)t$	$2 + 0.6\cos(2\pi 100t)$

Table 2.1: *The parameters of the synthetic signal.*

10 msec. It should be noted that the amplitudes of the signal components are high-frequency modulated and thus, the local amplitude linearity is violated. The time-varying amplitudes $a_i(t)$ and the time-varying frequencies $F_i = f_i + \frac{1}{2\pi}\frac{d}{dt}\phi_i(t)$, for $i = 1, 2$, are to be estimated. In Figure 2.5, the parameters as they are estimated by aQHM are depicted, whereas in Figure 2.6, the same information is depicted for the eaQHM algorithm. As it can be seen in Figures 2.5 and 2.6, the

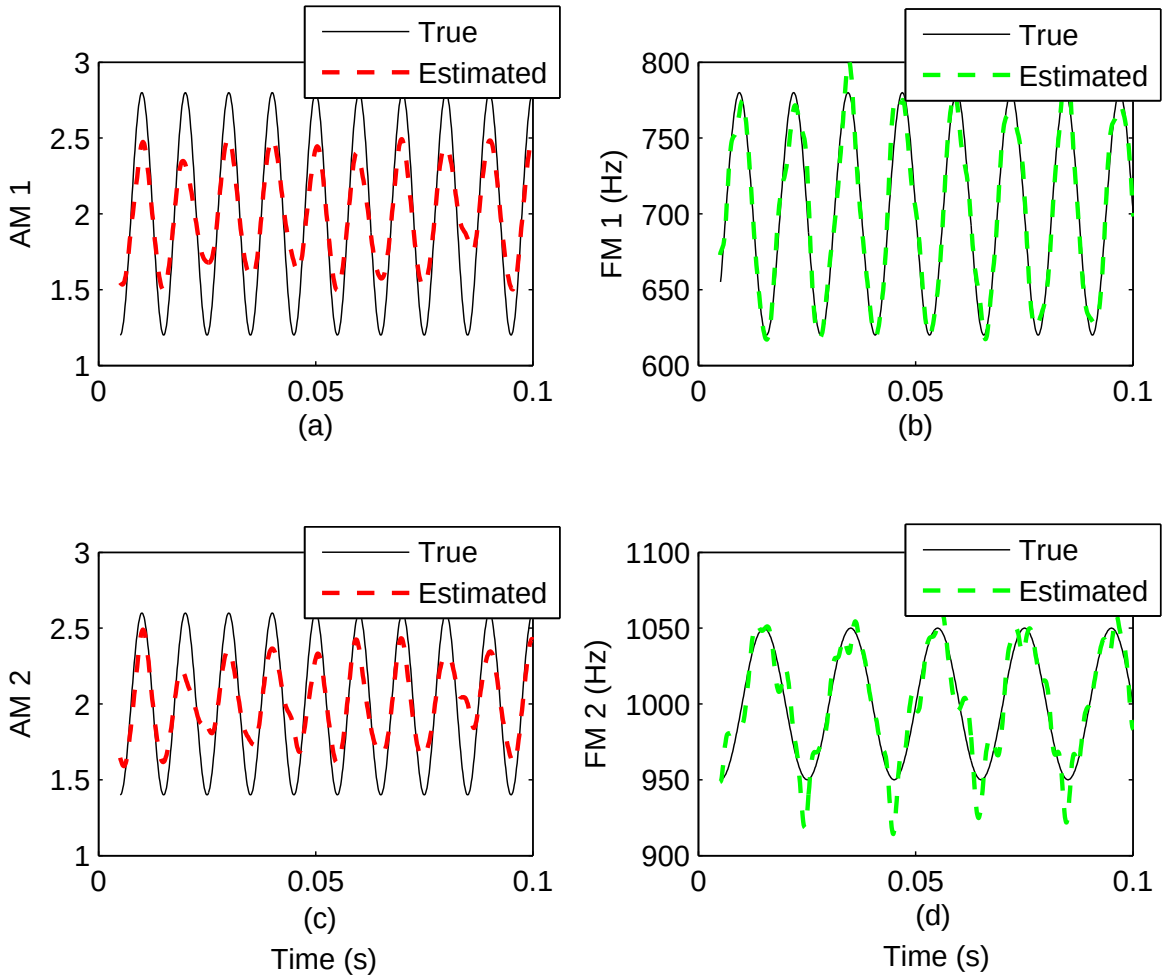


Figure 2.5: *Parameter estimation for aQHM. Upper panel: Amplitude (a) and Frequency (b) estimation for first component. Lower panel: Amplitude (c) and Frequency (d) estimation for second component.*

eaQHM performs better than aQHM in the estimation of the time varying frequencies and, especially, of the time varying amplitudes.

To test the robustness of the estimations provided by the eaQHM, additive white Gaussian noise of 20 and 10 dB SNR was added to the synthetic signal $x(t)$ described above. For comparison purposes, results for the aQHM are also provided. The performance of the algorithms is measured through the Mean Absolute Error (MAE) for amplitudes and frequencies.

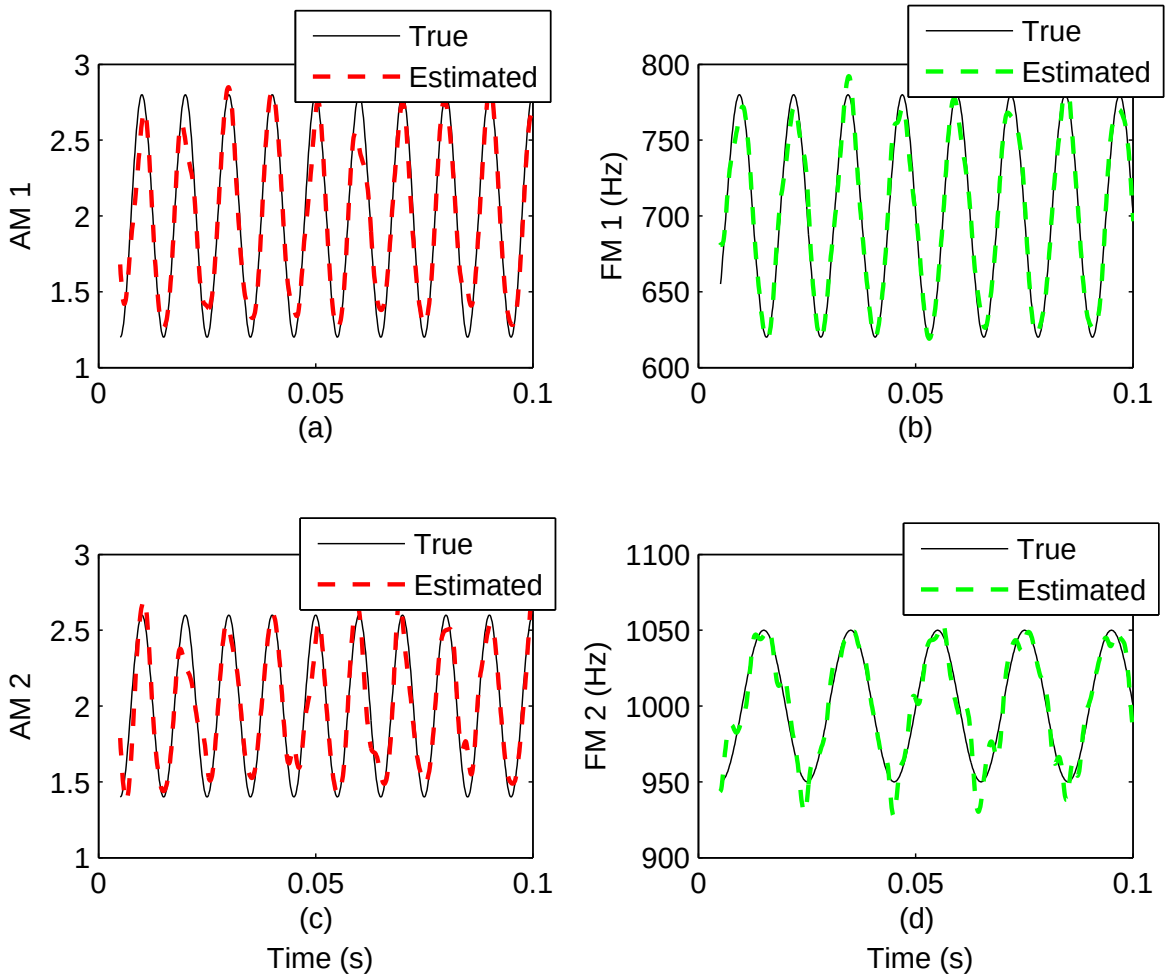


Figure 2.6: *Parameter estimation for eaQHM. Upper panel: Amplitude (a) and Frequency (b) estimation for first component. Lower panel: Amplitude (c) and Frequency (d) estimation for second component.*

The MAE of a parameter θ is defined as:

$$MAE\{\hat{\theta}\} = \frac{1}{M}\sum_{i=1}^{M}|\hat{\theta}^{(i)} - \theta| \tag{2.55}$$

where $\theta^{(i)}$ is the estimated parameter at the i^{th} simulation, and M is the number of Monte Carlo simulations. The results shown in this section are based on $M = 10000$ Monte Carlo simulations and the length of a Hamming analysis window for both models was 10 ms. The analysis step size was set to 1 sample. Table 2.2 presents the MAE and SRER scores for the aforementioned levels of noise.

MAE scores and SRER						
SNR	Model	$a_1(t)$	$a_2(t)$	$F_1(t)$	$F_2(t)$	SRER(dB)
∞	aQHM	0.2380	0.1842	7.6105	9.1731	22.6
	eaQHM	0.0889	0.0949	5.9217	7.0505	42.0
20 dB	aQHM	0.2235	0.1735	7.2704	7.8563	18.2
	eaQHM	0.1036	0.1079	6.1682	7.1241	20.3
10 dB	aQHM	0.2317	0.1860	8.6071	9.0302	10.7
	eaQHM	0.1490	0.1476	8.0513	8.1022	10.9

Table 2.2: *MAE scores and SRER for aQHM and eaQHM for* 10^4 *Monte Carlo simulations.*

2.6.2 Validation on Voiced Speech

The next step is to test the proposed model on real speech, and in particular, on voiced speech signals. The suggested iterative AM-FM decomposition algorithm based on aQHM/eaQHM can be applied on voiced speech signals in a straightforward way. Actually, the aQHM/eaQHM algorithm can be applied on a large voiced speech segment. Indeed, assuming that voiced speech is quasi-periodic and that the frequency content of voiced speech signals does not change very fast, then we only need to provide the fundamental frequency of the first voiced frame at the beginning of the voiced segment, $f_0(t_1)$, and then assume $\hat{f}_k^0(t_1) = kf_0(t_1)$. Applying QHM analysis on the first voiced frame, an updated set of $\hat{f}_k$ can be obtained for that frame. The updated set of frequencies can then be used as initial estimates for the next voiced frame. Continuing in this way, the whole voiced region will be analyzed by providing just the fundamental frequency for the first frame of the voiced segment. It is worth noting that the accuracy of the fundamental frequency estimator is not crucial for aQHM/eaQHM, since frequency mismatches are easily corrected.

For our purpose, we consider a voiced speech signal from the CMU-ARCTIC database with sampling frequency $F_s = 16$ kHz and duration of about 0.35 sec. For both algorithms, the number of harmonics was set to $K = 40$ and an estimate of the fundamental frequency of the beginning of the segment was given to the algorithm. At most 10 adaptations were allowed to the models. The analysis window size was 2.5 pitch periods and the analysis step size was 1 sample. In the following, the signals are considered up to a fixed maximum voiced frequency (5500 Hz). The original signal, along with the aQHM/eaQHM reconstructed signals and corresponding reconstruction errors, are shown in Figure 2.7.

To objectively compare the performance of both algorithms, the SRER defined in Eq. (2.53) was used. The SRER was

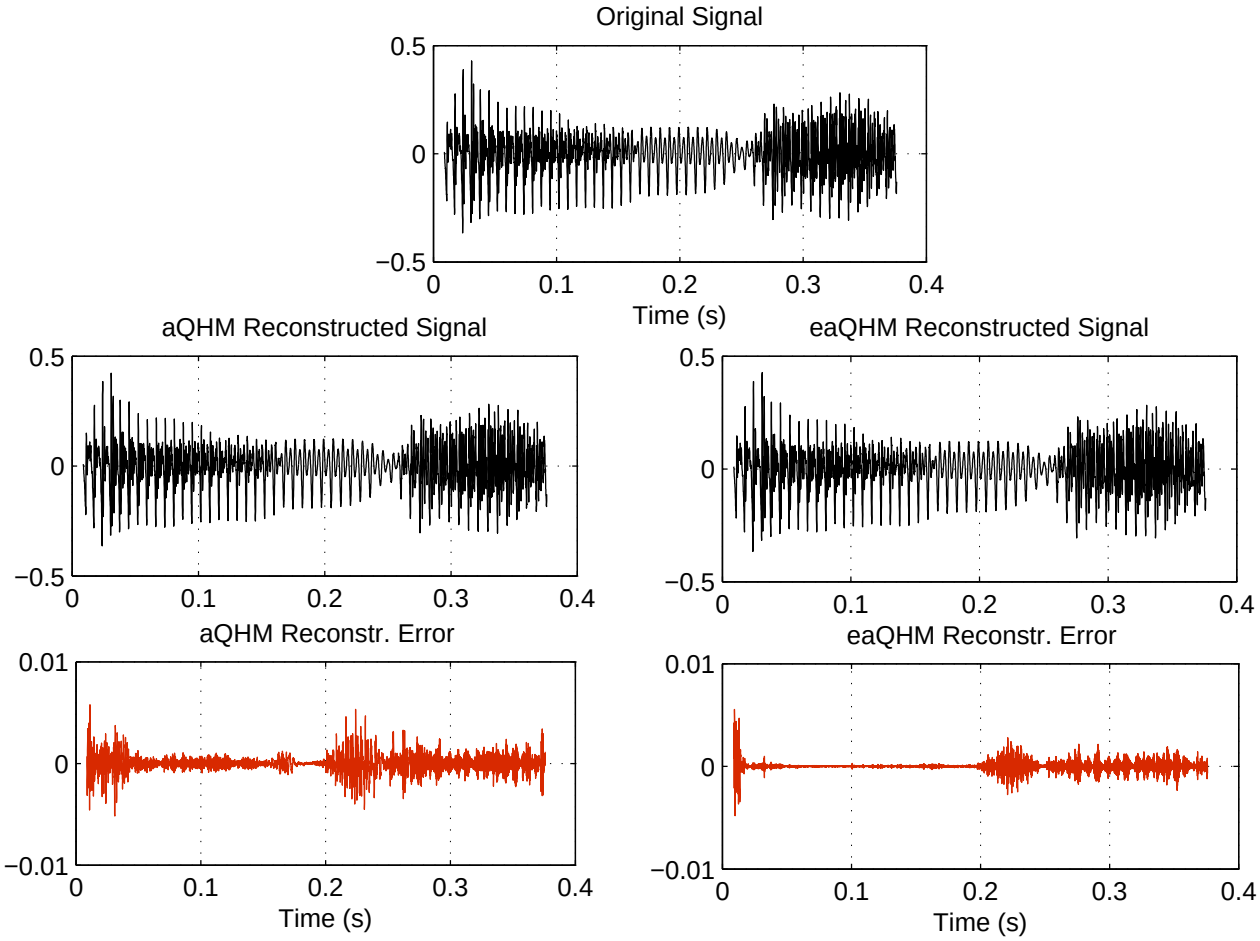


Figure 2.7: *Upper Panel: Original signal. Middle panel: aQHM (left) and eaQHM (right) reconstructed signal. Lower panel: aQHM (left) and eaQHM (right) reconstruction error.*

41.2607 dB for aQHM and 45.2166 dB for eaQHM. Two adaptations for aQHM and three adaptations for eaQHM were necessary for the models to converge.

To confirm these results, a large-scale objective test was performed. Using three different step sizes, namely 1ms, 2ms, and 4ms, we analyzed and reconstructed about 50 minutes of voiced speech from 3 speakers in the ARCTIC database. The sampling frequency of the speech signals was downsampled to 16kHz. A Hamming window of fixed length was

used; 3 times the average pitch period of the speaker. The same window was used for both aQHM and eaQHM. The number of components was set to $K = 30$. The average and standard deviation of the SRER (in dB) is provided in Table 2.3, along with various time-steps. Table 2.3 also presents the average number of adaptations (NoA) needed for the algorithms to converge. It is evident that, on average, eaQHM scores higher in terms of SRER, requiring, however,

ARCTIC database evaluation				
Step	Method	Mean (dB)	Std (dB)	NoA
1 msec	aQHM	34.5	4.6	2.9
	eaQHM	35.8	5.7	3.8
2 msec	aQHM	31.0	4.0	3.5
	eaQHM	33.2	5.0	3.9
4 msec	aQHM	30.8	3.4	3.6
	eaQHM	32.8	4.6	6.1

Table 2.3: *Mean and Standard Deviation of SRER (in dB) for approximately* 50 *minutes of voiced speech from the ARCTIC database.*

slightly more iterations than aQHM.

2.7 The adaptive Harmonic Model - aHM

For completeness in this section, a very brief review of the adaptive Harmonic Model (aHM) is presented [DS12]. The adaptive Harmonic Model can be mathematically described as:

$$x(t) = \sum_{k=-K}^{K} a_k(t) e^{jk\phi_0(t)} w(t) \tag{2.56}$$

where $a_k(t)$ is a time-varying complex function that copes with the amplitude and the instantaneous phase of the k^{th} harmonic component, $w(t)$ is the analysis window with support in $[-T, T]$, while K is the number of the harmonics, and $\phi_0(t)$ is a real function defined as the integral of the fundamental frequency $f_0(t)$:

$$\phi_0(t) = \int_0^t 2\pi f_0(u) du \tag{2.57}$$

The aHM has been successfully applied in speech analysis. Although it supports a strictly harmonic representation, the adaptation part comes from the aQHM-based theory: aHM projects a speech segment on a time-varying, *harmonically-related* set of exponential basis functions:

$$x(t) = \sum_{k=-K}^{K} (a_k + t b_k) e^{jk\phi_0(t)} w(t) \tag{2.58}$$

where $x(t)$ is the modeled analytic signal of the speech segment, and a_k, b_k are the already introduced parameters of the QHM. To obtain the a_k, b_k parameters, a Least-Squares minimization is set up:

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = (\mathbf{E_h}^H \mathbf{W}^H \mathbf{W} \mathbf{E_h})^{-1} \mathbf{E_h}^H \mathbf{W}^H \mathbf{W} \mathbf{s} \tag{2.59}$$

where $\mathbf{W}$ is the matrix containing the window values in the diagonal, $\mathbf{s}$ is the input signal vector, the matrix $\mathbf{E_h} = [\mathbf{E_{h0}} | \mathbf{E_{h1}}]$, and the submatrices $\mathbf{E_{hi}}$, $i = 0, 1$ have elements given by

$$(E_{h0})_{n,k} = e^{jk\phi_0(t_n)} \tag{2.60}$$

and

$$(E_{h1})_{n,k} = t_n e^{jk\phi_0(t_n)} = t_n (E_{h0})_{n,k}, \tag{2.61}$$

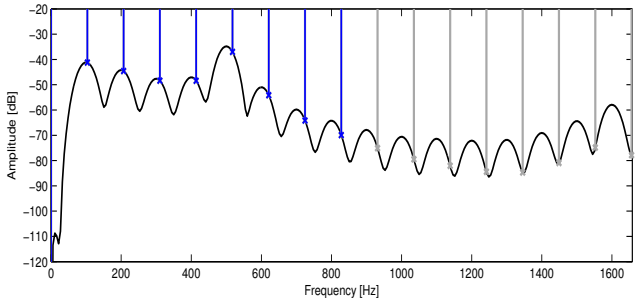
a solution structure similar to the one described earlier for the Quasi-Harmonic Models. The parameters $\hat{a}_k, \hat{b}_k$ provide a frequency correction estimate $\hat{\eta}_k$ for each frequency. These corrections combine together to form a correction term

related to the fundamental frequency, as a mean of the correction terms $\hat{\eta}_k$ relative to f_0:

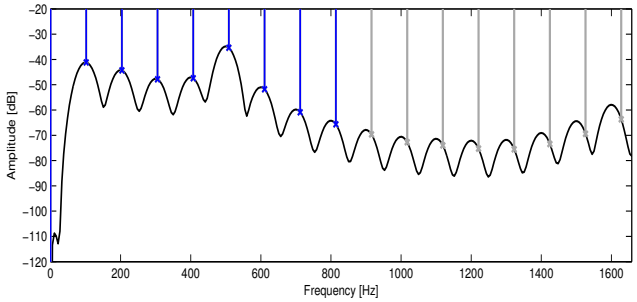
$$\hat{f}_0^{corr} = \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{\eta}_k}{k} \tag{2.62}$$

where K is the number of harmonics, and K is small at first ($K \approx 4-5$). The $\hat{f}_0^{corr}$ term updates the f_0 estimation (and thus, the ϕ_0 of Eq. (2.57)). When $|\hat{f}_0^{corr}|$ is small, the current set of harmonics are considered to have converged to the actual frequency values, and thus more harmonics can be added, and a new $\hat{f}_0^{corr}$ can be estimated. Based on this iterative process, a dedicated algorithm that successively refines the fundamental frequency curve to accurately localize the frequencies of the harmonics up to the Nyquist frequency is built. This algorithm is named as the *Adaptive Iterative Refinement - AIR* of f_0. Further information on AIR can be found in [DS13], but an illustrative example will be presented here.

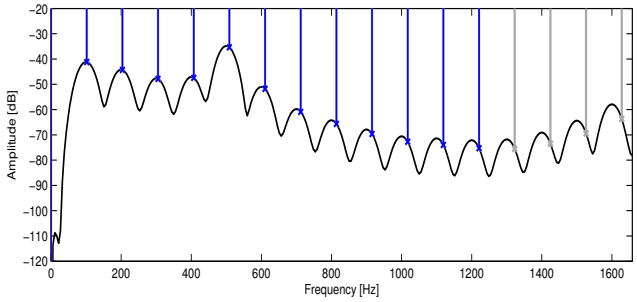
Let us consider the spectrum of a speech segment, which is purely harmonically modeled in its low band, where the frequency mismatch is considered small, as in Fig. 2.8a. Using the frequency correction mechanism of the QHM, a better estimate of the fundamental is obtained, and thus the first few harmonics in the low band are refined (Fig. 2.8b). Then, more harmonics are added (Fig. 2.8c) in the process, and further correction is applied on the fundamental frequency, resulting in more precise estimates of the harmonics up to the mid-band (Fig. 2.8d). This scheme continues until all harmonics are well localized up to the Nyquist Frequency (Figs. 2.8e, 2.8f). Compared to aQHM and eaQHM, the aHM



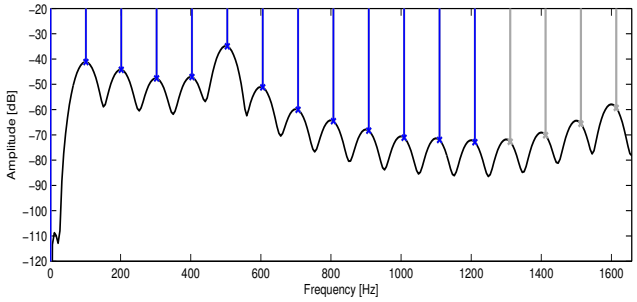
(a) At first, a small number of harmonics is used, where the frequency error is considered small.



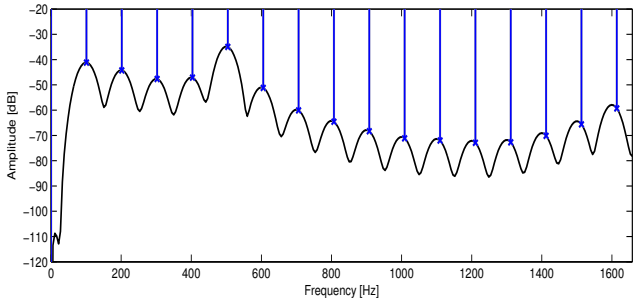
(b) Using the frequency correction mechanism, the f_0 and all its multiples are refined.



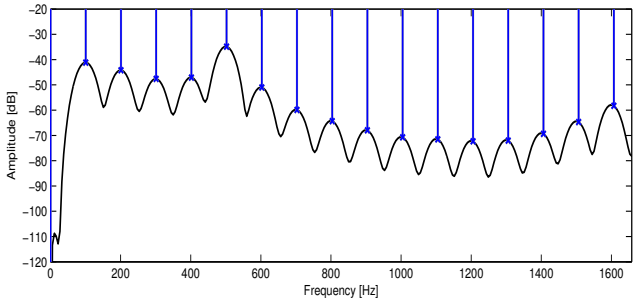
(c) A few more harmonics are considered for f_0 correction, up to the mid-band.



(d) Using the frequency correction mechanism, the f_0 and all its multiples are refined.



(e) All harmonics are considered for f_0 correction, up to the full-band.



(f) A final refinement of the f_0 is obtained and consequently all harmonics are well localized.

Figure 2.8: *Illustration of the Adaptive Iterative Refinement - AIR algorithm of aHM.*

adapts the parameters of the model only once during analysis. However, during AIR-f_0, the a_k and b_k parameters of aQHM are iteratively estimated, since they are involved in the refinement of the f_0. The aHM will be presented in more depth in the following chapter.

2.8 Number of parameters

For analysis and synthesis of a single frame, Table 2.4 presents an overview of the analysis and synthesis complexity of the SM, aQHM, eaQHM, and aHM to allow comparison. Complexity is considered as the number of parameters per frame each model requires to estimate (analysis) and represent (synthesis) K sinusoids. Please notice that the SM has the same complexity in the analysis and synthesis stages, while aQHM and eaQHM fit more parameters iteratively (a few times until convergence) during the analysis stage than for resynthesis. Finally, although the aHM is a harmonic model and consequently has less parameters for synthesis ($2K + 2$), when the AIR-f_0 algorithm is used, the a_k and b_k parameters of QHM are involved in the fundamental frequency refinement, and that implies an increase in complexity.

	Real parameters per frame			
	SM	**aQHM**	**eaQHM**	**AIR-aHM**
Analysis	$3K+1$	$5K+1$	$5K+1$	$(4K+1)+1$
Synthesis	$3K+1$	$3K+1$	$3K+1$	$(2K+1)+1$

Table 2.4: *Comparison of model complexity between SM, aQHM, eaQHM, and AIR-aHM for the analysis and synthesis stages. The table presents the number of real parameters per frame as a function of the number of sinusoids K to estimate (analysis complexity) and to represent (synthesis complexity) signals. Please note that the $+1$ term in all models corresponds to the mean value (DC component) of the signal.*

2.9 Conclusions

In this chapter, the adaptive sinusoidal models were presented. Three variants were discussed: the aQHM, the aHM, and the eaQHM which is proposed in this thesis. In the aQHM, the frequency (and thus, the phase) of the signal is adapted to the local characteristics of the analyzed signal. In the eaQHM, the amplitude, along with the frequency of the signal, is included in the adaptation process in a straightforward way. In the aHM, the instantaneous phases of the basis functions are integer multiples of the instantaneous phase of the fundamental frequency, which is iteratively estimated via a dedicated algorithm. The parameters of all models are computed using Least-Squares minimization. Experiments on synthetic signals showed that eaQHM performs better than aQHM in terms of MAE and SRER. The robustness of the eaQHM in the presence of white Gaussian noise is also demonstrated. Experiments on voiced speech using the ARCTIC database showed that eaQHM outperforms aQHM in terms of signal reconstruction.

Part II

Speech Analysis, Synthesis, and Modifications

Chapter 3

Related Work

In this Chapter, related work on the subject of speech analysis, synthesis, and modifications is presented. For the sake of clarity and convenience, related work is divided into two subsections: parametric and non-parametric techniques. A presentation of the most important schemes will be carried out in this work. All methods of speech analysis, modification, and synthesis will be described starting from earlier approaches up until the latest ones, in order to show the evolution of the scientific area throughout the years. Also, this way it is easier to highlight major improvements over the methods. However, emphasis will be given in parametric approaches, since they are closer to the models-in-hand.

3.1 Non Parametric Techniques

Non-parametric techniques are mostly based on the Short-Time Fourier Transform (STFT), with the so-called *Phase Vocoder* being the most well-known representative. Phase Vocoder is almost totally a frequency based technique. Other approaches include time domain techniques, such as the Overlap-Add method (OLA) and its variants, with Pitch-Synchronous OLA (PSOLA) being a milestone. The most significant ones from each perspective are discussed next.

3.1.1 The Phase Vocoder

The Short-time Fourier Transform (STFT) is the basis for the Phase Vocoder. Here, the basic mechanisms behind STFT analysis, modifications, and synthesis are presented.

Analysis-Synthesis

The idea behind STFT analysis is that a Fourier Transform is performed on a windowed speech segment, then a shift in time is made, and another Fourier Transform is applied, and so on, as it is depicted in Figure 3.1. The successive window locations are called *analysis time instants*, denoted by $t_a(k)$. Usually, the time shift is constant, and called *frame rate*. That means the analysis time instants are regularly spaced, $t_a(k) = lR$, where R is the frame rate.

Thus, the STFT can be mathematically described as

$$X(t_a(k), \omega) = \sum_{n=-\infty}^{\infty} h_u[n] x[t_a(k) + n] e^{-j\omega n} \tag{3.1}$$

where $h_u[n]$ is the analysis window, and $x[n]$ is the speech signal. The role of the window is to select and weighten the speech segment to be analyzed. The window is of finite duration and symmetric, and is of low-pass type. $X(t_a(k), \omega)$ is also called the short-time analysis spectrum around time instant $t_a(k)$. It can be easily observed that for a given value of ω, the STFT $X(t_a(k), \omega)$, can be viewed as the output of a complex band-pass filter centered at ω, and sampled at the successive time-instants $t_a(k)$. The STFT can also be described in terms of polar coordinates, magnitude and phase (hence the name):

$$X(t_a(k), \omega) = M(t_a(k), \omega) e^{j\phi(t_a(k), \omega)} = M(k) e^{j\phi(k)} \tag{3.2}$$

In practice, the STFT is evaluated at discrete frequencies, $\Omega_l = \frac{2\pi l}{N}$, using the FFT, where N is the length of the FFT.

Having this time-frequency representation of speech, the modification stage consists of applying the desired modifications to the stream of short-time analysis spectra, to obtain the short-term synthesis spectra, $Y(t_a(k), \omega)$, and then to

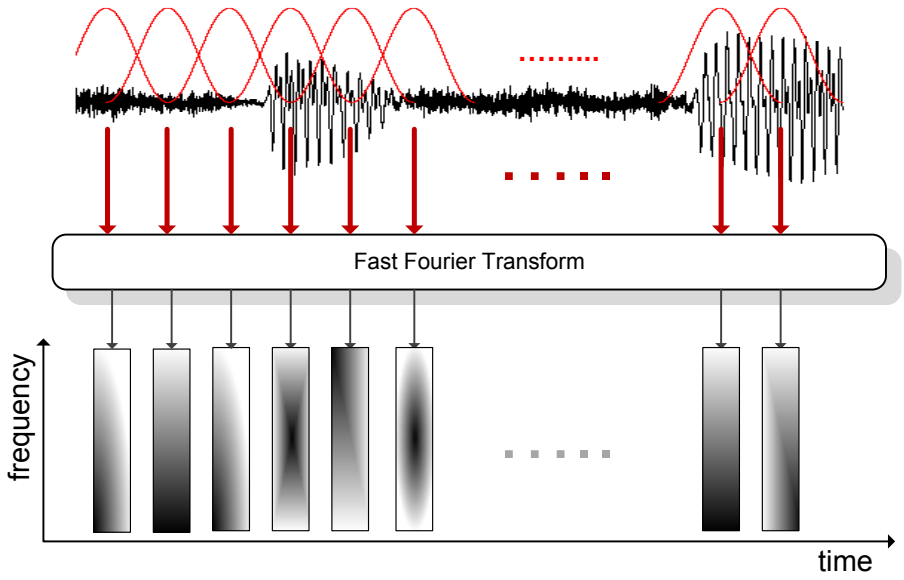


Figure 3.1: *The Short Time Fourier Transform*

synchronize them on a new set of time instants, called *the synthesis time instants*, $t_s(k)$. This stream of synthesis time instants is derived from the analysis time instants, according to the desired modification, as it will be explained later. Finally, the synchronized short-time synthesis spectra should be combined to produce the desired modified signal. However, it should be noted that, in general, there is no time-domain signal $y[n]$ that has as its STFT the $Y(t_s(k), \omega)$. The stream of short-time synthesis spectra must satisfy strong consistency conditions, described in [Por80]. That is because the stream of short-time synthesis spectra correspond to *overlapping* short-time signals. A solution to that was proposed by Griffin [GL84], by using the euclidean norm to estimate the signal $y[n]$ whose STFT $\hat{Y}(t_s(k), \omega)$ around time instants $t_s(k)$ best fits the modified STFT $Y(t_s(k), \omega)$. Solution to this minimization problem can be derived in closed form and is similar to the overlap-add procedure.

However, for time and pitch scale modifications, it is convenient to refer to a parametric model of speech production, although its parameters are not explicitly estimated. Portnoff [Por81] introduced a flexible quasi-stationary model of speech, and this will be discussed next. The connection between STFT analysis and this model will be presented, as well as the time and pitch scale modifications, with reference to this speech production model.

Stationary representation of speech As it is generally accepted, the speech production model is based on a time-varying linear filter driven by an excitation signal, which is either a sum of narrow-band signals with nearly harmonic instantaneous frequencies (voiced speech), or a stationary random sequence, with a flat power spectrum (unvoiced speech). If we denote as $g_n[m]$ the response of the system at time n to a unit-sample applied at time $(n-m)$, then $G(n,\omega)$ is the Fourier transform of $g_n[m]$ with respect to index m, i.e.

$$\sum_{m=-\infty}^{\infty} g_n[m]e^{-j\omega m} = G(n,\omega)e^{j\psi(n,\omega)} \tag{3.3}$$

$G(n,\omega)$ and $\psi(n,\omega)$ are referred to as the time-varying amplitude and phase of the system, respectively. Thus, $g_n[n]$ can be considered as nearly constant, since it represents the movements of physical articulators and it is usually relatively slow, compared to the variation of the speech waveform. In voiced speech, the excitation waveform $e[n]$ is represented as a sum of harmonic complex exponentials with unit amplitudes, zero inital phase, and slowly varying fundamental frequency function $2\pi/P[n]$, where $P[n]$ is the local pitch period. Mathematically,

$$e[n] = \sum_{k=0}^{P[n]-1} e^{j\Phi_k[n]} \tag{3.4}$$

where $\Phi_k[n]$ is the excitation phase of the k^{th} harmonic. That phase is defined as the integral of the time-varying frequency $\omega_k[n] = 2\pi k/P[n]$:

$$\Phi_k[n] = \sum_{m=0}^{n} \omega_k[m] = \sum_{m=0}^{n} \frac{2\pi k}{P[m]} \tag{3.5}$$

Considering $P[n]$ as nearly constant around time instant n, the following approximation can be derived:

$$\Phi_k[m] \approx \Phi_k[n] + \omega_k[n][m-n],\ |m-n| < \epsilon \tag{3.6}$$

where ϵ is a small constant. Now, the output of the time-varying filter is given by the convolution of the excitation and the filter impulse response:

$$x[n] = \sum_{m=-\infty}^{\infty} g_n[m]e[n-m] \tag{3.7}$$

Using the assumption of constant $P[n]$ for the duration of $g_n[m]$, the excitation signal can be replaced by its local harmonic representation:

$$x[n] = \sum_{k=0}^{P[n]-1} G(n,\omega_k[n])e^{j(\Phi_k[n]+\psi(n,\omega_k[n]))} = \sum_{k=0}^{P[n]-1} A_k[n]e^{j\theta_k[n]} \tag{3.8}$$

where $A_k[n]$ is the system amplitude at the harmonic frequency $\omega_k[n]$. The phase $\theta_k[n]$ of the k^{th} harmonic is the sum of the excitation phase $\Phi_k[n]$ and the system phase $\psi_k[n] = \psi(n,\omega_k[n])$:

$$\theta_k[n] = \Phi_k[n] + \psi(n,\omega_k[n]) = \Phi_k[n] + \psi_k[n] \tag{3.9}$$

The term $\theta_k[n]$ is referred to as *the instantaneous phase* of the k^{th} harmonic. Finally, since the system phase is slowly varying, the instantaneous phase can be expressed as

$$\theta_k[m] = \theta_k[n] + \omega_k[n](m-n),\ |m-n| < \epsilon \tag{3.10}$$

Further information on how these parameters are estimated are given in [ML95]

Pitch and Time Scale Modifications

With this brief description of the speech production model in hand, time and pitch scaling modifications can be addressed.

The object of time scaling is to alter the rate of articulation without affecting the spectral content. The pitch and the time evolution of the formant structure should be time-scaled, but not modified in any other way. A time-scale warping function would be useful, to map time instants in the original signal and time instants in the modified signal. This mapping $t \rightarrow t' = D(t)$ is called *the time-scale warping function*, and is often convenient to be expressed as an integral:

$$D(t) = \int_0^t \beta(\tau)d\tau \tag{3.11}$$

where $\beta(\tau)$ is positive and is called *the time modification rate*. For a fixed $\beta(\tau) = \beta$, the time scaling warping function is linear, and if $\beta > 1$, then speech is slowed down (time-scale expansion), whereas if $\beta < 1$, speech is sped up (time-scale compression). For time-varying time-modification rates, the time-scale warping function is non linear. Now, with reference to the speech production model introduced in the previous section, the speech parameters should be modified as follows:

$$n' \rightarrow P'[n'] = P[D^{-1}(n')] \tag{3.12}$$

$$n' \rightarrow A'_i[n'] = G'_i[n'] = G[D^{-1}(n'), \omega_i[D^{-1}(n')]] \tag{3.13}$$

$$n' \rightarrow \theta'_i[n'] = \Phi'_i[n'] + \psi\Big(D^{-1}(n'), \frac{2i\pi}{P[D^{-1}(n')]}\Big) \tag{3.14}$$

$$n' \rightarrow \Phi'_i[n'] = \sum_{m=0}^{n'} \frac{2i\pi}{P[D^{-1}(n')]} \tag{3.15}$$

The object of pitch scaling is to alter the fundamental frequency of speech without affecting the spectral envelope. A time-varying pitch-modification factor $t \rightarrow a(t) > 0$ is defined, which affects the pitch contour. Hence, the new pitch contour will be

$$n \rightarrow P'[n] = \frac{P[n]}{a[n]} \tag{3.16}$$

For $a(t) > 1$, the local pitch is increased by a factor of $a(t)$, whereas for $a(t) < 1$, the pitch is lowered by the same factor. Thus, the speech parameters are modified as follows:

$$n' \rightarrow P'[n'] = a[n']P[n'] \tag{3.17}$$
$$n' \rightarrow A'_i[n'] = G'_i[n'] = G[n', a[n']\omega_i(n')] \tag{3.18}$$
$$n' \rightarrow \theta'_i[n'] = \Phi'_i[n'] + \psi(n', a[n']\omega_i[n']) \tag{3.19}$$
$$n' \rightarrow \Phi'_i[n'] = \sum_{m=0}^{n'} a[m]\omega_i[m] \tag{3.20}$$

3.1.2 The Speech Transformation and Representation using Adaptive Interpolation of Weighted Spectrum (STRAIGHT) model

STRAIGHT is a hight-quality vocoder that uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region, and an excitation source design based on phase manipulation. It preserves the bilinear surface in the time-frequency region and allows for over 600% manipulation of such speech parameters as pitch, vocal tract length, and speaking rate, without further degradation due to the parameter manipulation. STRAIGHT is developed based on human speech perception system, which decomposes input sounds in terms of excitation (source) and resonant (filter) characteristics and it performs a periodic/aperiodic decomposition of the source signal and the estimation of pitch, as well as the spectrum. The quality of the resynthesized speech is high. However, the computational complexity is also very high and other parameters related to voice quality are not explicitly estimated.

More specifically, STRAIGHT decomposes the speech signal into three terms

- A smooth spectrogram, free from periodicities in time and frequency
- An f_0 contour, and
- A time-frequency periodicity map, which captures the spectral shape of the noise and also its temporal envelope

During the analysis, an f_0 contour is accurately estimated using a fixed-point algorithm. Then, this f_0 estimate is used to smooth out periodicity in the short-time spectrum using an f_0-adaptive filter and a surface reconstruction method. The result is a smooth spectrogram that captures vocal tract and glottal filters, but is free from the influence of f_0. During synthesis, pulses or noise with a flat spectrum are generated in accordance with voicing information and the f_0 contour. Speech is resynthesized from the smoothed spectrum and the pulse/noise component using an inverse FFT with an OLA technique.

From a modifications point of view, the step to be followed are very simple:

- Time-scale modification reduces to duplicating/removing ST slices from the STRAIGHT spectrogram and aperiodicity map.
- Pitch-scale modification reduces to modifying the f_0 contour

The three terms in STRAIGHT can be manipulated independently, which provides increased flexibility. STRAIGHT allows very high-factor prosodic modifications while maintaining the naturalness of the synthesized speech. The main disadvantage of STRAIGHT is its computational intensity. A schematic diagram for STRAIGHT is given in Figure 3.2.

3.1.3 The Overlap-Add (OLA) Methods

Pitch-Synchronous Overlap-Add (PS-OLA) is the basis for most time-domain techniques for time and pitch scale modifications. A simple description of this technique follows next, along with its most important variants.

Analysis-Synthesis

The analysis step consists of decomposing the speech signal $x[n]$ into a stream of short-time analysis signals, which can be denoted as $x[t_a(k), n]$, where $t_a(k)$ is the index of the short-time signal. An analysis window $h[n]$ is applied on

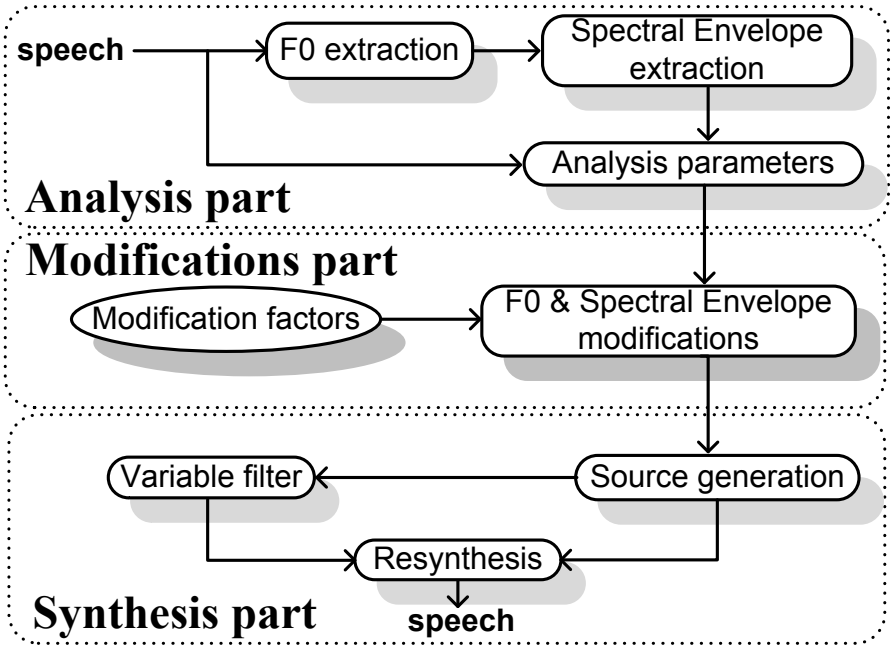


Figure 3.2: *STRAIGHT-based Analysis, Modification, and Synthesis.*

the signal in order to extract the short-time waveform, i.e.

$$x[t_a(k), n] = x[n]h[n - t_a(k)] \tag{3.21}$$

The time instants $t_a(k)$ are called *analysis pitch-marks*, and they are set in a pitch-synchronous manner on the voiced segments of speech and in a constant manner on the unvoiced ones. The length of the analysis window, denotes as T, varies between different PSOLA schemes, but it is proportional to the local pitch period, $P(k)$. For example, for Time-Domain PSOLA (TD-PSOLA), T is around $2P(k)$, whereas for Frequency-Domain PSOLA (FD-PSOLA), T is around $4P(k)$. Usually, a Hanning window is used in most implementations, although other choices can be made, according to mainlobe bandwidth and side-lobe attenuation.

The synthesis step is about transforming the stream of short-time analysis signals into a stream of short-time synthesis signals. These synthesis signals are synchronized on a new set of time instants, called synthesis time instants, $t_s(k)$, which are referred to as the *synthesis pitch-marks*. The stream of synthesis pitch-marks is determined from the analysis pitch-marks, according to the desired prosodic modification. Also, a mapping $t_s(k) \rightarrow t_a(k)$ is determined, which controls which analysis short-time signals should be selected for any given synthesis pitch-mark. A simple case consists of elimination or duplication of the analysis short-time signals, under the assumption that there is a one-to-one mapping between analysis and synthesis pitch-marks. In a more sophisticated approach, where there is no longer a one-to-one pitch-mark mapping, interpolation between successive short-time analysis signals lying closest to the synthesis pitch-mark is performed.

Finally, the synthetic output $y[n]$ is obtained by applying a weighted least-squares OLA procedure on the synchronized short-time synthesis signals. The analysis and synthesis windows are the same for TD-PSOLA, whereas in FD-PSOLA, the synthesis window is different, in order to account for the inherent changes in the frequency domain.

Pitch and Time Scale Modifications

The modifications can be considered as an intermediate step between analysis and synthesis. At first, the synthesis pitch-marks should be generated according to the desired pitch or/and time-scale modification, and then each synthesis pitch-mark is mapped with one or more analysis pitch-mark. Finally, the synthesis is carried out as described in the previous paragraph.

For pitch-scale modifications, the analysis pitch-marks are positioned pitch-synchronously, i.e.

$$t_a(k+1) - t_a(k) = P(t_a(k)) \tag{3.22}$$

where $P(t)$ is the pitch contour function $t \rightarrow P(t)$, and is considered constant within the analysis pitch-marks:

$$P(t) = P(t_a(k)), \;\; t_a(k) \leq t < t_a(k+1) \tag{3.23}$$

A new pitch contour, $P'(t)$, is specified according to the desired pitch modification. The stream of synthesis pitch-marks should be positioned pitch-synchronously with respect to this new pitch contour, that is

$$t_s(k+1) = t_s(k) + P'(t_s(k)) \tag{3.24}$$

For each synthesis pitch-mark, $t_s(k)$, we have

$$P'(t_s(k)) \approx \frac{P(t_s(k))}{\beta(t_s(k))} \tag{3.25}$$

where $\beta(t_s(k))$ is the pitch-scale modification factor. A recursive equation is available to determine these synthesis pitch-marks:

$$t_s(k+1) - t_s(k) = \frac{1}{t_s(k+1) - t_s(k)} \int_{t_s(k)}^{t_s(k+1)} \frac{P(t)}{\beta(t)} dt \tag{3.26}$$

and

$$\beta(t) = \beta(t_a(k)) = \beta_s, \;\; t_a(k) \leq t < t_a(k+1) \tag{3.27}$$

So, the synthesis pitch period $t_s(k+1) - t_s(k)$ is equal to a mean scaled pitch period in the original signal calculated over $[t_s(k+1), t_s(k)]$. Figures 3.3 and 3.4 show an example of pitch modifications based on TD-PSOLA.

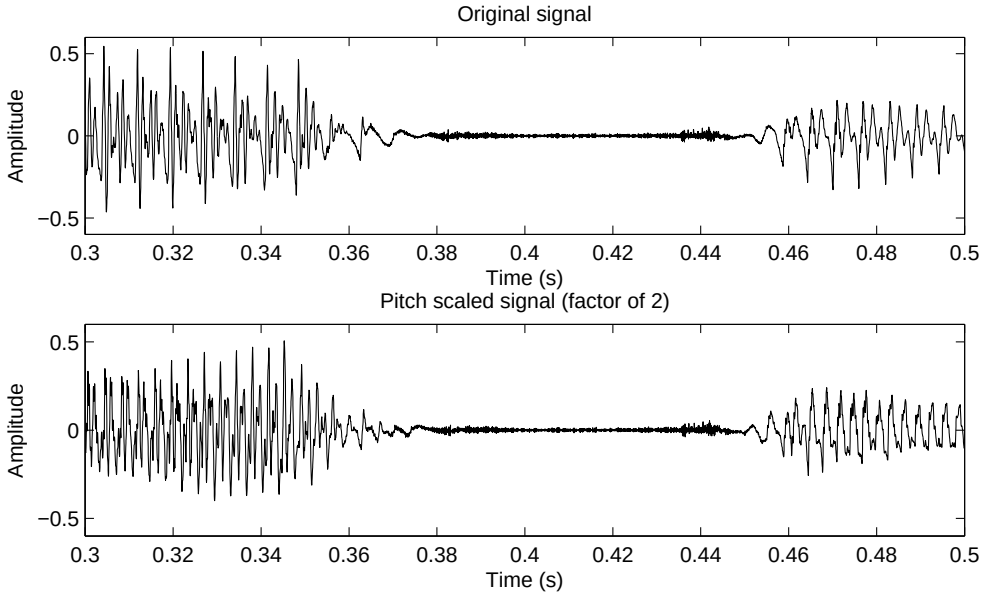


Figure 3.3: *Real signal TD-PSOLA pitch-scaling. Upper panel: Original waveform. Lower panel: Pitch-scale modified waveform. The pitch-scale modification factor is* 2*.*

For time-scaling, a time-scale modification factor a_s is associated with each analysis pitch-mark, from which a time-scale warping function can be formulated:

$$D(t_a(1)) = 0 \tag{3.28}$$

and

$$D(t) = D(t_a(k)) + a_s(t - t_a(k)), \;\; t_a(k) \leq t < t_a(k+1) \tag{3.29}$$

Since the pitch contour must be preserved in time-scaling, a stream of synthesis pitch-marks is obtained from the analysis pitch-marks, using the time-scale warping function. The pitch in the time-scaled signal at time t should be close to the pitch in the original signal at time $t' = D^{-1}(t)$. Now, a stream of synthesis pitch-marks should be found, such that

$$t_s(k+1) = t_s(k) + P'(t_s(k)) \tag{3.30}$$

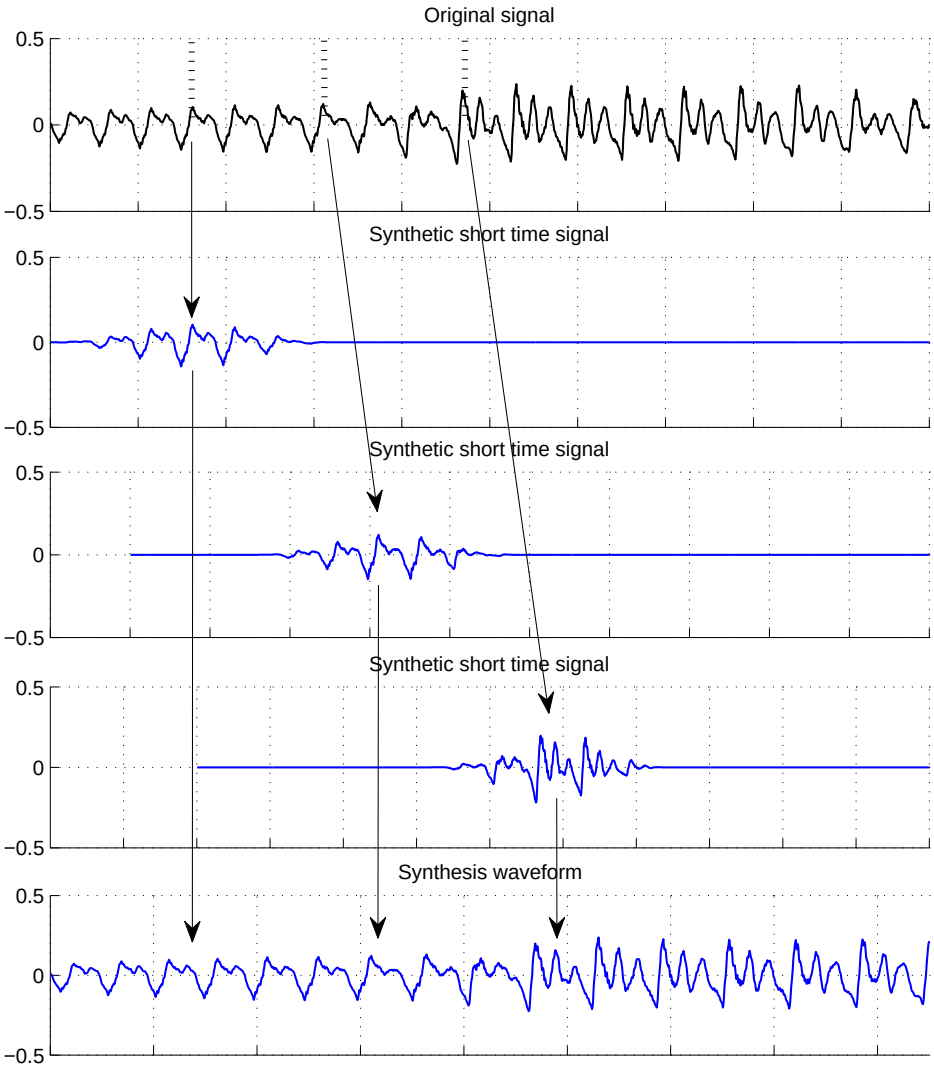


Figure 3.4: *Pitch-scale modification scheme with TD-PSOLA method. Upper panel: Original waveform. Middle panel: Three short time synthetic signals. Lower panel: Pitch-scale modified waveform.*

For this purpose, the introduction of the so-called *virtual pitch-marks*, $t_v(k)$, in the original signal is convenient, such that

$$t_s(k) = D(t_v(k)), \;\; t_v(k) = D^{-1}(t_s(k)) \tag{3.31}$$

Hence, we need to specify $t_s(k+1)$ so as $t_s(k+1) - t_s(k)$ is approximately equal to the original pitch at time $t_v(k)$. Mathematically, this can be expressed as

$$t_s(k+1) - t_s(k) = \frac{1}{t_v(k+1) - t_v(k)} \int_{t_v(k)}^{t_v(k+1)} P(t)dt \tag{3.32}$$

with $t_s(k+1) = D(t_v(k+1)$. In other words, the synthesis pitch period $t_s(k+1) - t_s(k)$ at time $t_s(k)$ equals to the mean value of the pitch in the original signal calculated over the "virtual" interval $[t_v(k), t_v(k+1)]$. Time scale modification for TD-PSOLA is shown in Fig. 3.5.

In Fig. 3.6, phase-vocoder and TD-PSOLA time-scaling are compared.

3.1.4 Other Approaches

Other variants of PSOLA include Frequency Domain PSOLA (FD-PSOLA) [CS86], which is only used for pitch-scaling and differs from TD-PSOLA in the definition of the short-time synthesis signals. Also, the modification is performed in the frequency domain of the short-time analysis spectrum, by adjusting the spacing between the harmonics

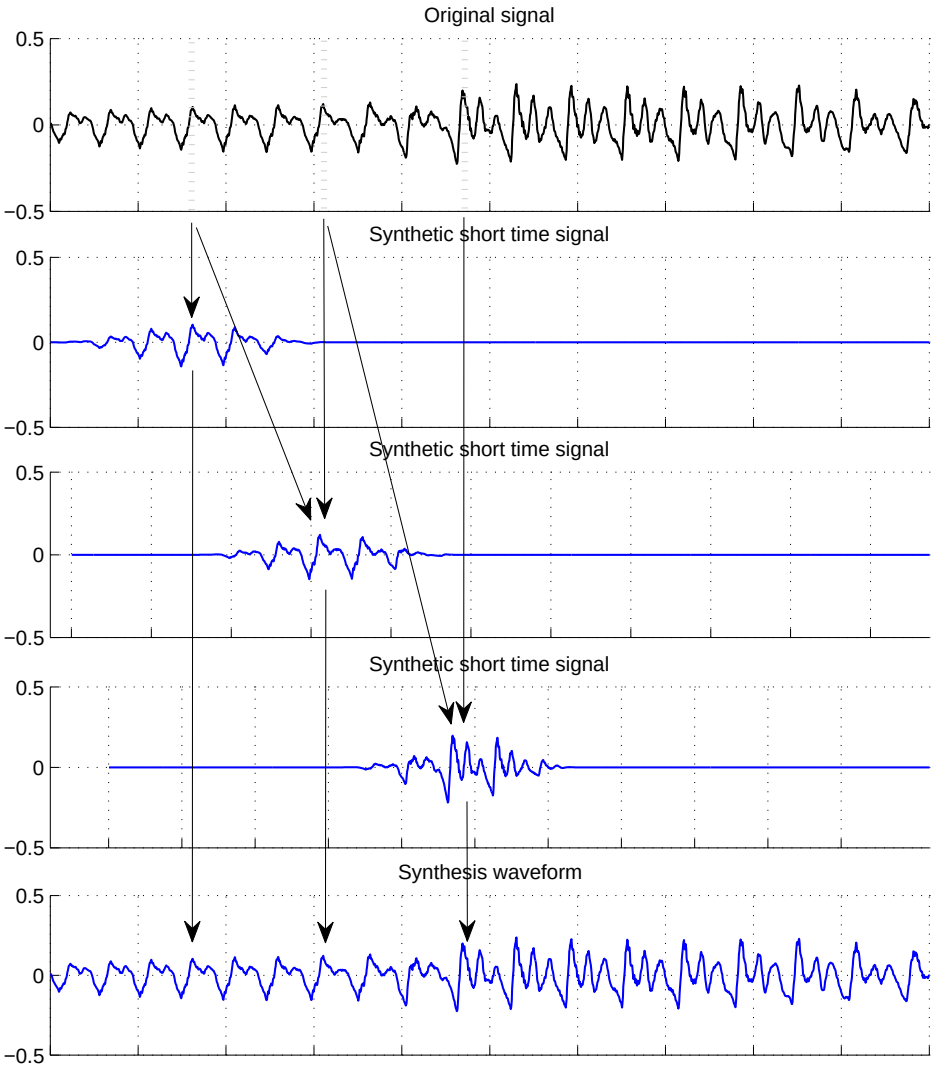


Figure 3.5: *Time-scale modification with TD-PSOLA method. Upper panel: Original waveform. Middle panel: Three short time synthetic signals. Lower panel: Time-scale modified waveform. Time-scale modification factor is* 2.

by a factor β_s.

Another approach is Linear Predictive PSOLA (LP-PSOLA) [MC90a]. The PSOLA scheme is embedded in a residual-excited vocoder, where AR models are fitted on the speech signal at a pitch-synchronous rate, and prosodic modifications are performed on the residual, which is obtained by methods of inverse filtering. The output is synthesized by filtering the modified residual with the time-varying synthesis filters. More details can be found in [MC90b] [MT92].

Similar to TD-PSOLA is the Waveform Similarity Overlap-Add (WSOLA) [VR93] method, where neighbouring frame similarities are exploited using autocorrelation techniques in order to avoid pitch period discontinuities or phase jumps at the synthesis boundaries of the traditional TD-PSOLA or SOLA [RW85].

The pioneering work of Griffin and Lim [GL88] on the Multiband Excitation (MBE) Vocoder should be mentioned. In MBE Vocoder, bands of the spectrum are separated as voiced or unvoiced. Methods to estimate the parameters of the speech model are presented and methods to synthesize speech from the model parameters are described. Specifically, the excitation and the spectral envelope parameters are estimated so that the synthesized spectrum fits, in the Least Squares sense, the original spectrum. However, the system is only capable of time-scaling speech signals.

Improvement over the Phase Vocoder have been proposed over the years. Laroche et al [LD99] [Puc95] proposed an explanation for the phasiness problem (often referred to as reverberation) and suggested an improvement for phase calculation that significantly reduces the problem. Also, the computational cost was reduced by more than a factor of two. Other improvements have been proposed in [Puc95].

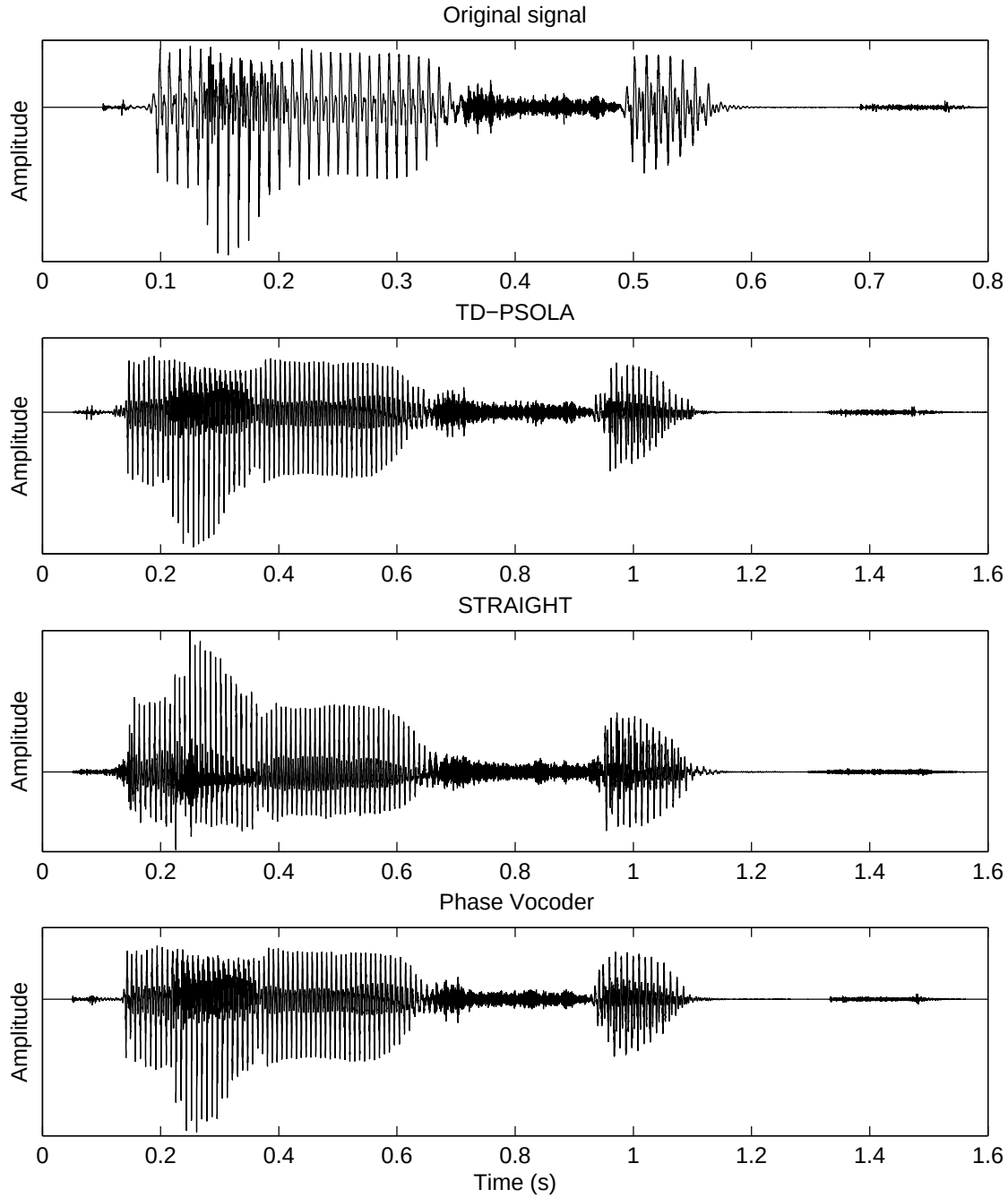


Figure 3.6: *Comparison between phase-vocoder, STRAIGHT, and TD-PSOLA time-scaling. First panel: Origial waveform. Second panel: Time-scaled signal obtained by the TD-PSOLA technique. Third panel: Time-scaled signal obtained by the STRAIGHT method. Fourth panel: Time-scaled singal obtained by the the phase-vocoder technique. Time-scale modification factor is* 2*.*

3.2 Parametric Techniques

Parametric techniques typically refer to methods which rely on a model of speech production, whose parameters are to be estimated. Two are the main representatives of such techniques: (1) a model with several in-series systems that represent the different stages of the human speech production system, i.e. excitation system, vocal tract system, and lips system, and (2) a model of time-series representation, i.e. a sum of frequency and/or amplitude modulated cosines.

Speech production based models have their origin in the work of Fant [FS66]. A simplification of the human speech production system is the following: (1) an excitation, which consists of a series of pulses, quasi-periodically placed in time, and represent the pressure signal that comes from the lungs, passes through the vocal cords. Models of the excitation, called *glottal models*, can be found in [Alk11]. (2) Subsequently, this excitation signal is modulated by the

vocal tract, which boosts or suppresses certain frequencies according to vocal tract resonances, the so-called *vocal tract formants*. The vocal tract is shown to be effectively represented by an *auto-regressive – AR* process. (3) Finally, the signal passes through the lips, which can be represented by a first order high-pass filter, and thus enforces high frequency components. The output of this system is the final speech signal. Although this model is simplified and is considered as a linear system – which is not the exact case for speech production – it can represent speech phenomena quite well. Parameters of such systems are typically the pitch period (distance between pulses of the excitation), the formant frequencies and bandwidths, and the order of the AR process, although other parameters may be added, according to a specific model. Modifications based on such systems include the variation of these parameters; pitch modification, formant and bandwidth increase or reduction, glottal parameter manipulations, etc.

On the other hand, time-series based parametric representations include the decomposition of speech into components: a *deterministic* part, which is usually modelled as a sum of frequency and/or amplitude modulated components, and a *stochastic* part, which is modelled as frequency-modulated Gaussian noise, usually weighted by a time-domain envelope. However, noise can be modelled by a sum of cosines as well, as in [MQ86]. Typically, the deterministic part represents voiced speech, whereas the stochastic parts represents unvoiced speech, friction noise, etc. More specifically, if the frequencies of the deterministic part are harmonically related, then the general model is called the *Harmonic* model. Therefore, various combinations have been made in literature: Deterministic plus Stochastic model [Ser89] [Sty96], Harmonic plus Noise model [Sty96], Sinusoidal plus Noise model [OdB99] and Quasi-Harmonic plus Noise model [PTRS10]. Moreover, because of the inability of such models to represent highly non-stationary parts of speech, such as stop consonants or transient speech areas, extended models have been suggested, generally called *Sinusoidal plus Noise plus Transients* models [Lev99] [Tho05].

Typical parameters of these models include the (harmonic or not) frequencies, amplitudes, and phases of the deterministic part, the number of sinusoids, whether an analysis frame is voiced or unvoiced, the time envelope of the noise, etc. Modifications using such models include parameter trajectory scaling. Interpolation schemes such as linear, cubic, or spline, are used to estimate the parameter trajectories in between analysis frames. However, attention should be paid both in phase coherence and in shape invariance of the resulting modified waveform.

The first approach was the Linear Predictive Vocoder [AH71], in which voiced speech is modelled as a convolution between a periodic train of pulses and a time-varying AR filter. Although quite popular at the beginning, it was soon abandoned due to its failure to provide high-quality modified speech. Later, Almeida and Marques [AS84] [MA89] were the first to propose the use of sinusoids for speech analysis and synthesis. In the late 80ies, Griffin [Gri87] and Serra and Smith [Ser89], had proposed a model where the sinusoids were harmonically related. However, milestones in sinusoidal modelling were the work of McAulay and Quatieri, the so-called *Sinusoidal Model*, which was followed by several variants, such as [GS92], and that of Stylianou[Sty96] on harmonic modelling, referred to as the *Harmonic Plus Noise Model*. The Sinusoidal Model is described next and the Harmonic Plus Noise Model follows. Finally, a recent source-filter model referred to as the *LF-ARX model* [VRC07] will be briefly discussed.

3.2.1 The Sinusoidal Model (SM)

In 1986, McAulay and Quatieri suggested their famous Sinusoidal Model (SM). In In this work, the speech waveform $s(t)$ is assumed to be the output of passing a vocal excitation waveform $e(t)$ through a linear system $h(t)$ representing the characteristics of the vocal tract. The system $h(t)$ is assumed to account for both the shape of the glottal pulse and the vocal tract impulse response.

Analysis-Synthesis

The commonly used binary unvoiced/voiced excitation model is replaced by a sum of sine waves of the form

$$e(t) = \sum_{k=1}^{N} a_k(t) \cos(\Omega_k(t)) \tag{3.33}$$

where N is the number of sinusoids, $a_k(t)$ is the time-varying amplitude associated with each sinewave, and the excitation phase $\Omega_k(t)$ is the integral of the time-varying frequency $\omega_k(t)$

$$\Omega_k(t) = \int_0^t \omega_k(\sigma) d\sigma + \phi_k \tag{3.34}$$

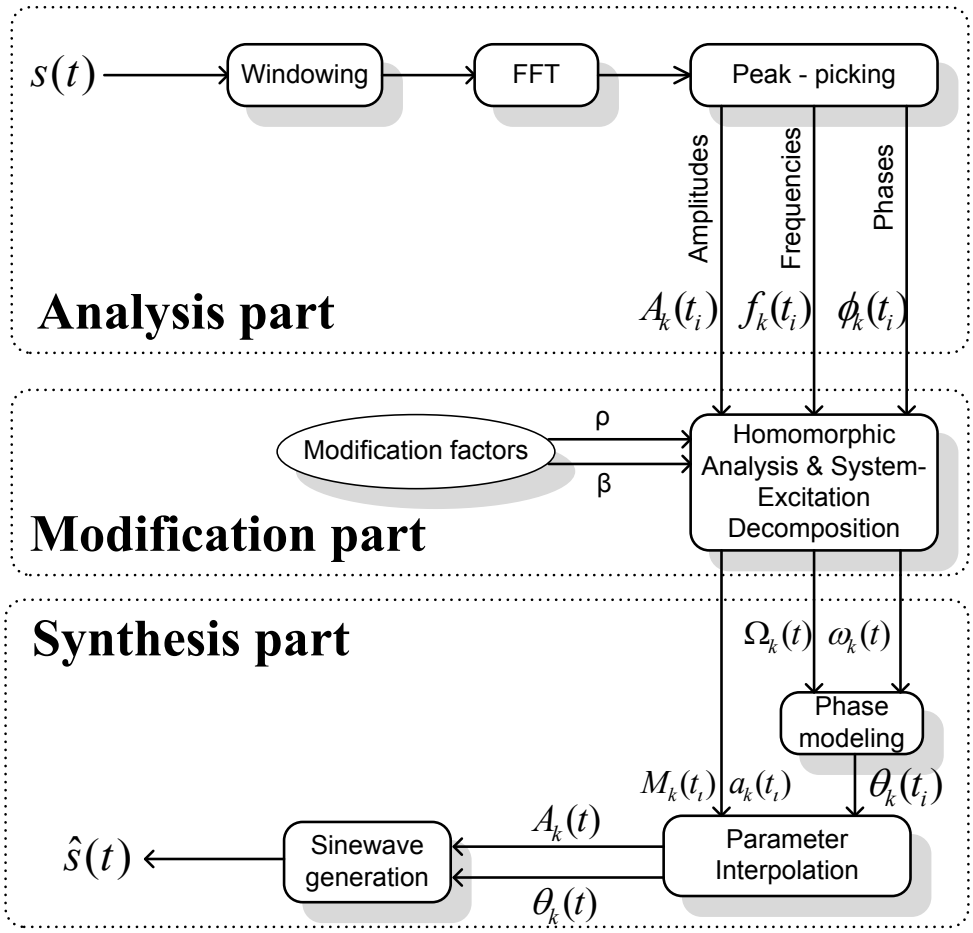


Figure 3.7: *Sinusoidal Analysis, Modification, and Synthesis.*

where ϕ_k is the fixed phase offset because the sinewaves are not generally in phase. The vocal tract transfer function $H(\omega; t)$ can be written as

$$H(\omega; t) = M(\omega; t)e^{j\psi(\omega;t)} \tag{3.35}$$

The dependence on t is due to the fact that the vocal tract impulse response is time-varying. The system amplitude and phase along each frequency trajectory $\omega_k(t)$ are given by

$$M_k(t) = M[\omega_k(t); t] \tag{3.36}$$

and

$$\psi_k(t) = \psi[\omega_k(t); t] \tag{3.37}$$

So, when the excitation signal $e(t)$ passes through the linear time-varying vocal tract $h(t)$, the output is the sinusoidal representation of the speech signal

$$s(t) = \sum_{k=1}^{N} A_k(t) \cos[\theta_k(t)] \tag{3.38}$$

where

$$A_k(t) = a_k(t)M_k(t) \tag{3.39}$$

and

$$\theta_k(t) = \Omega_k(t) + \psi_k(t) + \phi_k \tag{3.40}$$

represent the amplitude and phase of each sinewave along the frequency trajectory $\omega_k(t)$. The validity of this representation is subject to the stationarity assumption of the excitation amplitudes and frequencies, compared to the vocal tract impulse response.

The analysis consists of two steps. First, the estimation of frequencies, composite amplitudes and phases is performed using a high-resolution Fourier Transform magnitude. This is done in a frame-by-frame scheme, after applying a window on the speech frame. The second step accounts for the separation of the system and excitation components.

For the first step, let $S(\omega, kR)$ be the short-time Fourier Transform of the speech signal, and let R be the frame rate, so the estimated values are taken at kR sample indices. Hence, if $\hat{\omega}_l^k$ is the l^{th} frequency estimate of the k^{th} analysis frame, then the corresponding amplitudes and phases are given as

$$\hat{A}_l^k = |S(\hat{\omega}_l^k, kR)| \tag{3.41}$$

and

$$\hat{\theta}_l^k = arg[S(\hat{\omega}_l^k, kR)] \tag{3.42}$$

where arg denotes the principal value.

For the second step, the separation of system and excitation parameters is done using homomorphic deconvolution, under the assumption of the vocal tract transfer function being minimum phase. Thus, the excitation parameters at each analysis frame boundary are obtained as

$$\hat{a}_l^k = \hat{A}_l^k / \hat{M}_l^k \tag{3.43}$$

and

$$\hat{\Omega}_l^k = \hat{\theta}_l^k - \hat{\psi}_l^k \tag{3.44}$$

The synthesis is performed in three steps. First, there is a matching procedure between parameter values computed at two consecutive frame boundaries. Then, an appropriate interpolation scheme is applied on the resulting pairs of amplitude and phase samples of excitation and vocal tract functions over each frame, and finally the sinewaves are generated using the interpolated components.

The first step is about matching the excitation frequencies measured on frame k with those measured on frame $k+1$. After matching these frequencies, the matching of all other parameters comes easily, since they are measured at the excitation frequencies. An algorithm for matching the location of the spectral peaks was proposed in [MQ86], which uses a purely sinewave based model. One basic concept of this algorithm is the "birth-death" of a sinewave frequency. A brief example of how this algorithm works is the following:

Let ω_l^k and ω_l^{k+1} be the excitation frequency estimates of the l^{th} sinewave over frame k and $k+1$, respectively. In order to match these two frequencies (i.e. to assume that they belong to the same frequency trajectory), frequency ω_l^{k+1} should lie in the interval $[\omega_l^k - \Delta, \omega_l^k + \Delta]$. If it is not, then the frequency track associated with ω_l^k is considered "dead" in frame $k+1$, its amplitude is zeroed in this frame, and frequency ω_l^k is not considered any more. If, however, frequency ω_l^{k+1} lies in this "matching interval", and is the closest one to ω_l^k (since there might be more than one that lie in that interval - called candidates for matching), then it is declared as a definite match. After all matching is done, there might be some unmatched frequencies in frame $k+1$. If so, then new frequencies are "born" in frame k, with zero amplitude.

After parameter matching, the second step is about parameter interpolation. This is based on the assumption that the excitation and system functions are slowly varying across each frame along frequency tracks $\omega_l(t)$. System amplitudes M_l^i and excitation amplitudes a_l^i can be linearly interpolated. The system phases ψ_l^i can also be linearly interpolated, but this is not the case for the excitation phases and frequencies. Thus, a cubic polynomial is fitted on the excitation phase [MQ86].

Finally, the synthetic waveform is given by

$$\hat{s}[n] = \sum_{l=1}^{L[n]} \hat{A}_l[n] \cos(\hat{\theta}_l[n]) \tag{3.45}$$

where

$$\hat{A}_l[n] = \hat{a}_l[n]\hat{M}_l[n] \tag{3.46}$$

and

$$\hat{\theta}_l[n] = \hat{\Omega}_l[n] + \hat{\psi}_l[n] \tag{3.47}$$

where $L[n]$ is the number of sinewaves estimated at time n.

Pitch and Time Scale Modifications

Having this sinewave based model, prosody modifications are straightforward.

For time-scale transformations, the parameters that are scaled are the system amplitudes and phases, $M(\omega, t)$, and $\psi(\omega, t)$, and the excitation amplitudes and frequencies, $a_l(t)$ and $\omega_l(t)$. The modification of these parameters correspond

to moving slower or faster the vocal tract articulators and to stretching or compressing the frequency trajectories, respectively. Specifically, for an arbitrary time scale transformation, the time t_0 of the original articulation rate is mapped to the transformed time t_0' through the mapping

$$t_0' = W(t_0) \tag{3.48}$$

where $W(t)$ is the so-called *time warping function*. Fixed rate time-scaling will be briefly discussed here, since it is more convenient for understanding and the extension to time-varying rate change is straightforward. For fixed rate change ρ, the time warping function becomes $t' = \rho t \Rightarrow t = \rho^{-1}t'$. The mathematical model for time-scaled speech $s'(t')$ is given below:

$$s'(t') = \sum_{l=1}^{L(t')} A_l'(t)\cos(\theta_l'(t')) \tag{3.49}$$

where

$$A_l'(t') = A_l(\rho^{-1}t') = a_l(\rho^{-1}t')M_l(\rho^{-1}t') \tag{3.50}$$

and

$$\theta_l'(t') = \Omega_l'(t') + \psi_l(\rho^{-1}t') \tag{3.51}$$

where

$$\Omega_l'(t') = \int_{t_l'}^{t'} \omega_l(\rho^{-1}\tau)d\tau + \phi_l \tag{3.52}$$

The initial phase offset ϕ_l is chosen to preserve the impulselike nature of the excitation function during voicing, but in practice the onset timing may be different among sinewave components. This leads to the *phase dispersion* problem, which was handled in [QM92], by onset timing detection.

For pitch modification, the excitation frequencies must be shifted to new frequencies, according to a pitch scale modification factor β. Thus, frequency track $\omega_l(t)$ in the original signal corresponds to frequency track $\beta\omega_l(t)$ However, the model does not account for changes in the vocal tract spectral characteristics. For that, the system amplitudes, $M(\omega, t)$, and phases, $\psi(\omega, t)$, must be computed at the new frequency track locations $\beta\omega(t)$. Hence, the mathematical model for pitch-scaled speech $s'(t)$ is the following:

$$s'(t) = \sum_{k=1}^{L(t)} \hat{a}_l(t)\hat{M'}_l(t)\cos(\hat{\Omega}_l'(t) + \hat{\psi}_l'(t)) \tag{3.53}$$

where

$$\hat{M}_l'(t) = \hat{M}_l(\beta\hat{\omega}_l, t),\ \ \hat{\psi}_l'(t) = \hat{\psi}_l(\beta\hat{\omega}_l, t),\ \ \hat{\Omega}_l'(t) = \beta\int_{t_l}^{t} \omega_l(\tau)d\tau + \phi_l \tag{3.54}$$

Fig. 3.7 shows schematically the analysis, modification, and synthesis procedures.

3.2.2 The Harmonic Plus Noise Model (HNM)

During the early and mid-90ies, Stylianou proposed a new model, called the Harmonic plus Noise model (HNM) [Sty96]. In this model, speech spectrum is separated into two parts: a deterministic and a stochastic part, delimited by a time-varying maximum voiced frequency. In the lower band (deterministic part), the signal is considered to be harmonic. The stochastic part, which is the residual of the original signal minus the deterministic part, is modelled by an AR model and its time domain behaviour is imposed by a parametric time domain envelope. In this model, both the analysis and synthesis is performed in a pitch-synchronous manner, inspired by PSOLA. Thus, it can provide flexible techniques for time and pitch scaling, which will be discussed shortly.

The harmonic part accounts for the quasi-periodic phenomena of speech, while the noise part models non-periodic components, which typically include friction noise, unvoiced speech, etc. The time-varying maximum voiced frequency is used to determine the limit between the two parts.

In the lower band, the signal can be modelled as a sum of harmonically related sinusoids with slowly varying amplitudes and frequencies:

$$h(t) = \sum_{k=1}^{K(t)} A_k(t)\cos(k\theta(t) + \phi_k(t)) \tag{3.55}$$

where

$$\theta(t) = \int_{-\infty}^{t} \omega_0(u)du \tag{3.56}$$

and where $A_k(t), \phi_k(t)$ denote the amplitude and phase at time t of the k^{th} harmonic respectively, $\omega_0(t)$ is the fundamental frequency and $K(t)$ is the number of harmonics included in the harmonic part. The upper band contains the noise part. In voiced speech, the noise part exhibits a specific time domain structure in terms of energy distribution, i.e. it is concentrated in the part of the pitch period where the glottis is open. Thus, the frequency components of the noise part is described by a time-varying AR model, and its time domain structure is formed by modulation using a parametric envelope.

Analysis-Synthesis

Before applying the model on speech, an estimation of the fundamental frequency and the maximum voiced frequency is required. Hence, a pitch estimator similar to the one used in [Gri87] is used. Then, a voicing decision is made, and finally a refined pitch is defined as the fundamental frequency whose harmonics better fit the voiced frequencies detected in the lower band. Using this stream of pitch values, the position and duration of the analysis frames are set at a pitch-synchronous rare on the voiced portions of speech and at a fixed rate on unvoiced parts.

On the voiced frames, an estimate of the parameters are obtained at the center t_i of the analysis window. Thus, the model can be rewritten as

$$h(t) = \sum_{k=-K}^{K} a_k(t_i)e^{jk\omega_0 t} = \sum_{k=1}^{K} A_k(t_i)\cos(k\omega_0 t + \phi_k(t_i)), \;\; t_i - N \leq t \leq t_i + N \tag{3.57}$$

where $2N+1$ represents the length of the analysis frame in samples, $K = \dfrac{F_M(t_i)}{\omega_0(t_i)}$ is the number of harmonics included in the harmonic part, and $F_M(t_i)$ denotes the maximum voiced frequency.

The estimation of the parameters is performed using weighted least squares, that is

$$\epsilon = \sum_{t=-N}^{N} w(t)(x(t) - h(t))^2 \tag{3.58}$$

where $x(t)$ is the original signal. This is a different approach than the previously discussed Sinusoidal Model (SM), which performs peak peaking over the speech spectrum. Since the parameter estimation is entirely done in the time domain, shorter windows can be used. It is reminded that in SM (and other approaches that use FFT methods), a typical analysis window has a length of three to four pitch periods, while in HNM two pitch periods are used. This is an important property of HNM, since it is convenient for modelling segments where speech exhibits high pitch or amplitude non-stationarity.

For the noise part, $n(t)$, the estimation of the parameters is as follows. In each analysis frame, the power density function of the original signal is modelled by a p^{th}-order all-pole filter ($p = 15$ for a 16 kHz signal), and the variance of the signal is calculated. Then, a parametric envelope is estimated in each frame. A triangular type time-domain envelope has proved to provide satisfactory results. However, in [PS08], it was shown that an energy based time domain envelope outperforms the triangle type approach.

The synthesis is performed in a pitch-synchronous way. In a plain synthesis (without modifications) scheme, the analysis time instants, t_i^a, coincide with the synthesis time instants, t_i^s. For the harmonic part, the amplitudes and phases are estimated via LS, as previously mentioned, and are linearly interpolated between successive frames. Please note that the phases are unwrapped before applying interpolation. This is done by a predicting the phase of the current frame, using the phase of the previous one and the average instantaneous frequency. The noise part is synthesized using an Overlap-Add (OLA) procedure, in order to avoid discontinuities at the frame boundaries. Given a synthesis time instant, t_s^i, two pitch periods are synthesized by filtering a unit variance, white Gaussian noise through a normalized lattice filter, and multiplying the output by the variance estimated at the analysis time instant, t_i^a. If the frame is voiced, then the lower part is synthesized using harmonics, up to the estimated maximum voiced frequency, $F_M(t_a^i)$. Thus, the noise part is filtered by a high-pass filter with a cut-off frequency $f_c = F_M(t_a^i)$. Then, the synthetic noise part is obtained by applying OLA on two noise parts, one synthesized at synthesis time instant t_s^i, and the other synthesized at t_s^{i-1}. Finally, for voiced frames, the triangular time domain envelope is applied directly on the synthetic noise part. The final synthetic speech signal is obtained by adding the two parts,

$$s(t) = h(t) + n(t) \tag{3.59}$$

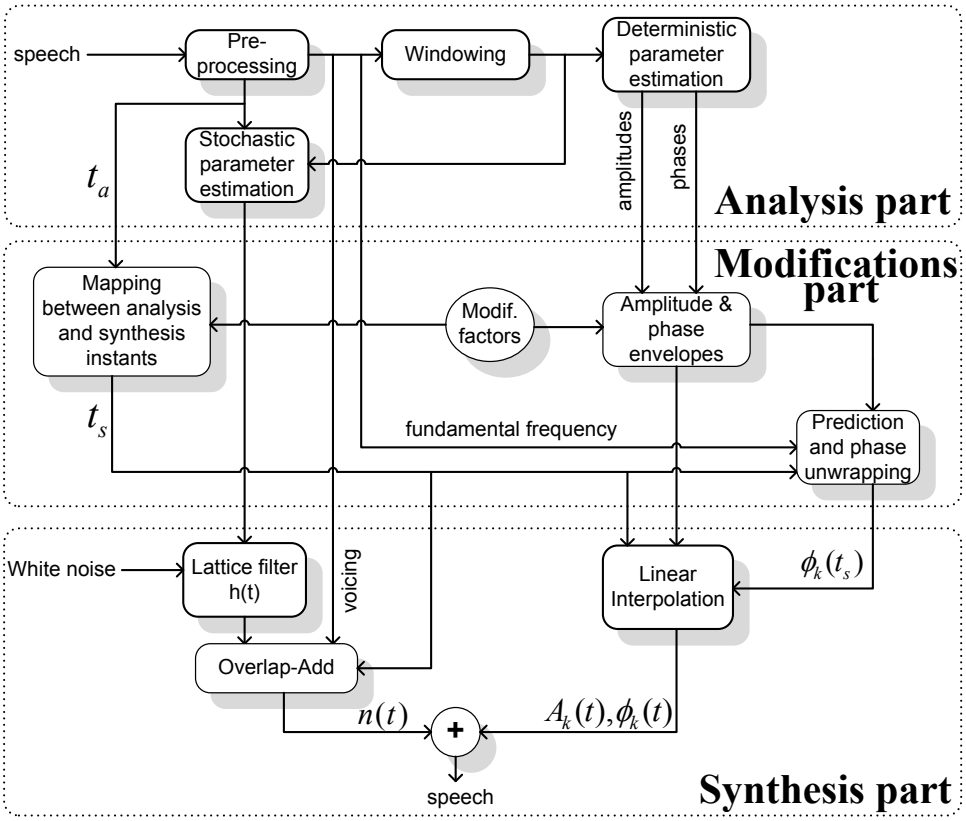


Figure 3.8: *Harmonic + Noise Analysis, Modification, and Synthesis.*

It is worth mentioning that since a harmonic assumption is held in this model, there is no need for frame-to-frame frequency matching, as in the Sinusoidal Model.

Pitch and Time Scale Modifications

For the sake of convenience, the case of joint time-scale and pitch-scale modifications is discussed here. As a first step, the new time synthesis instants should be found, according to the analysis time instants and the desired pitch and time scale modifications. HNM follows a pitch-synchronous scheme, inspired by PSOLA, which was briefly discussed in the previous section. A mapping between the synthesis and analysis instants is determined, specifying which analysis instant should be selected for a given synthesis one. This is performed according to the following constraints: for pitch scaling, the time evolution must be preserved, and for time scaling, the pitch contour must be preserved.

In case of pitch scaling, one should compute the amplitude and phases for the modified harmonics. For this, a spectral and phase envelope estimation is necessary. In HNM context, a regularized cepstrum technique is used [LM95], where discrete cepstral coefficients are calculated, with a frequency domain LS criterion, combined with a regularization method to increase robustness of the estimation. The amplitudes are then obtained by sampling the estimated envelope at the new harmonic frequencies. For the phase envelope, another approach is followed. Consider a voiced frame of a voiced portion of speech. The phase is unwrapped in the frequency domain by adding integer multiples of 2π, in order to keep the frequency slope variation, $d\phi_k = \phi_{k+1} - \phi_k$, as smooth as possible, where k is the k^{th} harmonic. In the next voiced frame, the phases are unwrapped by using the frequency slopes from the previous frame and not the frequency slopes of the current one. This way, phase continuation is guaranteed both in time and in frequency domain. Finally, the new phases are obtained by sampling the phase envelope at the modified pitch harmonics.

Having the new amplitudes and phases, the synthesis of modified speech is performed in the same manner as in synthesis without modifications. It is observed that modified speech is free of artefacts like “buzziness” or “metallic” notion, like in SM or other approaches. An overall flow diagram for HNM is depicted in Figure 3.8

3.2.3 The LF+ARX model

A common approach in speech processing is to represent the speech production system as a source-filter model. In such representations, the source is referred to as the glottal flow derivative (GFD) (as the convolution of the glottal flow and the lip radiation). A well-known model for GFD representation is the Liljencrants-Fant model (LF model) [FLL85], which characterizes the GFD using a number of parameters - usually five: one for the location of the glottal source, one for the amplitude, and three for the shape of the glottal flow. Figure 3.9 shows a typical LF waveform. Usually, the

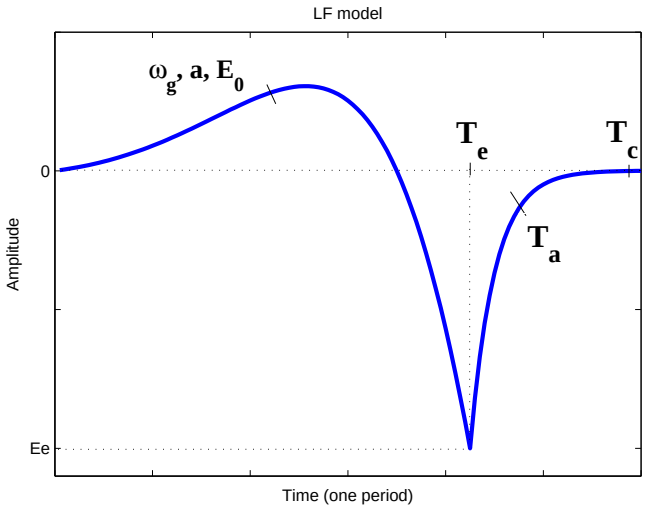


Figure 3.9: *The LF-model.*

parameters that define the shape of the waveform include the Open Quotient ($O_q = \frac{T_c}{T_0}$), the Asymmetry Coefficient ($a_m = \frac{T_z}{T_e}$), and the Return Phase Quotient ($Q_a = \frac{T_a}{(1-O_q)T_0}$). So, a period of the LF model can be mathematically described as

$$u_{LF}(t) = \begin{cases} E_1 e^{at}\sin(\omega t) & 0 \le t \le T_e \\ -E_2[e^{-b(t-T_e)} - e^{-b(T_0-T_e)}] & T_e \le t \le T_0 \end{cases} \tag{3.60}$$

where the parameters a, b, ω are related to the previously mentioned coefficients that define the shape of the GFD.

Hence, the speech signal can be represented by means of an Auto-regressive model with an exogenous input (ARX model)[DKA95]:

$$s[n] = -\sum_{k=1}^{p} a_k[n]s[n-k] + b_0 u_{LF}[n] + e[n] \tag{3.61}$$

where $a_k[n]$ are the time-varying coefficients of the order p AR model that describes the vocal tract, and $e[n]$ is the residual signal that carries information that remains uncaptured by the ARX-LF model.

Analysis-Synthesis

The analysis procedure consists of two steps: One, the estimation of the ARX-LF model parameters, and two, the decomposition of the residual signal. The residual signal has been proposed to decompose in three different ways [AR09]: a HNM-based decomposition, a modulated noise model, and a harmonic model.

At first, the parameters of the ARX-LF model are estimated in the following way: initially, an estimate for the f_0 is provided. Second, an estimation of the glottal closure instant (GCI) is performed [VRC07]. Third, a Viterbi algorithm is applied to regularize the LF source. Finally, the AR parameters are estimated. After this procedure, the residual can be modelled using any of the aforementioned ways. This analysis is for voiced parts – for unvoiced parts, the above analysis scheme is replaced by a simple Warped Linear Predictive (WLP) analysis.

The synthesis is done pitch-synchronously by passing the reconstructed glottal source through the time-varying AR filter. The reconstructed glottal source is given as

$$u[n] = w[n]u^l[n] + (1 - w[n])u^{l+1}u[n] \tag{3.62}$$

where $w[n]$ is a Hanning window whose length is twice the local period, and $u^l[n]$ and $u^{l+1}[n]$ denote the short term glottal signals obtained from the l^{th} and the $(l+1)^{th}$ analysis instants, respectively.

Pitch and Time Scale Modifications

When modifications are applied, then two intermediate steps are involved: i) the determination of the sequence of the analysis frame indices given a stream of f_0 and time modification coefficients, and ii) the modification and synthesis of the selected speech frames. The first step is very similar to the PSOLA analysis scheme. For time scaling, a typical duplication or elimination of frames is performed, so the focus will be on pitch scaling. So, the LF waveform and the residual signal need to be modified, since AR parameters that describe the vocal tract remain untouched.

The modified LF waveform is related to the original one by

$$\hat{u}_{LF}\left(\frac{t}{\hat{T}_0}\right) = u_{LF}\left(\frac{t}{T_0}\right) \tag{3.63}$$

which states that the spectrum of the modified glottal source is a stretched version of the original one with a duration of $\hat{T}_0$.

Another way to modify the pitch is to change the shape parameters of the LF model:

$$\hat{T}_e = T_e \Leftrightarrow \hat{O}_q = O_q \frac{T_0}{\hat{T}_0} \tag{3.64}$$

$$\hat{T}_a = T_a \Leftrightarrow \hat{Q}_a = Q_a \frac{1 - O_q}{1 - \hat{O}_q} \frac{T_0}{\hat{T}_0} \tag{3.65}$$

where $\hat{O}_q, \hat{Q}_a$ represent the modified open and return phase quotients, respectively. It should be noted that the shape of the spectral envelope of the LF waveform is preserved in this way.

3.2.4 Other Approaches

Parametric modelling has many other representatives and in this paragraph, some of them will be briefly discussed. In parallel to the work of McAulay and Quatieri on the Sinusoidal Model, Serra [Ser89] suggested a hybrid system for the analysis, transformation, and synthesis of sound based on a deterministic plus stochastic decomposition. This system is designed to obtain musically useful intermediate representations for sound transformations. The method is similar to the HNM described earlier in terms of decomposition, but the deterministic component is represented by a series of not-necessarily-harmonic sinusoids calculated by a STFT-based peak-picking method, as in SM. The stochastic component is represented by a series of magnitude-spectrum envelopes that work as a time-varying filter excited by white noise, similarly to HNM. This approach is able to create new sounds out of the representation of a particular sound. The deterministic signal is obtained by synthesizing a sinusoid from each trajectory. Then, the residual between the deterministic component and the original sound is modelled by a series of envelopes. Finally, the stochastic signal is generated by an inverse STFT. This system is very flexible and allows for transformations of the sound by manipulating each component separately.

George and Smith [GS97] proposed a novel speech analysis/synthesis system based on the combination of an overlap-add (OLA) sinusoidal model with an analysis-by-synthesis technique to determine the model parameters. An equivalent frequency-domain algorithm that takes advantage of the computational efficiency of the FFT is introduced and a refined OLA sinusoidal model is presented, which can offer shape-invariant speech modifications. The quality of the system is very high when the underlying frequencies are accurately estimated. Shape invariance and phase coherence are explicitly controlled in the modification process.

The Exponentially Damped Sinusoidal Model (EDSM) has also been proposed [NHD98a, JJH99, JHJ04] for modeling transient parts of speech or audio, along with more powerful parameter estimation schemes based on either Matching Pursuit or Subspace Methods. Subspace methods have good spectral properties and do not suffer from the time-frequency trade-off embedded in other methods. However, they are computationally intensive.

Furthermore, Degottex [Deg10] supported that a model which is more dedicated to voice production better respects physiological or acoustic constraints. Towards this direction, the Separation of the Vocal tract with the Liljencrants - Fant (LF) model plus Noise (SVLN) is suggested. The Transformed LF (TLF) glottal model is used to represent the deterministic component of the source. Instead of using the standard, tri-parametric version of LF [FLG85], a single parameter, named R_d, is used to control the shape of the source. The reason for this is that, according to Fant [Fan95], the R_d is the most effective parameter to describe voice qualities into a single value. Zero-mean Gaussian noise is used for representing the random component of the source. Amplitude modulation is applied on this noise during synthesis, to improve naturalness. Due to the different spectral properties of the deterministic and random components, the estimation of the Vocal Tract Filter (VTF) is adapted by taking into account this mixed source model. Transformations such as breathiness, time-scaling, and pitch transposition are available on this model, with very good results compared to the

state of the art. However, as in almost all glottal-based models, there is a stability problem when estimating glottal parameters in frames that are in-between voiced and unvoiced, i.e. transient frames.

Finally, refinements of the proposed models using more powerful parameter estimation schemes have been presented. The Fan-Chirp Transform (FChT) [KW06, WK07] is a recently introduced transform that employs an adaptive analysis basis composed of quadratic chirps. A sinusoidal analysis of speech similar to [MQ86] but with the FChT instead of the FFT has been conducted [DQ07, DQM09] with very satisfactory results. However, speech modifications are yet to be developed.

3.3 Conclusions and Discussion

The methods discussed so far yield high quality for speech resynthesis and moderate modifications. Parametric methods, such as SM or HNM, work well for well-estimated frequencies and under the assumption that speech is short-term stationary. That is, sinusoids that represent voiced speech have constant amplitudes and constant frequencies for a short time analysis window, typically $20 - 30$ ms. It is already shown in [PRS11] that this is not the case in speech, where there are rapid, non linear amplitude and frequency changes during short time intervals. It is essential for high-quality speech analysis, synthesis, and modifications, to be able to capture these short-time fluctuations. Furthermore, parametric models usually represent speech in a two-fold process: they estimate the model parameters on a "deterministic part" of speech, and the "randomness" is then modelled differently. Separation of components has been proved practical and convenient for processing, synthesis, and manipulation of speech, under the assumption that the components are well-separated and accurately estimated. Finally, voice production-based models, such as ARX-LF and SVLN, although providing high-quality output, they are complicated and very sensitive in parameter estimation, especially in unvoiced or transient parts. Hence, it would be desirable to have a parametric speech model that is relatively simple, flexible, high-quality, and robust in resynthesis and modifications.

Towards this direction, hybrid systems of speech analysis based on eaQHM (extended adaptive Quasi-Harmonic + Noise Model - eaQHNM) and aHM (adaptive Harmonic + Noise Model - aHNM) will be presented in this thesis, with very satisfactory results in terms of perceptual quality. However, it will be shown that certain assumptions in component separation of hybrid systems are not necessary, and *all speech sounds* can be very accurately represented as AM-FM components only. Thus, speech can be uniformly represented very accurately as a sum of AM-FM sinusoids, providing compactness, uniformity, and simplicity of speech representation. Modifications will be applied on this representation of speech as well, and their performance will be discussed and compared with the corresponding hybrid systems, as well as with other state-of-the-art systems.

Regarding transformations, the models described in this chapter have their limitations. Modifications based on non parametric frequency-domain methods introduce an undesirable reverberation effect, known as "chorusing", as well as other effects, such as transient smearing (loss of percussiveness) and phasiness (coloration of signal). Also, a number of them are computationally intensive. On the other hand, non parametric time-domain methods, while free of reverberation or chorusing artefacts and computationally efficient, they rely heavily on the quasi-periodic assumption of speech. Furthermore, parametric modelling is highly dependent on the performance of the analysis and synthesis algorithm. Hence, it is believed that since adaptivity provides a high quality analysis/synthesis scheme, prosodic modifications would be of superior quality compared to standard sinusoidal modeling techniques. However, certain aspects in speech modifications should be taken care of. An important aspect is the spectral and phase envelope estimation during pitch scaling. An accurate estimation of these envelopes is necessary in order to evaluate the magnitude and phase values of the shifted frequencies. While there is a variety of approaches for the magnitude envelope estimation, the phase envelope estimation is a difficult problem. In this thesis, phase modifications for time and pitch scaling is handled via very simple mathematical properties. Moreover, shape invariance is a property of analysis/synthesis systems that plays an important role. Shape invariance refers to the ability of a system to preserve the temporal structure of the speech waveform. The inability to maintain the shape of the waveform is caused by the so-called *phase dispersion* problem, which is due to that the reconstructed signal has the same frequency information as the original signal but the relationship of the phases between the different sinusoids has changed. This effect is audible and can be described as "chorusing". In this work, phase dispersion effects will be minimized using a very simple method that utilizes phase properties, and specifically, the notions of *relative phase* for the harmonic models and the *relative phase delays* for the quasi-harmonic models.

Chapter 4

Speech Analysis and Synthesis based on Adaptive Sinusoidal Models

In a Chapter 2, the members of Adaptive Sinusoidal Models family have been analytically described. However, their application on running speech is not straightforward. In literature, there are two different approaches in speech analysis systems. The first one is based on hybrid systems, that is, analysis and synthesis systems that decompose speech into more then one components, usually a deterministic and a stochastic one. The second one is based on full-band Systems, that is, analysis and synthesis systems that treat all parts of speech the same way, as a sum of AM-FM components. In this chapter, speech analysis and synthesis systems based on the aSMs will be presented, and among them, the newly suggested eaQHM will be discussed and compared to the other aSMs and the state-of-the-art. Discussion and motivation on both hybrid and full-band approaches will be presented in this chapter as well.

4.1 Hybrid Systems

Hybrid systems are considered well suited for resynthesis and prosodic modifications, since a well-mastered separation of speech into a deterministic and a stochastic component leads to a better manipulation of them and that aids to an enhanced quality of speech synthesis and modifications. A typical flowchart of a hybrid system is shown in Figure 4.1. Let us briefly discuss the elements of a general hybrid system. First of all, in the analysis part, the *pre-processing* stage often includes actions such as pitch estimation, voiced/unvoiced decision, maximum voiced frequency estimation, filtering, enhancement, or noise cancellation. The *deterministic analysis* part is responsible of modelling the deterministic characteristics of speech, whereas the *stochastic analysis* models the random component of speech, such as friction noise, unvoiced speech, etc. Except for the usual deterministic and stochastic component, recent speech models often include a *transient* part [AR09, Lev99], which captures (but not necessarily models) the transient parts of speech (vowel-to-consonant frames, and vice versa, as well as stop sounds) and is handled differently than the other two components. However, the identification of a transient frame is not an easy task, and the most convenient choice is to account them in either the deterministic or the stochastic part.

When the analysis parameters for all speech components are estimated, they are passed to the synthesis step, where a pre-processing of the parameters is performed, as for example parameter interpolation or spectral envelope estimation, in case of speech modifications. Finally, each component is synthesized separately and all components are summed up to form the synthesized speech signal.

Typical examples of such systems include the Harmonic + Noise Model (HNM) [Sty96], the STRAIGHT method [Kaw97], and the LF+ARX model [AR09]. In the following sections, a hybrid approach will be described based on a two-component paradigm: a deterministic and a stochastic component. The choice of such an approach is justified by its successful application in earlier models, and the convenience in manipulation for modification purposes. After that, drawbacks and misconceptions on hybrid sinusoidal-based systems will be discussed, and simple, full-band schemes of speech analysis, synthesis, and modifications will be proposed, where the term *full-band* refers to a uniform AM-FM decomposition of *all parts* of speech.

4.2 Pre-processing in hybrid systems

As discussed, hybrid systems consist of two components: a deterministic and a stochastic one. Thus, a preprocessing step is necessary to help the separation of components and the estimation of some crucial parameters. In most hybrid

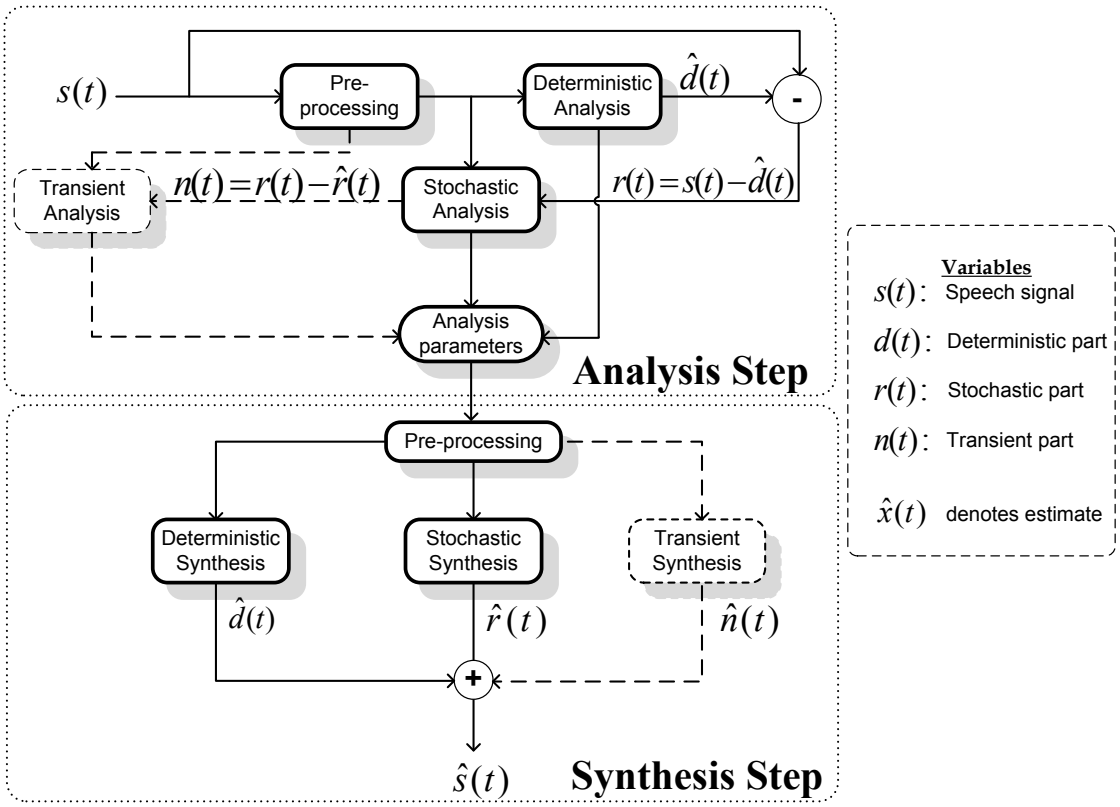


Figure 4.1: *A flowchart for the analysis and synthesis part of a general hybrid system framework. Upper panel: Analysis part. Lower part: Synthesis part.*

sinusoidal models in the literature, this preprocessing step allows the separation of speech in voiced/unvoiced/silence parts, in parallel to a f_0 estimation for voiced parts. We will now briefly describe these processes which apply to both hybrid systems presented in this chapter.

4.2.1 Voiced/Unvoiced/Silence Discrimination

Since the system in hand is a hybrid system, a voiced/unvoiced/silence (V/UV/S) discrimination algorithm needs to be run as a pre-processing step. Although any V/UV/S algorithm can be chosen, in our experiments the approach described in [Pan10] is selected. The V/UV detection is performed in a frame-by-frame procedure, with frame size of 30 ms and step size of 5 ms. The energy of each frame is computed and if it is above a threshold B_E, then it is assigned as speech. Otherwise, a silence flag is assigned on this frame. For the separation of voiced and unvoiced speech, once a frame has been flagged as speech, two conditions should hold to declare it as voiced:

1. the energy of the speech frame minus the energy of a low-pass-filtered version of it should be below B_d, and
2. the energy of the low-pass-filtered frame should be above B_s

As a final step, a median filter of order 5 is applied on the V/UV/S estimation in order to eliminate outliers.

4.2.2 Fundamental frequency estimation

For the f_0 estimation, any robust technique can be used, although the estimation is not critical for our systems due to the adaptation mechanisms. In the systems in hand, the recently proposed SWIPE pitch estimator is used. SWIPE estimates the pitch as the fundamental frequency of a sawtooth waveform whose spectrum best matches the spectrum of the input signal. The spectra are compared by computing a normalized inner product between the signal spectrum and a modified cosine. For details, please see [CH08].

4.3 The eaQHNM analysis and synthesis system

The speech signal is decomposed into two parts, i.e.

$$s(t) = d(t) + n(t) \tag{4.1}$$

where $d(t)$ denotes the deterministic part and $n(t)$ denotes the stochastic part. The separation of components is performed using the schemes presented in [Sty96], i.e. an initial analysis is performed in order to discriminate speech into voiced and unvoiced regions, and then the pitch for the voiced regions is determined, as explained earlier. The following subsections present the deterministic and the stochastic part decompositions in detail.

4.3.1 Analysis of the Deterministic Part

Assuming a Hamming analysis window $w(t)$ with support in $[-T_l, T_l]$, a frame of the deterministic part is initially modeled using QHM as:

$$d(t) = \Big(\sum_{k=-L}^{L} (a_k + tb_k) e^{j2\pi \hat{f}_k t} \Big) w(t) \tag{4.2}$$

where a_k, b_k are the complex amplitudes and the complex slopes of the k^{th} component respectively, $\hat{f}_k = k\hat{f}_0$ are the analysis frequencies, $\hat{f}_0$ is an initial estimation of the fundamental frequency, and L is the number of quasi-harmonics, as specified by the maximum voiced frequency. The estimation of the model parameters is obtained via Least Squares, as described in [Sty96]. As it was mentioned in the previous section, QHM is able to correct, in the least square sense, frequency mismatches that are due to inaccurate estimation of the fundamental frequency. Let η_k denote the frequency mismatch of the k^{th} component:

$$\eta_k = f_k - \hat{f}_k \tag{4.3}$$

where f_k is the actual frequency and $\hat{f}_k$ is an estimation of f_k. It was shown in [PRS08] that a projection of b_k onto a_k results in an estimate of the frequency mismatch, which is:

$$\hat{\eta}_k = \frac{1}{2\pi} \frac{\Re\{a_k\}\Im\{b_k\} - \Im\{a_k\}\Re\{b_k\}}{|a_k|^2} \tag{4.4}$$

where $\Re\{a_k\}, \Re\{b_k\}$ and $\Im\{a_k\}, \Im\{b_k\}$ are the real and imaginary parts of the complex amplitudes and complex slopes, respectively. Using Eq. (4.4), the analysis frequencies can be updated and the signal can be modelled again using QHM, but with a new set of analysis frequencies, $\tilde{f}_k = \hat{f}_k + \hat{\eta}_k$, and thus resulting in a more accurate signal representation.

However, only accurate frequency estimation is addressed via QHM. The stationary model principle is still valid within an analysis frame. In order to confront this issue, the projection of the signal onto a set of time-varying basis functions is suggested in [PRS11] and [KPRS12]:

$$d(t) = \left(\sum_{k=-L}^{L} \Big(a_k + tb_k\Big) \Big(\hat{\alpha}_k(t) e^{j\hat{\phi}_k(t)}\Big) \right) w(t) \tag{4.5}$$

with

$$\hat{\alpha}_k(t) = \frac{\hat{A}_k(t + t_i)}{\hat{A}(t_i)} \tag{4.6}$$

and

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \int_{t_i}^{t} \big(2\pi \hat{f}_k(\tau) + c(\tau)\big) d\tau, \;\; t \in [-T, T] \tag{4.7}$$

where $\hat{A}_k(t)$, $\hat{f}_k(t)$, $\hat{\phi}_k(t)$ are estimates of the instantaneous amplitudes, frequencies, and phases of the k^{th} component, respectively, $c(\tau)$ is the phase coherence term as explained in Section 2.3, Eq. (2.42), and t_i is the center of the analysis window.

The adaptation is completed by using the frequency correction mechanism first introduced in [PRS08], and states that an estimate of the mismatch between the actual k^{th}-frequency and the estimated one, termed $\eta_k = f_k - \hat{f}_k$, is given by

$$\hat{\eta}_k = \frac{1}{2\pi} \frac{\Re\{a_k\}\Im\{b_k\} - \Im\{a_k\}\Re\{b_k\}}{|a_k|^2} \tag{4.8}$$

Hence, at the first adaptation, for the analysis time instant t_i, the instantaneous frequencies are $\hat{f}_k(t_i) = k\hat{f}_0(t_i) + \hat{\eta}_k(t_i)$

and the instantaneous phases become

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \int_{t_i}^{t} (2\pi \hat{f}_k(\tau) + c(\tau))d\tau \tag{4.9}$$

Then, a Least Squares solution for the a_k, b_k using these refined frequencies (and phases) leads to a better estimate of the instantaneous amplitudes $\hat{A}_k(t) = |a_k(t)|$ and the $\hat{\eta}_k$ terms. By iteratively adding the $\hat{\eta}_k$ term of the current adaptation on the k^{th}-frequency track of the previous adaptation, the frequency tracks represent the underlying actual frequencies better.

This adaptation mechanism stops according to a reconstruction criterion related to the Signal-to-Reconstruction-Error Ratio (SRER), redefined here for convenience:

$$\frac{SRER^{i-1} - SRER^{i}}{SRER^{i-1}} < \epsilon \tag{4.10}$$

where $SRER^i$ is the Signal-to-Reconstruction-Error Ratio of the resynthesized signal in the i^{th} adaptation, defined as

$$SRER = 20 \log_{10} \frac{\sigma_{x(t)}}{\sigma_{x(t)-\hat{x}(t)}} \tag{4.11}$$

where σ_x denotes the standard deviation of $x(t)$, $x(t)$ is the actual signal and $\hat{x}(t)$ is the reconstructed signal, and ϵ is a threshold for convergence, typically set to 0.02.

Finally, it is essential to describe the estimation of the time-varying parameters. The k^{th} instantaneous amplitude track, $\hat{A}_k(t)$, is computed via linear interpolation of the successive estimates. Spline interpolation could be an alternative but it does not guarantee positiveness of the tracks. For that reason, linear interpolation is preferred. The k^{th} instantaneous frequency track, $f_k(t)$, is computed via spline interpolation, because splines offer smooth transitions between frequency estimates. Also, it is worth noting that frequency matching is trivial since the analysis frequencies are integer multiples of a fundamental. As for the k^{th} instantaneous phase track, $\hat{\phi}_k(t)$, similar interpolation schemes are not suitable; thus, a non parametric approach is followed based on the integration of instantaneous frequency. In addition, phase coherence over frame boundaries is addressed via the addition of an extra term in order to guarantee phase continuation over frame boundaries as in Eq. (4.9). Finally, the deterministic part can be approximated by its time-varying components using:

$$\hat{d}(t) = \sum_{k=-L}^{L} \hat{A}_k(t) e^{j\hat{\phi}_k(t)}. \tag{4.12}$$

4.3.2 Analysis of the Stochastic Part

The stochastic part, $n(t)$, defined as the residual between the analyzed signal and the deterministic part, does not only account for the unvoiced parts of speech but also for friction noise in voiced parts. It is modelled as:

$$n(t) = e(t)(u_G(t) * q(t)) \tag{4.13}$$

where $u_G(t)$ denotes a Gaussian white noise component that is convolved by a time-varying auto-regressive (AR) filter with impulse response $q(t)$, and $e(t)$ denotes the time-domain envelope. Standard LPC analysis is used for the estimation of the AR filter, with an order of $p = 18$, for a sampling frequency of $F_s = 16$ kHz. The time-domain envelope is a very important factor, since it is essential for efficient fusion between the deterministic and the stochastic part [Sty96]. In [PS08], it was shown that an energy-based time envelope is a good choice. The energy envelope is given by:

$$e(t) = \sum_{u=-T}^{T} |n(t+u)| \tag{4.14}$$

where T equals to 1 ms. Since the energy envelope has a pitch-synchronous behaviour [PTRS10], it can be approximated within a frame using a sum of few sinusoids:

$$\hat{e}(t) = \left(\sum_{k=-M}^{M} A_k e^{j2\pi f_k t + \phi_k} \right) w(t) \tag{4.15}$$

where M is a small integer, typically 3 or 4, and f_0 is the fundamental frequency of the frame. Amplitude estimation is performed using peak picking on the short-time spectrum of the energy envelope signal. The latter is obtained using a 30

ms Hamming analysis window and a frame rate of 5 ms.

4.3.3 Synthesis

During synthesis, for the deterministic part, the k^{th} instantaneous amplitude track, $\hat{A}_k(t)$, is computed via either linear or spline interpolation of the successive estimates from the last adaptation step. The k^{th} instantaneous frequency track, $f_k(t)$, is also computed via spline interpolation. As for the k^{th} instantaneous phase track, $\hat{\phi}_k(t)$, the non parametric approach based on the integration of instantaneous frequency is followed, as it is shown in the adaptation steps of the analysis. Then, the deterministic part can be approximated as:

$$\hat{d}(t) = \sum_{k=-L}^{L} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \tag{4.16}$$

The stochastic part $\hat{n}(t)$, is obtained by a simple Overlap-Add (OLA) method, using the parameters from the analysis part. Finally, the synthetic part is given by

$$\hat{s}(t) = \hat{d}(t) + \hat{n}(t) \tag{4.17}$$

4.3.4 Examples

Two examples are shown in Figures 4.2, 4.3 for a male and female speaker, respectively.

4.4 The aHNM analysis and synthesis system

In this section, the analysis and synthesis scheme of aHNM is presented, along with an application in time-scaling of speech. While the original adaptive Harmonic Model has been developed by Degottex and Stylianou in 2013 [DS13], the aHNM has been developed during this thesis as a first step towards speech modifications based on adaptive models. Only the analysis and synthesis part will be described here.

4.4.1 Analysis

In the analysis part, the deterministic and the stochastic part are separated and modelled. In general, the former can be described as

$$s(t) = s_d(t) + s_s(t) \tag{4.18}$$

where $s(t)$ denotes the speech signal, and $s_d(t), s_s(t)$ denote the deterministic and stochastic part, respectively. The deterministic part models the quasi-periodicities of voiced speech as a sum of time-varying harmonic components, thus

$$s_d(t) = \sum_{k=-K}^{K} A_k(t) e^{jk\phi_0(t)} \tag{4.19}$$

where

$$\phi_0(t) = \int_0^t 2\pi f_0(u) du \tag{4.20}$$

K is the number of components, and $A_k(t), k\phi_0(t)$ are the instantaneous amplitudes and the instantaneous phases of the k^{th} component, respectively. Please note that the instantaneous phase of the k^{th} component is an integer multiple of the instantaneous phase of the first harmonic, f_0, and that the analysis in voiced speech is full-band.

Deterministic Part

In the analysis step for the deterministic part, a parametrization of the speech signal at each analysis time instant t_a^i is undertaken. At first, a sequence of the analysis time instants is created in the voiced parts of speech using the provided $f_0(t)$ track, so as to have one analysis time instant per pitch period. Moreover, if the distance between t_a^i and t_a^{i+1} is short enough, aHM can model the amplitude variations of the unvoiced signal (like in plosives). Thus, the upper limit of the size of the analysis window is 20 ms and the lower limit comes from the provided $f_0(t)$ track, and is therefore set to 50 Hz. Around each analysis time instant t_a^i, a Blackman window with a length of 3 local pitch periods is applied to the speech signal. The phase track $\phi_0(t)$ is then computed by means of spline interpolation of f_0^i and using the integration formula in Eq.(4.20).

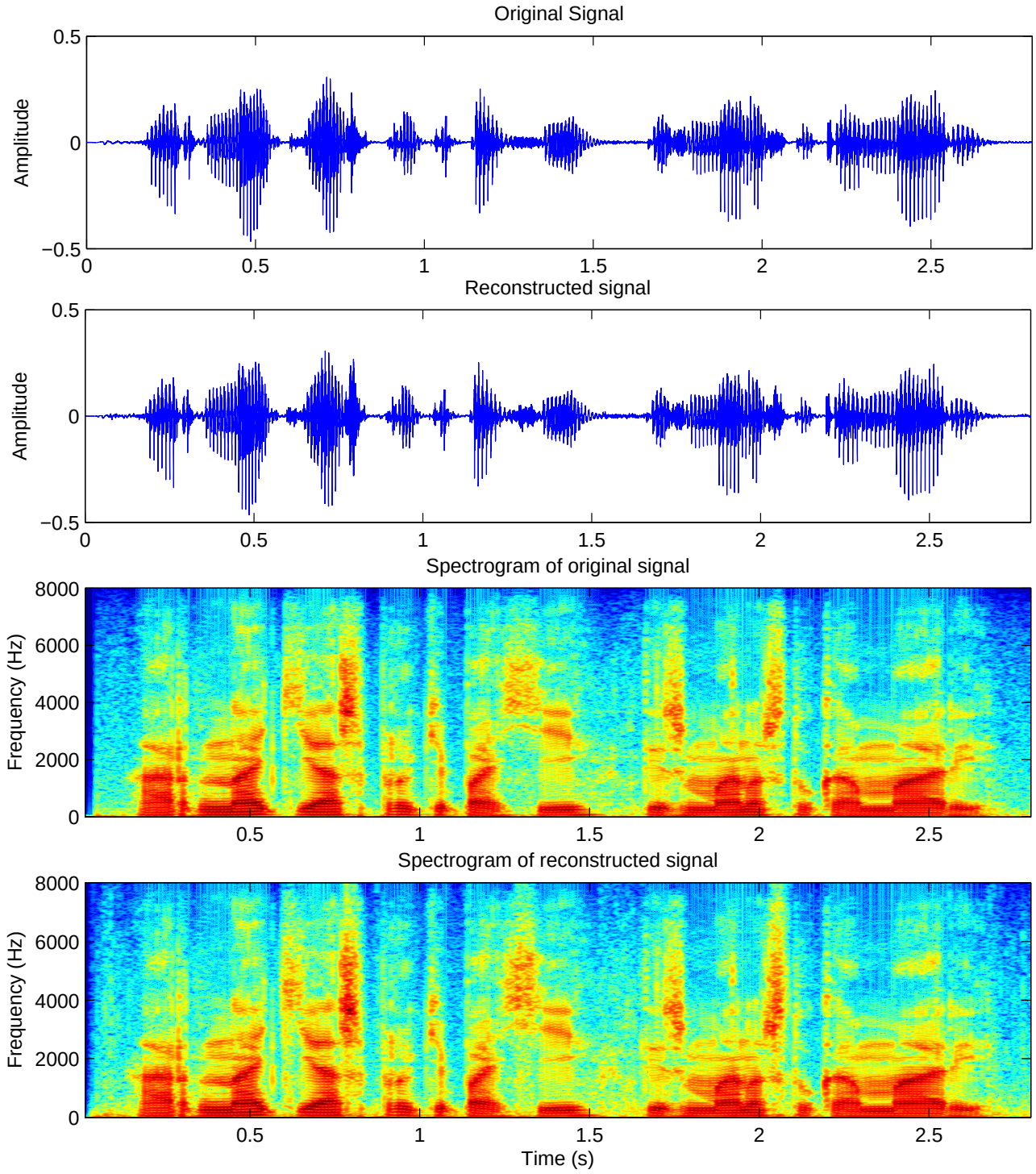


Figure 4.2: *extended adaptive Quasi-Harmonic + Noise Model: Original signal (first panel) and reconstructed signal (second panel) along with their corresponding spectrograms (third and fourth panel) for a male speaker.*

Adaptive Iterative Refinement - AIR

The fundamental frequency track of Eq.(4.20) is assumed to be known beforehand and can have a potential error, i.e.

$$\eta_0 = f_0 - \hat{f}_0 \tag{4.21}$$

that is called *frequency mismatch*, where f_0 is the actual fundamental frequency at a certain time instant and $\hat{f}_0$ is an estimate of the latter. Following the adaptive scheme presented in [PRS11], the amplitude $a_k(t)$ and fundamental

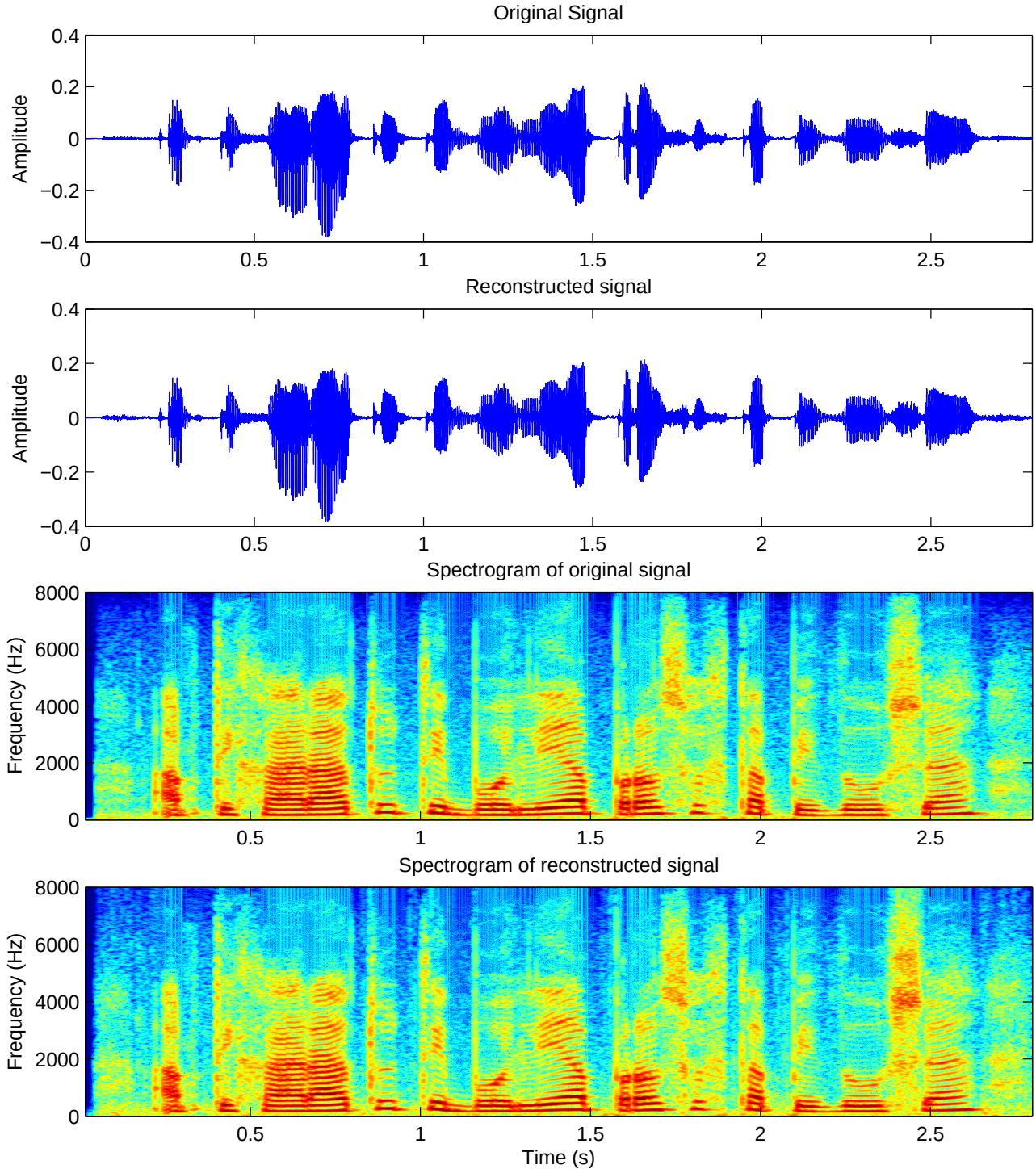


Figure 4.3: *extended adaptive Quasi-Harmonic + Noise Model: Original signal (first panel) and reconstructed signal (second panel) along with their corresponding spectrograms (third and fourth panel) for a female speaker.*

frequency $f_0(t)$ values are obtained by a linear interpolation, respectively, of their values, a_k^i and f_0^i, at the analysis time instants, t_a^i. In order to have an estimate of these values, the *adaptive Quasi-Harmonic Model - aQHM* is used, that is given by the following equation:

$$s(t) = \sum_{k=-K}^{K} (a_k + tb_k)e^{jk\phi_0(t)} \tag{4.22}$$

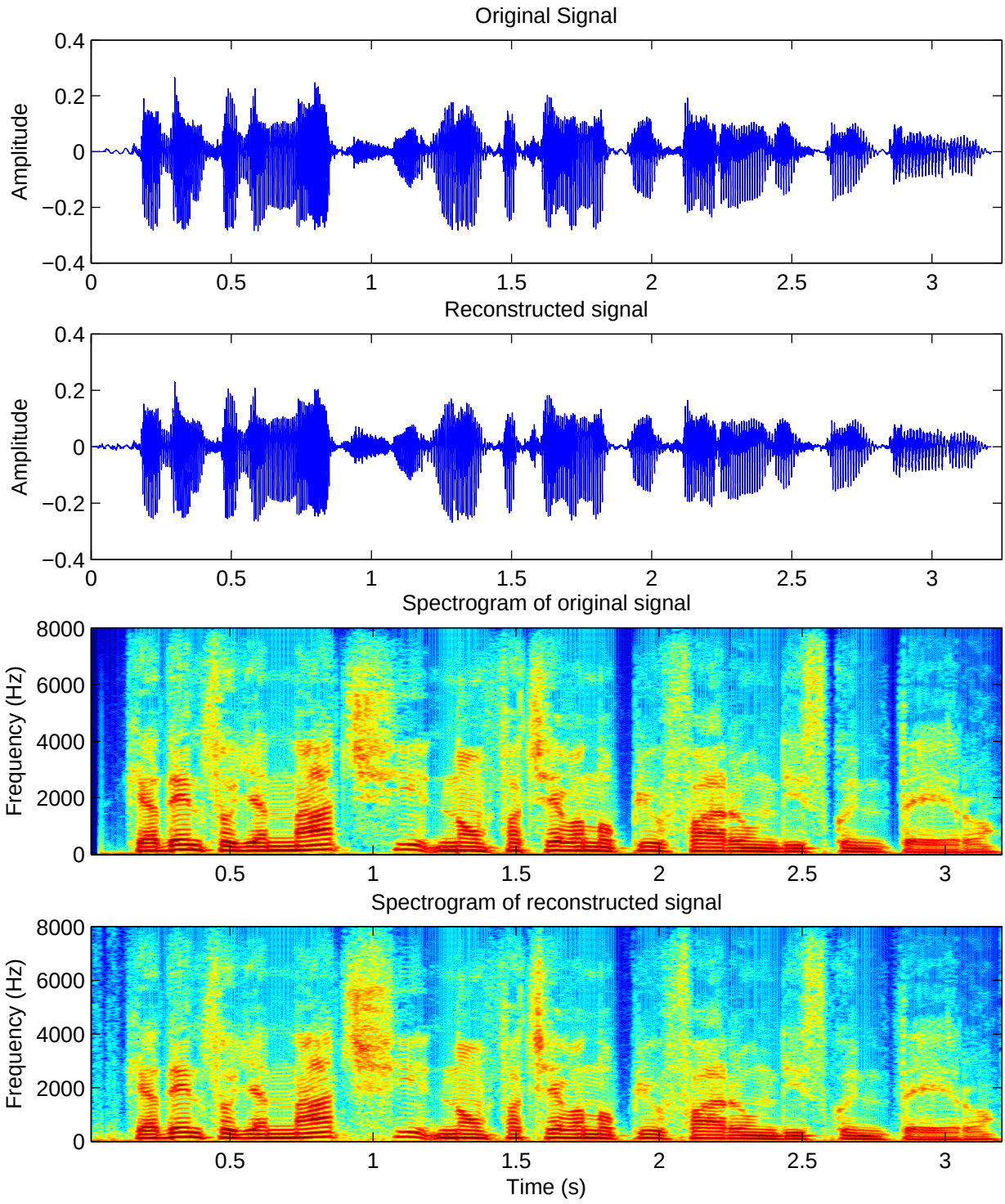


Figure 4.4: *adaptive Harmonic + Noise Model: Original signal (first panel) and reconstructed signal (second panel) along with their corresponding spectrograms (third and fourth panel) for a male speaker.*

where $\phi_0(t)$ is the same as in Eq.(4.20), a_k and b_k are the complex amplitude and the complex slope of the model, respectively, and K is again the number of the components. It has been shown in [PRS08] that a_k and b_k, that are obtained via a Least Squares minimization, can be used to provide an estimate, $\hat{\eta}_0$, for the frequency mismatch of Eq.(4.21). Thus, for the k^{th} component in general, this can be computed as:

$$\hat{\eta}_k = \frac{1}{2\pi} \frac{\Re\{a_k\}\Im\{b_k\} - \Im\{a_k\}\Re\{b_k\}}{|a_k|^2} \tag{4.23}$$

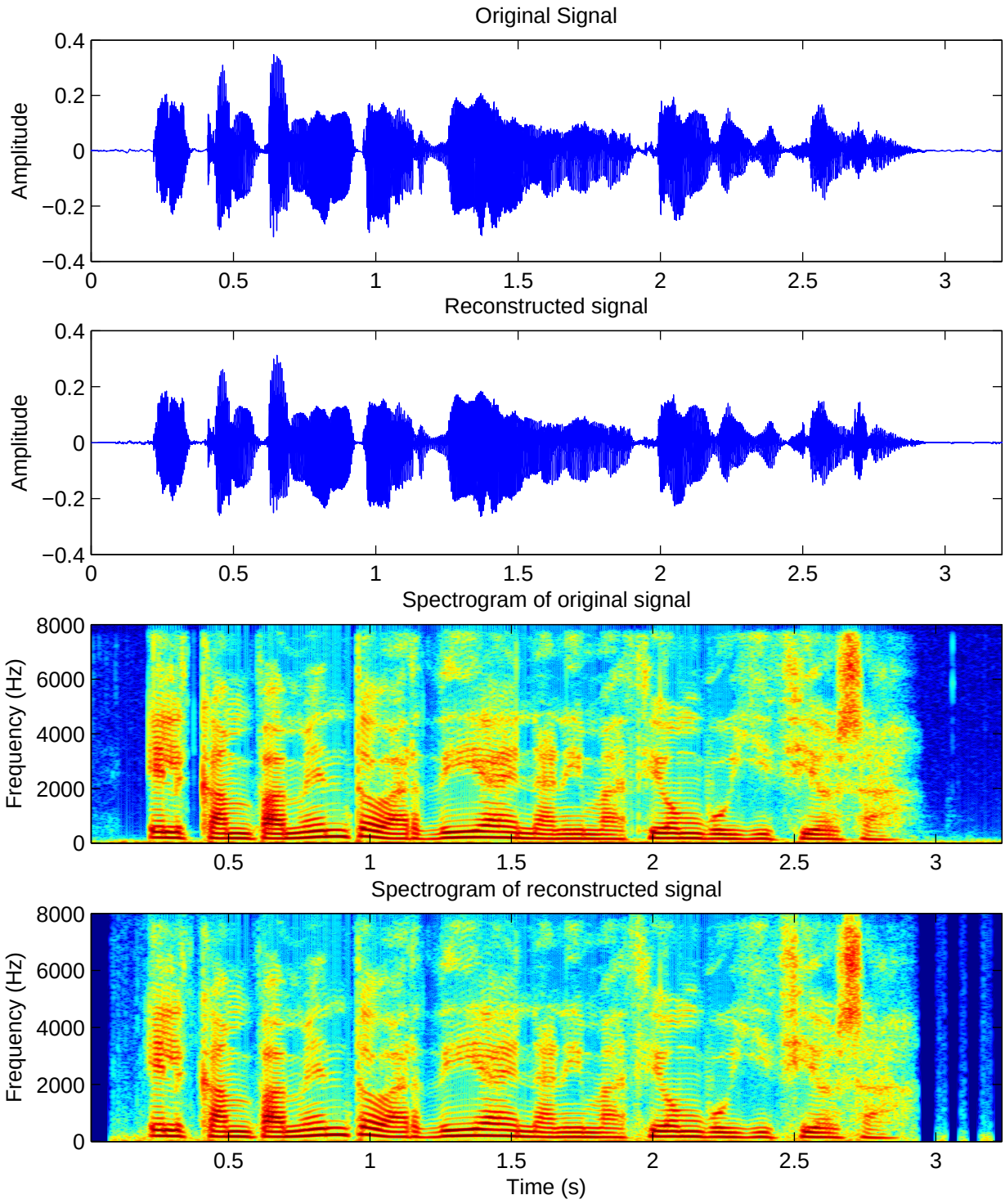


Figure 4.5: *adaptive Harmonic + Noise Model: Original signal (first panel) and reconstructed signal (second panel) along with their corresponding spectrograms (third and fourth panel) for a female speaker.*

where $\Re\{a_k\}, \Re\{b_k\}$ and $\Im\{a_k\}, \Im\{b_k\}$ are the real and imaginary parts, respectively, of the complex amplitude and the complex slope of the model.

Using this estimate, the fundamental frequency values f_0^i can be updated in an iterative manner. However, as it is shown in [PRS11], this term cannot be larger than the main lobe of the analysis window.

In [DS12], an iterative algorithm has been proposed to update the frequencies. Its main idea is discussed here. In a single analysis window, an arbitrary small number of harmonics K (e.g. 4) can be assumed. These harmonics are considered not to vary too much from their actual values, i.e. the mismatch η_k is small. By computing the LS solution for Eq.(4.22),

the correction term, η_0, related to the fundamental frequency f_0 can be then estimated by the following equation:

$$\hat{\eta_0} = \frac{1}{K}\sum_{k=1}^{K}\frac{\hat{\eta}_k}{k} \tag{4.24}$$

This estimation can be furtherly used to update the number of harmonics, K. If $\hat{\eta_0}$ is small, this means that the current set of harmonics have converged very close enough to their actual values. Then, K can be further increased to add new harmonics in a new set of harmonics. If $\hat{\eta_0}$ is large, then the current set of harmonics have not converged to their actual values and further iterations are necessary to successively reduce $\hat{\eta_0}$. The number of harmonics that are added in each iteration are given by the following equation:

$$K = \left\lfloor \frac{\frac{1}{2}N_w}{|\hat{\eta_0}|} \right\rfloor \tag{4.25}$$

where $N_w = min\{B_w, f_0\}$, where B_w is the bandwidth of the main lobe of the analysis window. Using the LS solution of Eq.(4.22), the local parameters a_k^i, b_k^i are computed, along with the k^{th} frequency mismatch, $\hat{\eta}_k$, and the fundamental frequency correction, $\hat{\eta}_0$. The number of harmonics, K^i, is then updated using Eq.(4.25). As a last step, the process is repeated for all frames until the Nyquist frequency is reached for all frames. This approach is termed as the *Adaptive Iterative Refinement - AIR* and a pseudocode can be found in [DS12].
It should be noted that the estimated amplitude and phase values that are obtained at the analysis step correspond to the aQHM model and not aHM which is used for synthesis. Therefore, the aHM model is used in a last iteration step to ensure the consistency between the models used in the analysis and the synthesis.

Stochastic Part

The stochastic part is modeled exactly the same way as in eaQHNM (See Section 4.3.2).

4.4.2 Synthesis

In the synthesis step for the deterministic part, each harmonic is generated in separate, one after the other, without using any window. Each harmonic component is synthesized by its parameters, namely its amplitudes $|a_k^i|$, phases $\angle a_k^i$, and fundamental frequency f_0^i. First, the instantaneous amplitude, $|a_k(t)|$, of the k^{th} harmonic is simply obtained by linearly interpolating the estimated $|a_k^i|$ on the analysis time instants t_a^i, on a logarithmic scale. The instantaneous phase $\angle a_k^i$ cannot be interpolated directly across time to obtain $a_k(t)$ because of its rotation due to the time advance between analysis time instants. Therefore, it is proposed to remove this effect using the integral of $f_0(t)$ from the start of the signal, and obtain the *relative phase - RP*:

$$\angle\tilde{a}_k^i = \angle a_k^i - k\phi_0(t_a^i) \tag{4.26}$$

Thus, assuming that the shape of the signal is changing smoothly, the phase values change also smoothly from one analysis time instant to the other. Then, the RP $\angle\tilde{a}_k^i$ can be interpolated to obtain its continuous counterpart, $\angle\tilde{a}_k(t)$. Additionally, a spline or cubic interpolation is necessary such as its time derivative, the frequency, is still continuous. All along the iterative process, and since the harmonic numbers K^i increase independently from one analysis time instant to the other, there are often missing components in the interpolations of amplitude and instantaneous phase. If this is the case, then the amplitude of the missing component is set to -300 dB and the corresponding phase $\angle\tilde{a}_k(t)$ is set to zero. For the stochastic part, it is resynthesized using the OLA method. For each frame, white noise is passed through the AR filter to obtain the frequency modulation of the stochastic part. Then, the energy envelope is computed from Eq. (4.15) and its multiplication with the frequency-modulated noise provides the reconstructed stochastic frame.

4.4.3 Examples

Two examples are shown in Figures 4.4, 4.5 for a male and female speaker, respectively.

4.5 An alternative for noise modeling

In [HDC02], it has been noted that overlap-add methods for unvoiced synthesis have certain drawbacks. Also, the modulated noise approach can model unvoiced speech well, but this is not the case for plosives. For this, another method is proposed in which unvoiced frames are synthesized using white noise as an input to a lattice implementation and a sample-by-sample interpolation of the reflection coefficients. Standard LPC techniques have been used to estimate the latter, with an AR filter of order 16, a Hanning window size of 20 ms length, and a step size of 10 ms. Then, one can

use the time-envelope of the original unvoiced speech to modulate the output of the filter. Although many algorithms are available to compute the time-envelope of a signal, the one proposed in [AR08] is selected. This algorithm is the time-domain analogue of the True-Envelope estimator and can be described as follows:

Initialize: $s(t) = |n(t)|$

1. lowpass-filter$(s(t), 500 \text{ Hz})$
2. $s(t) =$max$(s(t), |n(t)|)$
3. Goto **1** and **2** for 50 iterations

Finalize: lowpass-filter$(s(t), 500 \text{ Hz})$

An example of application is given in Figure 4.6.

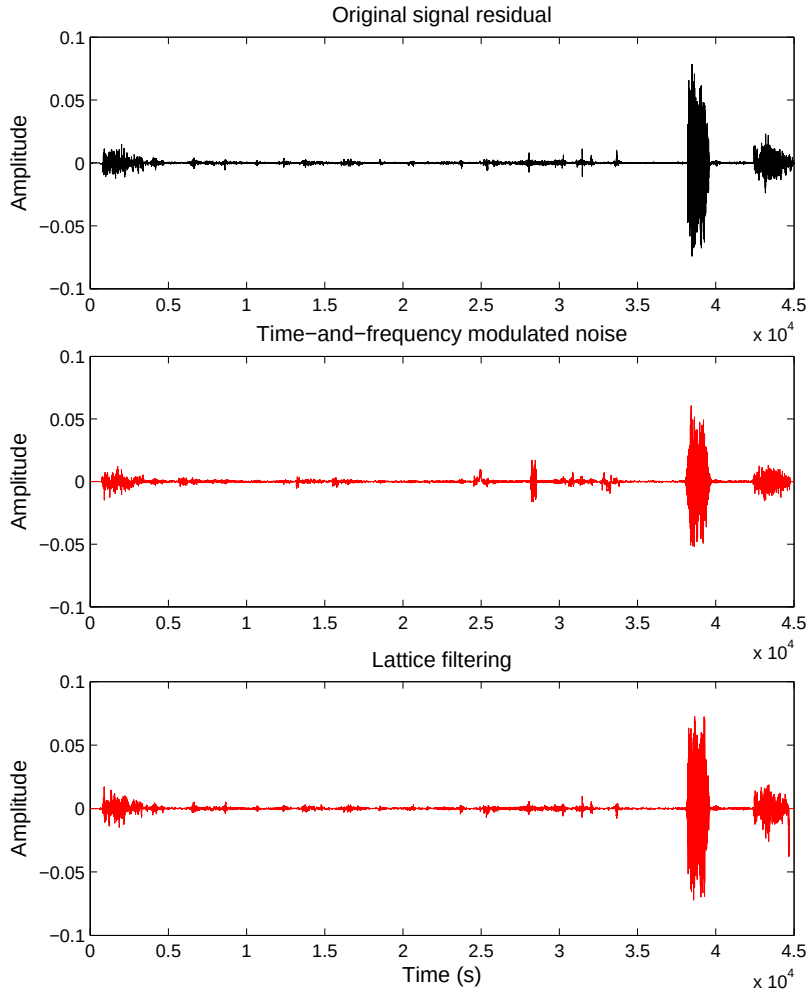


Figure 4.6: *An example of modeling the noise part of a speech signal using (a) time-and-frequency modulated noise and (b) sample by sample lattice filtering*

4.6 Discussion

In Sections 4.1-4.4, we presented a novel and a refined hybrid system for analysis and synthesis of speech based on the extended adaptive Quasi-Harmonic model and the adaptive Harmonic Model. The former decomposes voiced speech in AM-FM components that are quasi-harmonically related whereas the latter is inspired from the theory of adaptivity to accurately estimate an f_0 which is used to model voiced speech. The unvoiced parts of speech are represented by a

stochastic component which is implemented as time and frequency modulated white Gaussian noise for both systems. Illustrative examples are given for each model that depict their performance in time and frequency domain.

Both models rely on a voiced/unvoiced (V/UV) estimator that separate the corresponding parts of speech. The importance of such an estimator is crucial for the performance of the systems. Although a very simple V/UV estimator is used in this work, no significant artefacts are present in the resynthesized speech waveforms. However, a binary decision on voicing in frames is often erroneous, since a frame can be a transient frame, that is, in certain cases it cannot be unequivocally categorized as either voiced or unvoiced. According to the decision of the estimator, the frame will be modeled by either the deterministic or the stochastic model. This may result in problems in the frame boundaries due to inappropriate fusion between different modeling of adjacent frames.

It could be suggested to drop the V/UV estimator, to both reduce the complexity of the overall system and to eliminate possible frame categorization errors that could influence system performance. Such a suggestion leads to *full-band* systems that perform AM-FM decompositions on the *full-length* of the waveform.

4.7 Full-band Systems

Although hybrid models have been proved to provide flexibility in manipulation, synthesis, and modifications of speech, in this section full band analysis and synthesis systems of speech are presented, using the adaptive sinusoidal models on the *full length* of a speech waveform. This means that the model describes both voiced *and* unvoiced parts of speech. A generalized full-band system is depicted in Figure 4.7.

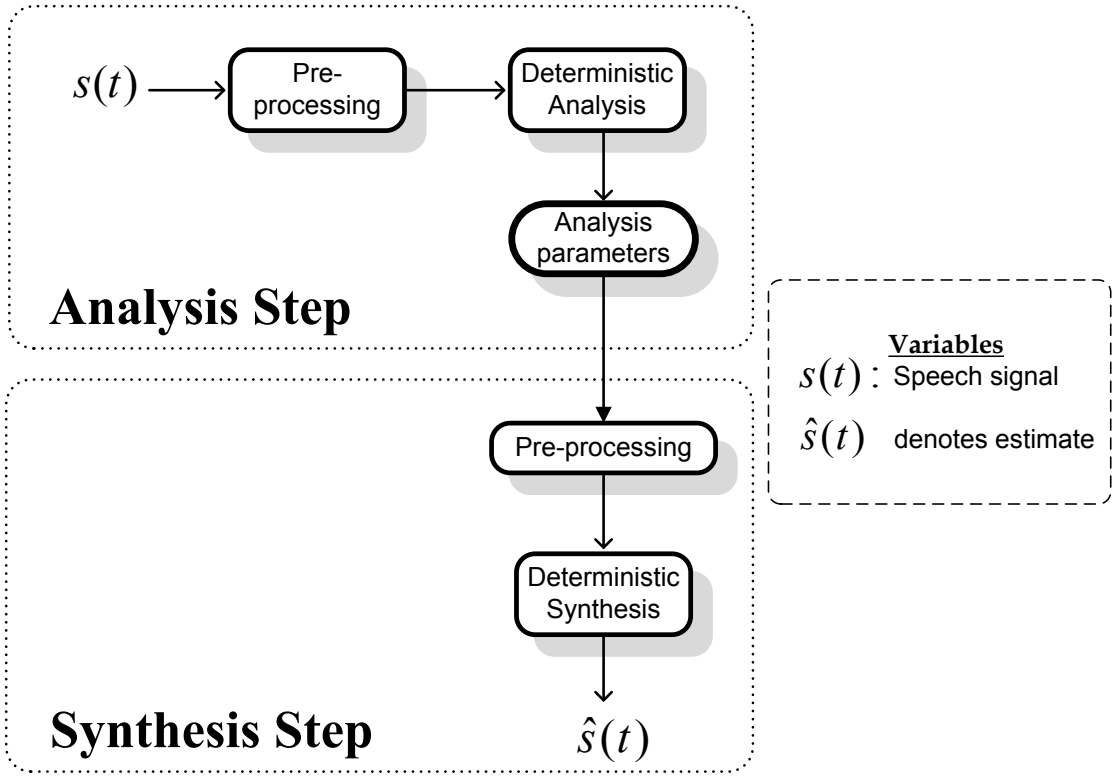


Figure 4.7: *A flowchart for a generalized full-band speech analysis system. Upper panel: analysis part. Lower part: synthesis part.*

4.7.1 Motivation

There are several reasons for suggesting such an approach: first of all, for voiced speech, a number of hybrid systems heavily rely on an accurate estimate of the so-called *maximum voiced frequency - MVF*, which divides the spectrum of voiced speech in a deterministic and a stochastic part. The efficient estimation of the MVF is critical for the performance of the system and its resulting modifications. Second, as it is described in [DS13], and supported by other researchers [DD97, DDH06] such a MVF is not necessary from a speech production point of view. In Figure 4.8, a series of glottal pulses is depicted on the upper panel, and its corresponding magnitude spectrum on the lower panel. As it can

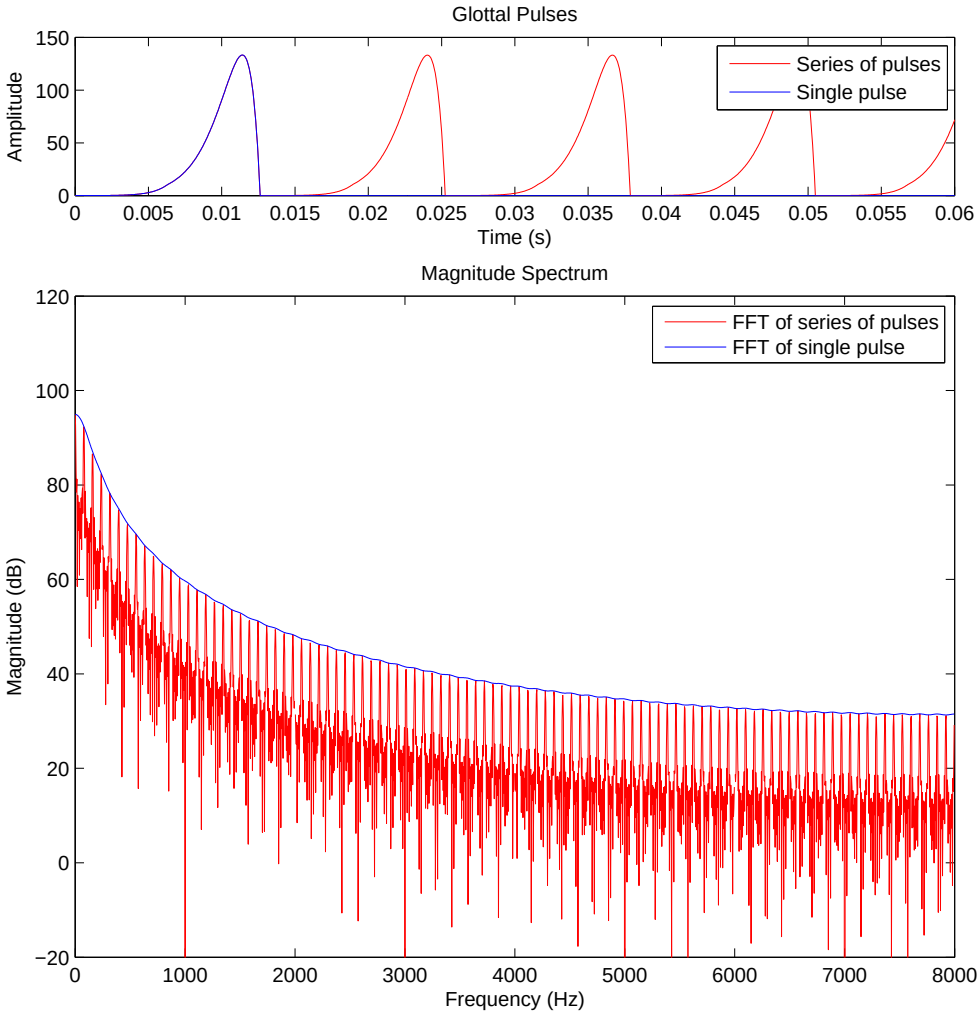


Figure 4.8: *Glottal pulses and corresponding magnitude spectra.*

be observed, the spectrum of the glottal pulse does not abruptly stop at some frequency but continues to decrease up to the Nyquist frequency. Moreover, the stationarity assumption of speech models states that speech is relatively stationary in a $20 - 30$ ms interval. However, even in purely harmonic speech segments, the stationarity assumption holds for low-to-medium range frequencies but not for higher frequencies. The variation of higher frequencies is also higher, since any small fluctuation of the fundamental frequency is propagated to the higher harmonics proportionally to the harmonic number. Thus, the stationarity assumption does *not* hold for these frequencies and the use of stationary basis functions, as in DFT for example, does not fit well to the high frequency region of the spectrum. The latter can be further supported by the recently suggested *Fan-Chirp transform - FChT* [KW06, WK07]. The FChT uses a chirp related frequency basis adapted to the input signal. An example is depicted in Figure 4.9, where in the upper panel the speech signal in time domain is shown, in the middle panel the spectrogram obtained by the DFT is shown, and in the lower panel the corresponding spectrogram obtained from the FChT is illustrated. Black colored parts denote voiced speech, green colored parts denote unvoiced speech. Although the low voiced frequencies in the DFT-based spectrogram seem to have a regular structure, this is not true for the mid- and high-range frequencies, where the frequency content is blurred. This is exactly because of the non-stationary nature of the frequency content. On the contrary, the use of the FChT reveals a regularity in the frequency content across all frequencies of voiced parts.

To show this in a more illustrative way, in Figure 4.10 we show one slice of voiced speech of each spectrogram in Figure 4.9. In standard hybrid speech analysis systems, a MVF separates the spectrum in a deterministic (left part of upper panel in Figure 4.10) and a stochastic (right part of upper panel in Figure 4.10) part. The former is mostly represented by a sum of sinusoids, and the latter is modelled by modulated noise. A careful inspection of corresponding Fan-Chirp Transform in the same figure reveals that harmonic structure is present in the frequency range which is supposed to be modelled with stochastic components. This observation clearly shows that current speech analysis systems often

overestimate the need of a MVF and a harmonic or sinusoidal representation could be applicable for all frequencies. Compared to the FChT, the aSMs provide more freedom in the instantaneous frequency curves, but the underlying principle is the same: local adaptivity.

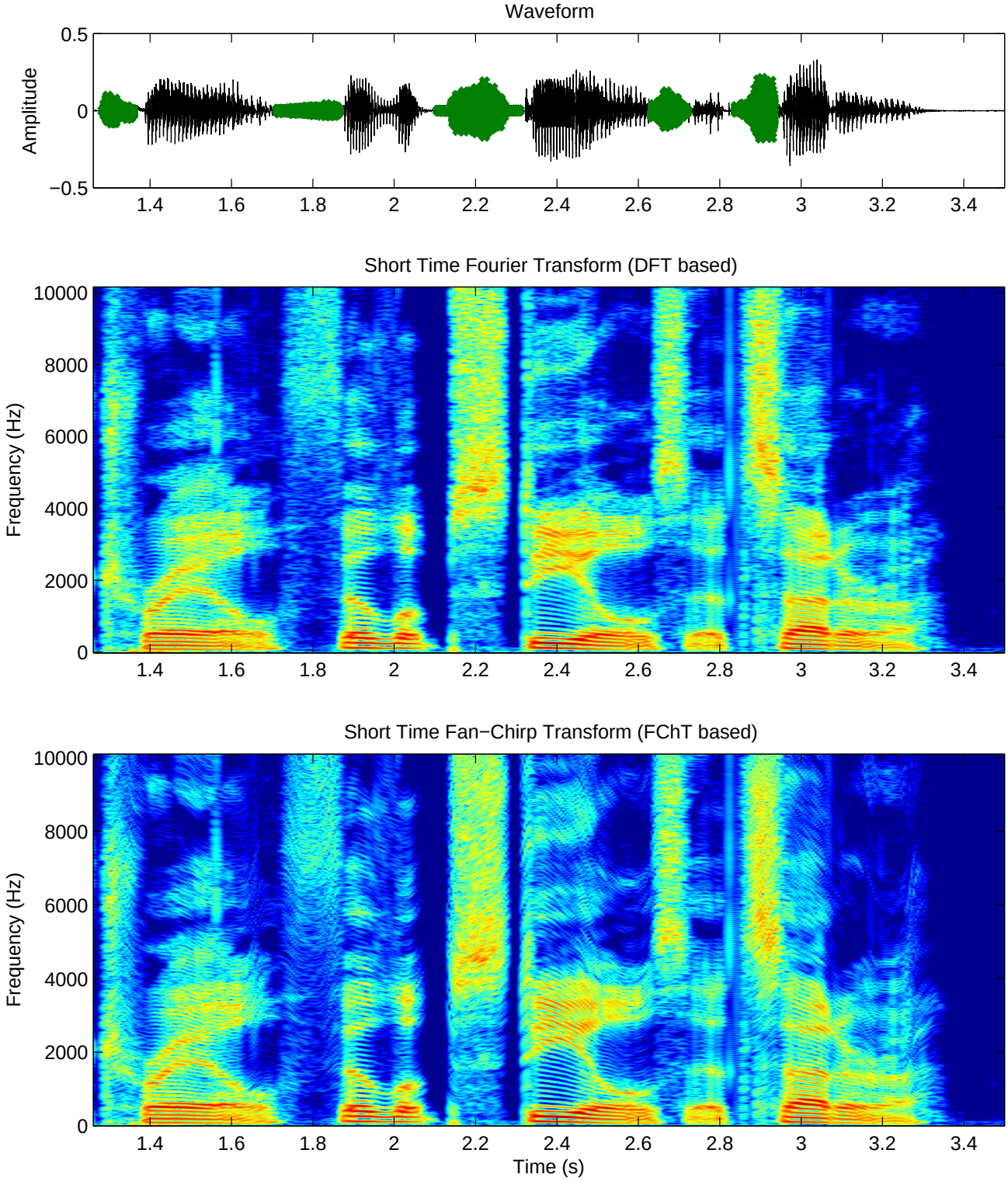


Figure 4.9: *Spectral analysis of speech. Black colored parts denote voiced speech, green colored parts denote unvoiced speech. Upper panel: Speech signal. Middle panel: DFT-based spectrogram. Lower panel: FChT-based spectrogram.*

However, it is questionable how and why sinusoids are appropriate when representing consonants, such as *fricatives*

or *voiceless/voiced stop sounds*. A *stop* sound is produced with complete closure of the articulators involved, so that the stream of air can not escape through the mouth. *Voiced stops* are produced with vibrating vocal folds whereas in *voiceless stops* vocal folds are apart. In voiced stops, there are oscillations right before the burst, whereas in voiceless stops, there is no oscillation and, in many languages, there is *aspiration* after the burst. A *fricative* is produced with close approximation of the two articulators, so that the stream of air is partially obstructed and turbulent airflow is produced. It is well-known that conventional sinusoidal or harmonic models cannot efficiently tackle this problem, due to the highly non-stationary nature of these parts of speech and the stationarity assumption inside the analysis window of the models. The standard Sinusoidal Model [MQ86] treats unvoiced parts, and hence stop sounds, the same way as voiced ones, based on the principle that periodogram peaks are close enough to satisfy the requirements imposed by the Karhunen-Loeve expansion [Tre68]. However, the perceptual quality of the reconstructed speech is rather mediocre, especially under modifications. The Harmonic + Noise Model, and other systems as well, utilizes a stochastic component that models unvoiced speech as time- and frequency-modulated noise. Although this representation is perceptually closer and allows for better manipulation in case of modifications, still it does not attain the quality of the original speech. To this direction,

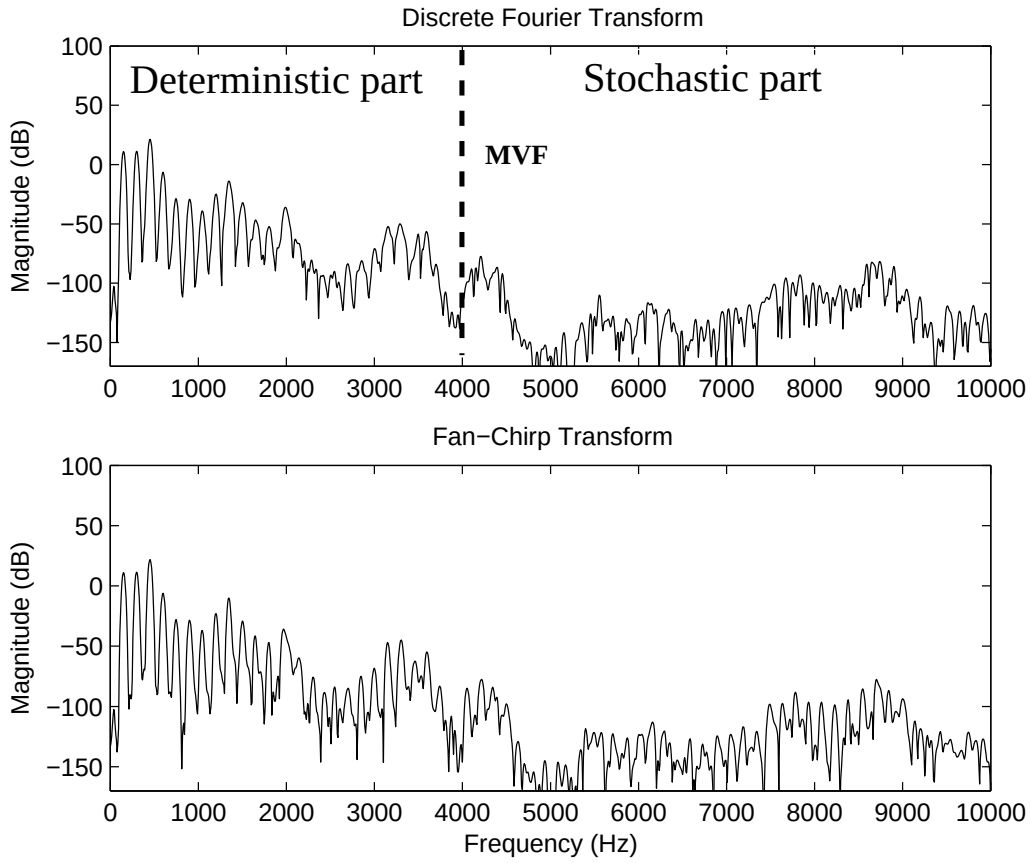


Figure 4.10: *Spectral analysis of voiced speech. Upper panel: FFT of a voiced speech segment. MVF denotes Maximum Voiced Frequency. Lower panel: Fan-Chirp Transform.*

let us examine more closely a consonant sample using the FFT and the FChT. In Figure 4.11, a fricative /s/ is depicted, along with its corresponding short time FFT and short time FChT obtained from a window centered in the middle of the sound, and the corresponding spectrograms based on the FFT and the FChT. Here, a similar conclusion can be drawn. Although there are not any prominent spectral peaks that can justify a sinusoidal model framework, intuitively, an adaptive decomposition of unvoiced speech should attempt to locate "optimal" frequency tracks that collectively minimize the mean-square error inside the frame. These "optimal" frequency tracks become more discernible in the FChT-based spectrogram, whereas in the DFT-based spectrogram severe blurring still exists.

In the next section, it will be shown how adaptivity can compensate the representation problem of fricatives and stop sounds, both voiceless and voiced. A complete and thorough study on the representation of unvoiced speech using adaptive speech models is beyond the scope of this thesis. For now, it is sufficient to show that adaptivity is capable of accurately representing stops and fricatives as AM-FM components.

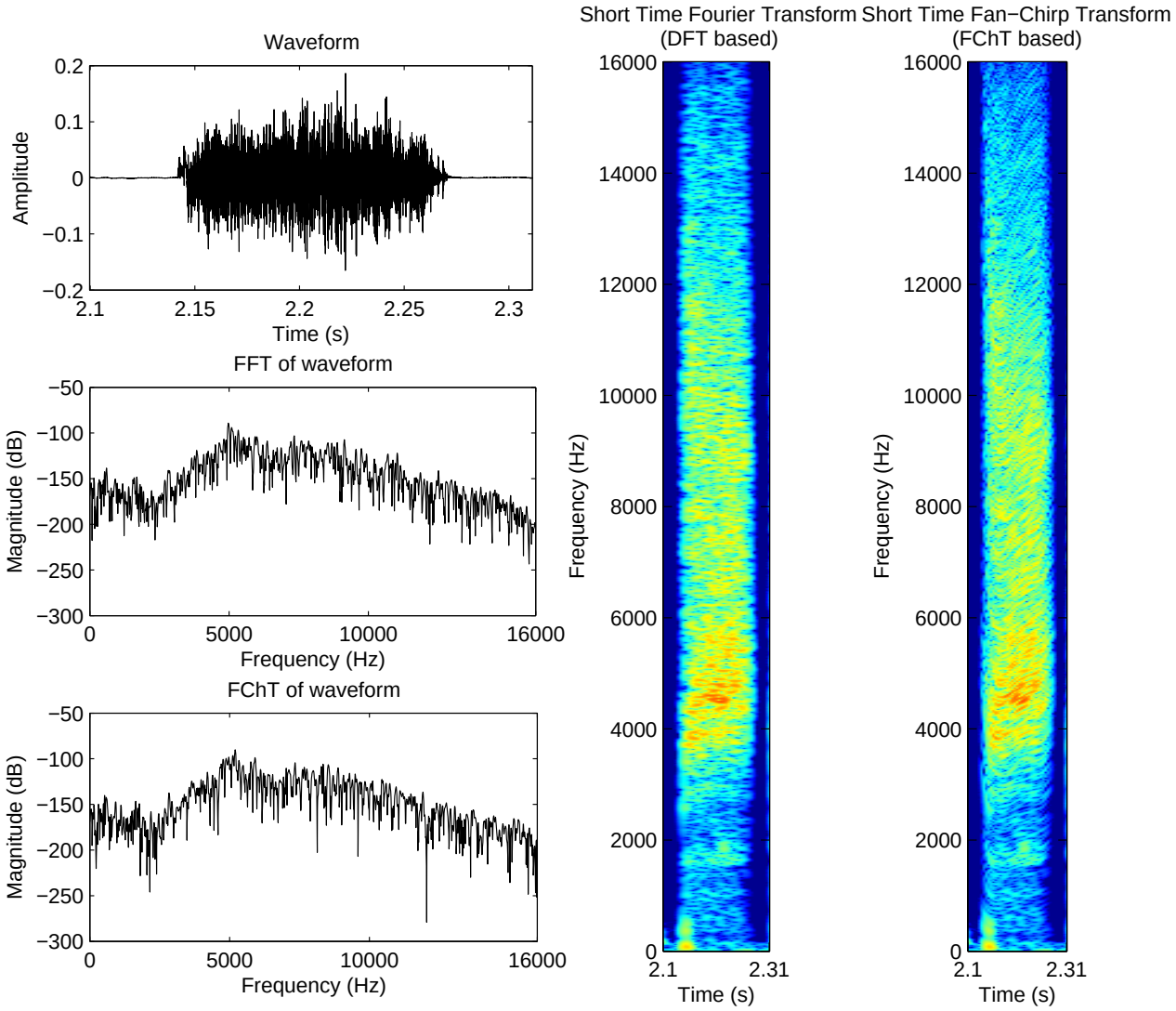


Figure 4.11: *Spectral analysis of unvoiced speech. First column: Unvoiced speech waveform, its FFT-based magnitude spectrum, and its FChT-based magnitude spectrum. Second column: FFT-based spectrogram slice of the corresponding waveform. Third column: FChT-based spectrogram slice of the corresponding waveform.*

4.8 Towards a uniform, adaptive, full-band AM-FM representation of speech

Let us roughly separate unvoiced speech into stop sounds (voiceless and voiced) and fricatives, and first deal with the former category. It has long been known that sinusoidal modelling is inefficient to model sounds well, since they are broadband signals and have noise-like frequency domain structure [Mac96]. Although sinusoidal models have been successfully applied for non-voiced speech, the nature of stops makes their modelling by a sum of stationary sinusoids inappropriate, because of the sudden change in amplitude during the release burst. An attempt to model the voiceless stops with a finite number of stationary sinusoids (i.e., one sinusoid every $80 - 100$ Hz) will manifest the Gibbs phenomenon just before the release time instant (*pre-echo* effect). This leads to an audible release energy smearing and therefore to a reconstructed signal with reduced intelligibility compared to the original signal. One could argue that an effort to model stops with a certain high amount of sinusoids would suffice; however, this is proved to be both insufficient and costly, since it requires a transient detection algorithm [Lev99][Tho05] and some proper handling (i.e., transform coding [Lev99] [Spa94]). The use of short analysis windows when stop sounds are detected as in [Lev99], does not alleviate the pre-echo effect as it will be also shown here. Other techniques such as multi-resolution sinusoidal analysis have failed to eliminate or alleviate the pre-echo effect [Lev99]. Because of the aforementioned problems, copy strategies or transform coding are mostly used over the short time region of the attack onset in speech and audio synthesis state-of-the-art systems.

However, SRER is commonly used as a global signal measure and thus small but pre-echo-related modelling errors at

the abrupt part of the reconstructed signal may be buried into the global modelling error. So, additionally, a *local* SRER will be used in order to reveal the efficiency of the reconstruction around the pre-echo area. Experiments show that the adaptive models provide a nearly pre-echo-free representation of stop sounds, without the necessity of using neither very short analysis windows for these sounds, nor a transient detector as in [Lev99]. Also, it is shown that for the adaptive sinusoidal models the overall quality in modelling stops is high in terms of SRER. Since voiced stop sounds exhibit some oscillatory behaviour, and thus their modelling is not as difficult as for their voiceless counterparts, our main focus will be on the voiceless stop sounds.

4.8.1 Adaptive Sinusoidal Modelling of Stop Sounds

In this section, a comparison between the conventional sinusoidal model [MQ86] and the adaptive sinusoidal models on a typical voiceless stop signal is presented. To this direction, a stop signal /t/ is extracted from a clear speech recording and is analyzed using the SM and the adaptive models. Since stops are broadband signals, attention should be paid in setting the parameters of the models. Both SM and adaptive models perform well under quasi-periodicity assumption, but this is not the case of this sound. SM performs peak picking on the spectrum of the input signal, so it does not need any initial frequency parameter values. On the other hand, adaptive models solve a least squares minimization problem, which requires a set of initial frequencies $\{f_k\}$ (i.e., harmonic frequencies for a voice sound). It is suggested that for a sampling frequency of $F_s = 16$ kHz, a low initial frequency value such as 80 Hz, which results in frequency values of $80k$ Hz, $k = -100, \cdots, 100$, is enough to span the frequency spectrum, i.e. it is a full band analysis. The QHM frequency mismatch correction mechanism will fine-tune the frequencies around the maxima of the spectrum, and thus the highest energy components will be modelled.

For all models, the Hamming window is used and it is set to 3 times the larger pitch period ($1/80$ s). A 2048-point FFT is computed for the analysis frame and a maximum of 100 spectral peaks are allowed for the SM. The number of components is also set to 100 and five adaptations are allowed at most for the adaptive models. The frame rate is 1 sample for all models. Global as well as local SRER measures are computed. Local SRER focuses only before the release (burst) time and is computed over an interval of $\frac{N_w}{2}$ samples right before the onset of the waveform, where N_w is half the analysis window length. Figure 4.12 shows the reconstructed signals for each case, with the aforementioned parameters, while Table 4.1 shows the global and local SRER evolution for all models.

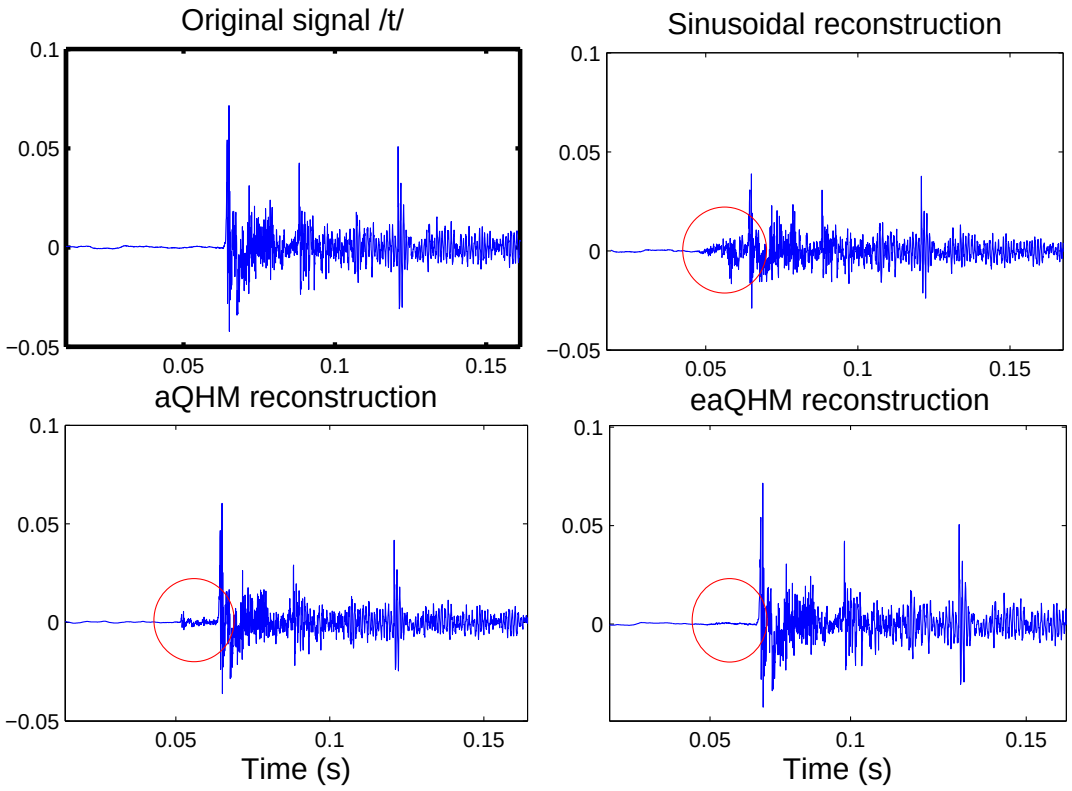


Figure 4.12: *Estimated waveforms for a stop sound. Upper panel: Original (left) and SM (right) reconstruction. Lower panel: aQHM (left) and eaQHM (right) reconstruction. The red ellipses mark the region where pre-echo occurs.*

Model	Global SRER (dB)	Local SRER (dB)
SM	6.9	4.1
aQHM	22.4	25.8
eaQHM	32.1	41.6

Table 4.1: *Global and Local Signal to Reconstruction Error Ratio values (dB) for all models on stop sound /t/.*

Based on the performance of the models in terms of local SRER, it is worth noticing that both adaptive models outperform the conventional sinusoidal model. Specifically, the eaQHM performs better than the aQHM, and both outperform SM in terms of reconstruction quality. Comparing the two adaptive sinusoidal models, the pre-echo effect is highly reduced for the aQHM, while it is mainly eliminated for the eaQHM. Moreover, the results from the global SRER show that both adaptive models produce high quality reconstruction of the stop sound compared with the conventional sinusoidal model. Experiments in manipulating the window length, the number of components, or both, did not provide any significant improvement for the SM representation. Therefore, it seems that the adaptation process is the key for accurate modelling of stops using long analysis windows. Moreover, it was observed that SM is unable to detect frequency components at the pre-echo area because of the stationary basis projection. This is not the case for the adaptive models, and it can be justified by the fact that adaptive modelling is a non parametric representation, taking into account local frequency (and amplitude, for the eaQHM) variation, which is pertinent for voiceless stop sounds. As a conclusion, adaptive sinusoidal modelling can represent highly non-stationary speech segments, like the voiceless stops, by projecting them on a set of also non-stationary basis functions that can capture the local characteristics of the signal. Thus, the pre-echo effect can be highly alleviated and sometimes eliminated, while a very high reconstruction performance is attained.

4.8.2 Database Validation for Stop Sounds

The next step is to strengthen the conclusions of the previous section using two databases of stop sounds.

Small Scale Validation

A small database of French speakers with male and female speakers is used for our purpose. Different voiceless stops corresponding to phonemes /p/, /t/, and /k/ are manually extracted from clean speech and are analyzed using the conventional sinusoidal models and the adaptive models, along with their voiced counterparts, for comparison purposes. The exact location of the burst release is manually identified, so as to compute local SRER accurately. The same metrics and parameters used in the previous section are also used here, i.e. a frame rate of 1 sample and an analysis window of 3 pitch periods. The sounds are categorized into classes of phonemes (20 waveforms for each class) and Table 4.2 shows mean value results for both global and local SRER. Apparently, adaptive modelling maintains its high SRER levels throughout different types of voiceless stops.

Small Scale Validation for Stop Sounds						
Global Signal to Reconstruction Error Ratio (dB)						
Model	$/p/$	$/t/$	$/k/$	$/b/$	$/d/$	$/g/$
SM	13.5	14.6	13.4	17.2	15.3	17.6
aQHM	20.8	23.2	23.2	28.9	27.9	28.2
eaQHM	27.1	31.2	28.4	35.5	33.5	33.1
Local Signal to Reconstruction Error Ratio (dB)						
Model	$/p/$	$/t/$	$/k/$	$/b/$	$/d/$	$/g/$
SM	7.5	4.4	7.2	12.6	12.8	13.1
aQHM	22.2	24.1	24.1	28.8	25.3	28.7
eaQHM	29.0	33.7	29.4	35.7	36.7	35.3

Table 4.2: *Global and Local Signal to Reconstruction Error Ratio values (dB) for all models on a small database of stops. Voiced stops are also included in this for comparison purposes.*

Large Scale Validation

A large scale validation is presented here. For convenience, only the eaQHM will be used in this validation, since it outperforms the aQHM and provides both amplitude and phase adaptation. A large database of both male and female French speakers is used. Phonetic labeling and manual segmentation is available in this database and thus stops can be easily extracted. In this experiment, 1000 stop sounds are considered. For such an amount of test signals, *the exact burst locations are not available* and consequently, the local SRER is not computed. Moreover, the frame rate of 1 sample used in the previous section, although providing high SRER values, is time consuming and is not realistic for applications. Hence, different frame rates are selected, namely 1ms, 2ms, and 4ms. Parameters other than the frame rate remain the same as in the previous sections. The interpolation schemes used in this experiment are described in [PRS11] and [MQ86] (i.e., for SM, linear interpolation between amplitudes and cubic interpolation between phases). Table 4.3 presents the results per phoneme, in terms of mean value of global SRER.

Large Scale Validation for Stop Sounds							
Global Signal to Reconstruction Error Ratio (dB)							
Step	Model	/p/	/t/	/k/	/b/	/d/	/g/
1ms	SM	12.7	12.8	12.4	16.6	14.9	15.3
	eaQHM	25.4	25.7	27.2	32.9	32.2	32.9
2ms	SM	12.8	12.7	12.3	16.5	15.0	15.4
	eaQHM	26.1	26.1	26.0	31.7	31.4	34.6
4ms	SM	12.9	12.6	12.2	16.7	15.0	15.3
	eaQHM	23.7	24.2	24.4	29.4	29.5	30.9

Table 4.3: *Global Signal to Reconstruction Error Ratio values (dB) for all models on a large database of stops. Voiced stops are also included in this for comparison purposes. Step denotes the analysis frame rate.*

As it can be observed from Table 4.3, the performance of the adaptive models sustains in high reconstruction levels, even with a frame rate up to 4 ms. The mean standard deviation per model is: 3 dB (SM) and 4.5 dB (eaQHM). No significant variations in standard deviation were observed across phonemes. Experiments with higher frame rates, such as 5 and 10 ms, showed an average decrease in performance of 3 and 7 dB respectively, compared to the 4 ms case, for all models and phonemes. Moreover, at higher frame rates the pre-echo effect was *partially* alleviated only for eaQHM modelling. Therefore it is suggested, as a rule of a thumb, the use of as low frame rate as possible. The average number of adaptations required for the convergence criterion in Eq.(2.52) is found to be 4.7 for the eaQHM, for all step sizes presented in Table 4.3.

4.8.3 Adaptive Sinusoidal Modelling of Fricative Sounds

As a reminder, fricatives are consonants produced by forcing air through a narrow passage made by placing two articulators close together. For modelling such sounds, a similar strategy as for stop sounds is followed for their analysis. A test case of a fricative /s/ is depicted in Figure 4.13. The signal is sampled at $F_s = 16$ kHz, and a low initial frequency value such as 80 Hz, which results in frequency values of $80k$ Hz, $k = -100, \cdots, 100$, is chosen. Hence, the frequencies cover the full-band of the spectrum. The other experimental settings are exactly the same as in the stop sound example discussed earlier, except that aQHM has been omitted for convenience.

In Table 4.4, the SRER performance of adaptive models compared to the standard SM is presented for our test case, only here there is no local SRER calculation. Clearly, adaptivity is able to represent fricatives very accurately, compared to stationary models, such as SM.

Model	Global SRER (dB)
SM	8.86
eaQHM	27.63

Table 4.4: *Signal to Reconstruction Error Ratio values (dB) for all models on a fricative sound /s/.*

4.8.4 Database Validation for Fricative Sounds

To validate our results, 485 voiced and voiceless fricatives have been automatically extracted from speech utterances. Voiced fricatives include /v/, /ð/, /s/, and /ʃ/, while unvoiced ones are /f/, /θ/, /z/, and /ʒ/. The number of fricatives

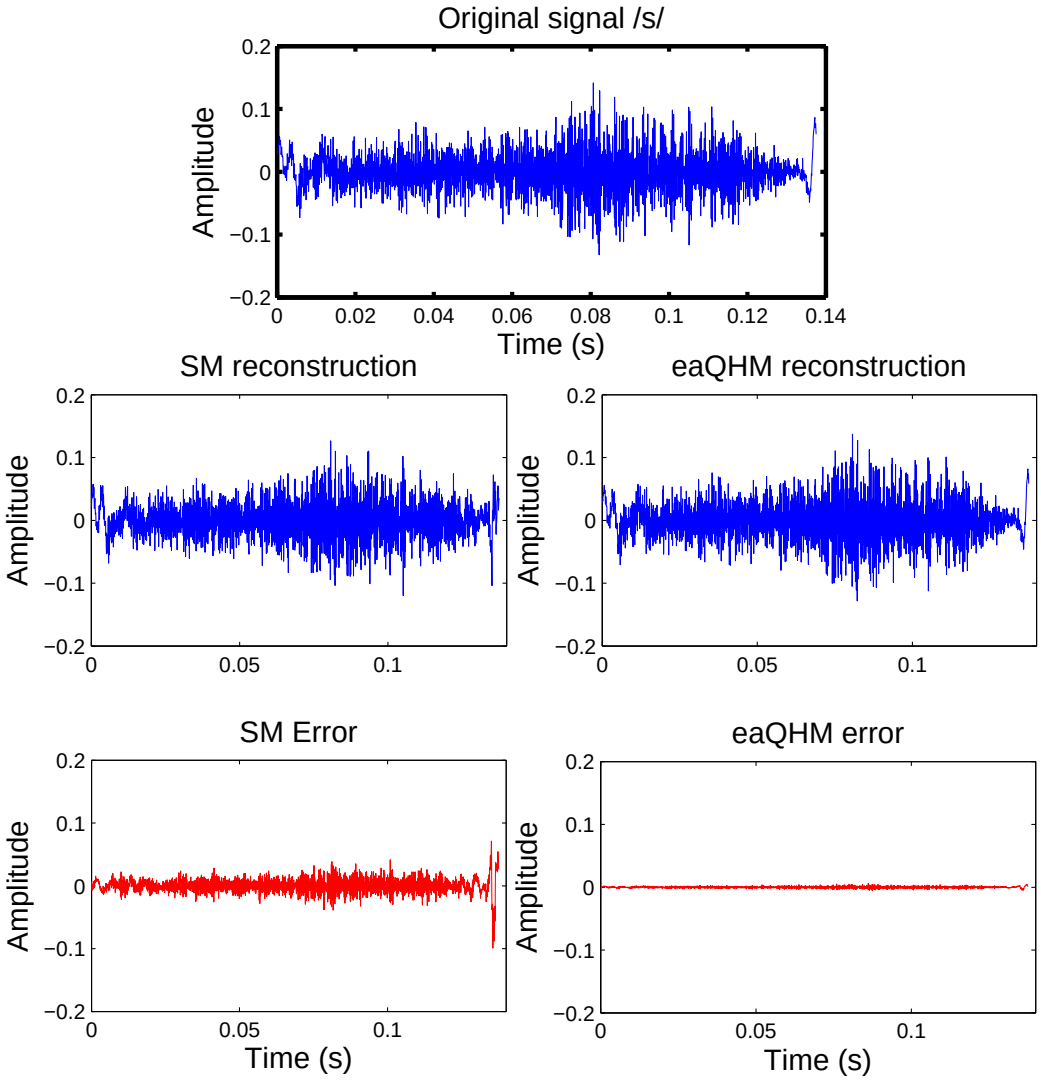


Figure 4.13: *Estimated waveforms for a fricative sound. Upper panel: Original signal. Middle panel: SM (left) reconstruction and eaQHM (right) reconstruction. Lower panel: SM (left) and eaQHM (right) reconstruction error.*

extracted from male speaker was almost the same with those from female speakers. As in the case of stop sounds, the frame rate of 1 sample used in the previous section is not realistic for applications. Hence, the same frame rates as previously are selected, namely 1 ms, 2 ms, and 4 ms. Parameters other than the frame rate remain the same as in the previous sections. Table 4.5 presents the results per fricative, in terms of mean value of SRER.

Large Scale Validation for Fricatives									
Signal to Reconstruction Error Ratio (dB)									
Step	Model	/v/	/ð/	/s/	/ʃ/	/f/	/θ/	/z/	/ʒ/
1 ms	SM	14.7	13.2	13.9	11.3	12.7	15.1	17.5	17.3
	eaQHM	26.4	25.6	24.1	26.4	25.8	24.3	29.5	28.9
2 ms	SM	13.1	11.3	12.1	10.4	10.2	14.7	15.9	15.2
	eaQHM	23.5	23.1	22.6	24.7	23.5	22.6	28.6	27.8
4 ms	SM	12.2	10.6	11.2	9.6	9.7	8.9	13.3	13.7
	eaQHM	22.4	22.2	21.9	23.1	22.6	21.7	27.5	27.1

Table 4.5: *Signal to Reconstruction Error Ratio values (dB) for all models on a large database of fricatives. Step denotes the analysis frame rate.*

As it can be observed from Table 4.5, the performance of the adaptive models sustains in high reconstruction levels, even with a frame rate up to 4 ms. The mean standard deviation per model is: 3.4 dB (SM) and 4.1 dB (eaQHM). No significant variations in standard deviation were observed across different fricatives. Experiments with higher frame rates were performed as well, such as 5 and 10 ms, that showed an average decrease in performance of 3.9 and 6.5 dB respectively, compared to the 4 ms case, for all models and fricatives. Therefore it is suggested, as a rule of a thumb, the use of as low frame rate as possible to attain a high enough perceptual and reconstruction quality. The average number of adaptations required for the convergence criterion in Eq.(2.52) is found to be 4.7 for the eaQHM, for all step sizes presented in Table 4.5.

4.8.5 Discussion

In this section, modelling of voiceless stop sounds and fricatives is presented and addressed via adaptive modelling. The well-known pre-echo effect of stop sounds in sinusoidal modelling is demonstrated and a solution is shown to be provided by the eaQHM. Pre-echo arises from the inability of sinusoidal models to represent highly non-stationary short time attacks, typically encountered in voiceless stop sounds. Using adaptive modelling, the pre-echo effect is greatly alleviated. The latter is demonstrated analytically using a characteristic example, where the limitations of sinusoidal modelling are also presented, and is validated on two different databases of stop sounds. Metrics such as global SRER for overall modelling and local SRER for a specific focus on the pre-echo effect are used and confirm the superiority of adaptive over stationary (conventional) sinusoidal modelling in representing highly non-stationary parts of speech. Moreover, fricatives are demonstrated to be represented very accurately using adaptive models. It is shown that local adaptation of the analysis parameters results in AM-FM components that are able to decompose and reconstruct fricative sounds effectively. SRER measures validate the latter for different fricative categories and different frame rates. Conclusively, it has been demonstrated that the adaptive models are capable of modelling not only voiced speech but also unvoiced parts as well with high accuracy. This is important to support the transition from hybrid systems to full-band systems that operate on the full-length of the speech signal, without any quality degradation, and thus providing a uniform representation of speech as AM-FM components.

4.9 The full-band eaQHM analysis and synthesis system

In the previous section, motivation towards full-band systems was presented in the context of analysis and synthesis of consonants. It is now apparent that the eaQHM can handle both categories of speech, voiced and unvoiced. However, care should be taken in the details of the implementation of such a system, since stability, robustness, and consistency are not only desirable for analysis and synthesis but also for modifications. The details of a full-band eaQHM-based analysis and synthesis system will be presented next.

The idea behind full-band eaQHM is that a first *purely harmonic* approximation of the speech signal is obtained, which successively - through adaptations and frequency corrections - converges to an adaptive, quasi-harmonic representation.

The full-band signal is described as an AM-FM decomposition

$$d(t) = \sum_{k=-K}^{K} A_k(t) e^{j\phi_k(t)} \tag{4.27}$$

where $A_k(t)$ is the instantaneous amplitude and $\phi_k(t)$ is the instantaneous phase of the k^{th} component, respectively. The instantaneous phase term is given by

$$\phi_k(t) = \phi_k(t_i) + \int_{t_i}^{t} 2\pi f_k(u) du \tag{4.28}$$

where $\phi_k(t_i)$ is the instantaneous phase value at the analysis time instant t_i, f_s is the sampling frequency, and $f_k(t)$ is the instantaneous frequency of the k^{th} component.

4.9.1 Analysis

At first, an initial and *continuous* f_0 estimation for all frames is obtained, noted by $\hat{f}_0$. Although there is no f_0 in unvoiced frames, a rough estimate can be useful for initialization. Then, the next step is to assume a full-band harmonicity to obtain a first estimate of the instantaneous amplitudes of all the harmonics. Using a Blackman analysis window $w(t)$ centered at t_i and with support in $[t_i - T, t_i + T]$, where $2T$ is of 3 local pitch periods length, a frame of the analyzed

speech is initially modelled using a simple Harmonic Model as:

$$d(t) = \Big(\sum_{k=-L}^{L} a_k e^{j2\pi \hat{f}_k t} \Big) w(t) \tag{4.29}$$

where a_k is the complex amplitude of the k^{th} harmonic, $\hat{f}_k = k\hat{f}_0$ are the analysis frequencies, and L is the number of harmonics that span the whole spectrum up to Nyquist frequency. The estimation of the model parameters is obtained via Least Squares, as described in [Sty96]. As opposed to [PTRS10], where the initial f_0 estimation is refined using an iterative QHM (iQHM), in our work no f_0 refinement is necessary, thus reducing the overall complexity of the algorithm, and a simple amplitude estimation for each component is performed. In [PTRS10], iQHM is operating as a means to refine the f_0 estimate for voiced frames. Since iQHM also holds the stationarity assumption, it was judged not to be crucial in frequency refinement, especially in unvoiced frames, where the refinement could lead to instability (e.g. abrupt jumps) of the estimated f_0. Thus, the iQHM estimation was dropped in this system, allowing reduced complexity without loss of accuracy. As a final step, the overall signal can be synthesized by interpolating the $|a_k|$ and $\hat{f}_k$ values over successive analysis time instants t_i, thus obtaining

$$\hat{d}(t) = \sum_{k=-L}^{L} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \tag{4.30}$$

where

$$\hat{A}_k(t) = |a_k(t)| \tag{4.31}$$

$$\hat{\phi}_k(t_i) = \angle a_k(t_i) \tag{4.32}$$

and

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \int_{t_i}^{t} (2\pi k \hat{f}_0(u) + c(u)) du \tag{4.33}$$

4.9.2 Adaptation

The above model is still harmonic and stationary within an analysis frame. Therefore, in order to converge to quasi-harmonicity and to confront the stationarity issue, the projection of the signal onto a set of time-varying basis functions is suggested in [KPRS12], by using the parameters a_k and b_k of the Quasi-Harmonic Model (QHM) [PRS08]. This yields the eaQHM model:

$$d(t) = \left(\sum_{k=-L}^{L} \Big(a_k + t b_k \Big) \Big(\hat{A}_k(t) e^{j\hat{\phi}_k(t)} \Big) \right) w(t) \tag{4.34}$$

with

$$\hat{A}_k(t) = \frac{\hat{A}_k(t + t_i)}{\hat{A}_k(t_i)} \tag{4.35}$$

and $\hat{\phi}_k(t)$ as in Eq. (4.33). In this model, a_k, b_k are the complex amplitude and the complex slope of the k^{th} component, and $\hat{A}_k(t)$, $\hat{f}_k(t)$, $\hat{\phi}_k(t)$ are estimates of the instantaneous amplitude, frequency, and phase of the k^{th} component, respectively, from the previous analysis step. The a_k, b_k parameters are obtained via Least Squares as shown in Section 2.4. It is apparent that the basis functions where the signal is projected are time-varying. The adaptation is completed by using the frequency correction mechanism first introduced in [PRS08], and states that an estimate of the mismatch between the actual k^{th}-frequency and the estimated one, termed $\eta_k = f_k - \hat{f}_k$, is given by

$$\hat{\eta}_k = \frac{1}{2\pi} \frac{\Re\{a_k\}\Im\{b_k\} - \Im\{a_k\}\Re\{b_k\}}{|a_k|^2} \tag{4.36}$$

Hence, at the first adaptation, for the analysis time instant t_i, the instantaneous frequencies are $\hat{f}_k(t_i) = k\hat{f}_0(t_i) + \hat{\eta}_k(t_i)$ and the instantaneous phases become

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \int_{t_i}^{t} (2\pi \hat{f}_k(u) + c(u)) du \tag{4.37}$$

Then, a Least Squares solution for the a_k, b_k using these refined frequencies (and phases) leads to a better estimate of the instantaneous amplitudes $\hat{A}_k(t) = |a_k(t)|$ and the $\hat{\eta}_k$ terms. By iteratively adding the $\hat{\eta}_k$ term of the current adaptation on

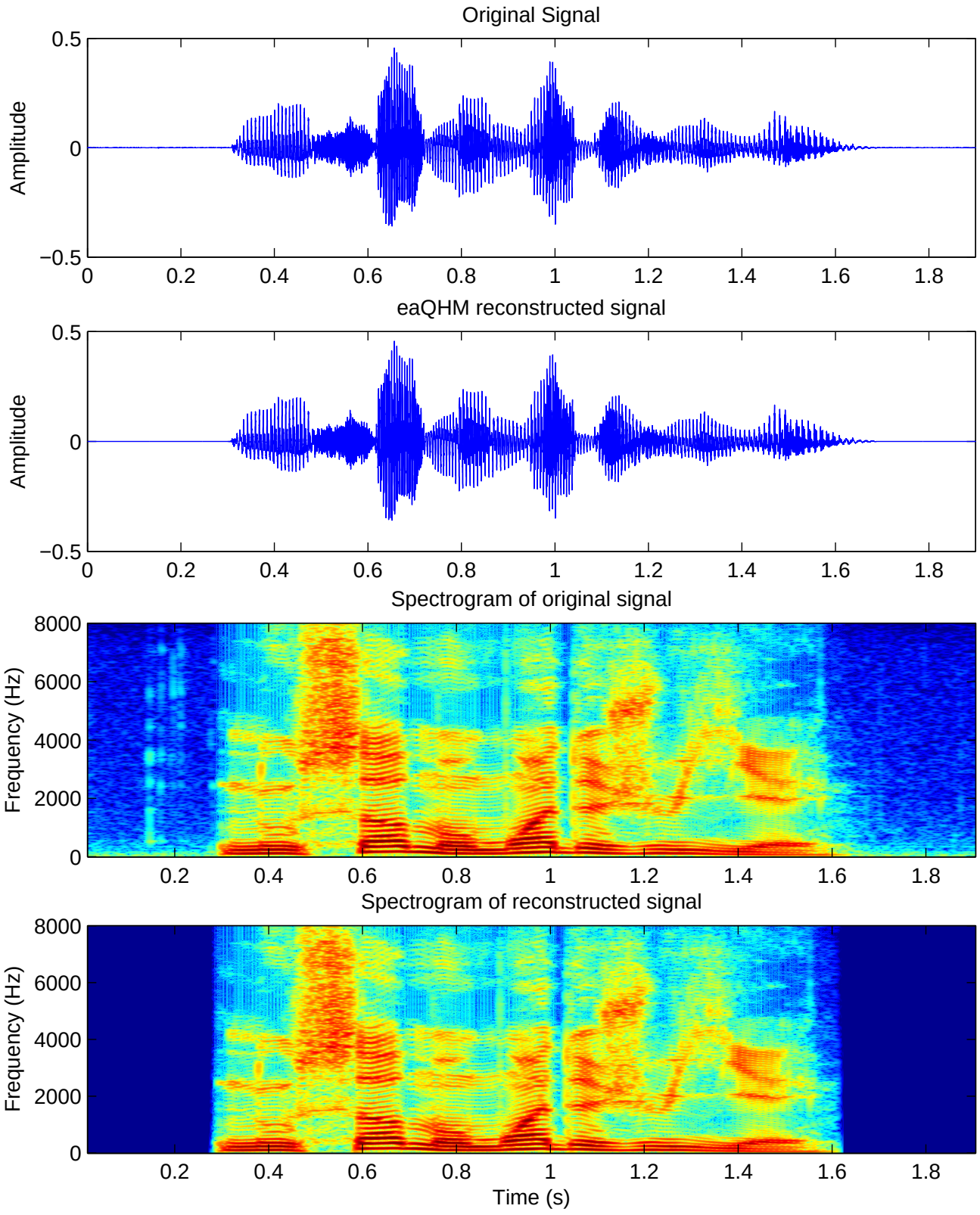


Figure 4.14: *extended adaptive Quasi-Harmonic Model: Original signal (first panel) and reconstructed signal (second panel) along with their corresponding spectrograms (third and fourth panel) for a male speaker.*

the k^{th}-frequency track of the previous adaptation, the frequency tracks deviate from strict harmonicity and represent the underlying actual frequencies better. Additionally, and on the contrary to previous works [PTRS10, PRS11], where the frequency correction estimation $\hat{\eta}_k$ on each adaptation should be less than $f_0/2$, in this approach it is supposed that after

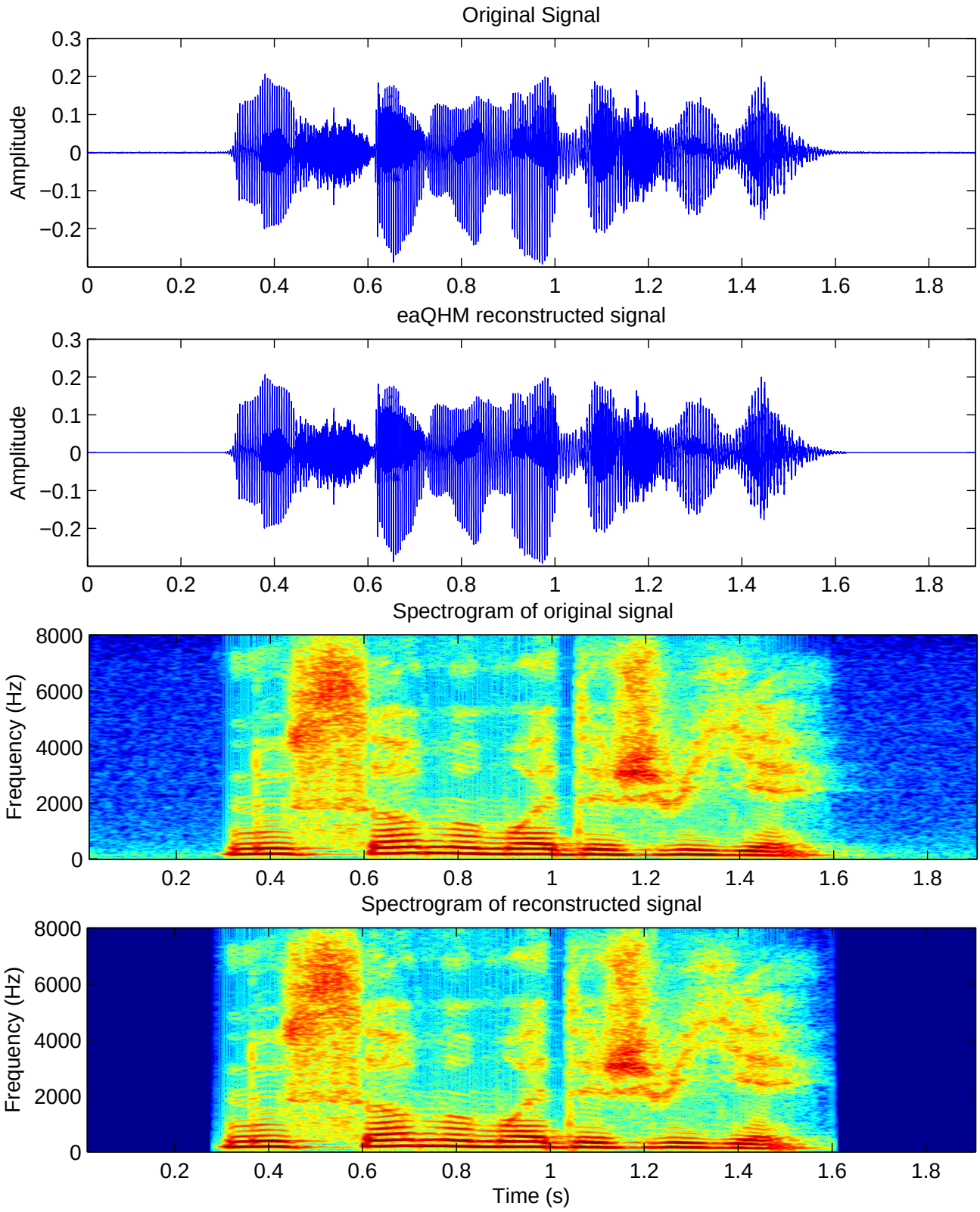


Figure 4.15: *extended adaptive Quasi-Harmonic Model: Original signal (first panel) and reconstructed signal (second panel) along with their corresponding spectrograms (third and fourth panel) for a female speaker.*

each adaptation the estimated frequencies become more and more localized to the actual frequencies, so the frequency

correction for a given analysis time instant t_i is constrained as in

$$|\hat{\eta}_k(t_i)| \leq \frac{\hat{f}_0(t_i)}{m+1} \tag{4.38}$$

where $m \in \{1, \cdots, M\}$ is the current adaptation number and M is the maximum number of allowed adaptations (in our experiment, $M = 6$). This way, any relatively large frequency correction value - which often leads to audible artefacts - that might be obtained in a higher adaptation step will be suppressed. This constraint is also motivated by the fact that unvoiced parts should be handled by the same model. This way, the AM-FM components are kept more "tight" in their variability in noise representation. Finally, this adaptation scheme continues until a convergence criterion is met, which is related to the overall Signal-to-Reconstruction-Error Ratio (SRER), that is, when the SRER stops increasing after each adaptation, then the algorithm is considered to have converged.

4.9.3 Synthesis

In the synthesis stage, the k^{th} instantaneous amplitude track, $\hat{A}_k(t)$, is computed via either linear or spline interpolation of the successive estimates from the last adaptation step. The k^{th} instantaneous frequency track, $f_k(t)$, is also computed via spline interpolation. Also, it is worth noting that a frequency matching mechanism is trivial, since the analysis frequencies are integer multiples of a fundamental and the number of components is constant. As for the k^{th} instantaneous phase track, $\hat{\phi}_k(t)$, the non parametric approach based on the integration of instantaneous frequency is followed, as it is shown in the adaptation steps of the analysis. Finally, the speech signal can be approximated by its time-varying components using:

$$\hat{d}(t) = \sum_{k=-L}^{L} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \tag{4.39}$$

A block diagram of the algorithm is depicted in Figure 4.16.

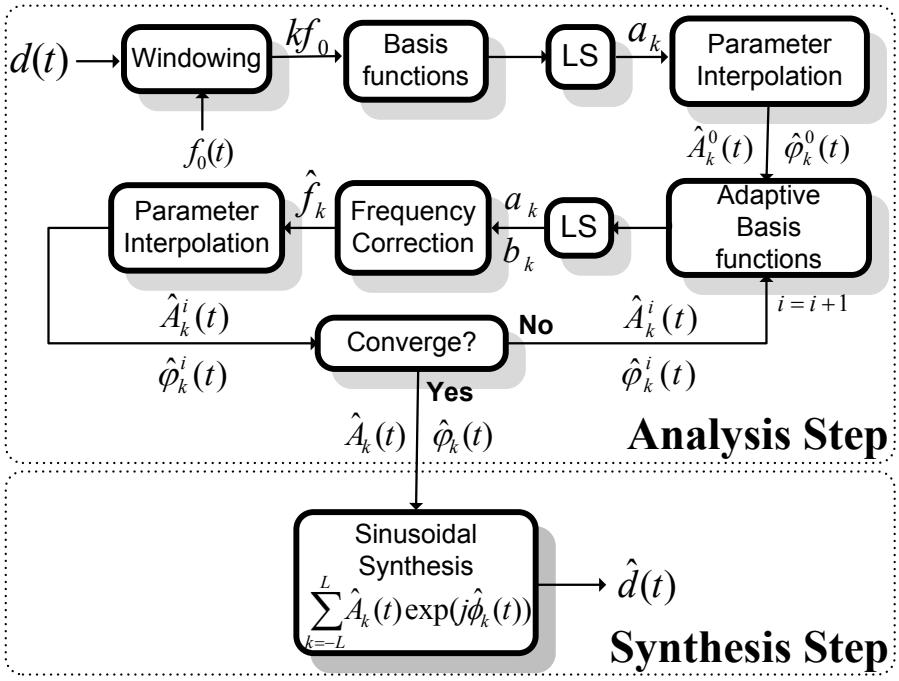


Figure 4.16: *Block diagram of the eaQHM system.*

4.9.4 Examples

Two examples are shown in Figures 4.14, 4.15 for a male and a female speaker, respectively.

4.10 The full-band aHM analysis and synthesis system

As discussed in Section 4.4, the adaptive Harmonic Model [DS13] was developed by Degottex in parallel to this thesis, and its original form suggested an analysis and synthesis in the full-band. For completeness and convenience, a very brief review of the analysis and synthesis schemes of the full-band adaptive Harmonic Model (aHM) is presented in this section. The aHM is actually nothing more than the aHNM discussed earlier, but with no noise component.

The adaptive Harmonic Model can be mathematically described as:

$$s(t) = \sum_{k=-K}^{K} a_k(t) e^{jk\phi_0(t)} \tag{4.40}$$

where $a_k(t)$ is a complex function that copes with the amplitude and the instantaneous phase of the k^{th} harmonic component, while K is the number of the components, and $\phi_0(t)$ is a real function defined as the integral of the fundamental frequency $f_0(t)$:

$$\phi_0(t) = \int_0^t 2\pi f_0(u) du \tag{4.41}$$

4.10.1 Analysis

In the analysis step, a parametrization of the speech signal at each analysis time instant t_a^i is undertaken. At first, a sequence of the analysis time instants are created in the voiced parts of speech using the provided $f_0(t)$ track, such we have one analysis time instant per pitch period. In unvoiced segments, even though the estimated $f_0(t)$ is meaningless, it can be used to generate the corresponding analysis time instants. Moreover, if the distance between t_a^i and t_a^{i+1} is short enough, aHM can model the amplitude variations of the unvoiced signal (like in plosives). Thus, the upper limit of the size of the analysis window is 20ms and the lower limit comes from the provided $f_0(t)$ track, and is therefore set to 50Hz. Around each analysis time instant t_a^i, a Blackman window with a length of 3 local pitch periods is applied to the speech signal. The phase track $\phi_0(t)$ is then computed by means of spline interpolation of f_0^i using the integration formula in Eq.(4.41).

4.10.2 Synthesis

In the synthesis step, each harmonic is generated in separate, one after the other, without using any window. Each harmonic component is synthesized by its parameters, namely its amplitudes $|a_k^i|$, phases $\angle a_k^i$, and fundamental frequency f_0^i. First, the instantaneous amplitude, $|a_k(t)|$, of the k^{th} harmonic is simply obtained by linearly interpolating the estimated $|a_k^i|$ on the analysis time instants t_a^i, on a logarithmic scale. The instantaneous phase $\angle a_k^i$ cannot be interpolated directly across time to obtain $a_k(t)$ because of its rotation due to the time advance between analysis time instants. Therefore, it is proposed to remove this effect using the integral of $f_0(t)$ from the start of the signal, and obtain the *relative phase - RP*:

$$\angle \tilde{a}_k^i = \angle a_k^i - k\phi_0(t_a^i) \tag{4.42}$$

Thus, by assuming that the shape of the signal is changing smoothly, the phase values change also smoothly from one analysis time instant to the other. Then, the RP $\angle \tilde{a}_k^i$ can be interpolated to obtain its continuous counterpart, $\angle \tilde{a}_k(t)$. Additionally, a spline or cubic interpolation is necessary such as its time derivative, the frequency, is still continuous. All along the iterative process, and since the harmonic numbers K^i increase independently from one analysis time instant to the other, there are often missing components in the interpolations of amplitude and instantaneous phase. If this is the case, then the amplitude of the missing component is set to -300 dB and the corresponding phase $\angle \tilde{a}_k(t)$ is set to zero.

4.10.3 Examples

Two examples are shown in Figures 4.17, 4.18 for a male and female speaker, respectively.

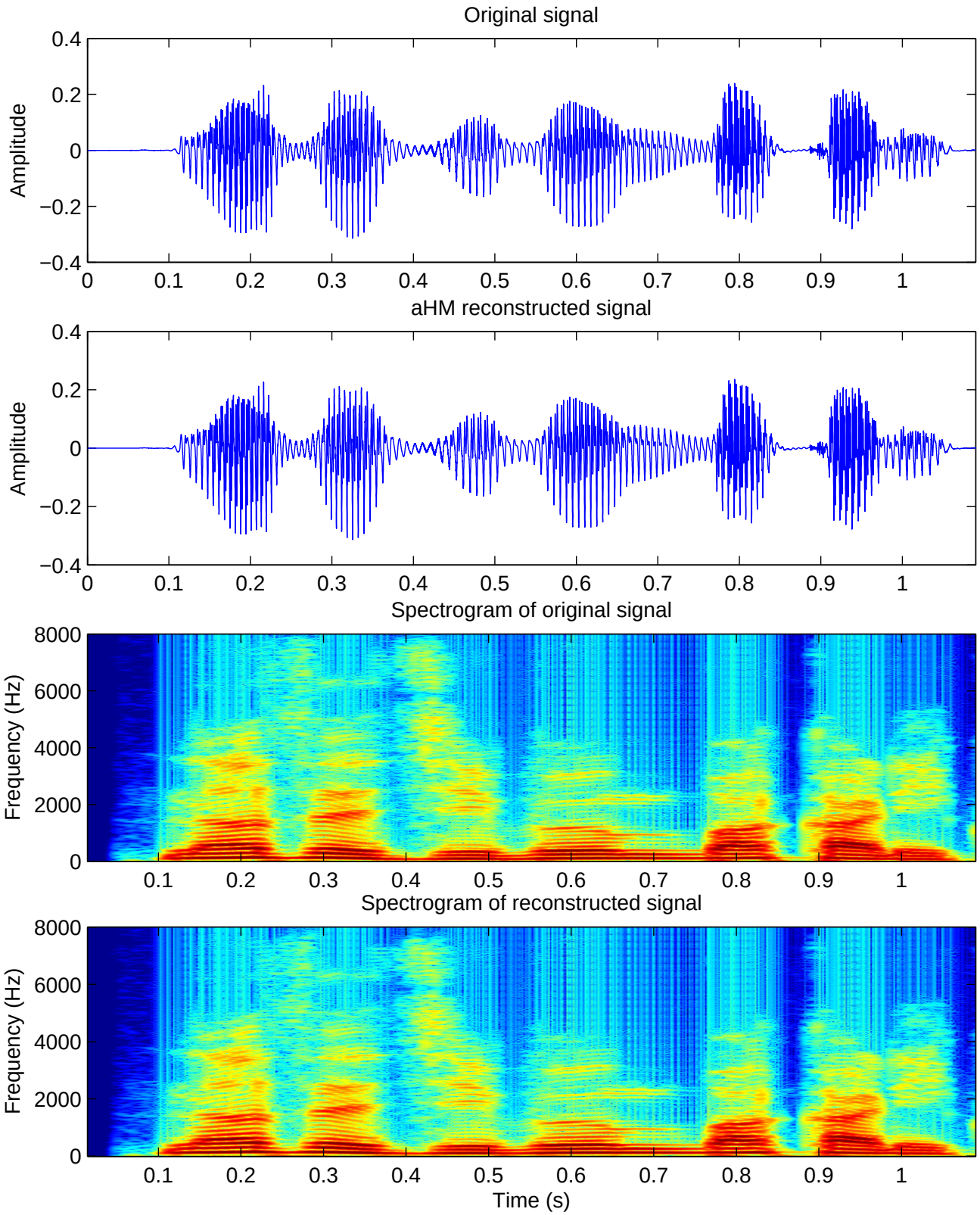


Figure 4.17: *adaptive Harmonic Model: Original signal (first panel) and reconstructed signal (second panel) along with their corresponding spectrograms (third and fourth panel) for a male speaker.*

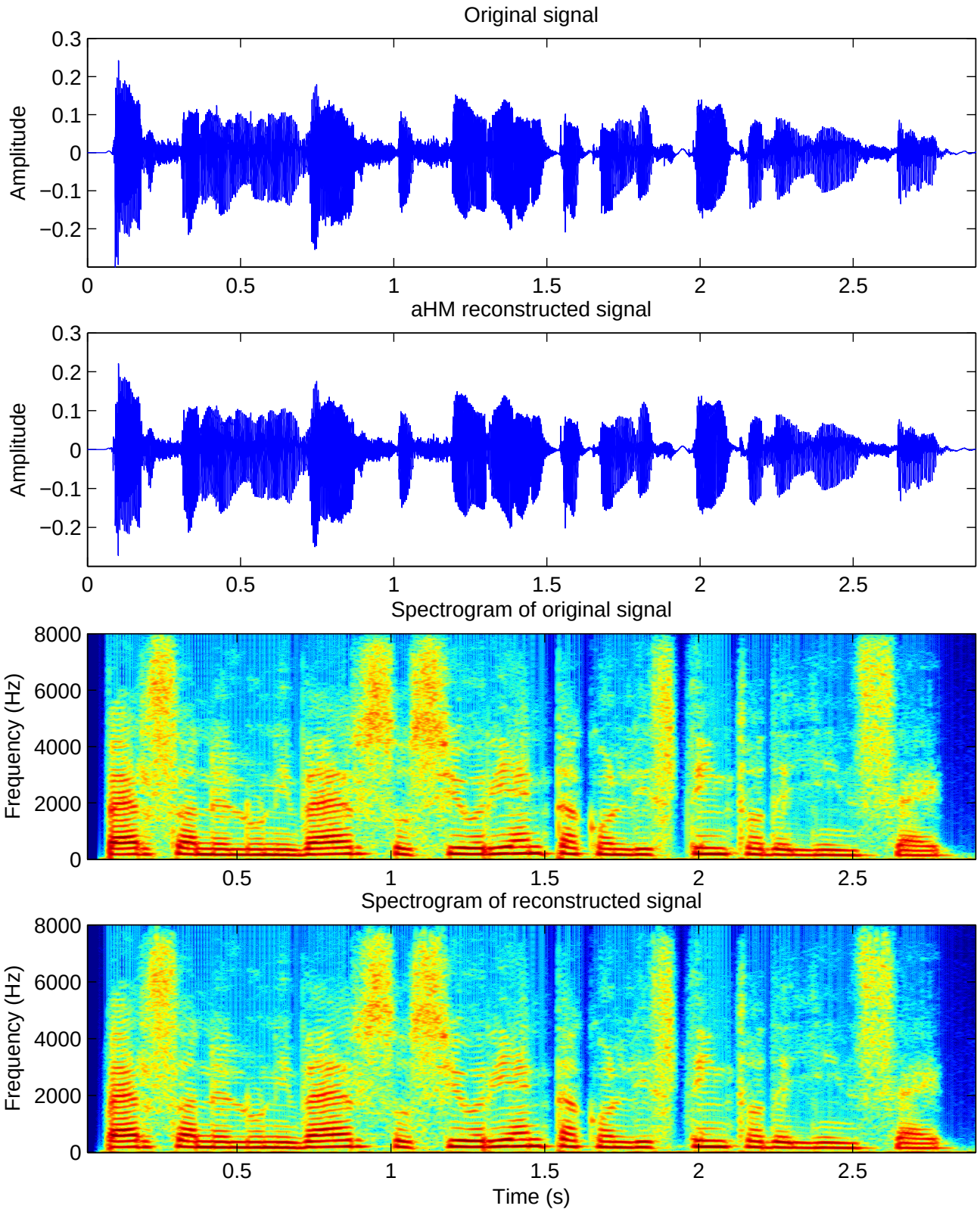


Figure 4.18: *adaptive Harmonic Model: Original signal (first panel) and reconstructed signal (second panel) along with their corresponding spectrograms (third and fourth panel) for a female speaker.*

4.11 Evaluation and Results

In this section, results will be shown on the resynthesis of the speech signal. Both objective and subjective measures will be presented. A comparison to the state-of-the-art will be discussed as well. Due to their availability, the following models will be selected for comparison to the adaptive models: Sinusoidal Model (SM), Harmonic + Noise Model (HNM), and STRAIGHT. In these experiments, a database of 32 speech utterances was used, including 16 male and 16 female speakers from 16 different languages: Greek, French, English, Spanish, Finnish, Chinese, Portuguese, Basque, Japanese, Italian, German, Korean, Russian, Arabic, Indonesian, and Turkish. All waveforms were sampled at 16 kHz.

In objective evaluation, the SRER is computed for the whole waveform, serving as an estimate of the total residual energy "missed" by each model. The higher the SRER value, the more information is captured by the model used. Since only the full-band models attempt to accurately reconstruct the speech signals, only those will be considered in the evaluation.

In subjective evaluation, a formal listening test has been conducted in order to measure perceptual quality. In this evaluation, all models are included: SM, HNM, STRAIGHT, aHM, aHNM, eaQHNM, eaQHM. The listening test has the form of Figure 4.21, and in most of the times, it was available on-line.

The parameters for the models were the following: for both pitch estimators, the pitch was estimated every 1 ms and their f_0 estimation limits were $[70, 220]$ Hz and $[120, 350]$ Hz for males and females, respectively. For AIR-f_0, which was used in the aHM model only, the analysis window is of Blackman type and its length is 3 local pitch periods, whereas the step size is pitch period synchronous. For the model parameter estimation, the analysis window is of Blackman type for aHM, and Hamming type for the eaQHM and SM. Their size is 3 times the local pitch period and the analysis step size was 2.5 ms, for *all* models. For the STRAIGHT method, the default parameters are used, and for the HNM, a synchronous analysis is considered, with a maximum voiced frequency of 5500 Hz.

4.11.1 Objective Evaluation

In objective analysis, the Signal-to-Reconstruction-Error Ratio (SRER) is chosen to measure the accuracy of the numerical representation between the original and the resynthesized speech. In Table 4.6, the mean and the standard deviation of the SRER for all utterances in our database are presented for both pitch estimators. It is clearly evident that quasi-harmonicity can capture more information of the underlying speech signal, with the same number of synthesis parameters. Figure 4.20 shows the first 16 frequency tracks in the analysis step for an utterance produced by Greek male

SRER Performance				
Model	Speakers			
	SWIPE		YIN	
	Males	Females	Males	Females
SM	18.6 (1.90)	18.6 (3.64)	14.3 (2.20)	16.2 (3.28)
aHM	23.9 (2.66)	18.9 (3.27)	23.9 (2.61)	19.9 (3.05)
eaQHM	34.5 (2.39)	30.9 (3.00)	34.4 (2.45)	30.7 (3.19)

Table 4.6: *Signal to Reconstruction Error Ratio values (dB) for all models on a database of* 32 *utterances (*16 *of male speakers,* 16 *of female speakers) using SWIPE and YIN pitch estimators. Mean and Standard Deviation are given.*

speaker, the local SRER for a sliding window of 30 ms, and the corresponding speech waveform for the two adaptive models. It should be noted that the overall SRER for the eaQHM is 34.67 dB whereas for the aHM is 25.60 dB for this sample, which contains both voiced and unvoiced areas. Intuitively, the eaQHM components in unvoiced speech attempt to locate "optimal" frequency tracks that collectively minimize the Mean-Square Error inside a frame. In this figure, it is obvious that in AIR-aHM all components are purely harmonic, and any slight fluctuation of the f_0 propagates in the higher harmonics. In the eaQHM however, the upper frequency components deviate from the multiples of the f_0 and their structure seems smoother. Based on the lower panel (time-varying SRER), it seems that the representation suggested by the eaQHM (middle panel) is more accurate compared to that one obtained by aHM (upper panel). Also, it should be mentioned that in our experiments, no manual refinement of the estimated f_0 is performed.

4.11.2 Subjective Evaluation

For perceptual quality evaluation, a formal listening test was designed. A part of it is currently available on-line[1]. The listeners were asked to evaluate the perceptual quality of the resynthesized speech compared to the original one, for

[1] http://www.csd.uoc.gr/~kafentz/listest/pmwiki.php?n=Main.EAQHM-LT

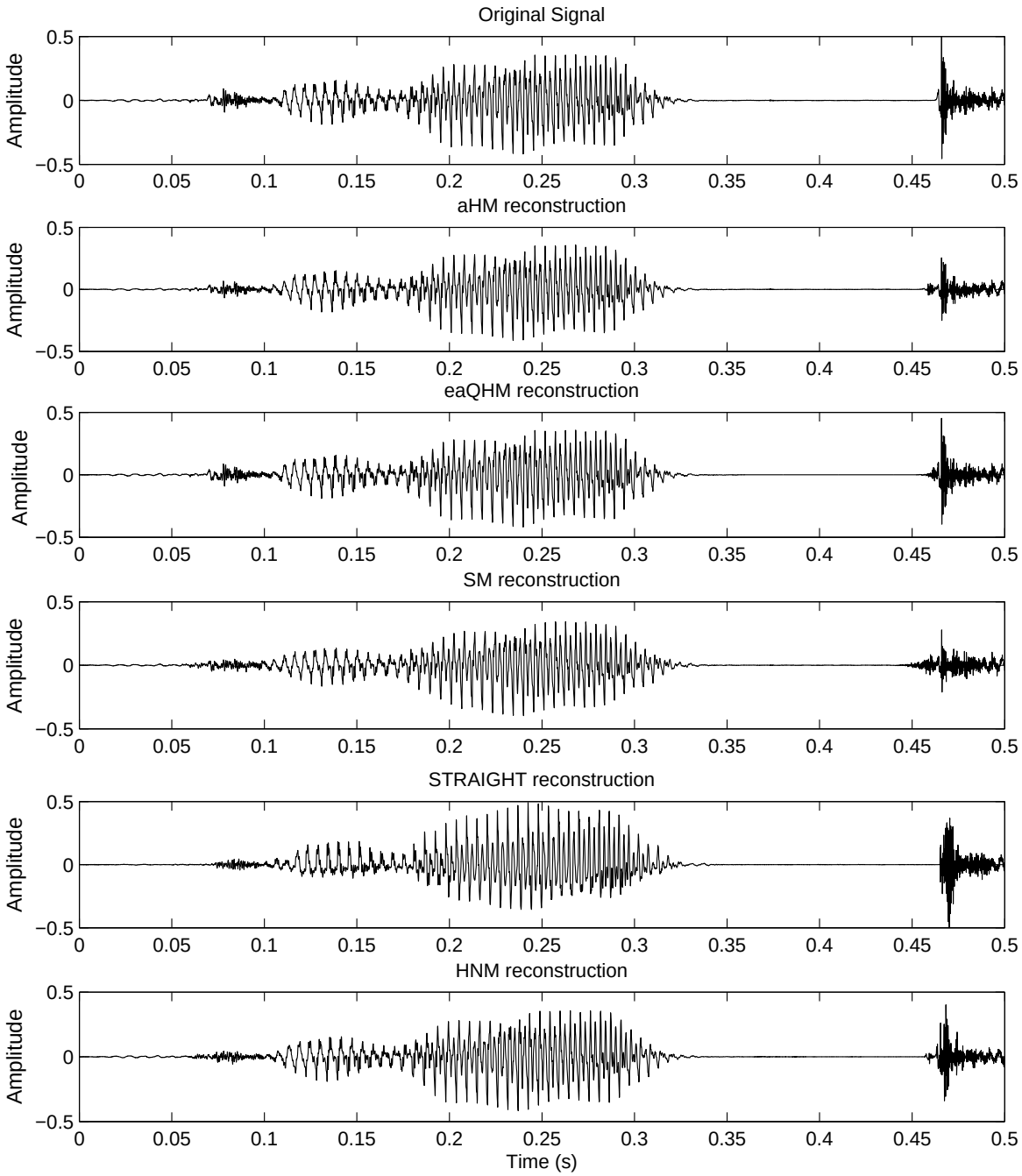


Figure 4.19: *Speech utterance (/krɔk^{h}ɛ/) in Korean language by a female subject. First panel: Original signal, Second panel: aHM reconstruction, Third panel: eaQHM reconstruction, Fourth panel: SM reconstruction, Fifth panel: STRAIGHT reconstruction, Sixth panel: HNM reconstruction.*

all different models. An $1-5$ scale was used in the evaluation according to the recommendation ITU-R BS [Ass03], with each scale being (1) "Very bad", (2) "Bad". (3) "Good", (4) "Very good", (5) "Perfect". The results from 34 listeners are depicted in Figures 4.22 and 4.23. In the same plot we show the 95% confidence interval. This shows that the obtained results are statistically significant. Please note that among these listeners, only 10 were familiar with signal processing and listening tests.

According to the listeners, the overall quality of all adaptive models is much higher than the state of the art. Moreover, perceptual differences between the adaptive models were not easy to find, and it was clearly stated that these differences are mostly present in the unvoiced parts, and especially in transients and sharp onsets of voiceless stop sounds (for example, in an aspirated velar /k/ in the utterance of Figure 4.19 by a Korean female). In general, it is acknowledged that the hybrid adaptive models (eaQHNM and aHNM) differ from the original in the unvoiced parts, where the modulated

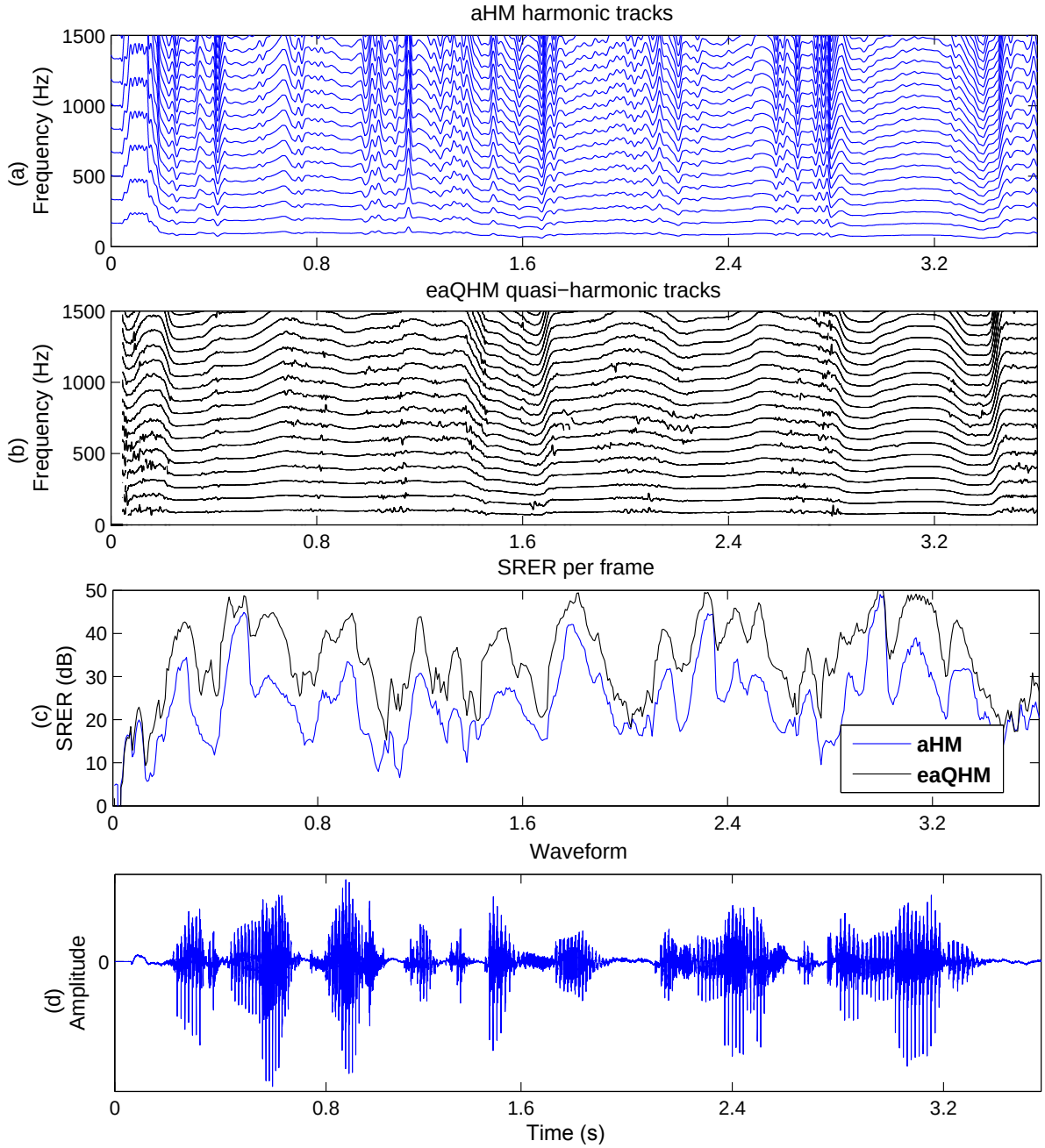


Figure 4.20: *Analysis data of a Greek male speaker for both adaptive models: (a) aHM tracks, (b) eaQHM tracks, (c) Local SRER for both models over time, (d) Speech waveform.*

noise representation does not attain the quality of the original signal. However, their quality is superior than the state-of-the-art (STRAIGHT, HNM, SM).

Moreover, by looking at the objective measures, it is interesting that although AIR-aHM performs significantly lower in terms of reconstruction, this does not translate to a respective quality degradation, as in the SM, where there is a substantial degradation, compared to the other two models. Finally, it is interesting that although the pitch estimators behave differently, both the adaptive models appear to be very stable in the reconstruction of output speech, as Table 4.6 shows.

4.12 Conclusions

In this chapter, we presented two hybrid and two full-band systems of analysis and synthesis of speech. The two hybrid systems have a deterministic and a stochastic component. The deterministic component is modeled either by the

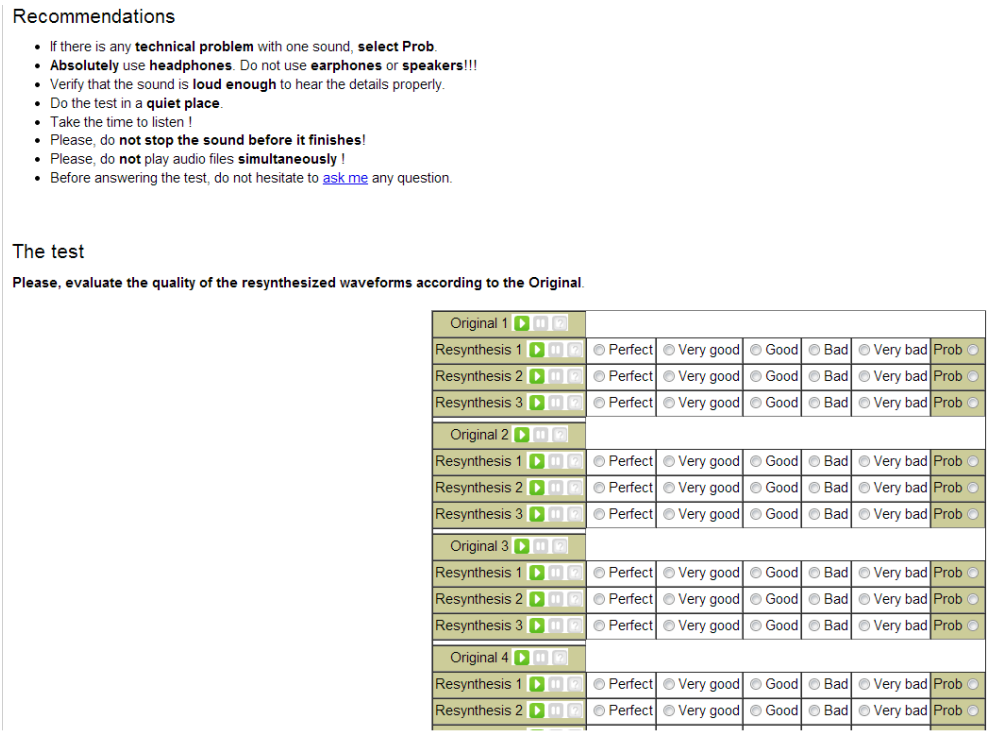


Figure 4.21: *Example of the listening test page.*

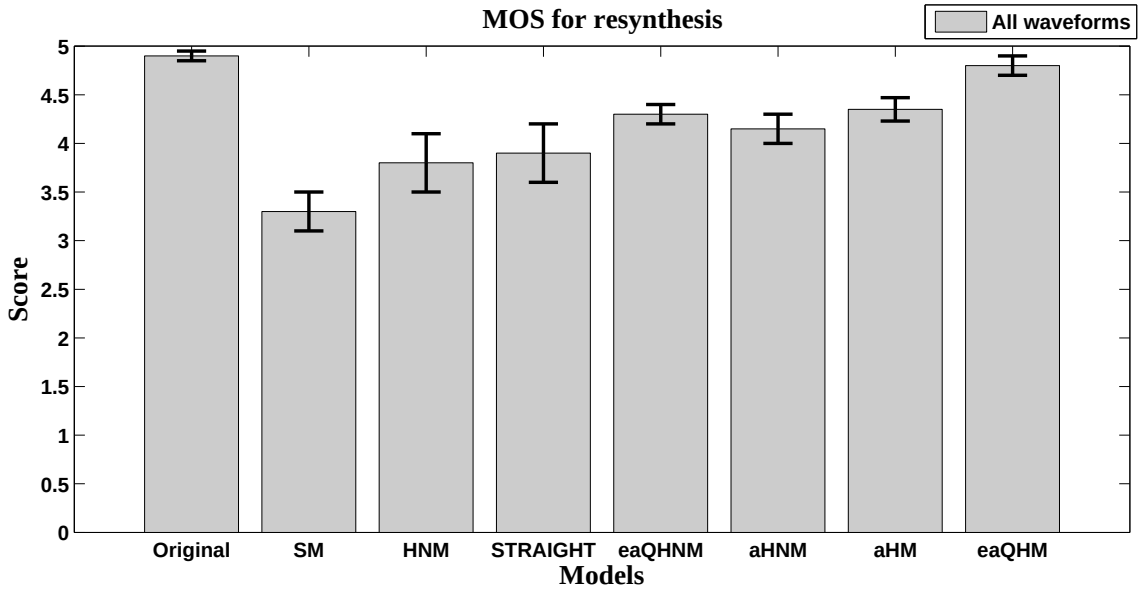


Figure 4.22: *Mean Opinion Score (MOS) of the resynthesis quality between the original recording and the reconstructions with all models, with the* 95% *confidence intervals.*

extended adaptive Quasi-Harmonic Model (eaQHM) or by the adaptive Harmonic Model (aHM). The former models speech as sum of AM-FM components that are quasi-harmonically related whereas the latter estimates its parameters using the theory of adaptivity but the resynthesis is purely harmonic, by iteratively refining a fundamental frequency estimate. The unvoiced parts of speech are modeled as time and frequency modulated Gaussian noise.

In addition, motivation for applying the models in the full-band of speech is proposed. These include (a) the questionable nature of the so-called maximum voiced frequency (MVF). More powerful analysis tools such as the Fan-Chirp Transform (FChT) have shown that there is structure in voiced speech segment up to the Nyquist frequency, and thus no noise component is necessary for high frequency representation, and (b) that fact that continuous, adaptive quasi-harmonic tracks have been proved to represent very accurately voiced and voiceless stop sounds as well as fricatives,

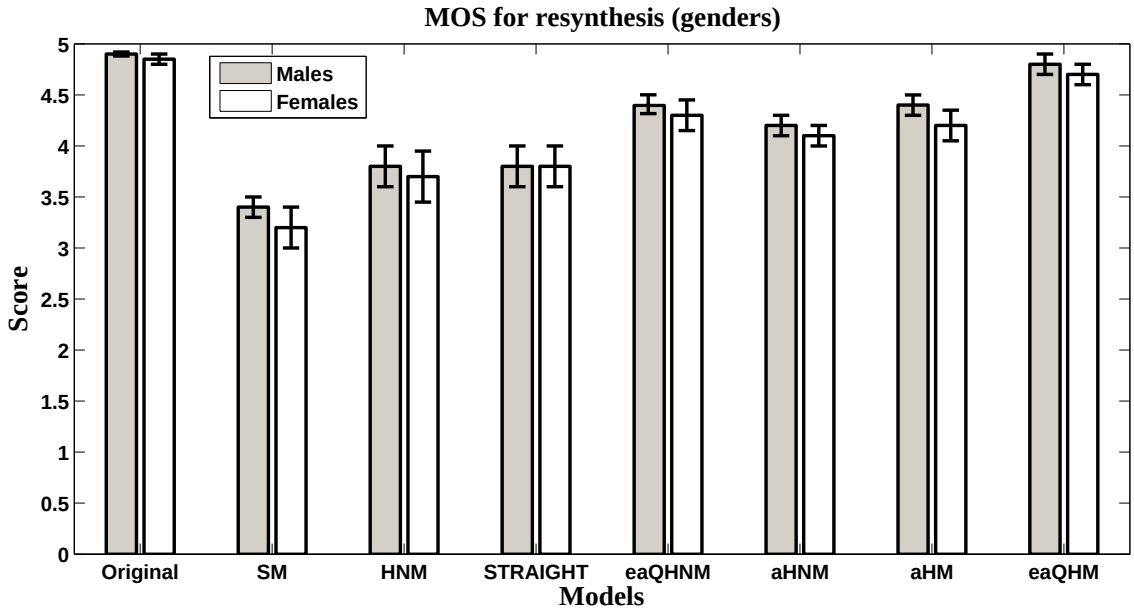


Figure 4.23: *Gender-based Mean Opinion Score (MOS) of the resynthesis quality between the original recording and the reconstructions with all models, with the* 95% *confidence intervals.*

provided that the frequency tracks span the spectrum adequately. The transient and stochastic nature of these sounds are well represented by quasi-harmonic tracks due to the frequency correction mechanism of QHM and the adaptation process. Suitable experiments for both stop and fricative sounds have confirmed this proposal, for different analysis step sizes.

Based on these observations, the full-band eaQHM has been developed, in parallel to the already developed aHM. The full-band eaQHM models all parts of speech as AM-FM sinusoids. Compared to the deterministic part of the eaQHNM, the full-band eaQHM starts from a strictly harmonic representation of speech and successively, through the frequency correction and adaptation mechanism, converges to quasi-harmonicity. The full-band aHM is identical to the deterministic part of the aHNM.

A comparison between available full-band models (SM, aHM, eaQHM) in terms of signal reconstruction is undertaken. The Signal-to-Reconstruction-Error Ratio (SRER) is a measure of closeness between the original and the reconstructed signal. It is shown that the eaQHM outperforms both aHM and SM in terms of SRER. From a perceptual point of view, a formal listening test revealed the superiority of the adaptive models (hybrid and full-band) compared to the state-of-the-art. Among all models, the eaQHM provides a transparent quality, indistinguishable from the original speech, whereas the aHM performs similarly well.

Chapter 5

Speech Modifications based on Adaptive Sinusoidal Models

Having analytically discussed the approaches, methods, and properties of the hybrid and full-band adaptive Sinusoidal Models, this chapter proposes methods for prosodic modifications of speech. First, modifications based on hybrid systems will be presented, following a similar approach as in milestone works [Ser89, Sty96]. This means that modifications must be applied on both components (deterministic and stochastic). Then, we focus more on modifications on the full-band systems which are proposed next, since the reconstruction quality of these models outperforms the corresponding of the hybrid models. Due to the purely deterministic (sinusoidal) representation of the full-band models, the most challenging part in modifications is the manipulation of the non-voiced parts of speech to attain a high perceptual quality.

A general flowchart for modifying speech using hybrid and full-band systems is given in Figure 5.1. We can observe that in full-band systems, the manipulation of a single component is performed. On the other hand, in hybrid systems both components (deterministic and stochastic) are to be manipulated differently.

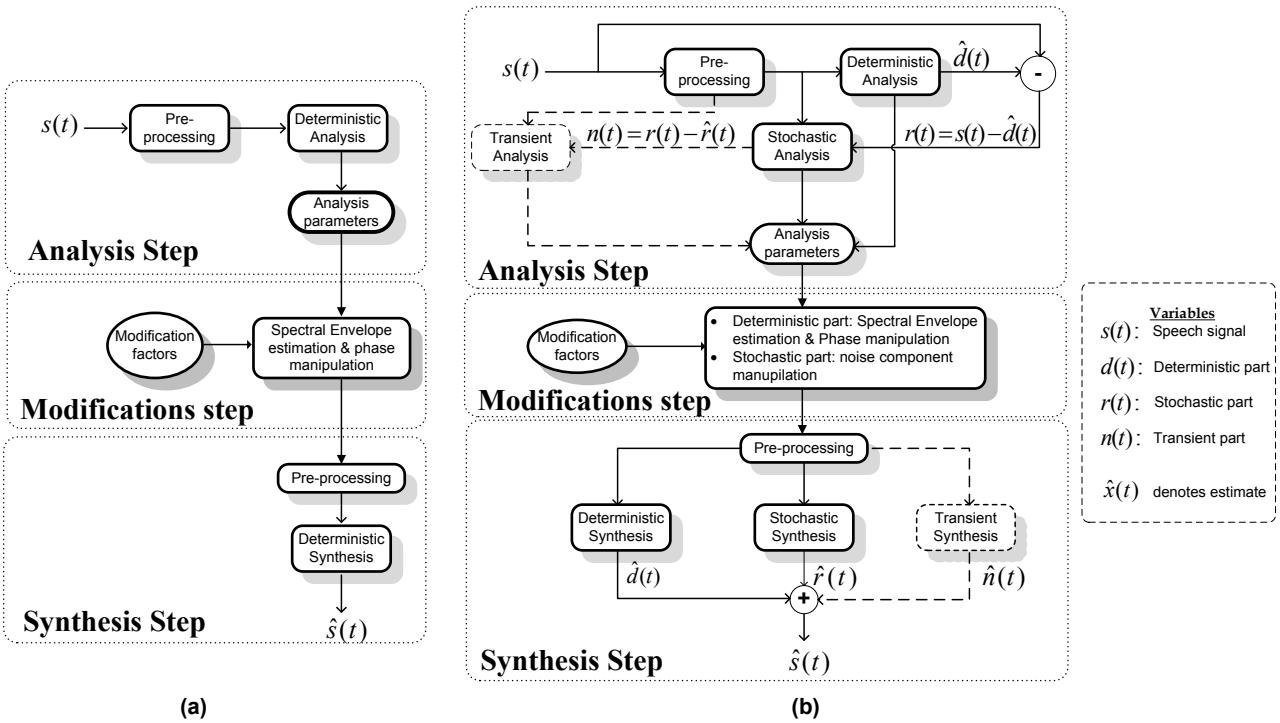


Figure 5.1: *A flowchart for the analysis, synthesis, and modifications part of (a) a full-band and (b) a general hybrid system framework. Upper panel: Analysis part. Middle part: Modifications part. Lower part: Synthesis part.*

Before going into the details, let us remind the purpose of time and pitch scaling in technical terms. The purpose of time-scale modification is to maintain the perceptual quality of the original speech signal while changing the apparent rate of articulation. The purpose of pitch-scale modification is to change the pitch contour of the original speech signal

while maintaining the apparent rate of articulation. The pitch contour (and thus the harmonics) should be shifted in frequency, and the formant structure should *not* be changed at a different rate than the rate of the input speech.

5.1 Time Scaling

When time scaling harmonically related sinusoids, the interpolation of instantaneous parameters is necessary to change the articulation rate. For instantaneous amplitudes and frequencies, this is performed via linear and spline interpolation, respectively. However, for the phase, the process is not straightforward because of its rotation due to the time advance across time instants. Therefore, it is proposed to remove this effect using the integral of kf_0 from the start of the signal, and obtain the *relative phase - RP* [DS13, KDRS13]. Thus, by assuming that the shape of the signal is changing smoothly, the phase values change also smoothly from one analysis time instant to the other. Then, the RP can be interpolated to obtain its continuous counterpart. Additionally, a spline or cubic interpolation is necessary such as its time derivative, the frequency, is still continuous. This way, the time scaled waveform is shape invariant.

5.1.1 Relative Phase

A theoretical basis for the relative phase notion follows next. Specifically, the linear phase term is sought to be removed during the resampling process in time scaling. Related work on linear phase removal has been suggested in other speech processing applications, such as concatenative speech synthesis [Sty01], speech transformations [Fed98], and speaker verification [LPY+12]. Let us consider a sinusoid

$$x_0(t) = \cos\left(2\pi \int_0^t f_0(u)du + \theta_0\right) \tag{5.1}$$

which we will consider as the reference sinusoid, and another sinusoid,

$$x_k(t) = \cos\left(2\pi k \int_0^t f_0(u)du + \theta_k\right), \quad k \in Z^+ \tag{5.2}$$

The instantaneous phases of the two sinusoids are

$$\phi_0(t) = 2\pi \int_0^t f_0(u)du + \theta_0, \quad \phi_k(t) = 2\pi k \int_0^t f_0(u)du + \theta_k \tag{5.3}$$

respectively. Let us consider that $\theta_0 = 0$, meaning that the time origin is set as the point where $\phi_0(0) = 0$. If we choose any analysis time instant t_a^i, the instantaneous phases become

$$\phi_0(t_a^i) = 2\pi \int_0^{t_a^i} f_0(u)du \quad , \quad \phi_k(t_a^i) = 2\pi k \int_0^{t_a^i} f_0(u)du + \theta_k \tag{5.4}$$

respectively. By changing variables, we get

$$\theta_k(t_a^i) = \phi_k(t_a^i) - k\phi_0(t_a^i) \tag{5.5}$$

Eq. (5.5) describes the relative phase in analysis time instants t_a^i. For time scaling, one would simply interpolate and time-scale $\theta_k(t_a^i)$ in successive analysis time instants to obtain $\theta_k'(t')$. Then, the instantaneous phase for the time scaled fundamental frequency would be

$$\phi_0'(t') = 2\pi \int_0^{t'} f_0'(u)du \tag{5.6}$$

and thus, the new instantaneous phase for the k^{th} harmonic would be

$$\phi_k'(t') = \theta_k'(t') + k\phi_0'(t') \tag{5.7}$$

5.1.2 Relative Phase Delay

For time scaling the instantaneous phase of quasi-harmonically related sinusoids, one would suggest to remove the integral of $f_k(t)$ (instead of $kf_0(t)$) from the k^{th} instantaneous phase track, also from the start of the signal, thus obtaining the relative phase which owns most of the randomness and all of the glottal pulse shape, and still changes smoothly from one time instant to the other. However, this approach from the well-known problem of *phase dispersion*, that is, while the

scaled signal has the same frequency content, the phases between the components change, resulting in a different wave shape. The phase dispersion problem has been addressed before in [QM92, Fed98], among others. A similar strategy will be followed here, using the concept of *relative phase delay*, first proposed in [Fed98].

The *phase delay* of the k^{th} sinusoid in the i^{th} analysis frame is defined as

$$\tau_k^i = \frac{\phi_k(t_a^i)}{\omega_k(t_a^i)} = \frac{\phi_k(t_a^i)}{2\pi f_k(t_a^i)} \tag{5.8}$$

where $\phi_k(t_a^i)$ is the phase value of the k^{th} sinusoid at analysis time instant t_a^i. The *relative phase delay* is defined as the difference between the phase lag of the k^{th} sinusoid and that of the first one, which corresponds to the fundamental frequency,

$$\Delta\tau_k^i = \tau_k^i - \tau_0^i \tag{5.9}$$

The signal is modified so as to preserve the relative phase delay. In a similar way, we define the phase delay at the synthesis time instants t_s^j as

$$\hat{\tau}_k^j = \frac{\hat{\phi}_k(t_s^j)}{\hat{\omega}_k(t_s^j)} = \frac{\hat{\phi}_k(t_s^j)}{2\pi f_k(t_s^j)} \tag{5.10}$$

In order to ensure shape invariance, equal relative phase delays at analysis and synthesis times are imposed:

$$\Delta\tau_k^i = \Delta\hat{\tau}_k^j \tag{5.11}$$

$$\tau_k^i - \tau_0^i = \hat{\tau}_k^j - \hat{\tau}_0^j \tag{5.12}$$

$$\hat{\phi}_k^j = (\hat{\tau}_0^j + (\tau_k^i - \tau_0^i))2\pi\hat{f}_k^j \tag{5.13}$$

The instantaneous phase $\phi_k(t_s^j)$ of the fundamental frequency at synthesis time instants is computed using the formula

$$\hat{\phi}_0^j = \hat{\phi}_0^{j-1} + \beta(\phi_0^i - \phi_0^{i-1}) \tag{5.14}$$

where β is the time scale factor, and ϕ_k^i is the unwrapped phase value at time instant i. Having the instantaneous phase values for all frequencies at the analysis time instants, the same strategy as in plain resynthesis is used, that utilizes frequency integration.

In hybrid systems, the above discussion is applied for the deterministic part, whereas for the stochastic part, a simple time stretching of the parametric noise envelope is sufficient. For sample-by-sample lattice filtering representation of noise, a simple interpolation of the reflection coefficients should be performed.

5.2 Pitch Scaling

In pitch scaling, the estimation of a new set of amplitude, frequency, and phase values is necessary due to pitch shifting. These values can be obtained by estimating the so-called *amplitude and phase envelopes* in the spectral domain. Spectral estimation is a field of study that has received increased attention because of the variety of its applications (voice conversion [GRC12], word recognition [BA09], speech recognition [WM05], speaker verification [HKS+12], speaker identification [RR95], to name a few), and many algorithms are available to achieve it in a robust manner, such as cepstrum-based techniques [GR90, CM96], AR models [EJM91, TKMI94, MG76], and multi-frame analysis [SK03, TT08].

Since pitch scaling requires the estimation of amplitudes in the new, shifted frequencies, the Discrete All-Pole method [EJM91] is used in this work for both models. For the phase, a simple approach is suggested for the aHM-based systems, which involves the computation and the interpolation of the *relative phases*, as mentioned in time scaling sections. For the eaQHM-based systems, the notion of *relative phase delays* is also employed, in order to minimize phase dispersion.

5.2.1 Amplitude Estimation

Amplitude estimation is performed via an all-pole technique, called the Discrete All-Pole method (DAP). This method utilizes a discrete version of the Itakura-Saito (IS) distortion measure as its error criterion, instead of a time-domain criterion that most of other all-pole models use. The IS error measure is given by

$$E_{IS} = \frac{1}{N}\sum_{m=1}^{N}\frac{X(\omega_m)}{\hat{X}(\omega_m)} - \log\frac{X(\omega_m)}{\hat{X}(\omega_m)} - 1 \tag{5.15}$$

where $X(\omega_m)$ is the given discrete spectrum defined at N frequency points, and $\hat{X}(\omega_m)$ is the all-pole model spectrum evaluated at the frequencies $\omega_m \in [0, f_s/2]$, where f_s is the sampling frequency. This method manages to overcome the well-known limitations of linear prediction [Mak75] and produces better fitting of spectra that are represented with a small set of discrete values, such as the case of sinusoidal models.

The DAP method works iteratively to solve a nonlinear set of equations, in order to converge to a global minimum. The order P of the method does not differ from the empirical choice that is employed in most all-pole methods, that is

$$P = \frac{f_s}{1000} + 2 \tag{5.16}$$

where f_s is in Hertz. The DAP method is used for spectral envelope estimation for all models. More details on DAP can be found in [EJM91].

5.2.2 Phase Estimation

For the aHM-based systems, the relative phase is once again used. After estimating and interpolating the relative phases, the integral of the shifted frequencies is added back, to obtain the instantaneous phases of each harmonic. Mathematically, each frequency track is modified as

$$kf_0(t) \rightarrow \rho k f_0(t) \tag{5.17}$$

where ρ is the pitch scale factor. Next, the relative phases are computed as in Eq. (5.5) and interpolated over time. Finally, the pitch-scaled frequencies are integrated and added back to the continuous relative phases, yielding the instantaneous phases of the pitch-scaled frequencies

$$\phi'_k(t) = \theta_k(t) + 2\pi \int_0^t \rho k f_0(u) du \tag{5.18}$$

The relative phase values allow the reconstruction of the shape of the signal, using the reference phase $\phi_0(t_a^i)$ in a synchronous reconstruction. For the purposes of pitch-scale modification, the f_0 track can be changed without any recomputation of the phase, because if the RPs are kept constant, the waveform will stretch or shrink accordingly without any other change.

For the eaQHM-based systems, an extension of the time scaling algorithm presented earlier is suggested. Provided that the pitch scaling factor ρ is constant over the duration of a frame, the phase variation induced by pitch-scaling is equivalent to that produced by time scaling using the same factor. Thus, Eq. (5.14) is changed into

$$\hat{\phi}_0^j = \hat{\phi}_0^{j-1} + \rho(\phi_0^i - \phi_0^{i-1}) \tag{5.19}$$

and based on this phase track, the rest of the instantaneous phase values at synthesis time instants t_s^j (which can be the same as the analysis ones, or different, if both time and pitch scale are applied) are generated using the relative phase delays as

$$\hat{\phi}_k^j = (\hat{\tau}_0^j + (\tau_k^i - \tau_0^i))2\pi\rho\hat{f}_k^j \tag{5.20}$$

The instantaneous phases are computed once again using the integration scheme of the analysis.

In hybrid systems, the above discussion is applied for the deterministic part, whereas for the stochastic part, no modification is performed.

For the modifications part, we will focus only on the full-band systems, since in hybrid systems, only the stochastic part is different, and its modification methods are well-known [Sty96].

5.3 Technical Definitions

As mentioned earlier, the purpose of time-scale modification is to maintain the perceptual quality of the original speech signal while changing the apparent rate of articulation. The pitch contour (and thus the harmonics) should be stretched or compressed in time, and the formant structure should be changed at a slower or faster rate than the rate of the input speech, but otherwise not modified. For an arbitrary time-scale modification, the time t in the original signal is mapped to a time t' in the modified signal. For that, a mapping function referred to as *the time-scale warping function* is

defined:

$$D(t) = \int_0^t \beta(\tau) d\tau \tag{5.21}$$

where $\beta(\tau) > 0$ is the time-varying time-scaling rate. When $\beta(\tau) > 1$, then the articulation rate is slowed down, whereas the opposite happens when $\beta(\tau) < 1$. Note that for a constant rate $\beta(\tau) = \beta$, then the time-scale warping function is reduced to a linear function of time, i.e. $D(t) = \beta t$.

Moreover, the purpose of pitch-scale modification is to change the pitch contour of the original speech signal while maintaining the apparent rate of articulation. The pitch contour (and thus the harmonics) should be shifted in frequency, and the formant structure should *not* be changed at a different rate than the rate of the input speech. For an arbitrary pitch-scale modification, the input $f_0(t)$ contour is mapped to a different one, $f_0'(t) = \rho(t) f_0(t)$ in the modified signal, where $\rho(t)$ is the pitch-scale factor function. When $\rho(t) > 1$, then the pitch increases, whereas the opposite happens when $\rho(t) < 1$. Note that for a constant $\rho(t) = \rho$, the pitch modification is invariant throughout the waveform.

5.4 Speech Modifications based on the aHM system

5.4.1 Time-Scale Modification Scheme

In the adaptive Harmonic model context, the parameters should be transformed in the way described next. Note that in an analysis window centered at t_a^i, the instantaneous components $\{a_k^i, f_0^i\}$, are known. From these, we can compute their continuous counterparts, which are the instantaneous amplitudes $A_k(t) = |a_k(t)|$ and frequencies $f_0(t)$, obtained by interpolating a_k^i and f_0^i, respectively. Then, the time-scaled waveform, $\hat{s}_{TS}(t')$ is given by:

$$\hat{s}_{TS}(t') = \sum_{k=-K}^{K} A_k'(t') e^{j\phi_k'(t')} \tag{5.22}$$

where $A_k'(t')$ and $\phi_k'(t')$ are computed using the following way:

1. The instantaneous amplitudes are time-scaled:

$$A_k'(t') = A_k(D^{-1}(t)) \tag{5.23}$$

2. In order to compute $\phi_k'(t')$, it is first necessary to compute the time-scaled frequencies. The instantaneous frequencies in the modified signal at time t' correspond to the instantaneous frequency in the original signal at time $D^{-1}(t')$:

$$k f_0'(t') = k f_0(D^{-1}(t')) \tag{5.24}$$

where $D^{-1}(t)$ is the inverse time-scale warping function.

3. Finally, to obtain a shape-preserving waveform, the relative phase (RP) values of the analysis need to be time-scaled. Therefore, we first compute the continuous time-scaled RPs, $\angle\tilde{a}_k(t)$, from the corresponding values, $\angle\tilde{a}_k^i$. For this, the RP is first computed by extracting the integral of the frequency from the phase information at analysis time instant t_a^i, as in Eq.(5.25):

$$\angle\tilde{a}_k^i = \angle a_k^i - k\phi_0(t_a^i) \tag{5.25}$$

Then, the RP values are interpolated, thus obtaining $\angle\tilde{a}_k'(t')$, as:

$$\angle\tilde{a}_k'(t') = \angle\tilde{a}_k(D^{-1}(t')) \tag{5.26}$$

and finally, the integrated time-scaled frequency is added back to the interpolated RP values:

$$\hat{\phi}_k'(t') = \angle\tilde{a}_k'(t') + \int_0^{t'} 2\pi k f_0'(u) du \tag{5.27}$$

This way the waveform retains its shape for voiced parts, i.e. the time-scaling is *shape-invariant*.

The time-scaling algorithm for aHM is given in Algorithm 2.

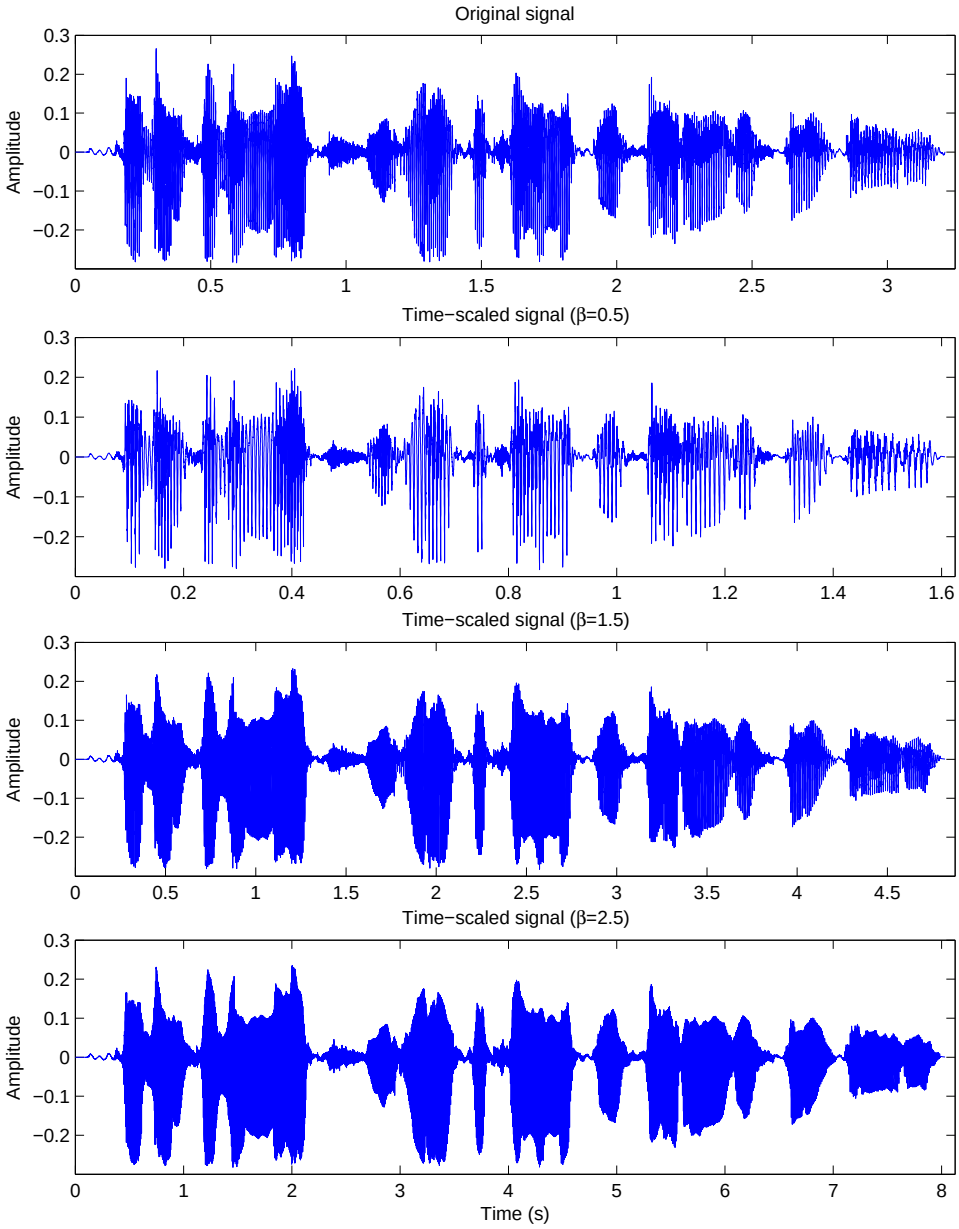


Figure 5.2: *Adaptive Harmonic Model time scaling: Original signal (first panel) and time-scaled signals (lower panels) for factors of* $0.5, 1.5$ *and* 2.5*, respectively.*

Algorithm 2 *Time-scaling using aHM*

Require: A set of parameters per analysis time instant (frame): $\{A_k(t_a^i), f_0(t_a^i), \angle a_k(t_a^i)\}$

Interpolate successive t_a^i to obtain t'

Interpolate $A_k(t)$ to obtain $A'_k(t')$ using Eq. (5.23)

Interpolate $kf_0(t)$ to obtain $kf'_0(t')$ using Eq. (5.24)

Estimate relative phases $\angle\tilde{a}_k^i$ using Eq. (5.25)

Interpolate relative phases to obtain $\angle\tilde{a}'_k(t')$ using Eq. (5.26)

Estimate instantaneous phase $\hat{\phi}'_k(t')$ using Eq. (5.27)

Time scaled speech

Synthesize $\hat{s}_{TS}(t')$ using harmonic synthesis

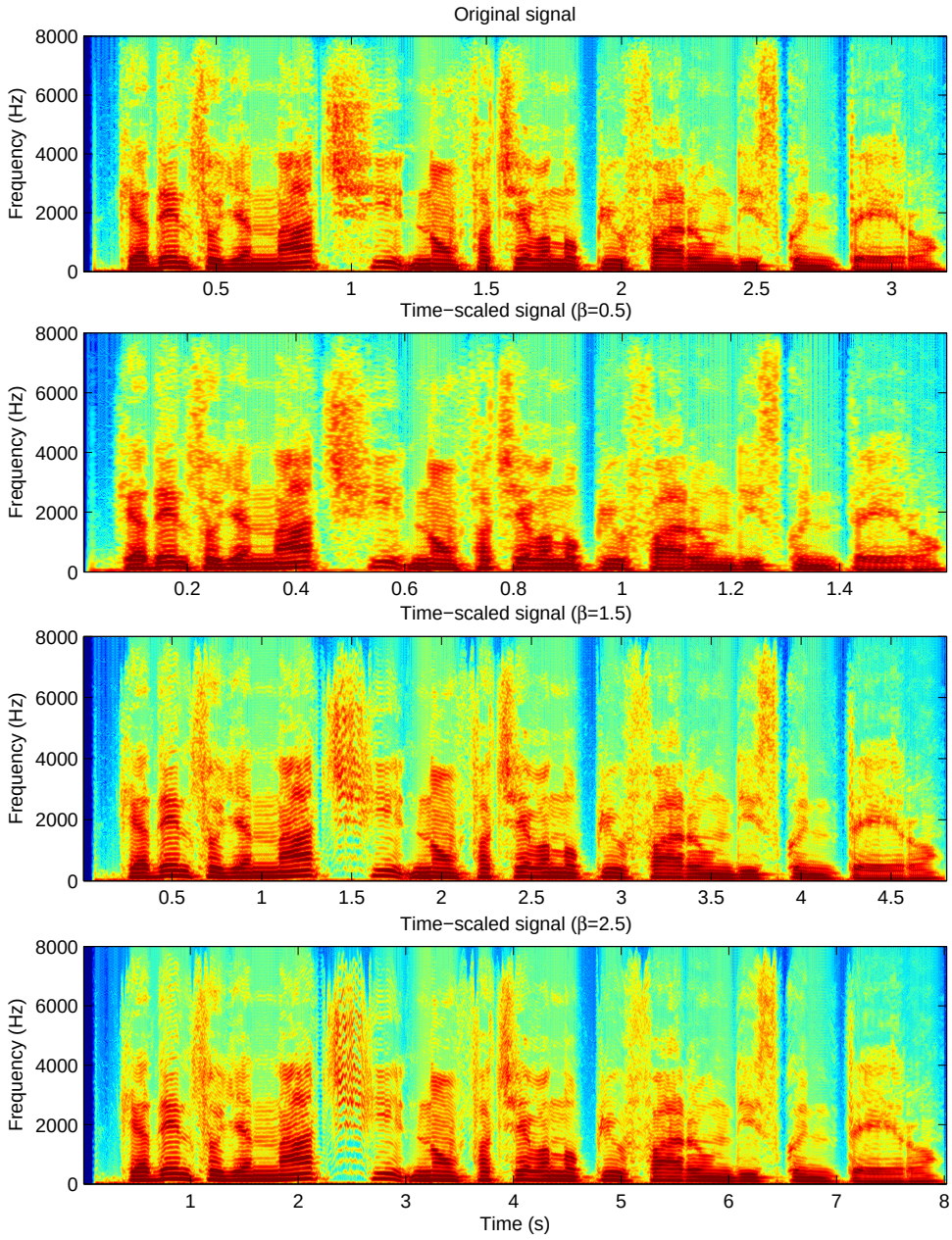


Figure 5.3: *Adaptive Harmonic Model time scaling spectra: Original signal spectrum (upper panel) and time-scaled signals spectra (lower panels) for factors of* $0.5, 1.5$ *and* 2.5*, respectively.*

Example of Application

In Figure 5.2 the shape invariance property of the aHM is clearly illustrated for different time-scale factors, 0.5, 1.5, and 2.5. A male speaker waveform is presented. In Figure 5.3 the corresponding waveforms of Figure 5.2 is depicted in the frequency domain.

5.4.2 Pitch-Scale Modification Scheme

In the adaptive Harmonic model context, the parameters should be transformed in the way described next. Note that in an analysis window centered at t_a^i, the instantaneous components $\{a_k^i, f_0^i\}$, are known. From these, we can compute their continuous counterparts, which are the instantaneous amplitudes $A_k(t) = |a_k(t)|$ and frequencies $f_0(t)$, obtained by interpolating a_k^i and f_0^i, respectively. Then, the pitch-scaled waveform, $s_{PS}(t)$, for a constant pitch-scale factor is

given by:

$$\hat{s}_{PS}(t) = \sum_{k=-K}^{K} A'_k(t) e^{j\phi'_k(t)} \tag{5.28}$$

where $A'_k(t)$ and $\phi'_k(t)$ are computed using the following way:

1. In order to compute $\phi'_k(t)$, it is first necessary to compute the pitch-scaled frequencies. Thus the new frequencies are given by:
$$kf_0(t) \leftarrow \rho k f_0(t) \tag{5.29}$$

2. The instantaneous amplitudes at analysis time instants t_a^i, $A'_k(t_a^i)$, are computed from sampling the spectral envelope at the corresponding frequencies $\rho k f_0$:
$$A'_k(t_a^i) = DAP(t_a^i, \rho k f_0) \tag{5.30}$$
where $DAP(t_a^i, f)$ is the Discrete All-Pole-based envelope constructed around time instant t_a^i. Then, the k^{th} instantaneous amplitude is linearly interpolated across successive time instants.

3. Then, the instantaneous phase should be re-computed. For this, the RP is first computed by extracting the integral of the initial fundamental frequency from the phase information at analysis time instant t_a^i, as in Eq. (4.42). Then, the RP values are interpolated, thus obtaining $\angle\tilde{a}_k(t)$, and finally, the integrated pitch-scaled frequency is added back to the interpolated RP values:
$$\hat{\phi}'_k(t) = \angle\tilde{a}_k(t) + \frac{2\pi}{f_s}\int_0^t \rho k f_0(u)du \tag{5.31}$$

Algorithm 3 summarizes the previous steps:

Algorithm 3 *Pitch-scaling using aHM*

Require: A set of parameters per analysis time instant (frame): $\{A_k(t_a^i), f_0(t_a^i), \angle a_k(t_a^i)\}$
Compute pitch-scaled frequencies using Eq. (5.29)
Compute the spectral envelope $DAP(t_a^i)$ around time instant t_a^i.
Sample the spectral envelope at the corresponding frequencies $\rho k f_0$ using Eq. (5.30)
Interpolate instantaneous amplitudes over successive time instants t_i^a to obtain $A'_k(t)$
Estimate relative phases $\angle\tilde{a}_k^i$ using Eq. (5.25)
Interpolate relative phases to obtain $\angle\tilde{a}'_k(t)$ using Eq. (5.26)
Estimate instantaneous phase $\hat{\phi}'_k(t)$ using Eq. (5.31)

Pitch scaled speech

Synthesize $\hat{s}_{PS}(t)$ using harmonic synthesis

Example of Application

In Figure 5.4, the aHM pitch shifting is demonstrated for different pitch-scale factors, 0.5, 1.5, and 2.0. A female speaker waveform is presented. In Figure 5.5 the corresponding waveforms of Figure 5.4 are depicted in the frequency domain.

5.5 Speech Modifications based on the eaQHM system

5.5.1 Time-Scale Modification Scheme

In the eaQHM context, the parameters should be transformed in the way described next. Let us first assume that in an analysis window centered at t_i, the instantaneous components $\{A_k(t_a^i), f_k(t_a^i), \phi_k(t_a^i)\}$, are known, and the component trajectories have been computed, i.e. $A_k(t), f_k(t), \phi_k(t)$, within this frame and up to the center of the next frame, t_a^{i+1}. Then, time scaling requires the following steps to be performed.

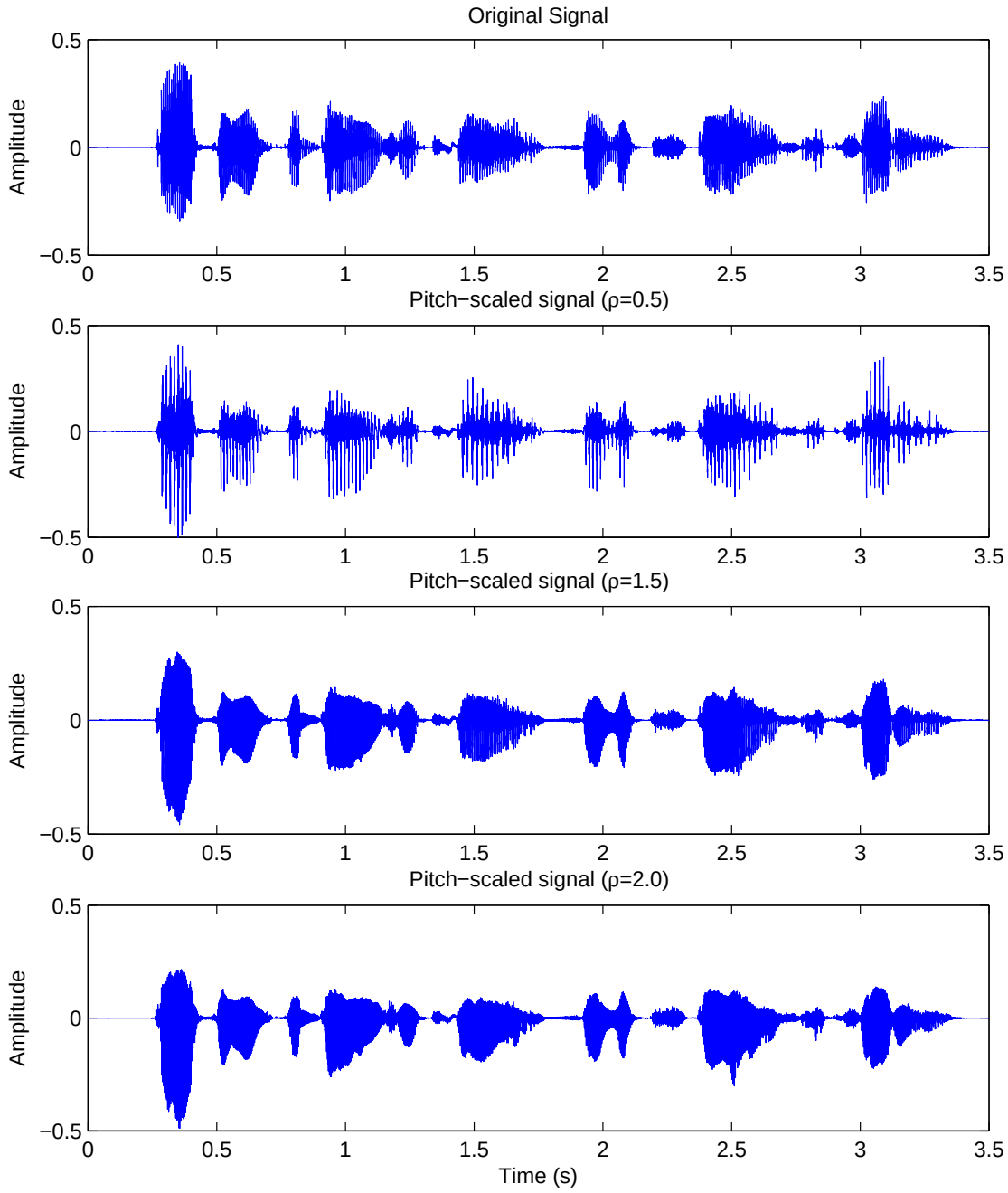


Figure 5.4: *adaptive Harmonic Model pitch scaling: Original signal (upper panel) and pitch-scaled signals (lower panels) for factors of* $0.5, 1.5$ *and* 2.0*, respectively.*

Let $\hat{s}(t)$ denote the AM-FM decomposed signal:

$$\hat{s}(t) = \sum_{k=-L}^{L} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \tag{5.32}$$

1. The instantaneous amplitudes are time-scaled:

$$A'_k(t') = A_k(D^{-1}(t')) \tag{5.33}$$

2. The instantaneous frequencies in the modified signal at time t' correspond to the instantaneous frequency in the original signal at time $D^{-1}(t')$:

$$f'_k(t') = f_k(D^{-1}(t')) \tag{5.34}$$

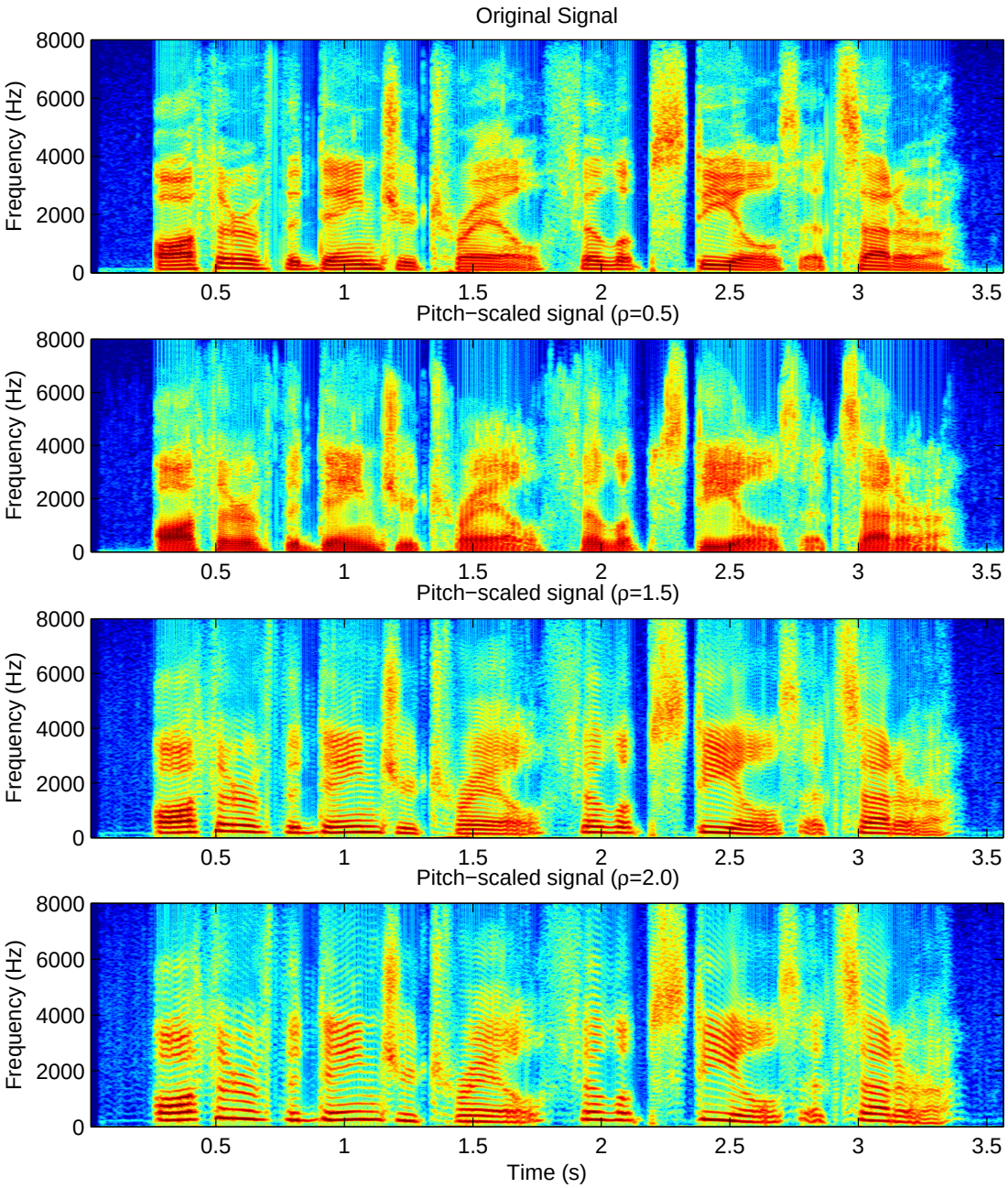


Figure 5.5: *adaptive Harmonic Model pitch scaling spectra: Original signal spectrum (upper panel) and pitch-scaled signals spectra (lower panels) for factors of* $0.5, 1.5$ *and* 2.0*, respectively.*

3. The instantaneous phase values of the fundamental phase for the synthesis frames are computed:

$$\phi_0(t_s^j) = \phi_0(t_s^{j-1}) + \beta(\phi_0(t_a^i) - \phi_0(t_a^{i-1})) \tag{5.35}$$

4. The instantaneous phase values of the k^{th} frequency for the synthesis frames are computed:

$$\phi_k(t_s^j) = (\hat{\tau}_0^j + (\tau_k^i - \tau_0^i))2\pi f_k^j \tag{5.36}$$

5. The instantaneous phase curves $\phi_k(t')$ are computed via instantaneous frequency integration:

$$\phi_k'(t') = \phi_k(t_s^j) + \int_0^{t'} (2\pi f_k'(u) + c(u))du \tag{5.37}$$

Thus, its time-scaled version, $s_{TS}(t')$, for a constant time-scale factor is given by:

$$\hat{s}_{TS}(t') = \sum_{k=-L}^{L} A'_k(t') e^{j\phi'_k(t')} \tag{5.38}$$

Algorithm 4 presents pseudocode for time-scaling using eaQHM:

Algorithm 4 *Time-scaling using eaQHM*

Require: A set of parameters per analysis time instant (voiced frames): $\{A_k(t_a^i), f_0(t_a^i), \phi_k(t_a^i)\}$

Interpolate successive t_a^i to obtain t'
Interpolate $A_k(t)$ to obtain $A'_k(t')$ using Eq. (5.33)
Interpolate $f_k(t)$ to obtain $f'_k(t')$ using Eq. (5.34)
Estimate $\phi_0(t_s^j)$ using Eq. (5.35)
Estimate $\phi_k(t_s^j)$ using $\hat{\tau}_0^j$ and Eq. (5.36)
Estimate $\phi'_k(t')$ using Eq. (5.37)

Time scaled speech

Synthesize $\hat{s}_{TS}(t')$ using sinusoidal synthesis

Example of Application

In Figure 5.6, the eaQHM-based time-scaling is demonstrated for different time-scale factors, 0.5, 1.5, and 2.5. A male speaker waveform is presented. The shape invariance property is clearly illustrated. In Figure 5.7 the corresponding waveforms of Figure 5.6 is depicted in the frequency domain.

5.5.2 Pitch-Scale Modification Scheme

Let us first assume that in an analysis window centered at t_a^i, the instantaneous components $\{A_k(t_a^i), f_k(t_a^i), \phi_k(t_a^i)\}$, are known, and the component trajectories have been computed, i.e. $A_k(t), f_k(t), \phi_k(t)$, within this frame and up to the center of the next frame, t_a^{i+1}. Then, pitch scaling requires the following steps to be performed.

1. In order to compute $\phi'_k(t)$, it is first necessary to compute the pitch-scaled frequencies. Thus the new frequencies are given by:
$$f_k(t) \leftarrow \rho f_k(t) \tag{5.39}$$

2. The instantaneous amplitudes at analysis time instants t_a^i, $A'_k(t_a^i)$, are computed from sampling the spectral envelope at the corresponding frequencies ρf_k:
$$A'_k(t_a^i) = DAP(t_a^i, \rho f_k) \tag{5.40}$$
where $DAP(t_a^i, f)$ is the Discrete All-Pole-based envelope constructed around time instant t_a^i. Then, the k^{th} instantaneous amplitude is linearly interpolated across successive time instants.

3. The instantaneous phase values of the fundamental phase for the synthesis frames are computed:
$$\phi_0(t_s^j) = \phi_0(t_s^{j-1}) + \rho(\phi_0(t_a^i) - \phi_0(t_a^{i-1})) \tag{5.41}$$

4. The instantaneous phase values of the k^{th} frequency for the synthesis frames are computed:
$$\phi_k(t_s^j) = (\hat{\tau}_0^j + (\tau_k^i - \tau_0^i))2\pi\rho f_k^j \tag{5.42}$$

5. The instantaneous phase curves $\phi'_k(t)$ are computed via instantaneous frequency integration:
$$\phi'_k(t) = \phi_k(t_s^j) + \int_0^t (2\pi\rho f_k(u) + c(u))du \tag{5.43}$$

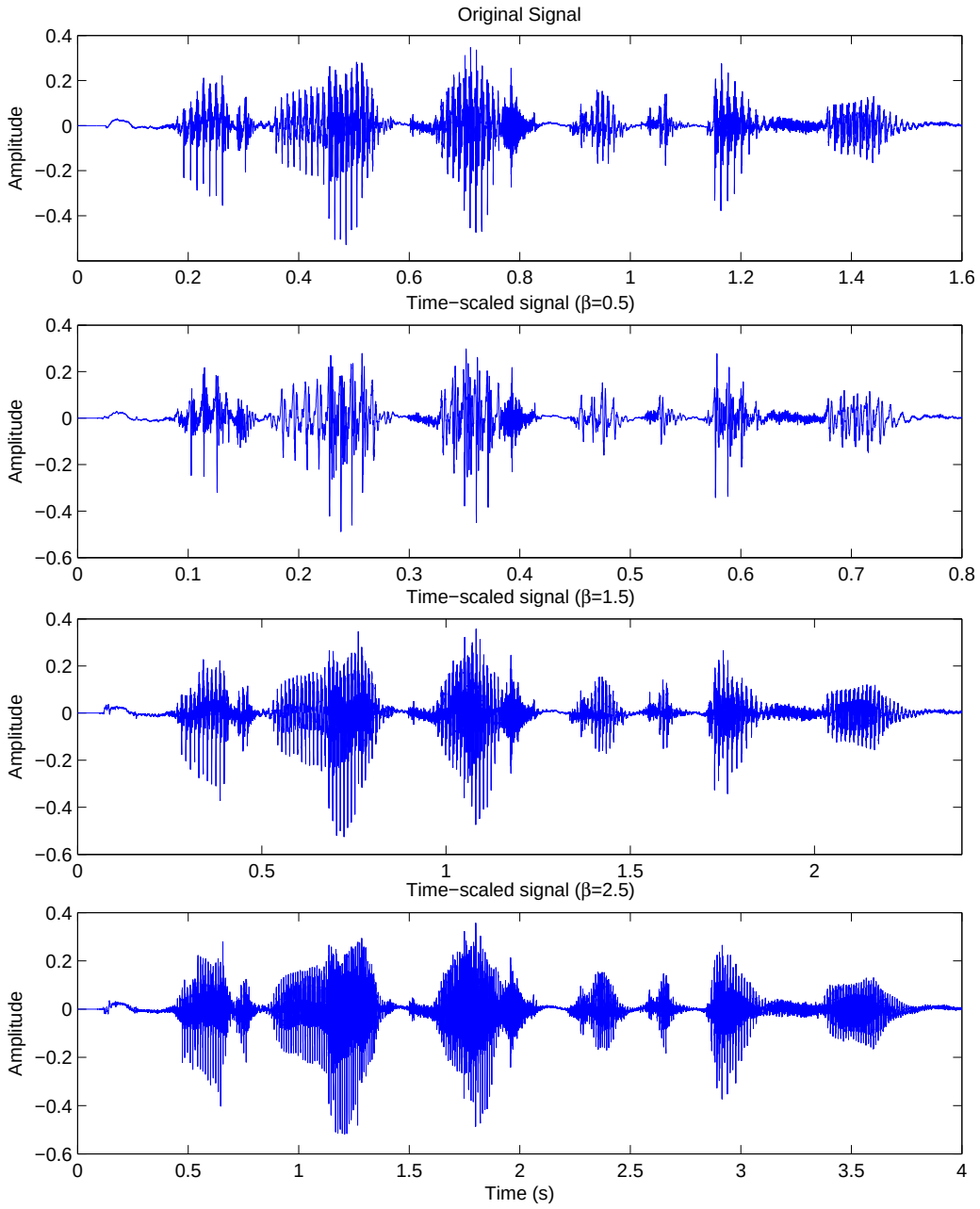


Figure 5.6: *extended adaptive Quasi-Harmonic Model time scaling: Original signal (upper panel) and time-scaled signals (lower panel) for factors of* $0.5, 1.5$ *and* 2.5*, respectively.*

Thus, its pitch-scaled version, $s_{PS}(t)$, for a constant pitch-scale factor is given by:

$$\hat{s}_{PS}(t) = \sum_{k=-L}^{L} A'_k(t) e^{j\phi'_k(t)} \tag{5.44}$$

Algorithm 5 summarizes the previous steps:

Example of Application

In Figure 5.8, the eaQHM pitch shifting is demonstrated for different waveforms and different pitch-scale factors, 0.5, 1.5, and 2.0. A female speaker waveform is presented. In Figure 5.9, the corresponding waveforms of Figures 5.8 are depicted in the frequency domain.

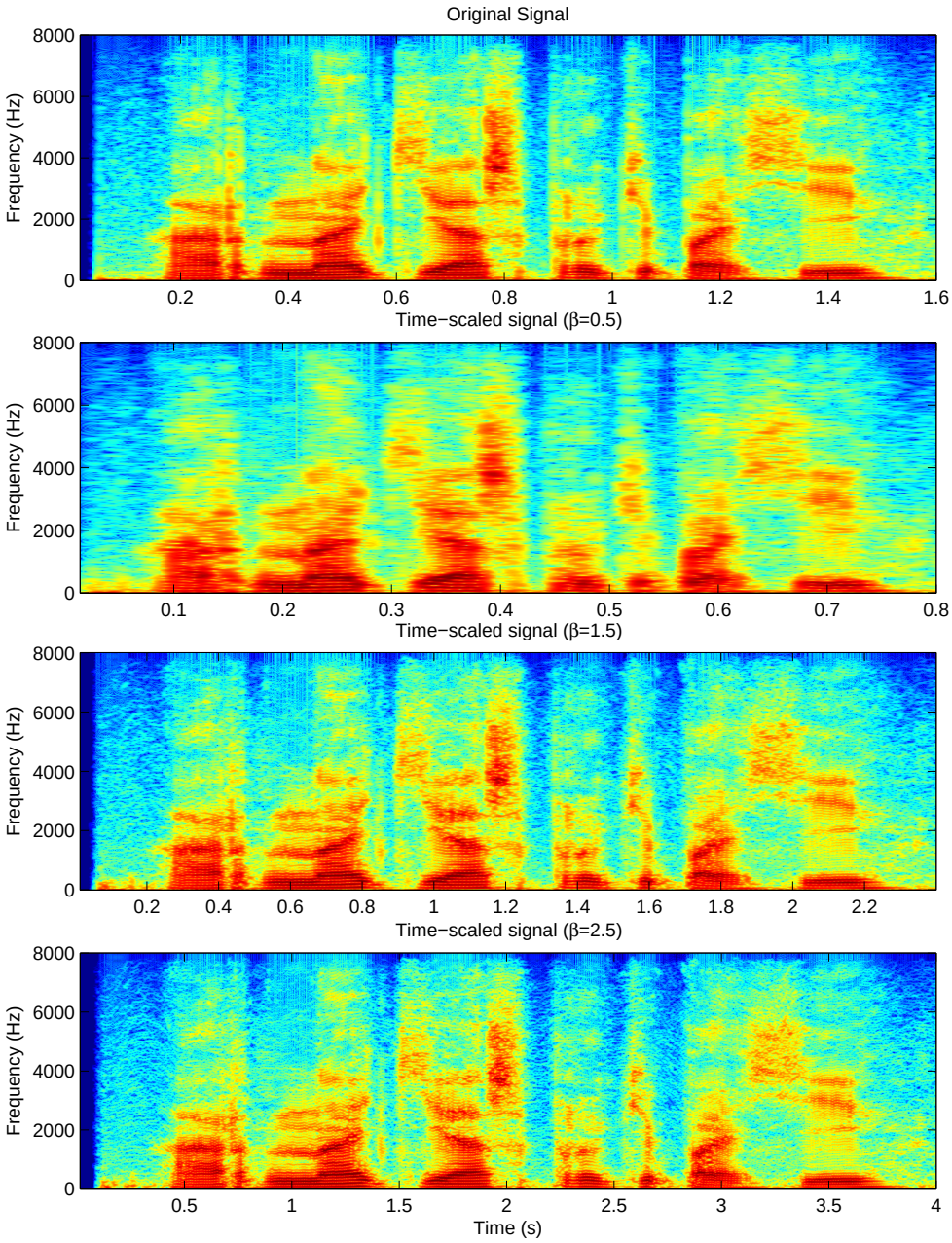


Figure 5.7: *extended adaptive Quasi-Harmonic Model time scaling spectra: Original signal spectrum (upper panel) and time-scaled signals spectra (lower panel) for factors of* $0.5, 1.5$ *and* 2.5*, respectively.*

5.6 Evaluation and Results

In this section, subjective evaluations of speech modifications using the adaptive models will be presented and compared to the state-of-the-art. Due to their availability, the following methods will be selected for comparison to the aSMs: HNM, STRAIGHT, and WSOLA for time scaling, and HNM and STRAIGHT for pitch-scaling. In this experiment, a database of 32 speech utterances was used, including 16 male and 16 female speakers from 16 different languages: Greek, French, English, Spanish, Finnish, Chinese, Portuguese, Basque, Japanese, Italian, German, Korean, Russian, Arabic, Indonesian, and Turkish. All waveforms were sampled at 16 kHz. Examples are available at http://www.csd.uoc.gr/~kafentz/listest.html.

Algorithm 5 *Pitch-scaling using eaQHM*

Require: A set of instantaneous parameters per analysis time instant: $\{A_k(t_a^i), f_k(t_a^i), \phi_k(t_a^i)\}$

Compute pitch-scaled frequencies $\rho f_k(t)$ using Eq. (5.39)
Compute the spectral envelope $DAP(t_a^i)$ around time instant t_a^i.
Sample the spectral envelope at the corresponding frequencies ρf_k using Eq. (5.40)
Interpolate instantaneous amplitudes over successive time instants t_i^a to obtain $A'_k(t)$
Estimate $\phi_0(t_s^j)$ using Eq. (5.41)
Estimate $\phi_k(t_s^j)$ using $\hat{\tau}_0^j$ and Eq. (5.42)
Estimate $\phi'_k(t)$ using Eq. (5.43)

Pitch scaled speech

Synthesize $\hat{s}_{PS}(t)$ using sinusoidal synthesis

5.6.1 Time-scaling

The time-scale modification factors were selected to be $0.5, 0.8, 1.2, 1.5, 2.0$, and 2.5, which are typical values for moderate speech prosodic modifications. For the HNM, the maximum voiced frequency is fixed to 5500 Hz, and the analysis is pitch synchronous. The analysis window size is two local pitch periods. The order of the AR filter for the noise part is set to 20. The parameters of the aHM and eaQHM are the ones described in the previous sections. For the WSOLA, an analysis window length of 15 ms is used. A tolerance variable Δ (a tolerance factor on the desired time-warping function to ensure signal continuity at segment joins) of 7 ms is selected, which according to [VR93], usually produces high-quality time-scaled speech. For the STRAIGHT method, the default parameters were used.
In general, the participants acknowledged the proposed methods natural. The aHM samples were considered of slightly higher quality than the eaQHM samples, and in general, both better than STRAIGHT. Also, common artefacts, such as "metallic" quality, chorusing, or musical noise do not appear more than in state-of-the-art methods. Although the models are simple, they are shown to perform similarly or even better than the - more complex - HNM or STRAIGHT methods, for time scale modifications, especially in voiced parts of speech, where the well-known problem of *lack of presence* present in the HNM is addressed. Note that the HNM decomposes speech into a deterministic and a stochastic component. As such, although it shares the harmonicity assumption in its deterministic component, it handles its stochastic part differently (modulated noise). In our WSOLA samples, a step effect in the amplitude of the time-scaled speech was observed, that led to audible artefacts. No such artefacts were present in the aHM or the eaQHM time-scaled samples. Finally, it should be noted that although the WSOLA technique performs quite close to the adaptive models and is much faster, it does not provide higher level representations of speech (i.e. spectral envelopes).

5.6.2 Pitch-scaling

The pitch-scale modification factors were selected to be 0.5, 0.8, 1.2, 1.5, and 2.0, which are typical values for speech. For both genders of speakers, a minimum and maximum value for the pitch estimation were posed: $f_{0(min,max)} = (120, 300)$ Hz for females, and $f_{0(min,max)} = (70, 200)$ Hz for males. For the HNM, the maximum voiced frequency is fixed to 5500 Hz, and the analysis is pitch synchronous. The analysis window size is set to two local pitch periods. The order of the AR filter for the noise part is set to 20. For the STRAIGHT, default parameters were used. The parameters of the adaptive models are the ones described in the previous sections. In general, first informal listenings acknowledged that common artefacts, such as "metallic" quality, chorusing, or musical noise do not appear in adaptive sinusoidal models more than they do in the state-of-the-art methods in hand. Once again, the aHM samples were considered slightly better than the eaQHM ones. However, for large pitch scale factor and due to the sinusoidal nature of the representations, the spectral area between the distant successive sinusoids manifests a sense of *tenseness* in voice. This is apparent especially in unvoiced parts, where the number of sinusoids is not high enough to represent these parts well. It should be noted that both the HNM and STRAIGHT use some kind of noise component, whereas adaptive models do not. The HNM uses time and frequency modulated noise to represent unvoiced parts and high-frequency components of voiced parts, whereas the STRAIGHT method uses all-pass filters to compensate for the buzz timbre of minimum-phase vocal tract filter.

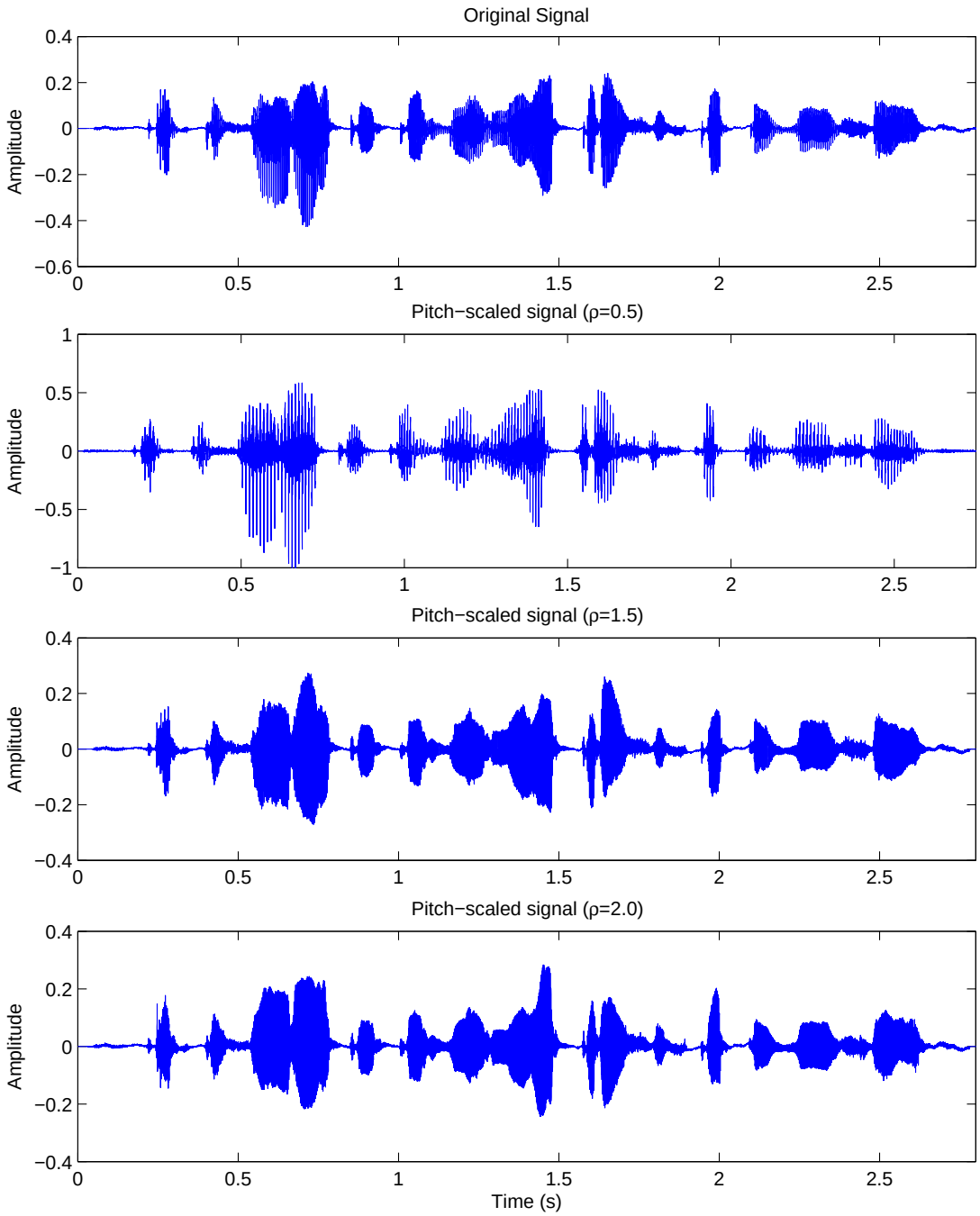


Figure 5.8: *extended adaptive Quasi-Harmonic Model pitch scaling: Original signal (upper panel) and pitch-scaled signals (lower panel) for factors of* $0.5, 1.5$ *and* 2.0*, respectively.*

5.7 Conclusions

In this chapter, speech modifications based on the adaptive Sinusoidal Models were presented. Modifications are governed by simple rules, making them highly attractive. For the aHM, each harmonic track is modified separately, using the notion of relative phase. Time scaling is achieved by interpolating the instantaneous amplitude, frequency, and relative phase tracks over time, and the instantaneous phase is computed by adding the integral of the time scaled instantaneous frequency from the start of the signal back to the interpolated relative phase track. Shape invariant waveforms are generated using this approach. For the eaQHM, since the model is quasi-harmonic, the notion of relative phase delays is employed to produce shape invariant waveforms. The instantaneous amplitude and frequency tracks are interpolated over time and for the instantaneous phase, the relative phase delays at the analysis time instants are forced to be valid at the synthesis time instants. This way, the well-known phase dispersion problem is minimized. In pitch scaling, the same principles are followed, only that the frequency tracks are resampled in new frequency values, and the corresponding

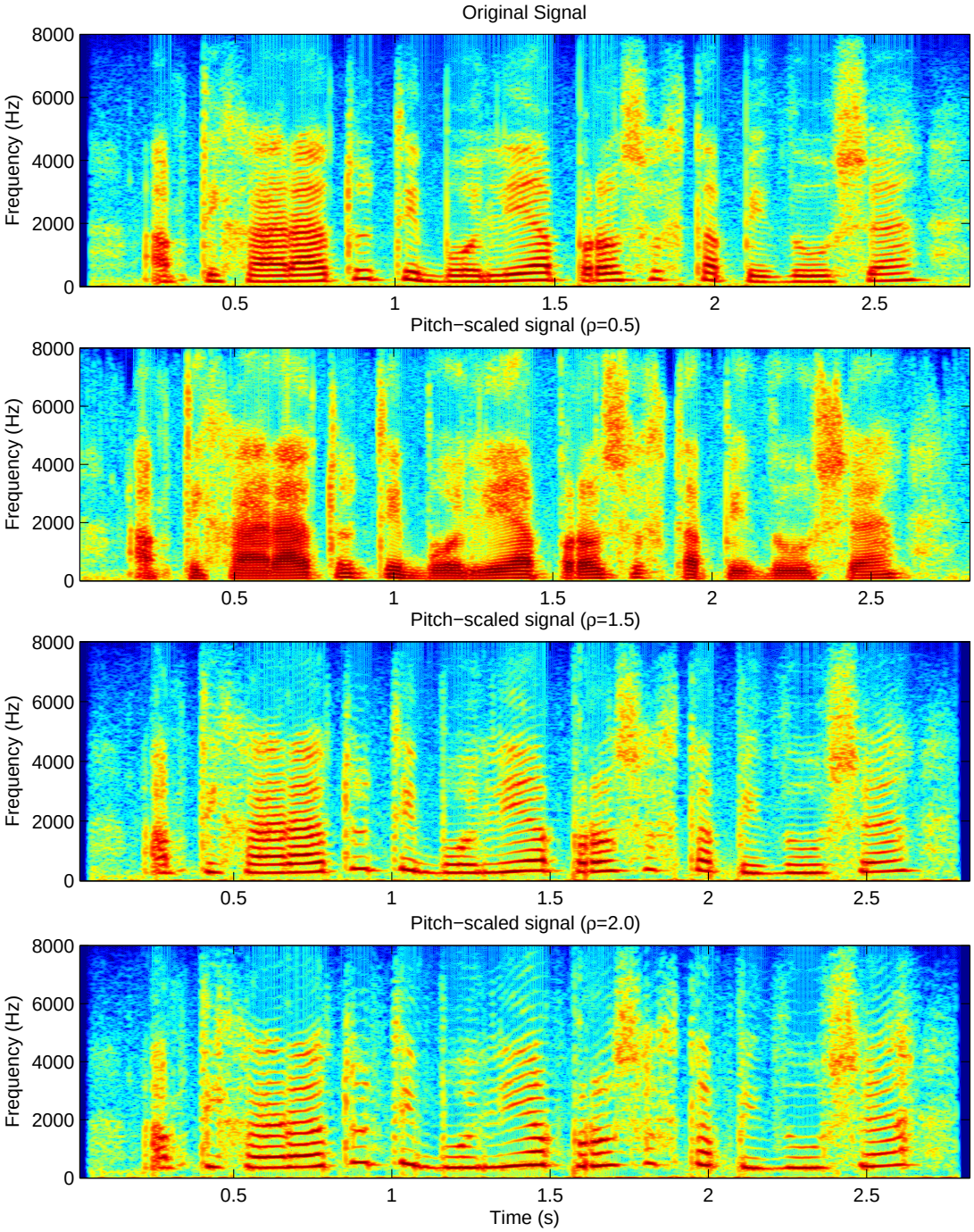


Figure 5.9: *extended adaptive Quasi-Harmonic Model pitch scaling spectra: Original signal spectrum (upper panel) and pitch-scaled signals spectra (lower panel) for factors of* $0.5, 1.5$ *and* 2.0*, respectively.*

amplitudes are obtained from a Discrete-All-Pole based spectral envelope. The instantaneous phases are computed using the relative phase and the relative phase delay for the aHM and the eaQHM, respectively in order to maintain waveform shape. Algorithms and examples are given for each model.

Part III

Applications

Chapter 6

Adaptive Sinusoidal Modelling of Musical Instrument Sounds

6.1 Introduction

Musical instrument sounds are notoriously challenging to represent with a single model because different instruments may feature radically different characteristics, such as sharp attacks, quasi-periodic (or quasi-stationary) oscillations, noise, and inharmonicity [FR98]. The attack is the most salient perceptual feature of musical instrument sounds that listeners use in dissimilarity judgments [GG77, SC64, Kru89]. It is well known that much of the characteristic quality of many musical sounds derives from the attack [IK93], although the harmonic structure of the quasi-stationary oscillations (when there is a steady state) is also clearly important [Han95]. Percussive sounds produced by plucking strings (such as harpsichords, harps, and the *pizzicato* playing technique) or striking percussion instruments (such as drums, idiophones, or the piano) feature a sharp onset with highly nonstationary oscillations that die out very quickly, called transients [BH09]. The reed in woodwind instruments presents a highly nonlinear behavior that also results in attack transients [Fle99], while the stiffness of piano strings results in a slightly inharmonic spectrum [BH09].

Ideally, the representation of musical instrument sounds should rely on a few parameters that capture most information in an intuitive way. Additionally, high fidelity is essential in synthesis [QM02], transformation [VZA06, SB98, CR12], model conversion [RVR07, VRR07], and estimation of various features, be they perceptual such as pitch [dK02, CH08] or acoustical such as size, material, fingering, among others [MIT+10, Abe13, TS01, MIK+12]. The musical instrument sound model can be used in a variety of related problems, from onset detection [BDA+05] and segmentation [PGV97, CR10, PV00] to musical instrument classification [MM99, HPD03], recognition [EK00, ERD06], identification [Bro99, KM95], and conversion [Gio09]. Parametric representations typically compare the signal to be represented with a template, such as wavelets [KM88], sinusoids [MQ86, SS90], exponentially damped sinusoids [HVL+05], or a dictionary of atoms [MZ93, SSDR08]. In general, sinusoids render a compact representation of quasi-stationary oscillations that is perceptually close to the original recording and whose parameters encode intuitive information. However, the quality of the representation depends on the parameter estimation techniques and on how the underlying model represents temporal variation in the parameters over the course of the sound.

There have been several proposals about how to estimate the parameters of sinusoidal models. The earliest models [MQ86, SS90] relied on straightforward estimation techniques, such as peak-picking and parabolic interpolation. Over the years, researchers have proposed improvements in parameter estimation [NMB07], partial tracking [LMR07, DGR93], and time-frequency resolution [AF95, KGV78, FF06]. Nonetheless, there are some intrinsic limitations to the representation of musical instrument sounds by quasi-stationary sinusoids, such as poor noise and transient modeling, smearing of the sharpness of attack, and limited temporal resolution. The standard sinusoidal model (SM) suffers from the time-frequency uncertainty principle when estimating the parameters because long analysis windows blur the temporal resolution to improve the frequency resolution and vice-versa. Moreover, the SM uses constant amplitude and frequency values inside the analysis window because the SM assumes that the sound is relatively stable locally. Consequently, the representation notoriously fails to capture short-time temporal variations such as transients or sharp percussive onsets [CKMS13]. Therefore, the SM tends to use a separate model for noise [SS90, CKD+13] and for transients [VM00, Dau11, DDS01, Dau06].

Exponentially Damped Sinusoidal Models (EDS) have been gaining popularity as the template (or basis functions) to represent many types of signals [KT82, UT96], including audio and musical instrument sounds [NHD98b, BBD02, HVL+05] and more recently for audio coding [DBR13]. Essentially, the EDS uses stationary sinusoids modulated by a real exponential function inside the analysis window to represent the partials. There are parameter estimation algorithms

based on iterative analysis-by-synthesis [Goo97] and subspace-based methods [RK89, VHPR96], sometimes referred to as "high-resolution methods" because they do not suffer from the time-frequency uncertainty of the FFT. Proponents of the EDSM to model musical instrument sounds claim that the EDS outperforms the SM [JH02] not only because the parameter estimation techniques are more powerful and robust, but also because the temporal envelope of the sinusoids is suitable to represent percussive sounds. However, the frequency of the partials in the EDS is still constant inside each analysis frame, and the beginning of the frame has to coincide with the onsets to take advantage of the shape of the temporal envelope. Synchronization of the frame boundaries with percussive onsets requires additional steps prior to modeling, namely onset detection and classification into percussive or not. The EDS notoriously requires additional partials [NHD98b] when the onsets are not synchronized. Another important drawback is the constant frequency of each partial inside the analysis frame.

Adaptive sinusoidal models have been applied in speech [KPRS12, PRS11] and musical instrument sounds [RÖ6, CKMS13] to address some of the issues with traditional sinusoidal models. Adaptation of sinusoidal partials *inside* the analysis window frees the algorithm from the inherent temporal limitation of the STFT and quasi-stationary sinusoids, allowing representation of changes in a temporal scale smaller than the hop size. In a preliminary study [CKMS13], we used an adaptive sinusoidal model dubbed the "extended adaptive Quasi-Harmonic Model" (eaQHM) [KPRS12] to represent percussive musical instrument sounds. We showed that the eaQHM outperformed the representation obtained with sinusoidal models that use stationary partials. Previously, the eaQHM had been used to model the speech counterparts of percussive audio sounds, namely *stop sounds*, outperforming quasi-stationary sinusoidal models [KRS13]. In the eaQHM, adaptation results from the iterative projection of the original waveform onto *nonstationary* basis functions that are locally adapted to its time-varying characteristics, rendering a flexible model capable of representing sudden changes such as transients or sharp onsets.

In this work, we model a large number of musical instrument sounds from different families with sinusoidal modeling algorithms and compare their modeling accuracy, defined as how much information the parametric model captures quantitatively. We use the "signal to reconstruction error rate" (SRER) as a measure of modeling accuracy. We compare both the local SRER, measured over a window just before the onset, and the global SRER, comprising the whole duration of the sound. The algorithms considered are the standard sinusoidal modeling algorithm [MQ86] (SM), exponentially damped sinusoids [DBR13] (EDS) using ESPRIT [RK89] for parameter estimation, and the eaQHM. First we show that adaptation significantly improves both the local and global SRER. Then we compare the modeling accuracy varying the size of the analysis window and the number of sinusoidal partials. We show that the eaQHM outperforms both the state of the art (EDS) and the baseline model (SM) for most instrumental families in all experiments with with the same number of analysis parameters as EDS and the same number of synthesis parameters as the SM.

In the next section, we describe the analysis and synthesis stages of eaQHM, focusing on adaptation. Then, we present the experimental setup, describe the musical instrument sound database used in this work and the analysis parameters. Next, we explain the experiments we performed, we present the results and evaluate the performance of the SM, EDS, and eaQHM in modeling musical instrument sounds. Finally, we discuss the results and present conclusions and perspectives for future work. The raw data will be available for reproducibility of the results along with Matlab code to obtain the figures and tables presented here. Sound examples and further information can be found at `http://www.csd.uoc.gr/~kafentz/listest/pmwiki.php?n=Main.AdaptiveSinMus`.

6.2 Experimental Setup

The aim of the experiment is to compare the modeling accuracy of the SM, EDS, and eaQHM for a broad range of musical instrument sounds, including percussive and nonpercussive. In this work, modeling accuracy is measured by the local and global SRER, calculated using Eq. (6.1).

$$SRER = 20\log_{10}\frac{\sigma_{x(t)}}{\sigma_{x(t)-\hat{x}(t)}} \tag{6.1}$$

The local SRER is measured over a window just before the onset to evaluate the smearing of the attack also known as pre-echo, a very common artifact among sinusoidal models that use quasi-stationary sinusoids. The global SRER measures the overall modeling accuracy, taking the whole sound into account.

The modeling accuracy depends on the number of partials K and the window size L for the sinusoidal algorithms. Traditionally, partials are not supposed to vary much inside the analysis window and thus are modeled with quasi-stationary sinusoids whose parameters are averaged over the window, changing between windows according to the step size. On the other hand, supposing that each sinusoid captures one partial, there is a minimum number of sinusoids required to represent the oscillatory energy in musical instrument sounds. Thus we will present a comparison of the SRER as a function of K and L for the SM, EDS, and eaQHM. First we describe the musical instrument sounds modeled and the selection of parameter values for the algorithms.

6.2.1 The Musical Instrument Sounds Used

In total, 90 musical instrument sounds were used in this work[1]. Table 6.1 lists the musical instruments divided in six classes, *Brass*, *Woodwinds*, *Strings*, *Percussion*, *Popular*, and *Keyboard*. The recordings were chosen to represent the range of musical instruments commonly found in traditional Western orchestras and in popular recordings. Some instruments feature different registers (alto, baritone, bass, etc) or different keys (pitched in C or B♭). All sounds used belong to the same pitch class (C), ranging in pitch height from C3 ($f0 \simeq 131$ Hz) to C6 ($f0 \simeq 1046$ Hz), but most are C3 or C4. The dynamics of all sounds is *forte*, while the duration was kept under 2 s. All sound files were edited so the first sample corresponds to the onset. Normal attack ("na") and no vibrato ("nv") were chosen whenever available. Presence of vibrato is indicated ("vib"), as well as different playing modes such as *staccato* ("stacc"), *sforzando* ("sforz"), and *pizzicato* ("pz"), achieved by plucking string instruments. Extended techniques were also included, such as *tongue ram* ("tr") for the flute and bowing ("bow") idiophones (vibraphone, xylophone, etc). Different materials such as metal, plastic and wood are also indicated (respectively by "me", "pl", and "wo".)

Brass	Bass Trombone (nv, na, stac), Bass Trumpet (na, vib), Cimbasso (nv, na, stac), Contrabass Trombone (stac), Contrabass Tuba (na, stac), Cornet, French Horn (nv, na, stac), Piccolo Trumpet (nv, na, stac), Tenor Trombone (nv, vib, na, stac), C Trumpet (nv, na, stac), Tuba (vib, na, stac), Wagner Tuba na, stac)
Woodwinds	Alto Flute (vib, na), Bass Clarinet (na, sforz, stac), Bassoon (na, stac), Clarinet (na, stac), Contra Bassoon (sforz, stac), English Horn (na, stac), Flute (nv, vib, na, stac, tr), Oboe (na, stac), Piccolo Flute (nv, vib, na, stac, sforz)
Strings	Cello (na, vib, pz), Double Bass (vib), Harp, Viola (na, nv, vib, stac, piz), Violin (na, nv, vib, stac, piz)
Percussion	Glockenspiel (wo, me, pl), Marimba, Vibraphone (me, pl, bow), Xylophone (wo, me)
Popular	Accordion, Acoustic Guitar, Baritone Sax, Bass Harmonica, Chromatic Harmonica, Classic Guitar, Mandolin, Pan Flute, Tenor Sax, Ukulele
Keyboard	Celesta (na, nv, stac), Clavinet, Piano

Table 6.1: *Musical instrument sounds used in all experiments. See text in 6.2.1 for a description of the terms in brackets*

6.2.2 Analysis Parameters

The parameter estimation for the SM follows [MQ86] with phase interpolation via cubic splines. The estimation of parameters for EDS used here is described in detail elsewhere [DBR13], while the estimation of the optimum number of poles (sinusoids) [BDR04] is used for comparison. In all experiments, the threshold for SRER convergence is set to 0.01, the size of the FFT is $N = 4096$ samples, and the sampling frequency for all sounds is $F_s = 16$kHz. The step size was $H = 1$ms (which corresponds to 16 samples). Prior to modeling with the SM, EDS, and the eaQHM, the fundamental frequency f_0 of all sounds was estimated using SWIPE [CH08] because in this work the window size L and the maximum number of sinusoidal partials K_{max} supposing harmonicity depend on f_0.

The number of (sinusoidal) partials K is an important input parameter which directly affects the modeling accuracy for the SM, EDS, and the eaQHM. In the SM, K dictates how many local maxima (harmonically related or not) the peak picking algorithm will retain. The parameter estimation algorithm for EDS [RK89] uses K to determine the separation between the dimension of the signal space and the noise space. In turn, the eaQHM initializes the template signal for QHM (see Section 2.4) with K harmonically related partials. For all the algorithms, we suppose the musical instrument sounds under investigation can be well represented as nearly harmonic, so we set the maximum number of partials K_{max} to the highest harmonic number below Nyquist frequency or equivalently the highest integer K that satisfies $Kf_0 \leq F_s/2$.

The window size L also directly affects the modeling accuracy of the SM, EDS, and the eaQHM. In the STFT, L determines the well known trade-off between temporal and spectral resolution which, in turn, directly affects the performance of the peak picking algorithm that the SM uses for parameter estimation. Moreover, the parameters estimated are averaged across the length L of the window and used to represent the center of the window. EDS estimates stationary (damped) sinusoids *inside* these frames, thus L limits the temporal modeling accuracy. Finally, the eaQHM uses SM-like frames and captures variations *inside* the analysis window. In this case, varying L does not directly affect temporal or spectral resolution because the eaQHM uses least squares (QHM) in the time domain to estimate the parameters.

[1] 'Popular' and 'Keyboard' musical instruments are from the RWC Music Database: Musical Instrument Sound http://staff.aist.go.jp/m.goto/RWC-MDB/. All other musical instruments are from Vienna Symphonic Library database of musical instrument samples http://www.vsl.co.at/en/65/71/84/1349.vsl

However, L will impact the modeling accuracy of the amplitude and frequency modulations inside the window. In the literature [RS78], $L = 3T_0$ (where $T_0 = 1/f_0$ is the fundamental period) is considered a reasonable value for speech and audio when using the SM. However, we are unaware of a systematic investigation of how L affects modeling accuracy for EDS. Next, we present the investigation on modeling accuracy (local and global SRER) as a function of the number of adaptations, K and L.

6.2.3 Adaptation Cycles

Figure 6.1 shows the *global* and *local* SRER as a function of the number of adaptation cycles (iterations). Each plot was averaged across the sounds indicated, while the plot "all instruments" is an average of the previously shown. Notice how the SRER increases quickly after a few iterations, slowly converging to a final value several orders of magnitude higher than before adaptation.

6.2.4 Number of Partials K

We studied the impact of K in the modeling accuracy of the SM, EDS, and eaQHM. We ran each algorithm with different numbers of partials as input parameter (the window size was kept at $L = 3T_0$) and recorded the resulting local and global SRER values. We started from K_{max} and decreased K by 2 partials each run. We expected the SM to quickly converge to a maximum value and stabilize because of the parameter selection algorithm. The literature on EDS [BDR04] suggests that there is an optimum number of partials for audio, thus we expected EDS to render a curve that would reach a maximum around that point and then decrease. Finally, adaptation allows the eaQHM to represent small temporal variations such as transients accurately as modulations of amplitude and frequency of the partials (no matter the number of partials). Therefore, we expected the eaQHM to yield higher values of SRER as the number of partials increased. Figures 6.2a and 6.2b shows the local and global SRER as a function of K for the SM, EDS, and the eaQHM. All curves are averaged across the sounds from the musical instruments indicated. Note that sounds with different f_0 values have different maximum number of partials K_{max}.

6.2.5 Window Size L

We investigated the impact of L in modeling accuracy for the SM, EDS, and the eaQHM. We ran each algorithm varying L from $3T_0$ to $8T_0$ with constant number of partials K_{max} and measured the resulting local and global SRER. We expected L to negatively impact all three algorithms differently. We expected L to have a greater impact on the SM because both parameter estimation and temporal resolution depend on L. We expected L to have a smaller impact on modeling accuracy for EDS because of the time-varying amplitude of the locally stationary sinusoids (despite the constant frequency value inside the window.) Finally, we conjectured that L will have a minor effect on the eaQHM because L mostly affects the eaQHM's ability to capture amplitude and frequency modulations inside the window. Figures 6.2c and 6.2d illustrate the results, showing the SRER as a function of L expressed as times T_0, so sounds with different f_0 values have different window size L in samples. All curves are averaged across the sounds from the musical instruments indicated.

However, Figure 6.2 is not enough evidence that the eaQHM outperforms the SM and EDS in average for all musical instrument sounds investigated. In what follows, we will compare the average modeling accuracy of the SM, EDS, and eaQHM using the curves from Figure 6.2 and defining the "mean SRER difference." For each musical instrument sound, we subtracted point by point the SRER values (in dB) corresponding to the SM and EDS from that of the eaQHM and averaged the result (across L or K). A positive "mean SRER difference" represents how much eaQHM outperforms the other method in average for that particular musical instrument, while a negative value means eaQHM was outperformed.

6.3 Analysis of Results

This section presents a systematic analysis of the experiments introduced in the previous section consisting of a comparison across musical instruments for all methods. Next, we present these mean SRER differences for all musical instrument sounds clustered by musical instrument family in Table 6.2 and Table 6.3.

6.3.1 Variation Across K Holding $L = 3T_0$

Table 6.2 shows the mean SRER difference between eaQHM and EDS and eaQHM and SM for the musical instruments clustered in families. The bottom row shows the average for all instruments labeled *Total*. The columns labeled *Local* and *Global* present the difference across K, while the column labeled K_{max} shows the difference in *global* SRER only for the maximum number of partials. The *Local* column is especially important to evaluate the attack since the SRER

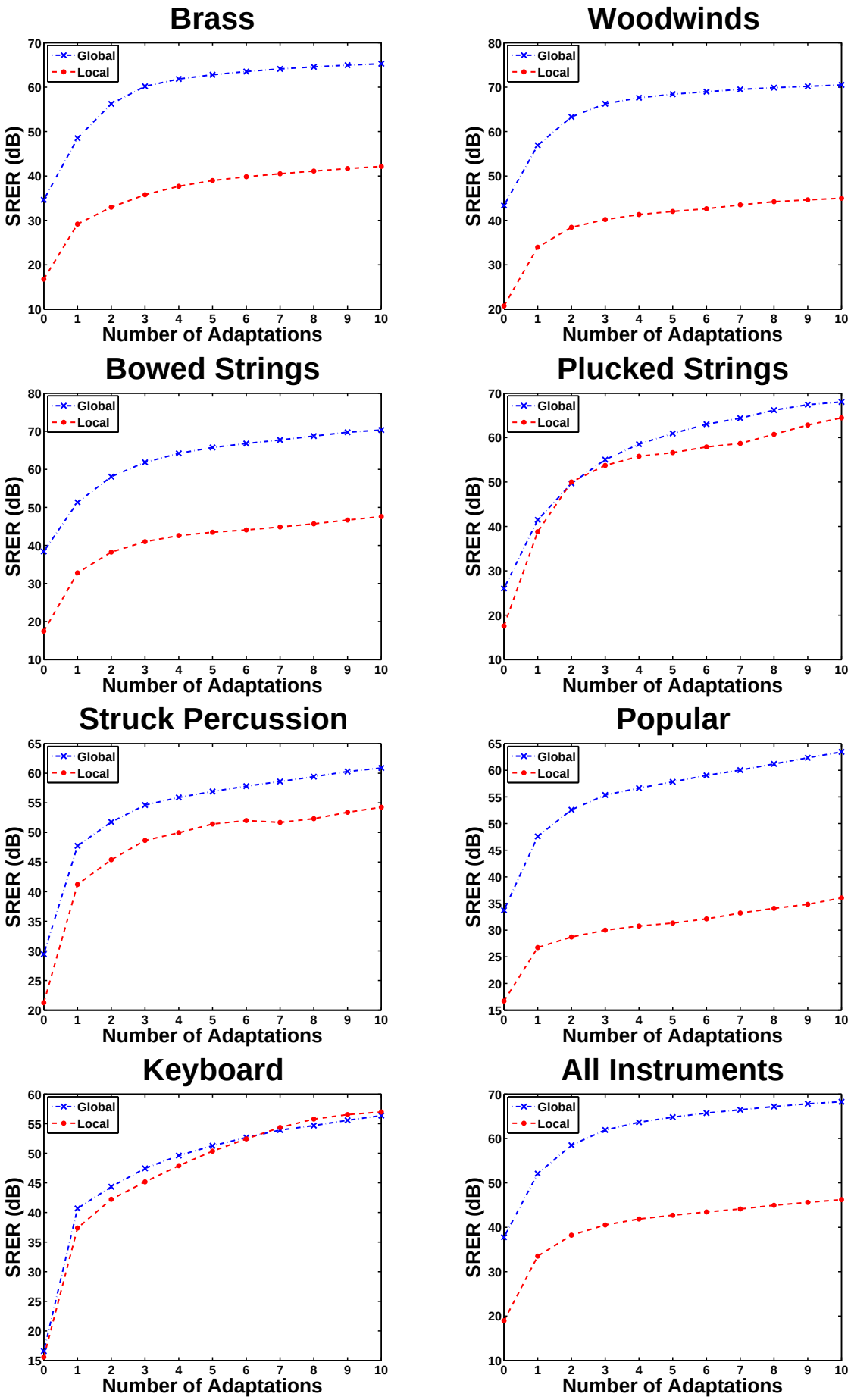
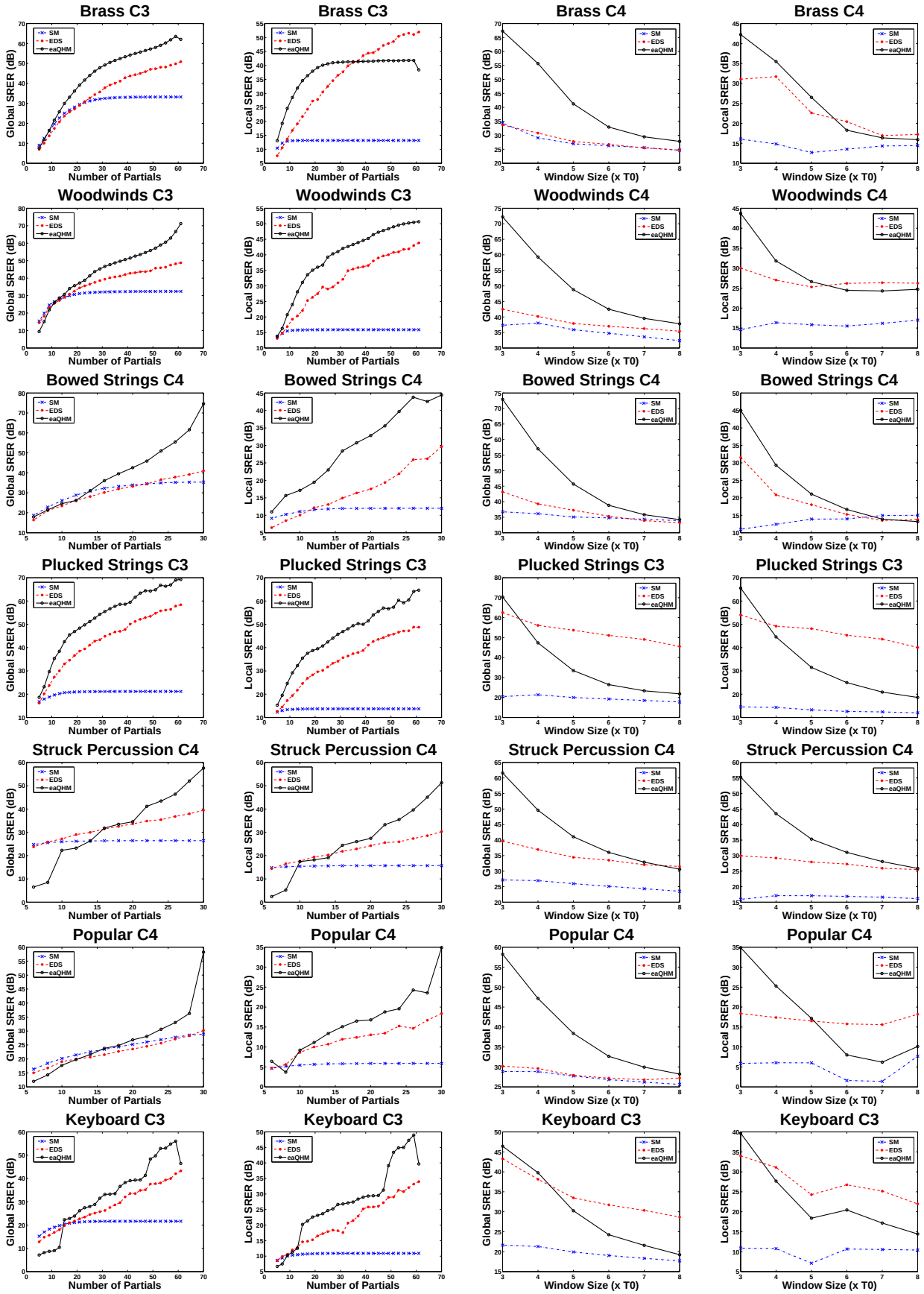


Figure 6.1: *Example of how adaptation increases the modeling accuracy. Plot of SRER as a function of number of adaptations.*



(a) Global SRER (b) Local SRER (c) Global SRER (d) Local SRER

Figure 6.2: *Comparison between global and local SRER as a function of the number of partials (a,b) and the size of the window (c,d) for the three models (SM, EDS, and eaQHM).*

does not include temporal information, the *Global* column assesses overall performance across K. The algorithms perform best at K_{max}, so this column will be used as reference measure of modeling accuracy per musical instrument family. The *Total* row will be used as reference measure of modeling accuracy per sinusoidal modeling algorithm. Hence the "mean SRER difference" in the K_{max} column and *Total* row gives the final modeling accuracy performance.

Locally, the eaQHM is only outperformed by EDS for *Bowed Percussion*, and globally for *Struck Percussion* and *Popular*. In turn, the SM only outperforms the eaQHM globally for *Popular*. Finally, the eaQHM presents a higher modeling accuracy than both SM and EDS with K_{max}. Both locally and globally, the eaQHM achieves the highest performance for *Plucked Strings*, outperforming EDS by over 10dB and SM by over 20dB. However, the column K_{max} reveals that the eaQHM outperforms EDS by more than 20dB for *Bowed Strings* and *Bowed Percussion*, while outperforming EDS by around 15dB for *Plucked Strings*. The row *Total* reveals that the eaQHM outperforms EDS and SM locally, globally, and for K_{max} in average when all musical instruments are clustered together.

6.3.2 Variation Across L Holding $K = K_{max}$

Similarly to the previous section, Table 6.3 shows the mean SRER difference between eaQHM and EDS and eaQHM and SM. The *Local* and *Global* columns present the mean SRER difference across L, while the $3T_0$ column shows the difference in *global* SRER for the window size that gives the best modeling accuracy for all algorithms. The local SRER is used to evaluate onset modeling accuracy and global SRER evaluates general performance. The column $3T_0$ will be used as reference measure per musical instrument family and the *Total* row as reference per model.

Under variation of L, the eaQHM only outperforms EDS locally for *Bowed Strings* and *Struck Percussion*. However, the eaQHM is only outperformed globally by EDS for *Plucked Strings* and *Keyboard*. Once again the eaQHM presented a consistently higher performance than EDS for $3T_0$. Note that the eaQHM outperforms the SM for every musical instrument cluster considered locally, globally, and for $3T_0$. When compared against EDS, eaQHM achieves the highest performance locally for *Struck Percussion*, globally for *Brass*, and using window size $3T_0$ for *Bowed Strings*. However, when we compare the performance with the SM, *Plucked Strings* is the highest performing cluster locally, globally, and for $3T_0$. Notice that the comparison across L with EDS is not consistent, while the comparison with the SM is much more consistent. This phenomenon will be analyzed in the following section.

	SRER(eaQHM-EDS)			**SRER(eaQHM-SM)**		
	Local	**Global**	$\mathbf{K_{max}}$	**Local**	**Global**	$\mathbf{K_{max}}$
Brass	7.69	9.22	*7.48*	22.44	9.05	*5.41*
Woodwinds	2.95	6.90	*19.60*	17.21	12.18	*29.93*
Bowed Strings	6.32	3.45	*21.52*	14.46	6.06	*31.21*
Plucked Strings	13.26	11.96	*15.75*	24.69	25.63	*42.99*
Bowed Percussion	-3.89	2.00	*21.80*	2.65	10.56	*37.63*
Struck Percussion	2.83	-1.52	*11.88*	7.68	2.77	*21.06*
Popular	1.09	-0.44	*11.93*	3.55	-1.21	*18.09*
Keyboard	6.33	4.41	*3.15*	16.99	11.87	*24.79*
Total	**4.55**	**4.50**	***14.14***	**14.15**	**9.07**	***26.39***

Table 6.2: *Mean SRER difference (dB) between eaQHM and EDS or SM across the number of partials K.*

6.4 Discussion

Figure 6.1 shows the impact of adaptation on modeling accuracy for the eaQHM, revealing that adaptation does lead to a significant improvement in modeling accuracy as measured by both local and global SRER. Figure 6.2 illustrates the behavior of the SM, EDS, and eaQHM under variation of the number of partials K and window size L, respectively. In general, there is no significant difference in local and global behavior for the curves, while the global SRER presents higher absolute values. The tendency for higher absolute global SRER seems natural for Figure 6.2 because the global SRER uses information from the whole sound duration, while the local SRER reflects the fit of the models right before the beginning of the sound.

Figures 6.2a and 6.2b show that the SM behaves as expected when varying the number of partials, not showing significant improvement in modeling accuracy after a certain number of partials. Comparison between the local SRER variation for the SM in Figure 6.2b and its global counterpart in Figure 6.2a shows that the global values are higher

	SRER(eaQHM-EDS)			SRER(eaQHM-SM)		
	Local	**Global**	**3T$_0$**	**Local**	**Global**	**3T$_0$**
Brass	-3.07	10.74	*27.27*	12.20	14.75	31.19
Woodwinds	-2.81	6.92	*18.21*	11.63	13.43	*30.44*
Bowed Strings	3.01	7.55	*28.11*	10.85	12.62	*38.18*
Plucked Strings	-12.38	-15.89	*7.78*	21.16	17.57	*49.88*
Bowed Percussion	-4.26	6.86	*21.80*	10.83	15.85	*37.63*
Struck Percussion	8.13	5.85	*10.82*	15.24	11.73	*19.25*
Popular	-0.91	5.57	*15.25*	9.00	9.02	*21.91*
Keyboard	-4.25	-4.02	*3.15*	12.90	10.60	*24.79*
Total	**-2.30**	**2.95**	***16.55***	**12.46**	**12.78**	***31.66***

Table 6.3: *Mean SRER difference between the eaQHM and EDS or SM across the window size L.*

for the same sounds, and the addition of partials tends to increase the global SRER more than the local SRER. This trend confirms that the SM has a tendency to fit onsets poorly and stable partials significantly better. Interestingly, EDS does not behave as predicted theoretically in the literature [BDR04]. Instead of reaching a maximum performance for a certain number of partials and decreasing after that, EDS presents a general trend to continuously improve the modeling accuracy when the number of partials increases. In fact, the optimum number of partials obtained with ESTER [BDR04] was much higher than the maximum number of partials K_{max} for all musical instrument sounds. Moreover, the difference in absolute value between local and global SRER for EDS is less significant than for the SM or the eaQHM, revealing tha EDS does not improve the fit with the presence of more information. Finally, the eaQHM behaved as expected, improving the modeling accuracy with more partials. In general, the absolute value for the global SRER is higher than for the local fit, but the difference between global and local is smaller than for the SM, indicating that the eaQHM presents a more consistent modeling performance throughout.

In turn, Figures 6.2c and 6.2d show that the modeling accuracy decreases with L for all algorithms. Once again the local and global SRER curves show a consistent behavior for each model, with the same general tendency for higher global than local SRER values for the same sounds. Interestingly, the window size affects the SM much less than EDS or the eaQHM, suggesting that the SM is more robust to variations of L than EDS or the eaQHM. On the other hand, it is depicted that the SM seldom outperforms the others (apparently, only when the performance of EDS and the eaQHM is compromised by very large values of L.) Again, the SM presents a higher difference between global and local SRER than EDS or the eaQHM. Figures 6.2c and 6.2d reveal that variation across L affected the performance of EDS significantly less than that of the eaQHM and the SM because of the different impact that L has in the eaQHM and SM against EDS. L is the length of the non-overlapping slice that EDS models. Internally, the effective length used by EDS to fit the parameters of the model is $L/2$ with a step size of 1 sample. Here we should notice that L more greatly affects the difference between global and local SRER for EDS than the number of partials. Not surprisingly, the modeling accuracy of the eaQHM suffers when L increases. In some cases, the general performance is still superior throughout (see Fig. 6.2c, *Brass*, *Woodwinds*, *Bowed strings*, and *Popular*). The difference between global and local SRER in absolute value for L is more significant than for K regarding the eaQHM. In the eaQHM, L affects frequency correction and interpolation mechanisms. Long windows have an averaging effect because the parameters are iteratively fit in the time domain. Frequency correction is applied an the center of the analysis window and the eaQHM uses interpolation to capture frequency modulations between windows. Thus, adaptation improves the fit more slowly for longer L, generally reaching a lower "roof" SRER value in fewer iterations.

6.4.1 Analysis and Synthesis Complexity

Here, complexity is considered as the number of parameters (or degrees of freedom) required to estimate and to represent each sinusoid. Table 6.4 shows a comparison for the SM, EDS, and eaQHM using real numbers because some parameters are complex. As is widely known, the SM requires the estimation of three real parameters for each sinusoid, namely the amplitude a_k, the phase ϕ_k, and the frequency f_k. In turn, EDS requires the estimation of two complex parameters [DBR13], amplitudes $\mathbf{a}_k$, and poles $\mathbf{z}_k$. These are converted into the four real parameters amplitude a_k, phase ϕ_k, frequency f_k, and damping coefficient δ_k (see [DBR13] for details). Finally, the eaQHM estimates two complex amplitude parameters $\mathbf{a}_k$ and $\mathbf{b}_k$. The frequencies f_k are initialized as integer multiples of a fundamental frequency f_0 and later corrected by the estimations $\mathbf{a}_k$ and $\mathbf{b}_k$, so f_k is not directly estimated by the eaQHM. Thus the eaQHM and EDS present a higher analysis complexity than the SM. Table 6.4 also shows the number of parameters

for the synthesis stage. Notice that the eaQHM has the same synthesis complexity as the SM, while both the SM and EDS need all analysis parameters also in the synthesis stage. The synthesis complexity of the eaQHM is lower than the analysis stage because the synthesis stage is essentially the same as the SM.

	Real numbers per sinusoid per frame		
	SM	**EDS**	**eaQHM**
Analysis	a_k, ϕ_k, f_k	$\Re\{\mathbf{a}_k\}, \Im\{\mathbf{a}_k\}$, $\Re\{\mathbf{z}_k\}, \Im\{\mathbf{z}_k\}$	$\Re\{\mathbf{a}_k\}, \Im\{\mathbf{a}_k\}$, $\Re\{\mathbf{b}_k\}, \Im\{\mathbf{b}_k\}$
Synthesis	a_k, ϕ_k, f_k	$\lvert a_k\rvert, \phi_k, f_k, \delta_k$	$\lvert a_k\rvert, \phi_k, f_k$

Table 6.4: *Comparison of model complexity for SM, EDS, and the eaQHM for the analysis and synthesis stages. The table presents the parameters (real numbers) to estimate (analysis complexity) and to represent (synthesis complexity) each sinusoid inside a frame.*

6.4.2 Modeling Accuracy and SRER

In this work, modeling accuracy is the ability of a model to capture information from a signal. The SRER defined in Eq. (6.1) is the ratio in dB of the energy in the original signal $x\,(t)$ and the modeling error or residual $\bar{x}\,(t)$. As such, the aim of maximizing the SRER is equivalent to minimizing the residual energy and therefore minimizing the information missed by the model. Here, we consider that some of the noise might be intrinsic to the musical instrument sound being modeled, such as breathing noise, and should be captured as well. Therefore, technically, the SRER does not measure the same as the more commonly known signal-to-noise ratio (SNR) because the SNR usually considers additive noise from an external source whose statistical properties differ from the signal's, such as background noise from quantization or transmission. Thus the aim of minimizing the residual energy in sinusoidal modeling can be interpreted as capturing as much signal energy as possible with sinusoids. Ideally, the sinusoids should capture all oscillatory energy (including transients) and leave a noisy residual with a flat spectrum, indicating that the residual is indeed statistically independent from the model.

An important feature of the SRER is that it uses the waveforms to estimate the energy, comparing $x\,(t)$ and $\hat{x}\,(t)$ directly. Consequently, $\bar{x}\,(t)$ will only have low energy when $\hat{x}\,(t)$ follows $x\,(t)$ closely. However, the SRER is blind to where the differences lie in the waveform. For discrete waveforms, if $\hat{x}\,(t)$ as an identical copy of $x\,(t)$ except for a single sample, the energy in $\bar{x}\,(t)$ results in a particular value of SRER. However, a different waveform $\hat{x}_2\,(t)$ created by adding a small perturbation to each sample in $x\,(t)$ could have the same SRER. Consequently, the SRER alone does not show that $\hat{x}_1\,(t)$ and $\hat{x}_2\,(t)$ are different waveforms. So, when comparing two models $\hat{x}_1\,(t)$ and $\hat{x}_2\,(t)$, it is only safe to say that a higher SRER indicates better modeling accuracy, which in turn suggests a better quality representation.

6.4.3 Percussive Musical Instruments

The musical instrument clusters that contain percussive sounds in this work are *Plucked strings*, *Struck percussion*, and *Keyboard*. The literature [NHD98b, BBD02, HVL+05] proposes that EDS is particularly suitable to model percussive sounds because of the exponential temporal envelope, commonly claiming that the parameter estimation technique [RK89] uses "high-resolution methods" that outperform traditional estimation methods based on Fourier analysis. However, the ability to adapt the amplitude of the sinusoidal partials to the local characteristics of the waveform makes the eaQHM extremely flexible to fit both percussive and nonpercussive musical instrument sounds. For example, Table 6.2 shows that the eaQHM outperformed EDS the most for *Plucked Strings*, a percussive sound that EDS should supposedly capture well due to the temporal envelope. Upon close examination, Table 6.2 reveals that the eaQHM in general outperforms EDS in local, global, and K_{max} for the percussive instruments. Table 6.3, on the other hand, shows better performance for the eaQHM than EDS only for $3T_0$ for these particular clusters.

In general terms, the experiment revealed that the eaQHM has a better modeling accuracy than EDS and SM in average across K, with a robust performance throughout. Table 6.2 shows that the eaQHM consistently achieved better modeling accuracy than EDS and SM for K_{max}. In some cases, the eaQHM significantly outperforms EDS, while

the *total* performance was 14.14 dB. Table 6.3 also reveals that the eaQHM outperforms EDS for $3T_0$ for all clusters of musical instruments with some significant differences of more than 20 dB, while comparison with the SM shows significant differences across all clusters. The *total* row indicates that the eaQHM outperforms the EDS in average by 16.55 dB and the SM by 31.66 dB.

6.5 Conclusion and Perspectives

Musical instrument sounds are challenging to represent accurately because different musical instruments may feature radically different characteristics, such as sharp onsets, attack transients, inharmonicity, or mechanical noise. The quality of the representation depends not only on accurate parameter estimation, but also on how the underlying model uses this information to capture and represent temporal variations of the model parameters. This work proposed to use an adaptive Sinusoidal Model dubbed eaQHM to represent percussive and nonpercussive musical instrument sounds as sinusoids modulated in amplitude and frequency. In general, the eaQHM renders a compact yet high-quality representation with intuitive parameters. The model represents well sharp onsets with attack transients, inharmonic spectra, and even mechanical noise.

We showed that adaptation of the sinusoids inside the analysis window allows the eaQHM to significantly increase the accuracy of representation without the need to increase model complexity. We used the signal to reconstruction error ratio (SRER) to compare the modeling accuracy of the eaQHM with exponentially damped sinusoids (EDS), considered here to be the state of the art, and the standard sinusoidal model (SM) as the baseline representation of a database of 90 percussive and nonpercussive musical instruments sounds. The experiments measured the local and global SRER as a function of the number of partials and size of the analysis window for the SM, EDS, and eaQHM. The local SRER is measured just before the onset to capture potential artifacts in attack transient modeling, and the global SRER measures the fit for the whole waveform to evaluate general modeling performance. The results showed that the eaQHM outperforms EDS and SM in average across both variations in all cases except local SRER for the EDS under variation of window size. Considering only the window size and number of partials for which the algorithms perform their best, the eaQHM consistently outperformed EDS by more than 10dB and SM by over 25dB in average.

Adaptive sinusoidal modeling can be used in parametric audio coding with low bit rates because the very high modeling accuracy potentially gives high fidelity with the same model complexity as the standard sinusoidal model. However, the eaQHM currently only handles monophonic sounds (speech or musical instruments). An important perspective of this work is to develop an adaptive algorithm for polyphonic audio. Therefore, future work should focus on improving the robustness of parameter estimation first. Presently, the use of least squares is costly and unstable, failing whenever the frequencies of two partials are closer together than a threshold value. We envision total least squares or singular value decomposition as good candidates to improve the robustness in parameter estimation and allow modeling of polyphonic audio.

Chapter 7

Expressive Speech Analysis and Classification

7.1 Introduction

Emotional (or stressed) speech can be defined as the speech style produced by an emotionally charged speaker. Such speech styles can be characterized as *happy, sad, angry, neutral* and *fearful* speech, among others. Analysis of emotional speech could provide information about the emotional state of the speaker, which can be useful in applications such as health care and emergency conditions, and is a necessary pre-processing step in applications such as recognition and classification. Also, speaker recognition and verification systems could benefit from such an analysis, as well as speech synthesis applications, like unit selection based text-to-speech synthesis or HMM-based speech synthesis.

Numerous approaches have been suggested in the literature in order to show the variation of speech characteristics among different emotion conditions. These variations can form *features* that are exploited to identify and/or classify different emotional speech styles [BGH00]. Womack and Hansen discussed the use of Linear Prediction (LP) coefficients and cepstral features in analyzing and classifying stressed speech [DH99, HW96, HWA94, WH95]. Zhou et al [ZHK01] have shown that the Teager operator can be used to obtain better results compared to LP-based features in classification of stressed speech. Moreover, it has been suggested that features related to the pitch mean and variance, as well as intensity features, are useful for discrimination among speaking styles [AR98, BN08]. Cummings et al [CCH89] have shown that the glottal pulse shape varies with different stressed conditions. Ruiz et al [RAH+96] discussed time and frequency related variabilities in stressed speech, whereas Castellanos et al [CBC96] provided an analysis of general acoustic-phonetic features in Lombard speech. Scherer [Sch03] investigated the intensity, duration, and spectral envelopes in stressed speech for speech and speaker recognition, whereas Bosch [Bos13] has discussed the importance of prosody for emotion recognition in speech. Ramamohan and Dandapat [RD06] suggested the use of a sinusoidal model (SM) to distinguish between different speaking styles, using its parameters (amplitude, frequency, phase) as features. For the recognition and/or classification of emotional speech, several classifiers have been suggested, such as Hidden Markov Models (HMM) [DH99, RD06, CH94, NMBM01, KCJL03, NFS03], Neural Networks (NN) [BGH00, HW96, NTN00, BWG04], Gaussian Mixture Models [LNHS05, AKK07], and Vector Quantization (VQ) [RD06, KK11] using a variety of feature vectors.

In spite of its wide range of applications [MBCM93], the Sinusoidal Model (SM) [MQ86] has not been thoroughly engaged in analysis and/or classification of stressed speech until recently [RD06, DTCT03]. In these approaches, the parameters of sinusoids (amplitude, frequency, and phase) over time are suggested as features for classification or conversion of speech using Hidden Markov Models, Vector Quantization, and Gaussian Mixture Model-based techniques. Although the use of amplitude and frequency contours was straightforward, the phase contours are either disregarded or could not be directly used in the analysis. Furthermore, the parameters obtained from sinusoidal analysis have a significant constraint; they are extracted under the assumption of *local stationarity*, that is, the speech signal is considered as *stationary* inside the analysis window. However, this is not the case for speech styles characterized as *"fast"* or *"angry"*. Recently, the adaptive Sinusoidal Models (aSMs) [PRS11, KPRS12, DS13] have managed to cope with this problem by projecting the signal onto a set of amplitude- and frequency-varying basis functions *inside* the analysis window. This way, the parameters represent the underlying signal more closely as an AM-FM decomposition. In brief, the adaptive Quasi-Harmonic Model (aQHM) [Pan10] adapts the phase of the basis function to the local characteristics of the signal, whereas the extended adaptive Quasi-Harmonic Model (eaQHM) [KPRS12] performs both amplitude and phase adaptation. More recently, the adaptive Harmonic Model (aHM) [DS13] assumes full-band harmonicity and iteratively adapts the fundamental frequency f_0 to localize harmonics up to the Nyquist frequency. All models have demonstrated their

ability to model adequately and accurately speech signals from different languages and different speakers. However, they have not been tested in emotional speech, where it is assumed that the AM-FM components of the speech signal behave differently compared to neutral or conversational speech.

In this work, the extended adaptive Quasi-Harmonic Model (eaQHM) is utilized to demonstrate its ability to analyze, resynthesize, and classify emotional speech. The speech corpus for the analysis and resynthesis is a high-quality, wideband database containing emotional running speech. Subjective listening tests have been conducted to prove the transparency of the resynthesized speech. It is also shown that eaQHM can efficiently model all styles of emotional speech in this database with high precision, and this is demonstrated via Signal-to-Reconstruction-Error Ratio (SRER) values, compared to the standard SM. Moreover, an emotion classification task is presented using the well-known Speech Under Simulated and Actual Stress (SUSAS) [HBG97] database, in which there are 11 pre-labelled emotional speech corpora. Details on the database are discussed in Section 3. Results show that the sinusoidal features of the eaQHM yield higher classification scores than those of the SM.

The rest of the work is organized as follows. In Section 7.2.1 presents the analysis parameters and the evaluation, both objective and subjective, of the eaQHM compared to SM. Section 7.3 describes the VQ-based classification experiment, and Section 7.4 discusses future perspectives. Finally, Section 7.5 concludes the work.

7.2 Analysis and Evaluation

In this section, the evaluation procedure is described, along with the dataset selection and the parameter estimation.

7.2.1 Objective Evaluation

At first, it is important to show that eaQHM can decompose high-quality running expressive speech signals into AM-FM components that represent the signal closer than SM. For this, a custom, small database of acted speech is used. This database consists of one male and one female subject, acting in four different speaking styles (*angry, sad, happy, neutral*), in a recording studio. A total number of 20 waveforms sampled at 16000 Hz are analyzed. All speech files in the database have been analyzed and resynthesized from their AM-FM components, and the corresponding SRER has been computed for each speech utterance. For this analysis, the window size was 30 ms for the SM and 3 local pitch periods for the eaQHM, both of Hamming type. A step size of 2.5 ms was selected for both models. The results are depicted in Table 7.1.

SRER Performance (Wideband Speech Database)				
Female Speaker				
Model	Speaking Styles			
	Angry	Happy	Neutral	Sad
SM	14.8 (1.36)	17.5 (3.0)	16.5 (1.36)	21.2 (1.64)
eaQHM	28.8 (1.24)	33.1 (1.81)	34.9 (2.23)	34.8 (3.60)
Male Speaker				
SM	17.0 (1.45)	14.3 (0.76)	16.0 (1.67)	16.5 (1.63)
eaQHM	35.7 (2.04)	31.6 (3.49)	33.3 (2.56)	33.1 (2.74)

Table 7.1: *Signal to Reconstruction Error Ratio values (dB) for both models on a small acted speech database. Mean and Standard Deviation are given.*

However, this database is not appropriate for classification purposes, since the containing data is too few. Another database will be used, named SUSAS (Speech Under Simulated and Actual Stress). The SUSAS database was developed in the 1990s and was the first emotional speech database ever created. It contains both actual and simulated stressed speech. In the simulated part, 9 U.S. English male speakers, of three main dialects (general USA, New England/Boston, and New York City accent), under different *simulated* stress conditions (*angry*, *clear*, *fast*, *lombard*, *loud*, *neutral*, *question*, *slow*, *soft*, and two conditions where the speaker was recorded during *medium and light activity*) have been recorded. Each speaking style corpus has 70 speech files per speaker, which consist of isolated, short communication words, such as *"hello", "break", "go"*, and *"destination"*. This amounts to about 1190 tokens per speaker, with a considerable subset of them being acoustically similar, such as (*six, fix*) and (*white, wide*). The simulated data in SUSAS database were sampled using a 16-bit A/D converter with sample rate of 8 kHz. Table 7.2 shows the mean and the standard deviation of SRER for all speakers, for most common speaking styles.

This clearly demonstrates the quality and the performance stability of the adaptive model compared to the SM on a large database of isolated words of different expressive speaking styles. It is interesting to note that both models appear

SRER Performance (SUSAS)				
Model	Speaking Styles			
	Angry	Loud	Clear	Fast
SM	16.6 (3.06)	16.8 (3.01)	16.8 (3.06)	16.7 (3.03)
eaQHM	32.3 (5.61)	32.8 (5.59)	32.6 (5.62)	32.9 (5.58)
	Question	Soft	Neutral	Slow
SM	16.8 (3.00)	16.7 (3.05)	16.8 (3.01)	16.8 (3.05)
eaQHM	32.8 (5.57)	32.9 (5.61)	32.9 (5.58)	32.9 (5.60)

Table 7.2: *Signal to Reconstruction Error Ratio values (dB) for both models on the SUSAS database. Mean and Standard Deviation are given.*

to be very stable around a mean of about 16.6 and 32.5 dB, for the SM and the eaQHM respectively. Although the distribution of SRERs is wider in eaQHM-analysis, the mean is high enough to show that in almost all cases the eaQHM manages to compactly capture most of the information present in the speech signal, for *all* speaking styles. Conclusively, it is evident that the adaptive model can handle word-isolated (i.e. SUSAS) and running expressive speech equally well.

7.2.2 Subjective Evaluation

For our subjective evaluation, a formal, on-line listening test was designed[1] using the small, high-quality database of emotional running speech. The listeners were asked to evaluate the overall quality of the resynthesized speech based on the two models. A total of 32 listeners participated in this test, and the results are depicted in Figure 7.1 along with the 95% confidence intervals. Please note that only 5 of them are familiar with signal processing. According to the preference test, almost all listeners noted eaQHM as being almost indistinguishable to the original one. It should be noted

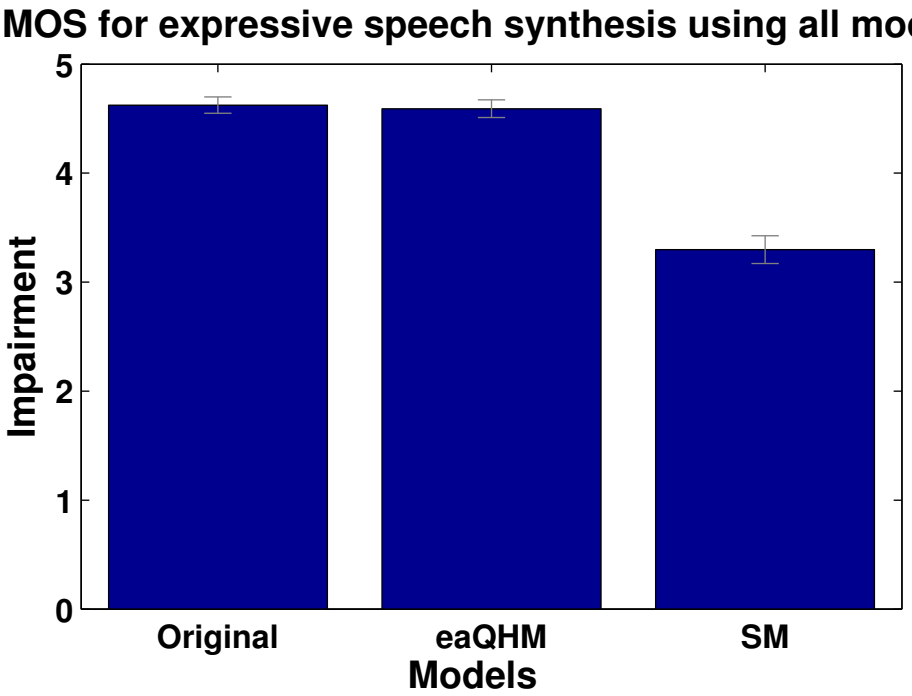


Figure 7.1: *Impairment evaluation of the resynthesis quality, with the 95% confidence intervals.*

that the SUSAS database was judged to perform poorly from a perceptual point of view due to the recording noise and the low sampling frequency. Informal listening tests showed that the eaQHM-based resynthesized speech samples were indistinguishable from the original ones, but this was the case for most samples obtained from the standard Sinusoidal Model as well. After careful listening, only a minority of waveforms demonstrated perceptual differences between the models but they were not enough in quantity to justify a listening test with this database. However, due to its pre-labelled data and its parallel corpora for each speaking style, this database was characterized as suitable for the classification task.

[1] http://www2.csd.uoc.gr/~kafentz/listest/pmwiki.php?n=Main.Exprtest

7.3 VQ-based Emotion Classification

As already discussed, a discrimination between different emotional speaking styles is of great interest. Considering a sinusoidal analysis, it has been reported that amplitude and frequency values of the sinusoidal components can be used successfully to characterize the different expressive classes (emotions) in a speech signal [RD06]. Since the eaQHM can compute these parameters more accurately, it is not surprising that their discrimination properties among different speaking styles are similar or better than those reported in the literature for the standard SM. An example is presented in Figure 7.2, where the parameters of two speech samples (of the same word: *"No"*) from the SUSAS database pronounced with different emotional content (*angry, neutral*) are depicted.

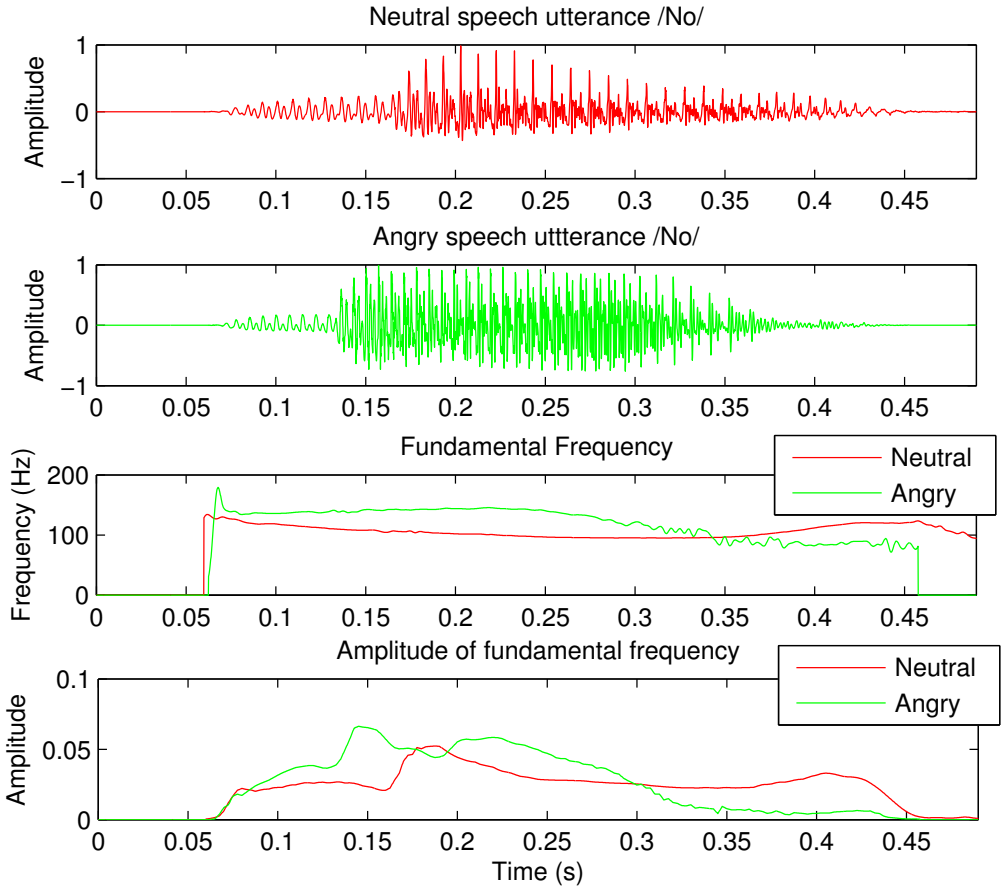


Figure 7.2: *An example of analysis of emotional speech: First panel, neutral speech. Second panel, angry speech. Third panel, $f_0(t)$ tracks for each sample. Fourth panel, $A_0(t)$ tracks for each sample.*

Clearly, the amplitudes and frequencies of the fundamental are different in each case, and this is the case for other sinusoidal components as well. Another example of a single word ("point") in four different emotions is depicted in Fig. 7.3, along with the corresponding spectrograms that partly reveal their differences. The signals are aligned according to the stop consonant /p/. It can be seen that these differences appear in amplitude strength, frequency variations, energy distributions, formant positioning, timings, duration of vowels and consonants, etc. Sinusoidal modeling can capture some of these differences in the form of AM-FM components [RD06]. Due to its adaptive processing, we propose that eaQHM can provide parameters that are highly accurate, which makes them more suitable for an emotion classification task than the same parameters obtained from a standard SM.

7.3.1 Feature Extraction

To evaluate our suggestion, a classification task based on a 128-bit Vector Quantizer (VQ) was designed using a subset corpus of the SUSAS, labelled as *Angry*, *Neutral*, *Soft*, and *Question*. A total number of 2520 waveforms (630 per emotion) were used. A number of 756 waveforms were kept for testing (189 per emotion), while the rest were used for

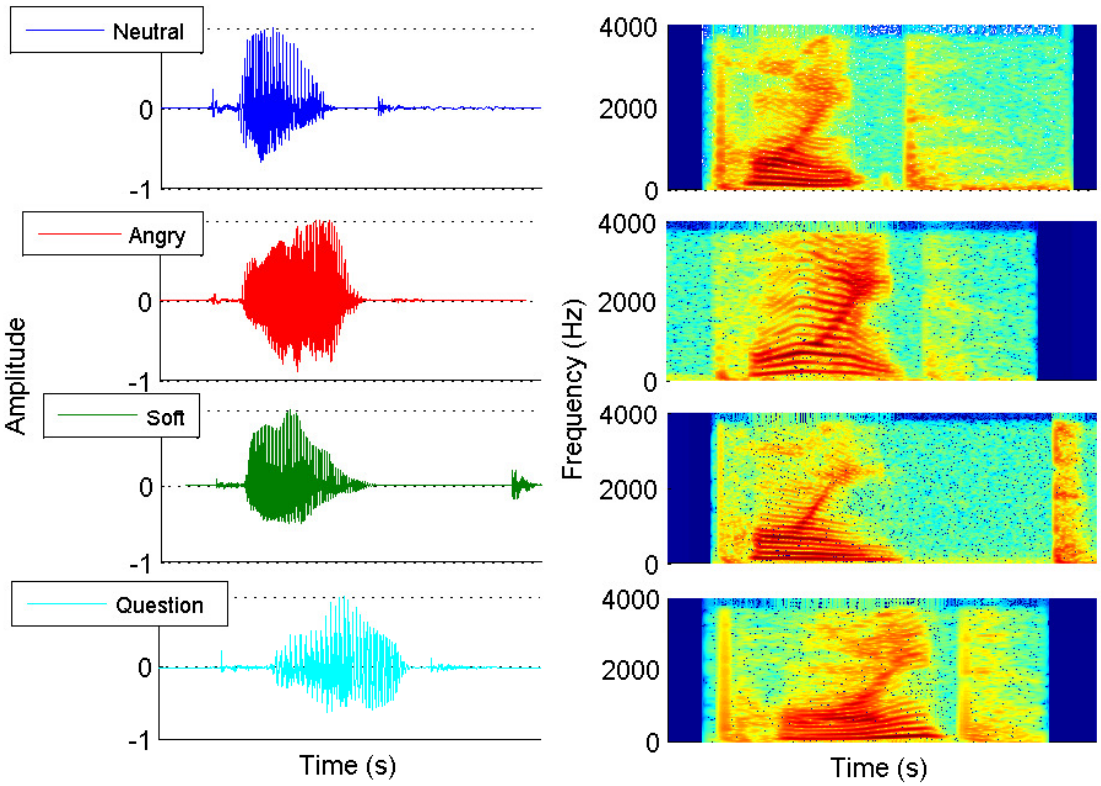


Figure 7.3: *An example of emotional speaking styles, in time and frequency: First panel, neutral. Second panel, angry. Third panel, soft. Fourth panel, question. The word "Point" is depicted in this example.*

training. All discrete-time waveforms were normalized to unit energy, as in

$$x[n] = \frac{x[n]}{\sqrt{\sum_{n=0}^{L-1} x^2[n]}} \tag{7.1}$$

where L is the signal length in samples. Both models used an analysis frame rate of 2.5 ms. The 10 strongest components of the magnitude spectrum of the FFT and the 10 highest sinusoidal amplitudes provided by the LS, along with their corresponding frequencies, were extracted from each analysis frame. The analysis window was set at 30 ms for the SM, and at 3 local pitch periods for the eaQHM. No distinction between voiced and unvoiced parts of speech was made in this work.

7.3.2 Classification - Single Feature

At first, two classification tasks were set, each one using different features (amplitudes and frequencies). Having M spectral vectors $\mathbf{x_i}$ containing the selected features (amplitudes or frequencies), the data matrix $\mathbf{X}$ is created as

$$\mathbf{X} = [\mathbf{x_1} \quad \mathbf{x_2} \quad \cdots \quad \mathbf{x_M}] \tag{7.2}$$

The codebooks are then designed based on the minimization of the Average Distortion (AD) between the training vectors and the codebook vectors in matrix $\mathbf{Y}$, where

$$\mathbf{Y} = [\mathbf{y_1} \quad \mathbf{y_2} \quad \cdots \quad \mathbf{y_C}] \tag{7.3}$$

and C is the codebook size. The AD is defined as

$$AD = \frac{1}{C} \sum_{k=1}^{C} \min_{\mathbf{y_i} \in \mathbf{Y}} d^2(\mathbf{x_k}, \mathbf{y_i}) \tag{7.4}$$

where $d(\mathbf{x}, \mathbf{y})$ is the Euclidean Distance (ED) between vectors $\mathbf{x}$ and $\mathbf{y}$. For each of the four emotions mentioned earlier, a codebook was designed using the LBG algorithm [LBG80]. The emotion is recognized by the minimum average distortion. The Confusion Matrix for the amplitude-based classification is given in Table 7.3, whereas for the corresponding frequency-based one is given in Table 7.4. It can be seen that in both cases the *angry* speaking style stands

		VQ Classification in % - Amplitudes			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	77(72)	14(14)	2(3)	7(11)
	Neutral	4(4)	64(63)	18(18)	14(15)
	Soft	3(5)	31(30)	56(50)	10(15)
	Question	6(4)	21(22)	13(20)	60(55)

Table 7.3: *Classification score (%) for four emotions of the SUSAS database, using amplitude features extracted from eaQHM and SM (in parenthesis).*

out of the rest of speaking styles. This is expected since this speaking style is very different than the others in terms of amplitude and frequency distributions [RD06].

		VQ Classification in %- Frequencies			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	71(70)	6(6)	7(5)	21(18)
	Neutral	6(6)	55(38)	24(28)	15(27)
	Soft	3(3)	13(25)	65(59)	14(13)
	Question	17(18)	18(24)	14(25)	50(33)

Table 7.4: *Classification score (%) for four emotions of the SUSAS database, using frequency features extracted from eaQHM and SM (in parenthesis).*

In general, the parameters obtained from the eaQHM lead to better classification scores in all cases. Furthermore, the *angry* speaking style has the highest correct classification percentage for both models and both sets of features. The *question* speaking style is the most difficult one to correctly classify when the frequencies are used as features, and we can see that it is mostly confused with the *neutral* speaking style. On the other hand, the *soft* speaking style has the lowest classification score when the amplitudes are used as features.

7.3.3 Classification - Combined Features

Since single-feature based classification leads to low classification scores, a combined classification scheme is suggested. The ADs obtained from amplitude and frequency based VQs are normalized by the highest corresponding AD. Then, the ADs of the corresponding emotions are added. Finally, the emotion with the minimum sum of ADs is selected as the recognized emotion. This way, when the VQs have decided differently, the VQ which is more "confident" in its decision (the minimum AD is far less than other ADs) can influence the final outcome. Figure 7.4 illustrates the proposed scheme.

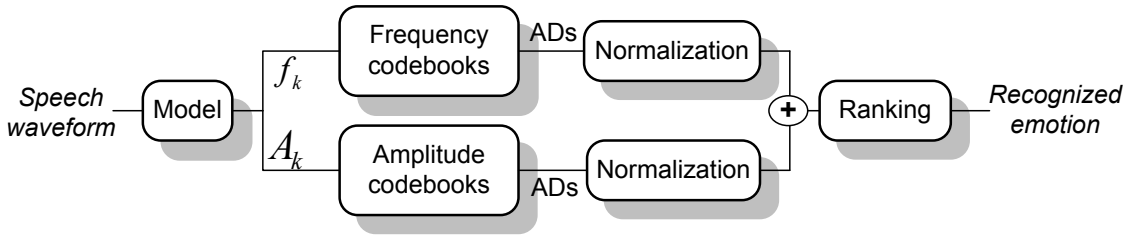


Figure 7.4: *The proposed classification scheme based on the combination of features. A_k and f_k denote the instantaneous amplitude and frequency components, and ADs denote the average distortion measures.*

Table 7.5 presents the corresponding classification scores for eaQHM and SM using the proposed scheme. Using

		VQ combined classification in %			
		Predicted Class			
		Angry	Neutral	Soft	Question
Class	Angry	83(77)	5(5)	1(5)	11(13)
	Neutral	15(4)	58(48)	12(24)	15(24)
	Soft	10(2)	18(29)	56(54)	16(15)
	Question	20(17)	6(24)	11(21)	63(38)

Table 7.5: *eaQHM and SM based Confusion Table in* % *based on amplitudes and frequencies for a* 128*-bit VQ classification between* 4 *emotions of the normalized SUSAS database.*

this scheme, on average, the eaQHM correctly classifies 65% of the utterances in the database, whereas the SM reaches 54%. Apparently, not all speaking styles were favoured by this combined scheme. Mostly the *angry* and the *question* speaking style achieved significant increase of their classification rates in both models. While the *angry* speaking style already had a relatively high percentage, the *question* speaking style has interestingly increased its correct classification score. However, the *soft* and *neutral* speaking style did not significantly change their percentages. This suggests that a weighted sum of the ADs before ranking may be more appropriate.

7.4 Discussion and Perspectives

In this work, we attempted to perform emotion classification from speech signals using instantaneous parameters of sinusoidal models. Although the database in hand contains short, isolated words with similar perceptual content, and this makes recognition and classification results rather difficult, results are encouraging. However there is room for improvement.

First of all, the use of phase information could be exploited in combination with amplitudes and frequencies. In [RD06], the number of *phase reversals* is suggested as a feature. However, a more intuitive measure could be suggested. In [SHE+09], the notion of relative phase shift (RPS) is revisited and a phase structure is shown to be revealed through RPS. It would be interesting to examine if there are different patterns in RPS structures that can help discriminate emotional content in speech, combined with the standard amplitude and frequency features.

Secondly, sinusoidal amplitudes provide an implicit information about the spectral envelope, and they have been shown to be important in emotion recognition [DH99, HW96, HWA94]. Nevertheless, when considering only a part of the full-band, such as the 10 highest spectral peaks, a significant part of the spectrum is not taken into account. The inclusion of that part may contribute to better recognition percentages. Moreover, higher frequency components were suggested to be disregarded in sinusoidal model-based emotion classification as inappropriate for the task [RD06]. However, the aSMs are able to follow the dynamics of speech in the upper bands, and thus to reveal the spectral details that are blurred due to the time-frequency trade-off of the FFT-based estimation.

Furthermore, vowels have received increasing attention when it comes to emotion recognition, however consonants are shown to be important as well (see for example [BVN10]). Since our model is full-band and models both voiced and unvoiced parts of speech using AM-FM components, it would be interesting to examine whether there is any useful information embedded in the sinusoidal representation of consonants that is able to distinguish emotions. Finally, different classifiers can be used, such as HMMs, SVMs, or GMMs, for a more efficient classification.

7.5 Conclusions

In this chapter, we presented an application of an adaptive sinusoidal model, named eaQHM, on the problem of emotional speech analysis and classification and compared it to the standard Sinusoidal Model. It was shown that different emotional speech styles can be effectively represented by the adaptivity mechanism of eaQHM, yielding very accurate AM-FM decomposition. This was demonstrated through resynthesis of the original speech signal from its AM-FM components and by evaluating the Signal-to-Reconstruction Error (SRER). A formal listening test was designed to evaluate the perceptual quality of the resynthesized speech and showed that eaQHM-resynthesized emotional speech is indistinguishable from the original. The instantaneous amplitude and frequency were used as features for the classification. Results showed that a Vector Quantization classification based on eaQHM achieves higher classification scores for a subset of the SUSAS database, both on single-feature classification based on the sinusoidal parameters and on their combination. Future work will focus on different classifiers, phase parameter exploitation, and transforming neutral speech into emotional.

Chapter 8

Conclusions and Future Work

8.1 Overview

In this work, we have presented the adaptive Sinusoidal Models (aSMs), with applications in speech modifications, speech classification, and musical instrument sound analysis. The focus of the thesis has been on the extended adaptive Quasi-Harmonic Model (eaQHM), which was introduced, thoroughly described, and evaluated for the aforementioned applications. The eaQHM has been shown to provide transparent speech quality and high Signal-to-Reconstruction Error Ratios (SRERs) by decomposing speech into AM-FM components.

First, the performance of the eaQHM was tested on analysis and resynthesis of speech. Hybrid and full-band systems were presented based on the eaQHM and the recently proposed adaptive Harmonic Model (aHM). According to this distinction, hybrid systems included the *adaptive Harmonic + Noise Model - aHNM* and the *extended adaptive Quasi-Harmonic + Noise Model - eaQHNM*, whereas full-band systems included the *adaptive Harmonic Model - aHM* and the *extended adaptive Quasi-Harmonic Model - eaQHM*, i.e. there was no noise component to model stochastic parts of speech. Full-band systems do not require voicing decision, which is often a difficult task, thus alleviating the overall complexity of the system. Results showed that the full-band eaQHM provides transparent speech quality and stands among the competition.

Hybrid systems based on the aSMs showed to be efficient in speech modifications, such as time and pitch scaling. The separate manipulation of components (deterministic, stochastic) provided convenience and flexibility, and modified speech turned out to be of high quality, given that voicing decision is well estimated. On the other hand, full-band systems were shown to provide equally high quality. Compared to the state-of-the art, the aSMs have performed better than well known methods, such as the Harmonic+Noise Model (HNM) and the Waveform Similarity Overlap-Add (WSOLA), and are comparable to the mostly used high quality method for speech modifications, named STRAIGHT. Moreover, the advantages of the aSMs include flexibility, simplicity, and compactness of representation.

The application of the aSMs in modelling musical instrument sounds was presented in this thesis. It was proved that eaQHM can outperform the standard Sinusoidal Model (SM) and the recently suggested Exponentially Damped Sinusoidal Model (EDSM) in terms of analysis and representation of the oscillatory, transient, and sustain behaviour of musical instrument sounds. Preliminary residual analysis has also shown that the "filtered white noise" approach in residual modelling of sounds is not attained by stationary sinusoidal models, such as the SM. On the contrary, the eaQHM was shown to leave a residual that when represented by filtered white noise, it is perceptually closer to the original residual signal, thus showing there is no oscillatory information left in it.

Finally, the eaQHM has been applied on the task of emotional speech analysis and classification. Although sinusoidal models have not been used much in the emotion classification literature, we first showed that the eaQHM outperforms the standard SM in analyzing and resynthesizing emotional speech in terms of reconstruction and perceptual quality. Then, the instantaneous parameters obtained from the analysis were used for the classification task, yielding higher classification scores for the eaQHM than for the SM.

8.2 Future Research Directions

Regarding the adaptive models themselves, further improvements and applications can be considered for the future. Improvements may include the complexity reduction of the models. Due to the adaptivity process, the computational time for all models is increased. Methods and techniques can be applied to increase speed. Also, other approaches for instantaneous phase estimation can be used and evaluated instead of frequency integration. Such an approach is cubic polynomials. Furthermore, the current convergence criterion for the eaQHM is based on the rate of increment of the

Signal-to-Reconstruction-Error Ratio (SRER) computed on the whole waveform, over successive adaptations. This way, some frames may not be optimally reconstructed in terms of their local SRER, that is, they need further adaptations to locally converge to their optimal SRER. To solve this, a frame-based convergence criterion should be imposed in a computationally efficient way. Finally, the Least-Squares-based parameter estimation scheme limits the application of the models to monophonic speech, an extension to polyphonic speech can be attempted, using more robust estimation schemes, such as Total Least Squares.

From an applications perspective, Text-to-Speech (TTS) synthesis using Unit-Selection from sinusoidal representations has been successfully suggested. A sinusoidal model which provides more accurate signal representation such as eaQHM can lead to more naturally sounding speech units. Moreover, sinusoidal models in statistical TTS are less common in the literature. HMM-based speech synthesis has gained increased attention over the last years, especially when low footprint and general speech domain are required. High resolution models such as eaQHM can provide parameters that can be proved useful. Another application is Voice Conversion (VC). VC is similar to speech modification but the modifications are with respect to a target speaker. Statistical methods such as GMMs have been successfully applied in VC using parameters from sinusoidal models. Furthermore, speech coding can be applied in the aSMs, and especially in the aHM, which provides almost transparent speech quality with less parameters than the eaQHM.

On speech modifications, there is room for improvement on pitch shifting. It is suggested that during pitch shifting, there is some sort of interaction between the vocal tract and the glottal source. In all state-of-the-art systems, this is not taken into account for. That is, the vocal tract remains unaffected during pitch shifting. Further research can be made on how the vocal tract changes according to different pitch values, and correlate it with formant frequencies and bandwidths. This will lead to perceptually better modification schemes. Additionally, some parts of speech need to be protected from time or pitch scaling both in full-band and in hybrid systems, such as stop sounds for time-scaling and unvoiced speech for pitch-scaling. A time-varying modification factor based on the characteristics of speech would improve naturalness even more. Finally, spectral envelopes other than DAP (which is computationally intensive due to the iteration process) can be tested and evaluated.

On musical instrument sounds analysis, the modification schemes discussed in this thesis can be applied to produce artificially time-stretched or pitch-shifted counterparts of the musical instrument sounds. Another very important application in the audio processing domain is audio coding. Since the eaQHM yields highly accurate estimates of the instantaneous components of a signal, then coding these parameters will probably result in better quality with the same bitrates compared to the state-of-the-art.

On emotional speech analysis, recognition, and classification, there is vast room for improvements. There is considerable uncertainty as to the best feature set for classifying emotional data. At first, more comparisons should be made, especially with classifications schemes that use MFCCs, which are the most common set of features when it comes to speech recognition. The advantage of sinusoidal models is that the instantaneous parameters - which are used as features for the recognition - are jointly "optimal" in the sense that they accurately represent the dynamics of speech over time and frequency using a reconstruction criterion. Thus, a sinusoidal approach offers a set of features that come from a single estimator, rather than many different ones, as is the case in the literature (MFCCs, energy, pitch, speaking rate, and statistics on them). Moreover, there is still a debate on which classifiers to use for the classification: Support Vector Machines (SVM), Vector Quantizers (VQ), Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Neural Networks (NN) are the most dominant candidates. However, the most interesting research direction is using the aSMs for statistical emotional speech synthesis and transformations between different emotions.

Part IV

Appendices

Appendix A

A Residual Analysis of Musical Instrument Sounds from Sinusoidal Modeling

A.1 Introduction

Sinusoidal modeling stands out among the models used to represent [SS90, NMB07, LMR07, DGR93] and transform musical instrument sounds [SB98, ABLS02, LSI98] due to the fidelity and flexibility of the representation. In essence, sinusoidal analysis models each partial with a time-varying sinusoid, capturing temporal variations in amplitude, frequency and phase (the parameters of the model). Sinusoidal modeling is considered to represent musical instrument sounds well because most musical instruments are designed to present very clear modes of vibration. However, there is noise present in virtually all musical instrument sounds, such as breathing noise in woodwinds or mechanical noise like the hammer striking the piano strings.

There have been improvements in sinusoidal modelling to address issues such as partial tracking [NMB07, LMR07, DGR93], transient modelling [VM00, LSI98], to augment the accuracy of parameter estimation as well as the temporal resolution by adapting partials trajectories inside the analysis window [DS12, KPRS12]. Nevertheless, the lack of noise is perceptually noticeable in the sinusoidal representation of musical instrument sounds [Goo96, DQ97]. Serra [Ser97] proposed to subtract the sinusoidal component (i.e., the result of sinusoidal analysis) from the original recording to estimate a "residual component". This residual is, by definition, whatever is left from sinusoidal modeling, and therefore, commonly assumed to be noise not captured by the sinusoidal model (usually because sinusoids are not a compact representation of noise). Considerably less effort has been made in residual modeling. It has become standard practice [Ser97, Goo96, DQ97] to model the residual component by filtering white noise with a time-varying filter that emulates the spectral characteristics of the residual signal. Naturally, there are different ways to model the spectral distribution of energy of the residual component. The basic assumption is that the residual signal does not contain perceptually relevant information in the phase spectrum, only in magnitude. Therefore, "psychoacoustic" filter banks are usually found in residual modelling [LSI98, Goo96]. Goodwin [Goo96] uses the short-time energy in equivalent rectangular bands (ERBs) of the magnitude spectrum for both the analysis and synthesis stage, and justifies stating that the ear is insensitive to energy distributions within each ERB. Levine [LSI98] uses Bark bands instead. Resynthesis commonly uses a piece-wise constant spectrum with magnitudes from the ERB (or Bark bands) energy and random phase. Goodwin remarks that temporal phase correlations can control the texture of the modelled residual, which has been studied further to synthesize environmental sounds (e.g., running water or crackling fire) [AE03]. Ding [DQ97] proposes to use multi-pulse excitation linear prediction (MPLP) to keep phase coherence with the sinusoidal component.

There have been no formal investigations on the filtered white noise model for residual from sinusoidal modelling of musical instrument sounds. In this work we present a systematic evaluation of how well filtered white noise models the residual from sinusoidal modelling of musical instrument sounds for different sinusoidal modelling algorithms. Each algorithm captures oscillatory behaviour differently and, consequently, leaves (perceptually) different residuals. We performed a subjective listening test to evaluate the perceptual similarity between filtered white noise and the residual of each sinusoidal algorithm. Then we use an objective measure of similarity to compare with the perceptual assessments. The next section briefly reviews the sinusoidal modelling algorithms used in this investigation. Next, we describe the framework used to decompose the musical instrument sounds into the blocks used in the evaluation, which is followed by a discussion and the conclusions and future perspectives.

A.2 Sinusoidal Modelling

Conceptually, traditional sinusoidal modelling supposes that the musical instrument sounds being modelled can be decomposed into quasi-harmonic oscillations and additive noise. In practice, the musical instrument sound $y(t)$ is separated into a sinusoidal component $y_s(t)$ plus a residual component $y_r(t)$, where $y_r(t)$ is obtained by subtraction of the purely sinusoidal component $y_s(t)$ from the original sound $y(t)$. The sinusoidal component is further represented as

$$y_s(t) = \left[\sum_{k=0}^{K} \alpha_k e^{j2\pi t f_k}\right] w(t) \tag{A.1}$$

where α_k and $\phi_k(t) = 2\pi f_k t$ are respectively the amplitude and phase of the k^{th} sinusoid inside the analysis window $w(t)$, and K is the number of sinusoids. The model assumes that the sinusoids describe stable partials of the sound so their parameters do not vary significantly inside the analysis window. Traditionally [Ser97], the parameters of the model α_k and $\phi_k(t)$ are estimated for each frame of the short-time Fourier transform, limiting the temporal resolution of the model to that of the STFT. In the rest of the text, SM stands for a sinusoidal model that imposes no restrictions on the frequencies of the partials [Ser97]. For most musical instrument sounds, a model where the sinusoids are harmonically related is a good approximation, giving rise to the Harmonic Model (HM) [Sty96], which uses sinusoids whose frequencies are multiple integers k of a fundamental frequency f_0 as $\phi_k(t) = 2\pi t k f_0$.

There have been proposals to improve the temporal resolution of the sinusoidal model by adapting the estimation of the parameters of the sinusoids *inside* the analysis window, resulting in *adaptive* sinusoidal models. In particular, the adaptive Harmonic Model (aHM) [DS12] used in this work modulates the frequency of each sinusoid inside the analysis window upon resynthesis. Recently, the extended adaptive Quasi-Harmonic Model (eaQHM) was developed [KPRS12]. The eaQHM algorithm adapts both the amplitudes and frequencies of the sinusoidal partials inside the analysis window, therefore it can be considered a full AM/FM model, as shown below

$$y_s(t) = \left[\sum_{k=0}^{K} \alpha_k(t)\, e^{j\phi_k(t)}\right] w(t)\,, \tag{A.2}$$

where $\alpha_k(t)$ denotes the time-varying amplitude and $\phi_k(t)$ denotes the instantaneous phase function of the k^{th} component inside the analysis window $w(t)$. Table A.1 summarizes the temporal representation of frequencies for the analysis and synthesis stages for the sinusoidal algorithms used.

	Analysis	**Synthesis**
SM	stationary	stationary (OLA)
HM	stationary	Splines
aHM	adaptive	Splines
eaQHM	adaptive	Splines

Table A.1: *Comparison of representations of frequency components for the analysis and synthesis stages of the sinusoidal algorithms used.*

A.3 Residual Modelling

The residual component $y_r(t)$ is modeled as

$$\hat{y}_r(t) = \int_0^t a(t-\tau)\, u(\tau)\, d\tau \tag{A.3}$$

where $\hat{y}_r(t)$ is the modeled residual component, $u(\tau)$ is white noise and $a(t,\tau)$ is the response of a time-varying filter. Serra [Ser97] wrote that "a stochastic, or noise, signal is fully described by its power spectral density which gives the expected signal power versus frequency. When a signal is assumed stochastic, it is not necessary to preserve either the instantaneous phase or the exact magnitude details of individual FFT frames," justifying the assumption that the residual component can be modeled as filtered white noise. There have been different proposals to estimate the filter $a(\tau)$ [Ser97, Goo96, DQ97]. In this work, we estimate the spectral envelope of each frame of the STFT of the residual component $y_r(t)$ using linear prediction (LPC) [Mak75] and use it as the time-varying filter coefficients, as has been previously proposed for speech [Sty96]. LPC is adequate for spectral envelope estimation of $y_r(t)$ because it tends to follow the average energy of noisy spectra rather than the peaks. Using Eq. (A.3) the model suposes that if we inverse

filter $y_r\,(t)$, we should obtain white noise (a signal with flat magnitude and no temporal phase coherence or random phase). In this work, we investigate if filtered white noise is perceptually close to the original residual signals with a listening test and further investigate if the inverse filtered residual component presents the characteristics of white noise with an objective measure based on the autocorrelation function.

A.4 Experimental Framework

Figure A.1 illustrates the steps of the experimental framework. Each musical instrument sound $y\,(t)$ is decomposed into sinusoidal $y_s\,(t)$ and residual $y_r\,(t)$ using the *SM*, the *HM*, the *aHM*, and the *eaQHM*. Each component, $y_s\,(t)$ and $y_r\,(t)$, is modeled with linear prediction, resulting in a time-varying spectral envelope $A_s\,(z)$ and $A_s\,(z)$ and an inverse filtered (whitened) signal $\bar{y}_s\,(t)$ and $\bar{y}_r\,(t)$, which are the prediction errors [Mak75]. In the listening test, we use white noise filtered with $A_r\,(z)$. The objective similarity measure compares $\bar{y}_r\,(t)$ with $\bar{y}_s\,(t)$ and $u\,(t)$.

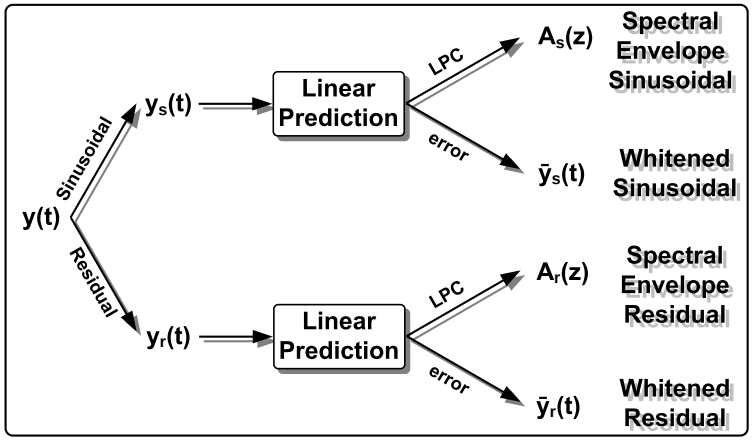


Figure A.1: *Illustration of the signal decomposition.*

Strings	Brass	Woodwinds
Double Bass	Bass Trombone	Bass Clarinet
Cello	Bass Trumpet	Bassoon
Viola	Cimbasso	Clarinet Bb
	Contrabass Tuba	English Horn
	Tenor Trombone	
	Tuba	
	Wagner Tuba	

Table A.2: *Musical instrument sounds used in the listening test.*

Table A.2 lists the 14 musical instruments used. The pitch of all sounds is $C3 \simeq 131$ Hz, the *dynamics* is *forte*, and the duration is under 2s. All sinusoidal algorithms used a window size equivalent to 3 times the period of the fundamental frequency $f_0 \simeq 131$ Hz, 50% overlap, and size of the FFT 4 times the window size. The linear prediction order used was 50 for both $y_s\,(t)$ and $y_r\,(t)$ to avoid smearing possible oscillatory energy left in $y_r\,(t)$ (missed by the sinusoidal model).

A.5 Evaluation

The evaluation consists of a listening test and an objective measure based on the autocorrelation function. However, firstly we estimate the residual energy to compare how well each sinusoidal algorithm models the musical instrument sounds. The less residual energy, the better the algorithm captured the oscillatory behavior. The signal to reconstruction error ratio (SRER) shown in Eq. (A.4) measures the ratio between the total energy and the energy in the residual component $y_r\,(t)$. The higher the ratio, the less residual energy there is in $y_r\,(t)$.

$$SRER = 20\log_{10}\frac{\sigma_{y(t)}}{\sigma_{y_r(t)}} \tag{A.4}$$

where $y(t)$ is the original signal, $\sigma(\cdot)$ is the standard deviation operator, and $y_r(t)$ is the residual component. Table A.3 shows the average SRER in dB across musical instrument sounds for each method, revealing that the eaQHM has a higher SRER than all other methods by roughly 15 dB.

SRER (dB)			
SM	HM	aHM	eaQHM
33.86	34.84	36.53	50.62

Table A.3: *Average Signal to Reconstruction Error Ratio (SRER) across musical instrument sounds.*

A.5.1 Listening Test

The purpose of the listening test is to evaluate the perceptual similarity between the residual signal $y_r(t)$ and its filtered-white-noise counterpart $\hat{y}_r(t)$ for the 14 musical instrument sounds listed in Table A.2 modeled with the four sinusoidal algorithms shown in Table A.1. For each participant, the listening test presented a subset of 16 pairs of sounds corresponding to $y_r(t)$ and $\hat{y}_r(t)$ from 4 musical instruments (times 4 algorithms) in random order to minimize cross comparison among methods. All sounds were normalized at -16dB RMS. The listener is instructed to listen to each pair as many time as they want and rate their perceptual similarity in a scale from 1 to 5 labeled with the terms 1) *Very different*, 2) *Different*, 3) *Fairly similar*, 4) *Very similar*, 5) *Identical*. The test can be found on-line [1]. Figure A.2 shows the result for 51 participants aged between 22 and 67, depicting the mean opinion score (MOS) and 95% confidence interval. In average, the eaQHM results in a residual signal that was considered between *fairly similar* and *very similar* to its filtered white noise counterpart. The other 3 algorithms (SM, HM, and aHM) produced residuals whose filtered white noise counterparts were considered practically *different*.

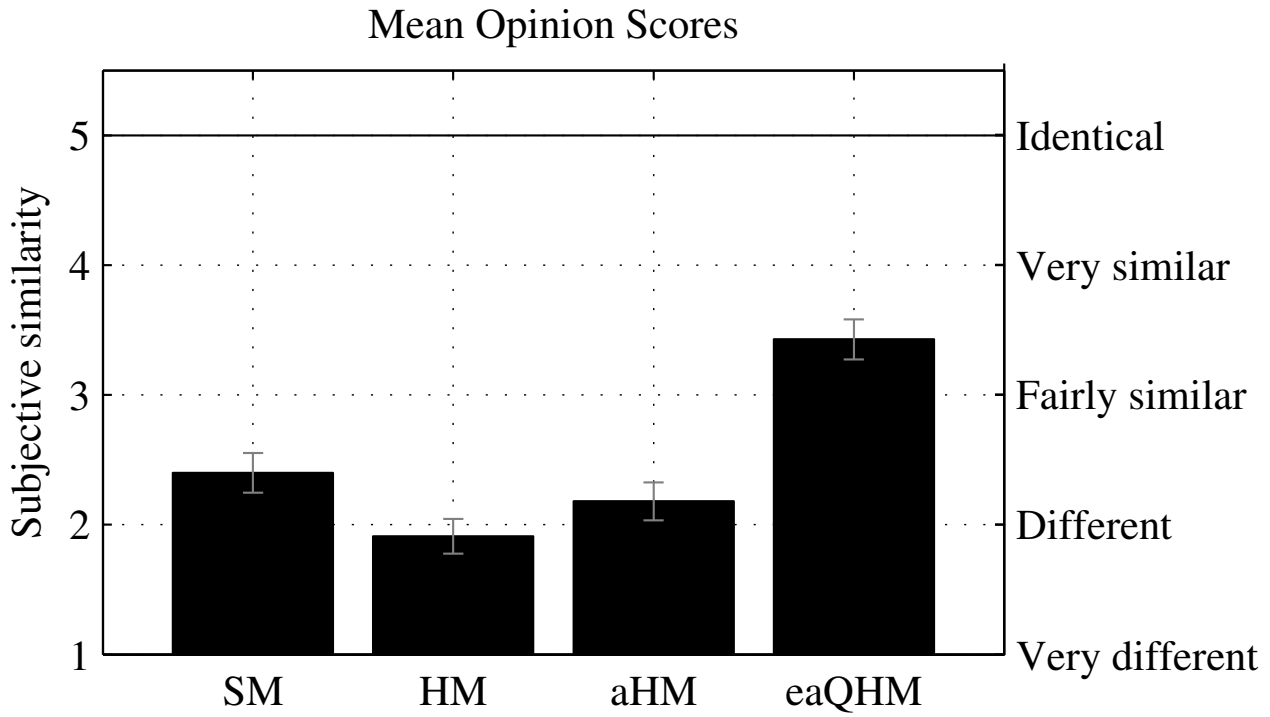


Figure A.2: *Result of the listening test. The figure shows the mean opinion score (MOS) and 95 % confidence interval for the four sinusoidal models tested.*

A.5.2 Objective Measure

The result of the listening test indicates that, in general, filtered white noise was not considered a perceptually similar representation of $y_r(t)$. However, the listening test gives no further evidence to help explain why. Ideally, we would like to identify what remains in the residual signal that departs from the conceptual filtered-white-noise hypothesis. In

[1] http://gillesdegottex.eu/ExCaetano2013simil

the listening test, the perceptual effect of the LPC spectral envelope of $y_r(t)$ is present in $\hat{y}_r(t)$. Thus we assume that the differences lie elsewhere, in the spectral fine structure or in the phase spectrum. To evaluate the importance of the fine structure between $y_r(t)$ and $\hat{y}_r(t)$, we compare the whitened residual component $\bar{y}_r(t)$ with the whitened sinusoidal component $\bar{y}_s(t)$ and with the model (i.e., white noise $u(t)$) with an objective similarity measure. We use the autocorrelation functions, shown in (A.5), which should provide a unique representation of both the white noise (zero except at zero lag) and the sinusoidal component (peaks at multiple integers of the fundamental frequency).

$$R(i) = \sum_{n=0}^{N-1-i} y(n)\, y(n-i) \tag{A.5}$$

The similarity measure is then the dot product between the autocorrelation functions, given by $\cos(\Theta\{\bar{y}_r, u\}) = R_{\bar{y}_r}(i) \cdot R_u(i)$ and $\cos(\Theta\{\bar{y}_r, \bar{y}_s\}) = R_{\bar{y}_r}(i) \cdot R_{\bar{y}_s}(i)$. The dot (or inner) product can be interpreted as the projection of $R_{\bar{y}_r}(i)$ onto $R_{\bar{y}_s}(i)$ and $R_u(i)$. Thus Θ is the angle between the autocorrelation functions interpreted as vectors, and it varies from 0 (identical) to 90° (orthogonal). Table A.4 shows the average of these values across all musical instruments to allow comparison per method. Following Fig. (A.2), we expected the eaQHM to give a significantly smaller $\Theta\{\bar{y}_r, u\}$ and larger $\Theta\{\bar{y}_r, \bar{y}_s\}$.

	SM	**HM**	**aHM**	**eaQHM**
$\Theta\{\bar{y}_r, u\}$	46.11°	51.63°	49.83°	50.95°
$\Theta\{\bar{y}_r, \bar{y}_s\}$	61.46°	67.25°	68.85°	67.48°

Table A.4: *Average angle in degrees across musical instrument sounds for each algorithm.*

A.6 Discussion

The extended adaptive Quasi-Harmonic Model (eaQHM) is tested to confront the notorious pre-echo effect in sinusoidal modelling and it is shown that highly accurate, pre-echo-free representations of percussive sounds are possible using the adaptive approach. Results on a database of percussive sounds such as plucked strings and percussion instruments show that, on average, the eaQHM improves by over 30 dB the Signal to Reconstruction Error Ratio (SRER) obtained by the standard sinusoidal model. A listening test showed that the percussive sounds modelled by the eaQHM are perceptually closer to the original recordings than the same sounds represented by a traditional sinusoidal model for more than 80% of the listeners in all cases.

We also notice that each sinusoidal modelling algorithm resulted in a different perceptual similarity, revealing that different algorithms leave different undesired information in the residual signal y_r. Therefore we suspect that there might be some oscillatory behaviour left in y_r. In other words, some sinusoidal modelling algorithms fail to capture all oscillatory energy such as frequency modulations or transients. The models that use slowly varying sinusoids (stable oscillations) plus additive noise might oversimplify the complexity of musical sounds. It has already been remarked [LSI98] that *sinusoids* plus *noise* plus *transients* might be a more realistic representation for musical instrument sounds. However, transients are characteristically present mostly during the attack, but there is no indication that the participants used the attack as perceptual cue. The listening test shows that the AM/FM modelling of the eaQHM captures most oscillatory energy, including transients.

On the other hand, Table A.4 reveals no significant difference across algorithms. The angles Θ do indicate that $\bar{y}_r$ is closer to u (white noise) than to $\bar{y}_s$ (sinusoidal) for all algorithms. But the similarities measured by Θ do no explain the results of the listening test. Our interpretation of this result is that the perceptual differences found in the listening test cannot be explained by fine spectral structure, rather, by phase coherence or transients.

Interestingly, one of the participants of the listening test remarked that, for each pair, one of them always sounded *brighter*. Indeed, $\hat{y}_r$ has more energy in high frequencies because pure white noise has a flat spectrum where energy is not equal per octave (let alone per ERB or Bark band). A possible course of investigation would be to use different types of noise (prior to applying the time-varying spectral envelope) to correctly balance the spectral energy, such as *pink* noise. Future perspectives also include using the eaQHM in transient detection and transient modelling for musical instrument recognition, segmentation, and sound transformations such as timbral variations, perceptually coherent time stretching and pitch shifting.

A.7 Conclusions and Future Perspectives

We presented a systematic investigation of the filtered white noise model for the residual from sinusoidal modelling of musical instrument sounds. Four different sinusoidal modelling algorithms were evaluated. We conducted a listening test and we developed an objective measure of spectral similarity. The listening test assessed the perceptual similarity between filtered white noise and the residual component for each sinusoidal algorithm. The results indicate that, in general, filtered white noise was considered *different* from the residual component. However, we determined that the eaQHM leaves a residual that is *fairly similar* to the filtered white noise counterpart. The objective measure compared the residual with both the sinusoidal component and their modelled counterpart across algorithms using the autocorrelation functions. The objective evaluation aimed to investigate the reason for the result of the listening test, trying to indicate whether there was "sinusoidal" energy left in the poorly modelled residuals, for example. The objective similarity measure did not indicate that the perceptual differences found can be explained by comparing spectral fine structure. However, the autocorrelation function only includes information from the power spectral density. Thus we suspect that the differences lie in the phase spectrum (possibly due to temporal phase coherence) or transients in the residual, confirming the conclusion of previous studies [Goo96, LSI98].

Future perspectives include using the eaQHM in transient detection and transient modelling for musical instrument recognition, segmentation, and sound transformations such as timbral variations, perceptually coherent time stretching and pitch shifting. Moreover, we should focus on determining the reason for the difference between the conceptual model of filtered white noise and what current sinusoidal modelling algorithms fail to model. The EDS model might be useful in such an analysis, and is planned to be included in the tests. Perspectives include using "colored" noise to correct the high-frequency energy content perceived as brightness (or some other more sophisticated psychoacoustic model). Further investigation on the temporal phase coherence should develop a measure for analysis and comparison with the sinusoidal component. Attack transients might account for some of the perceptual difference we found for most sinusoidal algorithms.

Appendix B

Publications

During this work, the following publications took place (in chronological order):

1. **Conferences**
   (a) **Kafentzis G. P.**, Pantazis Y., Rosec O., Stylianou Y.,
   An Extension of the Adaptive Quasi-Harmonic Model,
   In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012
   (b) **Kafentzis G. P.**, Rosec O., Stylianou Y.,
   On the Modeling of Voiceless Stop Sounds of Speech using Adaptive Quasi-Harmonic Models,
   In Conference of International Speech Communication Association (INTERSPEECH), 2012
   (c) **Kafentzis G. P.**, Degottex G., Rosec O., Stylianou Y.,
   Time-scale Modifications based on an Adaptive Harmonic Model,
   In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013
   (d) Caetano M., **Kafentzis G. P.**, Mouchtaris A., Stylianou Y.,
   Adaptive Sinusoidal Modeling of Percussive Musical Instrument Sounds,
   In European Signal Processing Conference (EUSIPCO), 2013
   (e) Caetano M., **Kafentzis G. P.**, Degottex G., Mouchtaris A., Stylianou Y.,
   Evaluating How Well Filtered White Noise Models the Residual from Sinusoidal Modeling of Musical Instrument Sounds,
   In Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013
   (f) **Kafentzis G. P.**, Degottex G., Rosec O., Stylianou Y.,
   Pitch-scale Modifications based on an Adaptive Harmonic Model,
   In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014
   (g) **Kafentzis, G. P.**, Rosec O., Stylianou Y.,
   Robust Full-Band Adaptive Sinusoidal Analysis and Synthesis of Speech,
   In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014
   (h) **Kafentzis, G. P.**, Yakoumaki T., Mouchtaris A., Stylianou Y.,
   Analysis of Emotional Speech using an Adaptive Sinusoidal Model,
   In European Signal Processing Conference (EUSIPCO), 2014
   (i) Yakoumaki T., **Kafentzis G. P.**, Stylianou Y.,
   Emotion Classification using adaptive Sinusoidal Modeling,
   In Conference of International Speech Communication Association (INTERSPEECH), 2014
2. **Journals**
   (a) Caetano M., **Kafentzis G. P.**, Mouchtaris A., Stylianou Y.,
   Adaptive Sinusoidal Modeling of Musical Instrument Sounds,
   In IEEE Transactions in Acoustics, Speech, and Language Processing (TASLP), 2014, under review.
   (b) **Kafentzis G. P.**, Rosec O., Stylianou Y.,
   Adaptive Sinusoidal Analysis, Synthesis, and Modifications of Speech,
   In IEEE Transactions in Acoustics, Speech, and Language Processing (TASLP), to be submitted.

Year	IEEE ICASSP	ISCA Interspeech	EURASIP EUSIPCO	IEEE WASPAA	IEEE Journals
2012	(1a)	(1b)	-	-	-
2013	(1c)	-	(1d)	(1e)	-
2014	(1f,1g)	(1i)	(1h)	-	(2a,2b)

Table B.1: *Publications over the years of the thesis*

Bibliography

[Abe13] J. Abeger. Automatic string detection for bass guitar and electric guitar. In Mitsuko Aramaki, Mathieu Barthet, Richard Kronland-Martinet, and SG'Elvi Ystad, editors, *From Sounds to Music and Emotions*, volume 7900 of *Lecture Notes in Computer Science*, pages 333–352. Springer Berlin Heidelberg, 2013.

[ABLS02] X. Amatriain, J. Bonada, A. Loscos, and X. Serra. Spectral processing. In Udo Zolzer, editor, *DAFX - Digital Audio Effects*, chapter 10, pages 373–438. John Wiley and Sons, 2002.

[AE03] M. Athineos and D. P. W. Ellis. Sound texture modelling with linear prediction in both time and frequency domains. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003.

[AF95] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *Signal Processing, IEEE Transactions on*, 43(5):1068–1089, 1995.

[AH71] B. Atal and S. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of Acoustical Society of America (JASA)*, 50:637–655, 1971.

[AKK07] M. M. H. El Ayadi, M. S. Kamel, and F. Karray. Speech emotion recognition using Gaussian Mixture Vector autoregressive models. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 957–960, 2007.

[Alk92] P. Alku. Glottal Wave Analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication*, 11:109–118, 1992.

[Alk11] P. Alku. Glottal inverse filtering analysis of human voice production : a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana - Academy Proceedings in Engineering Sciences*, 36:623–650, 2011.

[AR98] N. Amir and S. Ron. Toward an automatic classification of emotions in speech. *International Conference on Spoken Language Processing*, pages 555–558, 1998.

[AR08] Y. Agiomyrgiannakis and O. Rosec. Towards Flexible Speech Coding for Speech Synthesis: an LF + Modulated Noise Vocoder. In *Interspeech*, Brisbane, Australia, September 2008.

[AR09] Y. Agiomyrgiannakis and O. Rosec. ARX-LF-based source-filter methods for voice modification and transformation. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009.

[AS84] L. B. Almeida and F. M. Silva. Variable-frequency synthesis: an improved harmonic coding scheme. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 1:2751–2754, 1984.

[Ass03] The ITU Radiocommunication Assembly. ITU-R BS.1284-1: EN-general methods for the subjective assessment of sound quality. Technical report, ITU, 2003.

[BA09] C. Magi J. Pohjalainen T. Backstrom and P. Alku. Stabilised Weighted Linear Prediction. *Speech Communication*, 51:401–411, 2009.

[BBD02] R. Badeau, R. Boyer, and B. David. EDS parametric modeling and tracking of audio signals. In *Proc. of 5th Int. conf. on Digital Audio Effects (DAFx02*, pages 26–28, 2002.

[BDA+05] J.P. Bello, L. Daudet, S. Abdullah, C. Duxbury, M. Davies, and M. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5), September 2005.

[BDDD05] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit. Zeros of Z-Transform Representation with Application to Source-Filter Separation in Speech. *IEEE Signal Processing Letters*, 12:344–347, 2005.

[BDR04] R. Badeau, B. David, and G. Richard. Selecting the modeling order for the esprit high resolution method: an alternative approach. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, volume 2, pages ii–1025–8, 2004.

[BGH00] S.E. Bou-Ghazale and J.H.L. Hansen. A comparative study of traditional and newly proposed features for recognitionof speech under stress. *IEEE Transactions on Speech and Audio Processing*, 8(4):429–442, July 2000.

[BH09] R. Bader and U. Hansen. Modeling of musical instruments. In David Havelock, Sonoko Kuwano, and Michael Vorländer, editors, *Handbook of Signal Processing in Acoustics*, pages 419–446. Springer New York, 2009.

[BMC05] J. Benesty, S. Makino, and J. Chen. *Speech Enhancement*. Signals and Communication Technology. Springer, 2005.

[BN08] Murtaza Bulut and Shrikanth Narayanan. On the robustness of overall f0-only modifications to the perception of emotions in speech. *Journal of Acoustical Society of America (JASA)*, 123(6):4547–4558, 2008.

[Bos13] L. T. Bosch. Emotions, speech, and the ASR framework. *Speech Communication*, 40:213–225, 2013.

[Bro99] Judith C. Brown. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *The Journal of the Acoustical Society of America*, 105(3):1933–1941, 1999.

[BVN10] D. Bitouk, R. Verma, and A. Nenkova. Class-level spectral features for emotion recognition. *Speech Communication*, 52(7-8):613–625, 2010.

[BWG04] M.W. Bhatti, Y. Wang, and L. Guan. A neural network approach for human emotion recognition in speech. In *Proc. of the International Symposium on Circuits and Systems*, volume 2, pages II–181–4 Vol.2, 2004.

[CBC96] A. Castellanos, J. M. Benedi, and F. Casacuberta. An analysis of general acoustic - phonetic features for spanish speech produced with lombard effect. *Speech Communication*, 20:23–36, 1996.

[CCH89] K. E. Cummings, M. A. Clements, and J. H. L. Hansen. Estimation and comparison of the glottal source waveform across stress styles using glottal inverse filtering. *In Proceedings of IEEE Southeastcon*, pages 776–781, 1989.

[CH94] D. A. Cairns and J.H.L. Hansen. Nonlinear analysis and classification of speech under stressed condition. *Journal of Acoustical Society of America (JASA)*, pages 3392–3400, 1994.

[CH08] A. Camacho and J. G. Harris. A sawtooth waveform inspired pitch estimator for speech and music. *Journal of Acoustical Society of America (JASA)*, 124:1628–1652, 2008.

[CKD+13] M. Caetano, G. Kafentzis, G. Degottex, A. Mouchtaris, and Y. Stylianou. Evaluating how well filtered white noise models the residual from sinusoidal modeling of musical instrument sounds. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[CKMS] M. Caetano, G. P. Kafentzis, A. Mouchtaris, and Y. Stylianou. Adaptive sinusoidal modeling of musical instrument sounds. *IEEE Transactions on Audio, Speech, and Language Processing*. under review.

[CKMS13] M. Caetano, G. P. Kafentzis, A. Mouchtaris, and Y. Stylianou. Adaptive sinusoidal modeling of percussive musical instrument sounds. In *Proc. European Signal Processing Conference (EUSIPCO)*, 2013.

[CM96] O. Cappe and E. Moulines. Regularization techniques for discrete cepstrum estimation. *IEEE Signal Processing Letters*, 3:100–102, 1996.

[COCM01] M. Campedel-Oudot, O. Cappe, and E. Moulines. Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach. *IEEE Transactions on Speech and Audio Processing*, 9(5):469–481, Jul 2001.

[CR10] M. Caetano and X. Rodet. Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues. In *Proceedings of the Digital Audio Effects (DAFx)*, Graz, Austria, September 2010.

[CR12] M. Caetano and X. Rodet. A source-filter model for musical instrument sound transformation. In *Proceedings of the International Conference on Audio, Speech, and Signal Processing*, 2012.

[CS86] F. Charpentier and M. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 1:2015–2018, 1986.

[Dau06] L. Daudet. A review on techniques for the extraction of transients in musical signals. In *Proceedings of the International Conference on Computer Music Modeling and Retrieval*, 2006.

[Dau11] L. Daudet. Transients modelling by pruned wavelet trees. In *Proceedings of the International Computer Music Conference*, 2011.

[DBR13] O. Derrien, R. Badeau, and G. Richard. Parametric audio coding with exponentially damped sinusoids. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(7):1489–1501, 2013.

[DD97] B. Doval and C. D'Alessandro. Spectral correlates of glottal waveform models : an analytic study. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 446–452, 1997.

[DDH06] B. Doval, C. D'Alessandro, and N. Henrich. The Spectrum of Glottal Flow Models. *Acta Acustica*, 92:1009–1025, 2006.

[DDS01] C. Duxbury, M. E. Davies, and M. B. Sandler. Separation on transient information in musical audio using multiresolution analysis techniques. In *Proceedings of Digital Audio Effects (DAFx)*, 2001.

[Deg10] G. Degottex. *Glottal source and vocal-tract separation*. PhD thesis, UPMC-Ircam, France, 2010.

[DGR93] P. Depalle, G. Garcia, and X. Rodet. Tracking of partials for additive sound synthesis using hidden markov models. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 1993.

[DH99] B. D.Womack and J. H. L. Hansen. N-channel hidden markov models for combined stressed speech classification and recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 7:668–676, 1999.

[dK02] A. de Cheveigne and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[DKA95] W. Ding, H. Kasuya, and S. Adachi. Simultaneous estimation of vocal tract and voice source parameters based on an ARX model. *IEICE Transactions on Information Systems*, E78-D:738–743, 1995.

[DQ97] Y. Ding and X. Qian. Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (quasar) signal model. *Journal of Audio Engineering Society*, 45(7/8):571–584, 1997.

[DQ07] R. B. Dunn and T. F. Quatieri. Sinewave Analysis/Synthesis Based on the Fan-Chirp Transform. *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October, 2007.

[DQM09] R. B. Dunn, T. F. Quatieri, and N. Malyska. Sinewave parameter estimation using the fast fan-chirp transform. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 349–352, Oct 2009.

[DS12] G. Degottex and Y. Stylianou. A full-band adaptive harmonic representation of speech. In *Interspeech*, Portland, Oregon, U.S.A, 2012.

[DS13] G. Degottex and Y. Stylianou. Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2085–2095, 2013.

[DTCT03] C. Drioli, G. Tisato, P. Cosi, and F. Tesser. Emotions and voice quality: Experiments with sinusoidal modeling. In *In Proceedings of VOQUAL'A03*, pages 127–132, 2003.

[DWBH06] H. Deng, R.K. Ward, M.P. Beddoes, and M. Hodgson. A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):445–455, March 2006.

[EJM91] A. El-Jaroudi and J. Makhoul. Discrete All-Pole Modeling. *IEEE Transactions on Signal Processing*, 39:411–423, 1991.

[EK00] A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages II753–II756, 2000.

[ERD06] S. Essid, G. Richard, and B. David. Musical instrument recognition by pairwise classification strategies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1401–1412, 2006.

[Fan70] G. Fant. *Acoustic Theory of Speech Production*. Mouton De Gruyter, 1970.

[Fan95] G. Fant. The LF-model revisited. Transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2:119–156, 1995.

[Fed98] R. Di Federico. Waveform preserving time stretching and pitch shifting for sinusoidal models of sound. In *In Proceedings of the COST-G6 Digital Audio Effects Workshop*, pages 44–48, 1998.

[FF06] S. A. Fulop and K. Fitz. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *The Journal of the Acoustical Society of America*, 119(1):360–371, 2006.

[Fle99] N. H. Fletcher. The nonlinear physics of musical instruments. *Reports on Progress in Physics*, 62:723–764, 1999.

[FLG85] G. Fant, Q. Lin, and C. Gobl. Notes on Glottal Flow Interaction. *STL-QPSR*, pages 21–45, 1985.

[FLL85] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4:1–13, 1985.

[FM06] Q. Fu and P. Murphy. Robust Glottal Source Estimation based on Joint Source-Filter Model Optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:492–501, 2006.

[FR98] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer, New York, 23nd ed edition, 1998.

[FS66] G. Fant and B. Sonesson. Indirect studies of glottal cycles by synchronous inverse filtering and photoelectric glottography. *STL-QPSR*, pages 1–3, 1966.

[GG77] J. M. Grey and J. W. Gordon. Multidimensional perceptual scaling of musical timbre. *Journal of Acoustical Society of America (JASA)*, 61(5):1270–1277, 1977.

[Gio09] P. Giotis. Instrument timbre transformation using gaussian mixture models. Master's thesis, Universitat Pompeu Frabra, 2009.

[GL18] H. Gray and W. H. Lewis. *Anatomy of the Human Body*. Lea & Febiger, 1918.

[GL84] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32:236–243, 1984.

[GL88] D. W. Griffin and J. S. Lim. Multiband Excitation Vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(8):1223–1235, 1988.

[Goo96] M. Goodwin. Residual modeling in music analysis-synthesis. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 1005–1008, 1996.

[Goo97] M. Goodwin. Matching pursuit with damped sinusoids. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, volume 3, pages 2037–2040, 1997.

[GR90] T. Galas and X. Rodet. An improved cepstral method for deconvolution of source filter systems with discrete spectra: Application to musical sound signals. *Proceedings of the International Computer Music Conference (ICMC)*, pages 82–84, 1990.

[GRC12] E. Godoy, O. Rosec, and T. Chonavel. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1313–1323, 2012.

[Gri87] D. W. Griffin. *Multiband Excitation Vocoder*. PhD thesis, M.I.T, 1987.

[GS92] E. B. George and M. J. T. Smith. Analysis-by-Synthesis Overlap-Add Sinusoidal Modeling Applied to the Synthesis of Musical Tones. *Journal of the Audio Engineering Society*, 40:497–516, 1992.

[GS97] E. B. George and M. J. T. Smith. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Transactions on Speech and Audio Processing*, 5(5):389–406, 1997.

[Han95] S. Handel. Timbre perception and auditory object identification. In B.C.J. Moore, editor, *Hearing*, pages 425–461. Academic Press, New York, 1995.

[HBG97] J. Hansen and S. Bou-Ghazale. Getting started with SUSAS: A speech under simulated and actual stress database. *EUROSPEECH*, 4:1743 – 1746, 1997.

[HD10] X. Huang and L. Deng. An overview of modern speech recognition. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.

[HDC02] P. Hanna and M. Desainte-Catherine. Adapting the Overlap-Add Method to the Synthesis of Noise. *International Conference on Digital Audio Effects (DAFx)*, pages 101–104, 2002.

[HJA02] M. Hasegawa-Johnson and A. Alwan. *Speech Coding: Fundamentals and Applications*, volume 5, pages 2340–2359. John Wiley & Sons, Inc., December 2002.

[HKS+12] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertas, J. Sandberg, and M. Hansson-Sandsten. Comparing spectrum estimators in speaker verification under additive noise degradation. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 4769–4772, 2012.

[HPD03] P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32, 2003.

[HVK02] R. Heusdens, R. Vafin, and W.B. Kleijn. Sinusoidal modeling using psychoacoustic-adaptive matching pursuits. *Signal Processing Letters, IEEE*, 9(8):262–265, Aug 2002.

[HVL+05] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S. V. Huffel. Perceptual audio modeling with exponentially damped sinusoids. *Signal Processing*, 85(1):163–176, 2005.

[HW96] J. H. L. Hansen and B. Womack. Feature analysis and neural network based classification of speech under stress. *IEEE Transactions on Audio, Speech, and Language Processing*, 4:307–313, 1996.

[HWA94] J. H. L. Hansen, B. D. Womack, and L. M. Arsian. A source generator based production model for environmental robustness in speech recognition. *In Proc. ICSLP*, pages 1003 – 1006, 1994.

[IK93] P. Iverson and C. L. Krumhansl. Isolating the dynamic attributes of musical timbre. *The Journal of the Acoustical Society of America*, 94:2595, 1993.

[JH02] J. Jensen and R. Heusdens. A comparison of sinusoidal model variants for speech and audio representation. 2002.

[JHJ04] J. Jensen, R. Heusdens, , and S. Jensen. A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids. *IEEE Transaction on Speech and Audio Processing*, 12(2):121–132, March 2004.

[JJH99] J. Jensen, S.H. Jensen, and E. Hansen. Exponential sinusoidal modeling of transitional speech segments. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, volume 1, pages 473–476 vol.1, 1999.

[Kaw97] H. Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 1303–1306, Munich, Apr 1997.

[KCJL03] O. W. Kwon, K. Chan, J.Hao, and T. W. Lee. Emotion recognition by speech signals. In *EUROSPEECH*, pages 125–128, 2003.

[KDRS13] G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou. Time-scale Modifications based on an Adaptive Harmonic Model. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Vancouver, CA, May 2013.

[KDRS14] G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou. Pitch modifications of speech based on an adaptive harmonic model. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2014. In Press.

[KGV78] K. Kodera, R. Gendrin, and C. Villedary. Analysis of time-varying signals with small bt values. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):64–76, 1978.

[KK11] P. Khanna and M. Sasi Kumar. Application of vector quantization in emotion recognition from human speech. In *Information Intelligence, Systems, Technology and Management*, volume 141, pages 118–125. Springer Berlin Heidelberg, 2011.

[KM88] R. Kronland-Martinet. The wavelet transform for analysis, synthesis, and processing of speech and music sounds. *Computer Music Journal*, 12(4):11–20, 1988.

[KM95] I. Kaminsky and A. Materka. Automatic source identification of monophonic musical instrument sounds. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, volume 1, pages 189–194, 1995.

[KPRS12] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou. An Extension of the Adaptive Quasi-Harmonic Model. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Kyoto, March 2012.

[KRS13] G. P. Kafentzis, O. Rosec, and Y. Stylianou. On the Modeling of Voiceless Stop Sounds of Speech using Adaptive Quasi-Harmonic Models. In *Interspeech*, Portland, Oregon, USA, September 2013.

[KRS14] G. P. Kafentzis, O. Rosec, and Y. Stylianou. Robust full-band adaptive sinusoidal analysis and synthesis of speech. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2014. In Press.

[Kru89] C. L. Krumhansl. Why is musical timbre so hard to understand? In S. NielzG•n. and O. Olsson, editors, *Structure and perception of electroacoustic sound and music*, pages 43–54. Excerpta Medica, New York, 1989.

[KT82] R. Kumaresan and D.W. Tufts. Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 30(6):833–840, 1982.

[KW06] M. Kepesi and L. Weruaga. Adaptive chirp-based time-frequency analysis of speech. *Speech Communication*, 48:474–492, 2006.

[KYMS14] G. P. Kafentzis, T. Yakoumaki, A. Mouchtaris, and Y. Stylianou. Analysis of emotional speech using an adaptive sinusoidal model. In *European Signal Processing Conference (EUSIPCO)*, 2014. under review.

[Lar89] J. Laroche. A new analysis/synthesis system of musical signals using Prony's method. Application to heavily damped percussive sounds. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 2053–2056, Glasgow, UK, May 1989.

[LBG80] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Communications*, 28(1):84–95, Jan. 1980.

[LD99] J. Laroche and M. Dolson. Improved Phase Vocoder Time Scale Modification of Audio. *IEEE Transactions on Audio and Speech Processing*, 7:1–10, 1999.

[Lev99] S. Levine. *Audio Representations for Data Compression and Compressed Domain Processing*. PhD thesis, Stanford University, 1999.

[LM95] O. Cappe J. Laroche and E. Moulines. Regularized estimation of cepstrum envelope from discrete frequency points. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1995.

[LMR07] M. Lagrange, S. Marchand, and J.-B. Rault. Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1625–1634, 2007.

[LNHS05] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez. Automatic emotion recognition using prosodic parameters. In *Interspeech*, pages 493–496, 2005.

[LPY+12] P. De Leon, M. Pucher, J. Yamagishi, I. Hernáez, and I. Saratxaga. Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2280–2290, 2012.

[LSI98] S. Levine and J. O. Smith III. A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modications. In *The Audio Engineering Society Convention*, 1998.

[MA89] J. S. Marques and L. B. Almeida. Frequency-varying sinusoidal modeling of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37:763–765, 1989.

[Mac96] M. Macon. *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology, 1996.

[Mak75] J. Makhoul. Linear Prediction: A Tutorial Review. *Proceedings of the IEEE*, 63:561–580, 1975.

[MBCM93] M. W. Macon, Dr. D. J. Blumenthal, Dr. M. A. Clements, and Dr. R. M. Mersereau. Applications of sinusoidal modeling to speech and audio signal processing. In *report in Georgia Institute of Technology*, 1993.

[MC90a] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.

[MC90b] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.

[MG76] J. Markel and A. Gray. *Linear prediction of speech*. Springer Verlag, 1976.

[MIK+12] A. Maezawa, K. Itoyama, K. Komatani, T. Ogata, and H. G. Okuno. Automated violin fingering transcription through analysis of an audio recording. *Computer Music Journal*, 36(3):57–72, 2012.

[MIT+10] A. Maezawa, K. Itoyama, T. Takahashi, K. Komatani, T. Ogata, and H. Okuno. Violin fingering estimation based on violin pedagogical fingering model constrained by bowed sequence estimation from audio input. In Nicolás García-Pedrajas, Francisco Herrera, Colin Fyfe, José Manuel Benítez, and Moonis Ali, editors, *Trends in Applied Intelligent Systems*, volume 6098 of *Lecture Notes in Computer Science*, pages 249–259. Springer Berlin Heidelberg, 2010.

[ML95] E. Moulines and J. Laroche. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16(2):175–205, February 1995.

[MM99] J. Marques and P. J. Moreno. A study of musical instrument classification using gaussian mixture models and support vector machines. Technical report, Compaq Corporation, Cambridge Research laboratory, 1999.

[MQ86] R. J. McAulay and T. F. Quatieri. Speech Analysis/Synthesis based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:744–754, 1986.

[MT92] H. Valbret E. Moulines and J. P. Tubach. Voice transformation using psola techniques. *Speech Communication*, 11:175–187, 1992.

[MZ93] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[NFS03] T. Nwe, S. Foo, and L. De Silva. Speech emotion recognition using Hidden Markov Models. *Speech Communication*, 41:603–623, 2003.

[NHD98a] J. Nieuwenhuijse, R. Heusdens, and E. Deprettere. Robust exponential modeling of audio signals. In *ICASSP 1998*, Seattle, WA, USA, May 1998.

[NHD98b] J. Nieuwenhuijse, R. Heusens, and Ed F. Deprettere. Robust exponential modeling of audio signals. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, volume 6, pages 3581–3584, 1998.

[NMB07] L. D. O. Nunes, R. Merched, and L. W. P. Biscainho. Recursive least-squares estimation of the evolution of partials in sinusoidal analysis. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2007.

[NMBM01] N. Nogueiras, A. Moreno, A. Bonafonte, and J. Marino. Speech emotion recognition using Hidden Markov Models. In *EUROSPEECH*, pages 2679–2682, 2001.

[NTN00] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. *Neural Computing & Applications*, 9:290–296, 2000.

[OdB99] W Oomen and A. C. den Brinker. Sinusoids plus noise modelling for audio signals. In *International Conference of Audio Engineering Society Conference*, 8 1999.

[Pan10] Y. Pantazis. *Adaptive AM-FM Signal Decomposition With Application to Speech Analysis*. PhD thesis, Computer Science Department, University of Crete, 2010.

[PGV97] P. Prandoni, M. Goodwin, and M. Vetterli. Optimal time segmentation for signal modeling and compression. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, volume 3, pages 2029–2032, 1997.

[Por80] M. R. Portnoff. Time-frequency representation of digital signals and systems based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28:55–69, 1980.

[Por81] M. R. Portnoff. Short-time fourier analysis of sampled speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29:364–373, 1981.

[PQR99] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7:569–587, 1999.

[PRS08] Y. Pantazis, O. Rosec, and Y. Stylianou. On the Properties of a Time-Varying Quasi-Harmonic Model of Speech. In *Interspeech*, Brisbane, September 2008.

[PRS09a] Y. Pantazis, O. Rosec, and Y. Stylianou. AM-FM estimation for speech based on a time-varying sinusoidal model. In *Interspeech*, Brighton, September 2009.

[PRS09b] Y. Pantazis, O. Rosec, and Y. Stylianou. Chirp rate estimation of speech based on a time-varying quasi-harmonic model. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 3985–3988, April 2009.

[PRS10] Y. Pantazis, O. Rosec, and Y. Stylianou. On the robustness of the quasi-harmonic model of speech. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 4210–4213, March 2010.

[PRS11] Y. Pantazis, O. Rosec, and Y. Stylianou. Adaptive AM-•FM signal decomposition with application to speech analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:290–300, 2011.

[PS08] Y. Pantazis and Y. Stylianou. Improving the Modeling of the Noise Part in the Harmonic plus Noise Model of Speech. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, April 2008.

[PTRS10] Y. Pantazis, G. Tzedakis, O. Rosec, and Y. Stylianou. Analysis/Synthesis of Speech based on an Adaptive Quasi-Harmonic plus Noise Model. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, March 2010.

[Puc95] M. S. Puckette. Phase-locked Vocoder. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, 1995.

[PV00] P. Prandoni and M. Vetterli. R/D Optimal Linear Prediction. *IEEE Transactions on Speech and Audio Processing*, 8(6):646–655, Nov 2000.

[QM86] T.F. Quatieri and R. McAulay. Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6):1449–1464, Dec 1986.

[QM92] T.F. Quatieri and R.J. McAulay. Shape-Invariant Time-Scale and Pitch Modifications of Speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 40:497–510, 1992.

[QM02] T.F. Quatieri and R.J. McAuley. Audio signal processing based on sinusoidal analysis/synthesis. In Mark Kahrs and Karlheinz Brandenburg, editors, *Applications of Digital Signal Processing to Audio and Acoustics*, chapter 9, pages 343–416. Kluwer Academic Publishers, 2002.

[Qua02] T. F. Quatieri. *Discrete-Time Speech Signal Processing*. Prentice Hall, Engewood Cliffs, NJ, 2002.

[RÖ6] Axel Röbel. Adaptive additive modeling with continuous parameter trajectories. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1440–1453, July 2006.

[RAH+96] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, and D. Poch. Time and spectrum related variabilities in stressed speech under laboratory and real conditions. *Speech Communication*, 20:111–130, 1996.

[RD01] C. Ravishankar and S. Dimolitsas. *Speech Coding*. John Wiley & Sons, Inc., 2001.

[RD06] S. Ramamohan and S. Dandapat. Sinusoidal model-based analysis and classification of stressed speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):737–746, 2006.

[RK89] R. Roy and T. Kailath. Esprit-estimation of signal parameters via rotational invariance techniques. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(7):984–995, 1989.

[RR95] D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[RS78] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[RVR07] A. Röbel, F. Villavicencio, and X. Rodet. On cepstral and all-pole based spectral envelope modeling with unknown model order. *Pattern Recognition Letters*, 28(11):1343–1350, 2007.

[RW85] S. Roucos and A. Wilgus. High-quality time-scale modification for speech. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 10:493–496, 1985.

[SB98] X. Serra and J. Bonada. Sound transformations based on the sms high level attributes. In *Proceedings of the Digital Audio Effects (DAFx*, 1998.

[SC64] E. L. Saldanha and John F. Corso. Timbre cues and the identification of musical instruments. *The Journal of the Acoustical Society of America*, 36(11):2021–2026, 1964.

[Sch03] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256, 2003.

[Ser89] X. Serra. *A System for Sound Analysis, Transformation, Synthsis based on a Determistic plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.

[Ser97] X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S. Pope, A. Picialli, and G. De Poli, editors, *Musical Signal Processing*. Swets & Zeitlinger, 1997.

[SHE+09] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez. Simple representation of signal phase for harmonic speech models. *Electronics Letters*, 7, 2009.

[SK03] Y. Shiga and S. King. Estimating the spectral envelope of voiced speech using multi-frame analysis. *EUROSPEECH*, pages 1737–1740, 2003.

[SMFS89] P. Stoica, R.L. Moses, B. Friedlander, and T. Soderstrom. Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements. *IEEE Transactions on Audio, Speech, and Language Processing*, 37(3):378–392, Mar 1989.

[Spa94] A. Spanias. Speech Coding: A tutorial review. *Proceeding of the IEEE*, 82:1541–1582, October 1994.

[SS90] X. Serra and J. O. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):49–56, 1990.

[SSDR08] B. Sturm, J. J. Shynk, L. Daudet, and C. Roads. Dark energy in sparse atomic decompositions. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):671–676, 2008.

[Sty96] Y. Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, E.N.S.T - Paris, 1996.

[Sty01] Y. Stylianou. Removing linear phase mismatches in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9:232–239, 2001.

[Sty08] Y. Stylianou. Voice transformation. In Jacob Benesty, M.Mohan Sondhi, and Yiteng(Arden) Huang, editors, *Springer Handbook of Speech Processing*, pages 489–504. Springer Berlin Heidelberg, 2008.

[Tho05] H. Thornburg. *Detection and Modeling of Transient Audio Signals with Prior Information*. PhD thesis, Stanford University, 2005.

[TKMI94] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel-generalized cepstral analysis : a unified approach to speech spectral estimation. *International Conference of Spoken Language Processing*, pages 1043–1045, 1994.

[Tre68] H. Van Trees. *Detection, Estimation, and Modulation Theory: Part I.* Wiley, New York, 1968.

[TS01] C. Traube and J.O. Smith. Extracting the fingering and the plucking points on a guitar string from a recording. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 7–10, 2001.

[TT08] T. Toda and K. Tokuda. Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory hmm. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 3925–3928, 2008.

[UT96] S. Umesh and D.W. Tufts. Estimation of parameters of exponentially damped sinusoids using fast maximum likelihood estimation with application to nmr spectroscopy data. *Signal Processing, IEEE Transactions on*, 44(9):2245–2259, 1996.

[VHPR96] S. Van Huffel, Haesun Park, and J.B. Rosen. Formulation and solution of structured total least norm problems for parameter estimation. *Signal Processing, IEEE Transactions on*, 44(10):2464–2474, 1996.

[VM00] T. S. Verma and T. H. Y. Meng. Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, 24(2):47–59, 2000.

[VR93] W. Verhelst and M. Roelands. An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 554–557, 1993.

[VRC07] D. Vincent, O. Rosec, and T. Chonavel. A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 525–528, 2007.

[VRR06] F. Villavicencio, A. Röbel, and X. Rodet. Improving lpc spectral envelope extraction of voiced speech by true envelope estimation. In *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2006.

[VRR07] F. Villavicencio, X. Rodet, and A. Röbel. On cepstral and all-pole based spectral envelope modeling with unknown order. *Pattern Recognition Letters*, 28:1343–1350, 2007.

[VZA06] V. Verfaille, U. Zölzer, and D. Arfib. Adaptive digital audio effects (a-dafx): A new class of sound transformations. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1817–1831, 2006.

[WH95] B. D. Womack and J. H. L. Hansen. Stress independent robust hmm speech recognition using neural network stress classification. *EUROSPEECH*, pages 1999–2002, 1995.

[WK07] Luis Weruaga and M. Kepesi. The fan-chirp transform for non-stationary harmonic signals. *Signal Processing*, 87(6):1504–1522, 2007.

[WM05] M. Wolfel and J. McDonough. Minimum variance distortionless response spectral estimation. *IEEE Signal Processing Magazine*, 22(5):117–126, 2005.

[YKS14] T. Yakoumaki, G. P. Kafentzis, and Y. Stylianou. Emotional speech classification using adaptive sinusoidal modelling. In *Interspeech*, 2014. under review.

[YV98] B. Yegnanarayana and R. Veldhuis. Extraction of vocal-tract system characteristics from speech signals. *Speech and Audio Processing, IEEE Transactions on*, 6(4):313–327, Jul 1998.

[ZHK01] G. Zhou, J. H. L. Hansen, and J. F. Kaiser. Nonlinear feature based classification of speech under stress. *IEEE Transactions on Audio, Speech, and Language Processing*, 9:201–216, 2001.

[ZTB09] H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039 – 1064, 2009.

This work has been realized as a co-tutelle (international joint supervision of a Ph.D.) between University of Crete and University of Rennes 1 and has been funded by Orange Labs.

TECH/ASAP/VOICE
Orange Labs