ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

Εκφώνηση κειμένου από Η/Υ (Text To Speech) για την Ελληνική Γλώσσα

Ιωάννης Δ.Ε. Σουρλαντζής

Μεταπτυχιακή Εργασία

Ηράκλειο, 1996

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

Εκφώνηση κειμένου από Η/Υ (Text To Speech) για την Ελληνική Γλώσσα

Εργασία που υποβλήθηκε από τον
ΙΩΑΝΝΗ Δ.Ε. ΣΟΥΡΛΑΝΤΖΗ
ως μερική εκπλήρωση των απαιτήσεων
για την απόκτηση
ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

Συγγραφέας:

Ιωάννης Δ.Ε. Σουρλαντζής
Τμήμα Επιστήμης Υπολογιστών

Εισηγητική Επιτροπή:

Απόστολος Τραγανίτης
Αναπληρωτής Καθηγητής, Επόπτης

Νίκος Αλβέρτος
Επίκουρος Καθηγητής, Μέλος

Πάνος Τραχανιάς
Επίκουρος Καθηγητής, Μέλος

Δεκτή:

Πάνος Κωνσταντόπουλος
Αναπληρωτής Καθηγητής
Πρόεδρος Επιτροπής Μεταπτυχιακών Σπουδών

Ηράκλειο, 1996

Ένα σύστημα εκφώνησης κειμένου από Η/Υ (Text-To-Speech) για την ελληνική γλώσσα

Ιωάννης Σουρλαντζής
Μεταπτυχιακή Εργασία

Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

Περίληψη

Τα τελευταία χρόνια, τα συστήματα εκφώνησης κειμένου από Η/Υ (Text-To-Speech) έχουν αρχίσει να εμφανίζονται σαν συνηθισμένο κομμάτι ολοκληρωμένων περιβαλλόντων (MS-Windows, Macintosh). Μερικές από τις εφαρμογές των συστημάτων TTS είναι η παροχή υπηρεσιών μέσω τηλεφώνου (e-mail, telex), η παροχή βοήθειας σε άτομα με προβλήματα στην όραση ή την ομιλία, η ακουστική προσπέλαση βάσεων κειμένου και εκφώνηση οδηγιών.

Τα υπάρχοντα συστήματα TTS στηρίζονται στη μοντελοποίηση του ανθρώπινου συστήματος παραγωγής φωνής ή στη μοντελοποίηση του σήματος φωνής. Στη δεύτερη κατηγορία ανήκουν και τα συστήματα που συνθέτουν φωνή συρράπτοντας στοιχειώδη κομμάτια της (πχ φθόγγους, συλλαβές) που ονομάζονται φωνητικές μονάδες.

Στα πλαίσια αυτής της εργασίας κατασκευάστηκε ένα σύστημα TTS για την ελληνική γλώσσα βασισμένο στη τεχνική συρραφής φωνητικών μονάδων. Οι φωνητικές μονάδες στο σύστημα μας ονομάστηκαν διασυλλαβές και ορίζονται σαν το τμήμα φωνής από το μέσο ενός φωνήεντος μέχρι το μέσο του επόμενου. Η επιλογή αυτή έχει άμεση σχέση με τα χαρακτηριστικά της ελληνικής γλώσσας η οποία περιέχει πολλά και καθαρά στη προφορά φωνήεντα. Έτσι η συρραφή γίνεται πάντα μεταξύ φωνηέντων γι' αυτό και οι ασυνέχειες που πρέπει να εξομαλυνθούν δεν απαιτούν μεγάλη επεξεργασία. Στο σύστημα ελέγχεται η προσωδία της φωνής (τονικότητα, διάρκεια, ένταση) για την έκφραση οριστικών και ερωτηματικών προτάσεων αλλά και την φυσικότερη έκφραση της ομιλίας. Επίσης περιγράφεται και υλοποιείται ένας συστηματικός τρόπος κατασκευής της βάσης διασυλλαβών για την σύνθεση των ελληνικών λέξεων.

Επόπτης: Απόστολος Τραγανίτης,
Αναπληρωτής καθηγητής,
Πανεπιστήμιο Κρήτης.

A Text-to-Speech System for Greek

John Surlantzis
Master's Thesis

Computer Science Department
University of Crete, Greece

Abstract

Text-to-Speech systems have recently become a standard part of integrated enviroments (MS-Windows, Macintosh). TTS systems are useful at telephone provided services (e-mail, telex, database transactions), for aiding people with hearing or speaking problems, for verifying documents etc.

Existing TTS systems are based on modeling human speech production system, or the produced speech signal. Concatenative TTS systems produce speech by concatenating speech elements (phomenes, phones, syllables etc).

Our TTS system for greek produces speech by concatenating intersyllables. Intersyllable is the portion of speech begining at the middle of a vowel and ending at the middle of the next one. The choice was driven by the characteristics of greek accent which uses many vowels. Approximately 6000 intersyllables are necessary for the synthesis of all greek words. Power smoothing is used during concatenation. Phonetic words are made up by rule-based concatenation of small words with their neighboring ones. Pitch contour is constructed for declarative, yes/no questions, and wh-questions.

A systematic method for the construction of the intersyllables-base is also described.

Advisor: Apostolos Traganitis
Associate Professor,
University of Crete.

Ευχαριστίες

Στο σημείο αυτό θα αναφέρω τους ανθρώπους που με στήριξαν με διάφορους τρόπους στη διάρκεια της εργασίας αυτής.

Ευχαριστώ κατ' αρχήν τον επόπτη καθηγητή μου Απόστολο Τραγανίτη ο οποίος με υπομονή με ενθάρρυνε και με βοήθησε με τις υποδείξεις του καθόλη τη διάρκεια της εργασίας. Ευχαριστώ επίσης το Γιάννη Στυλιανού ο οποίος μου έδειξε εμπιστοσύνη και μου έδωσε τη δουλειά του για τον έλεγχο της προσωδίας.

Επίσης πρέπει να ευχαριστήσω το Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης και το Ινστιτούτο Πληροφορικής του Ιδρύματος Τεχνολογίας και Έρευνας για την υλικοτεχνική υποστήριξη και την οικονομική ενίσχυση που μου παρείχαν κατά τη διάρκεια των μεταπτυχιακών μου σπουδών.

Ευχαριστώ τους συμφοιτητές μου Θοδωρή Χατσιούλη, Βαρβάρα Μαστή, Δέσποινα Βαμβακά που τους τελευταίους μήνες με βοήθησαν ψυχολογικά και ουσιαστικά στη περάτωση της εργασίας και την προετοιμασία της παρουσίασης. Ευχαριστώ επίσης την Ιωάννα Χουβαρδά και τη Γεωργία Φλουρή που με στήριξαν ηθικά.

Ευχαριστώ τους γονείς μου για την εμπιστοσύνη που μου έδειξαν και τον αδερφό μου που υπήρξε σημαντικός παράγοντας για την έναρξη των μεταπτυχιακών μου σπουδών.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Τα τελευταία χρόνια, η ευρεία διάδοση των υπολογιστών έχει προκαλέσει μια στροφή των ερευνητών της επιστήμης υπολογιστών σε τρόπους επικοινωνίας ανθρώπου-μηχανής περισσότερο ανθρωποκεντρικούς. Παραδείγματα είναι η ανάπτυξη διαφόρων τύπων περιβάλλοντος, η διαλογική επικοινωνία με τις εφαρμογές, η αυτόματη αναγνώριση ομιλίας, η αυτόματη ανάγνωση γραπτών κειμένων (Optical Character Recognition). Στην ίδια κατηγορία ανήκουν και τα συστήματα εκφώνησης κειμένου από Η/Υ (Text-To-Speech, TTS).

Ένα σύστημα TTS έχει σαν είσοδο ένα κείμενο σε ηλεκτρονική μορφή και σαν έξοδο ένα αναλογικό σήμα που είναι η προφορά αυτού του κειμένου. Πολλές εφαρμογές απαιτούν την ανάπτυξη συστημάτων TTS που η έξοδος τους να πλησιάζει ικανοποιητικά την ανθρώπινη ομιλία. Παραδείγματα είναι η ανάπτυξη βάσεων δεδομένων με κείμενα και ήχο, η ανάγνωση κειμένων σε άτομα με προβλήματα στην όραση ή στην ομιλία, η παροχή υπηρεσιών μέσω τηλεφώνου. Επιπλέον, τα συστήματα TTS σε συνδυασμό με συστήματα αυτόματης αναγνώρισης ομιλίας λύνουν το πρόβλημα της αμφίδρομης επικοινωνίας ανθρώπου-μηχανής με ομιλία (*διαλογικά* συστήματα). Γενικά, η ακρόαση ενός κειμένου πολλές φορές είναι πιο ξεκούραστη διαδικασία από την ανάγνωση του που μερικές φορές δεν είναι δυνατή (ανακοινώσεις προς πολλούς ακροατές, παράλληλη ενασχόληση με άλλη εργασία).

Τα προβλήματα που απαιτείται να λυθούν κατά τη σχεδίαση και κατασκευή ενός συστήματος TTS καλύπτουν πολλές και διαφορετικές γνωστικές περιοχές. Ένας ερευνητής πρέπει να δανειστεί γνώσεις από τη γλωσσολογία, την επιστήμη υπολογιστών, την ψυχοακουστική, τη διάδοση ηχητικών κυμάτων, τη φυσιολογία, τη ψηφιακή επεξεργασία σημάτων, τη τεχνητή νοημοσύνη και τη θεωρία πιθανοτήτων.

Έχουν προταθεί συστήματα που μοντελοποιούν το ανθρώπινο σύστημα παραγωγής φωνής ή το φωνητικό σήμα που παράγει. Μία πολύ διαδεδομένη μέ-

θοδος είναι η συρραφή μικρών κομματιών φωνής για το σχηματισμό των λέξεων και των προτάσεων. Η μέθοδος αυτή έχει δώσει πολύ καλά αποτελέσματα αλλά πολλοί ερευνητές υποστηρίζουν ότι δεν επιδέχεται επιπλέον βελτίωση και επιμένουν σε μοντέλα για το ανθρώπινο σύστημα παραγωγής φωνής. Τα περισσότερα συστήματα TTS αγνοούν το σκοπό της ομιλίας που είναι η επικοινωνία και εστιάζουν τη προσοχή τους στο ακουστικό αποτέλεσμα του ανθρώπινου συστήματος παραγωγής φωνής. Αυτό οφείλεται κυρίως στη περιορισμένη γνώση που έχουμε για τη δομή της ομιλίας. Ωστόσο, είναι κοινά παραδεκτό ότι απαραίτητη προϋπόθεση για την παραγωγή ομιλίας που πλησιάζει πολύ την ανθρώπινη είναι η κατανόηση του κειμένου. Έτσι πολλοί ερευνητές έχουν στραφεί στη τεχνητή νοημοσύνη για να βελτιώσουν τη *φυσικότητα* των συστημάτων TTS.

Οριοθέτηση εργασίας: Στην εργασία αυτή περιγράφεται η σχεδίαση και υλοποίηση ενός συστήματος TTS για την ελληνική γλώσσα. Ιδιαίτερο βάρος κατά τη σχεδίαση έχει δοθεί στις ιδιαιτερότητες της ελληνικής γλώσσας. Το σύστημα προφέρει ομιλία συρράπτοντας προηχογραφημένα κομμάτια ομιλίας (φωνητικές μονάδες) μεταβλητού μεγέθους. Η επιλογή φωνητικών μονάδων έχει άμεση σχέση με τα χαρακτηριστικά της ελληνικής γλώσσας. Η βάση φωνητικών μονάδων περιέχει τις φωνητικές μονάδες που απαιτούνται για την εκφώνηση όλων των ελληνικών λέξεων. Περιγράφουμε ένα συστηματικό τρόπο κατασκευής της βάσης με τον οποίο δίνεται η δυνατότητα και σε άλλους χρήστες να κατασκευάσουν τη δική τους βάση πχ. για άλλη γλώσσα όπως την ιταλική που έχει χαρακτηριστικά παρόμοια με τη δική μας αλλά και την ενημέρωση της υπάρχουσας. Το σύστημα είναι τμηματοποιημένο (modular), έτσι η μετατροπή του για άλλη γλώσσα δεν απαιτεί συνολικές τροποποιήσεις (μεταφερσιμότητα).

Για τη συρραφή των φωνητικών μονάδων γίνεται επεξεργασία με σκοπό την εξομάλυνση ασυνεχειών στα σημεία συνένωσης, και έλεγχος στα χαρακτηριστικά της *προσωδίας* της ομιλίας (τονικότητα, διάρκεια, ένταση). Το σύστημα μπορεί να εκφωνήσει οριστικές και ερωτηματικές προτάσεις ελέγχοντας τη προσωδία της ομιλίας. Η χρονική εξέλιξη της τονικότητας για κάθε φράση υπολογίζεται με συρραφή τμημάτων τονικότητας από πραγματική ομιλία. Ο ρυθμός ομιλίας ελέγχεται με την αναγνώριση των λέξεων που προφέρονται σαν μία.

Οργάνωση της εργασίας: Στο κεφάλαιο 2 περιγράφουμε την τυπική δομή σχεδόν όλων των υπαρχόντων συστημάτων TTS. Επίσης αναφέρουμε μερικές από τις πολλές εφαρμογές των συστημάτων TTS (παράγραφος 2.2) και περιγράφουμε το ανθρώπινο σύστημα παραγωγής φωνής (2.3) και τα χαρακτηριστικά της ελληνικής γλώσσας (2.4). Τέλος αναφερόμαστε σε κριτήρια αξιολόγησης συστημάτων TTS (2.7). Στο κεφάλαιο 3 γίνεται μια ανασκόπηση των προσπα-

θειών που έχουν γίνει στο παρελθόν για την κατασκευή συστημάτων TTS ή τμημάτων τους. Στο κεφάλαιο 4 περιγράφουμε τη σχεδίαση και κατασκευή του συστήματος TTS για την ελληνική γλώσσα. Στη παράγραφο 4.2 περιγράφουμε τη διαδικασία κατασκευής της βάσης φωνητικών μονάδων και στη παράγραφο 4.3 τη λειτουργία του συστήματος κατά την εκφώνηση κειμένου. Τέλος αναφέρουμε συμπεράσματα και μελλοντικές επεκτάσεις του συστήματος (κεφ. 5) όπως η χρήση της κωδικοποίησης GSM για τη συμπίεση της βάσης και τον έλεγχο της προσωδίας.

Κεφάλαιο 2

Συστήματα TTS

Στο κεφάλαιο αυτό θα ορίσουμε το πρόβλημα της *εκφώνησης κειμένου από Η/Υ* (*Text To Speech, TTS*) και θα αναφερθούμε στις διάφορες εφαρμογές που καθιστούν επιτακτική την αντιμετώπισή του, καθώς και στους λόγους για τους οποίους προσπαθούμε να το επιλύσουμε (πιθανές χρησιμότητες). Θα αναλύσουμε το πρόβλημα σε επιμέρους προβλήματα, πιο απλά και επομένως πιο εύκολα επιλύσιμα. Στη συνέχεια θα εξετάσουμε χωριστά τα υποπροβλήματα και τις δυσκολίες που αντιμετωπίζει κανείς για να τα λύσει. Στο τέλος του κεφαλαίου θα αναφέρουμε κριτήρια αξιολόγησης συστημάτων TTS.

2.1 Ορισμός προβλήματος

Ένα σύστημα TTS έχει σαν σκοπό να μετατρέψει ένα κείμενο[1] που είναι γραμμένο με το αλφάβητο και τους κανόνες γραμματικής κάποιας γλώσσας σε ομιλία (δηλαδή ήχο) αυτής της γλώσσας. Ένας άνθρωπος το κάνει αυτό με σχετική ευκολία (εκφώνηση-ανάγνωση γραπτού κειμένου). Επειδή ένα σύστημα TTS έχει σα στόχο να προσομοιώσει ανθρώπινη συμπεριφορά είναι χρήσιμο να δούμε πως λειτουργεί το ανθρώπινο σύστημα παραγωγής φωνής.

Εδώ πρέπει να κάνουμε το διαχωρισμό *προφοράς* και *ομιλίας* (προφορικού λόγου). Με τον όρο προφορά αναφερόμαστε στην ικανότητα του ανθρώπου να παράγει πολλούς, ποικίλους και καθαρούς ηχούς (τα ζώα αντίθετα παράγουν λιγότερους ηχούς από τον άνθρωπο) ενώ με τον όρο ομιλία εννοούμε τη χρήση της προφοράς για επικοινωνία, δηλαδή τη παραγωγή ήχων σύμφωνα με τους κανόνες κάποιας γλώσσας και επομένως τη μεταφορά κωδικοποιημένης πληροφορίας.

[1] Υποθέτουμε ότι το κείμενο εισάγεται στο σύστημα TTS σε ηλεκτρονική μορφή (δηλαδή ακολουθία χαρακτήρων-συμβόλων).

Μία εγγενής δυσκολία στη κατασκευή ενός συστήματος TTS είναι το ότι τα δεδομένα εισόδου και εξόδου αναπαριστούν πληροφορία με τελείως διαφορετικό τρόπο. Η πληροφορία που περιέχει το κείμενο είναι από τη φύση της ψηφιακή και κωδικοποιημένη κατά τους κανόνες της γλώσσας. Αντίθετα η ομιλία είναι αναλογικό σήμα (ήχος) με ιδιότητες που έχουν άμεση σχέση με το πως το παράγει ο άνθρωπος.

Το κείμενο έχει ελλειπείς οδηγίες για το πως πρέπει να εκφωνηθεί. Ακόμα και ένας άνθρωπος δεν είναι σίγουρος για το πως πρέπει να διαβάσει ένα γραπτό κείμενο. Σχεδόν πάντα ο τρόπος εκφώνησης ενός κειμένου εξαρτάται από το νοηματικό του περιεχόμενο το οποίο είναι πολύ δύσκολο να εκτιμηθεί από υπολογιστή.

2.2 Εφαρμογές

Η χρησιμότητα ενός συστήματος TTS φαίνεται σε πάρα πολλές εφαρμογές. Η χρήση του είναι ενδεδειγμένη σε περιπτώσεις στις οποίες το κείμενο που πρέπει να εκφωνηθεί δεν είναι συγκεκριμένο και γνωστό εκ των προτέρων. Σε περιπτώσεις που το κείμενο είναι γνωστό και σταθερό (πχ συνήθεις ανακοινώσεις σε μέσα μαζικής μεταφοράς) τότε είναι προτιμότερη η χρήση ηχογραφημένης ομιλίας. Κανένα σύστημα TTS μέχρι σήμερα δε παράγει ομιλία που να μοιάζει πάρα πολύ με ανθρώπινη.

Παραδείγματα περιπτώσεων που το κείμενο που πρέπει να εκφωνηθεί δεν είναι από πριν γνωστό, είναι η ανάγνωση κειμένων σε άτομα με πρόβλημα στην όραση, η εκφώνηση κειμένων για επαλήθευση και διόρθωση ή η μετάδοση ηλεκτρονικών μηνυμάτων (e-mail, telex) μέσω τηλεφώνου. Μία πιο φιλόδοξη εφαρμογή θα ήταν στη διδασκαλία ξένων γλωσσών δείχνοντας στο μαθητή τη σωστή προφορά οποιασδήποτε πρότασης. Αυτό θα απαιτούσε βέβαια πολύ σωστή άρθρωση από το συγκεκριμένο TTS σύστημα.

Η χρήση συστημάτων TTS είναι πολλές φορές προτιμότερη και σε περιπτώσεις που το κείμενο είναι γνωστό από πριν. Για παράδειγμα σε μια βάση δεδομένων θα ήταν χρήσιμο να μπορούμε να προσπελάσουμε τα κείμενα που περιέχει ακούγοντάς τα. Η ηχογράφηση όλων των κειμένων που περιέχει η βάση είναι πολύ χρονοβόρα και απαιτεί τεράστιο αποθηκευτικό χώρο. Επιπλέον για να γίνει μια αλλαγή στη βάση πρέπει να γίνει και η αντίστοιχη ηχογράφηση το οποίο είναι ακριβή και αργή διαδικασία. Η χρήση ενός συστήματος TTS μας απαλλάσει από τέτοια προβλήματα ενώ ταυτόχρονα μειώνει κατά πολύ το χώρο αποθήκευσης[2].

[2] Όταν αποθηκεύουμε multimedia πληροφορία το μεγαλύτερο ποσοστό της χωρητικότητας

Ένα άλλο πεδίο εφαρμογής είναι σε περιπτώσεις που η όραση είναι απασχολημένη με άλλες λειτουργίες πχ οδήγηση. Έτσι γίνεται δυνατή η διαρκής ενημέρωση του οδηγού χωρίς να αποσπάται η προσοχή του (από το δρόμο).

Πολύ χρήσιμο είναι ένα σύστημα TTS σε άτομα που δεν μπορούν να μιλήσουν αλλά μπορούν να χρησιμοποιήσουν τα χέρια τους για δακτυλογράφηση. Τα περισσότερα από τα υπάρχοντα συστήματα TTS είναι πραγματικού χρόνου δηλαδή εκφωνούν το ίδιο γρήγορα όσο και ένας άνθρωπος.

Γενικά, η ζήτηση TTS συστημάτων αυξάνει ραγδαία εξαιτίας της τάσης να γίνει η επικοινωνία χρήστη-υπολογιστή περισσότερο ανθρωποκεντρική. Έτσι η ακρόαση αντί για την ανάγνωση από το χρήστη αποτελεσμάτων ή οδηγιών οποιασδήποτε εφαρμογής (πχ τηλεφωνικές πληροφορίες) κάνει τη χρήση TTS συστημάτων ευρύτατη.

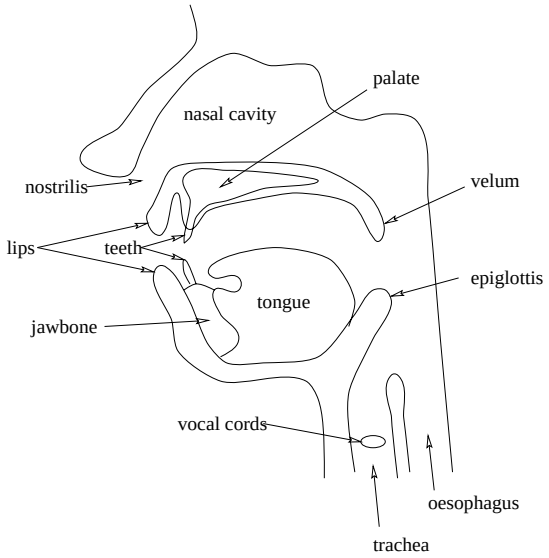
2.3 Το ανθρώπινο σύστημα παραγωγής φωνής

Η παραγωγή του προφορικού λόγου βασίζεται σε δύο μηχανισμούς. Τη *φώνηση* (*phonation*) που γίνεται από το λάρυγγα και την *άρθρωση* (*articulation*) που γίνεται από τα στοιχεία του στόματος. Στο σχήμα 2.1 φαίνεται απλοποιημένα μια τομή του ανθρώπινου αναπνευστικού συστήματος.

Η λειτουργία της φώνησης γίνεται ώς εξής : Οι πνεύμονες λειτουργούν σαν φυσερό και τροφοδοτούν με αέρα την τραχεία. Όταν το διάφραγμα κατεβαίνει τα πνευμόνια γεμίζουν αέρα ενώ με την αντίθετη κίνηση ο αέρας αναγκάζεται να φύγει με ταχύτητα από τη τραχεία. Ο λάρυγγας, που βρίσκεται μετά την τραχεία, είναι ένα οστό που καλύπτεται από δέρμα. Στο κέντρο βρίσκονται οι φωνητικές χορδές, δύο πτυχές οι οποίες τεντώνονται και παίρνουν ορισμένες θέσεις με την βοήθεια εξειδικευμένων μυών. Καθώς ο αέρας περνάει με ταχύτητα ανάμεσα από τις φωνητικές χορδές η πίεση που ασκείται σ'αυτές μειώνεται λόγω του φαινομένου *Bernulli*[3]. Οι φωνητικές χορδές κλείνουν με αποτέλεσμα η πίεση από κάτω τους να αυξάνει. Πάνω από κάποιο όριο πίεσης οι φωνητικές χορδές ανοίγουν και το φαινόμενο επαναλαμβάνεται. Τελικά στην πάνω πλευρά του λάρυγγα παράγονται ηχητικοί παλμοί σαν αυτούς του σχήματος 2.2. Η συχνότητά τους εξαρτάται από τη τάση με την οποία τεντώνονται οι φωνητικές χορδές και από τη θέση τους. Τη συχνότητα αυτή ονομάζουμε *θεμελιώδη συχνότητα* (*fundamental frequency, pitch frequency*) και αντίστοιχα τη περίοδο *θεμελιώδη περίοδο* (*fundamental period, pitch period*).

καταλαμβάνει η ομιλία.

[3]Είναι το φαινόμενο μείωσης της πίεσης που ασκεί ένα ρευστό σε μια επιφάνεια όταν ρέει με μεγάλη ταχύτητα



Σχήμα 2.1: Αναπνευστικό σύστημα ανθρώπου.

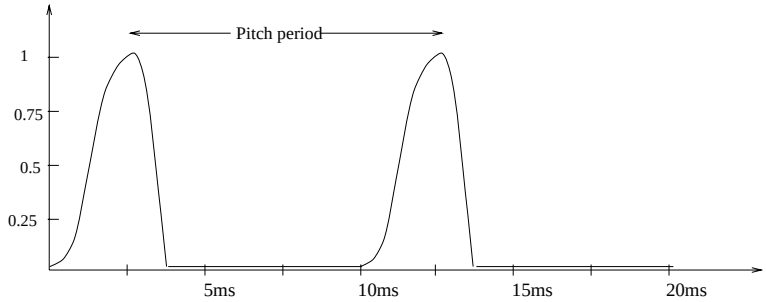
Ο αέρας τροφοδοτείται από την τραχεία (trachea) ανάμεσα από τις φωνητικές χορδές (vocal cords). Η επιγλωτίδα (epiglottis) φράζει τη τραχεία όταν καταπίνουμε και οδηγεί τη τροφή προς τον οισοφάγο (oesophagus). Το ηχητικό σήμα που παράγεται διαμορφώνεται στη συνέχεια από τα στοιχεία του στόματος, τη γλώσσα (tongue), τα δόντια (teeth) και τον σκληρό ουρανίσκο (palate). Ο πίσω (μαλακός) ουρανίσκος καταλήγει στη σταφύλη (velum) που ελέγχει τη δίοδο προς τη ρινική κοιλότητα (nasal cavity). Η εκπομπή του ήχου γίνεται από τα χείλια (lips) και τα ρουθούνια (nostrilis). Στο σχήμα σημειώνεται και τα κάτω σαγόνι (jawbone).

Το ηχητικό κύμα που παράγεται στο λάρυγγα περνάει από το φάρυγγα και τη στοματική κοιλότητα και εκπέμπεται με τη βοήθεια των χειλέων. Τα στοιχεία αυτά λειτουργούν σαν ακουστικό φίλτρο ενώ η ρινική κοιλότητα σαν ακουστικό αντηχείο[4].

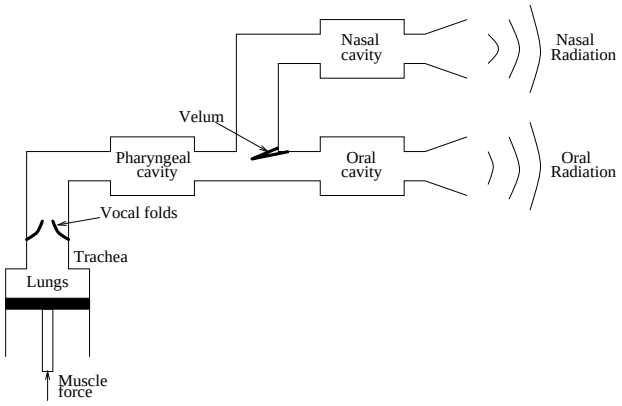
Πιο πάνω από τον λάρυγγα συναντάμε τη γλωττίδα η οποία φράζει το λάρυγγα όταν καταπίνουμε τροφή. Μετά βρίσκεται ο φάρυγγας που καταλήγει σε μια διακλάδωση που οδηγεί στη στοματική και τη ρινική κοιλότητα. Η δίοδος προς τη ρινική κοιλότητα ελέγχεται από μια επίφυση, τη σταφύλη (velum), που βρίσκεται στο πίσω μέρος του μαλακού ουρανίσκου (υπερώο). Όταν η σταφύλη αφήνει ανοικτή τη δίοδο ο ήχος που παράγεται είναι ένρινος. Τα στοιχεία που καθορίζουν τη γεωμετρία της φωνητικής οδού είναι η γλώσσα, ο σκληρός και ο μαλακός ουρανίσκος, τα δόντια και τα χείλια. Στην εκπομπή των ηχητικών κυμάτων κύριο ρόλο παίζει το σχήμα των χειλιών. Ένα ποσοστό του εκπεμπόμενου ήχου βγαίνει και από τη μύτη μέσω της ρινικής κοιλότητας. Ένα απλοποιημένο διάγραμμα του ανθρώπινου συστήματος παραγωγής φωνής

[4] Ακόμα και η θωρακική κοιλότητα λειτουργεί ώς αντηχείο

φαίνεται στο σχήμα 2.3



Σχήμα 2.2: Παλμοί διέγερσης



Σχήμα 2.3: Σχηματικό διάγραμμα του ανθρώπινου συστήματος παραγωγής φωνής

Ο αέρας από τα πνευμόνια (lungs) με τη βοήθεια του διαφράγματος (muscle force) περνάει από τη τραχεία (trachea) και δονεί τις φωνητικές χορδές (vocal chords). Ο ήχος διαμορφώνεται από τη φαρυγγική κοιλότητα (pharyngeal cavity) τη στοματική κοιλότητα (oral cavity) και τη ρινική κοιλότητα (nasal cavity) όταν το επιτρέπει η σταφύλη (velum).

Όπως φαίνεται από την παραπάνω περιγραφή η φώνηση παίζει το ρόλο της *διέγερσης* ενώ η στοματική και η ρινική κοιλότητα το ρόλο του *φίλτρου*. Πολλές φορές όμως διέγερση δεν έχουμε μόνο λόγο της φώνησης αλλά σχηματίζοντας ένα στένεμα σε κάποιο ύψος της φωνητικής οδού. Στο στένεμα δημιουργείται στροβιλοειδής κίνηση του αέρα που ακούγεται σαν θόρυβος. Ο θόρυβος και πάλι φιλτράρεται από το υπόλοιπο (μετά το στένεμα) της φωνητικής οδού. Συχνά τα δύο είδη διέγερσης συνυπάρχουν.

2.4 Χαρακτηριστικά ομιλίας και καταγραφή της

Ο πιο διαδεδομένος τρόπος καταγραφής ομιλίας είναι ο *γραπτός λόγος* (*ορθογραφική γραφή* ή *συμβατική γραφή*). Ο γραπτός λόγος όμως δεν παριστάνει τους ηχούς που παράγουμε όταν μιλάμε, δηλαδή τη προφορά, γιατί δεν υπάρχει μοναδική αντιστοιχία μεταξύ γραμμάτων του αλφάβητου και των ήχων που παράγουμε όταν τα προφέρουμε. Η αντιστοιχία αποκαθίσταται χρησιμοποιώντας τους κανόνες γραμματικής της γλώσσας. Για παράδειγμα στην ελληνική γλώσσα το γράμμα «*υ*» άλλες φορές προφέρεται σαν /*ι*/ άλλες σαν /*β*/ και άλλες σαν /*φ*/. Οι στοιχειώδης φωνές (κομμάτια προφοράς) με τις οποίες φτιάχνουμε τις λέξεις ονομάζονται *φθόγγοι* (*phonemes*). Έτσι ενώ γράφουμε «*ευγενείς*» προφέρουμε /*εβγενίς*/[5]. Ένας άλλος ορισμός για τον φθόγγο έρχεται από τη γλωσσολογία : αν αντικαταστήσουμε ένα φθόγγο μ'έναν άλλο σε μια λέξη τότε πιθανόν αυτή να αλλάξει σημασία. Η ελληνική γλώσσα έχει 25 φθόγγους που αναφέρονται στη παράγραφο 2.5.

Ένας φθόγγος δεν αντιπροσωπεύει ένα συγκεκριμένο ήχο αλλά μια κλάση ήχων που παράγονται με παρόμοια άρθρωση (κίνηση ή θέση των στοιχείων άρθρωσης). Το πως θα αρθρωθεί κάθε φθόγγους εξαρτάται από τους γειτονικούς του και κυρίως από το φθόγγο που ακολουθεί. Πχ αλλοιώς αρθρώνεται ο φθόγγος /*χ*/ στη λέξη «*χάρτης*» και αλλοιώς στη λέξη «*χέρι*» (αν προσπαθήσουμε να πούμε τη πρώτη συλλαβή αργά παρατηρούμε ότι η γλώσσα έχει άλλη θέση γι' αυτό και προφέρεται διαφορετικός ήχος). Αυτές οι διαφορετικές εκδόσεις του ίδιου φθόγγου ονομάζονται *αλλόφωνα* (*allophones*). Μία πιο αναλυτική καταγραφή των ήχων που φτιάχνουν τις λέξεις γίνεται με τη χρήση *φωνημάτων* (*phones*) στην οποία καταγράφονται-κωδικοποιούνται και οι αλλοφωνικές αλλοιώνεις.

Οι παραπάνω μέθοδοι γραφής επιδιώκουν να περιγράψουν τους ηχούς της ομιλίας κωδικοποιώντας τη θέση ή κίνηση των στοιχείων άρθρωσης. Η ομιλία έχει όμως κι άλλες ιδιότητες που δεν φαίνονται στις μεθόδους που είδαμε. Οι κυριότερες είναι η *τονικότητα* (θεμελιώδης συχνότητα), ο *ρυθμός* (χρονική εξέλιξη) και η *ένταση*. Οι τρεις αυτές οι παράμετροι χαρακτηρίζουν τη *προσωδία* της ομιλίας. Υπάρχουν μέθοδοι γραφής που καταγράφουν-κωδικοποιούν και την εξέλιξη της προσωδίας (International Phonetic Alphabet - IPA και ARPAbet).

Η προσωδία του κάθε φωνήματος εξαρτάται από το πόσο το *φορτίζουμε* (*stress*). Η *φόρτιση* εξαρτάται από τη θέση του φωνήματος μέσα στη λέξη (αν είναι πρώτο ή τελευταίο) και από το αν τονίζεται. Η προσωδία της λέξης (άρα και των φωνημάτων που τη σχηματίζουν) εξαρτάται από τη θέση της λέξης μέσα στη πρόταση και τη συντακτική της σημασία αλλά και από το είδος της

[5] Θα σημειώνουμε τη ορθογραφική γραφή μέσα σε « » και τη φωνητική μέσα σε / /

πρότασης (καταφατική, ερωτηματική κλπ). Τέλος η προσωδία μιας πρότασης εξαρτάται άμεσα από το νόημά της αλλά και από την έκφραση ή διάθεση του ομιλητή (θυμός, σαρκασμός κλπ).

Οι μέθοδοι γραφής που αναφέραμε δεν καταγράφουν όλα τα στοιχεία της ομιλίας. Σαν παράδειγμα θα αναφέρουμε το φαινόμενο της *συνάρθρωσης*. Κάθε φώνημα αντιστοιχεί σε μια συγκεκριμένη θέση ή κίνηση των στοιχείων άρθρωσης. Κατά τη συνεχή ομιλία τα στοιχεία άρθρωσης αλλάζουν θέσεις για να προφερθεί ο επιθυμητός ήχος. Ανάμεσα στη άρθρωση δύο φωνημάτων προφέρονται ήχοι που δεν ανήκουν ούτε στον αρχικό φώνημα ούτε στο τελικό γιατί τα στοιχεία άρθρωσης βρίσκονται σε ενδιάμεσες θέσεις. Οι ήχοι αυτοί όμως δε καταγράφονται. Το φαινόμενο αυτό ονομάζεται *συνάρθρωση* (*coarticulation*) και το ακουστικό αποτέλεσμα είναι η *συμπροφορά*. Η επίδοση ενός συστήματος TTS εξαρτάται σημαντικά από την σωστή απόδοση της συνάρθρωσης.

Όπως είδαμε οι ήχοι που προφέρουμε εξαρτώνται από κανόνες πολύ συγκεκριμένους (πχ ποιό γράμμα προφέρουμε) έως πολύ αφηρημένους (τί έκφραση θα έχουμε). Αυτό μας υποχρεώνει να κρατάμε πληροφορία για το κείμενο σε διαφορετικά επίπεδα αφαίρεσης.

2.5 Άρθρωση των φθόγγων της ελληνικής γλώσσας

Οι φθόγγοι της ελληνικής γλώσσας είναι συνολικά 25 και διακρίνονται στα σύμφωνα και στα φωνήεντα [Τρι93]. Δε μπορούμε να τα διαχωρίσουμε βάσει των παραμέτρων προφοράς (θέσεις στοιχείων άρθρωσης, τύπος διέγερσης) πχ να πούμε ότι τα φωνήεντα προφέρονται με ανοικτά τα χείλια και για διέγερση έχουν το μηχανισμό της φώνησης, οι φθόγγοι /ν/, /ζ/ προφέρονται με τον ίδιο τρόπο. Ο διαχωρισμός λοιπόν είναι περισσότερο καταχρηστικός αν εξαιρέσουμε τον τρόπο που ακούγονται.Τα φωνήεντα είναι

α, ε, ο, ι, ου

και τα σύμφωνα

β, γ, δ, ζ, θ, κ, λ, μ, ν, π, ρ, σ, τ, φ, χ, μπ, ντ, γκ, τσ, τζ.

Μιλώντας για τα σύμφωνα παρακάτω θα αναφέρουμε και μερικές αλλοφωνικές εκδόσεις τους. Αυτές είναι :

/λ'/ , ουρανικό /λ/, όπως στο «*ελιά*»

/ν'/ , ουρανικό /ν/, όπως στο «*εννιά*»

/μ'/ , ουρανικό /μ/, όπως στο «*μια*»

/μ̈/ , χειλοδοντικό /μ/, όπως στα «*αμφίβιο, σύμβολο*»

/ν̈/ , υπερωικό /ν/, όπως στα «*άγγελος, αγκαλιά*»

/κ'/, /γ'/, /γκ'/, /χ'/ , ουρανικά /κ/, /γ/, /γκ/, /χ/, όπως στα «*κυρία, γέρος, γκέμι,*

ως προς το τρόπο άρθρωσης τους		ως προς τη διάρκεια			
		στιγμιαία	εξακολουθητικά		
		ηχηρά	άηχα	τριβόμενα ηχηρά	τριβόμενα άηχα	ρινικά υγρά (ηχηρά)
χειλικά	διχειλικά	/π/	/μπ/			/μ/ /μ'/
	χειλοδοντικά			/φ/	/β/	
οδοντικά	μεσοδοντικά			/θ/	/δ/	
	γλωσσοδοντικά	/τ/	/ντ/			
	διπλοδοντικά	/τσ/	/τζ/	/σ/	/ζ/	
λαρυγγικά	ουρανικά	/κ'/	/γκ'/	/χ'/	/γ'/	/ν̇/
	υπερωικά	/κ/	/γκ/	/χ/	/γ/	
γλωσσικά	ουρανικά					/ν'/ /λ'/
	υπερωικά					/ν/ /λ/ /ρ/

Πίνακας 2.1: Πίνακας συμφώνων

Κατάταξη των συμφώνων ώς προς το τρόπο που αρθρώνονται, τη διάρκειά τους και την έντασή τους. Φαίνονται και οι αλλοφωνικές εκδόσεις των συμφώνων που αναφέρονται στο κείμενο.

χήρα»

Όλα τα φωνήεντα προφέρονται από το στόμα και ώς διέγερση λειτουργεί ο μηχανισμός της φώνησης. Όταν συμετέχει και η ρινική κοιλότητα τότε τα φωνήεντα ακούγονται ένρινα (αλλοφωνικές εκδόσεις). Χωρίζουμε τα φωνήεντα, ανάλογα με τις θέσεις των στοιχείων άρθρωσης τη στιγμή που τα προφέρουμε, σε ουρανικά και υπερωικά.

Ουρανικά : είναι τα /α/ και /ε/ κατά την προφορά των οποίων η γλώσσα είναι απλωμένη με τη μύτη της προς τον σκληρό ουρανίσκο, τα χείλια κλείνουν και οι άκρες τους είναι προς τα μέσα.

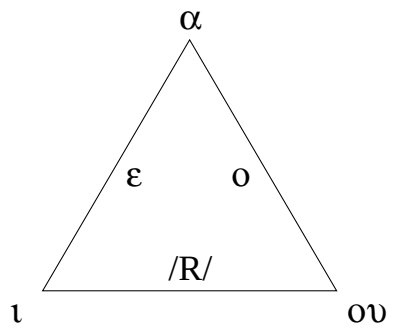
Υπερωικά : είναι τα /ο/ και /ου/. Κατά τη προφορά τους πίσω μέρος της γλώσσας υψώνεται προς την υπερώα ενώ το μπροστά μέρος της είναι χαλαρό. Οι άκρες των χειλιών τεντώνονται προς τα εμπρός και κλείνουν.

Το /α/ αρθρώνεται με τα χείλια ορθάνοιχτα και τη γλώσσα στη φυσική αδιάφορη θέση. Επειδή το στόμα είναι περισσότερο ανοικτό έχουμε και μεγαλύτερη εκπομπή ενέργειας. Γι'αυτό το προφέρουμε όταν θέλουμε να φωνάξουμε δυνατά. Μία σχηματική αναπαράσταση των φωνηέντων που χρησιμοποιείται στη γλωσσολογία φαίνεται στο σχήμα 2.4 και δείχνει ποιά φωνήεντα είναι κοντινά ώς προς τον τρόπο που αρθρώνονται.

Τα σύμφωνα μπορούν να χωριστούν σύμφωνα με τα εξής κριτήρια : α) το τρόπο άρθρωσης τους, β) τη διάρκειά τους και γ) την ένταση τους.

Στο πίνακα 2.1 φαίνεται συνολικά ο διαχωρισμός των συμφώνων. Στη συνέχεια θα δώσουμε με μικρές επεξηγήσεις το διαχωρισμό τους.

α) <u>ώς προς το τρόπο άρθρωσης τους</u> :



Σχήμα 2.4: Σχηματική αναπαράσταση φωνηέντων

Τα ελληνικά φωνήεντα παριστάνονται συνήθως σ' ένα τρίγωνο. Στη πάνω γωνία είναι το /*α*/ που προφέρεται με ανοικτό το στόμα. Στις δύο πλάγιες πλευρές σημειώνονται τα /*ε*/ και /*ο*/ που προφέρονται όταν το στόμα κλείνει κατά διαφορετικούς προς την προφορά των /*ι*/ και /*ου*/ αντίστοιχα. Ο φθόγγος /*R*/ (φωνητική γραφή ARPAbet) δεν υπάρχει στην ελληνική γλώσσα και προφέρεται στην αγγλική λέξη "*heard*" και στη γαλλική "*muse*".

Χειλικά : σχηματίζονται με τα χείλια, το σχήμα ή τη κίνηση τους ανάλογα με το αν είναι στιγμιαία ή εξακολουθητικά (διαχωρισμός (β)).

Οδοντικά : σχηματίζονται με την επαφή της γλώσσα πάνω στα δόντια ή το πάνω μέρος των πάνω δοντιών.

Λαρυγγικά : σχηματίζονται με στένεμα στο λάρυγγα.

Γλωσσικά : σχηματίζονται με τη γλώσσα να ακουμπάει με κάποιο τρόπο στον ουρανίσκο ή να πάλλεται, κραδαίνεται πάνω στον ουρανίσκο.

β) <u>ώς προς τη διάρκειά τους</u> :

Στιγμιαία : σχηματίζονται από συγχρονισμένες κινήσεις κάποιων στοιχείων άρθρωσης τα οποία φράζουν την ακουστική οδό σε κάποιο σημείο. Ακούγονται στιγμιαία την ώρα που η ακουστική οδό ανοίγει ξαφνικά.

Διαρκείας ή **εξακολουθητικά** : σχηματίζονται όταν τα στοιχεία άρθρωσης στενεύουν κάπου τη ακουστική οδό. Αυτά μπορούμε να τα προφέρουμε συνέχεια.

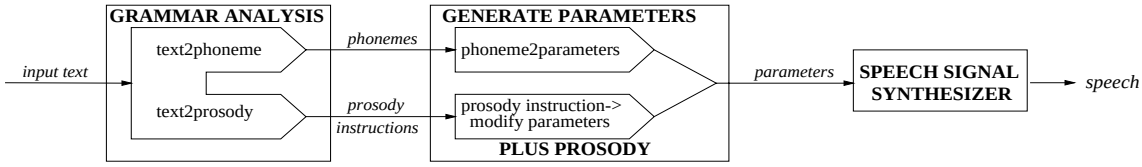
γ) <u>ώς προς την ένταση τους</u> :

Άηχα : σχηματίζονται όταν οι φωνητικές χορδές είναι ανοιχτές και χαλαρωμένες, γι'αυτό η διέγερση έχει χαμηλή ενέργεια.

Ηχηρά : σχηματίζονται όταν οι φωνητικές χορδές είναι τεντωμένες έτσι ώστε να δημιουργούν στένεμα. Έτσι η στροβιλοειδής κίνηση του αέρα είναι πιο δυνατή. Μερικές φορές λειτουργεί και η φώνηση γιατί με τη ροή του αέρα οι φωνητικές χορδές πάλλονται.

2.6 Τυπική δομή των συστημάτων TTS

Όπως αναφέραμε σε προηγούμενη παράγραφο ένα σύστημα TTS δέχεται σαν είσοδο κείμενο και παράγει στην έξοδο φωνητικό σήμα. Αν δούμε τα ενδιάμεσα αποτελέσματα μέσα στο TTS σύστημα τότε κινουμενοι από την είσοδο προς την έξοδο παρατηρούμε ότι σ' αυτά χάνονται τα χαρακτηριστικά του ψηφιακού-κωδικοποιημενου σήματος εισόδου και εμφανίζονται χαρακτηριστικά του αναλογικού σήματος εξόδου. Στις προηγούμενες παραγράφους αναφέραμε τέτοια ενδιάμεσα αποτελέσματα όπως τους φθόγγους, τα φωνήματα, τις θέσεις των στοιχείων άρθρωσης. Θα χωρίσουμε σε *τμήματα* (*modules*) ένα TTS σύστημα έχοντας υπόψη τα ενδιάμεσα αποτελέσματα (σχήμα 2.5). Το πρώτο τμήμα αναλαμβάνει τη μετατροπή του κειμένου εισόδου σε ακολουθία συμβόλων που αναπαριστούν κλάσεις ήχων (πχ φωνήματα, φθόγγους). Το δεύτερο τμήμα αντιστοιχίζει τις κλάσεις ήχων σε διανύσματα παραμέτρων του σύνθετη σήματος. Το τρίτο τμήμα είναι ένας συνθέτης σήματος στον οποίο δίνουμε μια αφηρημένη περιγραφή του ήχου που θέλουμε να παράγει (πχ μια περιγραφή του φάσματος) και στην έξοδο παίρνουμε τον αντίστοιχο ήχο. Τα όρια των τμημάτων που αναφέραμε δεν ορίζονται αυστηρά και εξαρτώνται από την αρχή λειτουργίας του συγκεκριμένου TTS συστήματος.



Σχήμα 2.5: Τμήματα ενός συστήματος TTS

Οι δύο λειτουργίες text2phoneme και *text2prosody* επικαλύπτονται κατά την αρχική επεξεργασία του κειμένου εισόδου.

Ο χωρισμός του συστήματος TTS σε τμήματα μας επιτρέπει να κάνουμε εύκολα αλλαγές στο σύστημα. Παραδείγματος χάρη αν θέλουμε να αλλάξουμε τη γλώσσα χρειάζεται να αντικατασταθεί μόνο το πρώτο τμήμα. Παρόμοια μπορούμε να δοκιμάσουμε διάφορους σύνθετες σήματος (αν υπάρχει κάποια συμβατότητα στις παραμέτρους) και να δούμε τη συμπεριφορά του καινούργιου TTS συστήματος.

Στη συνέχεια θα δούμε αναλυτικά τα τρία τμήματα ενός TTS συστήματος.

2.6.1 Γραμματική ανάλυση κειμένου

Το κομμάτι μεταφράζει την ακολουθία γραμμάτων εισόδου σε μια ακολουθία συμβόλων που παριστάνουν *φωνητικές μονάδες* (πχ φωνήματα, φθόγγους, λέξεις κλπ). Αναφερόμαστε γενικά σε φωνητικές μονάδες γιατί εξαρτώνται απ'τη σχεδίαση του TTS συστήματος. Για να γίνει η μετάφραση πρέπει να είναι γνωστοί οι φθόγγοι της γλώσσας για την οποία κατασκευάζεται το κομμάτι αυτό. Επίσης χρειάζεται να εισαχθούν με κάποιο τρόπο οι κανόνες προφοράς και γραμματικής της γλώσσας (πχ στην ελληνική το «*αι*» προφέρεται /ε/).

Απ'την γραμματική ανάλυση του κειμένου βγάζουμε στοιχεία για την προσωδία της ομιλίας. Πχ ο τονισμός είναι εύκολο να βρεθεί στην ελληνική γλώσσα και μας καθοδηγεί στο ποιός φθόγγος μέσα σε μια λέξη πρέπει να τονιστεί περισσότερο. Αυτό συνήθως μεταφράζεται σε μεγαλύτερη ένταση, τονικότητα και χρονική διάρκεια. Στην αγγλική γλώσσα είναι πιο δύσκολο να βρούμε ποιός φθόγγος τονίζεται μια και δεν υπάρχει σημάδι τονισμού.

Πληροφορία για τη προσωδία εξάγουμε και από τα σημεία στίξης. Καταλαβαίνουμε τη θέση της κάθε λέξης μέσα στη πρόταση (αν είναι πρώτη, τελευταία) καθώς και το είδος της πρότασης (αν τελειώνει σε τελεία, ερωτηματικό). Επίσης τα κόμματα μας δίνουν πληροφορία για τη χρονική εξέλιξη της προφοράς.

Υπάρχει συχνά πληροφορία για τη προσωδία που είναι χρήσιμη αλλά δεν μπορούμε να την προσδιορίσουμε απ'τη γραμματική ανάλυση. Πχ στη πρόταση «*Αύριο έρχεται ο Δημήτρης*» μπορούμε να φορτίσουμε τη λέξη «*αύριο*» αν θέλουμε να τονίσουμε το χρόνο άφιξης ή τη λέξη «*Δημήτρης*» αν θέλουμε να τονίσουμε το πρόσωπο της άφιξης.

Οι διαδικασίες που προσπαθούν να βρούν πληροφορία για τη προσωδία έχουν και τη μεγαλύτερη δυσκολία σε αντίθεση με τη μετάφραση από γράμματα σε φωνητικά κομμάτια που γενικά λύνεται πιο εύκολα. Όπως είπαμε και στη παράγραφο 2.4 αρκετή πληροφορία για τη προσωδία περιέχεται στο νόημα των λέξεων. Αυτό απαιτεί συντακτική ανάλυση του κειμένου (εύρεση ρήματος, υποκείμενου, δευτερεύουσας πρότασης) ή ακόμα και σημασιολογική ανάλυση, παραδείγματος χάρη να αξιολογούμε τη δραστικότητα ενός ρήματος (αλλιώς προφέρουμε μια πρόταση της οποίας το ρήμα είναι το σκοτώνω και αλλιώς αν είναι το χορεύω). Η πλούσια προσωδία είναι σημαντικός παράγοντας για να ακούγεται η συνθετική ομιλία φυσική. Γι'αυτό το λόγο η πληροφορία που εξάγουμε από το κείμενο για τη προσωδία επιδρά σημαντικά στο τελικό αποτέλεσμα.

2.6.2 Συνθέτης σήματος

Ο *συνθέτης φωνητικού σήματος* (*speech synthesizer*) δέχεται σαν είσοδο διανύσματα που περιγράφουν με κάποια αφαίρεση το σήμα φωνής και στην έξοδο δίνει το σήμα στο χρόνο. Έχουμε ήδη περιγράψει έναν σύνθετη φωνητικού σήματος στη παράγραφο 2.3 αυτό που χρησιμοποιεί ο άνθρωπος. Στο σύνθετη αυτό αν καθορίσουμε τις θέσεις των στοιχείων άρθρωσης και το τύπο διέγερσης τότε παράγεται ο αντίστοιχος ήχος (αρκεί να διοχετεύσουμε αέρα εκπνέοντας).

Ο συνθέτης συνήθως βασίζεται σε κάποιο *μοντέλο φωνητικού σήματος* (*speech signal modeling*). Το μοντέλο καθορίζεται πλήρως από τις παραμέτρους του. Είναι επιθυμητό το μοντέλο φωνητικού σήματος να είναι πλήρες (επαρκές) αλλά όχι πολύ γενικό.

Λέγοντας πλήρες εννοούμε ότι αν οποιοδήποτε σήμα φωνής το αναλύσουμε με βάση το μοντέλο και υπολογίσουμε τις παραμέτρους που το περιγράφουν και μετά το ξανασυνθέσουμε απ'τις παραμέτρους, τότε η ακουστή διαφορά του αρχικού και του συνθετικού σήματος πρέπει να είναι μικρή. Αυτό σημαίνει ότι το μοντέλο θα είναι ικανό να περιγράψει σωστά οποιοδήποτε σήμα φωνής.

Το μοντέλο φωνητικού σήματος δε χρειάζεται και δε συμφέρει να είναι γενικό γιατί θέλουμε να περιγράψει σήματα που παράγονται με συγκεκριμένο και γνωστό τρόπο. Το γεγονός ότι ξέρουμε τη πηγή του σήματος καθώς και τις ιδιότητες της πρέπει να μας καθοδηγεί στη σχεδίαση του μοντέλου. Ωστόσο το μοντέλο πρέπει να είναι παραστατικό δηλαδή οι παράμετροι να έχουν κάποιο φυσικό φωνητικό νόημα για να μπορούμε να τις ελέγχουμε. Σ'αυτή τη φάση δεν μας ενδιαφέρει η συμπίεση του σήματος φωνής η οποία όμως μας ενδιαφέρει απ'τη πλευρά της υλοποίησης. Υπάρχει ακόμα ένας λόγος που δε θέλουμε ο συνθέτης φωνητικού σήματος να είναι γενικός. Ο συνθέτης συνήθως δέχεται διανύσματα 20 έως 50 παραμέτρων και είναι πολύ δύσκολο να ξέρουμε αν ένα διάνυσμα παραμέτρων αντιστοιχεί σε φωνητικό ήχο. Δεν ξέρουμε δηλαδή το χώρο των παραμέτρων μέσα στον οποίο ο συνθέτης παράγει φωνητικά σήματα. Το ιδανικό θα ήταν ο συνθέτης να παράγει όλα τα δυνατά φωνητικά σήματα και μόνο αυτά.

Ο ρυθμός με τον οποίο ανανεώνονται οι παράμετροι του σύνθετη είναι σχετικά γρήγορος (τυπική τιμή κάθε $10msec$). Όπως θα φανεί από τη περιγραφή του ενδιάμεσου τμήματος ενός συστήματος TTS, δεν έχουμε συνεχώς πληροφορία για τις παραμέτρους. Έτσι σε πολλές χρονικές στιγμές τις υπολογίζουμε από παραμέτρους σε γειτονικές χρονικές στιγμές με παρεμβολή. Είναι σημαντικό λοιπόν οι παράμετροι να επιδέχονται παρεμβολή. Για παράδειγμα στους συντελεστές γραμμικής παρεμβολής δε μπορούμε να κάνουμε παρεμβολή γιατί το ψηφιακό φίλτρο που προκύπτει δεν είναι σίγουρα ευσταθές. Επίσης θέλουμε

με τη παρεμβολή ο συνθέτης να παράγει φωνητικά σήματα δηλαδή ο χώρος φωνητικής λειτουργίας του σύνθετη να είναι κυρτός.

2.6.3 Μετάφραση των φωνημάτων σε παράμετρους του συνθέτη σήματος

Το ενδιάμεσο τμήμα ενός συστήματος TTS αναλαμβάνει να μετατρέψει την ακολουθία φωνητικών μονάδων (φωνήματα , φθόγγοι) που κατασκευάζεται από τη γραμματική ανάλυση σε ακολουθία παραμέτρων για να τροφοδοτήσει το σύνθετη φωνητικού σήματος. Ένα βασικό πρόβλημα στη μετατροπή αυτή είναι ότι ο ρυθμός εισόδου και εξόδου είναι διαφορετικός. Τα φωνητικά σύμβολα έχουν ρυθμό παρόμοιο μ'αυτόν που τα παράγει ο άνθρωπος όταν μιλάει[6] ενώ οι παράμετροι του σύνθετη φωνητικού σήματος ανανεώνονται τυπικά κάθε 5 ή $10msec$.

Παρόλο που για τα καθαρά φωνητικά σύμβολα (απομονωμένοι φθόγγοι) συνήθως έχουμε τα αντίστοιχα διανύσματα παραμέτρων από ανάλυση πραγματικής φωνής, ωστόσο στο συνεχή λόγο δεν ισχύουν οι ίδιες αντιστοιχίες. Γενικά, είναι δύσκολο να βρούμε μια συνάρτηση αντιστοίχισης που να αποδίδει τα φωνητικά σύμβολα στο συνεχή λόγο. Ο βαθμός της δυσκολίας εξαρτάται άμεσα από το είδος του σύνθετη φωνητικού σήματος. Ακόμα πιο δύσκολο είναι να βρεθεί η αντιστοιχία για τους ηχούς που προφέρονται μεταξύ των φωνημάτων, δηλαδή να αποδοθεί το φαινόμενο της συμπροφοράς[7].

Για τη λύση αυτών των προβλημάτων είναι σημαντικό οι παράμετροι του σύνθετη φωνητικού σήματος να επιδέχονται παρεμβολή όπως αναφέραμε και στη προηγούμενη παράγραφο. Αν για δύο φωνήματα ο συνθέτης παράγει τις προφορες δύο φωνητικών συμβόλων τότε είναι επιθυμητό για τις τιμές ανάμεσα στα δύο διανύσματα ο συνθέτης να παράγει ηχούς ενδιάμεσους στις δύο προφορες. Στην αντίθετη περίπτωση πρέπει να βρούμε τη τροχιά των παραμέτρων για να πάμε από τη μια προφορά στην άλλη αποδίδοντας τη συμπροφορά. Ο εύκολος χειρισμός των παραμέτρων εξαρτάται απ'το αν έχουν κάποια φυσική σημασία (πχ θέση στοιχείων άρθρωσης, μήκος ή πλάτος ηχητικού σωλήνα κλπ).

Το τμήμα αυτό εκτός από φωνητικά σύμβολα από τη γραμματική ανάλυση του κειμένου δέχεται και οδηγίες για τη προσωδία τους. Με βάση τις οδηγίες αυτές πρέπει να μεταβάλλει τις παραμέτρους του σύνθετη. Η ευκολία της λειτουργίας αυτής εξαρτάται από τη σημασία των παραμέτρων. Πχ αν στη

[6]ο ρυθμός αυτός επιπλέον δεν είναι σταθερός γιατί η χρονική διάρκεια προφοράς των φωνημάτων δεν είναι ίδια για όλα τα φωνήματα

[7]Μία λύση που έχει δοθεί σ'αυτό το πρόβλημα είναι να αποθηκευτούν οι μεταβάσεις των φωνημάτων από πραγματική ομιλία

προσωδία αναφέρεται αύξηση της τονικότητας και μια παράμετρος είναι η θεμελιώδης συχνότητα, η αντιστοίχιση είναι άμεση. Στην αντίθετη περίπτωση απαιτείται επιπλέον επεξεργασία για να βρεθεί η τροχιά των παραμέτρων που αποδίδει την επιθυμητή προσωδία.

2.7 Κριτήρια αξιολόγησης TTS συστήματων

Προφανής χρησιμότητα μιας μεθόδου αξιολόγησης συστημάτων TTS είναι ότι μπορούμε να συγκρίνουμε διαφορετικά συστήματα TTS αλλά και να αξιολογήσουμε διαδοχικές εκδόσεις του ίδιου συστήματος για να μετρήσουμε την βελτίωσή του.

Ένας τρόπος να προσεγγίσουμε το πρόβλημα είναι να αξιολογήσουμε τη αξία ένας συστήματος TTS από τη μεριά του ακροατή. Αυτό μας δίνει ένα μέτρο του πόσο καλό (κακό) είναι ένα σύστημα TTS χωρίς να γνωρίζουμε το γιατί είναι καλό (κακό) [BP92]. Αυτό μπορεί να γίνει κάνοντας πειράματα με ακροατές σε διάφορες συνθήκες. Οι συνθήκες επηρεάζονται από πολλούς παράγοντες πχ θόρυβος περιβάλλοντος (εργαστηριακό περιβάλλον, σε κάποια μηχανή αυτόματης συνδιαλλαγής, ακρόαση μέσω τηλεφώνου), τη ικανότητα του ακροατή (είναι παιδί, δεν μιλάει στη μητρική του γλώσσα, προβλήματα ακοής), την παράλληλη ενασχόληση του ακροατή με κάτι άλλο (οδήγηση οχήματος, συνομιλία με άλλο άτομο) αλλά και τη κατάσταση του ακροατή (κόπωση, άγχος).

Όλα τα συστήματα TTS που έχουν φτιαχτεί παράγουν σε κάποιο βαθμό κατανοητή ομιλία. Όμως, ένα μήνυμα είναι πιο δύσκολα καταληπτό όσο πιο πολύπλοκο και περιεκτικό σε νόημα είναι[8]. Έτσι, όσο πιο πολύπλοκο είναι ένα μήνυμα τόσο περισσότερη ακουστική ποιότητα χρειάζεται για να γίνει καταληπτό από τον ακροατή. Αντίστοιχα από τη μεριά του ακροατή, όσο η ποιότητα ακρόασης με την οποία ακούει μειώνεται, τα μηνύματα που μπορεί να αντιληφθεί πρέπει να είναι πιο απλά. Έτσι όταν συζητάμε πρέπει η ακουστική ποιότητα που ”εκπέμπουμε” να περνάει κάποιο κατώλφι ακουστικής ποιότητας που θέτει ο ακροατής για το συγκεκριμένης πολυπλοκότητας μήνυμα. Ένας άνθρωπος όταν μιλάει ρυθμίζει αυτή την ”εκπεμπόμενη” ποιότητα (πχ μιλάει πιο αργά και πιο καθαρά) ανάλογα με το περιεχόμενο της ομιλίας του και τις ικανότητες του ακροατή. Δεδομένου ότι τα συστήματα TTS έχουν μια σταθερή ”εκπεμπόμενη” ποιότητα μπορούν να προφέρουν μηνύματα μέχρι κάποιο όριο πολυπλοκότητας.

[8] δηλαδή δεν είναι προβλέψιμο και δεν έχει περίσσεια πληροφορίας όπως για παράδειγμα ένας δυσνόητος μαθηματικός ορισμός.

Κριτήρια αξιολόγησης που έχουν προταθεί μετρούν το ποσοστό σωστής αντίληψης από διάφορους ακροατές, των μηνυμάτων που εκφωνεί ένα σύστημα TTS. Τα μηνύματα μπορεί να έχουν κάποιο νόημα ή να είναι χωρίς νόημα. Ένα σημαντικό πρόβλημα που προκύπτει στη πρώτη περίπτωση για συστήματα TTS διαφορετικής γλώσσας είναι ότι η κατανοητικότητα των μηνυμάτων των διαφορετικών γλωσσών πρέπει να είναι ίδια.

Κεφάλαιο 3

Επισκόπηση πεδίου-State of the Art

Στο κεφάλαιο αυτό θα αναφερθούμε σε σχετική δουλειά που έχει γίνει στο χώρο των συστημάτων TTS. Συγκεκριμένα θα κάνουμε μια σύντομη ιστορική αναδρομή μια και προσπάθειες υλοποίησης έχουν γίνει πολύ πριν την εξέλιξη των υπολογιστών. Στη συνέχεια θα περιγράψουμε τρεις κατηγορίες συστημάτων TTS αναφέροντας πρώτα τέσσερις σύνθετες φωνητικού σήματος που χρησιμοποιούνται κατά κόρο και στις τρεις κατηγορίες. Συνήθως η αρχή λειτουργίας του δεύτερου τμήματος (μετατροπή φθόγγων σε παραμέτρους του σύνθετη σήματος) καθορίζει και τη κατηγορία που ανήκει το σύστημα TTS. Στο τέλος θα αναφέρουμε μεθόδους παραγωγής προσωδίας και μετατροπής ορθογραφικής γραφής σε φωνητική, προβλήματα τα οποία αφορούν το πρώτο τμήμα ενός συστήματος TTS.

3.1 Ιστορική αναδρομή

Η πρώτη προσπάθεια σύνθεσης ομιλίας έγινε το 18ο αιώνα απ' τον Wolfgang von Kempelen. Η μηχανική κατασκευή του έμοιαζε με το ανθρώπινο σύστημα ομιλίας. Ο μηχανισμός διέγερσης είχε ένα γλωσσίδι (όπως στα πνευστά μουσικά όργανα) το οποίο πάλλοταν όταν διοχέτευε αέρα με τη βοήθεια ενός φυσερού. Για ακουστικό φίλτρο είχε ένα λαστιχένιο σωλήνα στον οποίο έδινε συγκεκριμένα σχήματα με το χέρι. Βοηθητικοί μηχανισμοί δημιουργούσαν στενέματα για τη παραγωγή συμφώνων. Η μηχανή μπορούσε να συνθέσει ολόκληρες φράσεις στα ιταλικά και στα γαλλικά. Προσπάθειες με μηχανικά μοντέλα συνεχίστηκαν μέχρι το 1920 οπότε οι ερευνητές στράφηκαν στα ηλεκτρικά αναλογικά κυκλώματα. Ο J.Q.Stewart παρουσίασε το 1939 τον Voder (voice demostrator). Ο χειριστής του Voder μπορούσε να επιλέξει ώς διέγερση θόρυβο ή αρμονικό σήμα (από έναν ταλαντωτή) για τα φωνήεντα καθώς και τη συχνότητα του αρμονικού

σήματος. Το φάσμα διαμόρφωνοταν από 10 διαδοχικά στη συχνότητα ζωνοπερατά φίλτρα με ελεγχόμενο απ' το χειριστή κέρδος. Στη συνέχεια προσπάθειες με ηλεκτρικά αναλογικά κυκλώματα έγιναν από τον Dum (1950) αλλά με την ανάπτυξη των υπολογιστών εγκαταλείφθηκαν.

3.2 Είδη TTS

Οι προσπάθειες που έχουν γίνει από τους διάφορους ερευνητές κατατάσσονται σε δύο μεγάλες κατηγορίες. Σ' αυτές που προσπαθούν να φτιάξουν μοντέλα για το ανθρώπινο σύστημα ομιλίας (*μοντέλα συστήματος, system models*) και σ' εκείνες που προσπαθούν να φτιάξουν μοντέλα για το φωνητικό σήμα (*μοντέλα σήματος, signal models*). Η πρώτη προσέγγιση είναι γνωστή και ώς *σύνθεση με άρθρωση* (*articulatory synthesis*) στην οποία οι παράμετροι του σύνθετη φωνητικού σήματος προσδιορίζουν τις θέσεις των στοιχείων άρθρωσης και το είδος διέγερσης. Στη δεύτερη κατηγορία ο συνθέτης φωνητικού σήματος δέχεται περιγραφές του φωνητικού σήματος. Υπάρχουν δύο υποκατηγορίες εδώ. Στη πρώτη οι παράμετροι παράγονται βάση κανόνων (*σύνθεση βάσει κανόνων, rule-based synthesis*) ενώ στη δεύτερη υπολογίζονται από ανάλυση πραγματικής φωνής. Η δεύτερη προσέγγιση αναφέρεται ώς *σύνθεση με συρραφή* (*concatenative synthesis*). Τα TTS συστήματα των δύο υποκατηγοριών που αναφέραμε διαφέρουν κυρίως ώς προς τη λειτουργία του δεύτερου τμήματος αλλά ο συνθέτης σήματος μπορεί να είναι ίδιος.

3.3 Είδη συνθετών φωνητικού σήματος

Οι σύνθετες σήματος των συστημάτων TTS βάσει κανόνων και των συστημάτων TTS με συρραφή, βασίζονται σε κάποιο μοντέλο φωνητικού σήματος (speech model). Συχνά όμως στα συστήματα TTS με συρραφή οι παράμετροι είναι απλά αφηρημένη περιγραφή του σήματος φωνής στο πεδίο της συχνότητας ή του χρόνου. Σύνθετες φωνητικού σήματος που χρησιμοποιούνται κατά κόρο είναι οι *formant σύνθετες* και οι σύνθετες γραμμικής πρόβλεψης που βασίζονται στο *μοντέλο «διέγερση-φίλτρο»* (*source-filter model*). Τεχνικές που βασίζονται στη περιγραφή του σήματος είναι η τεχνική *PSOLA* (Pitch Synchronization with OverLap and Add) και η *HNM* (Harmonic plus Noise Model).

3.3.1 Σύνθετες φωνητικού σήματος «διέγερση-φίλτρο»

Στο μοντέλο αυτό θεωρούμε ότι ο μηχανισμός διέγερσης (φώνηση ή θόρυβος) λειτουργεί ανεξάρτητα από την άρθρωση. Έτσι αν $s(n)$ είναι ένα σήμα φωνής τότε για το Z μετασχηματισμό του υποθέτουμε :

$$S(z) = U(z)H(z)$$

όπου $U(z)$ είναι η πηγή διέγερσης και $H(z)$ το φίλτρο του φωνητικού διαύλου και της εκπομπής των χειλιών. Το $H(z)$ αναλύεται :

$$H(z) = V(z)R(z)$$

όπου $V(z)$ είναι η απόκριση του φωνητικού διαύλου και $R(z)$ η απόκριση της εκπομπής των χειλιών. Η διέγερση $U(z)$ γράφεται :

$$U(z) = P(z)G(z)$$

όπου το $P(z)$ είναι μια παλμοσειρά από δέλτα ή/και λευκός θόρυβος ενώ το $G(z)$ υπάρχει μόνο στους έμφωνους φθόγγους και αντιπροσωπεύει τη δυναμική των φωνητικών χορδών και της φασματικής επίδρασης του λάρυγγα. Τελικά για τους έμφωνους φθόγγους έχουμε :

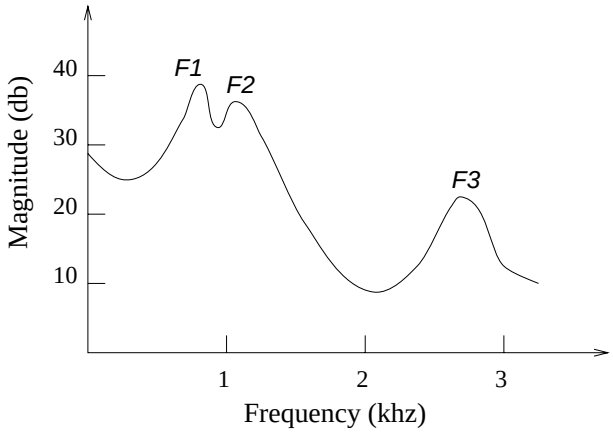
$$S(z) = P(z)G(z)V(z)R(z)$$

Τα P(z), G(z), V(z), R(z) μεταβάλλονται στο χρόνο. Συνήθως τα θεωρούμε σταθερά για χρονικό διάστημα 5 ή $10msec$ κάνοντας τη παραδοχή ότι τα στοιχεία άρθρωσης είναι σταθερά ή έχουν μετακινηθεί λίγο στο χρονικό αυτό διάστημα. Όταν όμως μιλάμε γρήγορα η παραδοχή δεν ισχύει για όλα τα χρονικά διαστήματα. Το μοντέλο υποθέτει την ανεξαρτησία των μηχανισμών διέγερσης και άρθρωσης. Η παραδοχή αυτή δεν είναι τελείως σωστή (Klatt 1987 [Kla87], Sorin 1994 [Sor94]).

3.3.1.1 Formant σύνθετες φωνητικού σήματος

Οι formant σύνθετες όπως αναφέραμε ανήκουν στην κατηγορία συνθετών «διέγερσης-φίλτρου» προσεγγίζουν το φωνητικό φάσμα ανακατασκευάζοντας συγκεκριμένες περιοχές του. Στο φάσμα της φωνής παρατηρούμε ζώνες συχνοτήτων με αυξημένο κέρδος (σχήμα 3.1). Τις συχνότητες στις οποίες εμφανίζεται αυτή η αύξηση της ονομάζουμε *formant συχνότητες*. Οι formant σύνθετες προσπαθούν να αποδώσουν το φωνητικό φάσμα σχηματίζοντας τέτοιες formant συχνότητες. Οι formant συχνότητες αντιστοιχούν στις συχνότητες συντονισμού των κοιλοτήτων κατά μήκος της φωνητικής οδού που λειτουργούν σαν

ακουστικά αντηχειά. Για τη προσέγγιση του φωνητικού φάσματος συνήθως χρησιμοποιούμε φίλτρα δεύτερης τάξης με μιγαδικούς πόλους. Η θέση των πόλων καθορίζει τη συχνότητα και το εύρος του συντονισμού. Μπορούμε να σχηματίσουμε το φωνητικό φάσμα συνδέοντας τέτοια φίλτρα (συνήθως 3 έως 6) εν σειρά ή παράλληλα. Ο klatt [Kla80] το 1980 παρουσίασε έναν formant σύνθετη σήματος που ελέγχονταν από 39 παραμέτρους οι οποίοι ενημερώνονταν κάθε 5 msec. Στο σύνθετη του υπήρχε η εν σειρά και η εν παραλλήλω σύνδεση των συντονισμένων φίλτρων. Οι παράμετροι έλεγχαν τη συχνότητα και το εύρος των συντονισμένων περιοχών καθώς και το είδος της διέγερσης. Ο συνθέτης του Klatt μπορούσε να συνθέσει ομιλία πολύ καλής ποιότητας αλλά ο έλεγχος των παραμέτρων ήταν δύσκολος. Ο συνθέτης αυτός από τότε χρησιμοποιήθηκε σε πολλά TTS συστήματα από διάφορους ερευνητές.



Σχήμα 3.1: Περιβάλλουσα φάσματος φωνής (φθόγγος /α/).

Το φάσμα των φθόγγων παρουσιάζει συχνότητες με αυξημένο κέρδος τις οποίες ονομάζουμε formant συχνότητες. Στο σχήμα φαίνονται τρεις οι $F1$, $F2$, $F3$.

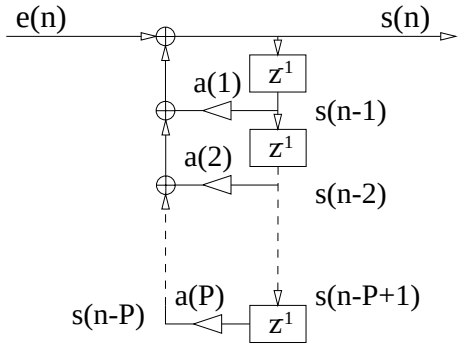
3.3.1.2 Σύνθετες φωνητικού σήματος γραμμικής πρόβλεψης

Στην ίδια κατηγορία συνθετών «διέγερση-φίλτρο» ανήκουν και οι σύνθετες γραμμικής πρόβλεψης. Στη μέθοδο αυτή συγκεκριμένα χρησιμοποιούμε τη συσχέτιση των γειτονικών δειγμάτων φωνής για να περιγράψουμε τη κυματομορφή. Θεωρούμε ότι μπορούμε να προβλέψουμε με κάποιο σφάλμα $e(n)$ ένα δείγμα $s(n)$ από τα P προηγούμενα :

$$s(n) = \sum_{i=1}^{P} \alpha(i) s(n-i) + e(n)$$

όπου $\alpha(i)$ είναι οι συντελεστές γραμμικής πρόβλεψης. Οι συντελεστές υπολογίζονται για ένα χρονικό χρονικό παράθυρο φωνής το οποίο υποθέτουμε

φασματικά σταθερό. Η συνθετική φωνή παράγεται από ένα φίλτρο P πόλων (σχήμα 3.2). Ο υπολογισμός των συντελεστών γίνεται με δύο τρόπους, από την *αυτοσυσχέτιση* ή την *συμμεταβλητότητα* του σήματος. Ο πρώτος τρόπος, σ' αντίθεση με τον δεύτερο, εξασφαλίζει ευστάθεια του φίλτρου. Αποδεικνύεται ότι το φίλτρο προσεγγίζει το φάσμα του σήματος με το μικρότερο δυνατό σφάλμα γι' αυτό και η φασματική του απόκριση είναι η περιβάλλουσα του σήματος ενώ το $e(n)$ είναι το θεμελιώδες σήμα διέγερσης της φωνής που ιδανικά είναι μια παλμοσειρά από δ και/ή λευκός θόρυβος. Παρόλο που το σήμα $e(n)$ δεν είναι παλμοσειρά δ, κατά την αναπαραγωγή διεγείρουμε το φίλτρο με μια παλμοσειρά δ της οποίας ελέγχουμε τη συχνότητα. Αν αντί για τους συντελεστές γραμμικής πρόβλεψης χρησιμοποιήσουμε *συντελεστές ανάκλασης* (*reflection coefficients*), οι οποίοι έχουν απόλυτη τιμή < 1, εξασφαλίζεται ευστάθεια για το φίλτρο που προκύπτει. Μπορούμε να αποφύγουμε τη μετατροπή των συντελεστών ανάκλασης σε συντελεστές γραμμικής πρόβλεψης υλοποιώντας lattice φίλτρα.



Σχήμα 3.2: Φίλτρο μη πεπερασμένης κρουστικής απόκρισης (IIR) με P πόλους.

Η ποιότητα της φωνής που παράγεται είναι μέτρια και οφείλεται κυρίως στην απλοποίηση που γίνεται στη διέγερση. Το σήμα $e(n)$ διαφέρει αισθητά από παλμοσειρά δ κι αυτό οφείλεται στη προσπάθεια να προσεγγίσουμε τη περιβάλλουσα της φωνής με φίλτρο που έχει μόνο πόλους.

3.3.2 Σύνθετες φωνητικού σήματος με περιγραφή του short-term σήματος

3.3.2.1 Σύνθεση φωνητικού σήματος με τη μέθοδο PSOLA

Με τη τεχνική PSOLA (Pitch Synchronous OverLap and Add, Charpentier και Stella 1986 [CS86]) μπορούμε να αναλύσουμε ανθρώπινη ομιλία ή κομμάτια ομιλίας και μπορούμε να επέμβουμε στη διάρκεια και τη τονικότητα της φωνής με

στόχο πχ. να συνενωσουμε δύο κομμάτια φωνής εξομαλυνοντας ασυνέχειες στη θεμελειωδη συχνότητα.

Η πρώτη έκδοση της τεχνικής PSOLA ήταν η τεχνική TD-PSOLA (Time Domain PSOLA) σύμφωνα με την οποία το σήμα φωνής χωρίζεται σε διαδοχικά επικαλυπτόμενα σήματα μικρής διάρκειας χρησιμοποιώντας ένα Hanning παράθυρο. Ο χωρισμός σε τμήματα γίνεται με ρυθμό σύγχρονο με τη θεμελιώδη συχνότητα για τα έμφωνα μέρη της φωνής και σταθερό για τα αφωνα. Ο έλεγχος της θεμελιώδους συχνότητας γίνεται με τη μεταβολή της διάρκειας των τμημάτων φωνής ενώ του ρυθμού με επανάληψη ή αφαίρεση κάποιων τμημάτων κατά την αναπαραγωγή. Με τη τεχνική TD-PSOLA δε μπορούμε να εξαλείψουμε φασματικές ασυνέχειες.

Στην έκδοση FD-PSOLA (frequency domain) υπολογίζεται η φασματική περιβάλλουσα των τμημάτων φωνής και διαιρώντας το συνολικό φάσμα με τη περιβάλλουσα υπολογίζεται και το φάσμα της διέγερσης. Ελέγχοντας τη διέγερση μεταβάλλεται η θεμελιώδης συχνότητα ενώ μεταβάλλοντας τη περιβάλλουσα εξαλείφονται φασματικές ασυνέχειες στα σημεία ένωσης διπλανών κομματιών φωνής.

Επόμενες εκδόσεις της τεχνικής αυτής είναι η τεχνική LP-PSOLA (Linear Prediction) (Moulines, Charpentier 1990 [MC90]) και η τεχνική MBR-PSOLA (Multi-Band Re-synthesis) (Dutoit, Leich [DL93] 1993)

3.3.2.2 Σύνθεση φωνητικού σήματος με τη μέθοδο HNM

Η τεχνική HNM (Harmonic plus Noise Model) (Στυλιανού 1996 [Sty96] [LM95]) μοιάζει ώς προς τη λειτουργία με τη τεχνική FD-PSOLA. Η φωνή χωρίζεται σε διαδοχικά τμήματα όπως και στη τεχνική PSOLA. Θεωρούμε ότι τα τμήματα φωνής είναι το άθροισμα

$$S(n) = H(n) + N(n)$$

όπου $H(n)$ είναι το αρμονικό τμήμα και $N(n)$ χρωματισμένος θόρυβος. Υποθέτουμε ότι το τμήμα κομμάτι καλύπτει το χαμηλό μέρος του φάσματος και ο θόρυβος το υψίσυχνο. Για τη περιγραφή του φάσματος αποθηκεύεται :

Για το αρμονικό τμήμα η θεμελιώδης συχνότητα F_0, το πλάτος και η φάση των αρμονικών της $F_0, 2F_0$ κλπ μέχρι μια μέγιστη αρμονική συχνότητα F_m που επίσης αποθηκεύεται. Μ' αυτό το τρόπο δειγματοληπτείται η περιβάλλουσα του αρμονικού κομματιού.

Για το θόρυβο οι συντελεστές γραμμικής πρόβλεψης ενός φίλτρου, που έχει μόνο πόλους, του οποίου η απόκριση προσεγγίζει το φάσμα του θορύβου.

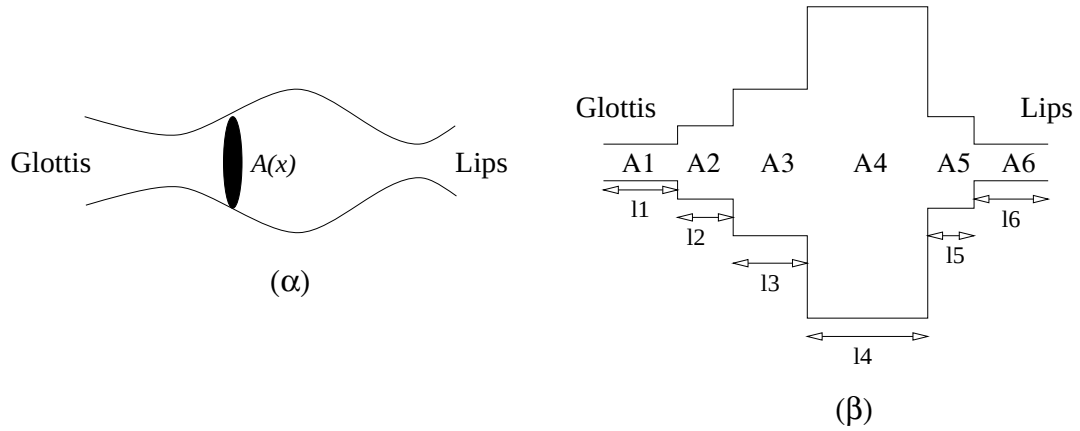
Κατά τη σύνθεση το αρμονικό κομμάτι δημιουργείται από τη άθροιση των αρμονικών συνιστωσών του και του θορύβου που παράγεται από το φίλτρο το

οποίο τροφοδοτείται με λευκό θόρυβο. Μεταβολή στη θεμελιώδη συχνότητα επιτυγχάνεται ανασυνθέτοντας το αρμονικό κομμάτι με διαφορετική F_0 αλλά την ίδια περιβάλλουσα. Το πλάτος και τη φάση των νέων συνιστωσών $F_0', 2F_0'$ κλπ υπολογίζονται με παρεμβολή από τα αρχικά. Ο έλεγχος της χρονικής εξέλιξης γίνεται μεταβάλλοντας τη χρονική απόσταση των τμημάτων φωνής επαναλαμβάνοντας/αγνοώντας κάποια όταν αυξάνεται/μειώνεται η χρονική διάρκεια.

Η τεχνική HNM συνθέτει φωνή πολύ καλής ποιότητας και, παρόλο που η ανάλυση είναι χρονοβόρα (αλλά για εφαρμογές TTS γίνεται μια φορά), η σύνθεση είναι γρήγορη.

3.4 Συστήματα TTS με άρθρωση

Τα σύστημα TTS με άρθρωση μοντελοποιούν το ανθρώπινο σύστημα ομιλίας. Όπως είπαμε η ομιλία βασίζεται σε δύο λειτουργίες, τη διέγερση (φώνηση και/ή παραγωγή θορύβου) και την άρθρωση. Προσπαθούμε να μοντελοποιήσουμε τις δύο αυτές λειτουργίες. Συνήθως μοντελοποιούμε μόνο το μηχανισμό άρθρωσης γιατί η φώνηση και ο θόρυβος μοντελοποιούνται και ελέγχονται εύκολα και με μοντέλα σήματος αν και έχουν γίνει μηχανικά μοντέλα και για το μηχανισμό διέγερσης.



Σχήμα 3.3: Γεωμετρικά μοντέλα προσέγγισης του φωνητικού διαύλου.

Στο (α) βλέπουμε ένα μοντέλο φωνητικού διαύλου όπου η διατομή είναι συνεχής συνάρτηση $A(x)$. Στο (β) η προσέγγιση γίνεται με περιοχές σταθερής διατομής που κάνει την εφαρμογή των εξισώσεων διάδοσης κυμάτων ευκολότερη. Πιο συγκεκριμένα καθώς περνάμε από το ένα τμήμα A_i στο επόμενο A_{i+1} ένα μέρος του ηχητικού κύματος διαδίδεται προς τα χείλια και ένα μέρος ανακλάται προς τα πίσω. Βρίσκοντας σε κάθε συναρμογή την διάδοση και την ανάκλαση υπολογίζουμε την απόκριση του προσεγγιστικού διαύλου.

Ο φωνητικός δίαυλος έχει μήκος περίπου 17 εκ. και διατομή που αλλάζει κατά μήκος του φωνητικού διαύλου και είναι συνάρτηση του χρόνου. Σχεδόν όλα τα μοντέλα που έχουν προταθεί καταλήγουν σε διδιάστατα μοντέλα. Συνήθως γίνεται η παραδοχή ότι η απόκριση του φωνητικού διαύλου δεν αλλάζει αισθητά αν ο δίαυλος γίνει ευθύγραμμος [DPH93]. Στο σχήμα 3.3α βλέπουμε ένα μοντέλο φωνητικού διαύλου όπου η διατομή είναι συνεχής συνάρτηση του μήκους. Μία προσέγγιση φαίνεται στο σχήμα 3.3β όπου η διατομή είναι σταθερή κατά περιοχές. Οι Proakis. et al. [DPH93] παρουσιάζουν μοντέλα με δύο περιοχές για τη παραγωγή διαφόρων φωνηέντων. Σ' αυτά τα μοντέλα οι παράμετροι ελέγχουν τη γεωμετρία της φωνητικής οδού.

Παρόμοιες εργασίες χαρτογραφούν το πραγματικό σχήμα της φωνητικής οδού από ακτινογραφίες X [Gab94]. Μία απλή προσέγγιση είναι η χωρική δειγματοληψία της χαρτογράφησης αυτής. Έτσι οι θέσεις των στοιχείων άρθρωσης αντιστοιχίζονται κατευθείαν στη γεωμετρία της φωνητικής οδού. Για να γίνει όμως αυτό χρειάζονται δεδομένα από κινηματογράφηση και φωτογράφηση με ακτίνες X ομιλητών που προφέρουν κάποιους φθόγγους. Μία άλλη προσέγγιση χρησιμοποιεί δείκτες για τη γεωμετρία της φωνητικής οδού (θέση γλώσσας, ουρανίσκου κλπ). Συνήθως το σχήμα της φωνητικής οδού το μετατρέπουμε σε μια συνάρτηση απόκρισης είτε εφαρμόζοντας τις εξισώσεις διάδοσης κυμάτων ή προσπαθώντας να μετατρέψουμε το φυσικό μοντέλο σε μοντέλο σήματος (πχ. διέγερση-φίλτρο) βρίσκοντας μαθηματικές εξαρτήσεις απ' το ένα στο άλλο.

Πολλοί ερευνητές περιγράφουν τη ομιλία σαν ένα συνεχές κυνηγητό των στοιχείων άρθρωσης να πάρουν κάποιες τελικές θέσεις που αλλάζουν με το χρόνο. Η τελικές θέσεις αντιστοιχούν στις θέσεις των στοιχείων άρθρωσης όταν αρθρώνουν καθαρούς φθόγγους. Οι τελικές αυτές θέσεις αλλάζουν καθώς πρέπει να αρθρώσουμε διαδοχικά τους φθόγγους. Η δυναμική των στοιχείων άρθρωσης όμως (μάζα, δύναμη των μυών, αλληλοεξαρτήσεις) εμποδίζουν την επίτευξη των διαδοχικών στόχων. Με το σκεπτικό αυτό μπορούμε να εφαρμόσουμε μηχανικά μοντέλα για την απόδοση των φαινομένων συμπροφοράς.

Παρόλο που με τη προσέγγιση αυτή προσομοιώνουμε το ανθρώπινο σύστημα ομιλίας και έχουμε πολύ καλά αποτελέσματα, η προσοχή των ερευνητών έχει στραφεί σε μοντέλα φωνητικού σήματος γιατί είναι πιο απλά και μπορούν να βρεθούν εύκολα μετρήσεις (φωνή). Αντίθετα οι μετρήσεις για το ανθρώπινο σύστημα ομιλίας είναι λίγες.

3.5 Συστήματα TTS με formant σύνθετη σήματος βάσει κανόνων

Στα συστήματα TTS αυτής της κατηγορίας η μετατροπή των φθόγγων ή φωνημάτων σε παραμέτρους του σύνθετη γίνεται *βάσει κανόνων*. Οι κανόνες οδηγούν με παραμέτρους το σύνθετη σήματος, ανάλογα με το φθόγγο ή φώνημα που πρέπει να προφερθεί κάθε χρονική στιγμή λαμβάνοντας υπόψη τους γειτονικούς φθόγγους (αλλοφωνικές αλλοιώσεις). Επίσης πρέπει να αποδίδουν και τα μεταβατικά φαινόμενα μεταξύ των φθόγγων. Οι κανόνες προκύπτουν από ανάλυση πραγματικής φωνής ώστε να συγκλίνουν στη σωστή προφορά συνεχούς ομιλίας. Συνήθως οι κανόνες αυτοί συνδυάζονται με τους κανόνες απόδοσης προσωδίας.

Οι Kelly και Gerstman [KG61] έκαναν τις πρώτες προσπάθειες χρησιμοποιώντας ένα formant σύνθετη σήματος και κανόνες που όρισαν βασιζόμενοι σε πειράματα. Στη δεκαετία του 80 πολλοί ερευνητές στράφηκαν προς αυτή τη κατηγορία των TTS συστημάτων. Απ' τα πιο γνωστά είναι το MITalk (Allen et al. 1987 [AHK87]), το Klattalk (Klatt 1982 [Kla82]) απ' το οποίο προήλθε το DECtalk (Bruckert 1983 [BT83]). Όλα τα παραπάνω χρησιμοποιούν τον formant σύνθετη του Klatt (1980)[Kla80] ή μεταγενέστερες εκδόσεις του.

3.6 Τεχνική συρραφής φωνητικών μονάδων

Με τη τεχνική αυτή η συνθετική φωνή προκύπτει από συρραφή τμημάτων πραγματικής ομιλίας. Θα ονομάσουμε αυτά τα τμήματα φωνής *φωνητικές μονάδες* που όπως θα δούμε έχουν διάφορα μεγέθη. Στη βιβλιογραφία βρίσκουμε εργασίες που χρησιμοποιούν φωνητικές μονάδες μεγέθους λέξης μέχρι μεγέθους φωνήματος ή και μικρότερες από φώνημα. Με τη τεχνική αυτή επιδιώκουμε τα φαινόμενα συμπροφοράς και τις αλλοφωνικές αλλοιώσεις να τις αντιγράψουμε από πραγματική φωνή.

Οι φωνητικές μονάδες συνήθως είναι αποθηκευμένες σε μορφή παραμέτρων του σύνθετη, η ανάλυση της πραγματικής φωνής δηλαδή γίνεται μια φορά. Έτσι η αντιστοίχιση των φωνητικών μονάδων σε παραμέτρους του σύνθετη γίνεται με ένα *ευρετήριο* (*lookup table*). Στη συνθετική φωνή που προκύπτει από τη συρραφή των φωνητικών μονάδων πρέπει να εξομαλυνθούν οι ασυνέχειες στα σημεία συρραφής και να αποδώσει η προσωδία της πρότασης το οποίο επιτυγχάνεται με τεχνικές επεξεργασίας σήματος σαν αυτές που αναφέρονται στη παράγραφο 3.3.2. Στη συνέχεια θα αναφέρουμε συστήματα TTS με συρραφή που λειτουργούν με διαφορετικό μέγεθος φωνητικών μονάδων.

3.6.1 Λέξεις

Ο άνθρωπος καταλαβαίνει το λόγο σαν μια ακολουθία λέξεων γι' αυτό και είναι το πρώτο που θα πρότεινε κάποιος μη ειδικός. Για τη συρραφή των λέξεων αντιμετωπίζουμε λίγα προβλήματα ασυνέχειας αν και συχνά δύο λέξεις προφέρονται σαν μια (*φωνητικές λέξεις*). Το φαινόμενο συμπροφοράς έχει καταγραφεί μέσα στη φωνητική μονάδα. Συμπροφορά όμως έχουμε και μεταξύ των λέξεων αν και είναι πιο χαλαρή. Έτσι και πάλι έχουμε να αντιμετωπίσουμε φασματικές ασυνέχειες και ασυνέχειες στη θεμελιώδη συχνότητα στα σημεία συρραφής που λύνονται με κάποια επεξεργασία σήματος. Ένα σημαντικό πρόβλημα είναι ότι οι λέξεις που έχουν προφερθεί απομονωμένες διαφέρουν πολύ απ' τις ίδιες λέξεις σε συνεχή ομιλία. Μπορεί να χρειαστεί λοιπόν η αποθήκευση της ίδιας λέξης σε διαφορετικές προτάσεις.

Ο Rabiner [RSF71] το 1971 έφτιαξε ένα TTS για τηλεφωνικά νούμερα αποθηκεύοντας προφορές λέξεων ενώ για τον έλεγχο της προσωδίας και την αντιμετώπιση των ασυνεχειών στα σημεία συρραφής χρησιμοποίησε έναν formant σύνθετη. Οι Fallside και Young (1978 [FY78]) χρησιμοποίησαν έναν σύνθετη σήματος γραμμικής πρόβλεψης και αποθήκευαν πολλές εκδόσεις της ίδιας λέξης.

Ένα TTS σύστημα που δεν έχει περιορισμούς στο λεξιλόγιο χρειάζεται τεράστιο λεξικό και πολύ χρόνο να κατασκευαστεί. Γι' αυτό και καταφεύγουμε σε μικρότερες φωνητικές μονάδες.

3.6.2 Συλλαβές

Οι *συλλαβές* της αγγλικής γλώσσας είναι περίπου 10000. Όπως και οι λέξεις απαιτούν μεγάλο χώρο αποθήκευσης και χρόνο να ηχογραφηθούν. Οι ασυνέχειες μεταξύ των συλλαβών είναι πιο έντονες. Η χρήση τους δεν παρέχει κανένα πλεονέκτημα γι' αυτό και δεν έχει τραβήξει το ενδιαφέρον των ερευνητών.

3.6.3 Ημισυλλαβές

Οι *ημισυλλαβές* είναι το πρώτο ή το δεύτερο μισό μιας συλλαβής. Και πάλι υπάρχουν ασυνέχειες μετά των ημισυλλαβών που πρέπει να αντιμετωπιστούν όπως επίσης και να αποδοθεί η συμπροφορά. Το πλήθος του όμως είναι μικρότερο. Η Bellcore (1996 [Bel96]) έχει παρουσιάσει το σύστημα ORATOR TTS που λειτουργεί με ημισυλλαβές.

3.6.4 Δίφωνα

Το *δίφωνο* ορίζεται σαν το κομμάτι φωνής που αρχίζει στη μέση ενός φωνήματος και τελειώνει στη μέση του επόμενου. Το δίφωνο επομένως περιέχει τη συμπροφορά μεταξύ των φωνημάτων. Τα σημεία συρραφής είναι στη μέση των φωνημάτων όπου το φάσμα τους είναι πιο σταθερό και οι φασματικές ασυνέχειες κατά τη συρραφή είναι μικρότερες. Αυτό διότι τη στιγμή αυτή τα στοιχεία άρθρωσης βρίσκονται πιο κοντά στις θέσεις στόχους (παράγραφος 3.4). Τα δίφωνα είναι < 1000 για την ελληνική γλώσσα οπότε η ηχογράφηση τους δεν είναι ιδιαίτερα χρονοβόρα. Για τη συρραφή έχουν χρησιμοποιηθεί διάφοροι σύνθετες φωνητικού σήματος (γραμμικής πρόβλεψης, PSOLA, formant σύνθετες). Ο συνθέτης χρησιμοποιείται και για τον έλεγχο της προσωδίας εκτός κι αν τα δίφωνα είναι πολλαπλά ηχογραφημένα με διαφορετική διάρκεια και τονικότητα. Επειδή τα δίφωνα καταγράφουν τη συμπροφορά παράγουν πολύ εύκολα κατανοητή φωνή γι' αυτό και πολλοί ερευνητές έχουν ασχοληθεί με TTS συστήματα συρραφής διφώνων.

Οι Dixon και Maxey (1968 [DM68]) χρησιμοποίησαν έναν formant σύνθετη για τον έλεγχο της διάρκειας και της τονικότητας και ένα λεξικό από 1000 δίφωνα

Ένα άλλο σύστημα αναπτύχθηκε το 1977 στα Bell Labs (Olive [Oli77]) στο οποίο δεν αποθηκεύεται ολόκληρο το δίφωνο αλλά μόνο το κομμάτι της μετάβασης απ' το αρχικό στο τελικό φώνημα δηλαδή το κέντρο του διφώνου. Η μέση ενός φωνήματος προκύπτει με παρεμβολή της αρχής του (που υπάρχει στο τέλος της προηγούμενης φωνητικής μονάδας) και του τέλους του (που υπάρχει στην αρχή της επόμενης). Ο τελικός αριθμός των φωνητικών μονάδων ήταν 600.

Ένα σύστημα TTS για γαλλικά αναπτύχθηκε στη Telecom to 1982 (Courbon και Emerard [CE82]) που χρησιμοποιεί 1200 δίφωνα και ένα σύνθετη γραμμικής πρόβλεψης.

Στη πρώτη έκδοση της τεχνικής PSOLA οι Charpentier και Stella 1986 [CS86] εφάρμοσαν συρραφή διφώνων.

3.6.5 Φωνήματα

Χρησιμοποιώντας φωνήματα σαν κομμάτια συρραφής αντιμετωπίζουμε περισσότερα προβλήματα από πριν. Η συρραφή τους είναι πιο δύσκολη αφού πρέπει να αποδοθεί η συμπροφορά και να εξομαλυνθούν φασματικές ασυνέχειες που είναι πιο έντονες μεταξύ διαφορετικών φωνημάτων. Επίσης πρέπει να αποθηκευθούν πολλές εκδόσεις του ίδιου φωνήματος με διαφορετικά γειτονικά φωνήματα.

Ο Hauptmann (1993 [Hau93]) έφτιαξε μια μεγάλη βάση από 3253 προτάσεις

που συνολικά είχαν 150000 φωνήματα(!!). Χρησιμοποιώντας ένα σύστημα αναγνώρισης ομιλίας και το ορθογραφικό κείμενο απομονώνει τα φωνήματα και στη συνέχεια κατατάσσει το καθένα ώς προς τη φόρτισή του, τους γείτονές του, τη θέση του μέσα στη λέξη και τη πρόταση. Κατά τη σύνθεση χρησιμοποιεί μια ευρεστική μέθοδο για να επιλέγει το πιο ταιριαστό φώνημα για το κείμενο εισόδου ενώ στα σημεία συρραφής εξομαλύνει τοπικά τη θεμελιώδη συχνότητα με τη τεχνική PSOLA και δεν επεμβαίνει στη συνολική τονικότητα και χρονική εξέλιξη. Παρόμοιο σύστημα ανέπτυξαν οι Black και Campbell (1995 [BC95]) στο οποίο η επιλογή του καλύτερου φωνήματος γίνεται με την ελαχιστοποίηση μιας συνάρτησης κόστους που εξαρτάται από τη ακουστική συνέχεια των διαδοχικών φωνημάτων και τη θέση του φωνήματος μέσα στη πρόταση. Η μόνη επεξεργασία κατά τη συρραφή είναι η επιλογή των ορίων των φωνημάτων με ένα κριτήριο ακουστικής συνέχειας. Αργότερα για τη συρραφή χρησιμοποίησε τη τεχνική PSOLA.

3.7 Ανάλυση κειμένου

Το πρώτο τμήμα ενός συστήματος TTS εκτελεί 2 λειτουργίες, α) αναγνωρίζει τους φθόγγους (ή γενικά φωνητικές μονάδες) που σχηματίζουν τις λέξεις του κειμένου (*text-to-phoneme*) και β) παράγει τη προσωδία για τη προφορά του κειμένου (*text-to-prosody*).

3.7.1 Μετατροπή γραμμάτων σε φθόγγους

Αναφέραμε ότι μια από τις λειτουργίες του πρώτου τμήματος είναι να μετατρέψει το κείμενο σε ακολουθία φωνητικών μονάδων. Αναφερθήκαμε στο κείμενο σε διαφορετικά επίπεδα αφαίρεσης (παραγράφου, πρότασης, φράσης, λέξης, φθόγγου κλπ). Η αναγνώριση των αντικειμένων αυτών μέσα στο κείμενο δεν είναι τετριμμένο πρόβλημα. Η δυσκολία του ποικίλει ανάλογα με τη διάλεκτο του κειμένου.

Στην αρχή μιας παραγράφου συνήθως βρίσκουμε ένα tab χαρακτήρα αλλά tab χαρακτήρες μπορεί να υπάρχουν και σε άλλα σημεία του κειμένου. Παρόμοια η τελεία υποδηλώνει το τέλος μιας πρότασης αλλά τελεία βάζουμε και σε συντμήσεις, σε αρχικά ονομάτων και στους δεκαδικούς αριθμούς. Αυτές οι ασάφειες λύνονται συνήθως κοιτώντας γειτονικές λέξεις. Επίσης οι συντμήσεις πρέπει να αντικατασταθούν από τις αντίστοιχες λέξεις πράγμα που σε πολλές περιπτώσεις δεν είναι προφανές. Πχ ένα «*Ν.*» μπορεί να σημαίνει «*Νίκος*» ή «*Νότια*» ή «*Νέα*». Εύκολα εντοπίζουμε φράσεις μέσα σε εισαγωγικά ή παρενθέσεις

αλλά δύσκολα αναγνωρίζουμε δευτερεύουσες προτάσεις. Στο κείμενο συχνά περιέχονται χαρακτήρες που αντιπροσωπεύουν λέξεις πχ τα %, $, °C. Παρόμοια μετατροπή πρέπει να γίνει και για τους αριθμούς μόνο που πρέπει να προσέξουμε σε τί αντιστοιχούν οι αριθμοί (ώρα, ημερομηνία, τηλεφωνικά νούμερα) και να μπούν στη σωστή πτώση.

Η επεξεργασία που περιγράφηκε παραπάνω έχει σαν σκοπό να παράγει μια ακολουθία λέξεων οργανωμένες σε φράσεις, προτάσεις. Οι λέξεις όμως βρίσκονται ακόμα στη ορθογραφική γραφή τους ενώ εμείς χρειαζόμαστε την φωνητική γραφή τους δηλαδή τους φθόγγους απ' τους οποίους αποτελούνται. Αυτό στην ελληνική γλώσσα είναι σχετικά εύκολο γιατί στις περισσότερες περιπτώσεις «διαβάζουμε ότι γράφουμε» και όταν δεν ισχύει αυτό οι κανόνες μετατροπής είναι απλοί και σαφείς (αν και υπάρχουν εξαιρέσεις όπως η συνίζηση). Σε άλλες γλώσσες (και η αγγλική) η μετατροπή από γράμματα σε φθόγγους είναι δύσκολη. Πολλοί ερευνητές έχουν αντιμετωπίσει το πρόβλημα για την αγγλική γλώσσα. Έχουν προταθεί συστήματα που μετατρέπουν βάσει κανόνων ένα γράμμα ή γράμματα σ' ένα φθόγγο ανάλογα με τα γειτονικά τους γράμματα (Klatt 1987 [Kla87]). Άλλη προσέγγιση είναι ο χωρισμός της λέξης στα συνθετικά της, προθέσεις, καταλήξεις (Allen 1987 [AHK87]). Οι Lucassen και Mercer (1984 [LM94]) χρησιμοποιούν κρυφές αλυσίδες Markov.

3.7.2 Εξαγωγή προσωδίας

Λέγοντας *προσωδία* (*prosody*) της ομιλίας εννοούμε τρία στοιχεία, τη *τονικότητα* (θεμελιώδη συχνότητα), την *ενέργεια* ή *ένταση* και τη *χρονική εξέλιξη* (χρονική διάρκεια των φθόγγων, καθορισμός παύσεων κλπ). Ονομάζουμε *μουσική καμπύλη* (*pitch contour*) τη συνάρτηση που περιγράφει τη θεμελιώδη συχνότητα ώς προς το χρόνο. Έχει παρατηρηθεί ότι στην ανθρώπινη ομιλία η τονικότητα και η ένταση είναι σχεδόν γραμμικά εξαρτημένες. Γι' αυτό όταν θέλουμε να πούμε κάτι σιγανά το προφέρουμε βαρύτονα και όσο ανεβάζουμε την ένταση της φωνής μας, ανεβαίνει και η τονικότητα. Έτσι από τη μουσική καμπύλη μπορούμε εύκολα να βρούμε την ένταση της φωνής κάθε χρονική στιγμή. Για τον προσδιορισμό της προσωδίας λοιπόν θέλουμε τη μουσική καμπύλη και τη χρονική εξέλιξη.

Ένα ιδανικό σύστημα TTS θα έπρεπε να παράγει προσωδία έχοντας τη δυνατότητα να κατανοήσει το κείμενο που προφέρει. Αυτό θα απαιτούσε κάποιο είδος τεχνητής νοημοσύνης κατά την επεξεργασία του κειμένου. Δεδομένου ότι μέχρι τώρα αυτό δε γίνεται, απαιτούμε από ένα TTS σύστημα να προφέρει το κείμενο χωρίς τη φόρτιση από το νόημά του το οποίο αν και πιο απλό πρόβλημα

είναι ακόμα δύσκολο.

Ο άνθρωπος προφέρει το κείμενο σαν μια ακολουθία λέξεων οι οποίες είναι οργανωμένες σε *φωνητικές φράσεις*. Τις φωνητικές φράσεις τις προφέρουμε συνήθως με μια ανάσα και νοηματικά είναι μια πρόταση ή μια φράση. Είναι σημαντικό να βρούμε που χωρίζονται οι φωνητικές φράσεις για να παράγουμε προσωδία ανεξάρτητα για τη κάθε μία. Επίσης πρέπει να τις ξεχωρίσουμε και χρονικά δηλαδή να αφήσουμε ησυχία ανάμεσα τους που αντιστοιχεί στις στιγμές που παίρνουμε αναπνοή. Συχνά τα όρια των φωνητικών φράσεων βρίσκονται στα σημεία στίξης (κόμμα, τελεία, θαυμαστικό, ερωτηματικό) αλλά αυτό δε συμβαίνει πάντα και σε όλες τις γλώσσες. Πολλά TTS συστήματα παράγουν προσωδία για μία πρόταση οπότε τα όρια των φωνητικών φράσεων είναι οι τελείες, τα θαυμαστικά κλπ (MITalk [AHK87]. Άλλα λαμβάνουν υπόψη τους και τα κόμματα ή αναζητούν συγκεκριμένες λέξεις που προσδιορίζουν αρχή ή τέλος φράσης.

Χρονική διάρκεια φθόγγων

Έχουν προταθεί πολλά μοντέλα που λειτουργούν βάσει κανόνων. Παράγοντες που επηρεάζουν τη διάρκεια των φθόγγων είναι οι γειτονικοί φθόγγοι, η συχνότητα μιας λέξης, η συντακτική της σημασία, η θέση της μέσα στη πρόταση. Το MITalk (Allen at al. 1987 [AHK87]) χρησιμοποιεί trial and error αλγορίθμους που συγκρίνουν συνθετική με πραγματική φωνή για τη σύγκλιση. Ο Campbell (1992 [Cam92]) χρησιμοποιεί νευρωνικά δίκτυα, ο Riley (1990 [Ril90], 1992 [Ril92]) εφαρμόζει έναν αλγόριθμο ομαδοποίησης των φθόγγων ώς προς τη διάρκειά τους. Στην ίδια κατηγορία με του Riley είναι και ο αλγόριθμος της AT&T (van Santen 1994 [vS94]).

Μουσική καμπύλη

Η τονικότητα ενός φθόγγου επηρεάζεται από όλα τα επίπεδα αφαίρεσης του κειμένου αλλά και από τους φυσικούς νόμους που διέπουν τη λειτουργία της φώνησης (Pierrehumbert 1981 [Pie81]). Η τονικότητα ενός φθόγγου εξαρτάται απ' το αν τονίζεται ή όχι και απ' τη θέση του φθόγγου μέσα στη λέξη. Επίσης εξαρτάται απ' τη θέση της λέξης μέσα στη πρόταση, τη συντακτική της σημασία καθώς και το είδος της πρότασης (ερωτηματική, καταφατική).

Στο MITalk (Allen at al. 1987 [AHK87]) η μουσική καμπύλη ακολουθούσε τρία πρότυπα ανάλογα με το είδος της πρότασης (καταφατική, ερωτηματική επιβεβαιωτική, ερωτηματική μη επιβεβαιωτική).

Κεφάλαιο 4

Το σύστημα σύνθεσης ομιλίας από κείμενο για ελληνική γλώσσα

Στο κεφάλαιο αυτό θα περιγράψουμε το σύστημα TTS για την ελληνική γλώσσα που κατασκευάστηκε στα πλαίσια αυτής της μεταπτυχιακής εργασίας. Θα αναφέρουμε αρχικά γενικές επιλογές που έγιναν κατά τη σχεδίαση, δηλαδή σε ποιά κατηγορία ανήκει το σύστημά μας, και επιγραμματικά τις λειτουργίες του κάθε τμήματος. Στη συνέχεια θα παρουσιάσουμε τη διαδικασία κατασκευής της βάσης φωνητικών μονάδων και τη λειτουργία του συστήματος κατά τη σύνθεση αναλυτικά.

4.1 Επιλογές στη σχεδίαση

Το σύστημα TTS για την ελληνική γλώσσα που κατασκευάστηκε ανήκει στη κατηγορία των συστημάτων TTS με συρραφή. Στο σύστημα μας μια φωνητική μονάδα αποτελείται από μια συστοιχία διφώνων. Τη φωνητική μονάδα την ονομάσαμε *διασυλλαβή* και ορίζεται σαν το κομμάτι φωνής που αρχίζει από το μέσο ενός φωνήεντος και τελειώνει στο μέσο του επόμενου. Έτσι η διασυλλαβή μπορεί να είναι δίφωνο όταν μεταξύ των δύο φωνηέντων δεν μεσολαβεί σύμφωνο, τρίφωνο όταν μεσολαβεί ένα σύμφωνο κοκ. Η επιλογή των διασυλλαβών σαν φωνητική μονάδα έγινε γιατί έχουν τα πλεονεκτήματα των διφώνων (εύκολη επίτευξη κατανοητής συνθετικής ομιλίας, μικρές ασυνέχειες στα σημεία συρραφής, φαινόμενα συμπροφοράς μεταξύ των φθόγγων είναι προηχογραφημένα) και παράλληλα λαμβάνοντας υπόψη τις ιδιαιτερότητες της ελληνικής γλώσσας.

Το 1ο τμήμα του συστήματος αναγνωρίζει τις διασυλλαβές μέσα στο κείμενο το οποίο γίνεται βρίσκοντας πρώτα τους φθόγγους με βάση τους κανόνες της ελληνικής γλώσσας. Στο τμήμα αυτό επίσης σχηματίζουμε λέξεις και προτά-

σεις και αναγνωρίζουμε το είδος μιας πρότασης που θα μας χρησιμεύσει στη παραγωγή προσωδίας. Το 2ο τμήμα βρίσκει τις ηχογραφημένες προφορές των διασυλλαβών χρησιμοποιώντας ένα ευρετήριο. Στη συνέχεια τις συρράπτει εξαλείφοντας ασυνέχειες στα σημεία συνένωσης. Τέλος στο τμήμα αυτό αποδίδουμε τη προσωδία μιας πρότασης η οποία παράγεται βάσει κανόνων. Σαν συνθέτης φωνητικού σήματος (3ο τμήμα) χρησιμοποιήθηκε η τεχνική περιγραφής σήματος HNM (Harmonic plus Noise Model [Sty96], [LM95]). Η λειτουργία της τεχνικής HNM αναφέρθηκε στη παράγραφο 3.3.2.2.

4.2 Βάση διφώνων

Για τη λειτουργία του συστήματος TTS για την ελληνική γλώσσα κατασκευάστηκε μια βάση από προφορές 5000 διασυλλαβών περίπου. Οι διασυλλαβές ηχογραφήθηκαν με ρυθμό δειγματοληψίας 16000 sampes/sec. Το επίπεδο θορύβου δεν ήταν ιδιαίτερα χαμηλό, της τάξης των 35 db, λόγω του θορύβου απ' το σκληρό δίσκο και το τροφοδοτικό του υπολογιστή. Η διαδικασία ηχογράφησης κράτησε περίπου 10 μέρες κυρίως λόγο έλλειψης προηγούμενης εμπειρίας.

4.2.1 Σχεδίαση βάσης διφώνων

Ώς διασυλλαβή ορίσαμε το τμήμα φωνής από τη μέση ενός φωνήεν ώς τη μέση του επόμενου. Θα συμβολίζουμε ένα φωνήεν με /V/ (Vowel), ένα σύμφωνο με /C/ (Consonant), ένα ή περισσότερα σύμφωνα με /Cs/ (Consonants) και το τέλος ή την αρχή μιας λέξης με #. Έτσι οι διασυλλαβές είναι της μορφής /V_1CsV_2/ ή /V_1V_2/ όπου τα φωνήεντα θεωρούμε ότι ξεκινάνε και σταματούν στο μέσον τους. Μαζί με τις διασυλλαβές προσθέτουμε και τα τμήματα φωνής που συμπληρώνουν την αρχή και το τέλος μιας λέξης και είναι της μορφής /#CsV/ ή /#V/ και /VCs#/ ή /V#/. Για κάθε διασυλλαβή λαμβάνουμε υπόψη μας και τον τονισμό, δηλαδή για τη διασυλλαβή /V_1CsV_2/ ηχογραφούμε ξεχωριστά τρεις διασυλλαβές όπου στη μια τονίζεται το /V_1/ στο άλλο το /V_2/ και στο τρίτο ούτε το /V_1/ ούτε το /V_2/. Για παράδειγμα οι διασυλλαβές που συνθέτουν τη λέξη «*καλημέρα*» είναι : /#*κα*/, /*αλη*/, /*ημέ*/, /*έρα*/, /*α*#/.

Αν θεωρήσουμε και το # σαν φωνήεν τότε από ένα φωνήεν μέχρι το επόμενο μεσολαβεί πάντα μια διασυλλαβή. Συμβολίζοντας με /(Cs)/ την μεσολάβηση ή όχι συμφώνων έχουμε δύο ειδών διασυλλαβές, τα /$V_1(Cs)V_2$/ και τα /#$(Cs)V$/ ή /$V(Cs)$#/. Στόχος κατά την σχεδίαση της βάσης είναι οι διασυλλαβές να προφέρονται σε συνεχή λόγο και στη συνέχεια να απομονώνονται παρά να προφέρονται μόνες τους. Με το σκεπτικό αυτό για κάθε διασυλλαβή προφέρουμε

μια λέξη με συγκεκριμένη μορφή. Για τη διασυλλαβή /$V_1(Cs)V_2$/ προφέρουμε τη λέξη «$V_{11}\tau V_1(Cs)V_2\tau V_{22}$», για η διασυλλαβή /#$(Cs)V_1$/ τη λέξη «#$(Cs)V_1\tau V_{11}$» και για τη διασυλλαβή /$V_1(Cs)$#/ τη λέξη «$V_{11}\tau V_1(Cs)$#». Το φωνήεν /V_{ii}/ είναι ίδιο με το /V_i/ $i = 1, 2$ αλλά μπορεί να διαφέρει στο τονισμό. Έτσι για τη διασυλλαβή /*αντρέ*/ προφέρουμε τη λέξη «*αταντρέτε*». Η μορφή αυτή των λέξεων έχει πολλά πλεονεκτήματα :

1) Τα φωνήεντα /V_1/ και /V_2/ των διασυλλαβών της μορφής /$V_1(Cs)V_2$/ δεν βρίσκονται στην αρχή ή στο τέλος της λέξης οπότε προφέρονται πιο ουδέτερα.

2) Μπορούμε να ελέγξουμε τον τονισμό. Όταν κάποιο από τα /V_1/, /V_2/ είναι τονισμένα τότε τα /V_{11}/, /V_{22}/ είναι άτονα. Στη περίπτωση που και τα δύο είναι άτονα τότε η λέξη που προφέρουμε τονίζεται στη λήγουσα δηλαδή στο /V_{22}/. Αυτό βοηθάει στην σωστότερη και ευκολότερη προφορά των λέξεων απ' τον εκφωνητή.

3) Το σύμφωνο /*τ*/ είναι στιγμιαίο και παρουσιάζει πολύ μικρή συμπροφορά με τα /V_1/, /V_2/ γι' αυτό και η επίδραση του στη φασματική περιβάλλουσά τους είναι πολύ μικρή. Αυτό διευκολύνει τη συρραφή των διασυλλαβών μια και η φασματική περιβάλλουσα ίδιων φωνηέντων[1] πλησιάζει περισσότερο αυτή των ιδανικών φθόγγων άρα μοιάζουν και περισσότερο μεταξύ τους.

4) Με σκοπό τη φασματική ομοιομορφία ίδιων φωνηέντων διαφορετικών διασυλλαβών έγινε και η επιλογή τα /V_{11}/, /V_{22}/ να είναι τα ίδια με τα /V_1/, /V_2/ αντίστοιχα έτσι ώστε οι κινήσεις των στοιχείων άρθρωσης που απαιτούνται να είναι μικρότερες. Γι' αυτό και τα στοιχεία άρθρωσης, κατά τη προφορά των φωνηέντων, είναι ποιό κοντά στις ιδανικές θέσεις.

4.2.2 Κατασκευή βάσης διφώνων

Θα εξηγήσουμε τώρα τη διαδικασία κατασκευής της βάσης διασυλλαβών η οποία γίνεται μια φορά στην αρχή αλλά και κάθε φορά που θέλουμε να προσθέσουμε νέες διασυλλαβές στη βάση. Για να συλλέξουμε τις διασυλλαβές που έπρεπε να προφέρουμε βρήκαμε όλες τις διασυλλαβές από λέξεις ενός ελληνικού λεξικού, συγκεκριμένα του προγράμματος *ispellH*. Το λεξικό περιείχε περίπου 4500 διασυλλαβές. Θεωρητικά οι δυνατές διασυλλαβές είναι σχεδόν πενταπλάσιες. Οι συνδυασμοί συμφώνων που βρήκαμε ήταν 242 ενώ τα δυνατά φωνήεντα 11 (/*α*/, /*ε*/, /*ι*/, /*ο*/, /*ου*/, /*ά*/, /*έ*/, /*ί*/, /*ό*/, /*ού*/, #). Δεν συμπεριλαμβάνονται οι συνδυασμοί όπου και τα δύο φωνήεντα είναι τονισμένα ή είναι το #. Έτσι ο θεωρητικός συνολικός αριθμός διασυλλαβών ανέρχεται σε $242 \cdot (11 \cdot 11 - 5 \cdot 5 - 1) =$

[1] προφανώς συρράπτουμε μόνο ίδια φωνήεντα, δηλαδή οι γειτονικές διασυλλαβές είναι πάντα της μορφής /$V_1(Cs_1)V_2$/, /$V_2(Cs_2)V_3$/. Το /V_2/ έχει διαφορετική φασματική περιβάλλουσα στις δύο διασυλλαβές γιατί προφέρεται μαζί με διαφορετικούς φθόγγους /Cs_1/, /Cs_2/.

22990 διασυλλαβές. Για τη κατασκευή της βάσης ηχογραφήθηκαν μόνο οι διασυλλαβές που συλλέξαμε από τις λέξεις του λεξικού οι οποίες αποδείχθηκαν αρκετές για τη σύνθεση σχεδόν όλων των λέξεων που συναντήσαμε σε άλλα κείμενα.

Μέχρι στιγμής έχουμε συλλέξει με αυτόματο τρόπο τις διασυλλαβές που θα περιέχει η βάση και έχουμε επίσης παράγει και τις λέξεις που πρέπει να προφέρουμε. Κατά την ηχογράφηση προφέραμε τις λέξεις σε προτάσεις των 50 (κυρίως για να μήν φτιάχνουμε μεγάλα αρχεία ήχου) και με χρονική απόσταση μεταξύ τους $1sec$. Οι παύσεις ανάμεσα στις λέξεις επιτρέπουν την εύκολη απομόνωση τους. Οι λέξεις προφέρονται όσο το δυνατόν πιο «ουδέτερα» σαν λέξεις στη μέση μιας αδιάφορης οριστικής πρότασης. Είναι σημαντικό η προσωδία να είναι παρόμοια σ' όλες τις λέξεις της βάσης. Δηλαδή ο ρυθμός ομιλίας να είναι σταθερός καθώς και η τονικότητα των έμφωνων φθόγγων. Επειδή η διάρκεια ηχογράφησης όλης της βάσης ήταν περίπου 10 μέρες (η οποία εξαρτάται αρκετά από την εμπειρία του εκφωνητή) είναι δύσκολο να έχουμε ίδια προσωδία σε διαφορετικές ηχογραφήσεις. Ένας τρόπος να το πετύχουμε αυτό είναι πριν αρχίσουμε την ηχογράφηση καινούριων λέξεων να ακούμε και να προφέρουμε ταυτόχρονα αρκετές φορές τις πρώτες λέξεις που ηχογραφήσαμε για να "συγχρονιζόμαστε" με την προσωδία τους.

4.2.2.1 Απομόνωση διφώνων

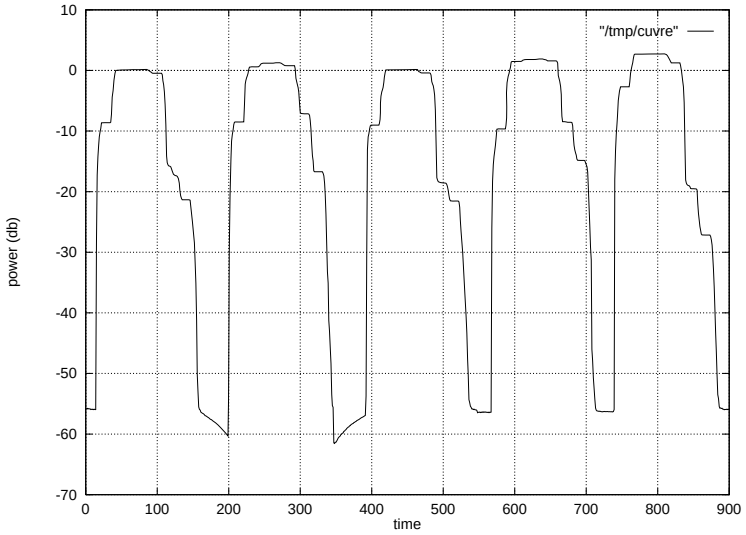
Μετά την ηχογράφηση όλων των προφορών των λέξεων για την κατασκευή της βάσης χρειάζεται να βρεθούν να όρια των διασυλλαβών μέσα στις λέξεις. Για το σκοπό αυτό χρησιμοποιείται και η φωνητική γραφή των λέξεων που για τις συγκεκριμένες λέξεις συμπίπτει με την ορθογραφική. Δε χρειάζεται δηλαδή να αναγνωρίσουμε ποιά φωνήεντα περιέχουν οι λέξεις (*speech recognition*) απλά να τα εντοπίσουμε γνωρίζοντας από πριν ποιά είναι. Αυτό γίνεται σε δύο φάσεις.

1. Απομόνωση λέξεων

Στη πρώτη φάση απομονώνουμε τις λέξεις της κάθε πρότασης ανιχνεύοντας τις παύσεις μεταξύ των λέξεων. Επειδή ξέρουμε από πριν τον αριθμό των λέξεων (τις μετράμε απ' την ορθογραφική γραφή) έχουμε κάποιου είδους επιβεβαίωση για τη σωστή οριοθέτηση των λέξεων. Η ανίχνευση των παύσεων γίνεται υπολογίζοντας την ενέργεια του σήματος σε διαδοχικά αλληλοεπικαλυπτόμενα τετραγωνικά παράθυρα[2] διάρκειας 0.6 sec (μια τυπική κυματομορφή ισχύος φαίνεται στο σχήμα 4.1). Στη συνέχεια υπολογίζουμε την ενέργεια του σήματος

[2]Ο υπολογισμός της ισχύος σε ένα τετραγωνικό παράθυρο γίνεται προσθέτοντας στην ισχύ του προηγούμενου παράθυρου την ισχύ των καινούριων δειγμάτων και αφαιρώντας την ισχύ των παλιών. Αυτό κάνει τον υπολογισμό της ισχύος πολύ γρήγορο.

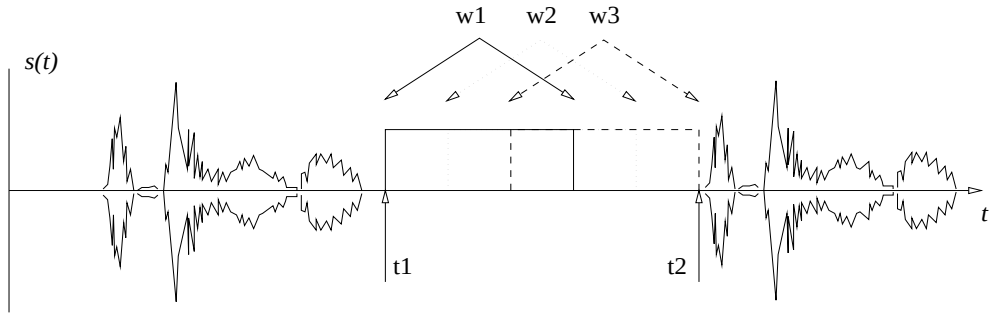
όταν αυτό βρίσκεται εξ' ολοκλήρου μέσα σε μια λέξη έστω E_{word}. θεωρούμε ότι ένα παράθυρο βρίσκεται εξ' ολοκλήρου μεταξύ δύο λέξεων όταν έχει ενέργεια μικρότερη από $E_{word} - 25\ db$. Η ενέργεια αυτή οφείλεται αποκλειστικά στο θόρυβο αφού η παύση μεταξύ των λέξεων έχει χρονική διάρκεια μεγαλύτερη απ' αυτή του παράθυρου. Έτσι η παύση έχει σαν αποτέλεσμα μια ακολουθία από παράθυρα με ενέργεια μικρότερη από $E_{word} - 25\ db$. Τα όρια αυτών των περιοχών ορίζουν το τέλος της προηγούμενης και την αρχή της επόμενης λέξης (σχήμα 4.2).



Σχήμα 4.1: Κυματομορφή ισχύος για παράθυρα $0.6sec$

Στο σχήμα φαίνεται η εξέλιξη της τοπικής ισχύος για μια πρόταση με 5 λέξεις. Η ισχύς σε παράθυρα που είναι εξ' ολοκλήρου σε παύσεις είναι σχεδόν $50db$ λιγότερη από τα παράθυρα που περιέχουν φθόγγους

Δύο σφάλματα θα μπορούσαν να συμβούν σ' αυτή τη φάση : α) μια λέξη να αναγνωριστεί σαν δύο (δηλαδή να σπάσει στη μέση), β) δύο λέξεις να αναγνωριστούν σαν μία (δηλαδή να συνενωθούν). Το πρώτο σφάλμα δε συμβαίνει ποτέ γιατί η διάρκεια 0.6 sec του παράθυρου είναι πολύ μεγαλύτερη από τη μέγιστη απόσταση μεταξύ ηχηρών φθόγγων, οι οποίοι έχουν υπολογίσιμη ισχύ. Έτσι τα παράθυρα μέσα στις λέξεις περιέχουν πάντα κάποιο φθόγγο που αυξάνει την ισχύ τους πάνω απ' το κατώφλι των $E_{word} - 25\ db$. Η δεύτερη περίπτωση σφάλματος συμβαίνει συνήθως όταν στη μέση των παύσεων εμφανίζεται κρουστικός θόρυβος, από κύμα αέρα προς το μικρόφωνο, το οποίο για να αποφευχθεί απαιτεί προσεκτική ηχογράφηση (θέση μικροφώνου, χρήση αντιανεμικού σφουγγαριού). Ο κρουστικός θόρυβος όταν δεν είναι στη μέση τις παύσης έχει σαν αποτέλεσμα την αύξηση της διάρκειας της λέξης στην οποία



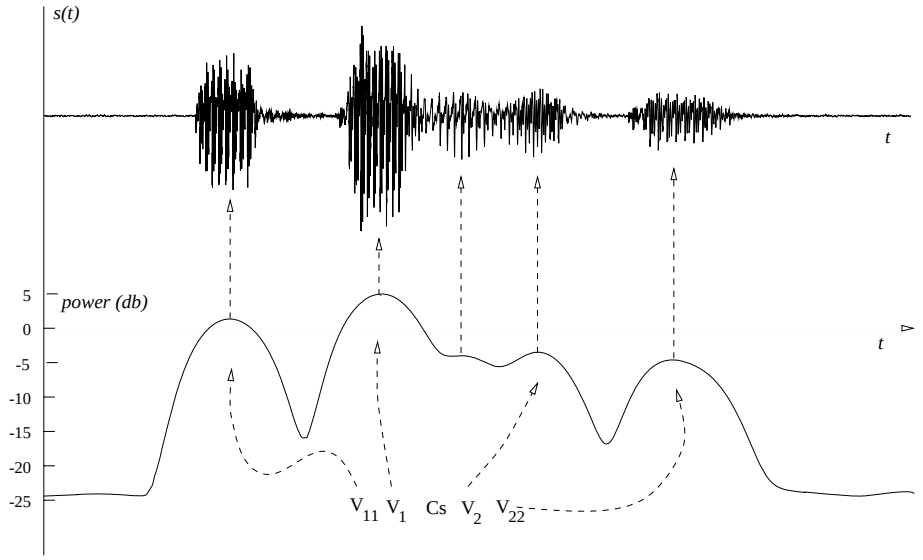
Σχήμα 4.2: Διαδοχικά παράθυρα σε διάστημα παύσης

Τα παράθυρα w_1, w_2, w_3 βρίσκονται εξ' ολοκλήρου σε διάστημα παύσης και έχουν ισχύ μικρότερη από $E_{word} - 25\ db$. Τα όρια της περιοχής που ορίζουν δείχνουν το τέλος της προηγούμενης (t_1) και την αρχή της επόμενης λέξης (t_2).

είναι πιο κοντά μια και στην ανίχνευση θα συμπεριλάβουμε το θόρυβο σαν φθόγγο της λέξης. Παρά την προσεκτική ηχογράφηση τέτοια φαινόμενα ήταν συχνά και διορθώθηκαν με τη βοήθεια ενός γραφικού επεξεργαστή σήματος. Άλλο ένα πρόβλημα προκύπτει με τις λέξεις της μορφής /#CsV/ και /VCs#/. Όταν το /Cs/ είναι χαμηλής ισχύος (πχ /σ/, /χ/, /θ/) ένα κομμάτι του μπορεί να θεωρηθεί παύση και να μη συμπεριληφθεί στη λέξη. Για να σιγουρευτούμε ότι η οριοθέτηση των λέξεων δεν έχει σαν αποτέλεσμα το χάσιμο κάποιου φθόγγου χαμηλής ισχύος, συμπεριλαμβάνουμε επιπλέον χρονικό περιθώριο γύρω της.

2. Οριοθέτηση διασυλλαβών μέσα σε κάθε λέξη

Στη δεύτερη φάση για κάθε απομονωμένη λέξη βρίσκουμε τα όρια των διασυλλαβών για το οποίο τη προφέραμε. Αυτό επιτυγχάνεται βρίσκοντας το μέσο των φωνηέντων, γνωρίζοντας τον αριθμό τους και ποιά είναι από την ορθογραφική γραφή της λέξης. Τα μέσα των φωνηέντων τα βρίσκουμε χρησιμοποιώντας ένα κριτήριο ενέργειας σαν αυτό της προηγούμενης φάσης σε συνδυασμό μ' ένα κριτήριο χρονικής απόστασης των μέσων των φωνηέντων. Κατασκευάζουμε μια καμπύλη ενέργειας του σήματος ώς προς το χρόνο υπολογίζοντας την ενέργεια του σήματος σε διαδοχικά αλληλοεπικαλυπτόμενα παράθυρα *Hamming*. Το μέγεθος του παράθυρου είναι ίσο με τη μέση διάρκεια των φωνηέντων. Σ' αυτή τη χρονική ανάλυση (time resolution) ξεχωρίζουν τα φωνήεντα αλλά δε διακρίνεται η θεμελιώδης περιοδικότητα της φωνής. Αν δούμε (σχήμα 4.3) τη κυματομορφή στο χρόνο της προφοράς μιας λέξης και της αντίστοιχης καμπύλης ενέργειας παρατηρούμε ότι το μέσο των φωνηέντων βρίσκεται πάντα σε τοπικό μέγιστο της καμπύλης ενέργειας. Η καμπύλη ενέργειας όμως συνήθως παρουσιάζει περισσότερα τοπικά μέγιστα από τον αριθμό των φωνηέντων, πρέπει λοιπόν να αντιστοιχίσουμε τα φωνήεντα στα σωστά τοπικά μέγιστα. Ο αλγόριθμος



Σχήμα 4.3: Κυματομορφή ισχύος μιας λέξης για παράθυρα $0.15sec$

Στο πάνω μέρος του σχήματος βλέπουμε στο πεδίο του χρόνου τη λέξη «*ετεβγουτου*» που προφέρθηκε για να εισάγουμε στη βάση τη διασυλλαβή /*εβγου*/. Η αντίστοιχη κυματομορφή ισχύος παρουσιάζει 5 τοπικά μέγιστα, 4 στο μέσο των φωνηέντων ενώ το 5ο προέρχεται από τον έμφωνο φθόγγο /*β*/.

αντιστοίχισης επιλέγει τα μεγαλύτερα τοπικά μέγιστα με την προϋπόθεση ότι διπλανά μέγιστα δεν απέχουν λιγότερο από $0.08sec$. Το αποτέλεσμα της παραπάνω διαδικασίας είναι δείκτες που συνδέουν τα φωνήεντα των λέξεων με το μέσο της προφοράς τους που αποθηκεύονται σ' ένα αρχείο κειμένου. Στο ίδιο αρχείο αποθηκεύουμε και τις χρονικές στιγμές που ξεκινάει και σταματάει η προφορά κάθε λέξης όταν οι διασυλλαβές είναι της μορφής /$\#CsV$/ και /$VCs\#$/ αντίστοιχα.

Πρόβλημα παρατηρείται στην οριοθέτηση διασυλλαβών του τύπου /V_1V_2/ όπου συχνά δεν εμφανίζονται δύο ξεχωριστά τοπικά μέγιστα για το κάθε φωνήεν λόγο της ισχυρής συμπροφοράς. Ένα άλλο πρόβλημα είναι ότι κάποιο φωνήεν παρουσιάζει δύο τοπικά μέγιστα και ο αλγόριθμος αντιστοίχισης επιλέγει το λάθος αν και μερικές φορές και τα δύο μέγιστα είναι σε λάθος θέση.

4.2.2.2 Επαλήθευση απομόνωσης διφώνων

Από τη διαδικασία οριοθέτησης που περιγράψαμε παραπάνω προκύπτουν λάθη σαν αυτά που μόλις αναφέραμε. Για να επαληθεύσουμε τη σωστή οριοθέτηση των διασυλλαβών μπορούμε να χρησιμοποιήσουμε έναν γραφικό επεξεργαστή σήματος, δηλαδή βλέποντας τη κυματομορφή στο χρόνο να επιβεβαιώνουμε

την οριοθέτηση των διασυλλαβών. Η διαδικασία αυτή είναι το ίδιο χρονοβόρα με το να κάναμε τον εντοπισμό όλων των διασυλλαβών με το χέρι (μη αυτόματα). Καταφεύγουμε λοιπόν σε άλλη μέθοδο πιο διαλογική και πιο γρήγορη. Χρησιμοποιούμε το ίδιο το σύστημα TTS για την εκφώνηση της βάσης, της οποίας όπως έχουμε πει έχουμε την ορθογραφική γραφή. Κατά την ακρόαση διασταυρώνουμε τις λέξεις τις βάσης με τις προφορές τους και διορθώνουμε όταν χρειάζεται το αρχείο με τους δείκτες "φωνήεν προς τοπικό μέγιστο". Με τη διαδικασία αυτή μπορούμε να επαληθεύσουμε τη σωστή αντιστοίχιση των φωνηέντων μιας λέξης στα σωστά τοπικά μέγιστα ισχύος. Δε μπορούμε όμως να είμαστε απόλυτα σίγουροι για την ακριβή οριοθέτηση των διασυλλαβών δηλαδή αν ξεκινούν/σταματούν ακριβώς στο μέσο των φωνηέντων.

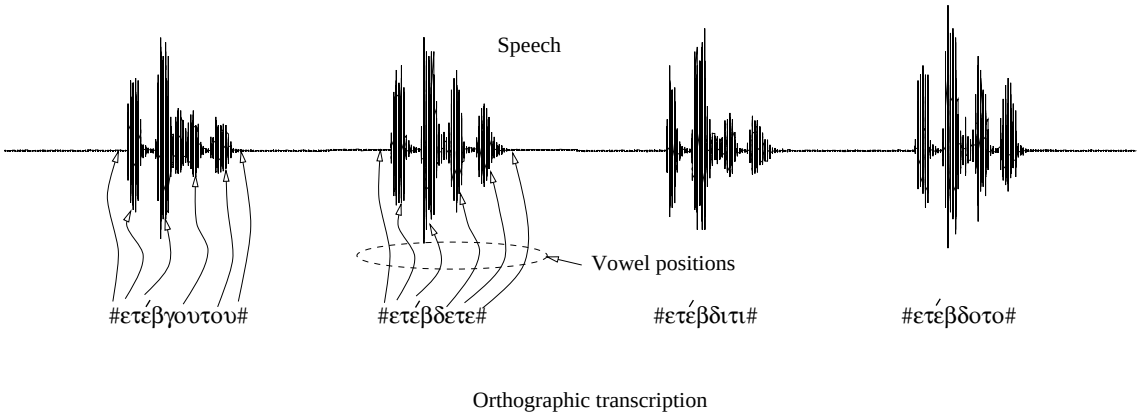
4.2.3 Διαχείριση της βάσης

Μετά τις διαδικασίες που περιγράψαμε παραπάνω η βάση διασυλλαβών είναι αποθηκευμένη σε τρία αρχεία, ένα αρχείο κειμένου που περιέχει την ορθογραφική γραφή των λέξεων, άρα και των διασυλλαβών, ένα αρχείο ήχου (το οποίο έχουμε χωρίσει σε πολλά μικρότερα λόγω μεγέθους) που περιέχει τις προφορές των λέξεων, και ένα αρχείο κειμένου που περιέχει της θέσεις των μέσων των φωνηέντων, άρα τα όρια των διασυλλαβών. Οι προτάσεις που απαρτίζουν τη βάση είναι περίπου 2:30 ώρες ομιλίας (τα αρχεία φωνής με συχνότητα δειγματοληψίας 16000$samples/sec$ είναι 160$Mbytes$). Η συνολική διάρκεια των διασυλλαβών αντίθετα είναι μόλις 16 λεπτά (20$Mbytes$). Για λόγους ταχύτητας και χωρητικότητας λοιπόν αποθηκεύουμε σε ξεχωριστό αρχείο οι απομονωμένες διασυλλαβές. Το αρχείο αυτό πρέπει να δημιουργείται ξανά όταν αλλάζουμε την οριοθέτηση κάποιας διασυλλαβής ή όταν ηχογραφούμε ξανά κάποια διασυλλαβή (σχήμα 4.4).

Η διαχείριση της βάσης όταν την έχουμε φορτώσει στη μνήμη γίνεται μέσω μιας συνάρτησης κατακερματισμού. Αναγνωριστικό κλειδί για κάθε διασυλλαβή είναι η φωνητική γραφή της.

4.3 Διαδικασία σύνθεσης

Για την εκφώνηση του κειμένου εισόδου εκτελούνται διαδοχικές λειτουργίες. Καταρχήν αναγνωρίζουμε τις διασυλλαβές που συνθέτουν τις λέξεις του κειμένου βρίσκοντας τους φθόγγους βάση των κανόνων της ελληνικής γραμματικής. Επίσης αναγνωρίζουμε τις λέξεις και τις προτάσεις του κειμένου. Στη συνέχεια ανακαλούμε απ' τη βάση τις προφορές των διασυλλαβών που χρειάζονται και



Σχήμα 4.4: Πληροφορία που αποθηκεύεται στη βάση διασυλλαβών.

Η βάση είναι αποθηκευμένη σε τρία αρχεία, ένα φωνής (speech), ένα με το κείμενο των λέξεων (orthographic transcripion), και ένα που περιέχει τη θέση των φωνηέντων (vowel positions).

τις συρράπτουμε εξομαλύνοντας ασυνέχειες στα σημεία συνένωσης. Τέλος για κάθε πρόταση παράγουμε τη μουσική καμπύλη της ανάλογα με το είδος της (κατάφαση, ερώτηση) και τις *φωνητικές λέξεις* που την αποτελούν. Οι *φωνητικές λέξεις* είναι ομάδες λέξεων που προφέρονται σαν μία (πχ το άρθρο με το ουσιαστικό ή το επίθετο που ακολουθεί).

4.3.1 Γραμματική ανάλυση κειμένου

Στη παράγραφο αυτή θα περιγράψουμε τις δύο λειτουργίες του πρώτου τμήματος του συστήματος TTS. Τη παραγωγή της φωνητικής γραφής για την αναγνώριση των διασυλλαβών που συνθέτουν τις λέξεις και την εξαγωγή της απαραίτητης πληροφορίας για τη παραγωγή της προσωδίας κάθε πρότασης.

4.3.1.1 Μετατροπή ορθογραφικής γραφής σε ακολουθία διφώνων

Για τη σύνθεση της κάθε λέξεις χρειαζόμαστε τα βρούμε από ποιές διασυλλαβές συντίθεται. Το κείμενο εισόδου λοιπόν πρέπει να μετατραπεί στην αντίστοιχη ακολουθία διασυλλαβών. Η μετατροπή αυτή γίνεται σε δύο φάσεις, πρώτα βρίσκουμε τους φθόγγους των λέξεων και στη συνέχεια τους ομαδοποιούμε σε διασυλλαβές.

Στη πρώτη φάση, όπου αντιστοιχίζουμε τα γράμματα σε φθόγγους, χρησιμοποιούμε τους κανόνες της ελληνικής γραμματικής. Στην ελληνική γλώσσα συνήθως «προφέρουμε ότι διαβάζουμε» εκτός από συγκεκριμένες περιπτώσεις που ισχύουν σαφείς κανόνες. Οι κανόνες αυτοί αφορούν :

1) Τα δίψηφα φωνήεντα. Το «*αι*» αντιστοιχεί στο φθόγγο /*ε*/, τα «*ει*», «*οι*», «*υι*» προφέρονται σαν το φθόγγο /*ι*/ και το «*ου*» προφέρεται σαν το φθόγγο /*ου*/.

2) Τα διπλά σύμφωνα «*ξ*» και «*ψ*» προφέρονται σαν /*κσ*/ και /*πσ*/ αντίστοιχα.

3) Τα όμοια σύμφωνα «*ββ*», «*κκ*», «*λλ*», «*μμ*», «*νν*», «*ππ*», «*ρρ*», «*σσ*», «*ττ*» προφέρονται σαν ένας φθόγγος /*β*/, /*κ*/, /*λ*/ κλπ αντίστοιχα[3]. Εξαίρεση είναι το «*γγ*» που προφέρεται σαν το δίψηφο σύμφωνο /*γκ*/.

4) Οι συνδυασμοί «*αυ*», «*ευ*» προφέρονται σαν /*αβ*/, /*εβ*/ αντίστοιχα όταν ακολουθεί φωνήεν ή ηχηρό σύμφωνο (πίνακας 2.1) αλλιώς προφέρονται σαν /*αφ*/, /*εφ*/, πχ «*σκεύος*», «*αυγό*» και «*αυτός*», «*ευτυχία*». Όταν ακολουθεί «*β*» ή «*φ*» προφέρουμε ένα /*β*/, /*φ*/ αντίστοιχα.

Δε χρειάζεται να διαχωρίσουμε τις αλλοφωνικές αλλοιώσεις του κάθε φθόγγου. Πχ το /*κ*/ προφέρεται διαφορετικά πριν από /*α*/ (υπερωικό) και διαφορετικά πριν από /*ε*/ (ουρανικό) αλλά κατά την ηχογράφηση των διασυλλαβών θα προφερθούν μαζί οπότε θα προφερθεί το σωστό αλλόφωνο, άρα δε χρειάζεται να αναγνωρίσουμε πιο αλλόφωνο είναι. Με το ίδιο σκεπτικό δεν αναγνωρίζουμε και πάθη συμφώνων πχ το «*γ*» πριν από «*χ*» προφέρεται πάντα σαν /*ν*/, αλλά αντικατάσταση του «*γ*» από «*ν*» δε γίνεται γιατί το «*γχ*» προφέρεται και ηχογραφείται όλο μαζί σαν /*νχ*/.

Πρόβλημα προκύπτει με το φαινόμενο της συνίζησης για το οποίο δεν υπάρχει συγκεκριμένος κανόνας. Όταν μετά από «*ι*» («*η*», «*υ*», «*ει*», «*οι*») ή «*ε*» («*αι*») ακολουθεί άλλο φωνήεν αυτά συχνά προφέρονται σαν ένα. Έτσι λέμε «*δι-ά-ρκει-α*» αλλά και «*α-ρα-δειά-ζω*» , «*συγ-χα-ρη-τή-ρι-α*» αλλά και «*χα-τή-ρια*».

Στη φάση αυτή επίσης απαλείφουμε τις αποστρόφους (το «*εξ' αρχής*» γίνεται «*εξαρχής*») και συνενώνουμε λέξεις που δε χωράνε ολόκληρες στο τέλος μιας γραμμής.

Στη δεύτερη φάση αναγνωρίζουμε τις διασυλλαβές που σχηματίζουν κάθε λέξη. Όπως έχουμε πει οι διασυλλαβές είναι της μορφής /$V_1(CS)V_2$/ και /#$(CS)V$/ ή /$V(CS)$#/. Για να τα βρούμε μέσα στις λέξεις αρκεί να κινούμαστε από ένα φωνήεν ή την αρχή μιας λέξης στο επόμενο φωνήεν ή όταν δεν υπάρχει στο τέλος της λέξης. Πχ η λέξη «*υπολογιστής*» συντίθεται απ' τις διασυλλαβές /#*ι*/, /*ιπο*/, /*όλο*/, /*ογι*/, /*ιστί*/, /*ίσ*#/. Θεωρούμε ότι οι χαρακτήρες που περιστοιχίζουν μια λέξη είναι το κενό (*space*), ο χαρακτήρας *tab*, η τελεία (.), το κόμμα (,) , το ερωτηματικό (; και ?), το θαυμαστικό (!), και η άνω τελεία (·). Οι λατινικοί χαρακτήρες αγνοούνται όπως και οι αριθμοί.

[3]Σε ορισμένες λέξεις τα διπλά σύμφωνα προφέρονται πιο έντονα πχ «παλλαϊκός», «υπερρεαλιστικός». Στο σύστημα TTS που κατασκευάστηκε τα διπλά σύμφωνα προφέρονται πάντα σαν τα απλά.

4.3.1.2 Ανάλυση κειμένου για τη παραγωγή προσωδίας

Στη παράγραφο 2.6.1 θα αναφέρουμε με ποιό τρόπο παράγουμε προσωδία για το κείμενο εισόδου. Η λειτουργία αυτή απαιτεί την αποθήκευση του κειμένου εισόδου σε διαφορετικά επίπεδα αφαίρεσης.

Όπως αναφέραμε στη παράγραφο 3.7.2 τα συστήματα TTS παράγουν προσωδία ανεξάρτητα για κάθε φωνητική φράση όπως αυτή ορίζεται στο συγκεκριμένο σύστημα TTS. Στο σύστημα TTS για τα ελληνικά σαν φωνητική πρόταση ορίζεται η πρόταση. Για την αναγνώριση προτάσεων πρέπει να βρούμε τα σημεία στίξης με τα οποία τερματίζονται προτάσεις που είναι η τελεία, το ερωτηματικό και το θαυμαστικό. Υποθέτουμε ότι αυτά τα σημεία στίξης σημαίνουν μόνο τέλος πρότασης και δε χρησιμοποιούνται για άλλο σκοπό, πχ για συντομογραφίες. Τα κόμματα αναγνωρίζονται και χρησιμεύουν για τον καθορισμό των παύσεων μεταξύ των λέξεων.

Στο επίπεδο της πρότασης αποθηκεύουμε το είδος της (οριστική, ερωτηματική) και των αριθμό των λέξεων που την αποτελούν. Στο επίπεδο της λέξης αποθηκεύουμε τον αριθμό συλλαβών της λέξης, τον τονισμό της (οξύτονη, παροξύτονη, προπαροξύτονη, διπλά τονισμένη[4]) και τη θέση της μέσα στη πρόταση.

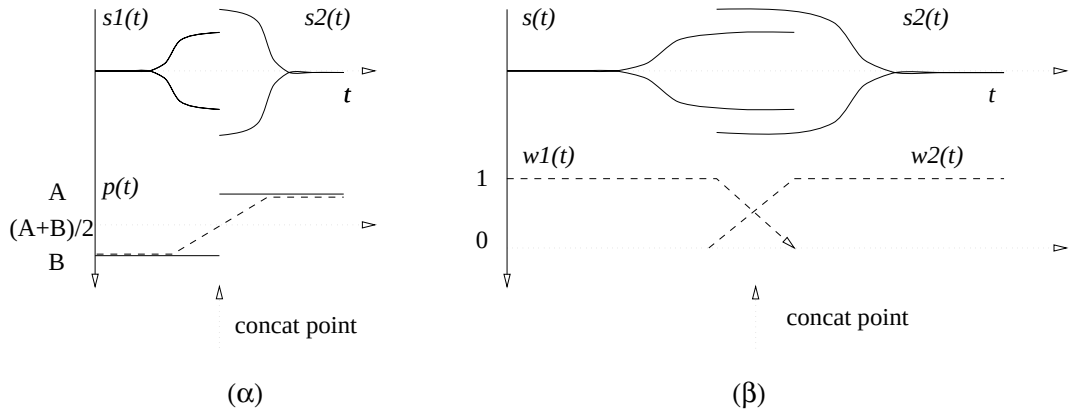
Για κάθε λέξη ελέγχουμε αν είναι άρθρο, σύνδεσμος, αντωνυμία, επίρρημα και μερικές φορές αν είναι όνομα συγκεκριμένου γένους και πτώσης από τη κατάληξη του ονόματος. Οι πληροφορίες αυτές χρησιμοποιούνται για το σχηματισμό φωνητικών λέξεων.

4.3.2 Τεχνικές συρραφής φωνημάτων

Όπως αναφέραμε και στη παράγραφο 3.6 στα συστήματα TTS με συρραφή απαιτείται κάποιου είδους επεξεργασία σήματος για την εξομάλυνση ασυνεχειών στα σημεία συνενώσεις των φωνητικών μονάδων. Η χρήση διφώνων κάνει ευκολότερη τη συρραφή γιατί οι φασματικές ασυνέχειες, που διορθώνονται πιο δύσκολα, είναι μικρότερες σε σύγκριση με άλλες φωνητικές μονάδες. Η χρήση διασυλλαβών έχει σαν αποτέλεσμα ακόμα μικρότερες ασυνέχειες αφού η συρραφή συμβαίνει μόνο σε φωνήεντα (προφανώς ίδια). Τα φωνήεντα έχουν κατά μέσο όρο μεγαλύτερη διάρκεια γι' αυτό είναι πιο σταθερά φασματικά.

Η πρώτη προσπάθεια συρραφής ήταν χωρίς να γίνεται καμιά επεξεργασία. Το αποτέλεσμα αυτής της πολύ γρήγορης μεθόδου είναι κατανοητή ομιλία αλλά είναι εμφανείς οι στιγμές της μετάβασης από τη μια διασυλλαβή στην

[4] Διπλά τονισμένες είναι οι προπαροξύτονες που ακολουθούνται από κτητικό, πχ «τα αποτελέσματά μας»



Σχήμα 4.5: Σχηματισμός φωνητικών λέξεων

Στο (α) οι διασυλλαβές ενώνονται χωρίς επικάλυψη απλά εφαρμόζοντας μια γραμμική περιβάλλουσα στην ισχύ. Αν A, B είναι η τοπική ισχύς του $s_1(t)$, $s_2(t)$ αντίστοιχα στο σημείο συρραφής (concat point) η ισχύς και των δύο θα είναι $(\mathrm{A}+\mathrm{B})/2$. Στο σχήμα (β) οι διασυλλαβές $s_1(t)$, $s_2(t)$ επικαλύπτονται και προστίθενται αφού σταθμιστούν με τις περιβάλλουσες $w_1(t)$ (fade out), $w_2(t)$ (fade in) αντίστοιχα.

επόμενη λόγω διαφορετικής τοπικής ισχύος. Η λειτουργία του συστήματος χωρίς επεξεργασία κατά τη συρραφή είναι πολύ χρήσιμη κατά την επαλήθευση οριοθέτησης των διασυλλαβών γιατί είναι πιο εμφανή ακουστικά τα όριά τους.

Για να ξεπεραστεί το πρόβλημα της ασυνέχειας ισχύος στα σημεία συνένωσης εφαρμόστηκαν δύο μέθοδοι : α) οι διασυλλαβές ενώνονται χωρίς επικάλυψη και πολλαπλασιάζονται με μια γραμμική περιβάλλουσα (σχήμα 4.5α), β) οι δύο διασυλλαβές ενώνονται με κάποια επικάλυψη και στην διασυλλαβή που τελειώνει εφαρμόζεται μια περιβάλλουσα εξασθένισης (fade out) ενώ στην διασυλλαβή που αρχίζει μια περιβάλλουσα ενίσχυσης (fade in) (σχήμα 4.5β). Οι μεταβάσεις των διασυλλαβών είναι ακουστές μόνο όταν υπάρχουν *ασυνέχειες τονικότητας* (*pitch discontinuties*).

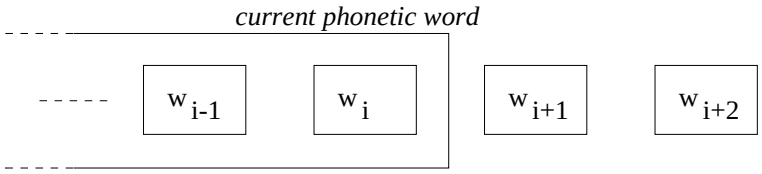
4.3.3 Έλεγχος προσωδίας

Με τον όρο προσωδία, όπως έχουμε αναφέρει, εννοούμε το ρυθμό, τη τονικότητα και την ένταση της ομιλίας. Θα παρουσιάσουμε τώρα τον έλεγχο που κάνουμε στις παραμέτρους αυτές.

4.3.3.1 Έλεγχος ρυθμού ομιλίας

Σαν έλεγχο ρυθμού εννοούμε τον έλεγχο της διάρκειας των φθόγγων και των παύσεων μεταξύ των λέξεων, δηλαδή το σχηματισμό φωνητικών λέξεων. Φωνη-

τικές λέξεις είναι συστοιχίες λέξεων που προφέρονται σαν μία ,πχ «*το ποτήρι μου*» προφέρεται σαν μία λέξη «*τοποτήριμου*».



Σχήμα 4.6: Σχηματισμός φωνητικών λέξεων

Η λέξη w_{i+1} συγχωνεύεται με τη τρέχουσα φωνητική λέξη (current phonetic word) βάσει κανόνων.

Η συνένωση των λέξεων γίνεται βάση κάποιων κανόνων που ελέγχουν αν μπορούν να ενωθούν γειτονικές λέξεις. Κατατάσσουμε τις λέξεις σε τρεις κατηγορίες.

1) Αυτές που μπορούν να συνενωθούν με την επόμενη. Τέτοιες είναι :
- τα άρθρα σε όλα τα γένη και όλες τις πτώσεις.
- οι αντωνυμίες κυρίως οι αδύνατοι τύποι σε όλα τα γένη και όλες τις πτώσεις εκτός από τις κτητικές και τις ερωτηματικές.
- οι σύνδεσμοι, συμπλεκτικοί («*και, ούτε, μήτε*»), διαζευκτικοί («*ή, είτε*»), αντιθετικοί («*μα, αλλά, παρά, ενώ, μόνο*»), ειδικοί («*ότι, που, πως*»), όλοι οι χρονικοί, αιτιολογικοί («*γιατί, επειδή, αφού*»), αποτελεσματικοί («*ώστε (να), που*»), διστακτικοί («*μήν, μήπως*»), συγκριτικής («*παρά*»).
- τα μόρια «*ας, θα, να, μα, για*».
- οι προθέσεις «*με, σε, για, ως, προς, κατά, μετά, παρά, αντί, από, χωρίς, δίχως, εκ, εξ, εν, επί, προ*» καθώς και για τις αριθμητικές πράξεις «*συν, πλην, επί, δια*».
- τα επιρρήματα όλα εκτός από τα ερωτηματικά.

2) Αυτές που μπορούν να συνενωθούν με την προηγούμενη. Τέτοιες είναι οι αδύναμοι τύποι των κτητικών αντωνυμιών «*μου, σου, του, της, μας, σας, τους*».

3) Λέξεις που δεν ενώνονται ούτε με την επόμενη ούτε με τη προηγούμενη.

Δυστυχώς οι λέξεις της δεύτερης κατηγορίας ανήκουν και στη πρώτη. Οι κτητικές αντωνυμίες «*μου, σου, μας, σας*» είναι και προσωπικές. Οι προσωπικές αντωνυμίες συνήθως ακολουθούνται από ρήμα το οποίο μερικές φορές μπορούμε να αναγνωρίσουμε από τη κατάληξή του. Οι κτητικές αντωνυμίες «*του, της, μας*» είναι και άρθρα τα οποία συνήθως ακολουθούνται από όνομα (ουσιαστικό, επίθετο, αντωνυμία) ίδιου αριθμού και πτώσης. Αυτό μπορούμε να το επαληθεύσουμε πάλι από την κατάληξη του ονόματος.

Ο μηχανισμός οριοθέτησης φωνητικών λέξεων είναι ο εξής : Έστω ότι μια

πρόταση έχει N λέξεις τις w_i $i = 1..N$ και ότι η τρέχουσα φωνητική λέξη περιλαμβάνει και την w_i (σχήμα 4.6). Στη τρέχουσα φωνητική λέξη ενώνεται η επόμενη όταν : α) η τελευταία λέξη w_i της τρέχουσας φωνητικής ανήκει στη πρώτη κατηγορία, δηλαδή ενώνεται με την επόμενη, β) η επόμενη λέξη w_{i+1} ανήκει στη δεύτερη κατηγορία, δηλαδή ενώνεται με την προηγούμενη.

Όταν δε συμβαίνουν τα α) και β) τερματίζουμε τη τρέχουσα φωνητική λέξη στη w_i και αρχίζουμε να σχηματίζουμε την επόμενη που στην αρχή περιέχει μόνο την w_{i+1}. Η πρώτη φωνητική λέξη στην αρχή περιέχει μόνο την w_1. Το κόμμα τερματίζει πάντα τη τρέχουσα φωνητική λέξη. Επίσης δεν επιτρέπουμε δύο διαδοχικές λέξεις της δεύτερης κατηγορίας γιατί θα είχαμε δύο κτητικά στη σειρά το οποίο δεν έχει νόημα. Όταν συμβαίνει αυτό η δεύτερη λέξη ενώνεται με την επόμενη (αφού ανήκει και στη πρώτη κατηγορία), πχ «*τα πράγματά σου μου ήρθαν*».

Η συνένωση δύο λέξεων γίνεται με δύο τρόπους. Έστω ότι η τελευταία διασυλλαβή της πρώτης λέξης είναι /$V_1(Cs_1)$#/ και το πρώτο της δεύτερη λέξης το /#$(Cs_2)V_2$/. Αντικαθιστούμε αυτά τα δύο με το /$V_1(Cs_1)(Cs_2)V_2$/ αν αυτό υπάρχει στη βάση διασυλλαβών οπότε οι δύο λέξεις πλέον προφέρονται σαν μία. Αν δεν υπάρχει η διασυλλαβή /$V_1(Cs_1)(Cs_2)V_2$/ στη βάση τότε συρράπτουμε τα /$V_1(Cs_1)$#/ και /#$(Cs_2)V_2$/ αλλά μειώνουμε τη παύση μεταξύ τους αφαιρώντας εξίσου δείγματα απ' το τέλος του /$V_1(Cs_1)$#/ και την αρχή του /#$(Cs_2)V_2$/. Το πρόβλημα που υπάρχει στη περίπτωση αυτή είναι ότι το φαινόμενο συμπροφοράς δεν αποδίδεται σωστά αλλά τις περισσότερες φορές η συμπροφορά στα συμπλέγματα των φθόγγων που σχηματίζονται είναι μικρή.

Στη διάρκεια των φθόγγων δεν επεμβαίνουμε καθόλου μια και τα αποτελέσματα κατά τη σύνθεση ήταν ικανοποιητικά.

4.3.3.2 Έλεγχος μουσικής καμπύλης

Με τον όρο μουσική καμπύλη ορίσαμε τη χρονική εξέλιξη της θεμελιώδους συχνότητας της φωνής. Η έκφραση μιας πρότασης (αν είναι καταφατική, αρνητική, ερωτηματική, θαυμαστική κλπ) καθορίζεται κυρίως από τη μουσική της καμπύλη. Θα παράγουμε μουσική καμπύλη, για κάθε πρόταση ανεξάρτητα, με μια τεχνική συρραφής προαποθηκευμένων μουσικών καμπύλων φωνητικών λέξεων. Τις πρότυπες μουσικές καμπύλες φωνητικών λέξεων τις υπολογίσαμε από ανάλυση πραγματικής φωνής.

Το πρότυπο της μουσικής καμπύλης που θα ακολουθήσει μια φωνητική λέξη εξαρτάται :
α) από το είδος της πρότασης (οριστική, ερωτηματική) στην οποία ανήκει,
β) από το ρόλο της φωνητικής λέξης μέσα στη πρόταση και

γ) από το είδος της λέξης που ορίζεται από το μέγεθος και τη θέση του τόνου.

Τα είδη των προτάσεων που αναγνωρίζουμε είναι τα εξής :
- οριστικές καταφατικές,
- ερωτηματικές επιβεβαιωτικές,
- ερωτηματικές μη επιβεβαιωτικές.

Επιβεβαιωτικές ερωτήσεις είναι αυτές που μπορούν να απαντηθούν με *ναι* ή *όχι* πχ «*σήμερα είναι δευτέρα;*». Οι μη επιβεβαιωτικές ερωτήσεις περιέχουν πάντα ερωτηματική αντωνυμία ή ερωτηματικό επίρρημα πχ «*τί μέρα είναι σήμερα;*».

Ως προς το ρόλο μια φωνητική λέξη μπορεί να είναι :
- πρώτη λέξη της πρότασης,
- τελευταία λέξη της πρότασης,
- προτελευταία λέξη της πρότασης,
- ερωτηματική λέξη μιας ερωτηματικής πρότασης,
- τίποτα από τα παραπάνω δηλαδή στη μέση κάποιας πρότασης.

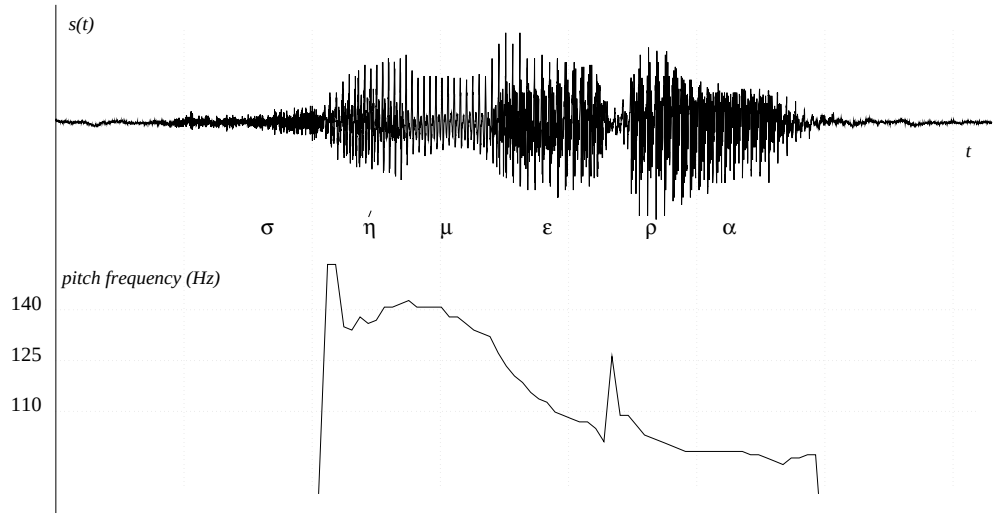
Ως προς το είδος της μια φωνητική λέξη μπορεί να είναι
- μονοσύλλαβη οξύτονη,
- δισύλλαβη οξύτονη ή παροξύτονη,
- τρισύλλαβη οξύτονη ή παροξύτονη ή προπαροξύτονη,
- τετρασύλλαβη παροξύτονη ή προπαροξύτονη.

ενώ όταν δεν είναι κάτι από τα παραπάνω κατατάσσεται στις λέξεις πλησιέστερου μεγέθους με τον ίδιο τονισμό. Έτσι μια πεντασύλλαβη οξύτονη κατατάσσεται στις οξύτονες τρισύλλαβες.

Θα περιγράψουμε τώρα πως δημιουργούμε τις πρότυπες μουσικές καμπύλες φωνητικών λέξεων και στη συνέχεια το μηχανισμό παραγωγής μουσικής καμπύλης για μια πρόταση.

Κατασκευή προτύπων μουσικών καμπύλων για φωνητικές λέξεις

Πρέπει να κατασκευάσουμε πρότυπα φωνητικών καμπύλων για κάθε πιθανό σενάριο φωνητικής λέξης, πχ μια δισύλλαβη οξύτονη φωνητική λέξη που είναι προτελευταία σε μια επιβεβαιωτική ερώτηση. Για το σκοπό αυτό προφέραμε προτάσεις που περιείχαν συγκεκριμένο είδος φωνητικών λέξεων. Επιλέξαμε 8 λέξεις μία από κάθε είδος έστω τις w_i $i = 1..8$. Χρησιμοποιώντας τις λέξεις αυτές σχηματίσαμε 8 προτάσεις των 5 λέξεων εναλλάσοντας κυκλικά τις λέξεις, δηλαδή $w_1w_2w_3w_4w_5$, $w_2w_3w_4w_5w_6$, ... , $w_8w_1w_2w_3w_4$. Έτσι έχουμε όλα τα είδη των λέξεων σε όλες τις δυνατές θέσεις. Η διαδικασία αυτή επαναλαμβάνεται και για τα τρία είδη προτάσεων. Στη συνέχεια εξάγουμε τη μουσική καμπύλη των προτάσεων χρησιμοποιώντας το πρώτο τμήμα της τεχνικής ΗΝΜ και απομονώνουμε τη μουσική καμπύλη των λέξεων. Στο σχήμα 4.7 βλέπουμε τη μουσική



Σχήμα 4.7: Χρονική εξέλιξη τονικότητας.

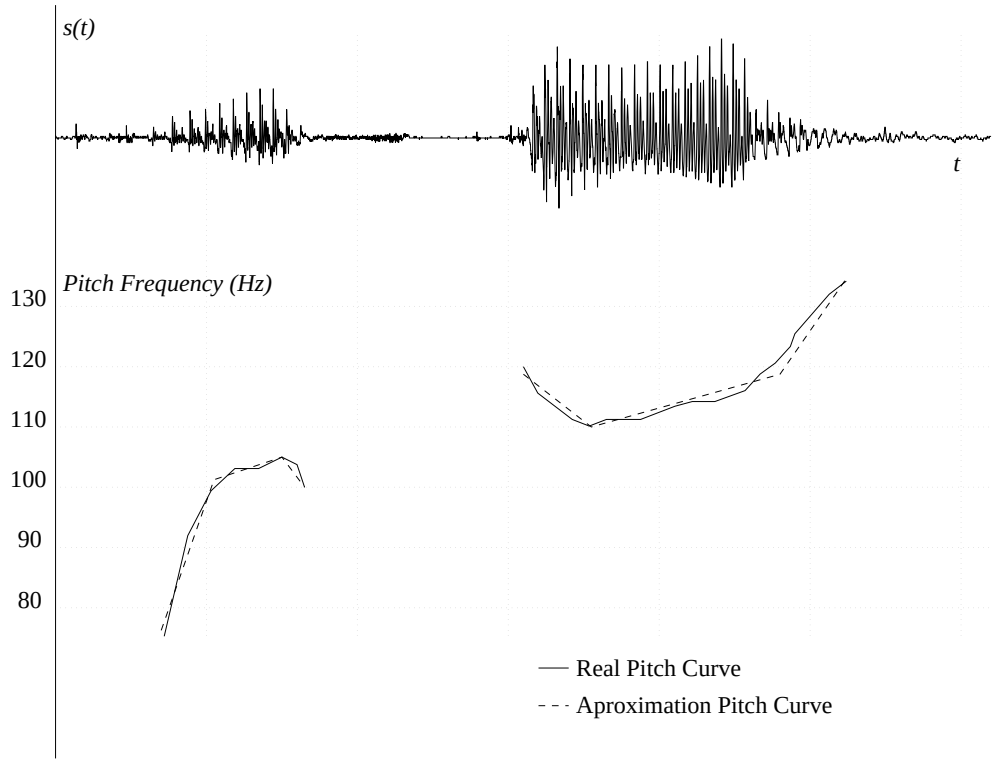
Πάνω φαίνεται η κυματομορφή στο χρόνο για τη λέξη «*σήμερα*» και κάτω τη αντίστοιχη μουσική καμπύλη. Η λέξη είχε το ρόλο πρώτης λέξης σε μια επιβεβαιωτική ερώτηση.

καμπύλη και τη κυματομορφή στο χρόνο της λέξης «*σήμερα*». Από τη μουσική καμπύλη αποθηκεύουμε μόνο τα τμήματα που αντιστοιχούν στα φωνήεντα προσεγγίζοντάς τα με τρεις γραμμικές περιοχές όπως φαίνεται στο σχήμα 4.8. Ορίζουμε, δηλαδή την αρχική και τελική τιμή της θεμελιώδους συχνότητας και δύο ενδιάμεσα σημεία καμπής, στο 20% και στο 80% της διάρκειας του φωνήεντος. Η επιλογή έγινε βάσει παρατήρησης της μουσικής καμπύλης διαφόρων φωνηέντων. Η τονικότητα των έμφωνων συμφώνων προκύπτει με γραμμική παρεμβολή από το τέλος του προηγούμενου φωνήεντος και την αρχή του επόμενου. Με την παρεμβολή προσεγγίζουμε ικανοποιητικά τα πραγματικά δεδομένα.

Μηχανισμός παραγωγής μουσικής καμπύλης

Όπως είπαμε η μουσική καμπύλη μιας πρότασης προκύπτει από συρραφή προτύπων μουσικών καμπύλων φωνητικών λέξεων αντιστοιχίζοντας σε κάθε φωνητική λέξη ένα πρότυπο. Η αντιστοίχιση αυτή εξαρτάται από το είδος της πρότασης, το ρόλο και το είδος της φωνητικής λέξης. Αναγνωρίζουμε αν μια πρόταση είναι οριστική ή ερωτηματική απ' το σημείο στίξης με το οποίο τερματίζεται. Στη περίπτωση που η πρόταση είναι ερωτηματική πρέπει να δούμε αν είναι επιβεβαιωτική ή όχι ανάλογα με το αν περιέχει ερωτηματική αντωνυμία ή ερωτηματικό επίρρημα. Αν δεν υπάρχει τέτοια λέξη τότε η πρόταση είναι επιβεβαιωτική.

Στις μη επιβεβαιωτικές προτάσεις η ερωτηματική αντωνυμία ή το ερωτηματικό επίρρημα παίζει το ρόλο της *ερωτηματικής λέξης*. Τη φωνητική λέξη στην οποία συμμετέχει την ονομάζουμε *ερωτηματική φωνητική λέξη*. Στις επιβεβαι-



Σχήμα 4.8: Προσέγγιση προτύπων μουσικών καμπύλων.

Πάνω φαίνεται η κυματομορφή στο χρόνο για τη λέξη «*αυτό*» ενώ κάτω με συνεχή γραμμή η πραγματική μουσική καμπύλη και με διακεκομμένη η προσέγγιση της με τρεις γραμμικές περιοχές. Η λέξη είχε το ρόλο τελευταίας λέξης σε μια μη επιβεβαιωτική ερώτηση.

ωτικές προτάσεις συνήθως υπάρχει ασάφεια ως προς το νόημα της ερώτησης απ' το οποίο καθορίζεται η προσωδία της. Πχ στην ερώτηση «*Αύριο έρχεται ο Δημήτρης;*» αν ρωτάμε για το χρόνο άφιξης τονίζουμε τη λέξη «*αύριο*» αν μας ενδιαφέρει το πρόσωπο της άφιξης τονίζουμε τη λέξη «*Δημήτρης*». Τη λέξη που τονίζουμε την ονομάζουμε ερωτηματική λέξη και τη φωνητική λέξη στην οποία συμετέχει ερωτηματική φωνητική λέξη. Στη παράγραφο 2.4 είχαμε πει ότι το πως θα προφέρουμε μια πρόταση εξαρτάται και απ' το νόημα της. Η ασάφεια που δείξαμε με το παράδειγμα στις επιβεβαιωτικές προτάσεις οφείλεται στο ότι αγνοούμε το νόημα των προτάσεων. Το σύστημα μας θεωρεί αυθαίρετα ως ερωτηματική φωνητική λέξη μιας επιβεβαιωτικής πρότασης τη πρώτη φωνητική λέξη της πρότασης. Μία εναλλακτική λύση θα ήταν ο χρήστης να σημειώνει με ειδικούς δεσμευμένους χαρακτήρες την ερωτηματική λέξη στις επιβεβαιωτικές ερωτήσεις.

Για να καθορίσουμε το είδος μιας λέξης πρέπει να βρούμε το μέγεθός της και τον τονισμό της. Το μέγεθος μιας λέξης ορίζεται σαν ο αριθμός των φωνηέντων

(φθόγγων) που περιέχει. Η κατάταξη μια λέξης λοιπόν ως προς το είδος της είναι εύκολη.

Τέλος πρέπει να καθοριστεί και ο ρόλος της κάθε λέξης γιατί μια λέξη μπορεί να έχει περισσότερους από έναν ρόλους. Αυτό συμβαίνει συχνά στις μικρές προτάσεις, πχ «*πότε θα έρθεις*» ή με φωνητικές λέξεις «*πότε θαέρθεις*». Η λέξη «*πότε*» είναι ερωτηματική αλλά και πρώτη λέξη της πρότασης και προτελευταία. Η ασάφεια λύνεται θέτοντας προτεραιότητες στους ρόλους. Μία λέξη είναι κατά προτεραιότητα ερωτηματική, πρώτη λέξη της πρότασης και τέλος τελευταία ή προτελευταία. Με αυτές τις προτεραιότητες η λέξη «*πότε*» είναι ερωτηματική.

4.3.3.3 Έλεγχος έντασης

Στη παράγραφο 3.7.2 είπαμε ότι η ένταση ακολουθεί τις μεταβολές της τονικότητας λόγο κατασκευής του μηχανισμού φώνησης. Επειδή στη βάση διασυλλαβών διακρίνουμε τονισμένα από άτονα φωνήεντα η πληροφορία για την ένταση είναι ηχογραφημένη. Έτσι δεν επεμβαίνουμε καθόλου στην ένταση εκτός από την κανονικοποίηση που κάνουμε κατά τη συρραφή των διασυλλαβών.

Κεφάλαιο 5

Επίλογος

Το σύστημα TTS για την ελληνική γλώσσα που κατασκευάσαμε ανήκει στη κατηγορία των συστημάτων TTS που βασίζονται στη συρραφή φωνητικών μονάδων. Στο σύστημα μας οι φωνητικές μονάδες ήταν διασυλλαβές, τμήματα φωνής που ξεκινούν από το μέσο ενός φωνήεντος και σταματούν στο μέσο του επόμενου. Οι διασυλλαβές σαν φωνητικές μονάδες έχουν όλα τα πλεονεκτήματα των διφώνων δηλαδή εύκολη συρραφή και εκ των προτέρων ηχογράφηση της συμπροφοράς. Ειδικότερα οι διασυλλαβές παρουσιάζουν ακόμα λιγότερες ασυνέχειες στα σημεία συρραφής ενώ παράλληλα ο αριθμός των απαραίτητων συρραφών είναι μικρότερος. Μοναδικό μειονέκτημα έναντι των διφώνων είναι ο μεγαλύτερος αριθμός τους ο οποίος όμως δεν είναι απαγορευτικός για τη γρήγορη κατασκευή της βάσης διασυλλαβών.

Στη συνέχεια θα αναφέρουμε συμπεράσματα και παρατηρήσεις από την ακρόαση του συστήματος TTS για την ελληνική γλώσσα και τέλος θα προτείνουμε βελτιώσεις και επεκτάσεις.

5.1 Συμπεράσματα

Η συνθετική ομιλία, όταν δεν γίνεται καμιά επεξεργασία κατά τη συρραφή και δεν ελέγχεται η προσωδία των προτάσεων, είναι εύκολα καταληπτή η οποία όμως δεν έχει καμιά έκφραση, όπως αν εκφωνούσαμε ένα κείμενο αγνοώντας τα σημεία στίξης. Αυτό οφείλεται στη σωστή απόδοση του φαινομένου της συμπροφοράς, των αλλοφωνικών αλλοιώσεων των φθόγγων και κυρίως στην ύπαρξη τονισμού αφού διαχωρίζουμε τονισμένα από άτονα φωνήεντα. Ωστόσο, είναι αντιληπτές οι ασυνέχειες ισχύος στα σημεία συνένωσης οι οποίες μειώνονται σημαντικά κάνοντας εξομάλυνση ισχύος (παράγραφος 3.6). Οι ασυνέχειες τονικότητας δεν αντιμετωπίζονται και γίνονται αντιληπτές όταν είναι μεγάλες.

Έντονα αντιληπτές είναι οι παύσεις μεταξύ λέξεων που προφέρονται μαζί. Η παρατήρηση αυτή μας οδήγησε στο σχηματισμό των φωνητικών λέξεων, δηλαδή προφέροντας τις σαν μια λέξη όπως το άρθρο με το ουσιαστικό. Η βελτίωση που παρατηρείται είναι φυσικότητα στο ρυθμό ομιλίας. Ο ρυθμός της ομιλίας όμως επηρεάζεται από τη διάρκεια των φθόγγων η οποία εξαρτάται από τη σωστή οριοθέτηση των διασυλλαβών κατά την κατασκευή της βάσης. Γενικά όταν η οριοθέτηση δεν είναι εσφαλμένη ο ρυθμός της ομιλίας είναι ομαλός.

Για την έκφραση οριστικών και ερωτηματικών προτάσεων ελέγχεται η τονικότητα της συνθετικής φωνής. Η προσέγγιση της μουσικής καμπύλης με γραμμικές περιοχές είναι αρκετή για να αποδοθεί έκφραση σε μια πρόταση (δηλ. να γίνει ερώτηση ή κατάφαση κλπ). Η φυσικότητα της συνθετικής ομιλίας είναι ικανοποιητική για απλές προτάσεις αλλά φτωχή για πολύπλοκες, πχ όταν υπάρχουν δευτερεύουσες προτάσεις. Αυτό οφείλεται κυρίως στη μικρή ποσότητα πληροφορίας που εξάγουμε από το γραπτό κείμενο για να παράγουμε προσωδία.

Ο συστηματικός τρόπος κατασκευής της βάσης δίνει τη δυνατότητα σε κάποιον να κατασκευάσει μία καινούρια βάση διασυλλαβών για την ελληνική γλώσσα ή για ξένη όπως την ιταλική ή την ισπανική. Η διαδικασία που ακολουθείται είναι : συλλογή των διασυλλαβών που απαιτούνται για τη γλώσσα από κάποιο ορθογραφικό λεξικό σε ηλεκτρονική μορφή, εκφώνηση και ηχογράφηση των λέξεων που περιέχουν τις διασυλλαβές, απομόνωση των διασυλλαβών και επαλήθευση της σωστής οριοθέτησής τους. Η ποιότητα της βάσης εξαρτάται αρκετά από την ουδέτερη εκφώνηση των λέξεων και τη σταθερότητα της τονικότητας και και του ρυθμού των λέξεων κατά τη διάρκεια ηχογράφησης. Τα στοιχεία αυτά συντελούν στη μείωση των ασυνεχειών κατά τη συρραφή και τη σωστή απόδοση της προσωδίας.

5.2 Βελτιώσεις, Επεκτάσεις

Το σύστημα TTS που κατασκευάστηκε είναι μια πρώτη προσπάθεια προσέγγισης του προβλήματος για την ελληνική γλώσσα. Επιδέχεται λοιπόν πολλές βελτιώσεις και επεκτάσεις.

Γραμματική Ανάλυση κειμένου : Απαραίτητη προσθήκη στο τμήμα αυτό είναι η αναγνώριση αριθμών, συντομογραφιών και ειδικών συμβόλων (πχ %, $, °C) για την ορθότερη εκφώνηση οποιουδήποτε κειμένου.

Ασυνέχειες κατά τη συρραφή : Οι ασυνέχειες στη τονικότητα κατά τη συρραφή των διασυλλαβών μπορούν να αντιμετωπιστούν αν κανονικοποιηθεί όλη η βάση σε μια συγκεκριμένη τιμή θεμελιώδους συχνότητας με τη τεχνική HNM. Με την ίδια μέθοδο είναι δυνατός ο συγχρονισμός ως προς τη θεμελιώδη περιοδικότητα των τμημάτων φωνής που συρράπτονται για την εξάλειψη ασυνεχειών στη φάση.

Παραγωγή προσωδίας : Το σύστημα επιδέχεται βελτίωση ως προς την πληροφορία που εξάγεται από το κείμενο για την παραγωγή προσωδίας. Μία επέκταση είναι η συντακτική αναγνώριση των λέξεων (ρήμα, υποκείμενο κλπ.) έτσι ώστε ο ρόλος των λέξεων να είναι πιο πλούσιος. Επίσης η αναγνώριση δευτερευουσών προτάσεων είναι χρήσιμη στην οριοθέτηση φωνητικών φράσεων. Πολλές βελτιώσεις μπορούν να γίνουν για το ρυθμό της συνθετικής ομιλίας π.χ. κατασκευάζοντας κάποιο μοντέλο για τη διάρκεια των φθόγγων και των παύσεων μεταξύ των φωνητικών λέξεων και φράσεων.

Βάση διασυλλαβών : Η διαδικασία κατασκευής της βάσης δεν παρουσιάζει ιδιαίτερα μειονεκτήματα εκτός από το χρόνο που απαιτείται για την ηχογράφηση που οφείλεται στο πλήθος των απαραίτητων διασυλλαβών. Ο αριθμός τους μπορεί να μειωθεί στο 1/3 αν δε διαφοροποιούμε τις διασυλλαβές ως προς τον τονισμό των φωνηέντων. Σ' αυτή τη περίπτωση οι διασυλλαβές που περιέχονται στο λεξικό του προγράμματος *ispellH* είναι περίπου 2000 ενώ όλες οι δυνατές διασυλλαβές είναι περίπου 8000 (με τη διαφοροποίηση του τονισμού είναι περίπου 22000, παράγραφος 4.2.2). Η κατασκευή μιας τέτοιας βάσης όμως έχει σαν συνέπεια ότι η προσωδία για τον τονισμό των φωνηέντων πρέπει να αποδίδεται με κάποια επεξεργασία πχ με την τεχνική HNM. Ο αριθμός των διασυλλαβών μπορεί να μειωθεί ακόμα αν μειώσουμε τον αριθμό των συμπλεγμάτων συμφώνων σχηματίζοντας τα σύνθετα συμπλέγματα από απλά. Για παράδειγμα στη λέξη «*εκπέμπω*» η συμπροφορά μεταξύ των φθόγγων /*κ*/ και /*π*/ είναι πολύ μικρή[1] οπότε η διασυλλαβή /*εκπέ*/ μπορεί να σχηματιστεί από τη συρραφή των διασυλλαβών /*εκ#*/ και /*#πέ*/.

Επίσης μπορεί να χρησιμοποιηθεί κάποιος άλλος αλγόριθμος για την απομόνωση των διασυλλαβών μέσα στις λέξεις της βάσης ώστε να μειωθούν τα σφάλματα οριοθέτησης πχ να ανιχνεύεται το μέσο των φωνηέντων από τη φασματική σταθερότητά τους.

[1] Γιατί και τα δύο σύμφωνα είναι στιγμιαία

Μία πιθανή αλλαγή είναι η χρήση διαφορετικού σύνθετη φωνητικού σήματος αντί της τεχνικής HNM όπως η τεχνική PSOLA [CS86] [MC90] [DL93] ή η κωδικοποίηση GSM που θα βοηθούσε και στη συμπίεση της βάσης για τη πιο οικονομική αποθήκευσή της. Επίσης το σύστημα θα μπορούσε να συνεργαστεί με κάποιο αλγόριθμο μετατροπής χαρακτηριστικών ομιλίας (speaker modification) ώστε να δίνεται η δυνατότητα στο χρήστη να επιλέξει τα χαρακτηριστικά του ομιλητή (άνδρας, γυναίκα κλπ).

Βιβλιογραφία

[Τρι93] Μ. Τριανταφυλλίδης. *Νεοελληνική Γραμματική της Δημοτικής*. Ο.Ε.Δ.Β., 1993.

[AHK87] J. Allen, S. Hunnicut, and D. H. Klatt. *From Text to Speech: The MITalk System*. Cambridge Studies in Speech Science and Communication. Cambridge University Press, Cambridge, UK, 1987.

[BC95] A.W. Black and N. Camobell. Optimising selection of units from speech databases for concatenative synthesis. In *EUROSPEECH*, pages 581--584, Madrid Spain, 1995.

[Bel96] Bellcore. , 1996. http://www.belcore.com/demotoo/ORATOR/index.html.

[BP92] C. Benoit and L. Pols. On the assessment of synthetic speech. In G. Bailly and C. Benoit, editors, *Talking Machines, theories, models and designs*, pages 435--441. North-Holland Elsevier Science Publishers, Amsterdam, 1992.

[BT83] E. Bruckert and W. Tetschner. Three-Tiered Software and VLSI Aid Developmental System to Read Text Aloud. *Electronics*, 56:133--138, 1983.

[Cam92] W.N. Campbell. Syllable-Based Segmental Duration. In G. Bailly and C. Benoit, editors, *Talking Machines, theories, models and designs*, pages 211--224. North-Holland Elsevier Science Publishers, Amsterdam, 1992.

[CE82] J.L. Courbon and F. Emerand. A text-to-speech machine using sinthesis by diphones. In *Proc. ICASSP*, SPARTE, pages 1597--1600, Paris, 1982.

[CS86] F.J. Charpentier and M.G. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *Proc. ICASSP*, pages 2015--2018, Tokyo, 1986.

[DL93] T. Dutoit and H. Leich. MBR-PSOLA: Text-to-Speech Synthesis Based on an MBE Re-synthesis of the Segments Databases. *Speech Communication*, 13:435--440, 1993.

[DM68] N.R. Dixon and H.D. Maxey. Terminal Analog synthesis of Continuous Speech Using the Diphone Method of Segmental Assemply. *IEEE Transactions on Audio and Electroacoustics*, AU-16(1):40--50, 1968.

[DPH93] John R. Deller, John G. Proakis, and John H.L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.

[FY78] F. Fallside and S.J. Young. Speech output from a computer-controlled water-supply network. In *Proc. IEE*, volume 125, pages 157--161, 1978.

[Gab94] B. Gabioud. Articularory Models in Speech Synthesis. In E. Keller, editor, *Fundamentals of Speech Synthesis and Speech Recognition*, pages 215--230. John Wiley & Sons, New York, 1994.

[Hau93] A.G. Hauptmann. A first experiment in concatenation synthesis from a large corpus. In *Eurospeech*, SpeakEZ, pages 1701--1704, Berlin, 1993.

[KG61] J. Kelly and L. Gerstman. An Artificial Talker Driven from a Phonetic Input. *Journal of the Acoustical Society of America*, 33(1):835, 1961.

[Kla80] D.H. Klatt. Software for a Cascade/Parallel Formant Synthesizer. *Journal of the Acoustical Society of America*, 67(3):971--995, 1980.

[Kla82] D.H. Klatt. Review of text-to-speech conversion for english. In *Proc. ICASSP*, pages 1589--1592, Paris, 1982.

[Kla87] D.H. Klatt. Review of Text-to-Speech Conversion for English. *Journal of the Acoustical Society of America*, 82(3):737--793, 87.

[LM94] J.M. Lucassen and R.L. Mercer. An Information Theoreric Approach to the Automatic Determination of Phonemic Baseforms. In *Proc. ICASSP*, pages 42.5.1--42.5.4, San Diego, 1994.

[LM95] Yannis Stylianou Jean Laroche and Eric Moulines. High-quality speech modification based on a harmonic+noise model. In *EUROSPEECH*, pages 451--454, Madrid Spain, 1995.

[MC90] E. Moulines and F. Charpentier. Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication*, 9:453--467, 1990.

[Oli77] J.P. Olive. Rule synthesis of speech from dyadic units. In *Proc. ICASSP*, pages 568--570, Hartford, 1977.

[Pie81] J. Pierrehumbert. Synthesizing Intonation. *Journal of the Acoustical Society of America*, 70(4):985--995, 1981.

[Ril90] M.D. Riley. Tree-based modelling of speech synthesis. In *ESCA Workshop on Speech Synthesis*, pages 229--232, Autrans, Grenoble, 1990.

[Ril92] M.D. Riley. Tree-Based Modelling of Segmental Durations. In G. Bailly and C. Benoit, editors, *Talking Machines, theories, models and designs*, pages 265--273. North-Holland Elsevier Science Publishers, Amsterdam, 1992.

[RSF71] L.R. Rabiner, R.W. Schafer, and J.L. Flanagan. Computer Synthesis of Speech by Concatenation of Formant-Coded Words. *Bell System Technical Journal*, 50:1541--1558, 1971.

[Sor94] C. Sorin. Towards high-quality multilingual text-to-speech. In *CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology*, pages 53--62, Munich, 1994.

[Sty96] Yannis Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Superieure des Telecommunications Paris, Jan 1996.

[vS94] J.P.H. van Santen. Assignment of Segmental Duration it Text-to-Speech Synthesis. *Computer Speech and Language*, 8(2):95--128, 1994.