ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

«Σύγχρονες τεχνικές σχεδίασης και υλοποίησης συστήματος παραγωγής συνθετικής ομιλίας με επεξεργασία στο πεδίο του χρόνου»

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Αθήνα, Ιούλιος 2011

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

«Σύγχρονες τεχνικές σχεδίασης και υλοποίησης συστήματος παραγωγής συνθετικής ομιλίας με επεξεργασία στο πεδίο του χρόνου»

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Συμβουλευτική Επιτροπή : Γεώργιος Καραγιάννης
Ανδρέας Σταφυλοπάτης
Στέφανος Κόλλιας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 5η Ιουλίου 2011.

Αθήνα, Ιούλιος 2011

.....

Αιμίλιος του Ηλία Χαλαμανδάρης

Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Η συγκεκριμένη διατριβή δεν είναι τίποτα περισσότερο από το αποτέλεσμα μιας μεγάλης διαδρομής και περιπέτειας, στην διάρκεια της οποίας συνάντησα ανθρώπους που θέλω να ευχαριστήσω από την καρδιά μου, για την εμπιστοσύνη, την βοήθεια αλλά και την υποστήριξη που προσέφεραν. Καταρχήν θα ήθελα να ευχαριστήσω τον καθηγητή μου, μέντορα και φίλο κ. Γιώργο Καραγιάννη που με ευκολία υπήρξε και τα τρία παραπάνω όποτε χρειαζόταν. Ευχαριστώ θερμά τους καθηγητές κ. Στέφανο Κόλλια, κ. Ανδρέα Σταφυλλοπάτη, κ. Τίμο Σελλή, κ. Παναγιώτη Τσανάκα και κ. Σίμο Ρετάλη που μου έκαναν την τιμή να είναι μέλη της συμβουλευτικής και εξεταστικής επιτροπής του διδακτορικού μου.
Τίποτα από όσα καταγράφονται εδώ, αλλά και τα πολλά περισσότερα που δεν μπορούν να περιγραφούν δεν θα ήταν δυνατά αν δεν είχα την τύχη να συνεργαστώ με την φανταστική ομάδα των Σπύρου Ράπτη, Πύρρου Τσιάκουλη και Σωτήρη Καραμπέτσου. Ευχαριστώ τους καλούς συνεργάτες στο ΙΕΛ, ιδιαίτερα την Νάσια Δήμου, την γλωσσολόγο από τα ξένα, τον Άκη Αθανασέλη, τον Μυτιληνιότερο όλων, τον Θανάση Πρωτόπαπα, την Ευαγγελία Τσιλιγιάννη τον Γιώργο Γιαννόπουλο, καθώς επίσης και την Μαρία Φουντά, με την οποία πρωτογνωρίσαμε την σύνθεση φωνής ως φοιτητές.
Δεν μπορώ παρά να είμαι ευγνώμων για την στήριξη, συμπαράσταση και βοήθεια που μου προσέφερε η οικογένειά μου και ειδικότερα οι γονείς μου, Μαρία και Ηλίας Χαλαμανδάρης, τα αδέρφια μου, Ρόη, Γιώργος, Δημήτρης και Ελίζα, όπως επίσης και η αγαπημένη μου σύντροφός Simran.
Η διαδρομή συνεχίζεται, το ίδιο και η περιπέτεια...

Περίληψη

Η παρούσα διατριβή παρουσιάζει την ερευνητική προσπάθεια και τα αποτελέσματα αυτής αναφορικά με την σχεδίαση και υλοποίηση ενός συστήματος συνθετικής ομιλίας με τεχνικές επεξεργασίας ψηφιακού σήματος στο πεδίο του χρόνου. Πιο συγκεκριμένα, η διατριβή, εκτός από την εκτενή αναφορά σε ερευνητικά αποτελέσματα παρόμοιων προσπαθειών στον ίδιο γνωστικό τομέα, εστιάζει στα εξής επιμέρους σημεία:

- Στην μελέτη της προσωδίας και στην έμμεση μοντελοποίηση αυτής μέσω του υποσυστήματος βέλτιστης επιλογής ακουστικών μονάδων. Προτείνεται ένας καινοτόμος μηχανισμός για την μοντελοποίηση, παραγωγή και εφαρμογή πρότυπων καμπυλών προσωδίας στο συνθετικό σήμα φωνής, με τρόπο που επιτρέπει την διατήρηση της μικροπροσωδίας και την ποικιλότητα στην προσωδία.
- Στον σχεδιασμό, υλοποίηση και επεξεργασία του πρωτογενούς υλικού για την βάση δεδομένων του συνθέτη φωνής, το οποίο βασίζεται σε ηχογραφήσεις ενός φυσικού ομιλητή. Η μέθοδος που προτείνεται αποτελεί μία καινοτόμο τεχνική για τον σχεδιασμό του σώματος κειμένου, λαμβάνοντας υπόψη τα ιδιαίτερα χαρακτηριστικά του συνθέτη φωνής και εξασφαλίζοντας μέγιστη κάλυψη διαφορετικών φαινομένων.
- Στα διαφορετικά υποσυστήματα του συνθέτη ομιλίας τα οποία είναι υπεύθυνα για την επεξεργασία φυσικής γλώσσας, την κανονικοποίηση κειμένου από Greeklish και την φωνητική μεταγραφή του κειμένου, προτείνοντας νέες μεθόδους για την φωνητική μεταγραφή για την Ελληνική γλώσσα, όπως επίσης και την μετατροπή από Greeklish σε ορθά Ελληνικά.
- Σε ειδικές προσαρμογές και τεχνικές που προτείνονται για την δημιουργία ενός συνθέτη ομιλίας ειδικά σχεδιασμένου για την σύμπραξη με υποστηρικτικά εργαλεία προσβασιμότητας για εμποδιζόμενα άτομα, λαμβάνοντας μέριμνα για τις ανάγκες χρήστη και τα σενάρια χρήσης.

Τα αποτελέσματα της συγκεκριμένης ερευνητικής προσπάθειας έχουν οδηγήσει στην υλοποίηση του ποιοτικότερου συστήματος συνθετικής ομιλίας για την Ελληνική γλώσσα, ενώ παράλληλα έβαλε τις βάσεις δημιουργίας συνθέτη ομιλίας για την Βουλγαρική γλώσσα με εξαιρετικά υψηλή φυσικότητα και καταληπτότητα.

Abstract

This thesis presents the research effort and its results regarding the design and implementation of a speech synthesis system based on time-domain techniques. More specifically, this thesis, apart from a detailed literature review, focuses on the following specific points:

- The study of prosody and its modeling for the Text-to-Speech system. The proposed mechanism provides a novel algorithm for modeling, producing and applying of prosodic curves onto the synthetic speech signal, in a manner that allows the preservation of microprosody and diversity in prosodic patterns.

- The design and development of primary material for the database of a voice synthesizer, which is based on recordings of a native speaker. The method proposed is an innovative one for the design of the textual corpus, taking into account the specific characteristics of the voice synthesizer and ensuring maximum coverage of different acoustic and linguist phenomena.

- The different subsystems of the speech synthesizer, which are responsible for natural language processing, text normalization and phonetic transcription of the text. New methods have been proposed for the phonetic transcription of the Greek language, as well as the conversion from Greeklish to Greek.

- Extra focus was given in the customization and adaptation of the speech synthesis system for optimal performance in the framework of accessibility, providing the ground for an optimized TTS system as an assistive tool.

The results of this research effort have led to the realization of the highest quality TTS system for Greek language, providing at the same time the basis for creating a TTS system for the Bulgarian language of similarly high quality and naturalness.

Κατάλογος Περιεχομένων

1. ΕΙΣΑΓΩΓΗ 1

1.2 Ερευνητικη συνεισφορα της διατριβης 2
1.2.1 Μοντελοποίηση της προσωδίας στην σύνθεση φωνής 2
1.2.2 Σχεδιασμός και ανάπτυξη της βάσης δεδομένων ενός συνθέτη φωνής 2
1.2.3 Ειδικά θέματα επεξεργασίας κειμένου για τον συνθέτη φωνής 3
1.2.4 Ειδικά θέματα σχεδιασμού και προσαρμογής συνθέτη φωνής για χρήση σε προσβάσιμα εργαλεία 4
1.3 Διαρθρωση της διατριβης 4
1.4 Βιβλιογραφια Κεφαλαιου 7

2. ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΤΕΧΝΟΛΟΓΙΑ ΣΥΝΘΕΣΗΣ ΦΩΝΗΣ 9

2.1 Τι ειναι η συνθεση φωνης 10
2.2 Ιστορικη αναδρομη της συνθεσης φωνης 12
2.3 Αρχιτεκτονικη ενος συνθετη φωνης 14
2.3.1 Το υποσύστημα Επεξεργασίας Φυσικής Γλώσσας 15
2.3.2 Το υποσύστημα Ψηφιακής Επεξεργασίας Σήματος 16
2.3.3 Κατηγορίες τεχνολογιών σύνθεσης φωνής 16
2.3.3.1 Παραμετρικές μέθοδοι σύνθεσης φωνής 18
2.3.3.1.1 Η μέθοδος σύνθεσης με μοντελοποίηση αρθρωτών 18
2.3.3.1.2 Η μέθοδος σύνθεσης φωνής με κανόνες 19
2.3.3.1.3 Η μέθοδος σύνθεσης φωνής με χρήση HMM 20
2.3.3.2 Data-driven μέθοδοι σύνθεσης φωνής 22
2.3.3.2.1 Η μέθοδος σύνθεσης φωνής με παράθεση ακουστικών μονάδων 22
2.3.3.3 Μέθοδοι ψηφιακής επεξεργασίας και σύνθεσης σήματος φωνής με παράθεση 24
2.3.3.3.1 Μέθοδος σύνθεσης MBROLA 25
2.3.3.3.2 Μέθοδος σύνθεσης WSOLA 25
2.3.3.3.3 Μέθοδος σύνθεσης HNM 26
2.3.3.3.4 Μέθοδος σύνθεσης TD-PSOLA 27
2.4 Βιβλιογραφια Κεφαλαιου 30

3. ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΑΙ ΠΑΡΑΓΩΓΗ ΠΡΟΣΩΔΙΑΣ 33

3.1 Εισαγωγη 33
3.1.1 Ορισμός 34
3.1.1.1 Τόνος και επιτονισμός(Pitch and Intonation) 36
3.1.1.2 Φραστικός τόνος 36
3.1.1.3 Λεξικός τόνος 36
3.1.1.4 Λέξη επιτονισμού 37
3.1.1.5 Κλίση 37
3.1.1.6 Ρυθμός 37
3.2 Μοντελοποιηση επιτονισμου 38
3.2.1 Αναλυτικές Μέθοδοι: Μοντέλο ToBI 40
3.2.2 Αναλυτικές Μέθοδοι: Μοντέλο Fujisaki 40
3.2.3 Γεννετικές Μέθοδοι: Μοντέλο IPO (Ολλανδική σχολή) 41
3.2.4 Γεννετικές Μέθοδοι: Μοντέλο Tilt RFC 42
3.2.5 Μοντελοποίηση Επιτονισμού – Πρώτα Συμπεράσματα 43
3.3 Η αξια του μοντελου επιτονισμου στην συνθεση φωνης 44
3.4 Η προσεγγιση μας 44
3.4.1 Θεωρία πίσω από την προσέγγιση μας 44
3.4.2 Υπολογισμός των πρότυπων καμπυλών θεμελιώδους συχνότητας 46
3.4.3 Υπολογισμός της καμπύλης επιτονισμού 48

3.4.4 Εφαρμογή της καμπύλης επιτονισμού *50*
3.4.5 Αποτελέσματα *53*
3.4.6 Συζήτηση – Θέματα προς διερεύνηση *54*
3.5 Βιβλιογραφια Κεφαλαιου 55

4. Η ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΣΥΝΘΕΣΗΣ ΦΩΝΗΣ 59

4.1 Εισαγωγη 60
4.2 Σχεδιαση σωματος κειμενου προς ηχογραφηση 61
4.2.1 Το μήκος της ακουστικής μονάδας *61*
4.2.2 Διαφορετικά πεδία αναφοράς σύνθεσης *63*
4.2.3 Μέθοδοι επιλογής βέλτιστων προτάσεων *65*
4.2.4 Αλγόριθμος επιλογής βέλτιστου υπο-σώματος κειμένου *66*
4.3 Η προσεγγιση μας 68
4.3.1 Στάδιο 1: Επιλογή του σώματος κειμένου S *69*
4.3.1.1 Κάλυψη συχνότερων λέξεων 70
4.3.1.2 Κάλυψη σε επίπεδο διφωνημάτων 70
4.3.1.3 Κάλυψη σε επίπεδο προσωδιακών φαινομένων 70
4.3.2 Στάδιο 2: Σταχυολόγηση του σώματος κειμένου S με την χρήση του συνθέτη φωνής *72*
4.3.3 Στάδιο 3: Εμπλουτισμός και διορθώσεις στο σώμα κειμένου S_{sub} *74*
4.3.4 Στάδιο 4: Περαιτέρω βελτίωση και εμπλουτισμός του σώματος κειμένου με νέες ακουστικές μονάδες *76*
4.4 Αποτελεσματα 76
4.5 Δημιουργια ηχογραφημενου σωματος κειμενου 80
4.5.1 Ηχογράφηση σώματος κειμένου *80*
4.5.2 Επεξεργασία ηχογραφήσεων *81*
4.5.2.1 Δυναμική συμπίεση του σήματος φωνής 81
4.5.3 Αυτόματη κατάτμηση ηχογραφημένου σώματος κειμένου *83*
4.6 Υπολογισμος των σημειων αναλυσης του αρχικου σηματος (Pitchmarks) 86
4.6.1 Εισαγωγικά *86*
4.6.2 Ο προτεινόμενος αλγόριθμος *90*
4.6.2.1 Μέθοδος υπολογισμού της καμπύλης της θεμελιώδους συχνότητας 91
4.6.2.2 Εκτίμηση εμφώνων και αφώνων τμημάτων φωνής 92
4.6.2.3 Εκτίμηση της θεμελιώδους συχνότητας με παρεμβολή 95
4.6.2.4 Επιλογή των σημείων ανάλυσης (pitchmarks) 96
4.6.2.5 Αποτελέσματα της μεθόδου 97
4.6.2.6 Πειραματική αξιολόγηση μεθόδου για σύνθεση φωνής 100
4.7 Βιβλιογραφια Κεφαλαιου 102

5. ΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΟΥ 105

5.1 Κανονικοποιηση κειμενου 106
5.2 Συστημα Φωνητικης Μεταγραφης 106
5.2.1 Φωνητική της Ελληνικής γλώσσας *107*
5.2.1.1 Τα φωνήεντα 107
5.2.1.2 Τα σύμφωνα 108
5.2.2 Γενικοί κανόνες φωνητικής μεταγραφής για την Ελληνική γλώσσα *109*
5.2.3 Σημαντικά θέματα και προβλήματα στην φωνητική μεταγραφή για τα ελληνικά *110*
5.2.4 Μέθοδοι βασισμένες σε λεξικά *111*
5.2.5 Μέθοδοι βασισμένες σε κανόνες *112*
5.2.6 Μέθοδοι βασισμένες σε δεδομένα (Data-driven methods) *113*
5.2.7 Η προσέγγισή μας *114*
5.2.7.1 Η αντιστοίχιση της ορθογραφικής στην φωνητική αναπαράσταση 117
5.2.7.2 Αυτόματη ανεύρεση κανόνων 119

5.2.8 Αποτελέσματα – Συμπεράσματα *122*
5.3 Αποτελεσματα - Θεματα για περαιτερω μελετη 125
5.4 Βιβλιογραφια Κεφαλαιου 126

6. ΤΟ ΦΑΙΝΟΜΕΝΟ ΤΩΝ GREEKLISH 129

6.1 Εισαγωγη 130
6.2 Στατιστικη αναλυση 131
6.3 Κανονες απεικονισης – χαρακτηριστικα της γραφης Greeklish 135
6.4 Διαφορετικοι τυποι γραφης Greeklish 136
6.5 Αναλυση των προτιμησεων στο δειγμα 139
6.6 Το συστημα αυτοματης μετατροπης Greeklish σε Ελληνικα 139
6.6.1 Η δική μας προσέγγιση *139*
6.6.2 Αξιολόγηση της αποτελεσματικότητας του συστήματος *141*
6.7 Βιβλιογραφια Κεφαλαιου 147

7. ΣΥΣΤΗΜΑ ΣΥΝΘΕΤΗΣ ΦΩΝΗΣ ΜΕ ΈΜΦΑΣΗ ΣΕ ΘΕΜΑΤΑ ΠΡΟΣΒΑΣΙΜΟΤΗΤΑΣ 149

7.1 Εισαγωγη 150
7.2 Απαιτησεις χρηστη 151
7.3 Ειδικες προσαρμογες συνθετη φωνης 152
7.3.1 Προσαρμογή επεξεργασίας φυσικής γλώσσας *152*
7.3.2 Βελτιστοποίηση ταχύτητας και απόκρισης συστήματος *154*
7.3.3 Πολυγλωσσικότητα και ακουστική ποιότητα *155*
7.4 Αξιολογηση και επικυρωση αποτελεσματων 158
7.4.1 Αξιολόγηση ακουστικής ποιότητας *158*
7.4.1.1 Ακουστικά πειράματα αξιολόγησης 158
7.4.1.1.1 Πείραμα 1: Αξιολόγηση σε επίπεδο πρότασης 158
7.4.1.1.2 Πείραμα 2: Αξιολόγηση σε επίπεδο λέξης 159
7.4.1.1.3 Πείραμα 3: Αξιολόγηση σε επίπεδο παραγράφου 160
7.4.2 Αξιολόγηση χρηστικότητας *161*
7.4.2.1 Ανευρετική (heuristic) αξιολόγηση 161
7.4.2.2 Πειραματική αξιολόγηση 162
7.4.2.2.1 Αποτελεσματικότητα 162
7.4.2.2.2 Αποδοτικότητα 162
7.4.2.2.3 Ικανοποίηση 163
7.5 Συμπερασματα – Συζητηση 163
7.6 Βιβλιογραφια Κεφαλαιου 164

8. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ 167

8.1 Κυριες συνεισφορες της διατριβης 168
8.1.2 Μοντελοποίηση της προσωδίας στην σύνθεση φωνής *168*
8.1.3 Σχεδιασμός και ανάπτυξη της βάσης δεδομένων του συστήματος σύνθεσης φωνής *169*
8.1.4 Ειδικά θέματα επεξεργασίας κειμένου για τον συνθέτη φωνής *169*
8.1.5 Ειδικά θέματα σχεδιασμού και προσαρμογής συνθέτη φωνής για χρήση σε προσβάσιμα εργαλεία *170*
8.2 Θεματα μελλοντικης ερευνας 170
8.2.1 Μελέτη της προσωδίας συναισθηματικού λόγου *171*
8.2.2 Μελέτη και ανάπτυξη προ-επεξεργασίας κειμένου για εκφραστικό λόγο *171*
8.2.3 Μελέτη για τον σχεδιασμό της βάσης δεδομένων ενός εκφραστικού συνθέτη και του αλγορίθμου βέλτιστης επιλογής *172*
8.2.4 Μελέτη και σχεδιασμός νέων καινοτόμων εφαρμογών για καθημερινή χρήση *172*
8.3 Βιβλιογραφια Κεφαλαιου 173

9. ΚΑΤΑΛΟΓΟΣ ΔΗΜΟΣΙΕΥΣΕΩΝ ΤΟΥ ΣΥΓΓΡΑΦΕΑ 176

9.1 ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΕ ΠΕΡΙΟΔΙΚΑ ΜΕ ΚΡΙΤΕΣ 177
9.2 ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΕ ΣΥΛΛΟΓΙΚΟΥΣ ΤΟΜΟΥΣ ΜΕ ΚΡΙΤΕΣ 177
9.3 ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΕ ΔΙΕΘΝΗ ΣΥΝΕΔΡΙΑ ΜΕ ΚΡΙΤΕΣ.....178

Κατάλογος Σχημάτων

Σχήμα 1: Αναπαράσταση της λειτουργίας ενός απλού συνθέτη φωνής ως ένα σύστημα με είσοδο το κείμενο και έξοδο την συνθετική φωνή. 10
Σχήμα 2: Αναπαράσταση της μηχανής σύνθεσης ήχων του Wolfgang von Kempelen. 12
Σχήμα 3: Η εξέλιξη των τεχνολογιών σύνθεσης και αναγνώρισης φωνής, από την αρχή τους μέχρι τον τελικό τους στόχο. 14
Σχήμα 4: Αναπαράσταση της λειτουργίας ενός απλού συνθέτη φωνής. 15
Σχήμα 5: Αναπαράσταση και ανάδειξη των επιμέρους υποσυστημάτων ενός σύγχρονου συνθέτη φωνής. 15
Σχήμα 6: Κατηγοριοποίηση των τεχνολογιών σύνθεσης φωνής με βάση την θεωρία σύνθεσης και την τεχνική σύνθεσης. 17
Σχήμα 7: Κατηγοριοποίηση των τεχνολογιών σύνθεσης με βάση το πρωτογενές υλικό και την τεχνική επεξεργασίας σήματος. 17
Σχήμα 8: Δομικό διάγραμμα μηχανής σύνθεσης με formants σύμφωνα με το μοντέλο του Klatt. [Klatt1976] 20
Σχήμα 9: Δομικό διάγραμμα συστήματος σύνθεσης φωνής από κείμενο με χρήση HMM. Πηγή: [Zen 2007] 21
Σχήμα 10: Παράδειγμα σύνθεσης με παράθεση λέξεων για εκφώνηση της ώρας. 23
Σχήμα 11: Παράδειγμα σύνθεσης της λέξης /έλα/ με χρήση διφωνημάτων. Στο συγκεκριμένο παράδειγμα απεικονίζεται και το υποσύστημα επιλογής βέλτιστης ακουστικής μονάδας (διφώνημα). (Πηγή: [Karabetsos2011]) 24
Σχήμα 12: Απεικόνιση της μεθόδου WSOLA όπου η επιλογή των επικαλυπτόμενων παραθύρων βασίζεται σε κριτήριο ομοιότητάς με βάση το SFFT των παραθύρων. (Πηγή: [Verhelst1992]) 26
Σχήμα 13: Σύνθεση φωνής με βάση την μέθοδο HNM (Πηγή:[Stylianou2001]) 27
Σχήμα 14: Αναπαράσταση σήματος φωνής και δημιουργία των ST σημάτων για επεξεργασία μέσω PSOLA. 28
Σχήμα 15: Κυματομορφή σήματος φωνής και επισημείωση πιθανών pitchmark σημείων ανάλυσης. Τα τελευταία σημειώνονται με κατακόρυφες διακεκομμένες γραμμές. 28
Σχήμα 16: Σχηματική αναπαράσταση του τρόπου αύξησης (α) και μείωσης (β) της βασικής συχνότητας με τη μέθοδο TD-PSOLA. 29
Σχήμα 17: Σχηματική απεικόνιση των διαφορετικών ερευνητικών πτυχών που έχει η μελέτη της προσωδίας. (Πηγή:[Bolinger1989]) 35
Σχήμα 18: Ο ρόλος των μοντέλων επιτονισμού ως σύνδεσμος μεταξύ των γλωσσικών δομών και των ακουστικών πραγματώσεων στην καμπύλη του F0. (Πηγή: [Butzberger1990]) 39
Σχήμα 19: Αναπαράσταση προσιδιακής επισήμειωσης με την μέθοδο ToBi για την φράση /Marianna made the mar,alade/. (Πηγή ['*Guidelines for TOBI labeling*' by Mary Beckman and Gayle Ayers) 40
Σχήμα 20: Η μοντελοποίηση της προσωδίας στην σύνθεσης φωνής με την μέθοδο του Fujisaki. 41
Σχήμα 21: Περιγραφή της θεμελιώδους συχνότητας στο σήμα φωνής με το μοντέλο IPO (Ολλανδική σχολή). 42
Σχήμα 22: Το διάγραμμα ροής για την εκπαίδευση της μηχανής προσωδίας (α) και της διαδικασίας παραγωγής μοντέλου προσωδίας κατά την εκτέλεση (β). 45
Σχήμα 23: Απεικόνιση συσταδοποίησης καμπυλών προσωδίας σε συλλαβές. Για κάθε συστάδα καμπυλών απεικονίζεται με σκούρα γραμμή η μέση καμπύλη. 47

Σχήμα 24: Δέντρο απόφασης για τον τύπο πρότυπης καμπύλης F0 που πρέπει να πάρει μία συλλαβή CCV ανάλογα με τα περικείμενα γλωσσικά χαρακτηριστικά της. 48
Σχήμα 25: Απεικόνιση συνθετικού σήματος φωνής με παράθεση (Γ), της καμπύλης Pitch όπως προκύπτει από την παράθεση των επιμέρους καμπυλών των ακουστικών μονάδων (Α) και η λειασμένη καμπύλη Pitch όπως προκύπτει από βαθυπερατό φίλτρο (Β). Οι κατακόρυφες ευθείες επισημειώνουν τα όρια των διφωνημάτων (ακουστικών μονάδων) που επιλέχθηκαν για την σύνθεση της συγκεκριμένης φράσης (/Golden gate A/). 50
Σχήμα 26: Η ενσωμάτωση της μικροπροσωδίας κατά το μήκος ενός φωνήματος στο συνθετικό σήμα φωνής. Η τοπική τιμή της προσωδίας του φυσικού σήματος προστίθεται με βάρος που εξαρτάται από την απόσταση του σημείου από το μέσο του φωνήματος, ώστε το μέσο όπου πραγματοποιείται η ένωση διαφορετικών ακουστικών μονάδων να μην περιέχει τοπικές ασυνέχειες στην F0. 51
Σχήμα 27: Τυπική καμπύλη για την μεταβλητή w(t) από την οποία εξαρτάται το μέγεθος της μικροπροσωδίας του φυσικού σήματος που ενσωματώνεται στην τελική καμπύλη της θεμελιώδους συχνότητας. Η χρονική στιγμή 0,5 αντιστοιχεί στο μέσο του φωνήματος και το σημείο όπου πραγματοποιείται η παράθεση των δύο διαδοχικών ακουστικών μονάδων από την βάση δεδομένων. 52
Σχήμα 28: Ο υπολογισμός της τελικής καμπύλης θεμελιώδους συχνότητας για ένα έμφωνο τμήμα φωνής (α). Η καμπύλη (β) έχει δημιουργηθεί από την παράθεση και λείανση των επιμέρους καμπυλών θεμελιώδους συχνότητας για κάθε συλλαβή. Η καμπύλη (γ) προκύπτει από την ενσωμάτωση της μικροπροσωδίας των έμφωνων ήχων στην λειασμένη τελική καμπύλη μοντελοποίησης. 53
Σχήμα 29: Το διάγραμμα ροής για τον σχεδιασμό του σώματος κειμένου προς ηχογράφηση. Με διακεκομμένες γραμμές ορίζονται τα διαφορετικά στάδια της διαδικασίας. Η ένωση των επιμέρους σωμάτων κειμένου Α, Β και Γ αποτελούν το τελικό σώμα κειμένου που κατασκευάστηκε για τον συνθέτη ομιλίας. 68
Σχήμα 30: Χάρτης φωνητικής κάλυψης του επιλεγμένου σώματος κειμένου. Για κάθε διφώνημα αναφέρεται η συχνότητα εμφάνισης. Η κατακόρυφη στήλη ορίζει το αριστερό φώνημα και η οριζόντια γραμμή το δεξί φώνημα για κάθε διφώνημα. 79
Σχήμα 31: Διάγραμμα συμπίεσης ήχου 82
Σχήμα 32: Σχηματική αναπαράσταση της κυματομορφής μιας ηχογράφησης όπως ηχογραφήθηκε (α) και κατόπιν της δυναμικής συμπίεσης της έντασης (β). Η συγκεκριμένη διαδικασία επιτρέπει την κανονικοποίηση των ηχογραφήσεων στα -20db RMS. 82
Σχήμα 33: Παράδειγμα κατάτμησης σήματος φωνής (εκφορά της λέξης /σημείωση/) με βάση φωνήματα. 83
Σχήμα 34: Απεικόνιση ιεραχικής συσταδοποίησης για το φώνημα /N/. Οι τελευταίες συστάδες, με τα λιγότερα μέλη ανά συστάδα και με σημαντική απόσταση από τις κοντινότερες συστάδες, αφαιρούνται αυτόματα από την βάση δεδομένων αφού περιέχουν ενδεχόμενα σφάλματα. 85
Σχήμα 35: Καμπύλη πυκνότητας πιθανότητας διαρκειών για το φώνημα /ε/ στην βάση δεδομένων. Τα στιγμιότυπα που χαρακτηρίζονται ως outliers αφαιρούνται αυτόματα από την βάση δεδομένων. 86
Σχήμα 36: Κυματομορφή ηλεκτρο-λαρυγγογράφου και η αντίστοιχη κυματομορφή του σήματος φωνής. Με κατακόρυφες γραμμές επισημειώνονται τα σημεία κλεισίματος της γλωττίδας (GCI). 87

Σχήμα 37: Λαρυγγογράφος και τοποθέτησή του στον ομιλητή. Πηγή: Inst. of Phonetic Sciences in Amsterdam. 88
Σχήμα 38: Παράθεση με TD-PSOLA δύο ακουστικών μονάδων φωνής με λάθος επισημειωμένα σημεία ανάλυσης (pitchmarks). Η διαφορά φάσης στα σημεία ανάλυσης προκαλεί παραμόρφωση του σήματος κατά την παράθεση με επικάλυψη. Πηγή: [Stylianou2001]. 90
Σχήμα 39: Σχηματική αναπαράσταση του αλγορίθμου εντοπισμού σημείων ανάλυσης Pitchmarks. 91
Σχήμα 40: Διάγραμμα ροής για την απόφαση έμφωνου/άφωνου ήχου στο σήμα φωνής. 93
Σχήμα 41: Απεικόνιση σήματος φωνής και το αποτέλεσμα από την αυτόματη επισημείωση έμφωνων και άφωνων ήχων μέσα στο σήμα φωνής. 94
Σχήμα 42: Ιστόγραμμα με βάση τα zero-crossings για ένα σήμα φωνής. Το κατώφλι για την διάκριση έμφωνων από άφωνους ήχους ορίζεται αυτόματα 0.05 όπου και παρατηρείται τοπικό ελάχιστο στην κατανομή πιθανότητας. 95
Σχήμα 43: Παράδειγμα υπολογισμού των σημείων ανάλυσης σε άφωνα τμήματα του σήματος φωνής. Τα σημεία ανάλυσης υπολογίζονται με βάση την θεμελιώδη συχνότητα του περιβάλλοντος. 96
Σχήμα 44: Σύγκριση των Pitchmarks όπως αυτά υπολογίζονται από τον αλγόριθμο DYPSA (α) και τον αλγόριθμό μας (β). 97
Σχήμα 45: Πρόσθεση δύο παραθύρων ST με διαφορά φάσης στα σημεία ανάλυσης PitchMarks. Στο γράφημα β) τα συνεπή PitchMarks οδηγούν σε πρόσθεση χωρίς παραμόρφωση, ενώ στην περίπτωση γ) εμφανίζεται παραμόρφωση λόγο διαφοράς φάσης στα σημεία ανάλυσης κατά 1msec. Τα ST στα γραφήματα β) και γ) έχουν πολλαπλασιασθεί με παράθυρο Hanning. 98
Σχήμα 46: Τα Pitchmarks ενός σήματος φωνής στο σήμα (α), στο λαρυγγογράφημα (β), και στην πρώτη παράγωγο του δεύτερου (γ). 99
Σχήμα 47: Η κατανομή της διαφοράς των υπολογισμένων σημείων ανάλυσης από τα αντίστοιχα σημεία GCI όπως προέκυψαν από το σήμα του λαρυγγογράφου. Άρεν ομιλητής με κωδικό KED. Πηγή CMU KED Database. (άξονας Υ σε χιλιάδες δείγματα, άξονας Χ σε δευτερόλεπτα της ώρας). 100
Σχήμα 48: Συνδυασμοί γραφημάτων όπου η προφορά του /i/ παρουσιάζει δυσκολίες. 111
Σχήμα 49: Διάγραμμα ροής του αλγορίθμου φωνητικής μεταγραφής. Για κάθε κανόνα που υπολογίζεται, ελέγχεται η επαλήθευσή του στο σύνολο του λεξικού. Αν αποτύχει η επαλήθευση, τότε το μήκος του περικειμένου αυξάνει σταδιακά και επαληθεύεται ο νέος κανόνας. 115
Σχήμα 50: Φωνητική μεταγραφή της λέξης *κυάλια* με χρήση του φωνήματος *epsilon*. 116
Σχήμα 51: Αντιστοίχιση των γραμμάτων στην φωνητική αναπαράσταση μέσω της δυναμικής στρέβλωσης. 118
Σχήμα 52: Φωνητική μεταγραφή της λέξης *διάγραμμα* χωρίς την χρήση του φωνήματος *epsilon* (προσέγγισή μας). 118
Σχήμα 53: Φωνητική μεταγραφή της λέξης *διάγραμμα* με την χρήση του φωνήματος *epsilon* (άλλες προσεγγίσεις). 119
Σχήμα 54: Πλήθος κανόνων ως προς το μήκος του περικειμένου. 120
Σχήμα 55: Ιστόγραμμα εμφάνισης κανόνων φωνητικής μεταγραφής για τα Ελληνικά όπως παρατηρήθηκαν στο σώμα εκπαίδευσης 122
Σχήμα 56: Ποσοστό επιτυχίας του συστήματος αυτόματης φωνητικής μεταγραφής σε σχέση με το σώμα λέξεων εκπαίδευσης. 124

Σχήμα 57: Ιστόγραμμα συχνότητας εμφάνισης των μοναδικών λέξεων στο σώμα κειμένου μελέτης. 131
Σχήμα 58: Κατανομή της προέλευσης του σώματος κειμένου ανά χώρα 132
Σχήμα 59: Κατανομή των μοναδικών επισκεπτών σε νέους και συχνούς επισκέπτες. 133
Σχήμα 60: Γεωγραφική προέλευση των χρηστών του συστήματος. 134
Σχήμα 61: Κατανομή διαφορετικών χρηστών ανά αριθμό μετατροπών. 135
Σχήμα 62: Μέσος όρος προτίμησης των τριών τύπων Greeklish. 137
Σχήμα 63: Κατανομή του ποσοστού ασυνέπειας ως προς τον προτιμητέο τύπο γραφής, ανά χρήστη. 138
Σχήμα 64: Σχηματικό διάγραμμα λειτουργίας του συστήματος. 140
Σχήμα 65: Σχηματική απεικόνιση του υποσυστήματος NLP που πραγματοποίηθηκε στο πλαίσιο προσαρμογής του συνθέτη φωνής για περιβάλλοντα ανάγνωσης οθόνης. 153
Σχήμα 66: Αναπαράσταση εκρηκτικού ήχου στο σήμα φωνής. Επισημειώνεται η περιοχή όπου πραγματοποιείται η "έκρηξη" του φωνήματος και χρησιμοποιείται ως περιοχή όπου δεν επιτρέπεται η επεξεργασία της διάρκειας του ήχου. 157

Κατάλογος Πινάκων

Πίνακας 1: Στατιστική ανάλυση των μοναδικών λέξεων για διαφορετικού ύφους εφημερίδες. Για κάθε ζευγάρι εφημερίδων απεικονίζεται το ποσοστό των μοναδικών λέξεων που έχουν κοινό. 64
Πίνακας 2: Ποσοστό των κοινών μοναδικών λέξεων για δύο διαφορετικές εφημερίδες (Πολιτική και Αθλητική) στη διάρκεια τριών διαδοχικών ετών. 65
Πίνακας 3: Ψευδοκώδικας ροής του αλγόριθμου επιλογής βέλτιστων προτάσεων. 67
Πίνακας 4: Αλγόριθμος σταχυολόγησης του σώματος κειμένου με χρήση των δεδομένων από το υποσύστημα βέλτιστης επιλογής ακουστικών μονάδων του συστήματος σύνθεσης φωνής. 73
Πίνακας 5: Τα αποτελέσματα της αξιολόγησης με την μέθοδο MOS, όπου οι ερωτηθέντες βαθμολόγησαν από το 1 έως το 5 τα δείγματα συνθετικού λόγου, με διαφορετικές μεθόδους υπολογισμού των σημείων ανάλυσης (pitch-marks). 101
Πίνακας 6: Τα αποτελέσματα της προτίμησης των δειγμάτων με διαφορετικούς αλγορίθμους υπολογισμού των σημείων ανάλυσης (pitch-marks). 101
Πίνακας 7: Οι συμφωνικοί φθόγγοι της Ελληνικής γλώσσας Έμφωνοι (Φ) και άφωνοι (Α), ανά τόπο και τρόπο άρθρωσης σύμφωνα με το διεθνές φωνητικό αλφάβητο. 107
Πίνακας 8: Η θέση των φωνηέντων της Ελληνικής γλώσσας στον ακουστικό χώρο σύμφωνα με το διεθνές φωνητικό αλφάβητο 107
Πίνακας 9: Πλήθος κανόνων ιεραρχημένοι με βάση το μήκος του περικειμένου. Οι τιμές στις παραγράφους αναφέρονται στο μήκος του αριστερού και δεξιού περικειμένου για κάθε κανόνα αντίστοιχα. 121
Πίνακας 10: Αποτελέσματα μεθόδου με διαφορετικά σύνολα εκμάθησης του συστήματος. 123
Πίνακας 11: Κατανομή της προέλευσης του σώματος κειμένου ανά χώρα στην διάρκεια 6 μηνών. 132
Πίνακας 12: Οι τρεις διαφορετικοί τύποι Greeklish και οι προτιμήσεις αυτών από τους χρήστες. 137
Πίνακας 13: Διαχωρισμός των λέξεων σε κατηγορίες προς έλεγχο. 141
Πίνακας 14: Δείγμα λέξεων που ελέγχθηκαν χειρωνακτικά για την αξιολόγηση του συστήματος. 142
Πίνακας 15: Αποτελέσματα αξιολόγησης του συστήματος για την πρώτη κατηγορία λέξεων. 143
Πίνακας 16: Αποτελέσματα αξιολόγησης του συστήματος για την δεύτερη κατηγορία λέξεων. . 144
Πίνακας 17: Αποτελέσματα αξιολόγησης του συστήματος για την τρίτη κατηγορία λέξεων. 145
Πίνακας 18: Η κλίμακα βαθμολόγησης και τα αποτελέσματα του πειράματος για την αξιολόγηση σε επίπεδο πρότασης. 159
Πίνακας 19: Αποτελέσματα της ακουστικής αξιολόγησης από τους χρήστες αναφορικά με την ποιότητα του τελικού συνθέτη φωνής. Εκτός από τηεν γενικότερη ποιότητα, εξετάστηκαν η ευκολία και ευχαρίστηση ακρόασης, η κατανόηση και η άρθρωση του συστήματος... 160

1. ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο που ακολουθεί γίνεται εισαγωγή στη διατριβή, περιγράφοντας το ερευνητικό πεδίο το οποίο πραγματεύεται. Πιο συγκεκριμένα, παρουσιάζεται η ερευνητική συνεισφορά της διατριβής στα διάφορα επιμέρους επίπεδα, παρέχοντας παράλληλα μια συνοπτική περιγραφή των συνεισφορών αυτών. Παράλληλα, παρέχεται η διάρθρωση της διατριβής, δίνοντας μια περίληψη για το κάθε κεφάλαιο που ακολουθεί.

1.2 Ερευνητική συνεισφορά της διατριβής

Η ερευνητική συνεισφορά της παρούσας διδακτορικής διατριβής κινείται σε τέσσερεις άξονες. Αρχικά, προτείνεται μία καινοτόμος μέθοδος για την μοντελοποίηση του επιτονισμού στο πλαίσιο της σύνθεσης φωνής. Νέες καινοτόμες τεχνικές και μέθοδοι προτείνονται για τον σχεδιασμό και την ανάπτυξη της βάσης δεδομένων του συστήματος σύνθεσης φωνής. Παράλληλα παρουσιάζουμε ειδικά θέματα αλλά και μια νέα προσέγγιση για την επεξεργασία του κειμένου σε έναν συνθέτη φωνής για τα Ελληνικά, εστιάζοντας κυρίως στο σύστημα φωνητικής μεταγραφής και στο φαινόμενο των Greeklish. Τέλος, περιγράφονται όλες οι καινοτομίες και ειδικές προσαρμογές που σχεδιάσαμε και ενσωματώσαμε στο σύστημα σύνθεσης φωνής για την χρήση του ως εργαλείο σε πλατφόρμα προσβασιμότητας για άτομα με προβλήματα.

1.2.1 Μοντελοποίηση της προσωδίας στην σύνθεση φωνής

Το πεδίο της μοντελοποίησης της προσωδίας είναι ένα ανοικτό ερευνητικά πεδίο, με πολλές διαφορετικές προσεγγίσεις και μεθοδολογίες [Dimou & Chalamandaris, 2006]. Στο πλαίσιο της συγκεκριμένης διατριβής προτείνεται ένας νέος ουσιαστικά αλγόριθμος ο οποίος επιχειρεί την μοντελοποίηση της προσωδίας εμμέσως, χωρίς την χρήση κάποιου ρητού μοντέλου προσωδίας, όπως περιγράφεται σε άλλα συστήματα σύνθεσης φωνής. Πιο συγκεκριμένα, η μέθοδος που προτείνουμε κάνει χρήση της συλλαβής ως ελάχιστη ακουστική μονάδα και στην συνέχεια επιχειρεί να εντοπίσει διαφορετικά προσωδιακά πρότυπα. Τα πρότυπα αυτά λαμβάνονται υπόψη κατά την επιλογή βέλτιστων ακουστικών μονάδων, επιτυγχάνοντας μία έμμεση μοντελοποίηση της επιθυμητής προσωδίας. Βασικό χαρακτηριστικό της προτεινόμενης μεθόδου είναι η διατήρηση της μικροπροσωδίας κατά την εφαρμογή της τελικής καμπύλης, γεγονός που διατηρεί σε υψηλά επίπεδα την ποικιλότητα και φυσικότητα του τελικού συνθετικού σήματος, όπως άλλωστε εξήχθει ως συμπέρασμα από τα πειράματα αξιολόγησης που πραγματοποιήσαμε στα πλαίσια της έρευνάς μας. [Giannopoulos et al., 2003]

1.2.2 Σχεδιασμός και ανάπτυξη της βάσης δεδομένων ενός συστήματος σύνθεσης φωνής

Μία από τις κυριότερες συνεισφορές της συγκεκριμένης διατριβής εστιάζεται στην διαδικασία σχεδιασμού και υλοποίησης της βάσης δεδομένων ενός συστήματος σύνθεσης φωνής. Οι βασικές συνιστώσες της μεθόδου μας στο συγκεκριμένο πεδίο είναι τρεις: α) η διαδικασία και ο αλγόριθμος σχεδιασμού του σώματος κειμένου προς ηχογράφηση, β) η ψηφιακή επεξεργασία και

επισημείωση των ακουστικών αρχείων και γ) η αυτοματοποιημένη σταχυολόγηση της βάσης δεδομένων για την άμεση ενσωμάτωσή της σε ένα σύστημα σύνθεσης φωνής, χωρίς την χειρονακτική διόρθωσή της [Chalamandaris et al., 2009a].

Η μέθοδος σχεδιασμού του σώματος κειμένου προς ηχογράφηση βασίζεται σε έναν απλήστο αλγόριθμο επιλογής, και ολοκληρώνεται με επιπλέον στάδια σταχυολόγησης και βελτίωσης του τελικού αποτελέσματος με βάση το υποσύστημα βέλτιστης επιλογής ακουστικών μονάδων. Η προσέγγιση αυτή ουσιαστικά οδηγεί στην δημιουργία ενός σώματος κειμένου συμπληρωματικού ως προς τις ιδιαιτερότητες του υποσυστήματος βέλτιστης επιλογής ακουστικών μονάδων, γεγονός που ουσιαστικά αντισταθμίζει σε μεγάλο βαθμό το ανεπίλυτο ακόμη και σήμερα πρόβλημα της βέλτιστης επιλογής βαρών στην συνάρτηση κόστους του αλγορίθμου επιλογής [Chalamandaris et al., 2011] [Tsiakoulis et al., 2008] και [Karabetsos et al., 2009]. Η ύπαρξη 2 επιπλέον σταδίων στην διαδικασία σχεδιασμού και ηχογράφησης του σώματος κειμένου για τον συνθέτη φωνής, διασφαλίζει περαιτέρω την πληρότητα του σώματος κειμένου και της τελικής βάσης δεδομένων, αντιμετωπίζοντας παράγοντες που έχουν να κάνουν με τις ιδιαιτερότητες της φωνής του ομιλητή, αλλά και τις συνθήκες ηχογράφησης [Founda et al., 2001a] και [Founda et al., 2001b]. Σημαντικό κομμάτι της προτεινόμενης διαδικασίας σχεδιασμού και ανάπτυξης της βάσης δεδομένων του συστήματος σύνθεσης φωνής, αποτελεί τόσο η ψηφιακή επεξεργασία των ηχογραφήσεων, όσο και το στάδιο του αυτόματου τεμαχισμού και επισημείωσης των ηχογραφήσεων αυτών. Τέλος, η συγκεκριμένη διατριβή προτείνει έναν καινοτόμο μηχανισμό για την αυτοματοποιημένη σταχυολόγηση των στοιχείων της βάσης δεδομένων, ούτως ώστε να αντιμετωπίζονται χωρίς ουσιαστική επίβλεψη προβλήματα που έχουν προκύψει από σφάλματα ή ανακριβή αποτελέσματα κατά την αυτόματη κατάτμηση των ηχογραφήσεων. Σημαντική συνεισφορά στην επεξεργασία των ηχογραφήσεων για την βάση δεδομένων του συστήματος αποτελεί και ο αλγόριθμος που αναπτύξαμε για την αποτελεσματική επισημείωση των σημείων ανάλυσης pitchmarks για την παράθεση στο πεδίο του χρόνου. [Chalamandaris et al., 2009b]

1.2.3 Ειδικά θέματα επεξεργασίας κειμένου για τον συνθέτη φωνής

Η συγκεκριμένη διατριβή πραγματεύεται επίσης θέματα που άπτονται της επεξεργασίας κειμένου και ειδικότερα στο πλαίσιο ανάπτυξης ενός συστήματος σύνθεσης φωνής για τα Ελληνικά [Protopapas et al., 2010]. Πιο συγκεκριμένα, μία από τις σημαντικότερες συνεισφορές της

διατριβής αποτελεί η ανάπτυξη ενός αυτοματοποιημένου συστήματος φωνητικής μεταγραφής για τα Ελληνικά. Βασιζόμενο σε έναν data-driven αλγόριθμο, ο συγκεκριμένος μηχανισμός κάνει χρήση απλών κειμενικών πόρων για να δημιουργήσει ένα σύνολο από κανόνες μεταγραφής από γραφήματα της Ελληνικής σε αντίστοιχα φωνήματα, προσφέροντας αποτελέσματα υψηλής ακρίβειας και ποιότητας [Chalamandaris et al., 2005]. Παράλληλα, προτείνεται και παορυσιάζεται ένας εξειδικευμένος αλγόριθμος για την αντιμετώπιση του φαινομένου των Greeklish και την αυτόματη μεταγραφή κειμένων που είναι γραμμένα σε ένα οποιοδήποτε τύπο Greeklish σε ορθά Ελληνικά, με ακρίβεια που ξεπερνάει το 98% [Chalamandaris et al., 2004a] [Chalamandaris et al., 2004β] και [Chalamandaris et al., 2006].

1.2.4 Ειδικά θέματα σχεδιασμού και προσαρμογής συστήματος σύνθεσης φωνής για χρήση σε προσβάσιμα εργαλεία

Τέλος, μία σημαντική συνεισφορά της συγκεκριμένης διατριβής αποτελεί και η συστηματική αντιμετώπιση θεμάτων που άπτονται της λειτουργίας του συστήματος σύνθεσης φωνής ως εργαλείου υποβοήθησης και επαύξησης της προσβασιμότητας για άτομα με μειωμένη όραση [Chalamandaris et al., 2010]. Σημαντικό ρόλο στην συγκεκριμένη διαδικασία διαδραμάτισε τόσο το στάδιο της καταγραφής των αναγκών χρήστη, όσο και το στάδιο της αξιολόγησης του τελικού συστήματος σύνθεσης φωνής και της επικύρωσης των αποτελεσμάτων, μέσω ειδικών μηχανισμών. Ως επιπλέον επικύρωση της συγκεκριμένης συνεισφοράς προστίθεται και η άμεση υιοθέτηση του τελικού συστήματος από ενδιαφερόμενες ομάδες χρηστών λογισμικών προσβασιμότητας, αλλά και οι πολύ θετικές εντυπώσεις που αναφέρουν σχετικά [Chalamandaris et al., 2009c].

1.3 Διάρθρωση της διατριβής

Η διάρθρωση της διατριβής είναι οργανωμένη σε κεφάλαια τα οποια παρουσιάζουν αναλυτικά τα ειδικά θέματα σύνθεσης φωνής τα οποία μελετήσαμε στο πλαίσιο της συγκεκριμένης έρευνας. Πιο συγκεκριμένα, στο κεφάλαιο 2 παρουσιάζεται η καταγραφή του πεδίου σύνθεσης φωνής, με ιστορική αναδρομή στο συγκεκριμένο πεδίο, αλλά και παρουσίαση και κατηγοριοποίηση των διαφορετικών τεχνολογιών και μεθόδων σύνθεσης φωνής που έχουν αναπτυχθεί μέχρι σήμερα.

Στο κεφάλαιο 3 παρέχεται το θεωρητικό υπόβαθρο για την μελέτη της προσωδίας στην σύνθεση φωνής, παρουσιάζονται οι κυριότερες προσεγγίσεις για την μοντελοποίησης της προσωδίας στην

σύνθεση φωνής, καθώς επίσης και η δική μας προσέγγιση η οποία βασίζεται σε μία δημιουργική μέθοδο έμμεσης μοντελοποίησης.

Στο κεφάλαιο 4 παρουσιάζουμε τις βασικές αρχές σχεδιασμού που διέπουν την βάση δεδομένων για έναν συνθέτη φωνής, εμβαθύνοντας ιδιαίτερα στον σχεδιασμό του σώματος κειμένου για την δημιουργία της βάσης, καθώς επίσης και στην ψηφιακή επεξεργασία του ηχογραφημένου σήματος φωνής, ούτως ώστε αυτό να ενσωματωθεί απρόσκοπτα στην τελική βάση δεδομένων του συστήματος σύνθεσης φωνής.

Στο κεφάλαιο 5 παρουσιάζουμε τις βασικές αρχές λειτουργίας του υποσυστήματος της επεξεργασίας κειμένου του συστήματος σύνθεσης φωνής, εστιάζοντας κυρίως στο σύστημα της φωνητικής μεταγραφής για τα Ελληνικά, το οποίο εύκολα μπορεί να προσαρμοστεί σε άλλες γλώσσες με την χρήση των κατάλληλων πόρων ως πρωτογεννές υλικό εκπαίδευσης του συστήματος.

Στο κεφάλαιο 6 παρουσιάζουμε μία ειδική κατηγορία γραπτών κειμένων που απαιτούν ιδιαίτερη επεξεργασία για την ορθή μετατροπή τους σε φωνή, τα Greeklish. Αφού περιγράψουμε το φαινόμενο και την μελέτη της βιβλιογραφίας, παρουσιάζουμε την προσέγγισή μας για την αποτελεσματική αντιμετώπιση του συγκεκριμένου φαινομένου, η οποία έχει οδηγήσει στην ανάπτυξη του αποτελεσματικότερου αλγορίθμου για την ορθή μετατροπή από Greeklish σε Ελληνικά.

Στο 7^{ο} κεφάλαιο παρουσιάζουμε την μελέτη που πραγματοποιήσαμε για την αποτελεσματική προσαρμογή του συστήματος σύνθεσης φωνής σε περιβάλλοντα ανάγνωσης οθόνης. Το συγκεκριμένο κεφάλαιο περιγράφει τα στάδια μελέτης, σχεδιασμού, ανάπτυξης και επαλήθευσης της προσαρμογής του συνθέτη φωνής, με έμφαση στις ανάγκες που παρουσιάζουν οι χρήστες με μειωμένη ή καθόλου όραση.

Τέλος, στο κεφάλαιο 8 παρουσιάζουμε τα συμπεράσματα της διατριβής, τα θέματα προς συζήτηση και μελλοντική διερεύνηση, ενώ παράλληλα πραγματοποιείται συνοπτική αναφορά στα ειδικότερα θέματα στα οποία συνεισφέρει η συγκεκριμένη διατριβή.

1.4 Βιβλιογραφία Κεφαλαίου

[Chalamandaris et al., 2004a] A. Chalamandaris, P. Tsiakoulis, S. Raptis, G. Giannopoulos and G. Carayannis, "Bypassing Greeklish!", in Proc. LREC 2004: 4th International Conference on Language Resources And Evaluation, May 26-28, Lisbon, Portugal, 2004

[Chalamandaris et al., 2004b] A. Chalamandaris, P. Tsiakoulis, S. Raptis and G. Giannopoulos, "An Efficient and Robust Algorithm for Bypassing Greeklish", in Proc. IC-SCCE 2004: 1st International Conference from Scientific Computing to Computational Engineering, 8-10 September, Athens, Greece, 2004

[Chalamandaris et al., 2005] A. Chalamandaris, S. Raptis and P. Tsiakoulis, "Rule-based grapheme-to-phoneme method for the Greek", in Proc. Interspeech'2005: 9th European Conference on Speech Communication and Technology, September 4-8, Lisbon, Portugal, 2005

[Chalamandaris et al., 2006] A. Chalamandaris, A. Protopapas, P. Tsiakoulis and S. Raptis, "All Greek to me! An automatic Greeklish to Greek Transliteration System," 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 24–26 May, 2006

[Chalamandaris et al., 2009a] A. Chalamandaris, P. Tsiakoulis, S. Raptis, and Sotiris Karabetsos, "Design of an Efficient Corpus for High-Quality Unit Selection TTS for Bulgarian", in Proc. 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poland, 2009

[Chalamandaris et al., 2009b] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and Spyros Raptis, "An Efficient and Robust Pitch Marking Algorithm on the Speech Waveform for TD-PSOLA", in Proc. Intl. IEEE Conference on Signal and Image Processing Applications (ICSIPA), Malaysia, 2009

[Chalamandaris et al., 2009c] A. Chalamandaris, S. Raptis, P. Tsiakoulis, and S. Karabetsos, "Enhancing Accessibility of Web Content for the Print-Impaired and Blind People", in USAB2009: Human-computer interaction for eInclusion, A. Holzinger and K. Miesenberger (Eds.), Lecture Notes in Computer Science, Springer, 2009

[Chalamandaris et al., 2010] A. Chalamandaris, S. Karabetsos, P. Tsiakoulis, S. Raptis, "A Unit Selection Text-to-Speech Synthesis System Optimized for Use with Screen Readers", IEEE Transactions on Consumer Electronics, Vol. 56, No. 3, pp. 1890-1897, August, 2010.

[Chalamandaris et al., 2011] Chalamandaris, P. Tsiakoulis, S. Raptis, S. Karabetsos, "Corpus Design for a Unit Selection TtS System with Application to Bulgarian" in Human Language Technology. Challenges for Computer Science and Linguistics, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011

[Dimou & Chalamandaris, 2006] A. L. Dimou and A. Chalamandaris, "Language identification from suprasegmental cues; examining the role of rhythm in the identification of a Greek dialect", in Proc. La comunicazione parlata / Spoken Communication, February, , Italy, 2006

[Dimou & Chalamandaris, 2008] A. L. Dimou and A. Chalamandaris, "Is idiom identification possible from prosodic information? An experimental approach for the Greek language", in Proc. 4th Intl Conf. Speech Prosody 2008, pp. 759-762 (2008)

[Founda et al., 2001a] M. Founda, A. Chalamandaris, G. Tambouratzis, and G. Carayannis, "Reducing Spectral Mismatches in Concatenative Speech Synthesis via Systematic Database Enrichment", in Proceedings of the Eurospeech-2001 Conference, Aalborg, Denmark, 4-7 September 2001, pp. 837-840.

[Founda et al., 2001b] M. Founda, A. Chalamandaris, G. Tambouratzis, and G. Carayannis, "Studying the Factors Affecting the Optimal Unit Selection Algorithm for a TTS System for the Greek Language", in Proceedings of the 4th European Conference on Noise Control EURONOISE2001, Patra, 14-17 January 2001, Vol. II, pp. 758-764.

[Giannopoulos et al., 2003] G. Giannopoulos, A. Chalamandaris, S-E. Fotinea, T. Athanaselis, G. Carayannis, "Analysis and modelling of the Carrier Declination for the Greek language", in Proc. of the 15th

International Congress of Phonetic Sciences - ICPhS03, 3-9 August 2003, Barcelona, pp. 555-558.

[Karabetsos et al., 2008] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, and S. Raptis, "HMM-based Speech Synthesis for the Greek Language," in Petr Sojka, Ivan Kopecek, and Karel Pala (eds.), Lecture Notes in Computer Science (LNCS), Springer – Verlag, 2008

[Karabetsos et al., 2009] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, S. Raptis, "Embedded Unit Selection Text-to-Speech Synthesis for Mobile Devices", IEEE Transactions on Consumer Electronics, Issue 2, Vol. 56, May 2009.

[Karabetsos et al., 2010] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, S. Raptis, "One-Class Classification for Spectral Join Cost Calculation in Unit Selection Speech Synthesis", IEEE Signal Processing Letters, Vol. 17, No. 8, pp. 746-749, August, 2010

[Protopapas et al., 2010] A. Protopapas, M. Tzakosta, A. Chalamandaris and P. Tsiakoulis, "IPLR: an online resource for Greek word-level and sublexical information", Language Resources & Evaluation, Springer, 2010.

[Raptis & Chalamandaris, 2005] S. Raptis and A. Chalamandaris, "IMUTUS - Interactive Music Tuition System", The 5th Open MusicNetwork Workshop, July 4-5, Vienna, Austria, 2005

[Raptis et al., 2005b] S. Raptis, A. Askenfelt, D. Fober, A. Chalamandaris, E. Schoonderwaldt, S. Letz, A. Baxevanis, K. Falkenberg Hansen and Y. Orlarey, "IMUTUS – An Effective Practicing Environment For Music Tuition", International Computer Music Conference (ICMC 2005), September 5-9, Barcelona, Spain, 2005

[Raptis et al., 2009a] S. Raptis, P. Tsiakoulis, A. Chalamandaris and S. Karabetsos, "High Quality Unit-Selection Speech Synthesis for Bulgarian", In Proc. 13th International Conference on Speech and Computer (SPECOM'2009), St. Petersburg, Russia, June 21-25, 2009

[Raptis et al., 2009b] S. Raptis, P. Tsiakoulis, A. Chalamandaris and S. Karabetsos, "User Interaction Design for a Home-Based Telecare System", in USAB2009: Human-computer interaction for eInclusion, A. Holzinger and K. Miesenberger (Eds.), Lecture Notes in Computer Science, Springer, 2009

[Raptis et al., 2010] S. Raptis, A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, "The ILSP Text-to-Speech System for the Blizzard Challenge 2010", In Proc. Blizzard Challenge 2010 Workshop, Kyoto, Japan, September 25, 2010

[Tsiakoulis et al., 2005] P. Tsiakoulis, S. Karabetsos, S. E. Fotinea and I. Dologlou, "Spectral Estimation for Speech Signals based on Decimation and Eigenanalysis", in Proc. HERCMA-2005 (7th Hellenic European Conf. on Computer Mathematics & its Applications), Athens, Greece, September, 2005

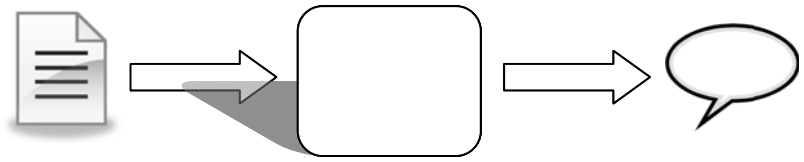
[Tsiakoulis et al., 2008] P. Tsiakoulis, A. Chalamandaris, S. Karabetsos and S. Raptis, "A Statistical Method for Database Reduction for Embedded Unit Selection Speech Synthesis," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Las Vegas, USA, 2008

2. ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΤΕΧΝΟΛΟΓΙΑ ΣΥΝΘΕΣΗΣ ΦΩΝΗΣ

Στο κεφάλαιο που ακολουθεί γίνεται εισαγωγή στην τεχνολογία σύνθεσης φωνής. Πραγματοποιείται ιστορική αναδρομή στην τεχνολογία του συνθετικού λόγου, ενώ παρουσιάζονται συνοπτικά και με την κατάλληλη κατηγοριοποίηση οι διάφορες τεχνικές δημιουργίας συνθετικού λόγου και οι κυριότερες διαφορετικές προσεγγίσεις που οδήγησαν σε ποικίλες τεχνολογίες που έχουν αναπτυχθεί. Από αυτές η τεχνολογία σύνθεσης φωνής με παράθεση ακουστικών μονάδων και επεξεργασία σήματος στο πεδίο του χρόνου παρουσιάζεται αναλυτικότερα, αφού αποτελεί και την βάση για το σύστημα στο οποίο η συγκεκριμένη διατριβή αναφέρεται. Το συγκεκριμένο κεφάλαιο θέτει τις βάσεις για την παρουσίαση και επεξήγηση του υπολοίπου της διατριβής, παρουσιάζοντας τις βασικές πτυχές του συγκεκριμένου επιστημονικού πεδίου που πρόκειται να αναλυθεί με λεπτομέρεια, ενώ παρουσιάζει με ευκρίνεια τους στόχους και την συνεισφορά της συγκεκριμένης εργασίας.

2.1 Τι είναι η σύνθεση φωνής

Ο λόγος είναι το σημαντικότερο μέσο επικοινωνίας μεταξύ των ανθρώπων. Η σύνθεση φωνής, δηλαδή η αυτόματη παραγωγή σήματος φωνής, αποτελεί αντικείμενο μελέτης εδώ και αρκετές δεκαετίες. Ειδικά τα τελευταία χρόνια, έχει πραγματοποιηθεί μεγάλη πρόοδος στη σύνθεση φωνής, με τη δημιουργία συνθετών που χαρακτηρίζονται από υψηλή καταληπτότητα. Ένας συνθέτης φωνής από κείμενο, (διεθνώς γνωστός με την ονομασία TTS synthesizer, από τα αρχικά της αντίστοιχης αγγλικής ορολογίας Text-To-Speech synthesizer), είναι ένα σύστημα που λειτουργεί με τη χρήση υπολογιστή, και μπορεί να διαβάζει οποιοδήποτε κείμενο στη συγκεκριμένη γλώσσα για την οποία το σύστημα αναπτύχθηκε. Θεωρώντας το κείμενο ως ένα σήμα εισόδου στο σύστημα TTS, η έξοδος είναι συνθετική ομιλία.



Σχήμα 1: Αναπαράσταση της λειτουργίας ενός απλού συνθέτη φωνής ως ένα σύστημα με είσοδο το κείμενο και έξοδο την συνθετική φωνή.

Η παραπάνω περιγραφή αν και συνοπτική υποδηλώνει τις δυσκολίες που ένα σύστημα συνθετικής ομιλίας έχει να αντιμετωπίσει. Ένας μέσος άνθρωπος, έχοντας αποκτήσει την απαραίτητη γνώση, είναι σε θέση να αναγνωρίσει ακρωνύμια και αριθμούς, να προβλέψει την σωστή κλίση ενός αριθμού, να αναγνωρίσει σημαντικά σημεία, λέξεις και συνδυασμούς αυτών μέσα σε ένα γραπτό κείμενο, τα οποία θα του επιτρέψουν να διαβάσει το συγκεκριμένο κείμενο με τον απαραίτητο χρωματισμό, μελωδία, να προσθέσει έμφαση, να συνδέσει λέξεις μεταξύ τους για να διευκολύνει την κατανόηση,. χωρίς ωστόσο το κείμενο το ίδιο να περιέχει ειδικές επισημειώσεις αναφορικά με τα παραπάνω ιδιαίτερα χαρακτηριστικά της ανάγνωσής του. Το συγκεκριμένο γεγονός γίνεται ακόμη πιο ξεκάθαρο αν αναλογιστεί κανείς τα διαφορετικά είδη κειμένων που μπορεί ένας αναγνώστης να συναντήσει και τους διαφορετικούς τρόπους ανάγνωσης που θα έπρεπε να υιοθετήσει για το καθένα από αυτά [Duggan2003].

Η σύνθεση φωνής έχει να αντιμετωπίσει το σημαντικό πρόβλημα της μίμησης του ανθρώπου ως προς τον τρόπο που ο τελευταίος ουσιαστικά «ερμηνεύει» ένα γραπτό κείμενο, όπως ακριβώς ένας

μουσικός αναλαμβάνει να ερμηνεύσει μία παρτιτούρα. Αυτό είναι ένα πρόβλημα που προσπαθεί να προσεγγίσει η σημερινή τεχνολογία σύνθεσης φωνής. Ωστόσο, αυτό που τα περισσότερα συστήματα σύνθεσης φωνής έχουν σχεδιασθεί να εκτελούν είναι η ανάγνωση κειμένων με έναν ουδέτερο και καταφατικό τρόπο, ποσοτικοποιώντας με συμβατικούς τρόπους τα ειδικά αυτά χαρακτηριστικά του κειμένου που θα επέτρεπαν μία πιο φυσική και συχνά συναισθηματική εκφορά αυτού.

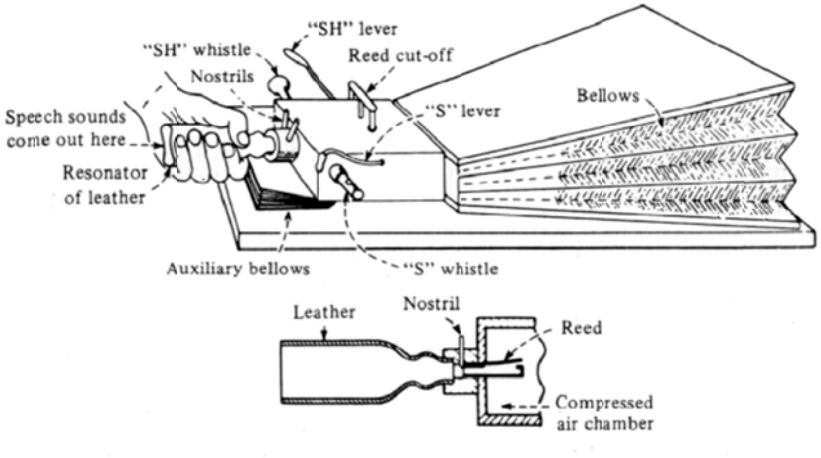
Έχοντας κανείς υπόψη του τα παραπάνω μπορεί να συμπεράνει τα τρία βασικά πεδία όπου ένα σύστημα σύνθεσης φωνής πρέπει να επιτύχει: α) η ικανότητα να επεξεργάζεται με αποτελεσματικότητα οποιοδήποτε κείμενο, β) η ικανότητα να παράγει κατανοητό λόγο και γ) η ικανότητα να παράγει λόγο με φυσικότητα [Dutoit1997]. Ειδικά το χαρακτηριστικό της φυσικότητας του λόγου είναι δύσκολο να προσδιοριστεί και για τον λόγο αυτόν συχνά συμπίπτει με τον όρο ευχαρίστηση (pleasantness) που λαμβάνει ο ακροατής. Τα πεδία αυτά παρουσιάζονται αναλυτικότερα στο κεφάλαιο της αξιολόγησης του συστήματος συνθετικής ομιλίας (κεφάλαιο 6) όπου παρουσιάζεται η διαδικασία αξιολόγησης για τον συνθέτη φωνής που αναπτύξαμε, ενώ πραγματοποιείται επιπλέον ειδική μελέτη για την σύμπραξη του συστήματος σύνθεσης φωνής με εργαλεία υποστήριξης ατόμων με προβλήματα όρασης. Ωστόσο είναι σημαντικό να αναφερθούν οι διαφορετικές αυτές πτυχές στο συγκεκριμένο σημείο της διατριβής, ώστε να επισημανθεί το μέγεθος των απαιτήσεων που υπάρχουν από ένα τέτοιο σύστημα [Bailly2003].

Σήμερα, αν και η τεχνολογία συνθετικής ομιλίας έχει να επιδείξει αλματώδη πρόοδο με αποτελέσματα που χαρακτηρίζονται αποδεκτά, παρουσιάζει σημαντικούς περιορισμούς, στην τελική ποιότητα του συνθετικού σήματος φωνής για τις εφαρμογές που στοχεύονται. Θα μπορούσε να πει κανείς ότι η ποιότητα που επιτυγχάνεται όσον αφορά την φυσικότητα ενός συνθέτη φωνής, είναι συχνά αντιστρόφως ανάλογη με το εύρος των εφαρμογών που μπορεί να εξυπηρετήσει το συγκεκριμένο σύστημα. Το ίδιο ισχύει και αναφορικά με το εύρος που παρουσιάζει μία εφαρμογή. Αυτό με άλλα λόγια σημαίνει ότι όσο πιο περιορισμένο είναι το πεδίο εφαρμογής τόσο πιθανότερο είναι το σύστημα να παρουσιάζει υψηλότερη ποιότητα στο τελικό αποτέλεσμα. Είναι επόμενο λοιπόν ένας συνθέτης φωνής για την αναγγελία της ώρας να παρουσιάζει υψηλότερη ποιότητα και συνέπεια από έναν συνθέτη φωνής για την εκφώνηση

ειδήσεων. Όλες αυτές οι παράμετροι καθορίζουν σε μεγάλο βαθμό τόσο τον σχεδιασμό, όσο και την λειτουργία του συνθέτη και των επιμέρους υποσυστημάτων του.

2.2 Ιστορική αναδρομή της σύνθεσης φωνής

Αν και η τεχνολογία της σύνθεσης φωνής έχει εξελιχθεί κυρίως τα τελευταία χρόνια, η ιδέα προσδιορίζεται στο μακρινό παρελθόν, με τις πρώτες αξιοσημείωτες προσπάθειες κοντά στα τέλη του 18ου αιώνα (1779) όπου ο Δανός Christian Kratzenstein δημιούργησε κατασκευές με σωλήνες ώστε να μοντελοποιήσει τον φωνητικό σωλήνα για τα 5 κύρια φωνήεντα. Το 1791 ο Αυστριακός Wolfgang von Kempelen δημιούργησε μία κατασκευή η οποία είχε την δυνατότητα αναπαραγωγής φωνηέντων και συμφώνων με τη χρήση διαφορετικών μοντέλων της γλώσσας και των χειλιών.



Σχήμα 2: Αναπαράσταση της μηχανής σύνθεσης ήχων του Wolfgang von Kempelen.

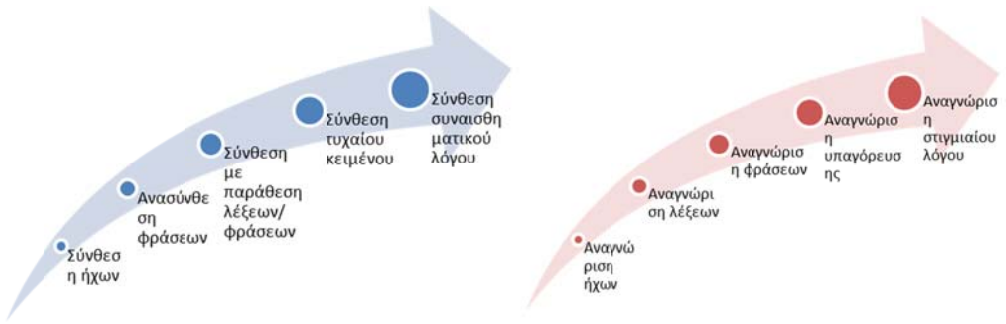
Ακολούθησαν πολλές άλλες κατασκευές όπως η «ομιλούσα μηχανή» του Charles Wheatstone το 1837 η οποία βασιζόταν στην μηχανή του von Kempelen και η "Euphonia" του M. Faber το 1857, μέχρι την πρώτη ηλεκτρονική συσκευή σύνθεσης φωνής που αναπτύχθηκε στα Bell Labs της Αμερικής και ονομάστηκε VOCODER[1] το 1939 [Dudley1939]. Η μηχανή αυτή με την χρήση

[1] Στόχος των VOCODER (VOice COder and DEcodeR) ήταν η συμπίεση της φωνής, ωστόσο έδωσαν μεγάλη ώθηση στην τεχνολογία συνθετικής ομιλίας.

πλήκτρων επέτρεπε την σύνθεση ιδιαίτερα καταληπτής φωνής. Ο πρώτος συνθέτης φωνής για υπολογιστή σχεδιάσθηκε και αναπτύχθηκε στα Bell Labs στις αρχές της δεκαετίας του 60, ο οποίος και σημείωσε την απαρχή ενός νέου επιστημονικού πεδίου και οδήγησε στην ανάπτυξη των πρώτων ολοκληρωμένων συστημάτων σύνθεσης φωνής μέσω υπολογιστή. Στα τέλη της δεκαετίας του 1940 ο Franklin S. Cooper με τους συνεργάτες του στα Haskins Laboratories δημιούργησαν την μηχανή "Pattern Playback" η οποία επιχειρούσε να μετατρέψει εικόνες από ακουστικά πρότυπα κωδικοποιημένα μέσω της φασματικής αναπαράστασης σε ήχους. Μέσω αυτής της συσκευής αργότερα ο Alvin Liberman κατάφερε να ανακαλύψει την σημασία των των φωνημάτων ως μοναδιαίων ακουστικών μονάδων στην αντίληψη του ήχου της φωνής. Αργότερα, στις δεκαετίες των 1980 και 1990 υπήρξαν τα συστήματα MITalk και KlatTalk του Dennis Klatt [Klatt1980], όπως επίσης και το σύστημα της Bell Labs.

Οι πρώτες προσπάθειες είχαν ως στόχο την μοντελοποίηση είτε του ανθρώπινου συστήματος παραγωγής φωνής σε φυσικό επίπεδο, είτε του ίδιου του σήματος της φωνής σε ακουστικό επίπεδο. Η μοντελοποίηση του συστήματος παραγωγής είχε ως αποτέλεσμα τους συνθέτες φωνής με αρθρωτικά μοντέλα, ενώ η μοντελοποίηση του ακουστικού σήματος φωνής οδήγησε στους formant και LPC συνθέτες φωνής. Με την πρόοδο της τεχνολογίας και την αύξηση της υπολογιστικής ισχύος, αναπτύχθηκαν συστήματα σύνθεσης φωνής που βασίζονται στην χρήση πραγματικών σημάτων φωνής και την αντιγραφή ή παράθεση των τμημάτων αυτών με διάφορες τεχνικές. Η τρέχουσα τεχνολογική τάση βασίζεται κυρίως στην χρήση εκτενών ηχογραφήσεων οι οποίες χρησιμοποιούνται για σύνθεση είτε με παράθεση βέλτιστων ακουστικών μονάδων, είτε με στοχαστική μοντελοποίηση μέσω Hidden Markov Models (HMM TTS) [Kim2006].

Η διαδρομή της τεχνολογίας συνθετικού λόγου, ακολουθεί μία παράλληλη πορεία με αυτήν της αναγνώρισης φωνής. Εκκινώντας από την σύνθεση και αναγνώριση βασικών ήχων, προχώρησαν στην σύνθεση και αναγνώριση λέξεων, στην σύνθεση ενός ουδέτερου και στην αναγνώριση ενός υπαγορευτικού στυλ, στοχεύοντας στην σύνθεση και αναγνώριση ενός πλήρως εκφραστικού ή στιγμιαίου λόγου αντίστοιχα [Eide2004].

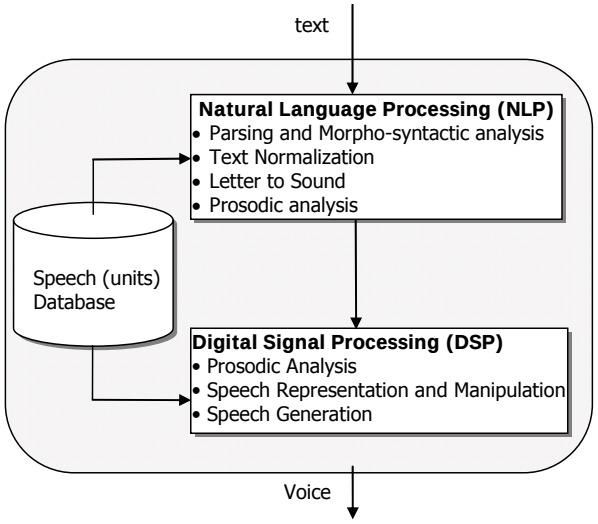


Σχήμα 3: Η εξέλιξη των τεχνολογιών σύνθεσης και αναγνώρισης φωνής, από την αρχή τους μέχρι τον τελικό τους στόχο.

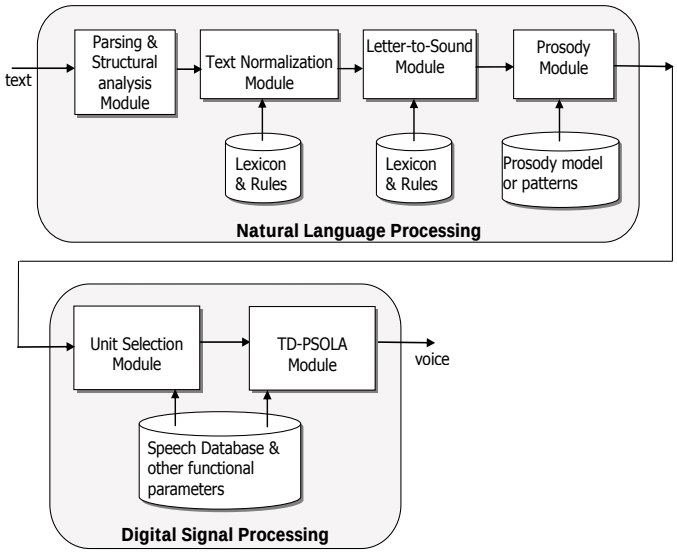
2.3 Αρχιτεκτονική ενός συνθέτη φωνής

Για να προσδιορίσουμε την αρχιτεκτονική ενός συνθέτη φωνής θα πρέπει να τονίσουμε και πάλι το πεδίο που σήμερα καλύπτει η συγκεκριμένη τεχνολογία και διαφέρει σημαντικά από την απλή παραγωγή ήχων που αρχικά επιχειρήθηκε από τους πρώτους ερευνητές του χώρου. Η ευθύνη του συνθέτη φωνής είναι να μετατρέπει ένα οποιοδήποτε κείμενο σε συνθετική φωνή με υψηλή καταληπτότητα και φυσικότητα, συμπεριλαμβάνοντας προφανώς όλες τις διεργασίες που απαιτούνται σε ενδιάμεσα στάδια, όπως είναι η κατάλληλη επεξεργασία του κειμένου μέχρι και την τελική παραγωγή του συνθετικού σήματος Rabiner[1993], [Santen1997].

Η γενική αρχιτεκτονική ενός συστήματος σύνθεσης φωνής παρουσιάζεται στο Σχήμα 4. Εκεί μπορεί κανείς να διακρίνει την ύπαρξη δύο διακριτών υποσυστημάτων και συγκεκριμένα του υποσυστήματος για την επεξεργασία φυσικής γλώσσας (*Natural Language Processing*) και του υποσυστήματος της ψηφιακής επεξεργασίας σήματος (*Digital Signal Processing*). Το συγκεκριμένο διάγραμμα παρουσιάζει την γενική αρχιτεκτονική ενός συνθέτη ομιλίας ανεξάρτητα από την διαθέσιμη τεχνολογία ανάλυσης, επεξεργασίας και σύνθεσης ψηφιακού σήματος φωνής. Ακολουθεί αναλυτικότερη περιγραφή των υποσυστημάτων αυτών.



Σχήμα 4: Αναπαράσταση της λειτουργίας ενός απλού συνθέτη φωνής.



Σχήμα 5: Αναπαράσταση και ανάδειξη των επιμέρους υποσυστημάτων ενός σύγχρονου συνθέτη φωνής.

2.3.1 Το υποσύστημα Επεξεργασίας Φυσικής Γλώσσας

Το συγκεκριμένο υποσύστημα είναι υπεύθυνο για την σάρωση, ανάλυση και μετατροπή του κειμένου εισόδου σε μια ενδιάμεση αναπαράσταση, η οποία είναι κατάλληλη για να τροφοδοτήσει

το επόμενο υποσύστημα με τις απαραίτητες πληροφορίες. Εκτός από την προεπεξεργασία του κειμένου και την κανονικοποίησή του, το συγκεκριμένο υποσύστημα είναι επίσης υπεύθυνο για την παραγωγή σημαντικών πληροφοριών οι οποίες στην συνέχεια θα τροφοδοτήσουν την μηχανή παραγωγής προσωδίας και επιτονισμού η οποία ανήκει στο DSP υποσύστημα. Συχνά περιέχει μορφοσυντακτικούς αναλυτές, ένα υποσύστημα κανονικοποίησης κειμένου, ένα υποσύστημα φωνητικής μεταγραφής και ένα αναλυτή προσωδιακών χαρακτηριστικών κειμένου. Στο πλαίσιο της συγκεκριμένης διατριβής πρόκειται να εμβαθύνουμε σε ειδικά χαρακτηριστικά αυτών των υποσυστημάτων [Bailly2003].

2.3.2 Το υποσύστημα Ψηφιακής Επεξεργασίας Σήματος

Το συγκεκριμένο υποσύστημα είναι υπεύθυνο για την κατάλληλη ανάλυση, επεξεργασία και μετατροπή του ψηφιακού σήματος φωνής για την σύνθεση του τελικού ψηφιακού σήματος συνθετικής ομιλίας. Αν το υποσύστημα Επεξεργασίας Φυσικής Γλώσσας του συστήματος αποτελεί το κοινό χαρακτηριστικό στα περισσότερα συστήματα σύνθεσης φωνής, το υποσύστημα Ψηφιακής Επεξεργασίας Σήματος αποτελεί το σημείο διαφοροποίησης των συστημάτων αυτών. Ανάλογα με τον τύπο συνθέτη, το συγκεκριμένο υποσύστημα περιλαμβάνει κοινά χρησιμοποιούμενους αλγορίθμους για την επεξεργασία και σύνθεσης φωνής, αλλά και εξειδικευμένες τεχνολογίες που σκοπό έχουν την περαιτέρω βελτίωση του τελικού συνθετικού σήματος. Στις παραγράφους που ακολουθούν παρουσιάζονται οι διαφορετικές θεωρήσεις και τεχνολογίες σύνθεσης φωνής, οι οποίες οδηγούν σε διαφορετικές υλοποιήσεις του υποσυστήματος Ψηφιακής Επεξεργασίας του Σήματος φωνής.

2.3.3 Κατηγορίες τεχνολογιών σύνθεσης φωνής

Οι τεχνολογίες σύνθεσης φωνής διαφέρουν μεταξύ τους σε δύο διαφορετικές πτυχές:

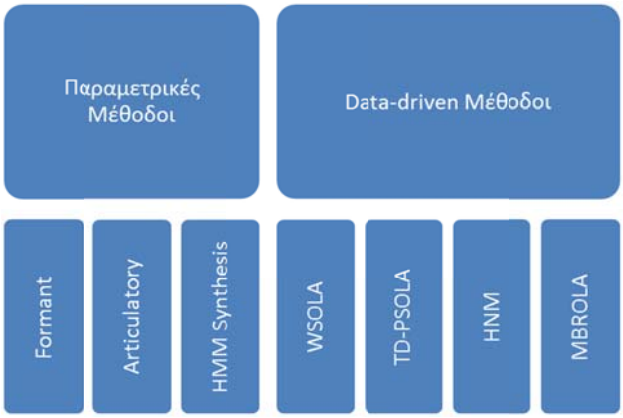
α) στην διαφορετική θεώρηση του τρόπου παραγωγής του σήματος φωνής και

β) στην προσέγγιση και στον τρόπο μοντελοποίησης και αναπαραγωγής του σήματος φωνής.

Η πρώτη διάσταση ουσιαστικά αναφέρεται στο επίπεδο το οποίο η σύνθεση φωνής επιχειρεί να μοντελοποιήσει, αν δηλαδή επιχειρεί να μοντελοποιήσει τον τρόπο παραγωγής της φωνής, ή το ίδιο το σήμα αυτής. Η δεύτερη διάσταση ουσιαστικά αποτελεί ένα διαφορετικό σημείο

διαφοροποίησης των τεχνολογιών, το οποίο ανάλογα με την απάντηση της πρώτης ερώτησης και επομένως το επίπεδο μοντελοποίησης, συναντάμε διαφορετικές μεθόδους για την μοντελοποίηση παραγωγής ή της ίδιας της ανθρώπινης φωνής. Στο σχήμα που ακολουθεί μπορεί κανείς να δει με εποπτικό τρόπο τις βασικότερες κατηγορίες.

Σχήμα 6: Κατηγοριοποίηση των τεχνολογιών σύνθεσης φωνής με βάση την θεωρία σύνθεσης και την τεχνική σύνθεσης.



Σχήμα 7: Κατηγοριοποίηση των τεχνολογιών σύνθεσης με βάση το πρωτογενές υλικό και την τεχνική επεξεργασίας σήματος.

Οι αλγόριθμοι που ανήκουν στις παραμετρικές μεθόδους, προσπαθούν να μιμηθούν τον τρόπο παραγωγής της ανθρώπινης φωνής, δημιουργώντας ένα παραμετρικό μοντέλο για την φυσική περιγραφή παραγωγής ανθρώπινης φωνής. Τέτοιες μέθοδοι είναι η σύνθεση με αρθρωτές

(Articulatory synthesis), η σύνθεση με formants, η LPC σύνθεση κ.α. Η ποιότητα των μεθόδων αυτών ακόμη βρίσκεται σε αρκετά χαμηλότερο επίπεδο από τις προηγούμενες μεθόδους, ενώ αποτελούν αντικείμενο ανοικτής έρευνας σε πολλά επίπεδα, ειδικά με την διαρκή εξέλιξη των υπολογιστικών μηχανών που επιτρέπουν την δημιουργία ολοένα και περισσότερο πολύπλοκων μοντέλων.

Οι αλγόριθμοι που επιχειρούν να μοντελοποιήσουν το ίδιο το σήμα της φωνής βασίζονται σε πλήθος δεδομένων φυσικής ομιλίας από κάποιον φυσικό ομιλητή, και προσπαθούν να συνδυάσουν τα δεδομένα αυτά με τον καλύτερο δυνατό τρόπο ώστε να αναπαράγουν ένα σήμα φωνής που θα έμοιαζε με φυσικό. Η μέθοδοι αυτές ονομάζονται data-driven μέθοδοι και μέχρι σήμερα αποτελούν την ποιοτικότερη τεχνολογία σύνθεσης φωνής στην βιομηχανία.

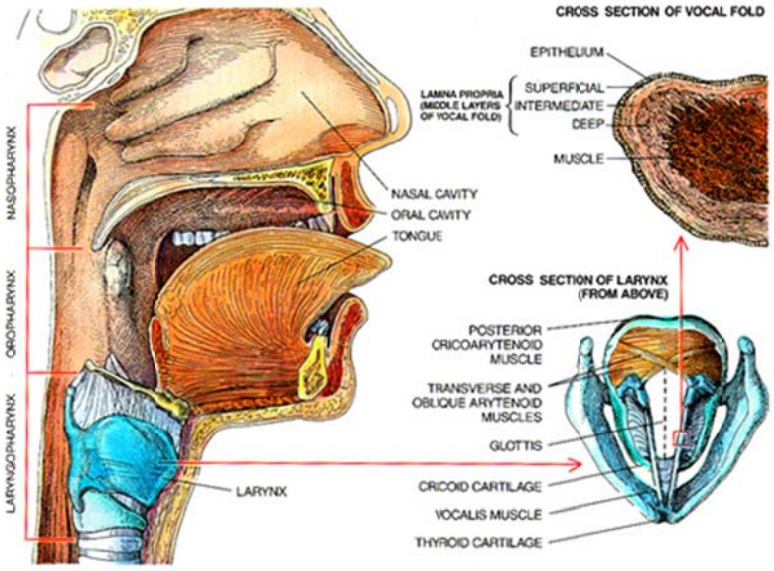
2.3.3.1 Παραμετρικές μέθοδοι σύνθεσης φωνής

2.3.3.1.1 Η μέθοδος σύνθεσης με μοντελοποίηση αρθρωτών

Η σύνθεση φωνής με αρθρωτές [Baer1991] αποτέλεσε την πρώτη προσέγγιση για σύνθεση φωνής και παραμένει ακόμη ανοικτό ερευνητικό πεδίο. Η βασική θεωρία πίσω από την συγκεκριμένη τεχνική είναι η μοντελοποίηση του συστήματος παραγωγής της ανθρώπινης φωνής σε δύο διακριτά επίπεδα:

α) στην μοντελοποίηση της φυσιολογίας του ανθρώπινου φωνητικού σωλήνα και των αρθρωτών αυτής, και

β) στην μοντελοποίηση του μηχανισμού ελέγχου αυτών.



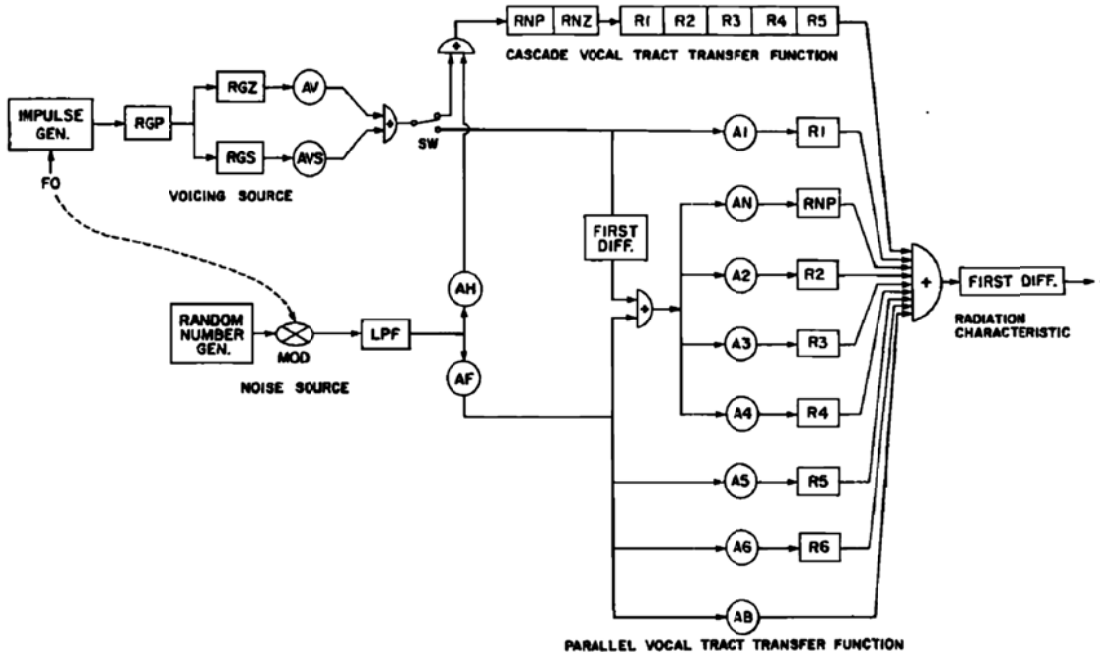
Εικόνα 1: *Το ανθρώπινο σύστημα παραγωγής φωνής. Διακρίνονται οι τρεις κύριες κοιλότητες του φωνητικού σωλήνα, φαρυγγική, ρινική και στοματική, καθώς επίσης και οι κυριότεροι αρθρωτές.*
Πηγή: http://scienceray.com/biology/anatomy-of-speech/

Η συγκεκριμένη προσέγγιση κάνει χρήση κυρίως μη γραμμικών μοντέλων για την μοντελοποίηση της παραγωγής και διάδοσης του ήχου μέσα στην φωνητική οδό, ενώ αποτελεί συχνά πεποίθηση των ερευνητών ότι η συγκεκριμένη προσέγγιση αποτελεί το μέλλον και την λύση στο πεδίο της συνθετικής φωνής. Παρ' όλα αυτά, η συγκεκριμένη προσέγγιση υποφέρει τόσο από την υπολογιστικό κόστος που έχει όσο και από τα ελλιπή μοντέλα που έχουν προταθεί μέχρι σήμερα.

2.3.3.1.2 Η μέθοδος σύνθεσης φωνής με κανόνες

Η σύνθεση φωνής με κανόνες βασίζεται στην θεώρηση που επιχειρεί να μοντελοποιήσει το σύστημα παραγωγής ανθρώπινου λόγου με ένα ευρύτερο μοντέλο πηγής-φίλτρου για το σήμα φωνής. Ο κυριότερος εκπρόσωπος της συγκεκριμένης προσέγγισης είναι η σύνθεση φωνής με formants. Η συγκεκριμένη τεχνική, αν και δεν επιχειρεί να μοντελοποιήσει την φυσιολογία του ανθρώπινου συστήματος παραγωγής φωνής, προσπαθεί να «γεννήσει» σήμα που να προσομοιάζει το σήμα της φωνής, μοντελοποιώντας τον φωνητικό σωλήνα με ένα σύνολο από φίλτρα δευτέρου βαθμού συνεδεμένα είτε σε σειρά, είτε παράλληλα, είτε με συνδυασμό και των δύο. Κάθε φίλτρο αντιπροσωπεύει ένα formant ή ένα anti-formant (για ένρινους ήχους). Στο ίδιο μοντέλο, η πηγή

του συστήματος εκπροσωπεί τις φωνητικές χορδές, των οποίων η διέγερση προσομοιάζεται με ένα περιοδικό σήμα το οποίο διαφοροποιείται ανάλογα με την φύση του ήχου που πρόκειται να παραχθεί. Η συγκεκριμένη τεχνολογία προτάθηκε από τους Holmes και Klatt στα [Klatt1980], [Olive1977], [Hanson2002].



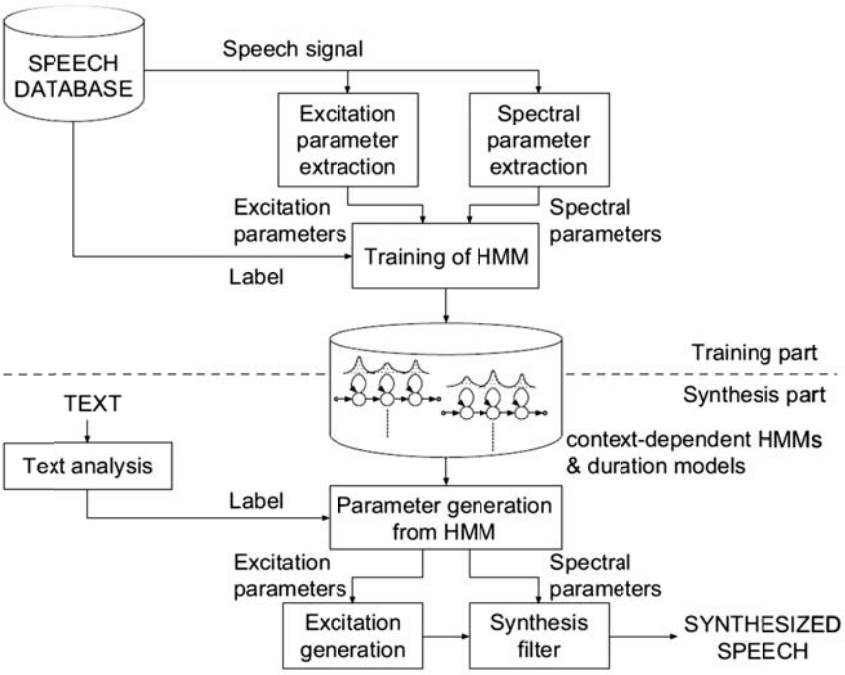
Σχήμα 8: Δομικό διάγραμμα μηχανής σύνθεσης με formants σύμφωνα με το μοντέλο του Klatt. [Klatt1976]

Η συγκεκριμένη προσέγγιση αποτέλεσε σημαντική συνεισφορά στον τομέα της σύνθεσης φωνής και ακόμη αποτελεί σημαντικό πεδίο έρευνας, αφού μεγάλη μερίδα των ερευνητών θεωρούν την συγκεκριμένη τεχνική ακόμη πολλά υποσχόμενη. Παρ' όλα αυτά, η ποιότητα του συνθετικού σήματος που παράγει είναι ακόμη σημαντικά χαμηλότερη από άλλες μεθόδους σύνθεσης φωνής και σπάνια χρησιμοποιείται σε εμπορικά συστήματα πλέον.

2.3.3.1.3 Η μέθοδος σύνθεσης φωνής με χρήση HMM

Η συγκεκριμένη μέθοδος αποτελεί μία ειδική προσέγγιση που διαφοροποιείται όχι μόνο στο πεδίο επεξεργασίας και παραγωγής του σήματος φωνής, αλλά και στο ίδιο το σύστημα της βέλτιστης επιλογής διφώνου. Η βασική ιδέα πίσω από την συγκεκριμένη τεχνική βασίζεται στην

ανάλυση και την παραμετρική αναπαράσταση της φωνής με στατιστικά μοντέλα Κρυφά Μαρκοβιανά Μοντέλα HMM (Hidden Markov Models). Αν και η συγκεκριμένη μεθοδολογία βρίσκει μεγάλη εφαρμογή στην αναγνώριση φωνής, προτάθηκε αρχικά το 1989 και αναπτύχθηκε το 1995 από τους Falaschi και Tokuda et al. αντίστοιχα [Tokuda1985], [Tokuda1995]. Το δομικό διάγραμμα ενός συστήματος σύνθεσης φωνής με HMM φαίνεται στο σχήμα 9.



Σχήμα 9: Δομικό διάγραμμα συστήματος σύνθεσης φωνής από κείμενο με χρήση HMM. Πηγή: [Zen 2007]

Αρχικά το σύστημα εκπαιδεύεται με βάση ένα ηχογραφημένο σώμα κειμένου, επισημειωμένο σε επίπεδο φωνήματος, από όπου και προκύπτουν τα μοντέλα HMM, τόσο για την φασματική όσο και για την προσωδιακή πληροφορία. Τα μοντέλα αυτά είναι εξαρτώμενα από το περικείμενο (context dependent HMM) τα οποία παρατηρούν τις παραμέτρους των καταστάσεων ως μια «λογική» σειρά ακολουθιών, όπως προκύπτει από το περιβάλλον τους. Κατά την σύνθεση, οι παράμετροι των HMM που εμπεριέχουν πληροφορία για το περικείμενο παίζουν τον ρόλο του αλγορίθμου της βέλτιστης επιλογής ακουστικών μονάδων στον συνθέτη φωνής, αφού ανάλογα με αυτές καθορίζονται και τα επιμέρους χαρακτηριστικά των HMM καταστάσεων μέσω των

κατανομών πιθανότητας. Το τελικό στάδιο της σύνθεσης περιλαμβάνει την συνένωση των διαδοχικών καταστάσεων HMM όπως έχουν προκύψει από τις κατανομές πιθανότητας, και την χρήση ενός συστήματος πηγής-φίλτρου για την γένεση του τελικού σήματος φωνής. Αν και η συγκεκριμένη τεχνολογία είναι πολλά υποσχόμενη και επιλύει σημαντικά προβλήματα υπολογιστικού κόστους σε σχέση με τις υπόλοιπες corpus-based τεχνολογίες σύνθεσης, ο συνθετικός λόγος που παράγεται υποφέρει από αφύσικη προσωδία και ρομποτική χροιά [Kim2006], [O'Shaughnessy2007].

2.3.3.2 Data-driven μέθοδοι σύνθεσης φωνής

2.3.3.2.1 Η μέθοδος σύνθεσης φωνής με παράθεση ακουστικών μονάδων

Η σύνθεση με παράθεση ακουστικών μονάδων (concatenative synthesis) βασίζεται στην κατάλληλη συρραφή τμημάτων φυσικής ομιλίας που έχει προηχογραφηθεί. Ανάλογα με την στοχευόμενη εφαρμογή αλλά και την διαθέσιμη τεχνολογία, η σύνθεση με παράθεση μπορεί να κατηγοριοποιηθεί σε τρεις βασικές κατηγορίες:

α) την σύνθεση με λέξεις,

β) την διφωνηματική σύνθεση και

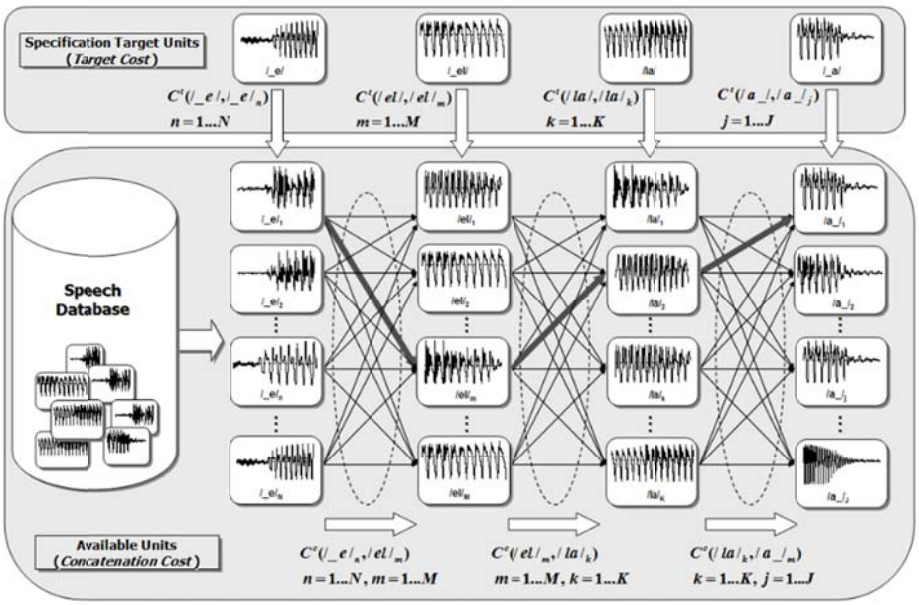
γ) την σύνθεση με βέλτιστη επιλογή ακουστικής μονάδας.

Η σύνθεση με παράθεση λέξεων βασίζεται στην παράθεση και κατάλληλη σύνδεση προηχογραφημένων λέξεων ή μεγαλύτερων φράσεων με σκοπό την δημιουργία συνθετικού λόγου περιορισμένου λεξιλογίου, που περιορίζεται ουσιαστικά στους δυνατούς συνδυασμούς λέξεων και φράσεων για τους οποίους έχει σχεδιασθεί [Fujimura1978]. Χωρίς να περιέχει εξελιγμένους αλγορίθμους επεξεργασίας ή επιλογής μονάδας, η συγκεκριμένη κατηγορία αποτελεί ακόμη και σήμερα λύση σε εξειδικευμένες περιπτώσεις όπου απαιτείται σύνθεση περιορισμένων φράσεων και λέξεων. Ωστόσο, ο εξαναγκασμός στο σύστημα συρραφής, αποκλειστικά λέξεων, δημιουργεί προσωδιακά σφάλματα και ασυνέχειες.

Σχήμα 10: Παράδειγμα σύνθεσης με παράθεση λέξεων για εκφώνηση της ώρας.

Η διφωνηματική σύνθεση φωνής βασίζεται στην χρήση διφωνημάτων (δύο διαδοχικών φθόγγων) που έχουν εξαχθεί από το σήμα φυσικής ηχογράφησης ενός ομιλητή. Αν και μπορεί να δημιουργήσει συνθετική ομιλία από οποιοδήποτε κείμενο χωρίς περιορισμό στο λεξιλόγιο ή στο πεδίο αναφοράς, παράγει συνθετική φωνή με ρομποτική χροιά [Coker2000].

Τέλος ***η σύνθεση με επιλογή βέλτιστης επιλογής ακουστικής μονάδας*** (Unit-selection Corpus-based) βασίζεται σε εκτενείς ηχογραφήσεις φυσικού ομιλητή, οι οποίες έχουν κατατμηθεί σε κατάλληλες ακουστικές μονάδες οι οποίες και αποτελούν την βάση για τον συνθέτη φωνής. Αυτές οι μονάδες μπορούν να είναι φωνήματα, διφωνήματα ή μεγαλύτερα τμήματα λόγου, οι οποίες κατά την διαδικασία της σύνθεσης επιλέγονται και παρατίθενται για να σχηματίσουν μεγαλύτερα κομμάτια ήχου. Η επιλογή τους βασίζεται σε πλήθος παραμέτρων και πραγματοποιείται μέσω μεθόδων δυναμικού προγραμματισμού.



Σχήμα 11: Παράδειγμα σύνθεσης της λέξης /έλα/ με χρήση διφωνημάτων. Στο συγκεκριμένο παράδειγμα απεικονίζεται και το υποσύστημα επιλογής βέλτιστης ακουστικής μονάδας (διφώνημα). (Πηγή: [Karabetsos2011])

Η τεχνολογία που χρησιμοποιεί το σύστημά μας είναι η σύνθεση με παράθεση βέλτιστων ακουστικών μονάδων, οι οποίες περιορίζονται σε διφωνήματα. Ωστόσο η χρήση ειδικά σχεδιασμένου μηχανισμού επιλογής βέλτιστων ακουστικών μονάδων, επιτρέπει την παράθεση μεγαλύτερων μονάδων από διφωνήματα, με αποτέλεσμα οι χρησιμοποιούμενες ακουστικές μονάδες να είναι συχνά μεταβλητού μήκους [Black1997], [Campel2005], [Hunt1996], [Vepa2004], [Beutnagel199].

2.3.3.3 Μέθοδοι ψηφιακής επεξεργασίας και σύνθεσης σήματος φωνής με παράθεση

Ανάμεσα στους αρκετούς διαφορετικούς αλγορίθμους που υπάρχουν για την ψηφιακή επεξεργασία και παραγωγή συνθετικού σήματος φωνής με παράθεση ακουστικών μονάδων, οι σημαντικότεροι που χρησιμοποιούνται σήμερα είναι η *Time Domain Pitch Synchronous Overlap Add* (TD-PSOLA), η *Harmonic plus Noise (HNM)*, η *Multiband Resynthesis Overlap Add* (MBROLA) και η WSOLA *Waveform Similarity Overap Add.*

2.3.3.3.1 Μέθοδος σύνθεσης MBROLA

Η μέθοδος MBROLA παρά το γεγονός ότι βασίζεται σε δίφωνα, η ποιότητα που προσφέρει θεωρείται υψηλότερη από τους περισσότερους συνθέτες δίφωνων. Αυτό οφείλεται μερικώς, στο γεγονός ότι βασίζεται στην προ-επεξεργασία διφώνων, επιβάλλοντας όμοιες διάρκειες, εντάσεις και φασματική λείανση μεταξύ των διαφορετικών φωνημάτων της βάσης δεδομένων. Βασίζεται ουσιαστικά στην εφαρμογή της αρχικής θεώρησης της διφωνηματικής σύνθεσης, όπου ουσιαστικά όλα τα φωνήματα της βάσης ιδανικά εκφέρονται με το ίδιο pitch, ένταση και διάρκεια. Στο πλαίσιο ανάπτυξης της συγκεκριμένης μεθόδου, αναπτύχθηκε μία πλατφόρμα η οποία υποστηρίχθηκε από πλήθος εργαστηρίων και ερευνητών ανά τον κόσμο, συμβάλλοντας στην δημιουργία ενός πολυγλωσσικού συστήματος, διαθέσιμου στην επιστημονική κοινότητα.

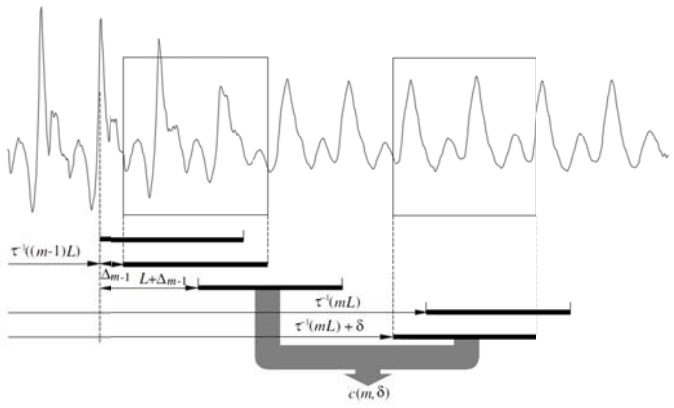
2.3.3.3.2 Μέθοδος σύνθεσης WSOLA

Η μέθοδος WSOLA (Waveform Similarity OLA) βασίζεται στις γενικότερες τεχνικές σύνθεσης στο πεδίο του χρόνου Overlap Add, κάνοντας ένα βήμα παραπάνω, προτείνοντας την χρήση σταθερού παραθύρου ανάλυσης και την προσπάθεια χρήσης σημάτων φωνής που ομοιάζουν μεταξύ τους το μέγιστο δυνατό. Με άλλα λόγια, στα σημεία σύνδεσης των διαδοχικών ακουστικών μονάδων, το σύστημα δεν κάνει παράθεση των γειτονικών στοιχειωδών τμημάτων ήχου, αλλά προσπαθεί να βρει παράθυρα ήχου που μοιάζουν μεταξύ τους καλύτερα, χρησιμοποιώντας ένα κριτήριο φασματικής απόστασης με βάση το Short Time FFT των σημάτων.

$$X(\omega, m) = \sum_{n=-\infty}^{+\infty} x(n+m)w(n)e^{-j\omega t}$$

Εξίσωση 1: Ορισμός του Short Time FFT.

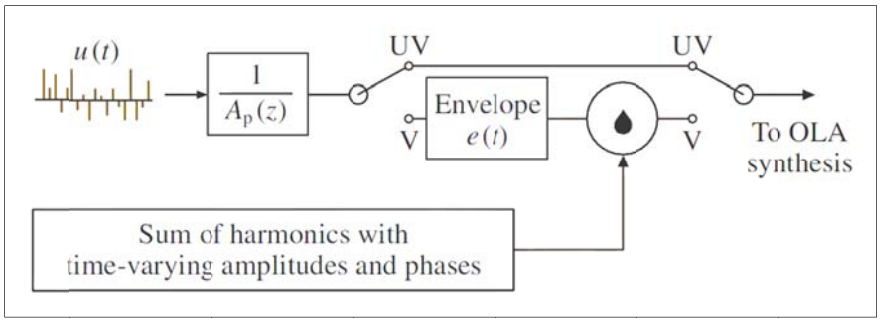
Η συγκεκριμένη μέθοδος προσφέρει υψηλής ποιότητας συνθετικό σήμα, χωρίς ωστόσο να επιτρέπει την μετατροπή της θεμελιώδους συχνότητας παράλληλα με την μετατροπή της διάρκειας των φωνημάτων.



Σχήμα 12: Απεικόνιση της μεθόδου WSOLA όπου η επιλογή των επικαλυπτόμενων παραθύρων βασίζεται σε κριτήριο ομοιότητάς με βάση το SFFT των παραθύρων. (Πηγή: [Verhelst1992])

2.3.3.3.3 Μέθοδος σύνθεσης HNM

Η μέθοδος HNM (Harmonic Plus Noise Model) αποτελεί ένα αντιπροσωπευτικό παράδειγμα των ημιτονοειδών μοντέλων (Sinusoidal models), τα οποία από μόνα τους αποτελούν έναν ειδικό τρόπο προσέγγισης και επεξεργασίας του σήματος φωνής για σύνθεση. Η βασική θεωρία πίσω από την συγκεκριμένη κατηγορία αλγορίθμων είναι η αποδόμηση του σήματος φωνής σε αρμονικές συνιστώσες, και η επεξεργασία του σήματος με βάση τις αρμονικές αυτές. Η μετατροπή της θεμελιώδους συχνότητας αλλά και της διάρκειας πραγματοποιείται στο πεδίο της συχνότητας, ενώ το σημαντικότερο πρόβλημα των μεθόδων αυτών είναι η δυνατότητα ανίχνευσης των αρμονικών καθώς και η ορθή ανίχνευση των σημείων όπου οι συγκεκριμένες αρμονικές απουσιάζουν από το σήμα. Η μέθοδος HNM διαχωρίζει το σήμα φωνής σε δύο συνιστώσες, μία χαμηλών συχνοτήτων όπου περιέχονται και οι αρμονικές, και μία υψηλών συχνοτήτων όπου ουσιαστικά αναπαρίσταται η «θορυβώδης» φύση του σήματος. Πραγματοποιώντας τον συγκεκριμένο διαχωρισμό, και εκτελώντας διαφορετικού τύπου επεξεργασία στις δύο συνιστώσες, η μέθοδος HNM επιτυγχάνει καλύτερη ποιότητα ήχου και λείανση φασματικών ασυνεχειών στα όρια των ακουστικών μονάδων [Stylianou1999], [Stylianou2002], [Olive1998].



Σχήμα 13: Σύνθεση φωνής με βάση την μέθοδο HNM
(Πηγή:[Stylianou2001])

2.3.3.3.4 Μέθοδος σύνθεσης TD-PSOLA

Η μέθοδος PSOLA (Pitch Synchronous Overlap Add) αναπτύχθηκε αρχικά στα εργαστήρια της France Telecom (CNET).

Αν και υπάρχουν πολλές διαφορετικές εκδόσεις της PSOLA όλες βασίζονται στην ίδια λογική. Η πιο γνωστή από αυτές είναι η Time Domain – PSOLA, όπου η επεξεργασία του σήματος εκτελείται στο πεδίο του χρόνου. Ο βασικός αλγόριθμος της TD-PSOLA μπορεί να διακριθεί σε τρία βήματα:

1. το βήμα ανάλυσης στο οποίο το αρχικό σήμα διαιρείται σε ξεχωριστά, αλλά συχνά επικαλυπτόμενα τεμάχια σήματος μικρού μήκους (short-term analysis signals ή ST)

2. την μετατόπιση κάθε σήματος ανάλυσης ώστε να πραγματοποιηθεί μετασχηματισμός της διάρκειας αλλά και της θεμελιώδους συχνότητας όπως καθορίζονται από την προσωδία στόχο

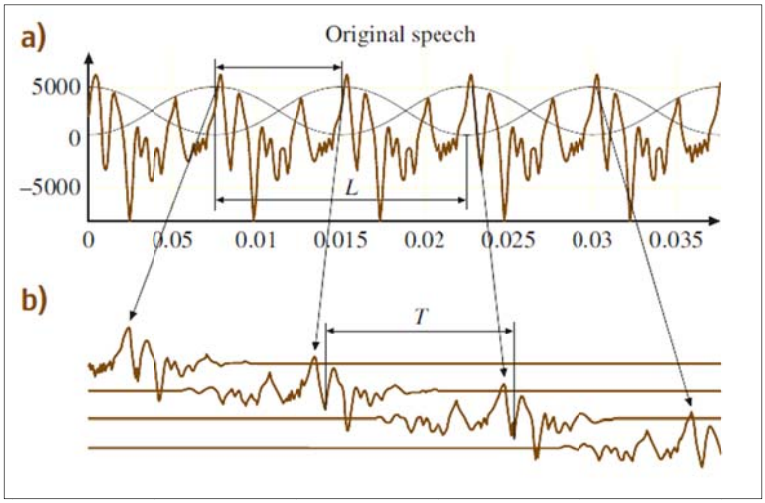
3. το βήμα σύνθεσης όπου τα τεμάχια παραθέτονται και συνενώνονται με τη μέθοδο της επικάλυψης (overlap-add).

Τα ST τμήματα $x_m(n)$ αποτελούν παραθυροποιημένα στοιχειώδη τμήματα σήματα από το αρχικό σήμα φωνής $x(n)$, τα οποία έχουν μήκος ίσο με 2 φορές την τοπική θεμελιώδη συχνότητα:

$$x_m(n) = h_m(t_m - n)x(n)$$

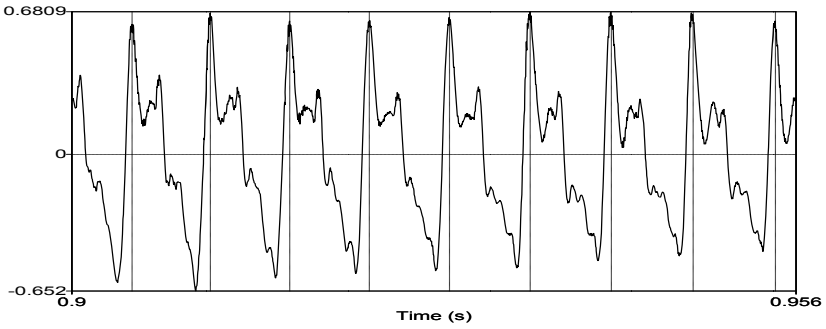
Εξίσωση 1: Ορισμός των ST-signals.

όπου m είναι ένας δείκτης που αναφέρεται στα ST σήματα.



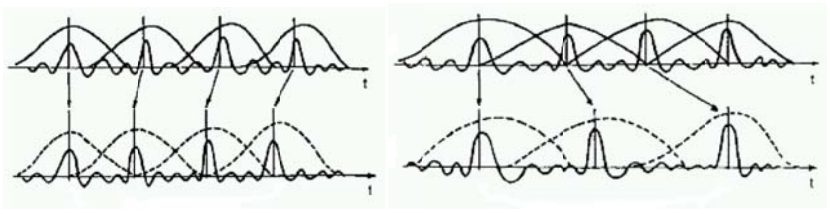
Σχήμα 14: Αναπαράσταση σήματος φωνής και δημιουργία των ST σημάτων για επεξεργασία μέσω PSOLA.

Το παράθυρο που χρησιμοποιείται είναι τύπου Hanning και έχει κέντρο τα pitch σημεία ανάλυσης (analysis pitch marks) τα οποία είναι σημεία μέσα στην κυματομορφή που απέχουν το ένα από το άλλο κατά μία περίοδο που αντιστοιχεί στην θεμελιώδη συχνότητα.



Σχήμα 15: Κυματομορφή σήματος φωνής και επισημείωση πιθανών pitchmark σημείων ανάλυσης. Τα τελευταία σημειώνονται με κατακόρυφες διακεκομμένες γραμμές.

Η αλλαγή στην θεμελιώδη συχνότητα του σήματος επιτυγχάνεται με την μετατόπιση των σημείων ανάλυσης κατά τρόπο τέτοιο που τα νέα σημεία απέχουν απόσταση ίση με την περίοδο που αντιστοιχεί στην νέα τοπική θεμελιώδη συχνότητα. Στην συνέχεια τα στοιχειώδη αυτά τμήματα σήματος (ST-signals) προστίθενται με την ανάλογη επικάλυψη μεταξύ τους. Μία μείωση ή αύξηση στο pitch περιγράφεται στο παρακάτω σχήμα:



(α) (β)

Σχήμα 16: Σχηματική αναπαράσταση του τρόπου αύξησης (α) και μείωσης (β) της βασικής συχνότητας με τη μέθοδο TD-PSOLA.

Μεταβολή στην διάρκεια του σήματος μπορεί να επιτευχθεί ταυτόχρονα με την μεταβολή στην θεμελιώδη συχνότητα, με την επανάληψη ή την παράλειψη στοιχειωδών τμημάτων σήματος, ανάλογα με το αν επιθυμούμε την επιμήκυνση ή τον περιορισμό της τελικής διάρκειας του σήματος.

Η συγκεκριμένη τεχνική επιτρέπει τον ταυτόχρονο χειρισμό των διαρκειών και της θεμελιώδους συχνότητας του σήματος φωνής, παρουσιάζοντας ωστόσο και περιορισμούς στο εύρος της επεξεργασίας του σήματος, αφού ανάλογα και με τα ιδιαίτερα χαρακτηριστικά της φωνής, μεγάλη σχετικά αλλαγή της θεμελιώδους συχνότητας επιφέρει παραμόρφωση στο σήμα της φωνής, ενώ σημαντική παράμετρος στην συγκεκριμένη μεθοδολογία αποτελεί η κατάλληλη επιλογή των σημείων ανάλυσης του σήματος, αφού αυτά επηρεάζουν σημαντικά το τελικό σήμα. Περισσότερα για την τεχνική επιλογής των σημείων ανάλυσης στο σήμα (PitchMarks) αναφέρουμε στο κεφάλαιο που περιγράφει την βάση δεδομένων και την διαδικασία δημιουργίας της [Moulines1990].

2.4 Βιβλιογραφία Κεφαλαίου

[Baer1991] Baer T., J.C. Gore, L.C. Gracco, P.W. Nye: Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels, J. Acoust. Soc. Am. 90, 799–828 (1991)

[Bailly2003] Bailly G., W.N. Campbell, and B. Mobius, "ISCA Special Session: hot topics in speech synthesis", Proc. Eurospeech 2003, pp. 37-40, Geneva, 2003.

[Beutnagel1999] Beutnagel M., Mohri, R., and Riley, M., "Rapid unit selection from a large speech corpus for concatenative speech synthesis," in Proc. Eurospeech 99, Budapest, 1999.

[Black1997] Black A., and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," Proc. of Eurospeech 97, vol. 2, pp. 601-604, Greece, 1997.

[Black1995] Black A.W., N. Campbell: Optimising selection of units from speech databases for concatenative synthesis, ESCA Eurospeech 95, 581–584 (1995)

[Campbell2005] Campbell N., "Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech," IEICE trans. Inf. & Syst., vol. E88-D, no. 3, pp.376-383, 2005.

[Coker2000] Coker C.H.: A model of articulatory dynamics and control, Proc. IEEE 64, 452–459 (1976)

Coorman G., Fackrell, J., Rutten, P., and Coile, B. V., "Segment selection in the LH realspeak laboratory TTS system," Proc. of the ICSLP 2000, vol. 2, pp. 395-398, 2000.

[Dudley1939] Dudley H., R.R. Riesz, S.A. Watkins: A synthetic speaker, J. Franklin Inst. 227, 739–764 (1939), http://www.bell-labs.com/org/1133/Heritage/Vocoder/

[Duggan2003] Duggan B. and M. Deegan, "Considerations in the usage of text to speech (tts) in the creation of natural sounding voice enabled web systems", In Proc. of the 1st Int. Symp. on Information and Communication technologies (ISICT '03:), pp. 433–438, Trinity College Dublin, 2003.

[Dutoit1997] Dutoit T.: An Introduction to Text-to-Speech Synthesis (Kluwer Academic, Dordrecht 1997)

[Dutoit2008] Dutoit T., "Corpus-based Speech Synthesis," Springer Handbook of Speech Processing, J. Benesty, M. M. Sondhi, Y. Huang (eds), Part D, Chapter 21, pp. 437-455, Springer, 2008.

[Eide2004] Eide E., R. Bakis, W. Hamza, J.F. Petrelli: Toward expressive synthetic speech. In: Text-to-Speech Synthesis – New Paradigms and Advances, Professional Technical Reference, ed. by S. Narayanan, A. Alwan (Prentice-Hall, Upper Saddle River 2004) pp. 219–248, Chap. 11

[Fujimura1978] Fujimura O., J. Lovins: Syllables as concatenative phonetic elements. In: Syllables and Segments, ed. by A. Bell, J.B. Hooper (North-Holland, New York 1978) pp. 107–120

[Fujisaki1992] Fujisaki H.: The role of quantitative modeling in the study of intonation, Proc. Int. Symp. Japanese Prosody (1992) pp. 163–174

[Giannopoulos et al., 2003] G. Giannopoulos, A. Chalamandaris, S-E. Fotinea, T. Athanaselis, G. Carayannis, "Analysis and modelling of the Carrier Declination for the Greek language", in Proc. of the 15th International Congress of Phonetic Sciences - ICPhS03, 3-9 August 2003, Barcelona, pp. 555-558.

[Hanson2002] Hanson H.M., K.N. Stevens: A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn, J. Acoust. Soc. Am. 112, 1158–1182 (2002)

[Heuven1995] Heuven V. J. van and R. van Bezooijen, "Quality Evaluation of Synthesized Speech," Speech Coding and Synthesis, W. B. Kleijn, and K. K. Paliwal (eds), Chapter 21, pp. 707-738, Elsevier Science, 1995.

[Hunt1996] Hunt A., A.W. Black: Unit selection in a concatenative speech synthesis system using a large speech database, Proc. ICASSP 96, 373–376 (1996)

[Karabetsos2010] Karabetsos S.,"Βελτίωση της ποιότητας της συνθετικής φωνής και εφαρμογή σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και υπηρεσίες", Διδακτορική διατριβή ΕΜΠ, 2010

[Kain2006] Kain A., M. Macon: Spectral voice conversion for text-to-speech synthesis, Proc. IEEE ICASPP 98, 285–288 (1998)

[Kim2006] Kim S.-J., Kim J.-J. and Hahn M.-S., "HMM-based Korean speech synthesis system for hand-held devices," IEEE Trans. Consumer Electronics, vol. 52, no. 4, pp. 1384-1390, 2006.

[Klatt1980] Klatt D.H.: Software for a cascade/parallel formant synthesizer, J. Acoust. Soc. Am. 67, 971–995 (1980)

[Macchi1993] Macchi M., M.J. Altom, D. Kahn, S. Singhal, M. Spiegel: Intelligibility as a function of speech coding method for template-based speech synthesis, Proc. Eurospeech 93, 893–896 (1993) Part D 19

[Markel1976] Markel J.D., A.H. Gray: Linear Prediction of Speech (Springer, New York 1976)

[Moore2007] Moore R. K., "PRESENCE: A human-inspired architecture for speech-based human-machine interaction," IEEE Trans. Computers, 56, pp. 1176-1188, 2007.

[Moulines1990] Moulines E., F. Charpentier: Pitch-synchronous waveform processing techniques for text-to speech synthesis using diphones, Speech Commun. 9(5-6), 453–467 (1990)

[O'Shaughnessy2007]O'Shaughnessy D., "Modern Methods of Speech Synthesis," IEEE Circuits and Systems Magazine, Third Quarter 2007, pp. 6-23, 2007.

[Olive1998] Olive J., J. van Santen, B. Mobius, C. Shih: Synthesis. In: Multilingual Text-to-Speech Synthesis – The Bell Labs Approach, ed. by R. Sproat (Kluwer Academic, Dordrecht 1998), Chap. 7

[Olive1977] Olive J.P.: Rule synthesis of speech from diadic units, Proc. ICASSP 77, 568–570 (1977)

[Quartieri1992] Quartieri T.F., R.J. McAulay: Shape invariant timescale and pitch modification of speech, IEEE Trans. Signal Process. 40(3), 497–510 (1992)

[Rabiner1993] Rabiner L., B.H. Juang: Fundamentals of Speech Recognition (Prentice-Hall, Englewood Cliffs 1993) pp. 339–341

[Richard1995] Richard G., M. Liu, D. Sinder, H. Duncan, Q. Lin, J. Flanagan, S. Levinson, D. Davis, S. Slimon: Numerical simulations of fluid flow in the vocal tract, Proc. of Eurospeech Madrid (1995) pp. 18–21

[Sagisaka1988] Sagisaka Y.: Speech synthesis by rule using an optimal selection of non-uniform synthesis units, Proc. ICASSP 88, 679–682 (1988)

[Santen1998] Santen J.P. van: Timing. In: Multilingual Text-to-Speech Synthesis – The Bell Labs Approach, ed. by R. Sproat (Springer, New York 1998) pp. 115–139

[Santen1997] Santen J.P.H. van: Combinatorial issues in text to speech synthesis, EuroSpeech '97 5th European Conference on Speech Communication and Technology 5, 2511–2514 (1997)

[Schroeter1991] Schroeter J., M.M. Sondhi: Speech coding based on physiological models of speech production. In: Advances in Speech Signal Processing, ed. by S. Furui, M.M. Sondhi (Marcel Dekker, New York 1991) pp. 231–268

[Silverman1992] Silverman K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg: TOBI: A standard for labeling English prosody, Proc. ICSLP'92 Banff (1992) pp. 867–870

[Stone1990] Stone M.: A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data, J. Acoust. Soc. Am. 87, 2207–2217 (1990)

[Stylianou1996] Stylianou I.: Modeles Harmoniques plus Bruit combines avec des Methodes Statistiques, pour la Modication de la Parole et du Locuteur, Doctoral Thesis (Ecole Nationole Superieure des Telecommunications, Paris 1996), in French

[Stylianou2001] Stylianou Y.: Applying the harmonic plus noise model in concatenative speech synthesis, IEEE Trans. Speech Audio Process. 9(1), 21–29 (2001)

[Syrdal2007] Syrdal A.: Development of a standard for the evaluation of intelligibility of text-to-speech synthesis systems by ANSI Accredited Standards Committee S3, Bioacoustics, working group S3/WG 91, Text-to-Speech Synthesis Systems, Personal communication (2007)

[Tokuda2004] Tokuda K., H. Zen, A.W. Black: An HMM-based approach to multilingual speech synthesis. In: Text-to-Speech Synthesis – New Paradigms and Advances, Professional Technical Reference, ed. by S. Narayanan, A. Alwan (Prentice-Hall, Upper Saddle River 2004) pp. 135–153, Chap. 7

[Tsiakoulis2008] Tsiakoulis P., A. Chalamandaris, S. Karabetsos and S. Raptis, "A statistical method for database reduction for embedded unit selection speech synthesis," in IEEE ICASSP 2008, pp. 4601-4604, 2008.

[Tsiakoulis2010] Tsiakoulis P., "Σύνθεση φωνής με υπολογιστική αεροδυναμική ανάλυση του ανθρωπινού ηχητικού σωλήνα και σύγκριση με κλασσικές μεθόδους", Διδακτορική διατριβή ΕΜΠ, 2010

[Vepa2004] Vepa J., S. King: Join cost for unit selection speech synthesis. In: Text-to-Speech Synthesis – New Paradigms and Advances, Professional Technical Reference, ed. by S. Narayanan, A. Alwan (Prentice-Hall, Upper Saddle River 2004) pp. 35–62, Chap. 3

[Viswanathan2005]Viswanathan M. and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," Computer Speech & Language, vol. 19, pp. 55-83, January 2005.

3. ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΑΙ ΠΑΡΑΓΩΓΗ ΠΡΟΣΩΔΙΑΣ

3.1 Εισαγωγή

Στο κεφάλαιο αυτό πρόκειται να αναφερθούμε στην προσωδία, στον ορισμό αυτής, στην σημασία που έχει και τον ρόλο που διαδραματίζει σε ένα σύστημα σύνθεσης φωνής, στις διαφορετικές μεθόδους και αλγορίθμους που υπάρχουν σχετικά με την μοντελοποίηση αυτής, όπως επίσης και στην προσέγγιση που έχουμε εμείς ακολουθήσει. Σημαντικό χαρακτηριστικό της μεθόδου μας αποτελεί η προσέγγιση της μοντελοποίησης μέσα από τα ίδια τα δεδομένα (data-driven method) γεγονός που μας επιτρέπει να δημιουργήσουμε με σχετική ευκολία διαφορετικά μοντέλα από διαφορετικά δεδομένα, μοντελοποιώντας ουσιαστικά τα προσωδιακά χαρακτηριστικά του συγκεκριμένου ομιλητή. Στο τέλος του κεφαλαίου συζητάμε για την αποτελεσματικότητα της προσέγγισής μας, ενώ παρουσιάζουμε συνοπτικά θέματα προς μελλοντική διερεύνηση, όπου η συγκεκριμένη μέθοδος θα έπρεπε να προσαρμοστεί ανάλογα ώστε να μπορεί να αντιμετωπίσει εξίσου αποτελεσματικά πολυπλοκότερες καταστάσεις.

3.1.1 Ορισμός

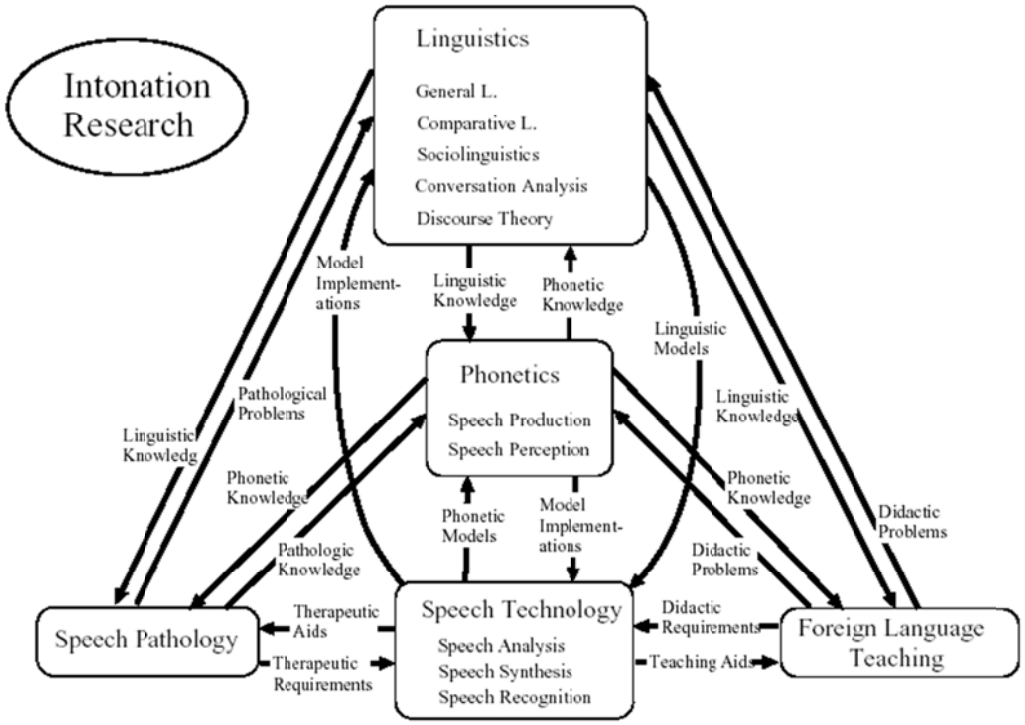
Ο όρος προσωδία περιλαμβάνει τον επιτονισμό - τον ρυθμό και τον λεκτικό τόνο (lexical stress) στην φωνή. Τα προσωδιακά χαρακτηριστικά μιας δομικής μονάδας λόγου, είτε αυτή είναι συλλαβή, είτε λέξη ή φράση, ονομάζονται υπερτεμαχιακά χαρακτηριστικά (suprasegmental) επειδή επηρεάζουν και αναφέρονται σε όλα τα τεμάχια της δομικής αυτής μονάδας. Αυτά τα χαρακτηριστικά εκδηλώνονται, ως μήκη συλλαβών (durations), τόνος στην φράση (tone) και λεξικός τόνος (stress).

Ο ρόλος της προσωδίας στις διαπροσωπική επικοινωνία είναι πολύπτυχος [Bolinger1978] [Leben1976]. Καταρχήν τα προσωδιακά χαρακτηρστικά χρησιμοποιούνται για την διάκριση λεκτικών μονάδων, αφού μέσω των κατάλληλων διαρκειών, διαφοροποίηση στον τόνο κ.α. δημιουργούνται επιτονικά πρότυπα που ενώ περιέχουν παρόμοια ή ίδια φωνήματα, είναι ικανά ώστε να επιτρέψουν στον ακροατή να κατανοήσει το νόημα ή ακόμη και να διακρίνει ομόηχες λέξεις [Bolinger1989]. Χαρακτηριστικό παράδειγμα είναι οι προτάσεις «*To bury treasure was exciting*» και «*To Barry, treasure was exciting*» οι οποίες έχουν σχεδόν ίδια προφορά και την ίδια αλληλουχία φωνημάτων, ωστόσο μπορεί κανείς να διακρίνει την μία από την άλλη βασιζόμενος στην διαφορετική προσωδία κατά την εκφορά τους [Chomsky1968] [Dimou & Chalamandaris, 2006].

Κατά δεύτερο λόγο, τα προσωδιακά χαρακτηριστικά επιτρέπουν την δημιουργία σχέσεων μεταξύ των φράσεων σε ακουστικό επίπεδο, δημιουργώντας τις απαραίτητες ακουστικές δομές και αντίστοιχα σχέσεις μεταξύ τους [Dimou & Chalamandaris, 2008]. Τέλος, η προσωδία είναι απαραίτητη για την ενσωμάτωση χαρακτηριστικών όπως έμφαση, αντίθεση, κ.α. σε λεκτικές δομές ή φράσεις [Crystal1969].

Μία απλουστευμένη θεώρηση της προσωδίας, όπως άλλωστε πολύ συχνά χρησιμοποιείται στον χώρο της σύνθεσης φωνής, αναφέρει ως βασικές συνιστώσες της προσωδίας την θεμελιώδη συχνότητα, δηλαδή την περιοδικότητα των έμφωνων ήχων στον λόγο, την ένταση και την διάρκεια των δομικών στοιχείων του λόγου, είτε αυτά είναι φθόγγοι, συλλαβές, λέξεις κ. ο. κ. [Mobius1993] Επομένως η «προσωδία» είναι μία πολύπλοκη έννοια που περιλαμβάνει τόσο πληροφορία συχνότητας ταλάντωσης των φωνητικών χορδών, όσο και ενεργειακών μεταβολών και χρονσιμού.

Η μελέτη της προσωδίας αποτελεί κοινό τόπο για πολλές επιστήμες και διαφορετικούς χώρους, όπως είναι η γλωσσολογία, η φωνητική παθολογία, η μεθοδολογία εκμάθησης ξένης γλώσσας, κ.α., ενώ ο ορισμός της προσωδίας λαμβάνει διαφορετικές περιγραφές όταν αναφερόμαστε στο πεδίο της παραγωγής λόγου, στην ακουστική και στο επίπεδο της αντίληψης [Kompe1997]. Στο σχήμα 17 φαίνεται μία γραφική απεικόνιση των διαφορετικών ερευνητικών επιπέδων που έχει η μελέτη της προσωδίας. Αν και η μελέτη της προσωδίας έχει την ίδια βάση για όλες σχεδόν τις γλώσσες, η κάθε γλώσσα παρουσιάζει διαφορετικά χαρακτηριστικά, τα οποία και προβάλλουν διαφορετικές ανάγκες και συμβάσεις. Στις τονικές γλώσσες π.χ. η θέση του τόνου συχνά ορίζει διαφορετικής σημασίας λέξεις (Μανδαρίνικα), ενώ ο τρόπος πραγμάτωσης του τόνου στο σήμα της φωνής μπορεί να παρατηρηθεί είτε με απλή αύξηση της θεμελιώδους συχνότητας τοπικά, είτε με επιμήκυνση των διαρκειών των συλλαβών (time-stretched languages), είτε με συνδυασμό και των δύο και ενδεχομένως και παράλληλη αύξηση της τοπικής έντασης.



Σχήμα 17: Σχηματική απεικόνιση των διαφορετικών ερευνητικών πτυχών που έχει η μελέτη της προσωδίας. (Πηγή:[Bolinger1989])

Η συγκεκριμένη διατριβή μελετά φαινόμενα της Ελληνικής γλώσσας - μία συλλαβική γλώσσα με διακεκριμένες τις τονισμένες συλλαβές από τις άτονες και χωρίς μεγάλη εξάρτηση του ρυθμού από τον τόνο.

Στην πράξη, οι επιμέρους έρευνες παρουσιάζουν στεγανά μεταξύ τους και αδυναμία συγχώνευσης των ερευνητικών αποτελεσμάτων από ένα επίπεδο σε ένα άλλο, λόγω διαφορετικών συμβάσεων και αντιλήψεων μεταξύ ανθρωπιστικών επιστημών (γλωσσολογία, εκμάθηση γλώσσας, φωνητική παθολογία) και φυσικών επιστημών (τεχνολογία φωνής).

Στην συνέχεια αναφέρουμε τις βασικότερες έννοιες και όρους που είναι απαραίτητοι για την κατανόηση του συγκεκριμένου πεδίου.

3.1.1.1 Τόνος και επιτονισμός(Pitch and Intonation)

Τόνο στο πλαίσιο της προσωδίας ονομάζουμε το αποτέλεσμα της αντίληψης της θεμελιώδους συχνότητας F0 από το ανθρώπινο αφτί. Αντίστοιχα επιτονισμός ορίζεται το σύνολο των διαφορετικών προτύπων του τόνου στο επίπεδο της αντίληψης, ή της θεμελιώδους συχνότητας F0 στο επίπεδο της παραγωγής. Η μελέτη του τόνου έχει πολύ μεγαλύτερες απαιτήσεις από ό,τι η μελέτη των διαρκειών η οποία απλά απαιτεί τον διαχωρισμό των φθόγγων στο σήμα φωνής, καθώς η αναπαράσταση του επιτονισμού γίνεται σε πολυδιάστατο χώρο, χωρίς κοινά αποδεκτό μέχρι σήμερα σύστημα παραμέτρων.

3.1.1.2 Φραστικός τόνος

Ο φραστικός τόνος αναφέρεται σε ολόκληρη την πρόταση και υποδηλώνει το σημείο όπου εκδηλώνεται η έμφαση. Αυτή μπορεί να εκδηλωθεί είτε ως απότομη αύξηση ή μείωση της θεμελιώδους συχνότητας, με παράλληλη συχνά αύξηση και της αντίστοιχης τοπικής ενέργειας του σήματος της φωνής και επομένως σημαντική αλλαγή στις καμπύλες της προσωδίας.

3.1.1.3 Λεξικός τόνος

Ο λεξικός (ή λεκτικός) τόνος είναι ο τόνος που αναφέρεται σε κάθε λέξη που τονίζεται. Σε ορισμένες γλώσσες όπως την Ελληνική, ο τόνος αυτός σημειώνεται στα γραφήματα (αναπαράσταση τονισμένων φωνηέντων μέσω του σημείου του τόνου), ενώ στις περισσότερες

γλώσσες ο τόνος εννοείται και αποστηθίζεται πρακτικά, κατά την εκμάθηση της γλώσσας. Συχνές είναι επίσης οι περιπτώσεις όπου ομόηχες λέξεις με διαφορετική σημασία, διακρίνονται μεταξύ τους αποκλειστικά από την θέση του τόνου μέσα στην λέξη αυτή. Χαρακτηριστικό παράδειγμα αποτελεί η Αγγλική γλώσσα, όπου λέξεις όπως /address/τονίζεται σε διαφορετική συλλαβή όταν πρόκειται για ρήμα και όταν πρόκειται για ουσιαστικό, ενώ αντίθετα σε γλώσσες όπως η Γαλλική, ο τονισμός της λέξης πραγματοποιείται υποχρεωτικά στην λήγουσα.

3.1.1.4 Λέξη επιτονισμού

Λέξη επιτονισμού ονομάζεται ένα υποσύνολο λέξεων ή χαρακτήρων στην φράση, το οποίο παρουσιάζει έναν και μοναδικό λεξικό τόνο. Σε περιπτώσεις όπου λέξεις δεν περιλαμβάνουν λεξικό τόνο, αυτές συγχωνεύονται σε γειτονικές λέξεις επιτονισμού. Στην πρόταση π.χ. «την Κυριακή που πέρασε», οι περιλαμβανόμενες λέξεις επιτονισμού είναι οι ΛΕ1: «την Κυριακή» και ΛΕ2: «που πέρασε», όπου περιλαμβάνονται στην μεν πρώτη η λέξη «την» και στην δε δεύτερη η λέξη «που».

3.1.1.5 Κλίση

Κλίση της καμπύλης επιτονισμού ονομάζουμε την κλίση της φέρουσας της καμπύλης της θεμελιώδους συχνότητας. Η μοντελοποίηση της γίνεται συνήθως με τμήματα διακεκριμένων ευθειών, από αρχή πρότασης μέχρι ενδιάμεσο σημείο στίξης, π.χ. κόμμα, και από ενδιάμεσο σημείο στίξης μέχρι το τελικό σημείο της περιόδου.

3.1.1.6 Ρυθμός

Η μελέτη του ρυθμού στο πλαίσιο της σύνθεσης φωνής και όχι μόνο, έχει κυρίως μετατοπισθεί στην μελέτη των διαρκειών των ακουστικών μονάδων (φωνήματα) και λιγότερο συχνά στην μελέτη των διαρκειών μεγαλύτερων μονάδων όπως είναι οι συλλαβές ή λέξεις. Ο λόγος είναι προφανής: ο ευκολότερος διαχωρισμός των μονάδων αυτών μέσα στο σήμα. Αν και η συγκεκριμένη συνιστώσα της προσωδίας αποτέλεσε αντικείμενο μελέτης και έρευνας μετά την θεμελιώδη συχνότητα, θεωρείται πλέον ότι η μοντελοποίηση των διαρκειών των φωνημάτων μπορεί να επιτευχθεί με ικανοποιητική ακρίβεια, τουλάχιστον όσον αφορά την ανθρώπινη αντίληψη των ήχων αυτών.

Αναφορικά με την μοντελοποίηση της θεμελιώδους συχνότητας, συνοπτικά μπορούμε να αναφέρουμε τα ευρήματα του Hawkins όπου τελικά περιγράφει τα συστηματικά λάθη που υπεισέρχονται στην μοντελοποίηση και παραγωγή της μονοδιάστατης μεταβλητής της θεμελιώδους συχνότητας, με χρήση πολυδιάστατων παραμέτρων και χαρακτηριστικών. Πιο συγκεκριμένα στην έρευνά του καταλήγει στα εξής συμπεράσματα:

1. *«Δεν υπάρχουν καθαρά διακεκριμένες γλωσσικές μονάδες ή οντότητες που να αποτελούν ταυτόχρονα και δομικές μονάδες του επιτονισμού.»*

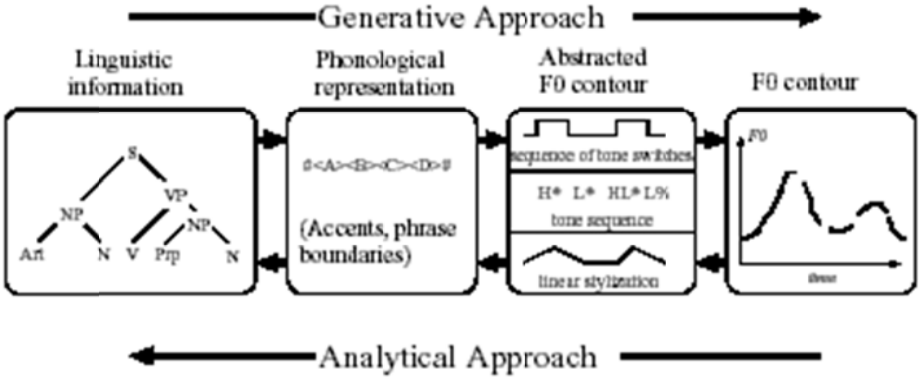
2. *«Γλωσσικά φαινόμενα και χαρακτηριστικά είναι γενικότερα ιδιαίτερα πολύπλοκα και επηρεάζουν μεγαλύτερα τμήματα της καμπύλης της θεμελιώδους συχνότητας.»*

3. *«Οι πολυδιάστατες αυτές παράμετροι γενικότερα συμβάλλουν σε περισσότερα από ένα γλωσσικά δομικά στοιχεία, και είναι ιδιαίτερα μεταβλητές.»*

Οι παρατηρήσεις αντικατοπτρίζουν σε μεγάλο βαθμό την δυσκολία δημιουργίας ενός γενικού μοντέλου προσωδίας. Για τον λόγο αυτό, είναι πολλές οι διαφορετικές προσεγγίσεις που έχουν επιχειρηθεί τις τελευταίες τρεις δεκαετίες για την μοντελοποίηση της προσωδίας, τις οποίες θα παρουσιάσουμε συνοπτικά στην συνέχεια.

3.2 Μοντελοποίηση επιτονισμού

Οι διαφορετικές προσεγγίσεις για την μοντελοποίηση του επιτονισμού μπορούν εύκολα να κατηγοριοποιηθούν σε δύο σημαντικές ομάδες: ***τις γεννετικές/δημιουργικές (generative)*** και στις ***αναλυτικές (analytical)***. Η διαφορά τους έγκειται στην διαδρομή που ακολουθεί κανείς από τα δεδομένα προς το αποτέλεσμα. Στο σχήμα 18 που ακολουθεί παρατηρείται η διαφορά αυτή.



Σχήμα 18: Ο ρόλος των μοντέλων επιτονισμού ως σύνδεσμος μεταξύ των γλωσσικών δομών και των ακουστικών πραγματώσεων στην καμπύλη του F0. (Πηγή: [Butzberger1990])

Σε γενικές γραμμές, η δημιουργική (ή γεννετική) προσέγγιση προσπαθεί να παράγει καμπύλες θεμελιώδους συχνότητας F0 αποκλειστικά από υψηλού επιπέδου γλωσσική πληροφορία, ενώ η αναλυτική προσέγγιση προσπαθεί μέσα από την παρατήρηση πραγματικών δεδομένων καμπυλών F0 να παράγει προσεγγιστικά μοντέλα, με την βοήθεια αναπαραστάσεων που έχουν φωνολογική και γλωσσική προέλευση. Η διάρθρωση της γλώσσας διαχωρίζεται σε δύο επίπεδα:

α) το πρωτεύον γραμματικό ή/και συντακτικό και

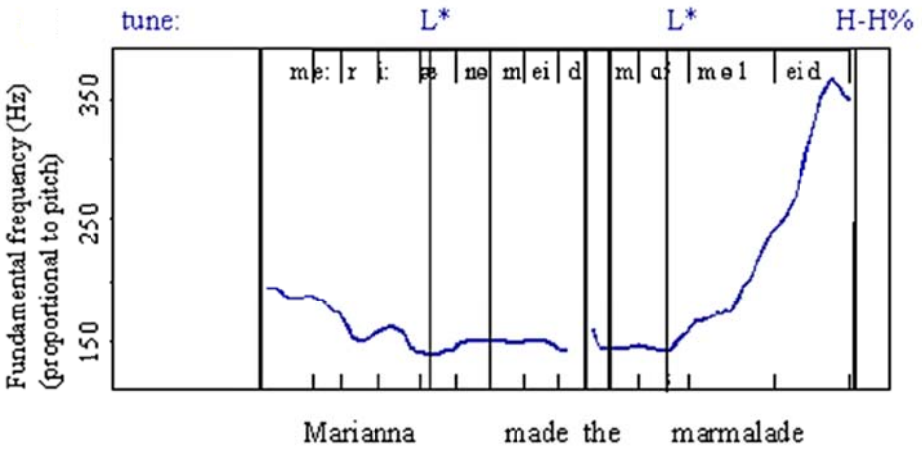
β) το δευτερεύον φωνολογικό.

Στο πρώτο επίπεδο οι μονάδες (λέξεις, γραμματικές/συντακτικές μονάδες) είναι φορείς έννοιας, σε αντίθεση με το δεύτερο επίπεδο όπου τα φωνήματα είναι μονάδες αναπαράστασης διακεκριμένων ήχων.

Στην συνέχεια θα παρουσιάσουμε συνοπτικά μερικές από τις βασικότερες μεθόδους, όπως έχουν ανακοινωθεί σε επιστημονικά συγγράμματα και έχουν υλοποιηθεί σε συστήματα παραγωγής συνθετικής φωνής. Αν και η γλώσσα στην οποία αναφέρονται οι περισσότερες προσεγγίσεις είναι διαφορετικές, συχνά η προσαρμογή τους σε μία άλλη γλώσσα είναι εφικτή χωρίς σημαντικά εμπόδια, εφόσον βέβαια οι γλώσσες παρουσιάζουν παρόμοια χαρακτηριστικά (συλλαβικές, τονικές κτλ.) [Bloomington1951], [Bolinger1978], [Hirst1998], [Leben1976].

3.2.1 Αναλυτικές Μέθοδοι: Μοντέλο ToBI

Η προσέγγιση αυτή [Pierrehumbert1980] είναι αναλυτική και επιχειρεί να περιγράψει όσο πιο αφηρημένα γίνεται τον τρόπο που τα φωνολογικά φαινόμενα επηρεάζουν τον επιτονισμό. Η γραμματική της μεθοδολογίας αυτής ορίζει ότι μία καμπύλη F0 μπορεί να περιγραφεί με μία σειρά υψηλών και χαμηλών τόνων της F0, αλλά και τόνων των ορίων των φράσεων. Το βασικό πλεονέκτημα είναι ότι επιτυγχάνει υψηλού επιπέδου ανάλυση των φωνολογικών φαινομένων που συμβάλλουν στην F0, ωστόσο η προσαρμογή της σε άλλες γλώσσες απαιτεί ιδιαίτερη προσπάθεια και γλωσσολογικές γνώσεις. Η συγκεκριμένη προσέγγιση έχει αποτελέσει για πολλά χρόνια την κυρίαρχη προσέγγιση στον χώρο των γλωσσολόγων, παρουσιάζοντας ένα προσαρμοσμένο μοντέλο για κάθε ξεχωριστή γλώσσα [Campbel1990], [Beckman1997], [Pierrehumbert1983].

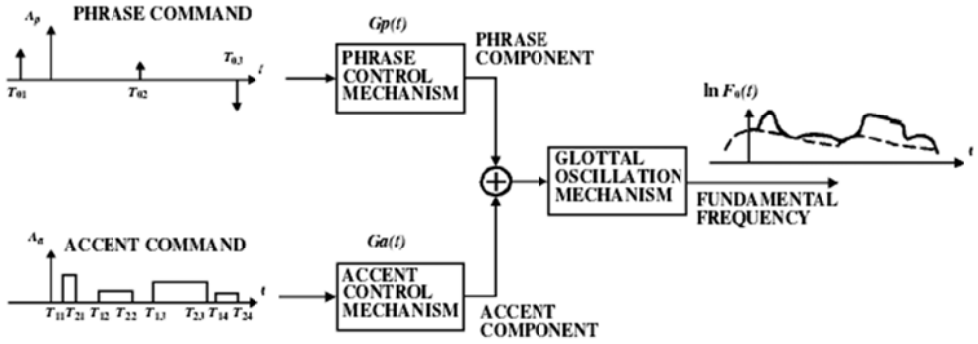


Σχήμα 19: Αναπαράσταση προσιδιακής επισήμειωσης με την μέθοδο ToBi για την φράση /Marianna made the mar,alade/. (Πηγή ['*Guidelines for TOBI labeling*' by Mary Beckman and Gayle Ayers)

3.2.2 Αναλυτικές Μέθοδοι: Μοντέλο Fujisaki

Το συγκεκριμένο μοντέλο (προτάθηκε το 1982) αναπτύχθηκε αρχικά για την μοντελοποίηση των προτύπων επιτονισμού για την Ιαπωνική γλώσσα, ωστόσο η υιοθέτησή του από ερευνητές για άλλες γλώσσες έχει πραγματοποιηθεί με επιτυχία, με αποτέλεσμα το συγκεκριμένο μοντέλο να είναι ένα από τα περισσότερο διαδομένα παγκοσμίως [Fujisaki1981-1992-1997]. Η λειτουργία

του μοντέλου έγκειται στην θεώρηση ότι η συνολική καμπύλη F0 μπορεί να δημιουργηθεί από μία φθίνουσα καμπύλη (φέρουσα), στην οποία υπερτίθενται τοπικά μέγιστα και ελάχιστα, λόγω λεκτικών τόνων. Κατά αυτόν τον τρόπο, γίνεται ένας διαχωρισμός της φέρουσας-κλίσης της καμπύλης F0 και των επιμέρους καμπυλών για κάθε λέξη επιτονισμού ξεχωριστά, ενώ στο τελικό αποτέλεσμα της καμπύλης της F0 συμβάλλουν με διαφορετικό τρόπο τόσο παράμετροι σε επίπεδο φράσης, όσο και παράμετροι σε επίπεδο λέξης αλλά και συλλαβής.

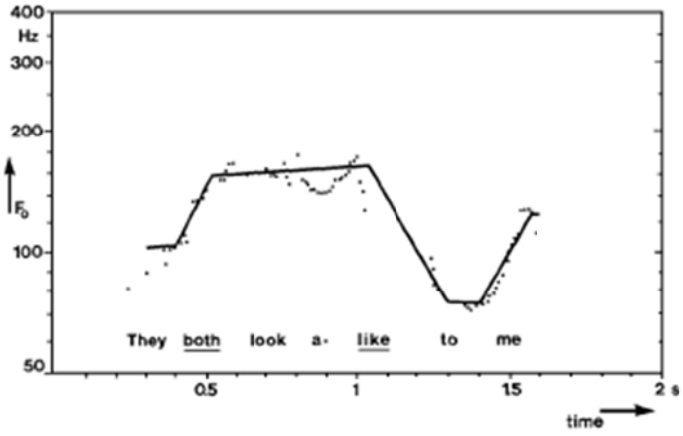


Σχήμα 20: Η μοντελοποίηση της προσωδίας στην σύνθεσης φωνής με την μέθοδο του Fujisaki.

Αν και αποτελεί μία ιδιαίτερα ελκυστική και αποδοτική προσέγγιση, πολλές φορές δημιουργούνται προβλήματα με την αυθαίρετη και συχνά μη γλωσσολογικά δικαιολογημένη ανάθεση των σημείων φραστικών και λεκτικών τόνων. Ακόμη και σήμερα ωστόσο η συγκεκριμένη μέθοδος αποτελεί πεδίο έρευνας ιδιαίτερα στον τρόπο υπολογισμού αλλά και ενσωμάτωσης συγκεκριμένων παραμέτρων της μεθόδου αυτής [Mixdorf2000], [Mobius1993].

3.2.3 Γεννετικές Μέθοδοι: Μοντέλο IPO (Ολλανδική σχολή)

Η συγκεκριμένη προσέγγιση παρουσιάστηκε το 1990 και σκοπό έχει την περιγραφή του επιτονισμού μέσα από το πρίσμα της αντίληψης από τον ακροατή. Η βασική ιδέα υποστηρίζει ότι μόνο οι σημαντικές αλλαγές στην καμπύλη της F0 έχουν νόημα να μοντελοποιηθούν αφού κυρίως αυτές γίνονται αντιληπτές από τον ακροατή, αλλά και αυτές βρίσκονται κυρίως υπό τον έλεγχο του ομιλητή. Κατά αυτόν τον τρόπο, η μοντελοποίηση γίνεται με την χρήση μεμονωμένων ευθύγραμμων τμημάτων που μοντελοποιούν την καμπύλη της θεμελιώδους συχνότητας του σήματος φωνής από μία αλλαγή μέχρι την επόμενη αλλαγή [Butzberger1990].



Σχήμα 21: Περιγραφή της θεμελιώδους συχνότητας στο σήμα φωνής με το μοντέλο IPO (Ολλανδική σχολή).

Τα ευθύγραμμα αυτά τμήματα στην συνέχεια προσαρμόζονται σε ένα πλέγμα τριών μονότονα φθινουσών ευθειών, που με την σειρά τους μοντελοποιούν την κλίση.

Σε γενικές γραμμές, η προσέγγιση αυτή είναι ιδιαίτερα απλή και ελκυστική αφού επιτρέπει την εύκολη υλοποίηση συστημάτων που βασίζονται στην συγκεκριμένη θεώρηση. Ωστόσο ένα από τα βασικά της μειονεκτήματα είναι η αδυναμία της να μοντελοποιήσει μικρές αλλαγές του F0, σε επίπεδο μικροπροσωδίας, οι οποίες διατηρούν σημαντικό ρόλο στην φυσικότητα του συνθετικού λόγου, όπως έχει γίνει φανερό από σειρά επιστημονικών εργασιών.

3.2.4 Γεννετικές Μέθοδοι: Μοντέλο Tilt RFC

Το μοντέλο αυτό παρουσιάστηκε το 1994 και η βασική του λειτουργία βασίζεται στην μοντελοποίηση της F0 σε τρία επίπεδα, ένα επίπεδο για την F0, ένα ενδιάμεσο επίπεδο και ένα φωνολογικό επίπεδο. Επιχειρείται η μοντελοποίηση και των δύο κατευθύνσεων, τόσο από την καμπύλη F0 προς το φωνολογικό επίπεδο όπου γίνεται ουσιαστικά αναπαράσταση του εκφερόμενου λόγου μέσω των φωνημάτων, όσο και από το φωνολογικό επίπεδο προς την F0. Το ενδιάμεσο επίπεδο εισάγει τρία βασικά στοιχεία, την ανύψωση, την πτώση και την σύνδεση (rise, fall and connection). Το φωνολογικό επίπεδο χρησιμοποιεί επίσης διαφορετικά στοιχεία, τον υψηλό και τον χαμηλό τόνο, το στοιχείο της σύνδεσης και το στοιχείο της ανύψωσης στα όρια των φράσεων.

Η βασική θεώρηση υποστηρίζει ότι κάθε καμπύλη F0 μπορεί να διαιρεθεί σε μία γραμμική σειρά μη επικαλυπτόμενων συνεχών τμημάτων που αντίστοιχα μπορούν να χαρακτηρισθούν ως rise, fall και connection. Έχοντας παράλληλα εκφράσει τις εξισώσεις ορισμού για τα διαφορετικά αυτά τμήματα, μπορεί κανείς να ορίσει διαφορετική κλίμακα τόσο σε χρόνο, όσο και πλάτος. Σε γενικές γραμμές, η προσέγγιση αυτή είναι αρκετά αποδοτική και βασίζεται περισσότερο σε μαθηματική προσέγγιση μηχανικής μοντελοποίησης παρά φωνολογικής, ξεπερνώντας αρκετά από τα ενδογενή προβλήματα γλωσσολογικών συμβάσεων. Ωστόσο στο συγκεκριμένο μοντέλο υπάρχουν παραλείψεις που προκύπτουν από την γενική θεώρηση της καμπύλης, παραβλέποντας ειδικές περιοχές περικειμένου (συμφραζόμενα) [Taylor1994].

3.2.5 Μοντελοποίηση Επιτονισμού – Πρώτα Συμπεράσματα

Όλα τα παραπάνω μοντέλα είναι οι βασικότεροι εκπρόσωποι ενός μεγάλου εύρους προσεγγίσεων και θεωριών γύρω από την μοντελοποίηση του επιτονισμού. Το κάθε ένα παρουσιάζει διαφορετικά πλεονεκτήματα και μειονεκτήματα, με αποτέλεσμα να μην έχει υπάρξει κάποιο μοντέλο που να έχει επικρατήσει ολοκληρωτικά. Άλλωστε η παράλληλη ύπαρξη τόσων διαφορετικών μεταξύ τους προσεγγίσεων για την ίδια εργασία υποδηλώνει το γεγονός ότι δεν έχει υπάρξει κάποια σύγκλιση ως προς μια κοινά αποδεκτή μέθοδο μοντελοποιήσης του επιτονισμού. Η χρήση του καθενός μοντέλου εξαρτάται άμεσα από τον σκοπό που ο ερευνητής επιθυμεί να εξυπηρετήσει. Στην περίπτωση της σύνθεσης φωνής εξάλλου, τα δεδομένα έχουν αλλάξει σημαντικά τα τελευταία χρόνια, με την τεχνολογία να έχει εξελιχθεί αλματωδώς ειδικότερα την τελευταία δεκαετία.

Η δική μας προσέγγιση [Giannopoulos et al., 2003] διαφέρει από τις προαναφερθείσες μεθόδους, στο ειδικό χαρακτηριστικό ότι «*κωδικοποιεί*» τα προσωδιακά χαρακτηριστικά ενός συνόλου ηχογραφήσεων, ανεξάρτητα του τρόπου εκφώνησης ή του ειδικού ύφους (genre) που αυτές μπορεί να περιέχουν, λαμβάνοντας υπόψη την απαραίτητη γλωσσολογική πληροφορία. Κατά αυτόν τον τρόπο, η συγκεκριμένη μέθοδος επιτρέπει την άμεση και αποδοτική μοντελοποίηση οποιουδήποτε υποκείμενου προσωδιακού ύφους, εφόσον φυσικά αυτό διατηρείται με συνέπεια καθ' όλο το μήκος των ηχογραφήσεων. Με άλλα λόγια, η συγκεκριμένη μέθοδος ουσιαστικά δημιουργεί διαφορετικά μοντέλα για κάθε ομιλητή αλλά και το ύφος της ομιλίας που διατηρεί στο σύνολο των ηχογραφήσεων του [Kamp1993], [Hirst2000].

3.3 Η αξία του μοντέλου επιτονισμού στην σύνθεση φωνής

Η μοντελοποίηση του επιτονισμού στο πλαίσιο της σύνθεσης φωνής, εκτός από ένα ανοικτό πεδίο έρευνας, αποτελεί και ένα από τα σημαντικότερα σημεία διαφοροποίησης των συνθετών φωνής. Συχνά η ποιότητα του τελικού συστήματος είναι συνάρτηση της επιτυχούς ή μη, μοντελοποίησης και εφαρμογής του επιτονισμού στο συνθετικό σήμα φωνής. Αν και αρχικά η τάση ήταν η αυστηρή μοντελοποίηση της προσωδίας και η εφαρμογή των παραγόμενων μοντέλων στο τελικό σήμα, σήμερα, η τάση αυτή έχει αλλάξει μαζί με την αλλαγή που έχει επέλθει στην τεχνολογία σύνθεσης φωνής. Η χρήση εκτενών βάσεων δεδομένων και αλγορίθμων επιλογής βέλτιστων ακουστικών μονάδων, έχει οδηγήσει στην υιοθέτηση αλγορίθμων που επιχειρούν την έμμεση μοντελοποίηση του επιτονισμού μέσα από τα ίδια τα δεδομένα εκπαίδευσης του συστήματος. Με άλλα λόγια η τάση σήμερα οδηγεί σε υιοθέτηση περισσότερο δημιουργικών τεχνικών με όσο το δυνατό μικρότερη ψηφιακή επεξεργασία και περισσότερη ελευθερία στο σύστημα για παραγωγή πλουσιότερων μοντέλων [Kohler1997], [Liberman1984].

3.4 Η προσέγγισή μας

Η προσέγγισή μας βασίζεται σε έναν αλγόριθμο που ανήκει στην κατηγορία των γεννετικών προσεγγίσεων [Cambell1994] [Campione2000], [Klatt1976], [Nakai1997].. Προσπαθεί να μοντελοποιήσει τα δεδομένα που έχει στη διάθεσή του το σύστημα και να προσφέρει μέσω του αλγορίθμου της βέλτιστης επιλογής την δυνατότητα στο σύστημα να παράγει πλούσια μοντέλα προσωδίας εμμέσως, με την βοήθεια του υποσυστήματος βέλτιστης επιλογής ακουστικών μονάδων. Η παραγόμενη καμπύλη προσωδίας δεν εφαρμόζεται αυστηρά στο τελικό σήμα φωνής, αλλά με τρόπο που επιτρέπει την διατήρηση της μικροπροσωδίας στο τελικό συνθετικό σήμα αλλά και με την ελάχιστη δυνατή παραμόρφωση στο τελικό αποτέλεσμα [Giannopoulos et al., 2003].

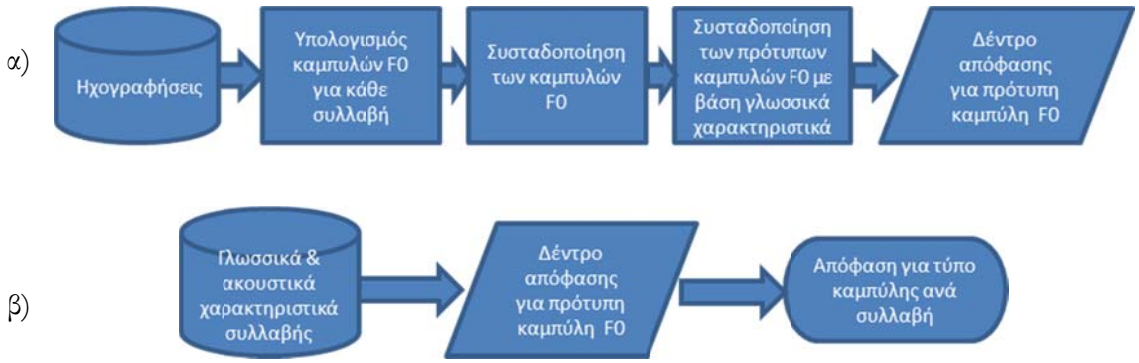
3.4.1 Θεωρία πίσω από την προσέγγιση μας

Αν και η προσωδία περιλαμβάνει τόσο τον επιτονισμό όσο και τις αντίστοιχες εντάσεις και διάρκειες των ακουστικών μονάδων, η προσέγγισή μας περιλαμβάνει άμεσα μόνο την μοντελοποίηση του επιτονισμού ενώ έμμεσα περιλαμβάνει τόσο τον ρυθμό όσο και την ένταση, μέσω της βέλτιστης επιλογής διφώνου που λαμβάνει υπόψη τα σημεία στίξης και άλλες παραμέτρους που συμβάλλουν στην ποικιλότητα των διαρκειών και της έντασης των συλλαβών και

τελικά στον καθορισμό του τελικού ρυθμού. Η συγκεκριμένη θεώρηση επιβεβαιώθηκε πειραματικά τόσο κατά την ανάλυση όσο και κατά την σύνθεση λόγου για διαφορετικούς ομιλητές.

Η προσέγγισή μας είναι κυρίως αναλυτική και βασίζεται στην μοντελοποίηση των δεδομένων της ίδιας βάσης που το τελικό σύστημα σύνθεσης φωνής χρησιμοποιεί ως πρωτογενές υλικό για την σύνθεση ομιλίας, βασιζόμενη ωστόσο σε θεωρήσεις που αφορούν σημαντικά σημεία τόσο στο πρωτεύον μορφοσυντακτικό επίπεδο, όσο και στο δευτερεύον φωνολογικό, γεγονός που της προσδίδει έναν επιπλέον γενετικό χαρακτήρα. Η συγκεκριμένη μεθοδολογία αποτελεί αναπόσπαστο κομμάτι της διαδικασίας της βέλτιστης επιλογής ακουστικών μονάδων και ουσιαστικά δεν μπορεί να αποτελέσει γενικό περιγραφικό μοντέλο επιτονισμού, αλλά συνδυασμό βέλτιστης επιλογής διφώνου με την χρήση ενός γενικότερου μοντέλου αναπαράστασης, που παρουσιάζει ωστόσο χαρακτηριστικά τα οποία είναι σε μεγάλο βαθμό ανεξάρτητα από την ίδια την γλώσσα.

Στο διάγραμμα που ακολουθεί περιγράφεται σχηματικά η ροή της μεθόδου μοντελοποίησης του επιτονισμού στο σύστημά μας στα διαφορετικά επίπεδα.



Σχήμα 22: Το διάγραμμα ροής για την εκπαίδευση της μηχανής προσωδίας (α) και της διαδικασίας παραγωγής μοντέλου προσωδίας κατά την εκτέλεση (β).

Ο συγκεκριμένος αλγόριθμος κάνει χρήση τόσο του γλωσσικού επιπέδου του σώματος κειμένου όσο και του ακουστικού επιπέδου, συνδυάζοντας τα ικανοποιητικά. Οι πρότυπες καμπύλες θεμελιώδους συχνότητας για τις συλλαβές εξάγονται από το ίδιο το ακουστικό επίπεδο και στην

συνέχεια χρησιμεύουν για την περαιτέρω συσταδοποίηση των γλωσσικών χαρακτηριστικών του σώματος κειμένου. Με την μέθοδο της συσταδοποίησης μέσω CART (Classification and Regression Trees) [Breiman et al. (1984)] μπορούμε στην συνέχεια να αναπαράγουμε τα μοντέλα αυτά από γλωσσικά χαρακτηριστικά και μόνο. Ο συγκεκριμένος μηχανισμός απαιτεί περισσότερη ανάλυση, η οποία ακολουθεί στις επόμενες παραγράφους.

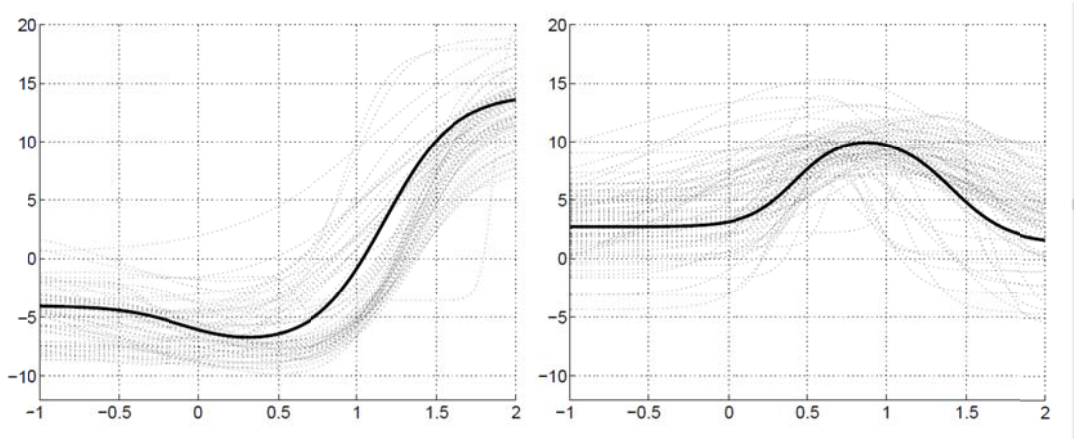
3.4.2 Υπολογισμός των πρότυπων καμπυλών θεμελιώδους συχνότητας

Στο μοντέλο που έχουμε αναπτύξει η βασική δομική μονάδα που χρησιμοποιούμε είναι η γραμματική συλλαβή η οποία αποτελεί και την βάση για το μοντέλο μας. Η εξαγωγή των συλλαβών έγινε αυτόματα μέσω της φωνητικής αναπαράστασης της ελληνικής γλώσσας. Πιο συγκεκριμένα, από το φωνητικά μετεγγραμμένο σώμα κειμένου εξάγουμε όλες τις δυνατές συλλαβές του τύπου CV, CCV, VC, CCCV, CCCCV CCCVC, CCVCC[2], οι οποίες και αποτέλεσαν την βάση για την ομαδοποίηση (clustering) των διαφορετικών καμπυλών F0. Η αναζήτηση των διαφορετικών συλλαβών έγινε εξαντλητικά με την μέθοδο της ανίχνευσης συλλαβών οι οποίες κρίνονταν ως υπαρκτές στην περίπτωση μόνο όπου λέξεις της ίδιας γλώσσας αρχίζουν με πρώτη την συλλαβή αυτή.

Κατόπιν της εξαγωγής όλων των δυνατών φωνητικών συλλαβών, για κάθε συλλαβή στο φωνητικό σώμα, υπολογίζουμε την λειασμένη καμπύλη F0, την οποία και στην συνέχεια κανονικοποιούμε τόσο ως προς το μήκος, όσο και προς το εύρος των τιμών. Έτσι, οι παραγόμενες καμπύλες έχουν τιμές από τον χρόνο 0 μέχρι τον χρόνο 1, και οι τιμές πλάτους που μπορεί να λάβει η F0 είναι όλες μέσα στο διάστημα [0,1]. Για την κανονικοποίηση των καμπυλών χρησιμοποιήσαμε την παρακάτω σχέση, όπου x(n) η τιμή της καμπύλης της F0 την χρονική στιγμή n.

$$x_{norm}(n) = 2\left(x(n) - \frac{\min(x)}{\max(x)} - \min(x)\right) - 1$$

[2] Όπου C εννοείται σύμφωνο (Consonant) και όπου V εννοείται φωνήεν (Vowel)



Σχήμα 23: Απεικόνιση συσταδοποίησης καμπυλών προσωδίας σε συλλαβές. Για κάθε συστάδα καμπυλών απεικονίζεται με σκούρα γραμμή η μέση καμπύλη.

Στην συνέχεια, ομαδοποιήσαμε τις καμπύλες αυτές με την τεχνική των K-means, με μετρική απόσταση την ευκλείδεια απόσταση, ενώ το K καθορίστηκε ανευρετικά, έτσι ώστε κάθε παραγόμενη κατηγορία να έχει σημαντικό αριθμό μελών. Στην συνέχεια, οι κατηγορίες αυτές αποτέλεσαν την βάση για μία εκ νέου εκπαίδευση ενός συστήματος CART[3]. Το CART αυτό, χρησιμοποιώντας τόσο φωνητικά χαρακτηριστικά (π.χ. φωνήματα που προηγούνται ή ακολουθούν) όσο και χαρακτηριστικά γλωσσικής σημασίας (π.χ. σημεία στίξης, απόσταση από αυτά, μέρος του λόγου, κτλ.) χρησιμεύει στην αυτόματη απόφαση για ποιο από τα παραπάνω clusters πρέπει να επιλεχθεί.

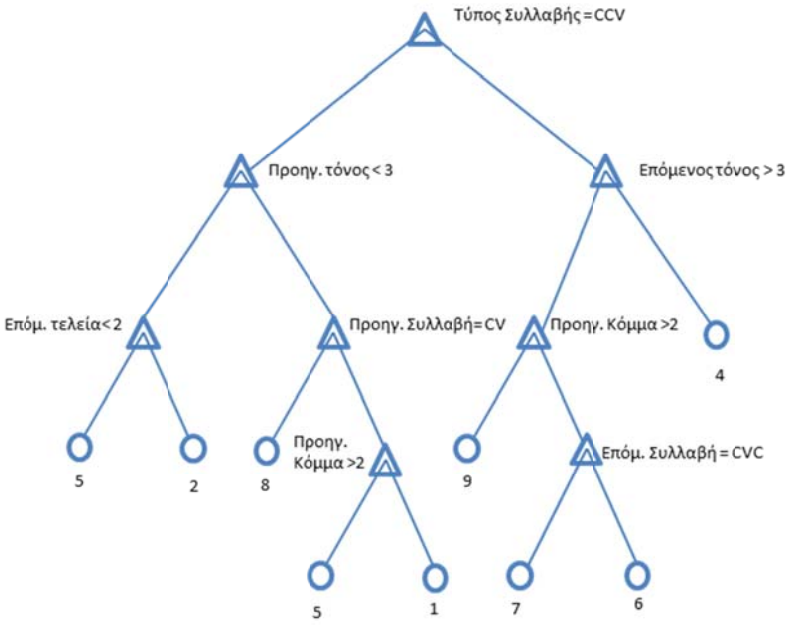
Τα χαρακτηριστικά που χρησιμοποιούμε για την «οδήγηση» αυτού του δένδρου απόφασης είναι τα παρακάτω:

- ο τύπος της τρέχουσας συλλαβής
- ο τύπος των δύο προηγούμενων και των δύο επόμενων συλλαβών
- η απόσταση από το προηγούμενο και επόμενο σημείο στίξης

[3] Classification and Regression Trees

- η απόσταση από τον προηγούμενο και επόμενο τόνο

Η εκπαίδευση του συστήματος απαιτεί την ύπαρξη ηχογραφήσεων επισημειωμένων σε φωνολογικό επίπεδο, έτσι ώστε να εξαχθούν τα διαφορετικά μοντέλα για κάθε συλλαβή. Η επισημείωση απαιτείται να είναι ακριβής, αν και η ύπαρξη μεμονωμένων παραπειστικών δεδομένων (outliers) αντιμετωπίζεται με την χρήση της μη γραμμικής μοντελοποίησης που επιτυγχάνει η μέθοδος CART. Αυτό άλλωστε το χαρακτηριστικό είναι ένα από τα βασικά πλεονεκτήματα της μεθόδου αυτής και ένας βασικός λόγος για τον οποίο η συγκεκριμένη μέθοδος επιτυγχάνει την αποδοτική μοντελοποίηση της καμπύλης επιτονισμού με συνέπεια.



Σχήμα 24: Δέντρο απόφασης για τον τύπο πρότυπης καμπύλης F0 που πρέπει να πάρει μία συλλαβή CCV ανάλογα με τα περικείμενα γλωσσικά χαρακτηριστικά της.

3.4.3 Υπολογισμός της καμπύλης επιτονισμού

Ο υπολογισμός της τελικής καμπύλης επιτονισμού πραγματοποιείται με παράθεση των επιμέρους καμπυλών επιτονισμού των ακουστικών μονάδων που έχουν επιλεγεί από τον αλγόριθμο βέλτιστης επιλογής. Οι επιμέρους καμπύλες παρεμβάλονται και κανονικοποιούνται ως προς την διάρκειά τους, έτσι ώστε η τελική τους διάρκεια να συμφωνεί με την διάρκεια των αντίστοιχων ακουστικών

μονάδων που παρατίθενται, ενώ η τελική καμπύλη επιτονισμού προκύπτει από την λείανση της τελικής καμπύλης μέσω ενός βαθυπερατού φίλτρου ενδιάμεσης τιμής.

Για την παρεμβολή των σημείων της καμπύλης της θεμελιώδους συχνότητας όπου δεν είναι δυνατόν να υπολογιστεί επειδή το σήμα τοπικά είναι απεριοδικό, χρησιμοποιούμε την σχέση της γραμμικής παρεμβολής

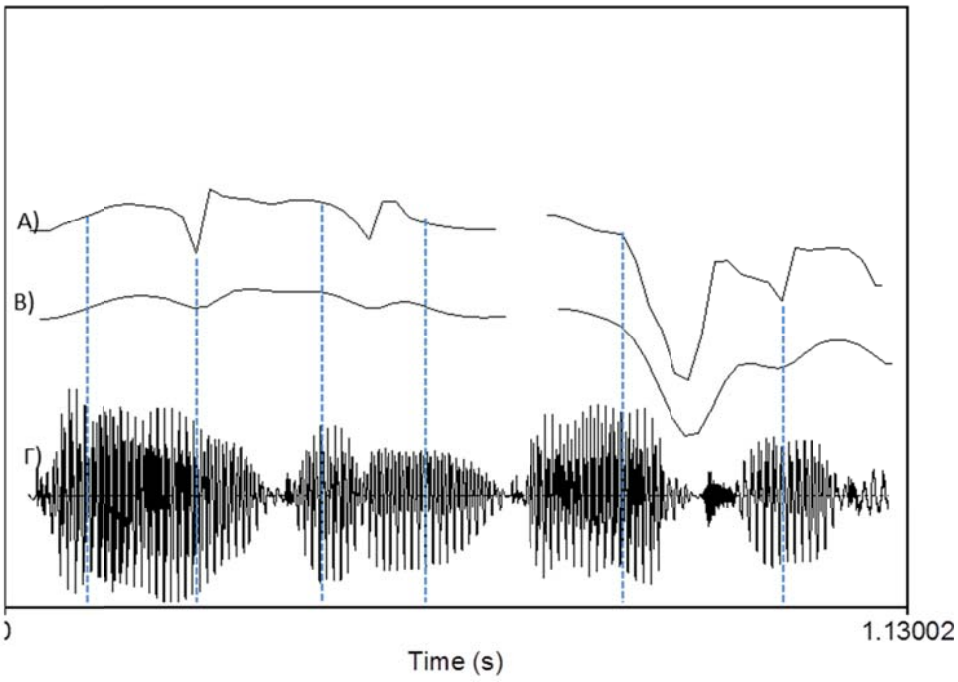
$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0}$$

Που οδηγεί στον υπολογισμό του σημείου y με βάση τα γνωστά σημεία$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$ και $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$

$$y = y_0 + \frac{(x - x_0)y_1 - (x - x_0)}{x_1 - x_0}$$

Το βαθυπερατό φίλτρο ενδιάμεσης τιμής που χρησιμοποιούμε χρησιμοποιεί παράθυρο 10msec, διατηρώντας την ενδιάμεση τιμή του σήματος για το μήκος του παραθύρου.

Παρακάτω μπορεί κανείς να δει την καμπύλη της θεμελιώδους συχνότητας ενός συνθετικού σήματος φωνής κατά την παράθεση και μετά την λείανση της F0.



Σχήμα 25: Απεικόνιση συνθετικού σήματος φωνής με παράθεση (Γ), της καμπύλης Pitch όπως προκύπτει από την παράθεση των επιμέρους καμπυλών των ακουστικών μονάδων (Α) και η λειασμένη καμπύλη Pitch όπως προκύπτει από βαθυπερατό φίλτρο (Β). Οι κατακόρυφες ευθείες επισημειώνουν τα όρια των διφωνημάτων (ακουστικών μονάδων) που επιλέχθηκαν για την σύνθεση της συγκεκριμένης φράσης (/Golden gate A/).

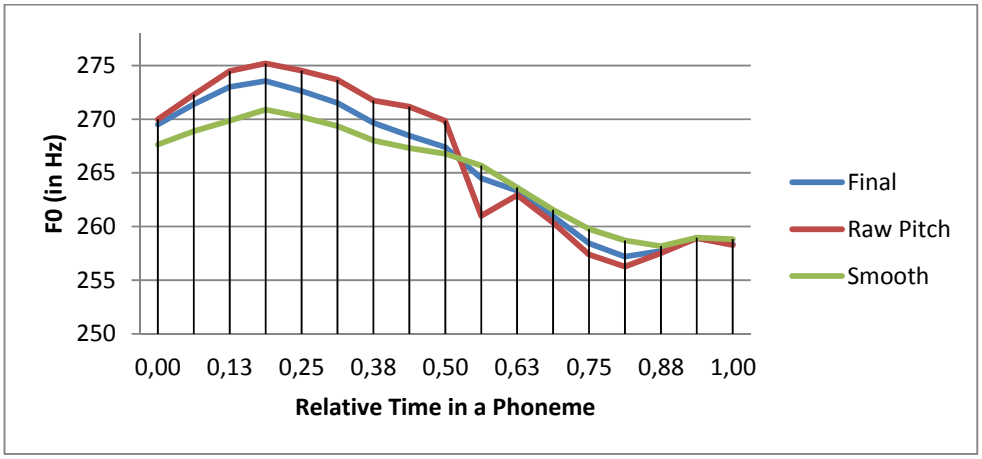
3.4.4 Εφαρμογή της καμπύλης επιτονισμού

Ένας από τους σημαντικότερους παράγοντες που ευθύνονται συχνότερα για την ρομποτική χροιά ενός συνθέτη φωνής είναι ο τρόπος εφαρμογής της καμπύλης επιτονισμού στο τελικό συνθετικό σήμα φωνής. Όπως έχουμε αναφέρει και προηγουμένως, ένας βασικός περιορισμός της μεθόδου σύνθεσης με παράθεση ακουστικών μονάδων στο πεδίο του χρόνου είναι το εύρος τιμών που μπορεί να λάβει η χρονική αλλά και τονική μετατροπή του σήματος. Με άλλα λόγια, υψηλές αλλαγές της θεμελιώδους συχνότητας ή των διαρκειών του σήματος φωνής επιφέρουν σημαντική παραμόρφωση στο σήμα [Vepa2004], [Wightman2000].

Παράλληλα, είναι κοινή γνώση ότι η μικροπροσωδία στο σήμα της φωνής αποτελεί συστατικό που σχετίζεται άμεσα με την φυσικότητα της συνθετικής φωνής. Αυτός άλλωστε είναι ο λόγος που σε πολλά συστήματα σύνθεσης φωνής που κάνουν χρήση ενός ρητού (explicit) μοντέλου επιτονισμού,

το επίπεδο της φυσικότητας στο τελικό σήμα φωνής υπολοίπεται σημαντικά της φυσικής ηχογράφησης. Η προσέγγιση που εμείς έχουμε επιλέξει είναι η εφαρμογή ενός εμμέσου μοντέλου επιτονισμού και όχι ρητού, λαμβάνοντας υπόψη και ενσωματώνοντας την μικροπροσωδία των φωνημάτων τοπικά. Η τελική καμπύλη μοντελοποίησης του επιτονισμού στο σήμα προκύπτει από την ενσωμάτωση της σχετικής μικροπροσωδίας των φωνημάτων στο σήμα, δίνοντας έμφαση στα έμφωνα τμήματα ήχου.

Η ενσωμάτωση της μικροπροσωδίας στην τελική καμπύλη θεμελιώδους συχνότητας λαμβάνει χώρα ανά φώνημα στο τελικό σήμα συνθετικής φωνής. Για την περιοχή κάθε φωνήματος, ο αλγόριθμός μας ενσωματώνει την τοπική μικροπροσωδία με ένα τοπικό βάρος που αυξάνει στα άκρα του φωνήματος και μειώνεται στο μέσο του, όπου άλλωστε υπάρχει και το σημείο της ένωσης μεταξύ των διαδοχικών ακουστικών μονάδων (διφωνήματα). Το βάρος της μικροπροσωδίας στο σημείο παράθεσης των ακουστικών μονάδων είναι μικρό έτσι ώστε να αποφεύγεται η εισαγωγή οποιασδήποτε ασυνέχειας στην θεμελιώδη συχνότητα (βλ. Σχήμα 26).



Σχήμα 26: Η ενσωμάτωση της μικροπροσωδίας κατά το μήκος ενός φωνήματος στο συνθετικό σήμα φωνής. Η τοπική τιμή της προσωδίας του φυσικού σήματος προστίθεται με βάρος που εξαρτάται από την απόσταση του σημείου από το μέσο του φωνήματος, ώστε το μέσο όπου πραγματοποιείται η ένωση διαφορετικών ακουστικών μονάδων να μην περιέχει τοπικές ασυνέχειες στην F0.

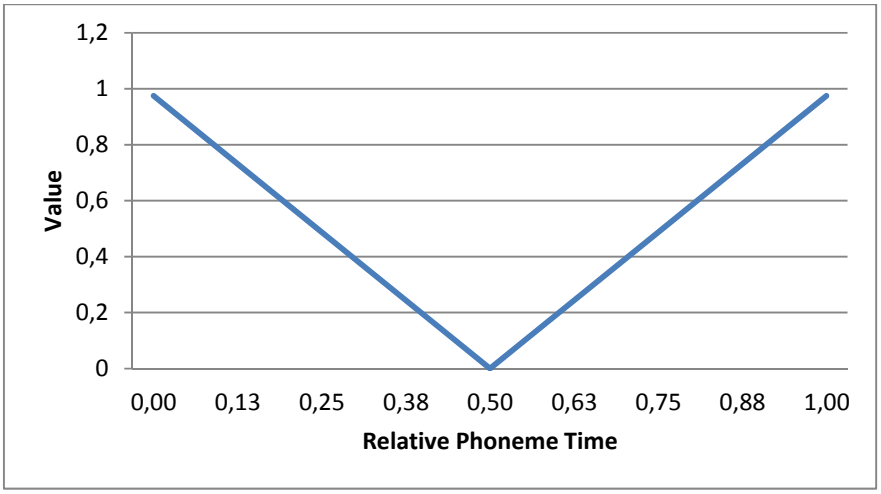
Η τελική καμπύλη της προσωδίας προκύπτει από την παρακάτω σχέση

$$F0_f^t = F0_S^t + w_u^t(F0_L^t - F0_s^t)$$

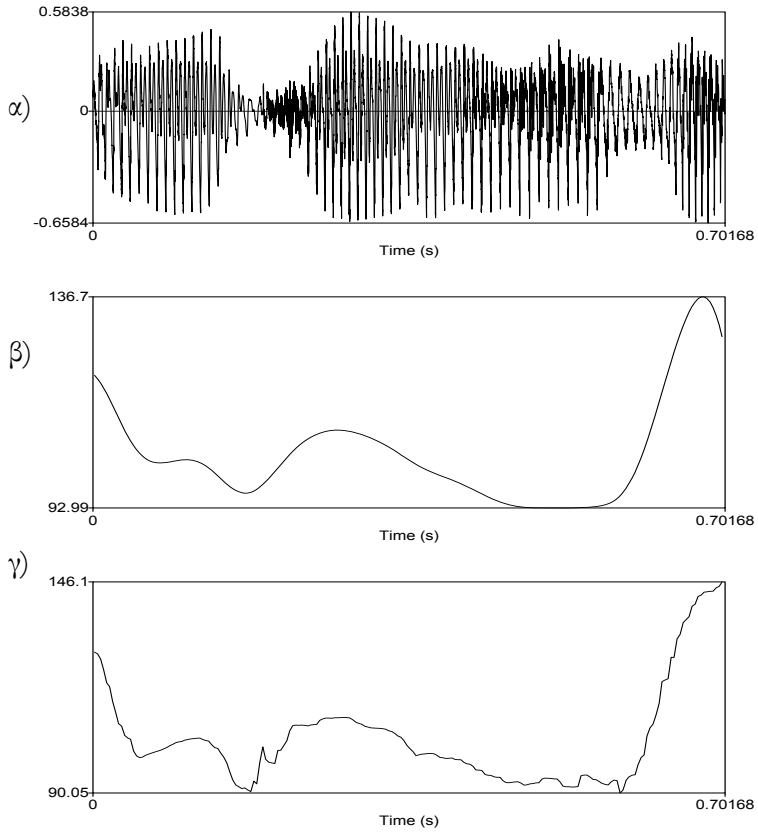
Όπου $F0_s$ είναι η τιμή της λειασμένης καμπύλης F0, $F0_L$ είναι η τοπική τιμή της F0 στο σήμα φωνής χωρίς καμία επεξεργασία, και $F0_s$ είναι η τιμή της τελικής θεμελιώδους συχνότητας, μετά την ενσωμάτωση της τοπικής μικροπροσωδίας. Το βάρος w_u^t πόσο κοντά στην αρχική F0 του φυσικού σήματος πρόκειται να είναι η τελική καμπύλη του F0 και μεταβάλλεται ανάλογα με την απόστασή του από το μέσο του φωνήματος σύμφωνα με την σχέση

$$w_u(t) = a_u|t - t_o|$$

Όπου a_u είναι μία σταθερά που εξαρτάται από το είδος του φωνήματος u και t_o είναι ο χρόνος που αντιστοιχεί στο μέσο του φωνήματος.



Σχήμα 27: Τυπική καμπύλη για την μεταβλητή w(t) από την οποία εξαρτάται το μέγεθος της μικροπροσωδίας του φυσικού σήματος που ενσωματώνεται στην τελική καμπύλη της θεμελιώδους συχνότητας. Η χρονική στιγμή 0,5 αντιστοιχεί στο μέσο του φωνήματος και το σημείο όπου πραγματοποιείται η παράθεση των δύο διαδοχικών ακουστικών μονάδων από την βάση δεδομένων.



Σχήμα 28: Ο υπολογισμός της τελικής καμπύλης θεμελιώδους συχνότητας για ένα έμφωνο τμήμα φωνής (α). Η καμπύλη (β) έχει δημιουργηθεί από την παράθεση και λείανση των επιμέρους καμπυλών θεμελιώδους συχνότητας για κάθε συλλαβή. Η καμπύλη (γ) προκύπτει από την ενσωμάτωση της μικροπροσωδίας των έμφωνων ήχων στην λειασμένη τελική καμπύλη μοντελοποίησης.

3.4.5 Αποτελέσματα

Η παραπάνω μέθοδος επιτρέπει στο σύστημά μας να παράγει καμπύλες επιτονισμού που αφενός δεν δημιουργούν παραμορφώσεις στο σήμα φωνής κατά την εφαρμογή τους, αλλά και αφετέρου διατηρούν την μικροπροσωδία στις περιοχές του σήματος φωνής όπου απαιτείται. Όπως έχει ήδη αναφερθεί, σημαντική καινοτομία του προτεινόμενου αλγορίθμου αποτελεί ο αποτελεσματικός συνδυασμός στοιχείων και χαρακτηριστικών και από τα δύο επίπεδα, το γλωσσικό και το ακουστικό, παρέχοντας έναν απλό τρόπο για την αναπαραγωγή των μοντέλων αυτών.

Ένα επίσης βασικό πλεονεκτήματα της μεθόδου αυτής είναι η εξ' ολοκλήρου αυτόματη διαδικασία που περιλαμβάνει, τόσο για την εκπαίδευση, όσο και για την παραγωγή των μοντέλων. Τα αποτελέσματα της χρήσης των μοντέλων αυτών που εξαρτώνται αποκλειστικά από τα πρωτογενή δεδομένα, οδηγούν σε ποιοτική συνθετική ομιλία, ενώ η προσαρμογή τους σε νέα δεδομένα επιτυγχάνεται με εντελώς αυτόματους τρόπους. Με αυτόν τον τρόπο, η συγκεκριμένη προσέγγιση είναι εύκολο να προασρμοστεί και αναμένεται εξίσου αποτελεσματική σε εφαρμογές που περιλαμβάνουν για παράδειγμα δεδομένα για εκφώνηση παραμυθιών, για εκφώνηση ειδήσεων, κ.ο.κ. Η χρήση μίας τόσο μικρής δομικής μονάδας όπως είναι η συλλαβή, φαίνεται ότι είναι ιδανική για την μοντελοποίηση της προσωδίας συλλαβικών γλωσσών όπως είναι η Ελληνική αλλά και η Βουλγαρική, η οποία επίσης αποτέλεσε αντικείμενο της ερευνητικής και αναπτυξιακής μας προσπάθειας, με εξίσου υψηλής ποιότητας αποτελέσματα [Chalamandaris et al., 2009a] [Raptis et al., 2009a] .

3.4.6 Συζήτηση – Θέματα προς διερεύνηση

Η μέθοδος που περιγράψαμε για την μοντελοποίηση της καμπύλης επιτονισμού για τον συνθέτη φωνής παρουσιάζει σημαντικές διαφοροποιήσεις από άλλες προσεγγίσεις ενώ προσφέρει σημαντική ευελιξία για εύκολη προσαρμογή σε άλλες γλώσσες και φωνές. Η συμπεριφορά του συγκεκριμένου μοντέλου είναι ιδιαίτερα αποτελεσματική στην περίπτωση όπου το πεδίο εφαρμογής και το ύφος ομιλίας είναι γενικά ουδέτερο, χωρίς ιδιαίτερη εκφραστικότητα ή χρωματισμό. Παρ' όλα αυτά είναι προφανές ότι η συγκεκριμένη μέθοδος δεν μπορεί να αντιμετωπίσει με την ίδια αποτελεσματικότητα ύφη ομιλίας με μεγαλύτερη εκφραστικότητα ή συναίσθημα αφού η γλωσσική πληροφορία που λαμβάνεται υπόψη για την μοντελοποίηση των προσωδιακών φαινομένων δεν περιλαμβάνει γραμματική, συντακτική ή σημασιολογική πληροφορία. Σε μια τέτοια περίπτωση απαιτείται περαιτέρω έρευνα και προσαρμογή της μεθόδου ούτως ώστε να περιλαμβάνει σημαντικά χαρακτηριστικά που καθορίζουν την εκφραστικότητα μιας λέξης ή φράσης, όπως είναι το μέρος του λόγου, το είδος της πρότασης, οι λέξεις κλειδιά αλλά και σημασιολογικά δεδομένα.

3.5 Βιβλιογραφία Κεφαλαίου

[Arvaniti2010] Arvaniti, A. & Baltazani, M. Intonational analysis and prosodic annotation of Greek spoken corpora. In S.-A. Jun (Ed.), Prosodic Models and Transcription: Towards Prosodic Typology. (2010) Oxford University Press.

[Beckman1997] Beckman, M. E. & Ayers, G. M. (1997). Guidelines for ToBI Labelling (version 3, March 1997). Technical report, Ohio State University.

[Beckman1996] Beckman, M. E. & Jun, S.-A. (1996). K-ToBI (KOREAN ToBI) Labeling Conventions (version 2.1, revised November 1996). Technical report, Ohio State University.

[Beckman1986] Beckman, M. E. & Pierrehumbert, J. B. (1986). Intonational structure in English and Japanese. Phonology Yearbook, 3, 255-310.

[Bloomington1951]Bloomington, Indiana. Distributed by Indiana Linguistics University Club Publications. Trager, G. & Smith, H. L. (1951). An Outline of English Structure. Norman, Oklahoma: Battenburg Press.

[Bolinger1978] Bolinger, D. (1978). Intonation across languages. In J. Greenberg (Ed.), Universals of Human Language, volume 2 (pp. 471-524). Palo Alto, CA: Stanford University Press.

[Bolinger1989] Bolinger, D. (1989). Intonation and its Uses. Palo Alto, CA: Stanford University Press.

[Butzberger1990]Butzberger, J., Ostendorf, M., Price, P., & Shattuck-Hufnagel, S. (1990). Isolated word intonation recognition using hidden Markov models. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 773-776). Albuquerque, NM.

[Campbell1996] Campbell, N. (1996). Autolabelling Japanese ToBI. In Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP) (pp. 2399- 2402). Philadelphia, PA.

[Campbell1994] Campbell, W. (1994). Combining the use of duration and F0 in an automatic analysis of dialogue prosody. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), volume 3 (pp. 1111-1114). Yokohama, Japan.

[Campione2000] Campione, E., Hirts, D., & VÈronis, J. (2000). Automatic stylisation and modelling of French and Italian intonation. In A. Botinis (Ed.), Intonation: Analysis, Modelling and Technology, volume 15 of Text, Speech and Language Technology (pp. 185-208). Dordrecht: Kluwer.

[Chomsky1968] Chomsky, N.& Halle, M. (1968). The Sound Pattern of English. New York: Harper and Row.

[Crystal1969] Crystal, D. (1969). Prosodic systems and intonation in English. Cambridge: Cambridge University Press.

[Dimou & Chalamandaris, 2006] A. L. Dimou and A. Chalamandaris, "Language identification from suprasegmental cues; examining the role of rhythm in the identification of a Greek dialect", in Proc. La comunicazione parlata / Spoken Communication, February, , Italy, 2006

[Dimou & Chalamandaris, 2008] A. L. Dimou and A. Chalamandaris, "Is idiom identification possible from prosodic information? An experimental approach for the Greek language", in Proc. 4th Intl Conf. Speech Prosody 2008, pp. 759-762 (2008)

[Fujisaki1992] Fujisaki H.: The role of quantitative modeling in the study of intonation, Proc. Int. Symp. Japanese Prosody (1992) pp. 163–174

[Fujisaki1981] Fujisaki, H. (1981). Dynamic characteristics of voice fundamental frequency in speech and singing - Acoustical analysis and physiological interpretations. In Proceedings of the 4th F.A.S.E Symposium on Acoustics and Speech, volume 2 (pp. 57-70).

[Fujisaki1983] Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. Mac-Neilage (Ed.), The Production of Speech (pp.39-55). Berlin: Springer.

[Fujisaki1997] Fujisaki, H. (1997). Prosody, models, and spontaneous speech. In Y. Sagisaki, N. Campbell, & N. Higuchi (Eds.), Computing Prosody (pp. 27-42). New York: Springer.

[Giannopoulos et al., 2003] G. Giannopoulos, A. Chalamandaris, S-E. Fotinea, T. Athanaselis, G. Carayannis, "Analysis and modelling of the Carrier Declination for the Greek language", in Proc. of the 15th International Congress of Phonetic Sciences - ICPhS03, 3-9 August 2003, Barcelona, pp. 555-558.

[Hirst1998] Hirst, D. & Di Christo, A. (1998). Intonation Systems: A Survey of Twenty Languages. Cambridge, UK: Cambridge University Press.

[Hirst1983] Hirst, D. J. (1983). Structures and categories in prosodic representations. In A. Cutler & D. R. Ladd (Eds.), Prosody: Models and Measurements (pp. 93-109). Berlin: Springer.

[Hirst2000] Hirst, D., Di Christo, A., & Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. Internet paper. http://194.57.187.30/_hirst/articles/2000 Hirst&al.pdf, (pp. 1-21).

[Kamp1993] Kamp, H. & Reyle, U. (1993). From Discourse to Logic. Dordrecht: Kluwer Academic Publishers.

[Klatt1976] Klatt D.H.: Linguistic use of segmental duration in English: Acoustic and perceptual evidence, J. Acoust. Soc. Am. 59, 1208–1221 (1976)

[Kohler1997] Kohler, K. J. (1997). Modelling prosody in spontaneous speech. In Y. Sagisaki, N. Campbell, & N. Higuchi (Eds.), Computing Prosody (pp. 187-210). New York: Springer.

[Kompe1997] Kompe, R. (1997). Prosody in Speech Understanding Systems. Lecture Notes in Arti_cial Intelligence, 1307. Berlin: Springer.

[Leben1973] Leben, W. (1973). Suprasegmental Phonology. PhD thesis, MIT.

[Leben1976] Leben, W. (1976). The tones in English intonation. Linguistic Analysis, 2, 69-107.

[Lehiste1970] Lehiste, I. (1970). Suprasegmentals. Cambridge, MA: MIT Press.

[Liberman1984] Liberman, M. & Pierrehumbert, J. B. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oerhle (Eds.), Language Sound Structure (pp. 157-233). Cambridge, MA: MIT Press.

[Liberman1977] Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. Linguistic Inquiry, 8, 249-336.

[Mixdorff2000] Mixdorff, H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. In Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 3 (pp. 1281-1284). Istanbul, Turkey.

[Mobius1993] Mobius, B. (1993). Ein quantitatives Model der deutschen Intonation: Analyse und Synthese von Grundfrequenzverl‰ufen. T¸bingen: Niemeyer.

[Mobius1993] Mobius, B., P‰tzold, M., & Hess,W. (1993). Analysis and synthesis of F0 contours by means of Fujisaki's model. Speech Communication, 13, 53-61.

[Nakai1995] Nakai, M., Singer, H., Sagisaka, Y., & Shimodaira, H. (1995). Automatic prosodic segmentation by F0 clustering using superposition modeling. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 624-627). Detroit, Michigan.

[Nakai1997] Nakai, M., Singer, H., Sagisaki, Y., & Shimodaira, H. (1997). Accent phrase segmentation by F0 clustering using superpositional modelling. In Y. Sagisaki, N. Campbell, & N. Higuchi (Eds.), Computing Prosody: Computational Models for Processing Spontaneous Speech chapter 22, (pp. 343-359). New York.

[Ostendorf1997] Ostendorf, M. & Ross, K. (1997). A multi-level model for recognition of intonation labels. In Y. Sagisaki, N. Campbell, & N. Higuchi (Eds.), Computing Prosody: Computational Models for Processing Spontaneous Speech chapter 19, (pp. 291-308). New York: Springer.

[Pierrehumbert1990] Pierrehumbert, J. & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse, chapter 14, (pp. 271-311). Intentions in Communication. MIT Press: Cambridge, MA.

[Pierrehumbert1980]Pierrehumbert, J. (1980). The Phonology and Phonetics of English Intonation. PhD thesis, MIT.

[Pierrehumbert1983]Pierrehumbert, J. B. (1983). Automatic recognition of intonation patterns. In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (pp. 85-90). Cambridge, Massachusetts: MIT.

[Schweitzer2004]Schweitzer A., B. Moebius: Exemplar-based production of prosody: Evidence from segment and syllable durations, Proc. Speech Prosody 2004 (Nara), ed. by B. Bel, I. Marlien (ISCA, Grenoble 2004)

[Silverman1992] Silverman K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg: TOBI: A standard for labeling English prosody, Proc. ICSLP'92 Banff (1992) pp. 867–870

[Taylor1994] Taylor, P. A. (1994). A Phonetic Model of Intonation in English. PhD thesis, University of Edinburgh,

[Vepa2004] Vepa J., S. King: Join cost for unit selection speech synthesis. In: Text-to-Speech Synthesis – New Paradigms and Advances, Professional Technical Reference, ed. by S. Narayanan, A. Alwan (Prentice-Hall, Upper Saddle River 2004) pp. 35–62, Chap. 3

[Wightman2000] Wightman, C.W., Syrdal, A. K., Stemmer, G., Conkie, A.,&Beutnagel, M. (2000). Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), volume 2 (pp. 71-74). Bejing, China.

4. Η ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΣΥΝΘΕΣΗΣ ΦΩΝΗΣ

Η βάση δεδομένων για τον συνθέτη ομιλίας με παράθεση ακουστικών μονάδων που κάνει χρήση τεχνικών επεξεργασίας στο πεδίο του χρόνου αποτελεί ένα από τα βασικότερα υποσυστήματα του συστήματος, η ποιότητα του οποίου εξαρτάται άμεσα από την ποιότητα της ίδιας της βάσης δεδομένων. Ο σχεδιασμός, αλλά και τα κριτήρια που πρέπει να πληροί η βάση δεδομένων εξαρτάται τόσο από την εφαρμογή που πρόκειται να υποστηρίξει το συγκεκριμένο σύστημα σύνθεσης φωνής, όσο και από την ίδια την υποκείμενη τεχνολογία σύνθεσης φωνής. Στο κεφάλαιο αυτό πραγματοποιείται αναφορά στις μελέτες που πραγματεύονται το συγκεκριμένο πεδίο της σύνθεσης φωνής, ενώ στην συνέχεια παρουσιάζουμε τόσο την δική μας προσέγγιση όσο και ειδικότερα θέματα που άπτονται της βάσης δεδομένων για το σύστημα συνθετικής ομιλίας στο οποίο αναφερόμαστε. Ειδική αναφορά γίνεται στην μέθοδο για την επιλογή των βέλτιστων

σημείων ανάλυσης (pitchmarks) για την επεξεργασία του σήματος στο πεδίο του χρόνου, καθώς επίσης και στον αυτόματο τεμαχισμό και κανονικοποίηση της ποιότητας των ηχογραφήσεων.

4.1 Εισαγωγή

Όπως έχει ήδη αναφερθεί στο κεφάλαιο 2 και στην γενική περιγραφή ενός συστήματος συνθετικής ομιλίας, η βάση δεδομένων αποτελεί αν όχι το βασικότερο, ένα από τα βασικότερα υποσυστήματα και διαδραματίζει καθοριστικό ρόλο στην τελική ποιότητα του συνθετικού σήματος φωνής. Το παραπάνω άλλωστε ισχύει σε μεγαλύτερο βαθμό για συνθέτες φωνής που βασίζονται στην χρήση προηχογραφημένου σώματος κειμένου και στη τεχνολογία παράθεσης ακουστικών μονάδων (corpus-based concatenative TTS) όπως είναι και το σύστημα που περιγράφουμε κι εμείς [Raptis et al., 2010] .

Η σχεδίαση του σώματος κειμένου προς ηχογράφηση, συχνά λέγεται ότι είναι η «αρχή και το τέλος», όσον αφορά στην ποιότητα του συνθετικού σήματος [Black1998] [Möbius2000]. Η τεχνολογία της σύνθεσης με παράθεση ακουστικών μονάδων (και στην συγκεκριμένη περίπτωση διφωνημάτων) παρουσιάζει περιορισμούς στην τελική ποιότητα της συνθετικής ομιλίας λόγω των παραμορφώσεων που συχνά επιφέρει η ψηφιακή επεξεργασία του σήματος φωνής. Μεγάλες φασματικές ή προσωδιακές ασυνέχειες είναι ιδιαίτερα δύσκολο να αντιμετωπιστούν με ψηφιακή επεξεργασία χωρίς παράλληλα να υπεισέρχονται παραμορφώσεις που είναι αισθητές από το ανθρώπινο αφτί. Κατά αυτόν τον τρόπο, η ποιότητα του τελικού συνθετικού σήματος εξαρτάται άμεσα από τα δεδομένα εκπαίδευσης του συστήματος, καθορίζοντας ουσιαστικά το εύρος της ψηφιακής επεξεργασίας που απαιτείται κατά την σύνθεση.

Το ηχογραφημένο σώμα κειμένου αποτελεί την βάση για το σύστημα σύνθεσης φωνής και σχεδιάζεται κατά τέτοιον τρόπο ώστε να παρουσιάζει την βέλτιστη κάλυψη σε διαφορετικά πεδία, όπως είναι η κάλυψη σε διφωνήματα, η επαρκής κάλυψη των προσωδιακών φαινομένων, η κάλυψη σημαντικών και συχνών λέξεων κ.α. [Möbius2000] [Zhu2002]. Από την αρχή της ανάπτυξης της τεχνολογίας σύνθεσης φωνής, το θέμα της σχεδίασης του σώματος κειμένου προς εκφώνηση αποτελεί ένα από τα σημαντικότερα ζητήματα στην διαδικασία αυτή. Οι προϋποθέσεις που απαιτούνται να πληρούνται σε ένα τέτοιο σώμα κειμένου έχουν αλλάξει σημαντικά με τον καιρό, ακολουθώντας την αλματώδη εξέλιξη που υπάρχει στην τεχνολογία σύνθεσης φωνής. Κατά αυτόν

τον τρόπο, και πάντα αναφορικά με την σύνθεση με παράθεση ακουστικών μονάδων (concatenation synthesis), οι απαιτήσεις άλλαξαν σημαντικά, και ενώ αρχικά η βασική απαίτηση ήταν η πλήρης κάλυψη σε διφωνήματα [Lenzo2000], με σκοπό την δημιουργία βάσης με ένα μοναδικό στιγμιότυπο από κάθε διφώνημα, σήμερα απαιτείται κάλυψη σε πολλά περισσότερα επίπεδα, όπως παραδείγματος χάριν στις συχνότερες λέξεις, προσωδιακά φαινόμενα κ.α. όπως αναφέραμε και προηγούμενως [Nagy2006].

4.2 Σχεδίαση σώματος κειμένου προς ηχογράφηση

Με την προϋπόθεση ότι η ακουστική μονάδα του συνθέτη φωνής είναι πλήρως ορισμένη, η σχεδίαση του σώματος κειμένου άμεσα ανάγεται σε ένα πρόβλημα μέγιστης κάλυψης ακουστικών μονάδων [Bozkurt2003]. Ορίζοντας ως *C* το σύνολο των ακουστικών μονάδων που πρέπει το σώμα κειμένου να καλύπτει, επιχειρούμε να υπολογίσουμε το σύνολο των ελάχιστων δυνατών προτάσεων που περιλαμβάνουν το μέγιστο δυνατό υποσύνολο του *C*.

Ανάλογα με τα χαρακτηριστικά του πεδίου αναφοράς του συνθέτη αλλά και τις ανάγκες του πεδίου εφαρμογής, η φύση αλλά και το πλήθος των ακουστικών μονάδων μπορεί να κυμαίνεται από μερικές δεκάδες έως αρκετές χιλιάδες λέξεις, συλλαβές, φωνήματα κ.ο.κ. Στην ακραία περίπτωση ενός πολύ περιορισμένου πεδίου αναφοράς (domain) έχει αποδειχθεί ότι η σχεδίαση ενός εξειδικευμένου σώματος κειμένου προς την συγκεκριμένη εφαρμογή παρέχει εύρωστα και συνεπή αποτελέσματα υψηλής ποιότητας. Σε αυτήν την περίπτωση το μόνο που αρκεί είναι η δημιουργία προτάσεων όλων των δυνατών συνδυασμών των απαραίτητων λέξεων σε όλα τα δυνατά προσωδιακά περιβάλλοντα. Κατά αυτόν τον τρόπο, απλές περιπτώσεις όπως η δημιουργία ενός ηχητικού ρολογιού ή περιορισμένων καιρικών προγνώσεων, είναι δυνατό να εξυπηρετηθούν με μικρή προσπάθεια. Σε τέτοιες περιπτώσεις η χειροκίνητη επιλογή ή δημιουργία των προτάσεων του σώματος κειμένου είναι συχνά η απλούστερη και αποτελεσματικότερη μέθοδος.

4.2.1 Το μήκος της ακουστικής μονάδας

Διαφορετικού τύπου ακουστικές μονάδες απαιτούν διαφορετικές στρατηγικές κάλυψης και υπολογισμού του σώματος κειμένου. Λέξεις, φράσεις ή ακόμη και προτάσεις διαφέρουν σημαντικά με τα τριφωνήματα, διφωνήματα ή φωνήματα όσον αφορά στον υπολογισμό του βέλτιστου σώματος κειμένου. Αρχικά λοιπόν η βασική απαίτηση ήταν η πλήρης κάλυψη των δυνατών

διφωνημάτων στην συγκεκριμένη γλώσσα. Η τεχνολογία της σύνθεσης με διφωνήματα[4], απαιτεί την ύπαρξη ενός μόνο στιγμιότυπου από κάθε διφώνημα σε κοινές συνθήκες, όσο ήταν αυτό δυνατόν. Κατά την ηχογράφηση του σώματος αυτού, ο ομιλητής προσπαθεί να διατηρήσει σταθερό, ουδέτερο και σχετικά αργό ρυθμό εκφώνησης. Οι συνθήκες αυτές απαιτούνται από την συγκεκριμένη τεχνολογία και τον τρόπο χειρισμού των προσωδιακών χαρακτηριστικών του σήματος φωνής [Black2003].

Η τεχνολογία της σύνθεσης με βέλτιστη επιλογή διφώνου κάνει διαφορετικές υποθέσεις, θεωρώντας ότι κατά την διαδικασία της σύνθεσης θα έχει στην διάθεσή της καταλληλότερα διφωνήματα και επομένως δεν περιλαμβάνει ανάγκη υψηλού βαθμού επεξεργασία σήματος. Ο χειρισμός των προσωδιακών χαρακτηριστικών γίνεται σε μικρότερη έκταση προσδίδοντας μικρότερη παραμόρφωση στο τελικό συνθετικό σήμα. Η συγκεκριμένη μεθοδολογία έχει υιοθετηθεί και από το δικό μας σύστημα, όπου περισσότερη έμφαση δίνεται στην επιλογή βέλτιστης ακουστικής μονάδας και επιτρέπεται σε μικρότερο βαθμό η επιπλέον ψηφιακή επεξεργασία του σήματος. Έτσι, η «πληρότητα» της βάσης αποκτά άλλη διάσταση, με ιδιάζουσα σημασία στο σύστημά μας.

Μαζί με τις διαφορετικές υποθέσεις και λειτουργίες της τεχνολογίας αυτής, άλλαξαν και οι ανάγκες που πρέπει το σώμα κειμένου να καλύπτει, και αν και η τεχνολογία αυτή έχει ωριμάσει αρκετά, η διαδικασία της σχεδίασης του αρχικού σώματος κειμένου δεν καθορίζεται κοινά από όλα τα μέλη της ερευνητικής κοινότητας [Francois2001] [Lambert2004] [Villasenor2003]. Πολλοί είναι αυτοί που υποστηρίζουν ότι η πλήρης κάλυψη σε διφωνήματα αρκεί, ενώ σε άλλες επιστημονικές ανακοινώσεις αναφέρεται ότι η πλήρης κάλυψη των συχνότερων λέξεων, σε διαφορετικά περιβάλλοντα και προτάσεις είναι η βασικότερη προϋπόθεση για ένα πλήρες σώμα κειμένου. Ωστόσο, λαμβάνοντας υπόψη τις διαφορετικές υποθέσεις που αρχικά κάνει η συγκεκριμένη τεχνολογία σύνθεσης ομιλίας φαίνεται ότι ούτε η μία ούτε η δεύτερη θέση είναι επαρκείς, αφού είναι λογικό να απαιτείται και κάλυψη, όσο αυτό είναι δυνατόν, σε διαφορετικά προσωδιακά φαινόμενα ανά φώνημα. Αυτή άλλωστε είναι και η άποψη στην οποία φαίνεται να συγκλίνουν οι περισσότεροι βασικοί παίκτες του τομέα της σύνθεσης φωνής σε παγκόσμιο επίπεδο, όπως

[4] Σύνθεση με διφωνήματα (diphone synthesis) ονομάζεται η τεχνολογία που απαιτεί μόνο ένα στιγμιότυπο από κάθε διφώνημα, ενώ σύνθεση που κάνει χρήση περισσότερων στιγμιοτύπων ανά διφώνημα ονομάζεται σύνθεση με βέλτιστη επιλογή διφώνου (unit selection synthesis).

άλλωστε φαίνεται και από την πλειοψηφία των πιο πρόσφατων επιστημονικών ανακοινώσεων. Παράλληλα, σημαντικές παρατηρήσεις αναφορικά με την επιλογή βέλτιστου σώματος κειμένου έχουν να κάνουν με την παρουσία πολυπληθών σπάνιων φαινομένων στην γλώσσα ή την προφορά που όμως επιδρούν σημαντικά στην γενικότερη εικόνα και απόδοση του συνθέτη φωνής (LNRE – Large Number of Rare Events) [Möbius2000], υποστηρίζοντας ότι η χρήση τριφωνημάτων ως ακουστική μονάδα υπερτερεί σημαντικά του διφωνήματος. Το φαινόμενο των LNRE ουσιαστικά αναφέρεται σε ακουστικά φαινόμενα που ενώ δεν συναντώνται συχνά σε μία γλώσσα, είναι δυνατό να προκαλέσουν σημαντικές ασυνέχειες σε ένα σύστημα σύνθεσης φωνής όταν κληθεί να τα αντιμετωπίσει. Τέτοιες περιπτώσεις π.χ. είναι συνδυασμοί φωνημάτων που δεν συναντώνται συχνά στην Ελληνική γλώσσα, αλλά απαιτούνται για την εκφορά ξένων λέξεων ή όρων. Ειδικότερα το φαινόμενο των LNRE φαίνεται ότι διαδραματίζει σημαντικό ρόλο στην ποιότητα ενός συστήματος σύνθεσης φωνής, αφού η εμφάνιση σπάνιων ακουστικών φαινομένων, αν δεν έχει προβλεφθεί ανάλογα, οδηγεί σε σημαντικές ασυνέχειες ή παραμορφώσεις στο τελικό αποτέλεσμα, μειώνοντας συνολικά την απόδοση του συστήματος. Το γεγονός αυτό οφείλεται ουσιαστικά στην ακουστική αντίληψη που έχει το ανθρώπινο αυτί, το οποίο δεν συγχωρεί ατέλειες που είναι ικανές να αποσπάσουν την προσοχή του ακροατή.

Η δική μας προσέγγιση [Chalamandaris et al., 2009a] βασίζεται σε διφωνήματα και επιχειρεί να σχεδιάσει ένα βέλτιστο σώμα κειμένου με βάση τα χαρακτηριστικά του συνθέτη φωνής, ενώ περιλαμβάνει ειδικό στάδιο αντιμετώπισης πολυπληθών σπάνιων φαινομένων [Chalamandaris et al., 2011] καθώς και βελτιώσεων αναφορικά με τα χαρακτηριστικά της φωνής του ομιλητή [Founda et al., 2001a] [Founda et al., 2001b].

4.2.2 Διαφορετικά πεδία αναφοράς σύνθεσης

Όπως ήδη αναφέραμε, διαφορετικά πεδία αναφοράς σύνθεσης φωνής (domains) επιτρέπουν διαφορετικές στρατηγικές σχεδιασμού σώματος κειμένου, ειδικά σε συνδυασμό με διφωνήματα ή τριφωνήματα. Ένα πεδίο αναφοράς μπορεί να χαρακτηριστεί ως περιορισμένο (restricted) εφόσον το λεξιλόγιο αλλά και η ορολογία που χρησιμοποιεί είναι θεωρητικά περιορισμένα, αλλά μπορεί να είναι και απείρως ορισμένο (unlimited) εφόσον οι συνδυασμοί των λέξεων και φράσεων που μπορεί να περιέχει είναι στην πράξη άπειροι σε πλήθος. Στα [Franois2002] και [Francois2001] οι συγγραφείς παρουσιάζουν τις διαφορετικές ακουστικές κατανομές σε διαφορετικά πεδία αναφοράς

ως ένα χαρακτηριστικό του γραπτού λόγου. Για τα Ελληνικά, εξετάσαμε την κατανομή των γλωσσικών αλλά και ακουστικών μονάδων σε διαφορετικού ύφους και στυλ κείμενα. Πιο συγκεκριμένα, εξετάσαμε κείμενα από τρεις διαφορετικές εφημερίδες, διαφορετικής θεματολογίας και συγκεκριμένα 2 πολιτικές, 2 οικονομικές και 2 αθλητικές εφημερίδες. Πραγματοποιώντας στατιστική ανάλυση των παραπάνω συνόλων κειμένων, τόσο σε επίπεδο λέξης όσο και σε επίπεδο διφωνήματος και τριφωνήματος καταλήξαμε στο συμπέρασμα ότι όντως οι κατανομές των ακουστικών αλλά και γλωσσικών χαρακτηριστικών των κειμένων αυτών διαφέρουν σημαντικά μεταξύ τους [Chalamandaris et al., 2009b].

Μοναδικές Λέξεις	***Πολιτική Α***	***Πολιτική Β***	***Οικονομική Α***	***Οικονομική Β***	***ΑθλητικήΑ***	***ΑθλητικήΒ***
Πολιτική Α	100	68,97	60,54	58,98	34,82	43,02
Πολιτική Β	64,56	100	56,78	58,02	32,03	36,81
Οικονομική Α	61,09	58,9	100	69,89	53,23	49,09
Οικονομική Β	60,01	63,02	73,4	100	41,13	43,41
Αθλητική Α	55,06	54,45	53,24	51,15	100	76,14
Αθλητική Β	51,03	51,23	49,85	53,21	78,83	100

Πίνακας 1: Στατιστική ανάλυση των μοναδικών λέξεων για διαφορετικού ύφους εφημερίδες. Για κάθε ζευγάρι εφημερίδων απεικονίζεται το ποσοστό των μοναδικών λέξεων που έχουν κοινό.

Όπως άλλωστε θα περίμενε κανείς, κείμενα με ίδιο στυλ ή θεματολογία παρουσιάζουν όμοιες γλωσσικές αλλά και ακουστικές κατανομές, ενώ κείμενα διαφορετικής θεματολογίας παρουσιάζουν διαφορετικά αντίστοιχα χαρακτηριστικά.

Προχωρώντας ένα βήμα παραπάνω, εξετάσαμε τα χαρακτηριστικά των κειμένων της ίδιας εφημερίδας σε 3 διαδοχικά έτη, με αποτέλεσμα να παρατηρήσουμε ότι ακόμη και στα κείμενα ίδιας θεματολογίας τόσο τα γλωσσικά χαρακτηρηστικά όσο και τα ακουστικά διαφοροποιούνται από εποχή σε εποχή, υποδεικνύοντας ότι το πεδίο αναφοράς είναι κάτι πολύ στενότερο από αυτό που θεωρείται γενικότερα (π.χ. αθλητικά, οικονομικά, πολιτικά κ.α.).

Μοναδικές Λέξεις	*Έτος 1*	*Έτος 2*	*Έτος 3*
Έτος 1	100	60,4	59,65
Έτος 2	61,9	100	61,23
Έτος 3	61,04	61,14	100

Μοναδικές Λέξεις	*Έτος 1*	*Έτος 2*	*Έτος 3*
Έτος 1	100	98,51	98,43
Έτος 2	98,63	100	98,59
Έτος 3	98,4	98,44	100

Πίνακας 2: Ποσοστό των κοινών μοναδικών λέξεων για δύο διαφορετικές εφημερίδες (Πολιτική και Αθλητική) στη διάρκεια τριών διαδοχικών ετών.

Από τα παραπάνω φαίνεται καθαρά ότι κανείς πρέπει να λαμβάνει σοβαρά υπόψη τα ιδιαίτερα χαρακτηριστικά του πεδίου αναφοράς, αλλά και του πεδίου εφαρμογής του συνθέτη φωνής πριν σχεδιάσει το σώμα κειμένου προς ηχογράφηση.

4.2.3 Μέθοδοι επιλογής βέλτιστων προτάσεων

Σε γενικές γραμμές, η τεχνική που ακολουθείται συχνότερα στην διαδικασία αυτήν, είναι η μεγαλύτερη δυνατή κάλυψη περισσότερων μοναδικών περιβαλλόντων, μέσω χρήσης «άπληστων» αλγορίθμων (greedy algorithms). Με τον όρο μοναδικά περιβάλλοντα περιγράφουμε τα μοναδικά διαφορετικά διανύσματα που προκύπτουν, αν μεταφέρουμε όλους τους βαθμούς ελευθερίας σε αντίστοιχες συντεταγμένες. Αν π.χ. έχουμε ως βαθμούς ελευθερίας την ταυτότητα του φωνήματος και το φώνημα που ακολουθεί, τότε ως διανύσματα μπορούμε να χρησιμοποιήσουμε τα $\begin{bmatrix} x \\ y \end{bmatrix}$ όπου x είναι η ταυτότητα του τρέχοντος φωνήματος, με τιμές από 1-36 (για τα Ελληνικά), και όπου y η ταυτότητα του επόμενου φωνήματος, με τιμές πάλι από 1-36.

Έχοντας ορίσει ως C το σύνολο των διαφορετικών ακουστικών μονάδων που επιθυμούμε να καλύψουμε, επιχειρούμε να επιλέξουμε ένα υποσύνολο προτάσεων S από ένα αρχικό σύνολο

προτάσεων *P*. Προφανώς πριν την επιλογή των προτάσεων, τόσο τα *C* και *P* πρέπει να έχουν προκαθορισθεί. Το αρχικό σώμα κειμένου *P* συνήθως είναι ιδιαίτερα εκτενές ενώ ιδανικά χαρακτηρίζεται από ένα στυλ, αυτό που επιθυμούμε να καλύψουμε με τον συνθέτη φωνής.

Οι μέθοδοι που προτείνονται στην αντίστοιχη βιβλιογραφία διαφοροποιούνται κυρίως σε δύο βασικά σημεία: α) στον τρόπο συσταδοποίησης των στοιχείων του συνόλου *C* και β) στην θεωρία πίσω από το σύνολο *C*.

Η πρώτη διαφορά έγκειται ουσιαστικά στην τροποποίηση του άπληστου αλγορίθμου επιλογής λαμβάνοντας υπόψη χαρακτηριστικά των ακουστικών μονάδων που ουσιαστικά επιτρέπουν την ομαδοποίησή τους και επομένως συχνά την κατά προταιρεότητα κάλυψή τους. Η δεύτερη διαφορά ωστόσο έγκειται σε σημαντικότερο επίπεδο της συγκριμένης διαδικασίας και έχει να κάνει ουσιαστικά με την ίδια την θεωρία πίσω από την διαδικασία. Άλλες π.χ. θεωρίες υποστηρίζουν ότι έμφαση πρέπει να δίνεται στην κατανομή των ακουστικών φαινομένων στο πεδίο αναφοράς και επομένως το σώμα στόχος πρέπει να ενσωματώνει παρόμοια κατανομή, άλλες ωστόσο υποστηρίζουν ότι έμφαση πρέπει να δίνεται στα σπάνια φαινόμενα (LNRE) που επηρεάζουν σημαντικά την ποιότητα και συνέπεια του συνθέτη. Σε κάθε περίπτωση, οι προτεινόμενες μέθοδοι στο σύνολό τους διατηρούν την ίδια αρχή κάλυψης μέσω προσαρμοσμένων ή μη άπληστων αλγορίθμων επιλογής.

4.2.4 Αλγόριθμος επιλογής βέλτιστου υπο-σώματος κειμένου

Ο αλγόριθμος επιλογής των βέλτιστων προτάσεων μέσα από ένα σώμα κειμένου δίνει λύση στο πρόβλημα ανεύρεσης του μικρότερου αριθμού προτάσεων που προσφέρουν μέγιστη κάλυψη σε μία δεδομένη παράμετρο [Black1997]. Ο αλγόριθμος παραμένει ο ίδιος και στις τρεις διαφορετικές παραμέτρους που χρησιμοποιούμε και έχουμε αναφέρει παραπάνω. Στην συνέχεια της ενότητας αυτής θα παρουσιάσουμε τον αλγόριθμο επιλογής με βάση τις συχνότερες λέξεις, ο οποίος γενικεύεται και στις υπόλοιπες περιπτώσεις.

Ξεκινώντας από μία λίστα *μ* συχνότερων λέξεων που επιθυμούμε να καλύψουμε, ιεραρχούμε το σύνολο των προτάσεων ανάλογα με την βαθμολογία που λαμβάνουν σύμφωνα με τις λέξεις που καλύπτουν. Η βαθμολογία που λαμβάνει κάθε πρόταση εξαρτάται από τον αριθμό των επιθυμητών λέξεων που καλύπτουν, και από την θέση που κατέχουν στην λίστα (σημαντικότητα)

ανάλογα με την σχετική συχνότητα. Η ταξινομημένη αυτή λίστα των προτάσεων χρησιμεύει στην συνέχεια στην διαρκή άντληση βέλτιστων προτάσεων με διαδοχικές επαναλήψεις. Σε κάθε μία επανάληψη απαιτούμε κάθε φορά ένα συγκεκριμένο αριθμό λέξεων να καλύπτεται εκ νέου. Ο αριθμός των επιθυμητών αυτών λέξεων μειώνεται κατά ένα στην περίπτωση που δεν προστίθεται καμία νέα πρόταση στο παραγόμενο σώμα κειμένου. Παρακάτω μπορεί κανείς να μελετήσει το διάγραμμα ροής του αλγορίθμου σε ψευδοκώδικα.

```
All_sentences = initial written corpus
lexicon = words to be covered
N = 15
while N>1
 Flag = 0
 i = 0
 while i<length(All_sentences)
   i = i+1
   if covered words in All_sentences(i)==N
     add_to_list All_sentences(i)
     remove_from_All_sentences All_sentences(i)
     remove_from_lexicon_covered_words
     Flag  = 1
   end
   if Flag==0
   N = N -1
 end
```

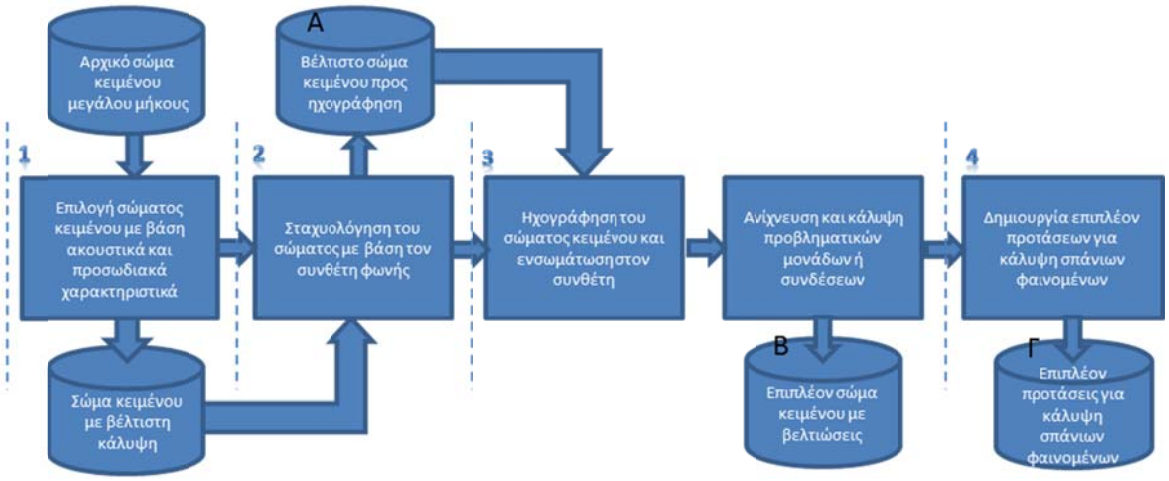
Πίνακας 3: Ψευδοκώδικας ροής του αλγόριθμου επιλογής βέλτιστων προτάσεων.

Έχοντας υπόψη τον αλγόριθμο αυτόν, είναι πλέον πολύ εύκολο να κατανοήσει κανείς ότι η παραπάνω διαδικασία μπορεί να πραγματοποιηθεί μέσω διανυσματικού κβαντισμού για περισσότερες από μία παραμέτρους. Έτσι, ενώ στην περίπτωση της κάλυψης των φωνημάτων έχουμε 36 διαφορετικά φωνήματα και επομένως 36 διαφορετικά διανύσματα προς αναζήτηση μέσα στο σώμα κειμένου, στην περίπτωση χρήσης περισσοτέρων παραμέτρων όπως είναι π.χ. τα διφωνήματα έχουμε 36 x 36 = 1296 διαφορετικά διανύσματα που επιθυμούμε να καλύψουμε κατά

την διαδικασία αυτήν. Παρομοίως, αν έχουμε περισσότερες παραμέτρους στο ίδιο πλαίσιο, απλά συνδυάζοντάς τες μέσω διανυσματικού κβαντισμού καταλήγουμε σε ένα πλήθος επιθυμητών διανυσμάτων, τα οποία αναζητάμε μέσα στο σώμα κειμένου. Η αναζήτηση γίνεται με τον τρόπο που περιγράψαμε παραπάνω, ο οποίος ουσιαστικά περιγράφει έναν «άπληστο» αλγόριθμο αναζήτησης.

4.3 Η προσέγγισή μας

Η προσέγγιση που αποφασίσαμε να ακολουθήσουμε συνδυάζει τις προαναφερθείσες μεθόδους σε μεγάλο βαθμό, αλλά λαμβάνοντας υπόψη τους περιορισμούς που παρουσιάζει η συγκεκριμένη τεχνολογία, προχωράμε ένα βήμα παραπέρα. Η βασική ιδέα της συγκεκριμένης μεθόδου ήταν η σχεδίαση ενός μηχανισμού που θα λαμβάνει υπόψη του εκτός από τα χαρακτηριστικά του πεδίου αναφοράς και εφαρμογής, τα ιδιαίτερα χαρακτηριστικά του ίδιου του συστήματος σύνθεσης φωνής για το οποίο προορίζεται το συγκεκριμένο σώμα κειμένου [Chalamandaris et al., 2009b], [Chalamandaris et al., 2011]. Πιο συγκεκριμένα, η διαδικασία σχεδίασης του σώματος κειμένου διαχωρίζεται σε τέσσερα διαφορετικά στάδια τα οποία αλληλοσυμπληρώνονται και στοχεύουν στην πλήρωση διαφορετικών απαιτήσεων και αναγκών. Στις παραγράφους που ακολουθούν παρουσιάζουμε αναλυτικά τα επιμέρους αυτά στάδια.



Σχήμα 29: Το διάγραμμα ροής για τον σχεδιασμό του σώματος κειμένου προς ηχογράφηση. Με διακεκομμένες γραμμές ορίζονται τα διαφορετικά στάδια της διαδικασίας. Η ένωση των επιμέρους σωμάτων κειμένου Α, Β και Γ αποτελούν το τελικό σώμα κειμένου που κατασκευάστηκε για τον συνθέτη ομιλίας.

4.3.1 Στάδιο 1: Επιλογή του σώματος κειμένου S

Για την κατασκευή ενός σώματος κειμένου *S* είναι απαραίτητη η ύπαρξη ενός μεγάλου σώματος *P* κειμένου, μέσα από το οποίο θα προκύψει το επιλεγμένο ως υπο-σώμα κειμένου, το οποίο και θα πληρεί τις ανάλογες προϋποθέσεις. Το αρχικό αυτό σώμα κειμένου μπορεί να προέλθει είτε από λογοτεχνικά κείμενα είτε από γενικού περιεχομένου κείμενα, ανάλογα με το πεδίο αναφοράς και/ή εφαρμογής. Έτσι, παραδείγματος χάριν, αν επιθυμούσαμε το σύστημα μας να ανταποκρίνεται καλύτερα σε δελτία καιρού, τότε το αρχικό μας σώμα κειμένου θα έπρεπε να περιέχει σε μεγάλο ποσοστό δεδομένα σχετικά με δελτία καιρού.[5] Στην συνέχεια δημιουργούμε ένα μικρότερο σώμα κειμένου *S*, προσπαθώντας ταυτόχρονα να επιτύχουμε βέλτιστη κάλυψη σε τρία διαφορετικά επίπεδα.

Σε επίπεδο συχνότερων λέξεων: εφόσον το αρχικό μας σώμα κειμένου περιέχει κείμενα σχετικά με το πεδίο ενδιαφέροντος, απαιτείται κάλυψη των συχνότερων λέξεων που συναντώνται σε συναφή κείμενα.

Σε επίπεδο διφωνημάτων: εφόσον οι μικρότερες δομικές μονάδες του συστήματός μας είναι το διφώνημα, οφείλουμε να έχουμε πλήρη κάλυψη όλων των δυνατών διφωνημάτων που μπορεί να απαντηθούν στην συγκεκριμένη γλώσσα. Στο σημείο αυτό πρέπει να αναφέρουμε ότι τα σημεία στίξης πρέπει να ερμηνεύονται σωστά από τον ομιλητή γιατί στην αντίθετη περίπτωση ενδεχομένως να μην έχουμε τελικά την κάλυψη των διφωνημάτων την οποία αρχικά υπολογίζουμε, ενώ ειδική μέριμνα πρέπει να δοθεί σε διφωνήματα που ενώ δεν συναντώνται σε Ελληνικές λέξεις, απαιτούνται για την εκφορά ξενικών λέξεων από Έλληνα ομιλητή.

Σε επίπεδο προσωδιακών φαινομένων: το πεδίο αυτό είναι από τα πολυπλοκότερα και ενδεχομένως η προσπάθεια πλήρους κάλυψής του να παράγει αποτελέσματα με απαγορευτικό μέγεθος για ηχογράφηση ή επεξεργασία. Είναι προφανές βέβαια ότι το συγκεκριμένο στάδιο προϋποθέτει ότι είναι καθορισμένος ο τρόπος μοντελοποίησης της προσωδίας σε μία φράση, είτε αυτός είναι βασισμένος σε δεδομένα (corpus-based modeling) είτε αυτός είναι βασισμένος σε

[5] Η περίπτωση στην οποία αναφερόμαστε στην συγκεκριμένη διατριβή πραγματεύεται την δημιουργία ενός γενικού περιεχομένου συνθέτη φωνής, με έμφαση σε δελτία ειδήσεων, οπότε και το αρχικό μας σώμα κειμένου περιλάμβανε έναν σημαντικό όγκο γραπτών δελτίων ειδήσεων.

κανόνες (rule-based modeling), έτσι ώστε να έχει νόημα και ο προσωδιακός χαρακτηρισμός των ακουστικών μονάδων.

Παρακάτω δίνεται αναλυτικότερη περιγραφή των επιμέρους αυτών επιπέδων.

4.3.1.1 Κάλυψη συχνότερων λέξεων

Η κάλυψη των συχνότερων λέξεων ενός σώματος κειμένου μπορεί να επιτευχθεί είτε άμεσα με τις N συχνότερες λέξεις που προκύπτουν από το αρχικό σώμα κειμένου, είτε, αν η γλώσσα που αναφερόμαστε χαρακτηρίζεται από πλούσια μορφολογία, όπως είναι π.χ. τα Ελληνικά, από τα N συχνότερα θέματα και τις M συχνότερες καταλήξεις. Το επίπεδο στο οποίο μπορεί κανείς να ανάγει το συγκεκριμένο πρόβλημα εξαρτάται από την γλώσσα, αλλά και από το εύρος του πεδίου ενδιαφέροντος όπου πρόκειται το τελικό σύστημα να αναφέρεται. Αν παραδείγματος χάριν κάποιος επιθυμεί να αναπτύξει έναν συνθέτη φωνής με στόχευση οποιοδήποτε δυνατό κείμενο, τότε θα πρέπει ενδεχομένως, αντί των συχνότερων λέξεων να απαιτήσει την κάλυψη των συχνότερων συλλαβών, ειδικότερα για την ελληνική γλώσσα που παρουσιάζει μία αμιγώς συλλαβική δομή.

4.3.1.2 Κάλυψη σε επίπεδο διφωνημάτων

Η κάλυψη σε επίπεδο διφωνημάτων απαιτείται εφόσον το τελικό σύστημα έχει ως ελάχιστη δομική μονάδα το διφώνημα. Αν ο συνθέτης φωνής βασίζεται σε φωνήματα ή ημιφωνήματα, τότε το στάδιο αυτό θα αναγόταν σε πλήρη κάλυψη των αντίστοιχων δομικών μονάδων. Για τα ελληνικά η πλήρης κάλυψη των διφωνημάτων απαιτεί 1134 μοναδικά διφωνήματα, και όχι $36^2 = 1296$. Πολλά από τα διφωνήματα αυτά, αν και δεν απαντώνται στο εσωτερικό ελληνικών λέξεων, συχνά απαντώνται μεταξύ λέξεων, ενώ απαντώνται και οι περιπτώσεις των διφωνημάτων που δεν είναι δυνατά στην ελληνική γλώσσα, αλλά χρησιμοποιούνται κατά την εκφορά ξένων λέξεων, όπως αναφέρθηκε και προηγουμένως.

4.3.1.3 Κάλυψη σε επίπεδο προσωδιακών φαινομένων

Η κάλυψη σε αυτό το επίπεδο προϋποθέτει ότι ο τρόπος μοντελοποίησης της προσωδίας είναι προκαθορισμένος, δηλαδή το υποσύστημα που είναι υπεύθυνο για την μοντελοποίηση της προσωδίας είναι υλοποιημένο.

Η αναγκαιότητα του σταδίου αυτού είναι προφανής στην περίπτωση που η μοντελοποίηση της προσωδίας γίνεται με χρήση δεδομένων και λιγότερο προφανής στην περίπτωση όπου η μοντελοποίηση γίνεται με χρήση κανόνων. Ωστόσο, και σε αυτήν την περίπτωση πρέπει να ληφθεί μέριμνα για την βέλτιστη κάλυψη των βασικότερων φαινομένων που μπορεί να αναπαράγει η μηχανή προσωδίας. Η τελική κάλυψη αναφέρεται είτε σε λέξεις, είτε σε μικρότερα δομικά στοιχεία όπως είναι οι συλλαβές ή τα διφωνήματα. Όσο μεγαλύτερο είναι το μήκος της δομικής μονάδας αναφοράς (φωνήματα, διφωνήματα, τροφωνήματα, συλλαβές, λέξεις κ.ο.κ.), τόσο μεγαλύτερου μεγέθους αποτελέσματα προκύπτουν κατά την κατασκευή του σώματος κειμένου και για τον λόγο αυτόν σπάνια π.χ. χρησιμοποιούνται λέξεις ως βάση κάλυψης.

Στην περίπτωση που πραγματευόμαστε, εφόσον ο αλγόριθμος παραγωγής μοντέλων επιτονισμού έχει ως βασική δομική μονάδα την φωνητική συλλαβή, ως βασική μονάδα αναφοράς χρησιμοποιήσαμε την φωνητική συλλαβή, ενώ ως επιπλέον παραμέτρους χαρακτηρισμού αυτής χρησιμοποιήσαμε την σχετική θέση της αναφορικά με τον λεξικό τόνο και τα σημεία στίξης, ούτως ώστε για να επιτύχουμε βέλτιστη κάλυψη προσωδιακών φαινομένων. Τα παραπάνω άλλωστε χαρακτηριστικά αποτελούν τις παραμέτρους που «οδηγούν» την μηχανή παραγωγής προσωδιακών μοντέλων του συστήματος σύνθεσης φωνής που έχουμε αναπτύξει και επομένως η βέλτιστη δυνατή κάλυψή τους επιτρέπει στην μηχανή παραγωγής προσωδιακών μοντέλων να λειτουργεί αποτελεσματικότερα. Ο τρόπος λειτουργίας του υποσυστήματος παραγωγής μοντέλων επιτονισμού παρουσιάζεται αναλυτικότερα στο κεφάλαιο που αναφέρεται στην προσωδία και την μοντελοποίησή της.

Για την ολοκλήρωση του συγκεκριμένου σταδίου είναι απαραίτητη η ύπαρξη δύο υποσυστημάτων του συνθέτη φωνής: α) το υποσύστημα γλωσσικής ανάλυσης και επεξεργασίας και β) το υποσύστημα παραγωγής προσωδίας. Τα υποσυστήματα αυτά είναι απαραίτητα ούτως ώστε να κανονικοποιηθεί το κείμενο, να αναλυθεί σε προτάσεις, να μετγραφεί σε φωνητική αναπαράσταση και να παραχθούν τα απαραίτητα προσωδιακά χαρακτηριστικά για κάθε φωνητική συλλαβή σε αυτό. Μόνον τότε θα έχουμε την δυνατότητα να χρησιμοποιήσουμε όλα τα παραπάνω στοιχεία για την δημιουργία των διανυσμάτων χαρακτηριστικών για κάθε μονάδα στο αρχικό σώμα κειμένου P.

Το παραγόμενο σώμα κειμένου *S* που προκύπτει από αυτό το στάδιο προσφέρει όσο είναι δυνατό μεγαλύτερη κάλυψη των μοναδικών διανυσμάτων χαρακτηριστικών που προέκυψαν από την ανάλυση του αρχικού σώματος κειμένου *P,* με υψηλή επικάλυψη και ιδιαίτερα μεγάλο μέγεθος, αφού απαιτήθηκε η πλήρης κάλυψη του συνόλου των μοναδικών διανυσμάτων, χωρίς κάποια συσταδοποίηση με βάση κάποια κριτήρια ομοιότητας.

4.3.2 Στάδιο 2: Σταχυολόγηση του σώματος κειμένου S με την χρήση του συνθέτη φωνής

Το δεύτερο στάδιο της διαδικασίας σχεδιασμού του σώματος κειμένου βασίζεται σε έναν προσαρμοσμένο μηχανισμό επιλογής βέλτιστων μονάδων του συνθέτη φωνής, ούτως ώστε να λειτουργεί ως μέτρο σύγκρισης των διαφορετικών ακουστικών μονάδων που έχουν συμπεριληφθεί στο σύστημα [Tsiakoulis et al., 2008]. Με βάση το συγκεκριμένο αυτό μέτρο σύγκρισης επιλέγεται στην συνέχεια το σταχυολογημένο σώμα κειμένου S'. Το συγκεκριμένο στάδιο απαιτεί περισσότερη ανάλυση.

Όπως περιγράψαμε στο πρώτο στάδιο της διαδικασίας, το υποσώμα κειμένου S περιέχει πλεονάζουσα πληροφορία με επικάλυψη και μεγάλο μέγεθος. Ως πρώτο βήμα του σταδίου, δημιουργούμε μία εικονική βάση δεδομένων, στην οποία απουσιάζει οποιαδήποτε ηχογράφηση ή σήμα φωνής και περιέχει μόνο στοιχεία που έχουν να κάνουν με την αντικειμενική πληροφορία των ακουστικών μονάδων. Με αυτόν τον τρόπο ουσιαστικά προσομοιώνουμε την βάση δεδομένων που θα προέκυπτε από το συγκεκριμένο σώμα κειμένου, με ταυτόχρονη ουδετεροποίηση του παράγοντα του σήματος φωνής. Στην παρακάτω συνάρτηση κόστους, όπου οι επιμέρους συναρτήσεις κόστους (target cost και concatenation cost) αθροίζονται, εμείς μηδενίζουμε τα επιμέρους κόστη που έχουν να κάνουν με το ίδιο το σήμα της φωνής.

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} W^t \cdot C^t(t_i, u_i) + \sum_{i=2}^{n} W^c \cdot C^c(u_{i-1}, u_i)$$

Πιο συγκεκριμένα, στην συνάρτηση του κόστους συνέχειας (concatenation cost) μηδενίζουμε τα κόστη που αναφέρονται στην φασματική ασυνέχεια και στην ασυνέχεια στην θεμελιώδη συχνότητα. Κατά αυτόν τον τρόπο, διατηρούμε μόνο τα κόστη που έχουν να κάνουν με «αντικειμενικά» κριτήρια των ακουστικών μονάδων, όπως είναι το περικείμενο, η προσωδιακή

ταυτότητα κ.λ.π, ενώ αγνοούμε κόστη «υποκειμενικά» που έχουν να κάνουν με τα τοπικά χαρακτηριστικά του σήματος φωνής και της εκάστοτε έκφανσής του.

Έχοντας τροποποιήσει κατάλληλα το υποσύστημα επιλογής βέλτιστων ακουστικών μονάδων του συνθέτη φωνής έτσι ώστε να λαμβάνει υπόψη μόνο τα αντικειμενικά χαρακτηριστικά των ακουστικών μονάδων, προσομοιώσαμε την σύνθεση του πλήρους αρχικού σώματος κειμένου *P*, διατηρώντας ταυτόχρονα την πληροφορία για την συχνότητα χρήσης κάθε ακουστικής μονάδας της βάσης. Η πληροφορία αυτή στην συνέχεια αποτέλεσε την βάση για την ιεράρχηση και αξιολόγηση κάθε πρότασης από το σώμα κειμένου *S*. Η αξιολόγηση και ιεράρχηση των προτάσεων του σώματος κειμένου *S* με βάση την συχνότητα χρησιμοποιήσης κάθε μονάδας από το σύστημα, αλλά και με βάση την συσταδοποίηση των μονάδων αυτών σε προηγούμενο στάδιο, βασίζεται σε προηγούμενη εργασία της ομάδας σύνθεσης φωνής, με την οποία ήταν δυνατή η σταχυολόγηση μια μεγαλύτερης βάσης δεδομένων συνθέτη φωνής σε μικρότερη, για την μεταφορά του συστήματος σε φορητές συσκευές [Tsiakoulis et al., 2008].

Ο συγκεκριμένος αλγόριθμος επιλέγει M προτάσεις από το δεδομένο σώμα κειμένου με βάση το κριτήριο της μεγιστοποίησης ενός διανύσματος καταλληλότητας που λαμβάνει υπόψη τις επιμέρους συχνότητες επιλογής για κάθε ακουστική μονάδα και τα συνολικά κόστη για κάθε πρόταση από τον αλγόριθμο βέλτιστης επιλογής του συστήματος σύνθεσης φωνής. Κατά την επιλογή μίας πρότασης από το σώμα κειμένου, το διάνυσμα καταλληλότητας για κάθε πρόταση επικαιροποιείται με τον πολλαπλασιασμό των μέσων διαφορών των συχνοτήτων για όμοιες ακουστικές μονάδες μέσα σε κάθε πρόταση με αυτές που περιέχει η πρόταση που επιλέχθηκε.

1. Initialize $F = [\, f_1^n, f_2^n \cdots f_K^n \,]$
2. Select $m = \arg\max_n F[n]$
3. Update $F[n] = F[n] \cdot D_{n,m}^i$ for $n = 1 \ldots K$
4. If #(*selected*) $< M$ goto step 2

Πίνακας 4: Αλγόριθμος σταχυολόγησης του σώματος κειμένου με χρήση των δεδομένων από το υποσύστημα βέλτιστης επιλογής ακουστικών μονάδων του συστήματος σύνθεσης φωνής.

Αποτέλεσμα του συγκεκριμένου σταδίου του μηχανισμού είναι ο υπολογισμός ενός υποσυνόλου του σώματος κειμένου S_{sub} το οποίο είναι σημαντικά μικρότερο από το σώμα *S* και ταυτόχρονα

περιλαμβάνει όλα τα απαράιτητα συστατικά που το σύστημα σύνθεσης φωνής θεωρεί απαραίτητα κατά την διαδικασία επιλογής βέλτιστων ακουστικών μονάδων. Το μέγεθος καθώς επίσης και ο βαθμός επικάλυψης του νέου σώματος κειμένου, το κατά πόσο δηλαδή οι μοναδικές ακουστικές συλλαβές που περιλαμβάνονται εμφανίζονται περισσότερες από μία φορές, εξαρτάται τόσο από το σύστημα επιλογής ακουστικών μονάδων του συνθέτη φωνής, όσο και από την αρχική συσταδοποίηση που έχουμε επιβάλει στο σώμα *S*, καθώς επίσης και στον αριθμό των όμοιων ακουστικών μονάδων που επιθυμούμε να διατηρήσουμε στο σώμα κειμένου, και καθορίζεται άμεσα μέσω της τιμής της μεταβλητής M.

4.3.3 Στάδιο 3: Εμπλουτισμός και διορθώσεις στο σώμα κειμένου S_{sub}

Κατά το τρίτο στάδιο της διαδικασίας και μετά την ηχογράφηση και ενσωμάτωση του σώματος S_{sub} στον συνθέτη φωνής, επιχειρούμε να εντοπίσουμε ασυνέχειες που θα παρουσιάσει το σύστημα σύνθεσης φωνής και οφείλεται κυρίως σε παράγοντες που είναι εξαρτώμενοι από τον ίδιο τον ομιλητή και τις ηχογραφήσεις που πραγματοποίησε. Με άλλα λόγια, ενώ στο προηγούμενο στάδιο υπολογίσαμε ένα βέλτιστο σώμα κειμένου βασιζόμενοι στα αντικειμενικά χαρακτηριστικά του, προσπαθώντας να απομονώσουμε τους παράγοντες που υπεισέρχονται από τα χαρακτηριστικά της φωνής του ομιλητή, στο συγκεκριμένο στάδιο ουστιαστικά προσπαθούμε να εντοπίσουμε ασυνέχειες και ελλείψεις που έχουν δημιουργηθεί από τα χαρακτηριστικά της φωνής ή των ηχογραφήσεων ειδικότερα [Founda et al., 2001b].

Η διαδικασία αυτή, όπως και η προηγούμενη, απαιτεί ότι οι βασικές λειτουργικότητες του υποσυστήματος της βέλτιστης επιλογής διφωνημάτων είναι παγιωμένες. Αυτό άλλωστε εξυπακούεται από το γεγονός ότι οποιαδήποτε αστοχία ή ασυνέχεια προσπαθούμε να αντιμετωπίσουμε με την διαδικασία αυτή, αναφέρεται στον συγκεκριμένο τρόπο λειτουργίας του συστήματος σύνθεσης φωνής σε συνδυασμό με την αρχική βάση δεδομένων. Για τον λόγο αυτό θα πρέπει ο τρόπος επιλογής των βέλτιστων διφωνημάτων να παραμένει ο ίδιος πριν και μετά την διαδικασία αυτή.

Για την ολοκλήρωση του συγκεκριμένου σταδίου, μέσω του ολοκληρωμένου συνθέτη, συνθέτουμε το σύνολο του αρχικού πλήρους σώματος *P* και εντοπίζουμε αυτόματα τοπικά μέγιστα στην συνάρτηση κόστους του υποσυστήματος βέλτιστης επιλογής ακουστικών μονάδων. Κατά αυτόν

τον τρόπο, εντοπίζουμε σημεία μέσα στο αρχικό σώμα κειμένου όπου ο συνθέτης θα παρουσίαζε μειωμένη απόδοση. Οι παράμετροι οι οποίες χρησιμοποιούνται για το στάδιο αυτό είναι οι ασυνέχειες που παρατηρούνται στο τελικό σήμα σε επίπεδο διφωνημάτων. Με τον όρο ασυνέχειες δεν εννοούμε ασυνέχειες μόνο στο φάσμα (φασματικές ασυνέχειες), αλλά και στην θεμελιώδη συχνότητα F0. Η αναζήτηση των ασυνεχειών αυτών γίνεται με βάση το υποσύστημα της βέλτιστης επιλογής διφωνημάτων, αναπτύσσοντας σε κάθε περίπτωση τα δένδρα απόφασης πλήρως. Όταν όλα τα επιμέρους κόστη ένωσης σε κάποιο επίπεδο του γράφου επιλογής βέλτιστης ακουστικής μονάδας ξεπερνούν τα προκαθορισμένα κατώφλια, τότε η συγκεκριμένη ακολουθία διφωνημάτων απαιτείται να υπάρχει στο επιπλέον αυτό σώμα κειμένου, σε συγκεκριμένο φωνητικό περιβάλλον. Με άλλα λόγια, αν κατά την σύνθεση μιας ακολουθίας φθόγγων παρατηρηθεί ότι δημιουργούνται ασυνέχειες λόγω απουσίας διφωνημάτων με κατάλληλα χαρακτηριστικά, τότε τα διφωνήματα αυτά θα πρέπει να σχεδιασθούν και να ενσωματωθούν εκ νέου στην βάση δεδομένων.

Αν παραδείγματος χάριν, προσπαθώντας να συνθέσουμε την λέξη «μύγα» και στο επίπεδο σύνδεσης /mi/ με το /iγ/ δεν υπάρχει κανένας συνδυασμός που να βρίσκεται κάτω από τα προκαθορισμένα κατώφλια αναφορικά με το περιβάλλον (context), την συνέχεια στην θεμελιώδη συχνότητα και την φασματική απόσταση, τότε απαιτείται το νέο σώμα κειμένου να περιέχει την συγκεκριμένη ακολουθία των διφωνημάτων σε όλα τα προκαθορισμένα προσωδιακά περιβάλλοντα, όπως αυτά έχουν ορισθεί από την μηχανή προσωδίας.

Στην συνέχεια, κάθε πρόταση του αρχικού σώματος P βαθμολογείται και ιεραρχείται ανάλογα με το κανονικοποιημένο κόστος που έφερε κατά την σύνθεση, αλλά και το σύνολο των τοπικών ελαχίστων που ενσωματώνει. Μέσω ενός άπληστου αλγορίθμου, επιλέγουμε τις N προτάσεις από το αρχικό σώμα P που έχουν την χειρότερη βαθμολόγηση και τα περισσότερα σημεία τοπικών ελαχίστων. Το νέο σώμα κειμένου S_{add} αποτελεί επιπλέον προτάσεις που στην συνέχεια ηχογραφούνται από τον ίδιο ομιλητή στις ίδιες συνθήκες με πριν και ενσωματώνονται στην βάση δεδομένων, ενώ το πλήθος N μπορεί να ορισθεί είτε άμεσα, είτε έμμεσα μέσω της ανάλυσης της κατανομής των υπολογισμένων βαθμολογιών για όλες τις προτάσεις.

4.3.4 Στάδιο 4: Περαιτέρω βελτίωση και εμπλουτισμός του σώματος κειμένου με νέες ακουστικές μονάδες

Το τελευταίο στάδιο της διαδικασίας, αν και μπορεί με μια πρώτη ματιά να φαίνεται μικρής σημασίας, είναι πολύ σημαντικό για την εξασφάλιση της υψηλής ποιότητας του συνθέτη σε σπάνια φαινόμενα της γλώσσας τα οποία εντοπίζονται σε διαφορετικά πεδία αναφοράς. Τέτοια φαινόμενα συναντάει κανείς σε ξένες λέξεις, σε ονόματα κ.α. Το στάδιο αυτό, αν και μπορεί να εκτελεστεί παράλληλα με τα υπόλοιπα στάδια, θεωρούμε ότι είναι καλύτερο να εκτελεστεί τελευταίο και μετά την ολοκλήρωση του $2^{ου}$ ή ακόμη και του $3^{ου}$ σταδίου, ούτως ώστε να συμπεριληφθούν όλες οι περιπτώσεις ασυνεχειών ή ελλείψεων που έχουν προκύψει από ατέλειες ή γενικότερα ασυνέπεια στις ηχογραφήσεις. Π.χ. ο ομιλητής είναι πιθανό να μην έχει εκφέρει σωστά διφωνήματα με πολύ μικρή συχνότητα, ή ακόμη και να τα έχει προσπεράσει, με αποτέλεσμα αυτά να μην συμπεριληφθούν στην αρχική βάση δεδομένων.

Η ολοκλήρωση του συγκεκριμένου σταδίου ολοκληρώνεται με ημιαυτόματα μέσα, αφού οι ακουστικές μονάδες που προκύπτουν ότι λείπουν από την αρχική βάση μάλλον δεν περιλαμβάνονται στο αρχικό σώμα κειμένου *P* και επομένως οι νέες προτάσεις θα πρέπει να δημιουργηθούν χειρονακτικά [Founda et al., 2001a].

4.4 Αποτελέσματα

Τα δεδομένα τα οποία χρησιμοποιήσαμε για το σύστημα σύνθεσης φωνής για την Ελληνική γλώσσα είναι δεδομένα ειδησεογραφικού χαρακτήρα, από εφημερίδες και από διαδικτυακές ειδησεογραφικές πύλες. Ο λόγος για τον οποίον επιλέξαμε το συγκεκριμένο πεδίο αναφοράς για τον συνθέτη φωνής ήταν επειδή το συγκεκριμένο πεδίο αν και περιορισμένο (restricted), είναι ταυτόχρονα και απείρως ορισμένο (unlimited) και το γενικότερα ουδέτερο στυλ ανάγνωσης του μπορεί να χρησιμοποιηθεί και σε άλλα διαφορετικά πεδία όπως είναι τα διαλογικά συστήματα ή άλλες φωνητικές εφαρμογές.

Το αρχικό σώμα κειμένου *P* προέκυψε από την συλλογή κειμένων από δύο ελληνικές ειδησεογραφικές πύλες στο διάστημα τριών διαδοχικών ετών. Αποφασίσαμε να μην συσταδοποιήσουμε τα κείμενα ανάλογα με την θεματική τους (π.χ. Αθλητικά, Οικονομικά Νέα κλπ), αλλά να ενσωματώσουμε όλα μαζί σε ένα σώμα κειμένου, δημιουργώντας έτσι ένα γενικότερο σώμα με μεγαλύτερη ποικιλότητα. Το αρχικό σώμα κειμένου *P* αποτελείτο αρχικά

από 42 περίπου μοναδικές διαφορετικές προτάσεις, από τις οποίες αφαιρέσαμε 6 περίπου εκατομμύρια προτάσεις που περιείχαν ξένες λέξεις, είτε ήταν ασυνήθιστα μεγάλες (μεγαλύτερες από 40 λέξεις μήκος). Το σώμα κειμένου των 36 εκατομμυρίων λέξεων περιείχε περίπου 216 λέξεις. Στο σημείο αυτό αξίζει να αναφέρουμε ότι παρόμοιες πρακτικές αναφέρουν ότι ρητά επιλέγουν προτάσεις σχετικά μικρού μήκους (10-12 λέξεις οι μεγαλύτερες) με το σκεπτικό ότι μεγάλες προτάσεις είναι δύσκολο να εκφωνηθούν χωρίς λάθος. Ωστόσο θεωρούμε ότι μεγαλύτερες προτάσεις επίσης πρέπει να συμπεριλαμβάνονται στο ηχογραφημένο σώμα κειμένου αφού τέτοιες προτάσεις είναι συχνά προσωδιακά πλούσιες, ενσωματώνοντας φαινόμενα που δεν συναντιούνται σε μικρότερες.

Όπως έχουμε ήδη αναφέρει και παραπάνω, οι ακουστικές μονάδες που χρησιμοποιήσαμε για την συγκεκριμένη διαδικασία ήταν τα διφωνήματα σε συνδυασμό με τα προσωδιακά τους χαρακτηριστικά. Το διάνυσμα για κάθε ακουστική μονάδα συγκεκριμένα περιλαμβάνει α) το διφώνημα, β) το ακουστικό περικείμενο που προηγείται και που ακολουθεί, γ) την προσωδιακή κατηγορία στην οποία ανήκει και δ) την intra-προσωδιακή σχετική κατηγορία. Τα προσωδιακά χαρακτηριστικά για κάθε ακουστική μονάδα παρήχθησαν από την μηχανή παραγωγής προσωδίας η οποία περιγράφεται στο κεφάλαιο 2 της διατριβής. Κατά αυτόν τον τρόπο, οι μοναδικές ακουστικές μονάδες που ανιχνεύσαμε στο συγκεκριμένο σώμα κειμένου ήταν περίπου 140 χιλιάδες.

Το σώμα κειμένου S που προέκυψε με την βοήθεια του άπληστου αλγορίθμου επιλογής, αριθμούσε περίπου 25 χιλιάδες προτάσεις με μέσο όρο 14,3 λέξεις ανά πρόταση, παρέχοντας πλήρη κάλυψη του συνόλου C των μοναδικών ακουστικών μονάδων που περιγράψαμε παραπάνω. Ένα από τα σημαντικότερα προβλήματα που συναντάει κανείς στην συγκεκριμένη διαδικασία, όπου ένα σύνολο προτάσεων επιλέγεται με βάση ένα διττό κριτήριο μεγιστοποίησης κάλυψης και ελαχιστοποίησης μεγέθους, είναι ότι συχνά επιλέγονται προτάσεις που περιλαμβάνουν ορθογραφικά ή τυπογραφικά λάθη, αφού παρέχουν κάλυψη σπάνιων ακουστικών φαινομένων. Για να ξεπεράσουμε το συγκεκριμένο πρόβλημα, συμπεριλάβαμε έναν επιπλέον μηχανισμό στο συγκεκριμένο στάδιο, με τον οποίο μας επιτρεπόταν η γρήγορη επίβλεψη του τελικού σώματος κειμένου και οι ακουστικές μονάδες για τις οποίες κάθε πρόταση είχε επιλεχθεί. Αν κάποια πρόταση περιείχε λάθη και έπρεπε να αφαιρεθεί, τότε ο συγκεκριμένος μηχανισμός πρότεινε

εναλλακτική πρόταση ή προτάσεις, οι οποίες συμπεριλάμβαναν τις ακουστικές μονάδες για τις οποίες οι αφαιρούμενες προτάσεις παρείχαν κάλυψη. Η συγκεκριμένη διαδικασία κυρίως εστιάστηκε σε ακουστικές μονάδες οι συχνότητες των οποίων ήταν σχετικά μικρή και οφειλόταν στο γεγονός ότι προέρχονταν κυρίως από τυπογραφικά σφάλματα.

Το επόμενο στάδιο περιελάμβανε την δημιουργία μίας βάσης δεδομένων για τον συνθέτη φωνής, χωρίς όμως να περιλαμβάνει τις αντίστοιχες ηχογραφήσεις. Το υποσύστημα βέλτιστης επιλογής ακουστικών μονάδων προσαρμόστηκε κατάλληλα ώστε να κάνει χρήση μόνο των αντικειμενικών κριτηρίων και όχι των κριτηρίων που αναφέρονται στο σήμα φωνής. Με την βοήθεια του συγκεκριμένου συστήματος, προσομοιώσαμε την σύνθεση όλων των προτάσεων του σώματος *P,* εξαιρώντας από αυτό ωστόσο τις προτάσεις που περιλαμβάνονταν στο σώμα *S*. Κατά την σύνθεση, στατιστικά στοιχεία για την χρήση κάθε ακουστικής μονάδας αποθηκεύτηκαν κατάλληλα και στην συνέχεια χρησιμοποιήθηκαν σε επίπεδο πρότασης, ούτως ώστε κάθε πρόταση να χαρακτηριστεί ανάλογα με τα στοιχεία αυτά. Στην συνέχεια, οι προτάσεις σταχυολογήθηκαν με βάση τις ακουστικές μονάδες που περιείχαν, αλλά και τα χαρακτηριστικά αυτών όπως προέκυψαν από τα στατιστικά στοιχεία χρήσης τους κατά την προσομοίωση σύνθεσης. Το σώμα των προτάσεων που επιλέχθηκε από το στάδιο αυτό ήταν 4.000 χιλιάδες προτάσεις περίπου, προσφέροντας κάλυψη τόσο σε φωνητικό, όσο και σε προσωδιακό επίπεδο.

Τα δύο τελευταία στάδια της διαδικασίας εκτελέστηκαν αφού είχαν ολοκληρωθεί οι ηχογραφήσεις και η ενσωμάτωσή τους στην βάση δεδομένων του συνθέτη φωνής, με βάση το σώμα κειμένου S_{sub} όπως προέκυψε από το αμέσως προηγούμενο στάδιο. Συνθέτοντας το σύνολο των προτάσεων του αρχικού σώματος κειμένου *P*, ανιχνεύσαμε αυτόματα τις προτάσεις που παρουσίαζαν τις μεγαλύτερες ή περισσότερες ασυνέχειες. Ως κριτήριο ανίχνευσης ασυνεχειών χρησιμοποιήσαμε το κανονικοποιημένο κόστος επιλογής κάθε πρότασης, όπως αυτό προέκυψε από υποσύστημα βέλτιστης επιλογής.

Εξίσωση 2: Συνάρτηση κόστους στον αλγόριθμο επιλογής βέλτιστων ακουστικών μονάδων.

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} W^t \cdot C^t(t_i, u_i) + \sum_{i=2}^{n} W^c \cdot C^c(u_{i-1}, u_i)$$

Κατά αυτόν τον τρόπο, ανιχνεύσαμε περίπου 2.000 ακουστικές μονάδες οι οποίες έπρεπε να καλυφθούν μέσω νέων ηχογραφήσεων, και οι οποίες περιλαμβάνονταν σε 220 διαφορετικές προτάσεις που επιλέχθηκαν αυτόματα από το αρχικό σώμα κειμένου *P*.

Το τελικό στάδιο της διαδικασίας περιελάμβανε την δημιουργία του φωνητικού χάρτη κάλυψης για την συγκεκριμένη γλώσσα, και την χειρονακτική παραγωγή προτάσεων που περιείχαν τα διφωνήματα που απουσίαζαν από τον συγκεκριμένο χάρτη. Αυτά καλύφθηκαν μέσω 32 νέων προτάσεων που περιείχαν είτε ανούσιες λέξεις, είτε ξένες λέξεις, είτε κατάλληλο συνδυασμό των δύο. Οι προτάσεις αυτές ηχογραφήθηκαν επίσης και ενσωματώθηκαν στην τελική βάση του συνθέτη φωνής.

	-	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	r	s	t	u	v	x	z	ά	έ	ί	ό	ύ
-	0	203	27	363	16	129	27	6	0	144	52	97	25	169	47	110	181	11	128	186	4	33	18	16	31	80	153	100	8
a	408	182	53	197	55	173	398	11	31	107	89	398	395	428	1087	52	920	628	1109	1233	15	151	68	85	20	101	81	51	14
b	4	50	1	0	2	31	1	1	0	38	3	1	12	2	2	81	1	22	4	1	16	1	2	1	39	24	25	14	12
c	0	18	0	0	0	1112	0	0	0	348	0	0	0	0	0	2	0	0	0	0	3	0	0	0	10	150	604	3	4
d	14	174	6	2	3	184	1	2	0	119	5	2	5	4	1	86	4	54	11	2	7	2	2	83	54	17	66	16	4
e	93	221	26	95	21	121	247	13	26	173	100	485	376	247	626	257	566	843	863	1026	5	192	81	70	66	67	28	87	5
f	20	132	1	9	1	81	1	2	0	137	0	45	5	3	8	218	5	51	37	264	6	1	5	2	52	88	107	55	17
g	10	28	1	1	3	1	1	1	0	1	1	2	14	4	1	11	1	33	2	5	8	1	2	1	23	1	1	18	7
h	0	0	0	2	0	0	0	44	0	0	0	8	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0
i	229	507	34	669	35	240	94	12	30	150	103	891	293	665	1120	353	553	342	2117	876	40	116	104	166	141	104	58	139	20
j	0	401	0	0	0	122	0	0	0	199	0	0	0	0	0	17	0	0	0	0	11	0	0	0	67	86	164	36	16
k	25	626	1	2	3	1	20	1	0	1	3	1	122	6	10	226	25	221	518	191	58	3	3	1	340	1	1	456	125
l	18	240	11	5	13	265	11	7	0	653	1	10	2	17	8	386	11	4	12	23	81	7	2	3	239	211	412	175	38
m	13	686	51	1	4	933	5	2	0	374	1	1	2	3	8	294	19	3	8	10	85	2	2	1	137	405	240	134	57
n	249	1130	22	91	601	696	33	9	2	830	25	102	27	101	50	621	232	15	195	274	115	27	38	6	158	224	375	286	31
o	178	125	31	119	75	112	141	14	3	99	182	255	461	394	1283	50	505	501	1156	536	14	123	87	41	30	47	64	23	2
p	2	469	1	2	1	385	1	1	0	418	1	2	159	2	14	584	2	558	154	116	356	2	1	1	165	141	252	428	33
r	32	765	15	59	20	208	15	10	0	1026	122	38	13	110	69	798	11	1	43	84	100	13	42	4	329	139	564	388	68
s	674	572	22	273	8	832	100	4	0	1816	64	168	47	148	75	319	381	26	269	1816	118	65	85	14	121	117	456	106	14
t	17	977	2	3	3	714	2	1	0	2602	3	2	11	15	3	1535	2	322	120	3	720	1	1	2	264	225	391	368	68
u	149	86	19	44	8	73	14	9	2	47	17	64	78	125	267	26	114	133	428	75	2	21	28	8	17	47	33	8	1
v	2	172	1	2	2	65	1	2	0	81	21	1	71	26	14	71	2	188	3	1	52	1	1	1	93	46	67	48	15
x	4	140	1	1	1	1	1	1	0	1	1	1	5	7	48	129	1	153	1	28	67	1	1	1	54	1	1	115	2
z	9	43	15	1	2	86	0	1	0	118	6	3	2	192	3	87	2	2	3	3	50	20	1	2	25	20	54	24	13
ά	100	34	12	79	13	26	67	14	14	57	38	82	256	127	362	31	118	223	311	277	12	69	33	109	6	7	4	7	1
έ	17	90	6	7	13	17	91	5	14	29	26	151	168	83	576	93	98	292	504	130	37	86	93	26	1	1	2	4	1
ί	156	594	14	97	15	146	37	4	20	74	55	119	100	242	521	193	229	155	729	415	87	31	49	121	18	21	11	12	2
ό	71	35	16	48	11	46	30	4	5	44	33	57	148	222	523	18	207	284	577	611	6	30	31	32	14	14	8	9	1
ύ	41	7	2	12	4	19	10	7	1	7	3	9	40	46	131	2	29	37	136	56	2	3	5	8	1	2	3	1	2

Σχήμα 30: Χάρτης φωνητικής κάλυψης του επιλεγμένου σώματος κειμένου. Για κάθε διφώνημα αναφέρεται η συχνότητα εμφάνισης. Η κατακόρυφη στήλη ορίζει το αριστερό φώνημα και η οριζόντια γραμμή το δεξί φώνημα για κάθε διφώνημα.

4.5 Δημιουργία ηχογραφημένου σώματος κειμένου

Ένα από τα βασικότερα στάδια της δημιουργίας της βάσης δεδομένων για τον συνθέτη φωνής, είναι αυτό της δημιουργίας του ηχογραφημένου σώματος κειμένου, όσο και ο τεμαχισμός αλλά και η επισημείωση των απαραίτητων χρονικών σημείων. Η συγκεκριμένη διαδικασία περιλαμβάνει 3 βασικά στάδια, τα οποία περιλαμβάνουν την ηχογράφηση του σώματος κειμένου, την επεξεργασία των ηχογραφήσεων για διόρθωση ατελειών, και την επεξεργασία του σήματος ώστε να επισημειωθούν κατάλληλα όλες τα απαραίτητα φαινόμενα και χρονικές στιγμές κατά μήκος του σήματος φωνής. Οι επόμενες παράγραφοι περιγράφουν με περισσότερη λεπτομέρεια τα στάδια αυτά.

4.5.1 Ηχογράφηση σώματος κειμένου

Η ηχογράφηση των προτάσεων του σώματος κειμένου που έχει σχεδιασθεί πραγματοποιείται σε ειδικό στούντιο, με ειδικό εξοπλισμό και κατάλληλες συνθήκες. Είναι προφανές ότι σημαντικό ρόλο στην ποιότητα των ηχογραφήσεων, εκτός από τον ίδιο τον ομιλητή, διαδραματίζει τόσο ο μηχανικός ήχου, όσο και ο εξοπλισμός και οι συνθήκες. Κατά την ηχογράφηση τρεις σημαντικές παράμετροι πρέπει να παρατηρούνται ώστε να διατηρούνται όσο το δυνατόν σταθεροί κατά την διάρκεια των ηχογραφήσεων:

α) η ποιότητα της φωνής του ομιλητή,

β) η ταχύτητα εκφώνησης και

γ) η διατήρηση των ρυθμίσεων του υλικού/λογισμικού ηχογράφησης καθ' όλη την διάρκεια των ηχογραφήσεων.

Η ποιότητα της φωνής του ομιλητή εξαρτάται συχνά από τον ίδιο τον ομιλητή, ο οποίος μπορεί είτε να παρουσιάζει υψηλή αντοχή σε εκτενείς ηχογραφήσεις, είτε να βραχνιάζει εύκολα, είτε να μεταβάλει την φωνή του ανάλογα με την κούραση. Σημαντικό εργαλείο για την παρακολούθηση της συνολικής ποιότητας της φωνής αποτελεί ένας μικρός αριθμός προτάσεων που καλείται ο ομιλητής να ακούσει και να εκφωνήσει, ελέγχοντας ταυτόχρονα την χροιά, αλλά και προσωδία, χαρακτηριστικά που οφείλει να διατηρεί σχετικά συνεπή ως προς τον τρόπο εκφοράς από τον ομιλητή. Στην ποιότητα των ηχογραφήσεων μπορεί κανείς να συμπεριλάβει και την ταχύτητα

εκφώνησης, η οποία επίσης θα πρέπει να διατηρείται όσο το δυνατόν σταθερή ή τουλάχιστον συνεπής κατά μήκος των ηχογραφήσεων. Οι ίδιες προτάσεις που χρησιμοποιούνται ως benchmark ελέγχου για την ποιότητα της φωνής του ομιλητή χρησιμοποιούνται και στην συγκεκριμένη περίπτωση, με απώτερο σκοπό να υπενθυμίζει στον ομιλητή την ταχύτητα ή το ειδικό ύφος εκφώνησης που ο ίδιος οφείλει να διατηρήσει.

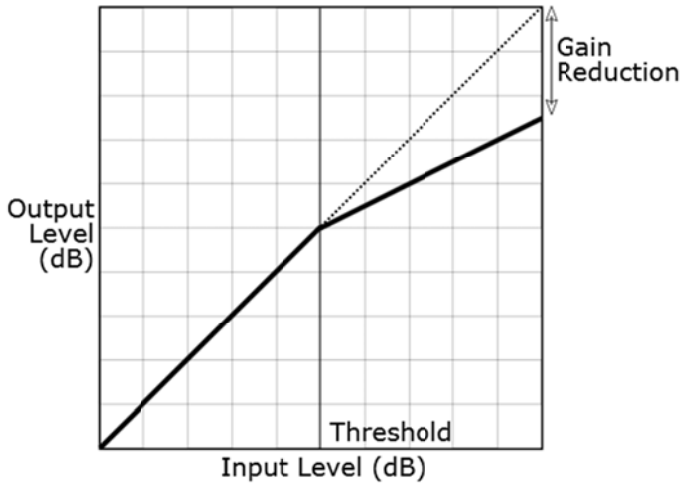
Τέλος σημαντικό ρόλο στην ποιότητα των ηχογραφήσεων αποτελεί η διατήρηση των ίδιων ρυθμίσεων, τόσο από πλευρά υλικού, όσο και από πλευρά λογισμικού, για όλες τις συνεδρίες ηχογραφήσεων, οι οποίες μπορεί να διαρκέσουν μερικές ημέρες ή ακόμη και εβδομάδες, ανάλογα με το μήκος του σώματος κειμένου αλλά και την ικανότητα του ομιλητή. Στην περίπτωση ωστόσο όπου ο ομιλητής καλείται να εκφωνήσει νέο σώμα κειμένου μετά από μεγάλο διάστημα, και συχνά ακόμη σε διαφορετική υποδομή, είναι σημαντικό οι συνθήκες ηχογράφησης των νέων δεδομένων, να μην διαφέρουν σημαντικά από τις αρχικές. Στην αντίθετη περίπτωση μπορεί κανείς να επιστρατεύσει μεθόδους φασματικής κανονικοποίησης για να φέρει το σύνολο των ηχογραφήσεων στα ίδια ποιοτικά επίπεδα.

4.5.2 Επεξεργασία ηχογραφήσεων

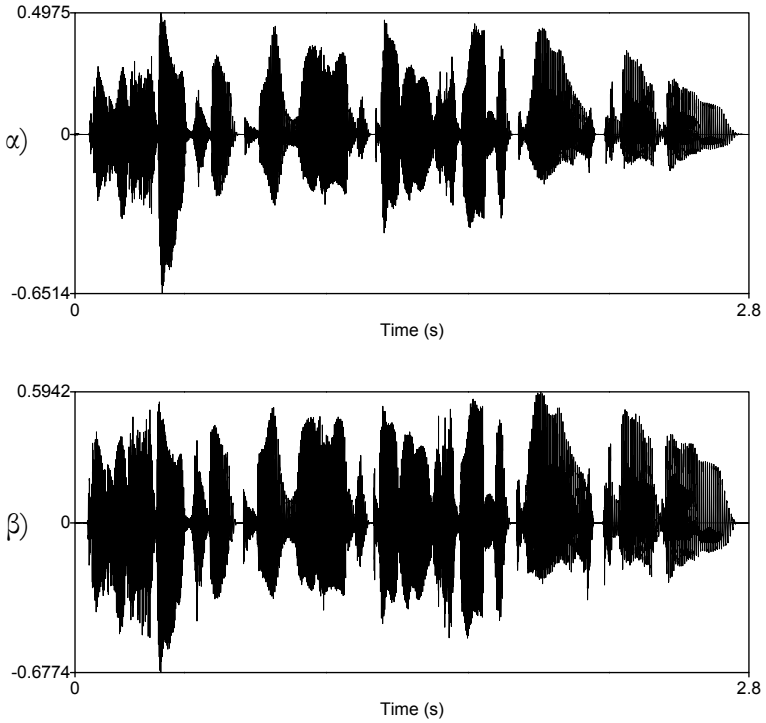
4.5.2.1 Δυναμική συμπίεση του σήματος φωνής

Το σύνολο των ηχογραφήσεων παρουσιάζει διακυμάνσεις στην ένταση του σήματος φωνής, αλλά και διαφοροποιήσεις στο φασματικό περιεχόμενο, από συνεδρία σε συνεδρία, ειδικά αν οι συνθήκες ή οι ρυθμίσεις του εξοπλισμού ηχογράφησης έχουν μεταβληθεί. Για την κανονικοποίηση της έντασης του σήματος φωνής, χρησιμοποιούμε έναν αλγόριθμο δυναμικής συμπίεσης της έντασης φωνής, η οποία αποσκοπεί στον περιορισμό του σήματος φωνής σε ένα συγκεκριμένο εύρος δυναμικής. Κατά αυτόν τον τρόπο, αποφεύγονται φαινόμενα clipping, ή ασύμμετρης έντασης στο τελικό συνθετικό σήμα.

Ο συγκεκριμένος αλγόριθμος μετατρέπει υψηλής ενέργειας ήχους σε χαμηλότερης, ενώ διατηρεί τους ήχους χαμηλής στάθμης αναλλοίωτους. Για να το επιτύχει αυτό, ο αλγόριθμος μειώνει την ένταση σε παράθυρα όπου ο ήχος ξεπερνάει συγκεκριμένο κατώφλι σε dB, πολλαπλασιάζοντας τα με ένα κέρδος μικρότερο της μονάδας, ανάλογα με την κλίμακα του συμπιεστή. Η κλίμακα που χρησιμοποιούμε είναι 4:1, δηλαδή μετατρέπει είσοδο των 4dB σε 1dB.



Σχήμα 31: Διάγραμμα συμπίεσης ήχου

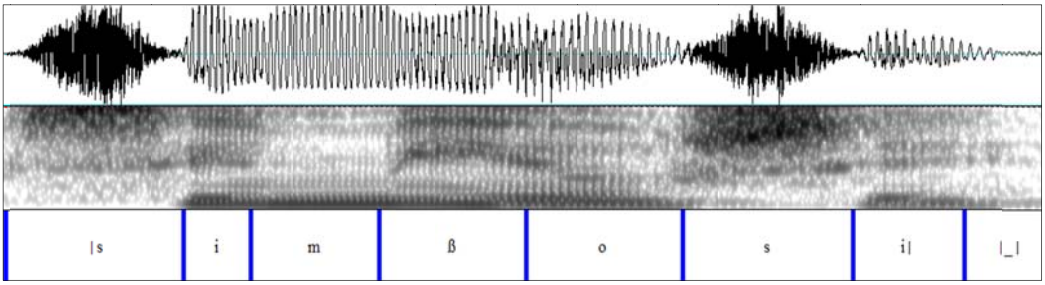


Σχήμα 32: Σχηματική αναπαράσταση της κυματομορφής μιας ηχογράφησης όπως ηχογραφήθηκε (α) και κατόπιν της δυναμικής συμπίεσης της έντασης (β). Η συγκεκριμένη διαδικασία επιτρέπει την κανονικοποίηση των ηχογραφήσεων στα -20db RMS.

Κατά αυτόν τον τρόπο, επιτυγχάνουμε να φέρουμε το σύνολο των ηχογραφήσεων σε παρόμοια ενεργειακή στάθμη, χωρίς να απαιτείται περαιτέρω κανονικοποίηση της έντασης του σήματος φωνής σε επίπεδο φωνήματος, όπως συχνά προτείνεται μέχρι σήμερα, με σημαντικά προβλήματα που προκύπτουν κατά την κατάτμηση των ηχογραφήσεων.

4.5.3 Αυτόματη κατάτμηση ηχογραφημένου σώματος κειμένου

Μία από τις σημαντικότερες διαδικασίες στον σχεδιασμό και ανάπτυξη ενός συνθέτη φωνής αποτελεί η διαδικασία των ηχογραφήσεων και η κατάτμηση αυτών, η οποία ουσιαστικά συνίσταται στην επισημείωση των ακουστικών αρχείων σε επίπεδο φωνήματος. Στην βιβλιογραφία έχουν υπάρξει πλήθος δημοσιεύσεων αναφορικά με την αυτόματη επισημείωση των ηχογραφήσεων με διάφορες μεθόδους, ωστόσο καμία μέχρι σήμερα δεν παρουσιάζει ολοκληρωτική υπεροχή σε σχέση με τις υπόλοιπες. Η μεθοδολογία που προτείνουμε για την συγκεκριμένη διαδικασία βασίζεται σε αυτόματη κατάτμηση με τον συνδυασμό εργαλείων εξαναγκασμένης αντιστοίχισης μέσω HMM ευριστικών μεθόδων για την ανίχνευση σημαντικών τοπικών προβλημάτων που οδηγούνται προς χειρωνακτική επόπτευση και διόρθωση.



Σχήμα 33: Παράδειγμα κατάτμησης σήματος φωνής (εκφορά της λέξης /σημείωση/) με βάση φωνήματα.

Για την εκπαίδευση του συστήματος HMM επιλέγουμε αυτόματα ένα σύνολο ηχογραφήσεων, το οποίο περιλαμβάνει τουλάχιστον 5 διαφορετικά στιγμιότυπα από κάθε φώνημα, με διαφορετικό περικείμενο κάθε φορά. Το σύνολο των προτάσεων αυτών τεμαχίζεται αυτόματα και διορθώνεται χειροκίνητα, ούτως ώστε τα όρια των φωνημάτων να είναι ακριβή και με ελάχιστα σφάλματα. Το σετ αυτό των προτάσεων, το οποίο συνήθως αποτελείται από 50 προτάσεις, αποτελεί στην συνέχεια το υλικό εκπαίδευσης για το σύστημα αυτόματης εξαναγκασμένης αντιστοίχησης μέσω

HMM. Σημαντικό ρόλο στην ορθότητα των αποτελεσμάτων διαδραματίζει η ορθότητα των σημείων στίξης στο αρχικό κείμενο και η ύπαρξη αυτών στις ηχογραφήσεις. Σε διαφορετική περίπτωση, όπου ένα σημείο στίξης διαφωνεί με την ηχογράφηση, το σύστημα εξαναγκασμένης αντιστοίχισης θα προσπαθήσει να επισημειώσει διαστήματα σιωπής ενώ στην πραγματικότητα δεν υπάρχουν, εισάγοντας σφάλμα στην βάση δεδομένων σε πολλαπλό επίπεδο.

Η ύπαρξη σφαλμάτων λόγω ασυμφωνίας του γραπτού κειμένου με τις αντίστοιχες εκφωνήσεις, αλλά και λόγω εγγενών προβλημάτων στην αναγνώριση φωνής, αποτελεί έναν από τους σημαντικότερους παράγοντες που επηρεάζουν αρνητικά την ποιότητα του συνθέτη φωνής και πρέπει να αντιμετωπίζονται με το βέλτιστο και αποδοτικότερο δυνατό τρόπο. Για τον λόγο αυτόν, αναπτύξαμε έναν μηχανισμό για την αυτοματοποιημένη ανίχνευση πιθανών σημείων σφάλματος στις επισημειωμένες ηχογραφήσεις, τα οποία αρχικά αφαιρούνται από την βάση δεδομένων, και σε επόμενη φάση διορθώνονται χειροκίνητα. Ο μηχανισμός αυτός περιλαμβάνει την σάρωση των ηχογραφήσεων, και την ιεραρχική συσταδοποίηση όλων των στιγμιοτύπων για κάθε διαφορετικό φώνημα. Για την συσταδοποίηση αυτή γίνεται χρήση της ευκλείδιας απόστασης των συντελεστών MFCC σε παράθυρο 15 msec στο μέσο κάθε φωνήματος.

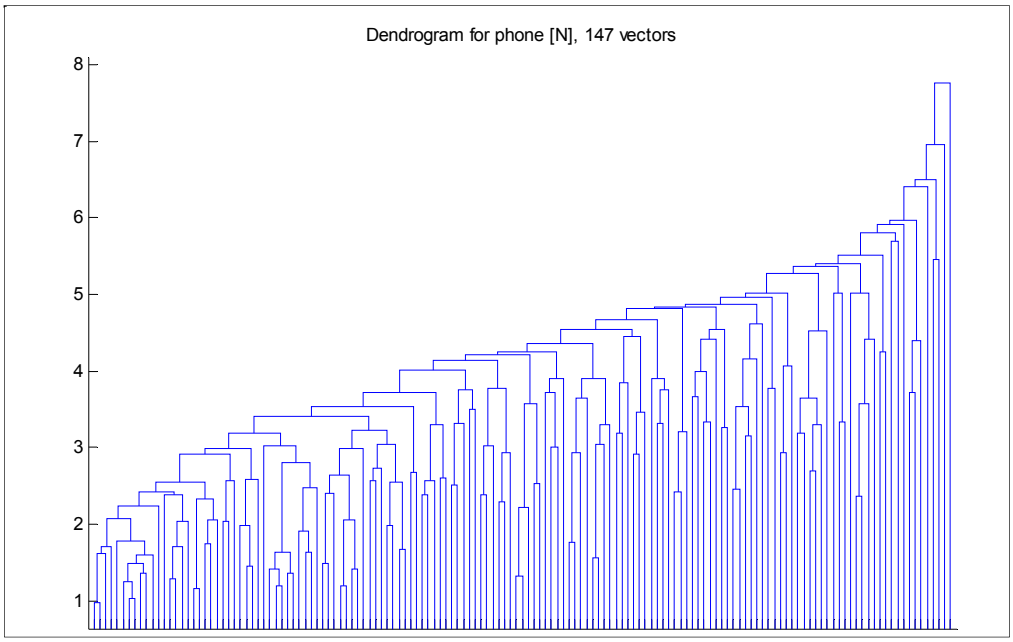
$$Power\ Cepstrum = |F\{log(|F\{f(t)\}|^2)\}|^2$$

Εξίσωση 3: Ο Cepstrum ενός σήματος.

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

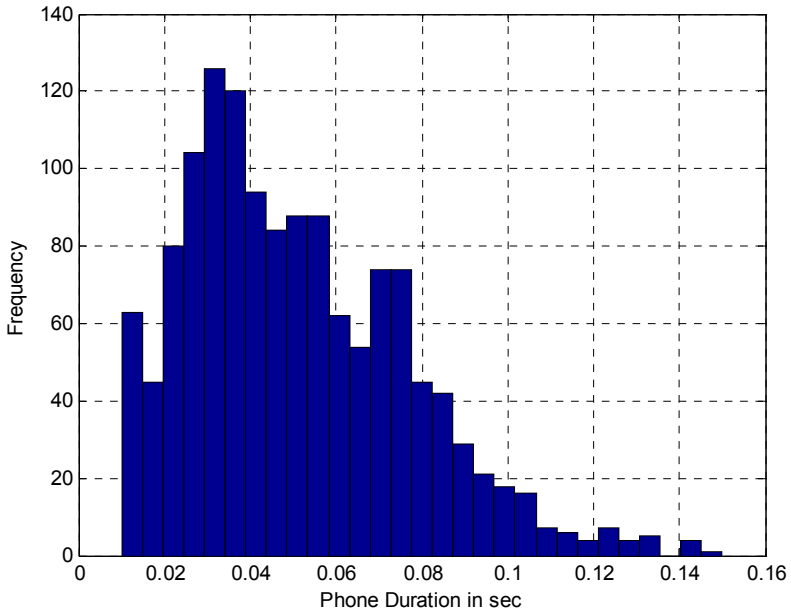
Εξίσωση 4: Ευκλίδεια απόσταση δύο διανυσμάτων.

Υπολογίζοντας δυναμικά ένα κατώφλι αποκοπής, οι συστάδες που διαφέρουν σημαντικά από το υπόλοιπο πλήθος επισημειώνονται και αφαιρούνται από την βάση δεδομένων. Το κατώφλι αποκοπής υπολογίζεται δυναμικά, ανάλογα με τον τύπο του φωνήματος και τον αριθμό των στιγμιοτύπων ανά φώνημα.



Σχήμα 34: Απεικόνιση ιεραχικής συσταδοποίησης για το φώνημα /N/. Οι τελευταίες συστάδες, με τα λιγότερα μέλη ανά συστάδα και με σημαντική απόσταση από τις κοντινότερες συστάδες, αφαιρούνται αυτόματα από την βάση δεδομένων αφού περιέχουν ενδεχόμενα σφάλματα.

Παράλληλα, υπολογίζοντας την πυκνότητα πιθανότητας για την διάρκεια κάθε φωνήματος, αφαιρούμε από την βάση δεδομένων ακουστικές μονάδες οι οποίες χαρακτηρίζονται outliers σε μία κατά σύμβαση κανονική κατανομή πιθανότητας.



Σχήμα 35: Καμπύλη πυκνότητας πιθανότητας διαρκειών για το φώνημα /ε/ στην βάση δεδομένων. Τα στιγμιότυπα που χαρακτηρίζονται ως outliers αφαιρούνται αυτόματα από την βάση δεδομένων.

Κατά αυτόν τον τρόπο, η διαδικασία τεμαχισμού και δημιουργίας της βάσης δεδομένων, με ταυτόχρονη σταχυολόγηση των δεδομένων, είναι πλήρως αυτόματη, οδηγώντας έτσι στην δημιουργία νέων φωνών και λειτουργικών βάσεων δεδομένων σε ελάχιστο χρόνο και με ελάχιστο κόπο. Η επίβλεψη και διόρθωση των εξαιρουμένων ακουστικών μονάδων από την βάση δεδομένων μπορεί να πραγματοποιηθεί σε μετέπειτα στάδιο, το οποίο και πρόκειται να επιφέρει περαιτέρω βελτίωση της ποιότητας της βάσης δεδομένων του συνθέτη φωνής. Στην περίπτωση στην οποία αναφερόμαστε, ο συγκεκριμένος μηχανισμός επέφερε μείωση στην βάση δεδομένων κατά 4% περίπου, βελτιώνοντας όμως ουσιαστικά τον επίπεδο ποιότητας του συνθέτη.

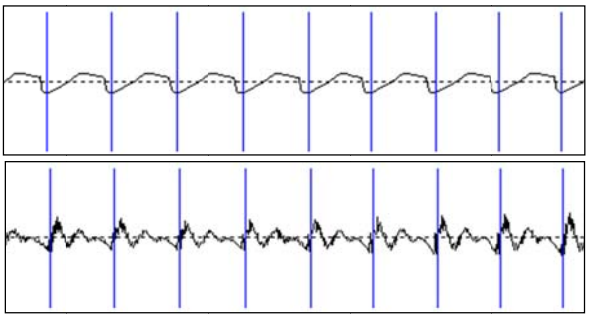
4.6 Υπολογισμός των σημείων ανάλυσης του αρχικού σήματος (Pitchmarks)

4.6.1 Εισαγωγικά

Όπως έχει ήδη αναφερθεί, τα συγκεκριμένα σημεία διαδραματίζουν σημαντικό ρόλο στην ποιότητα του τελικού σήματος, αλλά και στον βαθμό στον οποίο επιτρέπεται στο σύστημα να εκτελέσει μετατροπές, τόσο στην θεμελιώδη συχνότητα, όσο και στην διάρκεια του σήματος. Τα

σημεία ανάλυσης ορίζουν το κέντρο των παραθύρων για κάθε ένα ST-Signal που χρησιμοποιείται από την μηχανή σύνθεσης φωνής για την δημιουργία ενός ομαλού συνθετικού σήματος φωνής. Η απόσταση μεταξύ δύο διαδοχικών σημείων ανάλυσης αντιστοιχεί στην τοπική τιμή της θεμελιώδους περιόδου για το σήμα φωνής και απαιτείται η ακριβής εκτίμηση των σημείων αυτών, αφού η ορθή ή μη εκτίμησή τους επηρεάζει σημαντικά την τελική ποιότητα του συνθετικού σήματος. Στην πραγματικότητα, όσο περισσότερο ασυνεπή και ανακριβή είναι τα σημεία ανάλυσης, τόσο περισσότερο παραμορφωμένη και με αφύσικη τραχύτητα παρουσιάζεται η τελική συνθετική φωνή.

Πολλές σύγχρονοι μέθοδοι βασίζονται στην χρήση του σήματος που καταγράφει ένας ηλεκτρολαρυγγογράφος (EGG) ο οποίος τοποθετείται εξωτερικά στον λάρυγγα του ομιλητή κατά την διάρκεια της ηχογράφησης και η εγγραφή του συγχρονίζεται με την εγγραφή του σήματος φωνής. Οι τεχνικές αυτές επιτυγχάνουν να εκτιμήσουν με ακρίβεια τα σημεία ανοίγματος και κλεισίματος της γλωττίδας (GCO και GCI χρονικές στιγμές), αφού το σήμα του λαρυγγογράφου είναι ιδιαίτερα απλό στην επεξεργασία σε σχέση με το αντίστοιχο σήμα της φωνής και η περιοδικότητά του εκτιμάται με ακρίβεια.



Σχήμα 36: Κυματομορφή ηλεκτρο-λαρυγγογράφου και η αντίστοιχη κυματομορφή του σήματος φωνής. Με κατακόρυφες γραμμές επισημειώνονται τα σημεία κλεισίματος της γλωττίδας (GCI).

Ωστόσο το σημαντικότερο μειονέκτημα των μεθόδων αυτών είναι η ανάγκη χρήσης ενός εξειδικευμένου οργάνου όπως είναι ο λαρυγγογράφος, αλλά και φυσικά η ενόχληση που ο ίδιος δημιουργεί στον ομιλητή κατά την διάρκεια των ηχογραφήσεων.

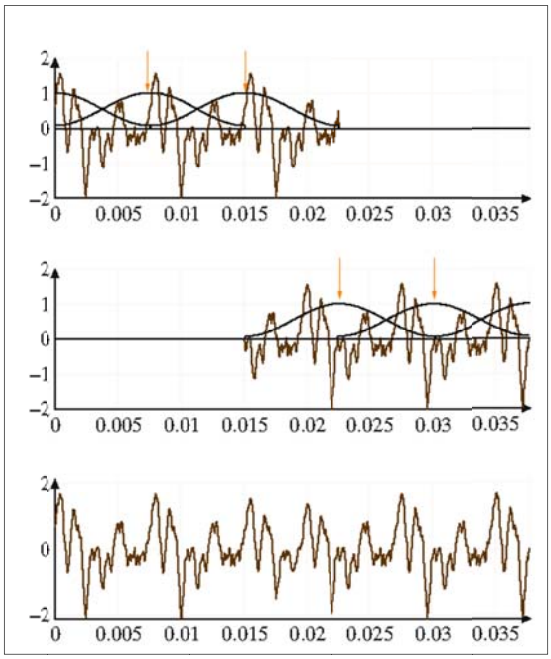
Σχήμα 37: Λαρυγγογράφος και τοποθέτησή του στον ομιλητή. Πηγή: Inst. of Phonetic Sciences in Amsterdam.

Στην βιβλιογραφία, άλλες προσπάθειες πολύ συχνά παραδέχονται ότι η επισημείωση των σημείων ανάλυσης στο σήμα φωνής επιτυγχάνεται χειρωνακτικά, διαδικασία που είναι τόσο επιρρεπής σε σφάλματα, όσο και ιδιαίτερα χρονοβόρα αλλά και επίπονη, αφού το αντικείμενο της επεξεργασίας είναι συχνά αρκετές ώρες ηχογραφήσεων.

Όσον αφορά στις μεθόδους που επιχειρούν να εντοπίσουν κατάλληλα σημεία ανάλυσης αυτόματα μέσα από το σήμα φωνής, είναι αρκετές αυτές που έχουν προταθεί στην βιβλιογραφία, με την πλειοψηφία τους να προσπαθούν να εντοπίσουν είτε το μέγιστο/ελάχιστο σημείο σε μία θεμελιώδη περίοδο [Dologlou1989], είτε τα ίδια τα GCI/GCO όπως αυτά θα υπολογιζόντουσαν μέσω του αντίστοιχου σήματος του λαρυγγογράφου. Πολλές από τις προαναφερόμενες μεθόδους ενσωματώνουν αλγορίθμους δυναμικού προγραμματισμού και επιτυγχάνουν να εντοπίζουν τα σημεία ανάλυσης pitchmarks με καλή ακρίβεια [Naylor2007]. Ειδικότερα ο αλγόριθμος DYPSA φαίνεται ότι επιτυγχάνει να εντοπίσει με σχετικά μεγάλη ακρίβεια τα σημεία CGI μέσα στο σήμα φωνής, βασιζόμενος κυρίως σε μεθόδους δυναμικού προγραμματισμού. Ωστόσο, ο πειραματισμός μας με τις περισσότερες από αυτές έδειξε ότι συχνά ήταν είτε ανεπαρκείς, είτε δύσκολα προσαρμόσιμες σε άλλες φωνές, με αποτέλεσμα να μην είναι δυνατή η χρήση τους σε

ένα σύστημα σύνθεσης φωνής που αποσκοπεί να εξυπηρετήσει πλήθος φωνών και γλωσσών εντελώς αυτοματοποιημένα.

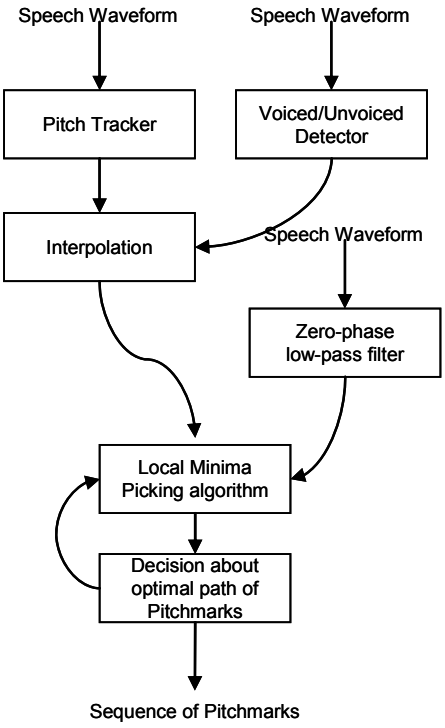
Όπως έχει διαφανεί τόσο από δικά μας πειράματα, όσο και από συναφή δημοσιεύσεις, το σημαντικότερο χαρακτηριστικό που πρέπει να έχουν τα σημεία ανάλυσης pitchmarks, εκτός από την ακρίβεια, είναι η συνέπεια (consistency) στον εντοπισμό των σημείων μέσα σε διαφορετικές ηχογραφήσεις. Με άλλα λόγια, τα σημεία ανάλυσης που εντοπίζονται για ένα /α/ σε μία ηχογράφηση, πρέπει να είναι συνεπή με τα αντίστοιχα σημεία ανάλυσης μιας εναλλακτικής πραγμάτωσης του ίδιου φωνήματος σε διαφορετικό σημείο της ηχογράφησης, ή ακόμη και σε διαφορετική ηχογράφηση. Σε διαφορετική περίπτωση, αν ο αλγόριθμος παραγωγής συνθετικού σήματος κληθεί να συρράψει φωνήματα των οποίων τα σημεία ανάλυσης παρουσιάζουν διαφορά φάσης, τότε το τελικό σήμα παρουσιάζει σημαντική παραμόρφωση η οποία γίνεται αντιληπτή από το ανθρώπινο αυτί. Στο παρακάτω σχήμα παρουσιάζονται δύο ακουστικές μονάδες με ασυνεπή σημεία ανάλυσης, η συρραφή των οποίων δημιουργεί σημαντική παραμόρφωση στο σήμα φωνής. Στο σχήμα, τα σημεία ανάλυσης (pitchmarks) επισημειώνονται με κατακόρυφα πορτοκαλί βέλη, όπου και μπορεί κανείς να διακρίνει την ασυνέπεια των σημείων μέσα στην νοητή περίοδο του σήματος.



Σχήμα 38: Παράθεση με TD-PSOLA δύο ακουστικών μονάδων φωνής με λάθος επισημειωμένα σημεία ανάλυσης (pitchmarks). Η διαφορά φάσης στα σημεία ανάλυσης προκαλεί παραμόρφωση του σήματος κατά την παράθεση με επικάλυψη. Πηγή: [Stylianou2001]

4.6.2 Ο προτεινόμενος αλγόριθμος

Η προσέγγιση μας [Chalamandaris et al., 2009b] αποτελεί μία εναλλακτική μέθοδο στο ειδικότερο πλαίσιο της σύνθεσης φωνής και στις απαιτήσεις τις οποίες η συγκεκριμένη τεχνολογία ορίζει. Ο προτεινόμενος αλγόριθμος ενσωματώνει τεχνικές δυναμικού προγραμματισμού για να εντοπίσει σημεία ανάλυσης pitchmarks μέσα στο σήμα, λαμβάνοντας ταυτόχρονα υπόψη γνώση της τροχιάς της θεμελιώδους συχνότητας του σήματος τοπικά. Στο παρακάτω σχήμα μπορεί κανείς να διακρίνει την ροή διαδικασιών του συγκεκριμένου αλγορίθμου.



Σχήμα 39: Σχηματική αναπαράσταση του αλγορίθμου εντοπισμού σημείων ανάλυσης Pitchmarks.

Η μέθοδος που περιγράφεται σχηματικά απαιτεί περισσότερη ανάλυση, η οποία ακολουθεί στις επόμενες παραγράφους όπου παρουσιάζονται και αναλύονται τα επιμέρους υποσυστήματα και στάδια αυτής.

4.6.2.1 Μέθοδος υπολογισμού της καμπύλης της θεμελιώδους συχνότητας

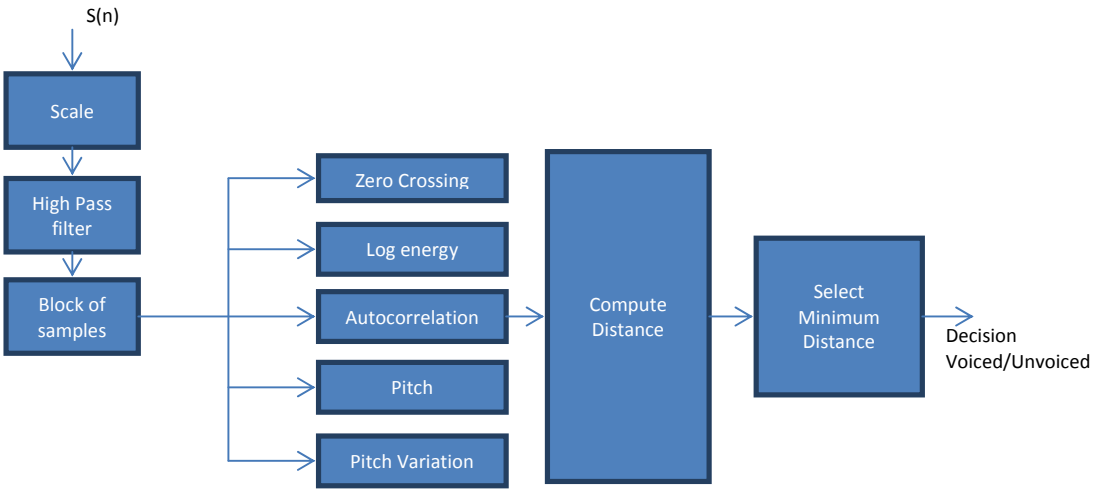
Το πρώτο στάδιο του αλγορίθμου μας περιλαμβάνει τον ακριβή υπολογισμό της καμπύλης της θεμελιώδους συχνότητας του σήματος της φωνής. Αν και η συγκεκριμένη διαδικασία είναι συχνά επιρρεπής σε σφάλματα, ειδικά όταν υπεισέρχεται θόρυβος στο σήμα φωνής, στην περίπτωση των δικών μας ηχογραφήσεων που προορίζονται για την δημιουργία συνθετικής φωνής, δεν υπάρχει το ενδεχόμενο να περιλαμβάνουν περιοχές με θόρυβο, αφού πραγματοποιούνται σε αναχοϊκό θάλαμο με αυστηρή επίβλεψη. Παρ' όλα αυτά και με στόχο την υψηλότερη δυνατή ακρίβεια και αποκλεισμό λανθασμένων εκτιμήσεων, χρησιμοποιούμε έναν εκτιμητή θεμελιώδους συχνότητας που βασίζεται στην απόφαση κατά πλειοψηφία από την εκτίμηση τριών διαφορετικών εκτιμητών συχνότητας. Πιο συγκεκριμένα χρησιμοποιούμε έναν εκτιμητή που κάνει χρήση της μεθόδου της

αυτοσυσχέτισης (autocorrelation) [Boersma1993], έναν εκτιμητή με την μέθοδο της πρόσθιας ετεροσυσχέτισης (cross correlation) [Boersma1997] και έναν εκτιμητή με βάση το φιλτραρισμένο σήμα από ένα βαθυπερατό φίλτρο μηδενικής φάσης (zero-phased low-pass filter) [Dologlou1989] ενώ το τελικό αποτέλεσμα προκύπτει από την πλειοψηφία των εκτιμήσεων. Έτσι επιτυγχάνεται υψηλότερη ευρωστία στην τελική εκτίμηση, αποφεύγοντας λάθη που οφείλονται σε ισχυρές αρμονικές του σήματος.

4.6.2.2 Εκτίμηση εμφώνων και αφώνων τμημάτων φωνής

Ο εντοπισμός των τμημάτων όπου το σήμα είναι έμφωνο ή άφωνο (όπου δηλαδή δεν υπάρχει περιοδικότητα στο σήμα) αποτελεί ένα ενδιάμεσο στάδιο στον αλγόριθμό μας, το οποίο εξυπηρετεί στην αποφυγή λανθασμένων εκτιμήσεων της θεμελιώδους συχνότητας στα συγκεκριμένα αυτά τμήματα. Ο αλγόριθμος που χρησιμοποιήσαμε για την απόφαση έμφωνου/άφωνου σήματος, λαμβάνει υπόψη την τοπική διακύμανση (variation) της θεμελιώδους συχνότητας, την ενέργεια, την ταχύτητα των zero-crossings, καθώς επίσης και τον συντελεστή αυτοσυσχέτισης του σήματος τοπικά. Η συνάρτηση απόφασης κάνει χρήση μιας συνάρτησης πιθανότητας ελαχίστου σφάλματος (minimum error probability) η οποία συνδυάζει τις 4 παραπάνω παραμέτρους. Για την εκπαίδευση της συνάρτησης απόφασης χρησιμοποιούμε τμήματα του σήματος φωνής του ομιλητή που έχουν επισημειωθεί ημιαυτόματα, με βάση την τοπική ενέργεια και τον τοπικό μέσο αριθμό περασμάτων του σήματος από το μηδέν (zero-corssings).

Το αποτέλεσμα του συγκεκριμένου σταδίου είναι ο εντοπισμός τμημάτων στο σήμα φωνής όπου το σήμα είναι έμφωνο (παρουσιάζει περιοδικότητα), είτε είναι άφωνο (χωρίς περιοδικότητα).



Σχήμα 40: Διάγραμμα ροής για την απόφαση έμφωνου/άφωνου ήχου στο σήμα φωνής.

Η τιμή των zero-crossings ανά δείγμα υπολογίζεται από το κλάσμα του πλήθους των εναλλαγών προσήμου στο σήμα, μέσα στο παράθυρο παρατήρησης, προς το μήκος του παραθύρου. Για να έχει νόημα η συγκεκριμένη μεταβλητή, θα πρέπει να έχουμε αφαιρέσει από το σήμα φωνής οποιαδήποτε DC συνιστώσα που θα επηρέαζε την ορθότητα της συγκεκριμένης προσέγγισης.

Η ενέργεια του σήματος E_s ορίζεται ως

$$E_s = 10log\left(\varepsilon + \frac{1}{N}\sum_{n=1}^{N} S^2(n)\right)$$

Όπου ε είναι μία μικρή σταθερά που μας επιτρέπει να αποφεύγουμε τον υπολογισμό του λογαρίθμου του μηδενός.

Ο κανονικοποιημένος συντελεστής αυτοσυσχέτισης C_1 ορίζεται ως:

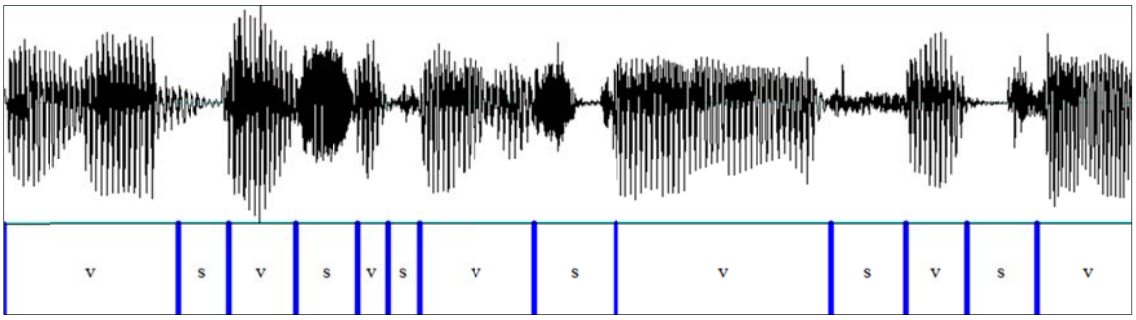
$$C_1 = \frac{\sum_{n=1}^{N} s(n)s(n-1)}{\sqrt{\sum_{n=1}^{N} s^2(n) \sum_{n=0}^{N-1} s^2(n)}}$$

Ενώ η διακύμανση της τιμής του Pitch στο παράθυρο παρατήρησης ορίζεται ως:

$$Cv = \frac{\sigma}{|\mu|}$$

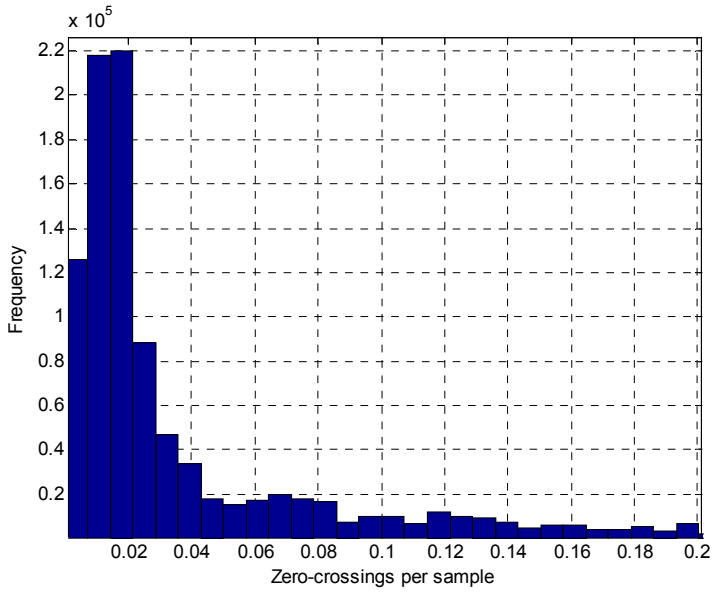
Όπου σ η τυπική απόκλιση και όπου |μ|, η μέση τιμή.

Τα αποτελέσματα του αλγορίθμου είναι ιδιαίτερα ακριβή και συνεπή, ενώ το παράθυρο ανάλυσης που χρησιμοποιούμε είναι μήκους 10 msec με επικάλυψη 50%.



Σχήμα 41: Απεικόνιση σήματος φωνής και το αποτέλεσμα από την αυτόματη επισημείωση έμφωνων και άφωνων ήχων μέσα στο σήμα φωνής.

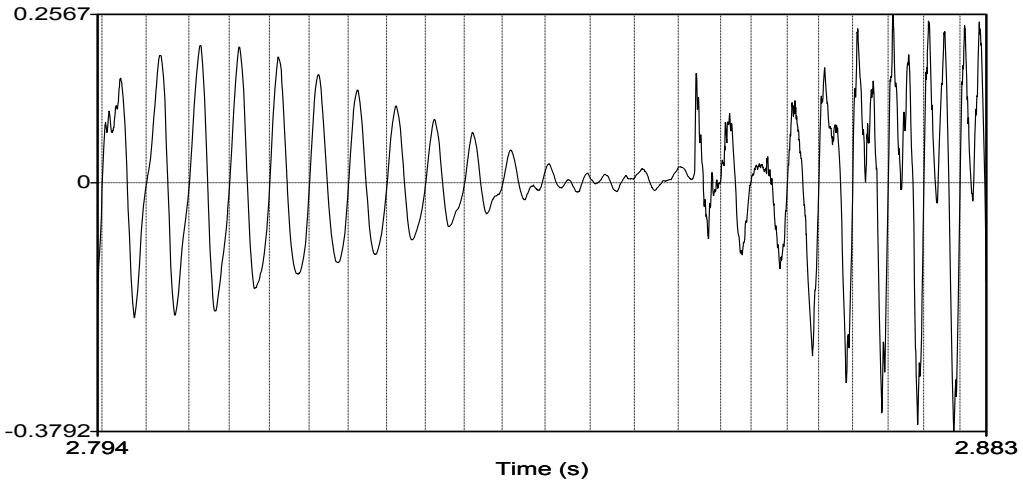
Τα δεδομένα για την εκπαίδευση του συστήματος προκύπτουν από την χειροκίνητη διόρθωση της αυτόματης επισημείωσης που έχει παραχθεί με βάση την ενέργεια του σήματος και την ταχύτητα των zero-crossings, τα κατώφλια απόφασης των οποίων έχουν υπολογισθεί αυτόματα μέσω της πυκνότητας πιθανότητας των αντίστοιχων τιμών.



Σχήμα 42: Ιστόγραμμα με βάση τα zero-crossings για ένα σήμα φωνής. Το κατώφλι για την διάκριση έμφωνων από άφωνους ήχους ορίζεται αυτόματα 0.05 όπου και παρατηρείται τοπικό ελάχιστο στην κατανομή πιθανότητας.

4.6.2.3 Εκτίμηση της θεμελιώδους συχνότητας με παρεμβολή

Ένα ενδιάμεσο στάδιο του αλγορίθμου μας είναι ο υπολογισμός σημείων της καμπύλης της θεμελιώδους συχνότητας στα άφωνα τμήματα με την μέθοδο της παρεμβολής (interpolation). Κατά αυτόν τον τρόπο, είναι δυνατός ο υπολογισμός «συνεπών» προς το περιβάλλον σημείων ανάλυσης σε τμήματα που έχουν χαρακτηριστεί άφωνα. Αν και η συγκεκριμένη προσέγγιση είναι κενή φυσικού νοήματος, αποτελεί στην ουσία έναν επιπλέον μηχανισμό αντιστάθμισης των περιπτώσεων όπου ο αλγόριθμος ανίχνευσης έμφωνων και άφωνων τμημάτων έχει αστοχήσει, χαρακτηρίζοντας έναν έμφωνο ήχο ως άφωνο.



Σχήμα 43: Παράδειγμα υπολογισμού των σημείων ανάλυσης σε άφωνα τμήματα του σήματος φωνής. Τα σημεία ανάλυσης υπολογίζονται με βάση την θεμελιώδη συχνότητα του περιβάλλοντος.

4.6.2.4 Επιλογή των σημείων ανάλυσης (pitchmarks)

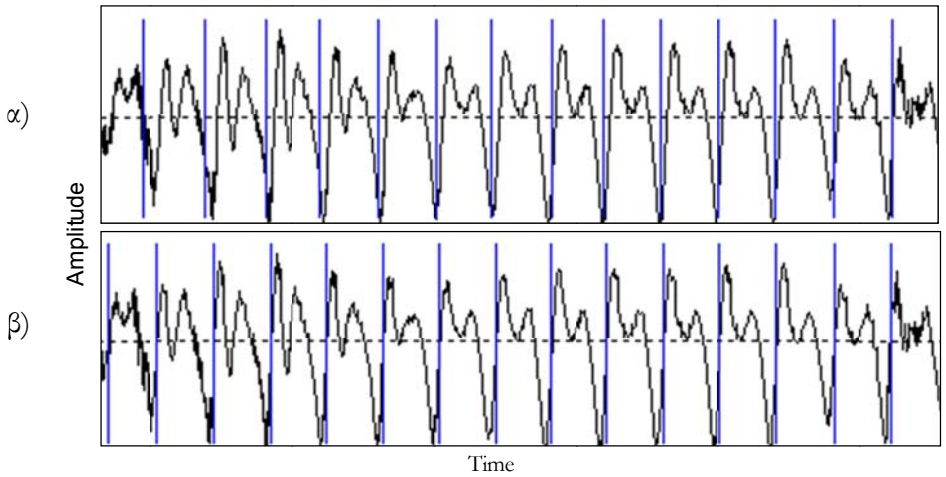
Το τελικό στάδιο της μεθοδολογίας μας [Chalamandaris et al., 2009b] αποτελεί η επιλογή των σημείων ανάλυσης στο σήμα φωνής με βάση τις παραμέτρους που έχουν υπολογιστεί σε προηγούμενα στάδια. Αρχικά το σήμα της φωνής επεξεργάζεται μέσω ενός FIR βαθυπερατού φίλτρου μηδενικής φάσης, με αποτέλεσμα την καταστολή των υψίσυχνων συνιστωσών του σήματος, οι οποίες σε διαφορετική περίπτωση θα μπορούσαν να εισάγουν σφάλμα κατά την διαδικασία επιλογής των τοπικών ελαχίστων στο σήμα της φωνής. Ο αλγόριθμος επιλογής των σημείων ανάλυσης αρχικά εντοπίζει τα ισχυρά ελάχιστα του σήματος στα τμήματα που έχουν χαρακτηριστεί ως έμφωνα, και στην συνέχεια μέσω μιας επαναληπτικής μεθόδου (iterative method) ανιχνεύει τα ισχυρά γειτονικά ελάχιστα για κάθε έμφωνο τμήμα σήματος. Τα ισχυρά ελάχιστα του σήματος εντοπίζονται μόνο σε σημεία όπου η RMS ένταση του σήματος φωνής είναι υψηλή.

$$x_{RMS} = \sqrt{\frac{\sum_{i=1}^{n} x_i}{n}}$$

Τα γειτονικά τοπικά ελάχιστα εντοπίζονται σε αποστάσεις από τα αρχικά ισχυρά σημεία ανάλυσης που είναι ανάλογες με την τοπική θεμελιώδη συχνότητα του σήματος. Αυτό επιτυγχάνεται με τον περιορισμό του παραθύρου ανίχνευσης τοπικών ελαχίστων ανάλογα με την τοπική περιοδικότητα του σήματος. Κατά αυτόν τον τρόπο τα σημεία ανάλυσης που επιλέγονται αντικατοπτρίζουν με συνέπεια την τοπική θεμελιώδη συχνότητα του σήματος.

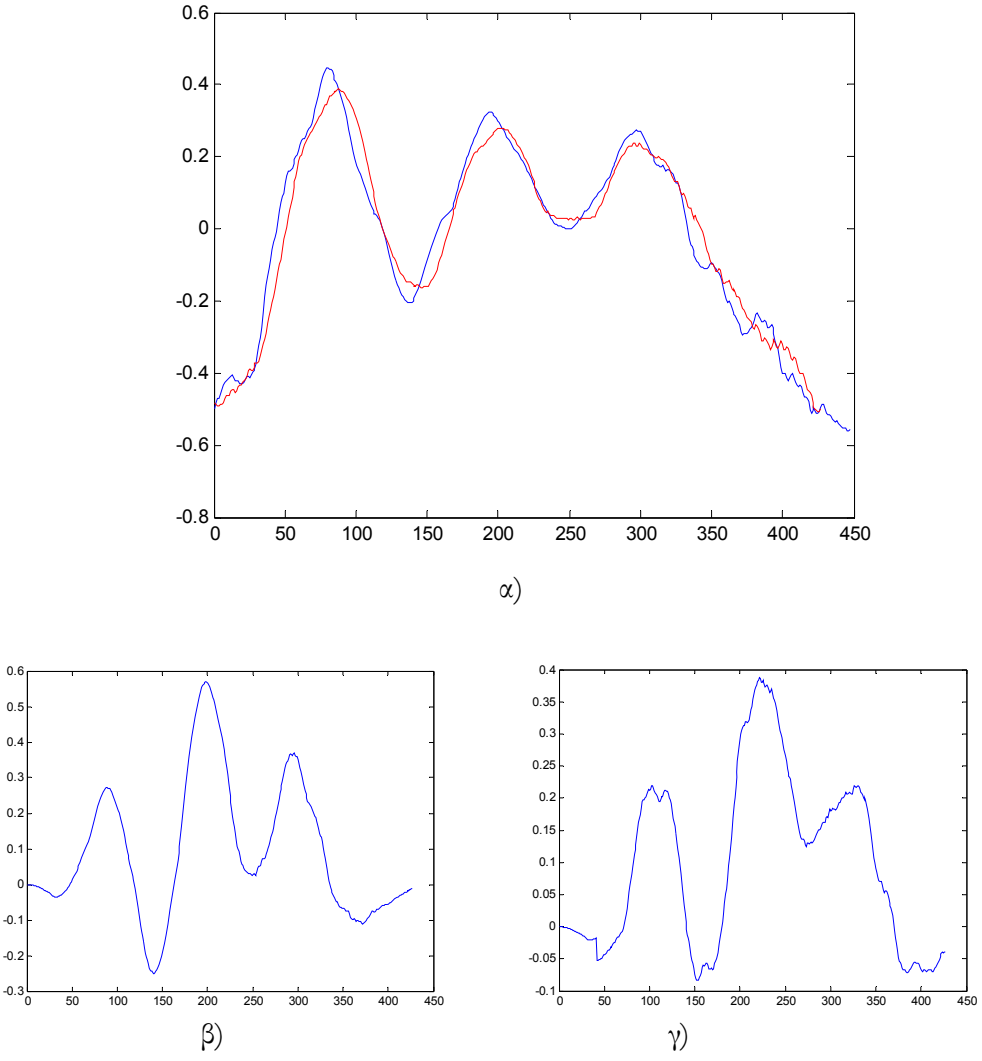
4.6.2.5 Αποτελέσματα της μεθόδου

Το σημαντικότερο χαρακτηριστικό της μεθόδου μας είναι η υψηλή ακρίβεια και συνέπεια των αποτελεσμάτων χωρίς επίβλεψη ή χειρωνακτική διόρθωση, γεγονός που αποτελεί και το πλεονέκτημα του σε σχέση με άλλους εξίσου ακριβείς αλγορίθμους όπως είναι ο DYPSA [Naylor2007]. Ιδιαίτερα σε σύγκριση με την μέθοδο DYPSA, η οποία εκτός των άλλων απαιτεί σημαντική προσπάθεια στην προσαρμογή της για μία νέα φωνή, η μέθοδός μας παρουσιάζει σημαντικά περισσότερο συνεπή αποτελέσματα, όσον αφορά στα σημεία που επιλέγει μέσα σε μία νοητή περίοδο σήματος φωνής. Το συμπέρασμα αυτό γίνεται περισσότερο ευκρινές από την παρακάτω εικόνα όπου σε ένα σήμα φωνής αντιπαραβάλλονται τα pitchmarks της μεθόδου DYPSA και της μεθόδου μας.



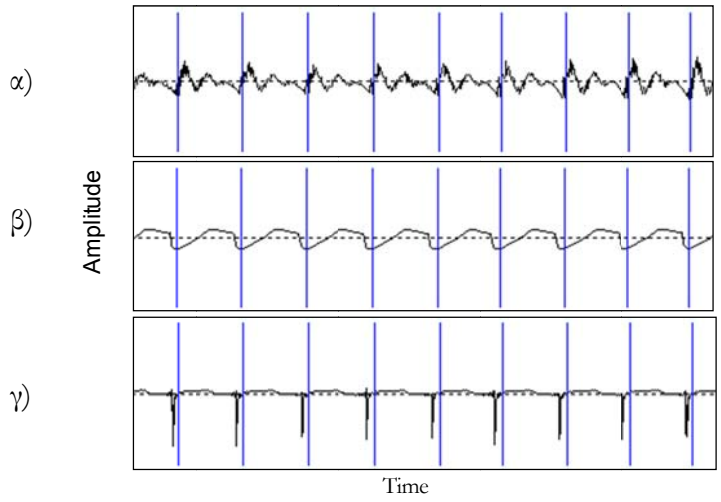
Σχήμα 44: Σύγκριση των Pitchmarks όπως αυτά υπολογίζονται από τον αλγόριθμο DYPSA (α) και τον αλγόριθμό μας (β).

Η συνέπεια των διαδοχικών σημείων pitchmarks μέσα στο σήμα φωνής, αλλά πολύ περισσότερο η συνέπεια μεταξύ μη διαδοχικών όμοιων φωνημάτων διαδραματίζει έναν από τους σημαντικότερους ρόλους στην χροιά αλλά και στην τελική ποιότητα της συνθετικής φωνής. Η αντίθετη περίπτωση, η επισημείωση pitchmarks με διαφορετική καθυστέρηση μέσα στην ίδια περίοδο, επιφέρει αλλοίωση στο συνθετικό σήμα κατά την πρόσθεση με επικάλυψη παραθύρων.



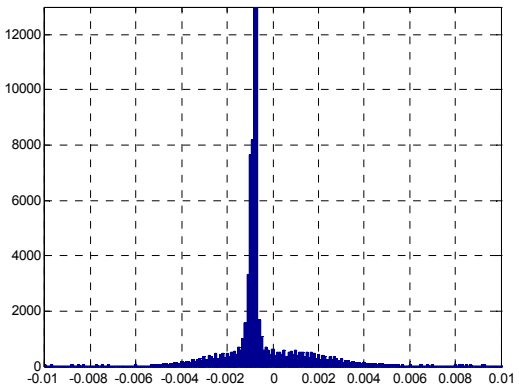
Σχήμα 45: Πρόσθεση δύο παραθύρων ST με διαφορά φάσης στα σημεία ανάλυσης PitchMarks. Στο γράφημα β) τα συνεπή PitchMarks οδηγούν σε πρόσθεση χωρίς παραμόρφωση, ενώ στην περίπτωση γ) εμφανίζεται παραμόρφωση λόγω διαφοράς φάσης στα σημεία ανάλυσης κατά 1msec. Τα ST στα γραφήματα β) και γ) έχουν πολλαπλασιασθεί με παράθυρο Hanning.

Όσον αφορά την σύγκριση των αποτελεσμάτων μας με τα σημεία CGI όπως αυτά υπολογίζονται από το αντίστοιχο σήμα του λαρυγγογράφου, παρατηρούμε ότι στην πλειοψηφία τους, τα σημεία που εμείς υπολογίζουμε, είτε προηγούνται, είτε καθυστερούν με σταθερό βήμα, καθ' όλο το εύρος των ηχογραφήσεων του ιδίου ομιλητή. Η προήγηση ή η καθυστέρηση αυτή εξαρτάται από τον ίδιο τον ομιλητή και τα χαρακτηριστικά της φωνής του. Το σημαντικότερο ωστόσο συμπέρασμα είναι ότι αυτή η διαφορά ως προς τα CGI σημεία του σήματος του λαρυγγογράφου δεν επηρεάζει αρνητικά την τελική ποιότητα του συνθέτη φωνής, όπως άλλωστε προκύπτει και από πειραματικά αποτελέσματα που εκτελέσαμε κατά την συγκεκριμένη διατριβή.



Σχήμα 46: Τα Pitchmarks ενός σήματος φωνής στο σήμα (α), στο λαρυγγογράφημα (β), και στην πρώτη παράγωγο του δεύτερου (γ).

Όπως φαίνεται άλλωστε και από την κατανομή των τιμών της διαφοράς των σημείων ανάλυσης από τα αντίστοιχα GCI σημεία του σήματος του λαρυγγογράφου, η απόκλιση των σημείων αυτών είναι της τάξης του ενός msec και με σταθερή σχεδόν τιμή για το σύνολο των ηχογραφήσεων.



Σχήμα 47: Η κατανομή της διαφοράς των υπολογισμένων σημείων ανάλυσης από τα αντίστοιχα σημεία GCI όπως προέκυψαν από το σήμα του λαρυγγογράφου. Άρεν ομιλητής με κωδικό KED. Πηγή CMU KED Database. (άξονας Υ σε χιλιάδες δείγματα, άξονας Χ σε δευτερόλεπτα της ώρας).

Πειραματική μελέτη με την χρήση ηχογραφήσεων από 6 διαφορετικούς ομιλητές έδειξε ότι το μήκος της χρονικής αυτής διαφοράς από τα σημεία GCI εξαρτάται άμεσα από τον ίδιο τον ομιλητή, παρουσιάζοντας μικρή μεν διαφορετική δε τιμή για τον καθένα από αυτούς.

4.6.2.6 Πειραματική αξιολόγηση μεθόδου για σύνθεση φωνής

Όπως αναφέραμε και προηγουμένως, η συγκεκριμένη μέθοδος αναπτύχθηκε με γνώμονα την χρήση της στο πλαίσιο της σύνθεσης φωνής με την μέθοδο TD-PSOLA. Για να αξιολογήσουμε την ποιότητα των αποτελεσμάτων της συγκεκριμένης μεθόδου, κάναμε χρήση ενός πειράματος αξιολόγησης με την μέθοδο MOS (Mean Opinion Score), όπου 6 διαφορετικοί ακροατές βαθμολόγησαν σύμφωνα με την προτίμησή τους δείγματα συνθετικής φωνής που έκαναν χρήση σημείων ανάλυσης, όπως αυτά είχαν προκύψει από τρεις διαφορετικές μεθόδους. Οι μέθοδοι που εξετάστηκαν είναι αυτή που περιγράψαμε, η μέθοδος DYPSA και η μέθοδος που προτείνεται στο λογισμικό ανάλυσης σήματος Praat, η οποία βασίζεται στον συντελεστή της αυτοσυσχέτισης του σήματος τοπικά. Τα δείγματα που εξετάστηκαν ήταν συνολικά 20 σε τρεις διαφορετικές εκδοχές, χρησιμοποιώντας τις τρεις διαφορετικές μεθόδους ανάλυσης που αναφέραμε. Οι ακροατές που αξιολόγησαν τα δείγματα ήταν όλοι τους ειδικοί σε θέματα σύνθεσης φωνής και τους ζητήθηκε να βαθμολογήσουν τα δείγματα που άκουσαν με βάση την ποιότητα του συνθετικού σήματος φωνής.

Η κλίμακα βαθμολόγησης που χρησιμοποιήθηκε ήταν από το 1-5, με 1 το χειρότερο και 5 το καλύτερο. Τα συνολικά αποτελέσματα φαίνονται στον πίνακα 5, όπου και η υπεροχή του αλγορίθμου μας είναι προφανής.

	MOS (1-5 Grade)	St. Deviation
Our approach	4.36	0.44
DYPSA	4.12	0.60
PointProcess (cc) Praat	3.23	0.58

Πίνακας 5: Τα αποτελέσματα της αξιολόγησης με την μέθοδο MOS, όπου οι ερωτηθέντες βαθμολόγησαν από το 1 έως το 5 τα δείγματα συνθετικού λόγου, με διαφορετικές μεθόδους υπολογισμού των σημείων ανάλυσης (pitch-marks).

Παράλληλα με το παραπάνω πείραμα, εκτελέσαμε ένα επιπλέον πείραμα αξιολόγησης όπου οι ερωτηθέντες άκουγαν 3 δείγματα, το καθένα κάνοντας χρήση διαφορετικού αλγορίθμου υπολογισμού σημείων ανάλυσης, και προσπαθούσαν στην συνέχεια να τα ιεραρχήσουν ανάλογα με την προτίμησή τους. Τα δείγματα αναπαράγονταν με τυχαία σειρά και επομένως οι ερωτηθέντες δεν γνώριζαν ποια μέθοδος αντιστοιχούσε σε ποιο δείγμα. Τα αποτελέσματα από αυτήν την αξιολόγηση φαίνονται στον πίνακα 6, όπου και πάλι η μέθοδος μας δείχνει να υπερτερεί των υπολοίπων.

	Preference %
Our approach	53.8
DYPSA	38.5
PointProcess (cc) Praat	7.7

Πίνακας 6: Τα αποτελέσματα της προτίμησης των δειγμάτων με διαφορετικούς αλγορίθμους υπολογισμού των σημείων ανάλυσης (pitch-marks).

Από τα αποτελέσματα μπορεί κανείς να συμπεράνει ότι ο συγκεκριμένος αλγόριθμος προσφέρει πολύ καλά αποτελέσματα όσον αφορά το πλαίσιο σύνθεσης φωνής με παράθεση ακουστικών μονάδων και την μέθοδο PSOLA.

4.7 Βιβλιογραφία Κεφαλαίου

[Beutnagel199] Beutnagel, M., Conkie, A. (1999), Interaction of Units in a Unit-Selection Database, Proceedings of the EUROSPEECH'99 Conference, pp. 1063-1066.

[Black1997] Black A. and P. Taylor, Festival speech synthesis system: system documentation (1.1.1), Tech. Rep. HCRC/TR-83, Human Communication Research Centre, 1997.

[Black1998] Black A., P. Taylor, R. Caley, 1998. The Festival Speech Synthesis System. http://festvox.org/festival

[Black2003] Black A.W., K.A. Lenzo,: Building Synthetic Voices, Language Technologies Institute, Carnegie Mellon University and Cepstral LLC. Retrieved from: http://festvox.org/bsv/ (2003)

[Black2000] Black, A., and Lenzo, K. Building voices in the Festival speech synthesis system. http://festvox.org, 2000.

[Boersma1993] Boersma P., (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 17: 97-110.

[Boersma1997] Boersma P., (1997). Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat/.

[Bozkurt2003] Bozkurt B., O. Ozturk and T. Dutoit, Text design for TTS speech corpus building using a modified greedy selection," in Proceedings of the Eurospeech'03, Geneva, Switzerland, 2003, pp. 277-180

[Campbell1997] Campbell N. and A. Black. Prosody and the selection of source units for concatenative synthesis. In: J. van Santen, R. Sproat, J. Olive and J. Hirschberg (eds.), Progress in Speech Synthesis, 1997, p279-292. New York: Springer Verlag.

[Charpentier1989]Charpentier F. and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," in Proc. EUROSPEECH, vol. 2, 1989, pp. 13–19.

[Chalamandaris et al., 2009a] A. Chalamandaris, P. Tsiakoulis, S. Raptis, and Sotiris Karabetsos, "Design of an Efficient Corpus for High-Quality Unit Selection TTS for Bulgarian", in Proc. 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poland, 2009

[Chalamandaris et al., 2009b] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and Spyros Raptis, "An Efficient and Robust Pitch Marking Algorithm on the Speech Waveform for TD-PSOLA", in Proc. Intl. IEEE Conference on Signal and Image Processing Applications (ICSIPA), Malaysia, 2009

[Chalamandaris et al., 2011] Chalamandaris, P. Tsiakoulis, S. Raptis, S. Karabetsos, "Corpus Design for a Unit Selection TtS System with Application to Bulgarian" in Human Language Technology. Challenges for Computer Science and Linguistics, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011

[Dologlou1989] Dologlou, I. G. Carayannis: "Pitch Detection based on zero-phase filtering", Speech Communication, Vol. 8, No 4, December 1989, pp. 309-318.

[Founda et al., 2001a] M. Founda, A. Chalamandaris, G. Tambouratzis, and G. Carayannis, "Reducing Spectral Mismatches in Concatenative Speech Synthesis via Systematic Database Enrichment", in Proceedings of the Eurospeech-2001 Conference, Aalborg, Denmark, 4-7 September 2001, pp. 837-840.

[Founda et al., 2001b] M. Founda, A. Chalamandaris, G. Tambouratzis, and G. Carayannis, "Studying the Factors Affecting the Optimal Unit Selection Algorithm for a TTS System for the Greek Language", in Proceedings of the 4th European Conference on Noise Control EURONOISE2001, Patra, 14-17 January 2001, Vol. II, pp. 758-764.

[Fotinea2005] Fotinea S., Tambouratzis G., A Methodology for Creating a Segment Inventory for Greek Time Domain Speech Synthesis International Journal of Speech Technology, Vol. 8, No. 2. (June 2005), pp. 161-172.

[Fotinea2001] Fotinea, S.-E., Tambouratzis, G., & Carayannis, G. 2001. Constructing a Segment Database for Greek Time-Domain Speech Synthesis. Proceedings of the Eurospeech-2001 Conference, Aalborg, Denmark, 3-7 September, Vol. 3, pp. 2075-2078.

[Founda2001] Founda, M., Chalamandaris, A., Tambouratzis, G. & Carayannis, G. 2001. Studying the Factors Affecting the Optimal Unit Selection Algorithm for a TTS System for the Greek Language. Proceedings of the 4th European Conference on Noise Control, Patra, 14-17 January, Vol. II, pp. 758-764.

[Founda2001] Founda, M., Tambouratzis, G., Chalamandaris, A.& Carayannis, G. 2001. Reducing Spectral Mismatches in Concatenative Speech Synthesis via Systematic Database Enrichment. Proceedings of the Eurospeech-2001 Conference, Aalborg, Denmark, 3-7 September, Vol. 2, pp. 837-840.

[Francois2001] Francois H. and O. Bo¨effard. Design of an Optimal Continuous Speech Database for Text-to-Speech Synthesis Considered as a Set Covering Problem. In Proceedings of Eurospeech, pages 829–832, Aalborg, Denmark, 2001.

[Franois2002] Franois H. and O. Boffard, The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database,î in 3rd International Conference on Language Resources and Evaluation (LREC 2002), 2002, vol. 5, pp. 1420-1426.

[Gauvain1990] Gauvain, J.F., Lamel, L.F., and Eskenazi, M., Design Considerations and Text Selection for BREF, a Large French Read Speech Corpus , Proc. of ICSLP., Kobe, Japan, 1990.

[Huckvale2000] Huckvale M., Speech Filing System: Tools for Speech Research, University College London, 2000, [Online] http://www.phon.ucl.ac.uk/resource/sfs/.

[Hunt1996] Hunt A., and A. Black, (1996). Unit selection in a concatenative speech synthesis system using a large speech database Proceedings of ICASSP 96, vol 1, pp 373-376, Atlanta, Georgia.

[Chen2001] Chen J.-H. and Y.-A. Kao, "Pitch marking based on an adaptable filter and a peak-valley estimation method," Computational Linguistics and Chinese Language Processing, vol. 6, no. 5, pp. 1–12, 2001.

[Santen1997] Santen Jan P. H. van and Adam L. Buchsbaum, Methods for optimal text selection, in Proc. Eurospeech '97, Rhodes, Greece, 1997, pp. 553-556.

[Klabbers2001] Klabbers E., K. St¨ober, R. Veldhuis, P. Wagner, and S. Breuer. Speech Synthesis Development Made Easy: The Bonn Open Synthesis System. In Proceedings of Eurospeech, volume 1, pages 521–524, Aalborg, Denmark, 2001.

[Karabetsos et al., 2009] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, S. Raptis, "Embedded Unit Selection Text-to-Speech Synthesis for Mobile Devices", IEEE Transactions on Consumer Electronics, Issue 2, Vol. 56, May 2009.

[Tsiakoulis et al., 2008] P. Tsiakoulis, A. Chalamandaris, S. Karabetsos and S. Raptis, "A Statistical Method for Database Reduction for Embedded Unit Selection Speech Synthesis," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Las Vegas, USA, 2008

[Kominek2000] Kominek J., Black A. The CMU Arctic Speech Databases, 5th ISCA Speech Synthesis Workshop – Pittsburgh

[Lambert2004] Lambert T. and A. Breen. A Database Design for a TTS Synthesis System Using Lexical Diphones. In 8th International Conference on Spoken Language Processing (ICSLP), pages 1381–1384, Korea, 2004.

[Lambert2006] Lambert T., Automatic Construction of a Prosodically Rich Text Corpus for Speech Synthesis Systems, International Speech Prosody Conference 2006 Dresden, Germany.

[Lemmetty1999] Lemmetty S., Review of Speech Synthesis Technology, Master's Thesis, Helsinki University of Technology, 1999

[Lenzo2000] Lenzo, K., and Black, A. Diphone collection and synthesis. In ICSLP200 (Beijing, China., 2000).

[Möbius2000] Möbius, Bernd (2000) Corpus-based speech synthesis: methods and challenges in Walter F. Sendlmeier, editor, Speech and Signals-Aspects of Speech Synthesis and Automatic Speech Recognition pp. 79-96 Hector, Frankfurt am Main dedicated to Wolfgang Hess on his 60th birthday.

[Nagy2006] Nagy A., P. Pesti, G. Németh, T. B hm, Design Issues of a Corpus-Based Speech Synthesizer, Hungarian Journal on Communications, 2005/6. special issue, pp. 18-24, Budapest, Hungary, 2005

[Naylor2007] Naylor P. A., A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," IEEE Trans. Audio, Speech and Language Processing, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[O'Brien2001] O'Brien D.and A.I.C.Monaghan, "Concatenative Synthesis Based On A Harmonic Model" in IEEE Transactions On Acoustics, Speech, And Signal Processing, Vol 9, No. 1, January 2001, pp. 11 - 20.

[Rabiner1993] Rabiner L. and B. H. Juang, Fundamentals of Speech Recognition. Prentice Hall, 1993.

[Rong2003] Rong Jon-Wei Yi.Corpus-Based Unit Selection for Natural-Sounding. Speech Synthesis. PhD Thesis, Massachusetts Institute of Technology, May 2003.

[Strube1974] Strube H.W., "Determination of the Instant of Glottal Closures from the Speech Wave," J. Acoust. Soc. Am., vol. 56, pp. 1625-1629, 1974.

[Tambouratzis2001]Tambouratzis, G., Fotinea, S.-E. & Carayannis, G. 2001. On the Systematic Construction of High-Quality Segment Databases for Greek TTS Systems. Proceedings of the 4th European Conference on Noise Control, Patra, 14-17 January, Vol. II, pp. 608-614.

[Taylor1998] Taylor P., A. Black, and R. Caley. The architecture of the festival speech synthesis system. In Proc. of the 3rd ESCA Workshop on Speech Synthesis, pages 305–310, 1998

[Villasenor2003] Villasenor -Pineda L., M. Montes y Gómez, M. A. Pérez-Coutino, and D. Vaufreydaz. 2003. A Corpus Balancing Method for Language Model Construction. In Computational Linguistics and Intelligent Text Processing, 4th International Conference, CICLing, pages 393–401, Mexico City, Mexico.

[Zhu2002] Zhu, W., Zhang, W., Shi,Q., Chen, F., Li, H., Ma, X. and Shen, L.,Corpus Building for Data-Driven TTS System, Proc. of the IEEE TTS 2002 Workshop, Santa Monica, USA 2002.

5. ΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΟΥ

Η επεξεργασία κειμένου αποτελεί μία από τις σημαντικότερες διεργασίες ενός συνθέτη φωνής και ευθύνεται για την εξαγωγή των απαραίτητων παραμέτρων από το κείμενο εισόδου για την σωστή τροφοδότηση του συστήματος παραγωγής συνθετικής ομιλίας. Η διεργασία αυτή περιλαμβάνει ένα σημαντικό αριθμό επεξεργασιών του κειμένου εισόδου και έχει την ευθύνη να αντιμετωπίσει ένα πλήθος φαινομένων που συναντώνται στον γραπτό λόγο. Στην συγκεκριμένη διατριβή, εστιάσαμε κυρίως σε δύο βασικές πτυχές της επεξεργασίας του κειμένου, την φωνητική μεταγραφή και την αυτόματη μετατροπή κειμένων Greeklish σε ορθά Ελληνικά. Η πρώτη αναφέρεται στην αυτόματη μεταγραφή οποιουδήποτε κειμένου στην αντίστοιχη φωνητική αναπαράσταση, ενώ η δεύτερη αναφέρεται σε κείμενα που ενώ είναι γραμμένα στην Ελληνική γλώσσα, είναι μετεγραμμένα με λατινικούς χαρακτήρες αντί για Ελληνικούς. Το φαινόμενο και η αντιμετώπιση των Greeklish αναλύονται διεξοδικά στο κεφάλαιο 5.

5.1 Κανονικοποίηση κειμένου

Η κανονικοποίηση του κειμένου αποτελεί μία από τις σημαντικότερες προκλήσεις που έχει να αντιμετωπίσει ένα συνθέτης φωνής κατά την λειτουργία του. Στόχος της διεργασίας αυτής είναι η αναγνώριση και η αποδοτική ανάπτυξη και ειδικών συμβολοσειρών όπως είναι τα ακρωνύμια, οι ημερομηνίες, οι αριθμοί, οι συντομογραφίες κ.α. Ιδιαίτερη σημασία έχει η συγκεκριμένη διαδικασία σε κλιτές γλώσσες όπως είναι τα Ελληνικά, όπου π.χ. η συντομογραφία «του κ. Πρωθυπουργού» θα πρέπει να αναπτυχθεί σε «του κυρίου Πρωθυπουργού» συμφωνώντας παράλληλα στην κλίση του ονόματος. Η Ελληνική γλώσσα, όντας μία γλώσσα με πλούσια κλίση και διαφορετικές πτώσεις [Petrounias1993], αποτελεί μία σχετικά δύσκολη περίπτωση όσον αφορά την κανονικοποίηση και ανάπτυξη συντομογραφιών, αριθμητικών κ.α. Κύριος μηχανισμός για την αντιμετώπιση της συγκεκριμένης πρόκλησης αποτελεί μία εσωτερική «γραμματική» στο σύστημα επεξεργασίας κειμένου, όπου ουσιαστικά τα αντικείμενα που πρόκειται να αναπτυχθούν επηρεάζονται γραμματικά από τις περικείμενες λέξεις, όσον αφορά στην κλίση, γένος και αριθμό. Αν και ο συγκεκριμένος μηχανισμός παρουσιάζει αρκετή πολυπλοκότητα, βασίζεται κυρίως σε ευρεστικούς κανόνες, η παρούσιαση των οποίων δεν αποτελεί αντικείμενο της συγκεκριμένης διατριβής.

5.2 Σύστημα Φωνητικής Μεταγραφής

Ένα βασικό υποσύστημα σε κάθε συνθέτη φωνής είναι αυτό που είναι υπεύθυνο για την αυτόματη μετατροπή του γραπτού κείμενου σε φωνητικό αλφάβητο. Με λίγα λόγια αυτό που περιγράφει πως μια ακολουθία γραμμάτων θα έπρεπε να προφερθεί από οποιονδήποτε Έλληνα ομιλητή. Η αναπαράσταση μιας φωνητικής λέξης γίνεται με την χρήση του διεθνούς φωνητικού αλφαβήτου, όπως αυτό έχει ορισθεί το 1996 από το IPA[6]. Ως φωνήματα μιας γλώσσας ορίζονται το σύνολο των στοιχειωδών ήχων (φθόγγων) που επαρκούν για την εκφορά οποιασδήποτε λέξης της γλώσσας αυτής. Στις παραγράφους που ακολουθούν παρουσιάζουμε την βασική θεωρία για την φωνητική αναπαράσταση της Ελληνικής γλώσσας μιας και το σύστημα στο οποίο αναφερόμαστε έχει ως πρωτεύουσα γλώσσα υποστήριξης την Ελληνική.

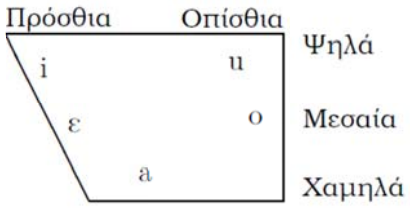
[6] International Phonetic Alphabet

5.2.1 Φωνητική της Ελληνικής γλώσσας

Το πλήθος των φωνημάτων διαφέρει όπως είναι φυσικό από γλώσσα σε γλώσσα και ενώ π.χ. στην Αγγλική γλώσσα συναντιόνται 44 φωνήματα, στην ελληνική συναντιόνται 29, εκ των οποίων τα 5 είναι φωνήεντα και τα 24 σύμφωνα [Petrounias1993]. Στους πίνακες που ακολουθούν φαίνονται όλα τα φωνήματα του Ελληνικού φωνητικού αλφαβήτου, τον τρόπο και τον τόπο όπου αυτά σχηματίζονται στην φωνητική κοιλότητα [Πρωτόπαπας2003].

	Φων.	Χειλικά	Χειλοδοντικά	Οδοντικά	Φατνιακά	Ουρανικά	Υπερωικά
Κλειστά	Φ	b		d		ɟ	g
	Α	p		t		c	k
Τριβόμενα	Φ		v	ð	z	j	ɣ
	Α		f	θ	s	ç	x
Προστριβόμενα	Φ				(ʤ)		
	Α				(ʦ)		
Ένρινα	Φ	m	(ɱ)	n		ɲ	ŋ
Πλάγια	Φ			l		ʎ	
Παλλόμενα	Φ				r ɾ		

Πίνακας 7: Οι συμφωνικοί φθόγγοι της Ελληνικής γλώσσας Έμφωνοι (Φ) και άφωνοι (Α), ανά τόπο και τρόπο άρθρωσης σύμφωνα με το διεθνές φωνητικό αλφάβητο.



Πίνακας 8: Η θέση των φωνηέντων της Ελληνικής γλώσσας στον ακουστικό χώρο σύμφωνα με το διεθνές φωνητικό αλφάβητο

5.2.1.1 Τα φωνήεντα

Στη φωνητική, φωνήεντα ονομάζονται οι φθόγγοι οι οποίοι κατά τον σχηματισμό τους στη φωνητική οδό στερούνται οποιουδήποτε φραγμού ή επαρκούς στένωσης, ώστε να παραχθεί

ακουστή τριβή. Με άλλα λόγια, ο εξερχόμενος αέρας για τον σχηματισμό ενός φωνήεντος περνά πάνω από το κέντρο της γλώσσας κατά ομοιόμορφο τρόπο, χωρίς, πρακτικά, να βρίσκει αντίσταση πουθενά. Σε όλες τις γλώσσες τα φωνήεντα αποτελούν τον πυρήνα της συλλαβής, καθώς μπορούν μόνα τους ή σε συνδυασμό με σύμφωνο ή σύμφωνα να σχηματίσουν μια συλλαβή, σε αντίθεση με τα σύμφωνα. Ανάλογα με τον τρόπο σχηματισμού τους, τα φωνήεντα ταξινομούνται ως προς τη θέση της γλώσσας κατά τον κάθετο και τον οριζόντιο άξονα, ως προς το σχήμα το οποίο παίρνουν τα χείλη, ενώ ακόμη εξετάζεται ο τόπος σχηματισμού (στοματική ή ρινική κοιλότητα, λάρυγγας), η διάρκεια εκφοράς, η μεταβολή ποιότητας κατά την άρθρωση και η ποσότητα της μυϊκής τάσης η οποία χρειάζεται για την παραγωγή τους. Τα φωνήεντα της ελληνικής είναι συνολικά εφτά και χωρίζονται σε τρεις κατηγορίες:

α) τα βραχύχρονα (ε, ο),

β) τα μακρόχρονα (η, ω) και

γ) τα δίχρονα (α, ι, υ).

Ορισμένα φωνήεντα τείνουν να είναι άηχα στην περίπτωση που συναντώνται σε άτονη τελική συλλαβή (λήγουσα) ανάμεσα σε δύο άφωνα σύμφωνα, όπως π.χ. στην λέξη *φάσης, /fásis/* όπου το φώνημα /i/ προφέρεται άφωνο.

5.2.1.2 Τα σύμφωνα

Το σύμφωνο είναι φθόγγος χαμηλής τονικότητας κατ' αντιδιαστολή προς τα φωνήεντα, που παράγεται όταν, κατά την ομιλία, ο εκπνεόμενος αέρας προσκρούει σε εμπόδιο και κατόπιν διέρχεται από τέλειο φραγμό (οπότε παράγονται τα «κλειστά» σύμφωνα) ή στενό της στοματικής κοιλότητας (παράγοντας τα «διαρκή»-«τριβόμενα» σύμφωνα), με αποτέλεσμα να δημιουργείται ένας αισθητός ήχος. Στα σύμφωνα υπάρχουν τρεις κατηγορίες ανάλογα με το κριτήριο που χρησιμοποιείται:

Κατά την φύση του ήχου που έχουν διαιρούνται σε: α) άηχα: θ, κ, π, τ, σ, τσ, φ, χ και β) ηχηρά: γ, γκ, β, δ, ζ, λ, μ, μπ, ν, ντ, ρ, τζ.

Κατά τη διάρκεια διαιρούνται σε: α) στιγμιαία: κ, π, τ, γκ, μπ, ντ, τσ, τζ και β) εξακολουθητικά: γ, β, δ, χ, φ, θ, σ, ζ, λ, μ, ν, ρ.

Κατά το μέρος όπου σχηματίζονται στο στόμα διαιρούνται σε: α) χειλικά: π, β, φ, μπ, β) οδοντικά: τ, δ, θ, ντ, γ) συριστικά ή διπλοδοντικά: σ, ζ, τσ, τζ, δ) λαρυγγικά: κ, γ, χ, γκ, ε) υγρά ή γλωσσικά: λ, ρ και στ) ρινικά: μ, ν.

Κατά την εκφώνηση των φωνηέντων παρατηρούνται αλλοφωνικές ποικιλίες ορισμένων φωνημάτων που εμφανίζονται σε περιπτώσεις που εξαρτώνται από το περικείμενο. Χαρακτηριστικό παράδειγμα του συγκεκριμένου φαινομένου αποτελούν οι περιπτώσεις CiV (Consonant /i/ Vowel) όπου το φώνημα /i/ τροποποιείται ανάλογα με το περικείμενο στο οποίο βρίσκεται το σύμπλεγμα CiV, με εξαιρέσεις σε κανόνες που συχνά οδηγούν και σε ομόγραφες λέξεις με διαφορετική προφορά και νόημα αντίστοιχα.

Όπως αναφέρει ο κ. Πετρούνιας [Petrounias1993]:

> *«Το γενικό συμπέρασμα είναι ότι η αντιπροσώπευση της προφοράς από τη γραφή τις περισσότερες φορές είναι απρόβλεπτη. Ο συμβολισμός της προφοράς από της ορθογραφία γίνεται συνήθως με τρόπο έμμεσο και με βάση περίπλοκες προϋποθέσεις.»*

Σε κάθε περίπτωση όμως, σε γενικές γραμμές η Ελληνική γλώσσα παρουσιάζει μια γενικότερη αμεσότητα όσον αφορά την αντιστοίχηση ορθογραφίας και προφοράς, με ένα επιπλέον χαρακτηριστικό το οποίο διευκολύνει σημαντικά την όλη διαδικασία, την ύπαρξη του τόνου στον γραπτό λόγο. Έτσι, ενώ για παράδειγμα η Ελληνική σε αντίθεση με την Ιταλική, δεν απαιτεί την πρόβλεψη του τόνου μέσα στην λέξη, παρά αυτός υποδεικνύεται εγγράφως. Μερικές γλώσσες όπως τα Ισπανικά, τα Φινλανδικά και τα Σουαχίλι παρουσιάζουν περισσότερο ευθείς συνδέσμους μεταξύ της γραπτής και της φωνητικής αναπαράστασης τους, ενώ άλλες, όπως τα Αγγλικά και τα Γαλλικά, παρουσιάζουν μόνο μερικούς γενικούς κανόνες μεταγραφής και πολλές εξαιρέσεις που εξαρτώνται από πολλές παραμέτρους όπως είναι το περικείμενο, το νόημα κ.λπ.

5.2.2 Γενικοί κανόνες φωνητικής μεταγραφής για την Ελληνική γλώσσα

Για τα ελληνικά, όπως προαναφέραμε, υπάρχει ένα συγκεκριμένο σετ κανόνων για αναπαραγωγή της φωνητικής αναπαράστασης από την ορθογραφική. Οι κανόνες ισχύουν σε όλο το εύρος των

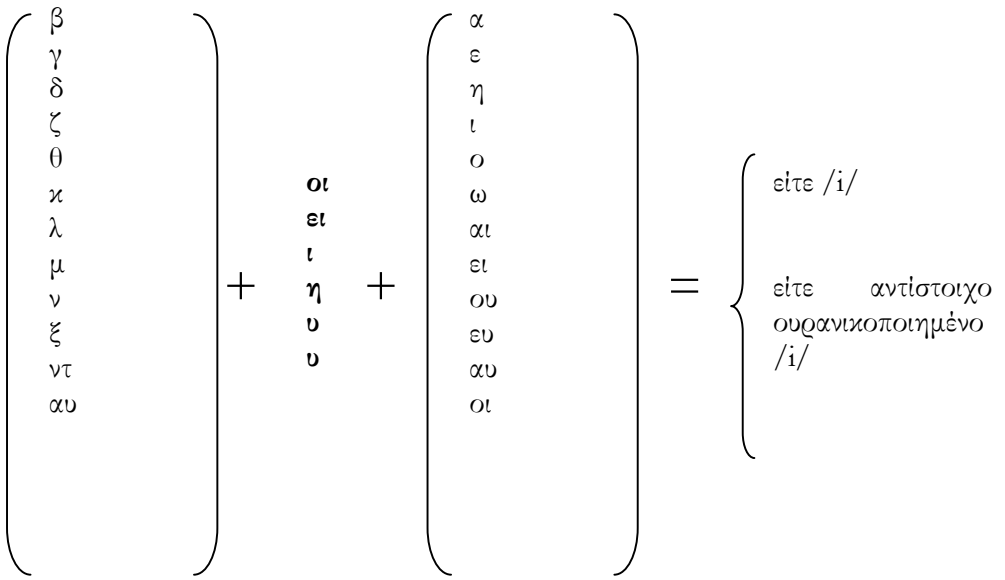
ελληνικών λέξεων, ωστόσο υπάρχει ένα αρκετά μεγάλο ποσοστό λέξεων (λόγιων) που δεν υπόκεινται στου κανόνες αυτούς και παρουσιάζουν μερικές εξαιρέσεις.

5.2.3 Σημαντικά θέματα και προβλήματα στην φωνητική μεταγραφή για τα ελληνικά

Για την ελληνική γλώσσα, παρ' όλο που δεν παρουσιάζει τόσες εξαιρέσεις όσο άλλες γλώσσες [Black1998], όσον αφορά την μεταγραφή από ορθογραφική σε φωνητική αναπαράσταση, υπάρχουν σημαντικά προβλήματα που ένα αυτόματο σύστημα πρέπει να αντιμετωπίσει. Οι ιδιαιτερότητες αυτές μπορούν να ομαδοποιηθούν σε τρεις βασικές κατηγορίες: στα φαινόμενα συνίζησης μεταξύ λέξεων, στα φαινόμενα προενρινοποίησης στους φθόγγους /μπ/, /ντ/ και /γκ-γγ/, και στα φαινόμενα συνίζησης όταν έχουμε το φώνημα /i/ ανάμεσα σε ένα σύμφωνο και ένα φωνήεν, δηλαδή στην μορφή /Σύμφωνο/-/άτονο i/-/Φωνήεν/. Προς χάριν συντομίας από εδώ και στο εξής τα φαινόμενα όπου εμφανίζεται ένα άτονο /i/ ανάμεσα σε δύο σύμφωνα, θα αναφέρονται ως *CiV*.[7]

Πιο συγκεκριμένα για τις περιπτώσεις *CiV* μπορεί να έχουμε οποιαδήποτε από τις περιπτώσεις που φαίνονται στο σχήμα που ακολουθεί. Τότε σε κάθε μία από τις περιπτώσεις αυτές, ανάλογα με την λέξη και το περικείμενο, το φώνημα /i/ είτε προφέρεται ως στρογγυλό /i/, είτε μετατρέπεται στο αντίστοιχο ουρανικοποιημένο φώνημα, ανάλογα με το αν το προηγούμενο σύμφωνο είναι έμφωνο ή άηχο.

[7] Consonant - /i/ - Vowel



Σχήμα 48: Συνδυασμοί γραφημάτων όπου η προφορά του /i/ παρουσιάζει δυσκολίες.

Από μετρήσεις που πραγματοποιήσαμε στο ηλεκτρονικό σώμα κειμένου του ΙΕΛ, [ΕΘΕΓ][8] υπολογίσαμε ότι τα φαινόμενα *CiV* εμφανίζονται σε ποσοστό 8,5%, παρατηρώντας παράλληλα το μέγεθος της δυσκολίας που υπεισέρχεται κατά την διαδικασία της φωνητικής μεταγραφής ενός ελεύθερου Ελληνικού κειμένου. Μπορεί εύκολα λοιπόν κανείς να συμπεράνει ότι ένα αυτόματο σύστημα φωνητικής μεταγραφής για τα Ελληνικά καλείται να αντιμετωπίσει σημαντικές δυσκολίες. Η αυτόματη φωνητική μεταγραφή γραπτών κειμένων αποτελεί σημαντικό κομμάτι της έρευνας εδώ και αρκετά χρόνια, αφού η ύπαρξη της είναι απαραίτητο συστατικό τόσο στην σύνθεση όσο και στην αναγνώριση φωνής. Υπάρχουν αρκετές διαφορετικές προσεγγίσεις και μελέτες γύρω από το συγκεκριμένο θέμα, ωστόσο όλες αυτές μπορούν ομαδοποιηθούν σε τρεις βασικές μεθοδολογίες. Αυτές παρουσιάζονται συνοπτικά στις επόμενες παραγράφους.

5.2.4 Μέθοδοι βασισμένες σε λεξικά

Οι μέθοδοι αυτές βασίζονται σε λεξικά που σκοπό έχουν να αποθηκεύσουν την μέγιστη δυνατή πληροφορία αναφορικά με την φωνητική μεταγραφή πλήθους λέξεων [Kim2004]. Συχνά αντί για

[8] Ο ΕΘΕΓ (Εθνικός Θησαυρός της Ελληνικής Γλώσσας) αποτελείται από κατάλληλα οργανωμένα σώματα κειμένου γραπτού που προέρχονται από πηγές με ποικίλη θεματολογία, όπως εφημερίδες, βιβλία, περιοδικά κ.α. Αριθμεί άνω των 34,000,000 λέξεων και αναβαθμίζεται συνεχώς.

λέξεις χρησιμοποιούνται μορφήματα, έτσι ώστε να μειωθεί αφενός η ανάγκη σε αποθηκευτικό χώρο, αλλά και να είναι δυνατή η αντιμετώπιση περισσοτέρων περιπτώσεων από απλά τις λέξεις που είναι αποθηκευμένες στο λεξικό. Η χρήση των μορφημάτων αποτελεί λίγο πολυπλοκότερη διαδικασία από την χρήση ενός λεξικού με λέξεις, αφού εμπεριέχει την γραφή μιας γραμματικής όσον αφορά τον τρόπο που τα μορφήματα τεμαχίζονται και τον τρόπο με τον οποίο συμπεριφέρονται όταν αυτά συνδυάζονται και αλληλεπιδρούν μεταξύ τους. Οι βασικότερες προσπάθειες που περιγράφουν την συγκεκριμένη προσέγγιση είναι αυτή των Coker (1985) και Allen et al. (1987) για τα Αγγλικά, ενώ για τα Γαλλικά αυτή του Laporte (1987), ο οποίος και κατασκεύασε ένα εκτενές φωνητικό λεξικό. Οι πρώτες προσπάθειες ενσωμάτωσης ενός τέτοιου λεξικού σε σύστημα σύνθεσης φωνής παρατηρούνται το 1987 από τον Allen et al. στο MITALK TTS σύστημα και από τον Levinson et al (1993) στο AT&T TTS σύστημα.

Τα φωνητικά λεξικά συχνά εκτός από την γραμματική και την φωνητική αναπαράσταση τους περιέχουν και άλλες πληροφορίες που συχνά είναι αναγκαίες για την αποσαφήνιση περιπτώσεων όπου η γραμματική αναπαράσταση δύο λέξεων είναι ίδια αλλά διαφορετικής προφοράς τους λόγω διαφορετικής λειτουργίας τους μέσα στην πρόταση. Το βασικό πλεονέκτημα τέτοιων λεξικών είναι ότι μπορούν να φανούν χρήσιμα και σε άλλες περιπτώσεις όπως είναι π.χ. η αυτόματη μετάφραση, η αναγνώριση φωνής κ.α. Η ελληνική γλώσσα είναι εξαιρετικά πλούσια γλώσσα όσον αφορά στην μορφολογία της και στις διαφορετικές μορφές που παρουσιάζουν οι λέξεις ανάλογα με την πτώση ή την κλίση τους, και για τον λόγο αυτό είναι αρκετά δύσκολο να δημιουργήσει κανείς ελληνικό φωνητικό λεξικό που να προσφέρει μεγάλη κάλυψη λέξεων, ειδικά με πρόβλεψη για όλες τις σύνθετες και παράγωγες λέξεις, οι οποίες είναι πολυάριθμες στην ελληνική γλώσσα.

5.2.5 Μέθοδοι βασισμένες σε κανόνες

Οι μέθοδοι αυτές χρησιμοποιούν ένα προκαθορισμένο σετ από κανόνες μεταγραφής από γραφήματα σε φωνήματα, ενώ παράλληλα κάνουν χρήση και ενός λεξικού λέξεων εξαιρέσεων που δεν μπορούν να αντιμετωπιστούν επιτυχώς από τους κανόνες [Pagel1998] [Caseiro2002]. Οι κανόνες αυτοί προκύπτουν από εξειδικευμένη γνώση (κυρίως από το πεδίο της γλωσσολογίας) και φυσικά διαφέρουν από γλώσσα σε γλώσσα. Από την εμφάνιση της μεθόδου αυτής έχουν πραγματοποιηθεί σημαντικές προσπάθειες για την εξαγωγή όσο το δυνατό γενικότερων κανόνων φωνητικής μεταγραφής για κάθε γλώσσα ξεχωριστά. Για τα Αγγλικά και τα Γαλλικά οι πρώτες

προσπάθειες πραγματοποιήθηκαν από τους Ainsworth (1973), McIlroy (1974), Elovitz et al. (1976), Hertz (1979), Hunnicutt (1980), Belrhali et al. (1992) και πολλούς άλλους. Για την ελληνική γλώσσα η πρώτη τεκμηριωμένη προσέγγιση πραγματοποιήθηκε από τον Πετρούνια ενώ η πρώτη προσπάθεια αυτοματοποίησης σε υπολογιστή για τις ανάγκες ενός συνθέτη φωνής συναντάμε από τον Μπακαμίδη και Καραγιάννη [Bakamides1985].

Η προσέγγιση αυτή έχει αρκετά κοινά σημεία με την μέθοδο χρήσης ενός λεξικού, μιας και γίνεται επιπλέον χρήση ενός λεξικού εξαιρέσεων. Παράλληλα, όπως αναφέραμε και στην προηγούμενη παράγραφο, η μέθοδος των λεξικών συχνά περιλαμβάνει εκτός από ολόκληρες λέξεις και μικρότερα συστατικά στοιχεία λέξεων, όπως είναι τα μορφήματα, οι ρίζες και οι καταλήξεις λέξεων που ουσιαστικά αποτελούν ένα σετ κανόνων φωνητικής μεταγραφής και όχι πραγματικών λέξεων. Η μέθοδος αυτή αποδίδει αρκετά καλά, ωστόσο το ποσοστό επιτυχίας και πολυπλοκότητας των κανόνων εξαρτώνται άμεσα από την ίδια την γλώσσα. Ενώ π.χ. για τα ελληνικά, εκτός των εξαιρέσεων, ο αριθμός των κανόνων που ισχύουν γενικά δεν ξεπερνούν τους 200, για άλλες γλώσσες όπως είναι τα αγγλικά ο αντίστοιχος αριθμός είναι σημαντικά μεγαλύτερος.

5.2.6 Μέθοδοι βασισμένες σε δεδομένα (Data-driven methods)

Στην κατηγορία αυτή ανήκουν οι τρεις σχετικά νεότερες μέθοδοι που προσπαθούν να αντιμετωπίσουν το συγκεκριμένο θέμα και αυτές είναι

α) η προφορά κατά αναλογία (pronunciation by analogy),

β) οι στατιστικές μέθοδοι βασισμένες σε στοχαστικές θεωρήσεις και στην θεωρία του κοντινότερου γείτονα και τα νευρωνικά δίκτυα.

Σε κάθε περίπτωση το σύστημα εκπαιδεύεται με βάση ένα λεξικό που έχει αντιστοιχισμένες τις ορθογραφικές αναπαραστάσεις όλων των λέξεων με τις φωνητικές, και ανάλογα εξάγει τις πληροφορίες που χρειάζεται αυτό για να λειτουργήσει.

Στην περίπτωση των νευρωνικών δικτύων συνήθως χρησιμοποιούνται πολύ-επίπεδα perceptrona τα οποία εκπαιδεύονται μέσω «back propagation» στα δεδομένα ενός προκατασκευασμένου

λεξικού. Οι στοχαστικές μέθοδοι κοντινότερου γείτονα βασίζονται στην συχνότητα εμφάνισης και τρόπου συμπεριφοράς συγκεκριμένων ακολουθιών χαρακτήρων, ανάλογα με το περιβάλλον στο οποίο αυτές βρίσκονται [Ravishankar1997] [Dermatas1999].

Ωστόσο, η μέθοδος που κερδίζει συνεχώς έδαφος είναι αυτή της προφοράς κατά αναλογία, η οποία προσπαθεί να εξάγει πρότυπα απεικόνισης που αποτελούνται από ακολουθίες χαρακτήρων και να δημιουργήσει κανόνες που ισχύουν καθολικά στο λεξικό εκπαίδευσης. Με λίγα λόγια, κατά την διάρκεια της εκπαίδευσης του συστήματος γίνεται αναζήτηση ακολουθιών ορθογραφικών χαρακτήρων που μεταγράφονται φωνητικά κατά έναν και μοναδικό τρόπο μέσα στο λεξικό. Αν κάποια ακολουθία χαρακτήρων μεταγράφεται με περισσότερους από έναν τρόπους μέσα στο ίδιο το λεξικό, τότε αναζητείται αυτόματα μεγαλύτερο μήκος ακολουθίας χαρακτήρων. Το μήκος αυξάνει μέχρι να ισχύει η μοναδικότητα του κανόνα.

Το βασικό πλεονέκτημα των μεθόδων αυτών είναι ότι είναι ανεξάρτητες από την γλώσσα και μπορούν να εφαρμοσθούν άμεσα για την μοντελοποίηση κάθε γλώσσας, εφόσον βέβαια υπάρχουν διαθέσιμοι οι απαραίτητοι πόροι (τα λεξικά εκπαίδευσης).

5.2.7 Η προσέγγισή μας

Ο αλγόριθμος που αναπτύξαμε [Chalamandaris2005] για την αυτόματη φωνητική μεταγραφή γραπτού κειμένου βασίζεται στην προφορά κατά αναλογία όπως περιγράφτηκε συνοπτικά παραπάνω, ενώ παράλληλα κάνει και επιπλέον χρήση καθολικά ισχυόντων κανόνων, που έχουν ορισθεί σε πρώτο στάδιο με χειρωνακτικό τρόπο. Στις επόμενες παραγράφους θα παρουσιάσουμε αναλυτικά την μέθοδό μας, όπως επίσης και τα σημεία που περιέχουν ιδιαιτερότητες στην διαδικασία αυτή.

Οι κανόνες αυτοί που υιοθετήσαμε είναι της μορφής:

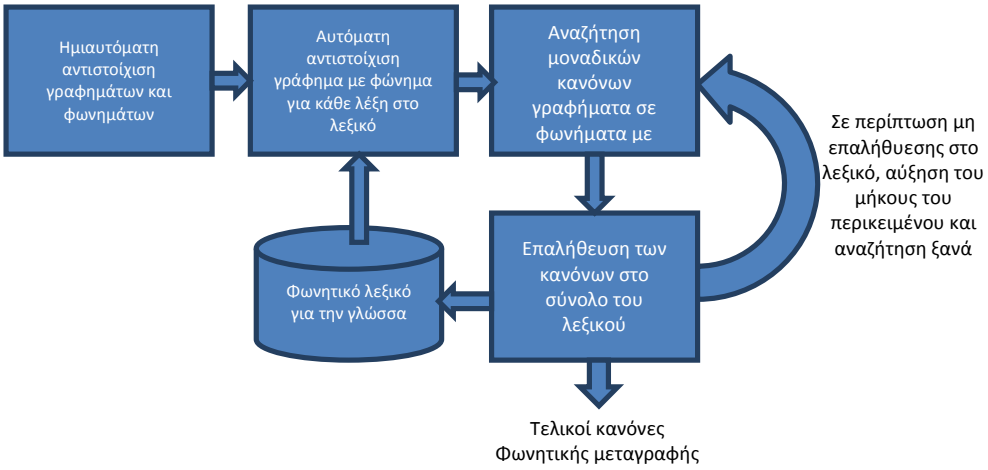
(*αριστερό περικείμενο*)-εστία-(*δεξί περικείμενο*) -> φωνητική μεταγραφή

Η μέθοδος συνοπτικά μπορεί να περιγραφεί με τα εξής τρία βήματα:

- Καθορισμός όλων των δυνατών αντιστοιχιών μεταξύ ορθογραφικών και φωνητικών συμπλεγμάτων.

- Αυτόματη αντιστοίχιση όλων των εγγραφών του λεξικού μεταξύ της ορθογραφικής και της αντίστοιχης φωνητικής αναπαράστασης για κάθε λέξη.

- Αυτόματη ανεύρεση από το σύστημα των περισσότερων δυνατών καθολικά ισχυόντων κανόνων μέσα στο λεξικό της παραπάνω μορφής, που έχουν το μικρότερο δυνατό μήκος, τόσο στο αριστερό, όσο και στο δεξί περικείμενο.



Σχήμα 49: Διάγραμμα ροής του αλγορίθμου φωνητικής μεταγραφής. Για κάθε κανόνα που υπολογίζεται, ελέγχεται η επαλήθευσή του στο σύνολο του λεξικού. Αν αποτύχει η επαλήθευση, τότε το μήκος του περικειμένου αυξάνει σταδιακά και επαληθεύεται ο νέος κανόνας.

Η εστία μπορεί να λάβει τιμές από ένα σύνολο ακολουθιών γραμμάτων όπως αυτά ορίζονται κατά τον σχεδιασμό του συστήματος και όχι αυθαίρετα από το σύστημα. Αυτό άλλωστε είναι και το πρώτο βήμα όπως αναφέρεται πιο πάνω. Η διαφορά σε σχέση με άλλες μεθόδους έγκειται στο γεγονός ότι οι περισσότερες από αυτές δεν επιχειρούν παρόμοια προσέγγιση στο σημείο αυτό, αλλά απλά προσπαθούν αυτόματα να εξάγουν πρότυπα που συνήθως δεν είναι μεγαλύτερα από ένα γράμμα ή μία δίφθογγο. Θεωρούν ότι ένα γράμμα ή μία δίφθογγος μπορεί να αντιστοιχιστεί είτε σε ένα ή περισσότερα φωνήματα, είτε σε εξαφάνιση, ή αλλιώς στο γνωστό στην βιβλιογραφία ως

φώνημα *epsilon*. Το φώνημα αυτό ουσιαστικά χρησιμοποιείται όπου γίνεται συνίζηση ή όταν κάποιο γράμμα δεν αντιστοιχεί σε κανένα φώνημα. Με άλλα λόγια το φώνημα *epsilon* συναντάται όπου κάποιο γράφημα δεν έχει αντίστοιχο φώνημα στην 1-1 αντιστοίχηση γραφημάτων και φωνημάτων και ουσιαστικά θεωρούμε ότι ένα γράφημα απλά «εξαφανίζεται» κατά την διαδικασία της φωνητικής μεταγραφής. Παρακάτω φαίνεται ένα παράδειγμα της χρήσης του φωνήματος *epsilon*.

κ	υ	ά	λ	ι	α
↓	↓	↓	↓	↓	↓
ç	*epsilon*	**á**	**ƛ**	*epsilon*	**a**

Σχήμα 50: Φωνητική μεταγραφή της λέξης *κυάλια* με χρήση του φωνήματος *epsilon*

Ωστόσο η μέθοδος αυτή έχει σημαντικά μειονεκτήματα αφού η αντιστοίχιση ενός γράμματος με το φώνημα *epsilon* συχνά οδηγεί σε περισσότερες της μίας λύσης με αποτέλεσμα η γενίκευση των κανόνων να είναι συχνά προβληματική. Αυτό μπορεί να συμβεί εύκολα όταν δύο ή περισσότερα γράμματα που μπορούν να οδηγήσουν στο φώνημα *epsilon* βρίσκονται σε διαδοχικές θέσεις σε μία λέξη όπου το συγκεκριμένο φαινόμενο μπορεί να αποδοθεί και στα δύο γράμματα εξίσου σωστά, χωρίς όμως να έχουν το ίδιο νόημα, ενώ συχνά με την χρήση του συγκεκριμένου φωνήματος δεν είναι δυνατόν να αποτυπωθεί η διαφορετική συμπεριφορά ορισμένων γραμμάτων σε διαφορετικά περικείμενα, κάνοντας την διαδικασία εντελώς μηχανική χωρίς φυσική σημασία.

Στην προσέγγιση μας δεν κάνουμε χρήση του φωνήματος *epsilon* [Ravishankar1997] αλλά ορίσαμε εξ' αρχής όλες τις δυνατές περιπτώσεις απεικόνισης γραμμάτων σε φωνήματα, καθώς επίσης ομάδων γραμμάτων σε ομάδες φωνημάτων, όπως π.χ. ότι το γράμμα /ν/ μετατρέπεται στο φώνημα /n/ και το σύμπλεγμα /νι/ μετατρέπεται σε /ni/ ή /ɲ/ Αυτό έγινε ημιαυτόματα και προέκυψαν άνω των 10.000 διαφορετικά πρότυπα, εκ των οποίων όμως πολλά δεν είναι δυνατά ή δεν συναντιόνται. Τα πρότυπα που παρατηρήθηκαν στο λεξικό εκπαίδευσης ήταν 1.161, χωρίς αυτό να σημαίνει ωστόσο ότι άλλα που δεν συναντήθηκαν δεν ισχύουν ή δεν υπάρχουν στην ελληνική γλώσσα, ενώ το σύνολο των στοιχειωδών μονάδων που χρησιμοποιήθηκαν (γραφήματα και ακολουθίες αυτών) ήταν 840.

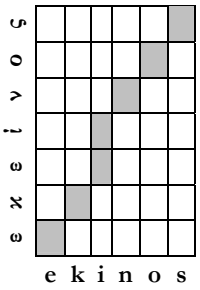
Κατόπιν της διαδικασίας αυτής, στο σύνολο των παραγόμενων κανόνων προστέθηκε επίσης και ένα σύνολο κανόνων που ισχύουν καθολικά στα ελληνικά και έχουν παρουσιαστεί στην βιβλιογραφία. Οι επιπλέον αυτοί κανόνες προστίθενται στο τελικό σύνολο γιατί έχουν καθολική ισχύ και υποδεικνύουν τον τρόπο προφοράς οποιασδήποτε λέξης με ελληνική προφορά, είτε υπάρχει ως λέξη είτε όχι. Ο λόγος που έγινε αυτό είναι κυρίως για την εξασφάλιση της ορθής προφοράς ακόμη και ανούσιων λέξεων που δεν υπάρχουν σε λεξικά. Παράδειγμα ενός τέτοιου κανόνα αποτελεί η περίπτωση του διφθόγγου *αυ* που προφέρεται /af/ όταν ακολουθείται από οποιοδήποτε άλλο άφωνο σύμφωνο ή όταν βρίσκεται στο τέλος της λέξης. Ωστόσο, σε καμία λέξη του λεξικού μας δεν εμφανίστηκε η περίπτωση *αυ-φ* ή *αυ-ζ* με αποτέλεσμα να μην συμπεριληφθούν οι ανάλογοι κανόνες φωνητικής μεταγραφής και επομένως την αδυναμία του συστήματος να αντιμετωπίσει ανάλογες περιπτώσεις σε λέξεις που ενδεχομένως να μην είναι Ελληνικές, αλλά θα περιέχουν τις ακολουθίες αυτές.

5.2.7.1 Η αντιστοίχιση της ορθογραφικής στην φωνητική αναπαράσταση

Ένα από τα σημαντικότερα θέματα στην διαδικασία αυτοματοποίησης φωνητικής μεταγραφής βρίσκεται στην διαδικασία της αυτόματης αντιστοίχησης μεταξύ της ορθογραφικής και της φωνητικής αναπαράστασης. Η διαδικασία αυτή δεν είναι μία απλή σύγκριση αλφαριθμητικών ακολουθιών, αλλά μία ιδιαίτερα πολύπλοκη διαδικασία που δεν αντιμετωπίζεται εντελώς αυτόματα. Η αυτόματη προσέγγιση προσπαθεί εντελώς αυτόματα να δημιουργήσει ζευγάρια γραφημάτων και φωνημάτων, δοκιμάζοντας όλους τους δυνατούς συνδυασμούς αντιστοίχησης για κάθε λέξη, όπου στο τέλος γίνεται χρήση του αλγορίθμου Viterbi για την ανεύρεση του βέλτιστου μονοπατιού αντιστοιχήσεων. Ο αλγόριθμος αυτός καταλήγει με πιθανές αντιστοιχήσεις μεταξύ γραφημάτων και φωνημάτων μαζί με τις αντίστοιχες πιθανότητες εμφάνισης. Όπως είναι φυσικό, ο συγκεκριμένος αλγόριθμος δεν είναι απόλυτα ακριβής και εκτός των άλλων απαιτεί μεγάλο όγκο δεδομένων εκμάθησης.

Η ημιαυτόματη προσέγγιση βασίζεται στις ίδιες αρχές με την προηγούμενη αυτόματη, ωστόσο υπάρχει ένα αρχικό σύνολο βασικών κανόνων αντιστοίχησης με βάση το οποίο γίνεται η αντιστοίχηση μεταξύ των γραφημάτων και των φωνημάτων. Η ημιαυτόματη προσέγγιση έχει καλύτερα αποτελέσματα από την αυτόματη, ωστόσο και αυτή παρουσιάζει αρκετά προβλήματα και λάθη στα αποτελέσματά της, τα οποία όπως είναι φυσικό άλλωστε μεταφέρονται και στα

υπόλοιπα στάδια της φωνητικής μετατροπής. Ο αλγόριθμος για την αυτόματη εύρεση βέλτιστων μονοπατιών συνήθως γίνεται με τη βοήθεια της μεθόδου της δυναμικής χρονικής στρέβλωσης (Dynamic Time Warping).



Σχήμα 51: Αντιστοίχιση των γραμμάτων στην φωνητική αναπαράσταση μέσω της δυναμικής στρέβλωσης.

Η μέθοδος αυτή γενικότερα ενδείκνυται για παρόμοιες περιπτώσεις όπου επιχειρείται η ανεύρεση ενός βέλτιστου μονοπατιού μεταξύ εναλλακτικών περιπτώσεων και η αντιστοίχηση όμοιων ακολουθιών που διαφέρουν μόνο στο μήκος. Γενικότερα, η προσέγγιση που προτιμάται για την εργασία αυτή είναι η ημιαυτόματη μέθοδος με την χρήση του φωνήματος *epsilon*.

Στην περίπτωσή μας, δεν κάνουμε χρήση κάποιου αυτόματου τρόπου απεικόνισης, ούτε χρήση του φωνήματος *epsilon*. Αντίθετα, ορίσαμε από την αρχή όλα τα δυνατά ζευγάρια απεικόνισης μονοσήμαντα, χωρίς να υπάρχει κάποια επικάλυψη ή ασάφεια μεταξύ διαφορετικών απεικονίσεων. Στην συνέχεια, η διαδικασία της αντιστοίχησης των γραφημάτων σε φωνήματα γίνεται με βάση τα πρότυπα αυτά, δίνοντας προτεραιότητα στα πρότυπα με το μεγαλύτερο μήκος. Κατά αυτόν τον τρόπο εξασφαλίζεται η μοναδικότητα των προτύπων απεικόνισης και η μη επικάλυψη τους. Ένα παράδειγμα εφαρμογής του αλγορίθμου που χρησιμοποιήσαμε φαίνεται στο παρακάτω διάγραμμα.

διά	γ	ρ	α	μμ	α
↓	↓	↓	↓	↓	↓
ðja	**ɣ**	**r**	**a**	**m**	**a**

Σχήμα 52: Φωνητική μεταγραφή της λέξης *διάγραμμα* χωρίς την χρήση του φωνήματος *epsilon* (προσέγγισή μας).

δ	ι	ά	γ	ρ	α	μ	μ	α
↓	↓	↓	↓	↓	↓	↓	↓	↓
ð	**j**	**a**	**ɣ**	**ɾ**	**a**	**m**	**epsilon**	**a**

Σχήμα 53: Φωνητική μεταγραφή της λέξης *διάγραμμα* με την χρήση του φωνήματος *epsilon* (άλλες προσεγγίσεις).

Στην περίπτωση που θα χρησιμοποιούσαμε το φώνημα *epsilon,* το δεύτερο /μ/ της λέξης θα αντιστοιχιζόταν σε αυτό, με αποτέλεσμα να υπάρχει ο δυνατός κανόνας /μ/ → epsilon όπως φαίνεται στο σχήμα 52.

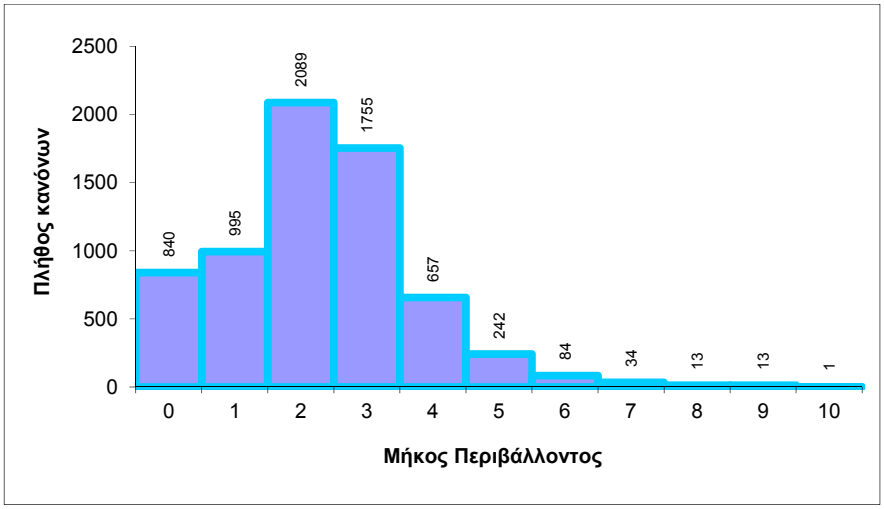
5.2.7.2 Αυτόματη ανεύρεση κανόνων

Το τελευταίο στάδιο της διαδικασίας περιλαμβάνει την αυτόματη ανίχνευση καθολικών κανόνων της μορφής *(αριστερό περικείμενο)*-εστία-*(δεξί περικείμενο)*, όπως περιγράφτηκε παραπάνω. Όπως αναφέραμε ήδη, η ιδέα είναι η εύρεση κανόνων που ισχύουν σε όλο το μήκος του λεξικού εκπαίδευσης, με το μικρότερο δυνατό συνολικό περικείμενο. Το μήκος του περικειμένου για κάθε κανόνα αυξάνει με σταθερό βήμα, έως ότου ο συγκεκριμένος κανόνας ισχύει για όλες τις λέξεις του λεξικού χωρίς να υπάρχουν άλλες λέξεις που να αντικρούουν τον συγκεκριμένο κανόνα. Στην περίπτωση που το περιβάλλον περικείμενο ενός κανόνα φτάσει στα όρια μιας λέξης και δεν παρουσιάζει καθολική ισχύ, τότε η λέξη αυτή αποτελεί εξαίρεση στους κανόνες του λεξικού και αποθηκεύεται στο λεξικό των εξαιρέσεων.

Ο αλγόριθμος που εκτελείται δίδεται παρακάτω σε μορφή ψευδοκώδικα όπως επίσης και σε διάγραμμα ροής διαδικασίας.

```
1.  i = 1 /* initialize the cursor */
2.  Starting from the letter at position i find the longest transcription pattern that can match with
    the next n letters (n=sizeof(best_unit))
3.  i = i+n/* proceed with the parsing */
4.  if ( i!=sizeof(string)) goto Step 2
```

Στο διάγραμμα που ακολουθεί φαίνεται το πλήθος όπως επίσης και η κατανομή των μηκών του συνολικού περικειμένου των κανόνων που προέκυψαν.



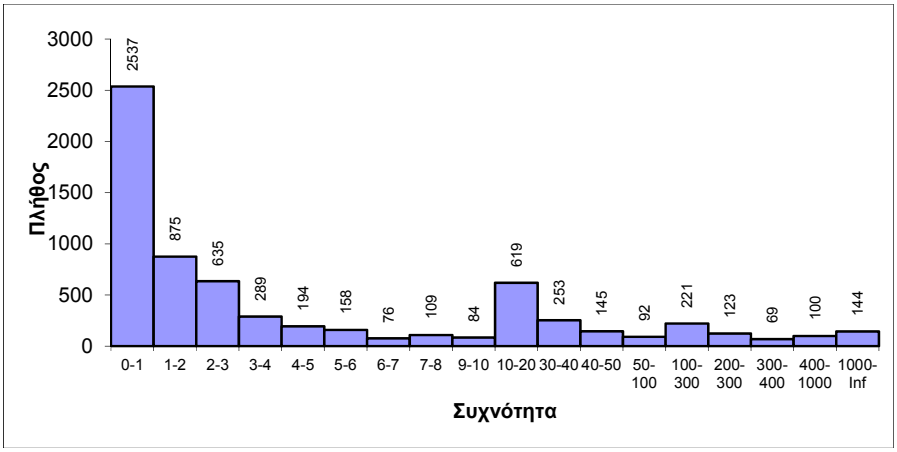
Σχήμα 54: Πλήθος κανόνων ως προς το μήκος του περικειμένου.

Συνολικό Μήκος	Σύνολο Κανόνων	(Αριστερό, Δεξί)										
0	840	**(0,0)**										
		840										
1	995	**(1,0)**	**(0,1)**									
		454	541									
2	2089	**(2,0)**	**(1,1)**	**(0,2)**								
		977	505	607								
3	1755	**(3,0)**	**(2,1)**	**(1,2)**	**(0,3)**							
		1011	205	170	369							
4	657	**(4,0)**	**(3,1)**	**(2,2)**	**(1,3)**	**(0,4)**						
		413	72	37	26	109						
5	242	**(5,0)**	**(4,1)**	**(3,2)**	**(2,3)**	**(1,4)**	**(0,5)**					
		159	3	11	6	20	43					
6	84	**(6,0)**	**(5,1)**	**(4,2)**	**(3,3)**	**(2,4)**	**(1,5)**	**(0,6)**				
		38	8	1	2	3	0	32				
7	34	**(7,0)**	**(6,1)**	**(5,2)**	**(4,3)**	**(3,4)**	**(2,5)**	**(1,6)**	**(0,7)**			
		7	2	4	1	0	9	2	9			
8	13	**(8,0)**	**(7,1)**	**(6,2)**	**(5,3)**	**(4,4)**	**(3,5)**	**(2,6)**	**(1,7)**	**(0,8)**		
		1	2	0	1	0	4	1	0	4		
9	13	**(9,0)**	**(8,1)**	**(7,2)**	**(6,3)**	**(5,4)**	**(4,5)**	**(3,6)**	**(2,7)**	**(1,8)**	**(0,9)**	
		1	0	0	0	0	0	6	1	0	5	
10	1	**(10,0)**	**(9,1)**	**(8,2)**	**(7,3)**	**(6,4)**	**(5,5)**	**(4,6)**	**(3,7)**	**(2,8)**	**(1,9)**	**(0,10)**
		1	0	0	0	0	0	0	0	0	0	0

Πίνακας 9: Πλήθος κανόνων ιεραρχημένοι με βάση το μήκος του περικειμένου. Οι τιμές στις παραγράφους αναφέρονται στο μήκος του αριστερού και δεξιού περικειμένου για κάθε κανόνα αντίστοιχα.

Ο παραπάνω πίνακας είναι πολύ χρήσιμος για την εξαγωγή ορισμένων συμπερασμάτων σχετικά με την πολυπλοκότητα των κανόνων και την κατανομή αυτών. Το γεγονός ότι το 94,2% των κανόνων περιέχει περικείμενο περιβάλλον συνολικού μήκους μέχρι 4, μας οδηγεί στο συμπέρασμα ότι η Ελληνική γλώσσα σε γενικές γραμμές παρουσιάζει συνέπεια στον τρόπο φωνητικής μεταγραφής με την πλειοψηφία των κανόνων μεταγραφής να ακολουθούν ένα απλοϊκό πρότυπο. Οι κανόνες δεν είναι κατά πλειοψηφία πολύπλοκοι, γεγονός που υποδηλώνει ότι η γενίκευση τους είναι σχετικά εύκολη και αποδοτική. Από τον ίδιο πίνακα επίσης παρατηρεί κανείς ότι οι κανόνες που έχουν μόνο αριστερό περιβάλλον (45,5%) είναι σημαντικά περισσότεροι από τους κανόνες που έχουν μόνο δεξί περιβάλλον (25,6%). Αυτό μας οδηγεί επίσης στο ποιοτικό συμπέρασμα ότι η συμπεριφορά των φωνημάτων εξαρτάται περισσότερο από περιβάλλον που προηγείται και λιγότερο από αυτό που έπεται. Αυτό άλλωστε είναι και κάτι που περιμέναμε εξαρχής αφού στην Ελληνική γλώσσα, ιδιαίτερα η προφορά δεν είναι τόσο άμεσα εξαρτώμενη από το περικείμενο που ακολουθεί. Το πλήθος των λέξεων που αποτέλεσαν εξαιρέσεις στο σύνολο των κανόνων είναι 113 μοναδικές λέξεις, οι οποίες παρουσιάζουν μοναδικότητα όσον αφορά την φωνητική τους μεταγραφή ενώ το σύνολο των κανόνων που εξήχθησαν από την διαδικασία είναι 6.723.

Στο επόμενο γράφημα κανείς μπορεί να δει την κατανομή των συχνοτήτων εμφάνισης των κανόνων φωνητικής μεταγραφής μέσα στο λεξικό εκπαίδευσης.



Σχήμα 55: Ιστόγραμμα εμφάνισης κανόνων φωνητικής μεταγραφής για τα Ελληνικά όπως παρατηρήθηκαν στο σώμα εκπαίδευσης

Από το παραπάνω γράφημα μπορεί κανείς να παρατηρήσει ότι το μεγαλύτερο ποσοστό των κανόνων έχει μικρή συχνότητα εμφάνισης, γεγονός που μας προβληματίζει αρχικά για την δυνατότητα γενίκευσης τους. Με μία πρώτη επαφή κανείς θα μπορούσε να υποθέσει ότι πολλοί κανόνες με μοναδιαία συχνότητα ενδεχομένως είναι λάθος. Προσεκτικός όμως έλεγχος των κανόνων αυτών έδειξε ότι περιείχαν ελάχιστα λάθη. Θα μπορούσε κανείς να ισχυρισθεί ότι το γεγονός αυτό υποδεικνύει ότι με ενδεχομένως μεγαλύτερο λεξικό να είχαμε ακόμα περισσότερους κανόνες και επομένως δεν έχουμε την μέγιστη δυνατή κάλυψη. Αυτό ωστόσο δεν είναι απολύτως αληθές αφού το ελάχιστο σύνολο λέξεων που καλύπτουν όλους τους παραπάνω κανόνες περιλαμβάνει μόλις 5.883 λέξεις, δηλαδή το 0.7% του υλικού που χρησιμοποιήσαμε για εκπαίδευση, γεγονός που μας επιβεβαιώνει ότι το λεξικό που χρησιμοποιήσαμε ήταν αρκετά πλήρες.

5.2.8 Αποτελέσματα – Συμπεράσματα

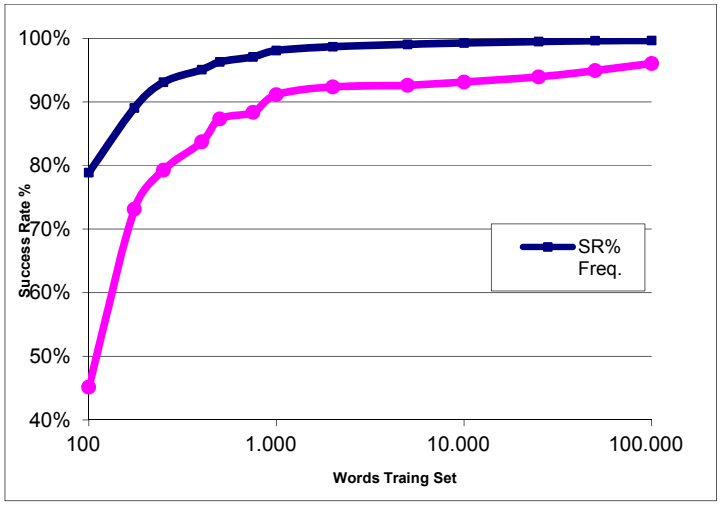
Σε ένα δεύτερο στάδιο μελέτης της πολυπλοκότητας της προφοράς της ελληνικής γλώσσας αλλά και της αποτελεσματικότητας της μεθόδου μας για τα ελληνικά, υπολογίσαμε την ευστοχία του συστήματος για διαφορετικού μεγέθους δεδομένα εκπαίδευσης του συστήματος. Πιο συγκεκριμένα, η διαδικασία που ακολουθήσαμε παρουσιάζεται παρακάτω.

Με την χρήση του ΕΘΕΓ (Εθνικός Θησαυρός Ελληνικής Γλώσσας) [HNC2005] εξάγαμε μία λίστα των συχνότερων ελληνικών λέξεων, μέσα από την οποία δημιουργήσαμε διαφορετικά σετ

εκπαίδευσης για το σύστημα, χρησιμοποιώντας κάθε φορά τις N συχνότερες λέξεις. Με βάση τις λέξεις αυτές εξάγονταν κάθε φορά διαφορετικοί κανόνες, η ευστοχία των οποίων ελεγχόταν κάθε φορά με βάση την λίστα των λέξεων που έχουν μεταγραφεί χειρωνακτικά. Στον παρακάτω πίνακα μπορεί κανείς να δει τα αποτελέσματα της παραπάνω διαδικασίας.

Αριθμός Λέξεων N	*Επιτυχία % (Με σχετική συχνότητα)*	*Επιτυχία (Χωρίς σχετική συχνότητα) %*	*Πλήθος Κανόνων*
100	78.88%	45.15%	49
175	89.06%	73.12%	72
250	93.10%	79.28%	94
400	95.08%	83.72%	117
500	96.31%	87.32%	127
750	97.09%	88.36%	145
1,000	98.10%	91.14%	176
2,000	98.69%	92.38%	226
5,000	99.06%	92.61%	342
10,000	99.28%	93.14%	495
25,000	99.50%	93.94%	888
50,000	99.63%	94.94%	1,379
100,000	99.66%	96.05%	2,243
200,000	99.70%	97.31%	3,362

Πίνακας 10: Αποτελέσματα μεθόδου με διαφορετικά σύνολα εκμάθησης του συστήματος.



Σχήμα 56: Ποσοστό επιτυχίας του συστήματος αυτόματης φωνητικής μεταγραφής σε σχέση με το σώμα λέξεων εκπαίδευσης.

Τα συμπεράσματα της διαδικασίας αυτής ήταν ιδιαίτερα σημαντικά αλλά και ταυτόχρονα χαρακτηριστικά της σχετικής απλότητας που διακρίνει την προφορά της ελληνικής γλώσσας. Παρατηρούμε πως με ιδιαίτερα μικρό αριθμό λέξεων μπορεί κανείς να παράγει σύστημα αυτόματης φωνητικής μεταγραφής με ανεκτά ποσοστά λάθους. Πιο συγκεκριμένα, μπορεί κανείς να παρατηρήσει ότι λαμβάνοντας μόλις τις 1,000 συχνότερες λέξεις, επιτυγχάνει ακρίβεια μεταγραφής 98.10% στο σύνολο του σώματος κειμένου που χρησιμοποιήθηκε (λαμβάνοντας υπόψη τις σχετικές συχνότητες των λέξεων), ενώ στο σύνολο του λεξικού μας των 890,000 διαφορετικών λέξεων επιτυγχάνει ευστοχία της τάξης του 91.14%. Ωστόσο παρατηρεί κανείς ότι ο υπολογιστικός φόρτος είναι δυσανάλογος της βελτίωσης της ακρίβειας του συστήματος, και με αύξηση π.χ. του αριθμού των λέξεων κατά 100%, η αντίστοιχη βελτίωση της ακρίβειας είναι μικρότερη της μίας ποσοστιαίας μονάδας. Θα μπορούσε κανείς να συμπεράνει ότι για την περίπτωση της φωνητικής μεταγραφής του συγκεκριμένου σώματος κειμένου θα αρκούσε η εκμάθηση του συστήματος με τις 5,000 συχνότερες λέξεις, αφού έτσι θα επιτύγχανε ακρίβεια της τάξης του 99.06%. Ωστόσο, το αντίστοιχο ποσοστό για το λεξικό που διαθέτουμε είναι σημαντικά χαμηλότερο με τιμή 92.61%. Η καμπύλη της αναλογίας του υπολογιστικού φόρτου με την ακρίβεια του συστήματος αυξάνει εκθετικά με μέγιστη τιμή το 100%, που όμως τείνει να λάβει σε πολύ μεγάλες τιμές υπολογιστικού κόστους.

Η γρήγορη βελτίωση της ακρίβειας του συστήματος στις χαμηλές τιμές του N είναι χαρακτηριστικό της απλότητας της προφοράς της ελληνικής γλώσσας, αφού υποδεικνύει ότι η προφορά διακρίνεται από σχετικά αυστηρούς κανόνες, με λίγες εξαιρέσεις. Το χαρακτηριστικό αυτό οφείλεται εκτός των άλλων και στο βασικό γνώρισμα της Ελληνικής γλώσσας όπου δεν απαιτείται πρόβλεψη του τόνου, αλλά αντίθετα αυτός σημειώνεται πάνω στην λέξη.

5.3 Αποτελέσματα - Θέματα για περαιτέρω μελέτη

Σύμφωνα με όλα τα προηγούμενα μπορεί κανείς να παρατηρήσει ότι το σύστημα που αναπτύξαμε είναι ιδιαίτερα αποδοτικό και ακριβές. Σε σύγκριση με αντίστοιχα άλλα συστήματα όπως το PHONEMIA [Bakamidis1987], παρουσίασε την υψηλότερη ακρίβεια στα αποτελέσματα, ενώ φαίνεται ότι γενικεύει ορθότερα από όλα τα υπόλοιπα άλλα συστήματα. Πιο συγκεκριμένα, σε δειγματοληπτική σύγκριση που έγινε με το σύστημα PHONEMIA έκδοση Οκτώβριος 2003, παρατηρήθηκε ότι τα δύο συστήματα είχαν ίδια απόκριση σε ποσοστό 92.32%, σε ποσοστό 5.57% το σύστημα μας υπερτερούσε, σε ποσοστό 1.57% το σύστημα μας υπολειπόταν, ενώ σε ποσοστό 0.54% τα δύο αποτελέσματα αν και διαφορετικά μπορούσαν να χαρακτηριστούν και τα δύο εξίσου σωστά.

Έχει παρατηρηθεί ότι ένας μικρός αριθμός λαθών υπεισέρχεται στο σύστημά μας από την χειρωνακτική μεταγραφή του υλικού εκπαίδευσης του συστήματος, και αυτό είναι άλλωστε επόμενο αφού το μέγεθος του υλικού εκπαίδευσης είναι πολύ μεγάλο. Μέσα στα σχέδια μας για το άμεσο μέλλον είναι ο εκτενής έλεγχος των δεδομένων μας με τη βοήθεια στατιστικών στοιχείων που προκύπτουν από τους κανόνες παραγωγής. Χαρακτηριστικό παράδειγμα είναι άλλωστε ότι σε αρχικό στάδιο προέκυπταν και κανόνες μεγαλύτερου μήκους από 10 μονάδες, που οφείλονταν σε ορθογραφικά λάθη.

5.4 Βιβλιογραφία Κεφαλαίου

[Andersen1995] Andersen O. and Dalsgaard P., "Multi-lingual testing of a self-learning approach to phonemic transcription of orthography", Eurospeech95, p. 1117-1120, 1995

[Andersen1996] Andersen O., Kuhn R., et al., "Comparison of two tree-structured approaches for grapheme-to-phoneme conversion", Proc. of ICSLP, pp. 1808-11, 1996.

[Anton1999] Anton Kiraz George, "Compressed storage of sparse finite-state transducers", Workshop on Implementing Automata, 1999.

[Bakamidis1987] Bakamidis S. and Carayannis G., "Phonemia: a phoneme transcription system for speech synthesis in modern Greek", Speech Communication 6 p. 159-169, 1987.

[Black1998] Black Alan W., Kevin Lenzo and Vincent Pagel. "Issues in Building general letter to sound rules.", 3rd ESCA Workshop on Speech Synthesis, proc. 77-80, 1998.

[Bosch2002] Bosch, Antal van den and Walter Daelemans. "Data-oriented methods fro grapheme-to-phoneme conversion", 6th European Conference of the Association for Computational Linguistics, proc. p45-53.

[Caseiro2002] Caseiro D., et al. "Grapheme-to-phone using finite-state transducers", IEEE Workshop on Speech Synthesis, 2002.

[Chalamandaris2005] Chalamandaris A., S. Raptis, and P. Tsiakoulis, "Rule-based grapheme-to-phoneme method for the Greek," *in Interspeech 2005*, pp. 2937-2940, 2005.

[Chotimongkol2000] Chotimongkol Ananlada and Alan W. Black., "Statistically trained orthographic to sound models for Thai", ICSLP, 2000.

[Dermatas1999] Dermatas E. and Kokkinakis G., "A Language-independent probabilistic model for automatic conversion between graphemic and phonemic transcription of words", Eurospeech99, p. 2071-2074, 1999.

[Gildea1995] Gildea Dan and Dan Jurasfsky, "Automatic induction of finite state transducers for simple phonological rules", 33rd Annual Meeting of the Association for Computational Linguistics, proc. p 9-15, 1995.

[HNC2005] Hellenic National Corpus™ (HNC) ILSP, http://hnc.ilsp.gr/ - Web Version 2.0, 2005

[Jansche2001] Jansche Martin, "Re-Engineering Letter-to-Sound Rules", The Second Meeting of the North American Chapter of the Association for Computational Linguistics, 2001.

[Kim2004] Kim Yeon-Jun, Ann Syrdal and Alistair Conkie, "Pronunciation Lexicon Adaptation for TTS Voice Building", ICLSP, 2004.

[Pagel1998] V., Lenzo K. and Black A. W., "Letter to sound rules for accented lexicon compression", ICSLP, vol. I, p 252-255, 1998.

[Petrounias1993] Petrounias E., "Grammatical and comparative analysis of modern Greek", University Studio Press, Thessaloniki, 1993.

[Ravishankar1997] Ravishankar M., Eskenazi M., "Automatic Generation of Context-Dependent Pronunciation", ESCA Eurospeech97, 1997.

[Rentzepopoulos1996] Rentzepopoulos P. and Kokkinakis G., "Efficient multi-lingual Phoneme-to-Grapheme conversion on HMM", Computational Linguistics, vol. 22, p.319-376, 1996.

[Riley1995] Riley Michael D., "A statistical model for generating pronunciation networks", International Conference on Acoustics Speech, and Signal Processing, p. 737-740, 1995.

[Schmidt1993] Schmidt M., et al, "Phonetic Transcription Standards for European Names (Onomastica)", European Conference on Speech Communication and Technology (Eurospeech), vol. 1, pp. 279—282, 1993.

[Sgarbas1998] Sgarbas K., Fakotakis N. and Kokkinakis G., "A PC-KIMMO-Based Bi-directional Graphemic/Phonetic Converter for Modern Greek", Literary and Linguistic Computing, Oxford University Press, vol.13 No.2 pp. 65-75, 1998.

[Torstensson2002] Torstensson Niklas, "Grapheme-to-phoneme conversion, a knowledge-based approach", Speech Music and Hearing TMH-QPSR-Fonetik, vol. 44, p.117-120, 2002.

[Walter1997] Walter M., P. Daelemans and Antal P. J. van den Bosch. "Language-Independent data-oriented grapheme-to-phoneme conversion", Progress in Speech Synthesis, p. 77-89, Springer, (1997) New York.

6. ΤΟ ΦΑΙΝΟΜΕΝΟ ΤΩΝ GREEKLISH

Τα Greeklish δεν αποτελούν μια νέα γλώσσα, διάλεκτο ή γραφή όπως συχνά λέγεται, αλλά ένας διαφορετικός τρόπος αναπαράστασης της ελληνικών λέξεων με λατινικούς χαρακτήρες. Τα βαθύτερα αίτια ύπαρξης του φαινόμενου αυτού ανιχνεύονται στην ελλιπή υποστήριξη της ελληνικής γλώσσας από τους υπολογιστές και άλλα μέσα ηλεκτρονικής επικοινωνίας. Θεωρήσαμε την δημιουργία συνθετικής ομιλίας από Greeklish μεγάλη πρόκληση λόγω της μεγάλης ποικιλότητας και για τον λόγο αυτόν αποφασίσαμε να ασχοληθούμε με το θέμα διεξοδικά. Στο συγκεκριμένο κεφάλαιο παρουσιάζουμε το φαινόμενο των Greeklish, την μελέτη και έρευνα που έχει πραγματοποιηθεί, καθώς επίσης και την προσέγγισή μας για την αυτόματη μετατροπή τους σε ορθά Ελληνικά. Τα πειραματικά αποτελέσματα του συστήματός μας αποδεικνύουν την αποδοτικότητα και αποτελεσματικότητα του αλγορίθμου μας, ειδικότερα αναφορικά και με πραγματικά δεδομένα που έχουν προκύψει από την διαδικτυακή έκδοση του συστήματός μας.

6.1 Εισαγωγή

Μελέτες που έχουν γίνει έχουν δείξει ότι σχεδόν όλοι οι χρήστες υπολογιστών έχουν αναγκαστεί να χρησιμοποιήσουν Greeklish τουλάχιστον μία φορά, ενώ το 50% των χρηστών άνω των 35 ετών θεωρούν τα Greeklish ως αναγκαίο κακό στην καθημερινή τους ενασχόληση με τους υπολογιστές. Σύμφωνα με προηγούμενη μελέτη έγινε φανερή η δυσκολία που τα Greeklish επιφέρουν στην ανάγνωση και κατανόηση ενός κειμένου, αφού κατά μέσο όρο ένα κείμενο γραμμένο σε Greeklish απαιτεί 40% επιπλέον χρόνο για την ανάγνωση και κατανόηση του από ότι αν ήταν γραμμένο σε ελληνικά.

Ιδιαίτερο χαρακτηριστικό των Greeklish είναι η ποικιλότητα και η ασυνέπεια που τα διακρίνει, όσον αφορά στους κανόνες απεικόνισης ελληνικών χαρακτήρων σε λατινικούς ή σύμβολα. Παραδείγματος χάριν, το ελληνικό γράμμα /θ/ συχνά μεταγράφεται με τα γράμματα /8/ /u/ /th/ /9/ /0/, ενώ π.χ. το γράμμα /δ/ αντίστοιχα στα γράμματα /d/ /th/ /6/. Πρόσφατη μελέτη υποστηρίζει ότι οι διαφορετικοί τύποι Greeklish μπορούν να ομαδοποιηθούν σε τρεις βασικές κατηγορίες [Androutsopoulos1999]:

1. στο φωνητικό Greeklish, όπου η ορθογραφία δεν λαμβάνεται υπόψη αλλά ο χρήστης προσπαθεί να αποδώσει την φωνητική διάσταση των λέξεων, χρησιμοποιώντας π.χ. το /i/ αντί των /η/, /ει/, /οι/ κ.ο.κ.

2. στο οπτικό Greeklish, όπου ο χρήστης προσπαθεί να αποδώσει όσο το δυνατό καλύτερα την οπτική διάσταση των λέξεων, χρησιμοποιώντας π.χ. το /3/ αντί του /ξ/, το /n/ αντί του /η/, το /w/ αντί /ω/ κ.ο.κ.

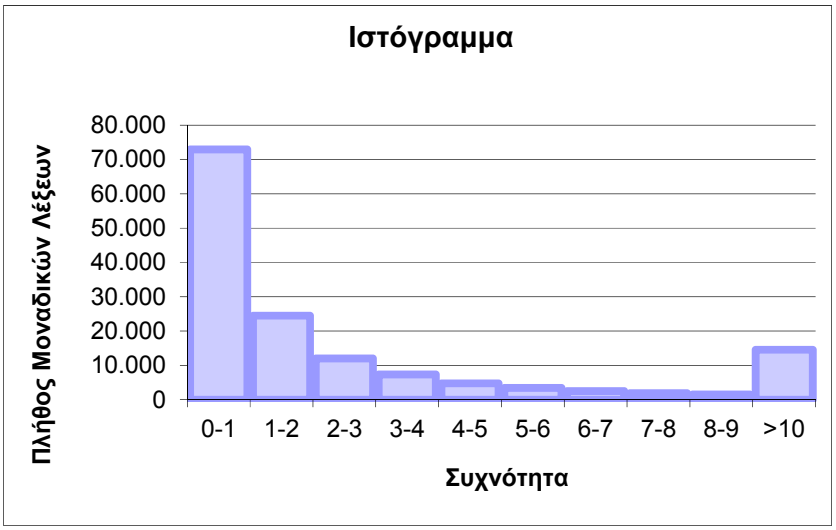
3. στο Greeklish σύμφωνα με την διάταξη του πληκτρολογίου, όπου ο χρήστης χρησιμοποιεί τους λατινικούς χαρακτήρες που συνυπάρχουν με τους αντίστοιχους ελληνικούς στο μέσο πληκτρολόγιο, χρησιμοποιώντας π.χ. τον χαρακτήρα /u/ αντί του /θ/, το /c/ αντί του /ψ/ κ.ο.κ.

Ωστόσο, είναι πλέον γνωστό ότι κανένας χρήστης δεν αρκείται σε έναν από αυτούς τους τύπους Greeklish, αλλά τους συνδυάζει παράλληλα με προσωπικές προτιμήσεις που κι αυτές ενδεχομένως

αλλάζουν ανάλογα με την διάθεση του χρήστη. Παρ' όλα αυτά θα ήταν υπερβολή να υποστηρίξει κανείς αυτό που έχει κατά καιρούς λεχθεί, ότι υπάρχουν τόσα διαφορετικά είδη Greeklish όσοι είναι και οι χρήστες υπολογιστών.

6.2 Στατιστική ανάλυση

Τα δεδομένα που χρησιμοποιήθηκαν για την στατιστική ανάλυση είναι τα κείμενα που συλλέχθηκαν από την δωρεάν υπηρεσία αυτόματης μετατροπής Greeklish σε Ελληνικά [Chalamandaris2004] (οι χρήστες είχαν το δικαίωμα να μετατρέψουν κείμενο μήκους μέχρι 255 χαρακτήρων) και συλλέχθηκε σε διάστημα 9,5 μηνών (από 1/1/2005 μέχρι 15/9/2005). Είναι απαραίτητο να τονίσει κανείς στο σημείο αυτό ότι τα κείμενα αυτά αποτελούν πραγματικά Greeklish από καθημερινή χρήση, ενώ το μέγεθος του δείγματος επιτρέπει την αβίαστη εξαγωγή όσο το δυνατό γενικών συμπερασμάτων. Το δείγμα των κειμένων Greeklish που συλλέχθησαν αποτελείται από 145.601 μοναδικές λέξεις που εμφανίζονται συνολικά 2.095.037 φορές. Το ιστόγραμμα συχνότητας εμφάνισης των μοναδικών λέξεων φαίνεται στο παρακάτω γράφημα.



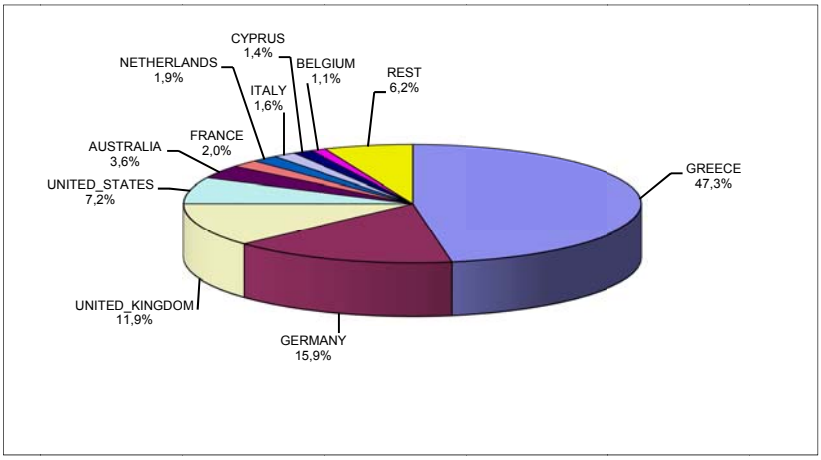
Σχήμα 57: Ιστόγραμμα συχνότητας εμφάνισης των μοναδικών λέξεων στο σώμα κειμένου μελέτης.

Το σώμα αυτό κειμένου αποτελείται από 171.698 διαφορετικές χρήσεις (requests), ενώ οι μοναδικές διευθύνσεις από όπου αυτό προέρχεται είναι 18.868. Οι χρήστες προέρχονται συνολικά

από 67 διαφορετικές χώρες. Οι πρώτες 10 χώρες, όσον αφορά την συχνότητα χρήσης της υπηρεσίας αυτής, φαίνονται στον πίνακα που ακολουθεί.

A/A	COUNTRY	Requests	Request %	Unique_IPs	Unique_IPs %
1	GREECE	81.187	47,28%	10710	53,75%
2	GERMANY	27.217	15,85%	3366	16,89%
3	UNITED_KINGDOM	20.425	11,90%	967	4,85%
4	UNITED_STATES	12.335	7,18%	1069	5,36%
5	AUSTRALIA	6.197	3,61%	404	2,03%
6	FRANCE	3.403	1,98%	401	2,01%
7	NETHERLANDS	3.196	1,86%	86	0,43%
8	ITALY	2.774	1,62%	244	1,22%
9	CYPRUS	2.382	1,39%	334	1,68%
10	BELGIUM	1.868	1,09%	229	1,15%
	REST	10.714	6,24%	1058	5,31%

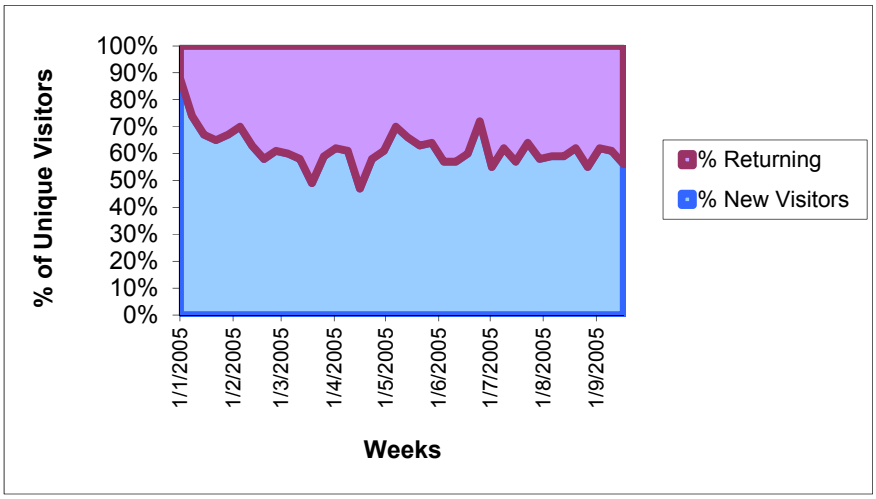
Πίνακας 11: Κατανομή της προέλευσης του σώματος κειμένου ανά χώρα στην διάρκεια 6 μηνών.



Σχήμα 58: Κατανομή της προέλευσης του σώματος κειμένου ανά χώρα

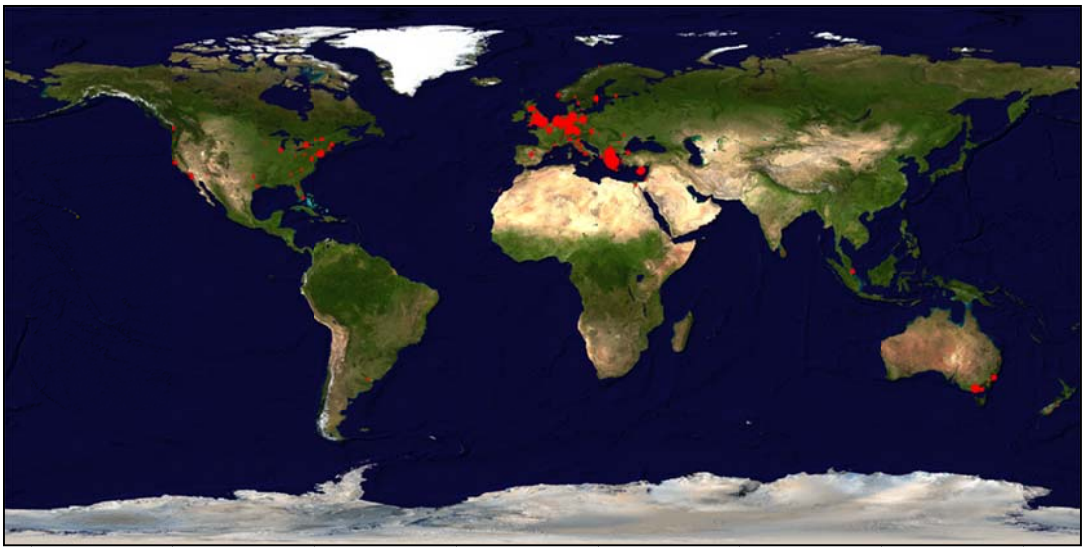
Όπως άλλωστε αναμενόταν, το μεγαλύτερο ποσοστό των χρηστών προέρχεται από την Ελλάδα, ενώ ακολουθούν χώρες όπου κατοικεί μεγάλο μέρος του απόδημου ελληνισμού. Με την χρήση της τεχνολογίας των cookies προκύπτει ότι κατά μέσο όρο το 61% των χρηστών είναι νέοι

χρήστες, ενώ το 39% είναι χρήστες που έχουν επισκεφθεί περισσότερες από μία φορές τον δικτυακό μας τόπο.



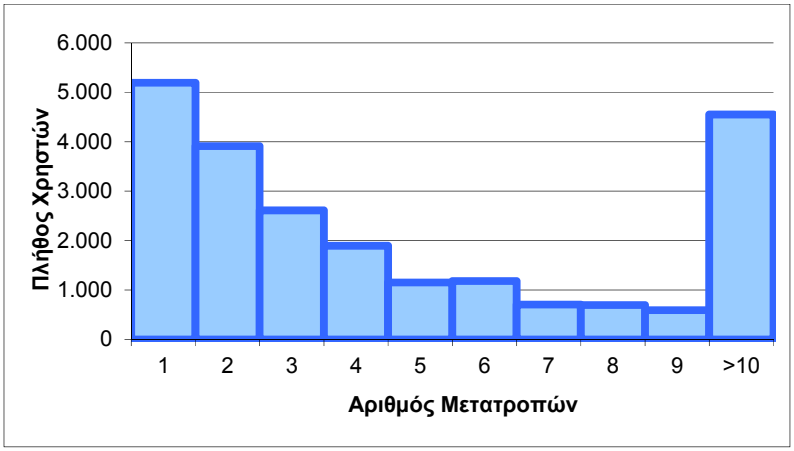
Σχήμα 59: Κατανομή των μοναδικών επισκεπτών σε νέους και συχνούς επισκέπτες.

Στην παρακάτω εικόνα απεικονίζεται η γεωγραφική προέλευση των χρηστών του συστήματος, οι οποίοι προέρχονται από περισσότερες από 284 διαφορετικές πόλεις. Το μέγεθος της κουκίδας υποδηλώνει το πλήθος των διαφορετικών διευθύνσεων που ανήκουν στην κάθε περιοχή.



Σχήμα 60: Γεωγραφική προέλευση των χρηστών του συστήματος.

Ένα σημαντικό στατιστικό στοιχείο επίσης αποτελεί η συχνότητα χρήσης από τους μοναδικούς χρήστες, αφού το στοιχείο αυτό υποδηλώνει εκτός από το αν οι χρήστες θεωρούν ότι είναι χρήσιμη η εφαρμογή και επομένως την χρησιμοποιούν συχνά, και το πόσο χαρακτηριστικά είναι τα δεδομένα που έχουμε συλλέξει για τους χρήστες ξεχωριστά. Πιο συγκεκριμένα, για την μελέτη των συνηθειών και των χαρακτηριστικών του τύπου Greeklish που κάθε χρήστης επιλέγει, είναι αναγκαία η ύπαρξη ενός αρκετά μεγάλου δείγματος και όχι απλά μερικές λέξεις ή φράσεις. Για την μελέτη που ακολουθεί χρησιμοποιήσαμε τα δεδομένα από χρήστες που είχαν περισσότερες από 100 διαφορετικές εγγραφές στο σύστημα και επομένως αξιοσημείωτο μέγεθος δεδομένων. Στο γράφημα που ακολουθεί φαίνεται το ιστόγραμμα των χρηστών ανά αριθμό χρήσεων. Ωστόσο το γράφημα δεν είναι απόλυτα ακριβές αφού μη στατικές διευθύνσεις που αντιστοιχούν σε ίδιο χρήστη καταγράφονται ως διαφορετικοί χρήστες και έτσι έχουμε μη πραγματική συσσώρευση παρατηρήσεων προς την αρχή του x-άξονα.



Σχήμα 61: Κατανομή διαφορετικών χρηστών ανά αριθμό μετατροπών.

6.3 Κανόνες απεικόνισης – χαρακτηριστικά της γραφής Greeklish

Κατά τον αρχικό σχεδιασμό του συστήματος αυτόματης μετατροπής από Greeklish σε Ελληνικά, ορίσθηκαν χειρωνακτικά περίπου 65 διαφορετικοί κανόνες [Chalamandaris2004a]. Η συστηματική χρήση όμως του συστήματος από πολλούς διαφορετικούς χρήστες έδειξε ότι οι κανόνες αυτοί δεν επαρκούσαν και ορίσθηκαν εκ νέου επιπλέον κανόνες, οι οποίοι στο σύνολο τους έφτασαν τους 165. Το μεγάλο πλήθος των κανόνων οφείλεται κατά μεγάλο ποσοστό και το γεγονός ότι συχνά οι χρήστες Greeklish δεν ξέρουν καλή ορθογραφία ή δεν την λαμβάνουν υπόψη τους όταν γράφουν Ελληνικά με λατινικό αλφάβητο, με αποτέλεσμα τα υποκείμενα ορθογραφικά λάθη να δημιουργούν νέα υποσύνολα κανόνων απεικόνισης. Στον πίνακα στο παράρτημα, μπορεί κανείς να παρατηρήσει τις πιο πιθανές απεικονίσεις από ελληνικούς χαρακτήρες σε λατινικούς, μαζί με τα αντίστοιχα ποσοστά εμφάνισης στο σώμα κειμένου που συλλέξαμε.

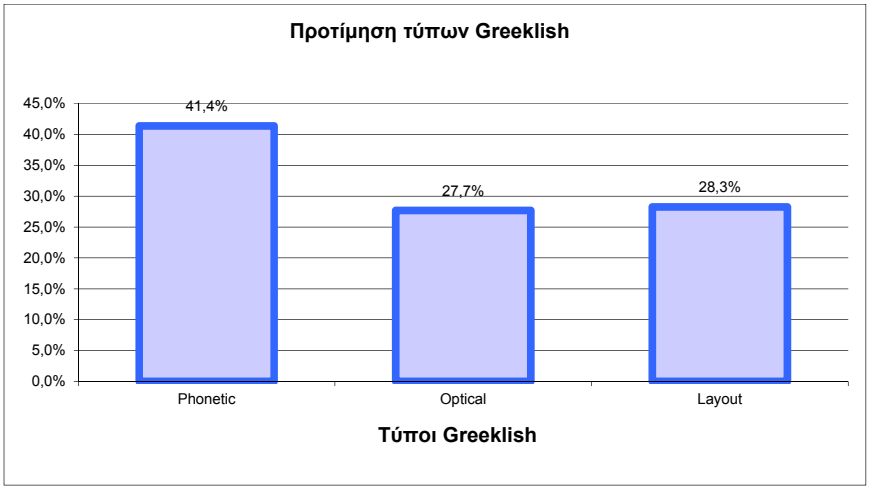
Για να αποκτήσει κανείς μία καλύτερη εικόνα του μεγέθους της πολυπλοκότητας που περιέχεται στο φαινόμενο των Greeklish, υπολογίσαμε όλους τους πιθανούς τρόπους αναπαράστασης σε Greeklish όλων των μοναδικών λέξεων που περιέχονται στο σώμα κειμένων ΕΘΕΓ. Οι πιθανοί τρόποι προέκυψαν λαμβάνοντας για κάθε γράμμα ή δίφθογγο όλους τους δυνατούς τρόπους αναπαράστασης που παρουσιάστηκαν στατιστικά σημαντικοί στο δείγμα που συλλέξαμε. Κατά αυτόν τον τρόπο, π.χ. το γράμμα /η/ παρουσιάστηκε να έχει 3 διαφορετικούς σημαντικούς

τρόπους μεταγραφής, η ένωση των οποίων παρείχε κάλυψη του συνολικού δείγματος μεγαλύτερη του 90%.

6.4 Διαφορετικοί τύποι γραφής Greeklish

Στο σημείο αυτό της έρευνάς μας [Chalamandaris2004b], αποφασίσαμε να μελετήσουμε τους διαφορετικούς τύπους Greeklish που συναντιόνται και που ενδεχομένως έχει νόημα η ομαδοποίηση τους. Σε προηγούμενη παράγραφο μιλήσαμε για τα τρία διαφορετικά είδη Greeklish, όπως αυτά μπορούν να ορισθούν ανάλογα με το τι προσπαθούν να εξυπηρετήσουν, αν δηλαδή προσπαθούν να ομοιάσουν οπτικά, φωνητικά ή ανάλογα με την διάταξη των ελληνικών γραμμάτων στο πληκτρολόγιο. Ωστόσο ο διαχωρισμός αυτός, αν και έχει νόημα και προσπαθεί να εκλογικεύσει την ποικιλότητα των Greeklish, δεν έχει μελετηθεί αν διέπει την πλειοψηφία των χρηστών Greeklish, και αν οι ίδιοι οι χρήστες είναι συνεπείς ως προς αυτές της κατηγορίες [Chalamandaris2006].

Σε πρώτο στάδιο μελετήσαμε την γενική προτίμηση των χρηστών όσον αφορά τους τρεις διαφορετικούς τύπους Greeklish, το οπτικό, το φωνητικό και ανάλογα με την διάταξη των γραμμάτων στο πληκτρολόγιο του υπολογιστή. Λαμβάνοντας υπόψη μόνο τους ελληνικούς χαρακτήρες που παρουσιάζουν την ανάλογη ποικιλότητα, έχουν δηλαδή τουλάχιστον τρεις διαφορετικές απεικονίσεις με λατινικά γράμματα και αντιστοιχούν στους παραπάνω τρεις τύπους, μετρήσαμε την προτίμηση των χρηστών. Πιο συγκεκριμένα τα γράμματα που μελετήθηκαν είναι τα /η/, /θ/, /ξ/, /υ/, /ψ/ και το /ω/, καθώς αυτά αποτελούν χαρακτηριστικούς αντιπροσώπους των διαφορετικών τύπων γραφής Greeklish. Στον παρακάτω πίνακα φαίνονται τα ποσοστά των αντίστοιχων μετρήσεων για καθένα από τα γράμματα, ενώ στο διάγραμμα φαίνεται ο συνολικός μέσος όρος για καθένα από τους τρεις τύπους Greeklish.



Σχήμα 62: Μέσος όρος προτίμησης των τριών τύπων Greeklish.

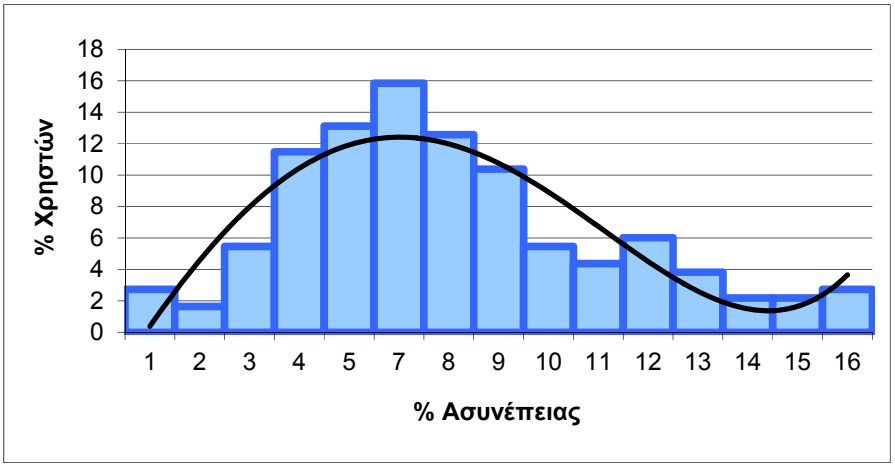
Ελληνικό Γράμμα		Phonetic		Optical		Layout	
ή	->	i	46.3%	n	8.8%	h	33.1%
αυ	->	ab,av,af,aph	42.7%	au	33.7%	ay	23.6%
αύ	->	ab,av,af,aph	39.9%	au	31.5%	ay	28.6%
ευ	->	eb,ef,ev	32.1%	eu	38.7%	ey	29.2%
εύ	->	eb,ef,ev	39.4%	eu	33.8%	ey	26.7%
η	->	i	38.4%	n	14.0%	h	37.4%
θ	->	th	45.1%	8,9,0,6	29.3%	u	14.9%
ντ	->	d,nd	19.9%	vt	4.3%	nt	75.8%
ξ	->	x,ks	74.1%	3	16.8%	j	9.1%
ου	->	u	21.1%	ou	47.6%	oy	31.2%
ού	->	u	17.0%	ou	53.3%	oy	29.7%
υ	->	i	38.7%	u	24.0%	y	35.2%
ψ	->	ps	89.8%	y	7.9%	c	2.3%
ω	->	o	41.3%	w	34.2%	v	20.6%
ύ	->	i	35.7%	u	29.1%	y	33.9%
ώ	->	o	40.4%	w	35.9%	v	20.8%
Μέσος Όρος			**41.4%**		**27.7%**		**28.3%**

Πίνακας 12: Οι τρεις διαφορετικοί τύποι Greeklish και οι προτιμήσεις αυτών από τους χρήστες.

Παρατηρούμε λοιπόν ότι ο φωνητικός τύπος Greeklish προτιμάται περισσότερο από τους χρήστες με ποσοστό 41,4%, ενώ οι δύο άλλοι τύποι χρησιμοποιούνται αμφότεροι με ένα ποσοστό

της τάξης περίπου των 28%. Οφείλουμε να τονίσουμε στο σημείο αυτό ότι για τον υπολογισμό των ποσοστών αυτών δεν λάβαμε υπόψη την συχνότητα των γραμμάτων μέσα στο σώμα κειμένων, αλλά τα σχετικά ποσοστά ανά γράμμα, π.χ. για το γράμμα /ψ/, παρόλο που δεν είναι τόσο συχνό όσο το γράμμα /η/, δεν υπολογίστηκε βάρος ανάλογο της συχνότητας του γράμματος στο σώμα κειμένου, αλλά θεωρήθηκε ισότιμο με τα υπόλοιπα γράμματα στον υπολογισμό του μέσου όρου.

Ένα σημαντικό χαρακτηριστικό που άξιζε να μελετήσουμε είναι και το ποσοστό ασυνέπειας του κάθε χρήστη ως προς τον τύπο που χρησιμοποιεί. Για τον υπολογισμό της παραμέτρου αυτής μετρήσαμε κατά μέσο όρο, για κάθε χρήστη ξεχωριστά, το ποσοστό των περιπτώσεων που ο χρήστης δεν κάνει χρήση του πιο συχνού προτύπου απεικόνισης κατά αυτόν. Στην περίπτωση π.χ. που ο χρήστης μεταγράφει το γράμμα /θ/ με το λατινικό γράμμα /u/ σε ποσοστό 80%, τότε το υπόλοιπο 20% λαμβάνεται ως ποσοστό ασυνέπειας.



Σχήμα 63: Κατανομή του ποσοστού ασυνέπειας ως προς τον προτιμητέο τύπο γραφής, ανά χρήστη.

Οι χρήστες κατά μέσο όρο παρουσιάζουν ασυνέπεια στον τρόπο μεταγραφής των ελληνικών γραμμάτων σε ποσοστό 7,65% κατά μέσο όρο, ενώ παρατηρήσαμε ότι η καμπύλη της κατανομής των μετρήσεων παρουσιάζει κανονικότητα στην μορφή της. Το ποσοστό αυτό, αν και δεν είναι μεγάλο, κάνει φανερή την ύπαρξη ασυνέπειας και την έλλειψη ενός μοναδικού τρόπου μεταγραφής ακόμα και μέσα στον ίδιο χρήστη. Αυτό άλλωστε είναι και ένα από τα χαρακτηριστικά γνωρίσματα των Greeklish.

6.5 Ανάλυση των προτιμήσεων στο δείγμα

Στο δείγμα πραγματικών δεδομένων που συλλέξαμε κατά την διαδικασία αυτή αποφασίσαμε να πραγματοποιήσουμε ανάλυση των διαφορετικών τρόπων αναπαράστασης ανά χρήστη, με σκοπό την επιβεβαίωση ή όχι των τριών διαφορετικών τύπων Greeklish που αναφέραμε παραπάνω [Chalamandaris2006]. Για τον λόγο αυτό, δημιουργήσαμε μία σειρά από ιεραρχικά λογαριθμικά μοντέλα χωρίς αλλά και με λανθάνουσες τάξεις (hierarchical log-linear models with and without latent classes) με σκοπό να ελέγξουμε την υπόθεση ότι οι χρήστες χρησιμοποιούν τους τρεις αυτούς τύπους αναπαράστασης. Για αυτό το πείραμα χρησιμοποιήσαμε τις μεταγραφές των ελληνικών γραμμάτων η, υ, ω, θ και του διφθόγγου ου, τα οποία παρουσιάζουν και τις τρεις διαφορετικές περιπτώσεις: n, u, w, 8 και ou για τον οπτικό, i, i, o, th και u για τον φωνητικό και h, y, u και oy για την τύπο του πληκτρολογίου αντίστοιχα. Στην ανάλυση αυτή χρησιμοποιήσαμε τους 50 συχνότερους μοναδικούς χρήστες (όπως αυτοί προέκυψαν από τον συνδυασμό μοναδικού αριθμού IP και αρχείου cookie). Συνολικά το δείγμα μας περιείχε 170,478 φορές τα προαναφερθέντα γράμματα, με 3,410 φορές κατά μέσο όρο ανά χρήστη.

6.6 Το σύστημα αυτόματης μετατροπής Greeklish σε Ελληνικά

Έχουν γίνει διάφορες προσπάθειες για την αυτόματη μετατροπή Greeklish σε ελληνικά, με περιορισμένες όμως δυνατότητες. Όλες σχεδόν οι προσεγγίσεις βασίζονται στην δημιουργία ενός συνόλου κανόνων απεικόνισης από έναν ή περισσότερους λατινικούς χαρακτήρες σε έναν ή περισσότερους ελληνικούς χαρακτήρες. Η τεχνική ωστόσο αυτή έχει σημαντικούς περιορισμούς λόγω του γεγονότος ότι η γραφή Greeklish παρουσιάζει εξαιρετικά μεγάλο εύρος διαφορετικών απεικονίσεων ενός γράμματος ή συμπλέγματος. Αυτό φάνηκε και από το προηγούμενο παράδειγμα όπου η λέξη /διεύθυνση/ μπορεί να γραφτεί με περίπου 200 διαφορετικούς τρόπους σε Greeklish γραφή. Στο διαδίκτυο υπάρχουν περισσότερες από 20 εφαρμογές και ιστοσελίδες αφιερωμένες στην αυτόματη μετατροπή Greeklish σε ελληνικά.

6.6.1 Η δική μας προσέγγιση

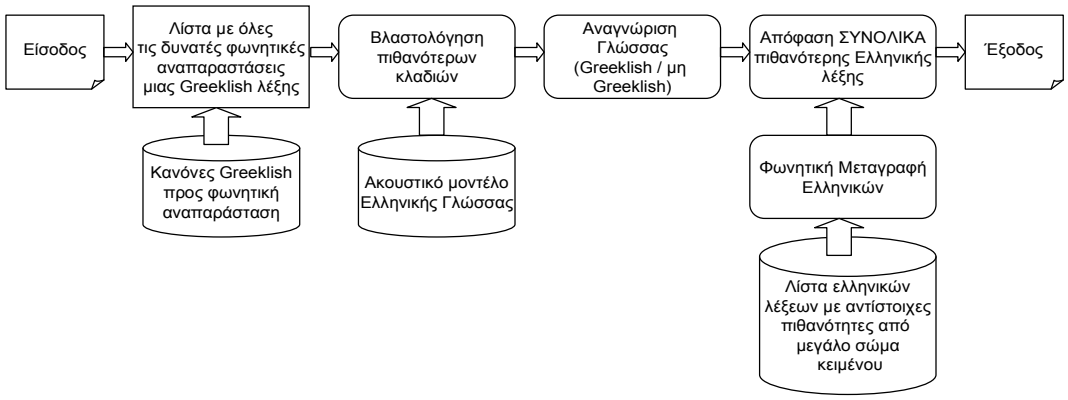
Η δική μας προσέγγιση [Chalamandaris2004a] [Chalamandaris2004b] διαφέρει σημαντικά σε διάφορα επίπεδα, ενώ κάνει επίσης χρήση ενσωματωμένου λεξικού, όπως και οι άλλες μέθοδοι που επιχειρούν να μεταγράψουν σε ορθά Ελληνικά. Η βασική λειτουργία του συστήματος μας ορίζεται από τρία διαφορετικά στάδια:

1. την ανάπτυξη μιας δενδρικής δομής με όλες τις δυνατές αναπτύξεις σε φωνητική αναπαράσταση κάθε Greeklish λέξης

2. την αναγνώριση γλώσσας, ένα απαραίτητο βήμα για μη Ελληνικές λέξεις [Dunning1994]

3. την απόφαση της πιο πιθανής για κάθε περίπτωση ελληνικής λέξης

Το κάθε ένα στάδιο από αυτά ενσωματώνει διαφορετικές διεργασίες και πληροφορίες. Στο ακόλουθο διάγραμμα φαίνεται σχηματικά η λειτουργία του συστήματος.



Σχήμα 64: Σχηματικό διάγραμμα λειτουργίας του συστήματος.

Οι κανόνες μετατροπής από Greeklish σε φωνητική αναπαράσταση, προέκυψαν, όπως είπαμε, από παρατήρηση πραγματικών δειγμάτων γραφής Greeklish στο διαδίκτυο και αριθμούν περί τους 170 διαφορετικούς κανόνες απεικόνισης. Κάνοντας χρήση αυτών των κανόνων, για κάθε λέξη παράγουμε μία δενδρική δομή με όλες τις δυνατές φωνητικές αναπαραστάσεις, οι οποίες στην συνέχεια σταχυολογούνται και απορρίπτονται οι λιγότερο πιθανές σύμφωνα με το ακουστικό μοντέλο που έχει δημιουργηθεί για την ελληνική γλώσσα από λεξικό 1.000.000 λέξεων. Το στάδιο αυτό είναι ιδιαίτερα σημαντικό αφού μας επιτρέπει να μειώνουμε τον υπολογιστικό φόρτο ουσιαστικά.

Κατόπιν, με βάση το ακουστικό μοντέλο, λαμβάνεται απόφαση για το αν η λέξη είναι όντως Greeklish ή μη Greeklish (π.χ. μη Ελληνική λέξη). Η απόφαση αυτή πραγματοποιείται με τον

έλεγχο των υψηλότερων συσσωρευμένων πιθανοτήτων όλων των φωνητικών αναπαραστάσεων. Αν όλες αυτές δεν ξεπερνούν ένα προκαθορισμένο κατώφλι, τότε η προς μετατροπή λέξη θεωρείται ως μη Greeklish και παραμένει ως έχει. Στην αντίθετη περίπτωση οι πιθανότερες φωνητικές αναπαραστάσεις ελέγχονται μέσω λεξικού (λίστας 1.000.000 συχνότερων λέξεων της ελληνικής) και επιλέγεται η συνολικά πιθανότερη λέξη από το λεξικό. Η τελική απόφαση επηρεάζεται από τρεις διαφορετικές παραμέτρους και για τον λόγο αυτό αναφέρουμε τον χαρακτηρισμό συνολικά πιθανότερη λέξη. Οι παράμετροι αυτές είναι η πιθανότητα της φωνητικής αναπαράστασης που προκύπτει από το ακουστικό μοντέλο, η πιθανότητα της αντίστοιχης ελληνικής λέξης μέσα στο λεξικό (όπως αυτή προέκυψε από την μέτρηση της συχνότητας αυτής σε μεγάλο σώμα κειμένου), και ένα επιπλέον κόστος που αναφέρεται στην συμφωνία της αρχικής λέξης Greeklish και της τελικής ελληνικής λέξης, όσον αφορά στην ορθογραφία αυτών. Η τελευταία είναι καθαρά μία ευριστική παράμετρος που παρατηρήθηκε ότι βελτιώνει σημαντικά την αποτελεσματικότητα του συστήματος, ιδιαίτερα σε λέξεις που παρουσιάζουν πλούσια ορθογραφική ποικιλότητα (παράγωγα, μορφήματα, κλίσεις, κ.λπ.), ενώ παράλληλα λαμβάνει υπόψη την ορθογραφία, παρόμοια με την ελληνική, που ενδεχομένως παρουσιάζεται.

6.6.2 Αξιολόγηση της αποτελεσματικότητας του συστήματος

Η αξιολόγηση του συστήματος είναι μία από τις περισσότερο χρονοβόρες, αλλά ταυτόχρονα από τις πιο σημαντικές εργασίες [Chalamandaris2006]. Για την αξιολόγηση του συστήματος χρησιμοποιήσαμε το σώμα κειμένου που αποκτήσαμε από την διαδικτυακή υπηρεσία. Για τον ορθότερο και αρτιότερο έλεγχο χωρίσαμε τις 145.601 μοναδικές λέξεις στις τρεις κατηγορίες τις οποίες χειρίζεται διαφορετικά το σύστημα. Πιο συγκεκριμένα οι κατηγορίες των λέξεων που ελέγχθηκαν είναι οι εξής:

Κατηγορίες λέξεων	***Αριθμός Μοναδικών λέξεων***	***Ποσοστό % των μοναδικών λέξεων***	***Αριθμός συνολικών λέξεων***	***Ποσοστό % των συνολικών λέξεων***
Γνωστή Ελληνική λέξη (Λεξικό)	109.900	75,48%	1.925.797	91,92%
Ξένη λέξη (μη Greeklish)	20.330	13,96%	144.270	6,89%
Άγνωστη Greeklish λέξη	1.470	10,11%	22.112	1,06%
Υπόλοιπες (κολλημένες λέξεις μεταξύ τους, ανορθόγραφες κ.λπ.)	661	0,45%	2.858	0,13%
Σύνολο	**145.601**	**100%**	**2.095.037**	**100%**

Πίνακας 13: Διαχωρισμός των λέξεων σε κατηγορίες προς έλεγχο.

Η τελευταία κατηγορία αν και δεν αποτελεί ειδική κατηγορία ως προς την λειτουργία του συστήματος, ωστόσο δεν μπορεί να περιληφθεί σε μία από τις προηγούμενες κατηγορίες και απαρτίζεται από συστηματικά λάθη όπως είναι π.χ. το κόλλημα δύο διαδοχικών λέξεων ή λέξεων και αριθμών (κυρίως σε περιπτώσεις διευθύνσεων ή τηλεφώνων).

Ο χειρωνακτικός έλεγχος της ορθότητας των μεταγραφών των λέξεων αυτών από το σύστημα έγινε σε ένα ποσοστό κάθε κατηγορίας, και όχι σε ολόκληρο το πλήθος των λέξεων, αφού αυτό είναι πολύ μεγάλο. Πιο συγκεκριμένα, για την πρώτη κατηγορία το δείγμα που ελέγξαμε ήταν ένα τυχαίο σύνολο 5.489 μοναδικών λέξεων, που αντιστοιχεί στο 5% του συνολικού αριθμού μοναδικών λέξεων της κατηγορίας, ενώ το άθροισμα των συχνοτήτων αυτών ήταν 4,74% των συνολικών λέξεων της κατηγορίας αυτής. Για την κατηγορία των ξένων λέξεων (μη Greeklish) ελέγξαμε τυχαίο δείγμα 2.031 μοναδικών λέξεων που αντιστοιχεί σε ποσοστό 10% επί των μοναδικών λέξεων της κατηγορίας αυτής, ενώ το αντίστοιχο άθροισμα των συχνοτήτων αυτών ήταν 11,07% των συνολικών συχνοτήτων της κατηγορίας. Για την τρίτη κατηγορία λέξεων ελέγξαμε τυχαίο δείγμα 1.470 μοναδικών λέξεων που αντιστοιχεί σε ποσοστό 10% των μοναδικών λέξεων της κατηγορίας και σε 9,77% των συνολικών λέξεων της ίδιας κατηγορίας. Η τελευταία κατηγορία δεν ελέγχθηκε καθόλου, αφού δεν έχει νόημα η μεταγραφή των μη ορθογραφημένων λέξεων.

Στον πίνακα που ακολουθεί φαίνονται τα προαναφερθέντα σύνολα περιληπτικά.

Κατηγορία	*Μοναδικές λέξεις κατηγορίας*	*Μοναδικές λέξεις που ελέγχθηκαν*	*% Μοναδικών λέξεων*	*Συνολικές λέξεις κατηγορίας*	*Συνολικές λέξεις που ελέγχθηκαν*	*% Συνολικών λέξεων*
Γνωστή Ελληνική λέξη (Λεξικό)	109.900	5.489	5,00%	1.925.797	91.240	4,74%
Ξένη λέξη (μη Greeklish)	20.330	2.031	10,00%	20.330	15.971	11,07%
Άγνωστη Greeklish λέξη	14.710	1.470	10,00%	22.112	2.161	9,77%
Υπόλοιπες (κολλημένες λέξεις μεταξύ τους, ανορθόγραφες κ.λπ.)	661	-	-	2.858	-	-

Πίνακας 14: Δείγμα λέξεων που ελέγχθηκαν χειρωνακτικά για την αξιολόγηση του συστήματος.

Όπως έχουμε ήδη αναφέρει, ο διαχωρισμός των λέξεων του σώματος κειμένου στις παραπάνω κατηγορίες μας διευκολύνει επειδή παρουσιάζουν ιδιαίτερα χαρακτηριστικά μεταξύ τους, αλλά και

γιατί είναι ευκολότερο να καθορίσουμε τους διαφορετικούς τύπους λαθών που ενδεχομένως εμφανίζονται.

Πιο συγκεκριμένα, για την πρώτη κατηγορία, ορίστηκαν τέσσερεις διαφορετικές μη επικαλυπτόμενες υποκατηγορίες, στις οποίες κάθε λέξη έπρεπε να αντιστοιχιστεί μία και μόνο φορά. Κατά αυτόν τον τρόπο, μία λέξη της πρώτης κατηγορίας θα μπορούσε να χαρακτηριστεί α) είτε ως σωστά μετεγγραμμένη, β) είτε ως λάθος μετεγγραμμένη λόγω εσφαλμένης καταχώρησης στο λεξικό, γ) είτε ως λάθος μετεγγραμμένη λόγω εσφαλμένου χειρισμού της μηχανής, γ) είτε ως ομόηχης λέξης Greeklish ή λέξης που υπάρχει και όμοια της γραμμένη και στην αγγλική γλώσσα. Ομόηχη λέξη Greeklish χαρακτηρίζουμε μία λέξη Greeklish η οποία μπορεί να προέλθει από περισσότερες από μία διαφορετικές ελληνικές λέξεις όπως π.χ. είναι η αναπαράσταση /poli/ που μπορεί να προέρχεται σχεδόν ισοπίθανα από τις λέξεις /πολύ/, /πολλοί/, /πολλή/, /πόλη/. Οι περιπτώσεις αυτές δεν μπορούν να χαρακτηρισθούν ως εσφαλμένες μετατροπές αφού η παραγόμενες ελληνικές λέξεις είναι σωστές, ωστόσο είναι φανερό ότι η ορθότητα των λέξεων αυτών εξαρτάται από το περικείμενο της λέξης και το νόημα που επιχειρεί να αποδώσει.

Στον πίνακα που ακολουθεί φαίνονται τα αποτελέσματα των μετρήσεων μας για την πρώτη κατηγορία.

	Σύνολο	**Σωστή Μεταγραφή**	**Αφηρημένη Greeklish ή όμοια ξένη**	**Λάθος Χειρισμός της μηχανής**	**Λάθος εγγραφή στο λεξικό**
Μοναδικές λέξεις	5.489	5.031	261	7	190
Μοναδικές λέξεις*Συχνότητα	**91.240**	**78.658**	**12.102**	**15**	**465**
% της πρώτης κατηγορίας	**4,74%**	**86,21%**	**13,26%**	**0,02%**	**0,51%**
% όλων των λέξεων	**91,92%**	**79,25%**	**12,19%**	**0,02%**	**0,47%**

Πίνακας 15: Αποτελέσματα αξιολόγησης του συστήματος για την πρώτη κατηγορία λέξεων.

Παρατηρούμε ότι στην συγκεκριμένη κατηγορία λέξεων το σύστημα παρουσιάζει πολύ καλή συμπεριφορά. Το ποσοστό των λαθών της μηχανής είναι πολύ μικρό, ενώ τα λάθη που προέρχονται από το ενσωματωμένο λεξικό είναι λιγότερα από το 0,5% των περιπτώσεων. Οι λάθος καταχωρήσεις στο λεξικό οφείλονται στο γεγονός ότι το λεξικό αυτό προέρχεται από σώμα

κειμένου που αποκτήθηκε μέσα από το διαδίκτυο, με σημαντική ωστόσο επεξεργασία για την απαλοιφή λαθών.

Ομοίως, για την δεύτερη κατηγορία των λέξεων ορίσαμε επτά διαφορετικές ομάδες λέξεων, στις οποίες κάθε λέξη της κατηγορίας αυτής μπορεί να αποδοθεί μοναδικά. Αυτές είναι οι εξής:

1. λέξη Greeklish που δεν μετατράπηκε για κάποιο λόγο
2. ανορθόγραφη λέξη Greeklish
3. μη Greeklish λέξη (π.χ. ξένη λέξη ή ακρώνυμο)
4. αριθμοί
5. Greeklish λέξη που όμως έχει την ίδια αναπαράσταση με γνωστή ξένη λέξη
6. Greeklish λέξεις που όμως αντιστοιχούν σε λέξεις της «αργκό» καθομιλουμένης
7. παράγωγα ξένων λέξεων που συναντιόνται συχνά σε ομάδες συζητήσεων στο διαδίκτυο

Από τις ομάδες αυτές λέξεων ουσιαστικά μόνο η πρώτη ομάδα μπορεί να θεωρηθεί ως λάθος του συστήματος, ενώ οι υπόλοιπες μπορούν να θεωρηθούν ότι σωστά δεν επιχειρήθηκε να μετατραπούν σε ελληνικά.

Στο ακόλουθο πίνακα φαίνονται συγκεντρωτικά τα αποτελέσματα των μετρήσεων μας για την κατηγορία αυτή.

	Σύνολο	**Greeklish Word**	**Ανορθόγραφο Greeklish**	**μη Greeklish λέξη**	**Αριθμοί**	**Ξενες και ελληνικες με την ιδια αναπαρασταση**	**Λέξεις «αργκό»**	**Chat ή Φορουμ λέξεις**
Μοναδικές λέξεις	2.031	26	237	1.499	145	45	64	15
Μοναδικές λέξεις*Συχνότητα	**15.971**	**109**	**368**	**11.610**	**3.225**	**541**	**96**	**22**
% της δεύτερης κατηγορίας	**11,07%**	**0,68%**	**2,30%**	**72,69%**	**20,19%**	**3,39%**	**0,60%**	**0,14%**
% όλων των λέξεων	**6,89%**	**0,05%**	**0,16%**	**5,01%**	**1,39%**	**0,23%**	**0,04%**	**0,01%**

Πίνακας 16: Αποτελέσματα αξιολόγησης του συστήματος για την δεύτερη κατηγορία λέξεων.

Όσον αφορά στην τρίτη κατηγορία λέξεων, τις άγνωστες λέξεις Greeklish που μεταγράφηκαν φωνητικά λόγω του γεγονότος ότι δεν υπάρχουν αντίστοιχες ελληνικές λέξεις στο ενσωματωμένο λεξικό, ορίσαμε έξι διαφορετικές ομάδες λέξεων, οι οποίες είναι οι εξής:

1. Greeklish λέξη η οποία όμως δεν έχει αντίστοιχη ελληνική στο λεξικό
2. ανορθόγραφα γραμμένη Greeklish λέξη
3. ανορθόγραφη ξένη λέξη
4. γνωστή ξένη λέξη
5. Greeklish λέξεις που όμως αντιστοιχούν σε λέξεις της «αργκό» καθομιλουμένης
6. παράγωγα ξένων λέξεων που συναντιόνται συχνά σε ομάδες συζητήσεων στο διαδίκτυο

Από τις ομάδες αυτές ουσιαστικά η πρώτη και η τέταρτη αποτελούν λάθη ή ελλείψεις του συστήματος. Στην πρώτη περίπτωση ως ελλείψεις στο λεξικό, στην δεύτερη ως λάθος αναγνώριση γλώσσας (Greeklish αντί ξένης λέξης).

Στο παρακάτω πίνακα φαίνονται τα αποτελέσματα συγκεντρωτικά.

	Σύνολο	**Greeklish λέξη/ έλλειψη στο λεξικό**	**Ανορθόγραφο Greeklish**	**Ανορθόγραφη ξένη λέξη**	**Ξένη λέξη**	**Λέξεις «αργκό»**	**Chat ή Φορουμ λέξεις**
Μοναδικές λέξεις	1.470	52	949	200	24	227	18
Μοναδικές λέξεις*Συχνότητα	**2.161**	**98**	**1.297**	**326**	**99**	**315**	**26**
% της δεύτερης κατηγορίας	**9,77%**	**4,53%**	**60,02%**	**15,09%**	**4,58%**	**14,58%**	**1,20%**
% όλων των λέξεων	**1,06%**	**0,05%**	**0,63%**	**0,16%**	**0,05%**	**0,15%**	**0,01%**

Πίνακας 17: Αποτελέσματα αξιολόγησης του συστήματος για την τρίτη κατηγορία λέξεων.

Λαμβάνοντας υπόψη τις κατηγορίες που αναφέραμε παραπάνω ως μοναδικά λάθη της μηχανής αυτόματης μετατροπής και προβάλλοντας ταυτόχρονα σε ολόκληρο το δείγμα τα αντίστοιχα ποσοστά, προκύπτει ότι το συνολικό ποσοστό σφάλματος του συστήματος είναι 0,63% ή διαφορετικά, το ποσοστό επιτυχίας του συστήματος είναι 99,37%. Ωστόσο, στο σημείο αυτό είναι αναγκαίο να τονίσουμε ότι το ποσοστό αυτό δεν είναι απόλυτα σωστό, αφού δεν λαμβάνουμε υπόψη τις Greeklish λέξεις που έχουν περισσότερες από μία διαφορετικές απεικονίσεις, όπως παρουσιάστηκε στην πρώτη κατηγορία λέξεων. Το πλήθος αυτών των λέξεων υπολογίστηκε ότι φτάνει το 12,19% των συνολικών λέξεων. Το ποσοστό αυτό είναι ιδιαίτερα μεγάλο και για τον λόγο αυτό θεωρείται σκόπιμη η ιδιαίτερη αντιμετώπιση της κατηγορίας αυτής από το σύστημα. Αυτό μπορεί να επιτευχθεί με την ενσωμάτωση ενός απλοϊκού γλωσσικού μοντέλου στο στάδιο

της επεξεργασίας των τελικών αποτελεσμάτων. Ένα τέτοιο μοντέλο, χωρίς να επιβαρύνει σημαντικά τον υπολογιστικό φόρτο, θα μπορεί να λύσει παρόμοια προβλήματα.

6.7 Βιβλιογραφία Κεφαλαίου

[Androutsopoulos1999] Androutsopoulos, J. (1999). Latin-Greek orthography in electronic mails: use and instances. Paper presented at the 20th Annual Meeting of the Linguistics Department, 23-25 April 1999, Aristotle University of Thessaloniki.

[Androutsopoulos2000]Androutsopoulos, J. (2000). From dieuthinsi to diey8ynsh. Orthographic variation in Latin-alphabeted Greek]. 4th International Conference on Greek Linguistics, September 1999, University of Nicosia,

[Chalamandaris2006]Chalamandaris A., A. Protopapas, P. Tsiakoulis, & S. Raptis (2006). All Greek to me! An automatic Greeklish to Greek transliteration system. 5th International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy, 24–26 May. in proceedings, E. Hinrichs, N. Ide, M. Palmer, & J. Pustejovsky (Eds.), pp. 1226–1229.

[Chalamandaris2006] Chalamandaris A., A. Protopapas, P. Tsiakoulis, and S. Raptis. "All Greek to me! An automatic Greeklish to Greek transliteration system." *5th Int. Conf. on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy, pp. 1226–1229, 2006.

[Chalamandaris2004]Chalamandaris A., P. Tsiakoulis, S. Raptis and G. Giannopoulos, "An Efficient and Robust Algorithm for Bypassing Greeklish", in Proc. IC-SCCE 2004: 1st International Conference from Scientific Computing to Computational Engineering, 8-10 September, Athens, Greece (2004)

[Chalamandaris2004]Chalamandaris A., P. Tsiakoulis, S. Raptis, G. Giannopoulos and G. Carayannis, "Bypassing Greeklish!", in Proc. LREC 2004: 4th International Conference on Language Resources And Evaluation, May 26-28, Lisbon, Portugal (2004)

[Dunning1994] Dunning, T. (1994). Statistical Identification of Language. Technical report CRLMCCS-94-273. Computing Research Lab, NewMexico State University.

[ELOT1982] ELOT (1982). Greek Organisation of Standardization

[Karakos2003] Karakos Alexandros (2003). Greeklish: An Experimental Interface for Automatic Transliteration. Journal Of The American Society For Information Science And Technology

[Koutsogiannis2003] Koutsogiannis, D. & Mitsikopoulou, B. (2003). Greeklish and Greekness: Trends and Discourses of 'Glocalness'. Proposal submitted the forthcoming special issue of Journal of Computer-Mediated Communication on "The Multilingual Internet"

[Lyras2010] Lyras, D., Kotinas, I., Sgarbas, K., Fakotakis, N., "A Stochastic Greek-to-Greeklish Transcriber Modeled by Real User Data, Artificial Intelligence: Theories, Models and Applications, Lecture Notes in Computer Science - 2010 Springer Berlin / Heidelberg

[Tseliga2003] Tseliga, T. (2003). A corpus-based study of discourse features in Roman-alphabeted Greek (i.e. Greeklish) emails. 1st International Conference on Internet and Language, Castellon, Spain, 18-20 September.

[Tseliga2003] Tseliga, T. and Marinis, T. (2003). On-line processing of Roman-alphabeted Greek: the influence of morphology in the spelling preferences of Greeklish. 6th International Conference in Greek Linguistics, Rethymno, Crete, 18-21 September, 2003.

[Converters2010]:http://www.translatum.gr/converter/greeklishconverter.htm

7. ΣΥΣΤΗΜΑ ΣΥΝΘΕΤΗΣ ΦΩΝΗΣ ΜΕ ΈΜΦΑΣΗ ΣΕ ΘΕΜΑΤΑ ΠΡΟΣΒΑΣΙΜΟΤΗΤΑΣ

Στο συγκεκριμένο κεφάλαιο περιγράφουμε όλες τις διαδικασίες που ακολουθήσαμε για την προσαρμογή του συστήματος σύνθεσης φωνής που περιγράφουμε για την χρήση του σε περιβάλλοντα επαυξημένης προσβασιμότητας. Πιο συγκεκριμένα, παρουσιάζουμε την διαδικασία σχεδιασμού και την συλλογή των απαραίτητων δεδομένων για την προδιαγραφή των αναγκών του χρήστη. Στην συνέχεια περιγράφουμε τα επιμέρους πεδία όπου απαιτήθηκε να προσαρμόσουμε επιμέρους υποσυστήματα αλλά και αλγορίθμους, έτσι ώστε να ικανοποιήσουμε τις προδιαγραμμένες ανάγκες χρήστη. Στην τελευταία ενότητα του κεφαλαίου παρουσιάζουμε τα αποτελέσματα μίας εκτενούς διαδικασίας επικύρωσης αλλά και αξιολόγησης του συστήματος ως βοηθήματος σε θέματα προσβασιμότητας, τα αποτελέσματα της οποίας επιβεβαιώνουν την

καταλληλότητα τόσο του σχεδιασμού όσο και των επιμέρους προσαρμογών που πραγματοποιήσαμε στο αρχικό σύστημα σύνθεσης ομιλίας.

7.1 Εισαγωγή

Η σημερινή τεχνολογία συνθετικής ομιλίας, έχοντας επιτύχει σημαντική πρόοδο στην φυσικότητα και στην ευκρίνεια και καταληπτότητα του συνθετικού λόγου, αλλά και σε συνδυασμό με την αύξηση της υπολογιστικής ισχύος που παρέχεται από τους υπολογιστές και τις φορητές συσκευές, έχει αποτελέσει σημαντικό εργαλείο για πολλές καθημερινές δραστηριότητές μας, με εφαρμογές από την ρομποτική, μέχρι τους σύγχρονους πλοηγούς GPS ή τους αυτόματους μεταφραστές τσέπης [Bailly2003]. Από την αρχή της γέννησής της, η τεχνολογία σύνθεσης φωνής αποτέλεσε μία φυσικότερη διεπαφή για την επικοινωνία του ανθρώπου με τους υπολογιστές, ενώ αποτέλεσε και αποτελεί ακόμη και σήμερα τον ακρογωνιαίο λίθο της προσβασιμότητας των υπολογιστών από άτομα με μειωμένη όραση.

Οι πρώτες τεχνολογίες σύνθεσης φωνής που υπήρξαν διαθέσιμες προς τελικούς χρήστες [O'Shaughnessy2007] ως βοήθημα ανάγνωσης οθόνης, υπέφεραν από έντονη ρομποτική ποιότητα με χαμηλή συχνά καταληπτότητα, ενώ συχνά οι υπολογιστικές απαιτήσεις τους δεν συμβάδιζαν με τους υπολογιστές που μπορούσαν να αγοράσουν τελικοί χρήστες. Σήμερα, τόσο η διαθέσιμη υπολογιστική ισχύς, όσο και η υψηλή ποιότητα της τεχνολογίας της σύνθεσης φωνής, έπαψαν να είναι πολυτέλεια για λίγους και ένας μέσος ή ακόμα και παλαιότερης γενιάς υπολογιστής τελικού χρήστη, μπορεί να εκτελέσει με άνεση οποιοδήποτε λογισμικό ανάγνωσης οθόνης.

Σήμερα, αν και θα νόμιζε κανείς ότι η φυσικότητα και καταληπτότητα ενός συνθέτη φωνής είναι τα βασικά και ίσως ακόμη και μοναδικά χαρακτηριστικά που αποζητούν οι χρήστες με μειωμένη όραση, εκτενής έρευνες έχουν δείξει ότι κάτι τέτοιο δεν είναι αληθές [Dutoit2008]. Αντίθετα όμως, όπως κάθε λογισμικό που προορίζεται για τελικούς χρήστες, έτσι και το συγκεκριμένο λογισμικό χρήζει ιδιαίτερων απαιτήσεων σχεδιασμού και ανάπτυξης. Πολύ δε περισσότερο στην συγκεκριμένη περίπτωση όπου οι τελικοί χρήστες παρουσιάζουν ιδιαίτερες ανάγκες και ο τρόπος χρήσης του λογισμικού και της τεχνολογίας διαφέρει σημαντικά από άλλες εφαρμογές.

Στο συγκεκριμένο κεφάλαιο παρουσιάζουμε την μελέτη που εκπονήσαμε σχετικά με την καταγραφή των απαιτήσεων χρήσης και σχεδιασμού για έναν συνθέτη φωνής ως βοήθημα, τις

απαραίτητες αλλαγές και προσθήκες που σχεδιάσαμε και αναπτύξαμε στο ίδιο πλαίσιο και τέλος την αξιολόγηση του τελικού συστήματος, τόσο από εξειδικευμένους επιστήμονες, όσο και από ομάδα τελικών χρηστών με προβλήματα όρασης.

7.2 Απαιτήσεις χρήστη

Όπως κάθε λογισμικό που προορίζεται για χρήση από συγκεκριμένες ομάδες χρηστών, έτσι και στην περίπτωσή μας, ήταν απαραίτητη από την αρχή της διαδικασίας η καταγραφή των ειδικών χαρακτηριστικών της ομάδας-στόχου, καθώς επίσης και οι απαιτήσεις που ορίζουν οι χρήστες αυτοί στον σχεδιασμό του συνθέτη φωνής [Hackos1997] [Chalamandaris et al., 2010].

Καθώς η χρήση του συγκεκριμένου συστήματος φωνής επρόκειτο να εκτελείται τοπικά από τελικούς χρήστες μέσω των προσωπικών τους υπολογιστών, τα χαρακτηριστικά του λογισμικού θα πρέπει να είναι τέτοια ώστε να πληροί τις τεχνικές προδιαγραφές που ορίζει ο μέσος υπολογιστή τέτοιου χρήστη [Earl1999]. Τα χαρακτηριστικά του μέσου υπολογιστή προέκυψαν από ερωτηματολόγια που τέθηκαν υπόψη δυνητικών χρηστών και περιλάμβαναν ερωτήσεις αναφορικά με το τρέχοντα υπολογιστή, την συχνότητα αναβάθμισης υπολογιστή καθώς επίσης και τα προγράμματα ανάγνωσης οθόνης που κατά κύριο λόγο χρησιμοποιούν στους υπολογιστές αυτούς.

Στην συνέχεια, ομάδα αποτελούμενη από 6 άτομα-δυνητικούς χρήστες, παρείχαν μέσω ερωτηματολογίου αλλά και μέσω απαντήσεων ελεύθερου κειμένου, πληροφορίες αναφορικά με τα χαρακτηριστικά που θεωρούν σημαντικότερα, όσον αφορά το λογισμικό σύνθεσης φωνής σε ρόλο βοηθήματος με αναγνώστη οθόνης. Από τις απαντήσεις προέκυψαν τρεις βασικές πτυχές του συνθέτη φωνής που απαιτείτο να προσαρμοστούν ανάλογα έτσι ώστε να εκπληρώσουν τις ιδιαίτερες ανάγκες του περιβάλλοντος χρήσης του: α) προσαρμοσμένη επεξεργασία φυσικής γλώσσας, β) βελτίωση της ταχύτητας απόκρισης και εκτέλεσης και γ) βελτίωση της ποιότητας σε διαφορετικά επίπεδα, όπως ήταν ο χειρισμός πολυγλωσσικών κειμένων και η ευκρίνεια του συνθετικού λόγου σε εξαιρετικά υψηλές ταχύτητες εκτέλεσης. Στις παραγράφους που ακολουθούν περιγράφουμε αναλυτικά τις προσαρμογές αυτές.

7.3 Ειδικές προσαρμογές συνθέτη φωνής

7.3.1 Προσαρμογή επεξεργασίας φυσικής γλώσσας

Μία από τις σημαντικότερες απαιτήσεις χρήστη [Chalamandaris et al., 2009c], όπως αυτές προέκυψαν, ήταν η δυνατότητα του συνθέτη φωνής:

α) να επεξεργάζεται και να χειρίζεται με επιτυχία γραπτό κείμενο που περιέχει όχι μόνο Ελληνικούς χαρακτήρες, αλλά μικτό,

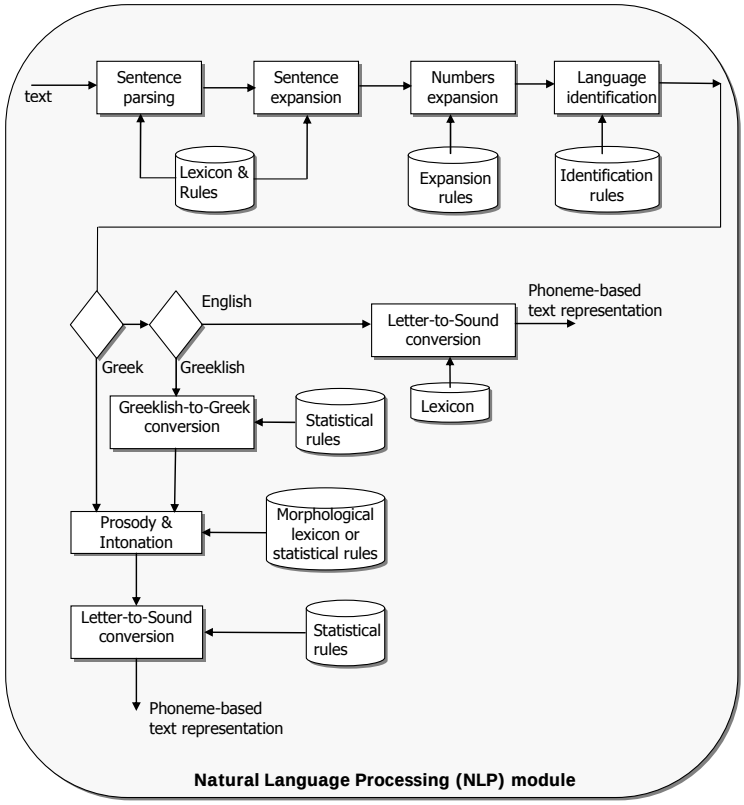
β) η δυνατότητα να αναγνωρίζει και να χειρίζεται συναισθηματεικονίδια (emoticons)[9] και

γ) η δυνατότητα αυτόματης διόρθωσης κατά την εκφώνηση κοινών τυπογραφικών λαθών που συναντώνται σε αναρτήσεις κυρίως του διαδικτύου όπου συχνά σε μία μη Ελληνική λέξη υπήρχαν Ελληνικοί χαρακτήρες.

Το τελευταίο φαινόμενο συνήθως οφείλεται στην συνειδητοποίηση του συγγραφέα ότι δεν έχει επιλέξει την κατάλληλη γλώσσα για πληκτρολόγηση αφού έχει πληκτρολογήσει κάποια γράμματα της λέξης που όμως γράφονται με τον ίδιο τρόπο στα Ελληνικά και στα Λατινικά. Η λέξη π.χ. "Nokia" πολύ συχνά στο Ελληνικό διαδίκτυο εμφανίζεται γραμμένη με τους δύο πρώτους χαρακτήρες στα Ελληνικά και τους υπόλοιπους με λατινικούς χαρακτήρες.

Όσον αφορά στον χειρισμό μικτών κειμένων, ως βασική προσθήκη στο υποσύστημα φυσικής επεξεργασίας του κειμένου εντάχθηκε η τεχνολογία μετατροπής των Greeklish σε Ελληνικά, με ενσωματωμένο το υποσύστημα αναγνώρισης γλώσσας. Τα υποσυστήματα αυτά έχουν περιγραφεί αναλυτικά στο κεφάλαιο 5 της συγκεκριμένης διατριβής. Κατά αυτόν τον τρόπο, το κείμενο εισόδου διέρχεται δι' ενός ενδιαμέσου σταδίου όπου πραγματοποιείται αναγνώριση και επεξεργασία των κειμένων που περιέχουν λέξεις ή φράσεις σε Greeklish [Chalamandaris2005], και κανονικοποίηση αυτών. Στο παρακάτω σχήμα φαίνεται το λογικό διάγραμμα ροής του υποσυστήματος επεξεργασίας φυσικής γλώσσας όπως αυτό προέκυψε μετά την συγκεκριμένη προσαρμογή.

[9] Emoticons είναι αναπαραστάσεις εκφράσεων προσώπου και συναισθημάτων με χρήση σημείων στίξης. Π.χ. ο συνδυασμός :) αναπαριστά το χαμόγελο και την εύθυμη διάθεση.



Σχήμα 65: Σχηματική απεικόνιση του υποσυστήματος NLP που πραγματοποίηθηκε στο πλαίσιο προσαρμογής του συνθέτη φωνής για περιβάλλοντα ανάγνωσης οθόνης.

Αναφορικά με τα συναισθηματεικονίδια (emoticons) καθώς επίσης και συντομογραφίες που χρησιμοποιούνται στην καθομιλουμένη του διαδικτύου σε διαλόγους, τα συνηθέστερα από αυτά ανιχνεύθηκαν και ενσωματώθηκαν μέσω λεξικού στο υποσύστημα κανονικοποίησης κειμένου του συνθέτη φωνής, έτσι ώστε ο τελικός χρήστης να λαμβάνει την πληροφορία αυτή ακουστικά. Ενώ για το φαινόμενο των ξένων λέξεων που περιέχουν Ελληνικά γράμματα, ολοκληρώθηκε υποσύστημα το οποίο αναγνώριζε τέτοιες περιπτώσεις, λαμβάνοντας υπόψη το σύνολο των Ελληνικών γραμμάτων που είναι ομόγραφα με τα αντίστοιχα του λατινικού αλφαβήτου, και στην συνέχεια μετέτρεπε τις συγκεκριμένες συμβολοσειρές στις αντίστοιχες έγκυρες λέξεις πριν την εκφώνηση από το σύστημα.

7.3.2 Βελτιστοποίηση ταχύτητας και απόκρισης συστήματος

Η ταχύτητα εκτέλεσης αλλά και απόκρισης του συστήματος σύνθεσης φωνής παραμένει μία από τις σημαντικότερες παραμέτρους για τους χρήστες λογισμικών ανάγνωσης οθόνης [Chalamandaris et al., 2009c] [Earl1999]. Φάνηκε ότι είναι εξαιρετικά σημαντικό ο χρήστης να ακούει σχεδόν αμέσως την πληροφορία για το κουμπί που πάτησε, ενώ οποιαδήποτε καθυστέρηση στην απόκριση ήταν αρκετή για να «κουράσει» τον χρήστη γρήγορα. Ήταν μεγάλη έκπληξη για εμάς όταν κατά την διαδικασία εξαγωγής των απαιτήσεων χρήστη, προτάθηκε από μέλη της ομάδας χρηστών η μείωση της ποιότητας της συνθετικής ομιλίας ώστε να εκτελείται ταχύτερα και αμεσότερα.

Για την βελτίωση της ταχύτητας εκτέλεσης και απόκρισης του συνθέτη φωνής, ακολουθήσαμε δύο παράλληλες τεχνικές:

α) την μείωση του μεγέθους της βάσης δεδομένων του συνθέτη [Tsiakoulis2008] με τρόπο που εξασφάλιζε υψηλή ποιότητα συνθετικής ομιλίας και

β) την ενσωμάτωση και προ-σύνθεση των συνηθέστερων λέξεων και φράσεων που χρησιμοποιούνται κατά την ανάγνωση της οθόνης του υπολογιστή.

Η μείωση του μεγέθους της βάσης δεδομένων πραγματοποιήθηκε με βάση προηγούμενη ερευνητική εργασία που εκτελέστηκε από την ομάδα σύνθεσης φωνής και στόχο είχε την αποδοτική σταχυολόγηση ακουστικών μονάδων από μία μεγάλη βάση δεδομένων, έτσι ώστε να μεταφερθεί το σύστημα σύνθεσης σε μικρότερα και υπολογιστικά ελαφρύτερα περιβάλλοντα, όπως είναι το κινητό τηλέφωνο [Tsiakoulis2008]. Ο αλγόριθμος αυτός βασίζεται στην επιλογή από την βάση ακουστικών μονάδων (διφωνημάτων) που καλύπτουν με επάρκεια το πλήθος των χαρακτηριστικών που διακρίνουν διαφορετικές πραγματώσεις ίδιων διφωνημάτων. Με την μέθοδο αυτήν, επιτύχαμε να ελαττώσουμε το μέγεθος της αρχικής βάσης δεδομένων κατά 35% περίπου, γεγονός που βελτίωσε τόσο την ταχύτητα εκτέλεσης όσο και τις απαιτήσεις σε χωρητικότητα από την τελική εφαρμογή.

Λαμβάνοντας δεδομένα κατά την χρήση λογισμικών ανάγνωσης οθόνης από τους χρήστες που μας βοήθησαν στην συγκεκριμένη εργασία, καταρτίσαμε μία λίστα από λέξεις και σύντομες

φράσεις που συναντούσε ο χρήστης συχνά. Τέτοιες λέξεις π.χ. είναι τα γράμματα του Ελληνικού και του Αγγλικού αλφαβήτου (κατά την φωνητική υποστήριξη της πληκτρολόγησης), οι επιλογές των συχνά χρησιμοποιούμενων προγραμμάτων, καθώς επίσης και συχνά εμφανιζόμενα μηνύματα από το λειτουργικό σύστημα. Έχοντας συλλέξει τα παραπάνω, ενσωματώσαμε στον συνθέτη φωνής προ-συντεθημένα ακουστικά αρχεία που κατά την εκτέλεση ο συνθέτης δεν χρειαζόταν να συνθέσει με τον κλασσικό τρόπο αλλά αρκούταν στο να τα αναπαράγει, προσαρμόζοντας παράλληλα τα προσωδιακά χαρακτηριστικά των ήχων αυτών ανάλογα με τις ρυθμίσεις του χρήστη. Η συγκεκριμένη τεχνική αποδείχτηκε ιδιαίτερα αποδοτική τόσο στην ταχύτητα εκτέλεσης, όσο και στην ταχύτητα απόκρισης με αποτέλεσμα ο συνθέτης φωνής να ανταποκρίνεται στα συγκεκριμένα ερεθίσματα του χρήστη κατά μέσο όρο κατά 2,4 φορές ταχύτερα από ό,τι πριν.

7.3.3 Πολυγλωσσικότητα και ακουστική ποιότητα

Ένα από τα σημαντικότερα σημεία που εντοπίστηκαν κατά την καταγραφή των απαιτήσεων χρήστη ήταν η ανάγκη για πολυγλωσσικότητα και η απαίτηση για συνθετικό λόγο υψηλής ευκρίνειας ακόμα και σε εξαιρετικά υψηλές ταχύτητες ανάγνωσης [Chalamandaris et al., 2010].

Το χαρακτηριστικό της πολυγλωσσικότητας, στο περιβάλλον ενός αναγνώστη οθόνης, έχει νόημα για τον συνθέτη φωνής αφού συχνά ο αναγνώστης οθόνης καλείται να διαβάσει στον χρήστη επιλογές, μενού ή ακόμη και ολόκληρα κείμενα τα οποία είναι γραμμένα στα Αγγλικά και δεν υπάρχουν διαθέσιμα στα Ελληνικά. Εξάλλου πολύ συχνή είναι η περίπτωση όπου σε ένα αμιγώς Ελληνικό κείμενο υπάρχουν όροι, λέξεις ή ακόμη και φράσεις γραμμένες στα Αγγλικά. Το τελευταίο παράδειγμα αποτελεί κανόνα πλέον σε αναρτήσεις του Ελληνικού διαδικτύου.

Λαμβάνοντας τα παραπάνω υπόψη, απαίτηση των χρηστών ήταν η ικανότητα του συνθέτη φωνής να μπορεί να διαβάσει με καταληπτότητα και με παρόμοια, αν όχι την ίδια φωνή, συμβολοσειρές γραμμένες στα Αγγλικά. Η συνηθέστερη πρακτική για την αντιμετώπιση του συγκεκριμένου προβλήματος είναι η απεικόνιση των φωνημάτων της μίας γλώσσας προς τα φωνήματα της γλώσσας του συνθέτη φωνής. Η τεχνική αυτή αποδίδει υποτυπωδώς, ενώ συχνά, εκτός από την γενικότερα χαμηλή ποιότητα που προσφέρει, οδηγεί σε αποτελέσματα με χαμηλή καταληπτότητα. Αυτό εξηγείται εύκολα αν λάβει κανείς υπόψη του το γεγονός ότι ειδικά στην Αγγλική γλώσσα, το πλήθος αλλά και ο συνδυασμός των φωνημάτων είναι σημαντικά μεγαλύτερος

και με εντελώς διαφορετικές κατά μέσο όρο διάρκειες σε σχέση με τα αντίστοιχα Ελληνικά φωνήματα. Κατά αυτόν τον τρόπο συχνοί και σύντομοι σε διάρκεια δίφθογγοι είναι δύσκολο να αναπαραχθούν μέσω Ελληνικών π.χ. φωνημάτων που έχουν εκφωνηθεί σε αμιγώς Ελληνικό σώμα κειμένου.

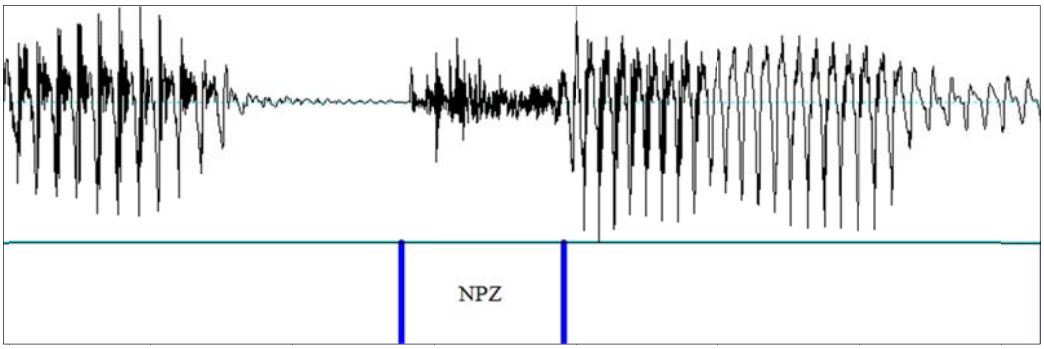
Για να αντιμετωπίσουμε με επιτυχία το συγκεκριμένο ζήτημα αποφασίσαμε ότι θα έπρεπε ο συνθέτης φωνής να μπορεί εξ αρχής να αναπαράγει τους διαφορετικούς αυτούς ήχους και να μην αρκείται σε απλή απεικόνιση τους σε όμοιους από άλλη γλώσσα. Για τον λόγο αυτόν, δημιουργήσαμε με τον ίδιο ομιλητή ένα επιπλέον ηχογραφημένο σώμα κειμένου, το οποίο ήταν αρκετό για να προσφέρει την μέγιστη δυνατή κάλυψη σε φωνήματα της Αγγλικής γλώσσας. Το σώμα κειμένου που επιλέξαμε προέρχεται από το σώμα κειμένου ARCTIC όπως αυτό δημοσιεύθηκε στο [Black2003]. Ζητήθηκε από τον ομιλητή να εκφωνήσει το συγκεκριμένο σώμα κειμένου χωρίς να προσπαθεί να μιμηθεί την αγγλική προφορά ώστε να μην παρουσιαστούν υπερβολές στον λόγο, ενώ για να διευκολυνθεί η συγκεκριμένη εργασία, ο ομιλητής άκουγε τα αντίστοιχα ηχητικά δεδομένα, εκφωνημένα από Άγγλο ομιλητή, έτσι ώστε να διασφαλιστεί η σωστή ανάγνωση και προφορά των λέξεων. Φυσικά η συγκεκριμένη στρατηγική δεν περιορίζεται μόνο στην ηχογράφηση ενός επιπλέον σώματος κειμένου για τα Αγγλικά, αλλά απαιτεί και σχεδίαση και ανάπτυξη επιπλέον εργαλείων τα οποία εξαρτώνται από την γλώσσα, όπως είναι η φωνητική μεταγραφή για την Αγγλική γλώσσα και η κανονικοποίηση κειμένου, επίσης για την Αγγλική γλώσσα. Το επίπεδο της κανονικοποίησης κειμένου για τα Αγγλικά δεν απαιτείτο να αναπτυχθεί πλήρως, αλλά μέχρι ενός συγκεκριμένου επιπέδου, όπου ο συνθέτης φωνής θα μπορούσε να διαβάσει με επιτυχία συντομογραφίες και συντμήσεις με Λατινικούς χαρακτήρες.

Είναι προφανές ότι η συγκεκριμένη τεχνική απαιτεί ιδιαίτερη προσπάθεια, τόσο από την πλευρά της ανάπτυξης του συστήματος, όσο και από την πλευρά του ίδιου του ομιλητή, από τον οποίο ζητείται η ηχογράφηση σε άλλη γλώσσα από την μητρική του. Παρ' όλα αυτά, τα αποτελέσματα της αξιολόγησης του συστήματος έδειξαν ότι η συγκεκριμένη προσπάθεια βελτίωσε σημαντικά την χρηστικότητα του συστήματος στο πλαίσιο ενός αναγνώστη φωνής.

Όσον αφορά την ποιότητα του συνθετικού λόγου κατά την εκτέλεση με εξαιρετικά υψηλές ταχύτητες εκφώνησης [Picheny1986], παρατηρήσαμε αρχικά ότι οι τα άτομα με προβλήματα

όρασης τείνουν να χρησιμοποιούν τον συνθέτη φωνής σε ταχύτητες εκφώνησης όπου ο λόγος συχνά είναι ακατάλληπτος για πολλούς. Η ταχύτητα αυτή εξυπηρετεί ιδιαίτερα όταν ο χρήστης διατρέχει τους τίτλους των επιλογών στα παράθυρα των εφαρμογών αλλά και όταν θέλει να διατρέξει γρήγορα ένα κείμενο. Παρατηρήσαμε ότι η ταχύτητα εκφώνησης που χρησιμοποιούσαν οι χρήστες ήταν συχνά μεγαλύτερη από 3 φορές της κανονικής ταχύτητας των φυσικών ηχογραφήσεων, με αποτέλεσμα η καταληπτότητα του λόγου συχνά να υποβαθμίζεται.

Λαμβάνοντας υπόψη τα χαρακτηριστικά του καθαρού λόγου (clear speech) σε σχέση με τα χαρακτηριστικά του διαλογικού (conversational), εξετάσαμε και παρατηρήσαμε επίσης ότι το επίπεδο της καταληπτότητας του λόγου μειωνόταν αισθητά με την απουσία των χαρακτηριστικών «εκρήξεων» στα εκρηκτικά σύμφωνα. Για να διασφαλίσουμε υψηλή ποιότητα του συνθετικού ήχου ακόμα και σε εξαιρετικά υψηλές ταχύτητες, ορίσαμε περιοχές στο σήμα φωνής, οι οποίες θεωρήθηκαν απαραίτητες κατά την χρονική συμπίεση του σήματος φωνής, απαγορεύοντας από τον αλγόριθμο της TD-PSOLA να αφαιρέσει pitch periods που περιέχονται σε αυτές τις περιοχές. Αυτές οι περιοχές κυρίως περιλαμβάνουν τις «εκρήξεις» των εκρηκτικών συμφώνων /p/ /t/ /k/ /g/ /b/ /d/, καθώς επίσης και το σύντομο ένρινο φώνημα /r/. Κατά αυτόν τον τρόπο, επιτύχαμε να εξασφαλίσουμε υψηλή καταληπτότητα του συνθετικού λόγου σε σχέση με τις συμβατικές μεθόδους χρονικής συμπίεσης σήματος φωνής.



Σχήμα 66: Αναπαράσταση εκρηκτικού ήχου στο σήμα φωνής. Επισημειώνεται η περιοχή όπου πραγματοποιείται η "έκρηξη" του φωνήματος και χρησιμοποιείται ως περιοχή όπου δεν επιτρέπεται η επεξεργασία της διάρκειας του ήχου.

Η μέθοδος αυτή αποδείχθηκε ότι συνεισφέρει σημαντικά στο παραπάνω φαινόμενο, όπως άλλωστε φάνηκε και από τα πειράματα αξιολόγησης του συστήματος.

7.4 Αξιολόγηση και επικύρωση αποτελεσμάτων

Ένα απαραίτητο και εξίσου σημαντικό με τα προηγούμενα στάδια σχεδίασης και ανάπτυξης του συγκεκριμένου λογισμικού είναι αυτό της αξιολόγησης και της επικύρωσης των αποτελεσμάτων της φάσης των απαιτήσεων χρήστη [Redish2003]. Η αξιολόγηση του συστήματος περιλάμβανε δύο βασικές πτυχές του συστήματος, την ποιότητα των ακουστικών αποτελεσμάτων και το επίπεδο της χρηστικότητας του συστήματος όσον αφορά τον σκοπό για τον οποίο σχεδιάσθηκε: την εκτέλεση σε συνεργασία με έναν αναγνώστη οθόνης για άτομα με προβλήματα όρασης. Ο σχεδιασμός, η εκτέλεση και τα αποτελέσματα των συγκεκριμένων πειραμάτων περιγράφονται στις επόμενες παραγράφους.

7.4.1 Αξιολόγηση ακουστικής ποιότητας

Το πρώτο πείραμα αφορά στην αξιολόγηση της ακουστικής ποιότητας του συστήματος. Ειδικότερα, στην συγκεκριμένη διαδικασία εξετάσθηκαν διαφορετικές διαστάσεις που ορίζουν την ποιότητα ενός συνθέτη φωνής, όπως είναι η αξιολόγηση σε επίπεδο λέξης, σε επίπεδο πρότασης και σε επίπεδο παραγράφου [Heuven1995] [Alvarez2002]. Το καθένα από αυτά εκτελέστηκε με την βοήθεια 12 ατόμων (8 άνδρες και 4 γυναίκες), με προβλήματα όρασης, οι οποίοι είχαν εκτενή εμπειρία στην χρήση υπολογιστή και προγράμματα ανάγνωσης οθόνης. Για όλους τους χρήστες η μητρική γλώσσα ήταν η Ελληνική και η ηλικία τους κυμαινόταν από 26 έως 40 έτη.

7.4.1.1 Ακουστικά πειράματα αξιολόγησης

7.4.1.1.1 Πείραμα 1: Αξιολόγηση σε επίπεδο πρότασης

Ο σκοπός του συγκεκριμένου πειράματος είναι η αξιολόγηση της φυσικότητας του συνθετικού λόγου (δηλαδή πόσο «φυσική» ακούγεται η φωνή), η ευκολία με την οποία ακούγεται το αποτέλεσμα (η προσπάθεια που απαιτείται για να παρακολουθήσει και να κατανοήσει κανείς τον λόγο) και η ποιότητα της άρθρωσης/καταληπτότητα (πόσο καθαρά αρθρώνεται ο λόγος). Για την εκτέλεση του πειράματος, οι ακροατές άκουσαν 35 διαφορετικά ερεθίσματα μέσου μήκους (μέση διάρκεια 13 λέξεις ανά πρόταση). Ζητήθηκε από τους συμμετέχοντες στο πείραμα να βαθμολογήσουν τις τρεις προαναφερθείσες διαστάσεις με βαθμό από το 1 έως το 5, με ένα το

χειρότερο και 5 το καλύτερο. Προς διευκόλυνση των ερωτηθέντων, κάθε βαθμός αντιστοιχίσθηκε και σε έναν παραφραστικό χαρακτηρισμό, όπως αυτά φαίνονται στον πίνακα 15.

	Quality	**Ease of listening**	**Pleasantness**	**Understandability**	**Pronunciation**
1	Bad	No meaning understood	Very unpleasant	Unclear all the time	Very often
2	Poor	Effort required	Unpleasant	Not very clear	Often
3	Fair	Moderate effort	Fair	Fairly clear	Few
4	Good	No appreciable effort required	Pleasant	Clear enough	Rarely
5	Excellent	No effort required	Very Pleasant	Very clear	No

	Quality	**Ease of listening**	**Pleasantness**	**Understandability**	**Pronunciation**
MOS	3.57	3.69	3.67	3.75	3.47
STD	0.76	0.83	0.86	0.70	0.78

Πίνακας 18: Η κλίμακα βαθμολόγησης και τα αποτελέσματα του πειράματος για την αξιολόγηση σε επίπεδο πρότασης.

Τα ευρήματα από το συγκεκριμένο πείραμα είναι ιδιαιτέρως σημαντικά, αφού αποδεικνύουν ότι το συγκεκριμένο σύστημα ανταποκρίνεται σε μεγάλο βαθμό στις προσδοκίες των χρηστών όσον αφορά την ευκολία ακρόασης και την άρθρωση του λόγου. Παράλληλα, η φυσικότητα του λόγου χαρακτηρίσθηκε επίσης από ιδιαίτερα υψηλή βαθμολογία αφού πλησιάζει σημαντικά την ποιότητα της σχεδόν φυσικής ομιλίας, κάτι που με δυσκολία επιτυγχάνεται στα περισσότερα συστήματα σύνθεσης φωνής.

7.4.1.1.2 Πείραμα 2: Αξιολόγηση σε επίπεδο λέξης

Ο σκοπός του συγκεκριμένου πειράματος [Viswanathan2005] είναι η αξιολόγηση της καταληπτότητας του συστήματος και της καθαρότητας άρθρωσής του. Χρησιμοποιήσαμε το διαγνωστικό τεστ ρίμας (Diagnostic Rhyme Test) το οποίο χρησιμοποιείται ευρέως σε συστήματα σύνθεσης φωνής για την μελέτη της άρθρωσης του συνθέτη όσον αφορά στα αρχικά ή τελικά σύμφωνα μιας λέξης. Κατά την εκτέλεση του πειράματος, οι ακροατές άκουσαν 33 διαφορετικά σετ ερεθισμάτων, αποτελούμενα από 2 ή τρεις λέξεις κάθε φορά, μερικές εκ των οποίων ήταν τεχνητές, χωρίς νόημα λέξεις. Οι λέξεις αυτές ανά σετ διαφοροποιόντουσαν μόνο σε

ένα γράμμα, και στο άκουσμα των οποίων ο ακροατής όφειλε να επιλέξει ποια λέξη άκουσε. Οι ακροατές κατάφεραν να επιλέξουν σωστά το ερέθισμα που άκουσαν σε ποσοστό μεγαλύτερο του 98%, αποδεικνύοντας το υψηλό επίπεδο ακουστικής ευκρίνειας του συνθετικού λόγου.

7.4.1.1.3 Πείραμα 3: Αξιολόγηση σε επίπεδο παραγράφου

Ο σκοπός του συγκεκριμένου πειράματος ήταν η αξιολόγηση του συνθέτη φωνής σε ένα γενικότερο επίπεδο [Viswanathan2005], επιχειρώντας να λάβουμε πληροφορία τόσο για την ροή του λόγου και την ευκολία κατανόησης του, όσο και για την καταλληλότητα του συστήματος για την συγκεκριμένη αποστολή. Στο συγκεκριμένο πείραμα επιχειρήσαμε να απαντήσουμε σε σημαντικά χαρακτηριστικά του συνθέτη φωνής, όπως είναι η φυσικότητα, η ευκολία κατανόησης κ.α. Οι ακροατές, ακούγοντας πέντε παραγράφους, αποτελούμενες κατά μέσο όρο από 83 λέξεις και 6 προτάσεις, απάντησαν στις ερωτήσεις που φαίνονται στον πίνακα 16.

	Quality	**Ease of listening**	**Pleasantness**	**Understandability**	**Pronunciation**
1	Bad	No meaning understood	Very unpleasant	Unclear all the time	Very often
2	Poor	Effort required	Unpleasant	Not very clear	Often
3	Fair	Moderate effort	Fair	Fairly clear	Few
4	Good	No appreciable effort required	Pleasant	Clear enough	Rarely
5	Excellent	No effort required	Very Pleasant	Very clear	No

EVALUATION RESULTS: EXPERIMENT 3

	Quality	**Ease of listening**	**Pleasantness**	**Understandability**	**Pronunciation**
MOS	3.57	3.69	3.67	3.75	3.47
STD	0.76	0.83	0.86	0.70	0.78

Πίνακας 19: Αποτελέσματα της ακουστικής αξιολόγησης από τους χρήστες αναφορικά με την ποιότητα του τελικού συνθέτη φωνής. Εκτός από την γενικότερη ποιότητα, εξετάστηκαν η ευκολία και ευχαρίστηση ακρόασης, η κατανόηση και η άρθρωση του συστήματος.

Ερωτώμενοι επίσης για την καταλληλότητα ή μη του συστήματος για τον συγκεκριμένο σκοπό, οι χρήστες είχαν την δυνατότητα να απαντήσουν με ελεύθερο κείμενο, παρέχοντας τα σχόλια που είχαν γενικότερα. Τα αποτελέσματα της συγκεκριμένης διαδικασίας δίνουν μία καθαρή εικόνα ότι το συγκεκριμένο σύστημα χαρακτηρίζεται καλό σε όλους σχεδόν τους τομείς, αποδεικνύοντας ότι

η υψηλή ποιότητα που προσφέρει γίνεται αποδεκτή με ικανοποίηση από τον τελικό χρήστη. Οι παραπάνω βαθμολογίες MOS σε σχέση με αντίστοιχες μετρήσεις για άλλους συνθέτες φωνής είναι ιδιαίτερα υψηλές καταδεικνύοντας την υψηλή ποιότητα του συστήματος μας παραγωγής συνθετικής ομιλίας.

7.4.2 Αξιολόγηση χρηστικότητας

Η αξιολόγηση της χρηστικότητας είναι ένας σημαντικός κρίκος της επαναληπτικής (iterative) διαδικασίας σχεδιασμού και ανάπτυξης. Η αξιολόγηση αυτή πραγματοποιήθηκε σε δύο διαφορετικές μεταξύ τους φάσεις, μία φάση που περιλάμβανε την ανευρετική αξιολόγηση – αξιολόγηση από ειδικούς ώστε να εντοπίζουν ενδεχόμενα προβλήματα από την οπτική γωνία των χρηστών – και μία πειραματική, στην οποία τελικοί χρήστες δοκίμασαν και αξιολόγησαν το τελικό σύστημα από την πλευρά της αποτελεσματικότητας, της αποδοτικότητας και της ικανοποίησης κατά την εμπειρία της εκτέλεσης.

7.4.2.1 Ανευρετική (heuristic) αξιολόγηση

Κατά την συγκεκριμένη φάση της διαδικασίας αξιολόγησης [Barnicle2000] [Moore2007], τρεις ειδικοί επιστήμονες σε θέματα προσβασιμότητας και χρήσης τεχνολογίας για επαύξηση αυτής, ανέλαβαν τον ρόλο του τελικού χρήστη και προσομοιώνοντας πλήρη τυφλότητα, επιχείρησαν να φέρουν εις πέρας ένα σύνολο από εργασίες (tasks) μέσω του υπολογιστή τους, χωρίς την χρήση οθόνης ή ποντικιού. Κατά αυτόν τον τρόπο, επιχειρήσαμε να εντοπίσουμε αρχικά, και πριν την ολοκλήρωση του τελικού συστήματος, θέματα που ενδεχομένως θα επηρέαζαν αρνητικά την χρηστική εμπειρία. Η φάση αυτή της αξιολόγησης υπήρξε ιδιαίτερα σημαντική, καθώς εκτός από τον εντοπισμό επιμέρους ατελειών του συστήματος, προτάθηκαν ιδέες που αφορούσαν τόσο στην ολοκλήρωση του συστήματος και στον τρόπο που θα έπρεπε να αλληλεπιδρά με τον χρήστη για την εγκατάστασή του, όσο και επιπλέον λειτουργικότητες, όπως η ενσωμάτωση λεξικού χρήστη και η ενεργοποίηση/απενεργοποίηση από τον χρήστη της εργονομίας για την αυτόματη μετατροπή των Greeklish σε Ελληνικά.

7.4.2.2 Πειραματική αξιολόγηση

Η φάσης της πειραματικής αξιολόγησης του συστήματος ολοκληρώθηκε με την βοήθεια 6 τυφλών χρηστών υπολογιστή, οι οποίοι είχαν εκτενή εμπειρία τόσο με υπολογιστή, όσο και με συστήματα ανάγνωσης οθόνης ειδικότερα. Ο καθένας χρήστης ολοκλήρωσε από τρία 45λεπτα διαστήματα χρήσης του συστήματος μέσω ενός αναγνώστη οθόνης. Ζητήθηκε από τους χρήστες να εκτελέσουν ένα σύνολο από διαφορετικές εργασίες με τον υπολογιστή τους, όπως να διαβάσουν ένα βιβλίο, να πλοηγηθούν στο διαδίκτυο, να συνομιλήσουν ηλεκτρονικά με άλλους κλπ. Τα αποτελέσματα για τον καθένα συλλέχθηκαν χειροκίνητα από μέλη της ομάδας που παρακολουθούσαν την εκτέλεση των συγκεκριμένων εργασιών, αλλά και από προφορικό διάλογο με τους αξιολογητές/χρήστες. Παρ' όλο που ο αριθμός των αξιολογητών ήταν μικρός, επιτύχαμε να εξάγουμε σημαντικά συμπεράσματα αναφορικά με την χρηστικότητα και αποτελεσματικότητα του συστήματος σε συνεργασία με αναγνώστες οθόνης.

7.4.2.2.1 Αποτελεσματικότητα

Οι συμμετέχοντες στην πειραματική αξιολόγηση, στο σύνολό τους, κατάφεραν να ολοκληρώσουν όλες τις εργασίες που τους είχαν δοθεί χωρίς κανένα πρόβλημα. Παρατήρησαν ότι τα εξελιγμένα γλωσσικής τεχνολογίας χαρακτηριστικά του συστήματος, όπως π.χ. η εκφώνηση των αγγλικών λέξεων με την ίδια φωνή, αποτελούν σημαντικό πλεονέκτημα του συστήματος όσον αφορά στην αλληλεπίδρασή του με αναγνώστες οθόνης. Παράλληλα, σημείωσαν ότι εργασίες που στο παρελθόν ήταν δύσκολο ή ακόμη και αδύνατο να εκτελεστούν, όπως π.χ. η ηλεκτρονική συνομιλία μέσω chat, ή η ανάγνωση ενός μηνύματος γραμμένο σε greeklish, με το συγκεκριμένο σύστημα μπορούσαν να εκτελεστούν απρόσκοπτα και χωρίς προσπάθεια.

7.4.2.2.2 Αποδοτικότητα

Αναφορικά με την αποδοτικότητα και το κατά πώς αυτή επηρεάστηκε με το νέο σύστημα, το σύνολο των ερωτηθέντων απάντησε θετικά αναφορικά με τον τρόπο που το συγκεκριμένο σύστημα συνεργάζεται με αναγνώστε οθόνες. Πιο συγκεκριμένα, πολύ θετικό χαρακτηρίστηκε το χαρακτηριστικό της υψηλής ταχύτητας και της άμεσης απόκρισης του συστήματος κατά την πληκτρολόγηση, καθώς επίσης και η ευκρίνεια της συνθετικής φωνής ακόμη και σε πολύ υψηλές ταχύτητες εκτέλεσης.

7.4.2.2.3 Ικανοποίηση

Τέλος, αναφορικά με τον τομέα της ικανοποίησης των χρηστών από το σύστημα, στο σύνολό τους απάντησαν ότι το επίπεδο της ικανοποίησης κατά την εμπειρία χρήσης ήταν σημαντικά υψηλότερο σε σχέση με άλλα συστήματα που χρησιμοποιούσαν μέχρι σήμερα, στο ίδιο περιβάλλον εκτέλεσης (ίδιος υπολογιστής και λογισμικό).

7.5 Συμπεράσματα – Συζήτηση

Στο συγκεκριμένο κεφάλαιο παρουσιάστηκε αναλυτικά ο ιδιαίτερος σχεδιασμός, η προσαρμογή και η αξιολόγηση του συνθέτη ομιλίας για τις ανάγκες χρήσης στο πλαίσιο βοηθημάτων προσβασιμότητας από άτομα με προβλήματα όρασης, ένα από τα σημαντικότερα πεδία χρήσης της συγκεκριμένης τεχνολογίας. Τόσο ο σχεδιασμός, όσο και η ανάπτυξη των απαραίτητων προσαρμογών του συστήματος συνθετικής ομιλίας για την συγκεκριμένη χρήση έγινε με γνώμονα τους εν δυνάμει τελικούς χρήστες και τις απαιτήσεις αυτών. Η διαδικασία της προσαρμογής του συστήματος ήταν επαναληπτική (iterative) με κύριο σκοπό τον εντοπισμό και αποτελεσματική αντιμετώπιση των ιδιαίτερων αναγκών χρήστη, αλλά και των πιθανών σεναρίων χρήσης. Η διαδικασία της αξιολόγησης του τελικού συστήματος απέδειξε οι προσαρμογές που πραγματοποιήθηκαν, αλλά και οι απαιτήσεις χρήστη που συλλέχθηκαν αρχικά, συνετέλεσαν σημαντικά στην δημιουργία ενός σύγχρονου συνθέτη ομιλίας για τα Ελληνικά που αντιμετωπίζει με αποτελεσματικότητα όλα τα προβλήματα που συναντάει ένας χρήστης υπολογιστή με μειωμένη όραση και σε συνεργασία με έναν αναγνώστη οθόνης. Οι προσαρμογές που απαιτήθηκαν ανήκαν τόσο στο υποσύστημα γλωσσικής επεξεργασίας του συστήματος, όσο και στο ίδιο το υποσύστημα της ψηφιακής επεξεργασίας σήματος, ενώ ιδιαίτερες παράμετροι χρειάσθηκε να ληφθούν υπόψη για την ολοκλήρωση του λογισμικού σε πακέτο για τον τελικό χρήστη. Τα αποτελέσματα της αξιολόγησης είναι θετικά και το συγκεκριμένο σύστημα αποτελεί πλέον την δημοφιλέστερη λύση ανάμεσα στην κοινότητα των χρηστών υπολογιστή με προβλήματα όρασης στην Ελλάδα.

7.6 Βιβλιογραφία Κεφαλαίου

[Alvarez2002] Alvarez Y.V., M. Huckvale: The reliability of the ITU-T P.85 standard for the evaluation of text-to speech systems, Proc. ICSLP 2002, 329–332 (2002)

[Bailly2003] Bailly G., W.N. Campbell, and B. Mobius, "ISCA Special Session: hot topics in speech synthesis", *Proc. Eurospeech 2003,* pp. 37-40, Geneva, 2003.

[Barnicle2000] Barnicle K., "Usability testing with screen reading technology in a windows environment," *Proc. of the Conf. on Universal Usability*, pp. 102-109, 2000.

[Beutnagel1999] Beutnagel M., Mohri, R., and Riley, M., "Rapid unit selection from a large speech corpus for concatenative speech synthesis," *in Proc. Eurospeech 99*, Budapest, 1999.

[Bigham2007] Bigham J., Cavender A. C., J. T. Brudvik, J. O. Wobbrock, and R. Ladner, "WebinSitu: A comparative analysis of blind and sighted browsing behaviour," *9th Intl. Conf. ACM SIGACCESS on Computers and Accessibility*, Arizona USA, pp. 51-58, 2007.

[Black1997] Black A., and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," *Proc. of Eurospeech 97*, vol. 2, pp. 601-604, Greece, 1997.

[Campbell2005] Campbell N., "Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech," *IEICE trans. Inf. & Syst.*, vol. E88-D, no. 3, pp.376-383, 2005.

[Chalamandaris2006] Chalamandaris A., A. Protopapas, P. Tsiakoulis, and S. Raptis. "All Greek to me! An automatic Greeklish to Greek transliteration system." *5th Int. Conf. on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy, pp. 1226–1229, 2006.

[Chalamandaris2005]Chalamandaris A., S. Raptis, and P. Tsiakoulis, "Rule-based grapheme-to-phoneme method for the Greek," *in Interspeech 2005*, pp. 2937-2940, 2005.

[Chu2000] Chu, Wai C. "*Speech coding algorithms: Foundation and evolution of standardized coders,*" John Wiley & Sons, 2003.

[Coorman2000] Coorman G., Fackrell, J., Rutten, P., and Coile, B. V., "Segment selection in the LH realspeak laboratory TTS system," *Proc. of the ICSLP 2000*, vol. 2, pp. 395-398, 2000.

[Duggan2003] Duggan B. and M. Deegan, "Considerations in the usage of text to speech (tts) in the creation of natural sounding voice enabled web systems", *In Proc. of the 1st Int. Symp. on Information and Communication technologies (ISICT '03:)*, pp. 433–438, Trinity College Dublin, 2003.

[Dutoit2008] Dutoit T., "Corpus-based Speech Synthesis*," Springer Handbook of Speech Processing,* J. Benesty, M. M. Sondhi, Y. Huang (eds), Part D, Chapter 21, pp. 437-455, Springer, 2008.

[Earl1999] Earl, C.L., Leventhal, J.D., "A survey of windows screen reader users: Recent improvements in accessibility," *Journal of Visual Impairment and Blindness*, vol. 93, no. 3, pp. 174-177, 1999.

[Hackos1997] Hackos J. T. and J. C. Redish, "*User and task analysis for interface design*," John Wiley & Sons, Inc., Chichester 1997.

[Heuven1995] Heuven V. J. van and R. van Bezooijen, "Quality Evaluation of Synthesized Speech," *Speech Coding and Synthesis,* W. B. Kleijn, and K. K. Paliwal (eds), Chapter 21, pp. 707-738, Elsevier Science, 1995.

[Karabetsos2009]Karabetsos S., P. Tsiakoulis, A. Chalamandaris, and S. Raptis, "Embedded unit selection text-to-speech synthesis for mobile devices," *IEEE Trans. on Consumer Electronics*, vol. 55, no. 2, pp. 613-621, 2009.

[Kim2006] Kim S.-J., Kim J.-J. and Hahn M.-S., "HMM-based Korean speech synthesis system for hand-held devices," *IEEE Trans. Consumer Electronics*, vol. 52, no. 4, pp. 1384-1390, 2006.

[Mohasi2006] Mohasi L. and D. Mashao, "Text-to-Speech Technology in Human-Computer Interaction", *5th Conference on Human Computer Interaction in Southern Africa, (CHISA 2006, ACM SIGHI)*, pp. 79-84, 2006.

[Moore2007] Moore R. K., "PRESENCE: A human-inspired architecture for speech-based human-machine interaction," *IEEE Trans. Computers*, 56, pp. 1176-1188, 2007.

[O'Shaughnessy2007] O'Shaughnessy Douglas, "Modern Methods of Speech Synthesis," *IEEE Circuits and Systems Magazine,* Third Quarter 2007, pp. 6-23, 2007.

[Picheny1986] Picheny et al. Speaking Clearly for the Hard of Hearing II: Acoustic Characteristics of Clear and Conversational Speech. J Speech Hear Res.1986; 29: 434-446

[Redish2003] Redish G., and Theofanos, M.F, "Observing Users Who Listen to Web Sites," *Usability Interface*, vol.9, no.4, 2003.

[Schnell2002] Schnell M., O. Jokisch, R. Hoffmann, and M. Kustner, "Text-to-speech for low-resource systems," *IEEE Workshop Multimedia Signal Processing (MMSP)*, St. Thomas, pp. 259-262, 2002.

[Schultz2006] Schultz T., A. W. Black, S. Vogel, and M. Woszczyna, "Flexible Speech Translation Systems," *IEEE trans. on Audio, Speech and Language Processing,* vol. 14, no. 2, pp. 403-411, 2006.

[Spiegel1990] Spiegel M.F., M.J. Altom, M.J. Macchi: Comprehensive assessment of the telephone intelligibility of synthesized and natural speech, Speech Commun. 9, 279–291 (1990)

[Tomko2005] Tomko S., T. K. Harris, A. Toth, J. Sanders, A. Rudnicky and R. Rosenfeld, "Toward Efficient Human Machine Speech Communication: The Speech Graffiti Project," *ACM Trans. on Speech and Language Processing*, vol. 2, no. 1, Article 2, pp. 1-27, 2005.

[Tsiakoulis2008] Tsiakoulis P., A. Chalamandaris, S. Karabetsos and S. Raptis, "A Statistical Method for Database Reduction for Embedded Unit Selection Speech Synthesis," *in IEEE ICASSP 2008*, pp. 4601-4604, 2008.

[Viswanathan2005]Viswanathan M. and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech & Language*, vol. 19, pp. 55-83, January 2005.

[Wai2003] Wai C Chu,. "*Speech coding algorithms: Foundation and evolution of standardized coders,*" John Wiley & Sons, 2003.

[W3C2003] W3C Speech Synthesis Markup Language Version 1.0: http://www.w3.org/TR/2003/CR-speechsynthesis-20031218/ (http://www.xml.com/pub/a/2004/10/20/ssml.html)

8. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Στο κεφάλαιο αυτό κάνουμε σύνοψη της παρούσας διατριβής, παρουσιάζοντας την ερευνητική συνεισφορά της συγκεκριμένης διατριβής στο πεδίο της σύνθεσης φωνής, καθώς επίσης και θέματα που άπτονται περαιτέρω έρευνας στο μέλλον. Τα αποτελέσματα της διατριβής εντάσσονται τόσο στο γενικότερο πεδίο της σύνθεσης φωνής, όσο και σε ένα ειδικότερο πλαίσιο της σύνθεσης φωνής για την Ελληνική γλώσσα, αλλά και την χρήση της συγκεκριμένης τεχνολογίας για επαυξημένη προσβασιμότητα σε περιπτώσεις μειωμένης όρασης. Έχοντας ένα αρκετά ευρύ πεδίο έρευνας, η συγκεκριμένη διατριβή εντοπίζει αρκετά και σημαντικά σημεία της συγκεκριμένης τεχνολογίας, τα οποία θεωρούνται ακόμη ανοικτά ερευνητικά, με αρκετά περιθώρια βελτίωσης, είτε με προσεγγίσεις που έχουν ήδη προταθεί, είτε με νέες. Στο ίδιο πλαίσιο περιγράφονται θέματα που αποτελούν κομβικό σημείο στην περαιτέρω εξέλιξη της τεχνολογίας της σύνθεσης φωνής και την πορεία της προς την παραγωγή ενός συνθέτη φωνής αδιαχώριστου από έναν φυσικό ομιλητή.

8.1 Κύριες συνεισφορές της διατριβής

Η συγκεκριμένη διατριβή κινείται στο πεδίο της συνθετικής ομιλίας, περιγράφοντας την ερευνητική μας δραστηριότητα στον χώρο και παρουσιάζοντας τα αποτελέσματα των προσπαθειών αυτών. Συνολικό αποτέλεσμα κοινής ερευνητικής προσπάθειας της ομάδας Σύνθεσης Φωνής, αποτελεί το τρέχον σύστημα συνθετικής ομιλίας για τα Ελληνικά που υπάρχει σήμερα. Η συγκεκριμένη διατριβή, αν και αποτέλεσε την βάση για σημαντική ερευνητική προσπάθεια σχεδόν σε όλους τους τομείς της συνθετικής ομιλίας, εστίασε κυρίως σε περιοχές του πεδίου όπου υπήρχε ανάγκη ουσιαστικής συνεισφοράς και προόδου, ώστε να βελτιωθεί η τεχνολογία συνθετικής ομιλίας. Η ερευνητική συνεισφορά της παρούσας διδακτορικής διατριβής κινήθηκε σε τέσσερεις άξονες.

8.1.2 Μοντελοποίηση της προσωδίας στην σύνθεση φωνής

Το πεδίο της μοντελοποίησης της προσωδίας είναι ένα ανοικτό ερευνητικά πεδίο, με πολλές διαφορετικές προσεγγίσεις και μεθοδολογίες. Στο πλαίσιο της συγκεκριμένης διατριβής προτάθηκε ένας νέος ουσιαστικά αλγόριθμος ο οποίος επιχειρεί την μοντελοποίηση της προσωδίας εμμέσως [Giannopoulos et al., 2003], χωρίς την χρήση κάποιου ρητού μοντέλου προσωδίας, όπως περιγράφεται σε άλλα συστήματα σύνθεσης φωνής. Πιο συγκεκριμένα, η μέθοδος που προτείναμε κάνει χρήση της συλλαβής ως ελάχιστη ακουστική μονάδα και στην συνέχεια επιχειρεί να εντοπίσει διαφορετικά προσωδιακά πρότυπα βασιζόμενο τόσο σε χαρακτηριστικά του γλωσσικού όσο και του ακουστικού επιπέδου. Τα πρότυπα αυτά λαμβάνονται υπόψη κατά την επιλογή βέλτιστων ακουστικών μονάδων, επιτυγχάνοντας μία έμμεση μοντελοποίηση της επιθυμητής προσωδίας [Dimou & Chalamandaris, 2006]. Βασικό χαρακτηριστικό της προτεινόμενης μεθόδου είναι η διατήρηση της μικροπροσωδίας κατά την εφαρμογή της τελικής καμπύλης, γεγονός που διατηρεί σε υψηλά επίπεδα την ποικιλότητα και φυσικότητα του τελικού συνθετικού σήματος, αλλά και το γεγονός ότι η συγκεκριμένη μέθοδος συνδυάζει αποτελεσματικά τόσο γλωσσικά, όσο και ακουστικά χαρακτηριστικά της γλώσσας και του ύφους, γεγονός που επιτρέπει την εύκολη προσαρμογή της σε νέες γλώσσες και στυλ εκφώνησης [Dimou & Chalamandaris, 2008].

8.1.3 Σχεδιασμός και ανάπτυξη της βάσης δεδομένων του συστήματος σύνθεσης φωνής

Μία από τις κυριότερες συνεισφορές της συγκεκριμένης διατριβής εστιάζεται στην διαδικασία σχεδιασμού και υλοποίησης της βάσης δεδομένων ενός συστήματος σύνθεσης φωνής [Chalamandaris et al., 2009a]. Οι βασικές συνιστώσες της μεθόδου μας στο συγκεκριμένο πεδίο ήταν τρεις: α) η διαδικασία και ο αλγόριθμος σχεδιασμού του σώματος κειμένου προς ηχογράφηση, β) η ψηφιακή επεξεργασία και επισημείωση των ακουστικών αρχείων και γ) η αυτοματοποιημένη σταχυολόγηση της βάσης δεδομένων για την άμεση ενσωμάτωσή της σε έναν συνθέτη φωνής, χωρίς την χειρονακτική διόρθωσή της.

Η μέθοδος σχεδιασμού του σώματος κειμένου προς ηχογράφηση όπως προτάθηκε βασίζεται σε έναν απλήστο αλγόριθμο επιλογής, και ολοκληρώνεται με επιπλέον στάδια σταχυολόγησης και βελτίωσης του τελικού αποτελέσματος με βάση το υποσύστημα βέλτιστης επιλογής ακουστικών μονάδων. Η προσέγγιση αυτή ουσιαστικά οδήγησε στην δημιουργία ενός σώματος κειμένου συμπληρωματικού ως προς τις ιδιαιτερότητες του υποσυστήματος βέλτιστης επιλογής ακουστικών μονάδων [Tsiakoulis et al., 2008], γεγονός που ουσιαστικά αντισταθμίζει σε μεγάλο βαθμό το ανεπίλυτο ακόμη και σήμερα πρόβλημα της βέλτιστης επιλογής βαρών στην συνάρτηση επιλογής. Η ύπαρξη 2 επιπλέον σταδίων στην διαδικασία σχεδιασμού και ηχογράφησης του σώματος κειμένου για έναν συνθέτη φωνής, διασφάλισε περαιτέρω την πληρότητα του σώματος κειμένου και της τελικής βάσης δεδομένων, αντιμετωπίζοντας παράγοντες που έχουν να κάνουν με τις ιδιαιτερότητες της φωνής του ομιλητή, αλλά και τις συνθήκες ηχογράφησης. Σημαντικό κομμάτι της προτεινόμενης διαδικασίας σχεδιασμού και ανάπτυξης της βάσης δεδομένων του συνθέτη φωνής, αποτέλεσε τόσο η ψηφιακή επεξεργασία των ηχογραφήσεων, όσο και το στάδιο του αυτόματου τεμαχισμού και επισημείωσης των ηχογραφήσεων αυτών. Τέλος, η συγκεκριμένη διατριβή πρότεινε έναν καινοτόμο μηχανισμό για την αυτοματοποιημένη σταχυολόγηση των στοιχείων της βάσης δεδομένων, ούτως ώστε να αντιμετωπίζονται χωρίς ουσιαστική επίβλεψη προβλήματα που έχουν προκύψει από σφάλματα ή ανακριβή αποτελέσματα κατά την αυτόματη κατάτμηση των ηχογραφήσεων.

8.1.4 Ειδικά θέματα επεξεργασίας κειμένου για τον συνθέτη φωνής

Η συγκεκριμένη διατριβή πραγματεύτηκε επίσης θέματα που άπτονται της επεξεργασίας κειμένου και ειδικότερα στο πλαίσιο ανάπτυξης ενός συστήματος σύνθεσης φωνής για τα Ελληνικά. Πιο

συγκεκριμένα, μία από τις σημαντικότερες συνεισφορές της διατριβής αποτέλεσε η ανάπτυξη ενός αυτοματοποιημένου συστήματος φωνητικής μεταγραφής για τα Ελληνικά [Chalamandaris et al., 2005]. Βασιζόμενο σε έναν data-driven αλγόριθμο, ο συγκεκριμένος μηχανισμός κάνει χρήση απλών κειμενικών πόρων για να δημιουργήσει ένα σύνολο από κανόνες μεταγραφής από γραφήματα της Ελληνικής σε αντίστοιχα φωνήματα, προσφέροντας αποτελέσματα υψηλής ακρίβειας και ποιότητας. Παράλληλα, προτάθηκε και παρουσιάστηκε ένας εξειδικευμένος αλγόριθμος για την αντιμετώπιση του φαινομένου των Greeklish [Chalamandaris et al., 2004a] [Chalamandaris et al., 2004b] και την αυτόματη μεταγραφή κειμένων που είναι γραμμένα σε ένα οποιοδήποτε τύπο Greeklish σε ορθά Ελληνικά, με ακρίβεια που ξεπερνάει το 98%.

8.1.5 Ειδικά θέματα σχεδιασμού και προσαρμογής συνθέτη φωνής για χρήση σε προσβάσιμα εργαλεία

Τέλος, μία σημαντική συνεισφορά της συγκεκριμένης διατριβής αποτελεί και η συστηματική αντιμετώπιση θεμάτων που άπτονται της λειτουργίας του συνθέτη φωνής ως εργαλείου υποβοήθησης και επαύξησης της προσβασιμότητας για άτομα με μειωμένη όραση [Chalamandaris et al., 2009a] [Chalamandaris et al., 2009c]. Σημαντικό ρόλο στην συγκεκριμένη διαδικασία διαδραμάτισε τόσο το στάδιο της καταγραφής των αναγκών χρήστη, όσο και το στάδιο της αξιολόγησης του τελικού συστήματος σύνθεσης φωνής και της επικύρωσης των αποτελεσμάτων, μέσω ειδικών μηχανισμών. Ως επιπλέον επικύρωση της συγκεκριμένης συνεισφοράς προστίθεται και η άμεση υιοθέτηση του τελικού συστήματος από ενδιαφερόμενες ομάδες χρηστών λογισμικών προσβασιμότητας, αλλά και οι πολύ θετικές εντυπώσεις που αναφέρουν σχετικά.

8.2 Θέματα μελλοντικής έρευνας

Στην παρούσα διατριβή παρουσιάσαμε την ερευνητική προσπάθεια και τα αποτελέσματα αυτής στο πλαίσιο της σχεδίασης αλλά και βελτίωσης ενός συστήματος σύνθεσης φωνής για τα Ελληνικά. Η συνεισφορά της συγκεκριμένης εργασίας σε επιμέρους θέματα που άπτονται του γενικότερου πλαισίου της σύνθεσης φωνής, είναι φανερά από τα αποτελέσματα του τελικού συστήματος στο οποίο κατέληξε η συνολική ερευνητική προσπάθεια. Ένα από τα σημαντικότερα αποτελέσματα της συγκεκριμένης διατριβής αποτελεί η αναγνώριση κύριων σημείων του ερευνητικού πεδίου της σύνθεσης φωνής που όχι μόνο επιδέχονται βελτιώσεων, αλλά η περαιτέρω μελέτη τους είναι

απαραίτητη για την εξέλιξη της τεχνολογίας σύνθεσης φωνής προς ένα σύστημα που θα είναι ικανό να παράξει συνθετική φωνή με υψηλή εκφραστικότητα και φυσικότητα, ανάλογα με το κείμενο. Οι επιμέρους τομείς προς άμεση μελλοντική έρευνα είναι οι εξής:

8.2.1 Μελέτη της προσωδίας συναισθηματικού λόγου

Τα κυριότερα χαρακτηριστικά του εκφραστικού λόγου μπορούν να εκφρασθούν ποσοτικά, μέσω της προσωδίας και των χαρακτηριστικών αυτής [Campbell 2003] [Eide 2004]. Πιο συγκεκριμένα, φαινόμενα όπως η έμφαση, η χαρά, η λύπη κ.α. μπορούν μέσω ποσοτικών αλλά και ποιοτικών χαρακτηριστικών να εντοπισθούν ή ακόμη και να μοντελοποιηθούν. Μέχρι σήμερα, αν και έχουν γίνει πολλές προσπάθειες για την μελέτη και μοντελοποίηση της προσωδίας στον εκφραστικό λόγο, δεν έχει επιτευχθεί η αποτελεσματική μοντελοποίηση των χαρακτηριστικών αυτών στο πλαίσιο της σύνθεσης φωνής. Απλοϊκές προσεγγίσεις που απλά επιταχύνουν ή επιβραδύνουν ανάλογα την ταχύτητα εκφώνησης ή/και την θεμελιώδη συχνότητα δεν έχουν προσφέρει τα αναμενόμενα αποτελέσματα. Είναι φανερό πλέον, ότι ποιοτικά χαρακτηριστικά του εκφραστικού και συναισθηματικού λόγου δεν μπορούν να μοντελοποιηθούν μόνο μέσω της θεμελιώδους συχνότητας και του ρυθμού εκφοράς, αλλά με μετρικές που ενδεχομένως περιλαμβάνουν ως υποσύνολό τους τις παραπάνω παραμέτρους.

8.2.2 Μελέτη και ανάπτυξη προ-επεξεργασίας κειμένου για εκφραστικό λόγο

Ο τρόπος ανάγνωσης ενός κειμένου με εκφραστικό τρόπο αποτελεί μια διαδικασία που εξαρτάται τόσο από το κείμενο, όσο και από τον ίδιο ομιλητή, όπως ακριβώς συμβαίνει και στην εκφραστικότητα της εκτέλεσης ενός μουσικού κομματιού από έναν μουσικό [Bulut 2004]. Οποιαδήποτε προσπάθεια αναζήτησης ενός πλήρους συνόλου λέξεων ή εκφράσεων που εκφέρονται με ιδιαίτερη εκφραστικότητα από όλους τους ομιλητές δεν θα μπορούσε παρά να αποτύχει. Ωστόσο, θεωρούμε δυνατή την μελέτη ενός βασικού συνόλου γλωσσικών χαρακτηριστικών και φαινομένων που κατά βάση φέρουν περισσότερη εκφραστικότητα κατά την ανάγνωση [Tsuzuki 2004]. Για τον λόγο αυτό, βασική προταιρεότητα στην μελλοντική μας έρευνα αποτελεί η μελέτη των γραμματικών και συντακτικών φαινομένων της Ελληνικής γλώσσας που φαίνεται ότι έχουν υψηλή πιθανότητα εκφραστικού περιεχομένου στην ανάγνωσή τους, με επόμενο άμεσο στόχο την δυνατότητα γενίκευσής τους σε περισσότερες γλώσσες.

8.2.3 Μελέτη για τον σχεδιασμό της βάσης δεδομένων ενός εκφραστικού συνθέτη και του αλγορίθμου βέλτιστης επιλογής

Τα αποτελέσματα των παραπάνω μελετών πρόκειται να προσφέρουν τον ακρογωνιαίο λίθο για την περαιτέρω μελέτη και σχεδίαση μιας πληρέστερης βάσης δεδομένων για τον συνθέτη φωνής [Tsuzuki 2004]. Έχοντας ποσοτικοποιήσει τα συγκεκριμένα χαρακτηριστικά, θα μπορεί κανείς να διερευνήσει την δυνατότητα κάλυψης των παραπάνω φαινομένων και χαρακτηριστικών μέσω ενός ηχογραφημένου σώματος κειμένου και της αντίστοιχης επισημείωσης αυτού. Οι παράμετροι αυτές είναι προφανές ότι θα απαιτηθούν για την ολοκήρωση του συστήματος βέλτιστης επιλογής ακουστικών μονάδων, ώστε να χρησιμοποιηθεί όχι μόνο κατά την εκτέλεση του συνθέτη φωνής, αλλά και κατά την διαδικασία σχεδιασμού του σώματος κειμένου, όπως προτάθηκε στην συγκεκριμένη διατριβή [Abe1990].

8.2.4 Μελέτη και σχεδιασμός νέων καινοτόμων εφαρμογών για καθημερινή χρήση

Μία σημαντική και απαραίτητη διαδικασία για την δυνατότητα περαιτέρω έρευνας αλλά και τον λόγο ύπαρξης αυτής αποτελεί η σύνδεση των αποτελεσμάτων αυτής με υπηρεσίες και προϊόντα της καθημερινής μας δραστηριότητας. Μέχρι σήμερα, έχει φανεί ότι η τεχνολογία σύνθεσης φωνής αποτελεί ένα απαραίτητο εργαλείο για εμποδιζόμενα άτομα, αλλά και όχι μόνο. Η ανάπτυξη νέων τεχνολογιών και η ταχύτατη διάθεσή τους μαζικά, όπως είναι οι έξυπνες φορητές συσκευές κ.α. μπορούν να αποτελέσουν την βάση για την ανάπτυξη και διάθεση υπηρεσιών και προϊόντων που μέχρι σήμερα δεν είχαμε ακόμη σκεφτεί. Σκοπός μας στο άμεσο μέλλον είναι η μελέτη των νέων δεδομένων που προσφέρει η νέα τεχνολογία για την ενσωμάτωση της σύνθεσης φωνής σε καθημερινές δραστηριότητες και διαδικασίες, με σκοπό την διευκόλυνση ή ακόμη και την εναλλακτική πρόσβαση στην πληροφορία μέσω ενός περισσότερο διαισθητικού τρόπου.

8.3 Βιβλιογραφία Κεφαλαίου

[Abe1990] Abe M., S. Nakamura, K. Shikano, H. Kuwahara: Voice conversion through vector quantization, Proc. IEEE ICASSP 88, 655–658 (1990), S14.1

[Bulut 2004] Bulut, M., Narayanan, S. and Johnson, J., "Synthesizing expressive speech: overview, challenges, and Open Questions," In S. Narayanan and A. Alwan. Text-to- Speech Synthesis: New Paradigms and Advances, pp. 175- 201, Prentice Hall.

[Campbell 2003] Campbell, N., "Towards Synthesizing Expressive Speech: Designing and Collecting Expressive Speech Data," Proc. Eurospeech 2003, pp. 1637-1640, 2003.

[Chalamandaris et al., 2004a] A. Chalamandaris, P. Tsiakoulis, S. Raptis, G. Giannopoulos and G. Carayannis, "Bypassing Greeklish!", in Proc. LREC 2004: 4th International Conference on Language Resources And Evaluation, May 26-28, Lisbon, Portugal, 2004

[Chalamandaris et al., 2004b] A. Chalamandaris, P. Tsiakoulis, S. Raptis and G. Giannopoulos, "An Efficient and Robust Algorithm for Bypassing Greeklish", in Proc. IC-SCCE 2004: 1st International Conference from Scientific Computing to Computational Engineering, 8-10 September, Athens, Greece, 2004

[Chalamandaris et al., 2005] A. Chalamandaris, S. Raptis and P. Tsiakoulis, "Rule-based grapheme-to-phoneme method for the Greek", in Proc. Interspeech'2005: 9th European Conference on Speech Communication and Technology, September 4-8, Lisbon, Portugal, 2005

[Chalamandaris et al., 2006] A. Chalamandaris, A. Protopapas, P. Tsiakoulis and S. Raptis, "All Greek to me! An automatic Greeklish to Greek Transliteration System," 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 24–26 May, 2006

[Chalamandaris et al., 2009a] A. Chalamandaris, P. Tsiakoulis, S. Raptis, and Sotiris Karabetsos, "Design of an Efficient Corpus for High-Quality Unit Selection TTS for Bulgarian", in Proc. 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poland, 2009

[Chalamandaris et al., 2009b] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and Spyros Raptis, "An Efficient and Robust Pitch Marking Algorithm on the Speech Waveform for TD-PSOLA", in Proc. Intl. IEEE Conference on Signal and Image Processing Applications (ICSIPA), Malaysia, 2009

[Chalamandaris et al., 2009c] A. Chalamandaris, S. Raptis, P. Tsiakoulis, and S. Karabetsos, "Enhancing Accessibility of Web Content for the Print-Impaired and Blind People", in USAB2009: Human-computer interaction for eInclusion, A. Holzinger and K. Miesenberger (Eds.), Lecture Notes in Computer Science, Springer, 2009

[Chalamandaris et al., 2010] A. Chalamandaris, S. Karabetsos, P. Tsiakoulis, S. Raptis, "A Unit Selection Text-to-Speech Synthesis System Optimized for Use with Screen Readers", IEEE Transactions on Consumer Electronics, Vol. 56, No. 3, pp. 1890-1897, August, 2010.

[Chalamandaris et al., 2011] Chalamandaris, P. Tsiakoulis, S. Raptis, S. Karabetsos, "Corpus Design for a Unit Selection TtS System with Application to Bulgarian" in Human Language Technology. Challenges for Computer Science and Linguistics, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011

[Craggs 2004] Craggs, R., "Annotating emotion in dialog: issues and approaches," In Lee, M. (Ed), Proceedings of the 7th Annual CLUK Research Colloquium, 2004.

[Dimou & Chalamandaris, 2006] A. L. Dimou and A. Chalamandaris, "Language identification from suprasegmental cues; examining the role of rhythm in the identification of a Greek dialect", in Proc. La comunicazione parlata / Spoken Communication, February, , Italy, 2006

[Dimou & Chalamandaris, 2008] A. L. Dimou and A. Chalamandaris, "Is idiom identification possible from prosodic information? An experimental approach for the Greek language", in Proc. 4th Intl Conf. Speech Prosody 2008, pp. 759-762 (2008)

[Eide 2004] Eide, E., Bakis, R., Hamza, W. and Pitrelli, J. F., "Toward expressive synthetic speech,", In Narayanan, S. and Alwan, A., Text-to-Speech Synthesis: New Paradigms and Advances, pp. 219-248, Prentice Hall.

[Eide2004] Eide E., R. Bakis, W. Hamza, J.F. Petrelli: Toward expressive synthetic speech. In: Text-to-Speech Synthesis – New Paradigms and Advances, Professional Technical Reference, ed. by S. Narayanan, A. Alwan (Prentice-Hall, Upper Saddle River 2004) pp. 219–248, Chap. 11

[Founda et al., 2001a] M. Founda, A. Chalamandaris, G. Tambouratzis, and G. Carayannis, "Reducing Spectral Mismatches in Concatenative Speech Synthesis via Systematic Database Enrichment", in Proceedings of the Eurospeech-2001 Conference, Aalborg, Denmark, 4-7 September 2001, pp. 837-840.

[Founda et al., 2001b] M. Founda, A. Chalamandaris, G. Tambouratzis, and G. Carayannis, "Studying the Factors Affecting the Optimal Unit Selection Algorithm for a TTS System for the Greek Language", in Proceedings of the 4th European Conference on Noise Control EURONOISE2001, Patra, 14-17 January 2001, Vol. II, pp. 758-764.

[Giannopoulos et al., 2003] G. Giannopoulos, A. Chalamandaris, S-E. Fotinea, T. Athanaselis, G. Carayannis, "Analysis and modelling of the Carrier Declination for the Greek language", in Proc. of the 15th International Congress of Phonetic Sciences - ICPhS03, 3-9 August 2003, Barcelona, pp. 555-558.

[Karabetsos et al., 2008] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, and S. Raptis, "HMM-based Speech Synthesis for the Greek Language," in Petr Sojka, Ivan Kopecek, and Karel Pala (eds.), Lecture Notes in Computer Science (LNCS), Springer – Verlag, 2008

[Karabetsos et al., 2009] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, S. Raptis, "Embedded Unit Selection Text-to-Speech Synthesis for Mobile Devices", IEEE Transactions on Consumer Electronics, Issue 2, Vol. 56, May 2009.

[Karabetsos et al., 2010] S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, S. Raptis, "One-Class Classification for Spectral Join Cost Calculation in Unit Selection Speech Synthesis", IEEE Signal Processing Letters, Vol. 17, No. 8, pp. 746-749, August, 2010

[Lee2002] Lee, C. M., Narayanan, S. and Pieraccini, R., "Combining acoustic and language information for emotion recognition", Proc. ICSLP 2002.

[Protopapas et al., 2010] A. Protopapas, M. Tzakosta, A. Chalamandaris and P. Tsiakoulis, "IPLR: an online resource for Greek word-level and sublexical information", Language Resources & Evaluation, Springer, 2010.

[Raptis & Chalamandaris, 2005] S. Raptis and A. Chalamandaris, "IMUTUS - Interactive Music Tuition System", The 5th Open MusicNetwork Workshop, July 4-5, Vienna, Austria, 2005

[Raptis et al., 2005b] S. Raptis, A. Askenfelt, D. Fober, A. Chalamandaris, E. Schoonderwaldt, S. Letz, A. Baxevanis, K. Falkenberg Hansen and Y. Orlarey, "IMUTUS – An Effective Practicing Environment For Music Tuition", International Computer Music Conference (ICMC 2005), September 5-9, Barcelona, Spain, 2005

[Raptis et al., 2009a] S. Raptis, P. Tsiakoulis, A. Chalamandaris and S. Karabetsos, "High Quality Unit-Selection Speech Synthesis for Bulgarian", In Proc. 13th International Conference on Speech and Computer (SPECOM'2009), St. Petersburg, Russia, June 21-25, 2009

[Raptis et al., 2009b] S. Raptis, P. Tsiakoulis, A. Chalamandaris and S. Karabetsos, "User Interaction Design for a Home-Based Telecare System", in USAB2009: Human-computer interaction for eInclusion, A. Holzinger and K. Miesenberger (Eds.), Lecture Notes in Computer Science, Springer, 2009

[Raptis et al., 2010] S. Raptis, A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, "The ILSP Text-to-Speech System for the Blizzard Challenge 2010", In Proc. Blizzard Challenge 2010 Workshop, Kyoto, Japan, September 25, 2010

[Schultz 2006] Schultz T., A. W. Black, S. Vogel, and M. Woszczyna, "Flexible Speech Translation Systems," IEEE trans. on Audio, Speech and Language Processing, vol. 14, no. 2, pp. 403-411, 2006.

[Tsiakoulis et al., 2005] P. Tsiakoulis, S. Karabetsos, S. E. Fotinea and I. Dologlou, "Spectral Estimation for Speech Signals based on Decimation and Eigenanalysis", in Proc. HERCMA-2005 (7th Hellenic European Conf. on Computer Mathematics & its Applications), Athens, Greece, September, 2005

[Tsiakoulis et al., 2008] P. Tsiakoulis, A. Chalamandaris, S. Karabetsos and S. Raptis, "A Statistical Method for Database Reduction for Embedded Unit Selection Speech Synthesis," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Las Vegas, USA, 2008

[Tsuzuki 2004] Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M. and Narayanan. S., "Constructing Emotional Speech Synthesizers with Limited Speech Database." Proc. ICSLP 2004.

9. ΚΑΤΑΛΟΓΟΣ ΔΗΜΟΣΙΕΥΣΕΩΝ ΤΟΥ ΣΥΓΓΡΑΦΕΑ

Ακολουθεί κατάλογος δημοσιεύσεων του συγγραφέα.

9.1 Δημοσιεύσεις σε περιοδικά με κριτές

[Chalamandaris et al., 2010]
A. Chalamandaris, S. Karabetsos, P. Tsiakoulis, S. Raptis, **"A Unit Selection Text-to-Speech Synthesis System Optimized for Use with Screen Readers"**, IEEE Transactions on Consumer Electronics, Vol. 56, No. 3, pp. 1890-1897, August, 2010.

[Karabetsos et al., 2010]
S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, S. Raptis, **"One-Class Classification for Spectral Join Cost Calculation in Unit Selection Speech Synthesis"**, IEEE Signal Processing Letters, Vol. 17, No. 8, pp. 746-749, August, 2010

[Protopapas et al., 2010]
A. Protopapas, M. Tzakosta, A. Chalamandaris and P. Tsiakoulis, **"IPLR: an online resource for Greek word-level and sublexical information"**, Language Resources & Evaluation, Springer, 2010.

[Karabetsos et al., 2009]
S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, S. Raptis, **"Embedded Unit Selection Text-to-Speech Synthesis for Mobile Devices"**, IEEE Transactions on Consumer Electronics, Issue 2, Vol. 56, May 2009.

9.2 Δημοσιεύσεις σε συλλογικούς τόμους με κριτές

[Chalamandaris et al., 2011]
Chalamandaris, P. Tsiakoulis, S. Raptis, S. Karabetsos, **"Corpus Design for a Unit Selection TtS System with Application to Bulgarian"** in Human Language Technology. Challenges for Computer Science and Linguistics, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011

[Chalamandaris et al., 2009c]
A. Chalamandaris, S. Raptis, P. Tsiakoulis, and S. Karabetsos, **"Enhancing Accessibility of Web Content for the Print-Impaired and Blind People"**, in USAB2009: Human-computer interaction for eInclusion, A. Holzinger and K. Miesenberger (Eds.), Lecture Notes in Computer Science, Springer, 2009

[Raptis et al., 2009b]
S. Raptis, P. Tsiakoulis, A. Chalamandaris and S. Karabetsos, **"User Interaction Design for a Home-Based Telecare System"**, in USAB2009: Human-computer interaction for eInclusion, A. Holzinger and K. Miesenberger (Eds.), Lecture Notes in Computer Science, Springer, 2009

[Karabetsos et al., 2008]
S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, and S. Raptis, **"HMM-based Speech**

Synthesis for the Greek Language," in Petr Sojka, Ivan Kopecek, and Karel Pala (eds.), Lecture Notes in Computer Science (LNCS), Springer – Verlag, 2008

9.3 Δημοσιεύσεις σε διεθνή συνέδρια με κριτές

[Raptis et al., 2010]
S. Raptis, A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, **"The ILSP Text-to-Speech System for the Blizzard Challenge 2010"**, In Proc. Blizzard Challenge 2010 Workshop, Kyoto, Japan, September 25, 2010

[Raptis et al., 2009a]
S. Raptis, P. Tsiakoulis, A. Chalamandaris and S. Karabetsos, **"High Quality Unit-Selection Speech Synthesis for Bulgarian"**, In Proc. 13th International Conference on Speech and Computer (SPECOM'2009), St. Petersburg, Russia, June 21-25, 2009

[Chalamandaris et al., 2009b]
A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and Spyros Raptis, **"An Efficient and Robust Pitch Marking Algorithm on the Speech Waveform for TD-PSOLA"**, in Proc. Intl. IEEE Conference on Signal and Image Processing Applications (ICSIPA), Malaysia, 2009

[Chalamandaris et al., 2009a]
A. Chalamandaris, P. Tsiakoulis, S. Raptis, and Sotiris Karabetsos, **"Design of an Efficient Corpus for High-Quality Unit Selection TTS for Bulgarian"**, in Proc. 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poland, 2009

[Tsiakoulis et al., 2008]
P. Tsiakoulis, A. Chalamandaris, S. Karabetsos and S. Raptis, **"A Statistical Method for Database Reduction for Embedded Unit Selection Speech Synthesis,"** in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Las Vegas, USA, 2008

[Dimou & Chalamandaris, 2008]
A. L. Dimou and A. Chalamandaris, **"Is idiom identification possible from prosodic information? An experimental approach for the Greek language"**, in Proc. 4th Intl Conf. Speech Prosody 2008, pp. 759-762 (2008)

[Chalamandaris et al., 2006]
A. Chalamandaris, A. Protopapas, P. Tsiakoulis and S. Raptis, **"All Greek to me! An automatic Greeklish to Greek Transliteration System,"** 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 24–26 May, 2006

[Dimou & Chalamandaris, 2006]
A. L. Dimou and A. Chalamandaris, "**Language identification from suprasegmental cues; examining the role of rhythm in the identification of a Greek dialect**", in Proc. La comunicazione parlata / Spoken Communication, February, , Italy, 2006

[Raptis & Chalamandaris, 2005]
S. Raptis and A. Chalamandaris, "**IMUTUS - Interactive Music Tuition System**", The 5th Open MusicNetwork Workshop, July 4-5, Vienna, Austria, 2005

[Chalamandaris et al., 2005]
A. Chalamandaris, S. Raptis and P. Tsiakoulis, "**Rule-based grapheme-to-phoneme method for the Greek**", in Proc. Interspeech'2005: 9th European Conference on Speech Communication and Technology, September 4-8, Lisbon, Portugal, 2005

[Raptis et al., 2005b]
S. Raptis, A. Askenfelt, D. Fober, A. Chalamandaris, E. Schoonderwaldt, S. Letz, A. Baxevanis, K. Falkenberg Hansen and Y. Orlarey, "**IMUTUS – An Effective Practicing Environment For Music Tuition**", International Computer Music Conference (ICMC 2005), September 5-9, Barcelona, Spain, 2005

[Chalamandaris et al., 2004b]
A. Chalamandaris, P. Tsiakoulis, S. Raptis and G. Giannopoulos, "**An Efficient and Robust Algorithm for Bypassing Greeklish**", in Proc. IC-SCCE 2004: 1st International Conference from Scientific Computing to Computational Engineering, 8-10 September, Athens, Greece, 2004

[Chalamandaris et al., 2004a]
A. Chalamandaris, P. Tsiakoulis, S. Raptis, G. Giannopoulos and G. Carayannis, "**Bypassing Greeklish!**", in Proc. LREC 2004: 4th International Conference on Language Resources And Evaluation, May 26-28, Lisbon, Portugal, 2004

[Giannopoulos et al., 2003]
G. Giannopoulos, A. Chalamandaris, S-E. Fotinea, T. Athanaselis, G. Carayannis, "**Analysis and modelling of the Carrier Declination for the Greek language**", in Proc. of the 15th International Congress of Phonetic Sciences - ICPhS03, 3-9 August 2003, Barcelona, pp. 555-558.

[Founda et al., 2001b]
M. Founda, A. Chalamandaris, G. Tambouratzis, and G. Carayannis, "**Studying the Factors Affecting the Optimal Unit Selection Algorithm for a TTS System for the Greek Language**", in Proceedings of the 4th European Conference on Noise Control EURONOISE2001, Patra, 14-17 January 2001, Vol. II, pp. 758-764.

[Founda et al., 2001a]
M. Founda, A. Chalamandaris, G. Tambouratzis, and G. Carayannis, "**Reducing Spectral**

Mismatches in Concatenative Speech Synthesis via Systematic Database Enrichment", in Proceedings of the Eurospeech-2001 Conference, Aalborg, Denmark, 4-7 September 2001, pp. 837-840.