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

Βελτίωση της ποιότητας συνθετικής φωνής και εφαρμογή σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και υπηρεσίες

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΣΩΤΗΡΙΟΣ Χ. ΚΑΡΑΜΠΕΤΣΟΣ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
& Μηχανικός Υπολογιστών Ε.Μ.Π. (2004)
Πτυχιούχος Ηλεκτρονικός Μηχανικός ΤΕΙ Αθήνας (2000)

Αθήνα, ΝΟΕΜΒΡΙΟΣ 2010

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

Βελτίωση της ποιότητας συνθετικής φωνής και εφαρμογή σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και υπηρεσίες

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΣΩΤΗΡΙΟΣ Χ. ΚΑΡΑΜΠΕΤΣΟΣ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
& Μηχανικός Υπολογιστών Ε.Μ.Π. (2004)
Πτυχιούχος Ηλεκτρονικός Μηχανικός ΤΕΙ Αθήνας (2000)

Τριμελής Συμβουλευτική Επιτροπή : καθ. Γεώργιος Καραγιάννης (επιβλέπων)
καθ. Πέτρος Μαραγκός
καθ. Νικόλαος Μήτρου

Επταμελής εξεταστική επιτροπή

.....	.....	.....
Γ. Καραγιάννης	Π. Μαραγκός	Ν. Μήτρου
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.
.....	.....	
Σ. Κόλλιας	Π. Τσανάκας	
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	
.....	.....	
Β. Μέρτζιος	Σ. Ράπτης	
Καθηγητής Δ.Π.Θ.	Ερευνητής Β' Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ) / Ε.Κ. "ΑΘΗΝΑ"	

Αθήνα, ΝΟΕΜΒΡΙΟΣ 2010

.....
ΣΩΤΗΡΙΟΣ Χ. ΚΑΡΑΜΠΕΤΣΟΣ

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

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΕΡΙΛΗΨΗ

Αντικείμενο της διδακτορικής διατριβής αποτελεί η τεχνολογία σύνθεσης φωνής από κείμενο. Έμφαση δίνεται στην μεθοδολογία σύνθεσης με επιλογή και ένωση ακουστικών μονάδων στο πεδίο του χρόνου (*unit selection concatenative speech synthesis*), εστιάζοντας κυρίως στον αλγόριθμο επιλογής ακουστικών μονάδων (*unit selection module*) και στην σχεδίαση της συνάρτησης του κόστους ένωσης. Εξετάζονται προσεγγίσεις που αφορούν τόσο την γενική περίπτωση συστήματος σύνθεσης φωνής από κείμενο (*general domain speech synthesis*), όσο και την περίπτωση προσαρμογής αυτής της τεχνολογίας σε περιβάλλον ενσωματωμένων συστημάτων με περιορισμένους υπολογιστικούς πόρους (*embedded speech synthesis*). Απώτερος στόχος είναι η βελτίωση της ποιότητας της συνθετικής ομιλίας με σκοπό την ευρεία υιοθέτηση συστημάτων σύνθεσης φωνής σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και τηλεπικοινωνιακές υπηρεσίες. Επιπρόσθετα, στην διατριβή εξετάζονται και νέες σύγχρονες εναλλακτικές παραμετρικές τεχνικές σύνθεσης φωνής (*statistical parametric speech synthesis*) για την περίπτωση της Ελληνικής γλώσσας. Πιο συγκεκριμένα, η διατριβή συνεισφέρει και πραγματεύεται τις ερευνητικές προσπάθειες στα εξής επιμέρους σημεία:

- Στην σχεδίαση και υλοποίηση του αλγόριθμου επιλογής ακουστικών μονάδων για γενικού σκοπού συστήματα Σύνθεσης Φωνής από κείμενο για την Ελληνική γλώσσα.

- Στην σχεδίαση, την αποδοτική αποκλιμάκωση και προσαρμογή συστήματος σύνθεσης φωνής από κείμενο με επιλογή και ένωση ακουστικών μονάδων, σε περιβάλλοντα περιορισμένων υπολογιστικών πόρων όπως είναι τα ενσωματωμένα συστήματα και ιδιαίτερα το περιβάλλον των κινητών τηλεφώνων.

- Στην υιοθέτηση και εφαρμογή ενός νέου μεθοδολογικού πλαισίου που βασίζεται σε δεδομένα (data driven), για την εκτίμηση και αποτίμηση των φασματικών ασυνεχειών που προκύπτουν στην ένωση των ακουστικών μονάδων και το οποίο μπορεί να επεκταθεί και στην συνολική συνάρτηση κόστους ένωσης ακουστικών μονάδων, προσφέροντας σημαντικά πλεονεκτήματα. Το νέο μεθοδολογικό πλαίσιο στηρίζεται σε τεχνικές μηχανικής μάθησης και συγκεκριμένα στην εφαρμογή ταξινομητών μιας τάξης (*one-class classification*).

- Στην μελέτη της παραμετρικής τεχνολογίας σύνθεσης φωνής από κείμενο με χρήση κρυφών Μαρκοβιανών μοντέλων (*HMM speech synthesis*) καθώς και στην υλοποίηση και προσαρμογή της στην περίπτωση της Ελληνικής γλώσσας. Η εν λόγω τεχνολογία δύναται να επιφέρει σημαντικά πλεονεκτήματα, τόσο για συστήματα γενικού σκοπού όσο και για ενσωματωμένα συστήματα.

Επιπλέον, περιγράφονται καινοτόμες εφαρμογές που έχουν σαν κύριο συστατικό την τεχνολογία σύνθεσης φωνής από κείμενο και απευθύνονται σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και τηλεπικοινωνιακές υπηρεσίες.

Λέξεις κλειδιά: Σύνθεση φωνής, Παραμετρική σύνθεση φωνής, ταξινόμηση μιας τάξης, αλγόριθμος επιλογής ακουστικών μονάδων, φασματική απόσταση, Κρυφά Μαρκοβιανά Μοντέλα, συναρτήσεις κόστους, τηλεπικοινωνιακές εφαρμογές, κινητό τηλέφωνο,

ABSTRACT

The subject of this thesis is speech synthesis technology and, in particular, the improvement of the quality of Text-to-Speech (TTS) systems for application in contemporary telecommunication environments and services. Emphasis is given in *Corpus-based Speech Synthesis* and in *Unit Selection Concatenative TTS* systems by focusing on the Unit Selection algorithm and the design of the cost functions which comprising it. Methods and approaches concerning the implementation of not only General Domain TTS systems, but also adapted scaled-down TTS systems for computational environments with limited resources and embedded systems in general, are explored and evaluated. In addition, contemporary Statistical Parametric Speech Synthesis based on Hidden Markov Models is explored for the case of the Greek language. More particularly, this thesis deals with research efforts and contributes to the following:

- The design and implementation of the unit selection algorithm for a general purpose Text-to-Speech system for the Greek language

- The design and implementation approaches for the efficient integration of Unit Selection technology in computational environments with limited resources, such as mobile devices, with no considerable speech quality degradation. In particular, the issues of database reduction, acoustic inventory compression and runtime computational load minimization are mainly addressed. Both objective and subjective assessments confirm the effectiveness of these approaches in terms of constructing a general purpose embedded unit selection TTS system and reducing the computational requirements while maintaining high speech quality.

- The introduction of one-class classification as a framework for the spectral join cost calculation in unit selection speech synthesis. A data-driven approach is adopted which exploits the natural similarity of consecutive speech frames in the speech database. Experimental results provide evidence on the effectiveness of the proposed method which clearly outperforms the conventional approaches currently employed. This method can be extended for designing the Join Cost function, offering many advantages.

- The adaption, implementation and the evaluation of a HMM speech synthesis framework for the case of the Greek language. This technology is capable of producing adequately natural speech in terms of intelligibility and intonation, offering many advantages and flexibility in constructing and manipulating general purpose TTS systems.

In addition, innovative applications for telecommunication systems and services are described, having TTS technology as a main component.

Key Words: Speech Synthesis, Parametric Speech Synthesis, One-Class classification, Unit Selection, Spectral distance, HMM, Cost functions, Join Cost, Target Cost, Telecommunication Systems, and Services.

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό, θα ήθελα να ευχαριστήσω όλους όσους συνέβαλλαν στην ολοκλήρωση αυτής της εργασίας. Ιδιαίτερα, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Γ. Καραγιάννη, για την ευκαιρία που μου παρείχε, την εμπιστοσύνη του και την άψογη επιστημονική του καθοδήγηση. Οφείλω να ευχαριστήσω τους καθηγητές μου, κ. Π. Μαραγκό και κ. Ν. Μήτρου που αποτελούν μέλη της συμβουλευτικής μου επιτροπής καθώς επίσης και τα μέλη της επταμελούς επιτροπής, καθηγητές κ. Σ. Κόλλια, κ. Π. Τσανάκα, κ. Β. Μέρτζιο και τον Ερευνητή Β' του ΙΕΛ/Ε.Κ. "ΑΘΗΝΑ" Δρ. Σ. Ράπτη.

Στα μέλη της ομάδας σύνθεσης φωνής του ΙΕΛ, Σπύρο Ράπτη, Αιμίλιο Χαλαμανδάρη, Πύρρο Τσιάκουλη, στην οποία έχω την ιδιαίτερη χαρά και τιμή να συμμετέχω, δεν οφείλω μόνο θερμές ευχαριστίες αλλά πολύ περισσότερα...

Η διδακτορική διατριβή δεν είναι μόνο αυτό, το τελικό κείμενο. Είναι μια διαδρομή, στην διάρκεια της οποίας διαμορφωνόμαστε, προσπαθούμε να γίνουμε καλύτεροι και, ίσως, να προσφέρουμε. Είναι μια πορεία στην οποία πορευόμαστε μόνοι μας. Είναι μια σύντομη ζωή.

Η πορεία αυτή αφιερώνεται στη μνήμη του Πατέρα μου.

Σωτήρης Καραμπέτσος

ΚΕΦΑΛΑΙΟ
-1-
ΕΙΣΑΓΩΓΗ

ΚΕΦΑΛΑΙΟ 1 – ΕΙΣΑΓΩΓΗ

Το κεφάλαιο αυτό αποτελεί αφενός εισαγωγή στην τεχνολογία σύνθεσης από κείμενο και τις εφαρμογές της και αφετέρου παρουσιάζει τόσο το αντικείμενο όσο και τους στόχους της διατριβής. Ειδικότερα, πραγματοποιείται μια εισαγωγική επισκόπηση στις υπάρχουσες προσεγγίσεις και τεχνικές της τεχνολογίας σύνθεσης φωνής από κείμενο. Δίνεται ιδιαίτερη έμφαση στις τεχνολογίες Επιλογής και Συρραφής (ή Ένωσης) Ακουστικών Μονάδων (unit selection speech synthesis) και Κρυφών Μαρκοβιανών Μοντέλων (HMM speech synthesis) οι οποίες αποτελούν τις επικρατέστερες προσεγγίσεις και εξετάζονται εκτενέστερα στην διατριβή. Έπειτα παρουσιάζονται τομείς στους οποίους η συνθετική ομιλία συναντά άμεση εφαρμογή. Στη συνέχεια, γίνεται μια σύνοψη των ερευνητικών ζητημάτων στην τεχνολογία σύνθεσης και παρουσιάζονται οι στόχοι και το αντικείμενο της διατριβής. Η ενότητα κλείνει με την οργάνωση της διατριβής.

1.1 ΤΕΧΝΟΛΟΓΙΕΣ ΣΥΝΘΕΣΗΣ ΦΩΝΗΣ

Η σύνθεση φωνής αποτελεί έναν από τους πρώτους τεχνολογικούς τομείς έρευνας και ανάπτυξης στην περιοχή της γλωσσικής επεξεργασίας και της επεξεργασίας σήματος, ενώ θεωρείται απαραίτητη συνιστώσα στην επικοινωνία ανθρώπου-μηχανής [O'Shaughnessy, 2003; Deng, 2005; Tomko, 2005; Moore, 2007; Mohasi, 2006]. Αν και ο όρος σύνθεση φωνής έχει ευρύτερη σημασία, συνήθως αναφέρεται στη δυνατότητα παραγωγής ή «κατασκευής» σήματος φωνής που αντιστοιχεί σε κάποιο κείμενο. Στην περίπτωση αυτή χρησιμοποιείται ο όρος **Σύνθεση Φωνής από Κείμενο** (Text-to-Speech) και αναφέρεται στην διαδικασία μετατροπής κειμένου σε συνθετική φωνή με την βοήθεια Η/Υ [Taylor, 2009; Benesty, 2008, Huang, 2001; Jurafsky, 2008; Dutoit, 1997]. Οι μέθοδοι παραγωγής συνθετικής ομιλίας διαφέρουν και εξαρτώνται κάθε φορά από την εκάστοτε εφαρμογή και τον σκοπό που εξυπηρετούν. Για παράδειγμα, κάποιο σύστημα θα μπορούσε να παράγει συνθετική ομιλία χρησιμοποιώντας κάποια από τις ακόλουθες λογικές:

- Αναπαραγωγή προηχογραφημένων μηνυμάτων
- Συρραφή (concatenation) προηχογραφημένων λέξεων
- Συρραφή προηχογραφημένων και κατάλληλα αποθηκευμένων στοιχειωδών ακουστικών μονάδων (ή τεμαχιδίων) ομιλίας (Diphone σύνθεση, Unit Selection σύνθεση). Τέτοιες μονάδες συνήθως είναι φωνήματα, διφωνήματα, τριφωνήματα κ.τ.λ. [Dutoit, 1997]
- Χρήση παραμετρικά ελεγχόμενου μοντέλου παραγωγής ομιλίας (π.χ. formant σύνθεση, LPC σύνθεση, HMM σύνθεση, articulatory σύνθεση) [Taylor, 2009]

Οι δύο πρώτες περιπτώσεις χαρακτηρίζονται από υψηλής ποιότητας σύνθεση αλλά βρίσκουν εφαρμογή μόνο σε περιπτώσεις και εφαρμογές που προϋποθέτουν περιορισμένο λεξιλόγιο. Οι επόμενες δύο είναι καταλληλότερες για γενικού σκοπού δυναμική σύνθεση και ορίζουν, όπως αναφέρθηκε, τα συστήματα που είναι γνωστά ως συνθέτες φωνής από κείμενο. Πράγματι, ο σκοπός ενός τέτοιου συστήματος

είναι να μπορεί να διαβάζει (με συνθετική φωνή) οποιοδήποτε κείμενο για τη γλώσσα που αναπτύχθηκε.

Στα σύγχρονα συστήματα σύνθεσης φωνής από κείμενο, είτε πρόκειται για συστήματα με επιλογή και συρραφή ακουστικών μονάδων είτε για συστήματα παραμετρικής σύνθεσης (κυρίως HMM σύνθεση), κοινό παρανομαστή αποτελεί η ύπαρξη μεγάλης βάσης δεδομένων προηχογραφημένης φυσικής ομιλίας. Τα συστήματα αυτά απαντώνται στην βιβλιογραφία με τον όρο **Corpus-based Speech Synthesis** και αποτελούν τα *συστήματα τρίτης γενιάς* [Benesty, 2008 (CH. 21), Taylor; 2009; Mobius, 2000; O'Shaughnessy, 2007; Campbell, 2005]. Επιτυγχάνουν αρκετά υψηλή ποιότητα, σχεδόν φυσική, ενώ μεθοδολογικά στηρίζονται πλήρως στα διαθέσιμα δεδομένα (data-driven) χαλαρώνοντας ή αποφεύγοντας εντελώς την απαίτηση χρήσης μοντέλων. Με άλλα λόγια, στηρίζονται στο γεγονός ότι τα ζητούμενα ποιοτικά χαρακτηριστικά της φωνής υπάρχουν διαθέσιμα στην βάση δεδομένων.

Στις τεχνολογίες δημιουργίας συνθετικής φωνής με χρήση προηχογραφημένων και κατάλληλα αποθηκευμένων στοιχειωδών ακουστικών μονάδων, συναντάμε τόσο την σύνθεση με διφωνήματα (**diphone synthesis**) όσο και την τεχνολογία σύνθεσης με επιλογή και ένωση ακουστικών μονάδων (**unit selection speech synthesis**).

Στη πρώτη περίπτωση, οι ακουστικές μονάδες είναι διφωνήματα. Ένα διφώνημα αποτελείται από δύο μισά φωνήματα. Τα όρια του διφωνήματος βρίσκονται στο μέσο της φασματικά σταθερής περιοχής κάθε φωνήματος. Η προηχογράφηση και αποθήκευση τους απαιτεί μία πραγμάτωση από κάθε επιτρεπτό συνδυασμό τους και ορίζει μια βάση δεδομένων η οποία περιέχει, εκτός άλλων, την χρονική οριοθέτηση τους. Η σύνθεση με διφωνήματα αποτελεί τεχνολογία *συστημάτων δεύτερης γενιάς* και σηματοδοτεί την αφετηρία της μεθοδολογικής προσέγγισης που στηρίζεται στα διαθέσιμα δεδομένα. Ωστόσο, σημαντικά μειονεκτήματα της τεχνολογίας είναι αφενός η ανάγκη για έντονες προσωδιακές τροποποιήσεις μέσω επεξεργασίας σήματος, με αποτέλεσμα την υποβάθμιση της ποιότητας, και αφετέρου η απαίτηση για επιλογή της κατάλληλης πραγμάτωσης κάθε διφώνου σε διάφορα επίπεδα προδιαγραφών, όπως για παράδειγμα το προσωδιακό περιβάλλον, το φασματικό ταίριασμα κτλ. [Beutnagel, 1998; Conkie, 1996; Dutoit, 1993; Dutoit, 1997, Chappell, 2002; Dixon, 1968; Campbell 1996, Iwahashi, 1993].

Στην δεύτερη περίπτωση, οι ακουστικές μονάδες προκύπτουν από την ηχογράφηση μεγάλου σώματος κειμένου και συναντώνται με περισσότερες από μια πραγματώσεις σε διαφορετικά φωνητικά και προσωδιακά περιβάλλοντα [O'Shaughnessy, 2007; Taylor, 2009; Hunt, 1996; Donovan, 1996]. Η τεχνολογία αυτή, η οποία όπως αναφέρθηκε εμπίπτει στα *συστήματα τρίτης γενιάς*, προέκυψε αφενός ως επέκταση και αφετέρου ως λογικό επακόλουθο για την άρση των περιορισμών που έθετε η σύνθεση με δίφωνα. Πράγματι, με την τεχνολογία επιλογής και ένωσης ακουστικών μονάδων υιοθετείται πλέον πλήρως το μεθοδολογικό πλαίσιο στήριξης στα διαθέσιμα δεδομένα (data-driven) αφού τα ποιοτικά χαρακτηριστικά και η μεταβλητότητα (variability) της φωνής σε διαφορετικά φωνητικά και προσωδιακά περιβάλλοντα (phonetic & prosodic context) εμπεριέχονται, κατά το δυνατό [Dutoit, 1997; Bozkurt, 2003], στην προηχογραφημένη βάση δεδομένων ενώ ταυτόχρονα περιορίζονται στο ελάχιστο οι προσωδιακές τροποποιήσεις και κατ' επέκταση η ανάγκη για έντονη επεξεργασία

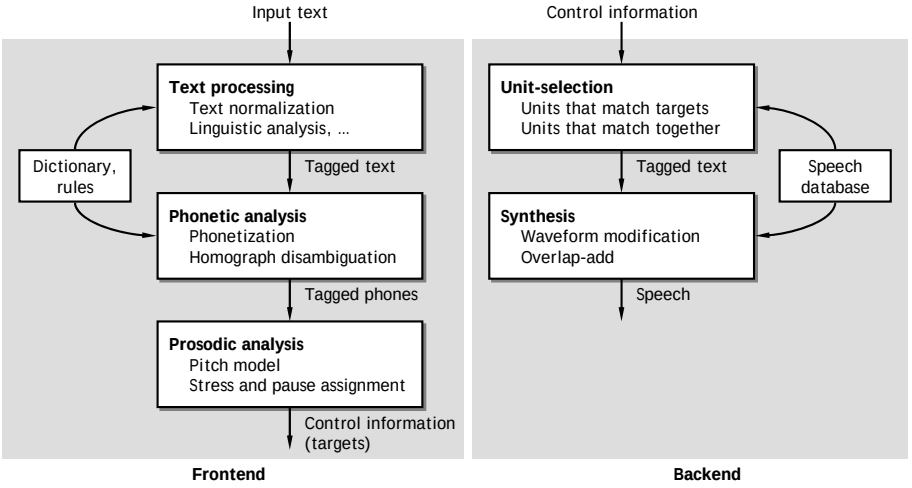
του σήματος. Το πρόβλημα πλέον έγκειται στην επιλογή και διαχείριση πλήθους πραγματώσεων από ακουστικές μονάδες με σκοπό να ικανοποιούνται οι επιμέρους προδιαγραφές που θέτει το κείμενο εισόδου, κυρίως σε επίπεδο προσωδίας, φυσικότητας κ.α.. Στα συστήματα με επιλογή και ένωση ακουστικών μονάδων, οι ακουστικές μονάδες μπορούν να είναι διαφορετικού τύπου, όπως φωνήματα, διφωνήματα κ.α. Επιπλέον, οι ακουστικές μονάδες μπορεί να είναι μεταβλητού μήκους με σκοπό την μείωση του αριθμού των απαιτούμενων συρραφών [Chu, 2001; Lee, 2003]. Το κυριότερο μειονέκτημα που παρουσιάζει η εν λόγω τεχνολογία, έγκειται στην περιορισμένη ευελιξία ως προς το συγκεκριμένο στυλ και ύφος ομιλίας που υιοθετείται και που ακολουθεί αυστηρά αυτό της ηχογραφημένης βάσης [Bailly, 2003; Taylor, 2009; Clark, 2007]. Παρά το μειονέκτημα αυτό, η τεχνολογία οδηγεί σε συστήματα σύνθεσης φωνής γενικού σκοπού (*general purpose* or *general domain speech synthesis systems*) υψηλής ποιότητας. Ιδιαίτερα δε σε περιπτώσεις ειδικού σκοπού (*limited domain*), όπως π.χ., καιρός, αθλητικά, η επιτευχθείσα ποιότητα είναι εξαιρετικά υψηλή [Black, 2000].

Αντίθετα, στην παραμετρική σύνθεση φωνής η παραγωγή του σήματος φωνής στηρίζεται στην υιοθέτηση κάποιου παραμετρικού μοντέλου και κατόπιν στον έλεγχο των παραμέτρων αυτού του μοντέλου είτε με κανόνες είτε με στατιστικό τρόπο. Σε γενικές γραμμές, αντιπρόσωποι της κατηγορίας αυτής είναι η σύνθεση με χρήση κανόνων στις φασματικές κορυφές (formants) του σήματος φωνής (**rule-based formant synthesis**) [Klatt, 1987], η σύνθεση με μοντελοποίηση του ανθρώπινου συστήματος παραγωγής ομιλίας (σύνθεση με αρθρωτές - **articulatory synthesis**) [Katsamanis, 2007] και, πρόσφατα, η στατιστικά ελεγχόμενη παραμετρική σύνθεση ή σύνθεση με χρήση Κρυφών Μαρκοβιανών Μοντέλων (**HMM-based speech synthesis**) [Zen, 2009]. Οι δύο πρώτες περιπτώσεις αποτέλεσαν τα συστήματα σύνθεσης *πρώτης γενιάς* αφού στηρίζονται στην αυστηρή μοντελοποίηση και την εύρεση (και κατασκευή) κανόνων. Η σύνθεση με χρήση Κρυφών Μαρκοβιανών Μοντέλων, η οποία αποτελεί τεχνολογία αιχμής με έντονο ερευνητικό ενδιαφέρον σε διεθνές επίπεδο, συμπεριλαμβάνεται στα συστήματα *τρίτης γενιάς* [Taylor, 2009]. Να τονιστεί ότι ο διαχωρισμός των συστημάτων γίνεται με βάση αφενός την τελική ποιότητα που επιτυγχάνουν και αφετέρου με τον τρόπο και την μεθοδολογία που αντιμετωπίζουν το πρόβλημα της σύνθεσης.

Το σχήμα 1.1 παρουσιάζει την γενική δομή ενός συστήματος σύνθεσης από κείμενο τρίτης γενιάς που βασίζεται στην επιλογή και συρραφή ακουστικών μονάδων. Αποτελείται από δύο βασικά υποσυστήματα, α) την γλωσσική επεξεργασία κειμένου (NLP: Natural Language Processing frontend) και β) την ψηφιακή επεξεργασία σήματος (DSP: Digital Signal Processing backend). Να σημειωθεί ότι η γενική δομή παραμένει ίδια και στην περίπτωση της παραμετρικής σύνθεσης φωνής με την διαφορά κυρίως στο υποσύστημα της ψηφιακής επεξεργασία σήματος.

Το υποσύστημα της γλωσσικής επεξεργασίας κειμένου είναι υπεύθυνο για την κατάλληλη μετατροπή, τον εμπλουτισμό και την συμβολική αναπαράσταση του κειμένου εισόδου ώστε να αυτό να διαβαστεί σωστά με συνθετική φωνή. Με άλλα λόγια παράγει την επαυξημένη αλληλουχία των φωνημάτων που πρέπει να παραχθούν. Στις βαθμίδες που το αποτελούν εμπλέκονται διάφορες γλωσσικές τεχνολογίες. Οι κυριότερες από αυτές περιλαμβάνουν, α) την επεξεργασία κειμένου (*text pre-processing*) [Dutoit, 1997], β) την φωνητική μεταγραφή (*grapheme to*

phoneme conversion or *letter to sound*) [Chalamandaris, 2005] και γ) την προσωδιακή ανάλυση (*Prosodic analysis and/or specifications*) [Taylor, 2009].



ΣΧΗΜΑ 1.1: Γενική δομή συστήματος Σύνθεσης Φωνής από Κείμενο με επιλογή και συρραφή ακουστικών μονάδων από βάση δεδομένων προηχογραφημένης φυσικής ομιλίας (Corpus-based unit selection concatenative speech synthesis).

Η επεξεργασία ή κανονικοποίηση κειμένου περιλαμβάνει και διαχειρίζεται κάθε μορφής κειμένου και γραφικών οντοτήτων όπως, αριθμοί, ημερομηνίες, ακρωνύμια, χαρακτήρες με ειδική σημασία κλπ., οι οποίες πρέπει να μετατραπούν σε προφορική μορφή. Επίσης, καθορίζει τα όρια των προτάσεων, λειτουργία ιδιαίτερα σημαντική αφού κάθε μετέπειτα επεξεργασία πραγματοποιείται σε επίπεδο πρότασης. Στην φωνητική μεταγραφή, ο γραπτός λόγος μετατρέπεται σε μια ενδιάμεση συμβολική αναπαράσταση, την φωνητική αναπαράσταση, που αντιστοιχεί στους στοιχειώδης ήχους που θα πρέπει να παραχθούν, δηλαδή την αντίστοιχη αλληλουχία των φωνημάτων. Η προσωδία είναι ένα σύνολο από γνωρίσματα της φωνής που περιλαμβάνει την μελωδία, τον ρυθμό, παύσεις, έμφαση σε λέξεις κλπ. Η φυσικότητα της παραγόμενης συνθετικής ομιλίας εξαρτάται σε μεγάλο βαθμό από την παραγωγή προσωδίας παρόμοιας με την ανθρώπινη φυσική προσωδία. Η προσωδιακή ανάλυση καθορίζει ή θέτει προδιαγραφές για την επιθυμητή προσωδία δηλαδή, τις διάρκειες, τον επιτονισμό, την ένταση κτλ. Το υποσύστημα της ψηφιακής επεξεργασία σήματος είναι υπεύθυνο για την παραγωγή του συνθετικού σήματος φωνής από την επαυξημένη συμβολική αναπαράσταση, δηλαδή είναι το τελικό στάδιο που μετατρέπει την αλληλουχία των φωνημάτων σε κυματομορφή φωνής με την ζητούμενη προσωδία. Οι κυριότερες βαθμίδες του υποσυστήματος είναι η μηχανή επιλογής των ακουστικών μονάδων από τη βάση δεδομένων [Hunt, 1996] και η βαθμίδα σύνθεσης ή παραγωγής της συνθετικής φωνής [Benesty, 2009 Ch 19]. Λεπτομέρειες σχετικά με αυτές τις βαθμίδες δίνονται σε επόμενες υποενότητες και κεφάλαια. Σε επίπεδο συστήματος, οι σχεδιαστικοί άξονες καθώς και οι κυριότεροι στόχοι αναφορικά με ένα σύστημα σύνθεσης φωνής από κείμενο, μπορούν να συνοψιστούν στα εξής κριτήρια, α) αριστοποίηση της ποιότητας της συνθετικής

φωνής, β) ελαχιστοποίηση της αλγοριθμικής πολυπλοκότητας, γ) ελαχιστοποίηση υπολογιστικού φορτίου και δ) οικονομία σε χρήση μνήμης. Φυσικά, αυτοί οι στόχοι είναι συχνά αντικρουόμενοι μεταξύ τους και οδηγούν σε συμβιβαστικές λύσεις

Στην παρούσα διατριβή επικεντρωνόμαστε στις τεχνολογίες τρίτης γενιά και συγκεκριμένα σε δύο τεχνικές σύνθεσης φωνής. Την τεχνική σύνθεσης με επιλογή και ένωση ακουστικών μονάδων όπου οι ακουστικές μονάδες είναι διφωνήματα και στην παραμετρική τεχνική σύνθεσης με χρήση Κρυφών Μαρκοβιανών Μοντέλων και στην προσαρμογή τους στην ελληνική γλώσσα. Στις υποενότητες που ακολουθούν περιγράφονται οι κυριότερες προσεγγίσεις για την παραγωγή συνθετικής φωνής δίνοντας έμφαση στις δύο παραπάνω τεχνικές.

1.1.1 Σύνθεση φωνής με μοντελοποίηση του ανθρώπινου συστήματος παραγωγής ομιλίας

Η μέθοδος σύνθεσης με μοντελοποίηση του ανθρώπινου συστήματος παραγωγής ομιλίας (articulatory synthesis) στηρίζεται στη φυσική του τρόπου παραγωγής φωνής. Στη μέθοδο αυτή γίνεται προσπάθεια μοντελοποίησης αφενός της κίνησης των ανθρώπινων οργάνων με χρήση χωροχρονικών συναρτήσεων που εξαρτώνται από φυσικές παραμέτρους όπως, θέση, ύψος, σχήμα, μεταβολές εμβαδού κ.α. και αφετέρου της γέννησης και διάδοσης του ήχου στη φωνητική οδό. Με άλλα λόγια μοντελοποιεί τους αρθρωτές του φωνητικού σωλήνα, και τις κινήσεις τους ενώ η παραγωγή φωνής γίνεται με αριθμητική επίλυση ακουστικών εξισώσεων. Γίνεται εμφανές ότι αυτή η μέθοδος αντιμετωπίζει δυσκολίες τόσο στη συγκέντρωση πρωτογενών δεδομένων όσο και στον ακριβή αναλυτικό υπολογισμό των συναρτήσεων. Ωστόσο, η συνεχής ραγδαία ανάπτυξη των υπολογιστικών συστημάτων αφήνει συνεχώς νέα περιθώρια για την περαιτέρω ανάπτυξη της [Coker, 1976; Maeda, 1982; Sondhi, 1987; Katsamanis, 2007]. Ως προσέγγιση θα μπορούσε να κατηγοριοποιηθεί μαζί με τις παραμετρικές τεχνικές, ωστόσο αντιμετωπίζεται ξεχωριστά λόγο της ιδιαιτερότητας με την οποία αντιμετωπίζει το πρόβλημα της σύνθεσης φωνής.

1.1.2 Σύνθεση φωνής με επιλογή και ένωση ακουστικών μονάδων

Η σύνθεση φωνής με επιλογή και συρραφή (ένωση) ακουστικών μονάδων αποτελεί την καθιερωμένη λύση για την δημιουργία σχεδόν φυσικών συνθετικών φωνών. Βασίζεται σε επιλογή και συρραφή ακουστικών μονάδων από βάση δεδομένων προηχογραφημένης φυσικής ομιλίας (Corpus-based Unit Selection Concatenative Speech Synthesis). Με αυτό τον τρόπο επιτυγχάνεται η διατήρηση των ποιοτικών χαρακτηριστικών της φωνής, αποφεύγοντας παράλληλα την αυστηρή μοντελοποίηση της. Αυτό οφείλεται στην πληθώρα των πραγματώσεων των ακουστικών μονάδων που προκύπτουν από την προηχογραφημένη βάση.

Η ιδέα της επιλογής και συρραφής διαφορετικών πραγματώσεων ακουστικών μονάδων είναι αρκετά παλιά. Στα [Harris, 1953; Peterson, 1958; Dixon, 1968; Olive, 1977; Olive, 1979] περιγράφονται ένα σύνολο από ιδέες που έθεσαν τα θεμέλια και συναντώνται και στα σημερινά συστήματα σύνθεσης. Ιδέες όπως η δημιουργία βάσης δεδομένων από διαφορετικές πραγματώσεις ακουστικών μονάδων, η χρήση δίφωνων, η ένωση των ακουστικών μονάδων συγκρίνοντας ποιοτικά τους

χαρακτηριστικά όπως η θεμελιώδης συχνότητα και οι φασματικές κορυφές (formants), η συσταδοποίηση ακουστικών μονάδων (clustering), είναι μερικά τέτοια παραδείγματα, βέβαια σε διαφορετικό υπολογιστικό πλαίσιο. Οι σύγχρονες προσεγγίσεις στην σύνθεση φωνής με επιλογή ακουστικών μονάδων, όπως οι διαφορετικού τύπου και οι μεταβλητού μήκους ακουστικές μονάδες, οι συναρτήσεις κόστους (ιδιαίτερα το κόστος 'στόχος'), πρωτοεμφανίστηκαν στο [Sagisaka, 1988; Sagisaka, 1992; Iwahashi, 1993]. Ωστόσο, το μεθοδολογικό πλαίσιο στο οποίο στηρίζεται η πλειοψηφία των σύγχρονων συστημάτων σύνθεσης φωνής εισήχθηκε φορμαλιστικά και εδραιώθηκε με την εργασία των Hunt και Black το 1996 [Hunt, 1996]. Στην εργασία αυτή παρουσιάζεται μια γενικευμένη προσέγγιση που σχετίζεται με τον μηχανισμό και την σχεδίαση του αλγόριθμου (ή αλλιώς την μηχανή) επιλογής ακουστικών μονάδων. Η μηχανή επιλογής ακουστικών μονάδων αποτελεί την καρδιά του συστήματος σύνθεσης, καθώς είναι υπεύθυνη για την επιλογή των κατάλληλων ακουστικών μονάδων από την βάση δεδομένων κατά την σύνθεση, που θα οδηγήσουν σε βέλτιστο ακουστικό αποτέλεσμα. Η λειτουργία του βασίζεται στην ελαχιστοποίηση σύνθετων συναρτήσεων κόστους, μέσω αλγορίθμων δυναμικού προγραμματισμού. Ο μηχανισμός υιοθετεί δύο συναρτήσεις κόστους οι οποίες αφορούν, α) το κόστος που αφορά την επιθυμητή ακολουθία ακουστικών μονάδων (**target cost**) και, β) το κόστος που αφορά τη σύνδεση και τη συνέχεια των ακουστικών μονάδων (**join** or **concatenation cost**). Οι συναρτήσεις κόστους και ο μηχανισμός επιλογής θα εξεταστούν αναλυτικά στο $2^{ο}$ κεφάλαιο της διατριβής. Σε γενικές γραμμές, η σχεδίαση των συναρτήσεων κόστους ενέχει σταθμισμένα κριτήρια που πρέπει να ελέγχονται και να πληρούνται κατά την σύνθεση και αφορούν τόσο το επίπεδο των προδιαγραφών που θέτει η προς σύνθεση ακολουθία, όσο και η αλληλουχία των ακουστικών μονάδων που θα επιλεγούν για συρραφή. Παραδείγματα τέτοιων κριτηρίων είναι ο εντοπισμός και η αποφυγή φασματικών ασυνεχειών κατά την συρραφή, το ταίριασμα των προσωδιακών χαρακτηριστικών, η στάθμιση των κριτηρίων κ.α. Επίσης σημαντικό παράγοντα αποτελεί ο σχεδιασμός των συναρτήσεων κόστους. Διάφορες προσεγγίσεις στα προηγούμενα ζητήματα παρουσιάζονται στα [Bulyko, 2001; Diaz, 2003; Diaza, 2006; Lee, 2003; Ding, 1998; Plumpe, 1998; Rouibia, 2005; Clark, 2007; Tihelka, 2007].

Εξελιγμένες ιδέες και τεχνικές σε αυτό το πλαίσιο, όπως είναι η σύνθεση με χρήση συσταδοποίησης ακουστικών μονάδων (clustered unit selection speech synthesis synthesis), οι οποίες δανείζονται και τεχνικές από την αναγνώριση φωνής, περιγράφονται στα [Donovan, 1996; Yi, 2003; Black, 1997; Black & Bennett, 2007; Clark, 2007]. Αυτή η περίπτωση διαφοροποιείται απο τις υπόλοιπες στο γεγονός οτι ομαδοποιεί τις ακουστικές μονάδες σε τάξεις (clustering), βάση πληροφοριών προσωδίας και περιβάλλοντος (context) απ' όπου προήλθαν. Οπότε, είναι δυνατό για το ίδιο διφώνημα ή άλλη μονάδα, να υπάρχουν πολλαπλές τάξεις όπου η κάθε μια από αυτές θα έχει συγκεκριμένα χαρακτηριστικά. Η μετέπειτα λογική είναι να ορίζονται οι συναρτήσεις κόστους ανάμεσα σε τάξεις ακουστικών μονάδων και όχι ανάμεσα σε μεμονωμένες πραγματώσεις τους. Αυτό έχει σαν αποτέλεσμα, σε μεγάλες βάσεις δεδομένων, να μειώνεται ο χρόνος και ο χώρος αναζήτησης.

Πολλές καινοτομίες και ιδέες στην αποδοτική σχεδίαση εισήχθησαν κατά την ανάπτυξη ολοκληρωμένων συστημάτων σύνθεσης φωνής από κείμενο. Μερικά από τα πιο αντιπροσωπευτικά παραδείγματα τέτοιων συστημάτων που αποτέλεσαν και τους πρώτους αντιπροσώπους της τεχνολογίας είναι τα εξής:

- ***CHATR*** [Black & Taylor, 1994]: το πρώτο σύστημα unit selection με χρήση φωνημάτων που αναπτύχθηκε στο ATR και προέκυψε ως εξέλιξη του v-talk [Sagisaka, 1992; Iwahashi, 1993].
- ***FESTIVAL*** [Black & Taylor, 1997, Clark, 2007]: αποτελεί σύστημα unit selection με χρήση διφωνημάτων και αναπτύχθηκε στο CSTR από τους Black και Taylor. Σε αυτό το σύστημα εφαρμόστηκαν οι ιδέες συσταδοποίησης [Black & Taylor, 1997] και αναζήτησης βέλτιστου σημείου ένωσης διφώνων [Conkie, 1996]. Πλέον, το festival διατίθεται ως ανοιχτή πειραματική πλατφόρμα ανάπτυξης συνθετών φωνής.
- ***AT&T NEXT-GEN*** [Syrdal, 2000; Beutnagel, 1998]: αποτελεί σύστημα unit selection με χρήση ημί-φωνημάτων και αναπτύχθηκε με βάση μια προηγούμενη έκδοση (*Flextalk*), το CHATR και το Festival. Στο σύστημα αυτό εισήχθη και η μέθοδος Harmonic plus Noise (HNM) [Stylianou, 2001] για την ένωση των ακουστικών μονάδων. Επίσης προτάθηκαν νέοι αλγόριθμοι για το ταχύτερο υπολογισμό των κριτηρίων στις συναρτήσεις κόστους και ιδιαίτερα σε αυτή του κόστους ένωσης [Beutnagel, 1999].
- ***BT LAURATE*** [Breen, 1998]: αποτελεί σύστημα unit selection με χρήση μεταβλητού μήκους ακουστικές μονάδες που αναπτύχθηκε από την British Telecom. Ιδιαίτερο χαρακτηριστικό του ήταν επίσης η χρήση μόνο φωνολογικών (phonological) κριτηρίων για την επιλογή των ακουστικών μονάδων.
- ***Microsoft WHISTLER*** [Huang, 2001]: αποτελεί σύστημα unit selection με χρήση ακουστικών μονάδων (senone) εξαρτώμενων από το περιβάλλον (context dependent) και οι οποίες ισοδυναμούν με μια κατάσταση (state) ενός τρι-φωνήματος που μοντελοποιείται με HMM. Κριτήριο επιλογής αποτέλεσε μια αντικειμενική συνάρτηση βασισμένη σε HMM scores σχετικά με την ένωση διαδοχικών τέτοιων ακουστικών μονάδων. Άλλη καινοτομία του συστήματος ήταν η χρήση στατιστικού μοντέλου βασισμένου στα δεδομένα (διαθέσιμη βάση δεδομένων φωνής) για την προσωδία.
- ***IBM*** [Donovan, 1996]: αποτελεί σύστημα unit selection με χρήση φωνημάτων. Εφαρμόζονται τεχνικές συσταδοποίησης με τη βοήθεια HMM και δένδρων απόφασης ώστε η βάση να αναλύεται και να αναπαριστάται από ομαδοποιημένα «φύλλα δένδρων» (leaf level representation of context dependent state-clustered HMMs) όπως προέκυψαν από την συσταδοποίηση. Κατά τη σύνθεση η ακολουθία φωνημάτων μετατρέπεται σε ακολουθία «φύλλων» και αλγόριθμος επιλογής πλέον εκτελείται στις συστάδες με τα «φύλλα δένδρων» με κριτήρια συνέχειας σε θεμελιώδη συχνότητα, διάρκεια και φάσμα. Η εκτίμηση των τελευταίων προκύπτει με ξεχωριστά μοντέλα για το καθένα.
- ***Lernout & Hauspie RealSpeak RSLAB*** [Coorman, 2000]: αποτελεί σύστημα unit selection με χρήση διφωνημάτων. Χρησιμοποιεί τεχνικές μάσκας (masking) στις συναρτήσεις κόστους με σκοπό την αποφυγή υψηλών τιμών σε τοπικά κόστη. Ουσιαστικά θέτει κατώφλι σε κάθε μετρική που χρησιμοποιεί στις συναρτήσεις κόστους.
- ***ATR XIMERA*** [Kawai, 2004; Iwahashi, 1993]: αποτελεί σύστημα unit selection με χρήση σύνθετων κριτηρίων στις συναρτήσεις κόστους και ιδιαίτερα στο target κόστος που χρησιμοποιεί ως χαρακτηριστικά την έξοδο από HMM σύνθεση. Στην πρώτη του έκδοση δινόταν ιδιαίτερη σε κριτήρια με έμφαση στις φασματικές αλλοιώσεις.

Τα σημερινά σύγχρονα συστήματα και οι τεχνικές που ακολουθούν είναι γνωστά κυρίως μέσου του διεθνή διαγωνισμού συστημάτων σύνθεσης φωνής Blizzard Challenge, ο οποίος διενεργείται κάθε χρόνο από το 2005 και έχει σαν στόχο την διερεύνηση τεχνολογιών και αλγορίθμων, μέσω της αξιολόγησης συστημάτων από διάφορους φορείς σε κοινές βάσεις δεδομένων [The Blizzard Challenge 2005, 2006, 2007, 2008, 2009, 2010]. Αντιπροσωπευτικά παραδείγματα, μεταξύ άλλων, αποτελούν τα συστήματα:

- ***IVONA*** [Kaszczuk, 2009]: αποτελεί εμπορικό σύστημα unit selection με χρήση διφωνημάτων, απλές συναρτήσεις κόστους αλλά ειδικά προσαρμοσμένες τεχνικές για προσωδιακές τροποποιήσεις κατά την παραγωγή του συνθετικού σήματος φωνής.
- ***TTS MARY*** [Schroeder, 2008]: αποτελεί σύστημα unit selection με χρήση διφωνημάτων και αναπτύχθηκε στο DFKI. Το MARY διατίθεται ως ανοιχτή πειραματική πλατφόρμα ανάπτυξης αλγορίθμων και συνθετών φωνής.
- ***iFLYTEK*** [Ling, 2007]: αποτελεί εμπορικό σύστημα unit selection με χρήση σύνθετων κριτηρίων στις συναρτήσεις κόστους οι οποίες βασίζονται σε τεχνικές μηχανικής μάθησης. Προσεγγίσεις με τεχνικές μηχανικής μάθησης για την σχεδίαση των συναρτήσεων κόστους ή επιμέρους στοιχείων τους, προσελκύουν έντονο ερευνητικό ενδιαφέρον στην παρούσα χρονική περίοδο.

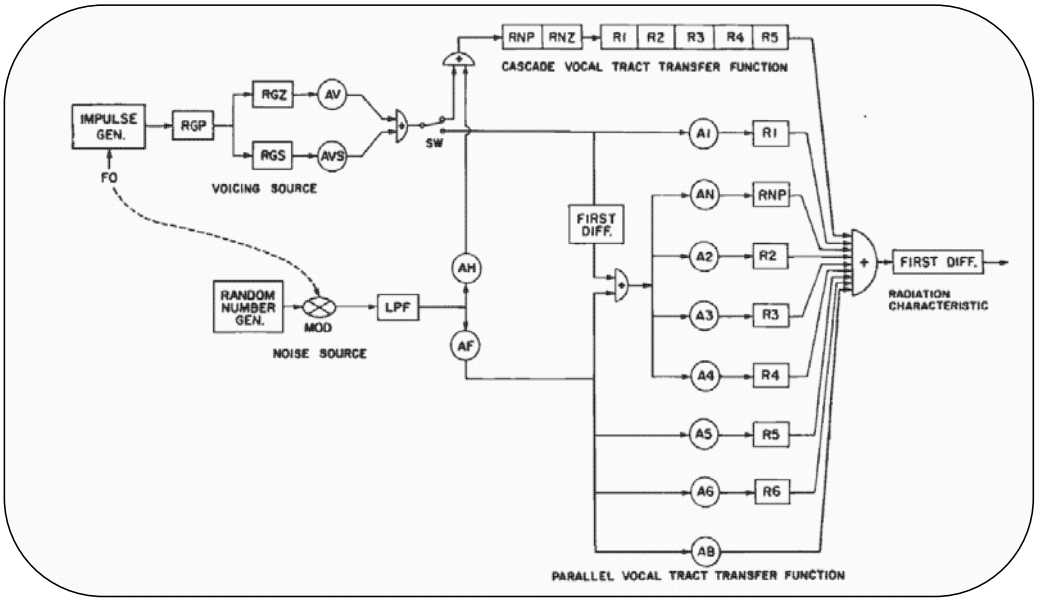
1.1.3 Παραμετρική σύνθεση φωνής

Όπως αναφέρθηκε, στην παραμετρική σύνθεση φωνής η παραγωγή του σήματος φωνής στηρίζεται στην υιοθέτηση κάποιου παραμετρικού μοντέλου και κατόπιν στον έλεγχο των παραμέτρων αυτού του μοντέλου είτε με κανόνες είτε με στατιστικό τρόπο. Σε σχέση με την τεχνολογία σύνθεσης με επιλογή και ένωση ακουστικών μονάδων, η παραμετρική σύνθεση δύναται να προσφέρει σημαντικά πλεονεκτήματα κυρίως λόγω της δυνατότητας που προσφέρει για εύκολη και ευέλικτη δημιουργία, τροποποίηση και διαχείριση, πλήθους ποιοτικών συνθετικών φωνών τόσο σε διαφορετικές γλώσσες όσο και σε διαφορετικά περιβάλλοντα και εφαρμογές [Zen, 2009; Black, 2007]. Πράγματι, εγγενείς περιορισμοί που προκύπτουν από τη τεχνολογία επιλογής και συρραφής ακουστικών μονάδων, όπως η έλλειψη ευελιξίας που προσφέρουν τόσο για την τροποποίηση του είδους και του τρόπου εκφώνησης όσο και για την κατασκευή νέων φωνών, αλλά και στην αυξημένη ανάγκη για υπολογιστικούς πόρους, μπορούν να ξεπεραστούν με την εδραίωση ενός αποδοτικού μεθοδολογικού πλαισίου παραμετρικής σύνθεσης. Το πιο επιτυχημένο μεθοδολογικό πλαίσιο παραμετρικής σύνθεσης είναι η σύνθεση φωνής από κείμενο με χρήση Κρυφών Μαρκοβιαννών Μοντέλων (HMM) το οποίο περιγράφεται παρακάτω μαζί με άλλες γνωστές προσεγγίσεις.

1.1.3.1 Σύνθεση με κανόνες

Η σύνθεση με έλεγχο των συντονισμών της φωνητικής οδού (formants) είναι ο κύριος εκπρόσωπος της σύνθεσης με κανόνες. Η μεθοδολογία σύνθεσης φωνής από κείμενο που βασίζεται στα formants (formant synthesis) αποτελεί παραμετρική τεχνική και στηρίζεται στο ευρύτερο μοντέλο πηγής-φίλτρου για το σήμα φωνής. Η τεχνική στηρίζεται στην εξαγωγή και χρήση κανόνων (rule based synthesis) που

αφορούν την λεπτομερή περιγραφή και ανάθεση τιμών σε ακουστικές παραμέτρους που συντελούν στην γέννηση του συνθετικού σήματος. Οι formant συνθέτες μοντελοποιούν τον φωνητικό σωλήνα και γενικά τα φαινόμενα της στοματικής κοιλότητας, με ένα σύνολο από φίλτρα δευτέρου βαθμού που μπορούν να συνδεθούν μεταξύ τους είτε σειριακά είτε παράλληλα. Κάθε φίλτρο αντιπροσωπεύει ένα formant ή κάποιο anti-formant στην περίπτωση ένρινων ήχων. Οι φωνητικές χορδές προσεγγίζονται από κάποιο περιοδικό σήμα που αποτελεί την πηγή. Το αποτέλεσμα της υπέρθεσης των φίλτρων με διέγερση το σήμα της πηγής αποτελεί το συνθετικό σήμα φωνής. Ο αριθμός των φίλτρων καθορίζει και τον αριθμό των formants που χρησιμοποιεί κάποιο σύστημα [Holmes, 1983; Klatt; 1987].



ΣΧΗΜΑ 1.2: Δομικό διάγραμμα της μηχανής σύνθεσης με βάση το μοντέλο του Klatt. Στο διάγραμμα φαίνονται τα φίλτρα (R - resonators) που μοντελοποιούν τα formants καθώς και οι παράμετροι που ελέγχουν την διαδικασία της σύνθεσης.

Στα πρώτα συστήματα οι εξαγωγή των κανόνων προέρχονταν από λεπτομερή και χειροκίνητη ανάλυση μεγάλου όγκου ηχογραφημένου σώμα κειμένου. Η ανάλυση αφορούσε τόσο την εξαγωγή και καταγραφή τιμών για τις ακουστικές παραμέτρους (π.χ. συχνότητα, πλάτος, εύρος ζώνης κτλ.) καθώς και την καταγραφή της χρονικής τους εξέλιξης σε ποικίλα περιβάλλοντα. Κατά την σύνθεση, κανόνες της μορφής IF-ELSE χρησίμευαν στην απόδοση τιμών για κάθε πλαίσιο (frame) συνθετικής φωνής. Το πιο γνωστό σύστημα σύνθεσης φωνής από κείμενο με βάση τα formants είναι η μηχανή Klatt [Klatt,1980 ; Klatt, 1987]. Το δομικό διάγραμμα της μηχανής φαίνεται στο σχήμα 1.2. Η μηχανή αποτελείται από φίλτρα τόσο σε σειριακή (cascade) όσο και παράλληλη σύνδεση (parallel). Όπως φαίνεται στο σχήμα, η μηχανή χρησιμοποιεί 6 formants (από τα αντίστοιχα φίλτρα που συμβολίζονται με R) και συνήθως λειτουργεί με συχνότητα δειγματοληψίας 8 και 10KHZ και με ρυθμό ανανέωσης τιμών στα πλαίσια ανά 10msec. Η είσοδος

αποτελείται από ένα σύνολο 39 παραμέτρων που αφορούν την περιγραφή των formants (συχνότητα, πλάτος, εύρος ζώνης) αλλά και τιμές που περιγράφουν την πηγή και την κατάσταση του ήχου (έμφωνο, άφωνο, τυρβώδες κτλ.) [Holmes, 1983; Klatt, 1987]. Στην περίπτωση της σειριακής σύνδεσης το πλάτος για κάθε formant καθορίζεται αυτόματα ενώ στην παράλληλη σύνδεση ελέγχεται εξωτερικά. Στο σύστημα χρησιμοποιούνται διάφορα μοντέλα πηγής όπως για παράδειγμα η πηγή Liljencrants-Fant (LF) που προσεγγίζει ικανοποιητικά την περιοδική διέγερση των φωνητικών χορδών [Fant, 1985].

Σε διεθνές επίπεδο, υπάρχει διαρκές ενδιαφέρον στην βελτιστοποίηση της σύνθεσης με βάση τα formants καθώς αποτελεί ένα ευέλικτο μοντέλο για την παραγωγή συνθετικής φωνής. Ως γνωστόν, αποτέλεσε πρόδρομο της μεθόδου σύνθεσης μέσω γραμμικής πρόβλεψης (***Linear Prediction Coding (LPC) synthesis***) που αν και τα formants δύναται να προσδιορίζονται αυτόματα, η τελική σύνθεση είναι χαμηλής ποιότητας ενώ παραμένει το πρόβλημα της μη αυτόματης εξαγωγής κανόνων. Η έρευνα στρέφεται κυρίως γύρω από συνιστώσες που αφορούν την επαρκή μοντελοποίηση των παραμέτρων της πηγής και των formants [Frölich, 2001; Vincent, 2005], όσο και σε θέματα υβριδικής χρήσης μεταξύ μηχανών σύνθεσης formant και επιλογής ακουστικών μονάδων [Carlson, 2005; Hertz, 2002]. Η εγγενείς δυσκολία στην πρωταρχική μορφή σύνθεσης με formants αφορά όχι τόσο στην δυνατότητα παραγωγής του σήματος φωνής από την παραμετρική του αναπαράσταση, αλλά από την παραγωγή και τον χειρισμό των ίδιων των παραμέτρων με τους κανόνες έτσι ώστε να πληρούν τις προδιαγραφές που θέτει το κείμενο. Πράγματι, ο μη αυτόματος προσδιορισμός κατάλληλων τιμών για τις παραμέτρους σε ποικίλα περιβάλλοντα παράλληλα με την εξαγωγή και χρήση κανόνων επιφέρει τεράστια πολυπλοκότητα και ξεπερνά ενδεχομένως τις ανθρώπινες δυνατότητες για τον χειρισμό τους. Το ζήτημα αυτό ξεπερνιέται με επιτυχία στη σύνθεση με χρήση Κρυφών Μαρκοβιανών Μοντέλων καθώς το πρόβλημα αντιμετωπίζεται από διαφορετική οπτική γωνία.

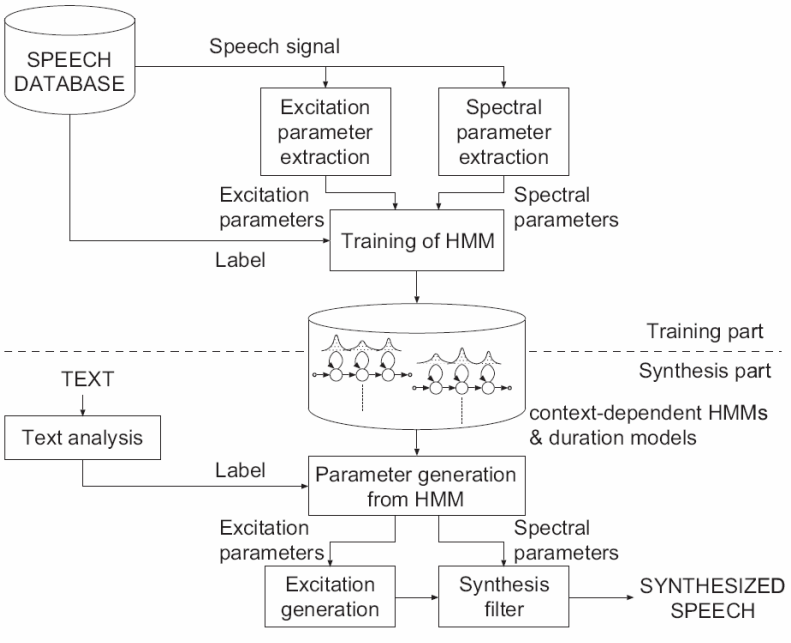
1.1.3.2 Σύνθεση με χρήση Κρυφών Μαρκοβιανών Μοντέλων

Η επιτυχία και η επικράτηση των Κρυφών Μαρκοβιανών Μοντελων (HMM) στο πεδίο της αναγνώρισης φωνής τείνει να επεκταθεί και στον τομέα της σύνθεσης φωνής καθώς τα HMM αποτελούν ένα γενικευμένο στατιστικό μοντέλο παραγωγής σήματος φωνής που προσφέρει την ευελιξία της παραμετρικής επεξεργασίας και διαχείρισης σε συνδυασμό με την επίτευξη ικανοποιητικής ποιότητας συνθετικής φωνής [Zen, 2009; Black, 2007; Taylor, 2009]. Σε γενικές γραμμές, η λειτουργία του στηρίζεται στην ανάλυση και την παραμετρική αναπαράσταση της φωνής, η οποία οδηγεί στην δυνατότητα στατιστικής μοντελοποίησης, με αποτέλεσμα να την καθιστά διαχειρίσιμη μέσω των Κρυφών Μαρκοβιανών Μοντέλων. Η ιδέα της εφαρμογής των HMM στο χώρο της σύνθεσης πρωτοεμφανίστηκε στο [Falaschi, 1989] αλλά αναπτύχθηκε ριζικά από τους Tokuda, Kobayashi και Imai [Tokuda, 1995], αν και ιδέες για την χρήση συσταδοποίησης μέσω δένδρων απόφασης είχαν εφαρμοστεί, όπως είδαμε, και στην τεχνολογία επιλογής και συρραφής ακουστικών μονάδων [Donovan, 1996].

Ειδικότερα, στο σχήμα 1.3, φαίνεται το δομικό διάγραμμα ενός συστήματος σύνθεσης φωνής από κείμενο με χρήση HMM. Όπως αναφέρθηκε, το σύστημα στηρίζεται στα διαθέσιμα δεδομένα (data driven) με την έννοια ότι χρειάζεται

εκπαίδευση από ήδη υπάρχουσα επισημειωμένη βάση δεδομένων φυσικής ομιλίας. Έτσι, το σύστημα χωρίζεται σε δύο στάδια: το στάδιο της εκπαίδευσης και το στάδιο της σύνθεσης [Zen, 2007].

Το στάδιο της εκπαίδευσης μοιάζει στην μεθοδολογία με αυτό της αναγνώρισης φωνής με την διαφορά ότι πραγματοποιείται από κοινού μοντελοποίηση τόσο της φασματικής όσο και της προσωδιακής πληροφορίας [Yoshimura, 1999]. Αναλυτικότερα, από την βάση δεδομένων εκπαίδευσης εξάγονται πληροφορίες που αφορούν την πηγή (προσωδία) και τον φωνητικό σωλήνα (φασματική αναπαράσταση), με τις οποίες εκπαιδεύονται τα μοντέλα των HMM. Με αυτό τον τρόπο, δημιουργούνται εξαρτώμενα από το «περιβάλλον» μοντέλα HMM (context dependent HMM) όπου μοντελοποιούν το σήμα φωνής σαν μια σειρά από τυχαίες παρατηρήσεις και εναλλαγές καταστάσεων (υπό περιορισμούς ως προς τις δυναμικές μεταβολές ώστε να πληρείται η συνέχεια [Tokuda, 1995; Tokuda, 2000; Toda, 2005]) που περιγράφονται με στατιστικό τρόπο. Έτσι, το τελικό σήμα φωνής αναπαρίσταται από την συνένωση των HMM ανά φώνημα [Zen, 2009; Zen, 2007].



ΣΧΗΜΑ 1.3: Δομικό διάγραμμα συστήματος σύνθεσης φωνής από κείμενο με χρήση HMM [Zen, 2007].

Κατά το στάδιο της σύνθεσης, αρχικά το κείμενο εισόδου αναλύεται και μετατρέπεται σε συμβολική αναπαράσταση που περιέχει πληροφορίες σχετικές με το «περιβάλλον» και κατόπιν πραγματοποιείται η συνένωση των HMM βάση αυτής της αναπαράστασης. Έπειτα, υπολογίζονται οι διάρκειες ανά κατάσταση των HMM μέσω των κατανομών πιθανότητας. Το επόμενο βήμα περιλαμβάνει την παραγωγή των διανυσμάτων για το φάσμα και την θεμελιώδη συχνότητα μέσω του μοντέλου και με βάση την μεγιστοποίηση των πιθανοτήτων (maximum likelihood) τους για την συγκεκριμένη σύνθεση, δηλαδή παράγεται ο συρμός διανυσμάτων των οποίων η πιθανότητα είναι μέγιστη δεδομένου των καταστάσεων και της σειράς των HMM για την προς σύνθεση πρόταση [Tokuda, 2000]. Τέλος, σύμφωνα με το μοντέλο

πηγής-φίλτρου, το συνθετικό σήμα φωνής προκύπτει μέσω του φίλτρου σύνθεσης. Συνήθως, το φίλτρο που χρησιμοποιείται είναι το MLSA (Mel log spectrum approximation filter for mel-cepstral coefficients) [Tokuda, 2000], ενώ εφαρμόζονται διάφορες τεχνικές για την διέγερση [Maia, 2007].

Το μοντέλο του σχήματος είναι αρκετά γενικευμένο και εύκολα προσαρμόσιμο σε άλλες γλώσσες. Συστήματα σύνθεσης με HMM έχουν ήδη αναπτυχθεί με επιτυχία για διάφορες γλώσσες όπως, Ιαπωνικά, Γερμανικά, Αγγλικά, Σουηδικά, Αραβικά, Κινέζικα, Κορεάτικα κ.α. [Abdel-Hamid, 2006; Gonzalvo, 2007; Krstulovic, 2007; Maia, 2003; Qian, 2006; Tokuda, 2002; Vesnicer, 2004]. Η εφαρμογή του μοντέλου σε κάποια γλώσσα εξαρτάται κυρίως από την σωστή αντιμετώπιση και μοντελοποίηση της ποικιλομορφίας του «περιβάλλοντος».

Το πιο γνωστό και αντιπροσωπευτικό παράδειγμα συστήματος σύνθεσης φωνής με HMM είναι το **HTS** [Zen, 2007; Yamagishi, 2007], το οποίο διατίθεται ως ανοιχτή πειραματική πλατφόρμα ανάπτυξης συνθετών φωνής και στο χώρο της σύνθεσης φωνής είναι πλέον τόσο δημοφιλές όσο και το **HTK** [Gales, 2007] για την αναγνώριση φωνής. Να σημειωθεί ότι η πλατφόρμα HTS χρησιμοποιεί το HTK για την εκπαίδευση των HMM. Στο ευρύτερο πλαίσιο της στατιστικής παραμετρικής σύνθεσης (στατιστικά ελεγχόμενη παραμετρική σύνθεση), παρόμοια προσέγγιση με τα HMM ακολουθείται στο σύστημα που παρουσιάζεται στο [Black, 2006] και ονομάζεται **CLUSTERGEN** και ενσωματώνεται στο Festival. Το σύστημα αυτό δεν χρησιμοποιεί HMM παρά μόνο έμμεσα για την αυτόματη κατάτμηση και επισημείωση της βάσης, αλλά χρησιμοποιεί δένδρα απόφασης (τύπου CART) για την εξαρτώμενη από το περιβάλλον (context-dependent) συσταδοποίηση κάθε διανύσματος χαρακτηριστικών. Ουσιαστικά φτιάχνεται ένα δένδρο απόφασης για κάθε ομάδα από διανύσματα χαρακτηριστικών που επισημειώθηκαν ότι ανήκουν στην ίδια HMM κατάσταση (same HMM state name). Παρόμοια λογική ακολουθείται κατά τη σύνθεση, στην οποία γίνεται εκτίμηση της κατάστασης (HMM state) και ανάσυρση του μέσου διανύσματος χαρακτηριστικών από κάθε φύλλο του δένδρου στο οποίο οδηγούν οι ερωτήσεις [Black, 2006].

Παρά τη σημαντική πρόοδο που έχει ήδη σημειωθεί, οι παραγόμενες φωνές μέσω HMM τείνουν να παρουσιάζουν μια μη αμελητέα υποβάθμιση στην ποιότητά τους, εντείνοντας την ανάγκη για περαιτέρω έρευνα. Σε επίπεδο τεχνολογίας η κυριότερες πηγές υποβάθμισης της τελικής ποιότητας οφείλονται αφενός στη δημιουργία του συνθετικού σήματος μέσω του μοντέλου πηγής-φίλτρου και αφετέρου λόγω της στατιστικής διαχείρισης της σύνθεσης την οποία ελέγχουν στατιστικά ασφαλή μοντέλα που συνήθως δεν αποτυπώνουν ούτε αναγεννούν πιο σύνθετα και πλούσια φαινόμενα που συναντώνται στη φωνή. Προς τη πρώτη κατεύθυνση έχουν ήδη εφαρμοστεί προχωρημένες τεχνικές vocoding όπως είναι η μέθοδος STRAIGHT [Kawahara, 1999] ενώ προς τη δεύτερη εφαρμόζονται τεχνικές όπως αυτές που περιγράφονται στα [Tokuda, 2003; Zen, 2004; Toda, 2005]. Παρά τις προηγούμενες δυσκολίες, το σημαντικότερο πλεονέκτημα είναι ότι με την προσέγγιση της παραμετρικής σύνθεσης αναδύονται δυνατότητες για προχωρημένες τεχνολογίες και εφαρμογές όπως,

- Εύκολη διαχείριση και προσαρμογή μοντέλων του στυλ και της προσωδίας της φωνής [Yamagishi & Kobayashi, 2007; Tachibana, 2006; Qin, 2006]

- Ρύθμιση των χαρακτηριστικών της φωνής, μετασχηματισμός φωνής με απουσία παράλληλων δεδομένων και δυνατότητα γρήγορης και εύκολης κατασκευής φωνών [Yamagishi, 2003; Yamagishi & Zen, 2007; Yamagishi, 2009]
- Προσαρμογής ομιλητή (speaker adaptation) και δημιουργία πολλαπλών φωνών με τρόπο αυτόματο με λιγοστά διαθέσιμα δεδομένα [Yamagishi, 2009; King, 2008; Gibson, 2009]. Η δυνατότητα αυτή συναντάται με τον όρο εύρωστη και μη-εποπτευόμενη σύνθεση φωνής (*unsupervised speaker adaptive robust speech synthesis*).
- Εφαρμογή και προσαρμογή σε όλες τις γλώσσες με δυνατότητα διαγλωσσικής σύνθεσης [Tokuda, 2004; Latorre, 2006]
- Χαμηλές απαιτήσεις σε υπολογιστικούς πόρους [Kim S.-J., 2006]

Περισσότερες λεπτομέρειες για την τεχνολογία σύνθεσης με HMM θα δοθούν σε επόμενο κεφάλαιο της διατριβής.

1.1.3.3 Υβριδικές τεχνικές

Οι υβριδικές τεχνικές αναφέρονται σε προσπάθειες αποδοτικού συνδυασμού των υπαρχόντων προσεγγίσεων με στόχο την εκμετάλλευση των πλεονεκτημάτων που προσφέρει η καθεμία. Οι γνωστότερες υβριδικές τεχνικές αφορούν προσπάθειες ενοποίησης: α) της σύνθεσης με formants με την βοήθεια HMM [Acero, 1999], της σύνθεσης με formants και της σύνθεσης με επιλογή και συρραφή ακουστικών μονάδων [Hertz, 2002; Carlson, 2005], γ) της articulatory σύνθεσης με την βοήθεια HMM [Nakamura, 2006, Toda, 2008] και δ) της σύνθεσης με HMM και της σύνθεσης με επιλογή και συρραφή ακουστικών μονάδων [Taylor, 2006].

1.2 ΕΦΑΡΜΟΓΕΣ ΣΥΝΘΕΣΗΣ ΦΩΝΗΣ

Τα τελευταία χρόνια, η σύνθεση φωνής από κείμενο έχει αξιοποιηθεί σε πληθώρα εφαρμογών. Βασικό κριτήριο για την ευρύτερη αποδοχή της τεχνολογίας αυτής σε προϊόντα και υπηρεσίες, αποτελεί η δυνατότητα να επιτυγχάνει υψηλό βαθμό ποιότητας (φυσικότητα και καταληπτότητα). Εφαρμογές της σύνθεσης φωνής σε διάφορα πεδία περιγράφονται στα [Gilbert, 2008; Deng, 2005; Duggan, 2003; Raptis, 2005; ; Raptis, 2009; Chalamandaris, 2009; Chalamandaris, 2010; O'Shaughnessy, 2003; Tomko, 2005; Moore, 2007; Peters, 2004; Mohasi, 2006; Eskenazi, 2009; Denby, 2010]. Ενδεικτικά, πεδία εφαρμογής της τεχνολογίας σύνθεσης φωνής είναι οι,

- τηλεπικοινωνιακές εφαρμογές και υπηρεσίες (τηλεφωνικά κέντρα, διαδίκτυο κα.).
- εκπαιδευτικές εφαρμογές (π.χ., εκπαίδευση μέσω υπολογιστή CALL: Computer aided language learning).
- εφαρμογές προσβασιμότητας για άτομα με ειδικές ανάγκες (αναγνώστες οθόνης, ομιλούντες ιστοτόποι κ.α.).
- ψυχαγωγικές υπηρεσίες (audio books, παιχνίδια, video games, εικονική πραγματικότητα, μεταγλώττιση κτλ.).
- αλληλεπιδραστικές εφαρμογές μεταξύ ανθρώπου και υπολογιστών (π.χ., πολυτροπική αλληλεπίδραση, avatars, ρομποτική).
- υπηρεσίες στη βασική και εφαρμοσμένη έρευνα κ.α. (π.χ., ακουστική ανάλυση, silent speech interfaces δηλαδή συστήματα διεπαφών φωνής χωρίς την παρουσία

ακουστικού σήματος, όπως για παράδειγμα παραγωγή φωνής με καταγραφή εικόνων video κ.α.) [Denby, 2010] .

Μερικές από τις παραπάνω εφαρμογές περιγράφονται και σε επόμενο κεφάλαιο της διατριβής.

1.3 ΕΡΕΥΝΗΤΙΚΑ ΖΗΤΗΜΑΤΑ

Μερικά από τα ερευνητικά θέματα που απασχολούν το σύνολο της ερευνητικής κοινότητας της σύνθεσης φωνής μπορούν να κατηγοριοποιηθούν σε αυτά που είναι ανεξάρτητα της εκάστοτε τεχνολογίας και σε αυτά που ενέχονται στην εκάστοτε τεχνολογία. Στην πρώτη περίπτωση συναντάμε ζητήματα που αποτελούν συνεχή στόχο στον τομέα της σύνθεσης φωνής. Μερικά από αυτά είναι,

- Η πολυτροπική σύνθεση φωνής (*multimodal speech synthesis*) [Tao, 2009; Huang, 2009]
- Η σύνθεση εκφραστικής/συναισθηματικής ομιλίας (*expressive/emotional speech synthesis*) [Schroeder, 2009]
- Η εύρωστη παραμετρική/υβριδική σύνθεση φωνής (*robust statistical parametric / hybrid concatenative-parametric speech synthesis*) [Zen, 2009; Taylor, 2006]
- Ο μετασχηματισμός και η προσαρμογή φωνής και ομιλητή (*Voice Transformation and Conversion*) [Stylllianou, 2009; Mouchtaris, 2006; Toda, 2007; Yamagishi, 2009b]
- Η μείωση των υπολογιστικών αναγκών και των αποθηκευτικών πόρων [Bailly, 2003]

Αναφορικά με τις επιμέρους τεχνολογίες, όπως τις συναντήσαμε στις προηγούμενες υποενότητες, τα ερευνητικά ζητήματα είναι πολλά και εξαρτώνται τόσο από την εκάστοτε τεχνολογική προσέγγιση όσο και από τον τομέα εφαρμογής. Πολλές πληροφορίες μπορούν να αναζητηθούν στα [Taylor, 2009; Benesty, 2008; Jurafsky, 2008; Huang, 2001; Bailly, 2003, Mobius, 2000]. Μερικά από αυτά, τα οποία εξετάζονται και στα πλαίσια της παρούσης διατριβής, παρουσιάζονται στη συνέχεια.

Στην τεχνολογία με επιλογή και συρραφή ακουστικών μονάδων, σημαντικό ζήτημα αποτελεί η σχεδίαση των συναρτήσεων κόστους καθώς και η επιλογή αλλά και ο τρόπος εφαρμογής των κριτηρίων από τα οποία αποτελούνται αυτές οι συναρτήσεις. Ειδικότερα, αναφορικά με το κόστος ένωσης ιδιαίτερο ενδιαφέρον παρουσιάζουν αποδοτικές μεθοδολογίες αναπαράστασης και τεχνικές εξαγωγής χαρακτηριστικών σε συνδυασμό με την εφαρμογή κατάλληλων μετρικών για τον προσδιορισμό και την εκτίμηση ασυνεχειών σε συνθετικά σήματα φωνής [Klabbers, 2001; Vepa, 2004; Vepa, 2006; Bellegarda, 2006]. Επιπλέον, σημαντική συνιστώσα αποτελεί η συνολική σχεδίαση του κόστους ένωσης για τον αλγόριθμο βέλτιστης επιλογής ακουστικών μονάδων [Hunt, 1996; Lee, 2001; Peng, 2002; Sakai, 2005; Syrdal, 2004; Toda, 2006; Tihelka, 2007].

Επίσης, η μείωση των απαιτήσεων σε υπολογιστικούς πόρους με ταυτόχρονη διατήρηση της υψηλής ποιότητας είναι κύριο μέλημα στην εν λόγω τεχνολογία σε κάθε τομέα εφαρμογής της. Στην παρούσα διατριβή αντιμετωπίζουμε το ζήτημα

αυτό για την περίπτωση της προσαρμογής της τεχνολογίας σε περιβάλλον ενσωματωμένων συστημάτων [Schnell, 2002].

Στην παραμετρική σύνθεση με HMM, πέρα από τους γνωστούς περιορισμούς όπως είναι για παράδιγμα η αναπαράσταση και ο μηχανισμός παραγωγής του σήματος φωνής, που οδηγούν σε υποβάθμιση της ποιότητας, ενδιαφέρον παρουσιάζει η εφαρμογή της τεχνολογίας σε κάθε γλώσσα λαμβάνοντας υπόψη τα χαρακτηριστικά και τις ιδιαιτερότητες της [Tokuda, 2004].

1.4 ΣΤΟΧΟΙ ΚΑΙ ΣΥΝΕΙΣΦΟΡΑ ΤΗΣ ΔΙΑΤΡΙΒΗΣ

Η φωνή αποτελεί τον φυσικότερο και τον αμεσότερο τρόπο στην ανθρώπινη επικοινωνία. Για το λόγο αυτό θεωρείται και απαραίτητη συνιστώσα στις σύγχρονες διεπαφές επικοινωνίας ανθρώπου-μηχανής. Στην ενότητα 1.2 σκιαγραφήσαμε μέρος του συνόλου των εφαρμογών που συνθέτουν την ολοκλήρωση της τεχνολογίας σύνθεσης φωνής σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και υπηρεσίες. Ωστόσο, όπως τονίστηκε στην ίδια ενότητα, η ευρύτερη αποδοχή της τεχνολογίας σύνθεσης φωνής από κείμενο σε κάθε είδους εφαρμογή, εξαρτάται κατά κύριο λόγο από την τελική ποιότητα που επιτυγχάνει. Η τελική ποιότητα χαρακτηρίζετε τόσο από την φυσικότητα όσο και από την καταληπτότητα. Όπως αναφέρει ο Dutoit στο [Benesty, 2009, Ch. 21], σε σχέση με το αντίστροφο πρόβλημα, δηλαδή αυτό της αναγνώρισης φωνής, η κριτική στην περίπτωση της σύνθεσης φωνής είναι συνήθως πιο αυστηρή και δύσκολα είναι επιεικής σε περιπτώσεις με αστοχίες. Η επίτευξη υψηλής τελικής ποιότητας σε ένα σύστημα σύνθεσης φωνής από κείμενο είναι συνάρτηση πολλών παραγόντων και εξαρτάται τόσο από την ίδια την τεχνολογία που χρησιμοποιείται όσο και από τις διάφορες βαθμίδες οι οποίες απαρτίζουν το εκάστοτε σύστημα. Αντικείμενο της διδακτορικής διατριβής αποτελεί η τεχνολογία σύνθεσης φωνής από κείμενο και η βελτίωση της ποιότητας της με σκοπό την ευρεία υιοθέτηση της σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και τηλεπικοινωνιακές υπηρεσίες. Παράδειγμα αυτών αποτελεί η ολοκλήρωση της τεχνολογίας τόσο σε **επίπεδο συσκευής** (π.χ., κινητό τηλέφωνο, PDA κ.α.) όσο και σε **επίπεδο δικτύου** και **διαδικτυακών υπηρεσιών** (π.χ., φωνητικά ηλεκτρονικά μηνύματα, ομιλούντες ιστοσελίδες κ.α.)

Ειδικότερα, ιδιαίτερη έμφαση δίνεται στην μεθοδολογία σύνθεσης με επιλογή και ένωση ακουστικών μονάδων στο πεδίο του χρόνου, εστιάζοντας κυρίως στον αλγόριθμο επιλογής ακουστικών μονάδων (***unit selection engine***) και στην σχεδίαση της συνάρτησης του κόστους ένωσης (***join*** or ***concatenation cost function***). Όπως θα δείξουμε στο 2^{ο} κεφάλαιο, η συνάρτηση του κόστους ένωσης είναι υπεύθυνη για τον έλεγχο της συρραφής (ένωσης) και της συνέχειας των υποψήφιων πραγματώσεων των ακουστικών μονάδων. Η συνάρτηση αποτελείται από διάφορα κριτήρια μεταξύ των οποίων τα συνηθέστερα που εξετάζονται είναι η θεμελιώδης συχνότητα, η ενέργεια και η φασματική συνέχεια [Hunt, 1996]. Η τεχνολογία σύνθεσης φωνής με επιλογή και συρραφή εξετάζεται υπό το πρίσμα συστημάτων που αφορούν τόσο την γενική περίπτωση, η οποία δεν προβάλλει ιδιαίτερους υπολογιστικούς περιορισμούς (όπως συστήματα σε προσωπικούς Η/Υ και σε εξυπηρετητές), όσο και την περίπτωση ολοκλήρωσης και προσαρμογής της σε περιβάλλον ενσωματωμένων συστημάτων (embedded systems), όπως είναι το περιβάλλον της συσκευής του κινητού τηλεφώνου.

Επίσης, στα πλαίσια της διατριβής εξετάζονται και οι νέες σύγχρονες εναλλακτικές παραμετρικές τεχνικές σύνθεσης φωνής όπως είναι η προσαρμογή της τεχνολογίας σύνθεσης φωνής με χρήση Κρυφών Μαρκοβιανών Μοντέλων για την περίπτωση της Ελληνικής γλώσσας. Τέλος, παρουσιάζονται κάποιες τηλεπικοινωνιακές εφαρμογές και τηλεπικοινωνιακές υπηρεσίες με κύριο συστατικό την τεχνολογία σύνθεσης φωνής.

Η συνεισφορά και οι ερευνητικές προσπάθειες της διατριβής επικεντρώνονται και συνοψίζονται στους εξής άξονες:

- **Σχεδίαση και υλοποίηση του αλγόριθμου επιλογής ακουστικών μονάδων για γενικού σκοπού συστήματα Σύνθεσης Φωνής από κείμενο**

  Οι ερευνητικές προσπάθειες αφορούν την σχεδίαση και την υλοποίηση του αλγόριθμου επιλογής ακουστικών μονάδων (Unit Selection Algorithm) για γενικού σκοπού σύστημα σύνθεσης φωνής από κείμενο για την Ελληνική γλώσσα. Ειδικότερα, περιγράφονται οι συναρτήσεις κόστους που αποτελούν τον αλγόριθμο και που είναι υπεύθυνες για τις ακουστικές μονάδες που τελικά θα επιλεγούν για την σύνθεση. Παράλληλα, εξηγούνται οι επιμέρους συναρτήσεις κόστους και τα κριτήρια που επιλέχτηκαν για να τις απαρτίσουν, και αφορούν πλήθος από χαρακτηριστικά που κρίνονται απαραίτητα στην διαδικασία σύνθεσης. Για παράδειγμα, τέτοια χαρακτηριστικά είναι οι προσωδιακές προδιαγραφές, οι προσωδιακές ασυνέχειες, οι φασματικές ασυνέχειες κτλ.. Επιπλέον, παρουσιάζεται η μεθοδολογία καθορισμού των παραγόντων στάθμισης που λειτουργούν ρυθμιστικά στις συναρτήσεις και προσδιορίζουν την σημαντικότητα των κριτηρίων ανά περίπτωση σύνθεσης.

- **Σύνθεση Φωνής σε περιβάλλον Ενσωματωμένων Συστημάτων**

  Σε επίπεδο συσκευής, το περιβάλλον ενός κινητού τηλεφώνου διαθέτει ιδιαίτερα περιορισμένους υπολογιστικούς πόρους προβάλλοντας σημαντικούς περιορισμούς τόσο σε μνήμη όσο και σε υπολογιστική ισχύ. Για την αποδοτική μεταφορά της τεχνολογίας σύνθεσης φωνής σε ένα τέτοιο περιβάλλον, είναι απαραίτητη η σημαντική μείωση των απαιτήσεων σε αποθηκευτική μνήμη και υπολογιστικών πόρων που προβάλουν οι σύγχρονες μέθοδοι σύνθεσης φωνής, χωρίς όμως σημαντική βλάβη της ποιότητας [Schnell, 2002; Nukaga, 2006]. Αναφορικά με το υποσύστημα της επεξεργασίας σήματος η βασική πρόκληση είναι η προσαρμογή της τεχνολογίας σύνθεσης φωνής στις υπολογιστικές ικανότητες του κινητού τηλεφώνου. Στα πλαίσια της διατριβής προτείνονται τεχνικές που αφορούν την σχεδίαση, την αποδοτική αποκλιμάκωση και προσαρμογή συστήματος σύνθεσης φωνής από κείμενο με επιλογή και ένωση ακουστικών μονάδων, σε περιβάλλοντα περιορισμένων υπολογιστικών πόρων όπως είναι τα ενσωματωμένα συστήματα και ιδιαίτερα το περιβάλλον των κινητών τηλεφώνων. Κύρια επιδίωξη είναι η βελτίωση του λόγου της ποιότητας προς τον απαιτούμενο όγκο δεδομένων, με σκοπό την ομαλή αποκλιμάκωση ενός υπάρχοντος συστήματος σύνθεσης φωνής στο περιβάλλον ενός κινητού τηλεφώνου χωρίς ανάλογη απώλεια σε ποιότητα ομιλίας. Οι ερευνητικές προσπάθειες κινούνται στις εξής βασικές κατευθύνσεις:

- Μείωση των υπολογιστικών απαιτήσεων που προκύπτουν από την βαθμίδα επιλογής ακουστικών μονάδων και αφορούν το κόστος φασματικής συνέχειας [Coorman, 2000; Beutnagel, 1999; Black, 1997]. Η διαδικασία επιλογής ακουστικών μονάδων είναι μια υπολογιστικά ιδιαίτερα απαιτητική και χρονοβόρα διαδικασία με αποτέλεσμα να απαιτούνται αφενός ανευρετικές τεχνικές στην διαδικασία αναζήτησης και αφετέρου προσαρμοσμένοι αλγόριθμοι στον υπολογισμό και την σύγκριση παραμέτρων κατά την εκτέλεση της. Για την ελαχιστοποίηση του υπολογιστικού φορτίου κατά την σύνθεση και για την περίπτωση του υπολογισμού του φασματικού κόστους (κόστος φασματικής συνέχειας) που απαιτεί τους περισσότερους υπολογιστικούς πόρους, υιοθετήθηκε η εφαρμογή συσταδοποίησης (clustering) μέσω διανυσματικής κβάντισης (vector quantization) για το διάνυσμα της φασματικής αναπαράστασης των ακουστικών μονάδων και τον σε μη πραγματικό χρόνο (offline) υπολογισμό των αποστάσεων ανάμεσα στα κέντρα των συστάδων (clusters). Με την τεχνική αυτή διατηρείται ο πλήρης χώρος αναζήτησης των ακουστικών μονάδων (διότι δεν πραγματοποιείται συσταδοποίηση στις ακουστικές μονάδες), γεγονός που είναι σημαντικό καθώς η βάση δεδομένων είναι ήδη μειωμένη στην περίπτωση των ενσωματωμένων συστημάτων (βλ. υποσημείωση 3). Η εφαρμογή της τεχνικής αυτής μείωσε σημαντικά τον χρόνο εκτέλεσης της μηχανής επιλογής χωρίς να επιφέρει ουσιαστική επίδραση στην τελική ποιότητα.

- Μείωση του μεγέθους της βάσης δεδομένων προηχογραφημένης φυσικής ομιλίας μέσω τεχνικών συμπίεσης και κωδικοποίησης της βάσης. Μελετήθηκαν οι επιπτώσεις της χρήσης απωλεστικών αλγορίθμων κωδικοποίησης φωνής για την συμπίεση των ακουστικών μονάδων που αποθηκεύονται στην βάση αναφορικά με την επίδραση που έχουν τόσο στην ποιότητα της παραγόμενης φωνής, όσο και κατά το στάδιο αποκωδικοποίησης και συρραφής τους (ανάγκη για τυχαία πρόσβαση στο ηχογραφημένο σήμα φωνής) για την παραγωγή συνθετικής ομιλίας. Ειδικότερα, υιοθετήθηκε και προσαρμόστηκε κατάλληλα η μέθοδος CELP (Code Excited Linear Prediction [Schroeder, 1985; Chu, 2003]). Το πλεονέκτημα της προτεινόμενης τεχνικής είναι ο ικανοποιητικός βαθμός συμπίεσης που επιτυγχάνει αναφορικά με το πρόβλημα και η εξασφάλιση της δυνατότητας τυχαίας πρόσβασης στις ακουστικές μονάδες μέσω ξεχωριστής κωδικοποίησης της καθεμίας από αυτές χωρίς συνέπειες στην γενικότερη διαδικασία της σύνθεσης.

- Μείωση μιας δεδομένης (υπάρχουσας) μεγάλης βάσης δεδομένων προηχογραφημένης φυσικής ομιλίας μέσω αποδοτικής περικοπής ακουστικών μονάδων οι οποίες αναμένεται να εμφανίζουν μικρή πιθανότητα χρησιμοποίησης και ταυτόχρονα διατήρηση μονάδων που είναι πιο «εύρωστες», με την έννοια του πιο πρόσφορες για μεταβολή και συρραφή με άλλες μονάδες του κλειστού συνόλου της βάσης με την μέγιστη δυνατή ποιότητα [Kumar, 2004; Krul, 2007]. Προτείνεται μια τεχνική που στηρίζεται σε στατιστικά αποτελέσματα που προκύπτουν από την σύνθεση, με χρήση της βαθμίδας επιλογής ακουστικών μονάδων (στην περίπτωσή μας δίφωνα), μεγάλου σώματος κειμένου. Η μέθοδος αξιοποιεί όχι μόνο την συχνότητα εμφάνισης των ακουστικών μονάδων αλλά και την βαθμολογία που αυτά λαμβάνουν από τις συναρτήσεις κόστους, έτσι ώστε

αφενός να επιλέγει τις πιο συχνά χρησιμοποιούμενες ακουστικές μονάδες και αφετέρου να περιορίζει την επιλογή πλεονάζουσων ακουστικών μονάδων. Επομένως, ο χαρακτηρισμός ως πλεονάζουσες μονάδες προκύπτει με κριτήριο το πως αντιμετωπίζονται αυτές οι μονάδες από τον ίδιο τον αλγόριθμο επιλογής.

Οι παραπάνω τεχνικές αξιολογήθηκαν από μια σειρά πειραμάτων με υποκειμενικά (ακουστικά πειράματα) και αντικειμενικά (τεχνικές επιδόσεις) κριτήρια τα αποτελέσματα των οποίων επικύρωσαν την λειτουργικότητα και την αποδοτικότητά τους. Η ολοκλήρωση και εφαρμογή των παραπάνω τεχνικών οδήγησε στην υλοποίηση ενός συστήματος σύνθεσης φωνής από κείμενο για την συσκευή του κινητού τηλεφώνου σημαντικά υψηλής ποιότητας [Karabetsos, 2009].

- **Προσαρμογή Σύνθεσης Φωνής με Κρυφά Μαρκοβιανά Μοντέλα**

Στο πλαίσιο της παρούσας εργασίας, διερευνήθηκαν και παραμετρικές μέθοδοι σύνθεσης φωνής που από την φύση τους απαιτούν σημαντικά χαμηλότερους υπολογιστικούς πόρους ενώ ταυτόχρονα παρέχουν μεγαλύτερη ευελιξία τόσο στη διαχείριση όσο και στην κατασκευή νέων συνθετικών φωνών. Όπως αναφέρθηκε, το πιο επιτυχημένο παράδειγμα τέτοιας μεθοδολογίας είναι η σύνθεση φωνής με βάση τα Κρυφά Μαρκοβιανά Μοντέλα. Η υιοθέτηση των Κρυφών Μαρκοβιανών Μοντέλων παρέχει ένα γενικευμένο στατιστικό μεθοδολογικό πλαίσιο (*HMM-based speech synthesis framework*) για την αποδοτική παραμετρική μοντελοποίηση και παραγωγή φωνής. Η μέθοδος αυτή θεωρείται πλέον ικανή να παράγει συνθετική φωνή υψηλής ποιότητας και αποτελεί τεχνολογία αιχμής [Zen, 2009; Black, 2007]. Στην παρούσα διατριβή, μελετήθηκε, προσαρμόστηκε, υλοποιήθηκε και αξιολογήθηκε το μεθοδολογικό πλαίσιο της παραμετρικής τεχνολογίας σύνθεσης φωνής από κείμενο με χρήση κρυφών Μαρκοβιανών μοντέλων για την περίπτωση της Ελληνικής γλώσσας. Η αποτίμηση του συστήματος πραγματοποιήθηκε με συγκριτική αξιολόγηση τόσο με ένα σύστημα σύνθεσης με διφωνήματα (diphone synthesis) με όσο και με το σύστημα επιλογής και συρραφής ακουστικών μονάδων (unit selection synthesis). Η ακουστική αξιολόγηση ανέδειξε ότι η εν λόγω τεχνολογία παράγει συνθετική φωνή ικανοποιητικής ποιότητας, η οποία αν και υπολείπεται ακόμα αυτής που επιτυγχάνεται με επιλογή και συρραφή ακουστικών μονάδων, δέχεται περαιτέρω τροποποιήσεις που μπορούν να επιφέρουν σημαντικές βελτιώσεις στο τελικό αποτέλεσμα. Τέτοιες τροποποιήσεις αφορούν αφενός το κομμάτι της προσωδίας και αφετέρου την άρση των περιορισμών που επιφέρουν οι κλασσικές τεχνικές αναπαράστασης, μοντελοποίησης και παραγωγής του συνθετικού σήματος φωνής (π.χ., LPC vocoders) [Karabetsos, 2008].

- **Κόστος Ένωσης με Ταξινομητές μιας Τάξης**

Στην τεχνολογία σύνθεσης φωνής από κείμενο που στηρίζεται σε επιλογή και συρραφή ακουστικών μονάδων από προηχογραφημένη βάση δεδομένων φυσικής ομιλίας, η φυσικότητα της παραγόμενης συνθετικής ομιλίας κρίνεται αρκετά υψηλή στις περισσότερες των περιπτώσεων. Συχνά όμως, παρουσιάζοντας ακουστικές ασυνέχειες λόγω των συρραφών που προκύπτουν από τις διάφορες πραγματώσεις των ακουστικών μονάδων. Αυτό έχει ως αποτέλεσμα την άμεση επίπτωση στην φυσικότητα του συνθετικού λόγου. Βασικό παράγοντα για αυτό, αποτελεί η

έλλειψη ενός αντικειμενικού κριτηρίου αξιολόγησης και επιλογής ακουστικών μονάδων που να αντικατοπτρίζει την ανθρώπινη αντίληψη για τις ακουστικές ασυνέχειες και κατ' επέκταση την φυσικότητα της ομιλίας [Pantazis, 2005; Vepa, 2004; Vepa, 2006; Stylianou, 2001; Klabbers, 2001; Founda, 2001].

Για το σκοπό αυτό, στα πλαίσια της διατριβής διερευνήθηκε η ανάπτυξη και αξιολόγηση ενός νέου κριτηρίου μέσω μίας νέας τεχνικής με στόχο την βελτίωση της φυσικότητας της συνθετικής ομιλίας. Συγκεκριμένα, η εκτίμηση και αποτίμηση των φασματικών ασυνεχειών που προκύπτουν στην ένωση των ακουστικών μονάδων βασίζεται στην υιοθέτηση και εφαρμογή ενός νέου μεθοδολογικού πλαισίου που βασίζεται στα **διαθέσιμα δεδομένα (data driven)**, και το οποίο μπορεί να επεκταθεί και στην σχεδίαση της συνολικής συνάρτηση κόστους ένωσης ακουστικών μονάδων, προσφέροντας σημαντικά πλεονεκτήματα όπως είναι για παράδειγμα η αποφυγή προσδιορισμού των συντελεστών στάθμισης (βάρη) στα επιμέρους κριτήρια της συνάρτησης κόστους ένωσης.

Ειδικότερα, για την εκτίμηση των φασματικών ασυνεχειών που προκύπτουν στην ένωση των ακουστικών μονάδων υιοθετήθηκαν μεθοδολογίες από τον χώρο της μηχανικής μάθησης και συγκεκριμένα η **χρήση ταξινομητών μιας τάξης (One-Class Classifiers)** [Tax, 2001; Hodge, 2004; Markou, 2003ab, Jain, 2000; Kennedy, 2009; Patcha, 2007]. Οι ταξινομητές μιας τάξης εφαρμόζονται κυρίως σε προβλήματα αναγνώρισης προτύπων δύο τάξεων για τα οποία διαθέσιμα δεδομένα εκπαίδευσης υπάρχουν μόνο για την μια τάξη ενώ για την άλλη τάξη είτε δεν υπάρχουν καθόλου δεδομένα είτε υπάρχουν λιγοστά. Ουσιαστικά εφαρμόζεται και σε προβλήματα πολλαπλών τάξεων εκ των οποίων μόνο μια θεωρείται η ζητούμενη τάξη ή τάξη στόχος (target class) ενώ οι υπόλοιπες θεωρούνται ως παραπειστικές ή εκτός τάξης (outlier class). Για την υιοθέτηση της εν λόγω μεθοδολογίας εκμεταλλευόμαστε το γεγονός ότι η διαθέσιμη βάση δεδομένων προηχογραφημένης φυσικής ομιλίας, που χρησιμοποιείται και κατά την σύνθεση, αποτελείται από πλήθος φυσικών "ενώσεων" όπως αυτές ορίζονται από τα ζεύγη συνεχόμενων πλαισίων φωνής. Η πληθώρα από φυσικές "ενώσεις" μπορούν να χρησιμοποιηθούν ως πρωτογενή δεδομένα περιγραφής του χώρου των "καλών" ενώσεων. Από την άλλη πλευρά, η συστηματική συλλογή από δεδομένα με φασματικές ασυνέχειες, ικανά να περιγράψουν τον χώρο των ακουστικών ασυνεχειών, αποτελεί μια δύσκολη, ακριβή και χρονοβόρα διαδικασία. Με αυτό τον τρόπο, το πρόβλημα της εκτίμησης των φασματικών ασυνεχειών μπορεί να ειδωθεί και να αντιμετωπιστεί ως **πρόβλημα μιας-τάξης** (one-class problem). Στα πλαίσια της διατριβής εξετάζεται η ανάλυση ανά φώνημα και η αναπαράσταση κάθε ζεύγους από συνεχόμενα πλαίσια φωνής με ένα διάνυσμα φασματικών αποστάσεων. Τα διανύσματα αυτά χρησιμοποιούνται για την εκπαίδευση του ταξινομητή μιας τάξης. Κατά την σύνθεση, δύο πλαίσια φωνής από διαφορετικές πραγματώσεις του ίδιου φωνήματος αναπαρίστανται με τον ίδιο τρόπο και το κόστος φασματικής ασυνέχειας προκύπτει από την βαθμολογία που δίνει ο ταξινομητής. Αναμένεται η πειραματική αξιολόγηση να αναδείξει ότι η προτεινόμενη μεθοδολογία επιτυγχάνει σημαντικά καλύτερη απόδοση στην εκτίμηση φασματικών ασυνεχειών από τις συνήθεις προσεγγίσεις. Σημαντικό πλεονέκτημα της προσέγγισης αυτής είναι αφενός ότι στηρίζεται στα διαθέσιμα δεδομένα και αφετέρου ότι μπορεί να υιοθετηθεί ως ευρύτερο μεθοδολογικό

πλαίσιο και να εφαρμοστεί με κάθε πιθανή αναπαράσταση ικανή να διακρίνει τις δύο τάξεις.

Τέλος, η προηγούμενη μεθοδολογία μπορεί να γενικευτεί και για την περίπτωση της συνολικής σχεδίασης της συνάρτησης κόστους ένωσης. Με την επαύξηση του διανύσματος χαρακτηριστικών με πληροφορία θεμελιώδους συχνότητας (pitch) και έντασης (intensity), η συνάρτηση κόστους ένωσης μετατρέπεται σε μια διαδικασία ταξινόμησης, μέσω του ταξινομητή μιας τάξης, και σύμφωνα με τα δεδομένα που συναντώνται στη διαθέσιμη βάση δεδομένων. Το σημαντικότερο πλεονέκτημα που προκύπτει από αυτή την προσέγγιση είναι η πλήρης αποφυγή προσδιορισμού των συντελεστών στάθμισης της συνάρτησης κόστους, μια διαδικασία που συνήθως γίνεται διαισθητικά και με χειρωνακτικό τρόπο.

- **Τηλεπικοινωνιακές και Διαδικτυακές εφαρμογές**

Όπως αναφέρθηκε, η επίτευξη υψηλής ποιότητας συνθετικής ομιλίας ανοίγει το δρόμο για ένα εύρος εφαρμογών στον ευρύτερο χώρο της επικοινωνίας ανθρώπου-μηχανής, που στην αντίθετη περίπτωση δεν θα ήταν εφικτές [Deng, 2005; Gilbert, 2008; Duggan, 2003]. Στα πλαίσια της διατριβής, θα περιγραφούν κάποιες καινοτόμες εφαρμογές, στις οποίες η σύνθεση φωνής έχει ουσιαστικό ρόλο και που αφορούν την ολοκλήρωση της τεχνολογίας είτε σε επίπεδο συσκευής είτε σε επίπεδο δικτύου.

1.5 ΟΡΓΑΝΩΣΗ ΤΗΣ ΔΙΑΤΡΙΒΗΣ

Η παρούσα έκθεση διαρθρώνεται σε πέντε κεφάλαια. Στο **κεφάλαιο 1**, πραγματοποιήθηκε μια εισαγωγική επισκόπηση των τεχνολογιών σύνθεσης φωνής με έμφαση σε αυτές που σήμερα θεωρούνται οι επικρατέστερες. Στην συνέχεια παρουσιάστηκαν μερικοί τομείς στους οποίους βρίσκει εφαρμογή η σύνθεση φωνής και κατόπιν συνοψίστηκαν μερικά από τα ζητήματα που συγκεντρώνουν το διεθνές ερευνητικό ενδιαφέρον. Τέλος, πραγματοποιήθηκε μια συνοπτική παρουσίαση των θεμάτων και της ερευνητικής συνεισφοράς της διατριβής.

Στο **κεφάλαιο 2**, δίνεται το θεωρητικό υπόβαθρο και αναλύεται το μεθοδολογικό πλαίσιο του αλγόριθμου επιλογής ακουστικών μονάδων. Επίσης πραγματοποιείται λεπτομερής βιβλιογραφική επισκόπηση για την συνάρτηση κόστους ένωσης και ειδικότερα για τις προσεγγίσεις που έχουν προταθεί για την εκτίμηση του κόστους φασματικής ασυνέχειας. Επιπλέον, παρουσιάζονται διάφορες προσεγγίσεις που έχουν προταθεί για την προσαρμογή της τεχνολογίας σε περιβάλλοντα μειωμένων υπολογιστικών πόρων. Τέλος, συνοψίζονται οι κυριότερες τεχνικές που αφορούν την παραγωγή του σήματος συνθετικής φωνής.

Στο **κεφάλαιο 3**, περιγράφονται οι τεχνικές που προτάθηκαν στα πλαίσια της διδακτορικής διατριβής για την προσαρμογή της τεχνολογίας σύνθεσης φωνής από κείμενο με επιλογή και συρραφή ακουστικών μονάδων στις υπολογιστικές ικανότητες του κινητού τηλεφώνου. Επιπλέον παρουσιάζεται η πειραματική αποτίμηση και τα αποτελέσματά της.

Στο **κεφάλαιο 4**, περιγράφεται το μεθοδολογικό πλαίσιο της παραμετρικής τεχνολογίας σύνθεσης φωνής από κείμενο με χρήση κρυφών Μαρκοβιανών

μοντέλων παρουσιάζεται η προσαρμογή και η συγκριτική αξιολόγηση του για την περίπτωση της Ελληνικής γλώσσας.

Στο **κεφάλαιο 5**, περιγράφεται το μεθοδολογικό πλαίσιο που προτείνεται στα πλαίσια της διατριβής για την σχεδίαση της συνάρτηση κόστους ένωσης και το οποίο βασίζεται στην χρήση ταξινομητών μιας τάξης (one-class classification). Παρουσιάζεται η μεθοδολογία υιοθέτησης του καθώς και η πειραματική του αξιολόγηση για την περίπτωση εκτίμησης των φασματικών ασυνεχειών. Συζητείται επίσης η περίπτωση της σχεδίασης της συνάρτησης κόστους ένωσης με χρήση ταξινομητών μιας τάξης συνολικά.

Στο **κεφάλαιο 6**, περιγράφονται διάφορες εφαρμογές που χρησιμοποιούν σύνθεση φωνής από κείμενο υψηλής ποιότητας. Οι εφαρμογές αναφέρονται τόσο σε επίπεδο συσκευών όσο και σε επίπεδο δικτύου και τηλεπικοινωνιακών υπηρεσιών.

Τέλος, στο **κεφάλαιο 7**, συνοψίζονται η συνεισφορά και τα συμπεράσματα που προέκυψαν στο πλαίσιο της διατριβής και στη συνέχεια αναγνωρίζονται μελλοντικές ερευνητικές κατευθύνσεις.

ΚΕΦΑΛΑΙΟ
-2-
Ο ΑΛΓΟΡΙΘΜΟΣ ΕΠΙΛΟΓΗΣ ΑΚΟΥΣΤΙΚΩΝ ΜΟΝΑΔΩΝ

ΚΕΦΑΛΑΙΟ 2 – Ο ΑΛΓΟΡΙΘΜΟΣ ΕΠΙΛΟΓΗΣ ΑΚΟΥΣΤΙΚΩΝ ΜΟΝΑΔΩΝ

Στο κεφάλαιο αυτό εξετάζεται το μεθοδολογικό πλαίσιο του αλγόριθμου επιλογής ακουστικών μονάδων (**unit selection algorithm**) στην σύνθεση φωνής από κείμενο με επιλογή και συρραφή ακουστικών μονάδων από βάση δεδομένων προηχογραφημένης φυσικής ομιλίας. Στόχος του κεφαλαίου είναι να θέσει τόσο το απαραίτητο θεωρητικό υπόβαθρο όσο και να αναδείξει τις προσεγγίσεις και τα ερευνητικά ζητήματα που προκύπτουν. Επιπρόσθετα, περιγράφονται και οι προσεγγίσεις που υιοθετεί το σύστημα σύνθεσης φωνής από κείμενο που χρησιμοποιήθηκε στα πλαίσια της διατριβής. Αρχικά, πραγματοποιείται μια σύντομη αναφορά στις επικρατέστερες προσεγγίσεις που αφορούν την σχεδίαση των συναρτήσεων κόστους που αποτελούν τον αλγόριθμο, με έμφαση στην συνάρτηση κόστους ένωσης. Ιδιαίτερη αναφορά γίνεται για τις προσεγγίσεις που αφορούν το κόστος φασματικής ασυνέχειας, πρόβλημα το οποίο εξετάζεται και σε επόμενα κεφάλαια της διατριβής. Στην συνέχεια πραγματοποιείται μια επισκόπηση των τεχνικών και των προσεγγίσεων που αφορούν την αποδοτική μεταφορά της τεχνολογίας σύνθεσης φωνής σε υπολογιστικά περιβάλλοντα μειωμένων υπολογιστικών πόρων όπως για παράδειγμα τα ενσωματωμένα συστήματα. Η ενότητα κλείνει με την συνοπτική παρουσίαση των κυριότερων τεχνικών παραγωγής και τροποποίησης φωνής που χρησιμοποιούνται στην σύνθεση.

2.1 ΕΙΣΑΓΩΓΗ – ΓΕΝΙΚΗ ΘΕΩΡΗΣΗ

Στα συστήματα σύνθεσης φωνής από κείμενο με επιλογή και συρραφή ακουστικών μονάδων από βάση δεδομένων προηχογραφημένης φυσικής ομιλίας, ο αλγόριθμος επιλογής ακουστικών μονάδων μαζί με τη μονάδα (module) παραγωγής και τροποποίησης σήματος φωνής, αποτελούν το υποσύστημα της Ψηφιακής Επεξεργασίας Σήματος. Η βασική ιδέα πίσω από τα συστήματα αυτά, έγκειται στο γεγονός ότι η διαθέσιμη βάση δεδομένων προηχογραφημένης φυσικής ομιλίας εμπεριέχει (ή από άποψη σχεδίασης θα πρέπει να εμπεριέχει) πολλαπλά παραδείγματα (πραγματώσεις) από κάθε ακουστική μονάδα (π.χ. δίφωνο) σε διαφορετικά προσωδιακά και φωνητικά περιβάλλοντα. Με αυτό τον τρόπο επιτυγχάνεται σημαντική ποικιλία και ευελιξία, αφού πλέον κατά τη σύνθεση προσφέρεται πλήθος από εναλλακτικές επιλογές για κάθε ακουστική μονάδα, ενώ τελικά επιλέγονται εκείνες που είναι βέλτιστες υπό κάποια κριτήρια. Συγκριτικά με τα προγενέστερα συστήματα σύνθεσης με διφωνήματα, τα συστήματα με επιλογή ακουστικών μονάδων διαφοροποιούν το πρόβλημα της σύνθεσης φωνής και ξεπερνούν εγγενείς υποθέσεις που συναντώνται σε αυτά. Πράγματι, η σύνθεση με επιλογή και συρραφή ακουστικών μονάδων καταργεί την (κατά τα άλλα ασθενή) υπόθεση ότι όλη η μεταβλητότητα (variability) που μπορεί να συναντηθεί σε μια ακουστική μονάδα είναι διαχειρίσιμη και επιτεύξιμη μόνο μέσω προσωδιακών τροποποιήσεων (κυρίως προσωδία και διάρκεια). Για παράδειγμα, το φωνητικό περιβάλλον (phonetic & prosodic context) είναι ιδιαίτερα σημαντικό χαρακτηριστικό και συντελεί καταλυτικά στην μεταβλητότητα, και κατ' επέκταση στα ιδιαίτερα χαρακτηριστικά, μιας ακουστικής μονάδας. Παράλληλα, ελαχιστοποιεί τις ανάγκες

επεξεργασία σήματος που προκύπτουν λόγω αυτών των προσωδιακών τροποποιήσεων, και οι οποίες σε πολλές περιπτώσεις οδηγούν σε σημαντική υποβάθμιση της τελικής ποιότητας. Με την δυνατότητα επιλογής ακουστικών μονάδων, το πρόβλημα της σύνθεσης καταλήγει σε ένα πρόβλημα ανάλυσης, διαχείρισης και αναζήτησης με κατάλληλα κριτήρια, του πλήθους των πραγματώσεων των ακουστικών μονάδων μέσω των χαρακτηριστικών τους.

Ο μηχανισμός της αναζήτησης και επιλογής των κατάλληλων ακουστικών μονάδων από τη βάση δεδομένων που πρόκειται να αποτελέσουν το τελικό συνθετικό σήμα φωνής παρέχεται από τον αλγόριθμο επιλογής ακουστικών μονάδων (**unit selection**). Διάφορες προσεγγίσεις έχουν προταθεί για το σκοπό αυτό όπως, η επιλογή με κανόνες, η επιλογή με ελαχιστοποίηση φασματικών κριτηρίων, η επιλογή με *Conditional Random Fields*, η επιλογή με στατιστικό μοντέλο κ.α. [Sagisaka, 1988; Sagisaka, 1992; Iwahashi; 1993; Weiss, 2006; Sakai, 2005]. Ωστόσο, η πιο γενικευμένη και ταυτόχρονα τυποποιημένη προσέγγιση, η οποία αποτελεί πλέον και το ευρύτερο μεθοδολογικό πλαίσιο για την σχεδίαση και υλοποίηση του αλγόριθμου επιλογής ακουστικών μονάδων, προτάθηκε το 1996 από τους Hunt και Black [Hunt, 1996]. Η προσέγγιση αυτή εξετάζεται στην ενότητα που ακολουθεί και υιοθετείται σε όλο το υπόλοιπο της διατριβής.

2.2 ΣΥΝΑΡΤΗΣΕΙΣ ΚΟΣΤΟΥΣ ΚΑΙ ΕΠΙΛΟΓΗ ΑΚΟΥΣΤΙΚΩΝ ΜΟΝΑΔΩΝ

Ο αλγόριθμος που προτάθηκε από τους Hunt και Black αντιμετωπίζει την επιλογή ακουστικών μονάδων ως ένα πρόβλημα **δυναμικής αναζήτησης** με την βοήθεια συναρτήσεων κόστους [Hunt, 1996]. Ειδικότερα, ορίζεται μια ολική συνάρτηση κόστους που συμπεριλαμβάνει κριτήρια που αντικατοπτρίζουν τις επιθυμητές ιδιότητες που θα πρέπει να καλύπτει η εκάστοτε σύνθεση τόσο σε επίπεδο προδιαγραφών όσο και σε επίπεδο αλληλουχίας των ακουστικών μονάδων που θα επιλεγούν για συρραφή. Τα κριτήρια μπορούν να αποτελούνται από διάφορα (μετρήσιμα) χαρακτηριστικά τόσο σε ακουστικό όσο και σε γλωσσολογικό επίπεδο. Το κριτήριο με το οποίο επιλέγεται η ακολουθία των ακουστικών μονάδων είναι η ελαχιστοποίηση αυτής της ολικής συνάρτησης κόστους.

Συγκεκριμένα, ορίζονται και χρησιμοποιούνται δύο επιμέρους συναρτήσεις που εξετάζουν διαφορετικές ιδιότητες και κριτήρια και οι οποίες αφορούν:

- Το **κόστος «στόχος»** (**target cost**), δηλαδή το κόστος που αφορά την επιθυμητή ακολουθία ακουστικών μονάδων. Το κόστος αυτό, που μπορεί να αποτελείται από επιμέρους κριτήρια (επιμέρους κόστη), εξετάζει την καταλληλότητα μιας ακουστικής μονάδας για την θέση που θα τοποθετηθεί στην συνθετική πρόταση. Η καταλληλότητα ορίζεται από τις προδιαγραφές που θέτει η προς σύνθεση πρόταση και οι οποίες προκύπτουν από κάποια εκτίμηση (π.χ., ανάλυση κειμένου). Εξετάζει δηλαδή την συμβατότητα μιας ακουστικής μονάδας με τα επιθυμητά φωνητικά και προσωδιακά χαρακτηριστικά για την θέση αυτή. Αποκλίσεις της ακουστικής μονάδας από τις επιθυμητές ιδιότητες και προδιαγραφές, σχηματίζουν το κόστος «στόχος».

- Το **κόστος «ένωσης»** (**join** or **concatenation cost**) που αφορά τη σύνδεση και τη συνέχεια των ακουστικών μονάδων. Το κόστος αυτό, που μπορεί επίσης να

αποτελείται από επιμέρους κριτήρια (επιμέρους κόστη) εξετάζει την συμβατότητα μιας ακουστικής μονάδας σε σχέση με τις γειτονικές της, σχετικά με τις ακουστικές ασυνέχειες που ενδεχομένως να προκύψουν από την ένωση τους. Οι ακουστικές ασυνέχειες σχετίζονται κυρίως με φασματικά και προσωδιακά χαρακτηριστικά. Μη επιτρεπτές αποκλίσεις των ακουστικών μονάδων σε αυτά τα μεγέθη σχηματίζουν το κόστος «ένωσης».

Και στις δύο περιπτώσεις το πλήθος και το είδος των κριτηρίων είναι ζήτημα σχεδίασης. Επιθυμητό είναι τόσο τα χαρακτηριστικά και το είδος των κριτηρίων όσο συνολικά η συνάρτηση κόστους αλλά και οι επιμέρους συναρτήσεις να συσχετίζονται με την ανθρώπινη αντίληψη (perceptually correlated) [Coorman, 2000; Bulyko, 2001; Diaz, 2003; Diaza, 2006; Lee, 2003; Ding, 1998; Plumpe, 1998; Rouibia, 2005; Clark, 2007; Tihelka, 2007; Toda, 2002; Toda, 2003; Toda, 2006; Syrdal, 2004; Syrdal, 2005]. Το σχήμα 2.1 δίνει την γενική σχηματική περιγραφή του αλγόριθμου επιλογής ακουστικών μονάδων ενώ στη συνέχεια ακολουθεί η μαθηματική του διατύπωση.

Υποθέτοντας για την υπόλοιπη ανάλυση ότι ο τύπος της ακουστικής μονάδας αναφέρεται σε διφώνημα, η παραπάνω διαδικασία περιέχει τον ορισμό δύο χρήσιμων ακολουθιών:
Έστω $\mathbf{t_1^n}$ η επιθυμητή ακολουθία σύνθεσης (target sequence), η οποία συμβολίζεται ως,

$$\mathbf{t_1^n = \{t_1, t_2, \ldots, t_n\}}$$

όπου ***n*** το πλήθος των διφώνων από τα οποία αποτελείται. Έστω επίσης $\mathbf{u_1^n}$ η ακολουθία των υποψήφιων διφώνων η οποία συμβολίζεται ως,

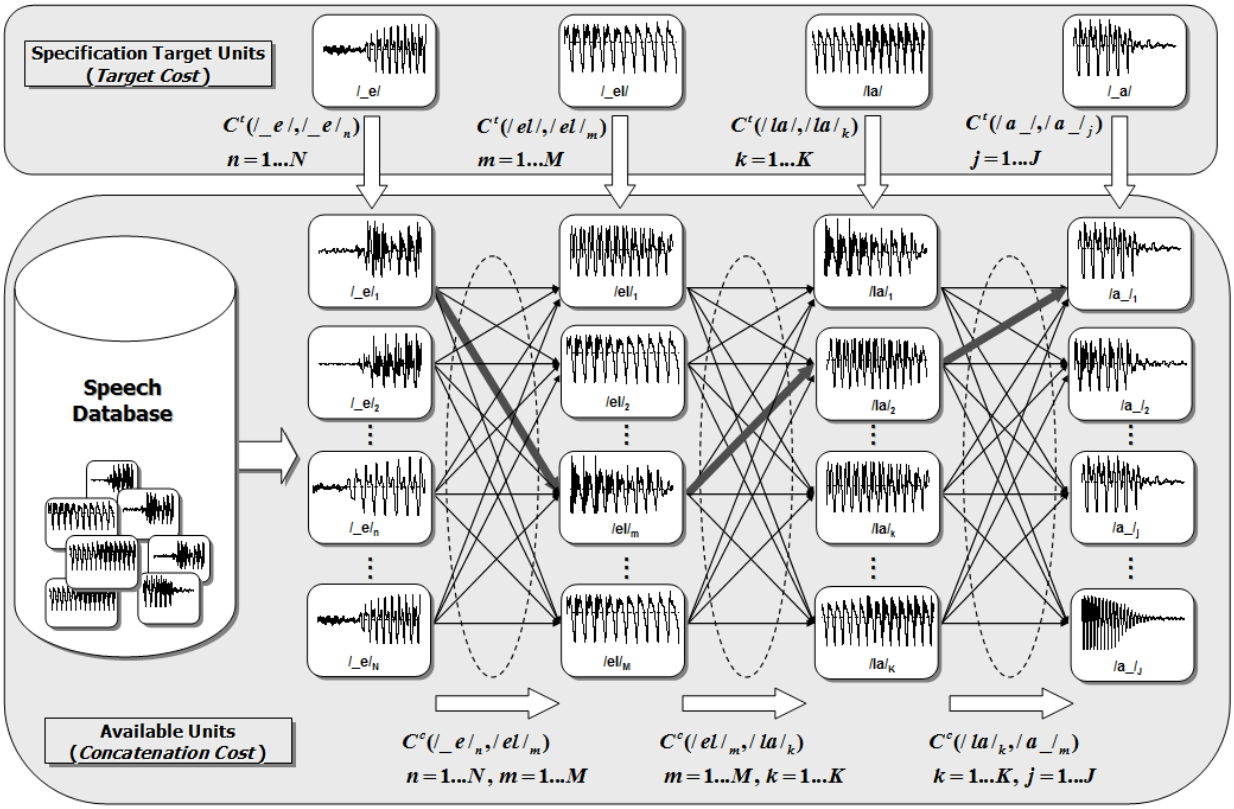
$$\mathbf{u_1^n = \{u_1, u_2, \ldots, u_n\}}$$

Οι δύο επιμέρους συναρτήσεις κόστους μπορούν να οριστούν βάση αυτών των ακολουθιών ώς εξής:
Το **κόστος «στόχος» (target cost)** ορίζεται από τη σχέση,

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t \cdot C_j^t(t_i, u_i) \qquad (2.1)$$

όπου,

p	:	Η διάσταση του διανύσματος χαρακτηριστικών
$C^t(t_i, u_i)$	:	Το υπό εξέταση κόστος «στόχος» όταν το $\mathbf{u_i}$ επιλεχθεί σαν όμοιο του $\mathbf{t_i}$.
$C_j^t(t_i, u_i)$	:	Επιμέρους συνάρτηση κόστους ή μερικό κόστος που αφορά συγκεκριμένο χαρακτηριστικό (έστω το j) του διανύσματος χαρακτηριστικών.
w_j^t	:	Παράγοντας στάθμισης (βάρος) που αντιστοιχεί στο συγκεκριμένο χαρακτηριστικό (έστω το j) του διανύσματος χαρακτηριστικών.



ΣΧΗΜΑ 2.1: Η διαδικασία και ο αλγόριθμος επιλογής ακουστικών μονάδων (unit selection process): Το κόστος «ένωσης» $\mathbf{C^c}$ εξετάζει και βαθμολογεί την ακουστική συμβατότητα των υποψήφιων ακουστικών μονάδων. Το κόστος «στόχος» $\mathbf{C^t}$ εξετάζει την συμβατότητα κάθε υποψήφιας ακουστικής μονάδας με την επιθυμητή ακουστική μονάδα όπως προκύπτει από τις (εκτιμώμενες) προδιαγραφές. Το παράδειγμα του σχήματος χρησιμοποιεί δίφωνα και αποτυπώνει την σύνθεση της λέξης "έλα" η οποία σε ακολουθία διφώνων αναλύεται ως **/_e/** - **/el/** - **/la/** - **/a_/** με N, M, K, J πραγματώσεις για κάθε δίφωνο αντίστοιχα. Η καλύτερη ακολουθία προκύπτει από το ελάχιστο σταθμισμένο συσσωρευτικό άθροισμα των δύο επιμέρους κοστών. Παράδειγμα επιλογής της καλύτερης ακολουθίας φαίνεται με το έντονα σκιασμένο μονοπάτι.

Αντίστοιχα, το **κόστος «ένωσης» (join** or **concatenation cost)** ορίζεται από τη σχέση,

$$C^c(u_{i-1},u_i)=\sum_{j=1}^{q} w_j^c \cdot C_j^c(u_{i-1},u_i) \qquad (2.2)$$

όπου,

q	:	Η διάσταση του διανύσματος χαρακτηριστικών
$C^c(u_{i-1},u_i)$	:	Το υπό εξέταση κόστος «ένωσης» όταν το $\mathbf{u_i}$ ενωθεί με το $\mathbf{u_{i-1}}$.
$C_j^c(u_{i-1},u_i)$	:	Επιμέρους συνάρτηση κόστους ή μερικό κόστος που αφορά συγκεκριμένο χαρακτηριστικό (έστω το j) του διανύσματος χαρακτηριστικών.
w_j^c	:	Παράγοντας στάθμισης (βάρος) που αντιστοιχεί στο συγκεκριμένο χαρακτηριστικό (έστω το j) του διανύσματος χαρακτηριστικών.

Η **συνολική συνάρτηση κόστους** ή το **συνολικό κόστος** (**total cost**) προκύπτει ως το σταθμισμένο άθροισμα των επιμέρους συναρτήσεων κόστους σε όλη την ακολουθία και δίνεται από τη σχέση,

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} W^t \cdot C^t(t_i, u_i) + \sum_{i=2}^{n} W^c \cdot C^c(u_{i-1}, u_i) \quad (2.3)$$

όπου,

$C(t_1^n, u_1^n)$: Το συνολικό ή ολικό κόστος.

W^t : Παράγοντας στάθμισης για τη σημαντικότητα του κόστους «στόχος».

W^c : Παράγοντας στάθμισης για τη σημαντικότητα του κόστους «ένωσης».

Οι δύο νέοι παράγοντες στάθμισης W^t και W^c λειτουργούν ρυθμιστικά για τη βαρύτητα που αποδίδεται στις δύο επιμέρους συναρτήσεις κόστους κατά τη διαδικασία της σύνθεσης. Η τελευταία εξίσωση μπορεί περαιτέρω να αναλυθεί με αντικατάσταση των σχέσεων (2.1) και (2.2) και να γραφεί ως εξής,

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} W^t \cdot \sum_{j=1}^{p} w_j^t \cdot C_j^t(t_i, u_i) + \sum_{i=2}^{n} W^c \cdot \sum_{j=1}^{q} w_j^c \cdot C_j^c(u_{i-1}, u_i) \quad (2.4)$$

Ο αλγόριθμος επιλογής, επιδιώκει την αναζήτηση και επιλογή εκείνης της ακολουθίας $\hat{u}_1^n$ που ελαχιστοποιεί τη συνάρτηση του ολικού κόστους, δηλαδή **ελαχιστοποιεί** τη ποσότητα,

$$\hat{u}_1^n = \min_{u_1 \ldots u_n} C(t_1^n, u_1^n) \quad (2.5)$$

Ο αλγόριθμος χρησιμοποιεί δυναμικό προγραμματισμό με βάση την αναζήτηση τύπου Viterbi, όπου κατά τη διάσχιση ενός γράφου με τις πολλαπλές εκδοχές των ακουστικών μονάδων, επιλέγεται εκείνη η διαδρομή που δίνει το **ελάχιστο κόστος**. Στο σχήμα 2.1 φαίνεται ένας τέτοιος γράφος. Το σκιασμένο μονοπάτι, που περιλαμβάνει συγκεκριμένες πραγματώσεις διφωνημάτων, είναι αυτό που κατά τη σύνθεση θα δώσει τα καλύτερα ακουστικά αποτελέσματα σύμφωνα με τις δεδομένες συναρτήσεις κόστους και εφόσον αυτές αντικατοπτρίζουν την ανθρώπινη αντίληψη. Συνήθως, όλα τα επιμέρους κόστη που χρησιμοποιούνται είναι κανονικοποιημένα στο διάστημα τιμών [0, 1] ώστε σε συνδυασμό με τα βάρη να εκφράζεται άμεσα η επίδραση κάθε χαρακτηριστικού. Σε κάθε περίπτωση, η τιμή μηδέν αντιπροσωπεύει πλήρες ταίριασμα ενώ η τιμή ένα δείχνει αντιπροσωπεύει απόκλιση.

Η τυποποίηση (φορμαλισμός) του αλγόριθμου επιλογής με αυτό τον τρόπο, παρέχει ένα γενικευμένο πλαίσιο σχεδίασης και υλοποίησης του. Η ιδέα χρήσης κοστών και συναρτήσεων κόστους ή γενικότερα προσεγγίσεων που να

αποφαίνονται σχετικά με τα δύο ζητούμενα, δηλαδή την συμβατότητα με τις προδιαγραφές και την συμβατότητα μεταξύ των ακουστικών μονάδων, παρέχει ένα γενικευμένο μεθοδολογικό πλαίσιο για την επιλογή των ακουστικών μονάδων. Το γεγονός αυτό συνεπικουρείται με την ιδιότητα πλήρους αναζήτησης στο διαθέσιμο χώρο των ακουστικών μονάδων και εξασφάλισης της βέλτιστης ακολουθίας σύμφωνα με τα δοσμένα κριτήρια και προδιαγραφές. Με βάση αυτό το μεθοδολογικό πλαίσιο, διάφορες τεχνικές και προσεγγίσεις έχουν προταθεί στη βιβλιογραφία σχετικά με την σχεδίαση τόσο των συναρτήσεων κόστους όσο και των επιμέρους συναρτήσεων που τις αποτελούν. Μερικές από αυτές τις προσεγγίσεις παρουσιάζονται στην επόμενη ενότητα.

2.3 ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΤΗΝ ΣΧΕΔΙΑΣΗ ΤΩΝ ΣΥΝΑΡΤΗΣΕΩΝ ΚΟΣΤΟΥΣ

Η σχεδίαση των συναρτήσεων κόστους αποτελεί αφενός έναν από τους σημαντικότερους και αφετέρου έναν από τους πιο πολύπλοκους σχεδιαστικούς παράγοντες σε ένα σύστημα σύνθεσης φωνής από κείμενο. Από τους σημαντικότερους επειδή η τελική ποιότητα της συνθετικής φωνής εξαρτάται άμεσα από το αποτέλεσμα τους και από τους πιο πολύπλοκους επειδή ενέχουν τόσο την επιλογή όσο και την διαχείριση ενός πλήθους από παραμέτρους. Στην γενική μεθοδολογία που παρουσιάσαμε στην προηγούμενη ενότητα, η επιλογή του κατάλληλου διανύσματος χαρακτηριστικών σε κάθε τύπο κόστους, η εύρεση των κατάλληλων μετρικών μεταξύ των χαρακτηριστικών και η στάθμιση τους συντελούν σε αυτή την πολυπλοκότητα. Στα προηγούμενα συντελεί το γεγονός ότι το τελικό αποτέλεσμα πρέπει να συμφωνεί και να συσχετίζεται με την ανθρώπινη αντίληψη.

Η σχεδίαση των συναρτήσεων κόστους αποτελεί ανοιχτό ερευνητικό ζήτημα ενώ οι τεχνικές που έχουν προταθεί συνήθως προσεγγίζουν κάθε επιμέρους συνάρτηση κόστους μεμονωμένα αν και υπάρχουν προσπάθειες για την από κοινού σχεδίαση. Από τις δύο επιμέρους συναρτήσεις κόστους, η συνάρτηση κόστους «στόχος» θεωρείται πιο δύσκολη περίπτωση καθώς έχει να κάνει με τις προδιαγραφές που θέτει ένα κείμενο, ή πιο συγκεκριμένα μια πρόταση προς σύνθεση, γεγονός που από μόνο του δεν παρέχει μοναδική λύση. Τα γενικά χαρακτηριστικά όπως, ο τρόπος, το στυλ και η προσωδία, που θα πρέπει να ικανοποιεί μια πρόταση δεν είναι μοναδικά από την άποψη της υποκειμενικής (ανθρώπινης) αξιολόγησης. Στην περίπτωση της συνάρτησης κόστους «ένωσης», ο σκοπός είναι πιο σαφής, ωστόσο και εδώ η έλλειψη ενός αντικειμενικού κριτηρίου ποσοτικής αξιολόγησης της ποιότητας των ενώσεων με βάση την ανθρώπινη αντίληψη, παράλληλα με την αναζήτηση των κατάλληλων χαρακτηριστικών και των κατάλληλών μετρικών μεταξύ τους, αποτελεί συνεχές σχεδιαστικό ζητούμενο.

Στις ενότητες που ακολουθούν, παρουσιάζονται μερικές από τις προσεγγίσεις που ακολουθούνται στη σχεδίαση των επιμέρους συναρτήσεων κόστους. Και στις δύο περιπτώσεις η επιλογή του κατάλληλου διανύσματος χαρακτηριστικών καθώς και το είδος των χαρακτηριστικών (π.χ., γλωσσολογικά χαρακτηριστικά, ακουστικά χαρακτηριστικά κ.α.) αποτελούν κρίσιμο παράγοντα. Επίσης, ζητήματα όπως έλλειψη μετρικών με φυσική ερμηνεία και άμεση σύνδεση με το ζητούμενο πρόβλημα, η πεπερασμένη φύση στους δυνατούς συνδυασμούς των χαρακτηριστικών, η δυνατότητα σύνδεσης και εξαγωγής των χαρακτηριστικών από

και με τις προδιαγραφές, η στάθμιση των χαρακτηριστικών αποτελούν ζητούμενο και στις δύο περιπτώσεις.

2.3.1 Η συνάρτηση κόστους "στόχος" (target cost)

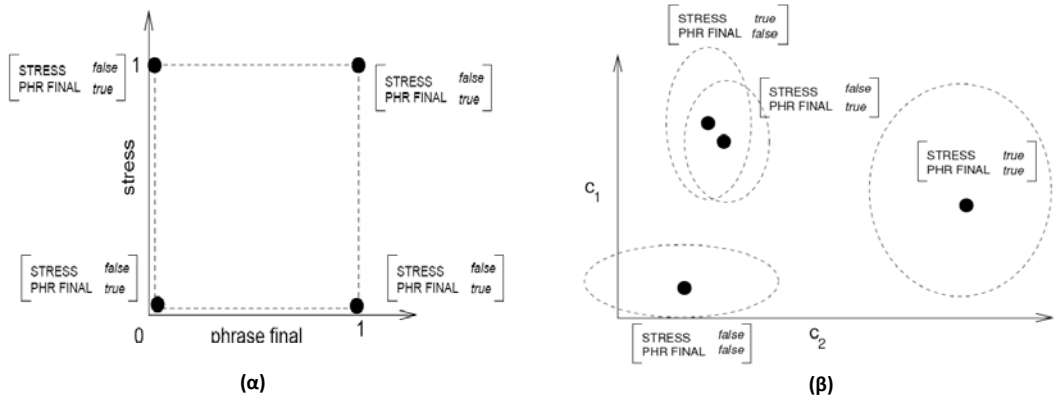
Όπως φαίνεται από τη σχέση 2.1, η ολική συνάρτηση κόστους «στόχος» αποτελείται από το σταθμισμένο άθροισμα επιμέρους συναρτήσεων κόστους η κάθε μια από τις οποίες αξιολογεί την απόκλιση των υποψηφίων ακουστικών μονάδων από την ιδεατή (ή επιθυμητή) ακουστική μονάδα βάση κάποιου χαρακτηριστικού και σύμφωνα με τις προδιαγραφές. Τόσο οι υποψήφιες ακουστικές μονάδες όσο και η επιθυμητή χαρακτηρίζονται και αναπαρίστανται από ένα σύνολο, ή αλλιώς ένα διάνυσμα, χαρακτηριστικών, το οποίο για τις πρώτες είναι μετρήσιμο ενώ για την δεύτερη εκτιμάται από τα δεδομένα της σύνθεσης (επιθυμητά χαρακτηριστικά). Οι παράμετροι που αποτελούν το διάνυσμα χαρακτηριστικών είναι ζήτημα του σχεδιαστή, ενώ μπορούν να χρησιμοποιηθούν χαρακτηριστικά είτε αριθμητικής είτε συμβολικής μορφής. Τα πιο συχνά χρησιμοποιούμενα χαρακτηριστικά είναι η επιθυμητή θεμελιώδης συχνότητα (pitch), ο επιτονισμός (intonation), η επιθυμητή διάρκεια, η επιθυμητή ένταση, η εξάρτηση από το περιβάλλον (phonetic context) κ.α. [Hunt, 1996; Strom, 2008; Taylor, 2006; Taylor, 2009; Fraser, 2007].

Ωστόσο, ο ακριβής αριθμητικός προσδιορισμός προσωδιακών μεγεθών (pitch, διάρκεια) από τις προδιαγραφές που θέτει η σύνθεση αποτελεί αφενός δύσκολη διαδικασία και αφετέρου δεν συνάδει με την ευρύτερη μεθοδολογία η οποία στηρίζεται στα διαθέσιμα δεδομένα και συνεπώς πρέπει κατά κάποιο τρόπο να τα «αντιγράφει» όπου αυτό είναι δυνατό. Η χρήση κατάλληλων χαρακτηριστικών τα οποία δεν προκύπτουν από κάποιο αυστηρό μοντέλο αποτελεί την πιο αποδεκτή προσέγγιση ώστε κατά τη σύνθεση να ξεσηκώνονται πρότυπα τα οποία υπάρχουν στην βάση της προηχογραφημένης φυσικής ομιλίας. Με αυτό τον τρόπο τα προσωδιακά χαρακτηριστικά δεν προσδιορίζονται αυστηρά (explicitly) αλλά προκύπτουν κατά τη σύνθεση. Για το λόγο αυτό, στα πιο σύγχρονα συστήματα συνήθως χρησιμοποιούνται χαρακτηριστικά κυρίως γλωσσολογικού τύπου (High-level linguistic features) αντί για ακουστικά χαρακτηριστικά (Low-level acoustic features) [Taylor, 1999; Taylor, 2006; Benesty, 2008 Ch. 21; Tihelka, 2005; Strom, 2008; Chu, 2001;]. Τα χαρακτηριστικά γλωσσολογικού τύπου ενέχουν κυρίως πληροφορία περιβάλλοντος για κάθε φώνημα όπως το φωνητικό περιβάλλον, ο τονισμός, η θέση στη συλλαβή, η θέση στη πρόταση, η έμφαση, το μέρος του λόγου κ.α., και η κύρια υπόθεση είναι ότι προσδιορίζουν ποιοτικά τον ζητούμενο επιτονισμό (intonation or pitch contour) και γενικότερα τα προσωδιακά χαρακτηριστικά. Εφόσον γίνει η σωστή επιλογή χαρακτηριστικών, ζήτημα υπό διερεύνηση ακόμα στη περιοχή της σύνθεση φωνής, τα επιθυμητά προσωδιακά χαρακτηριστικά αναμένεται ότι θα προκύπτουν ως αποτέλεσμα (by-product) από τον αλγόριθμο επιλογής ακουστικών μονάδων.

Το κυριότερο μειονέκτημα που προκύπτει από τη χρήση του σταθμισμένου αθροίσματος (σχέση 2.1) είναι ότι υποθέτει ότι κάθε χαρακτηριστικό λειτουργεί ανεξάρτητα (γραμμική σχέση) από τα υπόλοιπα. Αυτό έχει ως αποτέλεσμα να μην λαμβάνονται υπόψη αλληλεπιδράσεις μεταξύ των χαρακτηριστικών που έχουν διαφορετικό αντίκτυπο στο ακουστικό αποτέλεσμα. Αυτό φαίνεται και από τους παράγοντες στάθμισης όπου χρησιμοποιείται ένας για κάθε χαρακτηριστικό ενώ

στη πραγματικότητα θα έπρεπε να χρησιμοποιείται ένας για κάθε συνδυασμό χαρακτηριστικών (π.χ., *για q=2 χαρακτηριστικά με, έστω, δυνατότητα 3 τιμών για το καθένα έχουμε 2 παράγοντες στάθμισης ενώ θα μπορούσαν να χρησιμοποιηθούν* $3^q=3^2=9$ *παράγοντες στάθμισης, ένας για κάθε συνδυασμό τιμών των χαρακτηριστικών. Ωστόσο αυτό είναι απαγορευτικό σε πρακτική υλοποίηση*). Στην περίπτωση της φωνής, υπάρχουν πολλές περιπτώσεις στις οποίες διαφορετικοί συνδυασμοί χαρακτηριστικών μπορούν να αντιστοιχούν στον ίδιο ακουστικό χώρο άρα και στον ίδιο (ή παρόμοιο ήχο).

Για να αντιμετωπιστεί αυτός ο περιορισμός, το κόστος «στόχος» μπορεί να σχηματιστεί με διαφορετική προσέγγιση στην οποία χρησιμοποιείται κάποια συνάρτηση ή μηχανισμός που εκπαιδεύεται από τα δεδομένα και χρησιμοποιείται για την προβολή των συνδυασμών των χαρακτηριστικών σε ένα σημείο στον ακουστικό χώρο (acoustic space). Σε αυτή τη περίπτωση, για τον προσδιορισμό του κόστους, χρησιμοποιείται πλέον κάποια ακουστικού τύπου μετρική μεταξύ των υποψήφιων ακουστικών μονάδων και των επιθυμητών όπως προκύπτουν από την εκπαίδευση. Ο μηχανισμός που χρησιμοποιείται συνήθως είναι τα δένδρα απόφασης με τα οποία πραγματοποιείται συσταδοποίηση εξαρτώμενη από γλωσσολογική πληροφορία περιβάλλοντος (context dependent clustering) [Hamza; 2001; Donovan, 1996]. Το μειονέκτημα αυτής της προσέγγισης έγκειται αφενός στον ορισμό και την περιγραφή του ακουστικού χώρου (συνήθως χρησιμοποιείται ο ακουστικός χώρος που ορίζεται από cepstral συντελεστές) και αφετέρου στην μη επαρκή συσταδοποίηση λόγω των πολλών συνδυασμών των χαρακτηριστικών τα οποία δεν συναντώνται κατά την εκπαίδευση (unseen feature descriptions) [Taylor, 2006]. Στο σχήμα 2.2, παρουσιάζονται οι δύο προσεγγίσεις για τον σχηματισμό της συνάρτησης κόστους «στόχος».



ΣΧΗΜΑ 2.2: Η συνάρτηση κόστους «στόχος», α) μέσω σταθμισμένου αθροίσματος με συνδυασμό χαρακτηριστικών και χρήση επιμέρους κοστών για κάθε χαρακτηριστικό ξεχωριστά και β) με προβολή των χαρακτηριστικών στον ακουστικό χώρο (συντελεστές cepstral) μέσω συσταδοποίησης (στο γράφημα εμφανίζονται ως παράδειγμα οι 2 πρώτοι συντελεστές cepstral) [Πηγή: Taylor, 2006].

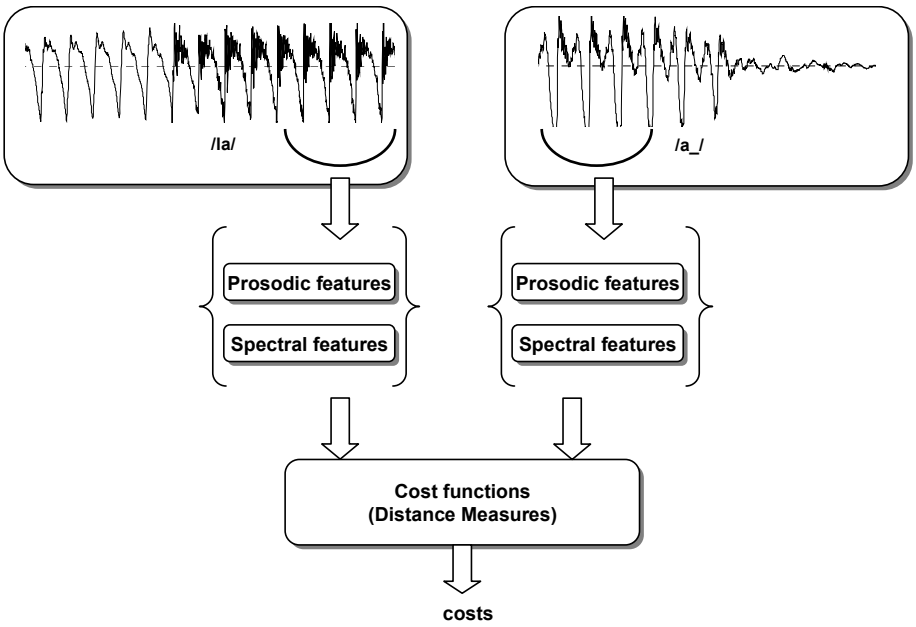
Το σύστημα που χρησιμοποιήθηκε στα πλαίσια της διατριβής ακολουθεί την προσέγγιση της σχέσης 2.1 και χρησιμοποιεί δύο χαρακτηριστικά (features) για την συνάρτηση κόστους «στόχος». Το ένα αφορά πληροφορία περιβάλλοντος και εξετάζει αν η υποψήφια ακουστική μονάδα προέρχεται από ίδιο φωνητικό

περιβάλλον με αυτό που συναντάται η επιθυμητή. Ο έλεγχος γίνεται σε βάθος ενός φωνήματος δεξιά και αριστερά του υποψήφιου. Το δεύτερο χαρακτηριστικό ελέγχει την επιθυμητή προσωδία όχι σε επίπεδο τιμών αλλά σε επίπεδο κινήσεων της καμπύλης επιτονισμού ανάμεσα σε σημεία ενδιαφέροντος. Τα σημεία ενδιαφέροντος προκύπτουν από γλωσσολογικά χαρακτηριστικά όπως από παύση σε κόμμα, από κόμμα σε τελεία, από κόμμα σε κόμμα κτλ. και στα οποία συνήθως διαμορφώνεται ο επιτονισμός. Η λογική της σχεδίασης εμπίπτει στη πρακτική που αναφέρθηκε προηγουμένως, και έχει ως ζητούμενο τα επιθυμητά προσωδιακά χαρακτηριστικά να ανασύρονται και να μιμούνται αυτά που συναντώνται στη διαθέσιμη βάση και, τελικά, κατά τη σύνθεση να προκύπτουν ως αποτέλεσμα (by-product) από τον αλγόριθμο επιλογής ακουστικών μονάδων.

2.3.2 Η συνάρτηση κόστους ένωσης (concatenation ή join cost)

Ο σκοπός της συνάρτησης κόστους «ένωσης» είναι να αξιολογεί την καταλληλότητα της συρραφής μεταξύ των υποψηφίων ακουστικών μονάδων. Φασματικές και προσωδιακές ασυμβατότητες στην ένωση των ακουστικών μονάδων οδηγούν σε σημαντική υποβάθμιση της ποιότητας [Klabbers, 2007; Plumpe, 1998]. Σύμφωνα με τη σχέση 2.2, προκύπτει και αυτή από το σταθμισμένο άθροισμα επιμέρους συναρτήσεων ή μετρικών που έχουν το ρόλο του επιμέρους κόστους. Συνεπώς, η αξιολόγηση στηρίζεται στα επιμέρους κόστη και το ζητούμενο είναι αυτά να συσχετίζονται όσο το δυνατόν περισσότερο με την ανθρώπινη αντίληψη. Και σε αυτή τη περίπτωση, οι υποψήφιες ακουστικές μονάδες χαρακτηρίζονται και αναπαρίστανται από ένα διάνυσμα χαρακτηριστικών (σχήμα 2.3). Οι παράμετροι που αποτελούν το διάνυσμα χαρακτηριστικών είναι ζήτημα του σχεδιαστή. Τα πιο συχνά χρησιμοποιούμενα χαρακτηριστικά είναι η θεμελιώδης συχνότητα (pitch) και κάποια φασματική αναπαράσταση ενώ πολλές φορές συμπεριλαμβάνονται μεγέθη όπως η ένταση (intensity) και η διάρκεια καθώς και οι πρώτες και δεύτερες παράγωγοι όλων των προηγούμενων μεγεθών [Hunt, 1996; Fraser, 2007; Plumpe, 1998; Blouin, 2002; Hirchfeld, 2000; Kirkpatrick, 2007]. Κάθε χαρακτηριστικό συνοδεύεται και από μια μετρική η οποία παίζει το ρόλο τους επιμέρους κόστους. Με αυτό τον τρόπο τα επιμέρους κόστη είναι υπεύθυνα για να προσδιορίσουν δύο κόστη, το **κόστος των προσωδιακών ασυνεχειών** και το **κόστος φασματικής ασυνέχειας**.

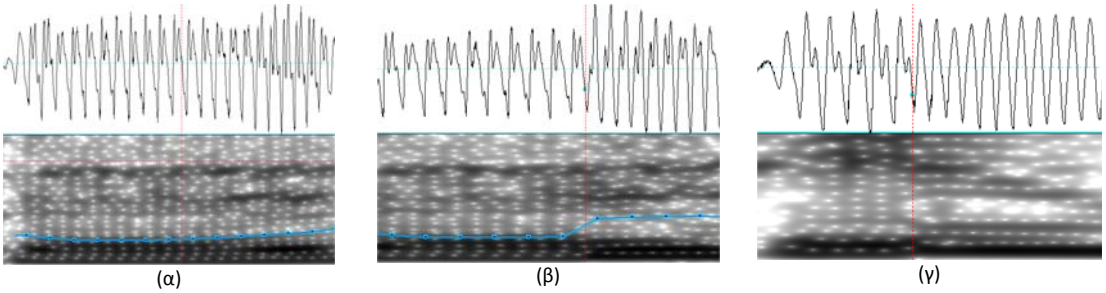
Μεταξύ των παραπάνω το κόστος προσωδιακών ασυνεχειών θεωρείται πιο εύκολη περίπτωση και συνήθως υπολογίζεται ως η διαφορά των τιμών της θεμελιώδους συχνότητας (pitch) στα όρια της ένωσης είτε σε γραμμική (*F0*) είτε σε λογαριθμική κλίμακα (*logF0*). Η ίδια απλή προσέγγιση ακολουθείται και για τις παραγώγους του pitch. Αντίθετα, το κόστος φασματικής ασυνέχειας, το οποίο εξετάζεται αναλυτικά παρακάτω, θεωρείται πιο σύνθετη περίπτωση η οποία προσεγγίζεται με διάφορους τρόπους [Klabbers, 2007; Pantazis, 2005b; Vepa, 2004; Jun Hu, 2006]. Λόγω αυτού του διαχωρισμού, σε πολλές περιπτώσεις το κόστος ένωσης αναφέρεται είτε μόνο στο κόστος των φασματικών ασυνεχειών είτε, σε πιο σύγχρονες προσεγγίσεις [Ling, 2007, Latacz, 2009; Sakai, 2005; Sakai, 2006; Sakai, 2009; Schroeder, 2008; Yoshida, 2008], τα δύο κόστη αντιμετωπίζονται από κοινού και το κόστος ένωσης δεν διαχωρίζεται ανάμεσα στα δύο.



ΣΧΗΜΑ 2.3: Η διαδικασία εξαγωγής χαρακτηριστικών και ο υπολογισμός των επιμέρους συναρτήσεων κόστους στη συνάρτηση κόστους «ένωσης».

2.3.2.1 Το Κόστος φασματικής ασυνέχειας

Ο προσδιορισμός του κόστους φασματικής ασυνέχειας αποτελεί σημαντικό ζήτημα στην σύνθεση φωνής με επιλογή και συρραφή ακουστικών μονάδων και έχει απασχολήσει ένα ευρύ σύνολο της ερευνητικής κοινότητας που ασχολείται με το αντικείμενο. Έχουν προταθεί διάφορες προσεγγίσεις ωστόσο δεν έχει επικρατήσει κάποια λύση και αποτελεί ανοιχτό ερευνητικό ζήτημα. Όπως φαίνεται στο σχήμα 2.3, ο υπολογισμός του ενέχει την παραμετροποίηση του σήματος φωνής των ακουστικών μονάδων στο όριο της συρραφής και την αναπαράσταση μέσω φασματικών χαρακτηριστικών. Στη συνέχεια υπολογίζεται κάποια φασματική απόσταση μεταξύ αυτών των χαρακτηριστικών η τιμή της οποίας προσδιορίζει και το ζητούμενο κόστος. Όπως θα δούμε, διάφορες παραμετροποιήσεις έχουν προταθεί για την φασματική αναπαράσταση του σήματος φωνής, με δημοφιλέστερη αυτή των MFCC (Mel-Frequency Cepstral Coefficients) [Davis, 1980] η οποία έχει επικρατήσει στην περίπτωση της αναγνώρισης φωνής. Σε αντιστοιχία, στα πλαίσια του υπολογισμού του κόστους φασματικής ασυνέχειας, έχουν δοκιμαστεί μεταξύ άλλων οι γνωστότερες φασματικές αποστάσεις από την περιοχή της επεξεργασίας φωνής [Gray, 1976; Rabiner, 1993]. Ωστόσο, το ζητούμενο από το κόστος φασματικής ασυνέχειας είναι να αντικατοπτρίζει την ανθρώπινη αντίληψη, οπότε τόσο η επιλογή των φασματικών χαρακτηριστικών όσο και η επιλογή της φασματικής απόστασης πρέπει να ικανοποιούν αυτό το αίτημα. Πράγματι, το αποτέλεσμα κάποιας φασματικής απόστασης (και εν γένει του κόστους ένωσης), πρέπει να πληρεί το αίτημα για υψηλή συσχέτιση σε σχέση με την ανθρώπινη αντίληψη (perceptually correlated) περί φασματικής ή γενικά ακουστικής ασυνέχειας. Για παράδειγμα, όταν το ανθρώπινο αυτί αντιλαμβάνεται ακουστική παραμόρφωση, λόγω της ένωσης των κυματομορφών, τότε η απόσταση πρέπει να δίνει μεγάλη τιμή ενώ στην αντίθετη περίπτωση χαμηλή.



ΣΧΗΜΑ 2.4: Παραδείγματα ακουστικών ασυνεχειών στην ένωση των ακουστικών μονάδων: α) συρραφή χωρίς ακουστικές ασυνέχειες, β) συρραφή με προσωδιακές και φασματικές ασυνέχειες, γ) συρραφή με έντονες φασματικές ασυνέχειες. Σε κάθε περίπτωση δίνεται η χρονική κυματομορφή με το αντίστοιχο ηχογράφημα. Η μπλε καμπύλη δείχνει την τροχιά του pitch, ενώ η κόκκινη κάθετη γραμμή δείχνει το σημείο ένωσης.

Οι πρώτες προσπάθειες για το σκοπό αυτό, επικεντρώθηκαν στην αναζήτηση τόσο της κατάλληλης αναπαράστασης όσο και της κατάλληλης (φασματικής) απόστασης. Μεταξύ άλλων, οι φασματικές αναπαραστάσεις που έχουν διερευνηθεί κατά καιρούς συμπεριλαμβάνουν τις εξής:

- **Cepstral συντελεστές μέσω FFT και μέσω LPC** [Rabiner, 1993; Rabiner, 1978] στα [Wouters, 1998; Bjorkan, 2005]
- **MFCC συντελεστές μέσω FFT και μέσω LPC** [Davis, 1980; Rabiner, 1993] στα [Vepa, 2002; Vepa, 2004; Vepa, 2006; Klabbers, 2001; Stylianou, 2001; Donovan, 2001; Kirkpatrick, 2006; Bjorkan, 2005; Tsuzaki, 2001]
- **Line Spectral Pairs (LSP)/ Line Spectral Frequencies (LSF)** [Krishnan, 1996; Chappell, 2002; So, 2007] στα [Vepa, 2002; Stylianou, 2001; Kirkpatrick, 2006]
- **Perceptual Linear Prediction (PLP)** [Hernansky, 1990] συντελεστές στο [Stylianou, 2001; Klabbers, 2001; Kirkpatrick, 2006; Kain, 2007]
- **LPC smooth spectrum και άλλα χαρακτηριστικά από LPC ανάλυση** [Quatieri, 2001; Rabiner, 1993; Rabiner, 1978] στα [Stylianou, 2001; Bjorkan, 2005; Vepa, 2002]
- **Auditory-based features** στα [Hansen, 1998; Tsuzaki, 2001; Tsuzaki, 2002]
- **Wavelets** [Quatieri, 2001] στα [Kirkpatrick, 2006]
- **Time-Frequency distributions (όπως Mellin, Vigner-Ville) και Bispectrum** [Quatieri, 2001] στο [Chen, 1999]
- **Formants** [Vepa, 2002; Vepa, 2004; Klabbers, 2001]
- **Multiple Centroids Analysis (MCA)** [Vepa, 2002; Vepa, 2004; Vepa, 2006]
- **Amplitude and Frequency modulation components (AM/FM)** [Quatieri, 2001] στα [Pantazis, 2005; Pantazis, 2005b]
- **Harmonic plus Noise** [Stylianou, 2001b] στα [Pantazis, 2005; Pantazis, 2005b]
- **FFT based log power spectrum** [Rabiner, 1978] στα [Klabbers, 2001; Stylianou, 2001]

Επιπρόσθετα, στη διεθνή αρθρογραφία έχουν μελετηθεί και δοκιμαστεί πλήθος αποστάσεων. Πολλές έχουν δώσει ικανοποιητικά αποτελέσματα ενώ άλλες δείχνουν να υπολείπονται σε απόδοση. Στις κυριότερες αποστάσεις που έχουν διερευνηθεί για το ρόλο των φασματικών αποστάσεων συμπεριλαμβάνονται οι εξής [Rabiner, 1993; Gray, 1976; Taylor, 2009]:

Αν $\boldsymbol{X}$, $\boldsymbol{Y}$ συμβολίζουν δύο διανύσματα χαρακτηριστικών διάστασης N τότε,

- **Manhattan ή City-Block**:

$$D(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^{N} |X_i - Y_i| \tag{2.6}$$

- **Euclidean**:

$$D(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{\sum_{i=1}^{N} (X_i - Y_i)^2} \tag{2.7}$$

- **Mahalanobis**:

  Η Mahalanobis αποτελεί γενίκευση της Ευκλείδειας απόστασης υπό την έννοια ότι λαμβάνει υπόψη την συμμεταβλητότητα (ή συνδιακύμανση) μεταξύ των χαρακτηριστικών. Υπολογίζεται ως,

$$D(\boldsymbol{X},\boldsymbol{Y}) = (\boldsymbol{X} - \boldsymbol{Y})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{Y}) \tag{2.8}$$

  όπου Σ ο πίνακας συμμεταβλητότητας (Covariance Matrix) ή

$$D(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{\sum_{i=1}^{N} \left(\frac{X_i - Y_i}{\sigma_i} \right)^2} \tag{2.9}$$

  με διαγώνιο πίνακα συμμεταβλητότητας, όπου σ_i η τυπική απόκλιση για το i χαρακτηριστικό.

- **Kullback-Leibler divergence**:

  Η συμμετρική απόσταση Kullback-Leibler προέρχεται από το χώρο της στατιστικής και χρησιμοποιείται για να μετρηθεί η απόσταση μεταξύ κατανομών πιθανότητας. Δίνεται από τη σχέση:

$$D = \int (f(x) - g(x)) \log\left(\frac{f(x)}{g(x)}\right) dx \tag{2.10}$$

  Στην σύνθεση φωνής συνήθως υπολογίζεται ως,

$$D(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^{N} (X_i - Y_i) \log\left(\frac{X_i}{Y_i}\right) \tag{2.11}$$

  όπου αποτελεί τη συμμετρική διακριτή προσέγγιση της.

- **Itakura (ή Likelihood Ratio)**:

  Η απόσταση Itakura φανερώνει την φασματική ομοιότητα μεταξύ δύο διανυσμάτων, που περιέχουν τους συντελεστές γραμμικής πρόβλεψης από ένα ζεύγος σημάτων. Επιδιώκει να δείξει κατά πόσο το φίλτρο της LPC ανάλυσης ενός διανύσματος μπορεί να προβλέψει το άλλο. Δίνεται από τη σχέση,

$$D(\boldsymbol{X},\boldsymbol{Y}) = \frac{\boldsymbol{X}^{\boldsymbol{T}} R_{\boldsymbol{Y}} \boldsymbol{X}}{\boldsymbol{Y}^{\boldsymbol{T}} R_{\boldsymbol{Y}} \boldsymbol{Y}} - 1 \tag{2.12}$$

όπου σε αυτή τη περίπτωση τα $\boldsymbol{X}$, $\boldsymbol{Y}$ αποτελούν διανύσματα που έχουν προέλθει από γραμμική πρόβλεψη (LPC), R_Y ο πίνακας αυτοσυσχέτισης (autocorrelation matrix) από τον οποίο προέρχεται ο προβλέπτης $\boldsymbol{Y}$.

Στην πλειοψηφία των περιπτώσεων, η αξιολόγηση των φασματικών αναπαραστάσεων και των αντίστοιχων αποστάσεων γίνεται με υλοποίηση ακουστικών πειραμάτων. Η αξιολόγηση πραγματοποιείται με χρήση διαφόρων στατιστικών μεγεθών όπως, το Mean Opinion Score (MOS), το Receiver Operating Characteristic (ROC), το Prediction Rate κ.α. [Taylor, 2009; Jurafsky, 2008; Huang, 2001]. Σε γενικές γραμμές, τα πειράματα αφορούν την παραγωγή συνθετικών λέξεων ή τριγραμμάτων τύπου CVC (Consonant-Vowel-Consonant) με χρήση πολλαπλών πραγματώσεων ίδιων φωνημάτων, όπου μια ομάδα ανθρώπων αξιολογεί το ακουστικό αποτέλεσμα των ενώσεων. Η αξιολόγηση πραγματοποιείται είτε με δυαδικού τύπου απόφαση (binary) [Klabbers, 2001; Stylianou, 2001; Pantazis, 2005 ; Kirkpatrick, 2006; Bjorkan, 2005] είτε με αποφάσεις σε κλίμακα 1-5 που εκφράζει την μέση ακουστική αξιολόγηση (Mean Opinion Score - MOS) [Vepa, 2004; Vepa, 2006; Wouters, 1998]. Τα αποτελέσματα του πειράματος συσχετίζονται με αυτά των φασματικών αποστάσεων, οπότε προκύπτουν συμπεράσματα σχετικά με το ποιά απόσταση προσεγγίζει καλύτερα την ακουστική αντίληψη των ακροατών. Στον πίνακα που ακολουθεί παρακάτω, φαίνεται συγκεντρωτικά η αξιολόγηση εκείνων των φασματικών αναπαραστάσεων και αποστάσεων, που περιλαμβάνονται σε περισσότερες από μία μελέτες που υπάρχουν στη διεθνή αρθρογραφία.

Επιπλέον, ένα σημαντικό στοιχείο που προκύπτει από τη μελέτη της αρθρογραφίας είναι τα αντικρουόμενα αποτελέσματα που προκύπτουν. Για παράδειγμα, στο [Klabbers, 2001] η αξιολόγηση της Euclidean MFCC προκύπτει χειρότερη απο αυτήν της Kullback-Leibler. Αντίθετα, στη μελέτη των Wouters και Macon [Wouters, 1998] η αξιολόγηση της Euclidean MFCC είναι αρκετά υψηλή. Επιπλέον, στο [Stylianou, 2001] τα αποτελέσματα είναι αρκετά ισορροπημένα, αφού και οι δύο αποστάσεις αξιολογούνται εξίσου θετικά. Επίσης, σημαντικός παράγοντας δείχνει να είναι και η φασματική αναπαράσταση. Για παράδειγμα, στη μελέτη που παρουσιάζεται στο [Stylianou, 2001], η απόσταση Kullback-Leibler φαίνεται να αποδίδει καλύτερα όταν υπολογίζεται από το φάσμα ισχύος που προκύπτει βάση του μετασχηματισμού Fourier. Αντίθετα, η αξιολόγηση της είναι χαμηλή, όταν χρησιμοποιείται η φασματική περιβάλλουσα που προκύπτει από το παραμετρικό μοντέλο της γραμμικής πρόβλεψης. Αυτό δεν συμβαίνει στην περίπτωση της μελέτης που παρουσιάζεται στο [Klabbers, 2001].

Η αξιολόγηση που αναφέρεται στο πίνακα αφορά τη σχετική σύγκριση με άλλες αναπαραστάσεις και αποστάσεις στην συγκεκριμένη μελέτη και δεν χαρακτηρίζει τη γενικότερη επίδοση ως κατάλληλο κόστος για τις φασματικές ασυνέχειες. Κοινός παρανομαστής τόσο στις προηγούμενες μελέτες όσο και σε άλλες που ακολούθησαν και αναφέρθηκαν προηγουμένως, αποτελεί το γεγονός ότι σε καμία περίπτωση δεν επιτυγχάνεται αρκετά υψηλό ποσοστό συσχέτισης με την ανθρώπινη ακουστική αντίληψη. Τίθεται δηλαδή ένα άνω όριο στο βαθμό επιτυχίας που επιτυγχάνεται από τη μεθοδολογική προσέγγιση αυτού του τύπου οπότε βελτίωση αναμένεται είτε μέσω κάποιας διαφορετικής προσέγγισης είτε μέσω κατάλληλου συνδυασμού των παραπάνω φασματικών αναπαραστάσεων και αποστάσεων.

ΠΙΝΑΚΑΣ 2.1: Βιβλιογραφική επισκόπηση αξιολόγησης φασματικών αναπαραστάσεων και αποστάσεων

ΤΥΠΟΣ ΑΠΟΣΤΑΣΗΣ	ΦΑΣΜΑΤΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ	ΤΡΟΠΟΣ ΥΠΟΛΟΓΙΣΜΟΥ	ΑΞΙΟΛΟΓΗΣΗ*	ΜΕΛΕΤΗ
EUCLIDEAN	MFCC	Με χρήση LPC.	- Καλή.	[Stylianou, 2001], [Wouters, 1998]
			- Κακή	[Klabbers, 2001], [Chen, 1999]
KULLBACK-LEIBLER	Περιβάλλουσα LPC	Με χρήση LPC.	- Καλή.	[Klabbers, 2001], [Founda, 2001]
			- Μέτρια	[Stylianou, 2001]
	FFT	Με χρήση FFT.	- Καλή.	[Stylianou, 2001]
ITAKURA	LPC	Με χρήση LPC.	- Καλή.	[Chen, 1999], [Klabbers, 2001], [Wouters, 1998]
			-Κακή	[Stylianou, 2001]
EUCLIDEAN	Log Power Spectra	Με χρήση LPC.	-Κακή	[Stylianou, 2001]
		Με χρήση FFT.	-Καλή	
EUCLIDEAN	Formants	Είτε με χρήση LPC είτε χειρωνακτικά.	- Κακή	[Founda, 2001], [Klabbers, 2001] [Vepa, 2004, 2006]
MAHALANOBIS	MFCC	Με χρήση FFT	- Καλή	[Donovan, 2001] [Vepa, 2004, 2006]

**Ο χαρακτηρισμός της αξιολόγησης (καλή – κακή – μέτρια κτλ.) αναφέρεται στη σχετική σύγκριση με άλλες αναπαραστάσεις και αποστάσεις στην συγκεκριμένη μελέτη και δεν χαρακτηρίζει τη γενικότερη επίδοση ως κόστος για τις φασματικές ασυνέχειες.*

Προς αυτή τη κατεύθυνση κινήθηκαν διάφορες μελέτες αλλά προτάθηκαν και νέες προσεγγίσεις. Για παράδειγμα, στο [Bulyko, 2001] προτείνεται η χρήση ενός επιπλέον κόστους που να αφορά και να αξιολογεί την τάξη (phone class) των ακουστικών μονάδων που πρόκειται να ενωθούν, αφού διαφορετικού τύπου ακουστικές μονάδες ενώνονται καλύτερα σε σχέση με άλλες [Syrdal, 2004; Syrdal, 2005]. Η προσέγγιση αυτή ταιριάζει καλύτερα σε συστήματα με μεταβλητού μήκους ακουστικές μονάδες. Στα [Coorman, 2000; Hirschfeld, 2000; Syrdal, 2004; Kawai, 2002] εξετάζεται η χρήση μόνο χαρακτηριστικών γλωσσολογικού τύπου αντί ακουστικών ή και συνδυασμοί των δύο. Τα γλωσσολογικού τύπου χαρακτηριστικά εκτός ότι ενσωματώνουν πληροφορία περιβάλλοντος (context), λαμβάνουν υπόψη ίδιες κατηγορίες πραγματώσεων των ακουστικών μονάδων (π.χ., ακουστικές μονάδες στο ίδιο context, ίδια συλλαβή, με ίδιο τύπο τονισμού κ.α.). Επίσης, η χρήση γλωσσολογικών χαρακτηριστικών αντισταθμίζει την βραχέως χρόνου εκτίμηση των ακουστικών χαρακτηριστικών στα όρια της ένωσης. Τα χαρακτηριστικά αυτά έχουν σημαντική συνεισφορά στην ακουστική αντίληψη. Οι προηγούμενες προσεγγίσεις ταιριάζουν καλύτερα τόσο σε συστήματα με

μεταβλητού μήκους ακουστικές μονάδες όσο και σε συστήματα που χρησιμοποιούν φωνήματα.

Επιπλέον, στην αρθρογραφία συναντώνται προσπάθειες που προσπαθούν να αντιμετωπίσουν εγγενώς το πρόβλημα της εκτίμησης βραχέως χρόνου των ακουστικών χαρακτηριστικών [Vepa, 2006; Taylor, 2006] αλλά και να ενσωματώσουν πληροφορία ανθρώπινης κρίσης [Vepa, 2004; Pantazis, 2005]. Στα [Vepa, 2002; Vepa, 2004] διερευνάται η χρήση γραμμικού συνδυασμού (σταθμισμένο άθροισμα) φασματικών αποστάσεων και φασματικών χαρακτηριστικών στο οποίο οι παράγοντες στάθμισης προσδιορίζονται με τη βοήθεια των αποτελεσμάτων αξιολόγησης που προέκυψαν από ακουστικά πειράματα, μέσω της επίλυσης ενός γραμμικού συστήματος εξισώσεων. Τα αποτελέσματα τους ανέδειξαν την υπόθεση ότι ο συνδυασμός φασματικών αναπαραστάσεων και αποστάσεων οδηγεί σε καλύτερα αποτελέσματα. Στις εργασίες [Pantazis, 2005; Pantazis, 2005b], διερευνάται η χρήση μη γραμμικών φασματικών αναπαραστάσεων (όπως είναι η αναπαράσταση AM/FM), για την σχεδίαση ενός ταξινομητή με στόχο τον διαχωρισμό μεταξύ κατάλληλων και ακατάλληλων ενώσεων. Τα δεδομένα εκπαίδευσης του ταξινομητή προέκυψαν από ακουστικά πειράματα τα οποία είχαν ζητούμενο το χαρακτηρισμό ενός πλήθους από ενώσεις σε «συνεχείς» και «ασυνεχείς». Η πειραματική τους αξιολόγηση ανέδειξε ότι η προσέγγιση αυτή είναι επιτυχής αφού πέτυχαν βελτίωση της τάξης του 90% σε σχέση με προηγούμενες τους μελέτες σε ίδια βάση δεδομένων. Το σημαντικότερο μειονέκτημα στις προηγούμενες προσεγγίσεις είναι ότι στηρίζονται σε αποτελέσματα που προέρχονται από ακουστικά πειράματα τα οποία αποτελούν μια χρονοβόρα και πολυδάπανη διαδικασία τόσο σε πόρους όσο και σε ανθρώπινο δυναμικό καθώς επίσης δεν είναι ξεκάθαρο αν είναι ανεξάρτητες του εκάστοτε ομιλητή και της βάσης της φυσικής ομιλίας.

Όπως και στην περίπτωση του κόστους «στόχος», η τάση για τη σχεδίαση του κόστους των φασματικών ασυνεχειών (όσο και συνολικά του κόστους «ένωσης») προσανατολίζεται στο να εκμεταλλεύεται όσο το δυνατόν καλύτερα τα διαθέσιμα δεδομένα (data driven), δηλαδή την διαθέσιμη βάση δεδομένων, έτσι ώστε η συνάρτηση κόστους να προκύπτει με εκπαίδευση πάνω σε αυτή, αποφεύγοντας τον περιορισμό που αναφέραμε, την εξάρτηση από τον ανθρώπινο παράγοντα. Προς αυτή τη κατεύθυνση έχουν ήδη κινηθεί διάφορες τεχνικές. Πριν τις αναφέρουμε, να σημειώσουμε το γεγονός ότι εγγενές χαρακτηριστικό της διαθέσιμης βάσης δεδομένων είναι ότι αποτελείται από ένα τεράστιο αριθμό από «φυσικές» ενώσεις για κάθε τύπο ακουστικής μονάδας. Πράγματι, κάθε πλαίσιο φωνής με το γειτονικό του μέσα στη βάση αποτελεί «φυσική» ένωση. Ωστόσο, η μη στάσιμη φύση του σήματος φωνής (quasi stationary) οδηγεί στο γεγονός αυτή η ιδιότητα να μην αντικατοπτρίζεται στις φασματικές αποστάσεις. Το γεγονός αυτό εκμεταλλεύονται διάφορα συστήματα σύνθεσης φωνής στις συναρτήσεις κόστους, στις οποίες ορίζουν να επιστρέφουν μηδενική τιμή (μηδέν κόστος) για γειτονικές ακουστικές μονάδες από τη βάση δεδομένων. Η φασματική απόσταση γειτονικών πλαισίων φωνής σπάνια είναι μηδενική. Ωστόσο, από την άποψη της ακουστικής αντίληψης η ένωση των δύο τέτοιων πλαισίων οδηγεί σε απόλυτα «φυσικό» αποτέλεσμα από την άποψη της ακουστικής αντίληψης. Την παρατήρηση αυτή εκμεταλλεύεται και η τεχνική που προτείνεται στα πλαίσια της διατριβής και η οποία εξετάζεται σε επόμενο κεφάλαιο.

Σε αυτό το σκεπτικό, στο [Bellegarda, 2006] εξετάζεται μια διαφορετική προσέγγιση που λειτουργεί στο πεδίο του χρόνου και εξετάζει ακολουθίες από πλαίσια φωνής για κάθε φώνημα. Συγκεκριμένα, τα πλαίσια, που συμμετέχουν στην ένωση, για κάθε φώνημα οργανώνονται σε ένα πίνακα και στη συνέχεια μετασχηματίζονται μέσω ανάλυσης σε ιδιάζουσες τιμές (Singular Value Decomposition - SVD). Οι γραμμές του πίνακα αντιπροσωπεύουν τα πλαίσια ενώ οι στήλες τις πραγματώσεις των ακουστικών μονάδων. Το φασματικό κόστος ένωσης υπολογίζεται από το συνημίτονο της γωνίας μεταξύ των διανυσμάτων του πίνακα. Η αξιολόγηση της τεχνικής με ακουστικά πειράματα ανέδειξε την υπεροχή της συγκριτικά με την Ευκλείδεια απόσταση ανάμεσα σε MFCC συντελεστές. Η προσέγγιση αυτή δεν έχει αξιολογηθεί σε άλλες εργασίες.

Στα [Vepa, 2002; Vepa, 2004; Vepa, 2006] προτείνεται η μοντελοποίηση της ακολουθίας από πλαίσια γύρω από την ένωση με χρήση γραμμικών δυναμικών μοντέλων και συγκεκριμένα με χρήση φίλτρου Kalman. Το φίλτρο Kalman χρησιμοποιήθηκε για την μοντελοποίηση και την εκτίμηση της τροχιάς των φασματικών παραμέτρων σε ακολουθία από πλαίσια γύρω από την ένωση. Η εκπαίδευση του φίλτρου πραγματοποιήθηκε σε φυσική ομιλία. Η φασματική αναπαράσταση που δοκιμάστηκε ήταν η LSF [Krishnan, 1996; Chappell, 2002; So, 2007] Σε γενικές γραμμές, το φασματικό κόστος ένωσης υπολογίζεται από το λάθος, ή αλλιώς από την απόκλιση των εκτιμώμενων τροχιών με τις τροχιές που συναντώνται στην ένωση. Η πειραματική αξιολόγηση ανέδειξε τη λειτουργικότητα της μεθόδου χωρίς ωστόσο να επιτυγχάνει σημαντικά ποσοστά συσχέτισης με την ανθρώπινη ακουστική αντίληψη. Η χρήση των τροχιών LSF εγείρει ερωτήματα, καθώς σε προηγούμενες μελέτες, τα LSF δεν συμπεριλαμβάνονται μεταξύ των καλύτερων φασματικών αναπαραστάσεων για το φασματικό κόστος.

Μια διαφορετική τεχνική, στην ίδια όμως λογική, περιγράφεται στο [Taylor, 2006] όπου το πρόβλημα αντιμετωπίζεται με πιθανοτική προσέγγιση. Συγκεκριμένα, εκτιμάται η πιθανότητα εμφάνισης από ακολουθίες πλαισίων φωνής γύρω από την ένωση. Τα πλαίσια φωνής αντιμετωπίζονται ως n-grams (συνήθως 2 πλαίσια πριν και μετά την ένωση) και η εκτίμηση της πιθανότητας της ακολουθίας των πλαισίων πραγματοποιείται από την υπάρχουσα βάση δεδομένων. Η φασματική συμβατότητα των ακουστικών μονάδων που θα ενωθούν εκτιμάται από την πιθανότητα εμφάνισης της ακολουθίας πλαισίων. Στο [Taylor, 2006] δεν πραγματοποιήθηκε αξιολόγηση της τεχνικής. Παρόμοιας φιλοσοφίας τεχνικές για τις συναρτήσεις κόστους προτείνονται στα [Sakai, 2005; Sakai, 2006; Sakai, 2009; Yoshida, 2008].

Ανάλογες προσεγγίσεις που στηρίζονται σε τεχνικές μηχανικής μάθησης έχουν πρόσφατα προταθεί στα [Ling, 2007, Latacz, 2009; Schroeder, 2008], στα πλαίσια της σχεδίασης των συστημάτων για το διεθνή διαγωνισμό Blizzard Challenge. Σε αυτές τις προσεγγίσεις τα συστήματα των [Latacz, 2009] και [Schroeder, 2008] δεν σημείωσαν σημαντική βελτίωση στη συνολική ποιότητα, σε αντίθεση με το σύστημα που περιγράφεται στο [Ling, 2007] που αξιολογήθηκε ανάμεσα στα δύο καλύτερα στο τελευταίο διαγωνισμό. Ωστόσο για το τελευταίο σύστημα δεν δίνονται επιμέρους αποτελέσματα ώστε να προκύψει συμπέρασμα σχετικά με το υποσύστημα που συνεισφέρει περισσότερο στη τελική ποιότητα. Στα δύο πρώτα συστήματα, η συνάρτηση κόστους ένωσης σχεδιάζεται από κοινού, με τη βοήθεια ενός επαυξημένου διανύσματος χαρακτηριστικών που αποτελείται από τις

διαφορές στη θεμελιώδη συχνότητα, την ένταση και τους συντελεστές MFCC στο σημείο της ένωσης. Η διαφορά των δύο συστημάτων είναι στο τύπο της ακουστικής μονάδας (διφώνημα και ημι-φώνημα αντίστοιχα). Με χρήση δένδρων απόφασης με πληροφορία περιβάλλοντος και με στατιστική μοντελοποίηση του διανύσματος με μίξη κατανομών Gauss (Gaussian Mixture Models) στα όρια της ένωσης των διφώνων επιδιώκεται η ανίχνευση φυσικών μεταβάσεων η οποία βαθμολογείται με την πιθανοφάνεια από τα εκπαιδευμένα μοντέλα. Τα μοντέλα δημιουργούνται με εκπαίδευση στην υπάρχουσα βάση δεδομένων. Παρά το γεγονός ότι το κόστος εκπαιδεύεται στην υπάρχουσα βάση δεδομένων, μια πιθανή εξήγηση για την απόδοση τους στην ανίχνευση των ακουστικών ασυνεχειών είναι ότι οι διαφορές στους συντελεστές MFCC δεν επαρκούν, ή με άλλα λόγια δεν εμπεριέχουν το σύνολο της πληροφορίας που είναι ικανή να διακρίνει τις, ακουστικά, φασματικές ασυνέχειες. Σε αντίθεση με την αναγνώριση φωνής που είναι αναγκαία εκείνα τα χαρακτηριστικά που δεν είναι επιρρεπή στην μεταβλητότητα της φωνής για τον ίδιο τύπο φωνήματος, στην σύνθεση φωνής είναι αναγκαία εκείνα τα φασματικά χαρακτηριστικά που θα διακρίνουν αυτήν την μεταβλητότητα [Klabbers, 2001].

Περισσότερες πληροφορίες για τη σχεδίαση τόσο για της συνάρτησης κόστους «στόχος» όσο και της συνάρτησης κόστους «ένωσης», σε σύγχρονα (εμπορικά και μη) συστήματα σύνθεσης φωνής από κείμενο είναι διαθέσιμες στο [Fraser, 2007], στο οποίο περιγράφονται οι παράμετροι των συστημάτων που συμμετείχαν στον διαγωνισμό Blizzard Challenge 2007.

Η συνάρτηση κόστους ένωσης του συστήματος που χρησιμοποιήθηκε στα πλαίσια της διατριβής αποτελείται από δύο επιμέρους κόστη, το κόστος προσωδιακών ασυνεχειών και το κόστος φασματικών ασυνεχειών. Για το κόστος φασματικών ασυνεχειών έχει χρησιμοποιηθεί τόσο η απόσταση Kullback-Leibler στο φάσμα ισχύος που προκύπτει από FFT όσο και η Ευκλείδεια απόσταση σε συντελεστές MFCC. Η απόδοση τους είναι σε γενικές γραμμές ισοδύναμη όπως θα δούμε και σε επόμενο κεφάλαιο στο οποίο προτείνεται μια νέα μέθοδος για την αποδοτικότερη εκτίμηση των φασματικών ασυνεχειών. Να σημειώσουμε ότι στο σύστημα δεν ακολουθείται η προσέγγιση μηδενισμού του φασματικού κόστους στη περίπτωση γειτονικών διφώνων, με σκοπό την αποφυγή περιπτώσεων με υψηλά τοπικά κόστη που ενδεχομένως να προκύψουν κατά τη σύνθεση. Αυτό είναι ένα μειονέκτημα του αλγόριθμου επιλογής ακουστικών μονάδων καθώς το μονοπάτι ολικού ελάχιστου κόστους δεν αποκλείει την ύπαρξη τοπικών κοστών με υψηλές τιμές.

2.3.3 Προσδιορισμός παραγόντων στάθμισης στις συναρτήσεις κόστους

Ο προσδιορισμός των παραγόντων στάθμισης στις συναρτήσεις κόστους αποτελεί από τις κρισιμότερες, και ταυτόχρονα από τις δυσκολότερες παραμέτρους στη σχεδίαση του αλγόριθμου επιλογής ακουστικών μονάδων. Η δυσκολία στον προσδιορισμό τους μπορεί να αποδοθεί τόσο στην πολυπλοκότητα της φυσικής ομιλίας (π.χ., δεν υπάρχει μοναδικά αποδεκτή εκφορά μιας πρότασης) όσο και σε επιμέρους λόγους όπως είναι η σύγκριση ετερογενών μεγεθών που εμπλέκονται στις συναρτήσεις (π.χ., σχ. 2.2) αλλά και οι ίδιες οι δυνατότητες που προσφέρει η βάση δεδομένων προηχογραφημένης φυσικής ομιλίας.

Παρά την ποικιλία προσεγγίσεων για τις συναρτήσεις κόστους, ο προσδιορισμός των παραγόντων στάθμισης συνήθως πραγματοποιείται είτε με αυτόματο είτε με μη αυτόματο τρόπο σύμφωνα με τρείς προσεγγίσεις: α) χειρωνακτικά [Founda, 2001; Coorman, 2000; Breen 1998; Clark, 2007], β) με σύγκριση αποτελεσμάτων σύνθεσης με φυσικές ηχογραφήσεις που θεωρούνται ως επιθυμητό αποτέλεσμα [Hunt, 1996; Kim, 2004; Alias, 2003], γ) με δεδομένα που προκύπτουν από ακουστικά πειράματα [Lee, 2003; Lee, 2001; Wouters, 1998; Toda, 2004].

Η μη αυτοματοποιημένη (χειρονακτική) ανάθεση τιμών, προκύπτει συνήθως από συμπεράσματα κάποιων μελετών ή/και παρατηρήσεων ως προς την επίδραση των παραμέτρων που χρησιμοποιούνται στην σύνθεση. Με άλλα λόγια η σημαντικότητα των παραμέτρων κρίνεται εμπειρικά και διαισθητικά από τον ανθρώπινο παράγοντα γεγονός που αν και εκ πρώτης όψεως φαίνεται περίεργο ωστόσο δεν είναι λίγες οι περιπτώσεις που οδηγεί σε υψηλής ποιότητας σύνθεση, ειδικά σε περιπτώσεις που ο αριθμός των παραγόντων στάθμισης είναι χαμηλός. Το γεγονός αυτό ενισχύεται και από τα αποτελέσματα που αναφέρονται σε συστήματα που κατά καιρούς συμμετέχουν στον διεθνή διαγωνισμό Blizzard Challenge.

Ο αυτόματος προσδιορισμός των παραγόντων στάθμισης με σύγκριση αποτελεσμάτων σύνθεσης με φυσικές ηχογραφήσεις συνήθως προκύπτει από την ελαχιστοποίηση μιας αντικειμενικής συνάρτησης (μετρικής) μεταξύ συνθετικού σήματος και σήματος φυσικής ομιλίας. Σε αυτή τη προσέγγιση, ένα μέρος των προτάσεων της βάσης δεδομένων χρησιμοποιείται για εκπαίδευση. Το σύστημα σύνθεσης παράγει αυτές τις προτάσεις, πραγματοποιείται σύγκριση με τις φυσικές και οι παράγοντες στάθμισης προσδιορίζονται (με επαναληπτικό τρόπο) και με γραμμική παλινδρόμηση (linear regression) με τη βοήθεια της ελαχιστοποίησης της αντικειμενικής συνάρτησης (συνήθως Ευκλείδεια απόσταση μεταξύ cepstral συντελεστών). Το εγγενές πρόβλημα της προσέγγισης αυτής είναι ο ορισμός της κατάλληλης αντικειμενικής συνάρτησης [Hunt, 1996].

Η προσέγγιση προσδιορισμού των παραγόντων στάθμισης με δεδομένα που προκύπτουν από ακουστικά πειράματα, ακολουθεί την ίδια λογική με την προηγούμενη, χωρίς όμως να ορίζει αντικειμενική συνάρτηση ως μετρική ομοιότητας. Το σκοπό της αντικειμενικής συνάρτησης εξυπηρετούν δεδομένα (βαθμολογίες) από ακουστικά πειράματα. Τα ακουστικά πειράματα για αυτές τις προσεγγίσεις αφορούν τον προσδιορισμό των παραγόντων στάθμισης σε επίπεδο πρότασης [Lee, 2003], σε επίπεδο ακουστικής μονάδας [Wouters, 1998] και σε επίπεδο επιμέρους συναρτήσεων κόστους [Toda, 2004]. Το σημαντικότερο μειονέκτημα σε αυτές την προσέγγιση είναι ότι στηρίζεται σε αποτελέσματα που προέρχονται από ακουστικά πειράματα τα οποία, όπως αναφέρθηκε, αποτελούν μια χρονοβόρα και πολυδάπανη διαδικασία τόσο σε πόρους όσο και σε ανθρώπινο δυναμικό.

Από τα προηγούμενα προκύπτει τόσο η πολυπλοκότητα όσο και η δυσκολία προσδιορισμού των παραγόντων στάθμισης, με αποτέλεσμα να κρίνεται αναγκαία κάποια σχεδιαστική προσέγγιση στις συναρτήσεις κόστους που είτε να καταργεί τον προσδιορισμό τους είτε να ελαχιστοποιεί τον απαιτούμενο αριθμό τους.

Στο σύστημα που χρησιμοποιήθηκε στα πλαίσια της διατριβής ο προσδιορισμός των παραγόντων στάθμισης πραγματοποιήθηκε με μη αυτόματο τρόπο και σύμφωνα με παρατηρήσεις επί του αποτελέσματος της σύνθεσης καθώς και

ανευρετικές διαδικασίες[1]. Στον πίνακα που ακολουθεί συνοψίζονται οι χρησιμοποιούμενοι παράγοντες στάθμισης και η ερμηνεία τους.

ΠΙΝΑΚΑΣ 2.2: Παράγοντες στάθμισης στις συναρτήσεις κόστους

ΠΑΡΑΓΟΝΤΑΣ ΣΤΑΘΜΙΣΗΣ	ΣΥΝΑΡΤΗΣΗ ΚΟΣΤΟΥΣ
W^t	Κόστος «στόχος» (επιθυμητής ακολουθίας)
W^c	Κόστος «ένωσης»
w_1^c	Επιμέρους κόστος ασυνέχειας της θεμελιώδους συχνότητας στην ένωση
w_2^c	Επιμέρους κόστος φασματικής ασυνέχειας στην ένωση
w_1^t	Επιμέρους κόστος επιθυμητής προσωδίας
w_2^t	Επιμέρους κόστος πληροφορίας περιβάλλοντος (context)

Στη συνέχεια, παρουσιάζονται ορισμένες περιπτώσεις (παραδείγματα) που αφορούν την ανάθεση τιμών στους παράγοντες στάθμισης. Αυτές αφορούν το εξεταζόμενο διφώνημα (u_i) σε σχέση με την ένωσή του με το προηγούμενο (u_{i-1}). Για παράδειγμα, όταν το εξεταζόμενο δίφωνο,

- έχει δύο έμφωνα τυρβώδη ή υγρά τότε το w_2^t και το w_2^c θεωρούνται πιο σημαντικά ενώ τα W^t και W^c θεωρούνται ισάξια
- έχει δύο εκρηκτικά τότε το w_2^c θεωρείται πιο σημαντικά όπως και το W^c. Τα βάρη γίνονται μηδέν όταν τα φωνήματα στα οποία αναφέρονται είναι άηχα εκρηκτικά. Το κόστος για φασματικές ασυνέχειες μηδενίζεται, αν το πρώτο φώνημα είναι άηχο εκρηκτικό
- είναι της μορφής /cr/ ή /rc/ αποτελείται δηλαδή από ένα σύμφωνο και το φώνημα /r/, τότε το W^t και το w_2^t θεωρούνται πιο σημαντικά όπως και το w_2^c.
- είναι της μορφής /vr/ αποτελείται δηλαδή από ένα φωνήεν και το φώνημα /r/ τότε W^t και W^c θεωρούνται εξίσου σημαντικά και τα επιμέρους κόστη
- είναι της μορφής /rv/ τότε το W^t με το w_2^t θεωρούνται πιο σημαντικά
- αποτελείται από δύο φωνήεντα τότε το w_1^t είναι πιο σημαντικό, τα w_1^c και w_2^c είναι ισάξια, ενώ το W^c είναι πιο σημαντικό
- αποτελείται από δύο σύμφωνα τότε το w_2^t είναι πιο σημαντικό, από τα w_1^c και w_2^c το δεύτερο είναι λίγο σημαντικότερο, ενώ το W^c είναι πιο σημαντικό

Να σημειώσουμε ότι σε άηχα εκρηκτικά το κόστος ένωσης δεν λαμβάνεται υπόψη. Το κόστος ασυνέχειας της θεμελιώδους συχνότητας για σύμφωνα ορίζεται εφόσον προσδιορίζεται τιμή της θεμελιώδους συχνότητας σε αυτά μέσω γραμμικής παρεμβολής (interpolation).

[1] Φουντά Μ., Χαλαμανδάρης Αιμ., *"Σύνθεση φωνής στο πεδίο του χρόνου: Δομή της Βάσης Σημάτων και Επιλογή του κατάλληλου Σήματος με Τεχνικές Δυναμικού Προγραμματισμού"*, Διπλωματική εργασία, Τμήμα ΗΜΜΥ, ΕΜΠ, 2000.

2.4 ΠΡΟΣΑΡΜΟΓΗ ΤΕΧΝΟΛΟΓΙΑΣ ΣΥΝΘΕΣΗΣ ΦΩΝΗΣ ΣΕ ΕΝΣΩΜΑΤΩΜΕΝΑ ΣΥΣΤΗΜΑΤΑ

Η επιτυχία της τεχνολογίας σύνθεσης φωνής από κείμενο με επιλογή και συρραφή ακουστικών μονάδων έφερε ως αποτέλεσμα την ευρεία αποδοχή της σε πλήθος εφαρμογών που αγγίζουν διάφορους τομείς της ανθρώπινης δραστηριότητας. Ωστόσο, όπως είδαμε, η τεχνολογία αυτή προβάλλει αυξημένες υπολογιστικές αλλά και αποθηκευτικές απαιτήσεις, που καθιστούν απαγορευτική την εφαρμογή της σε φορητά ή ενσωματωμένα συστήματα. Όμως, η διαρκής ανάγκη για επαυξημένες πολυμεσικές εφαρμογές και η ευρεία χρήση συσκευών τύπου κινητών τηλεφώνων ή PDA (personal digital assistants) έδρασε καταλυτικά για την υιοθέτηση υψηλής ποιότητας τεχνολογίας σύνθεσης φωνής από κείμενο σε περιβάλλοντα περιορισμένων υπολογιστικών πόρων. Εφαρμογές όπως, βοηθήματα για άτομα με ειδικές ανάγκες [Peters, 2004], μετάφραση από φωνή-σε-φωνή (speech to speech translation) σε κινητές συσκευές [Schultz, 2006], ρομποτική [Tomko, 2005] και γενικότερα επικοινωνία ανθρώπου μηχανής μέσω φωνής [Moore, 2007; Mohasi, 2006] αποκτούν προστιθέμενη αξία και γίνονται ελκυστικότερες με συνθετική φωνή υψηλής ποιότητας. Ως επακόλουθο, η επιδίωξη αυτή προσέλκυσε έντονο ερευνητικό αλλά και αναπτυξιακό ενδιαφέρον στην δυνατότητα προσαρμογής της τεχνολογίας σύνθεσης με επιλογή και ένωση ακουστικών μονάδων σε περιβάλλοντα περιορισμένων υπολογιστικών δυνατοτήτων όπως είναι τα ενσωματωμένα συστήματα και κυρίως τα φορητά τηλέφωνα (κινητά τηλέφωνα) και οι φορητές συσκευές PDA. Προσεγγίσεις με προηγούμενης γενιάς τεχνολογίες (όπως π.χ., ένωση διφωνημάτων, σύνθεση με formants) χαρακτηρίζονται από χαμηλές απαιτήσεις σε υπολογιστικό φορτίο αλλά επιτυγχάνουν όχι ιδιαίτερα υψηλή τελική ποιότητα [Hoffman, 2003; Sheikhzadeh, 2002]. Παράλληλα, σε αναλογία με τα πλήρη συστήματα σύνθεσης, έχουν προταθεί και προσεγγίσεις που αφορούν ειδικού σκοπού συστήματα σύνθεσης (π.χ., καιρός, αθλητικά) [Ishikawa, 1999].

Η μεταφορά και προσαρμογή της τεχνολογίας σύνθεσης με επιλογή και ένωση ακουστικών μονάδων σε περιβάλλοντα μειωμένων υπολογιστικών πόρων καθίσταται αποδεκτή εφόσον διατηρείται η υψηλή τελική ποιότητα. Για το σκοπό αυτό, πρόσφατες ερευνητικές εργασίες στο τομέα αυτό επικεντρώνονται κυρίως στην αποδοτική αποκλιμάκωση της τεχνολογίας και ειδικότερα στη βελτιστοποίηση και προσαρμογή σε ζητήματα όπως, η μείωση υπάρχουσας βάσης δεδομένων φυσικής ομιλίας, η συμπίεση αυτής της βάσης, η εναλλακτική αναπαράσταση, παραμετροποίηση και παραγωγή του σήματος φωνής και η ελαχιστοποίηση των υπολογιστικών απαιτήσεων κατά τη σύνθεση [Aylett, 2006; Schnell, 2002; Nukaga, 2006; Chazan, 2002; Kedia, 2007; Rutten, 2002; Tsiakoulis, 2008; Van der Vrecken, 1997; Kumar, 2004]. Συνοψίζοντας, τα σημαντικότερα ζητήματα στη τεχνολογία σύνθεσης φωνής από κείμενο για ενσωματωμένα υπολογιστικά περιβάλλοντα, σχετίζονται με την επίτευξη ισορροπίας μεταξύ των υπολογιστικών αναγκών, της αποδοτικής αποκλιμάκωσης, και της τελικής ποιότητας. Στο κεφάλαιο 3 περιγράφονται τεχνικές που αντιμετωπίζουν αυτά τα ζητήματα, με σκοπό την υλοποίηση ενός γενικού σκοπού συστήματος σύνθεσης φωνής από κείμενο για το περιβάλλον της συσκευής κινητού τηλεφώνου.

2.5 ΜΕΘΟΔΟΙ ΠΑΡΑΓΩΓΗΣ ΣΗΜΑΤΟΣ ΦΩΝΗΣ

Στο υποσύστημα της Ψηφιακής Επεξεργασίας Σήματος, το δεύτερο κύριο δομικό στοιχείο είναι η βαθμίδα παραγωγής του συνθετικού σήματος φωνής. Η βαθμίδα αυτή, επεξεργάζεται τις πληροφορίες από τον αλγόριθμο επιλογής ακουστικών μονάδων και, σε συνδυασμό με άλλα δεδομένα από τη βάση δεδομένων, είναι υπεύθυνη για την παραγωγή ομιλίας. Παρακάτω, γίνεται μια συνοπτική παρουσίαση των σύγχρονων και πιο συχνά εφαρμοζόμενων αλγόριθμων παραγωγής φωνής με τα κυριότερα χαρακτηριστικά τους ενώ ακολουθεί μια πιο εκτενής ποιοτική περιγραφή της μεθόδου TD-PSOLA η οποία εφαρμόζεται στο σύστημα που χρησιμοποιήθηκε στα πλαίσια της διατριβής. Επιγραμματικά, οι κυριότερες τεχνικές είναι,

- **PSOLA (Pitch Synchronous Overlap Add)** [Moulines, 1990]

Η μέθοδος αυτή εφαρμόζεται είτε στο πεδίο του χρόνου (TD) είτε στο πεδίο της συχνότητας (FD). Συνήθως, λόγω της απλότητας της, εφαρμόζεται η πρώτη τεχνική. Στηρίζεται στη μη παραμετρική αναπαράσταση του σήματος φωνής και τα στάδια επεξεργασίας που ακολουθεί είναι, 1) Ανάλυση του αρχικού σήματος, 2) Επιθυμητές προσωδιακές μετατροπές και 3) Σύνθεση τελικού σήματος. Ιδιαίτερη σημασία στη μέθοδο αυτή έχει η συνεπής (consistent), αυτόματη και εύρωστη εκτίμηση και επισημείωση των χρονικών στιγμών που σηματοδοτούν την θεμελιώδη περίοδο της κυματομορφής του σήματος φωνής, σε όλη τη βάση δεδομένων. Τα σημεία αυτά ονομάζονται pitch-marks και συντελούν σημαντικά στην τελική ποιότητα του τελικού συνθετικού σήματος [Chalamandaris, 2009b].

- **RELP (Residual Excited Linear Prediction)** [Dutoit, 1997]

Στην τεχνική RELP η βάση δεδομένων αναλύεται και αναπαριστάται με το μοντέλο της γραμμικής πρόβλεψης (LPC). Η παραγωγή του σήματος φωνής πραγματοποιείται όπως και στην TD-PSOLA με τη διαφορά ότι η επεξεργασία δεν γίνεται απευθείας στο σήμα φωνής αλλά στο σήμα που αποτελεί το σφάλμα της γραμμικής πρόβλεψης (residual). Η τεχνική αυτή συναντάται και με τον όρο **LP-PSOLA (Linear Prediction-PSOLA)** και το κύριο πλεονέκτημα που προσφέρει είναι στην εξοικονόμηση αποθηκευτικών πόρων αφού πλέον το αρχικό σήμα αποθηκεύεται σε μορφή συντελεστών ενώ το σήμα σφάλματος χαρακτηρίζεται από χαμηλότερο εύρος τιμών.

- **HNM (Harmonic plus Noise Model)** [Stylianou, 2001b]

Η τεχνική Harmonic plus Noise είναι μεταγενέστερη της TD-PSOLA και βασίζεται στην ανάλυση της φωνής σε αρμονικό και θορυβώδες μέρος. Ειδικότερα, η μέθοδος θεωρεί πως κάθε σήμα φωνής αποτελείται από ένα χρονικά μεταβαλλόμενο αρμονικό κομμάτι στο οποίο υπερτίθεται διαμορφωμένος θόρυβος. Αναλυτικότερα, το αρμονικό κομμάτι αποτελεί την περιοδική συνιστώσα του σήματος φωνής, ενώ ο θόρυβος αντιστοιχεί στο μη περιοδικό σήμα. Η τεχνική χωρίζει το φάσμα του σήματος σε δύο ζώνες για τις οποίες θεωρεί ότι η μεν πρώτη περιέχει τις αρμονικές συνιστώσες, ενώ η δεύτερη το διαμορφωμένο θορυβώδες τμήμα. Ο διαχωρισμός των ζωνών είναι χρονικά μεταβαλλόμενος, και καθορίζεται ως συνάρτηση του πλαισίου φωνής που επεξεργάζεται. Για το αρμονικό κομμάτι, αφού γίνει εκτίμηση της θεμελιώδους συχνότητας και της συχνότητας διαχωρισμού των ζωνών,

εκτιμώνται και αποθηκεύονται και οι υπόλοιπες αρμονικές της θεμελιώδους συχνότητας σε μορφή πλάτους και φάσης. Αυτές θα χρησιμοποιηθούν για την σύνθεση του τελικού σήματος. Για το θορυβώδες μέρος, χρησιμοποιείται λευκός θόρυβος που οδηγεί κάποιο φίλτρο μόνο με πόλους (AR), του οποίου οι συντελεστές έχουν προκύψει από ανάλυση του αυθεντικού σήματος. Τέλος, αναφορικά με το στάδιο σύνθεσης του τελικού σήματος, σε γενικές γραμμές ο αλγόριθμος εντοπίζει με χρήση κάποιων κριτηρίων (π.χ. μέθοδος ροπών) τα σωστά σημεία για την ένωση των κυματομορφών, προσπαθώντας να ελαχιστοποιήσει τυχόν φασματικές ασυνέχειες καθώς και ασυνέχειες φάσης. Κάθε πλαίσιο φωνής συντίθεται βάση των παραμέτρων που είχαν προκύψει από την ανάλυση, τόσο για το αρμονικό όσο και για το διαμορφωμένο θορυβώδες κομμάτι. Η τελική σύνθεση γίνεται με σύγχρονη, ως προς τη θεμελιώδη συχνότητα, πρόσθεση με επικάλυψη των επιμέρους πλαισίων. Τα πλεονεκτήματα της μεθόδου Harmonic plus Noise είναι ότι μέσω της παραμετρικής αναπαράστασης των συνιστωσών, επιτρέπει καλύτερη και πιο εύρωστη επεξεργασία του σήματος. Αυτό έχει σαν αποτέλεσμα τη μείωση ακουστικών παραμορφώσεων στο τελικό σήμα φωνής. Ουσιαστικά ξεπερνά το εμπόδιο της συνεπούς εκτίμησης των pitch-marks σε όλη τη βάση δεδομένων. Ωστόσο είναι πιο πολύπλοκη και υπολογιστικά πιο απαιτητική μέθοδος.

- **MBROLA (Multi-Band Re-synthesis Pitch Synchronous OverLap Add)** [Dutoit, 1993; Dutoit, 1997]

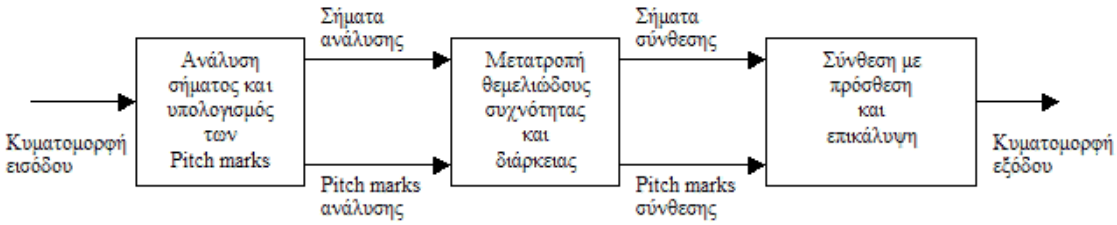
Η τεχνική MBROLA (Multi-Band Re-synthesis Pitch Synchronous OverLap Add) διαφοροποιείται στο γεγονός ότι προσπαθεί να εξασφαλίσει τη μείωση φασματικών ασυνεχειών και ασυνεχειών φάσης, δημιουργώντας όμοια πλαίσια φωνής (κατά το στάδιο της ανάλυσης) με την έννοια να είναι απόλυτα συγχρονισμένα μεταξύ τους. Έτσι, εφαρμόζει ένα στάδιο προεπεξεργασίας για εξασφάλιση σταθερής θεμελιώδους συχνότητας στα 100Hz ενώ ελέγχει περισσότερες παραμέτρους του σήματος φωνής εφαρμόζοντας μια τεχνική που ονομάζεται MBE (multiband excitation) και η οποία μοιάζει με την ανάλυση του Harmonic plus Noise μοντέλου που είδαμε προηγουμένως. Σε γενικές γραμμές, η αρχική βάση δεδομένων (σταθερού pitch) αναλύεται κατά MBE και χωρίζεται σε έμφωνα και άφωνα μέρη. Η τελική σύνθεση προκύπτει μέσω της MBE. Η τεχνική δεν έχει επικρατήσει στα σύγχρονα συστήματα σύνθεσης φωνής από κείμενο.

2.5.1 Η μέθοδος TD-PSOLA (Time Domain - Pitch Synchronous Overlap Add)

Η πιο διαδεδομένη μέθοδος παραγωγής σήματος φωνής είναι η TD-PSOLA (Time Domain – Pitch Synchronous Overlap Add) η οποία στηρίζεται στη πρόσθεση με επικάλυψη στο πεδίο του χρόνου, συγχρονισμένων στην θεμελιώδη περίοδο, κυματομορφών σήματος φωνής. Επιπλέον, η μέθοδος επιτρέπει τον έλεγχο και τη μετατροπή, τόσο της θεμελιώδους συχνότητας (pitch) όσο και της διάρκειας των κυματομορφών που θα ενωθούν, με αποτέλεσμα η ένωση των διφωνημάτων να γίνεται σύμφωνα με τα επιθυμητά προσωδιακά χαρακτηριστικά που καθορίζονται από προηγούμενα στάδια της διαδικασίας σύνθεσης. Σε γενικές γραμμές η μέθοδος χωρίζεται σε τρία βήματα:

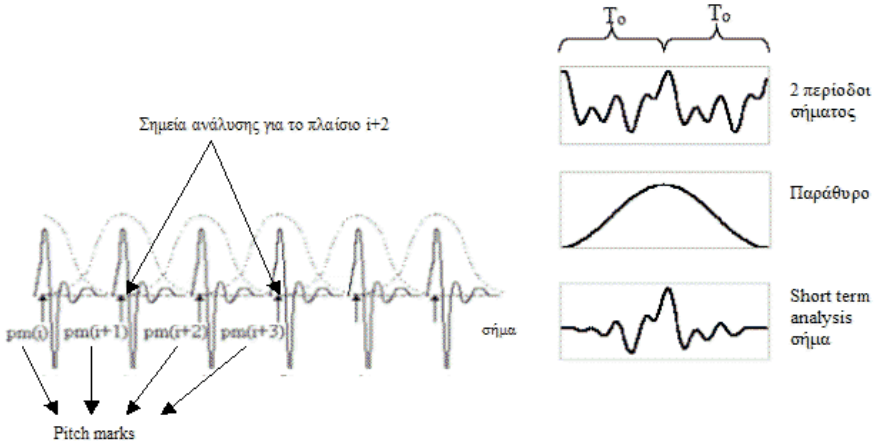
1) Το στάδιο ανάλυσης του αρχικού σήματος, στο οποίο παράγεται μια ενδιάμεση μη παραμετρική αναπαράσταση του σήματος.
2) Το στάδιο τροποποιήσεων, στο οποίο γίνονται υπολογισμοί και προσωδιακές μετατροπές στο αρχικό σήμα με σκοπό τη παραγωγή του ζητούμενου συνθετικού σήματος.
3) Το στάδιο σύνθεσης, στο οποίο πραγματοποιείται η σύνδεση των κυματομορφών σύμφωνα με τους κανόνες του δεύτερου βήματος.

Στο σχήμα 2.5 δίνεται σε μορφή δομικού διαγράμματος, η μεθοδολογία της μεθόδου TD-PSOLA.



ΣΧΗΜΑ 2.5: Δομικό διάγραμμα της μεθόδου TD-PSOLA.

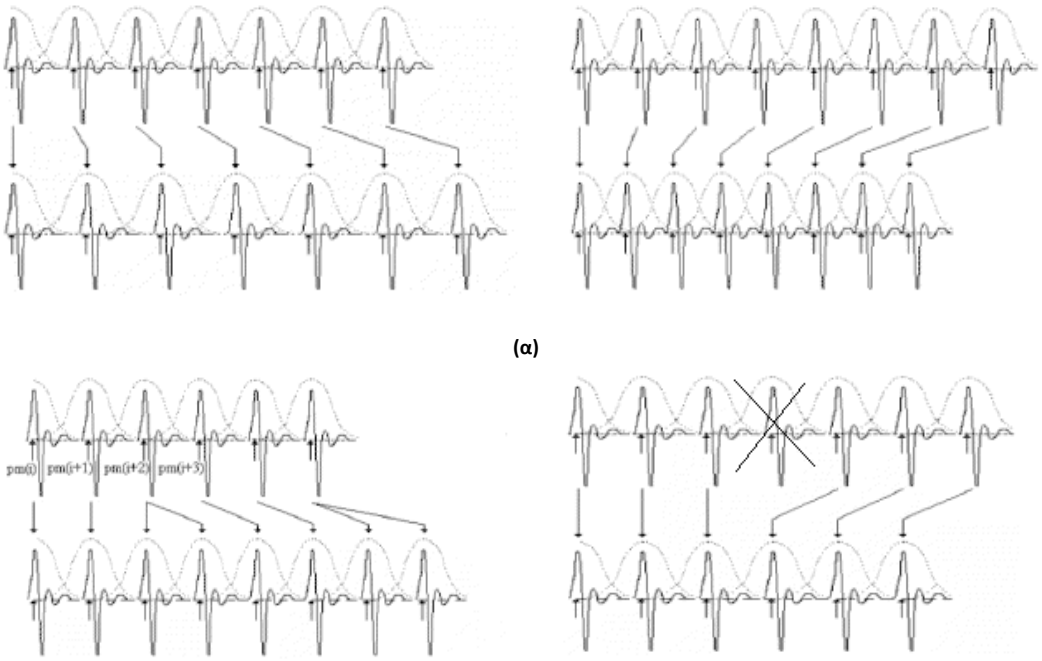
Το πρώτο βήμα στη διαδικασία ανάλυσης είναι ο υπολογισμός των χρονικών στιγμών σήμανσης της θεμελιώδους συχνότητας του σήματος εισόδου. Ουσιαστικά αναφέρονται στις χρονικές στιγμές αρχής και τέλους της θεμελιώδους περιόδου σε όλη τη χρονική εξέλιξη του σήματος. Τα σημεία αυτά ονομάζονται γενικά pitch-marks ενώ στο συγκεκριμένο στάδιο αναφέρονται ως pitch-marks ανάλυσης. Με τη βοήθεια των pitch-marks, το αρχικό σήμα (φωνή) μπορεί να διαχωριστεί σε πλήθος επικαλυπτόμενων σημάτων μικρού μήκους, που ονομάζονται st-signals (short–term analysis signals). Αυτά προκύπτουν με πλαισίωση και παραθύρωση του αρχικού σήματος (σχ. 2.6).



ΣΧΗΜΑ 2. 6: Διαδικασία διαχωρισμού σε short-term σήματα ανάλυσης.

Στο στάδιο τροποποίησης πραγματοποιείται η μετατροπή του συνόλου των st-signals σε ένα νέο τροποποιημένο σύνολο που ονομάζεται st-signals σύνθεσης, τα οποία είναι συγχρονισμένα σε ένα νέο σύνολο απο pitch marks. Το νέο σύνολο ονομάζεται pitch marks σύνθεσης. Η μετατροπή βασίζεται σε δύο λειτουργίες,
α) διαφοροποίηση και επιλογή στον αριθμό των st-signal ανάλυσης που θα χρησιμοποιηθούν και β) διαφοροποίηση των καθυστερήσεων ανάμεσα στα st-signal ανάλυσης που θα χρησιμοποιηθούν. Το πλήθος των pitch marks σύνθεσης εξαρτάται απο τους παράγοντες αλλαγής κλίμακας ως προς τη θεμελιώδη συχνότητα (pitch) και ως προς το χρόνο (διάρκεια). Η λειτουργία του αλγόριθμου είναι η σωστή αντιστοίχηση των pitch mark ανάλυσης σε pitch mark σύνθεσης και η σωστή επιλογή των st-signal ανάλυσης που θα αποτελέσουν τα st-signal σύνθεσης.

Στο τελευταίο στάδιο της διαδικασίας πραγματοποιείται η κατασκευή του συνθετικού σήματος με την σύγχρονη παράθεση πρόσθεση των επικαλυπτόμενων st-signal που επιλέχθηκαν για τη σύνθεση. Στο σχήμα 2.7, φαίνεται η διαδικασία μείωσης και αύξησης της θεμελιώδους συχνότητας και της διάρκειας με χρήση της τεχνικής TD-PSOLA.



(α)

(β)

ΣΧΗΜΑ 2.7: Τροποποίηση προσωδιακών χαρακτηριστικών με τη μέθοδο TD-PSOLA: α) Μείωση (αριστερά) και αύξηση (δεξιά) του pitch με μετατόπιση των σημάτων ανάλυσης, β) Αύξηση με επανάληψη (αριστερά) και μείωση με παράληψη (δεξιά) της διάρκειας

Το κυριότερα πλεονεκτήματα της μεθόδου TD-PSOLA είναι η απλότητα και η αποτελεσματικότητα της, ενώ στα μειονεκτήματα της περιλαμβάνονται αφενός κάποιες ακουστικές παραμορφώσεις που δημιουργούνται σε μεγάλες αλλαγές είτε του pitch, είτε της διάρκειας, είτε και των δύο και αφετέρου στην απαίτηση που προβάλλει για συνεπή (consistent) εκτίμηση και επισημείωση των pitch-marks, γεγονός που συντελεί σημαντικά στην τελική ποιότητα του τελικού συνθετικού σήματος [Chalamandaris, 2009b].

ΚΕΦΑΛΑΙΟ
-3-
ΣΥΝΘΕΣΗ ΦΩΝΗΣ ΣΕ ΠΕΡΙΒΑΛΛΟΝ ΕΝΣΩΜΑΤΩΜΕΝΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 3 - ΣΥΝΘΕΣΗ ΦΩΝΗΣ ΜΕ ΕΠΙΛΟΓΗ ΚΑΙ ΕΝΩΣΗ ΑΚΟΥΣΤΙΚΩΝ ΜΟΝΑΔΩΝ ΣΕ ΠΕΡΙΒΑΛΛΟΝ ΕΝΣΩΜΑΤΩΜΕΝΩΝ ΣΥΣΤΗΜΑΤΩΝ

Στο κεφάλαιο αυτό εξετάζονται οι ερευνητικές προσπάθειες και οι μεθοδολογικές προσεγγίσεις που υιοθετήθηκαν για την αποδοτική αποκλιμάκωση, την μεταφορά και την ολοκλήρωση τεχνολογίας σύνθεσης φωνής υψηλής ποιότητας σε περιβάλλον συσκευής κινητού τηλεφώνου, με στόχο την μέγιστη δυνατή αξιοποίηση των οφελών της στο περιβάλλον αυτό. Οι προσεγγίσεις σχετίζονται με την σχεδίαση και την υλοποίηση γενικού σκοπού συστήματος σύνθεσης φωνής με επιλογή και ένωση ακουστικών μονάδων (general domain unit selection speech synthesis system) και συντελούν τόσο στην αποδοτική μεταφορά της τεχνολογίας όσο και στην μείωση των υπολογιστικών απαιτήσεων, διασφαλίζοντας ταυτόχρονα υψηλή τελική ποιότητα. Τέλος, παρουσιάζονται τα αποτελέσματα της πειραματικής αξιολόγησης που επικυρώνουν τις παραπάνω προσεγγίσεις.

3.1 ΕΙΣΑΓΩΓΗ

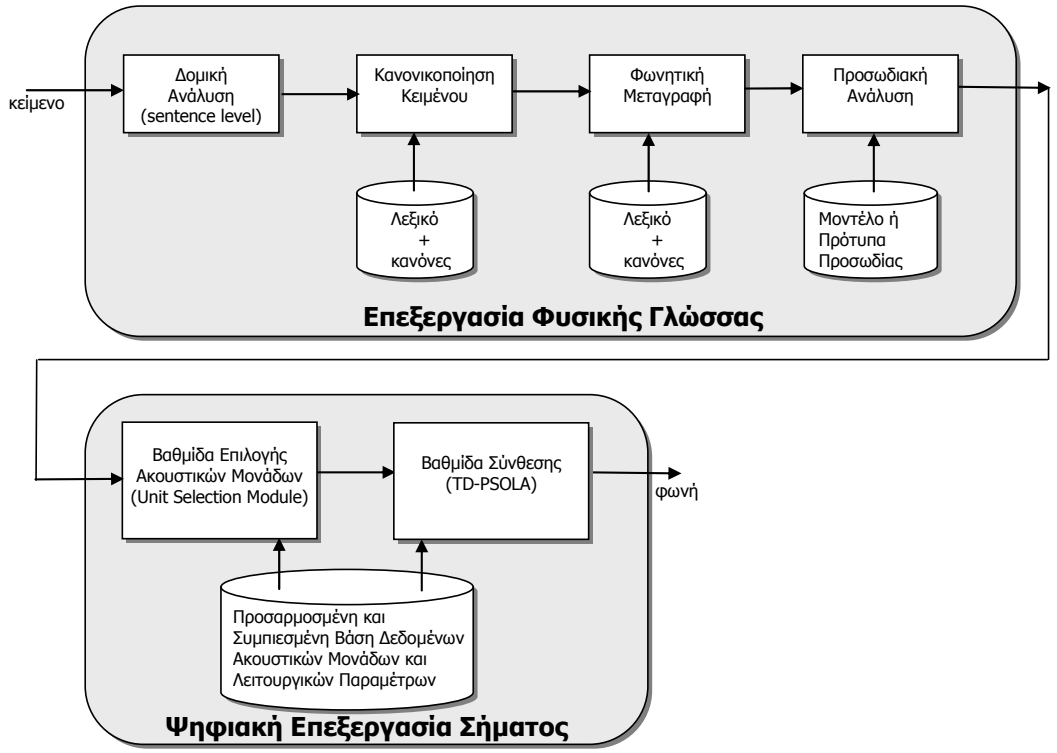
Όπως είδαμε στην βιβλιογραφική επισκόπηση στο κεφάλαιο 2, η αποδοτική μεταφορά και προσαρμογή της τεχνολογίας σύνθεσης φωνής από κείμενο σε υπολογιστικό περιβάλλον ενσωματωμένων συστημάτων, και ιδιαίτερα στο περιβάλλον συσκευής κινητού τηλεφώνου, αποτελεί έντονο σημείο ερευνητικού και αναπτυξιακού ενδιαφέροντος σε διεθνές επίπεδο. Ιδιαίτερο ενδιαφέρον παρουσιάζει η προσαρμογή της τεχνολογίας σύνθεσης με επιλογή και ένωση ακουστικών μονάδων, η οποία αποτελεί την επικρατούσα τεχνική για την επίτευξη υψηλής ποιότητας συνθετικής ομιλίας. Πρόσφατες ερευνητικές προσπάθειες στο χώρο αυτό, αφορούν την βελτιστοποίηση όλου του εύρους των διαδικασιών που εμπλέκονται στην σχεδίαση και υλοποίηση ενός γενικού σκοπού συστήματος σύνθεσης φωνής για ενσωματωμένα συστήματα. Ωστόσο, επικεντρώνονται κυρίως στις διαδικασίες ελάττωσης της απαιτούμενης βάσης προηχογραφημένης ομιλίας (*speech database reduction*), στην αναπαράσταση και παραμετροποίηση του σήματος φωνής και στην ελαχιστοποίηση του υπολογιστικού φορτίου κατά την σύνθεση, με ταυτόχρονη διασφάλιση υψηλής ποιότητας συνθετικής φωνής. Παράλληλα, σημαντικό ζητούμενο αποτελεί η δυνατότητα λειτουργίας σε **πραγματικό χρόνο** (*real time*) και με **μικρό χρόνο απόκρισης** (*response time*) [Schnell, 2002; Kim, 2006; Nukaga, 2006; Chazan, 2002; Rutten, 2002; Tsiakoulis, 2008].

Η αρχιτεκτονική του συστήματος σύνθεσης φωνής για το υπολογιστικό περιβάλλον της συσκευής κινητού τηλεφώνου, φαίνεται στο σχήμα 3.1. Σε γενικές γραμμές δεν διαφέρει από ένα σύστημα που προορίζεται για χρήση σε προσωπικούς Η/Υ (*desktop based TTS*) ή σε εξυπηρετητές (*server based TTS*), ωστόσο οι βαθμίδες από τις οποίες αποτελείται είναι ειδικά σχεδιασμένες και προσαρμοσμένες για λειτουργία σε περιβάλλοντα μειωμένων υπολογιστικών και αποθηκευτικών πόρων.

Αναφορικά με το υποσύστημα γλωσσικής επεξεργασίας κειμένου (NLP), η επεξεργασία πραγματοποιείται σε επίπεδο πρότασης (*sentence level parsing and*

structural analysis) και κατόπιν ακολουθεί η κανονικοποίηση κειμένου, η φωνητική μεταγραφή και η προσωδιακή ανάλυση.

Πιο αναλυτικά, το εισερχόμενο κείμενο υπόκειται σε επιφανειακή ή πλήρη συντακτική ανάλυση, με σκοπό τον καθορισμό των ορίων των προτάσεων. Το στάδιο αυτό έχει ιδιαίτερη σημασία καθώς όλα τα επόμενα στάδια της μηχανής σύνθεσης φωνής, πραγματοποιούν επεξεργασία σε επίπεδο πρότασης οπότε η έξοδος της βαθμίδας παρέχει το κείμενο εισόδου στις επόμενες βαθμίδες ανά πρόταση. Ο γραπτός λόγος περιλαμβάνει σειρά γραφικών οντοτήτων όπως αριθμοί, ημερομηνίες, ακρωνύμια και χαρακτήρες με ειδική σημασία, οι οποίες πρέπει να μετατραπούν σε προφορική μορφή. Επιπλέον, στο περιβάλλον της συσκευής του κινητού τηλεφώνου παρουσιάζονται ιδιαιτερότητες όπως, η διαχείριση του τονισμού, η διαχείριση εναλλακτικών τρόπων γραφής (*Greeklish*) και η σωστή ανάγνωση του μενού επιλογών της συσκευής. Η έξοδος της βαθμίδας κανονικοποίησης κειμένου τροφοδοτεί τις επόμενες βαθμίδες με τις προτάσεις που προκύπτουν έπειτα από την σωστή ανάπτυξή τους.



ΣΧΗΜΑ 3.1: Αρχιτεκτονική συστήματος Σύνθεσης Φωνής από Κείμενο με επιλογή και ένωση ακουστικών μονάδων από βάση δεδομένων προηχογραφημένης φυσικής ομιλίας (Unit Selection concatenative speech synthesis) για το υπολογιστικό περιβάλλον της συσκευής του κινητού τηλεφώνου.

Το στάδιο της φωνητικής μεταγραφής (*letter to sound module*), μετατρέπει το γραπτό λόγο σε ένα ενδιάμεσο συμβολικό επίπεδο φωνητική γραφής και αποτελεί βασικό συστατικό στην αλυσίδα της σύνθεσης φωνής από κείμενο. Η βαθμίδα της φωνητικής μεταγραφής τροφοδοτεί τα επόμενα στάδια της μηχανής με τις προτάσεις σε φωνητική γραφή.

Η φυσικότητα της παραγόμενης συνθετικής ομιλίας εξαρτάται σε μεγάλο βαθμό από την παραγωγή προσωδίας παρόμοιας με την ανθρώπινη φυσική προσωδία.

Υπάρχουν αρκετές τεχνικές για τον χειρισμό της προσωδίας και χωρίζονται, κυρίως βάσει του τρόπου προσέγγισης του ζητήματος, σε τεχνικές μοντελοποίησης (*model based*) και σε τεχνικές που βασίζονται σε δεδομένα (*data driven methods*). Στην δεύτερη περίπτωση, ουσιαστικά δεν εφαρμόζεται μοντέλο προσωδίας, αλλά πραγματοποιείτε προσωδιακή ανάλυση σε επίπεδο πρότασης, η οποία βασίζεται σε επεξεργασία που έχει υποστεί το αυθεντικό ηχογραφημένο σώμα κειμένου. Από την ανάλυση αυτή, εξάγονται παράμετροι που υποβοηθούν σε επόμενο στάδιο την διαδικασία επιλογής ακουστικών μονάδων. Το στάδιο της προσωδιακής ανάλυσης, αποτελεί το τελευταίο βήμα της γλωσσικής επεξεργασίας και τροφοδοτεί το στάδιο της ψηφιακής επεξεργασίας σήματος με τις προτάσεις (σε φωνητική γραφή) και τα επιθυμητά προσωδιακά χαρακτηριστικά τους.

Όπως φαίνεται και στο σχήμα 3.1, οι παραπάνω βαθμίδες λειτουργούν με την αυτόματη εξαγωγή και εφαρμογή κανόνων ενώ, όπου είναι απαραίτητο, συνεπικουρούνται με την χρήση λεξικών. Αυτή η προσέγγιση οδηγεί σε μείωση των απαιτήσεων σε υπολογιστική ισχύ και καθιστά αποδοτικότερη την εφαρμογή των βαθμίδων σε ενσωματωμένα συστήματα. Η λεπτομερής παρουσίαση των ζητημάτων που αφορούν την γλωσσική επεξεργασία κειμένου ξεφεύγει από τα πλαίσια της παρούσης διατριβής. Περισσότερες λεπτομέρειες μπορούν να αναζητηθούν στις αναφορές [Chalamandaris, 2005; Chalamandaris, 2006].

Το υποσύστημα της Ψηφιακής Επεξεργασίας Σήματος (DSP), περιλαμβάνει την επιλογή της τεχνικής ή του μοντέλου που θα χρησιμοποιηθεί για την γέννηση της συνθετικής ομιλίας (π.χ. επιλογή και ένωση/συρραφή ακουστικών μονάδων από βάση δεδομένων προηχογραφημένης φυσικής ομιλίας, παραμετρικό μοντέλο κ.α.) καθώς και όλων των αλγορίθμων που συνδέονται με την εκάστοτε τεχνική. Στην περίπτωση της σύνθεσης φωνής με επιλογή ακουστικών μονάδων, καρδιά του συστήματος είναι η βαθμίδα επιλογής ακουστικών μονάδων (*unit selection engine*) η έξοδος της οποίας καθορίζει τον συρμό από τις μονάδες που επιλέχθηκαν από την βάση δεδομένων και θα ενωθούν (συρραφή) με κατάλληλο τρόπο, στην βαθμίδα σύνθεσης, για την παραγωγή του συνθετικού σήματος ομιλίας. Όπως είδαμε στο κεφάλαιο 2, όπου παρουσιάστηκε η αρχή λειτουργίας, σκοπός της βαθμίδας είναι η παροχή ενός αποδοτικού μηχανισμού για την αυτόματη επιλογή της βέλτιστης ακολουθίας από ακουστικές μονάδες που θα αποτελέσουν το τελικό σήμα της συνθετικής φωνής. Ως βέλτιστη ακολουθία ακουστικών μονάδων ορίζεται εκείνη που οδηγεί στην ελαχιστοποίηση της ολικής συνάρτησης κόστους, η οποία αποτελείται από διάφορα επιμέρους κριτήρια στα οποία λαμβάνονται υπ’ όψιν διάφορες παράμετροι και γνωρίσματα τόσο από την πρόταση που πρόκειται να συντεθεί όσο και από τις υπάρχουσες αποθηκευμένες μονάδες.

Στην περίπτωση του συστήματος σύνθεσης φωνής που εξετάζουμε στην διατριβή, το υποσύστημα της Ψηφιακής Επεξεργασίας Σήματος αποτελείται από τρείς βαθμίδες: α) την βαθμίδα επιλογής ακουστικών μονάδων, β) την βαθμίδα σύνθεσης που στηρίζεται στην τεχνική TD-PSOLA (*Time Domain Pitch Synchronous Overlap Add*) [Moulines; 1990] και, γ) την βαθμίδα διαχείρισης της κατάλληλα προσαρμοσμένης και συμπιεσμένης βάσης δεδομένων. Επιπλέον, η βάση αναπαριστάται με συχνότητα δειγματοληψίας 16KHz, ενώ ως βασική ακουστική μονάδα χρησιμοποιείται το διφώνημα. Η βάση προέρχεται από προηχογραφημένη ομιλία, από γυναίκα ομιλητή, σύμφωνα με την τεχνική που θα εξετάσουμε στην επόμενη ενότητα.

Στις ενότητες που ακολουθούν, περιγράφονται οι τεχνικές και οι αλγόριθμοι που αναπτύχθηκαν με σκοπό την αποδοτική σχεδίαση και προσαρμογή ενός πλήρους συστήματος σύνθεσης φωνής από κείμενο με επιλογή ακουστικών μονάδων στο περιβάλλον της συσκευής του κινητού τηλεφώνου χωρίς σημαντική επίπτωση στην ποιότητα της ομιλίας. Οι ερευνητικές προσπάθειες προσανατολίστηκαν κυρίως σε τρεις δραστηριότητες που σχετίζονται τόσο με την βελτιστοποίηση του απαιτούμενου όγκου δεδομένων όσο και με την αποκλιμάκωση και μείωση των υπολογιστικών αναγκών.

Πιο συγκεκριμένα, σημαντικό παράγοντα αποτελεί η αποδοτική αποκλιμάκωση ή μείωση μιας υπάρχουσας (για χρήση σε Η/Υ ή εξυπηρετητή) βάσης δεδομένων με σκοπό την δημιουργία μικρότερης βάσης για εφαρμογή σε κινητό τηλέφωνο. Στην επόμενη ενότητα παρουσιάζεται μια τέτοια τεχνική η οποία στηρίζεται σε στατιστική ανάλυση των δεδομένων που προκύπτουν από την επιλογή των ακουστικών μονάδων που προκύπτουν από ανάλυση μεγάλου σώματος κειμένου και για τις οποίες λαμβάνεται υπόψη τόσο η συχνότητα εμφάνισης όσο και η «βαθμολογία» που πετυχαίνουν στην διαδικασία αναζήτησης από την μηχανή επιλογής. Η τεχνική διατηρεί την μέγιστη δυνατή ποιότητα λόγω επαυξημένης κάλυψης και αποφυγής επιλογής παρόμοιων ακουστικών μονάδων (*redundant units*). Η ποιότητα και φυσικότητα της συνθετικής ομιλίας που επιτυγχάνεται εξαρτάται άμεσα από τον αριθμό των πραγματώσεων που βρίσκονται αποθηκευμένες στην βάση για κάθε ακουστική μονάδα με αποτέλεσμα να απαιτούνται ιδιαίτερες τεχνικές για τον εμπλουτισμό της βάσης δεδομένων με τις κατάλληλες μονάδες.

Επιπλέον, σημαντικό παράγοντα αποτελεί η μείωση του μεγέθους της βάσης δεδομένων προηχογραφημένης φυσικής ομιλίας, μέσω τεχνικών συμπίεσης και κωδικοποίησης της βάσης. Μελετάται η χρήση και η προσαρμογή απωλεστικών αλγορίθμων κωδικοποίησης φωνής τύπου CELP (*Code Excited Linear Prediction*) [Chu Wai, 2003], για την συμπίεση των ακουστικών μονάδων που αποθηκεύονται στην βάση και εξετάζεται η επίδραση που έχουν στην ποιότητα της παραγόμενης φωνής κατά το στάδιο αποκωδικοποίησης και συρραφής τους. Σημαντικό μέλημα για την προσαρμογή στις ιδιαίτερες απαιτήσεις του συστήματος σύνθεσης φωνής, είναι η δυνατότητα παροχής μηχανισμού τυχαίας πρόσβασης και αποκωδικοποίησης των ακουστικών μονάδων χωρίς να προκαλείται αλλοίωση στις κρίσιμες λειτουργικές παραμέτρους που είναι απαραίτητες για την σύνθεση.

Τέλος, περιγράφεται μια τεχνική που επιτυγχάνει σημαντική μείωση των υπολογιστικών αναγκών της βαθμίδας επιλογής ακουστικών μονάδων. Η διαδικασία επιλογής ακουστικών μονάδων είναι ιδιαίτερα απαιτητική και χρονοβόρα διαδικασία με αποτέλεσμα να απαιτούνται αφενός ανευρετικές τεχνικές στην διαδικασία αναζήτησης και αφετέρου ιδιαίτεροι αλγόριθμοι στον υπολογισμό και την σύγκριση παραμέτρων κατά την εκτέλεση της. Για την ελαχιστοποίηση του υπολογιστικού φορτίου κατά την σύνθεση, υιοθετήθηκε η εφαρμογή συσταδοποίησης (*clustering*) μέσω διανυσματικής κβάντισης (*vector quantization*) για το διάνυσμα της φασματικής αναπαράστασης των ακουστικών μονάδων και τον (*offline*) υπολογισμό των αποστάσεων. Βασικός γνώμονας σε κάθε περίπτωση ήταν η επίτευξη συνθετικού λόγου υψηλής ποιότητας (φυσικότητας και καταληπτότητας). Τα αποτελέσματα της πειραματικής αξιολόγησης επικυρώνουν τις προτεινόμενες προσεγγίσεις [Karabetsos, 2009].

3.2 ΔΗΜΙΟΥΡΓΙΑ ΒΑΣΗΣ ΑΚΟΥΣΤΙΚΩΝ ΜΟΝΑΔΩΝ

Στην ενότητα αυτή παρουσιάζεται η μέθοδος που αναπτύχθηκε για την μείωση του μεγέθους μιας υπάρχουσας (μεγαλύτερης) βάσης δεδομένων ακουστικών μονάδων, με σκοπό την αξιοποίησή της σε γενικού σκοπού συστήματος σύνθεσης φωνής από κείμενο με επιλογή ακουστικών μονάδων στο υπολογιστικό περιβάλλον της συσκευής των κινητών τηλεφώνων. Η τεχνική στηρίζεται σε στατιστικά αποτελέσματα που προκύπτουν από την σύνθεση, με χρήση της βαθμίδας επιλογής ακουστικών μονάδων (στην περίπτωσή μας δίφωνα), μεγάλου σώματος κειμένου. Η μέθοδος αξιοποιεί όχι μόνο την συχνότητα εμφάνισης των ακουστικών μονάδων αλλά και την βαθμολογία, υπό την έννοια των συναρτήσεων κόστους, που αυτά λαμβάνουν έτσι ώστε αφενός να επιλέγει τις πιο συχνά χρησιμοποιούμενες ακουστικές μονάδες και αφετέρου να περιορίζει την επιλογή πλεονάζουσων ακουστικών μονάδων. Ο χαρακτηρισμός ως πλεονάζουσες μονάδες προκύπτει με κριτήριο το πως αντιμετωπίζονται αυτές οι μονάδες από τον αλγόριθμο επιλογής.

Όπως αναφέρθηκε, τα σύγχρονα συστήματα σύνθεσης φωνής γενικού σκοπού στηρίζονται σε μεγάλες βάσεις δεδομένων που απαρτίζονται από πληθώρα ακουστικών μονάδων προσπαθώντας να καλύψουν κατά το δυνατό, κάθε πιθανή πραγμάτωση σε κάθε περιβάλλον (context). Αυτή η προσέγγιση επιφέρει σημαντικό κόστος αναφορικά με τις υπολογιστικές απαιτήσεις. Επίσης, δεν είναι κατάλληλη για περιβάλλοντα μειωμένων υπολογιστικών δυνατοτήτων τόσο λόγω των αυξημένων αποθηκευτικών αναγκών όσο και λόγω της αυξημένης υπολογιστικής ισχύος που απαιτείται για την επεξεργασία τέτοιας βάσης κατά την σύνθεση [Rutten, 2002]. Οι προσεγγίσεις που ακολουθούνται συνήθως για την προσαρμογή των συστημάτων σε ενσωματωμένα περιβάλλοντα, στηρίζονται στην αποκλιμάκωση των υπαρχόντων συστημάτων ακολουθώντας τεχνικές είτε top-down είτε bottom-up [Rutten, 2002]. Στην πρώτη περίπτωση, λαμβάνεται υπόψη μόνο η υπάρχουσα βάση ακουστικών μονάδων και συνήθως πραγματοποιείται συσταδοποίηση (clustering) των ακουστικών μονάδων σύμφωνα με τις ακουστικές και προσωδιακές τους ιδιότητες. Αυτό έχει ως αποτέλεσμα να συρρικνώνεται ο χώρος αναζήτησης αφού πλέον πραγματοποιείται αναζήτηση σε ομάδες ακουστικών μονάδων [Black, 1997; Kim, 2001]. Επιπλέον , υπάρχουν τεχνικές που στηρίζονται σε ανευρετικά ακουστικά και προσωδιακά κριτήρια για την μείωση της υπάρχουσας βάσης [Kumar, 2004]. Αυτές οι προσεγγίσεις έχουν ως μειονέκτημα από την μια πλευρά την έλλειψη κάποιου αντικειμενικού κριτηρίου ή μετρικής, ώστε να αναγνωρίζουν και να ταξινομούν τις ακουστικές μονάδες, και από την άλλη πλευρά ότι η ύπαρξη τέτοιου κριτηρίου δεν συμβαδίζει απαραίτητα με τους μηχανισμούς της βαθμίδας επιλογής ακουστικών μονάδων προκαλώντας έκπτωση της τελικής ποιότητας [Aylet, 2006]. Στην δεύτερη περίπτωση, συναντώνται τεχνικές που βασίζονται σε δεδομένα (data driven) και συγκεκριμένα στην στατιστική συμπεριφορά της βαθμίδας επιλογής ακουστικών μονάδων. Ειδικότερα, η στατιστική επεξεργασία από την συμπεριφορά (έξοδο) της βαθμίδας μετά από σύνθεση μεγάλου σώματος κειμένου, οδηγεί σε επιλογή μονάδων που εμφανίζονται συχνότερα από τα υπόλοιπα και έτσι αυτά διατηρούνται στη νέα μειωμένη βάση [Rutten, 2002; Kim, 2001]. Ωστόσο, για μεγάλα ποσοστά μείωσης, η διατήρηση των συχνότερα εμφανιζόμενων ακουστικών μονάδων δεν αποκλείει και την ταυτόχρονη διατήρηση πλεονάζουσων μονάδων, με αποτέλεσμα να μην επιτυγχάνεται ευρύτερη κάλυψη

για γενικού σκοπού συστήματα σύνθεσης φωνής. Τέλος, τεχνικές όπως αυτή που παρουσιάζεται στο [Krul, 2007] εφαρμόζονται σε συστήματα συγκεκριμένου τομέα εφαρμογής (π.χ., καιρός, αθλητικές ειδήσεις) και δεν αποτελούν την ορθότερη λύση για γενικού σκοπού συστήματα σύνθεσης.

Η μέθοδος που αναπτύχθηκε, εμπίπτει στις μεθόδους της δεύτερης περίπτωσης και αντιμετωπίζει τα παραπάνω ζητήματα συνδυάζοντας τόσο την επιλογή των συχνότερων ακουστικών μονάδων όσο και την αξιολόγηση κάθε ακουστικής μονάδας, όπως αυτή προκύπτει από την βαθμίδα επιλογής και τις συναρτήσεις κόστους. Αξιολογείται δηλαδή τόσο η συχνότητα εμφάνισης κάθε μονάδας όσο και ο βαθμός διαφοροποίησης των μονάδων μεταξύ τους, έτσι ώστε να επιτυγχάνεται ευρύτερη κάλυψη και να αποκλείονται παρόμοιες μονάδες που έχουν ήδη επιλεγεί και να ελαχιστοποιείται ο πλεονασμός. Αυτό επιτυγχάνεται μέσω της διαφοράς μεταξύ των μονάδων όπως αυτή προκύπτει από τις συναρτήσεις κόστους της βαθμίδας επιλογής [Tsiakoulis, 2008; Karabetsos, 2009].

3.2.1 Περιγραφή του αλγόριθμου δημιουργίας βάσης ακουστικών μονάδων

Το σύστημα σύνθεσης φωνής, που εξετάζεται στην παρούσα διατριβή, χρησιμοποιεί δίφωνα ως ακουστικές μονάδες και η βαθμίδα επιλογής στηρίζεται στην μεθοδολογία που αναπτύχθηκε στο 2^{ο} κεφάλαιο. Η ιδέα της τεχνικής στηρίζεται στο συλλογισμό να συλλεχθούν τα πιο συχνά επιλεγόμενα δίφωνα, ενώ ταυτόχρονα να διασφαλίζεται ότι αγνοούνται παρόμοια δίφωνα. Για το σκοπό αυτό, ο αλγόριθμος επιλογής ακουστικών μονάδων εκτελείται σε μεγάλο σώμα κειμένου ώστε να συλλεχθούν επαρκή δεδομένα για την μετέπειτα επεξεργασία. Ειδικότερα, ορίζεται η παρακάτω συνάρτηση βαθμολόγησης (*score function*) για κάθε δίφωνο που χρησιμοποιείται κατά την σύνθεση. Η συνάρτηση ορίζεται σαν το συνδυασμένο τοπικό κόστος «στόχος» και «ένωσης» (target cost και join cost),

$$S(u_j^n) = C^t(t_n, u_j^n) + \min_k (C^c(u_j^n, u_k^{n-1})) \qquad (3.1)$$

όπου, u_j^n η j-οστή εκδοχή του n-οστού διφώνου και το t_n αποτελεί το δίφωνο στόχο. Ο δεύτερος όρος αποτελεί την συνάρτηση κόστους που εκφράζει και επιστρέφει το καλύτερο κόστος ένωσης του διφώνου u_j^n, από όλες τις πραγματώσεις του προηγούμενου διφώνου u^{n-1}, σε ευθύ τρόπο αναζήτησης Viterbi (forward Viterbi search).

Στην διαδικασία του αλγόριθμου υποθέτουμε ότι αν δύο πραγματώσεις του ίδιου διφώνου βαθμολογούνται κοντά, τότε είναι παρόμοια, από την άποψη της συμπεριφορά της βαθμίδας επιλογής, χωρίς να εξετάζεται κάποιο άλλο (συνήθως) υποκειμενικό κριτήριο. Αυτό πιστοποιείται και από τον ίδιο τον αλγόριθμο επιλογής μονάδων, αφού με αυτό τον τρόπο αποδίδει το καλύτερο μονοπάτι. Συνεπώς, μετά από την σύνθεση ενός αρκετά μεγάλου σώματος κειμένου, οι μέσες διαφορές στις βαθμολογίες μπορούν να χρησιμοποιηθούν σαν μετρική αξιολόγησης μεταξύ πραγματώσεων για το ίδιο δίφωνο.

Πιο αναλυτικά, για όλες τα προτάσεις που συντέθηκαν από το μεγάλο σώμα κειμένου, τα στατιστικά δεδομένα που συλλέγονται αφορούν τα εξής μεγέθη:

- Την συχνότητα επιλογής, f_j^n για κάθε u_j^n ή αλλιώς τον ολικό αριθμό που δείχνει πόσες φορές επιλέχθηκε κάποια πραγμάτωση.

- Την μέση διαφορά βαθμολογίας $D_{j,k}^n = \overline{|S(u_j^n) - S(u_k^n)|}$ για όλα τα ζεύγη πραγματώσεων του ίδιου διφώνου (με τις βαθμολογίες να αναφέρονται κάθε φορά στην ίδια πρόταση).

Η μέθοδος βασίζεται και στα δύο παραπάνω μεγέθη ώστε να διαλέξει τις κατάλληλες πραγματώσεις ενός διφώνου.

Έστω K ο αριθμός των πραγματώσεων ενός συγκεκριμένου διφώνου στην διαθέσιμη βάση δεδομένων και M ο αριθμός των επιθυμητών πραγματώσεων του στην νέα (μειωμένη) βάση με M < K. Ο αλγόριθμος επιλογής δίνεται στον παρακάτω πίνακα:

ΠΙΝΑΚΑΣ 3.1: Αλγόριθμος επιλογής πραγματώσεων ακουστικών μονάδων για την δημιουργία μειωμένης βάσης δεδομένων

1. Αρχικοποίηση $F = [f_1^n, f_2^n \ldots f_K^n]$
2. Επιλογή $m = \arg\max_n F[n]$
3. Ανανέωση $F[n] = F[n] \cdot D_{n,m}^i$ για *n = 1 … K*
4. Αν #(*επιλεγμένων πραγματώσεων*) < *M* τότε επιστροφή στο βήμα 2
5. Τερματισμός

Στον αλγόριθμο, η συνάρτηση *F* ορίζεται σαν το διάνυσμα καταλληλότητας (*fitness vector*) το οποίο αρχικοποιείται από τις συχνότητες επιλογής των πραγματώσεων. Στην συνέχεια, με επαναληπτικό τρόπο, επιλέγεται η πραγμάτωση με το καλύτερο διάνυσμα καταλληλότητας. Σημαντικό σημείο στην εκτέλεση του αλγόριθμου αποτελεί το βήμα 3, στο οποίο η ανανέωση του διανύσματος γίνεται με τέτοιο τρόπο ώστε να αποφεύγονται όμοιες πραγματώσεις. Η ιδέα στηρίζεται σε λογική παρόμοια με αυτή από τον χώρο των γενετικών αλγόριθμων (π.χ. fitness sharing), λόγω της ανανέωσης του διανύσματος σε κάθε βήμα επανάληψης. Η καταλληλότητα κάθε πραγμάτωσης πολλαπλασιάζεται με την μέση διαφορά της βαθμολογίας από την τελευταία πραγμάτωση που επιλέχθηκε. Με αυτό τον τρόπο, η καταλληλότητα όμοιων πραγματώσεων σε σχέση με αυτές που ήδη έχουν επιλεγεί μειώνεται, ενώ το αντίθετο συμβαίνει για διαφορετικές πραγματώσεις. Επιπλέον, η *F[m]* μηδενίζεται στην περίπτωση το ήδη επιλεχθέντων μονάδων.

Η τιμή στόχος για τον αριθμό Μ των επιθυμητών πραγματώσεων ενός διφώνου στην νέα βάση δεδομένων καθορίζεται από τον ρυθμό μείωσης ή αλλιώς το τελικό ποσοστό μείωσης της καινούριας βάσης ακουστικών μονάδων. Συνήθως χρησιμοποιείται κάποιο κριτήριο σχετικό με την επιθυμητή κάλυψη (coverage meeting criterion) [Rutten, 2002], το οποίο όμως για τον προτεινόμενο αλγόριθμο δεν αποτελεί την καλύτερη λύση αφενός επειδή οι συχνότερες ακουστικές μονάδες μπορεί να έχουν απορριφθεί και αφετέρου επειδή μια συγκεκριμένη κάλυψη δεν διασφαλίζει ομοιόμορφο ρυθμό μείωσης σε όλα τα είδη των ακουστικών μονάδων.

Για τον συγκεκριμένο αλγόριθμο, προτείνεται η χρήση της σχέσης 3.2 ώστε να διασφαλίζεται ομοιομορφία στην μείωση των πραγματώσεων για κάθε δίφωνο.

$$M = \min(M_{\max}, \max(M_{\min}, \log_b(K))) \qquad (3.2)$$

όπου τα μεγέθη $M_{\max}$ και $M_{\min}$, με $M_{\max} > M_{\min}$ ορίζουν τον μέγιστο και ελάχιστο αριθμό πραγματώσεων σε κάθε ακουστική μονάδα (δίφωνο) αντίστοιχα, ενώ η παράμετρος *b* προσδιορίζει λογαριθμική σχέση στον ρυθμό μείωσης.

3.2.2 Πειραματική αξιολόγηση και αποτελέσματα

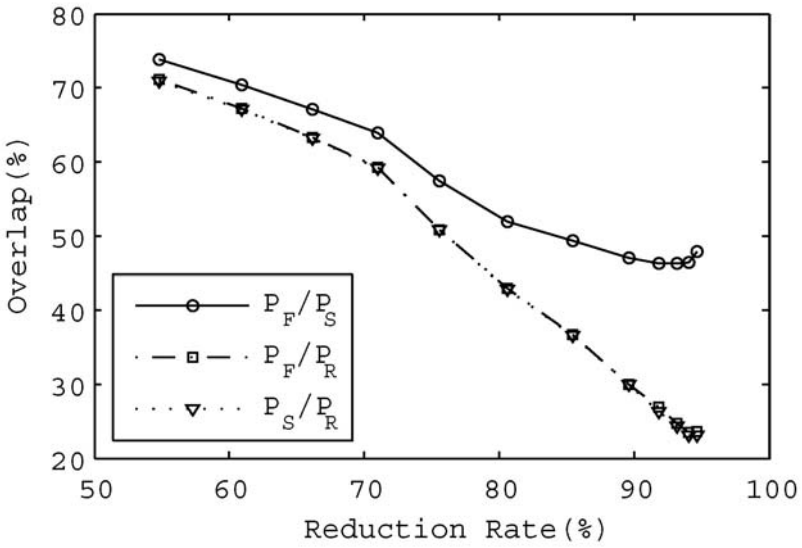
Η πειραματική αξιολόγηση του αλγόριθμου δημιουργίας μειωμένης βάσης δεδομένων ακουστικών μονάδων, στηρίχθηκε στην σύνθεση μεγάλου σώματος κειμένου γενικού περιεχομένου και κατόπιν στην επεξεργασία των αποτελεσμάτων που προέκυψαν από την βαθμίδα επιλογής ακουστικών μονάδων. Τα πειράματα πραγματοποιήθηκαν σε σύστημα σύνθεσης φωνής από κείμενο με αρχική (πλήρη) βάση δεδομένων που αποτελείται από περίπου 1291 ηχογραφημένες προτάσεις από γυναίκα ομιλητή. Η βάση αντιπροσωπεύει 1098 διαφορετικά δίφωνα και περιέχει περίπου 115.000 συνολικές πραγματώσεις. Σύμφωνα με τις υπολογιστικές δυνατότητες μιας 'μέσης' συσκευής κινητού τηλεφώνου, κρίνεται σκόπιμο να επιτευχθεί ρυθμός μείωσης έως και 95% της αρχικής βάσης, δηλαδή η νέα βάση να αποτελείται συνολικά από περίπου 5.750 πραγματώσεις διφώνων.

Το σώμα κειμένου που χρησιμοποιήθηκε για σύνθεση και καταγραφή των αποτελεσμάτων της βαθμίδας επιλογής ακουστικών μονάδων, αποτελείτο από 12.500 προτάσεις γενικού περιεχομένου και περιείχε 1.5Μ από πραγματώσεις διφώνων. Το 95% αυτού του σώματος χρησιμοποιήθηκε για την συλλογή δεδομένων και την στατιστική επεξεργασία της μηχανής επιλογής ακουστικών μονάδων ενώ το υπόλοιπο 5% χρησιμοποιήθηκε για την αξιολόγηση της μεθόδου.

Επιπρόσθετα, η προτεινόμενη μέθοδος, η οποία θα συμβολίζεται εφεξής ως *Pf*, συγκρίθηκε πειραματικά με την τεχνική που προτείνεται στο [Rutten, 2002] και στηρίζεται στην επιλογή πραγματώσεων με μοναδικό κριτήριο την συχνότητα εμφάνισης. Η μέθοδος των [Rutten, 2002] θα συμβολίζεται εφεξής ως *Ps*. Ως σημείο αναφοράς, παρατίθενται αποτελέσματα και για την τεχνική που στηρίζεται σε τυχαία επιλογή πραγματώσεων και η οποία συμβολίζεται ως *Pr*. Σε κάθε περίπτωση, η πειραματική σύγκριση αφορά τον ίδιο τελικό αριθμό (Μ) πραγματώσεων για κάθε δίφωνο.

Για την αξιολόγηση των παραπάνω τεχνικών, αρχικά κατασκευάστηκαν αρκετές βάσεις μεταβάλλοντας το τελικό ποσοστό μείωσης. Στο σχήμα 3.2, φαίνεται το ποσοστό επικάλυψης (για ίδιες πραγματώσεις), δηλαδή το ποσοστό επιλογής ίδιων

διφώνων, μεταξύ των μεθόδων για τον ίδιο ρυθμό μείωσης. Από τα διαγράμματα παρατηρείται η διαφοροποίηση των μεθόδων, ιδιαίτερα για μεγάλους ρυθμούς μείωσης. Για μικρά ποσοστά μείωσης, οι *Pf* και *Ps* δεν διαφοροποιούνται σημαντικά οπότε καταλήγουν σε δημιουργία βάσης με ίδια δίφωνα. Όμως, για ποσοστά μείωσης μεγαλύτερα του 80%, το ποσοστό επικάλυψης σταθεροποιείται γύρω στο 50%, που σημαίνει ότι οι βάσεις που προκύπτουν είναι διαφορετικές μεταξύ τους. Επιπλέον, από το διάγραμμα προκύπτει ότι οι δύο μέθοδοι σε σχέση με την *Pr* παρουσιάζουν χαμηλά ποσοστά επικάλυψης. Ωστόσο, πρέπει να σημειωθεί ότι σε σχετικά μεγάλους ρυθμούς επικάλυψης και για τα λιγότερο συχνά δίφωνα, δηλαδή στην περίπτωση που δεν υπάρχουν αρκετές πραγματώσεις σε μεμονωμένα δίφωνα όσες, δηλαδή όταν $K \leq M_{min}$, υπάρχει ασυμφωνία μεταξύ του ολικού ποσοστού μείωσης και αυτού για συγκεκριμένο δίφωνο. Σε αυτή την περίπτωση και οι δύο μέθοδοι επικαλύπτονται πλήρως.



ΣΧΗΜΑ 3.2: Ποσοστό επικάλυψης μεταξύ των βάσεων δεδομένων που προκύπτουν από τις τεχνικές *Pf* , *Ps και Pr* για διαφορετικά ποσοστά μείωσης.

Για την αξιολόγηση της προτεινόμενης μεθόδου καθώς και της βάσης δεδομένων που τελικά προκύπτει, χρησιμοποιήθηκαν μετρικές που προέρχονται από στατιστικές παραμέτρους που περιγράφουν την συμπεριφορά της βαθμίδας επιλογής διφώνων. Με άλλα λόγια, διερευνάται η συμπεριφορά της βαθμίδας για κάθε βάση που προκύπτει. Οι παράμετροι που χρησιμοποιούνται για το σκοπό αυτό, είναι οι μέσες τιμές των συναρτήσεων κόστους που αφορούν το κόστος επιθυμητής ακολουθίας (*target cost*), το κόστος ένωσης (*join cost*) και το ολικό κόστος (*total cost*), όπως αυτά προκύπτουν από την επιλογή του καλύτερου μονοπατιού ή με άλλα λόγια από τα διφωνα που επελέγησαν σε κάποια σύνθεση. Επιπλέον, λαμβάνονται υπόψη και οι μέγιστες τιμές αυτών των παραμέτρων έτσι ώστε να είναι δυνατή η ανίχνευση ασυνεχειών (είτε προσωδιακών είτε φασματικών) αφού μεγάλα κόστη δείχνουν πιθανά σημεία ασυνεχειών γεγονός που έχει μεγάλη πιθανότητα να συμβεί σε συστήματα με μικρές (μειωμένες) βάσεις από

ακουστικές μονάδες. Όλες οι προηγούμενες παράμετροι υπολογίζονται ανά συντεθειμένη πρόταση και υπολογίζεται η μέση τιμή τους σε όλο το σώμα κειμένου που χρησιμοποιήθηκε για αξιολόγηση (test corpus).

Το σχήμα 3.3, δείχνει τα συγκριτικά αποτελέσματα μεταξύ των μεθόδων *Pf* και *Ps*. Η μέθοδος *Pr* δεν εμφανίζεται καθώς δεν επιτυγχάνει αξιόλογα αποτελέσματα. Σαν σημείο αναφοράς, οι τιμές των παραμέτρων στην πλήρη (αρχική) βάση δεδομένων είναι,

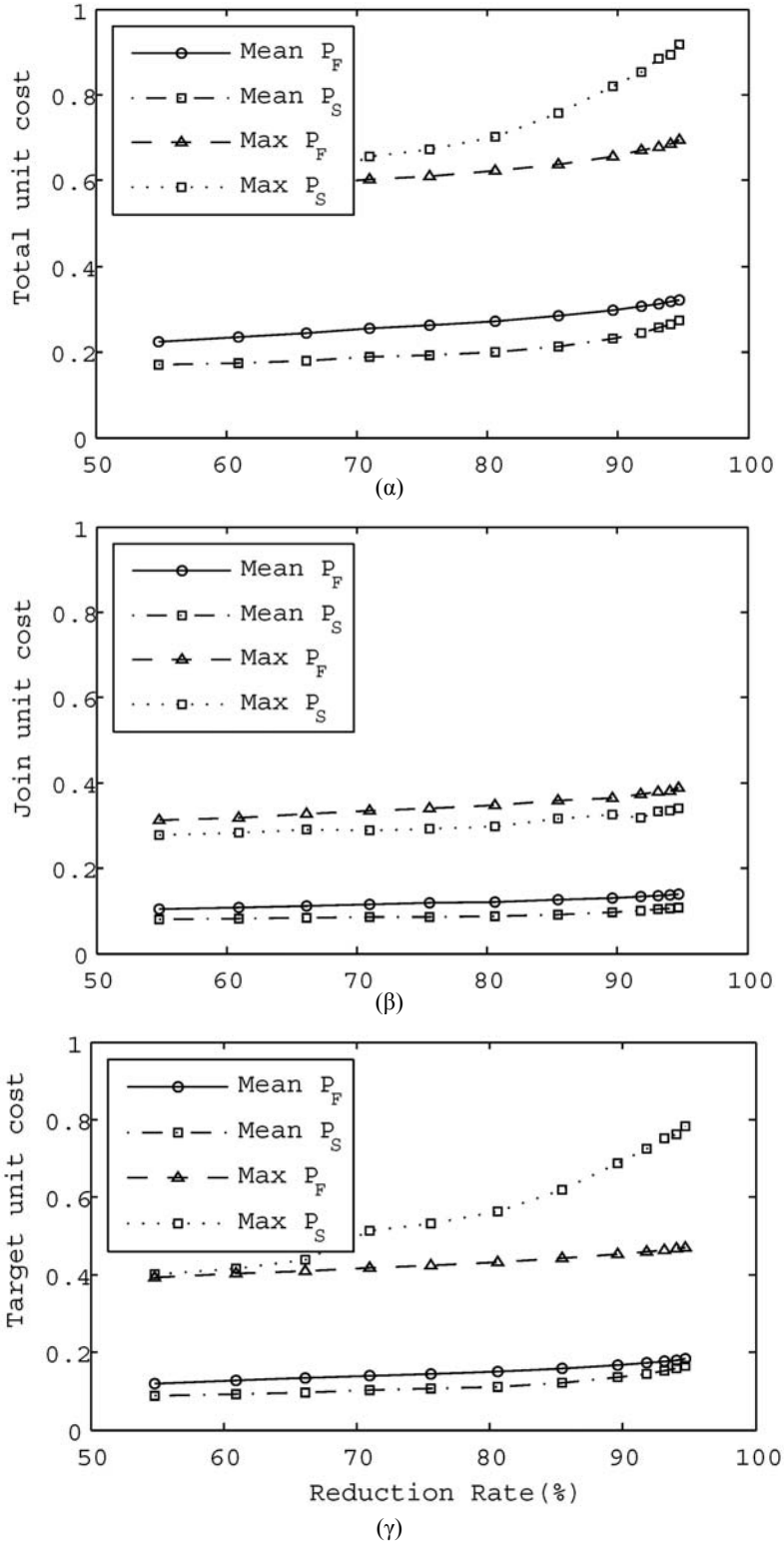
- **$\{total, join, target\}_{mean} = \{0.15, 0.07, 0.07\}$**
- **$\{total, join, target\}_{max} = \{0.\ 5, 0.27, 0.34\}$**

Από το σχήμα φαίνεται ότι αν και η μέθοδος *Ps* πετυχαίνει καλύτερα αποτελέσματα υπό την έννοια των μέσων κοστών, η μέθοδος *Pf* έχει μικρότερη μέση τιμή σε μέγιστα κόστη ανά πρόταση, γεγονός πιο εμφανές για μεγάλα ποσοστά μείωσης. Από αυτό συμπεραίνουμε ότι η *Ps* οδηγεί σε βάσεις που χαρακτηρίζονται από καλύτερα κόστη σε μέση τιμή αλλά περιλαμβάνει και πραγματώσεις που οδηγούν σε μεγάλο κόστος. Από την άλλη πλευρά, η *Pf* οδηγεί σε βάσεις με καλύτερο target κόστος αλλά με υψηλότερο join κόστος.

Για περαιτέρω αξιολόγηση της προτεινόμενης μεθόδου πραγματοποιήθηκαν μικρής κλίμακας ακουστικά πειράματα. Για το σκοπό αυτό έγινε τυχαία επιλογή 35 προτάσεων, 2 με 16 λέξεις η καθεμία, από το σώμα κειμένου για έλεγχο (test corpus). Η σύνθεση των προτάσεων έγινε και με τις δύο μεθόδους, χρησιμοποιώντας βάσεις δεδομένων που προέκυψαν για ποσοστό μείωσης της τάξης του 93%. Η ακουστική αξιολόγηση πραγματοποιήθηκε από μεικτή ομάδα 15 ανθρώπων που περιελάμβανε τόσο άτομα με εμπειρία σε τεχνολογίες σύνθεσης φωνής από κείμενο, όσο και άτομα χωρίς τέτοια εμπειρία. Η ακουστική αξιολόγηση αφορούσε την βαθμολόγηση σε κλίμακα από 1 έως 5 (1: κακή ποιότητα, 5: άριστη ποιότητα) των προτάσεων. Η ακουστική αξιολόγηση αυτού του είδους ονομάζεται ***Mean Opinion Score (MOS)*** και χρησιμοποιείται ευρέως για αξιολόγηση συστημάτων σύνθεσης φωνής από κείμενο [Alvarez, 2004; SpeechWorks, 2006]. Η παρουσίαση των προτάσεων στους ακροατές γινόταν ανά ζεύγη (ίδια πρόταση από κάθε μέθοδο) χωρίς όμως να γνωστοποιείται η μέθοδος για κάθε πρόταση. Τα αποτελέσματα φαίνονται στον πίνακα 3.2, που παράλληλα δείχνει και τις τιμές από τις στατιστικές παραμέτρους που αναφέρθηκαν προηγουμένως.

ΠΙΝΑΚΑΣ 3.2: Συγκριτική ακουστική αξιολόγηση MOS των μεθόδων *Pf* και *Ps* σε ζεύγη προτάσεων με δείγμα 15 ατόμων.

	MOS	Mean Costs {total, join, target}	Max Costs {total, join, target}
Pf	4.01	0.32, 0.14, 0.18	0.61, 0.34, 0.40
Ps	3.92	0.27, 0.11, 0.16	0.84, 0.30, 071



ΣΧΗΜΑ 3.3: Συγκριτική αξιολόγηση για τις τεχνικές *Pf* και *Ps*. Συναρτήσει του ποσοστού μείωσης φαίνονται, (α) η μέση τιμή του μέσου και του μέγιστου ολικού κόστους ανά πρόταση για την *Pf* και *Ps*, (β) οι τιμές του μέσου και του μέγιστου κόστους ένωσης για τις *Pf* και *Ps*, και (γ) οι τιμές του μέσου και του μέγιστου κόστους επιθυμητής ακολουθίας για τις *Pf* και *Ps*.

Από τα αποτελέσματα προκύπτει ότι η μέθοδος *Pf* φαίνεται να οδηγεί σε υψηλότερης ποιότητα συνθετική φωνή σε σχέση με την *Ps*. Ωστόσο, οι διαφορές των δύο μεθόδων δεν είναι ιδιαίτερα σημαντικές. Από την άλλη πλευρά, οι τιμές MOS συμφωνούν με την μέση τιμή του ολικού κόστους ανά πρόταση. Έτσι, επιβεβαιώνεται η αρχική υπόθεση ότι η *Ps* οδηγεί σε βάσεις με υψηλό πλεονασμό καθώς επιλέγει παρόμοιες πραγματώσεις και δεν αφήνει περιθώριο να επιλεγούν σπανιότερες αλλά εξίσου σημαντικές πραγματώσεις, οι οποίες πρέπει να υπάρχουν σε κάποιο σύστημα σύνθεσης φωνής από κείμενο γενικού περιεχομένου για να καλύπτουν, με υψηλή ποιότητα, ποικίλες περιπτώσεις σύνθεσης. Επιπλέον, η καλύτερη συσχέτιση που φαίνεται μεταξύ των τιμών MOS και των μέγιστων κοστών (*Max Costs*) σε αντίθεση με τα μέσα κόστη (*Mean Costs*) συνάδει με το συμπέρασμα για περαιτέρω βελτίωση της ολικής συνάρτησης κόστους, όπως για παράδειγμα προτείνεται στο [Toda, 2006].

Συμπερασματικά, η προτεινόμενη μέθοδος πετυχαίνει αρκετά υψηλή τιμή MOS στα ακουστικά πειράματα γεγονός που υποδεικνύει ότι η τελική ποιότητα (φυσικότητα και καταληπτότητα) της συνθετικής ομιλίας, παρά την μείωση της αρχικής βάσης δεδομένων κατά 93%, είναι εξίσου αποδεκτή, με αποτέλεσμα η σύνθεση φωνής στο περιβάλλον της συσκευής του κινητού τηλεφώνου να μην υστερεί σημαντικά από την αντίστοιχη σε υπολογιστικά περιβάλλοντα τύπου Η/Υ ή εξυπηρετητή. Επιπλέον, η τεχνική διασφαλίζει την ικανότητα σύνθεσης κειμένων γενικού περιεχομένου λόγω της ευρείας κάλυψης της κατανομής των πιθανών πραγματώσεων που μπορούν πρακτικά να προκύψουν.

3.3 ΣΥΜΠΙΕΣΗ ΚΑΙ ΚΩΔΙΚΟΠΟΙΗΣΗ ΤΗΣ ΒΑΣΗΣ ΑΚΟΥΣΤΙΚΩΝ ΜΟΝΑΔΩΝ

Τα συστήματα σύνθεσης φωνής από κείμενο με επιλογή και ένωση ακουστικών μονάδων από βάση δεδομένων προηχογραφημένης φυσικής ομιλίας, έχουν σημαντικές απαιτήσεις τόσο σε μνήμη όσο και σε αποθηκευτικούς χώρους με αποτέλεσμα να μην είναι άμεση η εφαρμογή τους σε υπολογιστικά περιβάλλοντα με μειωμένους πόρους, όπως είναι για παράδειγμα τα κινητά τηλέφωνα ή οι φορητές συσκευές τύπου PDA. Για παράδειγμα, μια βάση δεδομένων ηχογραφημένης φυσικής ομιλίας διάρκειας 20 λεπτών απαιτεί περίπου 39MByte για δειγματοληψία στα 16KHz, και με 16bits για κάθε δείγμα. Φυσικά τέτοια μεγέθη είναι απαγορευτικά για την συσκευή του κινητού τηλεφώνου κρίνοντας επιτακτική την ανάγκη για αποδοτική συμπίεση [Schnell, 2002].

Ωστόσο, σε σχέση με το γενικό πρόβλημα συμπίεσης και κωδικοποίησης του σήματος φωνής, η περίπτωση της σύνθεσης φωνής, παρουσιάζει σημαντικές ιδιαιτερότητες τόσο σε επίπεδο σχεδίασης όσο και στο τελικό ζητούμενο αποτέλεσμα. Ειδικότερα, η διατήρηση της φυσικότητας και καταληπτότητας του αρχικού σήματος είναι πρωταρχικός παράγοντας. Επιπλέον, σημαντικό ζήτημα αποτελεί η υπολογιστικά γρήγορη και αποδοτική διαδικασία αποκωδικοποίησης, αντίθετα με την διαδικασία κωδικοποίησης, η οποία μπορεί να είναι αρκετά σύνθετη και χρονοβόρα, καθώς πραγματοποιείται κατά την σχεδίαση της βάσης σε συνθήκες μη πραγματικού χρόνου (*offline*). Σημαντικό ζητούμενο επίσης αποτελεί η δυνατότητα, η τεχνική συμπίεσης να επιτρέπει και να παρέχει τυχαία πρόσβαση στα

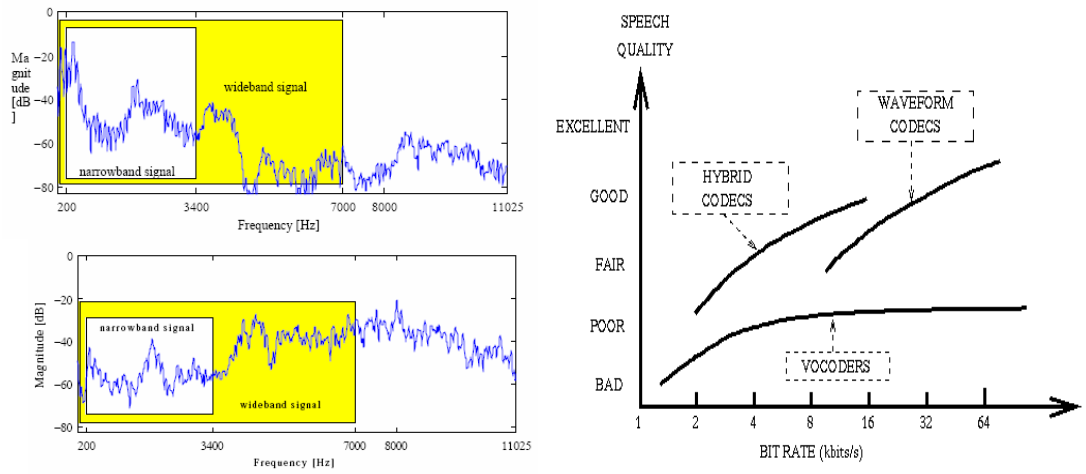
κωδικοποιημένα δεδομένα, ώστε να επιτυγχάνεται η τυχαία προσπέλαση στις ακουστικές μονάδες [Chang-Heon Lee, 2007; Van der Vrecken, 1997].

Για τη μηχανή σύνθεσης φωνής από κείμενο για κινητό τηλέφωνο μελετήθηκε και προσαρμόστηκε κατάλληλα, η τεχνική συμπίεσης CELP (Code Excited Linear Prediction) που αποτελεί την πλέον διαδεδομένη τεχνική κωδικοποίησης στην οικογένεια των κωδικοποιητών ανάλυσης μέσω σύνθεσης [Schroeder, 1985; Chu Wai C., 2003]. Η τεχνική εφαρμόστηκε με τέτοιο τρόπο ώστε να διασφαλίζεται η τυχαία πρόσβαση στις ακουστικές μονάδες και να επιτυγχάνεται υψηλή τελική ποιότητα με σημαντική αύξηση του συντελεστή συμπίεσης. Η τελική βάση που προέκυψε για το κινητό τηλέφωνο, έχει μέγεθος περίπου 4MByte, νούμερο που είναι άμεσα συγκρίσιμο με αντίστοιχες (εμπορικές) μηχανές σύνθεσης διεθνώς. Η τεχνική επιλέχθηκε κατόπιν αξιολόγησης και σύγκρισης και με άλλες τεχνικές τόσο εξειδικευμένες για το σήμα της φωνής όσο και γενικά στο χώρο του ήχου (*audio*). Ειδικότερα στο χώρο της φωνής, σε σχέση με άλλες τεχνικές, η CELP επιτυγχάνει τα καλύτερα αποτελέσματα. Στα πλαίσια του έργου δοκιμάστηκε και η μέθοδος VORBIS (www.vorbis.com), η οποία όμως πετυχαίνει χαμηλότερο βαθμό συμπίεσης για το ίδιο επίπεδο ποιότητας, συγκριτικά με την CELP στο σήμα φωνής. Η αποτελεσματικότητα και η επιτυχής προσαρμογή της CELP πιστώνεται από τα ακουστικά πειράματα που είχαν σαν στόχο την αξιολόγηση της ως προς την διατήρηση της αρχικής ποιότητας.

3.3.1 Προσαρμογή της μεθόδου CELP στην μηχανή σύνθεσης φωνής

Ο σκοπός της κωδικοποίησης φωνής είναι η αναπαράσταση του σήματος της φωνής με ελάχιστο αριθμό από bits, ενώ ταυτόχρονα να διατηρείται η τελική ποιότητα, δηλαδή η καταληπτότητα και η φυσικότητα. Η μείωση του απαιτούμενου αριθμού από bits οδηγεί σε λιγότερη απαίτηση στον απαιτούμενο ρυθμό μετάδοσης (*bit rate*), άρα και στο απαιτούμενο εύρος ζώνης. Βέβαια, η διατήρηση της καταληπτότητας και της φυσικότητας και η κωδικοποίηση με ελάχιστο αριθμό από bits, είναι στόχοι αντικρουόμενοι (*tradeoff*) και ο σκοπός της σχεδίασης ενός επιτυχημένου κωδικοποιητή είναι είτε η εύρεση βέλτιστης λύσης μεταξύ των παραπάνω στόχων, γεγονός αρκετά δύσκολο, είτε η προσαρμογή του στις απαιτήσεις κάποιας εφαρμογής με θέσπιση συγκεκριμένων προδιαγραφών. Στην περίπτωση των συστημάτων κωδικοποίησης φωνής έχουν προταθεί αρκετές τεχνικές και έχουν προκύψει αρκετές τυποποιήσεις για πληθώρα εφαρμογών [Gibson, 2005; Chu Wai C., 2003]. Οι αρχικές προσπάθειες ήταν εστιασμένες στην ανάπτυξη κωδικοποιητών φωνής στενής ζώνης φάσματος (*narrowband speech coding*) λόγω της ανάγκης για εξοικονόμηση φάσματος (π.χ., σε τηλεφωνικά δίκτυα - PSTN). Η τεχνολογική πρόοδος και οι απαίτηση για υπηρεσίες υψηλής ποιότητας, οδήγησαν στην ανάπτυξη συστημάτων κωδικοποίησης ευρείας ζώνης (*wideband speech coding*) καθώς και σε νέες τυποποιήσεις προς τον σκοπό αυτό. Με τον όρο στενή ζώνη εννοούμε ότι το σήμα φωνής υπόκειται σε δειγματοληψία με ρυθμό 8KHz και φιλτράρεται έτσι ώστε το φάσμα του να περιορίζεται στην ζώνη συχνοτήτων 200-3400Hz. Αντίθετα, στην επεξεργασία ευρείας ζώνης, η δειγματοληψία γίνεται στα 16KHz, ενώ το εύρος ζώνης που διατηρείται είναι στην περιοχή 50-7000Hz. Στις επικοινωνίες φωνής, η διαφορά μεταξύ ευρείας και στενής ζώνης έγκειται στην καλύτερη ποιότητα και καταληπτότητα καθώς στην πρώτη

περίπτωση η πληροφορία που περιέχεται στις υψηλές συχνότητες διατηρείται. Ένα παράδειγμα του φάσματος στενής και ευρείας ζώνης του σήματος φωνής φαίνεται στο σχήμα 3.4(α). Ο απαιτούμενος ρυθμός είτε μετάδοσης (για streaming πραγματικού χρόνου) είτε αποθήκευσης (στην περίπτωση της σύνθεσης φωνής) είναι 64kbps στην περίπτωση στενής ζώνης για 8bit/sample και 256kbps στην περίπτωση ευρείας ζώνης για 16bit/sample, για συχνότητα δειγματοληψίας 8KHz. Σημειώνουμε ότι τα συστήματα σύνθεσης φωνής χρειάζονται τουλάχιστον 11KHz συχνότητα δειγματοληψίας, ώστε να διατηρείται η ποιότητα σε υψηλό επίπεδο. Είναι φανερή λοιπόν η ανάγκη για μείωση του απαιτούμενου ρυθμού μετάδοσης και αποθήκευσης μέσω αποτελεσματικών, όσο αναφορά την τελική ποιότητα, τεχνικών συμπίεσης.

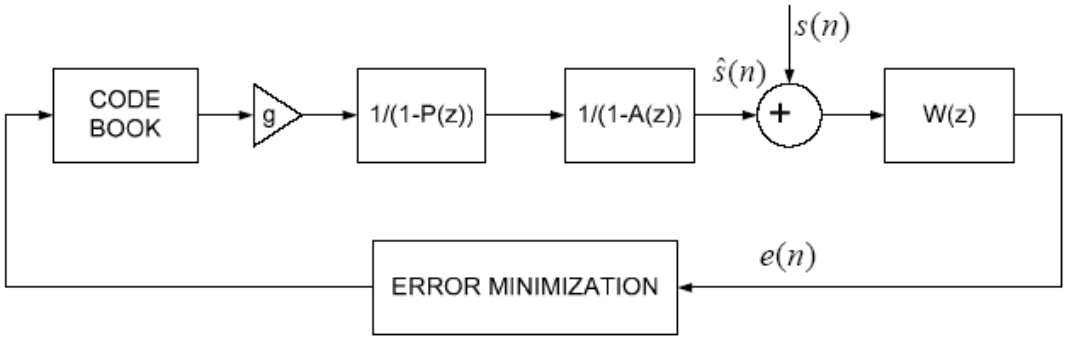


ΣΧΗΜΑ 3.4: α) Κωδικοποίηση στενής και ευρείας ζώνης, β) Σχέση ποιότητας και απαιτούμενου ρυθμού μετάδοσης για τις οικογένειες των τεχνικών κωδικοποίησης (Πηγή: [Gibson, 2005]).

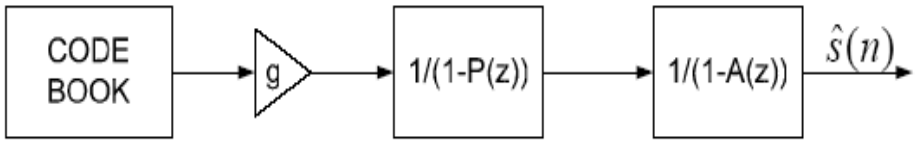
Η κωδικοποίηση και η συμπίεση γενικότερα, στηρίζεται σε πλεοναστικές ιδιότητες που εμπεριέχονται στα δεδομένα. Στην περίπτωση του σήματος φωνής τέτοια ιδιότητα είναι η «ψευδό» στασιμότητα (*quasi-stationary*) η οποία έχει σαν αποτέλεσμα την εμφάνιση προβλεψιμότητας. Με άλλα λόγια, το σήμα φωνής είναι δυναμικό αλλά με σχετικά αργές μεταβολές. Τα συστήματα κωδικοποίησης φωνής χωρίζονται σε τρεις κατηγορίες: α) κωδικοποιητές κυματομορφής (*waveform coders*), b) κωδικοποιητές πηγής (*source coders / vocoders*), c) υβριδικοί κωδικοποιητές (*hybrid coders*) ή κωδικοποιητές ανάλυσης μέσω σύνθεσης (*AbS – Analysis-by-Synthesis*). Από άποψη ποιότητας, η πρώτη κατηγορία είναι καλύτερη αλλά με υψηλούς ρυθμούς μετάδοσης ενώ η τρίτη δίνει συγκρίσιμα αποτελέσματα σε χαμηλότερους ρυθμούς. Χειρότερη είναι η δεύτερη κατηγορία, η οποία όμως επιτυγχάνει πολύ μικρούς ρυθμούς μετάδοσης [Gibson, 2005; Chu Wai C., 2003]. Στο σχήμα 3.4(β) φαίνεται ένα συγκριτικό γράφημα για τις τεχνικές .

Οι κωδικοποιητές ανάλυσης μέσω σύνθεσης αποτελούν τεχνολογία αιχμής, καθώς επιτυγχάνουν χαμηλούς ρυθμούς μετάδοσης με σχετικά υψηλή ποιότητα. Η κυριότερη διαφορά με τους κωδικοποιητές πηγής έγκειται στην αντιμετώπιση του

σήματος διέγερσης καθώς αυτή προσδιορίζεται μέσω ανάλυσης του σήματος φωνής και ελαχιστοποίησης της απόστασης (ή ελαχιστοποίηση του σφάλματος) του αυθεντικού από το ανακατασκευασμένο σήμα. Η ιδέα πρωτοεμφανίστηκε με την πολυπαλμική διέγερση (*MPE – Multi Pulse Excitation*) και κατόπιν αναπτύχθηκε μέσω των τεχνικών RPE (*Regular Pulse Excitation*) και CELP (*Code Excited Linear Prediction*) (σχήμα 3.5). Σε γενικές γραμμές, η αρχή λειτουργίας έχει ως εξής. Το σήμα φωνής χωρίζεται σε πλαίσια από τα οποία μέσω ανάλυσης (είτε γραμμική πρόβλεψη είτε άλλη ανάλυση) προκύπτουν οι συντελεστές του φίλτρου σύνθεσης και η κατάλληλη διέγερση μέσω της οποίας επιτυγχάνεται πιστότερη (υπό κάποιο κριτήριο) παραγωγή του αρχικού σήματος (*short-term analysis*). Αυτές οι παράμετροι τελικά μεταδίδονται κωδικοποιημένες. Επιπλέον, στον κωδικοποιητή χρησιμοποιείται είτε φίλτρο είτε κάποιο κωδικό-βιβλίο (*codebook*) ώστε να προβλέπεται και η θεμελιώδης συχνότητα του σήματος (*long-term analysis*). Πρακτικά, αφενός λόγω του μεγάλου αριθμού δοκιμών για την εύρεση της κατάλληλης διέγερσης και αφετέρου για αποδοτικότερη κωδικοποίηση, χρησιμοποιείται διανυσματική κβάντιση (*vector quantization*) και κατασκευάζονται κωδικό-βιβλία με πιθανές διεγέρσεις. Για παράδειγμα, 1024 πιθανές διεγέρσεις χρειάζονται 10 bit για αναπαράσταση. Έτσι στο δέκτη στέλνεται μόνο ο δείκτης στο κωδικό-βιβλίο. Η προσέγγιση της κατάλληλης διέγερσης είναι ο παράγοντας που διαχωρίζει την MPE από την RPE και την CELP. Στην MPE χρησιμοποιείται καθορισμένος αριθμός παλμών, οπότε κωδικοποιείται η θέση και το πλάτος, ενώ στην RPE καθορίζεται και η θέση σε σχέση με το πρώτο παλμό άρα κωδικοποιείται μόνο η τελευταία μαζί με τα πλάτη. Στην CELP χρησιμοποιείται, όπως αναφέρθηκε, διανυσματική κβάντιση.



(α)



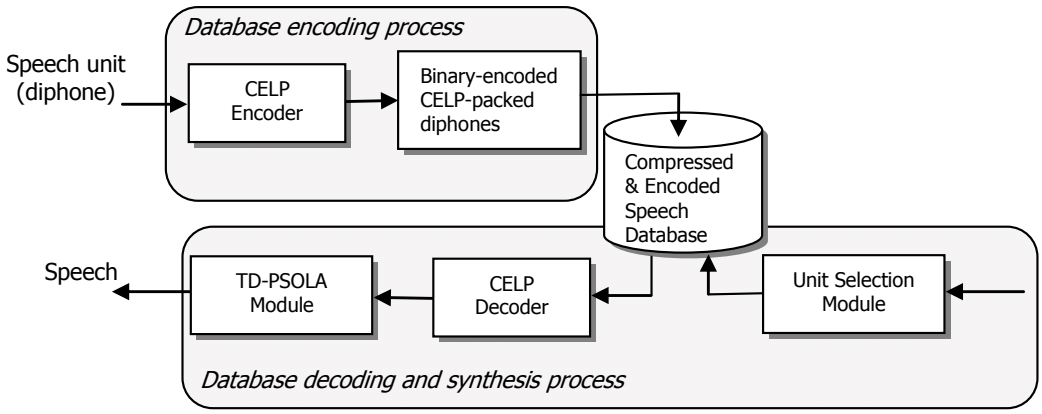
(β)

ΣΧΗΜΑ 3.5: Γενικό δομικό διάγραμμα της τεχνικής CELP: α) Κωδικοποιητής, β) αποκωδικοποιητής.

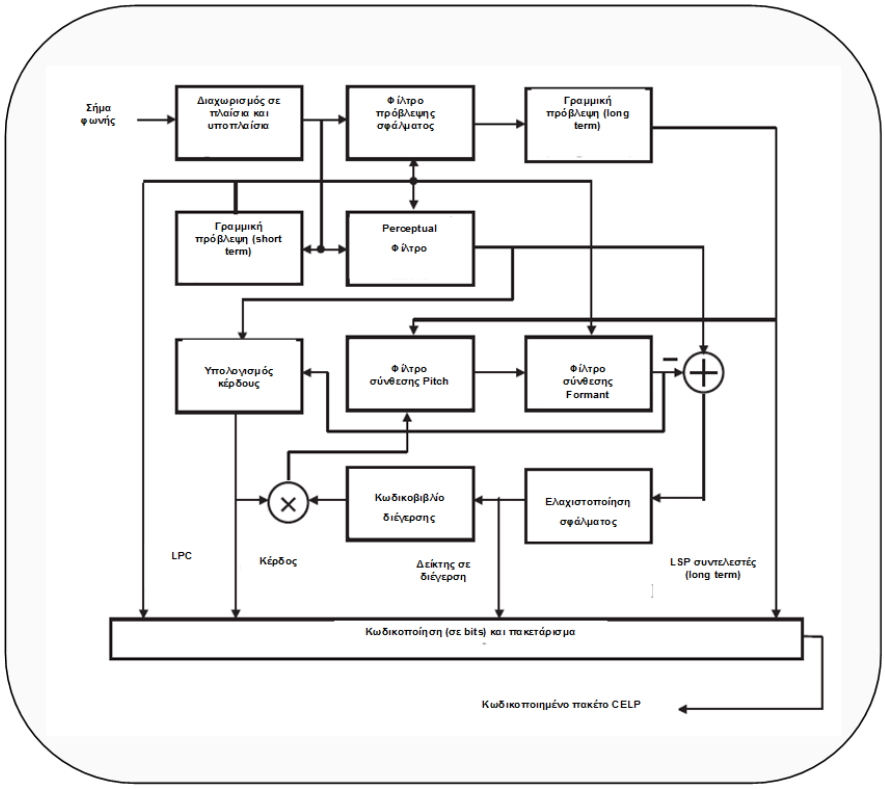
Στο σχήμα 3.6, απεικονίζεται η προσαρμογή της μεθόδου CELP για την περίπτωση σύνθεσης φωνής στο υπολογιστικό περιβάλλον του κινητού τηλεφώνου. Σύμφωνα με το σχήμα, η τυχαία πρόσβαση στις ακουστικές μονάδες (δίφωνα) επιτεύχθηκε μέσω της ξεχωριστής κωδικοποίησης κάθε ακουστικής μονάδας. Με αυτό τον τρόπο, η βάση δεδομένων αποτελείται από CELP παραμέτρους που αναπαριστούν πραγματώσεις διφώνων. Οι CELP παράμετροι κωδικοποιούνται με δυαδικό τρόπο (*binary encoding*) για αποτελεσματικότερη οργάνωση της βάσης. Είναι σημαντικό να παρατηρήσουμε ότι με αυτό τον τρόπο δεν προκύπτει αλλοίωση ούτε στα χρονικά όρια των διφώνων, ούτε στην θεμελιώδη συχνότητα (pitch) και στα χρονικά σημεία που την ορίζουν (*pitch marks*). Επιπλέον, τα σήματα δεν υπόκεινται σε ιδιαίτερη φασματική αλλοίωση (ή τουλάχιστον σε αλλοίωση που να είναι ακουστικά αντιληπτή) με αποτέλεσμα να μην επηρεάζονται τόσο οι ακουστικές μονάδες και η αναπαράστασή τους, όσο και η μετρική που καθορίζει την ύπαρξη φασματικών ασυνεχειών και χρησιμοποιείται ως κόστος στην βαθμίδα επιλογής ακουστικών μονάδων. Κατά την σύνθεση, μόνο οι ακουστικές μονάδες που επιλέχτηκαν ανασύρονται από την βάση και αποκωδικοποιούνται, οδηγώντας στην συνέχεια την βαθμίδα TD-PSOLA.

Το σχήμα 3.7, απεικονίζει το λεπτομερές δομικό διάγραμμα της CELP όπως εφαρμόστηκε και στην μηχανή σύνθεσης του κινητού τηλεφώνου. Η βασική διαδικασία πίσω από την κωδικοποίηση CELP έγκειται στα εξής:

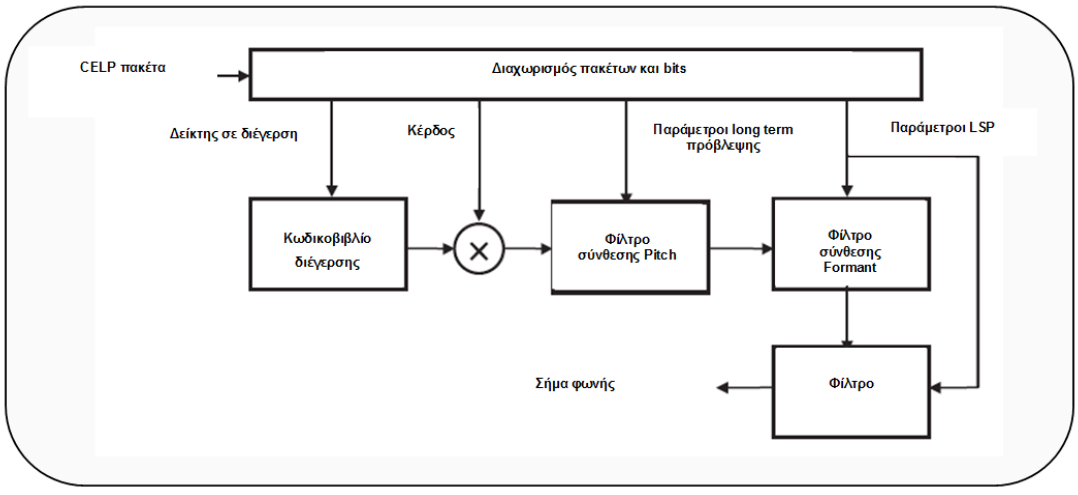
- Στην εφαρμογή της γραμμικής πρόβλεψης (*Linear Prediction Coding - LPC*) για την μοντελοποίηση του φωνητικού σωλήνα ή αλλιώς των συντονισμών του μέσω της εύρεσης των φασματικών κορυφών (*formants*).
- Στην χρήση κωδικό-βιβλίων (*codebooks*) τόσο προσαρμοσμένων (*adaptive*) όσο και στατικών, για την μοντελοποίηση της πηγής που θα χρησιμοποιηθούν ως είσοδος στο φίλτρο σύνθεσης της LPC.
- Στην βέλτιστη αναζήτηση των παραμέτρων της πηγής με ανάλυση μέσω σύνθεσης, για την ελαχιστοποίηση του τετραγωνικού σφάλματος μεταξύ αρχικού και κωδικοποιημένου σήματος.



ΣΧΗΜΑ 3.6: Διαδικασία συμπίεσης και κωδικοποίησης της βάσης δεδομένων ακουστικών μονάδων με προσαρμογή της μεθόδου CELP. Παράλληλα, απεικονίζεται και η διαδικασία αποκωδικοποίησης και σύνθεσης.



(α)



(β)

ΣΧΗΜΑ 3.7: Σύστημα κωδικοποίησης CELP για την συμπίεση της βάσης της μηχανής σύνθεσης φωνής για το κινητό τηλέφωνο: α) Κωδικοποιητής, β) Αποκωδικοποιητής.

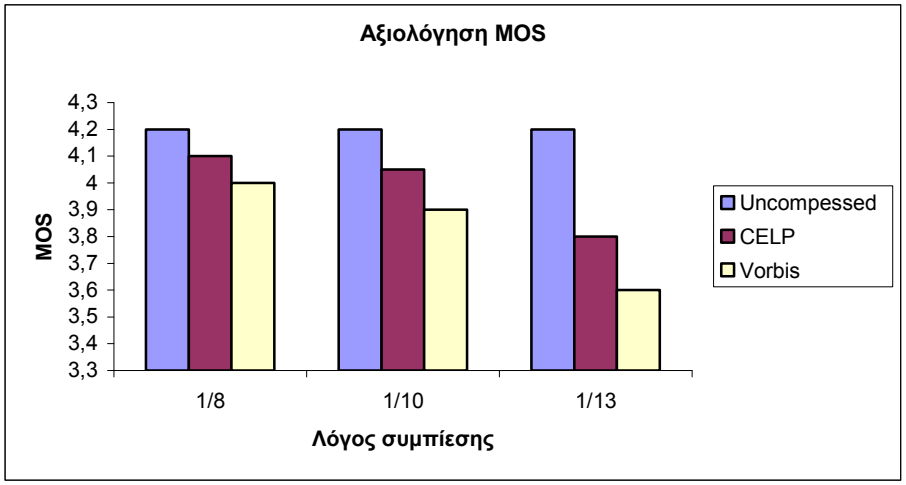
Η σύνθεση του σήματος εμπεριέχει και ψυχοακουστικά κριτήρια με την έννοια ότι η σύγκριση δεν δίνει έμφαση στις περιοχές που δεν θα ακουστούν από το ανθρώπινο αυτί. Στην περίπτωση μας η αναπαράσταση του φωνητικού σωλήνα επιτυγχάνεται με χρήση LSP (*Line Spectrum Pairs*) συντελεστών οι οποίοι είναι εύρωστοι στην κβάντιση [So, 2007]. Οι συντελεστές προκύπτουν με ανάλυση στο πλήρες πλαίσιο (όπως φαίνεται στο σχήμα). Κάθε υποπλαίσιο χρησιμεύει στον ακριβή προσδιορισμό των παραμέτρων που αφορούν την θεμελιώδη συχνότητα και την πηγή, καθώς και στην δημιουργία των κωδικό-βιβλίων. Σημειώνουμε ότι στην CELP τα περισσότερα bits δίνονται στο τελικό σήμα διέγερσης που ονομάζεται innovation καθώς δεν μπορεί να προκύψει από τεχνικές γραμμικής πρόβλεψης.

Η συχνότητα δειγματοληψίας της μηχανής σύνθεσης είναι στα 16KHz οπότε η μηχανή κωδικοποίησης λειτουργεί σε σήμα φωνής ευρείας ζώνης. Η διάρκεια ενός πλαισίου είναι 320 δείγματα για συχνότητα δειγματοληψίας 16KHz. Κάθε πλαίσιο χωρίζεται σε υποπλαίσια διάρκειας 40 δειγμάτων. Η ανάθεση των bits μπορεί να είναι μεταβλητή. Η τελική ανάθεση των bits ανά παράμετρο επηρεάζει τον τελικό βαθμό συμπίεσης αλλά και την τελική ποιότητα. Προφανώς, όσο λιγότερα bits χρησιμοποιούνται για την κωδικοποίηση των παραμέτρων, τόσο μειώνεται η ακρίβεια άρα και η ποιότητα. Στην περίπτωση μας, δοκιμάστηκαν διάφορες τιμές ανά παράμετρο ενώ και η σχεδίαση επιτρέπει την επιλογή διαφόρων τρόπων λειτουργίας, με αποτέλεσμα να υπάρχουν διάφορα επίπεδα συμπίεσης και τελικής ποιότητας. Ενδεικτικά, ο αριθμός των bits για τις LSP παραμέτρους κυμαίνεται από 18 έως 30 ανά πλαίσιο. Η θεμελιώδης συχνότητα κυμαίνεται σε 7 με 9 bits και τα κέρδη με 5 έως 7 bits ενώ τα κωδικο-βιβλία κωδικοποιούνται μέχρι και με 64 bits. Επιπλέον, υπάρχει ανάθεση από bits και για άλλες λειτουργίες όπως για παράδειγμα, αν το πλαίσιο ανήκει σε έμφωνο ή άφωνο ήχο κτλ. Η υλοποίηση της μηχανής κωδικοποίησης έγινε χρησιμοποιώντας αριθμητική σταθερής υποδιαστολής (*fixed point arithmetic*) η οποία είναι πολύ πιο γρήγορη σε σχέση με την αντίστοιχη κινητής υποδιαστολής (*floating point*) και στην περίπτωσή μας δεν φάνηκε να είχε αρνητική επίδραση στην τελική ποιότητα της συνθετικής ομιλίας.

3.3.2 Πειραματική αξιολόγηση και αποτελέσματα

Η κωδικοποίηση CELP, εφαρμόστηκε στην βάση δεδομένων για το κινητό τηλέφωνο που προέκυψε σύμφωνα με τη μέθοδο που περιγράφεται στην ενότητα 3.2. Το αρχικό μέγεθος της βάσης ήταν περίπου 40MByte. Η βάση που προκύπτει μετά από την κωδικοποίηση CELP είναι περίπου 4MByte. Επομένως, ο βαθμός συμπίεσης είναι περίπου 1/10. Ο τελικός βαθμός συμπίεσης προέκυψε κατόπιν διαφόρων δοκιμών και με γνώμονα την διατήρηση της τελικής ποιότητας της συνθετικής ομιλίας. Το τελικό μέγεθος της βάσης κρίνεται ιδιαίτερα ικανοποιητικό για σύστημα σύνθεσης φωνής που βασίζεται σε επιλογή και ένωση ακουστικών μονάδων. Όπως αναφέρθηκε, η επιτυχής εφαρμογή της κωδικοποίησης στη μηχανή σύνθεσης φωνής απαιτεί μειωμένη υπολογιστική πολυπλοκότητα στην αποκωδικοποίηση. Το αντίθετο δεν δημιουργεί πρόβλημα διότι πραγματοποιείται κατά την κατασκευή της βάσης. Το συνολικό υπολογιστικό φορτίο του αποκωδικοποιητή στην μηχανή σύνθεσης είναι της τάξης του 70% και, όπως θα φανεί και σε επόμενη ενότητα, δεν επιφέρει πρόβλημα για λειτουργία σε πραγματικό χρόνο και με χαμηλό χρόνο απόκρισης.

Το σχήμα 3.8 δείχνει τα αποτελέσματα της αξιολόγησης για τρεις διαφορετικούς βαθμούς συμπίεσης. Η επίδραση της συμπίεσης στην τελική ποιότητα του συνθετικού σήματος αξιολογήθηκε με την διεξαγωγή ακουστικών πειραμάτων που αφορούσαν την σύνθεση 52 προτάσεων. Η αξιολόγηση έγινε με βαθμολόγηση από μεικτή ομάδα 15 ανθρώπων με και χωρίς εμπειρία σε τεχνολογίες σύνθεσης φωνής από κείμενο. Οι προτάσεις ήταν γενικού περιεχομένου και δεν υπήρχαν στην βάση δεδομένων του συστήματος. Η ακουστική αξιολόγηση αφορούσε το mean opinion score (MOS) δηλαδή την βαθμολόγηση σε κλίμακα από 1 έως 5 (1: κακή ποιότητα, 5: άριστη ποιότητα) των προτάσεων. Η συγκριτική αξιολόγηση βασίστηκε στην σύνθεση των προτάσεων που πραγματοποιήθηκε α) δίχως συμπίεση, β) με συμπίεση CELP και γ) με συμπίεση Vorbis. Η παρουσίαση των προτάσεων στους ακροατές γινόταν ανά τριάδες (μία από κάθε μέθοδο) και με τυχαία σειρά χωρίς όμως να γνωστοποιείται η μέθοδος για κάθε πρόταση. Η αξιολόγηση στην περίπτωση που δεν εφαρμόζεται συμπίεση αξιολογήθηκε μία φορά ώστε να λειτουργήσει ως αναφορά για τις υπόλοιπες τεχνικές.



ΣΧΗΜΑ 3.8: Αποτελέσματα ακουστικών πειραμάτων για την αξιολόγηση της τεχνικής συμπίεσης της βάσης δεδομένων της μηχανής σύνθεσης.

Όπως φαίνεται από τα διαγράμματα, η τεχνική CELP επιτυγχάνει μέση τιμή βαθμολόγησης κοντά στο τέσσερα για βαθμό συμπίεσης της τάξης του 1/10 και γίνεται συγκρίσιμη, σε τελική ποιότητα, με την περίπτωση που δεν εφαρμόζεται καθόλου συμπίεση. Επιπλέον, για μεγαλύτερο λόγο συμπίεσης η υποβάθμιση της ποιότητας είναι αντιληπτή, όπως φαίνεται και από το διάγραμμα στο οποίο η βαθμολογία είναι στις 3,8 μονάδες. Η επιλογή βαθμού συμπίεσης μικρότερου από το 1/10 οδηγεί σε αύξηση της τελικής ποιότητας εις βάρος όμως της υπολογιστικής πολυπλοκότητας. Σε γενικές γραμμές, έχοντας ως αναφορά την ασυμπίεστη βάση των 40MB, μπορούμε να επιλέξουμε βαθμούς συμπίεσης από 1/5 έως και 1/10, με αποτέλεσμα η τελική βάση να κυμαίνεται από 8 έως 4MB αντίστοιχα. Επίσης, από το σχήμα, η κωδικοποίηση Vorbis φαίνεται να υστερεί στην τελική ποιότητα σε σχέση με την CELP για τον ίδιο βαθμό συμπίεσης.

3.4 ΥΠΟΛΟΓΙΣΜΟΣ ΤΟΥ ΚΟΣΤΟΥΣ ΦΑΣΜΑΤΙΚΗΣ ΑΣΥΝΕΧΕΙΑΣ

Στην ενότητα αυτή παρουσιάζεται η μέθοδος που αναπτύχθηκε για την μείωση του υπολογιστικού και αποθηκευτικού φόρτου που προκύπτει λόγω της βαθμίδας επιλογής ακουστικών μονάδων σε γενικού σκοπού συστήματα σύνθεσης φωνής από κείμενο με επιλογή και ένωση ακουστικών μονάδων [Hunt, 1996]. Η προσπάθεια είναι εξαιρετικά σημαντική για την επιτυχή προσαρμογή αυτών των συστημάτων σε περιβάλλοντα περιορισμένων υπολογιστικών δυνατοτήτων όπως είναι οι συσκευές των κινητών τηλεφώνων. Από την άλλη πλευρά, τέτοιου είδους τεχνικές πρέπει παράλληλα να διασφαλίζουν ότι δεν θα υποβαθμίσουν το σύστημα από άποψη τελικής ποιότητας τόσο σε φυσικότητα όσο και σε καταληπτότητα [Nukaga, 2006; Peters, 2004; Schnell, 2002]. Όπως έγινε φανερό σε προηγούμενη ενότητα, η βαθμίδα επιλογής ακουστικών μονάδων πραγματοποιεί πλήρη αναζήτηση στον χώρο των ακουστικών μονάδων με σκοπό να προσδιοριστεί η βέλτιστη (υπό την έννοια του ελάχιστου ολικού κόστους) ακολουθία ακουστικών μονάδων για την σύνθεση. Η αναζήτηση αυτή είναι αρκετά ακριβή σε υπολογιστικό κόστος και ουσιαστικά αποτελεί μια από τις πιο χρονοβόρες διαδικασίες της μηχανής σύνθεσης. Συνεπώς, η υπολογιστική απόδοση του αλγόριθμου επιλογής είναι ζωτικής σημασίας καθώς καθορίζει την ικανότητα λειτουργίας του συστήματος σύνθεσης αφενός σε πραγματικό χρόνο και αφετέρου με μικρό χρόνο απόκρισης.

Αναλυτικότερα, η εφαρμογή μηχανών σύνθεσης φωνής από κείμενο με επιλογή ακουστικών μονάδων σε περιβάλλοντα μειωμένων υπολογιστικών δυνατοτήτων προβάλλει ιδιαίτερες απαιτήσεις τόσο σε υπολογιστική ισχύ όσο και σε αποθηκευτική δυνατότητα. Ο δεύτερος παράγοντας καλύπτεται αποτελεσματικά με χρήση τεχνικών όπως αυτές που παρουσιάστηκαν στις ενότητες 3.2 και 3.3. Ο πρώτος παράγοντας σχετίζεται άμεσα με την υπολογιστική δυνατότητα διεκπεραίωσης όλων των απαραίτητων λειτουργιών της μηχανής σύνθεσης, για λειτουργία σε συνθήκες πραγματικού χρόνου ή και ακόμα λιγότερο [Schnell, 2002]. Η βαθμίδα επιλογής ακουστικών μονάδων σε σύγχρονές μηχανές σύνθεσης με βάσεις δεδομένων της τάξης των ~Gb, με πολύ μεγάλο αριθμό πραγματώσεων για κάθε ακουστική μονάδα, απαιτεί μεγάλη υπολογιστική ισχύ για λειτουργία πραγματικού χρόνου και με μικρό χρόνο απόκρισης. Στα συστήματα σύνθεσης φωνής σε προσωπικούς Η/Υ ή σε εξυπηρετητές, αυτό αντιμετωπίζεται όχι μόνο λόγω της υπάρχουσας υπολογιστικής ισχύος αλλά τόσο και με την χρήση τεχνικών περικοπής ακουστικών μονάδων κατά την αναζήτηση (*pruning*) όσο και με τεχνικές συσταδοποίησης μεταξύ παρόμοιων ακουστικών μονάδων [Founda, 2001; Beutnagel, 1999; Black, 1997]. Αυτές οι τεχνικές είναι ακατάλληλες στο περιβάλλον της συσκευής του κινητού τηλεφώνου καθώς βασίζονται στην μεγάλη ποικιλία σε ακουστικές μονάδες που θα απομείνουν. Η υπόθεση αυτή δεν υφίσταται στο κινητό τηλέφωνο διότι οι βάσεις δεδομένων είναι ήδη μειωμένες και κατ' επέκταση ο χώρος αναζήτησης είναι ήδη συρρικνωμένος.

Η τεχνική που παρουσιάζεται, βασίζεται στην συσταδοποίηση (*clustering*) μεταξύ των διανυσμάτων που αποτελούν το κόστος ένωσης και ειδικότερα εκείνων που απαρτίζουν το κόστος που αφορά τις φασματικές ασυνέχειες. Έτσι, είναι εφικτός ο offline υπολογισμός των αποστάσεων μεταξύ των κέντρων από κάθε συστάδα (*cluster*) μειώνοντας έτσι, τόσο το υπολογιστικό φορτίο όσο και τις απαιτήσεις σε αποθηκευτικούς πόρους. Η συσταδοποίηση (ή διαφορετικά η

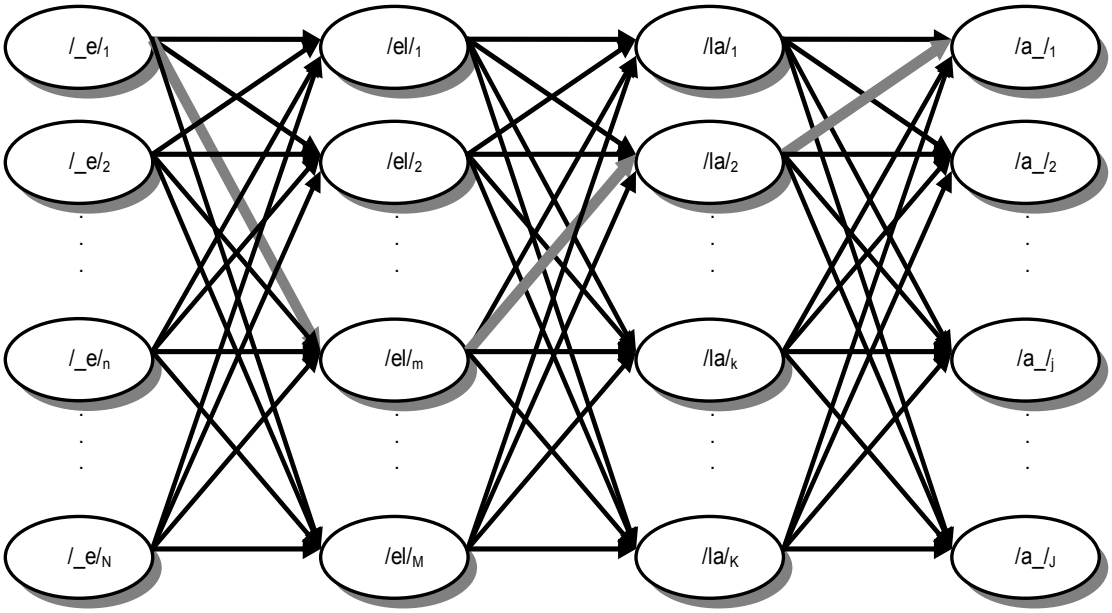
διανυσματική κβάντιση) δεν εφαρμόζεται στις ακουστικές μονάδες αλλά στο διάνυσμα χαρακτηριστικών για την φασματική συνέχεια για κάθε φώνημα. Με αυτή την προσέγγιση διατηρείται ο αρχικός χώρος αναζήτησης ενώ μειώνεται η διακριτική ικανότητα του κόστους που χαρακτηρίζει την ύπαρξη ασυνέχειας. Τα πειραματικά αποτελέσματα αναδεικνύουν ότι αυτή η μείωση δεν έχει συνέπεια στην τελική ποιότητα της συνθετικής ομιλίας ενώ βελτιώνει σημαντικά το υπολογιστικό κόστος [Karabetsos, 2009]. Παρόμοια προσέγγιση έχει προταθεί και στο [Coorman, 2000] για την περίπτωση όμως ευρείας κλίμακας συστημάτων (συστήματα σύνθεσης φωνής σε προσωπικούς Η/Υ ή/και σε εξυπηρετητές), χωρίς ωστόσο να εξετάζεται η περίπτωση τόσο των ιδιαίτερων χαρακτηριστικών προσαρμογής (μειωμένη βάση κτλ.) όσο και η υλοποίηση και η επίπτωση, σε περιβάλλον μειωμένων υπολογιστικών πόρων.

3.4.1 Αλγόριθμος ελαχιστοποίησης υπολογιστικών αναγκών με χρήση διανυσματικής κβάντισης

Η βαθμίδα επιλογής ακουστικών μονάδων αποτελεί τον μηχανισμό που θα καθορίσει την βέλτιστη ακολουθία διφώνων που θα αποτελέσουν το τελικό συνθετικό σήμα. Όπως αναφέρθηκε σε προηγούμενη ενότητα (Κεφ. 2), η ακολουθία προκύπτει μέσω αναζήτησης που βασίζεται σε συναρτήσεις κόστους που συνυπολογίζουν διάφορα κριτήρια. Ένα από τα πιο σημαντικά κριτήρια είναι αυτό που αφορά τον χαρακτηρισμό πιθανής φασματικής ασυνέχειας. Για τον υπολογισμό του, γίνεται ανάκτηση του διανύσματος αναπαράστασης του φάσματος για κάθε φώνημα και κατόπιν υπολογίζεται κάποια απόσταση (συνήθως η Ευκλείδια απόσταση μεταξύ MFCC συντελεστών) (βλ. Κεφ. 2) [Davis, 1980]. Για παράδειγμα, το σχήμα 3.9, απεικονίζει την σύνθεση της λέξης «*έλα*», η οποία έστω ότι αποτελείται από τα δίφωνα {/_e/,/el/,/la/,/a_/} και έστω ότι είναι διαθέσιμες N, M, K και J πραγματώσεις αντίστοιχα για το καθένα δίφωνω. Ο γράφος που προκύπτει απαιτεί τον υπολογισμό N·M+M·K+K·J αποστάσεων που συμβολίζονται ως C^c. Γίνεται λοιπόν αντιληπτό ότι στην σύνθεση σε κινητό τηλέφωνο, χρειάζεται ιδιαίτερη (υπολογιστική) αντιμετώπιση ο χειρισμός του κόστους ένωσης και ιδιαίτερα του κόστους που αφορά τις φασματικές ασυνέχειες.

Αναλυτικότερα, αν ορίσουμε ως P = {d} το σύνολο των φωνημάτων με |P| = N συνολικά στοιχεία και ως D = {*pq : p,q* $\in$ P και *pq* είναι επιτρεπτό} το σύνολο των διφώνων με |D| = M στοιχεία τότε θα ισχύει $N^2 \geq$ |D| = M $\sim$ O(N^2). Επιπλέον, η βάση δεδομένων με τις πραγματώσεις διφώνων ορίζει το σύνολο R = { u_k^{pq} : η $k_{\text{-οστή}}$ πραγμάτωση του *pq* $\in$ D }. Αν θεωρήσουμε ως K το μέσο αριθμό πραγματώσεων ανά δίφωνο, τότε κάποιο φώνημα μπορεί να σχηματίσει έως N δίφωνα σαν το αριστερό μέρος ενός διφώνου και έως N δίφωνα σαν το δεξί μέρος ενός διφώνου, και καθένα δίφωνο θα έχει αντίστοιχα K πραγματώσεις. Για παράδειγμα, το φώνημα /a/ σχηματίζει τα σετ διφώνων /aX/ και /Xa/, όπου X $\in$ P (δηλαδή το X μπορεί να είναι οποιοδήποτε φώνημα). Αφού τα δίφωνα του τύπου /Xa/ ενώνονται μόνο με δίφωνα του τύπου /aX/, και από την στιγμή που υπάρχουν K πραγματώσεις για το καθένα από αυτά, οι δυνατές ενώσεις για κάποιο φώνημα (μέσα στο σετ των διφώνων) είναι της τάξης N^2K^2 . Οπότε, αφού το σύνολο (σετ) των φωνημάτων έχει

Ν στοιχεία, τότε οι πιθανές ενώσεις στο R είναι της τάξης $O(N^3K^2)$. Επίσης, για κάθε $u \in R$ χρειάζονται δύο ζεύγη διανύσματος χαρακτηριστικών, ένα για το αριστερό φώνημα $v_L(u)$ (αριστερή ένωση) και ένα για το δεξί $v_R(u)$ (δεξιά ένωση), και μια μετρική $d(v_R(u_i^{aq}), v_L(u_j^{qb}))$ ή $d(v_R(u_i), v_L(u_j))$ ανάμεσα σε δύο δίφωνα που θα χρησιμεύσει ως συνάρτηση κόστους ένωσης για τις φασματικές ασυνέχειες.



$$C^c(/\_e/_n, /el/_m) \qquad C^c(/el/_m, /la/_k) \qquad C^c(/la/_k, /a\_/_j)$$
$$n=1...N,\ m=1...M \qquad m=1...M,\ k=1...K \qquad k=1...K,\ j=1...J$$

ΣΧΗΜΑ 3.9: Υπολογισμός του κόστους ένωσης στην βαθμίδα επιλογής ακουστικών μονάδων. Η βέλτιστη ακολουθία διφώνων αποτελεί το καλύτερο μονοπάτι του γράφου και προκύπτει από το ελάχιστο (συσσωρευτικό) κόστος ανάμεσα στα υποψήφια δίφωνα που αποτελούν τους κόμβους του γράφου. Το υπολογιστικό φορτίο αυξάνει όσο αυξάνει ο αριθμός των πραγματώσεων για κάθε δίφωνο. Σαν παράδειγμα, το καλύτερο μονοπάτι φαίνεται από την γκρίζα γραμμή του σχήματος.

Συνεπώς, προκύπτουν δύο δυνατότητες για την αποθήκευση και τον υπολογισμό. Είτε να αποθηκευτούν τα διανύσματα $v_L(u)$, $v_R(u)$ για κάθε $u \in R$ και ο υπολογισμός της απόστασης *d* να γίνεται κατά την ώρα της σύνθεσης, είτε να υπολογιστεί κάθε απόσταση από πριν (offline) και να αποθηκευτεί, υπό μορφή πίνακα, για κάθε πιθανή ένωση στο *R*. Η πρώτη περίπτωση οδηγεί σε υψηλό υπολογιστικό κόστος την ώρα της σύνθεσης ενώ η δεύτερη οδηγεί σε μεγαλύτερες αποθηκευτικές απαιτήσεις.

Για να μειωθούν οι προηγούμενες απαιτήσεις προτείνεται η offline συσταδοποίηση για το ίδιο φώνημα των διανυσμάτων και ο offline υπολογισμός και αποθήκευση των αποστάσεων ανάμεσα στα κέντρα καθεμιάς συστάδας (cluster). Η τεχνική οδηγεί σε σημαντική μείωση των υπολογιστικών και αποθηκευτικών απαιτήσεων διότι μειώνει το πλήθος των απαιτούμενων υπολογισμών του κόστους ένωσης, ενέχει όμως τον κίνδυνο πιθανώς να οδηγήσει σε χαμηλότερης ποιότητας

σύνθεση λόγω της μείωσης της διακριτικής ικανότητας του κόστους ένωσης για τις φασματικές ασυνέχειες.

Η αλγοριθμική οργάνωση της τεχνικής φαίνεται στον πίνακα 3.3. Από τον αλγόριθμο προκύπτει ότι, αν C είναι ο αριθμός συστάδων ανά φώνημα, ο ολικός αριθμός των πιθανών ενώσεων είναι της τάξης $O(NC^2)$ και παράλληλα ισχύει $O(NC^2)$ << $O(N^3K^2)$ με την προϋπόθεση ότι ο αριθμός των cluster είναι επαρκώς μικρός. Επιπρόσθετα, πρέπει να σημειωθεί το πλήθος των ενώσεων ανά δίφωνο είναι C^2 με χρήση συσταδοποίησης, ενώ στην περίπτωση που δεν χρησιμοποιείται συσταδοποίηση το πλήθος είναι NK^2 . Συνεπώς, η έκφραση $C^2 < NK^2$ είναι αληθής στην περίπτωση των ενσωματωμένων συστημάτων, εφόσον ο αριθμός των πραγματώσεων ανά δίφωνο είναι επαρκής και το C επιλεγεί επαρκώς μικρό. Για παράδειγμα, στο σύστημα σύνθεσης φωνής που υλοποιήθηκε, το σετ φωνημάτων για την ελληνική γλώσσα χρησιμοποιεί N = 34 στοιχεία και ο αριθμός των πραγματώσεων ανά δίφωνο είναι τουλάχιστον 10. Οπότε, αν επιλεγεί C = 32 (*cluster size*) το παραπάνω κριτήριο αληθεύει.

ΠΙΝΑΚΑΣ 3.3: Αλγόριθμος συσταδοποίησης και υπολογισμού του κόστους ένωσης φασματικών ασυνεχειών για το υπολογιστικό περιβάλλον της συσκευής του κινητού τηλεφώνου (offline).

1. $\forall p \in P$ επανέλαβε τα βήματα 2 έως 5
2. Βρες όλες τις πραγματώσεις των διφώνων που έχουν το *p* σαν αριστερό ή δεξί φώνημα δηλ., βρες R^p_{left} = $\{u^{lr} : u \in$ R και *p=l*$\}$ και R^p_{right} = $\{u^{lr} : u \in$ R και *p=r*$\}$
3. Πραγματοποίησε clustering των $\{v_L(u): u \in R^p_{left}\} \cup \{v_R(u): u \in R^p_{right}\}$ σε *C* clusters με κέντρα c_i όπου *i = 1…C*
4. Υπολογισμός των $M^p(i, j) = d(c_i, c_j)$, i,j = 1…C
5. Αποθήκευση των $M^p(i, j)$ καθώς και των δύο δεικτών σε cluster για κάθε δίφωνο
6. Τερματισμός

Κατά την σύνθεση, η απόσταση ανάμεσα σε ζεύγη διφώνων ανακτάται και υπολογίζεται ως *M(a, b)* αντί $d(v_R(u_i^{yp}), v_L(u_j^{px}))$, όπου a, b είναι οι αντίστοιχοι δείκτες σε συστάδα στο οποίο ανήκει κάθε φώνημα από κάθε δίφωνο. Η μέθοδος δεν μειώνει τον χώρο αναζήτησης καθώς είναι σαφές ότι δεν μπορεί να εφαρμοστεί συσταδοποίηση στις ακουστικές μονάδες αφού η βάση δεδομένων έχει ήδη μειωθεί (βλ. ενότητα 3.2). Αντίθετα, ο χώρος αναζήτησης παραμένει ίδιος και η συσταδοποίηση πραγματοποιείται στα χαρακτηριστικά (features) που συντελούν στον υπολογισμό του κόστους ένωσης που αφορά τις φασματικές ασυνέχειες, με τίμημα την μείωση της διακριτικής ικανότητας στην περίπτωση αυτή. Σημειώνεται ότι η μέθοδος βασίζεται στην υπόθεση ότι τα κόστη που ανήκουν στην ίδια συστάδα, έχουν αφενός μικρή διαφορά μεταξύ τους (και ότι αυτή η διαφορά δεν είναι αντιληπτή και σε ακουστικό επίπεδο) και αφετέρου ότι η μείωση της

διακριτικής ικανότητας αντισταθμίζεται τόσο από το κόστος «στόχος» που ορίζει την επιθυμητή ακολουθία ακουστικών μονάδων, όσο και από τα υπόλοιπα κόστη που εμπλέκονται στον υπολογισμό του ολικού κόστους ένωσης.

3.4.2 Πειραματική αξιολόγηση και αποτελέσματα

Ο αλγόριθμος του πίνακα 3.3 υλοποιήθηκε και αξιολογήθηκε στη μηχανή σύνθεσης φωνής από κείμενο για το κινητό τηλέφωνο. Η μηχανή σύνθεσης αποτελείται από βάση δεδομένων φυσικής ομιλίας της τάξης των 11000 διφώνων στα 16KHz. Τα διανύσματα χαρακτηριστικών αποτελούν συντελεστές MFCC (Mel-Frequency Cepstral Coefficients) [Davis, 1980] και ο αριθμός των συστάδων (*cluster*) επιλέχθηκε C = 32. Η συσταδοποίηση πραγματοποιήθηκε με τον αλγόριθμο *k-means* χρησιμοποιώντας την Ευκλείδεια απόσταση [Rabiner, 1993]. Η πειραματική αξιολόγηση της τεχνικής βασίζεται τόσο σε αντικειμενικά (υπολογιστικές επιδόσεις) όσο και σε υποκειμενικά (ακουστικά πειράματα) κριτήρια.

Ο πίνακας 3.4 που ακολουθεί, δείχνει τις μέσες υπολογιστικές επιδόσεις που επιτυγχάνονται από την βαθμίδα επιλογής ακουστικών μονάδων με χρήση του αλγόριθμου, για τη περίπτωση σύνθεσης 52 προτάσεων μέσου μήκους. Οι επιδόσεις δείχνουν την βελτίωση που προκύπτει σε σύγκριση με την βαθμίδα επιλογής που δεν χρησιμοποιεί την προτεινόμενη τεχνική. Σημειώνεται ότι με C_{US} (***C_{US} – clustered join cost unit selection***) και με F_{US} (***F_{US} – full unit selection***), συμβολίζεται η βαθμίδα επιλογής με χρήση ή χωρίς της προτεινόμενης τεχνικής, αντίστοιχα.

ΠΙΝΑΚΑΣ 3.4: Μέσες υπολογιστικές επιδόσεις της βαθμίδας επιλογής ακουστικών μονάδων με την μέθοδο C_{US}.

Βελτίωση ταχύτητας της βαθμίδας επιλογής ακουστικών μονάδων (*Unit selection speed improvement*)	> 3.5 times
Συνολική βελτίωση ταχύτητας (*Total speed improvement*)	> 29%
Μέσος χρόνος απόκρισης (*Mean response time*)	0.25sec
Λειτουργία σε πραγματικό χρόνο (Real time factor)	> 2.4
Ποσοστό επί του συνολικού χρόνου (*Percentage of total TTS time*)	F_{US}: 32.5% **C_{US}: 13.1%**

Τα αποτελέσματα αναδεικνύουν την σημαντική βελτίωση στις υπολογιστικές επιδόσεις τόσο της βαθμίδας επιλογής ακουστικών μονάδων όσο και συνολικά στην μηχανή σύνθεσης, με χρήση της τεχνικής C_{US}. Πράγματι, το υπολογιστικό φορτίο της βαθμίδας μειώνεται παραπάνω από τρεις φορές ενώ η μηχανή σύνθεσης βελτιώνει τις υπολογιστικές τις επιδόσεις κατά 29% σε σχέση με πριν. Επιπλέον, η μηχανή

σύνθεσης λειτουργεί με περισσότερο από 2,4 φορές σε πραγματικό χρόνο στο υπολογιστικό περιβάλλον που δοκιμάστηκε. Επιπλέον, η βαθμίδα επιλογής ακουστικών μονάδων αντιστοιχεί πλέον στο 13,1% του υπολογιστικού φόρτου κατά τη σύνθεση, έναντι του 32,5% χωρίς την εν λόγω τεχνική. Τέλος, η συνολική βελτίωση μειώνει σημαντικά τον χρόνο απόκρισης της μηχανής σύνθεσης. Σημειώνεται ότι ο χρόνος απόκρισης είναι ο χρόνος που χρειάζεται μέχρι να ξεκινήσει η σύνθεση από την στιγμή που έχει δοθεί το κείμενο εισόδου και αποτελεί σημαντικό παράγοντα στην ευρύτερη αποδοχή του συστήματος σύνθεσης και στην εφαρμογή του σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και υπηρεσίες.

Η επίδραση της τεχνικής στην τελική ποιότητα του συνθετικού σήματος αξιολογήθηκε και με υποκειμενικά κριτήρια, με την διεξαγωγή ακουστικών πειραμάτων που αφορούσαν την σύνθεση 52 προτάσεων τόσο με χρήση $\mathbf{F_{US}}$ όσο και με χρήση $\mathbf{C_{US}}$, και την βαθμολόγηση τους από μεικτή ομάδα 15 ανθρώπων με και χωρίς εμπειρία σε τεχνολογίες σύνθεσης φωνής από κείμενο. Οι προτάσεις ήταν γενικού περιεχομένου και δεν υπήρχαν στην βάση δεδομένων του συστήματος. Η ακουστική αξιολόγηση πραγματοποιήθηκε με το κριτήριο της μέσης ακουστικής αξιολόγησης (mean opinion score-MOS) δηλαδή την βαθμολόγηση σε κλίμακα από 1 έως 5 (1: κακή ποιότητα, 5: άριστη ποιότητα) των προτάσεων [Speechworks, 2006]. Η παρουσίαση των προτάσεων στους ακροατές γινόταν ανά ζεύγη (μία από κάθε μέθοδο) και με τυχαία σειρά χωρίς όμως να γνωστοποιείται η μέθοδος για κάθε πρόταση. Τα αποτελέσματα των μέσων τιμών των βαθμολογήσεων και η τυπική τους απόκλιση φαίνονται στον επόμενο πίνακα.

ΠΙΝΑΚΑΣ 3.5: Ακουστική αξιολόγηση της τεχνικής C_{us}

	Μέση Ακουστική Αξιολόγηση (MOS)	**Τυπική απόκλιση**
$\mathbf{F_{US}}$	3.98	0.45
$\mathbf{C_{US}}$	4.01	0.39

Τα αποτελέσματα δείχνουν ότι η C_{US} πετυχαίνει καλύτερα αποτελέσματα σε συνολική ποιότητα σε σχέση με την F_{US}. Από τις τιμές της τυπικής απόκλισης όμως γίνεται φανερό ότι, από την άποψη της τελικής ποιότητας, δεν προκύπτουν σημαντικές διαφορές και ότι τα αποτελέσματα είναι ισοδύναμα. Οπότε το συμπέρασμα που προκύπτει είναι ότι η C_{US} επιτυγχάνει τελική ποιότητα συνθετικής φωνής, πρακτικά μη διαχωρίσιμης σε σχέση με την F_{US}, αλλά με σημαντικό υπολογιστικό κέρδος. Να σημειώσουμε ότι οι σχετικά υψηλές MOS οφείλονται στο γεγονός ότι οι συμμετέχοντες στα ακουστικά πειράματα γνώριζαν ότι πρόκειται για συνθετική φωνή σε περιβάλλον κινητού τηλεφώνου. Επιπλέον, οι τιμές MOS δεν μπορούν να συγκριθούν με αντίστοιχες άλλων συστημάτων παρά μόνο μεταξύ τους.

Επιπρόσθετα, σημειώνεται ότι η επιλογή του αριθμού των συστάδων, επιφέρει κάποια ισορροπία (tradeoff) μεταξύ του χρόνου επεξεργασίας, του κόστους αποθήκευσης και της υποβάθμισης της διακριτικής ικανότητας του φασματικού κόστους. Πράγματι, όσο αυξάνει ο αριθμός των συστάδων ανά φώνημα, τόσο

αυξάνει και το κόστος αποθήκευσης. Από την άλλη πλευρά, μικρός αριθμός από συστάδες σημαίνει ότι πολλά φωνήματα αντιπροσωπεύονται από ένα διάνυσμα χαρακτηριστικών με αποτέλεσμα η βαθμίδα επιλογής να μην επιτυγχάνει σωστή αξιολόγηση των ακουστικών ασυνεχειών. Τέλος, πέρα από την συσκευή του κινητού τηλεφώνου, ο αλγόριθμος μπορεί να εφαρμοστεί και σε συστήματα σύνθεσης φωνής σε προσωπικούς Η/Υ ή εξυπηρετητές, καθώς ο μειωμένος χρόνος απόκρισης προσφέρει τη δυνατότητα ταυτόχρονης εξυπηρέτησης περισσότερων παράλληλων αιτήσεων για σύνθεση. Η λειτουργικότητα αυτή είναι ιδιαίτερα σημαντική σε αυτόματες τηλεφωνικές υπηρεσίες με σύνθεση φωνής.

3.5 ΣΥΝΟΛΙΚΗ ΑΠΟΤΙΜΗΣΗ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΕΠΙΔΟΣΕΩΝ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Ο πίνακας που ακολουθεί συγκεντρώνει τις υπολογιστικές επιδόσεις (benchmarks) της μηχανής σύνθεσης από κείμενο για το περιβάλλον της συσκευής του κινητού τηλεφώνου, μετά από την ολοκλήρωση των τεχνικών που περιγράφηκαν στις προηγούμενες ενότητες.

ΠΙΝΑΚΑΣ 3.6: Υπολογιστικές επιδόσεις της μηχανής σύνθεσης φωνής σε περιβάλλον κινητού τηλεφώνου

ΒΑΘΜΙΔΑ	**ΠΟΣΟΣΤΟ ΕΠΙ ΤΟΥ ΟΛΙΚΟΥ ΧΡΟΝΟΥ ΣΥΝΘΕΣΗΣ**
Γλωσσική επεξεργασία (NLP)	3%
Αποκωδικοποίηση CELP	69%
Επιλογή Ακουστικών Μονάδων	13%
TD-PSOLA	15%
ΓΕΝΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ	
Μέγεθος Βάσης	4-8MB
Συντελεστής Πραγματικού Χρόνου	$\geq$ 2,5
Χρόνος Απόκρισης	0.25sec

**Ο χρόνος απόκρισης καθορίζεται κυρίως από την βαθμίδα επιλογής ακουστικών μονάδων και δευτερευόντως από τις βαθμίδες αποκωδικοποίησης CELP και TD-PSOLA.*

Από τον πίνακα προκύπτει ότι η προσαρμοσμένη μηχανή σύνθεσης λειτουργεί σε πραγματικό χρόνο και επιτυγχάνει χαμηλό χρόνο απόκρισης. Επιπλέον, ο αρχικός στόχος της αποδοτικής αποκλιμάκωσης χωρίς σημαντική επίπτωση στην τελική ποιότητα έχει επιτευχθεί. Τα χαρακτηριστικά αυτά είναι ιδιαίτερα σημαντικά, καθώς η επίτευξή τους συντελεί στην τελική αποδοχή του συνθέτη τόσο από άποψη ποιότητας όσο και λειτουργικότητας. Με αυτό τον τρόπο, επιτυγχάνεται συνεχής ροή συνθετικής ομιλίας ανεξάρτητα από τον όγκο του κειμένου που πρόκειται να διαβαστεί, ενώ δεν προκύπτουν διακοπές και μεγάλος χρόνος αναμονής μέχρι να ξεκινήσει η σύνθεση. Οι δοκιμές έγιναν σε τυπική συσκευή κινητού τηλεφώνου με επεξεργαστή ταχύτητας 220MHz.

Συνοψίζοντας, στην ενότητα αυτή περιγράψαμε τη γενική αρχιτεκτονική ενός γενικού σκοπού συστήματος σύνθεσης φωνής από κείμενο με επιλογή και συρραφή ακουστικών μονάδων για ενσωματωμένα συστήματα. Επιπλέον, εξετάσαμε τρείς

τεχνικές που συντελούν στην αποδοτική αποκλιμάκωση και προσαρμογή αυτού του είδους τεχνολογίας σύνθεσης σε αυτά υπολογιστικά περιβάλλοντα χωρίς σημαντική επίπτωση στην τελική ποιότητα. Οι τεχνικές αυτές προσεγγίζουν διαφορετικές σχεδιαστικές πτυχές του συστήματος όπως είναι η δημιουργία της βάσης δεδομένων, η συμπίεση και διαχείριση της βάσης δεδομένων και η μείωση των απαιτούμενων υπολογιστικών αναγκών που προβάλλει ο αλγόριθμος επιλογής ακουστικών μονάδων.

Πιο συγκεκριμένα, ο αλγόριθμος δημιουργίας της βάσης δεδομένων φυσικής ομιλίας στηρίζεται στη συμπεριφορά του αλγόριθμου επιλογής και έχει ως χαρακτηριστικό ότι οδηγεί σε μειωμένη βάση δεδομένων που μετριάζει τον πλεονασμό σε ακουστικές μονάδες, διατηρώντας ταυτόχρονα την απαραίτητη ποικιλία από αυτές. Τα χαρακτηριστικά αυτά αποτελούν σημαντικό ζητούμενο για γενικού σκοπού συστήματα σύνθεσης φωνής. Για την συμπίεση και διαχείριση της βάσης δεδομένων, υιοθετήθηκε και προσαρμόστηκε κατάλληλα στη διαδικασία σύνθεσης η μέθοδος CELP. Τέλος, το υπολογιστικό φορτίο που επιφέρει ο υπολογισμός του κόστους φασματικών ασυνεχειών μειώθηκε σημαντικά με εφαρμογή διανυσματικής κβάντισης στο διάνυσμα των φασματικών χαρακτηριστικών και τον υπολογισμό του κόστους μεταξύ των κέντρων των συστάδων.

Η πειραματική αξιολόγηση, τόσο με αντικειμενικά όσο και με υποκειμενικά κριτήρια, ανέδειξε την αποδοτική αποκλιμάκωση και προσαρμογή του συστήματος σε περιβάλλον κινητού τηλεφώνου, την σημαντική βελτίωση στην αξιοποίηση των μειωμένων υπολογιστικών πόρων, διατηρώντας παράλληλα υψηλή τελική ποιότητα, τόσο σε φυσικότητα όσο και σε καταληπτότητα.

ΚΕΦΑΛΑΙΟ
-4-
ΠΑΡΑΜΕΤΡΙΚΗ ΣΥΝΘΕΣΗ ΦΩΝΗΣ

ΚΕΦΑΛΑΙΟ 4 – ΠΑΡΑΜΕΤΡΙΚΗ ΣΥΝΘΕΣΗ ΦΩΝΗΣ

Στο κεφάλαιο αυτό εξετάζεται η τεχνολογία της παραμετρικής σύνθεσης φωνής από κείμενο και συγκεκριμένα το μεθοδολογικό πλαίσιο της σύνθεσης φωνής από κείμενο με χρήση κρυφών Μαρκοβιανών μοντέλων. Αρχικά πραγματοποιείται μια εισαγωγή στην τεχνολογία αυτή περιγράφοντας τα βασικά της χαρακτηριστικά. Στη συνέχεια, αναλύεται η προσαρμογή, η υλοποίηση και η αξιολόγηση της τεχνολογίας για την περίπτωση της Ελληνικής γλώσσας. Συγκεκριμένα, περιγράφεται η υλοποίηση ενός συστήματος σύνθεσης φωνής για τα Ελληνικά με χρήση κρυφών Μαρκοβιανών μοντέλων (σύστημα HMM), δίνοντας έμφαση στις σχεδιαστικές παραμέτρους που χαρακτηρίζουν τις ιδιαιτερότητες της ελληνικής γλώσσας. Κατά τη πειραματική αξιολόγηση, πραγματοποιείται σύγκριση τόσο με το αντίστοιχο σύστημα με επιλογή και συρραφή ακουστικών μονάδων (unit selection) όσο και με ένα σύστημα σύνθεσης με διφωνήματα (diphone synthesis). Τα αποτελέσματα δείχνουν πως αν και το σύστημα με HMM υπολείπεται σε τελική ποιότητα σε σχέση με το αντίστοιχο με επιλογή και συρραφή ακουστικών μονάδων, ωστόσο επιτυγχάνει υψηλή τελική ποιότητα, γεγονός ενθαρρυντικό καθώς πρόκειται για μια πρώτη προσπάθεια υλοποίησης συστήματος με την εν λόγω τεχνολογία.

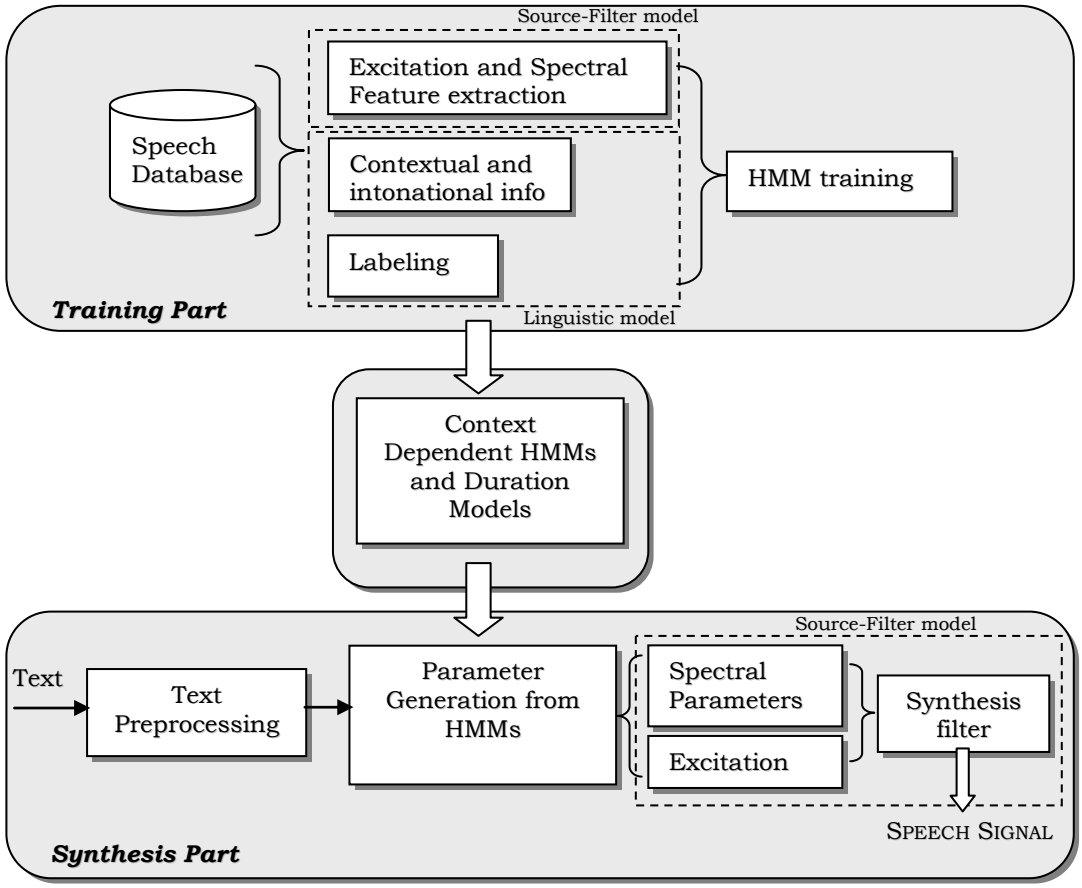
4.1 ΕΙΣΑΓΩΓΗ

Όπως αναφέρθηκε και στο πρώτο κεφάλαιο, η παραμετρική σύνθεση φωνής προσφέρει σημαντικά πλεονεκτήματα κυρίως λόγω της δυνατότητας που προσφέρει για εύκολη και ευέλικτη δημιουργία, τροποποίηση και διαχείριση, πλήθους ποιοτικών συνθετικών φωνών τόσο σε διαφορετικές γλώσσες όσο και σε διαφορετικά περιβάλλοντα και εφαρμογές [Zen, 2009; Black, 2007]. Σε αντίθεση με παλαιότερες προσεγγίσεις που στηρίζονταν στη γνώση (knowledge based) και έκαναν χρήση κανόνων, στις σύγχρονες προσεγγίσεις παραμετρικής σύνθεσης φωνής, η παραγωγή του σήματος φωνής στηρίζεται επίσης στην υιοθέτηση κάποιου παραμετρικού μοντέλου ωστόσο ο έλεγχος και η διαχείριση των παραμέτρων αυτού του μοντέλου πραγματοποιείται πλέον με στατιστικό τρόπο [Black, 2007]. Το πιο επιτυχημένο μεθοδολογικό πλαίσιο (στατιστικής) παραμετρικής σύνθεσης, είναι η σύνθεση φωνής από κείμενο με χρήση Κρυφών Μαρκοβιανών Μοντέλων (HMM speech synthesis). Τα HMM αποτελούν ένα ικανό και γενικευμένο στατιστικό μοντέλο παραγωγής σήματος φωνής που προσφέρει την ευελιξία της παραμετρικής επεξεργασίας και διαχείρισης σε συνδυασμό με την επίτευξη ικανοποιητικής ποιότητας συνθετικής φωνής [Zen, 2009; Black, 2007; Taylor, 2009]. Η αρχή λειτουργίας της σύνθεσης με HMM, στηρίζεται στην ανάλυση και την παραμετρική αναπαράσταση της φωνής, η οποία οδηγεί στην δυνατότητα στατιστικής μοντελοποίησης της, με αποτέλεσμα να την καθιστά διαχειρίσιμη μέσω των Κρυφών Μαρκοβιανών Μοντέλων. Το μεθοδολογικό πλαίσιο σύνθεσης φωνής με HMM είναι αρκετά γενικευμένο και εύκολα προσαρμόσιμο σε άλλες γλώσσες. Όπως είδαμε, συστήματα σύνθεσης με HMM έχουν ήδη αναπτυχθεί με επιτυχία για διάφορες γλώσσες (βλ. κεφ. 1). Η προσαρμογή της τεχνολογίας σε κάποια γλώσσα εξαρτάται κυρίως από την σωστή αντιμετώπιση, μοντελοποίηση και ενσωμάτωση της πληροφορίας του «περιβάλλοντος» (context) παράλληλα με τη σωστή

διαχείριση των ιδιαιτεροτήτων κάθε γλώσσας. Η πληροφορία περιβάλλοντος εκτείνεται σε διάφορα (γλωσσολογικά) επίπεδα, όπως θα δούμε παρακάτω. Όπως αναφέραμε, λόγω της ποιότητας που επιτυγχάνουν, τα συστήματα σύνθεσης φωνής με HMM ανήκουν στην **τρίτη γενιά συστημάτων** σύνθεσης φωνής από κείμενο.

4.2 ΣΥΝΘΕΣΗ ΦΩΝΗΣ ΜΕ ΧΡΗΣΗ ΚΡΥΦΩΝ ΜΑΡΚΟΒΙΑΝΩΝ ΜΟΝΤΕΛΩΝ

Στο σχήμα 4.1, φαίνεται το γενικό δομικό διάγραμμα της διαδικασίας που ακολουθεί ένα σύστημα σύνθεσης φωνής από κείμενο με χρήση HMM. Όπως αναφέρθηκε, το σύστημα στηρίζεται στα διαθέσιμα δεδομένα (data driven) με την έννοια ότι χρειάζεται εκπαίδευση από ήδη υπάρχουσα επισημειωμένη βάση δεδομένων φυσικής ομιλίας (corpus-based). Επιπλέον, η παραμετρική αναπαράσταση του σήματος φωνής (συνήθως μέσω του μοντέλου πηγής-φίλτρου [Quatieri, 2001]), συνοδεύεται από πληροφορία γλωσσολογικού και φωνητικού τύπου (linguistic and phonetic model) για την από κοινού στατιστική διαχείριση τους. Το σύστημα χωρίζεται σε δύο στάδια: το στάδιο της εκπαίδευσης και το στάδιο της σύνθεσης.



ΣΧΗΜΑ 4.1: Το μεθοδολογικό πλαίσιο σύνθεσης φωνής από κείμενο με χρήση κρυφών Μαρκοβιανών μοντέλων (HMM-based speech synthesis framework).

Στη μαθηματική του διάσταση, το πρόβλημα της σύνθεσης φωνής από κείμενο με HMM διατυπώνεται με την ακόλουθη λογική [Zen, 2009]: Χρησιμοποιώντας την πλήρως επισημειωμένη βάση δεδομένων, το **στάδιο της εκπαίδευσης** είναι υπεύθυνο για την παραμετροποίηση του σήματος φωνής και την εξαγωγή της κατάλληλης πληροφορίας τόσο σε ακουστικό όσο και σε γλωσσολογικό επίπεδο. Κατόπιν, πραγματοποιείται η εκτίμηση των παραμέτρων των μοντέλων HMM, σύμφωνα με το **κριτήριο μέγιστης πιθανοφάνειας** (ML - maximum likelihood criterion), δηλαδή πραγματοποιείται η εκτίμηση:

$$\hat{\lambda} = \arg\max_{\lambda} \{ p(O/W,\lambda) \} \tag{4.1}$$

όπου, λ το σετ από τις παραμέτρους του μοντέλου (π.χ., το κλασικό μοντέλο HMM $\lambda = (\mathrm{A},\mathrm{B},\pi)$ [Gales, 2007]), O ένα σύνολο από δεδομένα εκπαίδευσης και W ένα σύνολο από λέξεις ή ένα σώμα εκπαίδευσης (δηλ. η βάση δεδομένων φωνής) πού αντιστοιχούν στο O. Σε αντιστοιχία με την αναγνώριση φωνής, η εκπαίδευση στη περίπτωση της σύνθεσης που αφορά την εκτίμηση της εξίσωσης 4.1, βασίζεται στον αλγόριθμο Expectation-Maximization [Gales, 2007], με τη διαφορά ότι το σύνολο της πληροφορίας μοντελοποιείται από ένα σύνολο με multi-stream εξαρτώμενα από πληροφορία περιβάλλοντος HMM (multi-stream context dependent HMMs).

Στο ακουστικό επίπεδο ανήκει η πληροφορία που αφορά τα φασματικά και προσωδιακά χαρακτηριστικά ενώ στο γλωσσολογικό επίπεδο, όπως και στην αναγνώριση φωνής, ανήκει η πληροφορία που αφορά το φωνητικό επίπεδο (low level context information) η οποία όμως στην περίπτωση της σύνθεσης με HMM επαυξάνεται και με χαρακτηριστικά που αφορούν υψηλότερου επιπέδου γλωσσική πληροφορίας «περιβάλλοντος» (high level context information). Κάθε φώνημα, και πιο συγκεκριμένα κάθε κατάσταση φωνήματος (HMM state), επισημειώνεται και συνοδεύεται με το σύνολο αυτής της πληροφορίας. Έτσι, κάθε HMM είναι εξαρτώμενο από πληροφορία περιβάλλοντος (context dependent HMM). Όπως θα δούμε, για λόγους αραιότητας δεδομένων (data sparsity) στα HMM που προκύπτουν εφαρμόζεται συσταδοποίηση μέσω δυαδικών δένδρων απόφασης.

Το **στάδιο της σύνθεσης**, είναι υπεύθυνο για την παραγωγή της πιο πιθανής ακολουθίας (**MAP κριτήριο – Maximum a posteriori**) παραμέτρων φωνής $\hat{\boldsymbol{o}}$, δεδομένου της πρότασης $\boldsymbol{w}$ (ή πιο γενικά της ακολουθίας λέξεων) από το σύνολο παραμέτρων $\hat{\lambda}$ που έχουν εκτιμηθεί:

$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}} \{ p(\mathbf{o}/\mathbf{w}, \hat{\lambda}) \} \tag{4.2}$$

Από τις παραμέτρους αυτές πραγματοποιείται η παραγωγή (ανακατασκευή) του συνθετικού σήματος φωνής σύμφωνα με το μοντέλο πηγής-φίλτρου.

4.2.1 Διαδικασία εκπαίδευσης

Το στάδιο της εκπαίδευσης, μοιάζει στην μεθοδολογία με αυτό της αναγνώρισης φωνής, με τη διαφορά ότι πραγματοποιείται από κοινού μοντελοποίηση τόσο της φασματικής όσο και της προσωδιακής πληροφορίας. Πιο συγκεκριμένα, από την επισημειωμένη βάση δεδομένων εκπαίδευσης εξάγονται πληροφορίες που

αφορούν τόσο την πηγή (προσωδία) όσο και τον φωνητικό σωλήνα (φασματική αναπαράσταση), με τις οποίες εκπαιδεύονται τα μοντέλα των HMM και συγκεκριμένα οι καταστάσεις τους (HMM states). Συνήθως η μοντελοποίηση πραγματοποιείται ανά φώνημα το οποίο αναπαριστάται (μοντελοποιείται) με τρείς καταστάσεις, μια για διαφορετικές περιοχές του φωνήματος. Οι καταστάσεις του φωνήματος αντιπροσωπεύουν την περιοχή μετάβασης λόγω του προηγούμενου φωνήματος (αρχή), τη «σταθερή» (φασματικά) περιοχή (μέση) και την περιοχή μετάβασης προς το επόμενο φώνημα (τέλος). Σε αρκετές περιπτώσεις χρησιμοποιείται και η τοπολογία με πέντε καταστάσεις εκ των οποίων η πρώτη και η τελευταία αντιπροσωπεύουν απλώς καταστάσεις μετάβασης σε προηγούμενο και επόμενο HMM [Yamagishi, 2007].

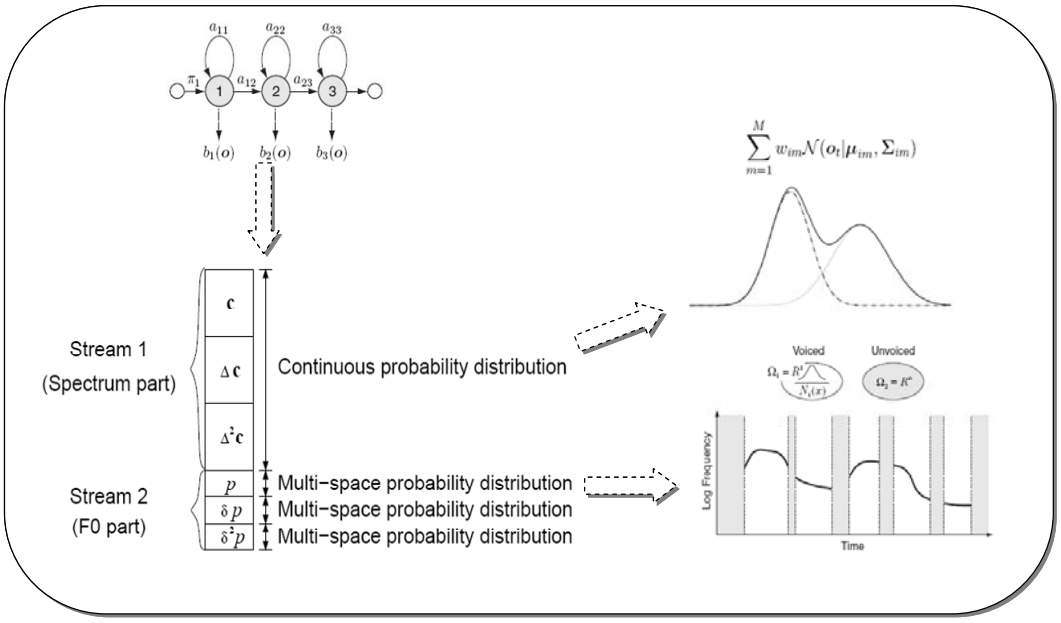
Κάθε κατάσταση χαρακτηρίζει είτε μια πολυμεταβλητή κατανομή Gauss είτε μίξη από κατανομές Gauss (Gaussian Mixture Models - GMM) η οποία μοντελοποιεί τα φασματικά χαρακτηριστικά. Επιπλέον, κάθε κατάσταση συνοδεύει και η αντίστοιχη κατανομή που μοντελοποιεί τον επιτονισμό, δηλαδή την θεμελιώδη συχνότητα. Η μοντελοποίηση του επιτονισμού ενέχει το εγγενές πρόβλημα του διαχωρισμού μεταξύ έμφωνων (voiced) και άφωνων (unvoiced) περιοχών, καθώς στη δεύτερη περίπτωση δεν ορίζεται οπότε χρειάζεται ειδική αντιμετώπιση. Στη περίπτωση αυτή χρησιμοποιούνται ειδικού τύπου κατανομές πιθανότητας (όπως η χρήση κατανομών πιθανότητας MSD – Multispace probability distributions [Yoshimura, 1999]) ή διαφορετικά ορίζεται η τιμή της θεμελιώδους συχνότητας με τη βοήθεια γραμμικής παρεμβολής (interpolation) ή διαφορετικού τύπου κατανομών [Yu, 2009]. Επιπρόσθετα, το προσωδιακό χαρακτηριστικό της διάρκειας μοντελοποιείται με το πίνακα μετάβασης των καταστάσεων των HMM, ο οποίος στη περίπτωση της σύνθεσης δεν αποτελείται από πιθανότητες αλλά από κατανομές πιθανοτήτων.

Για τη φασματική περιγραφή του σήματος φωνής μπορούν να χρησιμοποιηθούν διάφορες φασματικές αναπαραστάσεις από τις οποίες οι συνηθέστερες είναι οι γενικευμένοι cepstral συντελεστές σε κλίμακα Mel (Mel generalized cepstral coefficients) αλλά και οι συντελεστές MFCC [Tokuda, 2000], τα LSF (Line Spectral Frequencies) που προκύπτουν από γραμμική πρόβλεψη [Qian, 2006] και πιο πρόσφατα οι cepstral συντελεστές που προκύπτουν από τη μέθοδο STRAIGHT [Zen, 2007]. Η χρήση της τελευταίας αναπαράστασης βελτίωσε σημαντικά τη τελική ποιότητα περιορίζοντας σημαντικά το φαινόμενο «βόμβου» (buzziness) που συμβαίνει λόγω του είδους της διέγερσης στο μοντέλο πηγή-φίλτρο.

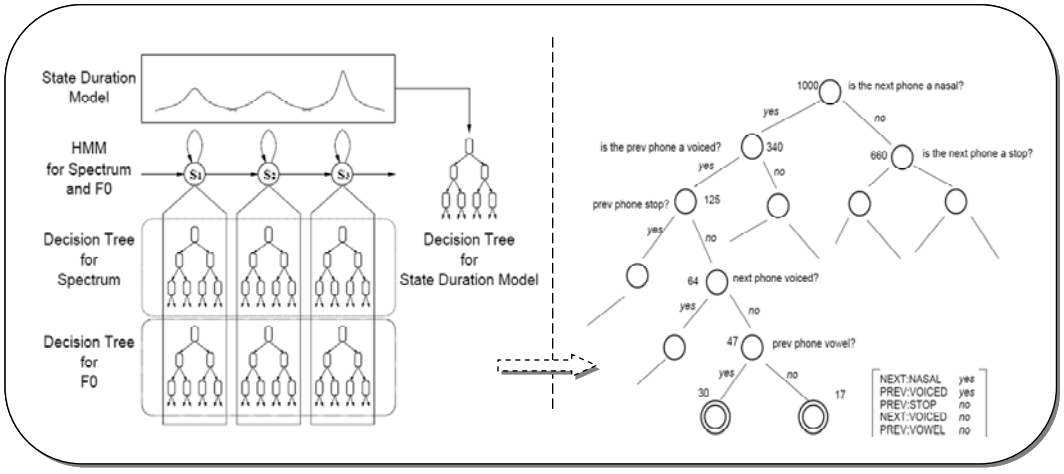
Σημαντικό γνώρισμα στη σύνθεση με HMM, αποτελεί η ενσωμάτωση πληροφορίας δυναμικών φασματικών και προσωδιακών χαρακτηριστικών (dynamic features). Για τα προσωδιακά χαρακτηριστικά η πληροφορία αυτή περιλαμβάνει μόνο τη θεμελιώδη συχνότητα. Ουσιαστικά συμπεριλαμβάνεται πληροφορία πρώτης και δεύτερης παραγώγου δηλαδή οι συντελεστές «ταχύτητας» (delta) και «επιτάχυνσης» (delta-delta) του διανύσματος χαρακτηριστικών. Χωρίς την ενσωμάτωση αυτού του είδους πληροφορίας, το παραγόμενο σήμα φωνής θα αποτελούνταν μόνο από τις μέσες τιμές για κάθε κατάσταση κάθε φωνήματος. Η ενσωμάτωση αυτής της πληροφορίας ουσιαστικά εξασφαλίζει ομαλές τροχιές και μεταβάσεις των παραμέτρων κατά τη σύνθεση [Zen, 2009; Tokuda, 1995].

Ιδιαίτερο επίσης χαρακτηριστικό στη σύνθεση με HMM, αποτελεί η επιπλεόν ενσωμάτωση πληροφορίας (γλωσσολογικού) «περιβάλλοντος» τόσο στη διαδικασία

της εκπαίδευσης όσο και της σύνθεσης. Για κάθε φώνημα πραγματοποιείται συσταδοποίηση με βάση το φωνητικό και γλωσσολογικό περιβάλλον (context dependent clustering) με τη βοήθεια δυαδικών δένδρων απόφασης (binary decision trees). Οπότε, κάθε κατάσταση των HMM είναι συνδεδεμένη με κάποιο φύλλο δένδρου συνεπώς, στη συνέχεια, κάθε φύλλο δένδρου αντιστοιχεί σε μια κατανομή η οποία μοντελοποιεί τα φασματικά και προσωδιακά χαρακτηριστικά αυτής της κατάστασης. Με αυτό τον τρόπο, δημιουργούνται εξαρτώμενα από το «περιβάλλον» μοντέλα HMM (context dependent HMMs) όπου μοντελοποιούν το σήμα φωνής σαν μια σειρά από τυχαίες παρατηρήσεις (υπό περιορισμούς) και εναλλαγές καταστάσεων που περιγράφονται με στατιστικό τρόπο. Έτσι, το τελικό σήμα φωνής αναπαρίσταται από την συνένωση των HMM ανά φώνημα. Οι παραπάνω διαδικασίες αποτυπώνονται στο σχήμα 4.2.



(α)



(β)

ΣΧΗΜΑ 4.2: Αναπαράσταση και μοντελοποίηση στην σύνθεση με HMM: (α) Η τοπολογία HMM, η διανυσματική αναπαράσταση με διαφορετικές ροές και η στατιστική μοντελοποίηση καθεμίας από αυτές. (β) Πλήρες μοντέλο φωνήματος με HMM και η συσταδοποίηση με πληροφορία «περιβάλλοντος».

Ένα HMM μοντέλο N καταστάσεων περιγράφεται ως $\lambda = (\mathrm{A}, \mathrm{B}, \pi)$ όπου, $\mathrm{A} = \{a_{ij}\}_{i,j=1}^{N}$ ο πίνακας με πιθανότητες μετάβασης στις καταστάσεις ή η κατανομές που περιγράφουν τις μεταβάσεις ανάμεσα στις καταστάσεις, $B = \{b_j(o_t)\}_{j=1}^{N}$ η κατανομή πιθανότητας των παρατηρήσεων τη χρονική στιγμή *t,* και $\pi = \{\pi_i\}_{i=1}^{N}$ οι αρχικές πιθανότητες καταστάσεων. Στην περίπτωση της σύνθεσης, ο A καθορίζει την διάρκεια κάθε κατάστασης σε ένα φώνημα και συνεπώς την διάρκεια του φωνήματος ενώ ο B καθορίζει ποιες παράμετροι θα παραχθούν και οι οποίες θα ανταποκρίνονται στα φασματικά και προσωδιακά χαρακτηριστικά της συγκεκριμένης κατάστασης του φωνήματος. Ειδικότερα, στο στάδιο της εκπαίδευσης τα μοντέλα καθορίζονται επιπλέον από

- το σύνολο των καταστάσεων $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$
- το σύνολο των παρατηρήσεων $\boldsymbol{O} = \{o_1, o_2, ..., o_T\}$

Σύμφωνα με την (4.1), στόχος της εκπαίδευσης είναι η εύρεση του μοντέλου που μεγιστοποιεί την $p(O/W)$ στα συγκεκριμένα δεδομένα. Η εκπαίδευση των HMM πραγματοποιείται με τους ίδιους αλγόριθμους που χρησιμοποιούνται και στη περίπτωση της αναγνώρισης φωνής. Οι βασικότερες τεχνικές που εφαρμόζονται είναι αρχικά οι αλγόριθμοι εκπαίδευσης forward-backward και ο αλγόριθμος αποκωδικοποίησης Viterbi, που βρίσκει την πιθανότερη ακολουθία καταστάσεων με δυναμικό προγραμματισμό. Κατόπιν ακολουθεί επανεκτίμηση των παραμέτρων μέσω του αλγόριθμου Baum–Welch (ή αλλιώς του αλγόριθμου Expectation Maximization). Ο αλγόριθμος Baum–Welch είναι ένας γενικευμένος αλγόριθμος μεγιστοποίησης της πιθανοφάνειας [Gales, 2007; Rabiner, 1999]. Ειδικότερα στη σύνθεση, όπως αναφέρθηκε, το σύνολο των παρατηρήσεων αποτελείται από διαφορετικές ροές δεδομένων (multi-stream) που αφορούν πληροφορία φασματικής και προσωδιακής αναπαράστασης. Τα φασματικά χαρακτηριστικά μοντελοποιούνται είτε με χρήση κανονικών κατανομών πολλών μεταβλητών (multivariate Gaussian distributions) είτε με μίξη πολλαπλών κανονικών κατανομών πολλών μεταβλητών (Gaussian Mixture Models) [Zen, 2009; Taylor, 2009]. Αντίθετα τα προσωδιακά χαρακτηριστικά, και ειδικότερα η θεμελιώδης συχνότητα και οι δυναμικές μεταβολές της, μοντελοποιούνται με ειδικού τύπου κατανομές ώστε να αντιμετωπίζουν το πρόβλημα μη ορισμού της θεμελιώδους συχνότητας σε άφωνες περιοχές. Συνήθως χρησιμοποιούνται κατανομές τύπου MSD (Multispace probability distributions) (σχήμα 4.2α), ώστε να διαχειρίζονται και να μοντελοποιούν μεγέθη που μπορούν να παίρνουν ταυτόχρονα διακριτές (discrete symbol) και συνεχείς τιμές (continuous probability density functions).

Συνοψίζοντας, τα συστήματα σύνθεσης φωνής με HMM χαρακτηρίζονται από τα παρακάτω χαρακτηριστικά:

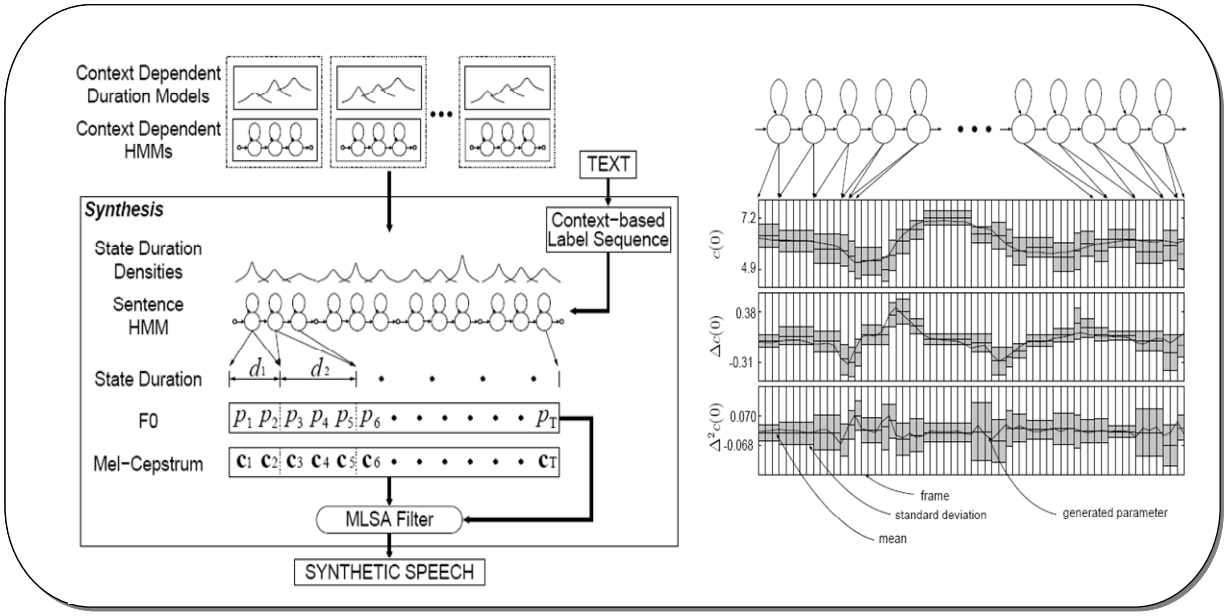
- Από κοινού περιγραφή της θεμελιώδους συχνότητας F0 (pitch) και του φάσματος με χρήση των χαρακτηριστικών μεγεθών logF0 και των cepstral συντελεστών σε κλίμακα mel αντίστοιχα, μαζί με την ταχύτητα (πρώτη χρονική παράγωγος) και την επιτάχυνση (δεύτερη χρονική παράγωγος) τους αντίστοιχα [Yoshimura, 1999].

- Την χρήση κατανομών πιθανότητας τύπου MSD – Multispace probability distributions για την περιγραφή και διαχείριση της F0 σε έμφωνους και άφωνους ήχους [Yoshimura, 1999; Zen, 2007].
- Την εκτενή αξιοποίηση πληροφορίας από το «περιβάλλον» (full context modelling), κατά την διαδικασία της οποίας λαμβάνονται υπόψη περισσότεροι παράγοντες τόσο σε φωνητικό όσο και σε γλωσσολογικό επίπεδο. Η αξιοποίηση της πληροφορίας γίνεται χρησιμοποιώντας τεχνικές συσταδοποίησης (cluster) με δένδρα αποφάσεων [Zen, 2009; Yamagishi, 2003].
- Την δημιουργία ξεχωριστών μοντέλων διάρκειας (ανά κατάσταση) για την μοντελοποίηση της χρονικής δομής (εξέλιξης) του σήματος φωνής [Yoshimura, 1999; Zen, 2004] μέσω HSMM (Hidden Semi Markov Models).

4.2.2 Διαδικασία σύνθεσης

Κατά το στάδιο της σύνθεσης, αρχικά το κείμενο εισόδου αναλύεται και μετατρέπεται σε συμβολική αναπαράσταση ανά φώνημα που περιέχει πληροφορίες σχετικές με το «περιβάλλον» και κατόπιν πραγματοποιείται η συνένωση των HMM βάση αυτής της αναπαράστασης. Έτσι σχηματοποιείται το συνολικό HMM που αναπαριστά την πρόταση (sentence HMM). Στη συνέχεια, εκτιμώνται οι διάρκειες ανά κατάσταση των HMM μέσω των αντίστοιχων κατανομών πιθανότητας. Το επόμενο βήμα περιλαμβάνει την παραγωγή των διανυσμάτων για το φάσμα και την θεμελιώδη συχνότητα μέσω του μοντέλου και με βάση την μεγιστοποίηση των πιθανοτήτων τους για την συγκεκριμένη σύνθεση (σχέση 4.2), δηλαδή παράγεται ο συρμός διανυσμάτων των οποίων η πιθανότητα είναι μέγιστη δεδομένου των καταστάσεων και της σειράς των HMM για την προς σύνθεση πρόταση. Τέλος, σύμφωνα με το μοντέλο πηγής-φίλτρου, το συνθετικό σήμα φωνής προκύπτει μέσω του φίλτρου σύνθεσης. Συνήθως, το φίλτρο που χρησιμοποιείται είναι το MLSA – Mel log spectrum approximation filter for mel-cepstral coefficients [Tokuda, 2000; Tokuda, 1995]. Η διαδικασία της σύνθεση περιγράφεται στο σχήμα 4.3. Στο ίδιο σχήμα φαίνεται και η επίδραση της ενσωμάτωσης των δυναμικών παραμέτρων στο στάδιο της σύνθεσης. Η διαδικασία παραγωγής φασματικών και προσωδιακών παραμέτρων υπόκειται σε περιορισμούς οι οποίοι προκύπτουν από την ενσωμάτωση των δυναμικών παραμέτρων τους στο στάδιο της σύνθεσης. Ο αλγόριθμος παραγωγής των παραμέτρων κατά τη σύνθεση υπό το κριτήριο της μέγιστης πιθανοφάνειας περιγράφεται στη συνέχεια και βασίζεται στα [Tokuda, 2000; Yoshimura, 1999; Zen, 2009; Benesty, 2008 Ch. 21]. Έστω ότι η ακολουθία του συνόλου των καταστάσεων είναι η $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$ και δεδομένου του συνόλου των παρατηρήσεων $\boldsymbol{o} = \{o_1, o_2, ..., o_T\}$ όπου T το μήκος της ακολουθίας, το πρόβλημα έγκειται στην αναζήτηση των παρατηρήσεων που μεγιστοποιούν την σχέση (4.2), η οποία προσεγγίζεται ως εξής:

$$\begin{aligned}\hat{\mathbf{o}} &= \arg\max_{\mathbf{o}}\{\, p(\mathbf{o}/\mathbf{w}, \hat{\lambda})\,\} = \arg\max_{\mathbf{o}}\{\sum_{q} p(\mathbf{o}, \mathbf{q}/\mathbf{w}, \hat{\lambda})\,\} \\ &\approx \arg\max_{\mathbf{o}}\{\sum_{q} p(\mathbf{o}, \mathbf{q}/\mathbf{w}, \hat{\lambda})\,\} = \arg\max_{\mathbf{o}}\max_{\mathbf{q}}\{\, p(\mathbf{o}, \mathbf{q}/\mathbf{w}, \hat{\lambda})\,\} \\ &= \arg\max_{\mathbf{o}}\max_{\mathbf{q}}\{P(\mathbf{q}/\mathbf{w}, \hat{\lambda}) \cdot\, p(\mathbf{o}/\mathbf{q}, \hat{\lambda})\}\end{aligned} \quad (4.3)$$



ΣΧΗΜΑ 4.3: Διαδικασία σύνθεσης φωνής με χρήση HMM. Στο σχήμα φαίνεται και η επίδραση των δυναμικών χαρακτηριστικών στη διαδικασία παραγωγής των διανυσμάτων.

Όπως είπαμε η ακολουθία του συνόλου των καταστάσεων $\hat{\boldsymbol{q}} = \{q_1, q_2, ..., q_T\}$ θεωρείται γνωστή και προκύπτει με μεγιστοποίηση της πιθανότητας της διάρκειας των καταστάσεων λόγω των προδιαγραφών του κειμένου εισόδου ως,

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}} \{ P(\mathbf{q} / \mathbf{w}, \hat{\lambda}) \} \tag{4.4}$$

Οπότε η (4.3) λόγω της (4.4) γίνεται,

$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}} \{ p(\mathbf{o} / \hat{\mathbf{q}}, \hat{\lambda}) \} \tag{4.5}$$

Η (4.5) εκφράζει την πιο πιθανή ακολουθία παρατηρήσεων δεδομένης της ακολουθίας καταστάσεων. Υποθέτοντας ότι το σύνολο των παρατηρήσεων σε κάθε κατάσταση ακολουθεί την κανονική πολυμεταβλητή κατανομή $b_j(\mathbf{o_t}) \sim N(\mathbf{o_t}; \boldsymbol{\mu}_\mathbf{j}, \boldsymbol{\Sigma}_\mathbf{j})$ και αφού το σύνολο των παρατηρήσεων εξαρτάται από την ακολουθία καταστάσεων, η (4.5) τελικά γίνεται,

$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}} \{ N(\mathbf{o} / \boldsymbol{\mu}_{\hat{\mathbf{q}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{q}}}) \} \tag{4.6}$$

όπου $\boldsymbol{\mu}_{\hat{\mathbf{q}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{q}}}$ το διάνυσμα της μέσης τιμής και ο πίνακας συμμεταβλητότητας της ακολουθίας καταστάσεων αντίστοιχα.

Αν και για τη διαδικασία της σύνθεσης απαραίτητοι είναι μόνο οι στατικοί συντελεστές, ωστόσο το πρόβλημα ανάγεται στην παραγωγή εκείνων των στατικών συντελεστών που να υπακούν σε περιορισμούς που τίθενται από τις (δυναμικές) μεταβολές τους. Χωρίς αυτούς τους περιορισμούς κάθε κατάσταση θα παρήγαγε τις μόνο τις μέσες τιμές (σχήμα 4.3). Οπότε το σύνολο των παρατηρήσεων (διανύσματα παρατήρησης) επαυξάνονται και με τους δυναμικούς συντελεστές ως εξής:

$$\boldsymbol{o_t} = [\boldsymbol{c}_t^T, \boldsymbol{\Delta}^{(1)}\boldsymbol{c}_t^T, ..., \boldsymbol{\Delta}^{(D-1)}\boldsymbol{c}_t^T] \quad (4.7)$$

όπου, $\boldsymbol{c}_t = [c_t(1), c_t(2), ..., c_t(M)]^T$ το διάστασης M διάνυσμα στατικών χαρακτηριστικών και $\boldsymbol{\Delta}^{(D)}\boldsymbol{c}_t = \sum_{\tau=-L_-^{(D)}}^{L_+^D} w^{(D)}(\tau)\boldsymbol{c}_{t+\tau}$ το τάξης D και διάστασης M διάνυσμα δυναμικών χαρακτηριστικών και $w^{(D)}$ παράγοντας στάθμισης (παράθυρο) υπολογισμού των δυναμικών χαρακτηριστικών. Συνήθως χρησιμοποιούνται δυναμικά χαρακτηριστικά μέχρι και δεύτερης τάξης που υπολογίζονται ως $\boldsymbol{\Delta c}_t = 0.5(\boldsymbol{c}_{t+1} - \boldsymbol{c}_{t-1})$ και $\boldsymbol{\Delta^2 c}_t = 0.5(\boldsymbol{c}_{t-1} - 2\boldsymbol{c}_t + \boldsymbol{c}_{t+1})$ οπότε το σύνολο των παρατηρήσεων εκφράζεται ως $\boldsymbol{o_t} = [\boldsymbol{c}_t^T, \boldsymbol{\Delta}^{(1)}\boldsymbol{c}_t^T, \boldsymbol{\Delta}^{(2)}\boldsymbol{c}_t^T]$.

Η σχέση μεταξύ του συνολικού διανύσματος παρατηρήσεων $\boldsymbol{o}$ και του συνόλου των στατικών χαρακτηριστικών $\boldsymbol{c}$ γράφεται σε μορφή πίνακα ως [Benesty, 2008 Ch. 21],

$$\boldsymbol{o} = \boldsymbol{W} \cdot \boldsymbol{c} \quad (4.8)$$

όπου,

$$\begin{aligned} &\boldsymbol{c} = [\boldsymbol{c}_1^T, \boldsymbol{c}_2^T, ..., \boldsymbol{c}_T^T]^T, \\ &\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, ..., \boldsymbol{W}_T] \otimes \boldsymbol{I}_{M\times M}, \\ &\boldsymbol{W_t} = [\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, ..., \boldsymbol{w}_t^{(D-1)}] \end{aligned} \quad (4.9)$$

όπου, $\boldsymbol{w}_t^{(d)}$ το εκάστοτε διάνυσμα με τους παράγοντες στάθμιση για τον υπολογισμό των δυναμικών χαρακτηριστικών. Οπότε το πρόβλημα της μεγιστοποίησης της σχέσης (4.6) εκφράζεται μέσω της (4.8) ως,

$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}} \{ N(\mathbf{W}\cdot\mathbf{c} \ / \boldsymbol{\mu}_{\hat{\mathbf{q}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{q}}}) \} \quad (4.10)$$

οπότε τελικά η $p(\mathbf{o}/\hat{\mathbf{q}}, \hat{\lambda}) = p(\mathbf{W}\cdot\mathbf{c}/\hat{\mathbf{q}}, \hat{\lambda}) = N(\mathbf{W}\cdot\mathbf{c} \ / \boldsymbol{\mu}_{\hat{\mathbf{q}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{q}}})$ μεγιστοποιείται ως,

$$\frac{\partial \log(p(\mathbf{W}\cdot\mathbf{c}/\hat{\mathbf{q}}, \hat{\lambda}))}{\partial \mathbf{c}} = 0 \quad (4.11)$$

Εφόσον έχουμε κανονικές κατανομές η (4.11) οδηγεί στην κλειστή λύση [Benesty, 2008 Ch. 21]:

$$\boldsymbol{R}_q \boldsymbol{c} = \boldsymbol{r}_q \quad (4.12)$$

όπου,

$$\begin{aligned} \boldsymbol{R}_q &= \boldsymbol{W}^T \boldsymbol{\Sigma}_q^{-1} \boldsymbol{W} \\ \boldsymbol{r}_q &= \boldsymbol{W}^T \boldsymbol{\Sigma}_q^{-1} \boldsymbol{\mu}_q \end{aligned} \quad (4.13)$$

Η σχέση (4.12) παράγει το στατικό σύνολο παρατηρήσεων $\boldsymbol{c}$ το οποίο μεγιστοποιεί την (4.10) και την (4.5), υπακούοντας παράλληλα στους περιορισμούς των δυναμικών χαρακτηριστικών που εκφράζονται από την (4.8).

4.3 ΠΡΟΣΑΡΜΟΓΗ ΣΤΗΝ ΕΛΛΗΝΙΚΗ ΓΛΩΣΣΑ

4.3.1 Σχεδίαση και υλοποίηση

Το σύστημα σύνθεσης φωνής με HMM που αναπτύχθηκε, στηρίζεται σε μεγάλο στο δομικό διάγραμμα του σχήματος 4.1, όσο και στις τεχνικές, εργαλεία και μεθοδολογίες που έχουν αναπτυχθεί και χρησιμοποιούνται σε διεθνές επίπεδο. Σε αυτό το στάδιο ανάπτυξης, η προσαρμογή στην ελληνική γλώσσα εστιάστηκε στην κατάλληλη ανάλυση και εξαγωγή των διανυσμάτων χαρακτηριστικών καθώς και στην κατάλληλη εξαγωγή, επισημείωση και μοντελοποίηση της πληροφορίας από το φωνητικό και γλωσσολογικό «περιβάλλον» (context), το οποίο εξαρτάται από την εκάστοτε γλώσσα.

Για την αναπαράσταση των φωνημάτων της Ελληνικής γλώσσας, υιοθετήθηκε ένα σύνολο από 37 φωνήματα. Αυτό χωρίζεται σε 26 σύμφωνα και αλλόφωνα, 5 τονισμένα φωνήεντα (έμφωνα), 5 άτονα φωνήεντα (έμφωνα) και ένα σύμβολο που αναπαριστά την σιωπή σε αρχή πρότασης, τέλος πρότασης και στις ενδιάμεσες παύσεις στη ροή της ομιλίας. Το σύνολο των 37 φωνημάτων ορίζει 9 τάξεις ως εξής:

- **άφωνα τυρβώδη (unvoiced fricatives)**: /f/, /x/, /X/, /T/, /s/
- **έμφωνα τυρβώδη (voiced fricatives)**: /v/, /J/, /j/, /D/, /z/
- **υγρά (liquids)**: /l/, /r/, /L/
- **ένρινα (nasals)**: /m/, /M/, /n/, /N/, /h/
- **άηχα εκρηκτικά (unvoiced stops)**: /p/, /t/, /k/, /c/
- **ηχηρά εκρηκτικά (voiced stops)**: /b/, /d/, /g/, /G/
- **σιωπή (silence)**: /-/
- **τονισμένα έμφωνα (stressed vowels)**: /a'/, /e'/, /i'/, /o'/, /u'/
- **άτονα έμφωνα (unstressed vowels)**: /a/, /e/, /i/, /o/, /u/

Σημαντικό γνώρισμα της Ελληνικής γλώσσας αποτελεί ο προσδιορισμός του τόνου, ο οποίος ορίζεται στο ελεύθερο κείμενο ή/και μπορεί να εξαχθεί από τους κανόνες της φωνητικής μεταγραφής. Οπότε, τα τονισμένα φωνήεντα αναπαρίστανται με ξεχωριστό σύμβολο. Το γνώρισμα αυτό έχει σημαντικό αντίκτυπο στα προσωδιακά χαρακτηριστικά καθώς τα τονισμένα φωνήεντα μοντελοποιούνται ως ξεχωριστά φωνήματα.

Για την εκπαίδευση των μοντέλων HMM μέσω συσταδοποίησης με δένδρα απόφασης, εκτός της πληροφορίας περιβάλλοντος που αναφέρεται στο φωνητικό επίπεδο (low level context information), τα υπόλοιπα χαρακτηριστικά που χρησιμοποιήθηκαν και που αφορούν υψηλότερου επιπέδου γλωσσική πληροφορία «περιβάλλοντος» (high level context information), συνοψίζονται ως εξής:

- **Επίπεδο φωνήματος**
  - ταυτότητα φωνήματος
  - θέση στην συλλαβή και προς τις δύο πλευρές
  - προηγούμενα φωνήματα δύο θέσεις πριν και μετά

- **Επίπεδο συλλαβής**
  - αριθμός φωνημάτων στην συλλαβή
  - αριθμός φωνημάτων στην προηγούμενη συλλαβή
  - αριθμός φωνημάτων στην επόμενη συλλαβή
  - έμφαση (accent) παρούσας, επόμενης και προηγούμενης συλλαβής
  - τόνος παρούσας, επόμενης και προηγούμενης συλλαβής
  - θέση συλλαβής στην λέξη και στη φράση
  - αριθμός προηγούμενων και επόμενων συλλαβών με έμφαση και τόνο στην παρούσα φράση
  - αριθμός συλλαβών μέχρι την επόμενη συλλαβή που περιλαμβάνει τόνο ή/και έμφαση
  - αριθμός από φωνήεντα στην συλλαβή

- **Επίπεδο λέξης**
  - μέρος του λόγου της παρούσας, προηγούμενης και επόμενης λέξης
  - αριθμός συλλαβών της παρούσας, προηγούμενης και επόμενης λέξης
  - θέση της λέξης στη φράση

- **Επίπεδο φράσης**
  - αριθμός συλλαβών της παρούσας, προηγούμενης και επόμενης φράσης
  - θέση στην κυρίως φράση
  - προσδιορισμός προσωδιακών χαρακτηριστικών σημείων της φράσης (πχ. κόμμα, τελεία κτλ.)

- **Επίπεδο πρότασης**
  - αριθμός συλλαβών στην παρούσα πρόταση
  - αριθμός λέξεων στην παρούσα πρόταση

Στα πλαίσια της συγκεκριμένης ανάπτυξης δεν χρησιμοποιήθηκε επισημείωση τύπου ToBI (Tones and Break Indices) ώστε τα προσωδιακά χαρακτηριστικά να προκύψουν κατά το δυνατό, μόνο από μέσω της εξάρτησης τους από τους φωνητικούς και γλωσσολογικούς παράγοντες.

Για την εκπαίδευση των μοντέλων HMM χρησιμοποιήθηκε μια βάση δεδομένων προηχογραφημένης φυσικής ομιλίας γενικού περιεχομένου με γυναίκα ομιλητή. Η βάση αποτελείται από 1200 προτάσεις με ευρεία κάλυψη σε πραγματώσεις φωνημάτων σε ποικίλα περιβάλλοντα, ενώ το στυλ της εκφώνησης θεωρείται ουδέτερο (στυλ «ανάγνωσης»). Η συχνότητα δειγματοληψίας επιλέχτηκε στα 16KHz με διακριτική ικανότητα 16bits. Η εξαγωγή των απαραίτητων παραμέτρων και χαρακτηριστικών πραγματοποιήθηκε με αυτόματο τρόπο και η κατάτμηση και επισημείωση έγινε με χρήση HMM [Gales, 2007] ενώ κατόπιν ελέγχθηκε και χειρονακτικά. Η διαδικασία ελέγχου είναι απαραίτητη καθώς τυχόν λάθη ή μικρές διαφοροποιήσεις οδηγούν σε λανθασμένη εκπαίδευση των μοντέλων οπότε κατ' επέκταση δεν θα υπάρχει αντιστοιχία κατά την διαδικασία της σύνθεσης. Επίσης, η ίδια βάση δεδομένων χρησιμοποιήθηκε και στο σύστημα με επιλογή και συρραφή ακουστικών μονάδων το οποίο χρησιμοποιήθηκε στη συγκριτική αξιολόγηση που περιγράφεται στην επόμενη ενότητα.

Για το σκοπό της σύνθεσης με HMM, σύμφωνα με τις διεθνής δημοσιευμένες πρακτικές η ανάλυση πραγματοποιήθηκε με παράθυρο τύπου Blackman σε πλαίσια των 25msec με ρυθμό πλαισίων 5msec (20msec επικάλυψη). Για κάθε πλαίσιο επιλέχθηκε η φασματική αναπαράσταση των συντελεστών cepstral σε κλίμακα Mel, επαυξημένους τόσο με τη πρώτη (delta) όσο και με τη δεύτερη (delta-delta) χρονική παράγωγο τους. Η θεμελιώδης συχνότητα αναπαραστάθηκε με την τιμή της σε λογαριθμική κλίμακα (logF0), και χρησιμοποιήθηκε τόσο με η πρώτη όσο και η δεύτερη χρονική παράγωγος της. Για κάθε HMM, Υιοθετήθηκε η τοπολογία με 5 καταστάσεις με κατεύθυνση από αριστερά προς τα δεξιά χωρίς παράλειψη κάποιας κατάστασης (5-state left-to-right with no skip). Σύμφωνα με τη βάση δεδομένων και τους παράγοντες πληροφορίας περιβάλλοντος που χρησιμοποιήθηκαν, το σύστημα που αναπτύχθηκε αποτελείτο από 116432 πλήρη μοντέλα που προήλθαν από ένα σύνολο 1293 ερωτήσεων για την κατασκευή των δένδρων απόφασης για τις συστάδες των καταστάσεων των HMM.

4.3.2 Πειραματική αξιολόγηση

Η αποτίμηση του συστήματος σύνθεσης φωνής με HMM πραγματοποιήθηκε με την εξαγωγή μικρής κλίμακας ακουστικών πειραμάτων και την εξαγωγή της μέσης ακουστικής αξιολόγησης (MOS), σε κλίμακα από 1 έως 5 (1: κακή ποιότητα, 5: άριστη ποιότητα) τόσο για την φυσικότητα όσο και για την καταληπτότητα του συνθετικού σήματος φωνής. Το σύστημα που αναπτύχθηκε συγκρίθηκε με δύο υπάρχοντα συστήματα σύνθεσης φωνής από κείμενο, ένα με επιλογή και συρραφή ακουστικών μονάδων (unit selection) και ένα που χρησιμοποιεί σύνθεση με διφωνήματα (μία πραγμάτωση για κάθε δίφωνο). Για το σκοπό αυτό έγινε τυχαία επιλογή 15 προτάσεων που αποτελούνταν από 2 έως 16 λέξεις η καθεμία, από γενικό σώμα κειμένου το οποίο δεν συμπεριλαμβάνεται στη βάση δεδομένων. Η σύνθεση των προτάσεων πραγματοποιήθηκε και με τα τρία συστήματα με σκοπό την συγκριτική αποτίμηση των δύο συστημάτων. Η ακουστική αξιολόγηση πραγματοποιήθηκε από μεικτή ομάδα, σχετικά με την εμπειρία τους σε τεχνολογίες σύνθεσης φωνής από κείμενο, 10 ανθρώπων. Η παρουσίαση των προτάσεων στους ακροατές γινόταν με τυχαία σειρά χωρίς όμως να γνωστοποιείται η μέθοδος για κάθε πρόταση. Κάθε τριάδα προτάσεων μπορούσε να ακουστεί από τους ακροατές όσες φορές επιθυμούσαν. Οι ακροατές ήταν ενημερωμένοι ότι πρόκειται για συνθετική ομιλία ενώ δεν τους δόθηκε κάποιο παράδειγμα φυσικής ομιλίας που να λειτουργήσει ως αναφορά στην βαθμολογία. Το γεγονός αυτό εξηγεί τις σχετικά υψηλές βαθμολογίες για όλα τα συστήματα. Τα αποτελέσματα της συγκριτικής αξιολόγησης φαίνονται στον πίνακα 4.1.

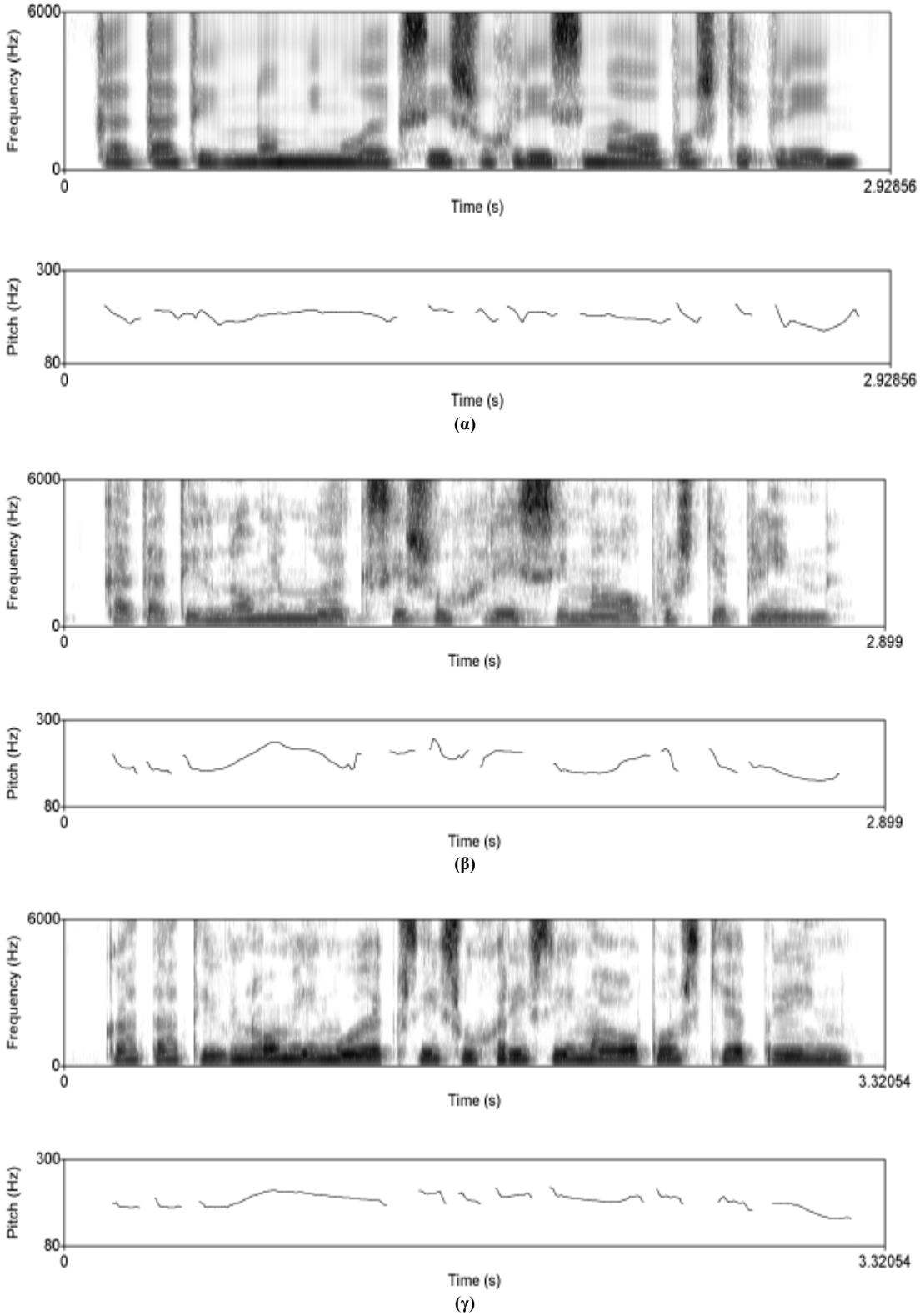
Από τα αποτελέσματα της αξιολόγησης προκύπτει ότι το σύστημα με επιλογή και συρραφή ακουστικών μονάδων υπερτερεί τόσο έναντι της μηχανής σύνθεσης με HMM όσο και της μηχανής σύνθεσης με διφωνήματα. Η μηχανή σύνθεσης με HMM, μειονεκτεί περισσότερο στην φυσικότητα της συνθετικής ομιλίας και λιγότερο στην καταληπτότητα. Το αποτέλεσμα αυτό οφείλεται περισσότερο στο γεγονός ότι το σύστημα με HMM χαρακτηρίζεται από το εγγενές μειονέκτημα της παραγωγής της συνθετικής φωνής μέσω ανακατασκευής από διανύσματα χαρακτηριστικών υιοθετώντας το μοντέλο πηγής φίλτρου. Το πρόβλημα έγκειται στο σήμα διέγερσης (σήμα πηγής) το οποίο αποτελείται από μια παλμοσειρά κρουστικού τυπου, με

αποτέλεσμα το τελικό σήμα να εμφανίζει «βόμβο», επηρεάζοντας άμεσα την τελική ποιότητα και κατ' επέκταση την ακουστική αξιολόγηση. Η υιοθέτηση πιο εξελιγμένων προσεγγίσεων στην μοντελοποίηση του σήματος διέγερσης έχει δείξει ότι οδηγεί σε σημαντική βελτίωση της ποιότητας [Zen, 2009; Maia, 2007; Kawahara, 1999]. Έτσι, παρά την ομαλή και καταληπτή παραγωγή συνθετικού σήματος φωνής, αυτή η ατέλεια επηρεάζει την τελική κρίση για την απόδοση του συστήματος. Ωστόσο, η καταληπτότητα που επιτυγχάνεται με την σύνθεση με HMM είναι αρκετά υψηλή και συγκρίσιμη με την μηχανή που βασίζεται σε επιλογή ακουστικών μονάδων. Το αποτέλεσμα αυτό δείχνει ότι η μοντελοποιήσεις, τόσο σε ακουστικό (φασματικό και προσωδιακό) επίπεδο όσο και σε γλωσσολογικό επίπεδο, επαρκούν για τον πλήρη έλεγχο της τελικής σύνθεσης και δύναται να οδηγήσουν σωστά τα μοντέλα για την παραγωγή ομαλής και καταληπτής φωνής με προσωδιακά χαρακτηριστικά που να τείνουν να 'ακολουθούν' εκείνα της διαθέσιμης βάσης δεδομένων που χρησιμοποιείται κατά την εκπαίδευση, δεδομένου ότι έχουν ληφθεί και επισημειωθεί σωστά οι κατάλληλοι παράγοντες που αφορούν την πληροφορία περιβάλλοντος (phonetic and linguistic context) ώστε να γενικεύουν κατάλληλα σε περιπτώσεις που δεν υπήρχαν κατά την εκπαίδευση (unseen context). Αυτό βέβαια εξαρτάται και από τα δεδομένα που εμπεριέχονται στη διαθέσιμη βάση, τα οποία πρέπει να καλύπτουν το σύνολο της απαιτούμενης πληροφορίας.

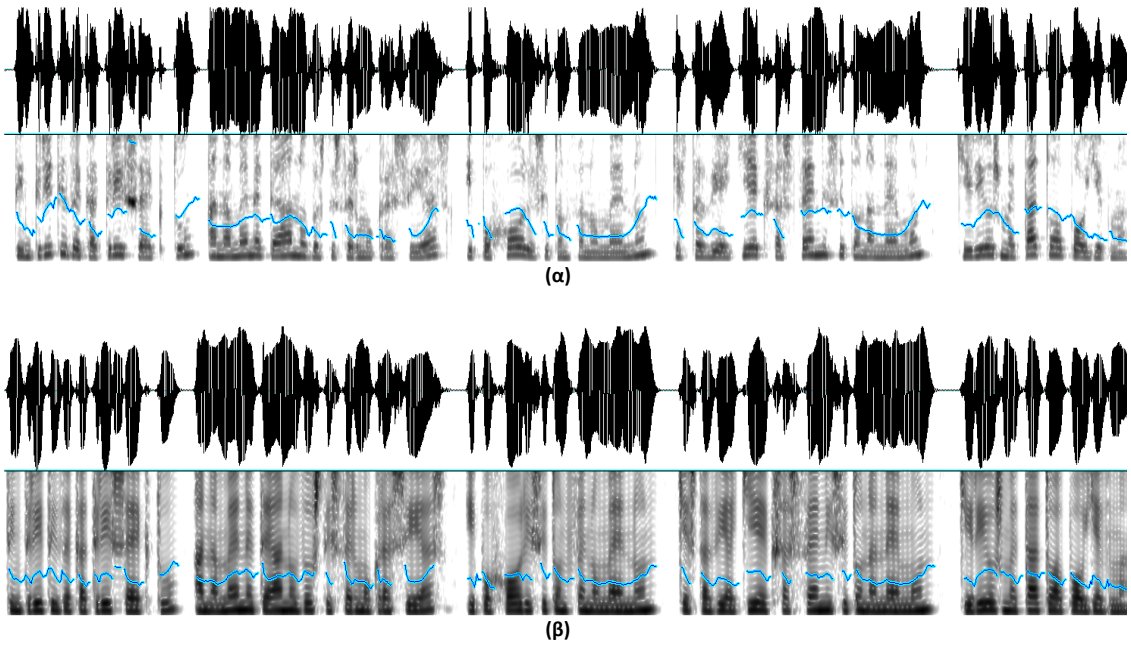
ΠΙΝΑΚΑΣ 4.1: Συγκριτική αξιολόγηση MOS του συστήματος σύνθεσης φωνής με HMM για την Ελληνική γλώσσα

ΣΥΣΤΗΜΑ	ΦΥΣΙΚΟΤΗΤΑ	ΚΑΤΑΛΗΠΤΟΤΗΤΑ
HMM	3.9	4.2
Diphone	3.7	3.8
Unit Selection	4.5	4.6

Αυτό είναι φανερό στο σχήμα 4.4 που απεικονίζεται το ηχογράφημα και η χρονική εξέλιξη της θεμελιώδους συχνότητας του συνθετικού σήματος της πρότασης "*κατά τη γνώμη μου εξίσου χρήσιμα όπως και πρίν (kata ti Jno'mi mu eksi'su xri'sima o'pos ce pri'n)*" όπως προκύπτει από την μηχανή σύνθεσης με HMM, την μηχανή με επιλογή και συρραφή ακουστικών μονάδων και την μηχανή σύνθεσης με δίφωνα. Όπως αναφέρθηκε, στο συγκεκριμένο σύστημα τα προσωδιακά χαρακτηριστικά δεν επισημειώθηκαν (π.χ. μέσω ToBI) παρά μόνο στην περίπτωση των τονισμένων φωνημάτων, γεγονός που επεξηγεί την λιγότερο πλούσια προσωδία που επιτυγχάνεται από το σύστημα σύνθεσης με HMM. Ένα παράδειγμα σύνθεσης με HMM και για τη περίπτωση πρότασης που υπάρχει στη βάση δεδομένων, φαίνεται στο σχήμα 4.5. "*Την Τρίτη δεκατρείς έκτου, αναμένεται άνοδος της θερμοκρασίας, υποχώρηση των φαινομένων, και σταδιακή εξασθένιση των ανέμων, κυρίως μετά το απόγευμα. - tin tríti Dekatrís éktu_ anaménete ánoDos tis Termokrasías_ ipoxórisi ton fenoménon_ ce staDiací eksasTénisi ton anémon_ cirios metá to apójevma.*". Παρατηρείται ότι στο συνθετικό σήμα φωνής η χρονική εξέλιξη της θεμελιώδους συχνότητας τείνει να ακολουθεί αυτή της φυσικής ομιλίας.



ΣΧΗΜΑ 4.4: Παράδειγμα ηχογραφήματος και επιτονισμού σε συνθετικό σήμα φωνής που προκύπτει από: α) το σύστημα σύνθεσης με HMM, β) το σύστημα σύνθεσης με επιλογή και συρραφή ακουστικών μονάδων, και γ) το σύστημα με χρήση διφώνων.



ΣΧΗΜΑ 4.5: Παράδειγμα συνθετικού σήματος φωνής που προκύπτει από το σύστημα σύνθεσης με ΗΜΜ. Στο (α) φαίνεται η φυσική πρόταση ενώ στο (β) το συνθετικό σήμα. Η μπλε καμπύλη σε κάθε περίπτωση δείχνει την χρονική εξέλιξη της θεμελιώδους συχνότητας.

4.4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Η χρήση κρυφών Μαρκοβιανών μοντέλων παρέχει ένα αποτελεσματικό και ευέλικτο μεθοδολογικό πλαίσιο για την παραμετρική προσέγγιση της σύνθεσης φωνής από κείμενο, μέσω μοντελοποίησης και διαχείρισης των παραμέτρων με στατιστικό τρόπο και με βάση τα διαθέσιμα δεδομένα. Στα πλαίσια της διατριβής, αναπτύχθηκε ένα τέτοιο σύστημα για την Ελληνική γλώσσα. Η προσαρμογή εστιάστηκε στην επισημείωση και εξαγωγή των απαραίτητων παραμέτρων και στην υιοθέτηση και εφαρμογή της φωνητικής και γλωσσολογικής μοντελοποίησης της πληροφορίας περιβάλλοντος. Η τελική ποιότητα που επιτυγχάνει το σύστημα αποτιμήθηκε με συγκριτικά ακουστικά πειράματα, τόσο έναντι ενός συστήματος με επιλογή και συρραφή ακουστικών μονάδων όσο και με ένα σύστημα σύνθεσης με διφωνήματα δεύτερης γενιάς. Η ακουστική αξιολόγηση ανέδειξε την υπεροχή του συστήματος με ΗΜΜ σε σχέση με το τελευταίο σύστημα. Από την άλλη πλευρά, η τελική ποιότητα υπολείπεται σε σχέση με το πρώτο σύστημα. Στο γεγονός αυτό συντελούν εγγενείς παράγοντες της διαδικασίας όπως είναι η μοντελοποίηση και εφαρμογή του σήματος διέγερσης. Συμπερασματικά, τα αποτελέσματα είναι ενθαρρυντικά καθώς το σύστημα αποτελεί μια πρώτη προσπάθεια, οπότε υπάρχει χώρος για πολλές βελτιώσεις. Τόσο η εφαρμογή πιο εξελιγμένων τεχνικών και μοντέλων στην διαδικασία παραγωγής του συνθετικού σήματος φωνής, όσο και η ενσωμάτωση περισσότερης πληροφορίας σχετικής με τις προσωδιακές ιδιότητες, αναμένεται να βελτιώσουν αισθητά την τελική ποιότητα του συστήματος.

ΚΕΦΑΛΑΙΟ
-5-
ΣΥΝΑΡΤΗΣΗ ΚΟΣΤΟΥΣ ΈΝΩΣΗΣ ΜΕ ΤΑΞΙΝΟΜΗΤΕΣ ΜΙΑΣ ΤΑΞΗΣ

ΚΕΦΑΛΑΙΟ 5 – ΣΥΝΑΡΤΗΣΗ ΚΟΣΤΟΥΣ ΕΝΩΣΗΣ ΜΕ ΤΑΞΙΝΟΜΗΤΕΣ ΜΙΑΣ ΤΑΞΗΣ

Στο κεφάλαιο αυτό, εξετάζεται ένα από τα βασικά ζητήματα σχετικά με τον αλγόριθμο επιλογής ακουστικών μονάδων, που είναι η εκτίμηση των φασματικών (ακουστικών) ασυνεχειών στην ένωση των φωνημάτων και οι οποίες συνθέτουν το κόστος ένωσης φασματικών ασυνεχειών. Πιο συγκεκριμένα, προτείνεται μια νέα προσέγγιση η οποία στηρίζεται σε δεδομένα (data driven), βασίζεται στη χρήση ταξινομητών μιας τάξης (one-class classifiers) και αφορά την εκτίμηση των φασματικών ασυνεχειών. Η προτεινόμενη τεχνική εκμεταλλεύεται την διαθέσιμη βάση δεδομένων ηχογραφημένης φυσικής ομιλίας, η οποία ουσιαστικά αποτελεί το πρωτογενές υλικό για την εκπαίδευση των ταξινομητών, δεδομένου ότι κάθε ζεύγος από γειτονικά πλαίσια σήματος φωνής (speech frames) αποτελεί μια πλήρως φυσική ένωση. Στο πλαίσιο της διατριβής, τα γειτονικά πλαίσια φωνής αναπαρίστανται από κοινού με ένα διάνυσμα χαρακτηριστικών που αποτελείται από διαφορετικού τύπου (φασματικές) αποστάσεις. Χρησιμοποιείται ένας ταξινομητής μιας τάξης για κάθε φώνημα και η εκπαίδευση του πραγματοποιείται στα παραπάνω διανύσματα χαρακτηριστικών. Η πειραματική αξιολόγηση της αναδεικνύει την λειτουργικότητά και την αποδοτικότητα της μεθόδου, η οποία υπερέχει έναντι των συμβατικών τεχνικών που χρησιμοποιούνται συνήθως.

5.1 ΕΙΣΑΓΩΓΗ

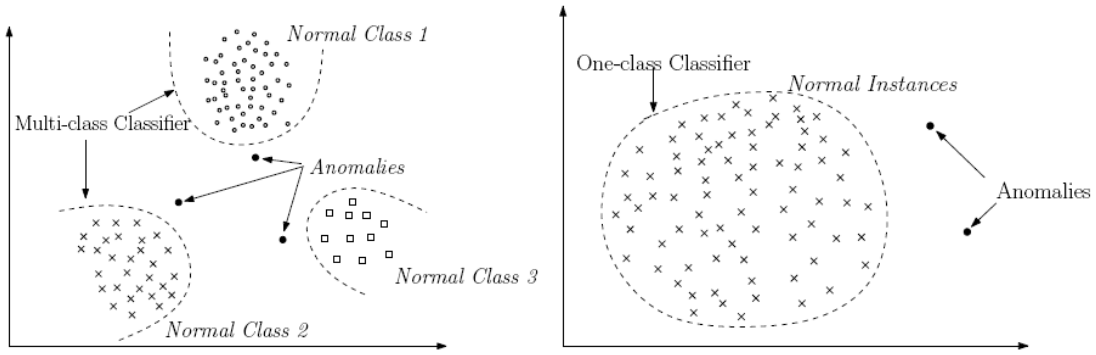
Οι ταξινομητές μιας τάξης εφαρμόζονται σε προβλήματα αναγνώρισης προτύπων δύο τάξεων για τα οποία διαθέσιμα δεδομένα εκπαίδευσης υπάρχουν μόνο για την μια τάξη ενώ για την άλλη τάξη είτε δεν υπάρχουν καθόλου δεδομένα, είτε υπάρχουν λιγοστά, είτε είναι δύσκολο και ασύμφορο να συλλεχθούν. Με παρόμοια λογική εφαρμόζεται και σε προβλήματα πολλαπλών τάξεων, αφού ένα πρόβλημα πολλών τάξεων αναλύεται σε πολλά προβλήματα δύο τάξεων [Duda, 1973; Bishop, 1995]. Στα παραπάνω προβλήματα, από την πλευρά της ταξινόμησης μιας τάξης, μόνο μια θεωρείται η ζητούμενη τάξη ή τάξη στόχος (**target class**) ενώ οι υπόλοιπες θεωρούνται ως παραπειστικές ή εκτός τάξης (**outlier class**) [Hodge, 2004; Markou, 2003ab, Jain, 2000; Kennedy, 2009; Patcha, 2007; Tax, 2001; Bakar, 2006; Chandola, 2009]. Η χρήση ταξινομητών μιας τάξης έχει πρόσφατα εφαρμοστεί σε διάφορους τομείς όπως,

- ταυτοποίηση/πιστοποίηση ομιλητή (speaker verification) [Brew, 2008]
- εξόρυξη δεδομένων (data mining) [Bakar, 2006]
- κατάτμηση ακουστικών (audio) σημάτων [Davy, 2002; Davy, 2006; Desobry, 2006]
- speaker diarization [Fergani, 2008]
- βιοϊατρική [Gardner, 2006]
- ταξινόμηση κειμένων [Manevitz, 2001]
- ανάλυση χρονοσειρών [Modenesi, 2009]
- ανίχνευση και ταξινόμηση ακουστικών (audio) σημάτων [Rabaoui, 2008]
- ανάλυση εκφράσεων προσώπου [Zeng, 2006]

Σε σχέση με την παραδοσιακή προσέγγιση που ακολουθείται στην Αναγνώριση Προτύπων, η οποία εστιάζει στην διάκριση και την ταξινόμηση αντικειμένων σε τάξεις, η ταξινόμηση μιας τάξης στοχεύει στην περιγραφή (και αναγνώριση) μιας (συγκεκριμένης) τάξης αντικειμένων. Η προσέγγιση αυτή προτείνει την περιγραφή και μοντελοποίηση της ζητούμενης τάξης με κλειστά όρια (enclosed boundary), ή με άλλα λόγια με κλειστό χώρο. Ως αποτέλεσμα, σε αυτήν την τάξη ανήκουν αντικείμενα μόνο εάν εμπίπτουν σε αυτόν τον κλειστό χώρο. Σε διαφορετική περίπτωση τα αντικείμενα θεωρούνται ως παραπειστικά (outliers) [Tax, 2001; Markou, 2003ab]. Ο όρος Ταξινόμηση-μιας-Τάξης (one-class classification) εισήχθη στην εργασία των Moya, Koch και Hostetler [Moya, 1993].

5.2 ΤΑΞΙΝΟΜΗΣΗ ΜΙΑΣ ΤΑΞΗΣ (ONE-CLASS CLASSIFICATION)

Η ταξινόμηση μιας τάξης αφορά και αντιμετωπίζει ευρύτερα το πρόβλημα της «Ανίχνευσης Καινοτομίας» (Novelty Detection) στο χώρο της μηχανικής μάθησης [Bishop, 1994]. Στόχος είναι η αναγνώριση νέων ή άγνωστων δεδομένων τα οποία δεν έχει συναντήσει το σύστημα μηχανικής μάθησης κατά την εκπαίδευση. Το πρόβλημα αυτό μπορεί να ειδωθεί και από την πλευρά της «Ανίχνευσης Ανωμαλιών ή Παραπειστικών δεδομένων» (Anomaly or Outlier Detection) [Markou, 2003ab; Chandola, 2009]. Στην περίπτωση αυτή στόχος είναι η διάκριση και μοντελοποίηση των δεδομένων (αντικειμένων) που ανήκουν σε συγκεκριμένη τάξη έναντι άλλων περιπτώσεων που μπορούν να τύχουν. Ένα γενικό παράδειγμα ταξινόμησης μιας τάξης δίνεται στο σχήμα 5.1 τόσο για πολλές όσο και για μια τάξη «στόχο».



ΣΧΗΜΑ 5.1: Παράδειγμα Ταξινόμησης-μιας-Τάξης [Chandola, 2009].

Ανεξαρτήτως μεθόδου, στην ταξινόμηση μιας τάξης στόχος είναι η απόρριψη των παραπειστικών δεδομένων και η αποδοχή των δεδομένων που ανήκουν στην ζητούμενη τάξη. Η αποδοχή αυτή στηρίζεται στη υπόθεση ότι ένα αντικείμενο που ανήκει στην ζητούμενη τάξη είναι παρόμοιο με τα υπόλοιπα αντικείμενα που ανήκουν σε αυτήν. Αν ως κριτήριο ομοιότητας $h(\mathbf{x}\,|\,X_T,\gamma)$ επιλεγεί είτε η απόσταση $d(\mathbf{x}\,|\,X_T,\gamma)$ από την τάξη «στόχος» είτε η πιθανότητα $p(\mathbf{x}\,|\,X_T,\gamma)$ να ανήκει στην τάξη «στόχος», τότε η ταξινόμηση πραγματοποιείται σύμφωνα με την εξής λογική:

$$h(\mathbf{x} \mid X_T, \gamma) = I(p(\mathbf{x} \mid X_T, \gamma) > \theta_p) = \begin{cases} 1, & \mathbf{x} \text{ is target} \\ 0, & \mathbf{x} \text{ is outlier} \end{cases} \quad (5.1)$$

ή

$$h(\mathbf{x} \mid X_T, \gamma) = I(d(\mathbf{x} \mid X_T, \gamma) < \theta_d) = \begin{cases} 1, & \mathbf{x} \text{ is target} \\ 0, & \mathbf{x} \text{ is outlier} \end{cases} \quad (5.2)$$

όπου $\mathbf{x}$ το διάνυσμα χαρακτηριστικών του αντικειμένου, X_T το σύνολο δεδομένων εκπαίδευσης δηλ. $X_T = \{\mathrm{x}_i \mid \mathrm{x}_i \in \Re^N, i = 1,...,N\}$, γ η πολυπλοκότητα ή τάξη του ταξινομητή (μεθόδου), θ_p και θ_d το προκαθορισμένο κατώφλι στην πιθανότητα ή την απόσταση αντίστοιχα και I μια δυαδική συνάρτηση ένδειξης (indicator function). Οι διάφορες μέθοδοι ταξινόμησης μιας τάξης διαφέρουν στον ορισμό και την βελτιστοποίηση τόσο του $p(\mathbf{x} \mid X_T, \gamma)$ όσο και του $d(\mathbf{x} \mid X_T, \gamma)$ καθώς και στην βελτιστοποίηση του κατωφλίου [Tax, 2001]. Εξίσου σημαντική σχεδιαστική παράμετρο αποτελεί η εξισορρόπηση του λάθους πρώτου είδους έναντι του λάθους δεύτερου είδους (βλ. , §5.2.2).

5.2.1 Ταξινομητές μιας τάξης

Η πιο συχνά ακολουθούμενη προσέγγιση στην ταξινόμηση μιας τάξης είναι η στατιστική μοντελοποίηση (Density or Statistical-based One-Class Classification) [Bishop, 1995; Tax, 2001, Markou, 2003]. Σε αυτή την προσέγγιση η ζητούμενη τάξη μοντελοποιείται από κάποια συνάρτηση πυκνότητα πιθανότητας και κατόπιν ορίζεται κάποιο κατώφλι πιθανότητας σε αυτήν την συνάρτηση για τον διαχωρισμό των αντικειμένων που ανήκουν στην ζητούμενη τάξη. Στην οικογένεια αυτή συναντάμε ταξινομητές μιας τάξης όπως, με κατανομές Gauss, με μείγμα κατανομών Gauss, με κατανομές μέσω Parzen κτλ.

Μια δεύτερη προσέγγιση στην ταξινόμηση μιας τάξης είναι η μοντελοποίηση του γεωμετρικού χώρου (Domain or Boundary One-Class Classification). Στην περίπτωση αυτή, η ζητούμενη τάξη μοντελοποιείται με την περιγραφή ενός κλειστού γεωμετρικού τόπου που περικλείει τα δεδομένα εκπαίδευσης. Η βελτιστοποίηση του χώρου αυτού οδηγεί στον ταξινομητή μιας τάξης. Επειδή στηρίζεται σε μετρικές (αποστάσεις) ανάμεσα στα δεδομένα, οι τεχνικές αυτές είναι ευαίσθητες στην κλίμακα των δεδομένων (δηλ. την κλίμακα στα στοιχεία του διανύσματος χαρακτηριστικών). Ωστόσο, συνήθως απαιτούν λιγότερο όγκο δεδομένων εκπαίδευσης για εύρωστη λειτουργία σε σχέση με την στατιστική μοντελοποίηση. Στην οικογένεια αυτή συναντώνται ταξινομητές SVDD (Support Vector Data Description), Nearest Neighbor κτλ.

Τέλος, στις μεθόδους ταξινόμησης μιας τάξης συναντάμε αυτές της ανακατασκευής (Reconstruction-based One-Class Classification), οι οποίες κυρίως βασίζονται σε υποθέσεις γύρω από τις ιδιότητες συσταδοποίησης (clustering) ή ανάλυσης σε υπο-χώρους (subspace) των δεδομένων εκπαίδευσης άρα και της τάξης «στόχος». Στην οικογένεια αυτή ανήκουν οι ταξινομητές όπως, k-means-based, Principal Component Analysis (PCA)-based, κτλ.

5.2.2 Αποτίμηση Ταξινομητών μιας τάξης

Η αξιολόγηση της επίδοσης της ταξινόμησης μιας τάξης αφορά ουσιαστικά τα λάθη ταξινόμησης τόσο ως προς το ποσοστό των δεδομένων της ζητούμενης τάξης που δεν αναγνωρίστηκαν ορθά, όσο και ως προς το ποσοστό των δεδομένων της παραπειστικής τάξης τα οποία λανθασμένα αναγνωρίστηκαν ότι ανήκουν στην ζητούμενη τάξη. Συνεπώς, η διαδικασία της ταξινόμησης καθορίζεται από τέσσερις περιπτώσεις οι οποίες συνοψίζονται στον πίνακα 5.1 μαζί με τα μεγέθη που τις αφορούν. Να σημειωθεί ότι τα παραπάνω λάθη εξαρτώνται αφενός από το κατώφλι και αφετέρου από την τάξη (ή πολυπλοκότητα) του μοντέλου του ταξινομητή. Όπως αναφέρθηκε το κατώφλι καθορίζεται από την ανοχή σε λάθη ως προς την ζητούμενη τάξη. Οι ανοχή αυτή, που εκφράζεται ως το ποσοστό FN, καθώς και η τάξη ορίζονται από τον σχεδιαστή. Ο καθορισμός της τάξης είναι πέρα του σκοπού της διατριβής και ο αναγνώστης παραπέμπεται στα [Tax, 2001; Tax, 2004] για περισσότερες πληροφορίες. Τα ποσοστά FN έναντι του FP αποτελούν κρίσιμες παραμέτρους, που πρέπει να ισορροπηθούν κατά την λειτουργία του ταξινομητή.

ΠΙΝΑΚΑΣ 5.1: Διάκριση των τεσσάρων περιπτώσεων ταξινόμησης στους Ταξινομητές Μιας-Τάξης και το αντίστοιχο λάθος ταξινόμησης.

		Πραγματική ετικέτα της τάξης (True class label)	
		Target Class	**Outlier Class**
Αποτέλεσμα Ταξινόμησης	**Target Class**	(True Positive - TP) $1-\varepsilon_t$ *target accepted*	(False Positive – FP) $\varepsilon_{II} = \varepsilon_o$ *outlier accepted*
(Estimated class label)	**Outlier Class**	(False Negative - FN) $\varepsilon_I = \varepsilon_t$ *target rejected*	(True Negative - TN) $1-\varepsilon_o$ *outlier rejected*

Το λάθος πρώτου είδους (*error of the first kind* - ε_I) δηλαδή το ποσοστό των FN είναι αυτό που μπορεί πρακτικά να ελαχιστοποιηθεί στην ταξινόμηση μιας τάξης ενσωματώνοντας όλο τον χώρο που ορίζουν τα παραδείγματα εκπαίδευσης. Ωστόσο η λύση αυτή είναι τετριμμένη και στην πράξη, κατά την σχεδίαση, ο ταξινομητής ανέχεται κάποιο ποσοστό λάθους στην τάξη «στόχο» ώστε να ισορροπήσει το άλλο πιθανό λάθος που ονομάζεται λάθος δεύτερου είδους (*error of the second kind* - ε_{II}) και αφορά το ποσοστό FP. Κατά την σχεδίαση δεν υπάρχουν παραδείγματα εκπαίδευσης της παραπειστικής τάξης ώστε να βελτιστοποιηθεί αυτό το λάθος. Πολλές φορές, για την αξιολόγηση των ταξινομητών μιας τάξης, είναι απαραίτητη η παραγωγή παραπειστικών δεδομένων με τεχνητό τρόπο. Τα παραπειστικά δεδομένα παράγονται συνήθως από ομοιόμορφη κατανομή γύρω από τα δεδομένα εκπαίδευσης της ζητούμενης τάξης [Tax, 2002].

Δεδομένου ενός ποσοστού (ή ρυθμού) TP, το κατώφλι απόφασης καθορίζεται από τα δεδομένα εκπαίδευσης έτσι ώστε να ικανοποιείται η σχέση:

$$\frac{1}{N}\sum_{i=1}^{N} I\{p(\mathbf{x}_i) > \theta_p\} \;=\; TP = 1-\varepsilon_t$$
$$or \qquad (5.3)$$
$$\frac{1}{N}\sum_{i=1}^{N} I\{d(\mathbf{x}_i) < \theta_d\} \;=\; TP = 1-\varepsilon_t$$

όπου $\mathrm{x}_i \in X_T$ (X_T : το σύνολο των δεδομένων εκπαίδευσης) και N το πλήθος των δεδομένων εκπαίδευσης. Η παραπάνω σχέση καθορίζει (μετρά) τα δεδομένα εκπαίδευσης που τελικά εσωκλείει η ζητούμενη τάξη για δεδομένο κατώφλι.

Η επίδραση του κατωφλίου αναδεικνύεται ποιοτικά από την γραφική παράσταση του ποσοστού TP έναντι του ποσοστού FP για διάφορες τιμές του κατωφλίου. Η καμπύλη αυτή ονομάζεται Receiver Operating Characteristic (ROC) [Fawcett, 2006; Davis, 2006]. Η καμπύλη ROC χαράσσεται είτε με τεχνητά παραπειστικά δεδομένα [Tax, 2002] είτε με πραγματικά δεδομένα (βλ. §5.3.3). Ένα ακόμα χρήσιμο μέγεθος, που συνήθως χρησιμοποιείται στην συγκριτική αξιολόγηση ταξινομητών, είναι το εμβαδό της επιφάνειας κάτω από την καμπύλη ROC. Το μέγεθος ονομάζεται AUC (Area under Curve) και ορίζεται ως [Tax, 2001],

$$AUC = 1-\int_0^1 \varepsilon_{\mathrm{II}}(\varepsilon_{\mathrm{I}})d\varepsilon_{\mathrm{I}} = 1-\int_0^1 \varepsilon_o(\varepsilon_t)d\varepsilon_t \qquad (5.4)$$

5.3 ΚΟΣΤΟΣ ΦΑΣΜΑΤΙΚΗΣ ΑΣΥΝΕΧΕΙΑΣ ΜΕ ΤΑΞΙΝΟΜΗΤΕΣ ΜΙΑΣ ΤΑΞΗΣ

Όπως αναφέρθηκε στο πρώτο και στο δεύτερο κεφάλαιο της διατριβής, η τάση για τη σχεδίαση του κόστους των φασματικών ασυνεχειών (όσο και συνολικά του κόστους «ένωσης») προσανατολίζεται στο να εκμεταλλεύεται όσο το δυνατόν καλύτερα τα διαθέσιμα δεδομένα (data driven), δηλαδή την διαθέσιμη βάση δεδομένων, έτσι ώστε η συνάρτηση κόστους να προκύπτει με εκπαίδευση πάνω σε αυτή, αποφεύγοντας όσο το δυνατόν περισσότερο, την εξάρτηση από τον ανθρώπινο παράγοντα τόσο στη σχεδίαση όσο και στη ρύθμιση. Όπως είδαμε στο δεύτερο κεφάλαιο, η εκπαίδευση των συναρτήσεων με δεδομένα που προκύπτουν από ακουστικά πειράματα είναι μεν αποδοτική ωστόσο είναι ασύμφορη στην εφαρμογή της.

Στην τεχνολογία σύνθεσης φωνής από κείμενο που στηρίζεται σε επιλογή και συρραφή ακουστικών μονάδων από προηχογραφημένη βάση δεδομένων φυσικής ομιλίας, η φυσικότητα της παραγόμενης συνθετικής ομιλίας κρίνεται αρκετά υψηλή στις περισσότερες των περιπτώσεων. Συχνά όμως, παρουσιάζοντας ακουστικές ασυνέχειες λόγω των συρραφών που προκύπτουν από τις διάφορες πραγματώσεις των ακουστικών μονάδων. Αυτό έχει ως αποτέλεσμα την άμεση επίπτωση στην φυσικότητα του συνθετικού λόγου. Βασικό παράγοντα για αυτό, αποτελεί η έλλειψη ενός αντικειμενικού κριτηρίου αξιολόγησης και επιλογής ακουστικών

μονάδων που να αντικατοπτρίζει την ανθρώπινη αντίληψη για τις ακουστικές ασυνέχειες και κατ' επέκταση την φυσικότητα της ομιλίας.

Ένα εγγενές χαρακτηριστικό της διαθέσιμης βάσης δεδομένων είναι ότι αποτελείται από ένα τεράστιο αριθμό από «φυσικές» ενώσεις για κάθε τύπο ακουστικής μονάδας. Πράγματι, κάθε πλαίσιο φωνής με το γειτονικό του μέσα στη βάση αποτελεί «φυσική» ένωση. Ωστόσο, η μη στάσιμη φύση του σήματος φωνής (quasi stationary) οδηγεί στο γεγονός αυτή η ιδιότητα να μην αντικατοπτρίζεται στις φασματικές αποστάσεις. Η φασματική απόσταση γειτονικών πλαισίων φωνής σπάνια είναι μηδενική. Ωστόσο, από την άποψη της ακουστικής αντίληψης η ένωση δύο τέτοιων πλαισίων οδηγεί σε απόλυτα «φυσικό» ακουστικό αποτέλεσμα. Την παρατήρηση αυτή εκμεταλλεύεται και η τεχνική που προτείνεται στο πλαίσιο της διατριβής.

Για το σκοπό αυτό, στο πλαίσιο της διατριβής διερευνάται η ανάπτυξη και αξιολόγηση ενός νέου κριτηρίου με στόχο την βελτίωση της φυσικότητας της συνθετικής ομιλίας. Συγκεκριμένα, για την εκτίμηση και αποτίμηση των φασματικών ασυνεχειών που προκύπτουν στην ένωση των ακουστικών μονάδων προτείνεται η υιοθέτηση και εφαρμογή ενός **νέου μεθοδολογικού πλαισίου** που βασίζεται στα **διαθέσιμα δεδομένα (data driven)** και συγκεκριμένα της βάσης δεδομένων φυσικής ομιλίας. Ειδικότερα, για την εκτίμηση των φασματικών ασυνεχειών που προκύπτουν στην ένωση των ακουστικών μονάδων υιοθετούνται μεθοδολογίες από τον χώρο της μηχανικής μάθησης και συγκεκριμένα η **χρήση των ταξινομητών μιας τάξης (One-Class Classifiers)** [Bishop, 1994; Tax, 2001; Markou, 2003ab; El-Yaniv, 2007].

Σύμφωνα με αυτή την προσέγγιση, η πειραματική αξιολόγηση αναμένεται να αναδείξει ότι η προτεινόμενη μεθοδολογία επιτυγχάνει σημαντικά καλύτερη απόδοση στην εκτίμηση φασματικών ασυνεχειών από τις συνήθεις προσεγγίσεις. Επιγραμματικά, τα σημαντικότερα πλεονεκτήματα της προσέγγισης αυτής μπορούν να συνοψιστούν στα εξής:

- Στηρίζεται στα διαθέσιμα δεδομένα οπότε το κόστος ένωσης για τις φασματικές ασυνέχειες εκπαιδεύεται και προκύπτει από πρότυπα που συναντώνται στην υπάρχουσα βάση δεδομένων

- Η εκπαίδευση δεν απαιτεί παρέμβαση από τον ανθρώπινο παράγοντα και δεν εξαρτάται από ακουστικά πειράματα

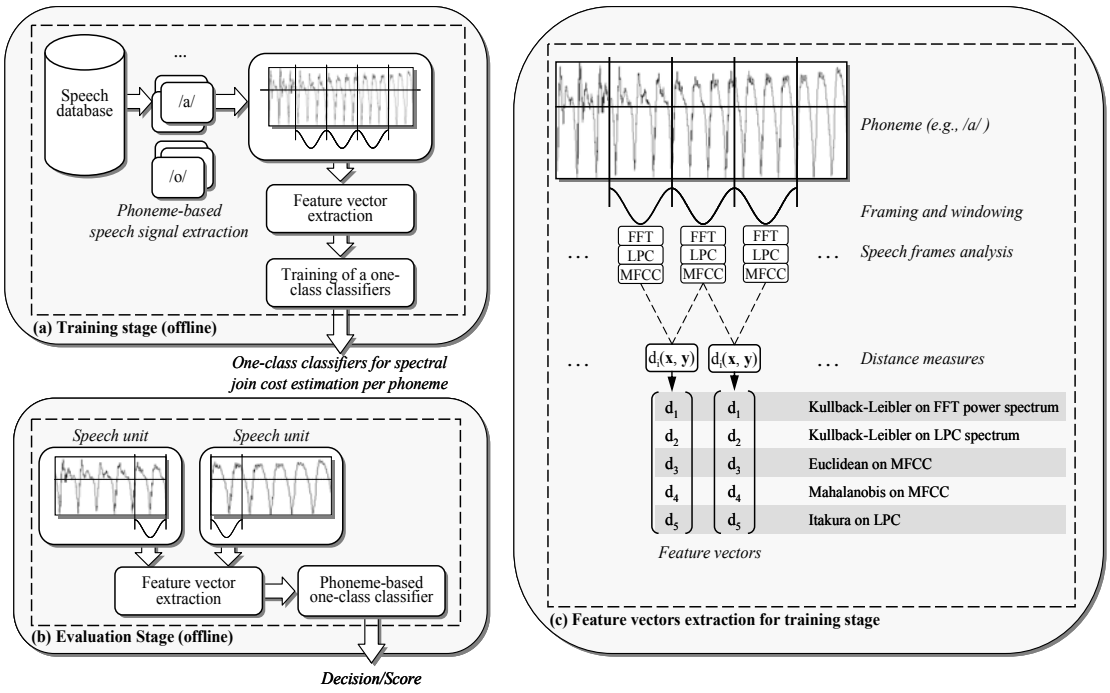
- Μπορεί να υιοθετηθεί ως ευρύτερο μεθοδολογικό πλαίσιο και να εφαρμοστεί με κάθε πιθανή αναπαράσταση που είναι ικανή να διακρίνει αποτελεσματικά τις δύο τάξεις.

- Μπορεί να επεκταθεί και στη σχεδίαση της συνολικής συνάρτησης κόστους ένωσης, με αποτέλεσμα την κατάργηση του προσδιορισμού των παραγόντων στάθμισης. Όπως είδαμε οι παράγοντες στάθμισης συνήθως προσδιορίζονται είτε χειρωνακτικά είτε με αυτόματες μεθόδους που στηρίζονται όμως σε αποτελέσματα από ακουστικά πειράματα.

5.3.1 Μεθοδολογικό πλαίσιο εκτίμησης φασματικής ασυνέχειας με ταξινομητές μιας τάξης

Για την υιοθέτηση της εν λόγω μεθοδολογίας στο πρόβλημα της επιμέρους συνάρτησης κόστους για τις φασματικές ασυνέχειες, εκμεταλλευόμαστε το γεγονός ότι η διαθέσιμη βάση δεδομένων προηχογραφημένης φυσικής ομιλίας, που χρησιμοποιείται και κατά την σύνθεση, αποτελείται από πλήθος φυσικών "ενώσεων" όπως αυτές ορίζονται από τα ζεύγη συνεχόμενων πλαισίων φωνής. Η πληθώρα από φυσικές "ενώσεις" μπορούν να χρησιμοποιηθούν ως πρωτογενή δεδομένα περιγραφής του χώρου των "καλών" ενώσεων (target class). Από την άλλη πλευρά, η συστηματική συλλογή από δεδομένα με φασματικές ασυνέχειες, ικανά να περιγράψουν τον χώρο των ακουστικών ασυνεχειών, αποτελεί μια δύσκολη, ασύμφορη και χρονοβόρα διαδικασία. Με αυτό τον τρόπο, το πρόβλημα της εκτίμησης των φασματικών ασυνεχειών μπορεί να προσεγγιστεί και να αντιμετωπιστεί ως **πρόβλημα μιας-τάξης** (one-class problem).

Το σχήμα 5.2 απεικονίζει το δομικό διάγραμμα που περιγράφει ευρύτερα το προτεινόμενο μεθοδολογικό πλαίσιο (σχήμα 5.2a-b) όσο και την εφαρμογή του στο πλαίσιο της διατριβής (σχήμα 5.2c). Όπως φαίνεται στο σχήμα, το προτεινόμενο μεθοδολογικό πλαίσιο για την εκτίμηση φασματικών ασυνεχειών με ταξινόμηση μίας τάξης χωρίζεται στο στάδιο εκπαίδευσης (training stage) και στο στάδιο εφαρμογής (evaluation stage). Τα δύο αυτά στάδια πραγματοποιούνται κατά τη σχεδίαση και υλοποίηση του συστήματος σύνθεσης (offline process).



ΣΧΗΜΑ 5.2: Δομικό διάγραμμα της χρήσης ταξινομητών μιας τάξης για τον υπολογισμό του φασματικού κόστους ένωσης στον αλγόριθμο επιλογής ακουστικών μονάδων: α) Το στάδιο εκπαίδευσης, β) το στάδιο εκτέλεσης και γ) η εξαγωγή του διανύσματος χαρακτηριστικών για την αναπαράσταση γειτονικών πλαισίων φωνής σε ένα φώνημα.

Στο στάδιο εκπαίδευσης, αρχικά πραγματοποιείται η εξαγωγή του σήματος φωνής ανά φώνημα. Για κάθε φώνημα, πραγματοποιείται διαχωρισμός σε πλαίσια και ακολουθεί η ανάλυση (αναπαράσταση/παραμετροποίηση) κάθε πλαισίου. Κατόπιν εξάγονται τα διανύσματα χαρακτηριστικών για την από κοινού αναπαράσταση διαδοχικών πλαισίων φωνής. Τα διανύσματα αυτά χρησιμοποιούνται στην εκπαίδευση του ταξινομητή μιας τάξης.

Στο στάδιο εφαρμογής, πραγματοποιείται παρόμοια διαδικασία με την διαφορά ότι το διάνυσμα χαρακτηριστικών αφορά την από κοινού αναπαράσταση για τα πλαίσια φωνής που δυνητικά πρόκειται να ενωθούν κατά την σύνθεση. Το διάνυσμα χαρακτηριστικών τροφοδοτεί τον ταξινομητή μιας τάξης, από τον οποίο προκύπτει ο χαρακτηρισμός μιας ένωσης ως φυσικής (target class) ή όχι (outlier class). Ο χαρακτηρισμός μπορεί να είναι είτε σε μορφή δυαδικής απόφασης (hard decision) είτε σε μορφή βαθμολογίας (soft decision) μέσω κάποιας απόστασης ή πιθανότητας, που εκφράζει τον βαθμό βεβαιότητας της απόφασης και επιστρέφεται από τον ταξινομητή. Και οι δύο περιπτώσεις είναι εφαρμόσιμες στο πλαίσιο της φασματικής συνάρτησης κόστους στον αλγόριθμο επιλογής ακουστικών μονάδων.

Το πρόβλημα εύρεσης κατάλληλης αναπαράστασης και εξαγωγής χαρακτηριστικών (features), ικανών αφενός για να περιγράψουν (αναπαραστήσουν) διαδοχικά πλαίσια φωνής και αφετέρου να αποτυπώνουν τόσο το φαινόμενο της ένωσης (concatenation) όσο και να επαυξάνουν την ικανότητα διάκρισης της ζητούμενης τάξης (target class) δεν είναι εύκολο. Στο πλαίσιο της διατριβής εξετάζεται η ανάλυση ανά φώνημα και η αναπαράσταση κάθε ζεύγους από συνεχόμενα πλαίσια φωνής με ένα διάνυσμα φασματικών αποστάσεων και διαφορετικών αναπαραστάσεων του σήματος φωνής. Ο αποτελεσματικός συνδυασμός διαφορετικών αναπαραστάσεων και διαφορετικών φασματικών αποστάσεων αναμένεται να επαυξήσει την συνολική απόδοση και να διακρίνει καλύτερα τις δύο τάξεις στο συγκεκριμένο πρόβλημα [Garau, 2008; Iyer, 2009; Vepa, 2004]. Τα διανύσματα αυτά χρησιμοποιούνται για την εκπαίδευση του ταξινομητή μιας τάξης. Κατά την σύνθεση, δύο πλαίσια φωνής από διαφορετικές πραγματώσεις του ίδιου φωνήματος αναπαρίστανται με τον ίδιο τρόπο και το κόστος φασματικής ασυνέχειας προκύπτει από την βαθμολογία που δίνει ο ταξινομητής. Οι λεπτομέρειες της διαδικασίας εφαρμογής περιγράφονται στην ενότητα που ακολουθεί.

5.3.2 Μεθοδολογία Εφαρμογής

Για την εφαρμογή του προτεινόμενου μεθοδολογικού πλαισίου, ακολουθήθηκε η διαδικασία που περιγράφεται στο σχήμα 5.2c. Τα πλαίσια σήματος φωνής σε κάθε φώνημα αναλύονται και αναπαρίστανται σε τρείς διαφορετικές φασματικές αναπαραστάσεις [Quatieri, 2001]:

- Βραχέως χρόνου φάσμα ισχύος Fourier (short-time FFT-based power spectrum)
- Συντελεστές cepstrum σε κλίμακα Mel (MFCC – Mel-Frequency cepstral coefficients)
- Συντελεστές γραμμικής πρόβλεψης (LPC – linear prediction coefficients)

Κάθε ζεύγος γειτονικών πλαισίων φωνής, αναπαρίσταται με πέντε διαφορετικές φασματικές αποστάσεις οι οποίες υπολογίζονται από τις παραπάνω φασματικές αναπαραστάσεις. Πιο συγκεκριμένα, οι φασματικές αποστάσεις που επιλέχθηκαν είναι (βλ. Κεφ. 2, §2.3.2.1):

- Η συμμετρική απόσταση Kullback-Leibler στο φάσμα ισχύος Fourier (***KL-FFT***) [Stylianou, 2001; Klabbers, 2001; Pantazis, 2005]

- Η συμμετρική απόσταση Kullback-Leibler στη φασματική περιβάλλουσα που υπολογίζεται από τους συντελεστές γραμμικής πρόβλεψης (***KL-LPC***) [Stylianou, 2001; Klabbers, 2001; Pantazis, 2005]

- Η Ευκλείδεια απόσταση στους συντελεστές cepstrum σε κλίμακα Mel (**E-MFCC**) [Vepa, 2006; Bellegarda, 2006; Stylianou, 2001; Klabbers, 2001]

- Η απόσταση Mahalanobis στους συντελεστές cepstrum σε κλίμακα Mel (**Mah-MFCC**) [Vepa, 2004; Donovan, 2001]

- Η απόσταση Itakura που υπολογίζεται από τους συντελεστές γραμμικής πρόβλεψης (***ITAK***) [Rabiner, 1993; Donovan, 2001]

Οι παραπάνω αναπαραστάσεις μαζί με τις αντίστοιχες φασματικές αποστάσεις έχουν χρησιμοποιηθεί και αξιολογηθεί κατά κόρο στον τομέα της επεξεργασίας φωνής [Rabiner, 1993]. Καθεμία από αυτές τις φασματικές αποστάσεις παρουσιάζει ένα σύνολο από πλεονεκτήματα και μειονεκτήματα αναφορικά με την ικανότητα διαχωρισμού των φασματικών ασυνεχειών. Ωστόσο, όπως αναφέρθηκε και στο κεφάλαιο 2, καμία από μόνη της δεν επιτυγχάνει την αποτελεσματική εκτίμηση και αποτίμηση των φασματικών ασυνεχειών. Το γεγονός αυτό θα φανεί και στην συνέχεια, καθώς αρχικά πειράματα εκπαίδευσης των ταξινομητών μιας τάξης με μια μόνο φασματική απόσταση δεν επέφερε επιπλέον κέρδος πέρα αυτού που επιτυγχάνεται με χρήση της φασματικής απόστασης από μόνη της ως κόστος στον αλγόριθμο επιλογής. Συνεπώς, για να επιτευχθεί επαυξημένη απόδοση με την χρήση ταξινομητών μιας τάξης, χρειάζεται ένα επαυξημένο και πιο αποδοτικό διάνυσμα χαρακτηριστικών, γεγονός που υποδεικνύει τον συνδυασμό των φασματικών αποστάσεων σε ένα διάνυσμα χαρακτηριστικών. Η σύμμειξη αυτή οδηγεί σε μια εμπλουτισμένη αναπαράσταση συνδυάζοντας τα επιμέρους πλεονεκτήματα τους, οδηγώντας με αυτό τον τρόπο σε μια πιο εύρωστη απόσταση (meta-measure). Επιπρόσθετα, με την μεθοδολογία και χρήση ταξινόμησης μιας τάξης δημιουργείται ένα αντικειμενικό μοντέλο αναφοράς της ζητούμενης τάξης (target class) όπως προκύπτει από τα υπάρχοντα δεδομένα. Η προσέγγιση αυτή συνάδει άμεσα με την τεχνολογία σύνθεσης φωνής με επιλογή και συρραφή ακουστικών μονάδων.

Στο πλαίσιο της διατριβής εστιάσαμε σε στατιστικούς ταξινομητές μιας τάξης και συγκεκριμένα διερευνήθηκε ο ταξινομητής μιας τάξης που βασίζεται σε μείγμα κατανομών Gauss – *Gaussian Mixture Model One-Class Classifier (**OCC-GMM**)*. Όπως αναφέρθηκε στην αρχή του κεφαλαίου, για αυτού του τύπου ταξινομητές απαιτείται η εκτίμηση ενός πιθανοτικού μοντέλου που θα περιγράφει την τάξη

«στόχο». Ο ταξινομητής OCC-GMM προκύπτει από γραμμικό συνδυασμό συναρτήσεων πυκνότητας πιθανότητας Gauss.

Πιο συγκεκριμένα, ο ταξινομητής OCC-GMM εκφράζεται ως:

$$p_{OCC-GMM}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N}} \sum_{k=1}^{K} a_k \frac{1}{\sqrt{\det(\Sigma_k)}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right) \quad (5.5)$$

όπου $\mathbf{x}$ το διάνυσμα χαρακτηριστικών ενός αντικειμένου (object), K ο αριθμός των κατανομών Gauss, $\boldsymbol{\mu}_k$ και Σ_k το διάνυσμα της μέσης τιμής και ο πίνακας συμμεταβλητότητας κάθε κατανομής Gauss αντίστοιχα, N η διάσταση των δεδομένων και a_k οι συντελεστές μίξης των κατανομών.

Η ταξινόμηση βασίζεται σε κάποιο κατώφλι θ_p , που σχετίζεται με την ποσότητα (κριτήριο) ομοιότητας (proximity measure) $p_{OCC-GMM}(\mathbf{x})$ και αντιστοιχεί στον πιθανοτικό βαθμό (κριτήριο) ομοιότητας (similarity measure) του $\mathbf{x}$ με την τάξη «στόχος». Η τιμή του κατωφλίου καθορίζεται προσδιορίζοντας ένα αποδεκτό ποσοστό λάθους στα δεδομένα εκπαίδευσης (target rejection error) που ορίζουν την ζητούμενη τάξη. Η τιμή αυτή καθορίζεται από τον σχεδιαστή και αποτελεί σημείο ισορρόπησης (trade-off) της τελικής απόδοσης.

Κάποιο αντικείμενο (δείγμα) ταξινομείται ότι ανήκει στην τάξη «στόχος» όταν η πιθανότητα του είναι μεγαλύτερη από το κατώφλι (hard decision),

$$f(\mathbf{x}) = I(p_{OCC-GMM}(\mathbf{x}) > \theta_p) \quad (5.6)$$

όπου η I αποτελεί συνάρτηση ένδειξης (indicator function) και ορίζεται ως:

$$I(A) = \begin{cases} 1, & \text{if A is true} \\ 0, & \text{otherwise} \end{cases} \quad (5.7)$$

όπου $A = p_{OCC-GMM}(\mathbf{x}) > \theta_p$.

Για το πρόβλημα της ανίχνευσης φασματικών ασυνεχειών, ένα αντικείμενο ανήκει στην τάξη «στόχος» όταν δεν υπάρχει φασματική ασυνέχεια ενώ χαρακτηρίζεται ως παραπειστικό αντικείμενο στην αντίθετη περίπτωση. Επίσης να σημειωθεί ότι η ποσότητα (πιθανότητα) $p_{OCC-GMM}(\mathbf{x})$ μπορεί να χρησιμοποιηθεί απευθείας ως συνάρτηση κόστους φασματικής ασυνέχειας (soft decision) στον αλγόριθμο επιλογής ακουστικών μονάδων.

5.3.3 Πειραματική αξιολόγηση - αποτελέσματα

Η βάση δεδομένων προηχογραφημένης φυσικής ομιλίας που χρησιμοποιήθηκε τόσο για την εκπαίδευση όσο και για τα πειράματα αξιολόγησης είναι μέρος του συστήματος σύνθεσης φωνής από κείμενο για την Ελληνική γλώσσα που χρησιμοποιήθηκε στην διατριβή. Πρόκειται για βάση με γυναίκα ομιλητή, με

προσεγμένη και ακριβή επισημείωση με χαρακτηριστικά δειγματοληψίας και κβάντισης 16KHz, 16bits. Στη παρούσα διατριβή η ανάλυση και αξιολόγηση περιορίζεται μόνο σε έμφωνα φωνήματα και συγκεκριμένα σε φωνήεντα, καθώς αυτά είναι περισσότερο επιρρεπή σε φασματικές ασυνέχειες που προκύπτουν από την ένωση τους [Syrdal, 2005; Stylianou, 2001; Pantazis, 2005; Klabbers, 2001; Vepa, 2006; Bellegarda, 2006].

Στο στάδιο εκπαίδευσης, κάθε πραγμάτωση φωνήματος αναλύεται σε πλαίσια διάρκειας 20 ms χωρίς επικάλυψη μεταξύ τους. Για κάθε πλαίσιο υπολογίζονται οι φασματικές αναπαραστάσεις που αναφέρθηκαν στην §5.3.2. Για κάθε πλαίσιο υπολογίζονται 18 συντελεστές γραμμικής πρόβλεψης, το φάσμα ισχύος Fourier με 512 σημεία μέγεθος FFT και 12 συντελεστές MFCC χωρίς να συμπεριλαμβάνεται ο μηδενικός συντελεστής. Η φασματική περιβάλλουσα που προκύπτει από τους συντελεστές γραμμικής πρόβλεψης υπολογίζεται με μέγεθος FFT 512 σημεία. Σε κάθε περίπτωση εφαρμόζεται παράθυρο τύπου Hamming και φίλτρο προ-έμφασης με συντελεστή 0.97.

Για τη εκπαίδευση του ταξινομητή OCC-GMM, το αποδεκτό ποσοστό λάθους στα δεδομένα εκπαίδευσης (*target rejection error* or *FN - False Negative*) ορίστηκε σε 15% (βλ. §5.2.2). Από αυτό το ποσοστό λάθους προκύπτει και το κατώφλι που θα χρησιμοποιηθεί κατά την ταξινόμηση. Η τάξη του μοντέλου για κάθε ταξινομητή σε κάθε φώνημα (δηλ. ο αριθμός των κατανομών Gauss ανά ταξινομητή ανά φώνημα) επιλέχθηκε με αυτόματο τρόπο και συγκεκριμένα μέσω του αλγόριθμου που περιγράφεται στο [Tax, 2004] (*consistency-based model order selection criterion*). Σύμφωνα με αυτόν τον αλγόριθμο, η τάξη του μοντέλου αυξάνεται διαρκώς μέχρι ο ταξινομητής να μην συμπεριφέρεται με συνέπεια (inconsistent), με την έννοια να μην σέβεται τις προδιαγραφές σχεδίασης οι οποίες θέτουν το αποδεκτό ποσοστό σε λανθασμένη ταξινόμηση αντικειμένων που ανήκουν στην τάξη «στόχο» (δηλ. το ποσοστό/ρυθμό FN - *False Negative fraction*). Με αυτό τον τρόπο, επιλέγεται εκείνη η (μεγαλύτερη) τάξη μοντέλου, στην οποία ο ταξινομητής παραμένει συνεπής. Στην περίπτωση μας, ο αριθμός των κατανομών Gauss που προέκυψε ήταν της τάξης των 6 έως 14 ανά φώνημα.

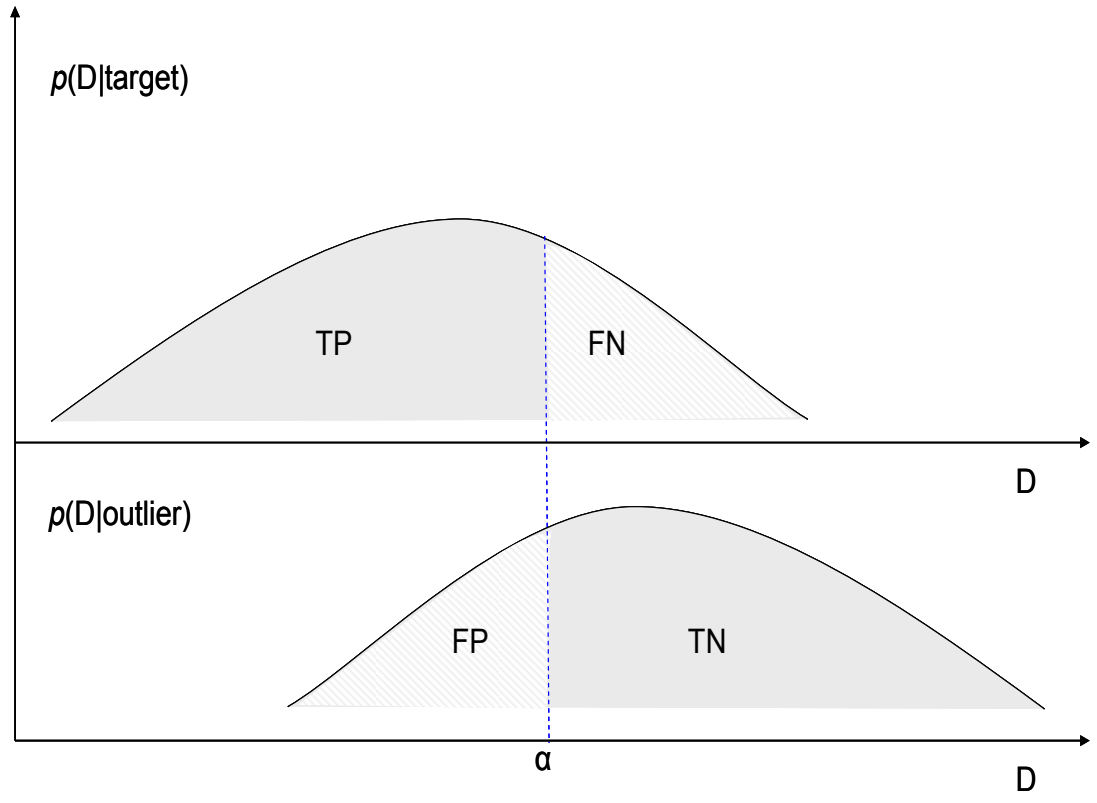
Για την αξιολόγηση της προτεινόμενης μεθόδου ακολουθήθηκε παρόμοια μεθοδολογία με αυτήν που περιγράφεται στα [Stylianou, 2001; Pantazis, 2005; Klabbers, 2001]. Πιο συγκεκριμένα, οργανώθηκαν και πραγματοποιήθηκαν ακουστικά πειράματα στα οποία συμμετείχαν 5 ακροατές με γνώσεις σε τεχνολογίες σύνθεσης φωνής από κείμενο. Στόχος του πειράματος ήταν οι ακροατές να ακούσουν ένα σύνολο από ενώσεις φωνημάτων σε φωνήεντα (vowel concatenations) και να αποφανθούν αν περιέχουν ή όχι φασματικές ασυνέχειες. Η απόκριση τους ήταν δυαδική (συνεχές/ασυνεχές). Σύμφωνα με τις παρατηρήσεις και τα συμπεράσματα από τις μελέτες στα [Stylianou, 2001; Pantazis, 2005; Klabbers, 2001], ο λόγος που επιλέχθηκαν μόνο ειδικοί σε τεχνολογίες σύνθεσης φωνής, είναι η ελαχιστοποίηση παραπειστικών και ασυνεπών αποτελεσμάτων λόγω της δυσκολίας της φύσης αυτού του τύπου ακουστικού πειράματος. Για την εκτέλεση του πειράματος δημιουργήθηκε ειδικό σώμα ακουστικού υλικού (test stimuli) αποτελούμενο από 964 C_iVC_j (*C: consonant set, V: vowel set*) πραγματώσεις, που δημιουργήθηκαν από ενώσεις διφώνων της μορφής C_iV and VC_j από την βάση δεδομένων, διασφαλίζοντας για κάθε περίπτωση ελάχιστη διάρκεια του φωνήεντος 150 ms καθώς και μικρές διακυμάνσεις στην ένταση (intensity) και την θεμελιώδη

συχνότητα (pitch). Σε καμία περίπτωση δεν πραγματοποιήθηκε κάποια τροποποίηση μέσω επεξεργασίας σήματος.

Η διαδικασία του ακουστικού πειράματος χωρίστηκε σε τρία μέρη και διεξήχθη σε τρείς ενότητες, διάρκειας μιας ώρας περίπου η καθεμία. Οι συμμετέχοντες έλαβαν οδηγίες να χρησιμοποιήσουν ακουστικά κατά τη διάρκεια του πειράματος και να εστιάζουν την προσοχή τους στον έμφωνο ήχο και συγκεκριμένα στις μεταβάσεις των φωνηέντων (vowel transitions).

Στο πείραμα, μια ένωση χαρακτηριζόταν με δυαδικό τρόπο είτε ως συνεχής είτε ως ασυνεχής και ο τελικός χαρακτηρισμός προέκυπτε από την πλειοψηφία των αποφάσεων των συμμετεχόντων.

Να σημειωθεί ότι οι συνεχείς ενώσεις αντικατοπτρίζουν την τάξη «στόχος» (target class), ενώ οι ασυνεχείς ενώσεις την παραπειστική τάξη (outlier class). Τα αποτελέσματα χρησιμοποιήθηκαν για την αξιολόγηση τόσο του ταξινομητή μιας τάξης (OCC-GMM) όσο και για κάθε απόσταση χωριστά για λόγους σύγκρισης.



ΣΧΗΜΑ 5.3 Διαδικασία προσδιορισμού καμπύλων ROC για την πειραματική αξιολόγηση του OCC-GMM και των επιμέρους φασματικών αποστάσεων.

Η παραπάνω αξιολόγηση στηρίχθηκε στα γραφήματα τύπου ROC (Receiver Operating Characteristic). Πιο συγκεκριμένα, για κάθε ποσότητα (είτε απόσταση είτε πιθανότητα) D, αρχικά εκτιμώνται οι εξής συναρτήσεις πυκνότητας πιθανότητας: η συνάρτηση πυκνότητας πιθανότητας της ποσότητας D δεδομένου ότι η ένωση χαρακτηρίστηκε ως συνεχής $p(D|target)$ και η συνάρτηση πυκνότητας πιθανότητας της ποσότητας D δεδομένου ότι η ένωση χαρακτηρίστηκε ως ασυνεχής $p(D|outlier)$. Οπότε, σύμφωνα με το δεδομένο πρόβλημα και για δεδομένη τιμή ενός κατωφλίου α, οι ποσότητες *True Positive fraction* ($TP(\alpha)$: περίπτωση στην οποία μια ένωση ορθά ταξινομείται ως συνεχής) και *False Positive fraction* ($FP(\alpha)$: περίπτωση στην οποία μια ένωση λανθασμένα ταξινομείται ως συνεχής) εκφράζονται ως:

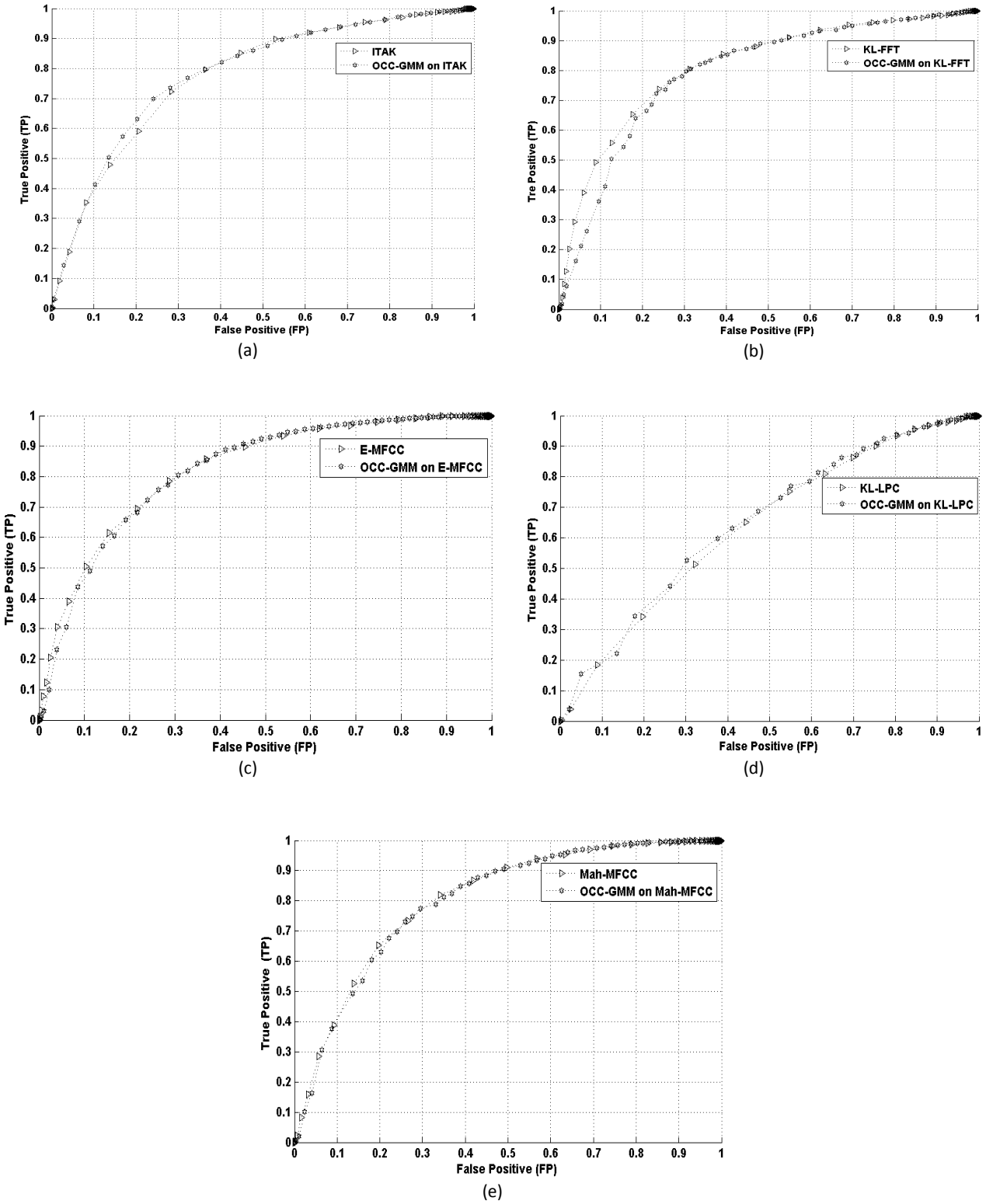
$$TP(\alpha) = \int_{-\infty}^{\alpha} p(D|target)\, dD \tag{5.8}$$

$$FP(\alpha) = \int_{-\infty}^{\alpha} p(D|outlier)\, dD \tag{5.9}$$

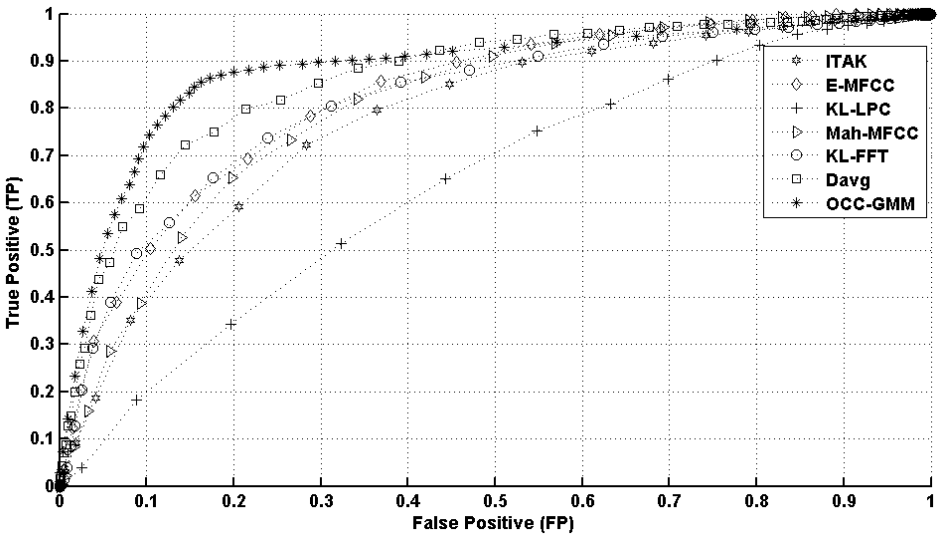
Η διαδικασία αυτή περιγράφεται ποιοτικά στο σχήμα 5.3. Επιπρόσθετα, η περίπτωση στην οποία μια ένωση λανθασμένα ταξινομείται ως ασυνεχής (*False Negative fraction*) είναι *FN=1-TP,* ενώ η περίπτωση στην οποία μια ένωση ορθά ταξινομείται ως ασυνεχής (*True Negative fraction*) είναι *TN=*1*-FP.* Η καμπύλη ROC σχηματίζεται από τη γραφική παράσταση του $TP(\alpha)$ με το $FP(\alpha)$, μεταβάλλοντας την τιμή του κατωφλίου α.

Στο σχήμα 5.4 παρουσιάζονται οι συνολικές (συνολικά για όλες τις ενώσεις) συγκριτικές καμπύλες ROC του ταξινομητή OCC-GMM έναντι κάθε φασματικής απόστασης, για την περίπτωση που ο ταξινομητής εκπαιδεύεται χρησιμοποιώντας μόνο μια φασματική απόσταση. Από τα γραφήματα φαίνεται ότι δεν προκύπτει κάποιο όφελος από την χρήση του ταξινομητή σε σχέση με την περίπτωση που η φασματική απόσταση χρησιμοποιείται σαν κόστος ως έχει. Εντούτοις, παρατηρούμε ότι ο ταξινομητής αποδίδει εξίσου καλά, αφού συμπεριφέρεται ισοδύναμα και ακολουθεί την απόδοση κάθε απόστασης. Συνεπώς, ο ταξινομητής δυνητικά «αφομοιώνει» την πληροφορία που παρέχεται από την απόσταση αναφορικά με το γεγονός να διακρίνει φασματικές ασυνέχειες. Από την άλλη πλευρά, ο ταξινομητής δεν μπορεί να παρέχει παραπάνω πληροφορία πέρα από αυτή που περιλαμβάνεται στα χαρακτηριστικά στα οποία εκπαιδεύεται.

Οι παρατηρήσεις αυτές παρέχουν σαφές κίνητρο για την δημιουργία ενός επαυξημένου διανύσματος χαρακτηριστικών όπως για παράδειγμα αυτό που προκύπτει από τον συνδυασμό πολλαπλών φασματικών αποστάσεων. Τα αποτελέσματα για την περίπτωση αυτή, στην οποία ο ταξινομητής έχει εκπαιδευθεί στο επαυξημένο διάνυσμα των φασματικών αποστάσεων, αποτυπώνονται στο σχήμα 5.5. Το σχήμα περιλαμβάνει και την απόδοση κάθε απόστασης καθώς και την καμπύλη ROC που προκύπτει από τον συνδυασμό των αποστάσεων παίρνοντας το σταθμισμένο άθροισμα τους με ίσους συντελεστές στάθμισης.



ΣΧΗΜΑ 5.4 Συγκριτικές καμπύλες ROC που αφορούν την αξιολόγηση του ταξινομητή OCC-GMM στην περίπτωση που εκπαιδεύεται σε μια μόνο απόσταση, έναντι της απόδοσης κάθε απόστασης: (a) Itakura vs. OCC-GMM στην Itakura, (b) KL-FFT vs. OCC-GMM στην KL-FFT, (c) E-MFCC vs. OCC-GMM στην E-MFCC, (d) KL-LPC vs. OCC-GMM στην KL-LPC και, (e) Mah-MFCC vs. OCC-GMM στην Mah-MFCC. Οι καμπύλες αφορούν την απόδοση συνολικά για όλα τα φωνήματα.



ΣΧΗΜΑ 5.5 Συγκριτικές καμπύλες ROC που αφορούν την αξιολόγηση του ταξινομητή OCC-GMM στην περίπτωση που εκπαιδεύεται στο διάνυσμα χαρακτηριστικών με όλες τις αποστάσεις. Για λόγους σύγκρισης, στο διάγραμμα φαίνονται τόσο οι καμπύλες ROC κάθε απόστασης χωριστά, όσο και του σταθμισμένου αθροίσματος (*Davg*) τους με ίσους συντελεστές στάθμισης. Οι καμπύλες αφορούν τα συνολικά αποτελέσματα που προκύπτουν από όλα τα φωνήματα (phoneme-independent).

Από τις καμπύλες ROC προκύπτει ότι η χρήση ταξινόμησης μιας τάξης μέσω του OCC-GMM επιφέρει σημαντική βελτίωση έναντι κάθε απόστασης ξεχωριστά. Για 10% ποσοστό FP προκύπτει ποσοστό TP περίπου 70%, μια βελτίωση της τάξης του 20% έναντι των αποστάσεων KL-FFT και E-MFCC οι οποίες είναι αυτές που αποδίδουν καλύτερα συγκριτικά με τις υπόλοιπες. Επιπρόσθετα, το αποτέλεσμα αυτό υποδεικνύει ένα ποσοστό της τάξης του 90% σωστής ανίχνευσης φασματικών ασυνεχειών (ποσοστό TN) με αστοχία 30% (trade-off) στην αναγνώριση συνεχών ενώσεων (ποσοστό FN). Επιπλέον, η αξιοποίηση του OCC-GMM επιφέρει βελτίωση της τάξης του 10% σε σχέση με το σταθμισμένο άθροισμα των αποστάσεων. Τα αποτελέσματα αυτά επικυρώνουν την αποτελεσματικότητα του OCC-GMM και κατά επέκταση την αποτελεσματικότητα του προτεινόμενου μεθοδολογικού πλαισίου για την εκτίμηση φασματικών ασυνεχειών στον αλγόριθμο επιλογής ακουστικών μονάδων.

Στον πίνακα 5.2, παρουσιάζονται τα αποτελέσματα που αφορούν την αξιολόγηση της απόδοσης τόσο του OCC-GMM όσο και των επιμέρους φασματικών αποστάσεων ανά φώνημα. Η αξιολόγηση αφορά το ποσοστό TP που επιτυγχάνεται για δεδομένο ποσοστό FP ίσο με 10%. Σε κάθε περίπτωση με χρήση του ταξινομητή OCC-GMM επιτυγχάνονται σημαντικά καλύτερα ποσοστά ανίχνευσης αφενός των φασματικών ασυνεχειών και αφετέρου των ενώσεων που είναι συνεχής. Επίσης, σύμφωνα με την ανθρώπινη αντίληψη, τόσο από τα αποτελέσματα ανά φώνημα όσο και από τα συνολικά αποτελέσματα που παρουσιάστηκαν παραπάνω, προκύπτει ότι καμία φασματική απόσταση από μόνη της δεν αποδίδει επαρκώς για

την ανίχνευση των φασματικών ασυνεχειών και κατ΄ επέκταση για την υιοθέτηση της ως κόστος στον αλγόριθμο επιλογής ακουστικών μονάδων. Το αποτέλεσμα αυτό είναι σε συμφωνία με μελέτες που έχουν παρουσιαστεί στην διεθνή βιβλιογραφία [Vepa, 2006; Bellegarda, 2006; Stylianou, 2001; Klabbers, 2001; Pantazis, 2005; Donovan, 2004].

ΠΙΝΑΚΑΣ 5.2 Συγκριτικά αποτελέσματα αξιολόγησης του OCC-GMM έναντι των επιμέρους φασματικών αποστάσεων ανά φώνημα, σε σχέση με το ποσοστό TP που επιτυγχάνεται για δεδομένο ποσοστό FP ίσο με 10%.

	Phoneme				
Measure	/a/	/e/	/o/	/i/	/u/
OCC-GMM	**74%**	**70%**	**59%**	**84%**	**86%**
E-MFCC	54%	55%	34%	58%	34%
KL-FFT	53%	51%	43%	42%	59%
ITAK	54%	30%	25%	62%	26%
Mah-MFCC	49%	52%	23%	51%	33%
KL-LPC	26%	15%	23%	36%	10%

ΠΙΝΑΚΑΣ 5.3 Εξάρτηση του ταξινομητή OCC-GMM από το μέγεθος της βάσης δεδομένων. Τα ποσοστά αφορούν το ποσοστό TP που επιτυγχάνεται για δεδομένο ποσοστό FP ίσο με 10%.

	Phoneme				
Corpus Size	/a/	/e/	/o/	/i/	/u/
100%	74%	70%	59%	84%	86%
50%	73%	65%	57%	81%	82%
20%	70%	63%	55%	79%	65%

Ενδιαφέρον παρουσιάζει η εξάρτηση της απόδοσης και η ευρωστία που επιτυγχάνει ο ταξινομητής OCC-GMM σε σχέση με το μέγεθος της διαθέσιμης βάσης δεδομένων φυσικής ομιλίας. Αναμένεται ότι όσο μεγαλύτερης διάρκειας είναι η διαθέσιμη βάση τόσο πιο εύρωστος να είναι ο ταξινομητής, καθώς εκπαιδεύεται σε περισσότερα δεδομένα με αποτέλεσμα να επιτυγχάνει την ακριβέστερη (στατιστική)

μοντελοποίηση του ζητούμενου χώρου (του χώρου των 'καλών' ενώσεων). Ο πίνακας 5.3, παρουσιάζει τα αποτελέσματα ανά φώνημα που αφορούν το ποσοστό TP που επιτυγχάνεται όταν μεταβάλλεται η διάρκεια (άρα και το μέγεθος) της βάσης δεδομένων, για δεδομένο ποσοστό FP ίσο με 10%.

Πράγματι από τα αποτελέσματα φαίνεται ότι όσο η διάρκεια της διαθέσιμης βάσης δεδομένων αυξάνει τόσο η απόδοση του ταξινομητή βελτιώνεται.

5.4 ΣΥΝΟΨΗ - ΣΥΜΠΕΡΑΣΜΑΤΑ

Στο κεφάλαιο αυτό εξετάστηκε το προτεινόμενο μεθοδολογικό πλαίσιο εκτίμησης φασματικών ασυνεχειών με χρήση ταξινόμησης μιας τάξης. Παρουσιάστηκε τόσο η γενική μεθοδολογία όσο και η εφαρμογή της στο πλαίσιο της διατριβής. Το προτεινόμενο πλαίσιο αποτελεί νέο παράδειγμα εκτίμησης και υπολογισμού του φασματικού κόστους στη τεχνολογία σύνθεσης φωνής με επιλογή και συρραφή ακουστικών μονάδων. Παρουσιάζει σημαντικά πλεονεκτήματα καθώς βασίζεται και εκμεταλλεύεται αμιγώς τα διαθέσιμα δεδομένα ενώ παράλληλα ξεπερνά τα μειονεκτήματα που χαρακτηρίζουν τις συνηθισμένες πρακτικές. Το μεθοδολογικό πλαίσιο αποτελεί γενικευμένη και επεκτάσιμη πρόταση. Τα συμπεράσματα αυτά ενισχύονται από την πειραματική αξιολόγηση που πραγματοποιήθηκε στο πλαίσιο της διατριβής και που ανέδειξε την σημαντική υπεροχή της προτεινόμενης προσέγγισης έναντι των συνήθων μεθοδολογιών που ακολουθούνται. Τα αποτελέσματα αυτά αποτελούν εφαλτήριο για περισσότερη έρευνα για την εφαρμογή των ταξινομητών μιας τάξης στο πλαίσιο τόσο του αλγόριθμου επιλογής ακουστικών μονάδων όσο και της τεχνολογίας σύνθεσης φωνής γενικότερα.

ΚΕΦΑΛΑΙΟ
-6-
ΕΦΑΡΜΟΓΕΣ

ΚΕΦΑΛΑΙΟ 6 – ΕΦΑΡΜΟΓΕΣ

Όπως αναφέρθηκε και στο πρώτο κεφάλαιο, η ανάπτυξη συστημάτων σύνθεσης φωνής, τροφοδοτεί πλήθος εργαλείων και εφαρμογών. Από βοηθήματα υποστηρικτικής τεχνολογίας, μέχρι εφαρμογές γλωσσικής εκπαίδευσης, τηλεπικοινωνιακές εφαρμογές, εφαρμογές μέσω διαδικτύου και μέσω τηλεφώνου, εργαλεία αυτοματισμού γραφείου, εργαλεία εναλλακτικής πρόσβασης σε πληροφορία, παιχνίδια και πολλά άλλα. Η τεχνολογία σύνθεσης φωνής συνεισφέρει σημαντικά σε τομείς όπως η υποστήριξη της πρόσβασης στην πληροφορία και το περιεχόμενο και η διευκόλυνση της καθημερινής επικοινωνίας και ενημέρωσης. Βασικός παράγοντας για τα παραπάνω είναι η επίτευξη υψηλής ποιότητας συνθετικής ομιλίας. Χωρίς αυτό το ιδιαίτερο ποιοτικό χαρακτηριστικό, η τεχνολογία βρίσκει κλειστό το δρόμο προς την αξιοποίηση της και κάθε σύστημα δεν έχει άλλο ρόλο πέρα από τη χρήση του ως ένα εργαστηριακό πρωτότυπο. Από την άλλη πλευρά, η επίτευξη υψηλής ποιότητας και σχεδόν φυσικής συνθετικής φωνής, ανοίγει το δρόμο προς την αξιοποίηση της σε πληθώρα καθημερινών εφαρμογών. Ωστόσο δεν αρκεί μόνο αυτή η συνιστώσα στο σύγχρονο περιβάλλον της ηλεκτρονικής αλληλεπίδρασης. Η ενότητα αυτή αφορά τις εφαρμογές της τεχνολογίας σύνθεσης φωνής από κείμενο. Εξετάζονται και περιγράφονται καινοτόμες εφαρμογές στις οποίες ενσωματώνεται το σύστημα σύνθεσης φωνής που εξετάστηκε στην παρούσα διατριβή, και οι οποίες καλύπτουν ένα εύρος από τηλεπικοινωνιακά συστήματα και τηλεπικοινωνιακές και διαδικτυακές υπηρεσίες.

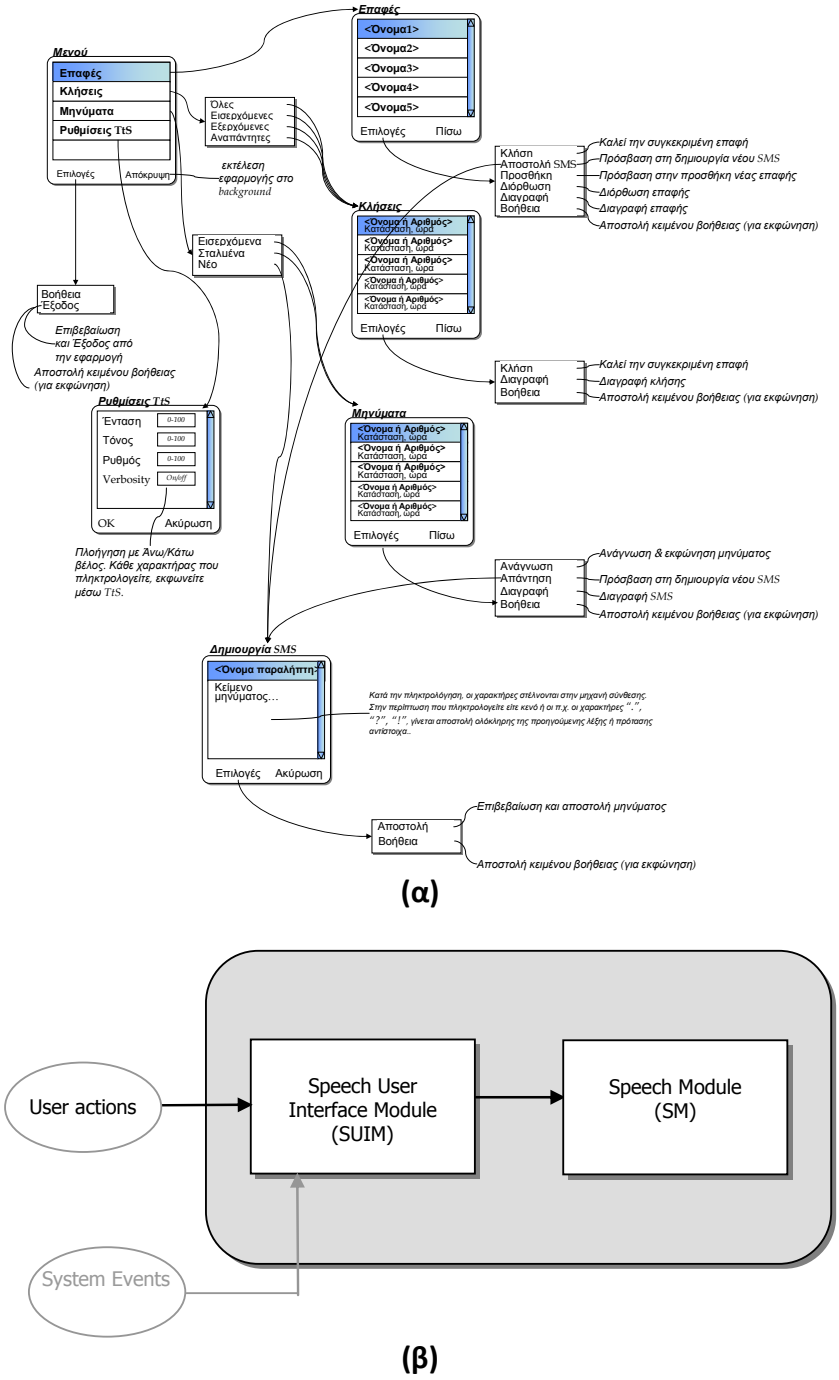
6.1 ΤΗΛΕΠΙΚΟΙΝΩΝΙΑΚΕΣ ΚΑΙ ΔΙΚΤΥΑΚΕΣ ΕΦΑΡΜΟΓΕΣ ΣΥΝΘΕΣΗΣ ΦΩΝΗΣ

Στη σημερινή εποχή του παγκόσμιου ιστού, των ενοποιημένων τηλεπικοινωνιακών υποδομών και δικτύων αλλά και των προσωπικών φορητών συσκευών, ιδιαίτερο ενδιαφέρον παρουσιάζει η διασύνδεση διαδικτυακού περιεχομένου και τηλεπικοινωνιακών υπηρεσιών με τεχνολογίες σύνθεσης φωνής, η ολοκλήρωση της τεχνολογίας σε πλήθος από υπολογιστικά περιβάλλοντα αλλά και η σχεδίαση αποτελεσματικών διεπαφών αλληλεπίδρασης ανθρώπου-μηχανής μέσω φωνής. Με γνώμονα την υψηλή ποιότητα, η τεχνολογία σύνθεσης φωνής πλέον συναντάται ως εφαρμογή τόσο σε φορητές ηλεκτρονικές συσκευές όσο και σε διαδικτυακές και τηλεπικοινωνιακές υπηρεσίες. Μερικές από αυτές, που υλοποιήθηκαν με τη βοήθεια της μηχανής σύνθεσης που χρησιμοποιήθηκε και στην παρούσα διατριβή, παρουσιάζονται στην συνέχεια.

6.6.1 Επίπεδο συσκευής

Με την μεταφορά και την αποδοτική ολοκλήρωση τεχνολογίας σύνθεσης φωνής υψηλής ποιότητας σε περιβάλλον συσκευής κινητής τηλεφωνίας, κατέστη δυνατή η υλοποίηση φωνητικής διεπαφής και ανάγνωσης οθόνης των περιεχομένων κινητού τηλεφώνου. Τέτοιο περιεχόμενο είναι για παράδειγμα τα γραπτά μηνύματα, η λίστα ονομάτων, το μενού πλοήγησης κτλ. Με αυτόν τον τρόπο, σε επίπεδο συσκευής η ολοκλήρωση της τεχνολογίας (βλ. Κεφ. 3) συνοδεύεται και από τον συστηματικό

σχεδιασμό περιβάλλοντος φωνητικής διεπαφής για κινητά τηλέφωνα, η οποία σε συνδυασμό με τις λειτουργίες και τα τεχνικά χαρακτηριστικά του λειτουργικού συστήματος οδηγεί σε ένα πλήρες σύστημα αναγνώστη οθόνης (screen reader) για κινητό τηλέφωνο.



ΣΧΗΜΑ 6.1 Παράδειγμα διάταξης και διασύνδεσης των επιλογών της φωνητικής διεπαφής

Για παράδειγμα, το σχήμα 6.1 παρουσιάζει την διασύνδεση και την δενδρική δομή των επιλογών που προσφέρει η φωνητική διεπαφή μέσω φωνητικής απόκρισης.

Η εφαρμογή της φωνητικής διεπαφής προσφέρει αλληλεπίδραση του χρήστη με τις βασικές λειτουργίες της συσκευής, μέσω ομιλούντων επιλογών (ομιλούντα μενού), παρέχοντας, μεταξύ άλλων, τις εξής βασικές δυνατότητες:

- Πρόσβαση και ανάγνωση του τηλεφωνικού καταλόγου. Αφορά τις βασικές λειτουργίες στον τηλεφωνικό κατάλογο, όπως κλήση, διαγραφή, προσθήκη και αποστολή μηνύματος. Στον χρήστη εκφωνούνται όλες οι βασικές πληροφορίες που αφορούν μια επαφή στον τηλεφωνικό κατάλογο (π.χ. όνομα και τηλέφωνο).

- Πρόσβαση και ανάγνωση στις λίστες εισερχομένων, εξερχόμενων και αναπάντητων κλήσεων. Αφορά τις βασικές λειτουργίες στην κατηγορία των κλήσεων καθώς και υποστήριξη επιπλέον λειτουργιών – μέσω επιπλέον ομιλούντων επιλογών – όπως προβολή, κλήση και διαγραφή.

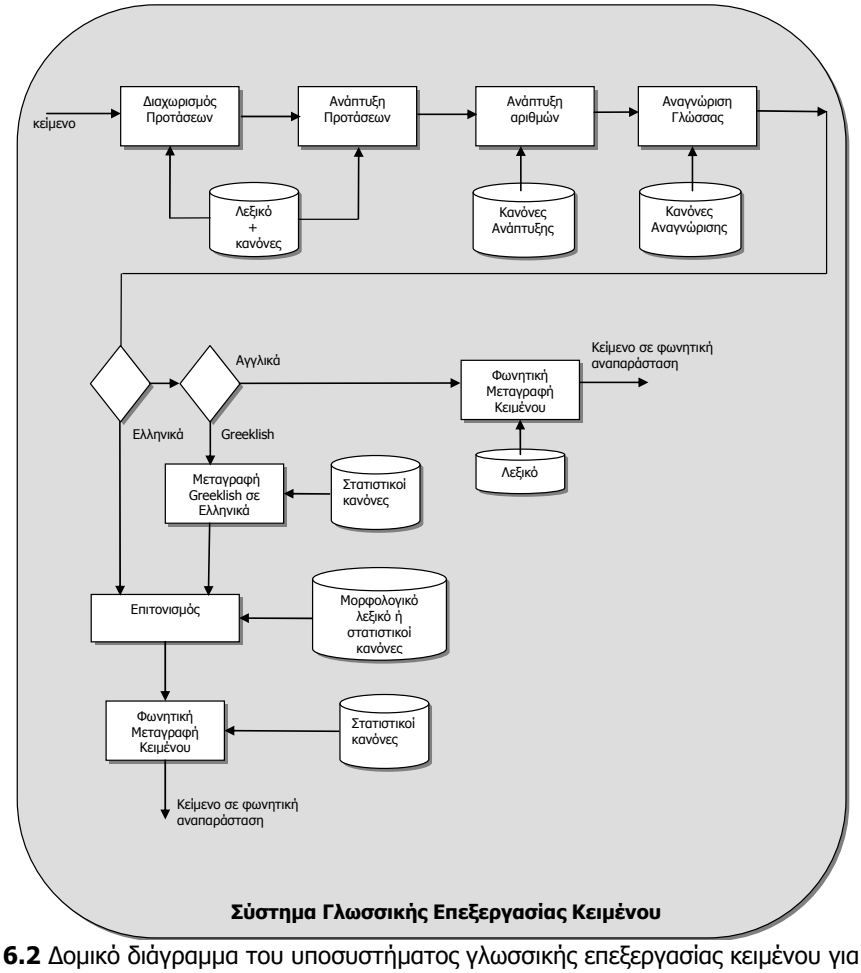
- Πρόσβαση και ανάγνωση στις λίστες γραπτών μηνυμάτων. Αφορά τις λίστες εισερχομένων και απεσταλμένων μηνυμάτων καθώς και την δημιουργία νέου μηνύματος. Σε κάθε επιλογή, παρέχονται επιπλέον δυνατότητες (επιπλέον ομιλούντα μενού) όπως ανάγνωση, διαγραφή και απάντηση των μηνυμάτων, καθώς και υποβοήθηση στη σύνταξη ενός νέου μηνύματος μέσω ανάγνωση κάθε λέξης που θα πληκτρολογεί ο χρήστης.

- Πρόσβαση και ανάγνωση των ρυθμίσεων των παραμέτρων της μηχανής σύνθεσης φωνής από κείμενο. Τέτοιες ρυθμίσεις είναι ο ρυθμός, η ένταση, ο τόνος και η ταχύτητα της συνθετικής φωνής.

- Πρόσβαση και ανάγνωση των ρυθμίσεων των παραμέτρων της γλωσσικής επεξεργασίας κειμένου (κανονικοποίηση κειμένου). Παρέχονται δυνατότητες για την ανάγνωση των σημείων στίξης και υποστηρίζεται η ύπαρξη λεξικού χρήστη για συντομογραφίες, ειδικές λέξεις κ.α.

Σε γενικές γραμμές, η εφαρμογή έχει την δυνατότητα παρακολούθησης, ανάκτησης και κατόπιν αποστολής σε μορφή κειμένου προς την μηχανή σύνθεσης φωνής, όλης την πληροφορία σχετικά με αλλαγή ή συμβάν στην γραφική διεπαφή (UI-user interface) της εφαρμογής. Τέτοιο συμβάν ή αλλαγή, είναι η πληκτρολόγηση χαρακτήρων, η μεταβολή της εστίασης σε λίστα επιλογών, η αλλαγή στην επιφάνεια εργασίας (new view) κ.α. Η εφαρμογή μπορεί να ενεργοποιείται αυτόματα αλλά και μέσω χρήσης ειδικού συνδυασμού πλήκτρων. Η απενεργοποίηση γίνεται επίσης και με χρήση συνδυασμού πλήκτρων. Η εφαρμογή της φωνητικής διεπαφής εγκαθίσταται στις εφαρμογές της συσκευής και κατά την εκτέλεση της, προβάλλει και θα αναγιγνώσκει στον χρήστη τις εστιασμένες επιλογές από το μενού επιλογών. Το τελευταίο χαρακτηρίζεται από δενδρική διάταξη ακολουθώντας έτσι την συνήθη πρακτική των εφαρμογών συσκευών κινητής τηλεφωνίας. Επιπλέον, η πλοήγηση του χρήστη στην εφαρμογή και τις λειτουργίες τις πραγματοποιείται σε συμφωνία με τα πλήκτρα της συσκευής (CBA buttons and Arrow keys). Σύμφωνα με αυτήν την διάταξη, η πλοήγηση και η πρόσβαση του χρήστη, πραγματοποιείται με την βοήθεια των πλήκτρων πλοήγησης της συσκευής του κινητού τηλεφώνου ενώ

παράλληλα εκφωνείται η εστιασμένη επιλογή. Η εκτέλεση κάθε επιλογής συνοδεύεται από επιπλέον μενού επιλογών, καθιστώντας έτσι λειτουργική την χρήση του κινητού τηλεφώνου μέσω της φωνητικής διεπαφής. Για παράδειγμα, αν ο χρήστης επιλέξει «μηνύματα» η επιλογή θα εκφωνηθεί και στο μενού επιλογών θα εμφανιστούν οι επόμενες δυνατότητες. Με χρήση των πλήκτρων πλοήγησης ο χρήστης εστιάζει και ακούει τις δυνατές επιλογές, ενώ με το πλήκτρο επιλογής, εκτελεί κάποια από αυτές. Συνεχίζοντας το παράδειγμα, η επιλογή «εισερχόμενα», θα εκφωνηθεί και θα οδηγήσει σε πρόσβαση της λίστας των εισερχομένων μηνυμάτων. Στη συνέχεια, θα εκφωνούνται στοιχεία του πρώτου μηνύματος στη λίστα (π.χ. όνομα αποστολέα και ημερομηνία). Όμοια, με χρήση των πλήκτρων πλοήγησης, ο χρήστης μετακινείται στα επόμενα μηνύματα. Η επιλογή κάποιου μηνύματος οδηγεί στην εμφάνιση και εκφώνηση επιπλέον μενού επιλογών που εκφωνούνται ανάλογα. Στην περίπτωση της λίστας εισερχομένων μηνυμάτων, προσφέρονται οι επιλογές «Ανάγνωση», «Διαγραφή», «Απάντηση» και «Βοήθεια» η εκτέλεση των οποίων θα ενεργοποιεί την ανάλογη δραστηριότητα.



ΣΧΗΜΑ 6.2 Δομικό διάγραμμα του υποσυστήματος γλωσσικής επεξεργασίας κειμένου για σύνθεση φωνής από κείμενο για αναγνώστες οθόνης [Chalamandaris, 2010].

Σε αντιστοιχία με το περιβάλλον συσκευής κινητού τηλεφώνου, η τεχνολογία σύνθεσης φωνής αποτελεί κύριο συστατικό και σε υπολογιστικά περιβάλλοντα Η/Υ. Πέρα από την απλή της εφαρμογή, η ολοκλήρωση τεχνολογίας σύνθεσης φωνής υψηλής ποιότητας με αναγνώστες οθόνης (screen readers) σε περιβάλλον Η/Υ,

αποτελεί και βασικό υποστηρικτικό εργαλείο για ΑΜΕΑ. Στην περίπτωση αυτή, πέρα από την υψηλή ποιότητα σύνθεσης (τόσο σε φυσικότητα όσο και σε καταληπτότητα), πρέπει να πληρούνται ένα σύνολο από επιπλέον προϋποθέσεις ώστε η τεχνολογία να καταστεί αποδεκτή και εφαρμόσιμη [Chalamandaris, 2010]:

- Εύρωστη γλωσσική επεξεργασία και διαχείριση κάθε είδους κειμένου και γραφής (π.χ., Greeklish)
- Ίδια φωνή για υποστήριξη ένθετων ξένων λέξεων (π.χ., Ελληνικά-Αγγλικά)
- Δυνατότητα για πολύ γρήγορη εκφώνηση
- Ταχύτητα εκτέλεσης και απόκρισης (Speed optimization and Response Latency)

Για την αντιμετώπιση των απαιτήσεων ταχύτητας και απόκρισης, εφαρμόζονται παρόμοιες τεχνικές με αυτές που περιγράφηκαν στο κεφάλαιο 3. Το σχήμα 6.2, δείχνει το δομικό διάγραμμα της βαθμίδας γλωσσικής επεξεργασίας κειμένου για την περίπτωση συστήματος σύνθεσης φωνής για αναγνώστες οθόνης.

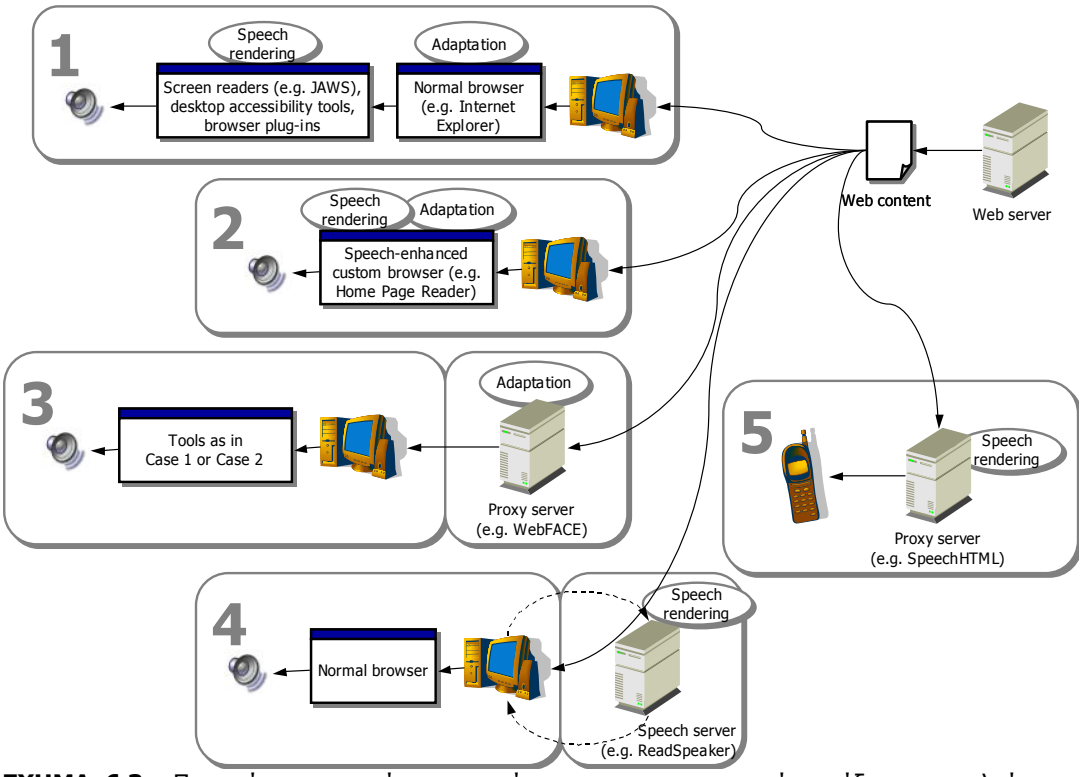
6.6.2 Επίπεδο τηλεπικοινωνιακών και δικτυακών εφαρμογών

Η ωρίμανση της τεχνολογίας σύνθεσης φωνής, έχει οδηγήσει πλέον και στην ευρεία αποδοχή της σε συστήματα τηλεφωνικών κέντρων. Χαρακτηριστικά παραδείγματα αποτελούν οι εφαρμογές IVR (interactive voice response) και η επέκταση τους μέσω προχωρημένων πρωτοκόλλων όπως είναι το MRCP (Media Resource Control Protocol), το οποίο χρησιμοποιείται από εξυπηρετητές (servers) για την παροχή υπηρεσιών αναγνώρισης και σύνθεσης φωνής. Επιπρόσθετα, χαρακτηριστική είναι η υιοθέτησή της τεχνολογίας σε πιο προχωρημένες διαδικτυακές εφαρμογές, ως βασική βαθμίδα για ενισχυμένη αλληλεπίδραση ανθρώπου-μηχανής. Στην περίπτωση αυτή, μερικές από τις πιθανές εφαρμογές αποτελούν,

- Η φωνητική επαύξηση και πλοήγηση σε διαδικτυακό περιεχόμενο (speech enabling web content)
- Το φωνητικό ηλεκτρονικό ταχυδρομείο (speech enabled -email)
- Η μετατροπή ηλεκτρονικών αρχείων σε ακουστικά αρχεία με τη βοήθεια απομακρυσμένων πρακτόρων και ηλεκτρονικού ταχυδρομείου (text-to-speech translation via remote e-mail agents/robots)

Η φωνητική επαύξηση και πλοήγηση σε διαδικτυακό περιεχόμενο αποτελεί αφενός σημαντική συνιστώσα προσβασιμότητας και αφετέρου επιτρέπει πιο φυσική αλληλεπίδραση ανθρώπου-μηχανής. Σε ότι αφορά την φωνητική επαύξηση και πλοήγηση σε διαδικτυακό περιεχόμενο, ένα σύστημα σύνθεσης φωνής μπορεί να είναι διαθέσιμο με δύο τρόπους. Είτε *Τοπικά* ως κομμάτι του λειτουργικού συστήματος, είτε Απομακρυσμένα Στην πρώτη περίπτωση δεν επιφέρει καμία επιβάρυνση στο εύρος ζώνης της σύνδεσης αφού η διαδικασία της σύνθεσης λαμβάνει χώρα τοπικά. Ωστόσο, απαιτεί την εγκατάσταση στον τοπικό υπολογιστή του συστήματος TTS. Στη δεύτερη περίπτωση, η υπηρεσία παρέχεται από ένα εξυπηρετητή φωνής. Αυτή η εναλλακτική λύση δεν απαιτεί τοπική εγκατάσταση, αλλά απαιτεί μεγαλύτερο εύρος ζώνης της σύνδεσης αφού πρέπει να μεταφέρεται σήμα φωνής. Μια σύντομη επισκόπηση των πιο συχνά χρησιμοποιούμενων προσεγγίσεων και περιπτώσεων για προσβασιμότητα, φωνητική επαύξηση και

πλοήγηση σε διαδικτυακό περιεχόμενο με βάση τη φωνή φαίνεται στο Σχήμα 6.3 και παρουσιάζεται στη συνέχεια [Raptis, 2005; Chalamandaris, 2009]:



ΣΧΗΜΑ 6.3 Περιπτώσεις τεχνικών προσεγγίσεων για την φωνητική επαύξηση και πλοήγηση σε διαδικτυακό περιεχόμενο με χρήση σύνθεσης φωνής από κείμενο [Raptis, 2005].

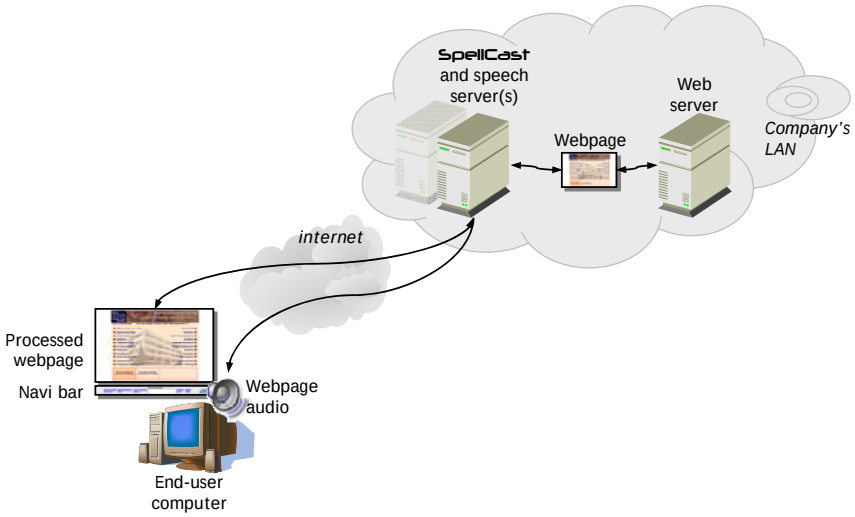
Περίπτωση 1: Αποτελεί την πιο απλή προσέγγιση στο ζήτημα. Χρησιμοποιείται ένας κανονικός πλοηγός (web browser) για την πρόσβαση και την ανάκτηση του διαδικτυακού περιεχομένου, ενώ ένα ξεχωριστό εργαλείο, όπως για παράδειγμα ένας αναγνώστης οθόνης ή ένα εργαλείο πρόσβασης (accessibility tool) ή ένα plug-in στον πλοηγό, αναλαμβάνει την απόδοση του περιεχομένου χρησιμοποιώντας ένα σύστημα TTS. Σε αυτή την περίπτωση, ο πλοηγός είναι υπεύθυνος για την απαραίτητη προσαρμογή του περιεχομένου σύμφωνα με τις ανάγκες και τις προτιμήσεις του χρήστη. Με αυτή την προσέγγιση δεν είναι δυνατή η αναδόμηση ή εξειδικευμένος τρόπος αλληλεπίδρασης αφού δεν δίνεται η δυνατότητα αξιοποίησης πληροφοριών για τη δομή της εκάστοτε σελίδας.

Περίπτωση 2: Μια λίγο διαφορετική προσέγγιση είναι αυτή του εξειδικευμένου πλοηγού. Ο πλοηγός είναι υπεύθυνος για όλες τις διεργασίες: πρόσβαση και ανάκτηση του περιεχομένου ιστού, προσαρμογή αυτού στις προτιμήσεις του χρήστη, και φωνητική απόδοση. Ένα τυπικό παράδειγμα εξειδικευμένου πλοηγού με προσαρμογή και φωνητικές ικανότητες είναι το Home Page Reader της IBM. Επιπλέον, παρέχει υποστήριξη για διαφορετικούς μορφότυπους (format) αρχείων. Κανονικά, ο Home Page Reader δεν υποστηρίζει αναδόμηση ιστοσελίδας, ωστόσο έχουν αναπτυχθεί ειδικές εκδόσεις του που παρέχουν ειδική υποστήριξη σε συγκεκριμένους ιστοτόπους (π.χ. W3C).

Περίπτωση 3: Στην προσέγγιση αυτή, η προσαρμογή μια ιστοσελίδας στην απαιτήσεις χρήστη διεκπεραιώνεται από ένα απομακρυσμένο εξυπηρετητή. Ο εν λόγω εξυπηρετητής κατέχει το προφίλ του εκάστοτε χρήστη και μεσολαβεί ανάμεσα στον παροχέα του περιεχομένου (content provider – web server) και του τοπικού υπολογιστή πελάτη (client). Η φωνητική απόδοση μπορεί να γίνει είτε από ένα αναγνώστη οθόνης ή άλλα εργαλείο προσβασιμότητας (περίπτωση 1), ή με τη χρήση ειδικού πλοηγού (περίπτωση 2). Αναδόμηση και ειδική αλληλεπίδραση δεν πραγματοποιείται σε αυτή την περίπτωση.

Περίπτωση 4: Στην περίπτωση αυτή η φωνητική απόδοση αναλαμβάνεται από κάποιο απομακρυσμένο «εξυπηρετητή φωνής» (“speech server”). Ένας κανονικός πλοηγός χρησιμοποιείται και μια έκδοση της σελίδας με συνθετική φωνής παράγεται δυναμικά με εντολή προς τον εξυπηρετητή. Το πλεονέκτημα είναι ότι δεν απαιτείται τοπική εγκατάσταση συστήματος TTS. Στα μειονεκτήματα περιλαμβάνονται η αυξημένη απαίτηση σε εύρος ζώνης της σύνδεσης, η μη υποστήριξη αναδόμησης των ιστοσελίδων και η αδυναμία αλληλεπίδρασης.

Περίπτωση 5: Η τελευταία προσέγγιση, που εμπίπτει στην κατηγορία της φωνητικής πλοήγησης (voice browsing), αφορά την περίπτωση στην οποία ο χρήστης συνδέεται μέσω φωνητικής διεπαφής με το διαδικτυακό περιεχόμενο, διαμέσου ενός proxy server. Το σύστημα SpeechHTML είναι τυπικό παράδειγμα αυτής της προσέγγισης.

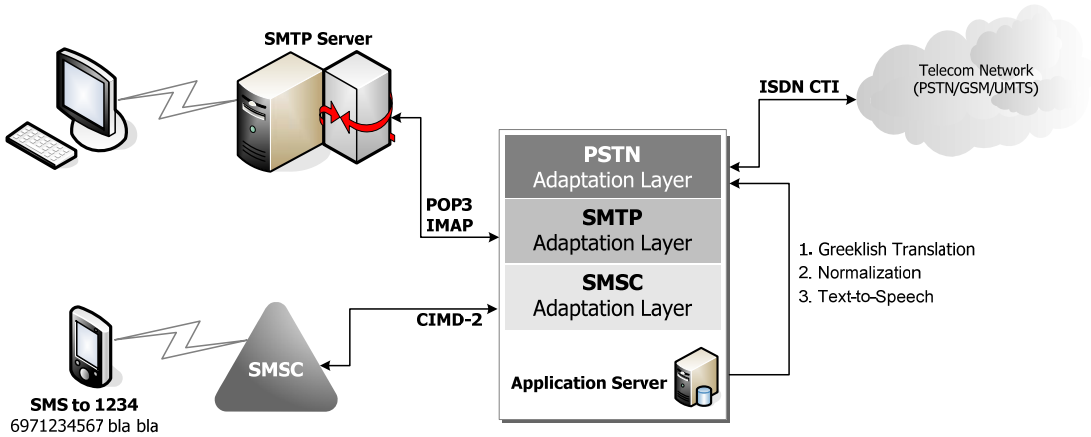


ΣΧΗΜΑ 6.4 Σύγχρονη αρχιτεκτονική συστήματος φωνητική επαύξησης και πλοήγηση σε διαδικτυακό περιεχόμενο [Chalamandaris, 2009].

Μια προσέγγιση, η οποία περιγράφεται στο σχήμα 6.4, δεν εστιάζει μόνο στο να καταστήσει το διαδικτυακό περιεχόμενο προσβάσιμο διαμέσου συνθετικής φωνής, αλλά ταυτόχρονα να προσφέρει πιο αποδοτική παρουσίαση και τρόπους αλληλεπίδρασης, διευκολύνοντας έτσι την διαδικασία πλοήγησης, καθιστώντας την πιο διαισθητική και κατανοητή. Το διαδικτυακό περιεχόμενο αναλύεται, αναδομείται και κατηγοριοποιείται και κατόπιν μετατρέπεται σε συνθετική ομιλία,

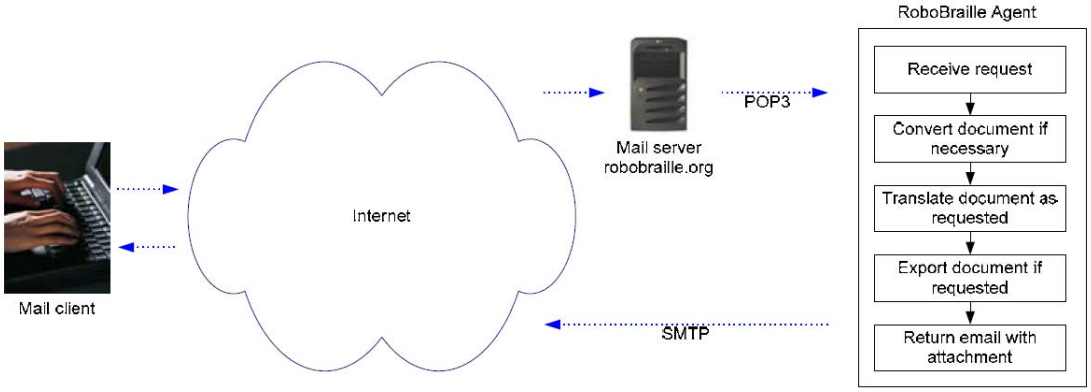
επιδιώκοντας την ακουστική απόδοση του διαδικτυακού περιεχομένου με δομημένο τρόπο. Επιπλέον, η προσέγγιση αυτή επιτρέπει την ενσωμάτωση λειτουργιών αλληλεπίδρασης και πλοήγησης, ενώ δεν απαιτεί τοπική εγκατάσταση λογισμικού σύνθεσης φωνής (server-based).

Η δυνατότητα παροχής υπηρεσιών σύνθεσης φωνής μέσω απομακρυσμένου «εξυπηρετητή φωνής» (speech server) βρίσκει εφαρμογή και στο δίκτυο της κινητής τηλεφωνίας. Σε επίπεδο δικτύου, η ολοκλήρωση και η διασύνδεση με τον εξυπηρετητή (server) καθώς και με τα υπόλοιπα εμπλεκόμενα στοιχεία του δικτύου κινητής τηλεφωνίας, καθιστούν εφικτές εφαρμογές όπως το φωνητικό ηλεκτρονικό ταχυδρομείο. Για παράδειγμα, ένας συνδρομητής θα μπορεί να αποστέλλει ένα γραπτό μήνυμα έχοντας την δυνατότητα επιλογής εναλλακτικών τρόπων παράδοσης, όπως, α) παράδοση μέσω αυτόματης κλήσης του παραλήπτη, στην οποία το δίκτυο θα αναλαμβάνει να πραγματοποιήσει αυτόματα μια κλήση στον αποδέκτη εκφωνώντας του τα περιεχόμενα του γραπτού μηνύματος, β) παράδοση του μηνύματος με την μορφή MMS που θα περιλαμβάνει το εκφωνημένο μήνυμα και γ) αποθήκευση του εκφωνημένου μηνύματος στο φωνητικό ταχυδρομείο του παραλήπτη ώστε να το λάβει όταν εκείνος επιλέξει. Τα παραπάνω προϋποθέτουν την δημιουργία, εγκατάσταση και λειτουργία λογισμικού σύνθεσης φωνής από κείμενο στα στοιχεία του δικτύου που διαχειρίζονται την λήψη και αποστολή των γραπτών μηνυμάτων δηλαδή, σύνδεση με τα κέντρα γραπτών μηνυμάτων κινητής τηλεφωνίας π.χ., Short Message Service Center – SMSC, το κέντρο μηνυμάτων πολυμέσων (Multimedia Message Service Center – MMSC) και το φωνητικό ταχυδρομείο (Voice MailBox – VMB) του αποδέκτη. Επιπλέον, απαιτείται και διεπαφή για υπηρεσίες ηλεκτρονικού ταχυδρομείου (π.χ., διεπαφή με SMTP Server και POP3 client). Στο σχήμα 6.5 παρουσιάζεται η αρχιτεκτονική ολοκλήρωσης σύνθεσης φωνής και εξυπηρετητή φωνής στο δίκτυο της κινητής τηλεφωνίας.



ΣΧΗΜΑ 6.5 Ενσωμάτωση υπηρεσιών σύνθεσης φωνής στο δίκτυο της κινητής τηλεφωνίας.

Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει η δυνατότητα μετατροπής ηλεκτρονικών αρχείων σε ακουστικά αρχεία, μέσω διαδικτύου, με τη βοήθεια απομακρυσμένων πρακτόρων και ηλεκτρονικού ταχυδρομείου [Christensen, 2006]. Η αρχιτεκτονική για την πραγματοποίηση της εφαρμογής αυτής, περιγράφεται στο σχήμα 6.6.



ΣΧΗΜΑ 6.6 Μετατροπή κειμένου σε φωνή μέσω e-mail agent [Christensen, 2006].

Στην εφαρμογή αυτή, ο χρήστης αποστέλλει ένα κείμενο υπό μορφή αρχείου, με ηλεκτρονικό ταχυδρομείο προς έναν απομακρυσμένο εξυπηρετητή ο οποίος με την βοήθεια κάποιου πράκτορα (agent) αναλαμβάνει την μετατροπή αυτού του αρχείου σε ακουστική μορφή μέσω συστήματος σύνθεσης φωνής από κείμενο. Στη συνέχεια, το ακουστικό αρχείο αποστέλλεται στον αρχικό χρήστη. Πέρα από την μετατροπή σε ακουστικό αρχείο, υποστηρίζονται και άλλες υπηρεσίες όπως, αποστολή μεταφρασμένου κειμένου, μετάφραση σε άλλη γλώσσα και μετατροπή σε ακουστικό αρχείο κτλ.

Πέρα από τις παραπάνω περιπτώσεις, η τεχνολογία σύνθεσης φωνής μπορεί να αξιοποιηθεί και σε πλήθος άλλων εφαρμογών. Ο ενδιαφερόμενος αναγνώστης, για περισσότερες πληροφορίες παραπέμπεται στα [O'Shaughnessy, 2003; Deng, 2005; Tomko, 2005; Moore, 2007; Mohasi, 2006; Eskenazi, 2009; Gilbert, 2008; Denby, 2010] και στις εκεί αναφορές.

6.2 ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην ενότητα αυτή παρουσιάστηκαν μερικές καινοτόμες εφαρμογές στις οποίες η σύνθεση φωνής από κείμενο αποτελεί βασική συνιστώσα. Οι εφαρμογές αφορούν τόσο σε επίπεδο Η/Υ και φορητών συσκευών όσο και σε επίπεδο σύγχρονων τηλεπικοινωνιακών και διαδικτυακών υπηρεσιών και εφαρμογών. Η σύνθεση φωνής από κείμενο υψηλής ποιότητας, αποτελεί πλέον βασική βαθμίδα όχι μόνο ως υποστηρικτικό εργαλείο σε εξειδικευμένες εφαρμογές από ειδικές ομάδες ανθρώπων, αλλά ως γενικό συστατικό για επαυξημένη και πιο φυσική αλληλεπίδραση στην επικοινωνία ανθρώπου-μηχανής.

ΚΕΦΑΛΑΙΟ
-7-
ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΚΑΤΕΥΘΥΝΣΕΙΣ ΜΕΛΛΟΝΤΙΚΗΣ ΈΡΕΥΝΑΣ

ΚΕΦΑΛΑΙΟ 7 – ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΚΑΤΕΥΘΥΝΣΕΙΣ ΜΕΛΛΟΝΤΙΚΗΣ ΕΡΕΥΝΑΣ

Η διατριβή πραγματεύεται την μελέτη, διερεύνηση και υλοποίηση τεχνικών και μεθοδολογιών στο αντικείμενο της σύνθεσης φωνής από κείμενο με σκοπό τόσο την βελτίωση της ποιότητας συνθετικής φωνής όσο και την εφαρμογή της σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και υπηρεσίες. Στην ενότητα αυτή συνοψίζονται οι ιδέες και τα σημαντικότερα σημεία συνεισφοράς καθώς και τα κυριότερα συμπεράσματα που προέκυψαν στο πλαίσιο της διδακτορικής διατριβής. Επιπλέον, προτείνονται κατευθύνσεις για περαιτέρω έρευνα στα αντικείμενα της διατριβής με στόχο την περαιτέρω βελτίωση τους.

7.1 ΣΥΝΕΙΣΦΟΡΑ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα διατριβή έχει ως αντικείμενο την τεχνολογία σύνθεσης φωνής από κείμενο και την βελτίωση της ποιότητας που αυτή επιτυγχάνει, με σκοπό την ευρεία υιοθέτηση της σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και τηλεπικοινωνιακές υπηρεσίες. Η συνεισφορά και οι ερευνητικές δραστηριότητες που πραγματοποιήθηκαν στο πλαίσιο της διατριβής, εστιάζουν σε τέσσερις βασικούς άξονες:

- Την σχεδίαση και την υλοποίηση του αλγόριθμου επιλογής ακουστικών μονάδων για σύστημα σύνθεσης φωνής από κείμενο για την Ελληνική γλώσσα (Unit Selection Speech Synthesis).

- Την σχεδίαση και ανάπτυξη τεχνικών και μεθοδολογιών καθώς και την υλοποίηση υψηλής ποιότητας γενικού σκοπού συστήματος σύνθεσης φωνής με επιλογή και ένωση ακουστικών μονάδων σε υπολογιστικό περιβάλλον κινητού τηλεφώνου (Embedded Unit Selection Speech Synthesis).

- Την διερεύνηση παραμετρικών μεθόδων σύνθεσης φωνής για την Ελληνική γλώσσα, που βασίζονται στη αξιοποίηση της μοντελοποίησης με Κρυφά Μαρκοβιανά Μοντέλα (HMM-based Speech Synthesis).

- Την διερεύνηση και ανάπτυξη ενός νέου μεθοδολογικού πλαισίου, βασισμένου σε ταξινομητές μιας τάξης (one-class classification), για την αξιολόγηση ύπαρξης φασματικών ασυνεχειών στην ένωση ακουστικών μονάδων και την εφαρμογή του ως συνάρτηση κόστους στον αλγόριθμο επιλογής ακουστικών μονάδων.

Επίσης, περιγράφονται και προτείνονται καινοτόμες εφαρμογές, στις οποίες η σύνθεση φωνής έχει ουσιαστικό ρόλο και που αφορούν την ολοκλήρωση της τεχνολογίας σε σύγχρονα τηλεπικοινωνιακά περιβάλλοντα και υπηρεσίες.

Πιο συγκεκριμένα, η συνεισφορά στις παραπάνω δραστηριότητες καθώς και τα συμπεράσματα που προέκυψαν έχουν ως εξής:

Αλγόριθμος Επιλογής Ακουστικών Μονάδων

Ο πρώτος άξονας αφορά την σχεδίαση και την υλοποίηση του αλγόριθμου επιλογής ακουστικών μονάδων καθώς και τις επιλογές των μηχανισμών και των κριτηρίων που απαρτίζουν τις συναρτήσεις κόστους που εμπεριέχει ο αλγόριθμος. Μελετήθηκε ενδελεχώς η σχετική βιβλιογραφία και προτάθηκαν αποδοτικοί μηχανισμοί, τεχνικές και μεθοδολογίες σε αυτά τα ζητήματα. Ειδικότερα, αναπτύχθηκαν και αξιολογήθηκαν κριτήρια που συνθέτουν τις συναρτήσεις κόστους και προτάθηκαν ευρετικές προσεγγίσεις για την αποδοτική τους ρύθμιση με στόχο την επίτευξη υψηλής ποιότητας συνθετικής ομιλίας. Σε κάθε περίπτωση και όπου ήταν δυνατό, έμφαση δόθηκε σε τεχνικές που βασίζονται σε διαθέσιμα δεδομένα (*data-driven techniques*), χαλαρώνοντας τις απαιτήσεις στη χρήση αυστηρών μοντέλων. Οι μέθοδοι, τα αποτελέσματα και τα συμπεράσματα υλοποιήθηκαν και ενσωματώθηκαν σε ένα ολοκληρωμένο σύστημα σύνθεσης φωνής από κείμενο τρέχουσας τεχνολογικής στάθμης για την Ελληνική γλώσσα, βελτιώνοντας σημαντικά την τελική του ποιότητα.

Σύνθεση Φωνής σε Υπολογιστικό Περιβάλλον Κινητού Τηλεφώνου

Ο δεύτερος άξονας, αφορά τις ερευνητικές προσπάθειες και τις μεθοδολογικές προσεγγίσεις που υιοθετήθηκαν για την αποδοτική αποκλιμάκωση, την μεταφορά και την ολοκλήρωση τεχνολογίας σύνθεσης φωνής υψηλής ποιότητας σε περιβάλλον συσκευής κινητού τηλεφώνου και γενικότερα σε περιβάλλοντα ενσωματωμένων συστημάτων, με στόχο την μέγιστη δυνατή αξιοποίηση των οφελών της στο περιβάλλον αυτό. Οι προσεγγίσεις σχετίζονται με την σχεδίαση και την υλοποίηση γενικού σκοπού συστήματος σύνθεσης φωνής (unlimited/general domain speech synthesis) με επιλογή και ένωση ακουστικών μονάδων και συντελούν, τόσο στην αποδοτική μεταφορά της τεχνολογίας, όσο και στην μείωση των υπολογιστικών απαιτήσεων, διασφαλίζοντας ταυτόχρονα υψηλή τελική ποιότητα. Στο πλαίσιο του συγκεκριμένου άξονα αναπτύχθηκε αλγόριθμος που συντελεί στην αποδοτική αποκλιμάκωση και μείωση μιας υπάρχουσας βάσης δεδομένων που χρησιμοποιείται σε πλήρες σύστημα σύνθεσης φωνής από κείμενο, με σκοπό την δημιουργία μικρότερης βάσης για εφαρμογή σε κινητό τηλέφωνο. Ο αλγόριθμος βασίζεται στη συμπεριφορά της μηχανής επιλογής ακουστικών μονάδων και στηρίζεται σε στατιστική ανάλυση των δεδομένων που προκύπτουν από την επιλογή των ακουστικών μονάδων που προκύπτουν από την σύνθεση και ανάλυση μεγάλου σώματος κειμένου και για τις οποίες λαμβάνεται υπόψη τόσο η συχνότητα εμφάνισης των ακουστικών μονάδων όσο και η «βαθμολογία» που αυτές πετυχαίνουν στην διαδικασία αναζήτησης από την μηχανή επιλογής. Η τεχνική διατηρεί την μέγιστη δυνατή ποιότητα λόγω επαυξημένης κάλυψης και αποφυγής επιλογής παρόμοιων ακουστικών μονάδων μειώνοντας παράλληλα τις αυξημένες υπολογιστικές και αποθηκευτικές απαιτήσεις . Επιπλέον, μελετήθηκε η χρήση και η προσαρμογή απωλεστικών αλγορίθμων κωδικοποίησης φωνής τύπου CELP, για την συμπίεση των ακουστικών μονάδων που αποθηκεύονται στην βάση δεδομένων και εξετάστηκε η επίδραση που έχουν στην ποιότητα της παραγόμενης φωνής κατά το στάδιο αποκωδικοποίησης και συρραφής τους. Η CELP προσαρμόστηκε κατάλληλα στις ιδιαίτερες απαιτήσεις του συστήματος σύνθεσης φωνής. Η προσαρμογή αυτή είναι ιδιαίτερα σημαντική διότι παρέχει τη δυνατότητα μηχανισμού τυχαίας πρόσβασης και αποκωδικοποίησης των ακουστικών μονάδων

χωρίς να προκαλείται αλλοίωση στις κρίσιμες λειτουργικές παραμέτρους που είναι απαραίτητες για την σύνθεση. Τέλος, αναπτύχθηκε αλγόριθμος που επιτυγχάνει σημαντική μείωση των υπολογιστικών αναγκών της βαθμίδας επιλογής ακουστικών μονάδων. Ο αλγόριθμος εφαρμόζεται στον υπολογισμό του κόστους φασματικής ασυνέχειας και επιτυγχάνει την ελαχιστοποίηση του υπολογιστικού φορτίου κατά την σύνθεση, εφαρμόζοντας συσταδοποίηση στο διάνυσμα χαρακτηριστικών της φασματικής αναπαράστασης των ακουστικών μονάδων και στον υπολογισμό των αποστάσεων μεταξύ τους σε μη πραγματικό χρόνο. Η εφαρμογή του αλγόριθμου οδηγεί σε σημαντική μείωση των υπολογιστικών και αποθηκευτικών απαιτήσεων διότι μειώνει το πλήθος των απαιτούμενων υπολογισμών για τις συναρτήσεις κόστους στην μηχανή επιλογής ακουστικών μονάδων. Η εφαρμογή της τεχνικής αυτής μείωσε σημαντικά τον χρόνο εκτέλεσης της μηχανής επιλογής χωρίς να επιφέρει ουσιαστική επίδραση στην τελική ποιότητα Συνοψίζοντας, βασικός γνώμονας σε κάθε περίπτωση ήταν η επίτευξη συνθετικού λόγου υψηλής ποιότητας. Οι παραπάνω τεχνικές αξιολογήθηκαν από μια σειρά πειραμάτων με υποκειμενικά και αντικειμενικά κριτήρια τα αποτελέσματα των οποίων επικύρωσαν την λειτουργικότητα και την αποδοτικότητά τους. Η ολοκλήρωση και εφαρμογή των παραπάνω τεχνικών οδήγησε στην υλοποίηση ενός συστήματος σύνθεσης φωνής από κείμενο τρέχουσας τεχνολογικής στάθμης για την συσκευή του κινητού τηλεφώνου σημαντικά υψηλής ποιότητας.

Σύνθεση Φωνής με Κρυφά Μαρκοβιανά Μοντέλα

Με γνώμονα συστήματα σύνθεσης φωνής από κείμενο τόσο γενικού σκοπού όσο και σε περιβάλλοντα ενσωματωμένων συστημάτων, ο τρίτος άξονας αναφέρεται στη διερεύνηση παραμετρικών μεθόδων σύνθεσης φωνής που από την φύση τους απαιτούν σημαντικά χαμηλότερους υπολογιστικούς πόρους ενώ ταυτόχρονα παρέχουν μεγαλύτερη ευελιξία τόσο στη διαχείριση όσο και στην κατασκευή νέων συνθετικών φωνών. Για το σκοπό αυτό, εξετάστηκε η τεχνολογία σύνθεσης φωνής με βάση τα Κρυφά Μαρκοβιανά Μοντέλα. Η υιοθέτηση των Κρυφών Μαρκοβιανών Μοντέλων παρέχει ένα γενικευμένο στατιστικό μεθοδολογικό πλαίσιο για την αποδοτική παραμετρική μοντελοποίηση, διαχείριση και παραγωγή φωνής. Στο πλαίσιο της διατριβής, μελετήθηκε, προσαρμόστηκε, υλοποιήθηκε και αξιολογήθηκε το μεθοδολογικό πλαίσιο της παραμετρικής τεχνολογίας σύνθεσης φωνής από κείμενο με χρήση κρυφών Μαρκοβιανών μοντέλων για την περίπτωση και τα ιδιαίτερα χαρακτηριστικά της Ελληνικής γλώσσας. Η πειραματική αποτίμηση του συστήματος πραγματοποιήθηκε με συγκριτική αξιολόγηση τόσο με ένα σύστημα σύνθεσης με διφωνήματα όσο και με το σύστημα επιλογής και συρραφής ακουστικών μονάδων. Η αξιολόγηση ανέδειξε ότι η εν λόγω τεχνολογία παράγει συνθετική φωνή ικανοποιητικής ποιότητας, η οποία αν και υπολείπεται ακόμα αυτής που επιτυγχάνεται με επιλογή και συρραφή ακουστικών μονάδων, δέχεται περαιτέρω τροποποιήσεις που μπορούν να επιφέρουν σημαντικές βελτιώσεις στο τελικό αποτέλεσμα. Τέτοιες τροποποιήσεις αφορούν αφενός το κομμάτι της προσωδίας και αφετέρου την άρση των περιορισμών που επιφέρουν οι κλασσικές τεχνικές αναπαράστασης, μοντελοποίησης και παραγωγής του συνθετικού σήματος φωνής.

Μεθοδολογικό Πλαίσιο Εντοπισμού και Αξιολόγησης Φασματικών Ασυνεχειών με χρήση Ταξινομητών μιας Τάξης

Στον τέταρτο άξονα, ο προτάθηκε και αναπτύχθηκε ένα νέο μεθοδολογικό πλαίσιο για την αξιολόγηση ύπαρξης φασματικών ασυνεχειών στην ένωση των ακουστικών μονάδων, το οποίο εφαρμόστηκε ως κριτήριο (συνάρτηση κόστους) στον αλγόριθμο επιλογής ακουστικών μονάδων. Στόχος είναι το κριτήριο να αντικατοπτρίζει την ανθρώπινη αντίληψη για τις ακουστικές ασυνέχειες και κατ' επέκταση την φυσικότητα της συνθετικής ομιλίας. Το νέο μεθοδολογικό πλαίσιο βασίζεται στα διαθέσιμα δεδομένα (data driven), στηρίζοντας το θεωρητικό του υπόβαθρο σε τεχνικές μηχανικής μάθησης και αναγνώρισης προτύπων και συγκεκριμένα στην χρήση ταξινομητών μιας τάξης (one-class classification). Η υιοθέτηση της εν λόγω μεθοδολογίας, στηρίχθηκε στο γεγονός ότι η διαθέσιμη βάση δεδομένων προηχογραφημένης φυσικής ομιλίας, που χρησιμοποιείται και κατά την σύνθεση, αποτελείται από πλήθος φυσικών "ενώσεων" όπως αυτές ορίζονται από τα ζεύγη συνεχόμενων πλαισίων φωνής. Η πληθώρα από φυσικές "ενώσεις" μπορούν να χρησιμοποιηθούν ως πρωτογενή δεδομένα περιγραφής και εκπαίδευσης του ταξινομητή μιας τάξης. Στο πλαίσιο της διατριβής, εξετάστηκε η ανάλυση ανά φώνημα και η αναπαράσταση κάθε ζεύγους από συνεχόμενα πλαίσια φωνής με ένα διάνυσμα φασματικών αποστάσεων. Η τάξη που ορίζουν τα δεδομένα εκπαίδευσης μοντελοποιήθηκε με μείγμα κατανομών Gauss (Gaussian Mixture Models - GMM). Η τεκμηρίωση της μεθοδολογίας που προτάθηκε και αναπτύχθηκε, επιβεβαιώθηκε από τα πειραματικά αποτελέσματα, όπου ανεδείχθη η αποτελεσματικότητα και η ευρωστία της προτεινόμενης μεθοδολογίας η οποία σε συγκριτική αξιολόγηση έχει σαφή προβάδισμα έναντι παγιωμένων τεχνικών και προσεγγίσεων που προτείνονται στη βιβλιογραφία. Βασικό πλεονέκτημα αυτού του μεθοδολογικού πλαισίου, είναι ότι στηρίζεται στα διαθέσιμα δεδομένα οπότε το κόστος ένωσης για τις φασματικές ασυνέχειες εκπαιδεύεται και προκύπτει από πρότυπα που δυνητικά συναντώνται στην υπάρχουσα βάση δεδομένων ενώ παράλληλα για την εκπαίδευση δεν απαιτεί παρέμβαση από τον ανθρώπινο παράγοντα και δεν εξαρτάται από ακουστικά πειράματα. Τέλος, μπορεί να υιοθετηθεί ως ευρύτερο μεθοδολογικό πλαίσιο και να εφαρμοστεί με κάθε πιθανή αναπαράσταση που είναι ικανή να διακρίνει αποτελεσματικά τις δύο τάξεις.

Συνοψίζοντας, οι στόχοι της έρευνας στο πλαίσιο της διδακτορικής διατριβής καλύφθηκαν σε μεγάλο βαθμό, ενώ οι τεχνικές και μεθοδολογίες που προτάθηκαν και αναπτύχθηκαν, οδήγησαν σε καινούρια αποτελέσματα και συμπεράσματα. Παράλληλα, τροφοδοτούν νέες ιδέες για την περαιτέρω έρευνα στο επιστημονικό πεδίο της διατριβής. Μερικές από αυτές τις ιδέες περιγράφονται στην ενότητα που ακολουθεί.

7.2 ΕΡΕΥΝΗΤΙΚΗ ΣΥΝΕΧΕΙΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Η βελτίωση της ποιότητας και η συνεχής αναβάθμιση των συστημάτων σύνθεσης φωνής από κείμενο, αποτελεί διαρκή στόχο στον τομέα της σύνθεσης φωνής. Η τρέχουσα τεχνολογία που στηρίζεται στην επιλογή και συρραφή ακουστικών μονάδων από προηχογραφημένη βάση δεδομένων φυσικής ομιλίας, έχει οδηγήσει σε σχεδόν φυσική συνθετική φωνή, παρουσιάζει ωστόσο εγγενείς αδυναμίες που

καθορίζονται από παράγοντες που χρήζουν περαιτέρω διερεύνησης. Σε γενικές γραμμές, οι απαιτήσεις σε υπολογιστικούς και αποθηκευτικούς πόρους, ο σχεδιασμός σώματος κειμένου για ηχογράφηση, ο αποδοτικός σχεδιασμός και τα κριτήρια του αλγόριθμου επιλογής ακουστικών μονάδων, η προσαρμοστικότητα και γενικότερα η δυνατότητα παραγωγής εκφραστικής/συναισθηματικής ομιλίας, είναι μερικοί από αυτούς [Bailly, 2003]. Στο πλαίσιο της διατριβής εξετάστηκαν αποτελεσματικά κάποια από τα παραπάνω ζητήματα. Ωστόσο τα ευρήματα που προέκυψαν, δημιουργούν και τροφοδοτούν προοπτικές για περαιτέρω έρευνα που δύναται να επιφέρει επιπρόσθετες βελτιώσεις.

Πιο συγκεκριμένα, κρίσιμο συστατικό στην τελική ποιότητα ενός συστήματος σύνθεσης φωνής είναι τα κριτήρια που αποτελούν τις συναρτήσεις κόστους του αλγόριθμου [Diaz, 2003]. Η διερεύνηση και υιοθέτηση τεχνικών και κριτηρίων που βασίζονται στα διαθέσιμα δεδομένα, φαίνεται να οδηγεί σε καλύτερα αποτελέσματα συγκριτικά με προσεγγίσεις που ακολουθούν αυστηρές μοντελοποιήσεις [Damper, 2001]. Το συμπέρασμα αυτό επιβεβαιώνεται και από το σύστημα σύνθεσης φωνής που εξετάστηκε στην διατριβή, οι συναρτήσεις κόστους του οποίου ακολουθούν κατά το δυνατόν αυτή την προσέγγιση. Η διερεύνηση και ο ορισμός εύρωστων και αποδοτικών κριτηρίων στις συναρτήσεις κόστους σύμφωνα με αυτή την προσέγγιση αποτελεί μια πρόσφορη μεθοδολογική νόρμα, και αναμένεται να συντελέσει σε ακόμα καλύτερα αποτελέσματα.

Προς αυτή την κατεύθυνση, το μεθοδολογικό πλαίσιο που προτάθηκε για τις φασματικές ασυνέχειες στην ένωση των ακουστικών μονάδων και στηρίχτηκε σε ταξινομητές μιας τάξης, μπορεί να επεκταθεί και να εφαρμοστεί συνολικά στην σχεδίαση της συνάρτησης κόστους ένωσης. Η από κοινού αναπαράσταση σε συνολικό διάνυσμα χαρακτηριστικών των μεγεθών που εμπλέκονται ως κριτήρια στην συνάρτηση κόστους ένωσης (π.χ. θεμελιώδης συχνότητα, ένταση, φασματικές αποστάσεις/αναπαραστάσεις κτλ.) και η εκπαίδευση των ταξινομητών σε αυτά με χρήση της υπάρχουσας βάσης δεδομένων προηχογραφημένης φυσικής ομιλίας δύναται να οδηγήσει σε συνάρτηση κόστους ένωσης που να στηρίζεται στα διαθέσιμα δεδομένα και που προκύπτει από εκπαίδευση (data driven trainable join cost function), καταργώντας παράλληλα την ανάγκη χρήσης και προσδιορισμού των παραγόντων στάθμισης (βάρη) για κάθε κριτήριο. Το γεγονός αυτό είναι ιδιαίτερα σημαντικό καθώς όπως αναφέρθηκε, ο προσδιορισμός των παραγόντων στάθμισης αποτελεί δύσκολη διαδικασία η οποία συνήθως αντιμετωπίζεται με ευρετικούς τρόπους [Black, 1996].

Στο πλαίσιο της διατριβής, η αποτελεσματικότητα του μεθοδολογικού πλαισίου που βασίζεται σε ταξινομητές μιας τάξης για τον εντοπισμό φασματικών ασυνεχειών, αναδείχθηκε χρησιμοποιώντας ένα διάνυσμα χαρακτηριστικών που αποτελείτο από φασματικές αποστάσεις για την περιγραφή της ένωσης, με ταξινομητή που βασίζεται σε μείγμα κατανομών Gauss. Τα πειραματικά αποτελέσματα έδειξαν ότι επιπλέον βελτίωση είναι εφικτή. Η αναζήτηση διαφορετικών τρόπων αναπαράστασης των ενώσεων ανάμεσα σε ακουστικές μονάδες (π.χ. εξάγοντας (φασματικά) χαρακτηριστικά αφού έχει προηγηθεί η ένωση ή μοντελοποιώντας τροχιές φασματικών μεγεθών στην ένωση) και η αξιολόγηση και άλλων ταξινομητών μιας τάξης ή και συνδυασμούς από ταξινομητές (combine classifiers) [Garau, 2008; Fergani, 2008; Davy, 2002; Davy, 2006], μπορεί να

οδηγήσει σε ακόμα καλύτερα αποτελέσματα και να βελτιώσει περαιτέρω την ποιότητα της συνθετικής φωνής.

Πολλές από τις τεχνικές που αναπτύχθηκαν για την προσαρμογή της τεχνολογίας επιλογής και συρραφής ακουστικών μονάδων σε υπολογιστικά περιβάλλοντα μειωμένων πόρων, μπορούν να εφαρμοστούν και σε πλήρες σύστημα σύνθεσης φωνής γενικού σκοπού, συντελώντας στην μείωση των υπολογιστικών και αποθηκευτικών απαιτήσεων χωρίς σημαντική επίπτωση στην τελική ποιότητα. Το αποτέλεσμα αυτό είναι ιδιαίτερα σημαντικό καθώς αποτελεί κρίσιμο παράγοντα σε πρακτικές εφαρμογές της τεχνολογίας. Ειδικότερα, η τεχνική συσταδοποίησης στο διάνυσμα χαρακτηριστικών της φασματικής αναπαράστασης των ακουστικών μονάδων και στον υπολογισμό των αποστάσεων μεταξύ τους σε μη πραγματικό χρόνο, μπορεί να εφαρμοστεί προσδιορίζοντας διαφορετικό αριθμό συστάδων ανά τύπο ακουστικών μονάδων. Επιπλέον, ο αριθμός συστάδων μπορεί να καθοριστεί με μέτρο τόσο την διασπορά των φασματικών διανυσμάτων σε κάθε τύπο ακουστικής μονάδας, όσο και από την εγγενή ευαισθησία των τύπων των ακουστικών μονάδων κατά την ένωση (π.χ. ενώσεις του /u/ είναι λιγότερο αντιληπτές από ενώσεις του /α/) [Syrdal, 2004; Syrdal, 2005]. Στον αλγόριθμο αποκλιμάκωσης και μείωσης μιας υπάρχουσας βάσης δεδομένων που χρησιμοποιείται σε πλήρες σύστημα σύνθεσης φωνής από κείμενο, και ο οποίος στηρίζεται στη συμπεριφορά των συναρτήσεων κόστους του αλγόριθμου επιλογής, περαιτέρω βελτίωση μπορεί να επιτευχθεί σχεδιάζοντας τις συναρτήσεις κόστους με τέτοιο τρόπο που να αντιμετωπίζουν αποτελεσματικά μεγάλα τοπικά κόστη που προκύπτουν κατά τη σύνθεση, χωρίς ωστόσο να αλλοιώνουν το τελικό ολικό κόστος. Αυτό φαίνεται και από τα ακουστικά πειράματα, η τιμή MOS των οποίων φαίνεται να συμβαδίζει (συσχετίζεται) καλύτερα με τα μέγιστα κόστη. Για το σκοπό αυτό μπορούν να εφαρμοστούν τεχνικές που δίνουν έμφαση και χειρίζονται κατάλληλα μεγάλες τιμές τοπικού κόστους [Toda, 2006].

Η διερεύνηση που πραγματοποιήθηκε στο πλαίσιο της διατριβής για την τεχνολογία παραμετρικής σύνθεσης φωνής από κείμενο με χρήση HMM για την Ελληνική γλώσσα, ανέδειξε την ικανότητα παραγωγής ικανοποιητικής ποιότητας συνθετικής φωνής. Τα πλεονεκτήματα που προσφέρει η εν λόγω τεχνολογία σε σχέση με αυτήν με επιλογή και συρραφή ακουστικών μονάδων είναι αρκετά, με σημαντικότερα την υψηλή ευελιξία στην δημιουργία και διαχείριση συνθετικών φωνών (στυλ, ύφος, χαρακτηριστικά ομιλητή κ.α. [Yamagishi, 2009b, Yamagishi, 2009]) και την ιδιαίτερα χαμηλή απαίτηση σε υπολογιστικούς πόρους [Zen, 2009]. Διαθέτοντας μια σημαντική γκάμα από τεχνολογικά εργαλεία που έχουν αναπτυχθεί στον τομέα της αναγνώρισης φωνής και μπορούν να εφαρμοστούν και στην περίπτωση της σύνθεσης, η σύνθεση φωνής με HMM συναντά σημαντικό εμπόδιο στο μοντέλο παραγωγής φωνής που στηρίζεται σε αυτό της πηγής-φίλτρου (source-filter model) το οποίο οδηγεί σε αλλοίωση της τελικής ποιότητας. Εύρωστες λύσεις αντιμετώπισης προς αυτή την κατεύθυνση, δηλαδή η εύρεση κατάλληλης αναπαράστασης, ικανής αφενός για στοχαστική μοντελοποίηση με την χρήση HMM, και αφετέρου για ανασύνθεσης του σήματος χωρίς (ακουστικές) αλλοιώσεις, αποτελεί σημαντικό ερευνητικό ζητούμενο [Zen, 2009; Black, 2007].

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Abdel-Hamid, O., Abdou, S., Rashwan, M., “Improving Arabic HMM based speech synthesis quality”, In proc. of Interspeech 2006, pp.1332--1335, Pittsburg, 2006.

[2] Acero A., “Formant analysis and synthesis using hidden markov models”, In Proc. of Eurospeech 1999, 1999.

[3] Alias, F., and Llora, X., “Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis”, in Proc. of Eurospeech 2003. 2003.

[4] Alvarez V. Y, Huckvale M, “The Reliability of the ITU-T P.85 Standard for the Evaluation of Text-to-speech Systems”, 2004.

[5] Aylett M. P., et al, “The cerevoice blizzard entry 2006: A prototype small database unit selection engine,” in Proc. Blizzard Challenge Workshop, 2006.

[6] Bailly G., Campbell N., Möbius B., "ISCA Special Session: Hot topics in speech synthesis", Proceedings of the Eurspeech 03, Geneva, Switzerland, pp. 37-40, 2003.

[7] Bakar Z., Mohemad R., Ahmad, A., and Deris, M., “A comparative study for outlier detection techniques in data mining”, in proc. 2006 IEEE Conference on Cybernetics and Intelligent Systems, pp. 1-6, 2006.

[8] Bellegarda J. R., “A Global Boundary-Centric Framework for Unit Selection Text-to-Speech Synthesis,” IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 3, pp. 990-997, May 2006.

[9] Benesty J., Sondhi M., Huang Y. (Eds.), Springer Handbook of Speech Processing. Springer, 2008.

[10] Beutnagel, M., Conkie, A., Syrdal, A.K., “Diphone synthesis using unit selection”, Proc. 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, pp. 185-190, 1998.

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

[12] Bishop, C., “Novelty detection and neural network validation”, in Proc. of IEE Conference on Vision and Image Signal Processing, pp. 217-222, 1994.

[13] Bishop, C., Neural Networks for Pattern Recognition. Oxford University Press, 1995.

[14] Bjørkan, I., Svendsen, T., Farner, S., “Comparing spectral distance measures for join cost optimization in concatenative speech synthesis”, 9th European Conference on Speech Communication and Technology Interspeech 2005, pp. 2577-2580, Lisbon, 2005.

[15] Black A. W. and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis,” in Proc. 5th Eurospeech, Rhodes, Greece, Sep. 1997, pp. 601–604, 1997.

[16] Black A.W., and Lenzo K., “Limited domain synthesis” in Proc. ICSLP, Vol. 2, pp. 411–414, Beijing, China, Sep., 2000.

[17] Black, A., “CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling”, Interspeech 2006 - ICSLP, Pittsburgh, PA, 2006.

[18] Black, A., Bennett, C., Blanchard, B., Kominek, J., Langner, B. Prahallad, K., Toth, A., “CMU Blizzard 2007: a hybrid acoustic unit selection system from statistically predicted parameters”, Blizzard Challenge 2007 Workshop, Bonn, Germany, 2007.

[19] Black, A.W., Zen H., Tokuda K., “Statistical parametric speech synthesis,” Proc. of IEEE ICASSP 2007, pp. 1229-1232, April, 2007.

[20] Black, A.W., Taylor, P., “Chatr: A genetic speech synthesis system” in Proc. Conf. Computational Linguistics, pp. 983-986, 1994.

[21] Blouin C., Rosec O., P. Bagshaw, and C. d’Alessandro, “Concatenation cost calculation and optimization for unit selection in TTS,” in Proc. IEEE 2002 Workshop on Speech Synthesis, September 2002.

[22] Bozkurt B., T. Dutoit T., O. Ozturk, "Text Design For TTS Speech Corpus Building Using A Modified Greedy Selection", Proc. Eurospeech, Geneva 2003, pp 277-280, 2003.

[23] Breen A., and Jackson P., “Non-uniform unit selection and the similarity metric within BT's laureate TTS system”, in Proc. 3rd ESCA/COCOSDA Int. Workshop on Speech Synthesis, Jenolan Caves, Blue Mountians, p. G.1, Australia, Nov., 1998.

[24] Brew A., Grimaldi M., Cunningham P., “An evaluation of one-class classification techniques for speaker verification”, Artificial Intelligence Review, Springer, 2008.

[25] Bulyko, I., and Ostendorf, M., “Unit selection for speech synthesis using splicing costs with weighted finite state transducers”, in Proceedings of Eurospeech 2001, 2001.

[26] 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.

[27] Campbell, W.N. and A.W. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis" in Progress in Speech Synthesis, J.V. Santen, et al., eds. 1996, pp. 279-292, Springer Verlag, 1996.

[28] Carlson, R., and Granström B., "Data-driven multimodal synthesis" Speech Communication, Volume 47, Issues 1-2, 2005, Pages 182-193, 2005.

[29] Chalamandaris A., P. Tsiakoulis, S. Karabetsos, S. Raptis, “An Efficient and Robust Pitchmarking Algorithm on the Speech Waveform for TD-PSOLA”, in Proc. of the IEEE ICSIPA 2009 (IEEE International Conference on Signal and Image Processing Applications 2009), paper 190, Malaysia, November, 2009 (b).

[30] Chalamandaris A., Protopapas A., Tsiakoulis P., 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, pp. 1226–1229, 2006.

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

[32] Chalamandaris A., S. Karabetsos, P. Tsiakoulis, S. Raptis, "A Unit Selection Text-to-Speech Synthesis System Optimized for Use with Screen Readers”, **accepted in** IEEE Transactions on Consumer Electronics, 2010.

[33] 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)

[34] Chandola, V., Banerjee, A., Kumar, V., “Anomaly detection: A survey” ACM Computing Surveys, 41, 3, article 15, pp. 15:1-15:58, 2009.

[35] Chang-Heon Lee, Sung-Kyo Jung, and Hong-Goo Kang, “Applying a Speaker-Dependent Speech Compression Technique to Concatenative TTS Synthesizers,” IEEE Trans. on Audio, Speech and Language Processing, vol. 15, no. 2, pp. 632–640, 2007.

[36] Chappell, D.T., Hansen, J.H.L, “A comparison of spectral smoothing methods for segment concatenation based speech synthesis”, Speech Communication, 36 (3-4), pp. 343-374, 2002.

[37] Chazan D., Hoory R., Kons Z., Silberstein D., and A. Sorin, “Reducing the footprint of the IBM trainable speech synthesis system,” in Proc. ICSLP 2002, Denver, CO, pp. 2381–2384. 2002.

[38] Chen J. D. and N. Campbell, “Objective distance measures for assessing concatenative speech synthesis,” in Proc. Eurospeech ‘99, Budapest, Hungary, 1999.

[39] Chu M., H. Peng, Hong-Yun Yang, E. Chang, “Selecting non-uniform units from a very large corpus for concatenative speech synthesizer”, in proc. IEEE ICASSP 2001, pp. 785-788, 2001.

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

[41] Christensen, L.B.: RoboBraille – Automated Braille Translation by Means of an Email Robot. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2006. LNCS, vol. 4061, pp. 1102–1109. Springer, Heidelberg (2006)

[42] Clark R.A.J., Richmond K., King S, “Multisyn: Open-domain unit selection for the Festival speech synthesis system”, Speech Communication, 49 (4), pp. 317-330, 2007.

[43] Coker C. H., “A model of articulatory dynamics and control”, Proc. IEEE, 64(4):452–460, April 1976

[44] Colotte, V., & Beaufort, R., “Linguistic features weighting for a text-to-speech system without prosody model”, In Proceedings of Interspeech 2005, 2005.

[45] Conkie A. and Isard S., “Optimal Coupling of Diphones”, Progress in Speech Synthesis, Springer-Verlag, pp. 293-305, 1996.

[46] Coorman, G., Fackrell, J., Rutten, P., and Coile, B. V., "Segment selection in the LH realspeak laboratory TTS system," Proc. of the International Conference on Spoken Language Processing (ICSLP 2000), vol. 2, pp. 395-398, 2000.

[47] Damper, R.I. (Ed.), “Data-Driven Techniques in Speech Synthesis,” ISBN: 978-0-412-81750-2, Springer, 2001.

[48] Davis J. and Goadrich M., “The Relationship between Precision-Recall and ROC Curves”, in Proc. of the 23rd International Conference on Machine Learning, Pittsburgh, PA, pp. 233-240, 2006.

[49] Davis, S., and Mermelstein, P., “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Trans. Acoustic Speech Signal Processing, 28(4), pp. 357–366, 1980.

[50] Davy M., F. Desobry, A. Gretton, and C. Doncarli, “An online Support Vector Machine for Abnormal Events Detection,” Signal Processing, vol. 86, no. 8, pp. 2009–2025, Aug., 2006.

[51] Davy, M. and Godsill, S., “Detection of abrupt spectral changes using support vector machines: an application to audio signal segmentation”, in proc. of IEEE ICASSP 2002, Vol. 2, pp. II-1313-II-1316, 2002.

[52] Denby B., Schultz T. et al. “Silent Speech Interfaces”, Speech Communication, 52 (2010), pp. 270–287, 2010.

[53] Deng Li, Wang K., Chou Wu, “Speech Technology and Systems in Human-Machine Communication”, IEEE Signal Processing Magazine, Editors’ Note, vol. 22, n. 5, pp. 12-14, 2005.

[54] Desobry F., Davy M., and C. Doncarli, “An online kernel change detection algorithm,” IEEE Transactions on Signal Processing, vol. 53, no. 5, May, 2005.

[55] Diaz F. C., Banga E. R. "On the design of cost functions for unit-selection speech synthesis", In EUROSPEECH 2003, 289-292, 2003.

[56] Díaza F. C., Banga E. R., “A method for combining intonation modeling and speech unit selection in corpus based speech synthesis systems”, Speech Communication, Vol.48, Issue 8, Pages 941-956, August, 2006.

[57] Ding, W., Fujisawa, K., Campbell, N. "Improving Speech Synthesis of CHATR Using a Perceptual Discontinuity Function and Constraints of Prosodic Modification", In 3rd Speech Synthesis Workshop SSW3-1998, pp. 191-194, 1998.

[58] Dixon, N. and Maxey, H., "Terminal analog synthesis of continuous speech using the diphone method of segment assembly", IEEE Transactions on Audio and Electroacoustics, 16(1), pp. 40–50, 1968.

[59] Donovan R. E., "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," in proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Pethshire, Scotland, pp. 59–62, 2001.

[60] Donovan, R. E. Trainable Speech Synthesis. Ph.D. thesis, Cambridge University Engineering Department. 1996.

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

[62] Duda, R. and Hart, P., Pattern Classification and Scene Analysis. John Wiley & Sons, New York, 1973.

[63] Dutoit, T., An Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, 1997.

[64] Dutoit, T., and Leich, H., "Text-to-speech synthesis based on an MBE re-synthesis of the segments database," Speech Communication, 13, pp. 435–440, 1993.

[65] El-Yaniv, R., and Nisenson, M., "Optimal Single-Class Classification Strategies" In Proc. of the 2006 Conference in Advances in Neural Information Processing Systems, MIT Press, 2007.

[66] Eskenazy M., "An Overview of Spoken Language Technology for Education", Speech Communication, 51 (2009), pp. 832–844, 2009.

[67] Falaschi A., M. Giustiniani, M. Verola, "A hidden Markov model approach to speech synthesis", Proc. Eurospeech, Vol. 1989, pp. 2187– 2190, 1989.

[68] Fant, G., Liljencrants, J., & Lin, Q., "A four parameter model of glottal flow," KTH, STL-QPS, No 4, pp. 1-13, 1985.

[69] Fawcett T., "An introduction to ROC analysis", Pattern Recognition Letters, 27, pp. 861–874, 2006.

[70] Fergani B., Davy M., Houacine A., "Speaker diarization using one-class support vector machines", Speech Communication, 50, pp. 355–365, 2008.

[71] Founda M., Tambouratzis G., Chalamandaris A. and G. Carayannis, "Reducing Spectral Mismatches in Concatenative Speech Synthesis via Systematic Database Enrichment", Eurospeech2001, pp. 837-840, 2001.

[72] Fraser, M., King, S., "The Blizzard Challenge 2007", in proc. Blizzard Challenge 2007 (in conjunction with the Sixth ISCA Workshop on Speech Synthesis), Bonn, Germany, paper 001, 2007.

[73] Frölich M., D. Michaelis and H.W. Strube, "SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals", J. Acoust. Soc. Amer., 110(1), 479-488, 2001

[74] Gales M. and Young S., "The Application of Hidden Markov Models in Speech Recognition", Foundations and Trends in Signal Processing, Vol. 1, No. 3, pp. 195–304, 2007.

[75] Garau G., and Renals S., "Combining Spectral Representations for Large-Vocabulary Continuous Speech Recognition", IEEE Trans. Audio, Speech and Language Processing, 16(3), pp. 508-518, 2008.

[76] Gardner A. B., Krieger A. M., Vachtsevanos G., "One-Class Novelty Detection for Seizure Analysis from Intracranial EEG", Journal of Machine Research, 7, pp. 1025--1044, 2006.

[77] Gibson J. D., "Speech Coding Methods, Standards, and Applications," IEEE Circuits and Systems Magazine, Vol. 5, No. 4, Fourth Quarter, 2005.

[78] Gibson M., “Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models,” in Proc. Interspeech 2009, Brighton, U.K., Sept., 2009.

[79] Gilbert M., Feng J., “Speech and Language Processing over the Web: Changing the way people communicate and access information” IEEE Signal Processing Magazine, vol. 25, n. 3, May, pp. 18-28, 2008.

[80] Gonzalvo, X., Iriondo, I., Socor, J., Alas, F., Monzo, C., “HMM-based Spanish speech synthesis using CBR as F0 estimator”, In proc. ISCA Tutorial and Research Workshop on Non Linear Speech Processing – NOLISP’07, 2007.

[81] Gray Jr., A.H., Markel, J.D., “Distance measures for speech processing”, IEEE trans. Acoustic Speech and Signal Proc., 24 (5), pp. 380-391, 1976.

[82] Hamza, W., Rashwan, M., & Afify, M., “A quantitative method for modeling context in concatenative synthesis using large speech database”, In Proceedings of the International Conference on Acoustics Speech and Signal Processing ICASSP 2001, 2001.

[83] Hansen J. H. L. and Chappell D. T., “An auditory-based distortion measure with application to concatenative speech synthesis,” IEEE Trans. Speech Audio Processing, vol. 6, pp. 489–495, 1998.

[84] Harris, C. M., “A study of the building blocks in speech” Journal of the Acoustical Society of America, 25(5), pp. 962–969, 1953.

[85] Hempstalk, K. and Frank, E., “Discriminating against New Classes: One-Class versus Multi-Class Classification”, Lecture Notes in Computer Science, Volume 5360/2008, Nov. 27, 2008.

[86] Hermansky H., “Perceptual linear predictive (PLP) analysis of speech”, The Journal of the Acoustical Society of America, 87(4), pp. 1738–1752, 1990.

[87] Hertz S., R. "Integration of Rule-based Formant Synthesis and Waveform Concatenation: A Hybrid Approach To Text-to-Speech Synthesis," IEEE 2002 Workshop On Speech Synthesis, 2002.

[88] Hirschfeld D., “Comparing static and dynamic features for segmental cost function calculation in concatenative speech synthesis”, in Proc. of the International Conference on Spoken Language Processing -ICSLP 2000, vol.2, pp. 435-438, 2000.

[89] Hodge, V. & Austin, J., “A survey of outlier detection methodologies”, Artificial Intelligence Review 22, 2, pp. 85-126”, 2004.

[90] Hoffmann R., Jokisch O., Hirschfeld D., Strecha G., H. Kruschke, and U. Kordon, “A Multilingual TTS System with Less than 1 Megabyte Footprint for Embedded Applications,” in Proc. of the IEEE ICASSP’03, Hong Kong, China, 2003.

[91] Holmes, J., "Formant synthesizers, cascade or parallel," Speech Communication, 2, pp. 251-273, 1983.

[92] Huang T. S., Mark A. Hasegawa-Johnson et al. “Sensitive Talking Heads”, IEEE Signal Processing Magazine, pp. 67-72, July, 2009.

[93] Huang X., Acero A., Hon Hsiao-Wuen, Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR, 2001.

[94] Hunt A. J. and Black A. W., “Unit selection in a concatenative speech synthesis system using a large speech database,” in IEEE Int. Conf. Acoustics, Speech and Signal Processing-ICASSP 1996, pp. 373–376, 1996.

[95] Ishikawa Y., Kisuki Y., Sakamoto T., and Hase T., “Speech Synthesis Method based on Application-Specific Synthesis Units and its Implementation on a 32-bit Microprocessor,” IEEE Trans. on Consumer Electronics, vol. 45, no. 3, pp. 980-985, 1999.

[96] Iwahashi, N., Kaiki, N., and Sagisaka, Y., “Speech segment selection for concatenative synthesis based on spectral distortion minimization”, trans. of IEICE, E76A, pp. 1942–1948, 1993.

[97] Iyer A. N., Ofoegbu U. O., Yantorno R., Smolenski B., “Speaker distinguishing distances: a comparative study”, International Journal of Speech Technology, Springer, 2009.

[98] Jain, A. K., Duin, R. P. W., Jianchang Mao, “Statistical pattern recognition: a review”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22(1), pp. 4 – 37, Jan., 2000.

[99] Jun Xu and Lianhong Cai, “Spectral Continuity Measures at Mandarin Syllable Boundaries”, in proc. ISCSLP'06 (5th Int. Symp. on Chinese Spoken Lang. Processing), 2006.

[100] Jurafsky D. & J. H. Martin. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. 2008.

[101] Kain A. and J. van Santen, "Unit-Selection Text-to-Speech Synthesis Using an Asynchronous Interpolation Model", Proc. of 6th ISCA Workshop on Speech Synthesis, August 2007.

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

[103] Kaszczuk M., Osowski, L. “The IVO Software Blizzard Challenge 2009 Entry: Improving IVONA Text-to-Speech”, in proc. Blizzard Challenge 2009.

[104] Katsamanis A., P. Tsiakoulis, P. Maragos, and A. Potamianos. Investigations in articulatory synthesis. In ICPhS, 2007

[105] Kawahara H., Masuda-Katsuse I., A. de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds”, Speech Communication, 27(3-4), pp. 187–207, 1999.

[106] Kawai H. and M. Tsuzaki, “Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative synthesis,” in proc. ICSLP 2002, Denver, USA, 2002.

[107] Kawai H., Toda T., Ni J., Tsuzaki M. and Tokuda K., “XIMERA: A new TTS from ATR based on Corpus-based Technologies”, in proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, pp. 179-184, 2004.

[108] Kedia D. S., M. Basu, “Architectural Optimizations for Text to Speech Synthesis in Embedded Systems Design Automation Conference”, in proc. ASP-DAC '07, Asia and South Pacific, Jan. 2007, pp. 298-303, 2007.

[109] Kennedy K., “Low Default Portfolio/One-Class Classification: A Literature review”, Technical Report, DITAIG, School of Computing, 2009.

[110] Kim N. S. and Park S. S., “Discriminative Training for Concatenative Speech Synthesis”, IEEE Signal Processing Letters, 11(1), pp. 40-43, 2004.

[111] Kim S., et al, “Pruning of redundant synthesis instances based on weighted vector quantization,” in Proc. EUROSPEECH, Aalborg, Denmark, pp. 2231–2234, 2001.

[112] 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.

[113] King S., Tokuda K., Zen H., and Yamagishi J., “Unsupervised adaptation for HMM-based speech synthesis,” in Proc. Interspeech 2008, Brisbane, Australia, pp. 1869–1872, Sept., 2008.

[114] Kirkpatrick, B., O'Brien D., Scaife, R., Errity A., "Spectral dynamics as a source of discontinuity in concatenative speech synthesis", in proc. 15th IEEE International Conference on Digital Signal Processing, DSP 2007, pp. 615-618, 2007.

[115] Kirkpatrick, B., O'Brien, D., Scaife, R., "Feature extraction for spectral continuity measures in concatenative speech synthesis", in Proc. ICSLP 2006, 2006.

[116] Klabbers E. and Veldhuis R., "Reducing audible spectral discontinuities," IEEE Trans. on Speech and Audio Processing, vol. 9, no. 1, pp. 39–51, Jan. 2001.

[117] Klabbers, E., van Santen, J.P.H., Kain, A., "The Contribution of Various Sources of Spectral Mismatch to Audible Discontinuities in a Diphone Database", IEEE Transactions on Audio Speech, and Language Processing, Vol. 15, Issue 3, pp. 949 – 956, March 2007.

[118] Klatt, D. H. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 67, 971–995, 1980

[119] Klatt, D. H., "Review of text-to-speech conversion for English", Journal of the Acoustical Society of America 82, 3, 1987.

[120] Krishnan S. and Rao P., "A comparative study of explicit frequency and conventional signal representation for speech recognition," Digital Signal Processing, vol. 6, pp. 249–284, 1996.

[121] Krstulovic, S., Hunecke, A., Schroeder, M., "An HMM-Based Speech Synthesis System applied to German and its Adaptation to a Limited Set of Expressive Football Announcements", In proc. of Interspeech 2007, Antwerp, 2007.

[122] Krul A., et al, "Approaches for adaptive database reduction for text-to-speech synthesis," in Proc. INTERSPEECH 2007, Antwerp, Belgium, pp. 2881–2884, 2007.

[123] Kumar R. and S. Prahallad Kishore, "Automatic pruning of unit selection speech databases for synthesis without loss of naturalness," in Proc. INTERSPEECH 2004, Jeju Island, Korea, pp. 1377–1380, 2004.

[124] Latacz L., Mattheyses W. and W. Verhelst, "The VUB Blizzard Challenge 2009 Entry", The Blizzard Challenge 2009 workshop, 4th September 2009, University of Edinburgh, 2009.

[125] Latorre J., Iwano K., and Furui S., "New approach to the polyglot speech generation by means of an HMM based speaker adaptable synthesizer," Speech Communication, Elsevier, vol. 48, no. 10, pp. 1227–1242, 2006.

[126] Lee Ki-Seung and Kim S. R., "Contect-Adaptive Smoothing for Concatenative Speech Synthesis", IEEE Signal Processing Letters, 9(12), pp. 422-425, 2002.

[127] Lee M., Lopresti D. P., Olive, J. P., "A text-to-speech platform for variable length optimal unit searching using perception based cost functions", International Journal of Speech Technology, 6(4), pp. 347-356, 2003.

[128] Lee, M., "Perceptual cost functions for unit searching in large corpus-based concatenative text-to speech", In Proc. EUROSPEECH 2001, Aalborg, Denmark, September 2001. pp. 2227–2230, 2001.

[129] Ling, Zhen-Hua, Qin, Long, et. al "The USTC and iflytek speech synthesis systems for Blizzard Challenge 2007", In Blizzard Challenge (BLZ3-2007), paper 017, 2007.

[130] Maeda S., "A digital simulation method of the vocal-tract system", Speech Communication, 1:199–229, 1982

[131] Maia R., Toda T., Zen H., Nankaku Y., Tokuda K., "An excitation model for HMM-based speech synthesis based on residual modeling, Proc. ISCA Speech Synthesis Workshop (SSW6), 2007.

[132] Maia, R., Zen, H., Tokuda, K., Kitamura, T., Resende F., G., Jr., ”Towards the development of a Brazilian Portuguese text-to-speech system based on HMM”, In proc. of Eurospeech 2003, pp.2465--2468, Geneva, 2003.

[133] Manevitz L. and Yousef M., “One-Class SVMs for Document Classification,” Journal of Machine Learning Research, vol. 2, pp. 139–154, 2001.

[134] Markou M. and Singh S., “Novelty detection: A review-part 1: Statistical approaches,” Signal Processing, vol. 83, no. 12, pp. 2481–2497, 2003a.

[135] Markou, M. and Singh, S., “Novelty detection: A review-part 2: Neural Network based approaches”, Signal Processing, vol. 83, no. 12, pp. 2499-2521, 2003b.

[136] Möbius B., "Corpus-based speech synthesis: methods and challenges", Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), pp. 87-116, 2000.

[137] Modenesi, A. P., and Braga, A. P., “Analysis of Time Series Novelty Detection Strategies for Synthetic and Real Data”, Neural Processing Letters, 30(1), pp. 1-17, 2009.

[138] Mohasi L. and D. Mashao, “Text-to-Speech Technology in Human-Computer Interaction”, 5th Conf. on Human Computer Interaction in Southern Africa, (CHISA 2006, ACM SIGHI), pp. 79-84, 2006.

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

[140] Mouchtaris A., J. V. Spiegel, P. Mueller, “Nonparallel Training for Voice Conversion Based on a Parameter Adaptation Approach,” IEEE Trans. Audio, Speech, & Language Processing vol.14 no.3 pp.952-963, May, 2006.

[141] Moulines, E. and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones," Speech Communication, 9(5), pp. 453-467, 1990.

[142] Moya, M., Koch, M., and Hostetler, L., “One-class classifier networks for target recognition applications”, In Proceedings World Congress on Neural Networks, pages 797–801, Portland, OR. International Neural Network Society, INNS, 1993.

[143] Nakamura K., T. Toda, Y. Nankaku, K. Tokuda, “On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum”, Proc. ICASSP, Vol. 06, pp. 93–96, 2006.

[144] Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., Zhang, J.-S., (...), Yamamoto, S., “The ATR multilingual speech-to-speech translation system”, IEEE Transactions on Audio, Speech and Language Processing, 14 (2), pp. 365-375, 2006.

[145] Nukaga N., Kamoshida R., Nagamatsu K., and Kitahara Y., “Scalable implementation of unit selection based text-to-speech system for embedded solutions,” Proc. of IEEE ICASSP 2006, pp. 849-852, Toulouse, 2006.

[146] Olive, J. and Liberman, M., “A set of concatenative units for speech synthesis”, Journal of the Acoustical Society of America, 65, S130. 1979.

[147] Olive, J. P., “Rule synthesis of speech from dyadic units”, In proc. IEEE ICASSP 1977, pp. 568–570, 1977.

[148] O'Shaughnessy, D., “Interacting with computers by voice: automatic speech recognition and synthesis”, Proceedings of the IEEE, Vol. 91, Issue 9, pp. 1272 - 1305, Sept. 2003.

[149] O'Shauqhnessy, D., “Modern methods of speech synthesis”, IEEE Circuits and Systems Magazine, Vol. 7, Issue 3, pp. 6 - 23, 2007.

[150] Pantazis Y., Stylianou Y., and E. Klabbers, “Discontinuity detection in concatenated speech synthesis based on nonlinear speech analysis,” in Proc. of Interspeech 2005, Lisbon, pp. 2817-2820, Portugal, 2005.

[151] Pantazis, Y., Stylianou Y., “On the detection of discontinuities in concatenative speech synthesis”, in proc. Workshop on Nonlinear Speech Processing, WNSP 2005, Heraklion, Crete, 2005, (b).

[152] Patcha, A. and Park, J.-M., “An overview of anomaly detection techniques: Existing solutions and latest technological trends”, Computer Networks, 51, 12, pp. 3448-3470, 2007.

[153] Pekalska E. and Duin R. THE DISSIMILARITY REPRESENTATION FOR PATTERN RECOGNITION Foundations and Applications. World Scientific Publishing Co. Pte. Ltd. 2005.

[154] Peng, H., Zhao, Y., Chu, M., “Perpetually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation”, In Proc. ICSLP 2002, Denver, USA, September 2002, pp. 2613–2616, 2002.

[155] Peters J.-P., Thillou C., and S. Ferreira, "Embedded Reading Device for Blind People: a User-Centred Design," Proc. of 33rd Applied Imagery Pattern Recognition Workshop (AIRP'04), 2004.

[156] Peterson, G. E., Wang, W. W.-Y., and Sivertsen, E., ”Segmentation techniques in speech synthesis”, Journal of the Acoustical Society of America, 30(8), pp. 739–742, 1958.

[157] Plumpe, M. and S. Meredith, "Which is More Important in a Concatenative Text-to-Speech System: Pitch, Duration, or Spectral Discontinuity," Third ESCA/COCOSDA Int. Workshop on Speech Synthesis, Jenolan Caves, Australia, pp. 231-235, 1998.

[158] Qian, Y., Soong, F., Chen, Y., Chu, M., “An HMM-based Mandarin Chinese text-to-speech system”, In: Q. Huo et al. (eds.) ISCSLP 2006, LNAI, vol. 4274, pp. 223—232, Springer, Heidelberg, 2006.

[159] Qin L., Ling Z., Wu Y., Zhang B., and Wang R., “HMM-based emotional speech synthesis using average emotion model,” Springer LNAI (Proc. ISCSLP-06), pp. 233–240, Dec., 2006.

[160] Quatieri, T., F. Discrete Time Speech Signal Processing, Principles and Practice. Prentice Hall, Upper Saddle River. 2002.

[161] Rabaoui A., Kadri H., Z. Lachiri, and N. Ellouze, “One-Class SVMs Challenges in Audio Detection and Classification Applications”, EURASIP Journal on Advances in Signal Processing, Volume 2008, Article ID 834973, 14 pages, 2008.

[162] Rabiner L., and B. H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, NJ, 1993.

[163] Rabiner, L. R., and Schafer, R. W. Digital processing of Speech signals. Prentice Hall, 1978.

[164] Raptis S., I. Spais and P. Tsiakoulis, “A Tool for Enhancing Web Accessibility: Synthetic Speech and Content Restructuring”, in Proc. HCII 2005: 11th International Conference on Human-Computer Interaction, 22-27 July, Las Vegas, Nevada, USA, 2005.

[165] Raptis S., 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.

[166] Rouibia S., Olivier Rosec, and Thierry Moudenc, “Unit Selection for Speech Synthesis Based on Acoustic Criteria” Text, Speech and Dialogue, LNCS Springer, Vol. 3658, pp. 281-287, 2005.

[167] Rutten P., Aylett M. P., Fackrell J., and Taylor P., “A statistically motivated database pruning technique for unit selection synthesis,” in Proc. ICSLP 2002, Denver, Colorado, USA, pp. 125–128, 2002.

[168] Sagisaka, Y., “Speech synthesis by rule using an optimal selection of non-uniform synthesis units”, In proc. IEEE ICASSP-88, pp. 679–682, 1988.

[169] Sagisaka, Y., Kaiki, N., Iwahashi, N., Mimura, K., “ATR ν-talk speech synthesis system”, in Proc. Int. Conf. Spoken Language Processing (ICSLP), pp. 483-486, 1992.

[170] Sakai S., Kawahara T., “Decision tree-based training of probabilistic concatenation models for corpus-based speech synthesis”, in Proc. Interspeech 2006, 2006.

[171] Sakai S., Shu H., “A probabilistic approach to unit selection for corpus-based speech synthesis”, Proc. Interspeech 2005, pp. 81-84, 2005.

[172] Sakai, S., Maia, R., Kawai, H., Nakamura, S., “A close look into the probabilistic concatenation model for corpus-based speech synthesis ”, Proc. of the 10th Annual Conference of the International Speech Communication Association, Interspeech 2009 , pp. 752-755, 2009.

[173] Schnell M., Jokisch O., Hoffmann R., and M. Kustner, “Text-to-speech for low-resource systems,” IEEE Workshop Multimedia Signal Processing (MMSP), St. Thomas, pp. 259-262, 2002.

[174] Schroeder M. R., and Atal B. S., “Code-excited linear prediction (CELP) high quality speech at very low bit rates,” in Proc. IEEE ICASSP, 85, pp. 934–940, 1985.

[175] Schroeder M., Charfuelan M., Sathish Pammi, Oytun Turk, "The MARY TTS entry in the Blizzard Challenge 2008", held in conjunction with Interspeech 2008, 2008.

[176] Schroeder M., “Expressive Speech Synthesis: Past, Present,and Possible Futures”, in J.H. Tao, T.N. Tan (eds.), Affective Information Processing, Springer Science+Business Media LLC 2009.

[177] Schultz T., Black A. W., Vogel S., and M. Woszczyna, “Flexible Speech Translation Systems,” IEEE trans. on Audio, Speech and Language Processing, vol. 14, no. 2, pp. 403-411, 2006.

[178] Sheikhzadeh H., Cornu E., Brennan R., Schneider T., “Realtime speech synthesis on an ultra low-resource, programmable DSP system”, Proc. IEEE ICASSP 2002, Orlando, vol. 1, pp. 433-436, 2002.

[179] So S., and Paliwal K. K., “A comparative study of LPC parameter representations and quantization schemes for wideband speech coding”, Digital Signal Processing, 17, pp. 114–137, 2007.

[180] Sondhi M., and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. IEEE Transactions on Acoustics, Speech, and Signal Processing, 35(7):955–967, 1987.

[181] SpeechWorks, “Assessing Text-to-speech System Quality”, White Paper, SpeechWorks inc., 2006.

[182] Strom V. and S. King, “Investigating Festival's target cost function using perceptual experiments”, In Proc. Interspeech, Brisbane, 2008.

[183] Stylianou Y. and A. K. Syrdal, “Perceptual and objective detection of discontinuities in concatenative speech synthesis,” in IEEE Int. Conf. Acoustics, Speech and Signal Processing-ICASSP 2001, pp. 837–840, 2001.

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

[185] Stylianou, Y., “Voice Transformation: A Survey”, in Proc. ICASSP 2009, Taiwan, 2009.

[186] Syrdal A. K. and Conkie A. D., “Data-Driven Perceptually based Join Costs,” in Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, pp. 49-54, June 2004.

[187] Syrdal, A. K., and Conkie, A. D., “Perceptually-based data-driven join costs: Comparing join types”, In Proc. of Interspeech 2005, Lisbon, Portugal, 2005.

[188] Syrdal, A. K., Wightman, C. W., Conkie, A., Stylianou, Y., Beutnagel, M., Schroeter, J., Strom, V., and Lee, K.-S., “Corpus-based techniques in the AT&T NEXTGEN synthesis system”, In ICSLP-00, Beijing, 2000.

[189] Tachibana M., Yamagishi J., Masuko T., and Kobayashi T., "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," IEICE Trans. Inf. Syst., vol. E89-D, no. 3, pp. 1092 1099, Mar., 2006.

[190] Tao J., Xin Le, Yin P., "Realistic Visual Speech Synthesis based on Hybrid Concatenation Method", IEEE Trans. on Audio Speech and Language Processing, 17(3), pp. 469-477, March, 2009.

[191] Tax D. M. J. and Duin R. P. W. "Uniform object generation for optimizing one-class classifiers" Journal of Machine Learning Research, 2, pp.155–173, 2002.

[192] Tax D. M. J. and K. R. Mueller, "A Consistency-Based Model Selection for One-class Classification", in Proc. ICPR 2004, vol. 2, Cambridge UK, IEEE Computer Society, , pp. 363-366, 22-26 August, 2004.

[193] Tax D. M. J., DDtools, "The Data Description Toolbox for Matlab", version 1.7.3, 2009.

[194] Tax D. M. J., One-class classification; Concept-learning in the absence of counter-examples. Ph.D. thesis, ISBN: 90-75691-05-x, Delft University of Technology, 2001.

[195] Taylor P., "Unifying unit selection and hidden Markov model speech synthesis", in Proc. of Interspeech 2006, paper 1456, Pittsburgh, USA, Sept. 2006.

[196] Taylor P., A.W. Black, "Speech synthesis by phonological structure matching", Proc. Eurospeech 1999, Vol. 99, pp. 623–626, 1999.

[197] Taylor P., Text-to-Speech Synthesis. Cambridge University Press, 2009

[198] Taylor, P. A., "The target cost formulation in unit selection speech synthesis", In Proc. of the Interspeech 2006, 2006.

[199] Taylor, P. A., "Unifying unit selection and hidden Markov model speech synthesis" In Proc. of the Interspeech 2006, 2006.

[200] The Blizzard Challenge 2005 (http:// festvox.org/ blizzard2005.html)

[201] The Blizzard Challenge 2006 (http:// festvox.org/ blizzard2006.html)

[202] The Blizzard Challenge 2007 (http:// festvox.org/ blizzard2007.html)

[203] The Blizzard Challenge 2008 (http:// festvox.org/ blizzard2008.html)

[204] The Blizzard Challenge 2009 (http:// festvox.org/ blizzard2009.html)

[205] Tihelka, D., "Symbolic prosody driven unit selection for highly natural synthetic speech", In Proceedings of Eurospeech-Interspeech 2005, 2005.

[206] Tihelka, D., Matoušek, J. and Kala, J., "Quality Deterioration Factors in Unit Selection Speech Synthesis", Text, Speech and Dialogue 2007, Lecture Notes in Artificial Intelligence LNAI, vol. 4629, pp. 508-515, Springer-Verlag Berlin Hiedelberg 2007, 2007.

[207] Toda T., Kawai H., and M. Tsuzaki., "Optimizing integrated cost function for segment selection in concatenative speech synthesis based on perceptual evaluations", in Proc. EUROSPEECH '03, pp. 297–300, Geneva, Switzerland, Sep., 2003.

[208] Toda T., Kawai H., Tsuzaki M., Shikano K., "An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis", Speech Communication, 48 (1), pp. 45-56, 2006.

[209] Toda T., Kawai H., Tsuzaki V, and K. Shikano, "Perceptual evaluation of cost for segment selection in concatenative speech synthesis", IEEE Workshop on Speech Synthesis, Santa Monica, U.S.A., Sep., 2002.

[210] Toda, T., an Tokuda, K., “Speech parameter generation algorithm considering global variance for HMM-based speech synthesis”, In Proceedings of Eurospeech-Interspeech 2005, Lisbon, Portugal, 2005.

[211] Toda, T., Black, A., and Tokuda, K., “Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model”, Speech Communiation, Vol. 50, No. 3, pp. 215-227, March, 2008.

[212] Toda, T., Black, A., and Tokuda, K., “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory” in IEEE Transactions of Audio, Speech and Language Processing, 15(8), pp. 2222-2236, 2007.

[213] Toda, T., Kawai, H., Tsuzaki, M., “Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis”, in proc. IEEE ICASSP 2004, pp. I657-I660, 2004.

[214] Tokuda K., H. Zen, A.W. Black, “HMM-based approach to multilingual speech synthesis”, Text to speech synthesis: New paradigms and advances, S. Narayanan, A. Alwan (Eds.), Prentice Hall, 2004.

[215] Tokuda K., Yoshimura T., Masuko T., Kobayashi T., Kitamura T., “Speech parameter generation algorithms for HMM-based speech synthesis”, in Proc. of IEEE ICASSP 2000, pp.1315-1318, June, 2000.

[216] Tokuda, K., Kobayashi, T., and Imai, S., “Speech parameter generation from HMM using dynamic features”, In Proceedings of the International Conference on Acoustics Speech and Signal Processing, ICASSP 1995, 1995.

[217] Tokuda, K., Zen, H., and Kitamura, T., “Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features,”, In Proc. of Eurospeech 2003, 2003.

[218] Tokuda, K., Zen, H., Black, A.,: An HMM-based speech synthesis system applied to English. In: proceedings of IEEE Speech Synthesis Workshop 2002 (IEEE SSW 2002), September, 2002.

[219] Tomko S., Harris T. K., A. Toth, J. Sanders, A. Rudnicky and R. Rosenfeld, “Toward Efficient Human Machine Speech Communication: The Speech Graffiti Project,” ACM Transactions on Speech and Language Processing, vol. 2, no. 1, Article 2, pp. 1-27, 2005.

[220] Tsiakoulis P., Chalamandaris A., Karabetsos S., Raptis S., “A Statistical Method for Database Reduction for Embedded Unit Selection Speech Synthesis,” in IEEE ICASSP 2008, pp. 4601-4604, 2008.

[221] Tsuzaki M. and H. Kawai, “Feature extraction of unit selection in concatenative speech synthesis: Comparison between AIM, LPC, and MFCC,” in proc. ICSLP 2002, Denver, USA, 2002.

[222] Tsuzaki M., "Feature extraction by auditory modeling for unit selection in concatenative speech synthesis", In proc. EUROSPEECH-2001, Denmark, pp. 2223-2226, 2001.

[223] Van der Vrecken O., Pierret N., Dutoit T., Pagel V., and F. Malfrere, “New techniques for the compression of synthesizer databases,” in Proc. 1997 IEEE Int. Symp. Circuits and Systems, Hong Kong, Jun. 9–12, 1997, pp. 2641–2644. 1997.

[224] Vepa J. and S. King, “Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis,” IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1763-1771, Sep. 2006.

[225] Vepa J., and S. King, “Join cost for unit selection speech synthesis,” in Text to Speech Synthesis: New Paradigms and Advances, S. Narayanan and A. Alwan, Eds. NJ: Prentice-Hall, pp. 35–62, 2004.

[226] Vepa, J.; King, S.; Taylor, P., “New objective distance measures for spectral discontinuities in concatenative speech synthesis”, Proc. of 2002 IEEE Workshop on Speech Synthesis, pp. 223-226, 2002.

[227] Vesnicer, B., Mihelic, F., “Evaluation of the Slovenian HMM-based speech synthesis system”, In: Petr Sojka, Ivan Kopeˇcek, and Karel Pala (eds.) TSD 2004, LNAI, vol. 3206, pp.513--520. Springer, Heidelberg, 2004.

[228] Vincent D., O. Rosec and T. Chonavel, “Estimation of LF glottal source parameters based on an ARX model”, Proc. of Interspeech, p. 333-336, Lisboa, Sep. 2005

[229] Weiss, C., Hess, W., “Conditional random fields for hierarchical segment selection in text-to-speech synthesis” in proc. Interspeech 2006, Pittsburgh, September 2006.

[230] Wouters J. and Macon M., “A perceptual evaluation of distance measures for concatenative speech synthesis,” in Proc. ICSLP’98, vol. 6, Sydney, Australia, pp. 2747–2750, 1998.

[231] Wouters J. and Macon M., “Control of spectral dynamics in Concatenative speech synthesis,” IEEE Trans. on Speech and Audio Processing, vol. 9, no. 1, pp. 30–38, Jan, 2001.

[232] Yamagishi J., Kobayashi T., Tachibana M., Ogata K., and Nakano Y., “Model adaptation approach to speech synthesis with diverse voices and styles,” in Proc. IEEE ICASSP-07, Apr. 2007, pp. 1233–1236, 2007.

[233] Yamagishi J., Nose T., Zen H., Ling Z., Toda T., Tokuda K., King S., Renals S., “A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis,'” IEEE Trans. Audio, Speech, & Language Processing, 17(6), pp. 1208-1230, August, 2009.

[234] Yamagishi J., Tamura M., Masuko T., Tokuda K., Kobayashi T., “A context clustering technique for average voice models”, IEICE Trans. Inf. & Syst., vol.E86-D, no.3, pp.534-542, March, 2003.

[235] Yamagishi, J., T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai “Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm,” IEEE Audio, Speech, & Language Processing, vol.17, issue 1, pp.66-83, January 2009, (b).

[236] Yamagishi, J., Zen, H., Toda, T., Tokuda, K., “Speaker-Independent HMM-based Speech Synthesis System - HTS-2007 System for the Blizzard Challenge 2007”, In proc. of Blizzard Challenge 2007 workshop, pp. 1-6, Bonn, 2007.

[237] Yi, J.R.-W., “Corpus-based unit selection for natural-sounding speech synthesis”, Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Mass. Inst. Technol, Cambridge, MA, 2003.

[238] Yoshida, A., Mizuno, H., Mano, K., “Segment selection method based on tonal validity evaluation using machine learning for concatenative speech synthesis”, in proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Las Vegas, pp. 4617-4620, 2008.

[239] Yoshimura T., Tokuda K., Masuko T., Kobayashi T., Kitamura T., “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”, in Proc. of Eurospeech 99, pp.2347-2350, Sept., 1999.

[240] Yu, K., Toda, T., Gasic, M., Keizer, S., Mairesse, F., Thomson, B., Young, S.J., “Probabilistic modelling of F0 in unvoiced regions in HMM-based speech synthesis”, In IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan, 2009.

[241] Zen H., Toda T., Nakamura M., Tokuda K., “Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005”, IEICE Trans. Inf. & Syst. vol.E90-D, No.1, pp.325-333, Jan., 2007.

[242] Zen H., Tokuda K. and A.W. Black, “Statistical parametric speech synthesis”, Speech Communication, 51 (11), pp. 1039–1064, 2009.

[243] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T., “Hidden semi-markov model based speech synthesis”, In Proc. of the 8th International Conference on Spoken Language Processing, Interspeech 2004, 2004.

[244] Zeng Z., Fu Y., Roisman G. I., Wen Z., Hu Y. and Huang T. S., “One-Class Classification for Spontaneous Facial Expression Analysis”, in Proc. of the IEEE 7th International Conference on Automatic Face and Gesture Recognition (FGR’06), 2006.

ΔΗΜΟΣΙΕΥΣΕΙΣ

ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΕ ΔΙΕΘΝΗ ΠΕΡΙΟΔΙΚΑ ΜΕ ΚΡΙΤΕΣ

[1] **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, August 2010, Vol. 17, No. 8, pp. 746-749, 2010.

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

ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΕ ΣΥΛΛΟΓΙΚΟΥΣ ΤΟΜΟΥΣ ΜΕ ΚΡΙΤΕΣ

[1] Aimilios Chalamandaris, Spyros Raptis, Pirros Tsiakoulis, **Sotiris Karabetsos**, "*Enhancing Accessibility of Web Content for the Print-Impaired and Blind People*", in A. Holzinger and K. Miesenberger (Eds.): Book: Human Computer Interaction (HCI) and Usability for e-Inclusion, Lecture Notes in Computer Science (LNCS) 5889, ISBN 978-3-642-10307-0, pp. 249–263, 2009, Springer-Verlag Berlin Heidelberg, 2009

[2] Spyros Raptis, Pirros Tsiakoulis, Aimilios Chalamandaris, **Sotiris Karabetsos**, "*User Interaction Design for a Home-Based Telecare System*", in A. Holzinger and K. Miesenberger (Eds.): Book: Human Computer Interaction (HCI) and Usability for e-Inclusion, ISBN 978-3-642-10307-0, Lecture Notes in Computer Science (LNCS) 5889, pp. 333–344, 2009, Springer-Verlag Berlin Heidelberg, 2009

[3] **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.), Book: Text, Speech and Dialogue, Book Series Chapter in Lecture Notes in Computer Science (LNCS), ISBN 978-3-540-87390-7, Springer – Verlag, Vol. 5246/2008, pp. 349 – 356, 2008

ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΕ ΔΙΕΘΝΗ ΣΥΝΕΔΡΙΑ ΜΕ ΚΡΙΤΕΣ

[1] A. Chalamandaris, P. Tsiakoulis, **S. Karabetsos**, S. Raptis, "*An Efficient and Robust Pitchmarking Algorithm on the Speech Waveform for TD-PSOLA*", in Proc. of the IEEE ICSIPA 2009 (IEEE International Conference on Signal and Image Processing Applications 2009), paper 190, Malaysia, November, 2009.

[2] A. Chalamandaris, P. Tsiakoulis, S. Raptis, **S. 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 (LTC 2009), Poznan, Poland, November 6-8, 2009.

[3] S. Raptis, P. Tsiakoulis, A. Chalamandaris, **S. Karabetsos**, "*High Quality Unit-Selection Speech Synthesis for Bulgarian*", in Proc. 13-th International Conference on Speech and Computer (SPECOM'2009), St. Petersburg, Russia, 2009.

[4] P. Tsiakoulis, A. Chalamandaris, **S. Karabetsos**, 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, April 2008, pp. 4601-4604, 2008.

[5] **S. Karabetsos**, P. Tsiakoulis, S. E. Fotinea and I. Dologlou, "*Formant Estimation of Speech Signals Using Subspace-Based Spectral Analysis*", in Proc. EUSIPCO 2006 (14th European Signal Processing Conference), Florence, Italy, September 2006

[6] **S. Karabetsos**, P. Tsiakoulis, S. E. Fotinea and I. Dologlou, "*On the Use of a Decimative Spectral Estimation Method based on Eigenanalysis and SVD for Formant and Bandwidth Tracking of Speech Signals*", in Proc. Interspeech 2005 (Int. Conf. on Speech Commun. Technol.), Lisbon, Portugal, September 2005, pp. 709-712

[7] 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.