Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων
και Μηχανικων Υπολογιστων

Τομεας Σηματων, Ελεγχου και Ρομποτικης

Εργαστηριο Ορασης Υπολογιστων, Επικοινωνιας Λογου και
Επεξεργασιας Σηματων

Εύρωστα Ακουστικά Χαρακτηριστικά για Αυτόματη Αναγνώριση Φωνής από Απόσταση

Διπλωματικη Εργασια

ΝΙΚΟΛΑΟΣ ΦΛΕΜΟΤΟΜΟΣ

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2016

Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων
και Μηχανικων Υπολογιστων

Τομεας Σηματων, Ελεγχου και Ρομποτικης

Εργαστηριο Ορασης Υπολογιστων, Επικοινωνιας
Λογου και Επεξεργασιας Σηματων

Εύρωστα Ακουστικά Χαρακτηριστικά για Αυτόματη Αναγνώριση Φωνής από Απόσταση

Διπλωματικη Εργασια

ΝΙΚΟΛΑΟΣ ΦΛΕΜΟΤΟΜΟΣ

Επιβλέπων: Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή επιτροπή στις 29 Μαρτίου 2016.

.....	.....	.....
Πέτρος Μαραγκός	Αλέξανδρος Ποταμιάνος	Γεράσιμος Ποταμιάνος
Καθηγητής	Αναπληρωτής Καθηγητής	Αναπληρωτής Καθηγητής
Εθνικό Μετσόβιο Πολυτεχνείο	Εθνικό Μετσόβιο Πολυτεχνείο	Πανεπιστήμιο Θεσσαλίας

Αθήνα, Μάρτιος 2016

.....
Νικόλαος Φλεμοτόμος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Με την εκπόνηση της παρούσης Διπλωματικής εργασίας κλείνει ένας μεγάλος κύκλος της ζωής μου, αυτός των προπτυχιακών μου σπουδών στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Θα ήθελα να ευχαριστήσω όλους τους καθηγητές μου που τα τελευταία χρόνια μου ενέπνευσαν τη δίψα για μάθηση και μου παρείχαν απλόχερα όλα τα απαραίτητα εφόδια για μια ολοκληρωμένη και πολύπλευρη εκπαίδευση.

Ιδιαίτερα θέλω να ευχαριστήσω τον καθηγητή κ. Πέτρο Μαραγκό, η διδασκαλία του οποίου στα μαθήματα Ψηφιακή Επεξεργασία Σήματος και Αναγνώριση Προτύπων με Έμφαση στην Αναγνώριση Φωνής αποτέλεσε καταλυτικό ρόλο στη μετέπειτα απόφασή μου να επικεντρωθώ στην περιοχή της Επεξεργασίας και Αναγνώρισης Φωνής. Στα μαθήματά του δεν παρέλειπε να παρουσιάζει τις σύγχρονες ερευνητικές τάσεις και να καλλιεργεί στο ακροατήριό του το ζήλο για εκ βάθους ενασχόληση και έρευνα. Η εν λόγω προσέγγισή του, σε συνδυασμό με τα πεδία ενδιαφέροντός του, με ώθησαν να απευθυνθώ σε αυτόν, ζητώντας του να επιβλέψει τη Διπλωματική μου εργασία, κάτι που αποδέχθηκε με προθυμία, προτείνοντάς μου το θέμα της και κατευθύνοντάς με ώστε να την περατώσω με επιτυχία.

Ιδιαίτερη συμβολή στην κατανόηση εννοιών απαραίτητων για την εκπόνηση της παρούσης εργασίας, όπως είναι η χρήση των Μηχανών Πεπερασμένης Κατάστασης με Βάρη στην Αυτόματη Αναγνώριση Φωνής, είχαν οι γνώσεις που αποκόμισα από το μάθημα Επεξεργασία Φωνής και Φυσικής Γλώσσας, όπως διδάχθηκε από τον καθηγητή κ. Αλέξανδρο Ποταμιάνο. Θέλω να ευχαριστήσω, ακόμα, τόσο τον κ. Ποταμιάνο, όσο και τον κ. Μαραγκό, για τις πολύτιμες συμβουλές τους και την έμπρακτη υποστήριξή τους κατά τις πρώτες μου προσπάθειες να ανοίξω τον επόμενο μεγάλο κύκλο της ακαδημαϊκής μου πορείας, αυτόν των μεταπτυχιακών και διδακτορικών μου σπουδών στις ΗΠΑ.

Θα πρέπει, επίσης, να ευχαριστήσω θερμά την Νάνσυ Ζλατίντση και τον Ισίδωρο Ροδομαγουλάκη, μέλη του Εργαστηρίου Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων, για το χρόνο τους και τις πολύτιμες συμβουλές τους σε όλα τα στάδια της παρούσης εργασίας.

Τέλος, οφείλω να ευχαριστήσω όλα τα αγαπημένα μου συγγενικά και φιλικά πρόσωπα, και κυρίως τους γονείς μου, Αντώνη και Σοφία, και την αδελφή μου Εβελίνα, για την αμέριστη στήριξή τους καθόλα τα χρόνια των σπουδών μου και για την κατανόησή τους όποτε θα ήθελα να αφιερώσω περισσότερο χρόνο μαζί τους, αλλά δεν μπορούσα λόγω των ακαδημαϊκών μου υποχρεώσεων.

Νίκος Φλεμοτόμος
Μάρτιος 2016

Η εργασία είναι αφιερωμένη στη γιαγιά μου Ευούλα και στη μνήμη της γιαγιάς μου Ζωής.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Περίληψη

Σκοπός της παρούσης Διπλωματικής εργασίας είναι η συγκριτική μελέτη διαφόρων μεθόδων εξαγωγής χαρακτηριστικών για χρήση στο πεδίο της Αναγνώρισης Φωνής από Απόσταση, με χρήση ενός μικροφώνου. Παρόλο που τις τελευταίες λίγες δεκαετίες υπάρχουν και εφαρμόζονται επιτυχημένα σύνολα χαρακτηριστικών στην περιοχή της Αυτόματης Αναγνώρισης Φωνής, με την απόδοση των συστημάτων να είναι ικανοποιητική σε καθαρές συνθήκες, στην περίπτωση που το μικρόφωνο απομακρύνεται από το στόμα του ομιλητή, η απόδοση πέφτει σε πολύ χαμηλά επίπεδα, καθώς εισάγονται παραμορφώσεις που οφείλονται σε μία ποικιλία παραγόντων, όπως είναι ο θόρυβος υποβάθρου και η αντήχηση.

Σημαντικό μέρος της εργασίας αφιερώνεται στη διεξοδική μελέτη, σε θεωρητικό και πειραματικό επίπεδο, των πλέον συχνά χρησιμοποιούμενων συνόλων χαρακτηριστικών που βασίζονται στην ενέργεια βραχέος χρόνου, των Αναφασματικών Χαρακτηριστικών στις Mel Συχνότητες (MFCCs), των συντελεστών Γραμμικής Πρόβλεψης βασισμένων στην Αντίληψη (PLPs), καθώς και παραλλαγών τους. Μέσω μιας σειράς πειραμάτων αναδεικνύεται η επίδραση που έχουν στην αναγνώριση διαφορετικές παραμετροποιήσεις κατά την εξαγωγή τους.

Ακόμα, μελετώνται οι πιο συνήθεις μέθοδοι μείωσης της διαστασιμότητας· η Ανάλυση Κύριων Συνιστωσών (PCA), η Γραμμική Διακριτική Ανάλυση (LDA) και η Ετεροσκεδαστική Γραμμική Διακριτική Ανάλυση (HLDA), όπως εφαρμόζονται μετά την ένωση διαδοχικών πλαισίων χαρακτηριστικών για την καλύτερη ανάδειξη της δυναμικής του σήματος.

Τέλος, εξετάζεται ο Τελεστής Teager Ενέργειας (TEO) υπό δύο σκοπιές. Πρώτον, προτείνεται ένα νέο πλαίσιο εργασίας όπου ο TEO χρησιμοποιείται στο πεδίο της συχνότητας για μείωση της υπολογιστικής πολυπλοκότητας και εισάγεται η έννοια του Φάσματος Teager Ισχύος (TPS), το οποίο μπορεί να χρησιμοποιηθεί στη ροή εργασίας γνωστών μεθόδων εξαγωγής χαρακτηριστικών, αντί του κλασικού Φάσματος Ισχύος ή σε συνδυασμό με αυτό, δίνοντας υποσχόμενα αποτελέσματα. Δεύτερον, χρησιμοποιείται στα πλαίσια του αλγορίθμου Gabor ESA για την εκτίμηση του στιγμιαίου πλάτους και της στιγμιαίας συχνότητας ενός σήματος και τη μετέπειτα εξαγωγή ποικίλων AM-FM χαρακτηριστικών. Όταν τα εν λόγω χαρακτηριστικά χρησιμοποιούνται σε συνδυασμό με τα MFCCs ή με τους Δέλτα-Φασματικούς Αναφασματικούς Συντελεστές (DSCCs) οδηγούν σε βελτιωμένα αποτελέσματα αναγνώρισης.

Όλα τα πειράματα στηρίζονται σε έναν αναγνωριστή χτισμένο στο σύστημα Kaldi, ενώ χρησιμοποιούνται πραγματικά δεδομένα για αναγνώριση από απόσταση. Για την αξιολόγηση των AM-FM χαρακτηριστικών γίνεται, ακόμα, χρήση προσομοιωμένων δεδομένων με ελεγχόμενες συνθήκες θορύβου.

Λέξεις-Κλειδιά: αναγνώριση φωνής από απόσταση, εξαγωγή ακουστικών χαρακτηριστικών, συντελεστές αναφάσματος στις mel συχνότητες, γραμμική πρόβλεψη βασισμένη στην αντίληψη, AM-FM χαρακτηριστικά, ευρωστία, Teager ενέργεια, φάσμα Teager ισχύος, μείωση διαστασιμότητας, μετατροπείς πεπερασμένης κατάστασης με βάρη

Abstract

The scope of this Diploma Thesis is the comparative study of various feature extraction methods used in the field of Distant Speech Recognition, using a single microphone. Although during the last few decades successful feature sets are being used in the area of Automatic Speech Recognition, with the final accuracy being satisfactory under clean conditions, when the microphone is moved away from the speaker's mouth, recognition accuracy is dropped down to very low levels, because of distortions which occur due to a variety of reasons, such as background noise and reverberation.

An important part of the Thesis is devoted to the meticulous study, both theoretically and experimentally, of the most often used feature sets which are based on the short-term energy, the Mel-Frequency Cepstrum Coefficients (MFCCs), the Perceptual Linear Prediction coefficients (PLPs), as well as certain variations. Through a series of experiments, the effect on the final recognition that different parametrizations have during the extraction process is highlighted.

Additionally, the most usual methods of dimensionality reduction are being investigated; Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Heteroscedastic Linear Discriminant Analysis (HLDA), as used after splicing successive feature frames in order to capture signal dynamics in a better way.

Finally, the Teager Energy Operator (TEO) is being studied under two different views. First, a new framework is being proposed where TEO is used in the frequency domain aiming at the reduction of computational complexity and the notion of Teager Power Spectrum (TPS) is being introduced, which can be used in the workflow of known feature extraction methods, instead of the classic Power Spectrum or in combination with it, giving promising results. Second, it is used as part of Gabor ESA for the estimation of the instantaneous amplitude and the instantaneous frequency of a signal and afterwards, for the extraction of a variety of AM-FM features. When the particular features are used in combination with MFCCs or with the Delta-Spectral Cepstrum Coefficients (DSCCs), they lead to the improvement of recognition results.

All the experiments are based on a recognizer built with Kaldi, while real data for distant recognition are being used. For the evaluation of AM-FM features, simulated data under controlled noise conditions are also being used.

Keywords: distant speech recognition, acoustic feature extraction, mel frequency cepstrum coefficients, perceptual linear prediction, AM-FM features, robustness, Teager energy, Teager power spectrum, dimensionality reduction, weighted finite state transducers

Κεφάλαιο 1

Εισαγωγή

1.1 Ομιλία και Άνθρωπος

"Ο λόγος που κάνει τον άνθρωπο κοινωνικότερο από τη μέλισσα ή τα άλλα ζώα που ζουν σε αγέλες είναι φανερός· επειδή τίποτα, όπως είπαμε, η φύση δεν κάνει χωρίς λόγο. Ο άνθρωπος είναι το μόνο ζώο με το χάρισμα του λόγου· οι άναρθρες κραυγές είναι εκδηλώσεις του ευχάριστου και του δυσάρεστου, γι' αυτό και υπάρχουν στα άλλα ζώα (μέχρι εκεί έφτασε η φύση τους, δηλαδή να αισθάνονται το ευχάριστο και το δυσάρεστο και να το εκφράζουν μεταξύ τους), ο λόγος όμως υπάρχει για να εκφράζεται το συμφέρον και το βλαβερό, συνεπώς το δίκαιο και το άδικο· τούτο ακριβώς είναι και το αποκλειστικό γνώρισμα του ανθρώπου σε σύγκριση με τα υπόλοιπα ζώα, δηλαδή ότι μόνο αυτός έχει το αίσθημα του καλού και του κακού και του δικαίου και του αδίκου και των άλλων συναφών αξιών· η δε επικοινωνία ως προς αυτά δημιουργεί την οικογένεια και την πόλη." [1].

Στο παραπάνω εδάφιο (*Πολιτικά, Α, 1253a 7-19*) ο Αριστοτέλης εκφράζει την άποψή του ότι είναι η ομιλία το χαρακτηριστικό εκείνο που ξεχωρίζει τον άνθρωπο από τα υπόλοιπα αγελαία ζώα, κάνοντας, μάλιστα, μια σαφή διάκριση μεταξύ του έναρθρου λόγου και της άναρθρης φωνής[1]. Η ιδέα αυτή διατηρήθηκε ανά τους αιώνες και υποστηρίχθηκε από πολλούς ακόμα στοχαστές και φιλόσοφους, με τον Thomas Hobbes (1588 - 1679) να αποτελεί ένα χαρακτηριστικό παράδειγμα, αναφέροντας πως *"η πιο ταπεινή και επικερδής εφεύρεση από όλες τις άλλες ήταν αυτή της Ομιλίας"* [2]. Σύμφωνα με τον Hobbes, η χρησιμότητα της ομιλίας δεν έγκειται απλά και μόνο στην επικοινωνία, αλλά στη μεταφορά της πνευματικής μας κατάστασης από τον κόσμο των ιδεών στον απτό κόσμο των λέξεων. Ερχόμενοι στη νεότερη εποχή, αξίζει να μνημονευτεί η ρήση του Ludwig Wittgenstein (1889 - 1951) *"τα όρια της γλώσσας μου ορίζουν τα όρια του κόσμου μου"* που αναδεικνύει τη γλώσσα ως το θεμέλιο στοιχείο της νόησης και της αντίληψης.

Η σχέση της νόησης με την ομιλία και η απάντηση στο ερώτημα εάν η σκέψη μάς επιτρέπει να μιλάμε ή εάν η γλώσσα μάς επιτρέπει να σκεφτόμαστε παραμένουν περιοχές αντιπαράθεσης ανάμεσα τόσο σε φιλοσόφους όσο και σε νευροεπιστήμονες και μελετητές του ανθρώπινου εγκεφάλου [3]· το σίγουρο πάντως είναι πως η ομιλία ήταν, είναι και θα είναι άρρηκτα συνδεδεμένη με την ανθρώπινη ύπαρξη. Ήδη από την ηλικία των 6 ετών ο άνθρωπος έχει στη διάθεσή του ένα λεξιλόγιο περίπου 600 λέξεων το οποίο μπορεί να διαχειριστεί με χρήση κατάλληλων

[1]Παρόλο που η εννοιολογική αυτή διάκριση είναι σημαντική, στη συνέχεια της εργασίας θα χρησιμοποιείται ο όρος φωνή εμπεριέχοντας τις έννοιες της ομιλίας και του λόγου, για λόγους συμβατότητας με την ελληνική βιβλιογραφία όπου έχει επικρατήσει ο όρος Αυτόματη Αναγνώριση Φωνής για να περιγράψει το διεθνώς εδραιωμένο όρο Automatic Speech Recognition.

συντακτικών δομών. Στον ανθρώπινο εγκέφαλο υπάρχουν διακριτές ημιανεξάρτητες περιοχές που εξειδικεύονται στα τρία συστατικά της γλώσσας: το λεξικό, που περιέχει τα δομικά εκείνα στοιχεία που καλούμε λέξεις, τη σημασιολογία, που προσδίδει εννοιολογική υπόσταση στις λέξεις και τις προτάσεις και τη σύνταξη ή γραμματική, που περιλαμβάνει τους κανόνες βάσει των οποίων μια αλληλουχία λέξεων μπορεί να σχηματίσει μία αποδεκτή πρόταση.

1.2 Αυτόματη Αναγνώριση Φωνής, Εφαρμογές και Ιστορικά Στοιχεία

Η ομιλία είναι συστατικό στοιχείο της ανθρώπινης φύσης και η συνεχής λειτουργία πολύπλοκων εγκεφαλικών δομών, σε συνεργασία με συγκεκριμένες ανατομικές ιδιαιτερότητες, την καθιστούν αναμφίβολα την πλέον εύκολη, άμεση και άκοπη μορφή επικοινωνίας, εκτός, φυσικά, από την περίπτωση που υπάρχουν συγκεκριμένα προβλήματα υγείας. Ήταν τουλάχιστον αναμενόμενο, συνεπώς, ο άνθρωπος να προσπαθήσει να ενσωματώσει το χαρακτηριστικό αυτό, τόσο υπό τη μορφή σύνθεσης, όσο και αναγνώρισης φωνής, στα μηχανήματα και τα υπολογιστικά συστήματα, με τις προσπάθειες να εντείνονται όσο η ψηφιακή πραγματικότητα ολοένα και περισσότερο εισβάλλει στη ζωή μας.

Με τον όρο Αυτόματη Αναγνώριση Φωνής εννοούμε τη μετατροπή, μέσω μιας αυτοματοποιημένης διαδικασίας, της ομιλίας σε κείμενο. Πρόκειται, με άλλα λόγια, για μια υπολογιστική διαδικασία, υλοποιημένη σε και εκτελούμενη από ηλεκτρονικό υπολογιστή, όπου η είσοδος είναι λέξεις, φράσεις ή συνεχής λόγος που εκφέρεται από έναν ή περισσότερους ομιλητές και η έξοδος είναι κείμενο που αντιστοιχεί στις λέξεις της εισόδου. Προφανώς, το ζητούμενο είναι η ακολουθία των λέξεων που παράγεται ως έξοδος να ταυτίζεται με την ακολουθία των λέξεων που δίνεται ως είσοδος, όπως τις είχε στο μυαλό του ο ομιλητής που τις εξέφερε. Πρόκειται για μία διαδικασία φαινομενικά πολύ απλή εάν αναλογιστούμε τα πολύ υψηλά ποσοστά επιτυχίας που θα είχε ένας άνθρωπος στη συγκεκριμένη εργασία. Η απόσταση, βέβαια, μεταξύ των δυνατοτήτων του ανθρώπινου εγκεφάλου και μιας υπολογιστικής μηχανής σε σύνθετα προβλήματα παραμένει τεράστια.

Οι πρώτες ερευνητικές προσπάθειες αναγνώρισης φωνής χρονολογούνται στο 1952, όταν ερευνητές των Bell Labs κατάφεραν να στήσουν ένα πρώτο σύστημα αναγνώρισης διακριτών ψηφίων που προφέρονται από έναν ομιλητή [4], βασισμένο σε συχνοτική πληροφορία και συγκεκριμένα στις θέσεις των δύο πρώτων formants. Το ποσοστό αναγνώρισης που ήταν περίπου ίσο με 98% έδωσε την ελπίδα και το έναυσμα για περαιτέρω μελέτη. Το 1959 εισήχθηκε για πρώτη φορά σε σύστημα αναγνώρισης η έννοια της στατιστικής πληροφορίας, καθώς προτάθηκε ότι η εκ των προτέρων γνώση στατιστικών στοιχείων της γλώσσας, όπως αποδεκτών ακολουθιών φωνημάτων, μπορεί να βελτιώσει την τελική απόδοση [5], ιδέα που δοκιμάστηκε και πειραματικά [6].

Παρόλο που οι ιδέες αυτές άργησαν σχετικά να βρουν την εφαρμογή που τους άρμοζαν [7], αξίζει να αναφερθεί πως το 1968, σε εργασία που δημοσιεύτηκε στη Σοβιετική Ένωση, έγινε για πρώτη φορά αναφορά σε πιθανή εφαρμογή αλγορίθμων δυναμικού προγραμματισμού για αναγνώριση φωνής [8] - οι απαρχές της τεχνικής που αργότερα θα ονομαζόταν Dynamic Time Wraping (DTW).

Ενώ οι προσπάθειες συνεχίζονταν, είχε αρχίσει να φαίνεται πως παρόλα τα χρόνια έρευνας που είχαν ήδη περάσει, τα αποτελέσματα ήταν πολύ περιορισμένα και αφορούσαν απλές εργασίες με εξαιρετικά περιορισμένο λεξιλόγιο. Σημαδιακή υπήρξε μία ανοιχτή επιστολή του John Pierce από τα Bell Labs το 1969 [9], όπου ασκούσε οξεία κριτική στην ερευνητική κοινότητα

που ασχολείται με την αναγνώριση φωνής. Χαρακτηριστικά ανέφερε στην εισαγωγή:

"Η αναγνώριση φωνής έχει γοητεία. Χρηματοδοτήσεις έχουν γίνει διαθέσιμες. Τα αποτελέσματα, όμως, είναι λιγότερο γοητευτικά. [...] Η αναγνώριση φωνής γενικού σκοπού φαντάζει πολύ μακρινή. Η αναγνώριση φωνής για ειδικούς σκοπούς είναι πολύ περιορισμένη. Θα φαινόταν καλό για τους ανθρώπους να ρωτήσουν τους εαυτούς τους γιατί εργάζονται πάνω στο πεδίο αυτό και τι περιμένουν να πετύχουν".

Η επόμενη δεκαετία, όμως, σημαδεύτηκε από μία σειρά επιτυχιών στο χώρο της Αυτόματης Αναγνώρισης Φωνής που έθεσαν τα θεμέλια των σημερινών συστημάτων. Εν πολλοίς σε αυτό βοήθησε η μεγάλη χρηματοδότηση που δόθηκε από τη DARPA (Defense Advanced Research Projects Agency) μέσω του 5-ετούς προγράμματος Speech Understanding Research (SUR) [10] και το οποίο οδήγησε σε αξιόλογες προσπάθειες από μεγάλες ερευνητικές ομάδες. Το 1979 γίνεται ίσως η πρώτη προσπάθεια αναγνώρισης φωνής ανεξάρτητα του ομιλητή (speaker-independent) στα Bell Labs [11].

Τη δεκαετία του 1980 γίνεται μία μεγάλη στροφή προς στατιστικές μεθόδους αναγνώρισης· μπαίνει, δηλαδή στο πεδίο η έννοια της πιθανότητας [12]. Η εφαρμογή των Κρυφών Μαρκοβιανών Μοντέλων (Hidden Markov Models - HMMs) για αναγνώριση φωνής και η χρήση γλωσσικών μοντέλων βασισμένων σε n-grams όχι μόνο βελτίωσε σημαντικά την απόδοση των συστημάτων, αλλά άνοιξε το δρόμο για αναγνώριση πολλών χιλιάδων λέξεων. Οι συνεχείς βελτιώσεις και οι νέες ιδέες τελικά οδήγησαν στην αύξηση του λεξικού μέχρι τις 65000 λέξεις [13], αλλά κυρίως στην αναγνώριση συνεχούς ομιλίας. Έτσι, τη δεκαετία του 1990 το όραμα του Large Vocabulary Continuous Speech Recognition (LVCSR) είχε γίνει πραγματικότητα. Προς αυτή την κατεύθυνση σαφώς βοήθησε η τεράστια τεχνολογική πρόοδος και η ευρεία παραγωγή ηλεκτρονικών υπολογιστών μεγάλης ισχύος, αλλά και ένας δεύτερος κύκλος χρηματοδοτήσεων για έρευνα από τη DARPA.

Το 1994 προτείνεται για πρώτη φορά ένα νέο πλαίσιο εργασίας που μπορεί να βρει εφαρμογή στην αναγνώριση φωνής, το οποίο συνδυάζει τη γνώση που είχε ήδη αποκτηθεί τα προηγούμενα χρόνια με αποτελεσματικό και αποδοτικό τρόπο. Πρόκειται για τα Weighted Finite State Transducers (WFSTs) [14]. Σήμερα, τα HMMs και τα WFSTs παραμένουν στο επίκεντρο των συστημάτων αναγνώρισης, τόσο για ερευνητικούς, όσο και εμπορικούς σκοπούς. Παράλληλα, σημαντικό μέρος της έρευνας έχει στραφεί προς τη χρήση Βαθέων Νευρωνικών Δικτύων (Deep Neural Networks). Τα νευρωνικά δίκτυα είχαν προταθεί και πιο νωρίς στη βιβλιογραφία, αλλά η αποτελεσματικότητά τους έγινε εμφανής μόλις πρόσφατα, τα τελευταία λίγα χρόνια [15].

Καθώς πολλά από τα θεμέλια συστατικά ενός συστήματος αναγνώρισης έχουν εδραιωθεί μέσα από μια σειρά ερευνών και πειραμάτων, είχε αρχίσει από νωρίς να διαφαίνεται η ανάγκη δημιουργίας ενός ευρέος αποδεκτού εργαλείου που θα αποτελούσε το baseline σύστημα για τις επόμενες ερευνητικές προσπάθειες. Έτσι, κατά καιρούς έχουν προταθεί διάφορα τέτοια εργαλεία, με τα πλέον αξιοσημείωτα να είναι το Julius [16], το CMUSphinx [17], το HTK [18], το RASR [19] και το Kaldi [20], τα οποία χρησιμοποιούνται σε μικρότερο ή μεγαλύτερο βαθμό σήμερα και βασίζονται όλα στη φιλοσοφία ανοιχτού κώδικα.

Τα συστήματα αναγνώρισης φωνής έχουν πλέον μπει ουσιαστικά στην καθημερινότητά μας. Όλοι έχουμε συνομιλήσει τηλεφωνικά με κάποιον αυτόματο τηλεφωνητή μιας δημόσιας (ή ιδιωτικής) υπηρεσίας ο οποίος "καταλαβαίνει" τι λέμε, ενώ οι περισσότεροι έχουμε πειραματιστεί με κάποια υπηρεσία φωνητικής αναγνώρισης στο κινητό μας τηλέφωνο, το αυτοκίνητο ή κάποια μηχανή αναζήτησης στο διαδίκτυο. Οι πιθανές χρήσεις της αναγνώρισης φωνής καλύπτουν μία ευρύτατη γκάμα εφαρμογών.

Μία από τις πιο ελπιδοφόρες χρήσεις της αναγνώρισης φωνής είναι για τη βοήθεια ατό-

μων με ειδικές ανάγκες. Για παράδειγμα, ένα αναπτυγμένο σύστημα λεκτικής επικοινωνίας ανθρώπου - μηχανής θα μπορούσε να αυξήσει σημαντικά την ποιότητα ζωής ενός ατόμου με σωματικές αναπηρίες και κινητικά προβλήματα [21]. Ένα αξιόπιστο σύστημα αναγνώρισης φωνής θα μπορούσε ακόμα να βοηθήσει ουσιωδώς άτομα με σοβαρά προβλήματα ακοής μέσω παραγωγής κειμένου σε πραγματικό χρόνο. Μπορεί η νοηματική γλώσσα και η ανάγνωση χειλιών να βοηθούν τα άτομα αυτά να ενταχθούν πλήρως στην καθημερινότητα και την κοινωνική ζωή, δεν είναι όμως πάντα αρκετά, όπως συμβαίνει παραδείγματος χάριν σε ένα αμφιθέατρο διδασκαλίας, όπου οι κωφάλαλοι βρίσκονται συνήθως σε μειονεκτική θέση [22, 23]. Η αναγνώριση φωνής χρησιμεύει, επίσης, σε άτομα με προβλήματα όρασης, καθώς δε χρειάζεται πλέον να χειρίζονται πολύπλοκα interfaces [24]. Επιπλέον, μαθητές με δυσλεξία ή άλλου είδους μαθησιακά προβλήματα μπορούν πλέον να χρησιμοποιήσουν στο έπακρο τις δυνατότητές τους στη συγγραφή κειμένων, υπαγορεύοντας άμεσα στον υπολογιστή το κείμενο που έχουν στο μυαλό τους και πετυχαίνοντας έτσι καλύτερα αποτελέσματα [25].

Ευρεία εμπορική εφαρμογή βρίσκουν τα τελευταία χρόνια συστήματα αναγνώρισης φωνής στον τομέα της αυτοκινητοβιομηχανίας για το χειρισμό ορισμένων λειτουργιών. Παραδείγματος χάριν, η αναζήτηση πληροφοριών κατά τη διάρκεια της οδήγησης, όπως είναι η αναζήτηση διευθύνσεων μέσω ενός συστήματος πλοήγησης, είναι ασφαλέστερη όταν γίνεται με φωνητικές εντολές σε σύγκριση με την πλέον ευρεία μέθοδο της χειροκίνητης εισαγωγής μέσω οθόνης αφής [26]. Παρόμοιες ιδέες έχουν βρει εφαρμογή τόσο στην πολεμική όσο και στην πολιτική αεροπορία με προτάσεις μάλιστα ακόμα και για χειρισμό μη επανδρωμένου αεροσκάφους από απόσταση με χρήση αναγνώρισης φωνής [27].

Με την επιτυχία της αναγνώρισης φωνής γίνεται πραγματικότητα η ιδέα των έξυπνων σπιτιών και των αυτοματισμών με χρήση φωνητικών εντολών [28], ενώ συστήματα αναγνώρισης έχουν χρησιμοποιηθεί επιτυχώς σε βιντεοπαιχνίδια, σε αυτόματη μετάφραση από φωνή σε φωνή (speech-to-speech translation) [29], σε ρομποτικά συστήματα [30]. Σίγουρα ένας μεγάλος δρόμος έχει ανοιχτεί, αλλά υπάρχει μεγάλη απόσταση ακόμα που πρέπει να διανυθεί, όπως θα φανεί και στη συνέχεια.

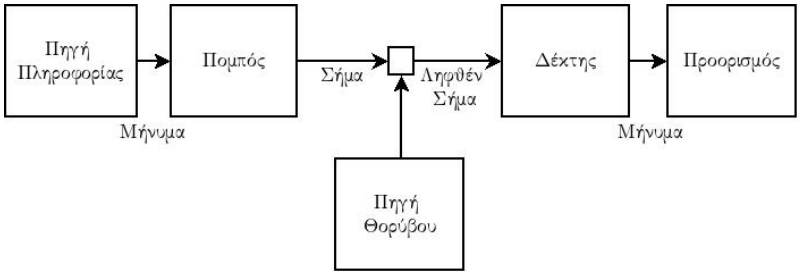
1.3 Αυτόματη Αναγνώριση Φωνής από Απόσταση

Τα συστήματα Αυτόματης Αναγνώρισης Φωνής έχουν φτάσει σε ένα αναμφισβήτητα πολύ καλό επίπεδο στις μέρες μας, έχοντας τη δυνατότητα να καλύπτουν ένα τεράστιο λεξιλόγιο - για την ακρίβεια απεριόριστο, δεδομένου ότι υπάρχουν οι κατάλληλες βάσεις δεδομένων και ο απαραίτητος χρόνος για εκπαίδευση - με εξαιρετικά ποσοστά απόδοσης. Το πρόβλημα, όμως, κάθε άλλο παρά λυμένο μπορεί να θεωρείται.

Η απόδοση ενός συστήματος αναγνώρισης μειώνεται δραματικά όταν το μικρόφωνο καταγραφής του σήματος της ανθρώπινης φωνής απομακρύνεται από το στόμα του ομιλητή. Πριν εξηγήσουμε αναλυτικά τους λόγους για τους οποίους αυτό συμβαίνει και τις προκλήσεις που έχει σήμερα να αντιμετωπίσει η ερευνητική κοινότητα που δραστηριοποιείται στο χώρο της αναγνώρισης φωνής, αξίζει να δούμε την Αυτόματη Αναγνώριση Φωνής υπό τη σκοπιά της Θεωρίας Πληροφορίας (Information Theory) [31], όπως την εισήγαγε ο Shannon στη μνημειώδη εργασία του [32].

Υπό αυτη την οπτική γωνία, κάθε σύστημα επικοινωνίας μπορεί να θεωρηθεί πως έχει τη γενική μορφή του Σχήματος 1.1. Η ομιλία, ως ένα μέσο επικοινωνίας, μπορεί κι αυτή να περιγραφεί υπό αυτό το πρίσμα. Η πηγή πληροφορίας είναι ο εγκέφαλος του ομιλητή (ή ένα μηχάνημα σε περίπτωση συνθετικής φωνής), το μήνυμα είναι η ακολουθία λέξεων και ο προορισμός είναι ο εγκέφαλος του συνομιλητή ή ένα μηχάνημα σε περίπτωση που θέλουμε να

έχουμε Αυτόματη Αναγνώριση Φωνής. Ρόλος του πομπού είναι η μετατροπή του μηνύματος σε σήμα, που δεν είναι άλλο από το ακουστικό κύμα που μεταφέρει την πληροφορία της φωνής, ενώ ρόλος του δέκτη είναι η ανακατασκευή του μηνύματος από το σήμα που λαμβάνει.



Σχήμα 1.1: Σχηματικό διάγραμμα ενός συστήματος επικοινωνίας. [*εικόνα προσαρμοσμένη από [32]*]

Όπως γίνεται προφανές, ζητούμενο είναι να έχουμε το μικρότερο δυνατό σφάλμα ανακατασκευής. Στην ιδανική περίπτωση που ξέραμε επακριβώς την παραμόρφωση που εισάγει η πηγή θορύβου στο σήμα, αλλά και το μηχανισμό παραγωγής του σήματος με δεδομένο το μήνυμα της πηγής πληροφορίας, το σφάλμα θα μπορούσε να είναι μηδενικό, αφού για κάθε πιθανή ακολουθία λέξεων θα ξέραμε εκ των προτέρων ποιο είναι το ληφθέν σήμα [33]. Προφανώς, κάτι τέτοιο δεν ισχύει, με αποτέλεσμα ο στόχος μας να γίνεται πολύ πιο δύσκολος. Ο στόχος αυτός γίνεται δυσκολότερος όσο πιο ισχυρή είναι η επίδραση της πηγής θορύβου που παρεμβάλλεται μεταξύ πομπού και δέκτη. Εκεί ακριβώς έγκειται η επιπλέον δυσκολία της Αναγνώρισης Φωνής από Απόσταση (Distant Speech Recognition DSR), εφόσον όταν το μικρόφωνο (ή η συστοιχία μικροφώνων) βρίσκεται μακριά από τον ομιλητή εισάγονται αλλοιώσεις που μπορούν να αποδοθούν σε τρεις κατηγορίες παραγόντων [34]:

- στο θόρυβο υποβάθρου, που αποτελείται από όλα τα ηχητικά σήματα τα οποία δεν περιλαμβάνονται στο επιθυμητό σήμα καταγραφής. Αυτά μπορεί να οφείλονται σε μηχανήματα που βρίσκονται στο χώρο, σε περιβαλλοντικούς θορύβους (θρόισμα φύλλων, βουητό αέρα, κ.ά.), σε ομιλία άλλων ομιλητών κ.ά.,
- στην ηχώ και την αντήχηση (reverberation), που οφείλονται σε ανακλάσεις του ήχου σε επιφάνειες του χώρου καταγραφής (κυρίως τοίχοι του δωματίου) και που αναπόφευκτα καταγράφονται,
- σε άλλους παράγοντες που σχετίζονται με το περιβάλλον καταγραφής, όπως είναι οι ιδιοσυχνότητες του δωματίου, η διεύθυνση του κεφαλιού του ομιλητή, το φαινόμενο Lombard [2], κ.ά.

Οι παραμορφώσεις αυτές εισάγουν έναν πολύ μεγάλο όγκο περιττής, ως προς το σκοπό της αναγνώρισης φωνής, πληροφορίας. Ένας ακόμη λόγος εισαγωγής περιττής πληροφορίας είναι τα ιδιαίτερα χαρακτηριστικά του ομιλητή που εκφέρει το μήνυμα, το οποίο σαφώς είναι ένα γενικό πρόβλημα της αναγνώρισης (και όχι μόνο από απόσταση), όταν αυτή θέλουμε να είναι ανεξάρτητη του ομιλητή. Το μήνυμα, όμως, που πρέπει να ανακατασκευαστεί, δηλαδή η ζητούμενη ακολουθία λέξεων, είναι η ίδια και δεν πρέπει να επηρεάζεται από οποιουδήποτε

[2]Με την ονομασία "φαινόμενο Lombart" είναι γνωστή η τάση των ανθρώπων να μεταβάλλουν, ακούσια και ασυναίσθητα, τα χαρακτηριστικά της φωνής τους όταν βρίσκονται σε θορυβώδεις συνθήκες ώστε να διασφαλίζουν ότι γίνονται κατανοητοί, κάτι που έχει αρνητικές επιπτώσεις κατά την Αυτόματη Αναγνώριση Φωνής [35].

είδους εξωτερικό παράγοντα. Μιλώντας με όρους θεωρίας πληροφορίας, ο ρυθμός πληροφορίας στο σήμα φωνής εκτιμάται στα $40 kbits/sec$, τη στιγμή που ο ρυθμός πληροφορίας του "καθαρού" (γραπτού) γλωσσικού μηνύματος, δηλαδή της χρήσιμης πληροφορίας, εκτιμάται σε μόλις $60 bits/sec$, 3 τάξεις μεγέθους πιο χαμηλά [36].

Οι επιπλέον αυτές δυσκολίες που εισάγει η Αναγνώριση Φωνής από Απόσταση, αλλά και το αναμφισβήτητα μεγαλύτερο εύρος εφαρμογών που ανοίγεται εάν ξεπεραστούν οι δυσκολίες, ωθούν όλο και περισσότερο την ερευνητική κοινότητα να στρέψει τις προσπάθειές της προς αυτή την κατεύθυνση, ώστε να επιτευχθεί το τελικό ζητούμενο της εύρωστης αναγνώρισης, της αναγνώρισης, δηλαδή, που δεν επηρεάζεται από τους διάφορους εξωτερικούς παράγοντες, όπως αυτοί που παρουσιάστηκαν. Αυτό είναι το πλαίσιο και το ερευνητικό πεδίο στο οποίο φιλοδοξεί να συνεισφέρει η παρούσα εργασία.

1.4 Στόχος της Εργασίας

Στόχος της παρούσης Διπλωματικής εργασίας είναι η συγκριτική μελέτη διαφορετικών μεθόδων εξαγωγής χαρακτηριστικών για χρήση στο πεδίο της Αναγνώρισης Φωνής από Απόσταση, με χρήση ενός μικροφώνου.

Μέχρι τα προηγούμενα χρόνια, η αξιολόγηση νέων χαρακτηριστικών γινόταν σε καθαρές συνθήκες, όπου πράγματι η απόδοση των συστημάτων κυμαίνεται πλέον σε ικανοποιητικά επίπεδα. Ωστόσο, όταν το μικρόφωνο απομακρύνεται από το στόμα του ομιλητή, η απόδοση μειώνεται αισθητά λόγω μιας ποικιλίας παραμορφώσεων, όπως είναι ο προσθετικός θόρυβος υποβάθρου και ο συνελικτικός θόρυβος που εισάγει το φαινόμενο της αντήχησης. Για το λόγο αυτό, πλέον δίνεται αρκετή έμφαση στον υπολογισμό εύρωστων χαρακτηριστικών, δηλαδή ακουστικών χαρακτηριστικών που δεν επηρεάζονται από τις συνθήκες του περιβάλλοντος. Η πλειοψηφία, όμως, των αποτελεσμάτων στη βιβλιογραφία εξάγονται από πειραματικούς ελέγχους σε ελεγχόμενες συνθήκες θορύβου, δηλαδή ύστερα από τεχνητή αλλοίωση των δεδομένων με προσθετικό ή συνελικτικό θόρυβο. Ιδιαίτερη έμφαση, λοιπόν, δόθηκε στην παρούσα εργασία στις πειραματικές συνθήκες, κάνοντας εκτενή χρήση της βάσης δεδομένων ATHENA, που έχει αναπτυχθεί για Αναγνώριση Φωνής από Απόσταση και αποτελείται από ηχογραφήσεις αλλοιωμένες από τις πιθανές παραμορφώσεις με τις οποίες μπορεί να έρθει αντιμέτωπο ένα σύστημα αναγνώρισης από απόσταση.

Επειδή στην πράξη η εκπαίδευση του συστήματος είναι αδύνατη σε πραγματικές συνθήκες, εφόσον αυτές είναι απρόβλεπτες, η αναντιστοιχία μεταξύ εκπαίδευσης και ελέγχου είναι σε όλα τα πειράματα επιτηδευμένα πολύ μεγάλη, με την πρώτη να λαμβάνει χώρα σε καθαρές συνθήκες, χωρίς πηγές θορύβου και χωρίς να λάβει χώρα κάποια προσαρμογή (adaptation). Εξάλλου, υπενθυμίζεται ότι στόχος είναι η συγκριτική μελέτη των χαρακτηριστικών και όχι η υλοποίηση ενός συστήματος με το ελάχιστο δυνατό σφάλμα αναγνώρισης. Στοχεύοντας, μάλιστα, σε μία άμεση σύγκριση των χαρακτηριστικών, χωρίς να υπεισέρχονται άλλοι παράγοντες, έγινε προσπάθεια να μειωθεί στο ελάχιστο η επίδραση λοιπών παραγόντων, όπως είναι για παράδειγμα το γλωσσικό μοντέλο[3].

1.5 Διάρθρωση της Εργασίας

Η παρούσα εργασία διαρθρώνεται ως εξής:

[3]Περισσότερες λεπτομέρειες μπορούν να βρεθούν στην Ενότητα 4.2.

Στο Κεφάλαιο 1 γίνεται μία συνοπτική παρουσίαση των θεμελιωδών εννοιών οι οποίες αποτελούν το πεδίο ενδιαφέροντος της εργασίας. Εξηγείται η μοναδικότητα της ανθρώπινης ομιλίας και η σημασία της για την ανθρώπινη ύπαρξη και δίνεται το ιστορικό πλαίσιο μέσα στο οποίο αναπτύχθηκε η περιοχή εκείνη που ονομάζουμε Αυτόματη Αναγνώριση Φωνής. Παρουσιάζονται τομείς όπου μπορεί να βρει εφαρμογή η συγκεκριμένη τεχνολογία, με πολλές εφαρμογές να έχουν ήδη όχι μόνο ερευνητικό, αλλά και εμπορικό ενδιαφέρον και αναφέρονται ποιοτικά τα διάφορα προβλήματα που συνδέονται με την Αναγνώριση Φωνής από Απόσταση.

Τα Κεφάλαια 2 και 3 συνοψίζουν το βασικό θεωρητικό υπόβαθρο πάνω στο οποίο στηρίζεται ο κορμός της εργασίας. Στο Κεφάλαιο 2 παρουσιάζονται τα βασικά βήματα που ακολουθεί ένα τυπικό σύστημα αναγνώρισης φωνής, από τη μοντελοποίηση της ακουστικής και γλωσσικής πληροφορίας και την εκπαίδευση του συστήματος μέχρι την αποκωδικοποίηση του σήματος προς αναγνώριση. Τα τελευταία χρόνια έχει αναπτυχθεί ένα νέο πλαίσιο εργασίας για τα συστήματα αναγνώρισης, που στηρίζεται στις Μηχανές Πεπερασμένης Κατάστασης με Βάρη (Weighted Finite-State Machines - WFSTs) και που πλέον αποτελεί τη state-of-the-art προσέγγιση. Η θεωρία πίσω από τα WFSTs, καθώς και ο αποδοτικός τρόπος χρήσης τους για τους σκοπούς της αναγνώρισης φωνής, παρουσιάζονται στο Κεφάλαιο 3. Το Κεφάλαιο αυτό κρίθηκε απαραίτητο για την ομαλή μετάβαση στα επόμενα, καθώς για όλα τα πειράματα και τις συγκρίσεις της παρούσης εργασίας χρησιμοποιείται το σύστημα Kaldi, το οποίο αποτελεί το βασικό εκπρόσωπο εργαλείων ανάπτυξης συστημάτων για Αυτόματη Αναγνώριση Φωνής στηριζόμενων στη λογική των WFSTs, κάνοντας εκτενή χρήση της σχετικής βιβλιοθήκης OpenFST [37].

Στο Κεφάλαιο 4 περιγράφεται το βασικό σύστημα αναγνώρισης που χρησιμοποιείται για το πειραματικό μέρος της εργασίας, όπου ακολουθείται το καθιερωμένο πρότυπο των HMMs/GMMs, με τελικό στόχο, όμως, την αναγνώριση φωνημάτων αντί λέξεων. Δίνονται οι κινητήριες γραμμές πίσω από την ιδέα της αναγνώρισης φωνημάτων χωρίς γλωσσικό μοντέλο, παρουσιάζονται οι βασικές παράμετροι που διατηρούνται σταθερές για όλα τα πειράματα και αναλύονται τα διακριτά στάδια της αναγνώρισης, σύμφωνα με τις βασικές αρχές του συστήματος Kaldi. Ακόμα, παρουσιάζονται οι ελληνικές βάσεις δεδομένων ATHENA και Logotypografia που θα χρησιμοποιηθούν για τα πειράματα, ενώ επίσης γίνεται ένας απλός πειραματισμός της επίδρασης που θα είχε ένα στατιστικό γλωσσικό μοντέλο στα τελικά αποτελέσματα της αναγνώρισης.

Στα Κεφάλαια 5 και 6 παρουσιάζονται τα πιο διαδεδομένα σύνολα χαρακτηριστικών που χρησιμοποιούνται στην πράξη σήμερα, οι Συντελεστές Αναφάσματος στις Mel-Συχνότητες (Mel-Frequency Cepstrum Coefficients - MFCCs) και οι συντελεστές Γραμμικής Πρόβλεψης βασισμένοι στην Αντίληψη (Perceptual Linear Prediction - PLP), αντίστοιχα. Αναλύονται τα απαραίτητα θεωρητικά στοιχεία που οδηγούν στην εξαγωγή των συγκεκριμένων χαρακτηριστικών και γίνεται μία σειρά πειραμάτων που αναδεικνύουν την επίδραση που έχουν διαφορετικές παραμετροποιήσεις κατά την εξαγωγή τους, όπως είναι το μήκος του χρονικού παραθύρου, το εύρος της εν χρήση συστοιχίας φίλτρων, η μέγιστη τάξη δυναμικών συντελεστών που προσδίδει χρήσιμη πληροφορία και το χρονικό παράθυρο που λαμβάνεται υπόψιν κατά τον υπολογισμό των δυναμικών αυτών συντελεστών. Τα MFCCs δίνουν γενικά καλύτερα αποτελέσματα από τα PLPs, ενώ σημαντική φαίνεται να είναι η επιλογή του κατάλληλου παραθύρου για την εξαγωγή των δυναμικών χαρακτηριστικών. Βοηθητική ως προς τα τελικά ποσοστά αναγνώρισης είναι η κανονικοποίηση τόσο των τελικών συντελεστών στο πεδίο του αναφάσματος (Cepstral Mean (& Variance) Normalization - CM(V)N), όσο και του αρχικού σήματος στο πεδίο του χρόνου.

Εκτός των MFCCs και PLPs, αναλύονται σε θεωρητικό και πειραματικό επίπεδο και

ορισμένες παραλλαγές τους που στοχεύουν σε μεγαλύτερη ευρωστία, όπως είναι οι Δέλτα-Φασματικοί Αναφασματικοί Συντελεστές (Delta-Spectral Cepstral Coefficients - DSCCs), τα PLPs με Σχετική Φασματική Ανάλυση (Relative Spectral Analysis - RASTA) και τα J-RASTA-PLPs. Η RASTA φαίνεται να βελτιώνει την αποτελεσματικότητα των PLPs, χωρίς, ωστόσο, να καταφέρνει να δώσει τελικά καλύτερα αποτελέσματα από τα MFCCs, ενώ η J-RASTA δεν επιφέρει επιπλέον βελτιώσεις. Τα DSCCs, από την άλλη, δίνουν πολύ υποσχόμενα αποτελέσματα, ιδιαίτερα σε συνθήκες αναγνώρισης από απόσταση, όταν χρησιμοποιούνται ως αυτόνομο σύνολο χαρακτηριστικών και όχι σε συνδυασμό με τα στατικά MFCCs, όπως έχουν προταθεί αρχικά.

Στο Κεφάλαιο 7 παρουσιάζονται οι πιο διαδεδομένες μέθοδοι που έχουν χρησιμοποιηθεί για τη μείωση της διαστασιμότητας των διανυσμάτων χαρακτηριστικών. Συγκεκριμένα, αναλύεται η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA), η Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis - LDA) και η Ετεροσκεδαστική LDA (Heteroscedastic LDA - HLDA), η οποία έχει κοινά σημεία με το Γραμμικό Μετασχηματισμό Μέγιστης Πιθανοφάνειας (Maximum Likelihood Linear Transform - MLLT). Οι μέθοδοι αυτές χρησιμοποιούνται αφότου το κάθε διάνυσμα χαρακτηριστικών συνενώνεται με γειτονικά του διανύσματα, σε μια προσπάθεια να αντικατοπτριστούν όχι μόνο τα στατικά, αλλά και τα δυναμικά χαρακτηριστικά ενός σήματος φωνής. Σε κάθε περίπτωση, η σύγκριση ανάμεσα σε PCA, LDA και LDA+MLLT ευνοεί την τελευταία μέθοδο. Ωστόσο, ενώ σε καθαρές συνθήκες η LDA+MLLT παρουσιάζει συγκεκριμένα πλεονεκτήματα, σε συνθήκες αναγνώρισης από απόσταση η απλή επαύξηση του διανύσματος των MFCCs με τους δυναμικούς (Δ και $\Delta\Delta$) συντελεστές δίνει σταθερά καλύτερα αποτελέσματα.

Η εξαγωγή των πιο συνηθισμένων συνόλων χαρακτηριστικών, δηλαδή των MFCCs και των PLPs, αλλά και πολλών άλλων συνόλων που έχουν προταθεί στη βιβλιογραφία, στηρίζεται στο φάσμα ισχύος των σημάτων, καθώς αυτό συνδέεται με την τετραγωνική τους ενέργεια. Στο Κεφάλαιο 8 αναλύεται ένας εναλλακτικός τρόπος υπολογισμού της ενέργειας ενός σήματος, που βασίζεται στη χρήση του Τελεστή Teager Ενέργειας (Teager Energy Operator - TEO). Για να διατηρηθούν τα υπολογιστικά πλεονεκτήματα της εργασίας στο πεδίο της συχνότητας, προτείνεται ένας νέος τρόπος χρήσης του TEO στο πεδίο αυτό, μέσω της εισαγωγής του Φάσματος Teager Ισχύος (Teager Power Spectrum), που μπορεί να ενταχθεί στη ροή εργασίας εξαγωγής των περισσότερων συνόλων χαρακτηριστικών που βασίζονται στην ενέργεια βραχέος χρόνου και στο Φάσμα Ισχύος. Επίσης, μέσα στο πλαίσιο αυτό, προτείνεται και αξιολογείται ο συνδυασμός των δύο τελεστών, του TEO και του Τελεστή Τετραγωνικής Ενέργειας (Squared Energy Operator - SEO), μέσω της προσέγγισης της ενέργειας στις χαμηλές συχνότητες με χρήση του TEO και στις υψηλές με χρήση του SEO. Μέσω αναλυτικής αναζήτησης του ορίου μέχρι το οποίο θα γίνει χρήση του TEO, μπορούν να επιτευχθούν σημαντικές βελτιώσεις στα τελικά αποτελέσματα αναγνώρισης. Ιδιαίτερα αξιοσημείωτη είναι η βελτίωση που επιτυγχάνεται όταν η εν λόγω μέθοδος εισάγεται στο πλαίσιο εργασίας των PLPs.

Στο ίδιο κεφάλαιο αναλύεται η δυνατότητα αποδιαμόρφωσης σημάτων με χρήση του TEO και του Αλγορίθμου Διαχωρισμού Ενέργειας (Energy Separation Algorithm - ESA), ώστε να εκτιμάται το στιγμιαίο πλάτος και η στιγμιαία συχνότητά τους. Βάσει αυτών, μπορούν να εξαχθούν διάφορα χαρακτηριστικά, τα οποία μελετάται κατά πόσο μπορούν να δράσουν ευεργετικά σε ένα σύστημα αναγνώρισης από απόσταση. Συγκεκριμένα, για την εξαγωγή των χαρακτηριστικών χρησιμοποιείται ο αλγόριθμος Gabor ESA, ενώ από τις παραμετροποιήσεις που δοκιμάστηκαν, φαίνεται πως 12 Gabor φίλτρα με 70% επικάλυψη αποτελούν την καλύτερη επιλογή. Μέσω μιας σειράς πειραμάτων, τόσο σε πραγματικά, όσο και σε συνθε-

τικά δεδομένα, φαίνεται πως ο συνδυασμός συγκεκριμένων χαρακτηριστικών διαμόρφωσης που σχετίζονται με τη συχνότητα οδηγούν σε βελτιωμένα αποτελέσματα αναγνώρισης όταν συνδυάζονται με τα MFCCs ή τα DSCCs, ιδίως σε συνθήκες συνελικτικού θορύβου λόγω αντήχησης. Τα καλύτερα αποτελέσματα παράγονται από το συνδυασμό των DSCCs με την πρώτη σταθμισμένη ροπή της στιγμιαίας συχνότητας με βάρος το τετραγωνικό πλάτος.

Τέλος, το Κεφάλαιο 9 αποτελεί μία σύνοψη του συνόλου της παρούσης εργασίας, όπου παρουσιάζονται τα πιο σημαντικά αποτελέσματα των διαφόρων πειραμάτων. Ακόμα, δίνονται κάποιες γενικές κατευθυντήριες ιδέες για περαιτέρω μελέτη και έρευνα πάνω στα θέματα με τα οποία η εργασία ασχολείται.

1.6 Συνεισφορά της Εργασίας

Πολύ συνοπτικά, οι κύριες συνεισφορές της παρούσης εργασίας είναι οι εξής:

- Ανάδειξη της σημασίας του μήκους του χρονικού παραθύρου που λαμβάνεται υπόψιν κατά των υπολογισμό των κλασικών δυναμικών (Δ και $\Delta\Delta$) χαρακτηριστικών, με την εν λόγω παράμετρο να πρέπει να παίρνει μικρές τιμές για αναγνώριση σε καθαρές συνθήκες και μεγαλύτερες για αναγνώριση από απόσταση.

- Ανάδειξη της σημασίας της κανονικοποίησης των σημάτων στο πεδίο του χρόνου, πριν την οποιαδήποτε περαιτέρω επεξεργασία τους.

- Πειραματικές ενδείξεις που συνηγορούν υπέρ της χρήσης CMVN ανά εκφορά και όχι ανά ομιλητή στην περίπτωση συνθηκών αναγνώρισης από απόσταση.

- Ανάδειξη της ευρωστίας των DSCCs και εξαγωγή βελτιωμένων αποτελεσμάτων όταν αυτά χρησιμοποιούνται ως αυτόνομο σύνολο ακουστικών χαρακτηριστικών, χωρίς να συνδυάζονται με στατική πληροφορία.

- Επιβεβαίωση της βελτιωτικής επίδρασης της RASTA ανάλυσης στα PLPs, με τη J-RASTA, όμως, να μη φαίνεται να δίνει επιπλέον βελτιώσεις, ενώ σε κάθε περίπτωση τα (απλούστερα) MFCCs φαίνεται να υπερτερούν ως προς την τελική απόδοση.

- Επιβεβαίωση της σχετικής σύγκρισης μεταξύ μεθόδων μείωσης της διαστασιμότητας μετά τη σύνδεση διαδοχικών πλαισίων χαρακτηριστικών, με την LDA+MLLT να υπερτερεί έναντι της απλής LDA, η οποία με τη σειρά της δίνει καλύτερα αποτελέσματα από την PCA. Ωστόσο, ειδικά σε συνθήκες αναγνώρισης από απόσταση, καμία από τις μεθόδους δεν κατάφερε να δώσει καλύτερα αποτελέσματα όταν συγκρίθηκε με την απλή χρήση των Δ και $\Delta\Delta$ συντελεστών.

- Εισαγωγή της έννοιας του TPS και πρόταση για συνδυασμό του με το κλασικό φάσμα ισχύος με τρόπο τέτοιο που οδηγεί σε καλύτερα αποτελέσματα όταν χρησιμοποιείται στη ροή εργασίας γνωστών μεθόδων εξαγωγής χαρακτηριστικών.

- Σύγκριση ποικίλων AM-FM χαρακτηριστικών, ορισμένα από τα οποία έχουν χρησιμοποιηθεί ξανά στη βιβλιογραφία και ορισμένα όχι. Εξ αυτών, ανάδειξη των MIF και F, δηλαδή της μέσης στιγμιαίας συχνότητας και της πρώτης ροπής της στιγμιαίας συχνότητας σταθμισμένης με το τετραγωνικό στιγμιαίο πλάτος, ως τα πλέον υποσχόμενα χαρακτηριστικά, κυρίως σε συνδυασμό με τα MFCCs ή DSCCs. Τελικά, αναλόγως των αναμενόμενων συνθηκών αναγνώρισης προτείνεται η χρήση των MFCCs+F (απουσία

προσθετικού θορύβου, ήπια ή καθόλου αντήχηση), των DSCCs+F (ήπιος προσθετικός θόρυβος με ή χωρίς αντήχηση) ή των σκέτων DSCCs (έντονος προσθετικός θόρυβος με ή χωρίς αντήχηση).

Κεφάλαιο 2

Αυτόματη Αναγνώριση Φωνής

2.1 Εξαγωγή Χαρακτηριστικών

Για την επιτυχή αναγνώριση φωνής, θα πρέπει πρώτα απ' όλα να βρούμε κατάλληλα σύνολα χαρακτηριστικών που διαχωρίζουν ένα τμήμα σήματος φωνής από ένα άλλο και ομαδοποιούν παρόμοια τμήματα, ώστε τελικά τα ίδια φωνήματα να ομαδοποιούνται μαζί και να διαχωρίζονται από τα διαφορετικά φωνήματα. Έτσι, όταν έρθει η ώρα να αναγνωριστεί ένα καινούριο φώνημα, τα χαρακτηριστικά του ελπίζουμε ότι θα είναι τέτοια, ώστε να ενταχθεί στην κατάλληλη ομάδα.

Πολλά τέτοια σύνολα χαρακτηριστικών έχουν προταθεί στη βιβλιογραφία, καθένα από τα οποία απαιτεί μία συγκεκριμένη διαδικασία επεξεργασίας του σήματος φωνής. Καθώς το μεγαλύτερο μέρος της εργασίας ασχολείται με την εξαγωγή και τη χρήση διαφορετικών συνόλων χαρακτηριστικών, καθώς και τη μεταξύ τους σύγκριση, στην Ενότητα αυτή θα αναφερθούμε στις βασικές αρχές που διέπουν το σύνολο των σχετικών διαδικασιών και κυρίως στη σημασία και την ποιοτική ερμηνεία της έννοιας "εξαγωγή χαρακτηριστικών". Αναλυτικά πληροφορίες για συγκεκριμένα σύνολα, μαζί με πειραματικά αποτελέσματα βρίσκονται στα Κεφάλαια 5 - 8, καθώς και στο Παράρτημα I.

Αρχικά, το σήμα φωνής δεν είναι παρά ένα διάμηκες ηχητικό κύμα, δηλαδή μία ακολουθία από πυκνώματα και αραιώματα που διαδίδονται στον αέρα, χάρη στην ελαστική ιδιότητα που έχει ο τελευταίος [38]. Για να αναπαρασταθεί ένα τέτοιο ηχητικό δεδομένο στον υπολογιστή το πρώτο βήμα είναι η καταγραφή του με ένα μικρόφωνο. Το μικρόφωνο είναι ένας μετατροπέας ο οποίος είναι επιφορτισμένος με τη μετατροπή του ακουστικού σήματος σε ηλεκτρικό ή, με άλλα λόγια, της ακουστικής ενέργειας σε ηλεκτρική. Στην περίπτωση ενός ιδανικού μικροφώνου (συσκευή η οποία δεν υπάρχει, αλλά μπορούμε να την υποθέσουμε για λόγους απλότητας), η παραγόμενη ηλεκτρική κυματομορφή θα έχει ακριβώς τα ίδια χαρακτηριστικά με την ακουστική κυματομορφή, με τα πυκνώματα του αέρα να αντιστοιχούν σε θετικά ηλεκτρικά δυναμικά και τα αραιώματα σε αρνητικά.

Στην απλούστερη περίπτωση, το ηχητικό κύμα προσκρούει σε ένα λεπτό σύρμα που ονομάζεται διάφραγμα, προκαλώντας την ταλάντωσή του με συχνότητα ίδια με αυτή του κύματος. Η ταλάντωση του διαφράγματος λαμβάνει χώρα εντός ενός στατικού μαγνητικού πεδίου. Όμως, η κίνηση ενός αγωγού κάθετη στις δυναμικές γραμμές ενός μαγνητικού πεδίου έχει ως αποτέλεσμα την ανάπτυξη ηλεκτρικού ρεύματος στον αγωγό. Έτσι, το διάφραγμα, και κατά συνέπεια το κλειστό ηλεκτρικό κύκλωμα του μικροφώνου, διαρρέεται από εναλλασσόμενο ηλεκτρικό ρεύμα με συχνότητα που ταυτίζεται με τη συχνότητα του ακουστικού κύματος. Το ηλεκτρικό αυτό ρεύμα είναι η έξοδος του μικροφώνου, βλέποντάς τον ως μετατροπέα με

είσοδο το σήμα φωνής.

Εν συνεχεία, ακολουθεί η μετατροπή του ηλεκτρικού σήματος από αναλογικό σε ψηφιακό μέσω της διαδικασίας της δειγματοληψίας, δηλαδή της μέτρησης της έντασής του ανά τακτά, σταθερά χρονικά διαστήματα [33] και της κβάντισης, δηλαδή της μετατροπής των πραγματικών τιμών σε διακριτές στάθμες που μπορούν να αναπαρασταθούν από πεπερασμένο αριθμό bits (συνήθως 8 ή 16). Όπως είναι γνωστό, η συχνότητα δειγματοληψίας θα πρέπει να ικανοποιεί τη συνθήκη Nyquist, δηλαδή να είναι τουλάχιστον διπλάσια της μέγιστης συχνότητας του σήματος, δεδομένου ότι το υπό εξέταση σήμα είναι ζωνοπεριορισμένο, ώστε το δειγματειλημμένο σήμα να αποτελεί μία πιστή και μοναδική αναπαράσταση του σήματος [31]. Παρόλο που το σήμα φωνής δεν είναι αυστηρά ζωνοπεριορισμένο, η συχνοτική του πληροφορία μειώνεται γρήγορα στις υψηλές συχνότητες, ενώ σχεδόν όλος ο όγκος πληροφορίας στην ανθρώπινη φωνή βρίσκεται κάτω από τα $10kHz$, οπότε μία συχνότητα δειγματοληψίας ίση με $20kHz$ είναι αρκετή. Στις τηλεφωνικές γραμμές, το σήμα φιλτράρεται με αποτέλεσμα όλη η συχνοτική πληροφορία να βρίσκεται κάτω από τα $4kHz$, οπότε αρκεί μια συχνότητα δειγματοληψίας ίση με $8kHz$. Η πλέον συνηθισμένη και αποδεκτή συχνότητα δειγματοληψίας για φωνή καταγεγραμμένη με μικρόφωνο και για τους σκοπούς της Αυτόματης Αναγνώρισης Φωνής είναι $16kHz$.

Κατόπιν της διαδικασίας αυτής, το σήμα φωνής αναπαριστάται ως μια ακολουθία κβαντισμένων δειγμάτων. Αυτό που κλασικά ακολουθείται κατά την επεξεργασία φωνής είναι ο διαχωρισμός του σήματος σε μικρά, σταθερής διάρκειας και συνήθως επικαλυπτόμενα πλαίσια (frames). Παρόλο που τελευταία γίνονται προσεγγίσεις με πλαίσια διάρκειας μέχρι 200-300$msec$, παραδοσιακά τα εν λόγω πλαίσια έχουν διάρκεια της τάξης των 10-30$msec$.

Η κεντρική ιδέα πίσω από την τεχνική της παραθύρωσης βασίζεται στο εξής. Σαφώς, η φωνή είναι ένα τυχαίο σήμα με έντονες και μεγάλες διακυμάνσεις στη διάρκεια του χρόνου, χωρίς κάποια περιοδικότητα και χωρίς να ικανοποιεί τις αναγκαίες συνθήκες της στασιμότητας στη γενική περίπτωση. Ωστόσο, τα χρονικά, αλλά και τα φασματικά του χαρακτηριστικά μπορούν να θεωρηθούν σταθερά για τμήματα διάρκειας $10-30msec$ [31]. Η τμηματική αυτή στασιμότητα της φωνής δίνει τη δυνατότητα χρήσης κλασικών τεχνικών που χρησιμοποιούνται στην ψηφιακή επεξεργασία σήματος, όπως είναι η ανάλυση Fourier, παρόλο που δε θα ήταν δυνατή η χρήση τους απευθείας στο σήμα χωρίς την προεπεξεργασία της παραθύρωσης.

Μέσω της εκάστοτε διαδικασίας εξαγωγής χαρακτηριστικών, υπολογίζεται ένα διάνυσμα στο χώρο $\mathbb{R}^d$ που χαρακτηρίζει το κάθε πλαίσιο ξεχωριστά. Ονομάζοντας $\mathbf{o}_i$ το διάνυσμα που χαρακτηρίζει το πλαίσιο i, δημιουργείται η ακολουθία $O = \mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \cdots$, η οποία ονομάζεται ακολουθία των παρατηρήσεων. Η εργασία της Αυτόματης Αναγνώρισης Φωνής ανάγεται πλέον στην εύρεση της πλέον πιθανής ακολουθίας λέξεων $\hat{W} = \hat{w}_1, \hat{w}_2, \hat{w}_3, \cdots$, δεδομένης της ακολουθίας παρατηρήσεων O [1]. Φορμαλιστικά, έχουμε

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmax}} P(W|O)\,, \tag{2.1}$$

όπου $\mathcal{W}$ είναι το σύνολο των πιθανών ακολουθιών λέξεων και ο υπολογισμός της πιθανότητας $P(W|O)$ γίνεται βάσει του ακουστικού και του γλωσσικού μοντέλου. Για την ακρίβεια, σύμφωνα με τον κανόνα του Bayes ισχύει

$$P(W|O) = \frac{p(O|W)P(W)}{P(O)}\,, \tag{2.2}$$

[1] Τονίζεται ότι προφανώς δεν υπάρχει ένα προς ένα αντιστοίχιση μεταξύ των δεικτών των στοιχείων της ακολουθίας O και της ακολουθίας W.

οπότε λαμβάνουμε

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmax}}\, p(O|W)P(W)\,, \tag{2.3}$$

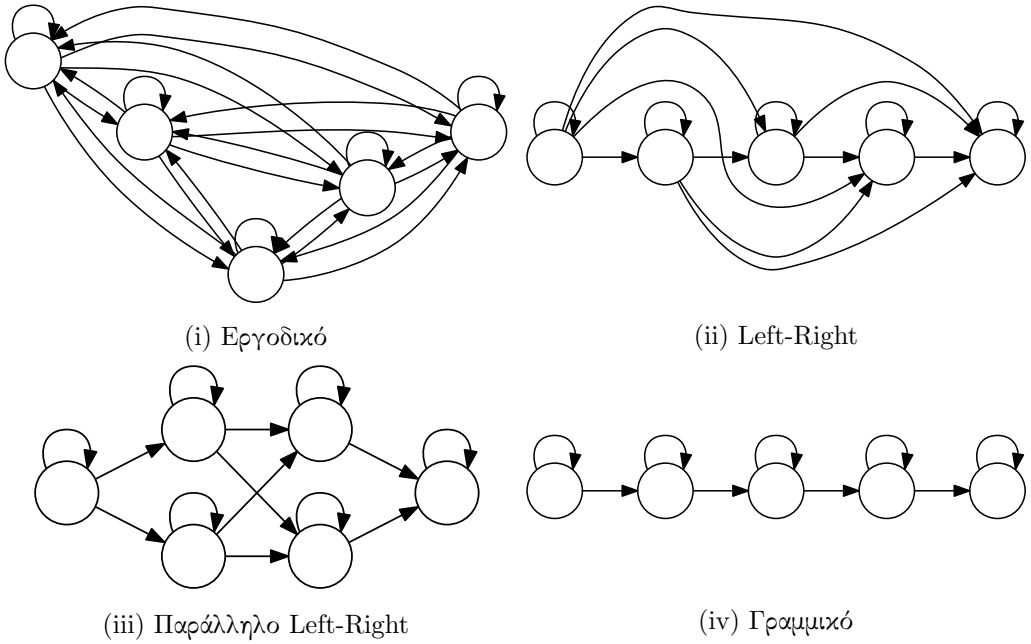
όπου $p(O|W)$ η ακουστική πιθανοφάνεια του O για το W και $P(W)$ η πρότερη πιθανότητα του W [13].

2.2 Ακουστικό Μοντέλο

Με τον όρο ακουστικό μοντέλο εννοούμε το στατιστικό εκείνο μοντέλο που χρησιμοποιείται κατά τη διαδικασία της Αναγνώρισης Φωνής για τον υπολογισμό της ακουστικής πιθανοφάνειας $p(O|W)$. Εάν η ακολουθία παρατηρήσεων O παράχθηκε κατά την εκφορά της ακολουθίας λέξεων W, τότε η εν λόγω πιθανοφάνεια αναμένεται να έχει υψηλή τιμή.

2.2.1 Κρυφά Μαρκοβιανά Μοντέλα

Στην πράξη, τα πιο ευρέως χρησιμοποιούμενα ακουστικά μοντέλα είναι τα Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models - HMMs), και συγκεκριμένα, left-right HMMs με επιτρεπτές μεταβάσεις μόνο μεταξύ διαδοχικών καταστάσεων ή με παραμονή στην ίδια κατάσταση (self-loop), που λέγονται και γραμμικά HMMs [39]. Η διαγραμματική μορφή ενός τέτοιου HMM παρουσιάζεται στο Σχήμα 2.1iv.



(i) Εργοδικό (ii) Left-Right

(iii) Παράλληλο Left-Right (iv) Γραμμικό

Σχήμα 2.1: Παραδείγματα τοπολογιών για HMMs. Τα γραμμικά HMMs είναι αυτά που κυρίως χρησιμοποιούνται στην Αυτόματη Αναγνώριση Φωνής.

Τυπικά, ένα οποιοδήποτε HMM ορίζεται πλήρως από το σύνολο των παραμέτρων [7]

$$\lambda = \{Q, V, A, B, \pi, F\}\,, \tag{2.4}$$

όπου

- $Q = \{q_1, q_2, \cdots, q_N\}$ ένα πεπερασμένο σύνολο καταστάσεων,
- $V = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M\}$ ένα πεπερασμένο σύνολο παρατηρήσεων που μπορεί να εξάγει το HMM (μία παρατήρηση ανά κατάσταση),
- $A = \{a_{ij}\}, i, j = 1, \cdots, N$ το σύνολο των πιθανοτήτων μετάβασης από μία κατάσταση i σε μια κατάσταση j, δηλαδή

$$a_{ij} = P(q_j \text{ για } t+1 | q_i \text{ για } t), \quad \sum_j a_{ij} = 1\,, \tag{2.5}$$

- $B = \{b_j(k)\}, j = 1, \cdots, N, k = 1, \cdots, M$ το σύνολο των πιθανοτήτων (ή κατανομών πιθανότητας) παραγωγής μιας παρατήρησης $\mathbf{v}_k$ από μια κατάσταση q_j, δηλαδή

$$b_j(k) = P(\mathbf{v}_k \text{ για } t | q_j \text{ για } t), \quad \sum_k b_j(k) = 1\,, \tag{2.6}$$

- $\pi = \{\pi_i\}, i = 1, \cdots, N$ το σύνολο των πιθανοτήτων αρχικής κατάστασης, δηλαδή

$$\pi_i = P(q_i \text{ για } t = 1), \quad \sum_i \pi_i = 1\,, \tag{2.7}$$

- F ένα πεπερασμένο σύνολο τελικών καταστάσεων, υποσύνολο του V.

Μια πιθανή παρατήρηση σε κάποια κατάσταση είναι ένα d-διάστατο διάνυσμα, όπως αυτό έχει υπολογιστεί από τη διαδικασία εξαγωγής χαρακτηριστικών. Εφόσον στη γενική περίπτωση, οι συντελεστές του διανύσματος μπορούν να λάβουν οποιαδήποτε τιμή στον πραγματικό άξονα, χωρίς να έχουν υποστεί κάποια διακριτοποίηση, τα μοντέλα που χρησιμοποιούνται συνήθως είναι συνεχή, οπότε η συνθήκη της σχέσης (2.6) γράφεται ως $\int_{-\infty}^{\infty} b_j(x)dx = 1$. Όπως έχει γίνει εμφανές, η εξαγόμενη παρατήρηση όταν το HMM βρίσκεται σε μια συγκεκριμένη κατάσταση εξαρτάται μόνο από την κατάσταση αυτή και όχι από πιθανές προηγούμενες καταστάσεις. Με άλλα λόγια, σε ένα HMM ισχύει η υπόθεση Markov.

Η ακουστική πιθανοφάνεια με ένα HMM, δηλαδή η πιθανότητα ένα HMM με παραμέτρους λ να παράξει μια ακολουθία παρατηρήσεων O, υπολογίζεται πρακτικά με χρήση του αλγορίθμου Viterbi, βάσει της πλέον πιθανής ακολουθίας καταστάσεων Q^*. Το λεγόμενο Viterbi score δίνεται ως

$$\begin{aligned} P_V &= \max_Q P(O, Q|\lambda) = P(O, Q^*|\lambda) && (2.8)\\ &= \max_{q_1, q_2, \cdots, q_T} \left\{ \pi_{q_1} b_{q_1}(\mathbf{o}_1) a_{q_1 q_2} b_{q_2}(\mathbf{o}_2) \cdots a_{q_{(T-1)} q_T} b_{q_T}(\mathbf{o}_T) \right\} \\ &= \max_i \delta_T(i)\,, && (2.9) \end{aligned}$$

όπου ορίζουμε την αναδρομική συνάρτηση

$$\delta_t(i) = \begin{cases} \pi_i b_i(\mathbf{o}_1) & , t = 1 \\ \max\limits_i \{\delta_{t-1}(i) a_{ij}\} b_j(\mathbf{o}_t) & , t = 2, 3, \cdots, T \end{cases}. \tag{2.10}$$

Το Viterbi score εύκολα φαίνεται πως δεν είναι μια ακριβής εκτίμηση της ακουστικής πιθανοφάνειας, καθώς αυτή θα προέκυπτε ως το άθροισμα όλων των πιθανοτήτων για τις

πιθανές αλληλουχίες μεταβάσεων που θα μπορούσε να ακολουθήσει το HMM. Οπότε, θα είχαμε

$$p(O|\lambda) = \sum_{Q} P(O, Q|\lambda)\,. \tag{2.11}$$

Ο υπολογισμός θα μπορούσε να γίνει και πάλι με τη βοήθεια δυναμικού προγραμματισμού, με την αντικατάσταση της συνάρτησης μεγίστου με άθροισμα στις σχέσεις (2.9), (2.10), δηλαδή βάσει του λεγόμενου αλγορίθμου Forward, ο οποίος όμως δε χρησιμοποιείται στην πράξη, εφόσον η απόδοση της αναγνώρισης δεν επηρεάζεται σημαντικά [13]. Περισσότερες λεπτομέρειες για τον αλγόριθμο Viterbi θα δοθούν στην Ενότητα 2.4.

Κάθε κατάσταση του HMM μπορούμε να θεωρήσουμε ότι αντιστοιχεί σε ένα φώνημα ή, γενικότερα, σε μια υπολεκτική μονάδα (Subword Unit - SU), όπως είναι τα phonelike units (PLUs), οι συλλαβές ή οι ημισυλλαβές. Τα PLUs βασίζονται στα φωνήματα, αλλά δεν ταυτίζονται με αυτά διότι μοντελοποιούνται βάσει της ακουστικής ομοιότητας και όχι της γλωσσολογικής [7]. Για προβλήματα με σχετικά μικρό λεξιλόγιο και απομονωμένες λέξεις, θα μπορούσαμε να χρησιμοποιήσουμε ως στοιχειώδη μονάδα ακόμα και την ίδια τη λέξη, αλλά σε συνεχή λόγο με μεγάλο λεξιλόγιο, κάτι τέτοιο δεν είναι εφικτό.

Στην πραγματικότητα, κάθε SU δεν παριστάνεται από μία μόνο κατάσταση ενός HMM, αλλά από ένα στοιχειώδες HMM με ένα σταθερό αριθμό καταστάσεων, της τάξης των 5-10 καταστάσεων [7]. Για παράδειγμα, θα μπορούσαμε να μοντελοποιήσουμε ένα φώνημα με ένα HMM 5 καταστάσεων, όπως αυτό που εικονίζεται στο Σχήμα 2.1iv. Τα self-loops στη συγκεκριμένη τοπολογία χρησιμεύουν ώστε να δίνεται η δυνατότητα σε κάθε φώνημα, κάθε SU και γενικότερα κάθε κατάσταση να έχει αυθαίρετα μεγάλη διάρκεια.

Λόγω του φαινομένου της συνάρθρωσης, οι ακουστικές ιδιότητες του κάθε φωνήματος επηρεάζονται από τα γειτονικά φωνήματα και δεν είναι ίδιες σε όλα τα ακουστικά περιβάλλοντα. Το φαινόμενο οφείλεται στη συνεχή φύση της ανθρώπινης ομιλίας και στην αναπόφευκτη αδράνεια των αρθρωτών (φωνητικές χορδές, γλώσσα, χείλη, κ.λπ.) που συμμετέχουν στη διαδικασία παραγωγής φωνής [40]. Για το λόγο αυτό, έχει καθιερωθεί να μη χρησιμοποιούνται φωνήματα (ή PLUs), αλλά SUs που λαμβάνουν υπόψιν τα προηγούμενα και τα επόμενα φωνήματα (ή PLUs), δηλαδή SUs εξαρτώμενα από τα συμφραζόμενα, όπως είναι τα τριφωνήματα (triphones) και σπανιότερα τα πενταφωνήματα (quinphones). Για παράδειγμα, η αγγλική φράση *get up* αποτελείται από τα φωνήματα /G/ /EH/ /T/ /AH/ /P/[2], αλλά από τα τριφωνήματα /(sil)G(EH)/ /(G)EH(T)/ /(EH)T(AH)/ /(T)AH(P)/ /(AH)P(sil)/.

Έχοντας, λοιπόν, επιλέξει τη στοιχειώδη υπολεκτική μονάδα που θα χρησιμοποιηθεί και αναπαριστώντας κάθε μία από τις μονάδες αυτές με ένα στοιχειώδες HMM, η αναπαράσταση μιας λέξης γίνεται με συνεχή concatenations τέτοιων στοιχειωδών HMMs, ενώ επιτρέποντας μεταβάσεις μεταξύ των HMMs που παριστάνουν λέξεις μπορεί να παρασταθεί μια πρόταση. Τη γνώση για το ποια HMMs πρέπει να συντεθούν για να παραχθεί μια λέξη την παρέχει το pronounciation lexicon, το οποίο παρέχει μία αντιστοιχία (όχι ένα προς ένα) ανάμεσα σε κάθε λέξη και την προφορά ή τις προφορές της.

2.2.2 Μοντέλα Μειγμάτων Γκαουσιανών

Όπως ήδη αναφέρθηκε, οι πιθανές παρατηρήσεις σε κάθε κατάσταση του HMM είναι διανύσματα που ανήκουν στον $\mathbb{R}^d$ και που δεν έχουν υποστεί κάποια διακριτοποίηση. Η πλέον συνήθης μοντελοποίηση της πιθανότητας μιας παρατήρησης σε μια συγκεκριμένη κατά-

[2]The CMU Pronouncing Dictionary http://www.speech.cs.cmu.edu/cgi-bin/cmudict

σταση είναι με χρήση Μοντέλων Μειγμάτων (d-διάστατων) Γκαουσιανών (Gaussian Mixture Models - GMMs) [41].

Ένα GMM δεν είναι παρά μια γραμμική υπέρθεση γκαουσιανών κατανομών, οπότε η πιθανότητα στην κατάσταση i να έχουμε μια παρατήρηση $\mathbf{x}$ μοντελοποιείται ως

$$b_i(\mathbf{x}) = \sum_{m=1}^{M_i} c_{im} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}) \,, \tag{2.12}$$

όπου M_i ο αριθμός των γκαουσιανών του μείγματος, ο οποίος μπορεί να διαφέρει για κάθε κατάσταση και οι συντελεστές c_{im} τηρούν τους απαραίτητους περιορισμούς ώστε ο συνδυασμός της σχέσης (2.12) να είναι κυρτός, δηλαδή

$$c_{im} \geq 0 \; \forall m \text{ και } \sum_{m=1}^{M_i} c_{im} = 1 \,. \tag{2.13}$$

Αξίζει να σημειωθεί πως πρακτικά όλες οι κατανομές πιθανότητας μπορούν να προσεγγιστούν με την επιθυμητή ακρίβεια από ένα GMM, δεδομένου αρκούντως μεγάλου αριθμού γκαουσιανών.

Συνήθως, για τους σκοπούς της Αυτόματης Αναγνώρισης Φωνής, επιλέγονται γκαουσιανές κατανομές με διαγώνιο πίνακα συμμεταβλητότητας. Παρόλο που οι πλήρεις πίνακες πετυχαίνουν ακριβέστερη μοντελοποίηση και οδηγούν δυνητικά σε καλύτερα αποτελέσματα αναγνώρισης, συνδέονται με δύο σημαντικά προβλήματα [33]. Πρώτον, χρειάζονται πολύ περισσότερο χρόνο και δεύτερον, απαιτούν πολύ μεγαλύτερο όγκο διαθέσιμων δεδομένων για την εκπαίδευσή τους. Με την υπόθεση της διαγώνιας μήτρας συμμεταβλητότητας, λοιπόν, κάθε γκαουσιανή κατανομή είναι ανεξάρτητη ως προς κάθε διάσταση και οπότε υπολογίζεται ως

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}) = \prod_{p=1}^{d} \frac{1}{\sqrt{2\pi\sigma_{imp}^2}} \exp\left\{ -\frac{(x_p - \mu_{imp})^2}{2\sigma_{imp}^2} \right\} \,, \tag{2.14}$$

όπου μ_{imp} και σ_{imp}^2 είναι αντίστοιχα η μέση τιμή και η διακύμανση της p-οστής διάστασης της m-οστής γκαουσιανής του μείγματος που αντιστοιχεί στην i-οστή κατάσταση του HMM.

2.2.3 Εκπαίδευση του Ακουστικού Μοντέλου

Ο συνδυασμός των HMMs και των GMMs παρέχει μία αποδοτική και κομψή στατιστική μοντελοποίηση για τους σκοπούς της αναγνώρισης, η οποία, όμως, στηρίζεται, όπως έχει γίνει εμφανές, σε μια σειρά παραμέτρων, οι οποίες απαιτούν κατάλληλη εκπαίδευση για τον υπολογισμό τους. Συγκεκριμένα, απαιτείται η εκπαίδευση των παραμέτρων a_{ij} και π, όπως εμφανίζονται στις σχέσεις (2.5), (2.7), καθώς και των παραμέτρων c_{im}, μ_{im}, Σ_{im} όπως εμφανίζονται στις σχέσεις (2.12) - (2.14). Μέσω της κατάλληλης εκπαίδευσης των GMMs, αυτόματα έχουμε στη διάθεσή μας το σύνολο πιθανοτήτων $B = \{b_j(k)\}$, αφού αυτό ακριβώς είναι που μοντελοποιούν. Ο αλγόριθμος που χρησιμοποιείται για τη ζητούμενη εκπαίδευση είναι ο αλγόριθμος Expectation - Maximization (EM).

Φυσικά, χρειαζόμαστε κάποιες αρχικές εκτιμήσεις και ο πιο απλός τρόπος να τις λάβουμε είναι με χρήση της τεχνικής που είναι γνωστή ως flat start [33]. Σύμφωνα με την τεχνική αυτή, θέτουμε ίσες με 0 τις πιθανότητες όσων μεταβάσεων θέλουμε εκ κατασκευής να είναι "ανενεργές" στο HMM, κάτι που μας εξασφαλίζει ότι θα παραμείνουν 0 και μετά το τέλος

της εκπαίδευσης, όπως θα φανεί στη συνέχεια. Όλες οι υπόλοιπες μεταβάσεις θεωρούνται αρχικά ισοπίθανες. Για παράδειγμα, για ένα μοντέλο 5 καταστάσεων, όπως του Σχήματος 2.1iv προκύπτει ο αρχικός πίνακας μεταβάσεων

$$A = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

Προφανώς, για μια τέτοια τοπολογία, θα έχουμε, επίσης,

$$\pi_i = \begin{cases} 1 & , i = 1 \\ 0 & , i \neq 1 \end{cases} .$$

Όσον αφορά τις απαραίτητες αρχικοποιήσεις για τα GMMs, όλες οι μέσες τιμές και οι μεταβλητότητες αρχικοποιούνται με τις αντίστοιχες δειγματικές μέσες τιμές και μεταβλητότητες, όπως μετρώνται στο σύνολο των δεδομένων εκπαίδευσης.

Στην περίπτωση των HMMs, ο αλγόριθμος EM είναι γνωστός με την ονομασία forward-backward αλγόριθμος [7] και βασίζεται στην πιθανότητα forward, όπως αυτή έχει αναφερθεί στην Υποενότητα 2.2.1, και την πιθανότητα backward. Το πρόβλημα κατά την εκπαίδευση ενός HMM είναι, δοθείσης μιας ακολουθίας παρατηρήσεων O και του συνόλου των πιθανών καταστάσεων του HMM, να προσαρμοστούν οι άγνωστες παράμετροι λ του μοντέλου ώστε να μεγιστοποιείται η πιθανότητα $P(O|\lambda)$.

Αναλυτικά, η πιθανότητα forward ορίζεται ως η πιθανότητα ένα HMM με παραμέτρους λ να παράξει μέχρι τη χρονική στιγμή t μια ακολουθία παρατηρήσεων $\mathbf{o}_1\mathbf{o}_2\cdots\mathbf{o}_t$ και να βρίσκεται στην κατάσταση i:

$$\alpha_t(i) = P(\mathbf{o}_1\mathbf{o}_2\cdots\mathbf{o}_t, q_t = i|\lambda) \tag{2.15}$$

$$= \begin{cases} \pi_i b_i(\mathbf{o}_1) & , t = 1 \\ \sum_i \{\alpha_{t-1}(i)a_{ij}\} b_j(\mathbf{o}_t) & , t = 2, 3, \cdots, T \end{cases} . \tag{2.16}$$

Όπως έχουμε δει, η τελευταία σχέση μπορεί να χρησιμοποιηθεί για τον ακριβή υπολογισμό της ακουστικής πιθανοφάνειας $p(O|\lambda)$ ενός HMM, αφού

$$\begin{aligned} p(O|\lambda) &= \sum_Q P(O, Q|\lambda) \\ &= \sum_{q_1, q_2, \cdots, q_T} \left\{ \pi_{q_1} b_{q_1}(\mathbf{o}_1) a_{q_1 q_2} b_{q_2}(\mathbf{o}_2) \cdots a_{q_{(T-1)} q_T} b_{q_T}(\mathbf{o}_T) \right\} \\ &= \sum_i \alpha_T(i) . \end{aligned} \tag{2.17}$$

Όμοια, η πιθανότητα backward ορίζεται ως η πιθανότητα ένα HMM με παραμέτρους λ που τη χρονική στιγμή t βρίσκεται στην κατάσταση i να παράξει μια ακολουθία παρατηρήσεων $\mathbf{o}_{t+1}\mathbf{o}_{t+2}\cdots\mathbf{o}_T$:

$$\beta_t(i) = P(\mathbf{o}_{t+1}\mathbf{o}_{t+2}\cdots\mathbf{o}_T|q_t = i, \lambda) \tag{2.18}$$

$$= \begin{cases} 1 & , t = T \\ \sum_j \{a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j)\} & , t = T-1, T-2, \cdots, 1 \end{cases} . \tag{2.19}$$

Ορίζουμε, τώρα, ως $\xi_t(i,j)$ την πιθανότητα το HMM να βρίσκεται στη θέση i τη χρονική στιγμή t και στη θέση j τη χρονική στιγμή $t+1$, και ως $\gamma_t(i)$ την πιθανότητα το HMM να βρίσκεται στη θέση i τη χρονική στιγμή t. Αποδεικνύεται [7] ότι

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(\mathbf{o}_{t+1})\beta_{t+1}(j)}{\sum_i \sum_j \alpha_t(i)a_{ij}b_j(\mathbf{o}_{t+1})\beta_{t+1}(j)}\,, \tag{2.20}$$

$$\gamma_t(i) = \sum_j \xi_t(i,j) \tag{2.21}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_i \alpha_t(i)\beta_t(i)}\,. \tag{2.22}$$

Οι εξισώσεις (2.20), (2.21) αποτελούν το βήμα του expectation στον αλγόριθμο EM. Βάσει αυτών, ακολουθεί το βήμα του maximization, όπου ουσιαστικά υπολογίζονται οι παράμετροι του μοντέλου σύμφωνα με το κριτήριο της μέγιστης πιθανοφάνειας (Maximum Likelihood - ML):

$$\begin{aligned}\hat{\pi}_i = &\text{ αναμενόμενος αριθμός φορών που το HMM βρίσκεται} \\ &\text{στην κατάσταση } i \text{ τη χρονική στιγμή } t=1 \\ = &\,\gamma_1(i)\,,\end{aligned} \tag{2.23}$$

$$\begin{aligned}\hat{a}_{ij} &= \frac{\text{αναμενόμενος αριθμός μεταβάσεων από την κατάσταση } i \text{ στην κατάσταση } j}{\text{αναμενόμενος αριθμός μεταβάσεων από την κατάσταση } i} \\ &= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}\,.\end{aligned} \tag{2.24}$$

Για λόγους πληρότητας, αλλά και για να είναι πιο εύκολη η γενίκευση στη συνεχή περίπτωση, παραθέτουμε και τη σχέση που δίνει τις παραμέτρους $b_j(k)$ στην περίπτωση διακριτών κατανομών πιθανοτήτων:

$$\begin{aligned}\hat{b}_j(k) &= \frac{\text{αναμενόμενος αριθμός φορών στη θέση } j \text{ με το σύμβολο } \mathbf{v}_k \text{ να παρατηρείται}}{\text{αναμενόμενος αριθμός φορών στη θέση } j} \\ &= \frac{\sum_{t=1 \text{ τ.ώ. } \mathbf{o}_t=\mathbf{v}_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}\,.\end{aligned} \tag{2.25}$$

Για τα GMMs, που χρησιμοποιούνται στην πράξη, γενικεύουμε τη σχέση (2.22) ώστε να εκφράζει την πιθανότητα το HMM να βρίσκεται στη θέση i τη χρονική στιγμή t με τη

m-οστή γκαουσιανή να δίνει την παρατήρηση o_t:

$$\gamma_t(i,m) = \frac{\alpha_t(i)\beta_t(i)}{\sum_i \alpha_t(i)\beta_t(i)} \cdot \frac{c_{im}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im})}{\sum_{m=1}^{M_i} c_{im}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im})} \,. \tag{2.26}$$

Σημειώνουμε ότι η σχέση (2.22) δεν ισχύει μόνο στην περίπτωση διακριτών κατανομών, αλλά και στην περίπτωση συνεχών, όταν επιλέξουμε να μοντελοποιήσουμε την πιθανότητα με μία μόνο γκαουσιανή.

Οι ανανεώσεις των παραμέτρων (βήματα maximization), τώρα, δίνονται ως εξής:

$$\begin{aligned}\hat{c}_{im} &= \frac{\text{αναμενόμενος αριθμός φορών στην κατάσταση } j \text{ με την } m\text{-οστή γκαουσιανή}}{\text{αναμενόμενος αριθμός φορών στην κατάσταση } j} \\ &= \frac{\sum_{t=1}^{T} \gamma_t(i,m)}{\sum_{t=1}^{T}\sum_{m=1}^{M_i} \gamma_t(i,m)} \triangleq \frac{N_{i,m}}{\sum_{m=1}^{M_i} N_{i,m}} \,.\end{aligned} \tag{2.27}$$

Κάνοντας χρήση και γενικεύοντας το κριτήριο μέγιστης πιθανοφάνειας για μία γκαουσιανή [41], προκύπτει ακόμα

$$\hat{\boldsymbol{\mu}}_{im} = \frac{1}{N_{i,m}} \sum_{t=1}^{T} \gamma_t(i,m)\mathbf{o}_t = \frac{\sum_{t=1}^{T} \gamma_t(i,m)\mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(i,m)} \,, \tag{2.28}$$

$$\hat{\boldsymbol{\Sigma}}_{im} = \frac{1}{N_{i,m}} \sum_{t=1}^{T} \gamma_t(i,m)(\mathbf{o}_t - \boldsymbol{\mu}_{im})(\mathbf{o}_t - \boldsymbol{\mu}_{im})^T = \frac{\sum_{t=1}^{T} \gamma_t(i,m)(\mathbf{o}_t - \boldsymbol{\mu}_{im})(\mathbf{o}_t - \boldsymbol{\mu}_{im})^T}{\sum_{t=1}^{T} \gamma_t(i,m)} \,. \tag{2.29}$$

Ο αλγόριθμος EM είναι επαναληπτικός με τα βήματα του expectation και του maximization να διαδέχονται συνεχώς το ένα το άλλο. Σε κάθε βήμα που ανανεώνονται οι παράμετροι του μοντέλου από λ_{old} σε λ_{new} αποδεικνύεται [41] ότι

$$P(O|\lambda_{new}) \geq P(O|\lambda_{old}) \,, \tag{2.30}$$

οπότε είναι εγγυημένο ότι ο αλγόριθμος θα συγκλίνει σε κάποιο τοπικό (όχι απαραίτητα ολικό) μέγιστο, μετά από αρκετές επαναλήψεις.

2.2.4 Δεμένες Καταστάσεις και Δέντρα Απόφασης

Η χρήση triphones ως SUs, παρά τα ευεργετικά αποτελέσματα που επιφέρει στην τελική απόδοση της αναγνώρισης, εμπεριέχει ένα βασικό "μειονέκτημα": τα triphones είναι πάρα

πολλά. Εάν υποθέσουμε μια γλώσσα με 40 φωνήματα, μια καθόλα ρεαλιστική προσέγγιση, προκύπτουν $40^3 = 64000$ triphones. Έτσι, πέραν της πολυπλοκότητας που εισάγουν, ελοχεύει ο κίνδυνος κάποια triphones να συναντώνται τόσο σπάνια που να μην εμφανίζονται αρκετές φορές στο σύνολο εκπαίδευσης, ώστε να γίνει μία αξιόλογη εκτίμηση των παραμέτρων που σχετίζονται με τα αντίστοιχα HMMs. Μπορεί ακόμα και να μην εμφανιστούν καμία φορά στο σύνολο εκπαίδευσης, αλλά να εμφανιστούν στο σύνολο ελέγχου, δηλαδή στις ακολουθίες λέξεων προς αναγνώριση.

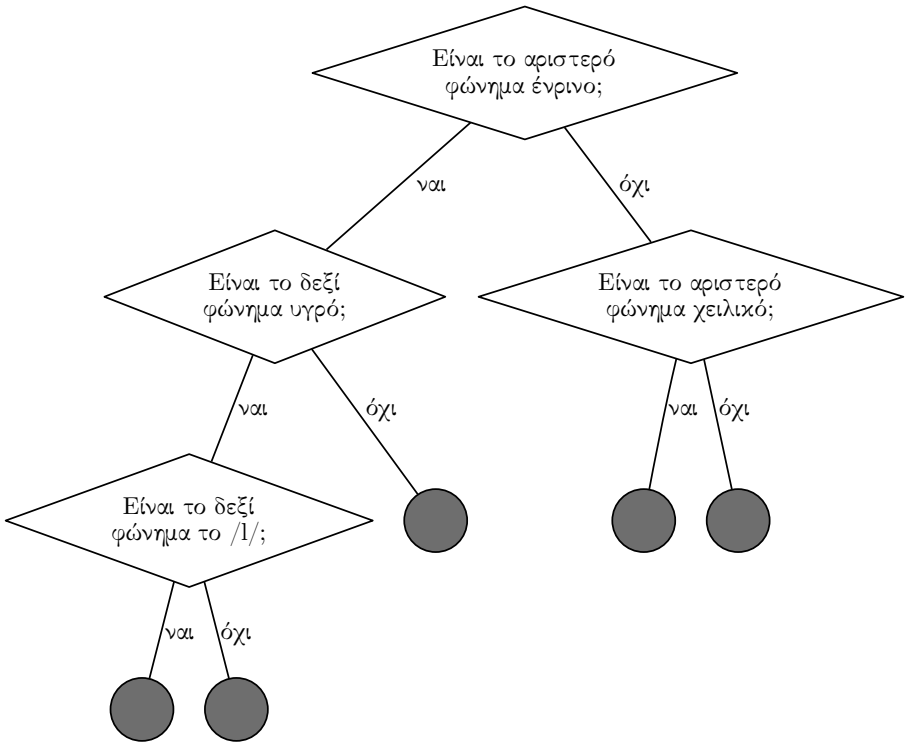
Λόγω της έλλειψης των τεραστίων ποσοτήτων δεδομένων εκπαίδευσης που θα απαιτούνταν, στην πράξη τα συστήματα αναγνώρισης χρησιμοποιούν την ιδέα του διαμοιρασμού παραμέτρων (parameter sharing) μεταξύ HMMs και καταστάσεων [42]. Μία πρώτη ιδέα είναι η χρήση ενός κοινού συνόλου συγκεκριμένου αριθμού γκαουσιανών κατανομών, οι οποίες χρησιμοποιούνται από όλες τις καταστάσεις στα HMMs και συνδυάζονται με κατάλληλους συντελεστές βαρύτητας, ώστε κάθε κατάσταση να μοντελοποιείται από ένα κατάλληλο GMM [43]. Ο διαμοιρασμός παραμέτρων μπορεί θεωρητικά να συμβαίνει σε οποιοδήποτε επίπεδο. Για παράδειγμα, μπορεί να συμβαίνει μεταξύ καταστάσεων ίδιου SU (συνήθως τριφωνήματος) ή διαφορετικών ή μπορεί να διαμοιράζεται όλο το state (δηλαδή το GMM), μόνο οι μεμονωμένες γκαουσιανές ή και μόνο συγκεκριμένες παράμετροι αυτών όπως π.χ. οι διακυμάνσεις, ενώ οι μέσες τιμές διαφέρουν. Στη συνέχεια, θα αναλυθεί μία τεχνική που εφαρμόζεται συνήθως στα σύγχρονα συστήματα και εμφανίζει αρκετά πλεονεκτήματα, τα HMMs με δεμένες καταστάσεις (tied-state HMMs), κάνοντας χρήση δέντρων απόφασης (decision trees) [44].

Σκοπός ενός tied-state HMM είναι να διασφαλίσει ότι υπάρχουν αρκετά δεδομένα εκπαίδευσης ώστε να εκτιμηθούν ορθά όλες οι απαραίτητες παράμετροι, ενώ παράλληλα διατηρούνται μεταξύ ίδιων φωνημάτων όλες οι σημαντικές διαφορές που σχετίζονται με τα συμφραζόμενα (context). Μάλιστα, με τη μέθοδο που κατασκευάζεται, δίνεται η δυνατότητα εκτίμησης παραμέτρων για HMMs τριφωνημάτων που δεν έχουν συναντηθεί καθόλου κατά την εκπαίδευση.

Για να περιγράψουμε τη διαδικασία, γίνεται αρχικά η υπόθεση ότι όλα τα τριφωνήματα μοντελοποιούνται από HMMs με τον ίδιο αριθμό καταστάσεων N. Για καθεμία από τις καταστάσεις αυτές και για κάθε φώνημα δημιουργείται ένα δέντρο απόφασης· οπότε, θα δημιουργηθούν συνολικά $N \cdot P$ δέντρα απόφασης, όπου P ο αριθμός φωνημάτων της γλώσσας. Όλα τα τριφωνήματα του ίδιου κεντρικού φωνήματος μοιράζονται τον ίδιο πίνακα μετάβασης A, οπότε τα δέντρα απόφασης χρησιμοποιούνται για την ομαδοποίηση των κατανομών πιθανότητας των παρατηρήσεων των καταστάσεων των HMMs.

Αρχικά, όλες οι καταστάσεις που πρόκειται να ομαδοποιηθούν σε ένα δέντρο απόφασης, για παράδειγμα η δεύτερη κατάσταση των HMMs όλων των τριφωνημάτων που αντιστοιχούν στο κεντρικό φώνημα x, τοποθετούνται στη ρίζα του δέντρου και θεωρείται ότι είναι tied, δηλαδή μοιράζονται τις ίδιες παραμέτρους, ενώ η αντίστοιχη κατανομή πιθανότητας μοντελοποιείται από μία μόνο γκαουσιανή. Ας είναι S το σύνολο αυτό των καταστάσεων και F όλα τα αντίστοιχα frames εκπαίδευσης. Εν συνεχεία, μετράται η λογαριθμική πιθανοφάνεια (log likelihood) $L(S)$ να παραχθούν τα F βάσει των S και επιλέγεται η ερώτηση εκείνη που θα διασπάσει τη ρίζα σε δύο κόμβους ώστε να μεγιστοποιηθεί η αύξηση της λογαριθμικής πιθανοφάνειας. Η διαδικασία επαναλαμβάνεται επαναληπτικά μέχρι η λογαριθμική πιθανοφάνεια να φτάσει σε ένα ελάχιστο κατώφλι, ενώ παράλληλα δεν έχουν δημιουργηθεί πάρα πολλά φύλλα στο δέντρο, ώστε να εξασφαλίζεται ότι όλα συνδέονται με επαρκή αριθμό δεδομένων εκπαίδευσης. Όλες οι ερωτήσεις είναι της μορφής "*Είναι το αριστερό ή το δεξί φώνημα μέλος του συνόλου X;*". Ένα παράδειγμα τέτοιου δέντρου απόφασης παρουσιάζεται στο Σχήμα 2.2.

Το τελευταίο στάδιο είναι η διάσπαση των γκαουσιανών κατανομών. Όπως αναφέρθηκε,



Σχήμα 2.2: Παράδειγμα δέντρου απόφασης για ομαδοποίηση καταστάσεων τριφωνημάτων. Η είσοδος στο δέντρο είναι η x-οστή κατάσταση των HMMs όλων των τριφωνημάτων που έχουν το ίδιο κεντρικό φώνημα. Εδώ δημιουργούνται πέντε ομάδες καταστάσεων τριφωνημάτων, όπου οι καταστάσεις κάθε ομάδας διαμοιράζονται τις ίδιες παραμέτρους, δηλαδή είναι δεμένες (tied).

η μοντελοποίηση ξεκινάει θεωρώντας μία γκαουσιανή ανά κατάσταση. Αφού έχουν δημιουργηθεί τα δέντρα απόφασης, οι παράμετροι των κατανομών πιθανότητας επαναληπτικά επανεκπαιδεύονται με συνεχείς διασπάσεις μίας ή περισσότερων γκαουσιανών κατανομών μέχρι να φτάσουμε στον επιθυμητό αριθμό γκαουσιανών από τις οποίες θα αποτελούνται τα GMMs.

Ο συνολικός αριθμός καταστάσεων, δηλαδή ο συνολικός αριθμός όλων των φύλλων όλων των δέντρων απόφασης, είναι συνήθως προκαθορισμένος και κυμαίνεται από κάποιες χιλιάδες μέχρι λίγες δεκάδες χιλιάδες καταστάσεις, με κάθε κατάσταση να αντιστοιχεί σε ένα GMM με 16 έως 64 γκαουσιανές. Τα νούμερα αυτά είναι εμπειρικά και εξαρτώνται εν πολλοίς από τον όγκο των διαθέσιμων δεδομένων εκπαίδευσης.

Σύμφωνα με την τεχνική που περιγράφηκε, ουσιαστικά ποτέ δεν εκπαιδεύονται, αλλά ούτε και αποθηκεύονται, τριφωνικά μοντέλα. Όταν κατά την αποκωδικοποίηση χρειαστεί κάποιο τριφωνικό μοντέλο, το αντίστοιχο HMM συντίθεται βάσει των κατάλληλων δέντρων απόφασης. Με τον τρόπο αυτό, μπορούν να συντεθούν τα απαραίτητα μοντέλα για όλα τα τριφωνήματα, ακόμα και για όσα δε συναντήθηκαν καθόλου κατά την εκπαίδευση.

2.2.5 Εξαναγκασμένη Ευθυγράμμιση

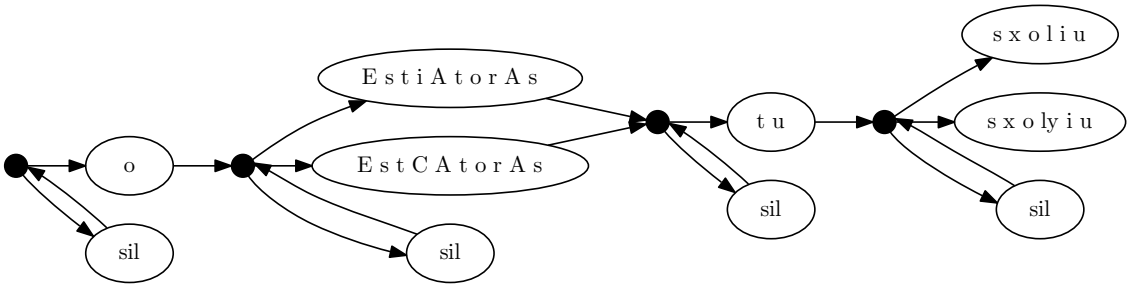
Η εκπαίδευση του ακουστικού μοντέλου γίνεται βάσει των διαθέσιμων για εκπαίδευση ηχογραφήσεων και των αντίστοιχων απομαγνητοφωνήσεων. Ωστόσο, στον προφορικό λόγο ενυπάρχουν κάποια στοιχεία, τα οποία στη γενική περίπτωση δεν είναι διαθέσιμα στο γραπτό κείμενο των απομαγνητοφωνήσεων. Τα κυριότερα από αυτά είναι οι πιθανές παύσεις μεταξύ

των λέξεων και η διαφορετική προφορά με την οποία εκφέρεται κάθε λέξη αναλόγως του ομιλητή.

Για την αντιμετώπιση των προβλημάτων αυτών και επειδή συνήθως δεν υπάρχουν διαθέσιμες ακριβείς απομαγνητοφωνήσεις σε επίπεδο φωνημάτων, παρά μόνο σε επίπεδο λέξεων, με εξαίρεση κάποιες λίγες βάσεις δεδομένων, είναι απαραίτητο ένα βήμα αυτόματης φωνητικής ευθυγράμμισης (phonetic alignment). Παρόλο που αυτό αποτελεί από μόνο του πεδίο ερευνητικής μελέτης και διαφορετικές μέθοδοι έχουν προταθεί [45], συνήθως, για τους σκοπούς της Αναγνώρισης Φωνής, χρησιμοποιείται η τεχνική της εξαναγκασμένης ευθυγράμμισης (forced alignment), μέσω εκπαίδευσης με τον αλγόριθμο Viterbi.

Σύμφωνα με τη μέθοδο αυτή, θεωρείται κατ' αρχήν ότι όλες οι πιθανές διαφορετικές προφορές μιας λέξης παρέχονται από το φωνητικό λεξικό, καθώς επίσης ότι περίοδοι σιωπής μπορεί να υπάρχουν προαιρετικά μετά από κάθε λέξη, ώστε να υπάρχει η δυνατότητα παύσης μεταξύ των διαδοχικά εκφερομένων λέξεων [33]. Αξίζει να σημειωθεί πως συνήθως τα "φωνήματα" σιωπής μοντελοποιούνται με μια πιο πολύπλοκη τοπολογία από τα υπόλοιπα φωνήματα, όπως για παράδειγμα με ένα left-right HMM πέντε καταστάσεων, όπου επιτρέπονται και μεταβάσεις μεταξύ απομακρυσμένων καταστάσεων και όχι μόνο μεταξύ διαδοχικών[3], όπως στο Σχήμα 2.1ii. Ακόμα, οι φασματικές ιδιότητες της σιωπής δεν επηρεάζονται από γειτονικά φωνήματα, οπότε εκπαιδεύονται μονοφωνικά μοντέλα (ανεξάρτητα από τα συμφραζόμενα).

Βάσει αυτών των υποθέσεων, δημιουργούνται για κάθε πρόταση HMMs με παράλληλες διαδρομές για κάθε πιθανή προφορά μιας λέξης και με όλα τα επιπρόσθετα "φωνήματα" σιωπής, όπως στο Σχήμα 2.3. Πάνω σε αυτά τα HMMs μπορεί να εφαρμοστεί ο αλγόριθμος Viterbi. Εφόσον ο Viterbi μπορεί να βρει στο HMM όπου εφαρμόζεται την πιο πιθανή ακολουθία καταστάσεων, από την ακολουθία αυτή μπορούν να εκτιμηθούν τα χρονικά όρια μεταξύ των λέξεων, αλλά και να βρεθεί ποια προφορά των λέξεων χρησιμοποιείται. Έτσι, γίνεται μία αυτόματη κατάτμηση και ευθυγράμμιση του λόγου σε καταστάσεις. Ο όρος "εξαναγκασμένη" ευθυγράμμιση προέρχεται από το γεγονός ότι ο αλγόριθμος Viterbi είναι αναγκασμένος να βρει τη βέλτιστη ευθυγράμμιση υπό αυστηρούς προκαθορισμένους περιορισμούς που του θέτει η δομή του HMM.



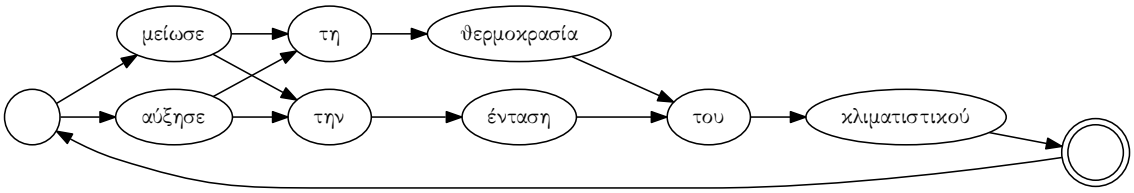
Σχήμα 2.3: Απλοποιημένο HMM για μοντελοποίηση της φράσης *ο εστιάτορας του σχολείου*. Φαίνεται πως υπάρχουν δύο εναλλακτικές προφορές για δύο λέξεις της φράσης, ενώ μπορεί να υπάρχουν παύσεις μεταξύ των λέξεων. Το HMM συντίθεται σε επίπεδο καταστάσεων και όχι λέξεων, οπότε κάθε έλλειψη εδώ αντιστοιχεί σε ένα επιμέρους HMM.

[3]Πρόκειται για τη σύμβαση που χρησιμοποιεί το σύστημα Kaldi, το οποίο θα χρησιμοποιηθεί κατά κόρον στην παρούσα εργασία.

2.3 Γλωσσικό Μοντέλο

Η απόδοση ενός συστήματος αναγνώρισης φωνής αυξάνεται κατά πολύ όταν χρησιμοποιείται ένα αξιόπιστο γλωσσικό μοντέλο. Με αυτό, μη γραμματικές και μη πιθανές προτάσεις μπορούν να απορριφθούν από τη διαδικασία της αναγνώρισης, με αποτέλεσμα τα σφάλματα να μειωθούν [13].

Τα μοντέλα που χρησιμοποιούνται κατά κόρον για LVCSR (Large - Vocabulary Continuous - Speech Recognition) είναι τα N-grams, τα οποία θα αναλυθούν στη συνέχεια. Για εφαρμογές όπου ξέρουμε εκ των προτέρων ένα περιορισμένο σύνολο φράσεων που μπορούν να ειπωθούν, όπως είναι η αναγνώριση φωνητικών εντολών [46], μπορεί να γίνει χρήση Γραμματικών Πεπερασμένης Κατάστασης (Finite State Grammars - FSGs), οι οποίες ορίζονται εξ αρχής από το σχεδιαστή του συστήματος. Ένα παράδειγμα FSG παρουσιάζεται στο Σχήμα 2.4.



Σχήμα 2.4: Παράδειγμα Γραμματικής Πεπερασμένης Κατάστασης.

2.3.1 Στατιστική Μοντελοποίηση Γλώσσας με N-gram Μοντέλα

Με τον όρο "n-gram" χαρακτηρίζουμε μια υποακολουθία από n ομοειδή αντικείμενα σε μια ακολουθία. Εδώ θα αναφερόμαστε σε n-grams λέξεων, αλλά αναλόγως της εφαρμογής στην Αυτόματη Αναγνώριση Φωνής, μπορεί να έχουμε n-grams γραμμάτων, φωνημάτων, κ.ά. Μία από τις πρώτες αναφορές σε n-grams μπορεί να βρεθεί στην εργασία του Shannon πάνω στη Θεωρία της Πληροφορίας [32], ο οποίος υποστήριξε ότι ένα συνθετικό κείμενο που παράγεται με βάση n-gram μοντελοποίηση γίνεται όλο και πιο κατανοητό και μοιάζει με φυσικό κείμενο όσο η τάξη του μοντέλου, δηλαδή το n, αυξάνεται.

Συμβολίζοντας μία ακολουθία από M λέξεις ως $W = w_1, w_2, \cdots, w_M \triangleq w_1^M$, ξέρουμε ότι ισχύει

$$P(W) = P(w_1)P(w_2|w_1)\cdots P\left(w_M|w_1^{M-1}\right) = P(w_1)\prod_{m=2}^{M} P\left(w_m|w_1^{m-1}\right) . \qquad (2.31)$$

Σύμφωνα με ένα n-gram μοντέλο, αντί να υπολογίζεται η πιθανότητα μιας λέξης δοθείσης όλης της προϊστορίας της, γίνεται μια προσέγγιση της προϊστορίας της λέξης μόνο από τις n προηγούμενες λέξης της:

$$P(W) \approx \prod_{m=1}^{M} P\left(w_m|w_{m-n+1}^{m-1}\right) = \prod_{m=1}^{M} P\left(w_m|w_{m-n+1}, w_{m-n+2}, \cdots, w_{m-1}\right) . \qquad (2.32)$$

Το μοντέλο αυτό είναι ισοδύναμο με ένα Μαρκοβιανό μοντέλο τάξης $n-1$, αφού κάθε κατάσταση εξαρτάται από την ιστορία μέχρι και $n-1$ "χρονικά επίπεδα" πίσω.

Στο συμβολισμό της σχέσης (2.32), για να έχουν νόημα αρνητικοί δείκτες, δηλαδή να μοντελοποιείται το περιβάλλον των πρώτων λέξεων, εισάγονται ειδικοί χαρακτήρες (<s>) στην

αρχή κάθε πρότασης του διαθέσιμου corpus. Για παράδειγμα, για $m = 1$, $n = 2$ (bigram), θα είναι $P(w_1|w_0) = P(w_1|<\text{s}>)$. Ακόμα, απαραίτητη είναι η εισαγωγή ειδικών χαρακτήρων στο τέλος κάθε πρότασης ($<\backslash\text{s}>$), ώστε οι πιθανότητες όλων των δυνατών ακολουθιών να αθροίζουν στη μονάδα. Σε διαφορετική περίπτωση, οι πιθανότητες όλων των ακολουθιών δεδομένου μήκους θα άθροιζαν στη μονάδα, οπότε οι πιθανότητες όλων των δυνατών ακολουθιών θα άθροιζαν στο άπειρο [47]. Για παράδειγμα, εάν το διαθέσιμο corpus περιελάμβανε την πρόταση *Εγώ παίζω*, για την εκπαίδευση ενός bigram μοντέλου, θα έπρεπε εκτιμηθούν οι πιθανότητες $P(\text{Εγώ}|<\text{s}>)$, $P(\text{παίζω}|\text{Εγώ})$ και $P(<\backslash\text{s}>|\text{παίζω})$.

Η εκπαίδευση ενός n-gram μοντέλου γίνεται βάσει ενός training corpus, δηλαδή ουσιαστικά ενός κειμένου που έχουμε στη διάθεση μας. Στα πλαίσια της Αυτόματης Αναγνώρισης Φωνής, το corpus αυτό δεν είναι άλλο παρά οι απομαγνητοφωνήσεις (transcriptions) των δεδομένων ομιλίας. Οι επιμέρους πιθανότητες μπορούν απλά να υπολογιστούν ως οι εκτιμήσεις μέγιστης πιθανοφάνειας (Maximum Likelihood Estimation - MLE) ως εξής:

$$P\left(w_m|w_{m-n+1}^{m-1}\right) = \frac{C\left(w_{m-n+1}^{n}\right)}{C\left(w_{m-n+1}^{n-1}\right)}, \tag{2.33}$$

όπου $C(s)$ ο αριθμός των εμφανίσεων που η ακολουθία s εμφανίζεται στο corpus.

Όσο αυξάνεται η τάξη του μοντέλου, τόσο καλύτερα αναμένονται τα αποτελέσματα. Ωστόσο, από ένα σημείο και μετά, η βελτίωση των αποτελεσμάτων δεν είναι τέτοια που να αντισταθμίζει τα τεράστια datasets, αλλά και τους χρόνους που απαιτούνται για την εκπαίδευση. Ακόμα, ένα πολύ "ισχυρό" γλωσσικό μοντέλο ελοχεύει τον κίνδυνο να γίνει πολύ biased στο σύνολο εκπαίδευσης και να μη γενικεύει καλά στο σύνολο ελέγχου. Τα πλέον συνηθισμένα μοντέλα είναι τα 3-grams, με τα μεγάλα σύγχρονα συστήματα να χρησιμοποιούν μέχρι και 5-grams.

Η διαισθητική βελτίωση του συστήματος αναγνώρισης με αύξηση της τάξης του μοντέλου μπορεί να διατυπωθεί φορμαλιστικά με τη βοήθεια της έννοιας του perplexity [33]. Το perplexity (PP) μιας ακολουθίας W με M λέξεις ορίζεται ως

$$PP(W) = P\left(W\right)^{-M} = \sqrt[M]{\frac{1}{P(w_1, w_2, \cdots, w_M)}}. \tag{2.34}$$

Συνδυάζοντας τις σχέσεις (2.34) και (2.32), προκύπτει ότι το perplexity όταν γίνεται χρήση n-gram μοντέλου είναι

$$PP(W) = \sqrt[M]{\prod_{m=1}^{M} \frac{1}{P\left(w_m|w_{m-n+1}^{m-1}\right)}}. \tag{2.35}$$

Όπως φαίνεται, η ελαχιστοποίηση του perplexity είναι ισοδύναμη με τη μεγιστοποίηση της πιθανότητας της ακολουθίας W σύμφωνα με το εν χρήσει μοντέλο.

Αξίζει στο σημείο αυτό να αναφερθεί ένα πρακτικό ζήτημα που σχετίζεται με τα n-gram μοντέλα. Πρόκειται για τη διάκριση μεταξύ κλειστού και ανοιχτού λεξιλογίου, όπου στην τελευταία περίπτωση, μπορεί στο σύνολο εκπαίδευσης να περιλαμβάνονται λέξεις οι οποίες δεν υπάρχουν στο υπό χρήση λεξιλόγιο (για παράδειγμα, δεν υπάρχουν στο φωνητικό λεξικό, ώστε να γνωρίζουμε πώς μπορούν να μετατραπούν σε ακολουθία φωνημάτων). Τότε, προηγείται μία φάση κανονικοποίησης του κειμένου, όπου όλες αυτές οι λέξεις, γνωστές ως λέξεις OOV (Out-Of-Vocabulary), αντικαθίστανται από μία κοινή, προκαθορισμένη "λέξη", η οποία συνηθίζεται να είναι η <UNK>. Για τους σκοπούς της n-gram μοντελοποίησης, η "λέξη" αυτή αντιμετωπίζεται όπως και κάθε άλλη λέξη του συνόλου εκπαίδευσης.

2.3.2 Smoothing με Χρήση της Μεθόδου Witten-Bell

Το εμφανές πρόβλημα στη χρήση MLE σύμφωνα με τη σχέση (2.33) είναι ότι ορισμένα n-grams μπορεί να μην εμφανίζονται καθόλου ή πολύ σπάνια στο corpus που χρησιμοποιείται για εκπαίδευση, χωρίς αυτό απαραίτητα να αντικατοπτρίζει μια πραγματική αδυναμία εμφάνισής του σε πραγματικές συνθήκες. Συνεπώς, δε θα πρέπει να του δώσουμε μηδενική πιθανότητα εμφάνισης. Η διαδικασία κατά την οποία μία μάζα πιθανότητας αφαιρείται από τις ακολουθίες με πολύ συχνές εμφανίσεις (discounting) και επιμερίζεται στις ακολουθίες με μηδενικές ή λίγες εμφανίσεις (backoff) ονομάζεται smoothing. Στη συνέχεια, θα αναλυθεί μία από τις πολλές μεθόδους που έχουν προταθεί για discounting, που είναι γνωστή ως μέθοδος Witten-Bell [48, 47].

Ο ορισμός της νέας πιθανότητας του εκάστοτε n-gram γίνεται αναδρομικά:

$$P^*\left(w_m|w_{m-n+1}^{m-1}\right) = \lambda_{w_{m-n+1}^{m-1}} P\left(w_m|w_{m-n+1}^{m-1}\right) + \left(1-\lambda_{w_{m-n+1}^{m-1}}\right) P^*\left(w_m|w_{m-n+2}^{m-1}\right) , \quad (2.36)$$

όπου το $P\left(w_m|w_{m-n+1}^{m-1}\right)$ υπολογίζεται βάσει της σχέσης (2.33). Για τον υπολογισμό των παραμέτρων $\lambda_{w_{m-n+1}^{m-1}}$ ορίζουμε τη βοηθητική συνάρτηση $N_{1+}\left(w_{m-n+1}^{m-1}\bullet\right)$ ως τον αριθμό των διαφορετικών n-grams που εμφανίζονται στο σύνολο εκπαίδευσης και ξεκινούν με το $(n-1)$-gram w_{m-n+1}^{m-1}:

$$N_{1+}\left(w_{m-n+1}^{m-1}\bullet\right) = \left|\left\{w_m : C\left(w_{m-n+1}^{m-1}w_m\right) > 0\right\}\right| , \quad (2.37)$$

όπου η συνάρτηση $|\cdot|$ δηλώνει τον πληθάριθμο ενός συνόλου. Θέλουμε, τώρα, να ικανοποιείται η σχέση

$$1-\lambda_{w_{m-n+1}^{m-1}} = \frac{N_{1+}\left(w_{m-n+1}^{m-1}\bullet\right)}{N_{1+}\left(w_{m-n+1}^{m-1}\bullet\right) + \sum_{w_m} C\left(w_{m-n+1}^{m}\right)} . \quad (2.38)$$

Χρησιμοποιώντας τις σχέσεις (2.36), (2.33) και (2.38), καταλήγουμε στην τελική σχέση υπολογισμού των πιθανοτήτων εμφάνισης των n-grams, σύμφωνα με τη μέθοδο Witten-Bell:

$$P^*\left(w_m|w_{m-n+1}^{m-1}\right) = \frac{C\left(w_{m-n+1}^{m}\right) + N_{1+}\left(w_{m-n+1}^{m-1}\bullet\right) P^*\left(w_m|w_{m-n+2}^{m-1}\right)}{\sum_{w_m} C\left(w_{m-n+1}^{m}\right) + N_{1+}\left(w_{m-n+1}^{m-1}\bullet\right)} . \quad (2.39)$$

2.4 Αναζήτηση και Αποκωδικοποίηση

Όπως έχει αναφερθεί, ζητούμενο στην Αναγνώριση Φωνής, σύμφωνα με το πλαίσιο όπου την έχουμε εισάγει, είναι η εύρεση της πλέον πιθανής ακολουθίας λέξεων $\hat{W}$ δεδομένης της ακολουθίας παρατηρήσεων O, όπως εισάγεται στη σχέση (2.3), όπου μια παρατήρηση δεν είναι παρά ένα διάνυσμα χαρακτηριστικών, όπως αυτό έχει εξαχθεί κατά το αντίστοιχο στάδιο.

Στην πραγματικότητα, κατά τη διαδικασία που έχει ακολουθηθεί, η ακουστική πιθανοφάνεια υποεκτιμάται, καθώς κατά τη δημιουργία του ακουστικού μοντέλου, όπου απλά ένα GMM υπολογίζει την πιθανότητα εμφάνισης μιας παρατήρησης δεδομένης μιας κατάστασης, δε λαμβάνεται υπόψιν η επίδραση του περιβάλλοντος (context) [33]. Ακόμα, τα δυναμικά εύρη των πιθανοτήτων που προκύπτουν από το ακουστικό και το γλωσσικό μοντέλο διαφέρουν σημαντικά μεταξύ τους. Για το λόγο αυτό, εισάγεται ένας παράγοντας, γνωστός ως LMSF (Language Model Scaling Factor), ο οποίος μειώνει την επίδραση του γλωσσικού μοντέλου ως εξής:

$$\hat{W} = \underset{W\in\mathcal{W}}{\operatorname{argmax}}\, p(O|W)P(W)^{LMSF} . \quad (2.40)$$

Προφανώς, εφόσον οι πιθανότητες είναι αριθμοί μικρότεροι της μονάδας, για να επιτευχθεί το επιθυμητό αποτέλεσμα, θα πρέπει $LMSF > 1$.

Για λόγους αριθμητικής ευστάθειας και υπολογιστικής πολυπλοκότητας, συνήθως δουλεύουμε στο λογαριθμικό πεδίο[4]. Οπότε

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmax}} \{\log p(O|W) + LMSF \cdot \log P(W)\} \ . \tag{2.41}$$

Υπό αυτή τη σκοπιά, ο LMSF δεν είναι παρά ένας συντελεστής βαρύτητας του γλωσσικού μοντέλου, όταν ο συντελεστής βαρύτητας του ακουστικού μοντέλου είναι σταθερός και ίσος με τη μονάδα [49].

Το γλωσσικό μοντέλο, όμως, ουσιαστικά παρέχει μια μοντελοποίηση της πιθανότητας των μεταβάσεων μεταξύ διαδοχικών λέξεων και υπό αυτή την έννοια εισάγει ένα πέναλτι για την εισαγωγή καινούριων λέξεων. Άρα, μειώνοντας την επίδραση του γλωσσικού μοντέλου, μειώνονται οι πιθανότητες εισαγωγής ή, εναλλακτικά, αυξάνεται το πέναλτι εισαγωγής νέων λέξεων. Με άλλα λόγια, το σύστημα αναγνώρισης δείχνει μια προτίμηση προς λίγες μεγάλες λέξεις, παρά προς πολλές μικρότερες [33]. Για να αντισταθμιστεί η αρνητική αυτή επίδραση του LMSF, εισάγεται ένα ξεχωριστό κόστος εισαγωγής λέξεων (Word Insertion Penalty - WIP):

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmax}}\, p(O|W) P(W)^{LMSF} WIP^{N(W)} \tag{2.42}$$

$$= \underset{W \in \mathcal{W}}{\operatorname{argmax}} \{\log p(O|W) + LMSF \cdot \log P(W) + N(W) \cdot \log WIP\} \ , \tag{2.43}$$

όπου $N(W)$ ο αριθμός λέξεων στην ακολουθία W.

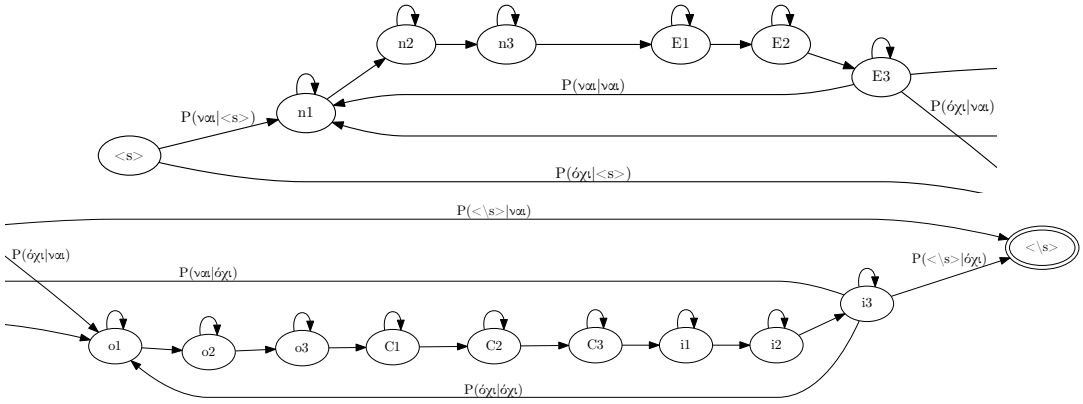
Η εύρεση, τώρα, της πλέον πιθανής ακολουθίας $\hat{W}$ είναι δουλειά του αποκωδικοποιητή (decoder). Για να γίνει πιο εύληπτη η δουλειά του decoder, θα θεωρήσουμε ένα απλό παράδειγμα. Έστω, λοιπόν, πως το λεξιλόγιο πάνω στο οποίο εργαζόμαστε είναι πολύ περιορισμένο και αποτελείται μόνο από τις λέξεις *ναι* και *όχι*. Ο λόγος είναι συνεχόμενος και μπορεί να ειπωθεί απεριόριστος αριθμός από τις δύο λέξεις αυτές στη σειρά. Σύμφωνα με το φωνητικό λεξικό, κάθε λέξη έχει μοναδική προφορά, ως εξής:

$$\text{ναι} \rightarrow \text{n E}$$
$$\text{όχι} \rightarrow \text{o C i}$$

Έστω ότι κάθε φώνημα μοντελοποιείται με ένα HMM τριών καταστάσεων, της τοπολογίας του Σχήματος 2.1iv. Έστω, ακόμα, ότι έχει χρησιμοποιηθεί ένα bigram μοντέλο που δίνει τις πιθανότητες μεταβάσεων μεταξύ διαδοχικών λέξεων. Σύμφωνα με τα όσα έχουν παρουσιαστεί έως τώρα, δημιουργείται ένα δίκτυο που καλείται δίκτυο αναζήτησης και που έχει τη μορφή του Σχήματος 2.5, όπου για απλότητα έχουν παραληφθεί οι καταστάσεις που μοντελοποιούν την πιθανή σιωπή μεταξύ των λέξεων.

Μέσα σε αυτό το δίκτυο αναζήτησης, λοιπόν, καλείται ο decoder να βρει το μονοπάτι εκείνο που, δεδομένης της ακολουθίας παρατηρήσεων, είναι το πλέον πιθανό. Για την εργασία αυτή, οι αποκωδικοποιητές βασίζονται στο δυναμικό αλγόριθμο Viterbi. Θεωρώντας το συνολικό δίκτυο αναζήτησης ως ένα ενιαίο HMM με παραμέτρους λ και σύνολο N καταστάσεων Q, ο αλγόριθμος Viterbi ψάχνει τη βέλτιστη ακολουθία καταστάσεων Q^* που θα δώσει το Viterbi score της εξίσωσης (2.8) και δουλεύει ως εξής [7]:

[4] Οποτε συναντάται λογάριθμος, θα θεωρείται φυσικός (με βάση το e), εκτός και αν ρητά αναφέρεται κάτι διαφορετικό.



Σχήμα 2.5: Παράδειγμα δικτύου αναζήτησης για αποκωδικοποίηση. Το λεξικό αποτελείται από τις λέξεις *ναι* και *όχι*, κάθε φώνημα μοντελοποιείται με HMM τριών καταστάσεων και έχει θεωρηθεί bigram γραμματική. Για λόγους ευκρίνειας, το δίκτυο έχει διασπαστεί σε δύο μέρη.

- Αρχικοποίηση

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(\mathbf{o}_1) && ,1 \leq i \leq N \\ \psi_1(i) &= 0 && ,1 \leq i \leq N \end{aligned}$$

- Αναδρομή

$$\begin{aligned} \delta_t(j) &= \max_{1 \leq i \leq N} \left\{ \delta_{t-1}(i) a_{ij} \right\} b_j(\mathbf{o}_t) && ,2 \leq t \leq T \,, 1 \leq j \leq N \\ \psi_t(j) &= \operatorname*{argmax}_{1 \leq i \leq N} \left\{ \delta_{t-1}(i) a_{ij} \right\} && ,2 \leq t \leq T \,, 1 \leq j \leq N \end{aligned}$$

- Τερματισμός

$$\begin{aligned} P_V &= \max_{1 \leq i \leq N} \left\{ \delta_T(i) \right\} \\ q_T^* &= \operatorname*{argmax}_{1 \leq i \leq N} \left\{ \delta_T(i) \right\} \end{aligned}$$

- Οπισθοδρόμηση (Backtracking)

$$q_t^* = \psi_{t+1} \left(q_{t+1}^* \right), t = T-1, T-2, \cdots, 1$$

Με T συμβολίζεται το τελευταίο frame της τελευταίας λέξης της υπό εξέταση ακολουθίας.

Ο αλγόριθμος βασίζεται στην αναδρομική συνάρτηση δ που ορίζεται ως

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P\left(q_1, q_2, \cdots, q_{t-1}, q_t = i, \mathbf{o}_1 \mathbf{o}_2 \cdots \mathbf{o}_t | \lambda\right) \tag{2.44}$$

και στη διατήρηση back pointers μέσω της συνάρτησης ψ ώστε να είναι δυνατό το τελικό backtracking. Σύμφωνα με τη φιλοσοφία του δυναμικού προγραμματισμού, αν σε κάθε κίνηση γίνει η βέλτιστη κίνηση, τότε η τελική, συνολική τροχιά θα είναι η βέλτιστη. Σημειώνεται ότι, ως συνήθως, οι υπολογισμοί γίνονται στο λογαριθμικό πεδίο, όπου ισοδύναμα χρησιμοποιείται η συνάρτηση

$$\tilde{\delta}_t(j) = \log \delta_t(j) = \max_{1 \leq i \leq N} \left\{ \tilde{\delta}_{t-1}(i) + \tilde{a}_{ij} \right\} + \tilde{b}_j(\mathbf{o}_t) \,. \tag{2.45}$$

Ο αλγόριθμος Viterbi είναι πολυπλοκότητας $O(N^2T)$, δηλαδή σίγουρα πολύ πιο γρήγορος από τον εκθετικό χρόνο που θα απαιτούταν για να τρέξει ο (ακριβής) αλγόριθμος Forward σε όλες τις δυνατές ακολουθίες λέξεων, ώστε εν τέλει να επιλεγεί η πιο ταιριαστή σε σχέση με τις διαθέσιμες παρατηρήσεις. Όπως, όμως, γίνεται εμφανές και από το υποτυπώδες παράδειγμα του Σχήματος 2.5, ο αριθμός καταστάσεων N μπορεί να γίνει πάρα πολύ μεγάλος σε πρακτικές εφαρμογές. Για λόγους περαιτέρω μείωσης, λοιπόν, της υπολογιστικής πολυπλοκότητας, χρησιμοποιείται μια παραλλαγή του αλγορίθμου Viterbi, που κάνει χρήση της ιδέας της ακτινωτής αναζήτησης και γι' αυτό καλείται Ακτινωτή Αναζήτηση Viterbi (Viterbi Beam Search ή και Time Synchronous Viterbi Beam Search) [13].

Σύμφωνα με τη μέθοδο αυτή, σε κάθε βήμα υπολογίζεται το επιμέρους Viterbi score, δηλαδή το Viterbi score της ακολουθίας καταστάσεων που θα ήταν η βέλτιστη εάν η αναζήτηση τελείωνε σε εκείνο το βήμα, έστω $L(t)$. Η τιμή αυτή χρησιμοποιείται για το κλάδεμα (pruning) όσων μονοπατιών δίνουν score μικρότερο του $\theta \cdot L(t)$ όπου θ μια σταθερά μεταξύ 0 και 1 που καλείται εύρος ακτίνας (beam width). Εφόσον δουλεύουμε στο λογαριθμικό πεδίο, το κατώφλι για το pruning είναι το $(\log L(t) - \eta)$, όπου $\eta = -\log\theta$. Εναλλακτικά, μπορεί να προκαθοριστεί ένας σταθερός αριθμός K από τις καλύτερες καταστάσεις που θα κρατιέται σε κάθε βήμα στο μέτωπο αναζήτησης ή και να συνδυαστούν οι δύο παραπάνω προσεγγίσεις. Σε κάθε περίπτωση, είναι απαραίτητο σε κάθε βήμα να αποθηκεύονται οι ενεργές καταστάσεις, δηλαδή αυτές που δεν έχουν ακόμα υποστεί pruning, σε ουρές. Το εύρος ακτίνας αποτελεί μια συμβιβαστική επιλογή ανάμεσα στα ζητούμενα της ακριβούς αναγνώρισης και του μικρού χρόνου αποκωδικοποίησης. Όσο πιο στενή είναι η ακτίνα, τόσο πιο γρήγορη είναι η αποκωδικοποίηση, αλλά ταυτόχρονα τόσο πιο πιθανά είναι τα λάθη λόγω pruning, οπότε τόσο περισσότερο μειώνεται η απόδοση.

Συχνά, χρειάζεται να βρεθεί όχι η βέλτιστη ακολουθία, αλλά ένα σύνολο των έστω N βέλτιστων ακολουθιών (N-best search) [13]. Για παράδειγμα, κατά την αποκωδικοποίηση πολλών περασμάτων, αφού δημιουργηθεί ένα σύνολο υποθέσεων, το τελικό αποτέλεσμα προκύπτει από αυτές μέσω επαναβαθμολόγησής τους (rescoring). Έτσι, μπορεί κατά το πρώτο πέρασμα της αποκωδικοποίησης να χρησιμοποιηθεί ένα απλό γλωσσικό μοντέλο, όπως ένα bigram μοντέλο και στη συνέχεια να γίνει rescoring βάσει ενός πιο πολύπλοκου μοντέλου. Ακόμα, για το rescoring μπορεί να χρησιμοποιηθούν διαφορετικοί συνδυασμοί LMSF και WIP, τα οποία πιθανώς δεν έχουν ληφθεί υπόψιν κατά το πρώτο πέρασμα.

Η αναπαράσταση των υποθέσεων αυτών γίνεται με ένα πλέγμα (word lattice), που ουσιαστικά πρόκειται για έναν κατευθυνόμενο ακυκλικό γράφο όπου κάθε ακμή αντιστοιχεί σε μια πιθανή λέξη, μαζί με το score της και κάθε κόμβος αντιστοιχεί στα όρια των λέξεων και περιέχει τη σχετική χρονική πληροφορία για την εξαναγκασμένη ευθυγράμμιση. Το πλέγμα αυτό μπορεί εύκολα να κατασκευαστεί με μια μικρή παραλλαγή του αλγορίθμου Viterbi. Σε κάθε βήμα πρέπει να κρατιέται ένα σύνολο από δείκτες προς τα πίσω από τις ενεργές καταστάσεις. Έτσι, στο τελευταίο βήμα του αλγορίθμου θα είναι δυνατή η οπισθοδρόμηση με στόχο τη δημιουργία πολλών πιθανών ακολουθιών, καθεμιά από τις οποίες θα είναι ένα μονοπάτι στο word lattice.

2.5 Αξιολόγηση της Αναγνώρισης

Ένα σύστημα αναγνώρισης φωνής εκπαιδεύεται πάνω σε ένα σύνολο δεδομένων εκπαίδευσης και αξιολογείται πάνω σε ένα σύνολο δεδομένων ελέγχου, ενώ πιθανώς χρησιμοποιείται και ένα σύνολο δεδομένων επαλήθευσης για την επιλογή κάποιων απαραίτητων παραμέτρων, γνωστό ως validation set ή held-out set. Η μετρική που έχει επικρατήσει για την αξιολό-

γηση των συστημάτων Αυτόματης Αναγνώρισης Φωνής είναι γνωστή ως Word Error Rate (WER) και δίνεται από τη σχέση (2.46).

$$WER = \frac{w_\epsilon \cdot \#\text{εισαγωγές} + w_\alpha \cdot \#\text{αντικαταστάσεις} + w_\delta \cdot \#\text{διαγραφές}}{\text{συνολικός αριθμός λέξεων στις απομαγνητοφωνήσεις}} \tag{2.46}$$

Θέτοντας όλους τους συντελεστές βαρύτητας που εμφανίζονται στη σχέση (2.46) ίσους με τη μονάδα, που είναι και η πλέον συνήθης πρακτική, παίρνουμε τη μετασχηματισμένη σχέση

$$WER = \frac{\#\text{εισαγωγές} + \#\text{αντικαταστάσεις} + \#\text{διαγραφές}}{\text{συνολικός αριθμός λέξεων στις απομαγνητοφωνήσεις}}\,. \tag{2.47}$$

Ο συνολικός αριθμός των εισαγωγών, των αντικαταστάσεων και των διαγραφών είναι τέτοιος ώστε να προκύπτει η ελάχιστη απόσταση Levenshtein μεταξύ του παραγόμενου από το σύστημα αναγνώρισης κειμένου και του κειμένου της απομαγνητοφώνησης που αφορά το σύνολο ελέγχου [33].

Εναλλακτικά, μπορεί να χρησιμοποιηθεί η ισοδύναμη μετρική της ακρίβειας των λέξεων που αναγνωρίστηκαν (Word Accuracy - WACC) [13], που ορίζεται ως

$$WACC = 100\% - WER\,. \tag{2.48}$$

Προφανώς, όσο υψηλότερο το WACC ή όσο χαμηλότερο το WER, τόσο καλύτερη θεωρείται η αναγνώριση.

Κεφάλαιο 3

Μετατροπείς Πεπερασμένης Κατάστασης με Βάρη

3.1 Βασικοί Ορισμοί

Οι Μετατροπείς Πεπερασμένης Κατάστασης με Βάρη (Weighted Finite-State Transducers - WFSTs) είναι η γενικότερη κατηγορία Αυτομάτων Πεπερασμένης Κατάστασης ή απλά Πεπερασμένων Αυτομάτων (Finite Automata - FA). Στην πιο απλή του μορφή, ένα FA είναι ένας Αποδοχέας Πεπερασμένης Κατάστασης (Finite-State Acceptor - FSA), ο οποίος φορμαλιστικά ορίζεται ως μια πεντάδα στοιχείων $(Q, \Sigma, \delta, q_0, F)$ [50], όπου

- Q ένα πεπερασμένο σύνολο καταστάσεων,
- Σ ένα πεπερασμένο σύνολο συμβόλων που καλείται αλφάβητο,
- $\delta : Q \times \Sigma \to Q$ η συνάρτηση μετάβασης που αναλόγως της τρέχουσας κατάστασης και του τρέχοντος συμβόλου καθορίζει ποια θα είναι η επόμενη κατάσταση,
- $q_0 \in Q$ η αρχική κατάσταση και
- $F \subseteq Q$ ένα σύνολο τελικών καταστάσεων ή καταστάσεων αποδοχής.

Στον παραπάνω ορισμό, μπορούμε αντί της συνάρτησης δ να ορίσουμε ισοδύναμα ένα πολυσύνολο επιτρεπτών μεταβάσεων $E \subseteq Q \times \Sigma \times Q$, καθώς επίσης η αρχική κατάσταση μπορεί να γενικευτεί σε ένα σύνολο $I \subseteq Q$ αρχικών καταστάσεων. Ακόμη, από τον παραπάνω ορισμό απουσιάζει το κενό σύμβολο ϵ, δηλαδή για κάθε μετάβαση στο αυτόματο απαιτείται ένα συγκεκριμένο σύμβολο του αλφαβήτου. Επίσης, εφόσον η δ παρουσιάζεται ως συνάρτηση από το $(Q \times \Sigma)$ στο Q, γίνεται η υπόθεση πως αναφερόμαστε σε Ντετερμινιστικό Πεπερασμένο Αυτόματο (Deterministic FA - DFA). Εάν από μία κατάσταση και με το ίδιο σύμβολο είναι επιτρεπτές περισσότερες από μία μεταβάσεις, αναφερόμαστε σε Μη-Ντετερμινιστικά Πεπερασμένα Αυτόματα (Non-deterministic FA - NFA). Ωστόσο, στην περίπτωση FSAs, αποδεικνύεται ότι για κάθε NFA υπάρχει ισοδύναμο DFA, αλλά και ότι για κάθε ϵ-NFA (όπου επιτρέπονται μεταβάσεις χωρίς κάποιο σύμβολο του αλφαβήτου) υπάρχει ισοδύναμο NFA [50].

Δουλειά ενός FSA, όπως υποδηλώνει και το όνομά του, είναι να αποδέχεται ή να απορρίπτει μια συμβολοακολουθία που δίνεται ως είσοδος και η οποία αποτελείται από σύμβολα που ανήκουν στο αλφάβητο Σ. Το σύνολο των συμβολοακολουθιών που αποδέχεται το FSA, δηλαδή το σύνολο των συμβολοακολουθιών για τις οποίες υπάρχουν επιτρεπτές μεταβάσεις

που οδηγούν από μια αρχική σε μια τελική κατάσταση, αποτελεί τη γλώσσα που αναγνωρίζει το FSA. Κάθε γλώσσα που μπορεί να παρασταθεί από ένα FSA αποτελεί μία κανονική γλώσσα (regular language).

Από την άλλη, δουλειά ενός Μετατροπέα Πεπερασμένης Κατάστασης (Finite-State Transducer - FST) είναι να μετασχηματίζει μία αναπαράσταση σε μία άλλη, δηλαδή να παίρνει ως είσοδο μία συμβολοακολουθία και αντί να παράγει ένα δυαδικό αποτέλεσμα αποδοχής ή μη, να παράγει μία νέα συμβολοακολουθία [33]. Κάθε τέτοια συσχέτιση που μπορεί να παρασταθεί από ένα FST αποτελεί μια ρητή σχέση (rational relation). Σημειώνεται ότι ένας αποδοχέας μπορεί να θεωρηθεί ειδική περίπτωση μετατροπέα όπου η συμβολοακολουθία εξόδου ταυτίζεται με τη συμβολοακολουθία εισόδου. Εάν εισάγουμε και την έννοια του βάρους, δηλαδή εάν κάθε μετάβαση από μία κατάσταση σε μία άλλη συνδέεται με κάποιο βάρος, το οποίο μπορεί να σχετίζεται με την πιθανότητα μετάβασης ή με κάποιο κόστος που επιφέρει η επιλογή της συγκεκριμένης μετάβασης, λαμβάνουμε τα WFSTs (και αντίστοιχα τους Αποδοχείς Πεπερασμένης Κατάστασης με Βάρη (Weighted Finite-State Acceptors - WFSAs)). Κάθε σχέση που μπορεί να παρασταθεί από ένα WFST αποτελεί μία ρητή δυναμοσειρά (rational power series).

Προτού δοθεί ο τυπικός ορισμός ενός WFSA ή WFST, θα πρέπει να οριστούν οι αφηρημένες αλγεβρικές δομές του μονοειδούς και του ημιδακτυλίου [51].

Ένα μονοειδές, συμβολιζόμενο ως $< \mathbb{M}, \circ, \bar{1} >$, αποτελείται από ένα σύνολο $\mathbb{M}$, μια προσεταιριστική δυαδική πράξη $\circ$ πάνω στο $\mathbb{M}$ και ένα ουδέτερο στοιχείο $\bar{1}$ ώστε $\bar{1} \circ a = a \circ \bar{1} = a \ \forall a \in \mathbb{M}$. Εάν επιπλέον ισχύει $a \circ b = b \circ a \ \forall a \in \mathbb{M}, \forall b \in \mathbb{M}$ τότε το μονοειδές καλείται αντιμεταθετικό.

Ένας ημιδακτύλιος, συμβολιζόμενος ως $< \mathbb{A}, \oplus, \otimes, \bar{0}, \bar{1} >$, αποτελείται από ένα σύνολο $\mathbb{A}$, εφοδιασμένο με δύο δυαδικές πράξεις $\oplus$ και $\otimes$ και δύο σταθερά στοιχεία $\bar{0}$ και $\bar{1}$, ώστε να ικανοποιούνται τα εξής αξιώματα:

(i) το $< \mathbb{A}, \oplus, \bar{0} >$ είναι ένα αντιμεταθετικό μονοειδές ,

(ii) το $< \mathbb{A}, \otimes, \bar{1} >$ είναι ένα μονοειδές,

(iii) ισχύει η επιμεριστική ιδιότητα ώστε
$a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$ και $(a \oplus b) \otimes c = (a \otimes c) \oplus (b \otimes c) \ \forall a \in \mathbb{A}, \forall b \in \mathbb{A}, \forall c \in \mathbb{A}$,

(iv) $\bar{0} \otimes a = a \otimes \bar{0} = \bar{0} \ \forall a \in \mathbb{A}$

Έτσι, ένα WFSA πάνω στο σύνολο στοιχείων $\mathbb{W}$ ενός ημιδακτυλίου ορίζεται ως μια επτάδα στοιχείων $(Q, \Sigma, I, F, E, \lambda, \rho)$ [52, 13], όπου

- Q ένα πεπερασμένο σύνολο καταστάσεων,
- Σ ένα πεπερασμένο σύνολο συμβόλων που αποτελεί το αλφάβητο εισόδου,
- $I \subseteq Q$ ένα σύνολο αρχικών καταστάσεων,
- $F \subseteq Q$ ένα σύνολο τελικών καταστάσεων,
- $E \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times \mathbb{W} \times Q$ ένα πεπερασμένο πολυσύνολο επιτρεπτών μεταβάσεων,
- $\lambda : I \to \mathbb{W}$ συνάρτηση που αντιστοιχεί μία τιμή βάρους σε κάθε αρχική κατάσταση και
- $\rho : F \to \mathbb{W}$ συνάρτηση που αντιστοιχεί μία τιμή βάρους σε κάθε τελική κατάσταση.

Ένα μονοπάτι π είναι μια ακολουθία πεπερασμένων (έστω n) διαδοχικών μεταβάσεων $t_0, t_1, \cdots, t_n$, όπου $t_i = (p(t_i), l(t_i), w(t_i), n(t_i)) \in E$, $i = 1, 2, \cdots, n$ και $p(t_{i+1}) = n(t_i)$. Εάν $p(t_0) \in I$ και $n(t_n) \in F$, τότε λέμε ότι το WFSA αποδέχεται τη συμβολοακολουθία $l(t_0), l(t_1), \cdots, l(t_n)$ με κόστος

$$w(\pi) = \lambda(p(t_0)) \otimes w(t_0) \otimes w(t_1) \otimes \cdots \otimes w(t_n) \otimes \rho(n(t_n)) \,. \tag{3.1}$$

Συνεπώς, ένα WFSA μπορεί να θεωρηθεί ως μια αντιστοίχιση μεταξύ συμβολοακολουθιών και βαρών. Αξίζει να σημειωθεί ότι τα HMMs μπορούν να ειδωθούν σαν μια ειδική κατηγορία WFSAs [52].

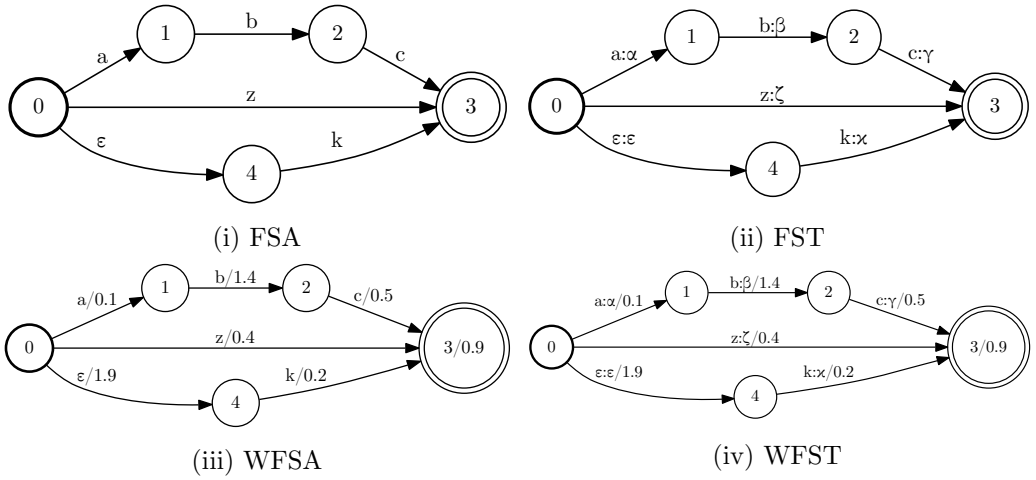
Παρόμοια, ένα WFST πάνω στο σύνολο στοιχείων $\mathbb{W}$ ενός ημιδακτυλίου ορίζεται στην πιο γενική του μορφή ως μια οκτάδα στοιχείων $(Q, \Sigma, \Delta, I, F, E, \lambda, \rho)$ [52, 13], όπου

- Q ένα πεπερασμένο σύνολο καταστάσεων,
- Σ ένα πεπερασμένο σύνολο συμβόλων που αποτελεί το αλφάβητο εισόδου,
- Δ ένα πεπερασμένο σύνολο συμβόλων που αποτελεί το αλφάβητο εξόδου,
- $I \subseteq Q$ ένα σύνολο αρχικών καταστάσεων,
- $F \subseteq Q$ ένα σύνολο τελικών καταστάσεων,
- $E \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times (\Delta \cup \{\epsilon\}) \times \mathbb{W} \times Q$ ένα πεπερασμένο πολυσύνολο επιτρεπτών μεταβάσεων,
- $\lambda : I \to \mathbb{W}$ συνάρτηση που αντιστοιχεί μία τιμή βάρους σε κάθε αρχική κατάσταση[1] και
- $\rho : F \to \mathbb{W}$ συνάρτηση που αντιστοιχεί μία τιμή βάρους σε κάθε τελική κατάσταση.

Τα FA παρουσιάζονται ως κατευθυνόμενοι γράφοι, όπου κατά σύμβαση κάθε κατάσταση συμβολίζεται με έναν κυκλικό κόμβο, ενώ οι τελικές καταστάσεις συμβολίζονται με κυκλικούς κόμβους με διπλή περιφέρεια. Κάθε κόμβος είναι επισημειωμένος με έναν μοναδικό αριθμό. Οι ακμές του γράφου παριστάνουν τις μεταβάσεις μεταξύ καταστάσεων. Στα WFSTs κάθε ακμή είναι επισημειωμένη με το χαρακτηριστικό $l_i(t) : l_o(t)/w(t)$, όπου $l_i(t)$ το σύμβολο εισόδου, $l_o(t)$ το σύμβολο εξόδου και $w(t)$ το βάρος. Σε περίπτωση που το $w(t)$ δεν εμφανίζεται, αυτό θεωρείται ίσο με $\bar{1}$, ενώ εάν όλες οι ακμές είναι επισημειωμένες ως $l(t)/w(t)$, τότε πρόκειται για WFSA. Οι αρχικοί και τελικοί κόμβοι επισημειώνονται και με το βάρος τους (αν αυτό δεν ισούται με $\bar{1}$). Παραδείγματα FA διαφορετικών τύπων παρουσιάζονται στο Σχήμα 3.1.

Συχνά, για τους σκοπούς της αναγνώρισης φωνής, τα βάρη παίζουν το ρόλο των πιθανοτήτων, οπότε ο κατάλληλος ημιδακτύλιος θα ήταν ο ημιδακτύλιος πιθανοτήτων (probability semiring) $< [0, 1], +, \cdot, 0, 1 >$. Στο λογαριθμικό πεδίο όπου γίνονται συνήθως οι υπολογισμοί για λόγους αριθμητικής ευστάθειας, χρησιμοποιούνται ως κόστη οι αρνητικοί λογάριθμοι των πιθανοτήτων. Επειδή στόχος είναι η εύρεση της πλέον πιθανής ακολουθίας λέξεων (με χρήση του αλγορίθμου Viterbi), ο πλέον κατάλληλος ημιδακτύλιος για αναγνώριση φωνής είναι ο τροπικός ημιδακτύλιος (tropical semiring) $< \mathbb{R}_+ \cup \{\infty\}, \min, +, \infty, 0 >$. Ορισμένες φορές χρησιμοποιείται και ο λογαριθμικός ημιδακτύλιος (log semiring) $< \mathbb{R}_+ \cup \{\infty\}, \oplus_{\log}, +, \infty, 0 >$, όπου $x \oplus_{\log} y = -\log(e^{-x} + e^{-y})$.

[1] Χωρίς βλάβη της γενικότητας, τα Πεπερασμένα Αυτόματα με Βάρη (Weighted FA - WFA) μπορούν να περιοριστούν ώστε να επιτρέπεται μία μόνο αρχική κατάσταση με βάρος $\bar{1}$. Πρόκειται για μία σύμβαση που ακολουθείται και στην πράξη από τα συστήματα που υλοποιούν WFA για λόγους απλότητας [53, 37].



(i) FSA (ii) FST

(iii) WFSA (iv) WFST

Σχήμα 3.1: Παραδείγματα διαφορετικών κατηγοριών Πεπερασμένων Αυτομάτων. Οι τρεις πρώτες μπορούν να θεωρηθούν ειδικές περιπτώσεις του WFST.

3.2 Κύριες Πράξεις και Λειτουργίες

Η πραγματική δύναμη των WFSTs έγκειται στη δυνατότητα αποδοτικής διαχείρισής τους και μετασχηματισμού τους μέσω ποικίλων πράξεων, εναδικών και δυϊκών. Η Ενότητα αυτή αποτελεί μία επισκόπηση των βασικότερων από τις πράξεις αυτές.

3.2.1 Ρητές Πράξεις

Σύμφωνα με τη Θεωρία Υπολογισμού, ορίζονται τρεις κανονικές πράξεις (regular operations) στις γλώσσες, η ένωση (union), η παράθεση (concatenation) και η κλειστότητα κατά Kleene (Kleene closure) ή άστρο του Kleene (Kleene star) [50], ως εξής: Αν L_1 και L_2 είναι κανονικές γλώσσες, τότε

- η ένωση των L_1, L_2 είναι η $L_1 \cup L_2 = \{x : x \in L_1 \text{ ή } x \in L_2\}$,
- η παράθεση των L_1, L_2 είναι η $L_1 \cdot L_2 = \{xy : x \in L_1 \text{ και } y \in L_2\}$,
- η κλειστότητα κατά Kleene της L_1 είναι η $L_1^* = \{x_1 x_2 \cdots x_k : k \geq 0 \text{ και κάθε } x_i \in L_1\}$.

Εφόσον κάθε κανονική γλώσσα μπορεί να παρασταθεί από το FSA που την αποδέχεται, οι αντίστοιχες πράξεις μπορούν να οριστούν υπό το πρίσμα των FSAs.

Αντίστοιχα με τα παραπάνω, για τα WFSTs ορίζονται οι παρακάτω τρεις ρητές πράξεις (rational operations) [54], όπου με $T(x, y)$ συμβολίζεται το συνολικό βάρος για τη μετατροπή της συμβολοακολουθίας x στη συμβολοακολουθία y μέσω του μετατροπέα T:

- η ένωση (ή το άθροισμα) δύο WFSTs T_1 και T_2

$$(T_1 \cup T_2)(x, y) = (T_1 \oplus T_2)(x, y) = T_1(x, y) \oplus T_2(x, y) \quad , \forall (x, y) \in \Sigma^* \times \Delta^* , \quad (3.2)$$

- η παράθεση (ή το γινόμενο) δύο WFSTs T_1 και T_2

$$(T_1 \cdot T_2)(x, y) = (T_1 \otimes T_2)(x, y) = \bigoplus_{\substack{x=x_1x_2 \\ y=y_1y_2}} T_1(x_1, y_1) \otimes T_2(x_2, y_2) \quad , \forall (x, y) \in \Sigma^* \times \Delta^* , \quad (3.3)$$

όπου το άθροισμα τρέχει για όλους τους πιθανούς τρόπους διάσπασης της συμβολοακολουθίας x σε x_1 και x_2 και όμοια για τη συμβολοακολουθία y,

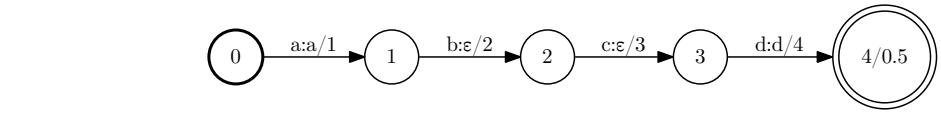
- η κλειστότητα κατά Kleene (ή απλά κλειστότητα) ενός WFST

$$T^*(x,y) = \bigoplus_{n=0}^{+\infty} T^n(x,y) \quad ,\forall (x,y) \in \Sigma^* \times \Delta^* \,, \tag{3.4}$$

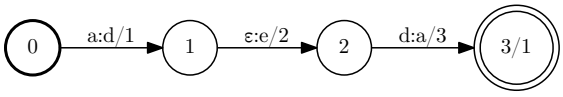
όπου

$$T^n(x,y) = \begin{cases} \overbrace{(T \otimes \cdots \otimes T)}^{\text{n φορές}}(x,y) & ,n > 0 \\ \begin{cases} \bar{1} & ,(x,y) = (\epsilon,\epsilon) \\ \bar{0} & ,\text{αλλιώς} \end{cases} & ,n = 0 \end{cases} . \tag{3.5}$$

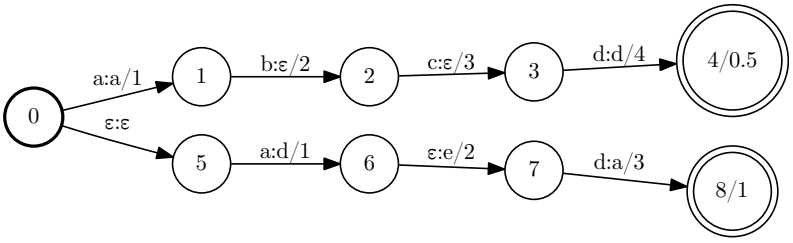
Μία εποπτική παρουσίαση των παραπάνω ρητών πράξεων αποτελεί το Σχήμα 3.2.



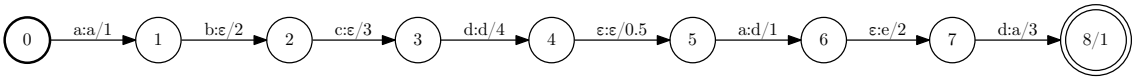
(i) WFST T_1



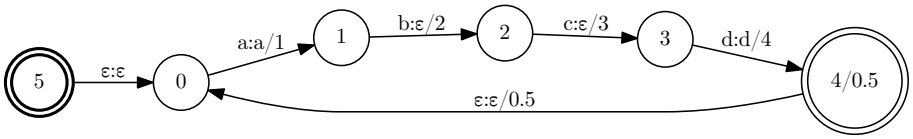
(ii) WFST T_2



(iii) Ένωση $T_1 \cup T_2$



(iv) Παράθεση $T_1 \cdot T_2$



(v) Κλειστότητα κατά Kleene T_1^*

Σχήμα 3.2: Οι ρητές πράξεις που ορίζονται στα WFSTs, υποθέτοντας τροπικό ημιδακτύλιο[2].

[2]Από εδώ και στο εξής, θα γίνεται χρήση τροπικού ημιδακτυλίου, εκτός εάν αναφέρεται ρητά το αντίθετο.

3.2.2 Προβολή, Αντιστροφή και Σύνθεση

Άλλες σημαντικές πράξεις σε ένα WFST είναι η προβολή (projection), η αντιστροφή (inversion) και η σύνθεση (composition) [13]. Προβολή ονομάζεται η διαδικασία κατά την οποία ένας μετατροπέας αντιστοιχίζεται σε έναν αποδοχέα, οι μεταβάσεις του οποίου επισημειώνονται είτε από τα σύμβολα εισόδου (ανώτερη ή πρώτη προβολή - upper ή first projection), είτε από τα σύμβολα εξόδου (κατώτερη ή δεύτερη προβολή - lower ή second projection) των αντίστοιχων μεταβάσεων του μετατροπέα [33]. Φορμαλιστικά, η πρώτη και δεύτερη προβολή ορίζονται αντίστοιχα ως [54]

$$\downarrow T(x) = \bigoplus_{y} T(x,y),, \tag{3.6}$$

$$T(x) \downarrow = \bigoplus_{x} T(x,y)\,. \tag{3.7}$$

Η αντιστροφή απλά αντιστρέφει τα σύμβολα εισόδου και εξόδου ενός μετατροπέα:

$$T^{-1}(x,y) = T(y,x)\,. \tag{3.8}$$

Κατά τη σύνθεση, από δύο WFSTs, έστω $T_1 = (Q_1, \Sigma_1, \Delta_1, I_1, F_1, E_1, \lambda_1, \rho_1)$ και $T_2 = (Q_2, \Delta_1, \Delta_2, I_2, F_2, E_2, \lambda_2, \rho_2)$, παράγεται ένα νέο WFST $T = (Q, \Sigma_1, \Delta_2, I, F, E, \lambda, \rho) = T_1 \circ T_2$, όπου διαισθητικά εάν η συμβολοακολουθία x μετατρέπεται στη z από το μετατροπέα T_1 και η z στην y από το μετατροπέα T_2, τότε ο T μετατρέπει τη x στην y [33]. Φορμαλιστικά, είναι

$$T_1 \circ T_2(x,y) = \bigoplus_{z \in \Delta_1^*} T_1(x,z) \otimes T_2(z,y)\,. \tag{3.9}$$

Αξίζει να σημειωθεί πως στην περίπτωση των acceptors, η σύνθεση ανάγεται στην πράξη της τομής (intersection).

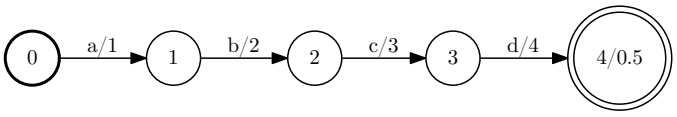
Για τον αλγοριθμικό υπολογισμό της σύνθεσης [13], κάθε κατάσταση $q \in Q$ στο T μπορεί να θεωρηθεί ως ένα ζεύγος $q = (q_1, q_2), \in Q_1 \times Q_2$. Εάν στο T_1 υπάρχει η μετάβαση t_1 από το q_1 στο q_1' με την επισημείωση $l_i(t_1) : l_0(t_1)/w(t_1)$ και στο T_2 υπάρχει η μετάβαση t_2 από το q_2 στο q_2' με την επισημείωση $l_i(t_2) : l_0(t_2)/q(t_2)$, τότε στο T υπάρχει η μετάβαση t από το (q_1, q_2) στο (q_1', q_2') με την επισημείωση $l_i(t_1) : l_o(t_2)/(w(t_1) \otimes w(t_2))$. Ακόμα, κάθε αρχική κατάσταση (i_1, i_2) έχει βάρος $\lambda(i_1) \otimes \lambda(i_2)$ και, ομοίως, κάθε τελική κατάσταση (f_1, f_2) έχει βάρος $\rho(f_1) \otimes \rho(f_2)$.

Κατά την παραπάνω ανάλυση, έγινε η υπόθεση ότι καμία κατάσταση του T_1 δεν έχει ϵ-έξοδο και καμία κατάσταση του T_2 δεν έχει ϵ-είσοδο. Στη γενικότερη περίπτωση, προηγείται ένα πρώτο βήμα όπου τα T_1, T_2 μετατρέπονται στα T_1', T_2', αντίστοιχα, και ο αλγόριθμος που περιγράφηκε υπολογίζει το $T_1' \circ T_2'$. Κατά το πρώτο αυτό βήμα, το T_1' υπολογίζεται από το T_1 αντικαθιστώντας το ϵ στις ϵ-εξόδους με ένα νέο σύμβολο, έστω ϵ_o, ενώ το T_2' υπολογίζεται από το T_2 αντικαθιστώντας το ϵ στις ϵ-εισόδους με ένα νέο σύμβολο, έστω ϵ_i. Ακόμα, σε όλες τις καταστάσεις του T_1' εισάγονται self-loops με επισημειώσεις $\epsilon : \epsilon\_i$ και σε όλες τις καταστάσεις του T_2' εισάγονται self-loops με επισημειώσεις $\epsilon\_o : \epsilon$ [13].

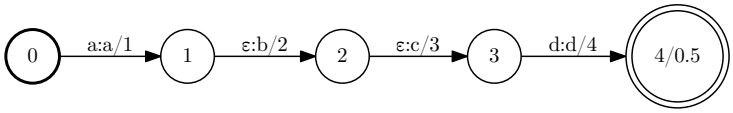
Οι βασικές πράξεις που αναλύθηκαν έως τώρα παρουσιάζονται εποπτικά στο Σχήμα 3.3.

3.3 Λειτουργίες Βελτιστοποίησης

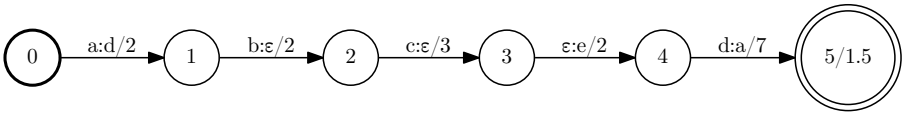
Οι λειτουργίες βελτιστοποίησης που μπορούν να εφαρμοστούν σε ένα WFST έχουν στόχο να αλλάξουν τη δομή τους κατά τρόπο τέτοιο ώστε να είναι πιο αποδοτική η διαχείρισή τους,



(i) Πρώτη Προβολή $\downarrow T_1$



(ii) Αντιστροφή T_1^{-1}



(iii) Σύνθεση $T_1 \circ T_2$

Σχήμα 3.3: Προβολή, αντιστροφή και σύνθεση στα WFSTs. Τα T_1 και T_2 ορίζονται στο Σχήμα 3.2.

όσον αφορά τους απαραίτητους υπολογιστικούς πόρους, τόσο σε χρόνο, όσο και σε μνήμη [13]. Προφανώς, η αλλαγή αυτή της δομής θα πρέπει να οδηγεί σε ένα ισοδύναμο WFST. Δύο WFSTs ονομάζονται ισοδύναμα εάν μετασχηματίζουν την ίδια ακολουθία εισόδου στην ίδια ακολουθία εξόδου με το ίδιο συνολικό βάρος [52].

Ίσως η πιο σημαντική από τις λειτουργίες βελτιστοποίησης ενός WFST είναι η διαδικασία μετατροπής του σε ντετερμινιστικό. Ένας μετατροπέας ονομάζεται ντετερμινιστικός ή ακολουθιακός εάν από κάθε κατάσταση υπάρχει μία μόνο μετάβαση δεδομένου ενός συμβόλου εισόδου και επιπλέον καμία κατάσταση δεν έχει ϵ-είσοδο [52]. Σε αντίθεση με την κλασική θεωρία των FA που ασχολείται με αυτόματα χωρίς βάρη, σύμφωνα με την οποία για κάθε NFA υπάρχει ένα ισοδύναμο DFA, για αυτόματα με βάρη δεν ισχύει κάτι τέτοιο. Ωστόσο, σχεδόν όλα τα WFSTs που χρησιμοποιούνται στην αναγνώριση φωνής μπορούν να μετατραπούν σε ντετερμινιστικά είτε απευθείας είτε μετά από κάποιους βοηθητικούς μετασχηματισμούς. Για παράδειγμα, κάθε ακυκλικό αυτόματο με βάρη έχει ντετερμινιστικό ισοδύναμο [52].

Ο αλγόριθμος για να βρεθεί το ντετερμινιστικό ισοδύναμο ενός WFST υποθέτει ότι ο ημιδακτύλιος που χρησιμοποιείται είναι ασθενώς αριστερά διαιρετός (weakly left-divisible), δηλαδή ότι για κάθε x και y στο σύνολο $\mathbb{A}$ του ημιδακτυλίου, τέτοια ώστε $x \oplus y \neq \bar{0}$, υπάρχει τουλάχιστον ένα $x \in \mathbb{A}$ ώστε $x = (x \oplus y) \otimes z$ [13]. Ο τροπικός ημιδακτύλιος είναι ασθενώς αριστερά διαιρετός.

Εάν η διαδικασία μετατροπής ενός αυτομάτου σε ντετερμινιστικό εφαρμοστεί σε ένα ϵ-NFA, θεωρώντας το ϵ ως ένα κανονικό σύμβολο, τότε το FA που θα προκύψει θα εξακολουθεί να έχει ϵ-μεταβάσεις, οπότε δε θα είναι ντετερμινιστικό. Η διαδικασία κατά την οποία εξαλείφονται οι ϵ-μεταβάσεις από ένα αυτόματο καλείται ϵ-απομάκρυνση (ϵ-removal) [13]. Πρόκειται για μία σημαντική διαδικασία, καθώς η ύπαρξη ϵs εισάγει καθυστερήσεις σε πολλές εφαρμογές. Το παραγόμενο WFST δεν περιέχει νέες καταστάσεις, αλλά περιέχει νέες μεταβάσεις, οι οποίες, ωστόσο, δε μεταβάλλουν τη σχέση που παριστάνει το αρχικό WFST. Οι μεταβάσεις που απομακρύνονται από ένα FST με το ϵ-removal είναι μόνο όσες είναι επισημειωμένες με $\epsilon : \epsilon$, που σημαίνει ότι μετά την εν λόγω διαδικασία, μπορεί να συνεχίσουν να υπάρχουν ϵs στην είσοδο ή στην έξοδο.

Για να εξαλειφθούν όσο το δυνατόν περισσότερα ϵs, προηγείται μια διαδικασία γνωστή ως συγχρονισμός (synchronization), κατά την οποία δοθέντος ενός WFST T, υπολογίζεται

ένα ισοδύναμο WFST T', το οποίο είναι συγχρονισμένο [54]. Ένα WFST καλείται συγχρονισμένο εάν η καθυστέρηση κάθε επιτυχούς μονοπατιού του είναι 0 ή μεταβάλλεται αυστηρά μονοτονικά. Ως καθυστέρηση (delay) $d(\pi)$ ενός μονοπατιού π ορίζεται η διαφορά μεταξύ του μήκους της συμβολοακολουθίας εξόδου και του μήκους της συμβολοακολουθίας εισόδου του μονοπατιού. Ο αλγόριθμος που χρησιμοποιείται για την εν λόγω διαδικασία απαιτεί μόνο το WFST πάνω στο οποίο εφαρμόζεται να έχει φραγμένες καθυστερήσεις, ικανή και αναγκαία συνθήκη για το οποίο είναι η καθυστέρηση κάθε κύκλου του WFST να ισούται με μηδέν. Διαισθητικά, κατά το synchronization, μειώνονται οι μεταβάσεις της μορφής $\epsilon : x$ και $x : \epsilon$ και αυξάνονται οι μεταβάσεις της μορφής $x : x$, που δεν περιέχουν ϵ, και της μορφής $\epsilon : \epsilon$, που μπορούν να εξαλειφθούν κατά το ϵ-removal [13].

Όπως με το synchronization προκύπτει μια αποδοτικότερη κατανομή των ϵs σε ένα WFST, έτσι και με το σπρώξιμο των βαρών (weight pushing) προκύπτει μια αποδοτικότερη κατανομή των βαρών. Συγκεκριμένα, με το weight pushing τα βάρη "σπρώχνονται" προς τις πιο αρχικές καταστάσεις του WFST. Με τον τρόπο αυτό, επιταχύνεται σημαντικά η διαδικασία της αναζήτησης των συντομότερων (υπό την έννοια των ελαχιστοβαρών σε τροπικό δακτύλιο) μονοπατιών στο WFST, αφού μπορούν από νωρίς να απορριφθούν μονοπάτια που φαίνεται ότι συνδέονται με μεγάλο βάρος [13].

Ο αντίστοιχος αλγόριθμος λειτουργεί σε δύο βήματα. Πρώτα, υπολογίζεται ένα δυναμικό $V(q)$ για κάθε κατάσταση q ενός WFST T. Συμβολίζοντας ως $\Pi(q, F)$ το σύνολο των μονοπατιών από την κατάσταση q σε κάποια τελική κατάσταση $q' \in F$, το $V(q)$ ορίζεται ως το συντομότερο μονοπάτι από την q στο F:

$$V(q) = \bigoplus_{\pi \in \Pi(q,F)} \{w(\pi) \otimes \rho(n(\pi))\}, \tag{3.10}$$

όπου $w(\pi)$ το συνολικό βάρος του μονοπατιού και $n(\pi)$ η κατάσταση όπου καταλήγει το μονοπάτι. Στη συνέχεια, συμβολίζοντας την αφετηρία μιας μετάβασης e με $p(e)$ και την κατάληξή της με $n(e)$, γίνεται ο εξής επανυπολογισμός των βαρών:

$$w(e) = V(p(e))^{-1} \otimes w(e) \otimes V(p(e)), \qquad \forall e : V(p(e)) \neq \bar{0} \tag{3.11}$$

$$\lambda(q) = \lambda(q) \otimes V(q), \qquad \forall q \in I \tag{3.12}$$

$$\rho(q) = V(q)^{-1} \otimes \rho(q), \qquad \forall q \in F : V(q) \neq \bar{0} \tag{3.13}$$

Ο αλγόριθμος υποθέτει ότι ο ημιδακτύλιος που χρησιμοποιείται είναι ασθενώς αριστερά διαιρετός, είναι k-κλειστός και δεν έχει μη-μηδενικά στοιχεία που αθροίζουν στο $\bar{0}$ (zero-sum free). Ένας ημιδακτύλιος με σύνολο $\mathbb{A}$ καλείται k-κλειστός εάν υπάρχει $k \geq 0$ ώστε $\bigoplus_{n=0}^{k+1} x^n = \bigoplus_{n=0}^{k} x^n$ για κάθε $x \in \mathbb{A}$, ενώ καλείται zero-sum free εάν $x \oplus y = \bar{0} \Rightarrow x = y = \bar{0}$ για κάθε $x, y \in \mathbb{A}$. Ο τροπικός ημιδακτύλιος ικανοποιεί και τις τρεις υποθέσεις, και μάλιστα είναι 0-κλειστός.

Τέλος, για αποφυγή πλεονασμών και καλύτερη διαχείριση της μνήμης, μεγάλης σημασίας είναι η διαδικασία ελαχιστοποίησης (minimization) ενός FA, δηλαδή της δημιουργίας ενός ισοδύναμου FA με τον ελάχιστο αριθμό καταστάσεων. Για την ελαχιστοποίηση ενός WFST [54] αρκεί πρώτα να εφαρμοστεί ο αλγόριθμος σπρωξίματος βαρών και στη συνέχεια ένας κλασικός αλγόριθμος ελαχιστοποίησης, όπως αυτός εφαρμόζεται σε ένα FSA, θεωρώντας την επισημείωση $l_i(e) : l_o(e)/w(e)$ μιας μετάβασης e ως ένα μοναδικό σύμβολο εισόδου, σαν να επρόκειτο για FSA. Σημειώνεται ότι οι κλασικοί αλγόριθμοι ελαχιστοποίησης υποθέτουν DFA και όχι NFA, οπότε σαν πρώτο βήμα προηγείται η διαδικασία μετατροπής ενός WFST σε ντετερμινιστικό.

Κατά την ελαχιστοποίηση ενός DFA πρώτα εξαλείφονται όλες τις απρόσιτες καταστάσεις και εν συνεχεία συγχωνεύονται όλες οι ισοδύναμες καταστάσεις. Δύο καταστάσεις λέγονται ισοδύναμες αν δεν είναι k-διακρίσιμες για κανένα k. Ο ορισμός της k-διακρισιμότητας είναι αναδρομικός: Δύο καταστάσεις είναι 0-διακρίσιμες αν η μία είναι τελική ενώ η άλλη όχι, ενώ λέγονται $(i+1)$-διακρίσιμες αν υπάρχει σύμβολο με το οποίο οδηγούν σε i-διακρίσιμες καταστάσεις.

Χαρακτηριστικά αποτελέσματα της εφαρμογής των λειτουργιών βελτιστοποίησης που παρουσιάστηκαν για τα WFSTs φαίνονται στο Σχήμα 3.4. Αξίζει να προσέξουμε ότι η σειρά με την οποία εφαρμόζονται διδοχικά αλγόριθμοι βελτιστοποίησης σε ένα WFST παίζει σημαντικό ρόλο. Για παράδειγμα, στο Σχήμα 3.4vi προκύπτει WFST μη-ντετερμινιστικό, παρόλο που έχει προηγηθεί διαδικασία μετατροπής του WFST του Σχήματος 3.4i σε ντετερμινιστικό. Παρατηρείται, ακόμα, ότι κατά την ελαχιστοποίηση του WFST του Σχήματος 3.4ii στο WFST του Σχήματος 3.4vii γίνεται και σπρώξιμο βαρών, όπως αναμενόταν σύμφωνα με τον αλγόριθμο που περιγράφηκε.

3.4 WFSTs και Αναγνώριση Φωνής

Τα WFSTs προσφέρουν ένα ενοποιητικό πλαίσιο εργασίας, τόσο για το γλωσσικό και ακουστικό μοντέλο, όσο και για άλλες πηγές γνώσεις σε ένα σύστημα αναγνώρισης, όπως το λεξικό που δίνει τις προφορές των λέξεων (pronunciation lexicon). Υπό την έννοια αυτή, εν τέλει έχουμε στη διάθεσή μας ένα μοναδικό στατικό WFST που έχει προκύψει από τη σύνθεση επιμέρους WFSTs και που απευθείας δημιουργεί ένα δίκτυο αναζήτησης για τη μετατροπή μίας ακολουθία ακουστικών χαρακτηριστικών, ή παρατηρήσεων όπως είχαν χαρακτηριστεί στο πλαίσιο των HMMs, σε μία πρόταση λέξεων. Με αυτό τον τρόπο επιτυγχάνεται ταχύτερη αποκωδικοποίηση, αλλά και δυνατότητα απομάκρυνσης των διαφόρων πλεονασμών μέσω τεχνικών βελτιστοποίησης, όπως περιγράφηκαν στην Ενότητα 3.3.

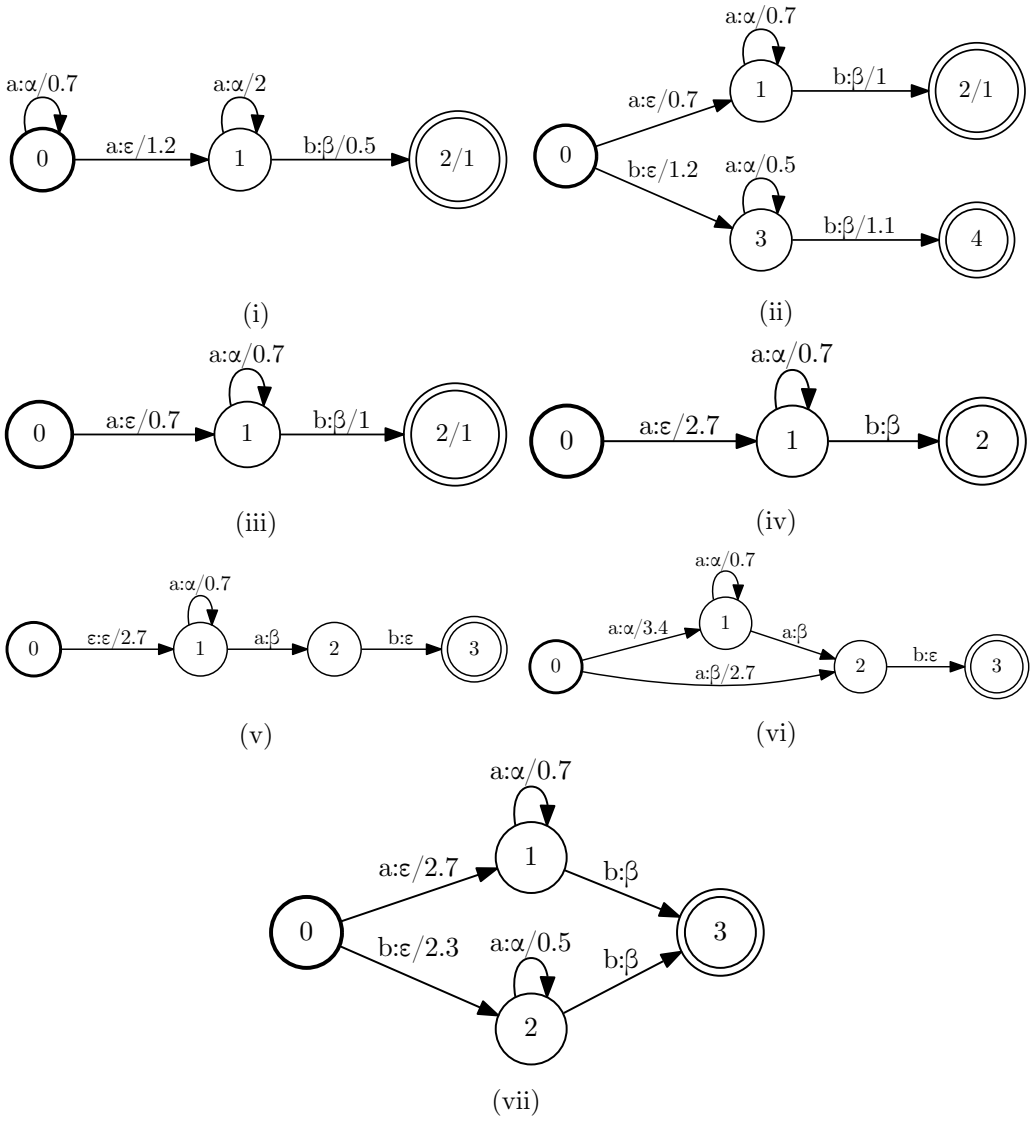
Αγνοώντας για απλότητα τις παραμέτρους LMSF και WIP, όπως εισήχθησαν στην Ενότητα 2.4, η πιθανοτική αναπαράσταση ενός μοντέλου για αναγνώριση φωνής δίνεται από τη σχέση (2.3). Στην πράξη, πολλές φορές το λεξικό δεν περιέχει μόνο μια αντιστοίχιση μεταξύ λέξεων και προφορών, δηλαδή αντιστοίχιση μεταξύ ακολουθίας γραμμάτων και ακολουθίας φωνημάτων για το σχηματισμό μιας λέξης, αλλά περιέχει επίσης και την πληροφορία πόσο πιθανή είναι η εμφάνιση μιας ακολουθίας φωνημάτων V δεδομένης της λέξης W [13]. Συνεπώς, η σχέση (2.3) μετατρέπεται στη σχέση

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmax}} \left\{ \sum_{V \in R(W)} p(O|V,W) P(V|W) P(W) \right\} \tag{3.14}$$

$$\approx \underset{W \in \mathcal{W}}{\operatorname{argmax}} \left\{ \sum_{V \in R(W)} p(O|V) P(V|W) P(W) \right\} . \tag{3.15}$$

, όπου $\mathcal{W}$ είναι το σύνολο των πιθανών ακολουθιών λέξεων και $R(W)$ το σύνολο των πιθανών ακολουθιών φωνημάτων για τη λέξη W.

Όπως έχει εξηγηθεί, για την αποκωδικοποίηση χρησιμοποιείται ο αλγόριθμος Viterbi, ο οποίος δίνει πάντα μια υποεκτίμηση, αλλά είναι υπολογιστικά βατός και στηρίζεται στην αντικατάσταση του αθροίσματος της τελευταίας σχέσης σε συνάρτηση μεγίστου. Έτσι, πρακτικά



Σχήμα 3.4: Λειτουργίες βελτιστοποίησης στα WFSTs. (i) Μη ντετερμινιστικό WFST. (ii) Μη ελάχιστο WFST. (iii) Ντετερμινιστικό WFST, ισοδύναμο του (i). (iv) Ισοδύναμο WFST του (iii) μετά από σπρώξιμο βαρών. (v) Ισοδύναμο WFST του (iv) μετά από συγχρονισμό. (vi) Ισοδύναμο WFST του (v) μετά από ϵ-απομάκρυνση. (vii) Ελάχιστο WFST, ισοδύναμο του (ii).

καταλήγουμε στη σχέση

$$\hat{W} \approx \underset{W \in \mathcal{W}}{\operatorname{argmax}} \left\{ \max_{V \in R(W)} p(O|V) P(V|W) P(W) \right\} \tag{3.16}$$

και, περνώντας στο λογαριθμικό πεδίο, στη σχέση

$$\hat{W} \approx \underset{W \in \mathcal{W}}{\operatorname{argmax}} \left\{ \max_{V \in R(W)} \{ \log p(O|V) + \log P(V|W) + \log P(W) \} \right\} . \tag{3.17}$$

Στη γλώσσα των WFSTs έχουμε το WFST H που μετατρέπει μια ακολουθία ακουστικών χαρακτηριστικών O σε μια ακολουθία φωνημάτων V με κόστος $w_H(O \to V) =$

$-\log P(O|V)$, το WFST L που μετατρέπει μια ακολουθία φωνημάτων V σε μια ακολουθία λέξεων W με κόστος $w_L(V \to W) = -\log P(V|W)$ και το WFSA G που αποδέχεται μια ακολουθία λέξεων W με κόστος $w_G(W) = -\log P(W)$. Αυτές οι μηχανές συντίθενται σε ένα μοναδικό WFST N ως

$$N = H \circ L \circ G\,. \tag{3.18}$$

Συνεπώς, το πρόβλημα της αναγνώρισης φωνής μετασχηματίζεται στο πρόβλημα εύρεσης του συντομότερου μονοπατιού πάνω στο WFST δοθείσας της ακολουθίας O:

$$\hat{W} \approx \underset{W \in \mathcal{W}}{\operatorname{argmin}} \left\{ \min_{V \in R(W)} \{(-\log p(O|V)) + (-\log P(V|W)) + (-\log P(W))\} \right\} \tag{3.19}$$

$$= \underset{W \in \mathcal{W}}{\operatorname{argmin}} \left\{ \min_{V \in R(W)} \{w_H(O \to V) \otimes w_L(V \to W) \otimes w_G(W)\} \right\} \tag{3.20}$$

$$= \underset{W \in \mathcal{W}}{\operatorname{argmin}}\, w_N(O \to W)\,, \tag{3.21}$$

σύμφωνα με τις πράξεις που ορίζονται στον τροπικό ημιδακτύλιο.

Για τη διαχείριση τριφωνικών μοντέλων, είναι απαραίτητο ένα ακόμα WFST C, το οποίο μετατρέπει μια ακολουθία τριφωνημάτων σε ακολουθία φωνημάτων, όπου το κάθε φώνημα είναι ανεξάρτητο των συμφραζομένων και ταυτίζεται με το κεντρικό του αντίστοιχου τριφωνήματος. Μετά την προσθήκη αυτή, το τελικό WFST μπορεί να γραφεί ως

$$N = H \circ C \circ L \circ G\,. \tag{3.22}$$

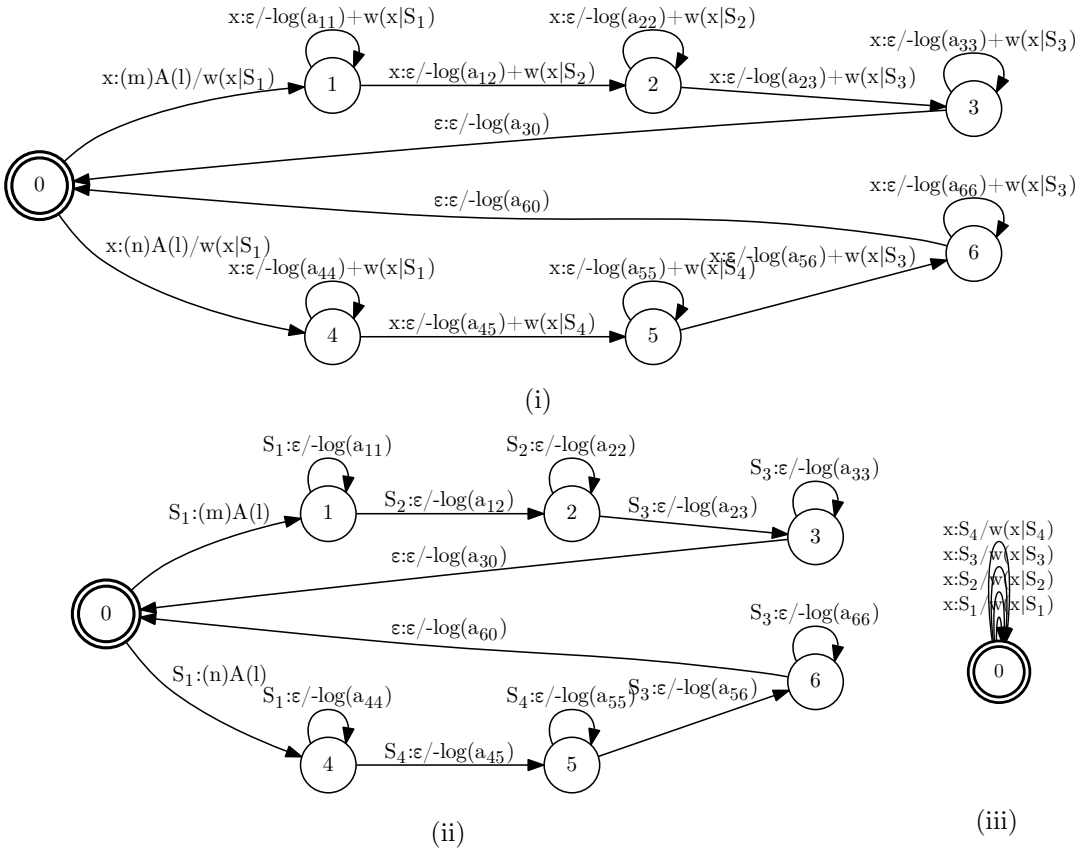
3.4.1 Κατασκευή των Επιμέρους Συνιστωσών

Ο μετατροπέας H μετατρέπει μια ακολουθία ακουστικών χαρακτηριστικών ή παρατηρήσεων O σε μια ακολουθία φωνημάτων ή γενικότερα μια ακολουθία SUs, όποια και αν έχουν επιλεγεί. Εφόσον συνήθως τα SUs που χρησιμοποιούνται στην πράξη είναι τα τριφωνήματα, αυτά είναι που θα θεωρήσουμε και στην πορεία. Το H, άρα, μπορεί να θεωρηθεί ως το σύνολο όλων των στοιχειωδών HMMs που μοντελοποιούν όλα τα τριφωνήματα, τα οποία ενοποιούνται σε ένα κοινό WFST μέσω των ρητών πράξεων της ένωσης και της κλειστότητας κατά Kleene [52].

Ωστόσο, οι πιθανές παρατηρήσεις σε κάθε κατάσταση των HMMs είναι διανύσματα που δεν έχουν υποστεί κάποια διακριτοποίηση. Συνεπώς, τα HMMs δεν μπορούν απευθείας να παρασταθούν ως FAs, αφού ο ορισμός των τελευταίων απαιτεί την ύπαρξη πεπερασμένου αλφαβήτου, τόσο για τα σύμβολα εισόδου, όσο και για τα σύμβολα εξόδου (στην περίπτωση των μετατροπέων). Για το λόγο αυτό, η πληροφορία που φέρει το σύνολο των HMMs μπορεί να διασπαστεί, ώστε να ληφθεί ένα WFST που φέρει την πληροφορία της τοπολογίας των HMMs και ένας μετατροπέας, που ξεφεύγει από τα πλαίσια ορισμού των WFSTs, που μοντελοποιεί την πιθανότητα ακουστικού ταιριάσματος. Για να γίνει η διαδικασία αντιληπτή, θα χρησιμοποιηθεί ένα παράδειγμα που σκιαγραφεί μία δυνατή τέτοια διάσπαση [13].

Έστω x ένα μετασύμβολο που παριστάνει οποιοδήποτε δυνατό διάνυσμα χαρακτηριστικών, δηλαδή οποιαδήποτε δυνατή παρατήρηση σε ένα HMM. Έστω, ακόμα, ότι όλα τα τριφωνήματα μοντελοποιούνται με left-right HMMs τριών καταστάσεων, όπου κάθε κατάσταση συμβολίζεται με S_i, $i = 1, 2, \cdots$. Όλες οι καταστάσεις S_i λαμβάνονται από ένα καθορισμένο, πεπερασμένο σύνολο δεμένων καταστάσεων, όπως αυτές προκύπτουν από τη διαδικασία που περιγράφηκε στην Υποενότητα 2.2.4. Η ένωση δύο τέτοιων HMMs με συνολικά 4 δεμένες καταστάσεις και το κλείσιμο της προκύπτουσας ένωσης φαίνονται στο Σχήμα 3.5i. Τόσο

οι πιθανότητες μετάβασης a_{ij}, όσο και οι πιθανότητες $b_{S_i}(x)$ εξαγωγής μιας παρατήρησης x από μια κατάσταση S_i, προκύπτουν από την εκπαίδευση του ακουστικού μοντέλου, όπως αναλύεται στην Υποενότητα 2.2.3. Κάνοντας χρήση τροπικού (ή λογαριθμικού) ημιδακτυλίου, λαμβάνονται ως βάρη των μεταβάσεων οι αρνητικοί λογάριθμοι των αντίστοιχων πιθανοτήτων, ώστε $w(x|S_i) = -\log b_{S_i}(x)$.

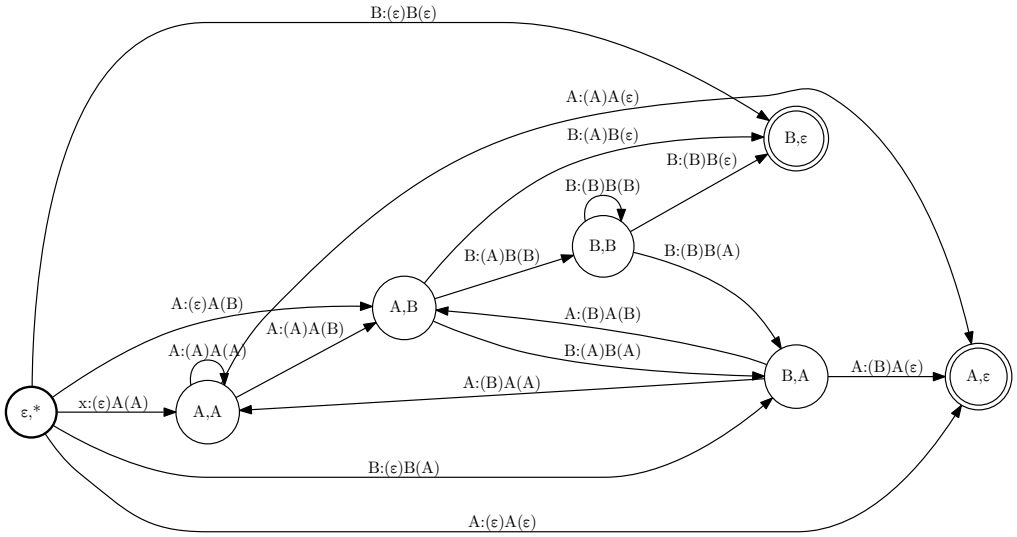


Σχήμα 3.5: Διάσπαση μετατροπέα από ακουστικές παρατηρήσεις σε τριφωνήματα σε δύο επιμέρους μετατροπείς. (i) Μετατροπέας που μετατρέπει μια ακολουθία ακουστικών παρατηρήσεων σε ακολουθία τριφωνημάτων. Πρόκειται για την κλειστότητα κατά Kleene της ένωσης των HMMs που μοντελοποιούν τα τριφωνήματα $(m)A(l)$ και $(n)A(l)$. (ii) WFST που μετατρέπει μια ακολουθία καταστάσεων HMMs σε ακολουθία τριφωνημάτων. (iii) Μετατροπέας που μετατρέπει μια ακολουθία ακουστικών χαρακτηριστικών σε ακολουθία καταστάσεων HMMs.

Ο μετατροπέας του Σχήματος 3.5i μπορεί να διασπαστεί στο WFST του Σχήματος 3.5ii και στο μετατροπέα του Σχήματος 3.5iii. Το πρώτο είναι ουσιαστικά το H που χρησιμοποιείται για την κατασκευή του WFST της σχέσης (3.22) και χρησιμοποιείται για τη μετατροπή μιας ακολουθίας καταστάσεων HMMs σε μια ακολουθία τριφωνημάτων (ή όποιων SUs εξαρτώμενων από τα συμφραζόμενα χρησιμοποιούνται). Ο δεύτερος, που μετατρέπει μια ακολουθία ακουστικών παρατηρήσεων σε μια ακολουθία καταστάσεων HMMs, χρησιμοποιείται απευθείας στη φάση της αποκωδικοποίησης. Σημειώνεται πως στην πράξη, για να είναι το N πιο αποδοτικό από άποψη απαιτούμενου χώρου, το H δεν περιέχει self-loops. Αυτά προσομοιώνονται κατά τη φάση της αποκωδικοποίησης [52].

Όσον αφορά στο WFST C που μετατρέπει μια ακολουθία τριφωνημάτων (ή γενικότερα SUs εξαρτώμενων από τα συμφραζόμενα) σε μια ακολουθία φωνημάτων (ή γενικότερα SUs

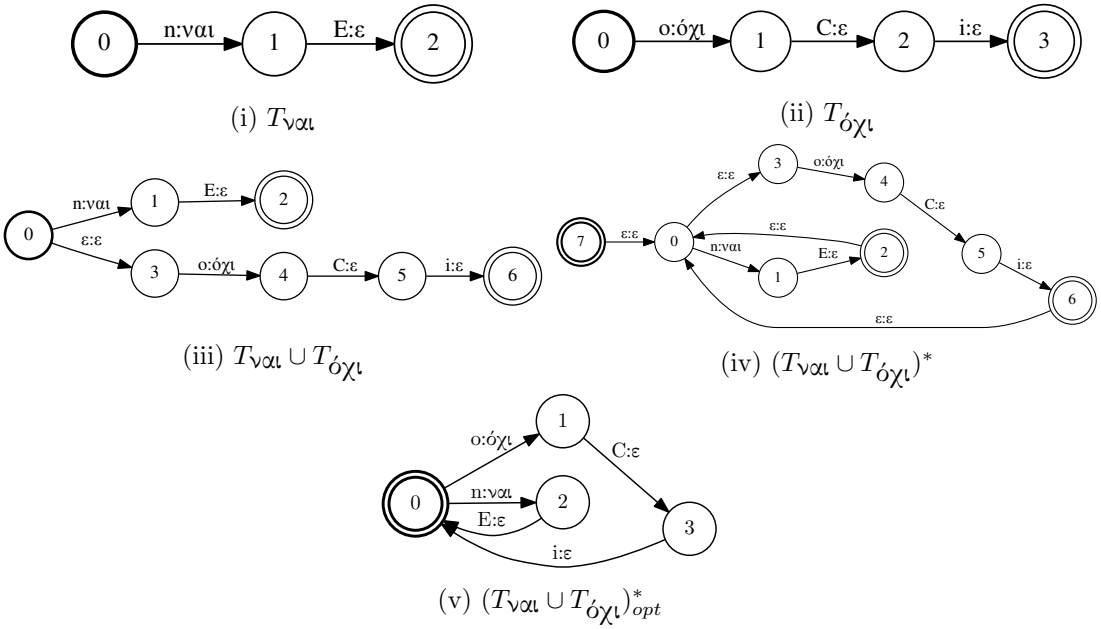
ανεξάρτητων από τα συμφραζόμενα), ας θεωρήσουμε ένα απλό παράδειγμα όπου υπάρχουν μόνο δύο φωνήματα, τα A και B. Είναι διαισθητικά πιο εύκολα αντιληπτό το WFST C' που μετατρέπει μια ακολουθία φωνημάτων σε ακολουθία τριφωνημάτων. Το C είναι απλά το αντίστροφο του C' [55, 52]. Έτσι, λοιπόν, το C' μετατρέπει, για παράδειγμα, την ακολουθία $A/B/A$ στην ακολουθία $(\epsilon)A(B)/(A)B(A)/(B)A(\epsilon)$. Λαμβάνοντας υπόψιν όλους τους πιθανούς συνδυασμούς, το σχετικό WFST απεικονίζεται στο Σχήμα 3.6. Επειδή υπάρχει μεγάλος αριθμός τριφωνημάτων τα οποία δε συναντώνται ποτέ, δεν υπάρχει λόγος να χρησιμοποιηθεί ολόκληρο το δίκτυο C. Αντ' αυτού, το C μπορεί να δημιουργθεί στη μνήμη τη στιγμή που χρειάζεται, ενσωματώνοντας στο N μόνο τα τμήματά του που πραγματικά χρειάζονται.



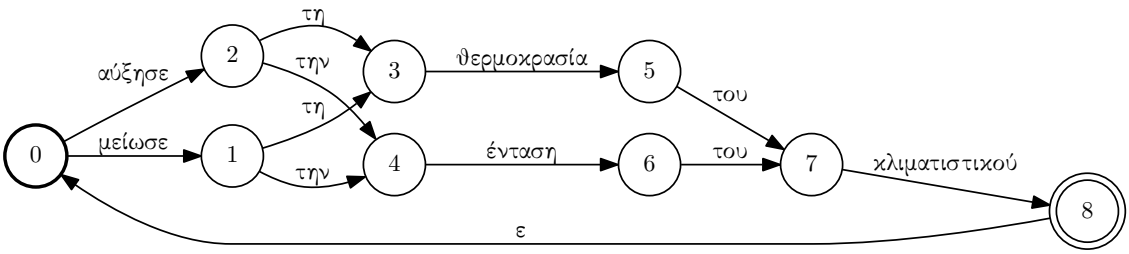
Σχήμα 3.6: WFST που μετατρέπει μια ακολουθία φωνημάτων ανεξάρτητων από τα συμφραζόμενα σε ακολουθία τριφωνημάτων. Θεωρείται πως υπάρχουν μόνο τα υποθετικά φωνήματα A και B.

Το φωνητικό λεξικό, τώρα, L, για τη μετατροπή μιας ακολουθίας φωνημάτων σε μια ακολουθία λέξεων, δημιουργείται ως εξής. Πρώτα, παράγεται ένα FST για κάθε εγγραφή του λεξικού που χρησιμοποιείται. Στη γενική περίπτωση όπου το λεξικό περιέχει και την πιθανοτική πληροφορία κάθε προφοράς, δηλαδή με ποια πιθανότητα κάθε λέξη εκφέρεται με μια συγκεκριμένη προφορά, τότε παράγονται WFSTs. Εν συνεχεία, λαμβάνεται η ένωση όλων αυτών των WFSTs και η κλειστότητα κατά Kleene της ένωσης [52]. Για παράδειγμα, θεωρώντας το στοιχειώδες λεξικό της σελίδας 46, η διαδικασία παραγωγής του L σκιαγραφείται στο Σχήμα 3.7. Η επέκταση για δυνατότητα παύσης ανάμεσα στις διαδοχικά εκφερόμενες λέξεις είναι άμεση, προσθέτοντας μία μετάβαση από κάθε τελική κατάσταση προς την αρχική με επισημείωση $sil : \epsilon$.

Τέλος, το WFSA G, που αποδέχεται μια ακολουθία λέξεων με κάποιο κόστος, δεν είναι παρά το γλωσσικό μοντέλο, όπως αυτό έχει παρουσιαστεί στην Ενότητα 2.3. Στην περίπτωση n-gram μοντέλου, που είναι και η συνηθέστερη περίπτωση, η αναπαράσταση με WFSA είναι άμεση, εφόσον πρόκειται για ένα Μαρκοβιανό μοντέλο τάξης $n-1$ [13]. Η αναπαράσταση μιας FSG ως FSA είναι επίσης άμεση, με τη μόνη τροποποίηση να είναι πως οι λέξεις της γραμματικής θα πρέπει να επισημειώνουν τις ακμές των μεταβάσεων και όχι τις καταστάσεις. Για παράδειγμα, η FSG του Σχήματος 2.4 μπορεί να παρασταθεί ως FSA όπως φαίνεται στο Σχήμα 3.8.



Σχήμα 3.7: Διαδικασία κατασκευής του WFST που μοντελοποιεί το φωνητικό λεξικό L. (i) FST για την προφορά της λέξης *ναι*. (ii) FST για την προφορά της λέξης *όχι*. (iii) Ένωση των $T_{\text{ναι}}$ και $T_{\text{όχι}}$. (iv) Κλειστότητα κατά Kleene της ένωσης των $T_{\text{ναι}}$ και $T_{\text{όχι}}$. (v) Ισοδύναμο WFST του (iv) μετά από ϵ-απομάκρυνση και ελαχιστοποίηση.



Σχήμα 3.8: Παράδειγμα FSG εκφρασμένης ως FSA. Πρόκειται για την FSG του Σχήματος 2.4.

3.4.2 Σύνθεση και Βελτιτοποίηση

Έχοντας κατασκευάσει τα απαραίτητα WFSTs H, C, L και G, αυτά μπορούν να συντεθούν σε ένα ενιαίο WFST N, σύμφωνα με τη σχέση (3.22). Ωστόσο, για να μην καταλήξει το N να είναι υπερβολικά μεγάλο, είναι απαραίτητο να προηγηθούν σταδιακές λειτουργίες βελτιστοποίησης, όπως αυτές που έχουν παρουσιαστεί στην Ενότητα 3.3. Η ανάγκη αυτή γίνεται εμφανής από το WFST του Σχήματος 3.7v, που είναι ισοδύναμο με αυτό του Σχήματος 3.7iv που μοντελοποιεί ένα στοιχειώδες φωνητικό λεξικό, αλλά αφότου έχουν λάβει χώρα κάποιες πράξεις βελτιστοποίησης.

Πρώτα απ' όλα, πρέπει να διασφαλιστεί ότι σε κάθε ενδιάμεσο βήμα παράγεται ένα ντετερμινιστικό WFST [52]. Θεωρώντας το G ντετερμινιστικό[3], πρέπει πρώτα να διασφαλιστεί η ντετερμινιστική φύση του $L \circ G$. Η ιδιότητα αυτή συχνά δεν ισχύει λόγω της ύπαρξης ομό-

[3]Πιθανά ες στο G αντιμετωπίζονται ως κανονικά σύμβολα. Εάν και πάλι το G δεν είναι ντετερμινιστικό, προηγείται η διαδικασία μετατροπής του σε ντετερμινιστικό.

ηχων λέξεων. Για το λόγο αυτό, το λεξικό επεκτείνεται ώστε στο τέλος κάθε ακολουθίας φωνημάτων να υπάρχει ένα ειδικό σύμβολο (για παράδειγμα, #1, #2, κ.λπ.). Ακόμα, όμως, και για τις ακολουθίες φωνημάτων που προσδιορίζουν μοναδικά μια λέξη, δηλαδή για λέξεις που δεν έχουν ομόηχες, είναι καλή πρακτική να τοποθετείται στο τέλος ένα ειδικό σύμβολο (για παράδειγμα το #1), διότι έτσι αποφεύγεται ο κίνδυνος σφαλμάτων σε περιπτώσεις ομόηχων φράσεων που αποτελούνται από μη ομόηχες λέξεις [13]. Για παράδειγμα, η λέξη *ταξίδια* δεν μπορεί να διαχωριστεί με βάση την προφορά της από τις λέξεις *τα ξίδια* παρά μόνο εάν εισαχθούν τα εν λόγω βοηθητικά σύμβολα. Έτσι, στη μία περίπτωση έχουμε την ακολουθία φωνημάτων {t A k s i D J A #1}, ενώ στην άλλη την ακολουθία {t A #1 k s i D J A #1}. Κατ' αυτόν τον τρόπο, παράγεται το τροποποιημένο WFST $\tilde{L}$, το οποίο συντίθεται με το G και το αποτέλεσμα της σύνθεσης μπορεί πλέον να μετατραπεί σε ντετερμινιστικό WFST, το οποίο συμβολίζεται ως LG:

$$LG = \det(\tilde{L} \circ G) \tag{3.23}$$

Με την παραπάνω τροποποίηση, όμως, του L, δημιουργούνται μονοπάτια που περιέχουν σύμβολα τα οποία δεν παράγει το C ως έχει. Συνεπώς, θα πρέπει να μετατραπεί ανάλογα και το C στο WFST $\tilde{C}$, με την προσθήκη self-loop μεταβάσεων σε κάθε κατάσταση του C[4], ώστε κάθε βοηθητικό σύμβολο να μπορεί να προστεθεί σε κάθε προφορά κάθε λέξης [52]. Εάν P είναι ο μέγιστος βαθμός ομοηχίας, δηλαδή ο μέγιστος αριθμός λέξεων που μπορούν να εκφραστούν με την ίδια προφορά, τότε σε κάθε κατάσταση πρέπει να εισαχθούν P self-loops με τις επισημειώσεις $\#1 : \#1, \#2 : \#2, \cdots \#P : \#P$. Το παραγόμενο WFST $\tilde{C}$ μπορεί τώρα να συντεθεί με το LG και να βρεθεί το ντετερμινιστικό ισοδύναμο του αποτελέσματος:

$$CLG = \det(\tilde{C} \circ LG) \tag{3.24}$$

Με όμοιο σκεπτικό με προηγουμένως, προστίθενται P self-loops στην αρχική κατάσταση του H, όπου η αρχική κατάσταση θεωρείται πως βρίσκεται στο σύνορο μεταξύ δύο οποιωνδήποτε SUs, όπως στο Σχήμα 3.5ii. Έτσι, το επόμενο βήμα είναι η σύνθεση του $\tilde{H}$ που μόλις δημιουργήθηκε με το CLG και η εκ νέου εύρεση ενός ντετερμινιστικού ισοδύναμου WFST:

$$HCLG = \det(\tilde{H} \circ CLG) \tag{3.25}$$

Στη συνέχεια, το $HCLG$ ελαχιστοποιείται και όλα τα βοηθητικά σύμβολα που έχουν εισαχθεί αντικαθίστανται με το κενό σύμβολο ϵ, με την αντίστοιχη πράξη να συμβολίζεται ως π_ϵ, ώστε να ληφθεί το WFST

$$N = \pi_\epsilon(\min(HCLG)) = \pi_\epsilon(\min(\det(\tilde{H} \circ \det(\tilde{C} \circ \det(\tilde{L} \circ G))))) \,. \tag{3.26}$$

Πρώτο βήμα της ελαχιστοποίησης, όπως έχουμε δει, είναι το σπρώξιμο των βαρών. Η επιλογή του ημιδακτυλίου στον οποίο θα λάβει χώρα η διαδικασία αυτή παίζει σημαντικό ρόλο, με τη χρήση του λογαριθμικού ημιδακτυλίου να φαίνεται να παρουσιάζει σημαντικά πλεονεκτήματα έναντι του τροπικού [52]. Και στις δύο περιπτώσεις το αποτέλεσμα της ελαχιστοποίησης ως προς τον αριθμό των καταστάσεων και των μεταβάσεων είναι ακριβώς το ίδιο, με τη διαφορά να έγκειται στη διαφορετική κατανομή των βαρών. Έχει παρατηρηθεί ότι η κατανομή που προκύπτει από το σπρώξιμο των βαρών στο λογαριθμικό ημιδακτύλιο κάνει το κλάδεμα (pruning) κατά την αποκωδικοποίηση με Viterbi beam search πιο αποδοτικό.

[4] Στην πράξη, στο σημείο αυτό αντί του C χρησιμοποιείται το $(\det(C^{-1}))^{-1}$ [13], επειδή το C δεν είναι ντετερμινιστικό, αλλά και ούτε μπορεί άμεσα να βρεθεί ντετερμινιστικό ισοδύναμό του.

Η βασική διαφορά όσον αφορά στον αλγόριθμο για την υλοποίηση του σπρωξίματος βαρών σχετίζεται με το δυναμικό $V(q)$ που αποδίδεται σε κάθε κατάσταση q του WFST, σύμφωνα με τη σχέση (3.10). Συγκεκριμένα, όταν χρησιμοποιείται τροπικός ημιδακτύλιος προκύπτει

$$V(q) = \min_{\pi \in \Pi(q,F)} \{w(\pi) + \rho(n(\pi))\}, \tag{3.27}$$

που μπορεί να υπολογιστεί με έναν κλασικό αλγόριθμο εύρεσης συντομότερων μονοπατιών σε γράφο. Αντίθετα, με χρήση λογαριθμικού ημιδακτυλίου προκύπτει

$$V(q) = -\log \sum_{\pi \in \Pi(q,F)} \left\{ e^{w(\pi)+\rho(n(\pi))} \right\}, \tag{3.28}$$

που ισούται με τον αρνητικό λογάριθμο της συνολικής πιθανότητας όλων των μονοπατιών από την κατάσταση q σε κάποια τελική κατάσταση, εφόσον τα εκάστοτε βάρη προκύπτουν ως αρνητικοί λογάριθμοι πιθανοτήτων. Η χρήση αυτής της συνάρτησης δυναμικού εξασφαλίζει ότι το παραγόμενο WFST θα διατηρεί τη συνήθη κανονικοποίηση των HMMs, όπου το άθροισμα των "βαρών" όλων των εξερχόμενων μεταβάσεων από οποιαδήποτε κατάσταση ισούται με 1.

Τέλος, αφού έχουν γίνει όλα τα παραπάνω βήματα, μπορεί να ακολουθήσει μια διαδικασία γνωστή ως παραγοντοποίηση (factoring) [52]. Κατά το factoring, κάθε αλυσιδωτή μετάβαση αντικαθίσταται με μία μόνο μετάβαση, η οποία είναι επισημειωμένη με την παράθεση όλων των συμβόλων των επιμέρους μεταβάσεων της αλυσίδας, ενώ το βάρος που της αποδίδεται ισούται με το γινόμενο[5] όλων των επιμέρους βαρών. Ως αλυσίδα ορίζεται ένα μονοπάτι του οποίου όλες οι καταστάσεις, εκτός της πρώτης και της τελευταίας, έχουν μία εισερχόμενη και μία εξερχόμενη μετάβαση.

Εφόσον το $\tilde{H}$ δεν περιέχει, όπως έχει αναφερθεί, self-loops, αλλά αυτά προσομοιώνονται κατά τη φάση της αποκωδικοποίησης, είναι αναμενόμενο ότι θα υπάρχουν αρκετές αλυσίδες στο τελικό WFST N που μετατρέπει μια ακολουθία καταστάσεων HMMs σε ακολουθία λέξεων. Από αυτές, όμως, δεν αντικαθιστώνται απαραίτητα όλες, αλλά γίνονται μόνο οι αντικαταστάσεις που θα οδηγήσουν σε μείωση του μεγέθους του μετατροπέα. Εξάλλου, το factoring δεν επηρεάζει το χρόνο αναγνώρισης, αλλά μόνο το μέγεθος του τελικού μετατροπέα.

Η απόφαση σχετικά με το εάν μια αλυσιδωτή μετάβαση θα συγχωνευτεί σε μία μπορεί να γίνει με χρήση της συνάρτησης κέρδους (gain). Έστω, λοιπόν, μια ακολουθία καταστάσεων HMMs $\sigma = S_k, S_{k+1}, \cdots$ και έστω $C(N)$ το σύνολο των αλυσίδων στο WFST . Τότε, συμβολίζοντας ως $l_i(\pi)$ και $l_o(\pi)$ τις συμβολοακολουθίες εισόδου και εξόδου, αντίστοιχα, ενός μονοπατιού π, το κέρδος $G(\sigma)$ της ακολουθίας σ ορίζεται ως

$$G(\sigma) = \sum_{\substack{\pi \in C(N) \\ l_i(\pi) = \sigma}} \{|\sigma| - |l_o(\pi)| - 1\}, \tag{3.29}$$

όπου $|x|$ το μήκος της συμβολοακολουθίας x. Η αντικατάσταση της ακολουθίας σ κατά το factoring οδηγεί σε μείωση του μεγέθους του μετατροπέα μόνο εάν $G(\sigma) > 0$.

Κατά το factoring, λοιπόν, ακολουθίες συμβόλων εισόδου που αντιστοιχούν στις ταυτότητες HMM καταστάσεων αντικαθίστανται από μία μοναδική συμβολοσειρά εισόδου που αντιστοιχεί στην ταυτότητα ενός HMM n καταστάσεων, όπου n ο αριθμός μεταβάσεων στην

[5] Για τροπικό ή λογαριθμικό ημιδακτύλιο το "γινόμενο" ταυτίζεται με την αλγεβρική πράξη της πρόσθεσης.

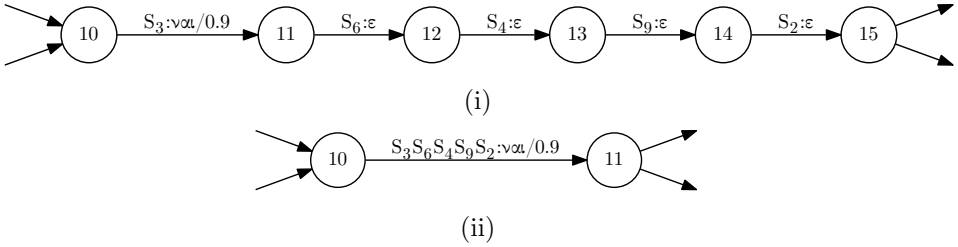
αλυσίδα προς αντικατάσταση. Ένα σχετικό παράδειγμα απεικονίζεται στο Σχήμα 3.9. Συνεπώς, η διαδικασία του factoring έχει ως αποτέλεσμα τη διάσπαση του N ως εξής:

$$N = H' \circ F\,, \tag{3.30}$$

όπου, συμβολίζοντας ως fact$(\cdot)$ την πράξη του factoring,

$$F = \text{fact}(N) = \text{fact}(\pi_\epsilon(\min(\det(\tilde{H} \circ \det(\tilde{C} \circ \det(\tilde{L} \circ G)))))) \tag{3.31}$$

και H' ένας μετατροπέας που αντιστοιχίζει ακολουθίες n HMM καταστάσεων σε HMMs n καταστάσεων. Το WFST που ουσιαστικά χρησιμοποιείται για την αναγνώριση είναι το F, εφόσον το H' μπορεί να προσομοιωθεί απευθείας στη φάση της αποκωδικοποίησης. Σημειώνεται ότι εναλλακτικά, μπορεί στην πράξη το factoring να προηγηθεί της ελαχιστοποίησης [52].



Σχήμα 3.9: Παράδειγμα παραγοντοποίησης σε τμήμα WFST. (i) Αλυσίδα 5 μεταβάσεων. (ii) Αντικατάσταση των αλυσιδωτών μεταβάσεων από μία μόνο μετάβαση. Στο παράδειγμα αυτό, το αντίστοιχο τμήμα του μετατροπέα H' της σχέσης (3.30) θα έπρεπε να αντιστοιχίζει την ακολουθία S_3, S_6, S_4, S_9, S_2 στη συμβολοσειρά $S_3S_6S_4S_9S_2$.

Κεφάλαιο 4

Χτίζοντας το Βασικό Σύστημα Αναγνώρισης

4.1 Κινητήριες Ιδέες

Όπως έχουμε δει στην Ενότητα 2.3, ένα αξιόπιστο και ακριβές γλωσσικό μοντέλο έχει ως αποτέλεσμα τη μείωση της μετρικής του perplexity. Το perplexity, όμως, είναι άμεσα συσχετισμένο με το WER [56]. Όπως φαίνεται, για παράδειγμα, στο [56], αύξηση της τάξης του n-gram μοντέλου κατά 2 έχει ως αποτέλεσμα μία σχετική μείωση του WER της τάξης του 10%. Τέτοιου είδους παρατηρήσεις έχουν ωθήσει την επιστημονική κοινότητα τα τελευταία χρόνια να αφιερώνει σημαντική ερευνητική δραστηριότητα στις δυνατότητες βελτίωσης του γλωσσικού μοντέλου για την όσο το δυνατό μεγαλύτερη αξιοποίηση των συμφραζομένων (context). Ακόμα, τα σύγχρονα συστήματα αναγνώρισης τα οποία βασίζονται σε τεράστια σύνολα δεδομένων, που προέρχονται από τους χρήστες του διαδικτύου, χρησιμοποιούν εξαιρετικά μεγάλα ποσά υπολογιστικών πόρων με στόχο τη στατιστική μοντελοποίηση της γλώσσας με μοντέλα όλο και μεγαλύτερης τάξης.

Παρόλο που η προσέγγιση αυτή αναμφίβολα έχει θετικά αποτελέσματα ως προς την απόδοση των σύγχρονων συστημάτων αναγνώρισης, μας κάνει ορισμένες φορές να ξεχνάμε το σημαντικότατο ρόλο που συνεχίζει να παίζει ένα αξιόπιστο ακουστικό μοντέλο, αλλά και η επιλογή των κατάλληλων ακουστικών χαρακτηριστικών. Η επικράτηση στατιστικών μεθόδων στην αναγνώριση φωνής πηγάζει κατά ένα μεγάλο βαθμό από τις ανεπαρκείς ή και λάθος γνώσεις που έχουμε όσον αφορά στον τρόπο με τον οποίο ο άνθρωπος αντιλαμβάνεται και αναγνωρίζει τους ήχους και συγκεκριμένα την ομιλία [36], με αποτέλεσμα πρακτικές που στηρίζονται σε χειροκίνητους κανόνες να είναι καταδικασμένες να αποτύχουν. Φαίνεται, ωστόσο, πως ένας συνδυασμός της στατιστικής μοντελοποίησης, σε επίπεδο για παράδειγμα γλωσσικού μοντέλου, και μιας προσπάθειας αξιοποίησης των όποιων γνώσεων έχουμε για τη Αναγνώριση Φωνής από τον Άνθρωπο (Human Speech Recognition - HSR) είναι προς τη σωστή κατεύθυνση.

Η επίδραση του περιεχομένου παίζει πράγματι πολύ σημαντικό ρόλο και στο HSR, εφόσον συχνά η σχετική πληροφορία χρησιμοποιείται από τον άνθρωπο για να επιλύσει τις όποιες αμφισημίες [57]. Ωστόσο, ακόμα και με χρήση ίδιου context και προσομοίωση ίδιων πειραματικών συνθηκών, φαίνεται πως το ASR εξακολουθεί να παρουσιάζει σημαντικά χαμηλότερες αποδόσεις σε σύγκριση με το HSR, με τα αντίστοιχα ποσοστά να διαφέρουν μέχρι και μία τάξη μεγέθους, ενώ παράλληλα, η συσχέτιση του perplexity με το WER φαίνεται να ακολουθεί την ίδια συμπεριφορά τόσο στο HSR όσο και στο ASR [58].

Θα πρέπει στο σημείο αυτό να σημειωθεί ότι το context ορίζεται σε πολλά διαφορετικά επίπεδα και μπορεί να επηρεάσει ποικιλοτρόπως την αναγνώριση [57]. Ως πληροφορία περιεχομένου (contextual information) ορίζεται η επίδραση που έχει το περιεχόμενο στην εντροπία της εργασίας της αναγνώρισης. Υπό αυτό το πρίσμα, η εντροπία των λέξεων είναι χαμηλότερη από την εντροπία των συλλαβών στο συνεχή λόγο, η οποία με τη σειρά της είναι χαμηλότερη από την εντροπία συλλαβών που δεν έχουν κάποια νοηματική υπόσταση ή από την εντροπία φωνημάτων. Με άλλα λόγια, για την ίδια παραμόρφωση που εισάγεται μέσω του καναλιού επικοινωνίας, η πληροφορία του περιεχομένου έχει ως αποτέλεσμα η αναγνώριση λέξεων να παρουσιάζει μεγαλύτερα ποσοστά επιτυχίας από την αναγνώριση συλλαβών ή φωνημάτων.

Από τα παραπάνω, λοιπόν, διαφαίνεται πως η πρόκληση που ανακύπτει για την Αυτόματη Αναγνώριση Φωνής είναι η επίτευξη χαμηλών ποσοστών λάθους κατά την αναγνώριση φωνημάτων, που αποτελούν άλλωστε και το βασικό στοιχείο αναγνώρισης στα περισσότερα σύγχρονα συστήματα. Η επίτευξη του στόχου αυτού θα είχε προφανείς άμεσες και βαθιές συνέπειες στα ποσοστά αναγνώρισης λέξεων σε συνεχή λόγο, όπου βεβαίως η εκμετάλλευση της πληροφορίας περιεχομένου μέσω κατάλληλων γλωσσικών μοντέλων κρίνεται αναγκαία.

Η αναγνώριση φωνημάτων, βέβαια, είναι μία εργασία που μπορεί να βρει εφαρμογές σε πολλά πεδία, πέραν της αναγνώρισης φωνής [59]. Για παράδειγμα, ένα αξιόπιστο σύστημα αναγνώρισης φωνημάτων είναι το βασικό συστατικό στοιχείο για τη φωνοτακτική ταυτοποίηση μιας γλώσσας (Language Identification - LID), δηλαδή την αυτόματη εξαγωγή της γλώσσας στην οποία εκφέρεται ένα συγκεκριμένο τμήμα κειμένου. Ακόμη, η αναγνώριση φωνημάτων μπορεί να χρησιμοποιηθεί για γρήγορη στόχευση λέξεων- και φράσεων-κλειδιών (Keyword Spotting - KWS), για εντοπισμό OOV λέξεων, κ.ά..

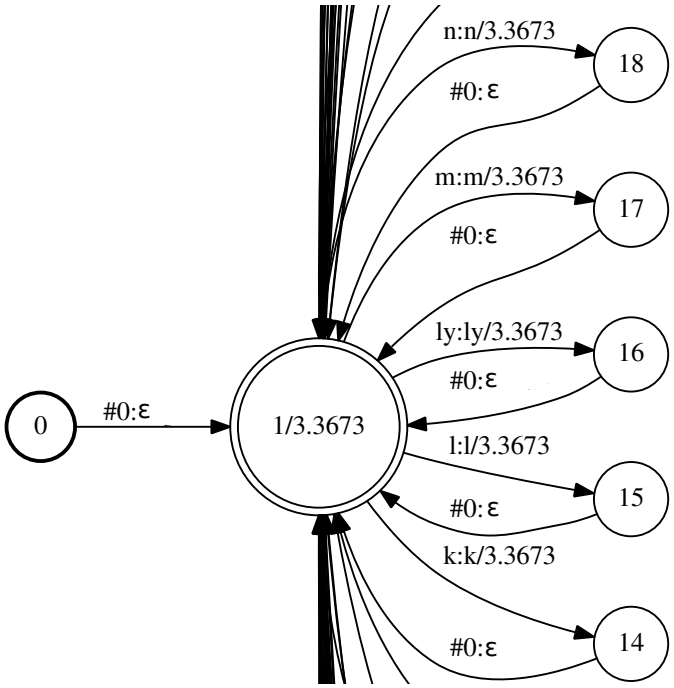
4.2 Κεντρικοί Άξονες του Συστήματος Αναγνώρισης

Η παρούσα εργασία, σύμφωνα με το πλαίσιο που σκιαγραφήθηκε στην Ενότητα 4.1, θα βασιστεί στην αναγνώριση φωνημάτων και όχι λέξεων. Στόχος είναι η συγκριτική μελέτη διαφορετικών συνόλων ακουστικών χαρακτηριστικών, όταν τα υπόλοιπα συστατικά στοιχεία του συστήματος, όπως το ακουστικό και το γλωσσικό μοντέλο παραμένουν αμετάβλητα.

Η διαδικασία για αναγνώριση φωνημάτων ακολουθεί τα βασικά βήματα της αναγνώρισης φωνής, όπως αναλύθηκαν στο Κεφάλαιο 2, με τις όποιες απαραίτητες αλλαγές. Οι "λέξεις" για το σύστημα αναγνώρισης είναι, τώρα, τα ίδια τα φωνήματα. Οπότε, ως είσοδος στο σύστημα αναμένεται και πάλι ένα σήμα φωνής, αλλά η έξοδος είναι μια ακολουθία φωνημάτων. Αυτό σημαίνει πως και οι απομαγνητοφωνήσεις που θα πρέπει να έχουμε στη διάθεσή μας για την εκπαίδευση, αλλά και για την αξιολόγηση του συστήματος, θα πρέπει να είναι σε επίπεδο φωνημάτων και όχι λέξεων. Η αξιολόγηση, έτσι, γίνεται με χρήση της μετρικής Phone Error Rate (PER), κατά αντιστοιχία της WER.

Η ακουστική μοντελοποίηση γίνεται βάσει του καθιερωμένου προτύπου των HMMs / GMMs, ενώ γίνεται χρήση τριφωνικών μοντέλων, λαμβάνοντας κατά τον τρόπο αυτό υπόψη την επίδραση του φαινομένου της συνάρθρωσης. Εφόσον στόχος είναι η εύρεση κατάλληλων ακουστικών χαρακτηριστικών, γίνεται προσπάθεια ελαχιστοποίησης της επίδρασης των υπόλοιπων παραγόντων στο τελικό αποτέλεσμα της αναγνώρισης. Για το λόγο αυτό, δε χρησιμοποιείται κάποιο στατιστικό γλωσσικό μοντέλο, όπως συμβαίνει στην πράξη. Αντ' αυτού, γίνεται χρήση μιας υποτυπώδους γραμματικής, όπου ορίζεται πως όταν βρισκόμαστε σε ένα οποιοδήποτε φώνημα, οι μεταβάσεις προς όλα τα υπόλοιπα φωνήματα ή προς το τέλος της πρότασης/φράσης είναι ισοπίθανες (π.χ. [40]). Τμήμα μιας τέτοιας γραμματικής, σε φορμαλισμό WFST, παρουσιάζεται στο Σχήμα 4.1. Φαίνεται πως το κόστος για όλες τις μεταβάσεις,

το οποίο προκύπτει ως ο αρνητικός φυσικός λογάριθμος της πιθανότητας μετάβασης, είναι το ίδιο. Το σύνολο των φωνημάτων που χρησιμοποιούνται εδώ είναι αυτό που παρουσιάζεται στον Πίνακα 4.1. Σημειώνεται πως το ειδικό σύμβολο #0 εισάγεται προς αντικατάσταση του ϵ ως σύμβολο εισόδου, για να μην υπάρχουν στην πορεία προβλήματα ντετερμινιστοποίησης.



Σχήμα 4.1: Τμήμα WFST που μοντελοποιεί μια γραμματική ισοπίθανων μεταβάσεων προς κάθε φώνημα.

4.3 Βάσεις Δεδομένων

Για τους σκοπούς των πειραμάτων της εργασίας, έγινε χρήση των ελληνικών βάσεων δεδομένων ATHENA [60] και Logotypografia [61]. Το σύνολο φωνημάτων για την ελληνική γλώσσα που χρησιμοποιήθηκε είναι αυτό που προτείνεται στο [61] και συνοψίζεται στον Πίνακα 4.1.

Ακολουθεί πρώτα μία παρουσίαση των δύο βάσεων και στη συνέχεια μία ανάλυση του ακριβούς τρόπου με τον οποίο έγινε χρήση των συγκεκριμένων βάσεων.

4.3.1 ATHENA

Η ATHENA είναι μια βάση δεδομένων για χρήση σε εφαρμογές έξυπνων σπιτιών και αυτοματισμών μέσω αναγνώρισης φωνής. Οι ηχογραφήσεις έγιναν σε χώρο δύο δωματίων με χρήση 20 πυκνωτικών μικροφώνων Shure MX391/O, 6 μικροφώνων MEMS (microelectromechanical systems), καθώς και μιας Kinect κάμερας για λήψη και αποθήκευση οπτικών δεδομένων. Για τη διατήρηση των καθαρών εκφορών χωρίς τις διάφορες αλλοιώσεις, έγινε ακόμη χρήση δύο close-talk μικροφώνων Sennheiser ew172G3. Όλες οι ηχογραφήσεις έχουν γίνει με ρυθμό δειγματοληψίας $48kHz$. Η βάση εμπεριέχει όλα τα στοιχεία εκείνα που αποτελούν εμπόδια σε μία πρακτική εφαρμογή αναγνώρισης από απόσταση, δηλαδή αντήχηση

Κατηγορία	Φώνημα	Παράδειγμα	Φώνημα	Παράδειγμα	Φώνημα	Παράδειγμα
Φωνήεντα						
	A	πουλ**ά**ω	i	ένεσ**η**	u	**ου**ρανός
	E	Αντρ**έ**ας	o	**ό**ρθιος		
Σύμφωνα						
Κλειστά	b	ε**μπ**όριο	g	ά**γκ**υρα	p	μελο**π**οιώ
	d	**ντ**ύσιμο	k	α**κ**ρίδα	t	**τ**υρί
Τυρβώδη	D	χορ**δ**ή	s	όρο**ς**	x	παι**χ**νίδι
	f	**φ**ιλία	T	αριθ**θ**μός	z	**ζ**έβρα
	G	τρα**γ**ωδία	v	κύ**β**ος		
Ένρινα	m	**μ**άνα	n	κολό**ν**α		
Υγρά	l	κύκ**λ**ος	r	**ρ**ίχνω		
Ουρανικά	c	**κ**ερί	J	λό**γι**α	N	**νι**όπαντρος
	C	**χ**ύτρα	ly	μα**λλι**ά		

Πίνακας 4.1: Σύνολο φωνημάτων της ελληνικής γλώσσας.

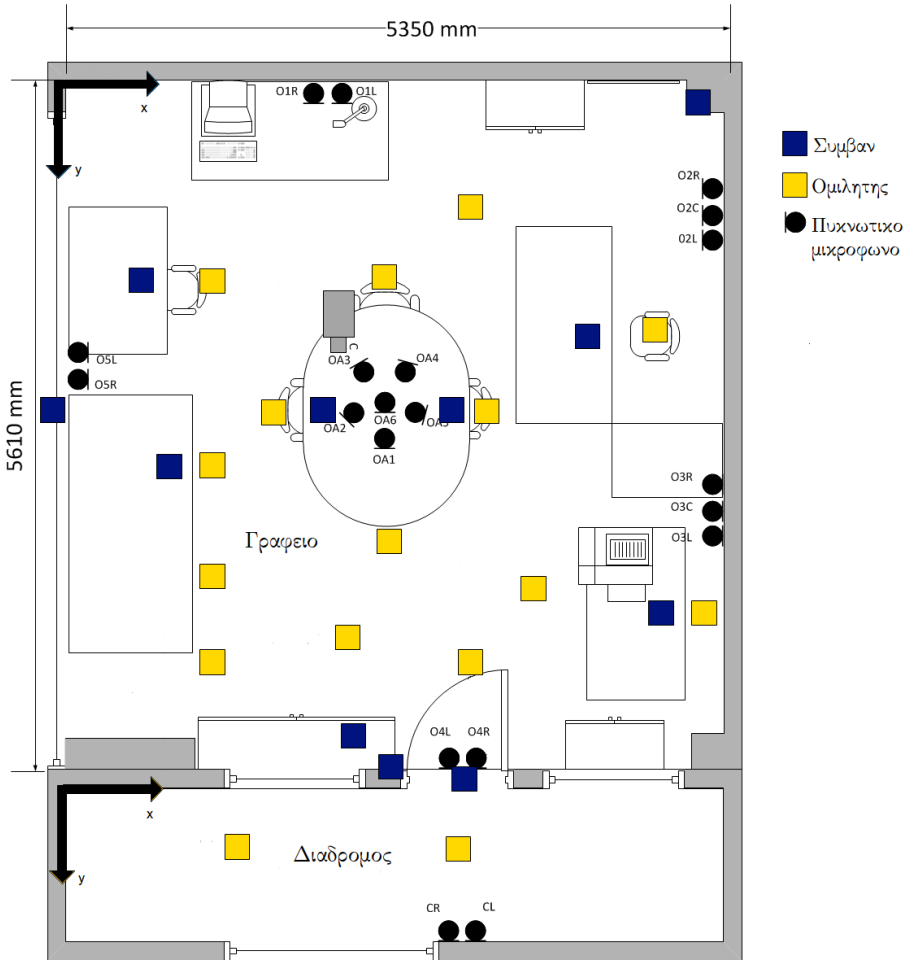
(reverberation), θόρυβο και διάφορα συμβάντα που λαμβάνουν χώρα είτε ανεξάρτητα, είτε σε επικάλυψη με τα φωνητικά δεδομένα. Ο χώρος ηχογραφήσεων παρουσιάζεται στο Σχήμα 4.2 με επισημειωμένες τις θέσεις των ομιλητών, των πυκνωτικών μικροφώνων και των διαφόρων συμβάντων.

Η βάση αποτελείται από 10 γυναίκες και 10 άντρες ομιλητές, καθένας από τους οποίους συμμετέχει σε 12 ηχογραφήσεις του ενός λεπτού. Συνεπώς, η συνολική διάρκεια της βάσης ισούται με 240 λεπτά. Αξίζει να σημειωθεί ότι 40% των συνολικών φωνητικών δεδομένων επικαλύπτονται με διάφορα συμβάντα υποβάθρου που αλλοιώνουν την ποιότητα, όπως είναι ήχοι που προέρχονται από βήματα, πόρτες, βήχα, τηλέφωνα, πληκτρολόγια, κ.λπ..

Τα φωνητικά δεδομένα διακρίνονται σε τέσσερις κατηγορίες και συγκεκριμένα σε φράσεις-κλειδιά, σε εντολές, σε πλούσιες φωνητικά προτάσεις και σε συζητήσεις. Οι φράσεις-κλειδιά είναι μικρές φράσεις που περιέχουν τη λέξη "Ντίρα" (π.χ. *Ντίρα σύνδεση*), ενώ οι εντολές αφορούν αυτοματισμούς έξυπνων σπιτιών (π.χ. *Κλείσε τα φώτα στο σαλόνι*). Οι πλούσιες φωνητικά προτάσεις είναι φράσεις που προέρχονται από τη Logotypografia, εκφερόμενες, βέβαια, από τους ομιλητές της ATHENA (π.χ. *στη συγκεκριμένη συνεδρίαση, κληρώνεται και ο αναπληρωτής του προέδρου που προαναφέραμε*). Τέλος, οι συζητήσεις είναι τμήματα διαλόγων μεταξύ δύο ομιλητών πάνω σε καθημερινά θέματα. Είναι συνολικά διαθέσιμες 240 εκφορές 12 διαφορετικών φράσεων-κλειδιών, 240 εκφορές 170 διαφορετικών εντολών και 140 εκφορές πλούσιων φωνητικά προτάσεων.

Όσον αφορά στους θορύβους υποβάθρου, αυτοί διακρίνονται σε 5 κατηγορίες, αναλόγως της πηγής τους. Έτσι, έχουμε θόρυβο λόγω μουσικής, λόγω ανεμιστήρα, λόγω ανοιχτού παραθύρου και λόγω ηλεκτρικής σκούπας. Η πέμπτη κατηγορία είναι η ησυχία, δηλαδή πρακτικά η απουσία θορύβου. Σε κάθε ηχογράφηση του ενός λεπτού υπάρχει μία μόνο κατηγορία θορύβου.

Η βάση δεδομένων είναι ισόποσα διαχωρισμένη σε σύνολο ανάπτυξης (dev set) και σύνολο ελέγχου (test set), με τις εκφορές 10 ομιλητών να ανήκουν στο πρώτο και τις εκφορές των υπολοίπων 10 να ανήκουν στο δεύτερο, ενώ παρέχονται μαζί τη βάση και οι απομαγνητοφωνήσεις των ηχογραφήσεων, με εξαίρεση τις συνομιλίες, οι οποίες δεν είναι απομαγνητοφωνημένες.



Σχήμα 4.2: Χώρος ηχογραφήσεων της βάσης δεδομένων ATHENA. Σημειώνονται οι θέσεις των ομιλητών, των πυκνωτικών μικροφώνων, αλλά και των διαφόρων συμβάντων. [*εικόνα προσαρμοσμένη από [60]*]

4.3.2 Logotypografia

Η Logotypografia αποτελεί την πρώτη ολοκληρωμένη ελληνική βάση δεδομένων. Αποτελείται από 70 γυναίκες και 55 άντρες ομιλητές, ενώ οι συνολική διάρκεια της βάσης είναι περίπου 72 ώρες ομιλίας, με 33136 εκφορές και μέση διάρκεια της κάθε εκφοράς 7.8 δευτερόλεπτα. Ένα μέρος των ηχογραφήσεων έχει γίνει με μικρόφωνα γραφείου Audio Technica ATM73a και οι υπόλοιπες με close-talk μικρόφωνα AKG C410. Όλες οι ηχογραφήσεις έχουν γίνει με ρυθμό δειγματοληψίας $16kHz$.

Οι ηχογραφήσεις διακρίνονται σε τρεις κατηγορίες αναλόγως του περιβάλλοντος ηχογράφησης. Συγκεκριμένα, ένα μέρος των ηχογραφήσεων έχει λάβει χώρα σε ηχομονωμένο δωμάτιο, ένα μέρος σε δωμάτιο χωρίς θορύβους και οι υπόλοιπες σε συνθήκες γραφείου, με όλες τις πηγές θορύβου που αυτό συνεπάγεται.

Μαζί με τη βάση παρέχονται οι απομαγνητοφωνήσεις για όλες τις εκφορές, επισημειωμένες με συγκεκριμένους περιγραφείς που υποδηλώνουν διάφορα φαινόμενα. Τα φαινόμενα αυτά είτε αφορούν ολόκληρη την εκφορά, όπως π.χ. ύπαρξη θορύβου υποβάθρου, είτε λαμβάνουν χώρα σε συγκεκριμένη χρονική στιγμή και με συγκεκριμένη διάρκεια, όπως π.χ. ήχος τηλεφώνου ή

παράπλευρη ομιλία από τρίτους ομιλητές που υπάρχουν στο χώρο ηχογράφησης. Οι εν λόγω περιγραφείς συνοψίζονται στον Πίνακα 4.2.

Φαινόμενα που αφορούν μέρος της εκφοράς

Περιγραφέας	Ερμηνεία
noise	Θόρυβος που δεν εμπίπτει σε κάποια από τις παρακάτω κατηγορίες
side_speech	Ομιλία από άλλον ομιλητή στο υπόβαθρο
phone_ring	Χτύπημα τηλεφώνου στο υπόβαθρο
breath	Ο ομιλητής εισπνέει ή εκπνέει
clear_throat	Ο ομιλητής καθαρίζει το λαιμό του
paper_rustle	Θόρυβος από χαρτιά
paff_noise	Ο ομιλητής μιλάει πολύ κοντά στο μικρόφωνο

Φαινόμενα που αφορούν ολόκληρη την εκφορά

Περιγραφέας	Ερμηνεία
TAG_RECHECK_NEEDED	Η απομαγνητοφώνηση είναι αβέβαιη και χρειάζεται επανεξέταση
TAG_BAD_AUDIO	Η εκφορά έχει πολύ κακή ακουστική ποιότητα
TAG_BAD_READING	Ο ομιλητής κάνει συχνές παύσεις
TAG_NON_NATIVE	Τα Ελληνικά δεν είναι η μητρική γλώσσα του ομιλητή
TAG_SPEECH_IN_NOISE	Θόρυβος υποβάθρου καθ' όλη τη διάρκεια της εκφοράς που σε κάποια σημεία καλύπτει την ομιλία
TAG_BACKGROUND_NOISE	Θόρυβος υποβάθρου καθ' όλη τη διάρκεια της εκφοράς

Πίνακας 4.2: Περιγραφείς των διαφόρων φωνητικών φαινομένων στη Logotypografia, μαζί με την ερμηνεία τους.

Εκτός από τους περιγραφείς του Πίνακα 4.2, γίνεται ακόμα χρήση ειδικών συμβόλων που υποδηλώνουν μη ολοκληρωμένες λέξεις, περικοπή της κυματομορφής, διαγραφές ή εισαγωγές λέξεων και λέξεις εκφερόμενες με λανθασμένη προφορά.

4.3.3 Επεξεργασία και Χρήση των Βάσεων

Γενικώς, η απόδοση ενός συστήματος αναγνώρισης φωνής είναι συνήθως καλύτερη όταν η εκπαίδευση γίνεται σε δεδομένα που έχουν ηχογραφηθεί σε ίδιες συνθήκες με αυτές υπό τις οποίες θα χρησιμοποιηθεί το σύστημα στην πράξη (π.χ. [62]). Ωστόσο, η συλλογή ενός ικανοποιητικά μεγάλου συνόλου δεδομένων υπό τις εκάστοτε κάθε φορά συνθήκες είναι πρακτικά αδύνατη. Για το λόγο αυτό, έχουν προταθεί μέθοδοι τεχνητής αλλοίωσης ενός καθαρού συνόλου δεδομένων για τη μερική μείωση της αναντιστοιχίας μεταξύ συνθηκών εκπαίδευσης και λειτουργίας [63]. Και στην περίπτωση αυτή, όμως, ενυπάρχουν μειονεκτήματα, καθώς για ένα σύστημα αναγνώρισης από απόσταση θα πρέπει για παράδειγμα να έχει εκ των προτέρων υπολογιστεί η κρουστική απόκρουση των δωματίων όπου πρόκειται να χρησιμοποιηθεί το σύστημα, καθώς και να υπάρχει κάποια γνώση για τους πιθανούς τύπους θορύβων υποβάθρου που μπορεί να υπεισέλθουν στην πράξη.

Αποφεύγοντας τέτοιου είδους περιορισμούς, τα πειράματα στην παρούσα εργασία θα εστιάσουν στην περίπτωση όπου υπάρχει έντονη αναντιστοιχία μεταξύ των συνθηκών κατά

την εκπαίδευση και κατά τον έλεγχο. Εξάλλου, έχουν προταθεί τεχνικές αποθορυβοποίησης που όταν εφαρμόζονται στα δεδομένα ελέγχου φαίνεται να είναι πιο αποτελεσματικές από την προσπάθεια αλλοίωσης των δεδομένων εκπαίδευσης [64].

Έτσι, λοιπόν, η εκπαίδευση του συστήματος γίνεται κάθε φορά αποκλειστικά με τις ηχογραφήσεις της Logotypografia που έχουν λάβει χώρα σε ηχομονωμένο δωμάτιο με close-talk μικρόφωνο, ενώ επίσης αγνοούνται όλες οι εκφορές που είναι επισημειωμένες με κάποιον περιγραφέα που αφορά ολόκληρη την εκφορά (Πίνακας 4.2). Έτσι, αποφεύγονται πολύ προβληματικές ή θορυβώδεις εκφορές, καθώς επίσης και εκφορές από αλλοδαπούς ομιλητές, εφόσον όλοι οι ομιλητές της βάσης ATHENA, όπου θα γίνει ο έλεγχος, έχουν ως μητρική γλώσσα την ελληνική. Η απόδοση μετράται βάσει των ηχογραφήσεων της βάσης ATHENA, από τις οποίες απομονώνονται τα τμήματα που περιέχουν ομιλία και ειδικότερα τα τμήματα που περιέχουν εντολές ή πλούσιες φωνητικά προτάσεις (αγνοούνται οι συνομιλίες και οι φράσεις-κλειδιά). Για να συμβαδίζουν οι δύο βάσεις μεταξύ τους, έγινε υποδειγματοληψία των ηχογραφήσεων της βάσης ATHENA από τα $48kHz$ στα $16kHz$.

Από τις εναπομείνασες εκφορές της Logotypografia αφαιρούνται, ακόμα, αυτές στις οποίες φαίνεται να έχει γίνει κάποια διαγραφή ή εισαγωγή λέξεων. Κι αυτό διότι η επισημείωση στις δύο αυτές περιπτώσεις είναι η ίδια, οπότε δεν μπορούμε να ξέρουμε τι πραγματικά είπε ο ομιλητής. Για παράδειγμα, εάν ο ομιλητής πρέπει να διαβάσει τη φράση *συμπεριφορά των φιλάθλων καταναλωτών* και αντ' αυτού πει *συμπεριφορά των ε καταναλωτών*, η εν λόγω απομαγνητοφώνηση θα είναι *συμπεριφορά των <φιλάθλων><ε> καταναλωτών*. Καταλήγουμε, έτσι, σε 6076 εκφορές μέσης διάρκειας $8.6sec$ στο σύνολο εκπαίδευσης, 190 εκφορές μέσης διάρκειας $2.7sec$ στο σύνολο ελέγχου και 190 εκφορές μέσης διάρκειας $2.9sec$ στο σύνολο ανάπτυξης.

Από τις απομαγνητοφωνήσεις και των δύο βάσεων αφαιρούνται όλα τα σημεία στίξης και τα ειδικά σύμβολα, ενώ από τη Logotypografia αφαιρούνται και τα τμήματα λέξεων που περιλαμβάνονται στις αρχικές απομαγνητοφωνήσεις, αλλά επισημειώνεται πως δεν εκφέρθηκαν από τον ομιλητή. Όσον αφορά στους περιγραφείς για φαινόμενα που αφορούν μέρος της εκφοράς, καθένας από αυτούς αντικαταστάθηκε από κάποιο ειδικό "φώνημα σιωπής", σύμφωνα με τις αντιστοιχίες που Πίνακα 4.3. Εξαίρεση αποτελεί ο περιγραφέας paff_noise, ο οποίος δε λήφθηκε υπόψιν εφόσον δηλώνει ότι μία ή δύο λέξεις δεν εκφέρθηκαν πολύ καθαρά, οπότε δεν έχει νόημα να μπει στη θέση του ένα φώνημα σιωπής. Επιπλέον, χρησιμοποιείται το φώνημα SIL που δηλώνει την προαιρετική σιωπή που μπορεί να υπάρξει μεταξύ των λέξεων, στην αρχή ή στο τέλος κάθε εκφοράς, τόσο για τη Logotypografia, όσο και για την ATHENA.

Περιγραφέας	Φώνημα
noise	NOI
side_speech	SID
phone_ring	PHO
breath	BRE
clear_throat	CLE
paper_rustle	PAP

Πίνακας 4.3: Αντιστοιχία μεταξύ περιγραφέων στη Logotypografia και φωνημάτων που χρησιμοποιήθηκαν στην πράξη.

Από τη συστοιχία των πυκνωτικών μικροφώνων που έχουμε στη διάθεσή μας μέσω της βάσης ATHENA, θα γίνει χρήση μόνο των μικροφώνων OA6 και CT1. Το πρώτο είναι το μεσαίο από την εξάδα μικροφώνων που βρίσκονται στο κέντρο του γραφείου, όπως φαίνεται

στην Εικόνα 4.2, ενώ το δεύτερο είναι το close-talk μικρόφωνο του πρωτεύοντος ομιλητή και χρησιμοποιείται για λόγους σύγκρισης.

Για να υπάρχει ένα ποσοτικό μέτρο της δυσκολίας της αναγνώρισης υπό τις συνθήκες των συγκεκριμένων ηχογραφήσεων, στον Πίνακα 4.4 δίνεται για κάθε μικρόφωνο η μέση τιμή (μ), η τυπική απόκλιση (σ), καθώς και η ελάχιστη (min) και μέγιστη (max) τιμή του Αριθμητικού Τμηματικού Σηματοθορυβικού Λόγου (Arithmetic Segmental Signal to Noise Ratio - SSNRA), όπως προκύπτει λαμβάνοντας υπόψιν όλες τις εκφορές του συνόλου ελέγχου της βάσης ATHENA που χρησιμοποιούνται, προτού να λάβει χώρα η υποδειγματοληψία στα $16kHz$. Το SSNRA προτιμάται από εναλλακτικές μεθόδους εκτίμησης του SNR για λόγους αριθμητικής ευστάθειας [65]. Δοθέντος ενός θορυβώδους σήματος s, αυτό διασπάται σε M μη επικαλυπτόμενα πλαίσια και για κάθε πλαίσιο i εκτιμάται η ισχύς του $\sigma^2_{s,i}$, καθώς και η ισχύς του θορύβου $\hat{\sigma}^2_{n,i}$ στο πλαίσιο αυτό. Το SSNRA υπολογίζεται, τότε, ως

$$\text{SSNRA} = 10\log\left(\frac{1}{M}\sum_{i=0}^{M-1}\frac{\sigma^2_{s,i}-\hat{\sigma}^2_{n,i}}{\hat{\sigma}^2_{n,i}}\right). \tag{4.1}$$

Επιλέχθηκε μήκος πλαισίου ίσο με $32msec$, ενώ για την εκτίμηση του θορύβου λήφθηκε υπόψιν σήμα συνολικής διάρκειας $0.8sec$ πριν και μετά την εκάστοτε εκφορά. Συγκεκριμένα, χρησιμοποιήθηκε το τμήμα από $0.5sec$ έως $0.1sec$ πριν την εκφορά και από $0.1sec$ έως $0.5sec$ μετά την εκφορά. Το διάρκειας $0.1sec$ σήμα αμέσως πριν και αμέσως μετά την εκφορά δε λήφθηκε υπόψιν, λόγω πιθανών σφαλμάτων του αλγορίθμου Εντοπισμού Δραστηριότητας Φωνής (Voice Activity Detection - VAD) που έχει χρησιμοποιηθεί (τα αποτελέσματα του οποίου ήταν εκ των προτέρων γνωστά). Η ισχύς εκτιμάται γενικά ως η μέση τετραγωνική ενέργεια. Σημειώνεται πως σε περίπτωση που ο αριθμητής της σχέσης (4.1) προκύψει για κάποιο πλαίσιο μη-θετικός, τότε ο εν λόγω όρος δε λαμβάνεται υπόψιν στο άθροισμα, ενώ εάν αυτό ισχύει για όλα τα πλαίσια του σήματος, τότε το SSNRA για το σήμα αυτό τίθεται ίσο με $-40dB$.

	μ	σ	min	max		μ	σ	min	max
O1L	9.66	5.34	-9.00	22.02	O5R	9.68	5.57	-10.59	21.08
O1R	9.43	5.30	-6.05	22.25	OA1	10.85	5.69	-1.28	24.81
O2C	9.86	5.80	-5.42	23.06	OA2	11.22	5.75	-2.58	24.67
O2L	8.42	5.67	-16.42	20.44	OA3	10.94	5.88	-6.50	23.75
O2R	9.81	5.85	-6.23	23.45	OA4	11.19	5.88	-4.48	23.90
O3C	10.01	6.20	-7.72	25.64	OA5	10.73	5.94	-4.95	23.86
O3L	10.01	6.10	-8.57	24.77	**OA6**	**11.06**	**5.92**	**-2.05**	**23.66**
O3R	10.00	6.14	-6.71	25.56	CCL	0.18	8.77	-40.00	19.32
O4L	9.18	6.63	-7.93	23.28	CCR	0.25	8.93	-40.00	20.11
O4R	8.84	6.72	-10.67	23.11	**CT1**	**33.68**	**5.82**	**16.53**	**46.35**
O5L	9.92	5.34	-7.31	22.65					

Πίνακας 4.4: Στατιστικά στοιχεία του SSNRA (σε dB) για κάθε μικρόφωνο της συστοιχίας, καθώς και για το close-talk μικρόφωνο, λαμβάνοντας υπόψιν όλες τις εκφορές του συνόλου ελέγχου της βάσης ATHENA που χρησιμοποιούνται. Με έντονη γραμματοσειρά παρουσιάζονται τα δύο μικρόφωνα που θα χρησιμοποιηθούν για τα πειράματα της εργασίας.

Όπως αναμενόταν, οι υψηλότερες τιμές SSNRA παρουσιάζονται για τα μικρόφωνα που βρίσκονται στο κέντρο του γραφείου, εφόσον οι ομιλητές βρίσκονται σε διάφορες τυχαίες

θέσεις και, άρα, στη μέση περίπτωση, τα κεντρικά μικρόφωνα είναι σε ευνοϊκή θέση. Σε κάθε περίπτωση, ωστόσο, το SSNRA φαίνεται ότι λαμβάνει τιμές σε ένα αρκετά μεγάλο εύρος. Για παράδειγμα, εκφορές όπου ο ομιλητής βρίσκεται στο διάδρομο οδηγούν ακόμα και σε αρνητικό SSNRA για το μικρόφωνο OA6, γεγονός που μας προϊδεάζει για γενικά υψηλά ποσοστά σφάλματος αναγνώρισης. Όμοια, φαίνεται πως υπάρχουν τμήματα φωνής που σχεδόν δε γίνονται αντιληπτά από τις ηχογραφήσεις των μικροφώνων του διαδρόμου, όπου το SSNRA λαμβάνει και την ελάχιστη τιμή των $40dB$. Αντιθέτως, όσον αφορά στο CT1, όπου το SSNRA λαμβάνει σταθερά υψηλές τιμές, τα αποτελέσματα αναμένεται να είναι σημαντικά καλύτερα.

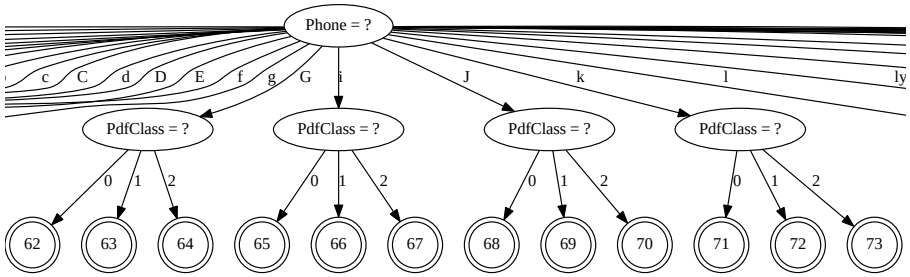
Ιδιαίτερη έμφαση δόθηκε στη δημιουργία κατάλληλου φωνητικού λεξικού, δηλαδή στην κατάλληλη αντιστοιχία μεταξύ λέξεων και ακολουθίας φωνημάτων. Θεωρήθηκε σκόπιμο να περιέχονται όλες οι λέξεις - ακόμα και οι κομμένες / ανορθόγραφες / μη-σωστά προφερόμενες κ.λπ. - που υπήρχαν στα διάφορα διαθέσιμα αρχεία απομαγνητοφώνησης. Κι αυτό γιατί εφόσον θα γίνει αναγνώριση φωνημάτων, καλό θα είναι να μην υπάρχουν "άγνωστες λέξεις" (OOV) στο τελικό σύστημα, γιατί απλά δε θα υπάρχουν "άγνωστα φωνήματα". Το εν λόγω λεξικό περιλαμβάνει τις αντιστοιχίες από λέξεις με λατινικούς χαρακτήρες στα φωνήματα του Πίνακα 4.1. Συνεπώς, το πρώτο βήμα της μετατροπής ήταν η μετατροπή όλων των απομαγνητοφωνήσεων ώστε να περιλαμβάνουν λατινικούς και όχι ελληνικούς χαρακτήρες. Τέλος, για τις λέξεις με πολλαπλές προφορές, διατηρήθηκε μόνο η (εμπειρικά) πλέον πιθανή.

4.4 Διαδοχικά Στάδια της Αναγνώρισης

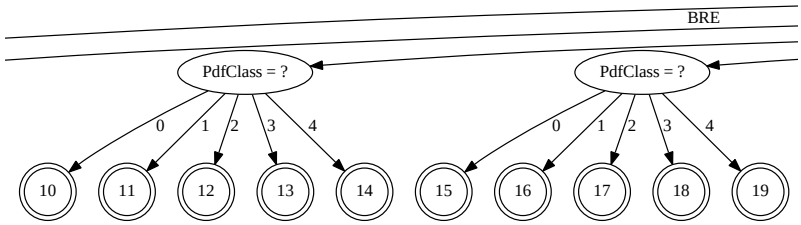
Για την εκπαίδευση του συστήματος αναγνώρισης, ακολουθήθηκαν οι βασικές αρχές των συνταγών του Kaldi. Οπότε, μετά την προετοιμασία των δεδομένων σε κατάλληλη μορφή ώστε να μπορούν να αποτελούν είσοδο για το Kaldi, η εκπαίδευση αποτελείται από τρία βασικά στάδια, την εκπαίδευση του μονοφωνικού μοντέλου και δύο περάσματα για την εκπαίδευση του τριφωνικού μοντέλου. Σημειώνεται ότι κατά την αναγνώριση φωνής είναι συνήθης πρακτική το σύνολο φωνημάτων να επαυξάνεται με επισημειωμένα φωνήματα που αφορούν διαφορετικές εκφορές και τονισμούς του ίδιου φωνήματος. Για παράδειγμα, ένα φώνημα επιτονίζεται συνήθως διαφορετικά όταν βρίσκεται στην αρχή, στη μέση ή στο τέλος της λέξης. Η επαύξηση αυτή γίνεται από προεπιλογή στις συνταγές του Kaldi και απενεργοποιήθηκε για τους σκοπούς των πειραμάτων της εργασίας, εφόσον αντιμετωπίζουμε τα φωνήματα ως τις ίδιες τις λέξεις.

Αρχικά, εκπαιδεύεται ένα μονοφωνικό μοντέλο. Για εξοικονόμηση υπολογιστικών πόρων, εφόσον οι παράμετροι που πρέπει να εκτιμηθούν για το μονοφωνικό μοντέλο είναι σχετικά λίγες και εφόσον τα δεδομένα εκπαίδευσης που έχουμε στη διάθεσή μας είναι πολλά, χρησιμοποιείται ένα υποσύνολο των δεδομένων αυτών. Για την ακρίβεια, από τις 6076 εκφορές, χρησιμοποιούνται για το στάδιο αυτό μόνο οι 2000 μικρότερης διάρκειας. Ακόμα, η χρήση των μικρότερων εκφορών οδηγεί σε πιο εύκολη και αποτελεσματική ευθυγράμμιση, ξεκινώντας από flat start. Δημιουργείται, έτσι, ένα στοιχειώδες δέντρο, όπως φαίνεται στο Σχήμα 4.3. Εκεί φαίνεται και η διαφορετική αντιμετώπιση μεταξύ πραγματικών φωνημάτων, που μοντελοποιούνται με τρεις καταστάσεις, και φωνημάτων που δηλώνουν σιωπή ή θόρυβο, που μοντελοποιούνται με 5 καταστάσεις.

Με βάση τη μοντελοποίηση που έχει προκύψει από το πρώτο στάδιο, ακολουθεί η εξαναγκασμένη ευθυγράμμιση 3000 τυχαίων εκφορών από το σύνολο των δεδομένων εκπαίδευσης. Για την ευθυγράμμιση, γίνεται χρήση του ακτινωτού αλγορίθμου Viterbi με εύρος ακτίνας (στο λογαριθμικό πεδίο) ίσο με 8. Εάν σε κάποια επανάληψη ο αλγόριθμος δεν καταφέρει να



(i)



(ii)

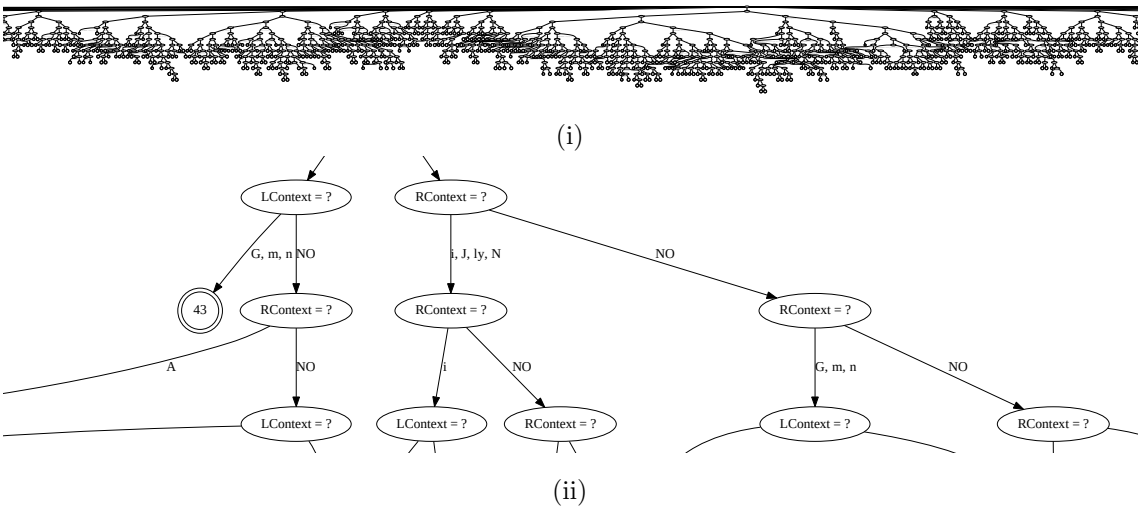
Σχήμα 4.3: (i) Τμήμα του δέντρου της μονοφωνικής μοντελοποίησης που αφορά φωνήματα σιωπής και θορύβου. (ii) Τμήμα του δέντρου που αφορά πραγματικά φωνήματα.

φτάσει σε τελική κατάσταση, ξαναπροσπαθεί με εύρος 40.

Στη συνέχεια, λαμβάνει χώρα η δημιουργία του τριφωνικού μοντέλου, με χρήση των 3000 ευθυγραμμισμένων εκφορών. Αυτό χτίζεται με τη λογική των δέντρων απόφασης, όπως αυτά έχουν αναλυθεί στην Υποενότητα 2.2.4, βάσει στατιστικών στοιχείων που συλλέγονται από τα μονοφωνικά μοντέλα. Για την εν λόγω διαδικασία, οι ερωτήσεις παράγονται αυτόματα και δεν έχουν απαραίτητα κάποια γλωσσολογική θεμελίωση. Ο μέγιστος αριθμός καταστάσεων, δηλαδή ο μέγιστος αριθμός φύλλων, ορίζεται στις 2000. Ο αριθμός γκαουσιανών ανά κατάσταση δεν είναι σταθερός, αλλά ο μέγιστος αριθμός γκαουσιανών για όλες τις καταστάσεις ορίζεται στις 10000. Σημειώνεται ότι χρησιμοποιούνται γενικά γκαουσιανές με διαγώνιους πίνακες συμμεταβλητότητας. Ένα τμήμα τέτοιου δέντρου απόφασης φαίνεται στο Σχήμα 4.4. Για την ακρίβεια, πρόκειται για πολλά δέντρα απόφασης, ενοποιημένα σε ένα μοναδικό δέντρο.

Τέλος, γίνεται εκ νέου εξαναγκασμένη ευθυγράμμιση του συνόλου αυτή τη φορά των δεδομένων, βάσει του τριφωνικού μοντέλου, και η διαδικασία της εκπαίδευσης επαναλαμβάνεται για να ληφθεί ένα βελτιωμένο μοντέλο που λαμβάνει υπόψιν όλο το σύνολο εκπαίδευσης. Έχοντας το τελικό τριφωνικό μοντέλο στη διάθεσή μας, σειρά έχει η δημιουργία του τελικού WFST, ακολουθώντας τα στάδια που έχουν περιγραφεί στην Υποενότητα 3.4.2.

Για την αποκωδικοποίηση χρησιμοποιείται ακτινωτή αναζήτηση Viterbi με εύρος ακτίνας ίσο με 13. Για το σύνολο ανάπτυξης δεν επιλέγεται άμεσα η πλέον πιθανή ακολουθία φωνημάτων, αλλά δημιουργείται ένα πλέγμα με τις πιθανότερες υποθέσεις. Στο πλέγμα αυτό γίνεται επαναβαθμολόγηση (rescoring) με διαφορετικές τιμές PIP (Phoneme Insertion Penalty κατά αντιστοιχία του WIP). Για κάθε επαναβαθμολόγηση επιλέγεται το πλέον πιθανό μονοπάτι του πλέγματος και υπολογίζεται το PER. Το PIP που δίνει το μικρότερο PER είναι αυτό που χρησιμοποιείται κατά την αποκωδικοποίηση του συνόλου ελέγχου. Οι τιμές που δοκιμάζονται



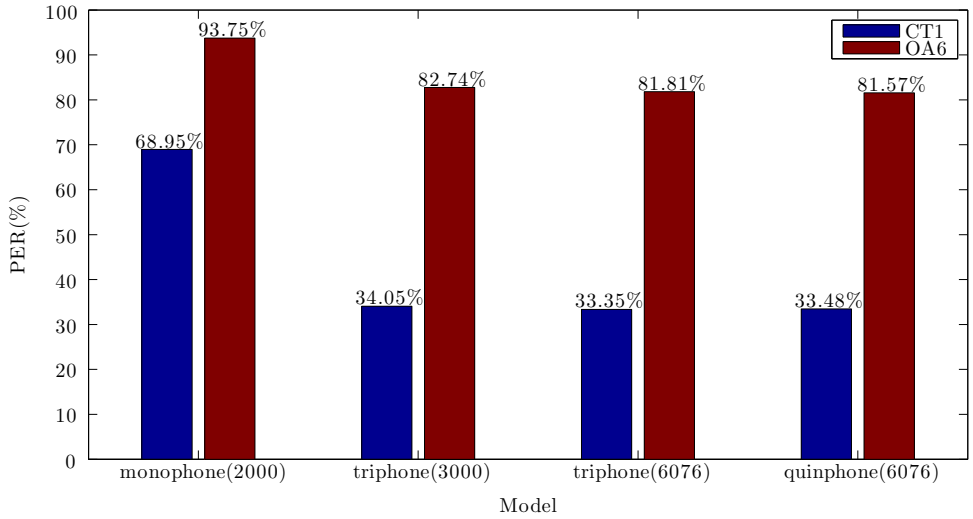
Σχήμα 4.4: (i) Τμήμα δέντρου απόφασης για την τριφωνική μοντελοποίηση. (ii) Μεγέθυνση τμήματος του (i).

για το PIP είναι οι $\{0, 0.5, 1\}$[1].

Στο Σχήμα 4.5 παρουσιάζεται το PER του συστήματος, εάν η αποκωδικοποίηση λάβει χώρα αμέσως μετά τη μοντελοποίηση σε ένα από τα τρία στάδια που περιγράφηκαν. Δίνεται, ακόμη, το PER στην περίπτωση που αντί για τριφωνική γίνει πενταφωνική μοντελοποίηση. Στην περίπτωση αυτή, τα δύο πρώτα στάδια παραμένουν ίδια και το πενταφωνικό μοντέλο εκπαιδεύεται πάνω στο σύνολο των δεδομένων, αφότου αυτά έχουν ευθυγραμμιστεί βάσει της τριφωνικής μοντελοποίησης του δεύτερου σταδίου. Παρατηρείται πως ενώ η τριφωνική μοντελοποίηση έχει ξεκάθαρα ευεργετικά αποτελέσματα στην απόδοση του συστήματος έναντι της μονοφωνικής, δε συμβαίνει το ίδιο όταν περνάμε από τριφωνήματα σε πενταφωνήματα (quinphones). Αυτό, σε συνδυασμό με το πολύ μεγαλύτερο υπολογιστικό κόστος, τόσο σε μνήμη, όσο και σε χρόνο, που επιφέρουν τα πενταφωνήματα, όπως παρατηρήθηκε από τα αντίστοιχα πειράματα, θέτει τη χρήση τριφωνημάτων μια συνετή επιλογή για τη μοντελοποίηση του περιβάλλοντος των φωνημάτων. Παρατηρείται, ακόμα, πως το σφάλμα αναγνώρισης είναι πολύ μεγαλύτερο, όπως άλλωστε αναμενόταν, στην περίπτωση της αναγνώρισης από απόσταση (μικρόφωνο OA6), σε σύγκριση με την αναγνώριση με το μικρόφωνο CT1 που βρίσκεται δίπλα στο στόμα του εκάστοτε ομιλητή.

Τα αποτελέσματα του Σχήματος 4.5 λήφθηκαν με χρήση 13 MFCCs, όπου το πρώτο (C_0) έχει αντικατασταθεί από την τετραγωνική ενέργεια, μαζί με τις πρώτες και δεύτερες παραγώγους τους. Έγινε χρήση παραθύρου Hamming μήκους $32msec$ με κίνηση ανά $10msec$ και συστοιχίας 40 φίλτρων που κάλυπταν το διάστημα $[0Hz, 8000Hz]$, ενώ οι παράγωγοι για κάθε πλαίσιο υπολογίστηκαν με βάση τα 8 γειτονικά πλαίσια. Οι λεπτομέρειες των MFCCs θα αναλυθούν στο Κεφάλαιο 5.

[1] Στα πειράματα όπου θα χρησιμοποιηθεί στατιστικό γλωσσικό μοντέλο, δοκιμάζονται επίσης διαφορετικές τιμές και για το LMSF, και για την ακρίβεια οι ακέραιες τιμές από 1 έως 20.



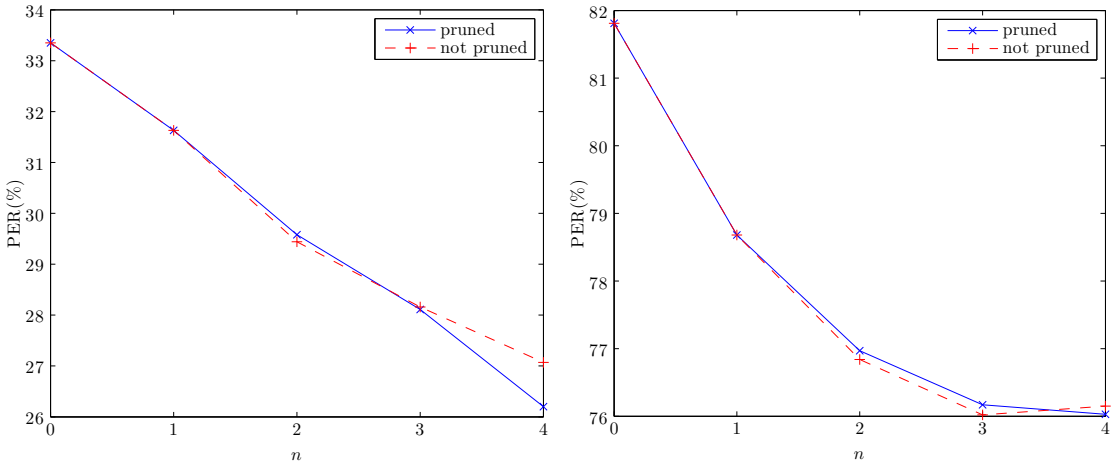
Σχήμα 4.5: Σφάλμα του συστήματος αναγνώρισης με χρήση διαφορετικών ακουστικών μοντέλων, για αναγνώριση από κοντά και από απόσταση. Στις παρενθέσεις αναγράφεται ο αριθμός των εκφορών, από τις συνολικά 6076, που χρησιμοποιήθηκαν για κάθε μοντέλο. Κάθε μοντέλο χτίζεται βάσει των ευθυγραμμίσεων που προκύπτουν από το προηγούμενό του από αριστερά προς τα δεξιά. Εξαίρεση αποτελεί το quinphone(6076) που βασίζεται στο triphone(3000).

4.5 Επίδραση του Γλωσσικού Μοντέλου

Όπως εξηγήθηκε στην Ενότητα 4.2, για τα πειράματα της εργασίας δε θα γίνει χρήση στατιστικού γλωσσικού μοντέλου, αλλά όλα τα φωνήματα θα θεωρούνται ανά πάσα στιγμή ισοπίθανα. Ωστόσο, είναι ενδιαφέρον στο σημείο αυτό να εξετάσουμε την επίδραση που θα είχε ένα απλό γλωσσικό μοντέλο στα αποτελέσματα της αναγνώρισης. Για τη μοντελοποίηση της γλωσσικής πληροφορίας, λοιπόν, χρησιμοποιήθηκε το IRSTLM [66], ένα εργαλείο ανοιχτού κώδικα για τη δημιουργία n-grams.

Το σώμα του κειμένου που χρησιμοποιήθηκε για τη δημιουργία των n-grams αποτελείται από όλες τις απομαγνητοφωνήσεις της Logotypografia, εκτός από εκείνες που αφορούν εκφορές επισημειωμένες με τους περιγραφείς TAG_NON_NATIVE, TAG_BAD_AUDIO, TAG_BAD_READING και TAG_RECHECK_NEEDED. Πρόκειται συνολικά για 30188 εκφορές. Αφότου οι εν λόγω απομαγνητοφωνήσεις μετατραπούν σε ακολουθίες φωνημάτων, όπου αγνοούνται τα φωνήματα σιωπής και θορύβου του Πίνακα 4.3, προκύπτει ένα σώμα κειμένου με 5,276,449 φωνήματα. Στην αρχή και στο τέλος της απομαγνητοφώνησης κάθε εκφοράς προστίθενται τα ειδικά σύμβολα $<s>$ και $<\backslash s>$ αντίστοιχα, για τους λόγους που αναλύονται στην Υποενότητα 2.3.1. Ακόμα, λαμβάνει χώρα smoothing με τη μέθοδο Witten-Bell που περιγράφεται στην Υποενότητα 2.3.2. Τέλος, για μείωση της πολυπλοκότητας του μοντέλου, αφαιρούνται τα n-grams εκείνα που εμφανίζονται με πολύ μικρή πιθανότητα και μπορούν να υπολογιστούν μέσω των backoff πιθανοτήτων τους χωρίς σημαντικές μεταβολές στο τελικό αποτέλεσμα. Για την ακρίβεια, επιλέγεται κάθε φορά να αφαιρούνται n-grams με πιθανότητα εμφάνισης μικρότερη του 10^{-5}. Στο Σχήμα 4.6 παρουσιάζονται τα σχετικά αποτελέσματα όταν όλες οι υπόλοιπες παράμετροι είναι ίδιες με αυτές που χρησιμοποιήθηκαν στα πειράματα του Σχήματος 4.5.

Φαίνεται πως η χρήση στατιστικού γλωσσικού μοντέλου πράγματι βελτιώνει την απόδοση



(i) Μικρόφωνο CT1 (ii) Μικρόφωνο OA6

Σχήμα 4.6: Επίδραση του γλωσσικού μοντέλου στην αναγνώριση με χρήση των μικροφώνων CT1 και OA6. Στον οριζόντιο άξονα φαίνεται η τάξη του γλωσσικού μοντέλου. Για $n = 0$, όλα τα φωνήματα θεωρούνται ισοπίθανα.

του συστήματος, χωρίς, ωστόσο, οι διαφορές να είναι τόσο θεαματικές όσο σε συστήματα αναγνώρισης λέξεων. Ακόμη, παρατηρείται πως η χρήση pruning όχι μόνο δεν επιφέρει σημαντικές αλλαγές, αλλά στις περισσότερες περιπτώσεις δίνει ελαφρώς καλύτερα αποτελέσματα απ' ό,τι όταν όλα τα n-grams διατηρούνται στο μοντέλο. Αυτό είναι πολύ σημαντικό από υπολογιστικής άποψης, εάν αναλογιστούμε πως στην περίπτωση της μοντελοποίησης με 3-grams, για παράδειγμα, με το pruning που κάνουμε αγνοούνται περίπου το 33.2% των αρχικά υπολογισθέντων 3-grams. Στην περίπτωση των 4-grams το αντίστοιχο ποσοστό είναι 58.7%.

Κεφάλαιο 5

Mel Frequency Cepstrum Coefficients (MFCCs)

5.1 Θεωρία των MFCCs

Τα πλέον καθιερωμένα σύνολα χαρακτηριστικών που χρησιμοποιούνται σε πρακτικές εφαρμογές, αλλά και από τα πρώτα που προτάθηκαν και έδωσαν αξιόλογα αποτελέσματα είναι τα λεγόμενα MFCCs (Mel-Frequency Cepstrum Coefficients - Συντελεστές Αναφάσματος στις Mel-Συχνότητες) [67].

Η απαραίτητη προεργασία που γίνεται είναι να περάσει κάθε σήμα από σύστημα προέμφασης με συνάρτηση μεταφοράς

$$H_{preemph}(z) = 1 - \tilde{a}z^{-1}, \tilde{a} \in (0.9, 1) \tag{5.1}$$

και εν συνεχεία να χωριστεί σε επικαλυπτόμενα, συνήθως, πλαίσια. Προκειμένου να ελαχιστοποιηθούν οι ασυνέχειες στα άκρα των πλαισίων, σε κάθε ένα πλαίσιο εφαρμόζεται συνήθως παραθύρωση Hamming. Το παράθυρο Hamming μήκους L ορίζεται ως

$$w(n) = 0.54 + 0.46 \cos\left(\frac{2\pi n}{L-1}\right) . \tag{5.2}$$

Η προέμφαση είναι απαραίτητη για την ενίσχυση των υψηλών συχνοτήτων, όπου συνήθως οι τιμές του φάσματος των ακουστικών σημάτων είναι χαμηλές και επηρεάζονται πιο έντονα από τυχόν θόρυβο. Η παραθύρωση είναι απαραίτητη, όπως και σε πολλές άλλες εφαρμογές της Επεξεργασίας Φωνής, διότι οι στατιστικές ιδιότητες του σήματος μεταβάλλονται με το χρόνο, αλλά θεωρούμε πως μένουν αμετάβλητες για μικρά διαστήματα (της τάξης των $10 - 30msec$) [31], οπότε μπορούν να εφαρμοστούν οι κλασικές τεχνικές ανάλυσης Fourier. Λόγω της μικρής διάρκειας των παραθύρων που χρησιμοποιούνται, τα MFCCs ανήκουν σε μια ευρεία κατηγορία χαρακτηριστικών που καλούνται short-term.

Τα MFCCs είναι μια αναπαράσταση που ορίζεται ως το πραγματικό cepstrum ενός παραθυρωμένου σήματος βραχέος χρόνου που προέρχεται από τον FFT μετασχηματισμό του σήματος. Το αξιοσημείωτο της μεθόδου είναι ότι γίνεται χρήση μιας μη-γραμμικής κλίμακας συχνοτήτων, η οποία προσεγγίζει τη συμπεριφορά της ανθρώπινης ακοής και έχει αποδειχθεί πως έχει σημαντικά πλεονεκτήματα στο πεδίο της αναγνώρισης φωνής. Η όλη ιδέα βασίζεται στο γεγονός πως ο άνθρωπος αντιλαμβάνεται πιο εύκολα ηχητικές μεταβολές στις χαμηλές παρά στις υψηλές συχνότητες.

Η μονάδα mel ορίζεται έτσι ώστε ζεύγη ήχων που κατά την ανθρώπινη αντίληψη απέχουν το ίδιο μεταξύ τους ως προς το pitch, διαχωρίζονται από έναν ίδιο αριθμό $mels$. Οι σχέσεις που συνδέουν μια συχνότητα εκφρασμένη σε $Hertz$ με την αντίστοιχη σε κλίμακα mel δεν είναι μοναδικές. Μία συχνά χρησιμοποιούμενη φόρμουλα είναι η εξής [68]:

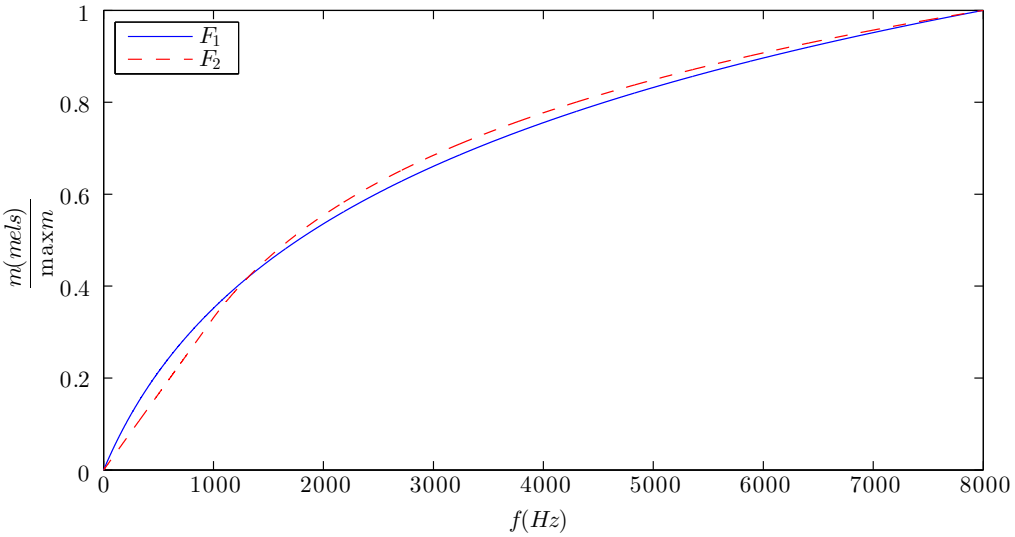
$$m = B(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{5.3}$$

$$f = B^{-1}(m) = 700\left(10^{m/2595} - 1\right) \tag{5.4}$$

Μία εναλλακτική φόρμουλα, η ιδέα της οποίας εμφανίστηκε αρχικά στο [69], ορίζει πως η αντιστοιχία μεταξύ $mels$ και $Hertz$ είναι γραμμική μέχρι τα $1000Hz$ και στη συνέχεια γίνεται λογαριθμική ως εξής:

$$m = \begin{cases} \dfrac{3f}{200} & , f < 1000Hz \\ 1000 + \dfrac{\log \frac{f}{1000}}{\log 1.0711703} & , f \geq 1000Hz \end{cases} \tag{5.5}$$

Οι δύο διαφορετικές προσεγγίσεις απεικονίζονται γραφικά στο Σχήμα 5.1.



Σχήμα 5.1: Μετατροπή των συχνοτήτων από την κλίμακα $Hertz$ στην κλίμακα mel. Η καμπύλη F_1 αντιστοιχεί στη φόρμουλα (5.3), ενώ η F_2 αντιστοιχεί στη φόρμουλα (5.5).

Η ιδέα, οπότε, είναι να χρησιμοποιήσουμε μια συστοιχία φίλτρων (filterbank), καθένα από τα οποία συλλέγει τις συχνότητες μίας συγκεκριμένης μπάντας, με τη συστοιχία να είναι πιο πυκνή στις χαμηλές και πιο αραιή στις υψηλές συχνότητες, καθώς η ανθρώπινη ακοή γίνεται λιγότερο ευαίσθητη.

Το πιο σύνηθες filterbank αποτελείται από $Q(\approx 20 - 40)$ τριγωνικά φίλτρα H^j, καθένα από τα οποία έχει εύρος ζώνης τέτοιο ώστε οι συχνότητες αποκοπής του να ταυτίζονται με τις κεντρικές συχνότητες των δύο γειτονικών του φίλτρων στην κλίμακα mel. Συμβολίζοντας την κεντρική συχνότητα του j φίλτρου ως f_c^j, όπου f_c^0 και f_c^{Q+1} οι κάτω και πάνω συχνότητες

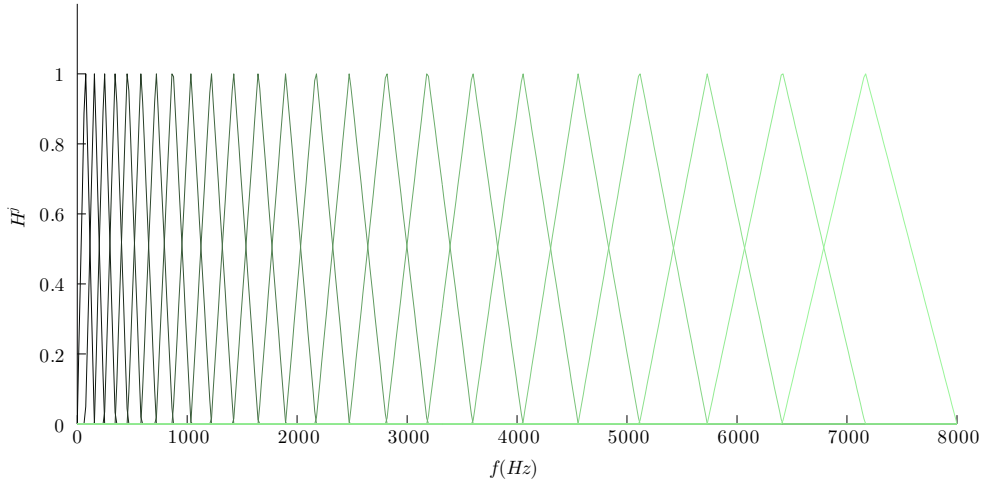
αποκοπής του πρώτου και τελευταίου φίλτρου της συστοιχίας αντίστοιχα, και θεωρώντας ακόμη πως $H^j(f_c^j) = 1$, το j φίλτρο περιγράφεται από την εξίσωση

$$H^j[k] = \begin{cases} 0 & , k < f_c^{j-1} \\ \dfrac{k - f_c^{j-1}}{f_c^j - f_c^{j-1}} & , f_c^{j-1} \leq k \leq f_c^j \\ \dfrac{f_c^{j+1} - k}{f_c^{j+1} - f_c^j} & , f_c^j \leq k \leq f_c^{j+1} \\ 0 & , k > f_c^{j+1} \end{cases} . \tag{5.6}$$

Οι κεντρικές συχνότητες είναι ισοκατανεμημένες με βάση την κλίμακα mel. Έτσι, συμβολίζοντας τώρα για ευκολία με f_h και f_l την υψηλότερη και χαμηλότερη αντίστοιχα συχνότητα της συστοιχίας, με N το μέγεθος του διακριτού FFT και με F_s τη συχνότητα δειγματοληψίας, οι κεντρικές συχνότητες προκύπτουν σύμφωνα με την εξίσωση

$$f_c^j = \frac{N}{F_s} B^{-1}\left(B(f_l) + j\frac{B(f_h) - B(f_l)}{Q+1}\right) . \tag{5.7}$$

Εποπτικά, η εν λόγω συστοιχία παρουσιάζεται στο Σχήμα 5.2, για 24 τριγωνικά φίλτρα και συχνότητα δειγματοληψίας $Fs = 16kHz$. Στο συγκεκριμένο παράδειγμα, θεωρείται ότι $f_c^0 = 0$ και $f_c^{Q+1} = Fs/2$. Αυτό, αν και αποτελεί συνηθισμένη πρακτική, δεν είναι απαραίτητο, εφόσον συχνά οι πολύ μικρές ή πολύ υψηλές συχνότητες αγνοούνται, ώστε να απομονωθεί χαμηλόσυχνος και υψίσυχνος θόρυβος.



Σχήμα 5.2: Συστοιχία τριγωνικών φίλτρων για την εξαγωγή των MFCCs. Θεωρείται συχνότητα δειγματοληψίας $16kHz$, ενώ η συστοιχία αποτελείται από 24 φίλτρα.

Αντί τα φίλτρα να είναι κανονικοποιημένα ώστε να έχουν μοναδιαίο ύψος, μπορεί να είναι κανονικοποιημένα ώστε να έχουν μοναδιαίο εμβαδόν, ώστε για κάθε φίλτρο να ισχύει

$$\sum_{k=0}^{N-1} H^j[k] = 1 . \tag{5.8}$$

Στην περίπτωση αυτή, το j φίλτρο της συστοιχίας περιγράφεται από την εξίσωση

$$H^j[k] = \begin{cases} 0 & ,k < f_c^{j-1} \\ \dfrac{2\left(k - f_c^{j-1}\right)}{\left(f_c^j - f_c^{j-1}\right)\left(f_c^{j+1} - f_c^{j-1}\right)} & ,f_c^{j-1} \leq k \leq f_c^j \\ \dfrac{2\left(f_c^{j+1} - k\right)}{\left(f_c^{j+1} - f_c^j\right)\left(f_c^{j+1} - f_c^{j-1}\right)} & ,f_c^j \leq k \leq f_c^{j+1} \\ 0 & ,k > f_c^{j+1} \end{cases} . \tag{5.9}$$

Υπολογίζουμε εν συνεχεία το λογάριθμο της ενέργειας της απόκρισης του κάθε φίλτρου j με είσοδο το παραθυρωμένο πλαίσιο $s_i(n)$, έστω $G_i(j)$. Γνωρίζουμε πως η έξοδος ενός συστήματος στο πεδίο της συχνότητας ισούται με το γινόμενο της απόκρισης συχνότητας του συστήματος με το Fourier Μετασχηματισμό της εισόδου. Έχοντας αυτό κατά νου και εκμεταλλευόμενοι το θεώρημα του Parseval που συνδέει το τετράγωνο του σήματος με το φάσμα ισχύος του σύμφωνα με τη σχέση

$$\sum_{n=0}^{N-1} s[n]^2 = \frac{1}{N} \sum_{k=0}^{N-1} |S[k]|^2 , \tag{5.10}$$

άμεσα καταλήγουμε στον εξης υπολογισμό:

$$G_i(j) = \log\left\{ \frac{1}{N} \sum_{k=0}^{N-1} |S_i[k] \cdot H^j[k]|^2 \right\} = \log\left\{ \frac{2}{N} \sum_{k=0}^{N/2} |S_i[k] \cdot H^j[k]|^2 \right\} , \tag{5.11}$$

όπου $S_i[k]$ ο DFT N σημείων του $s_i[n]$. Η προσθετική σταθερά $\log\{2/N\}$ παραλείπεται από τους υπολογισμούς, ενώ συχνά χρησιμοποιείται και η ελαφρώς τροποποιημένη φόρμουλα υπολογισμού

$$G_i(j) = \log\left\{ \sum_{k=0}^{N/2} |S_i[k]|^2 \cdot |H^j[k]| \right\} . \tag{5.12}$$

Τέλος, τα MFCCs λαμβάνονται ως οι πρώτοι N_c συντελεστές του DCT μετασχηματισμού της ενέργειας $G_i(j)$ του κάθε πλαισίου. Πρακτικά, χρησιμοποιείται $N_c = 13$ ως μια τυπική επιλογή για εφαρμογές αναγνώρισης φωνής και για την ακρίβεια κρατουνται οι συντελεστές $2-13$. Στα παραπάνω 12 χαρακτηριστικά προστίθεται και η τετραγωνική ενέργεια του σήματος, όπως δίνεται από τη σχέση (5.10), αφού πρώτα περάσει στο λογαριθμικό πεδίο, η οποία επίσης μπορεί να παίξει σημαντικό ρόλο στη διαδικασία της Αναγνώρισης.

Η ιδέα του να πάρουμε cepstral χαρακτηριστικά βασίζεται στο γεγονός ότι για την Αναγνώριση θέλουμε όσο το δυνατόν χαρακτηριστικά τέτοια που να διαχωρίζουν τα διαφορετικά φωνήματα μεταξύ τους. Αυτά σχετίζονται με τη θέση των διαφόρων αρθρωτών της φωνητικής οδού και στοματικής και ρινικής κοιλότητας (με άλλα λόγια του φίλτρου του αρχικού σήματος της πηγής) και όχι από την πηγή (π.χ. ταλαντώσεις φωνητικών χορδών). Όπως ξέρουμε, το cepstrum είναι ένας αξιόπιστος τρόπος διαχωρισμού του φάσματος της πηγής από το φάσμα του φίλτρου [31].

Από την άλλη, η χρήση του DCT δικαιολογείται για λόγους αποσυσχέτισης και συμπίεσης των δεδομένων (αφού όπως είδαμε κρατάμε τις πρώτες μόνο συνιστώσες). Συγκεκριμένα, για

σήματα φωνής, ο DCT προσεγγίζει σε μεγάλο βαθμό τον Karhunen–Loève Μετασχηματισμό (KLT), που είναι βέλτιστος ως προς την ελαχιστοποίηση της ενέργειας του σφάλματος συμπίεσης [70]. Ακόμα, ο DCT παράγει διανύσματα πραγματικών τιμών, γεγονός που διευκολύνει πολύ την αποδοτική αποθήκευση και το χειρισμό τους.

Δύο πρακτικές λεπτομέρειες στην παραπάνω διαδικασία είναι οι τεχνικές του dithering και του liftering. Ως dithering είναι γνωστή η διαδικασία της προσθήκης τυχαίου θορύβου ενός bit στην κυματομορφή, που είναι σχεδόν ισοδύναμο με την προσθήκη μίας σταθεράς στο φάσμα της κυματομορφής. Η διαδικασία αυτή λαμβάνει χώρα ώστε να εξαλειφθεί η πιθανότητα ύπαρξης μηδενικών τιμών στο φάσμα, που θα οδηγήσει σε προβλήματα κατά τον υπολογισμό των αντίστοιχων λογαριθμικών τιμών.

Όσον αφορά στο liftering, πρόκειται για μια διαδικασία όπου κάθε MFCC x_i, όπως έχει εξαχθεί από την παραπάνω διαδικασία, πολλαπλασιάζεται με ένα βάρος w_i, ώστε να προκύψει ένα νέο χαρακτηριστικό y_i [71]. Πρόκειται, λοιπόν, για ένα γραμμικό μετασχηματισμό που σε μητρική μορφή περιγράφεται από την εξίσωση

$$\mathbf{y} = \mathbf{W}\mathbf{x}\,, \tag{5.13}$$

όπου $\mathbf{W} = \text{diag}\{w_1, w_2, \cdots, w_{N_c}\}$. Οι τιμές της διαγωνίου του πίνακα μετασχηματισμού $\mathbf{W}$ επιλέγονται με τρόπο τέτοιο ώστε όλα τα προκύπτοντα MFCCs να ανήκουν στο ίδιο εύρος τιμών. Κάτι τέτοιο είχε ιδιαίτερα οφέλη τις πρώτες μέρες της Αναγνώρισης Φωνής, όταν αυτή βασιζόταν στον αλγόριθμο Δυναμικής Χρονικής Στρέβλωσης (Dynamic Time Warping - DTW) και στις Ευκλείδειες αποστάσεις μεταξύ των διανυσμάτων χαρακτηριστικών. Πλέον, το liftering δεν έχει ιδιαίτερη πρακτική αξία, αλλά εξακολουθεί συχνά να εφαρμόζεται για ιστορικούς κυρίως λόγους.

5.2 Παραγώγιση των MFCCs

Για χρονικά παράθυρα μικρότερα των $100msec$, η δυνατότητα κατηγοριοποίησης των φωνητικών χαρακτηριστικών από τον άνθρωπο δεν είναι ικανοποιητική [34]. Έτσι, καθώς για τα short-term χαρακτηριστικά το μέγεθος του πλαισίου δεν ξεπερνά τα $32msec$, φαίνεται πως υπάρχει η ανάγκη το διάνυσμα χαρακτηριστικών να δίνει πληροφορία για ένα μεγαλύτερο χρονικό τμήμα του σήματος.

Ο πιο απλός τρόπος να αντιμετωπιστεί το παραπάνω θέμα είναι το διάνυσμα χαρακτηριστικών i να είναι επαυξημένο ώστε να περιλαμβάνει χαρακτηριστικά που αφορούν τόσο το πλαίσιο i, όσο και τα γειτονικά του πλαίσια ή τη σχέση του με τα γειτονικά του. Στο Κεφάλαιο 7 θα μελετηθεί το ζήτημα αυτό υπό άλλη σκοπιά και με μεγαλύτερη λεπτομέρεια, αλλά εδώ θα αναφέρουμε την πιο συχνά χρησιμοποιούμενη μέθοδο, που προτάθηκε για πρώτη φορά στο [72] και αφορά την επαύξηση του διανύσματος των MFCCs με δυναμικά χαρακτηριστικά που προσεγγίζουν τη συμπεριφορά των MFCCs στο χρόνο. Παρόλο που εδώ αναφερόμαστε σε MFCCs, η μέθοδος είναι πολύ γενικότερη.

Έτσι, λοιπόν, τα δυναμικά χαρακτηριστικά πρώτης τάξης (που συχνά ονομάζονται συντελεστές ταχύτητας και συμβολίζονται ως Δ) λαμβάνονται ως οι ορθογώνιοι πολυωνυμικοί συντελεστές πρώτης τάξης. Θεωρώντας τα διαδοχικά πλαίσια $\{\cdots, i-2, i-1, i, i+1, i+2, \cdots\}$, και υποθέτοντας ότι έχουν εξαχθεί 13 MFCCs $x(k),\ k = 1, 2, \cdots, 13$ για κάθε πλαίσιο, οι

συντελεστές ταχύτητας για το πλαίσιο i υπολογίζονται ως

$$\Delta x_i(k) = \frac{\sum_{m=-M}^{M} m \cdot x_{i+m}(k)}{\sum_{m=-M}^{M} m^2}. \tag{5.14}$$

Η σταθερά M δηλώνει το περιβάλλον της παραγώγισης, δηλαδή πόσα γειτονικά πλαίσια λαμβάνονται υπόψιν κατά τον υπολογισμό των δυναμικών χαρακτηριστικών. Συνήθως το M κυμαίνεται στο διάστημα $[1, 10]$ [72].

Η σχέση (5.14) μπορεί να επαναχρησιμοποιηθεί για να ληφθούν δυναμικά χαρακτηριστικά δεύτερης τάξης (συντελεστές επιτάχυνσης ή ΔΔ συντελεστές), τρίτης τάξης, κ.λπ. Για παράδειγμα, εάν ένα σύστημα αναγνώρισης χρησιμοποιεί 13 MFCCs, μαζί με τους αντίστοιχους Δ και ΔΔ συντελεστές, τότε το μέγεθος του διανύσματος χαρακτηριστικών είναι $3 \cdot 13 = 39$.

5.3 Επίδραση των Επιμέρους Παραμέτρων κατά την Εξαγωγή των MFCCs

Στο σημείο αυτό θα εξετασθεί η επίδραση συγκεκριμένων παραμέτρων στην τελική απόδοση του συστήματος όταν γίνεται χρήση MFCCs. Για την εξαγωγή των χαρακτηριστικών έγινε χρήση του κώδικα που παρέχεται στο [73] για MATLAB, ενώ κάθε περαιτέρω μετασχηματισμός στα χαρακτηριστικά (όπως ο υπολογισμός των Δ) γίνεται απευθείας μέσω των προγραμμάτων που παρέχονται στο Kaldi.

Τα 13 MFCCs υπολογίζονται βάσει του DCT των συντελεστών που προκύπτουν από τη σχέση (5.11) (χωρίς τη σταθερά $\log\{2/N\}$), ενώ η συστοιχία φίλτρων προκύπτει από φίλτρα κανονικοποιημένα ως προς το εμβαδόν, σύμφωνα με τη σχέση (5.9). Επίσης, για την εξαγωγή των χαρακτηριστικών λαμβάνει χώρα liftering σύμφωνα με τη σχέση (5.13), όπου $w_1 = 1$ και $w_i = (i-1)^{0.6}$ για $2 \leq i \leq 13$. Όποτε δεν αναφέρεται διαφορετικά, γίνεται χρήση παραθύρου Hamming μήκους $32msec$ που κινείται ανά $10msec$, η συστοιχία αποτελείται από 40 φίλτρα και επιδρά στο εύρος συχνοτήτων $[0Hz, 8000Hz]$. Ακόμα, λαμβάνονται οι Δ και ΔΔ συντελεστές μέσω της σχέσης (5.14), όπου $M = 4$.

Στον Πίνακα 5.1 παρουσιάζεται η απόδοση του συστήματος καθώς μεταβάλλεται το μήκος του παραθύρου σε τυπικές τιμές που προτείνονται και όταν όλες οι άλλες παράμετροι παραμένουν σταθερές. Φαίνεται πως η ακριβής τιμή του μήκους του παραθύρου, όταν βρίσκεται στο συνηθισμένο εύρος των $15msec - 32msec$ δεν επηρεάζει σημαντικά τα ποσοστά σφάλματος του συστήματος, τόσο για αναγνώριση από κοντά, όσο και από απόσταση. Ωστόσο, οι τιμές κοντά στα $32msec$ φαίνεται να αποτελούν ασφαλέστερη επιλογή.

	15	20	25	32
CT1	33.86	33.13	**32.95**	33.35
OA6	82.04	84.40	83.14	**81.81**

Πίνακας 5.1: PER (%) καθώς το μήκος του παραθύρου λαμβάνει τυπικές τιμές ανάμεσα στα $15msec$ και τα $32msec$.

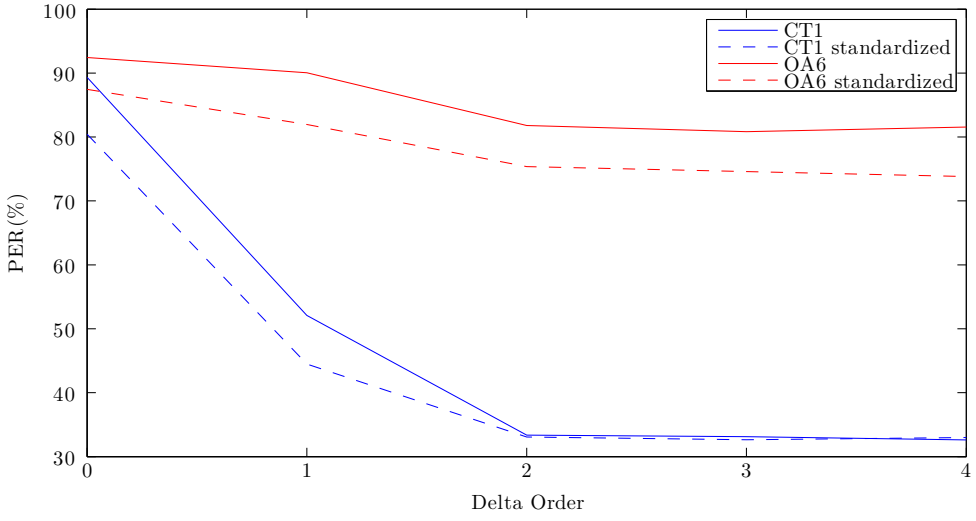
Από τη στιγμή που έχει αποφασιστεί εάν τα φίλτρα της συστοιχίας θα είναι κανονικοποιημένα ως προς το ύψος ή ως προς το εμβαδόν, κάτι που πρακτικά εξάλλου δεν έχει μεγάλη

σημασία, η συστοιχία ορίζεται πλήρως μέσω τριών παραμέτρων· του αριθμού των φίλτρων, της ελάχιστης συχνότητας αποκοπής (του πρώτου φίλτρου) και της μέγιστης συχνότητας αποκοπής (του τελευταίου φίλτρου). Η ελάχιστη συχνότητα αποκοπής ορίζεται κοντά στο 0 και η μέγιστη κοντά στη συχνότητα Nyquist. Ωστόσο, πολλές φορές προτείνεται αυτές οι χαρακτηριστικές συχνότητες να μην ταυτίζονται με τις προαναφερθείσες τιμές, ώστε να αποκόπτονται τμήματα του φάσματος που πιθανώς να επικρατεί ο θόρυβος. Ωστόσο, σύμφωνα με τον Πίνακα 5.2, φαίνεται πως, τουλάχιστον για τα υπό εξέταση δεδομένα, κάτι τέτοιο δεν είναι καλή πρακτική, εφόσον τα καλύτερα αποτελέσματα λαμβάνονται όταν η συστοιχία καλύπτει όλο το φάσμα.

	[0, 8000]	[20, 7800]	[130, 6800]
CT1	**33.35**	33.60	34.75
OA6	**81.81**	84.03	84.61

Πίνακας 5.2: PER (%) καθώς μεταβάλλονται σε τυπικές τιμές η ελάχιστη και η μέγιστη συχνότητα αποκοπής (σε $Hertz$) της συστοιχίας των φίλτρων για την εξαγωγή των MFCCs. Η συχνότητα Nyquist ισούται με $8kHz$.

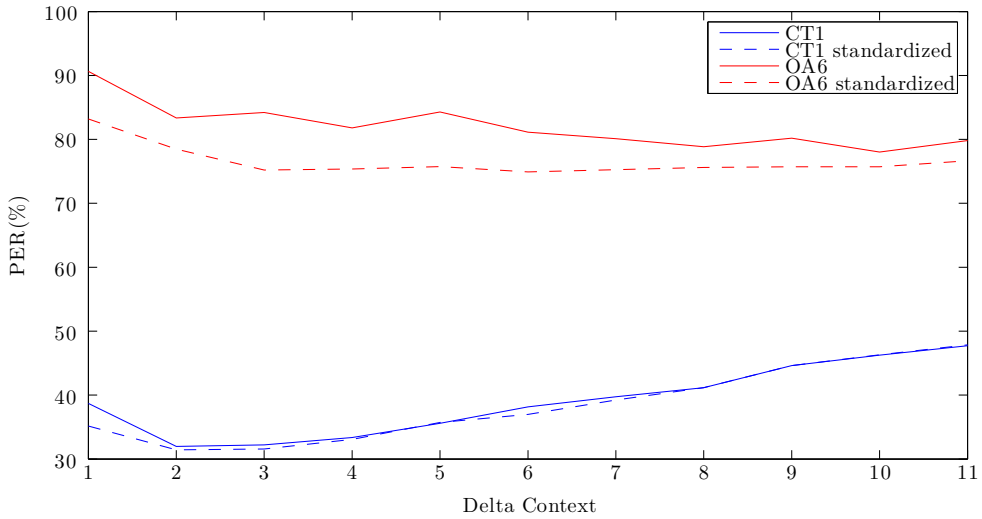
Εν συνεχεία, μελετάται η επίδραση της παραγώγισης στην απόδοση του συστήματος. Για κάθε πείραμα, λήφθηκαν δύο ξεχωριστά σύνολα χαρακτηριστικών, το ένα εκ των οποίων εξήχθη αφότου κάθε σήμα πριν το στάδιο της προέμφασης κανονικοποιήθηκε ώστε να έχει μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση. Στα Σχήματα 5.3 και 5.4 παρουσιάζονται αντίστοιχα η επίδραση που έχει ο βαθμός της παραγώγισης των χαρακτηριστικών και το χρονικό παράθυρο που λαμβάνεται υπόψη κατά την παραγώγιση. Σημειώνεται πως βαθμός παραγώγισης 3, για παράδειγμα, σημαίνει πως το διάνυσμα χαρακτηριστικών αποτελείται από τα MFCCs, τους Δ, τους ΔΔ και τους ΔΔΔ συντελεστές.



Σχήμα 5.3: PER (%) καθώς αυξάνεται η μέγιστη τάξη δυναμικών χαρακτηριστικών που περιλαμβάνονται στο τελικό διάνυσμα χαρακτηριστικών.

Φαίνεται πως η προσθήκη τόσο των Δ, όσο και των ΔΔ συντελεστών είναι εξαιρετικά ευεργετική για την αναγνώριση. Ωστόσο, η προσθήκη των ΔΔΔ ή και μεγαλύτερης τάξης δυναμικών συντελεστών δίνει πολύ μικρή βελτίωση στην απόδοση του συστήματος σε σύ-

γκριση με τις επιπλέον απαιτησεις σε μνήμη που εισάγει. Μην ξεχνάμε πως η προσθήκη κάθε τάξης δυναμικών συντελεστών συνεπάγεται αύξηση του διανύσματος χαρακτηριστικών κατά τον αριθμό των MFCCs που έχουν εξαχθεί, δηλαδή στην περίπτωσή μας κατά 13. Σε κάθε περίπτωση, η κανονικοποίηση των σημάτων φωνής φαίνεται να είναι ευεργετική, καθώς οδηγεί σε μικρότερα σφάλματα.



Σχήμα 5.4: PER (%) καθώς αυξάνεται το χρονικό παράθυρο που λαμβάνεται υπόψιν κατά τον υπολογισμό των δυναμικών χαρακτηριστικών. Ως Delta Context συμβολίζεται η σταθερά M της σχέσης (5.14).

Ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη του βέλτιστου παραθύρου για τον υπολογισμό των δυναμικών χαρακτηριστικών. Όσον αφορά στα καθαρά σήματα, φαίνεται πως παρουσιάζεται ένα σαφές βέλτιστο για παράθυρο μήκους 5 πλαισίων (δηλαδή για $M = 2$ στη σχέση (5.14)), ενώ αυξάνοντας το παράθυρο να αποτελέσματα συνεχώς χειροτερεύουν. Αντιθέτως, κάτι τέτοιο δε συμβαίνει στην περίπτωση των δεδομένων όπου υπεισέρχεται θόρυβος. Όταν τα σήματα δεν είναι κανονικοποιημένα στο πεδίο του χρόνου, τότε το ποσοστό σφάλματος ακολουθεί μία γενικώς πτωτική τάση όσο αυξάνεται το παράθυρο υπολογισμού. Όταν, όμως, έχει προηγηθεί κανονικοποίηση, τότε ένα παράθυρο μήκους 7 πλαισίων φαίνεται να είναι το "βέλτιστο", χωρίς, ωστόσο αξιοσημείωτες μεταβολές στο PER αυξάνοντας περαιτέρω το παράθυρο.

5.4 Εφαρμογή Cepstral Mean (& Variance) Normalization

Μία πολύ απλή τεχνική που έχει προταθεί στη βιβλιογραφία για τη μείωση των αρνητικών επιδράσεων του θορύβου είναι η Αναφασματική Κανονικοποίηση Μέσου (Cepstral Mean Normalization - CMN), που πολλές φορές ονομάζεται και Αναφασματική Αφαίρεση Μέσου (Cepstral Mean Subtraction - CMS). Αρχικά, χρησιμοποιήθηκε για να αντισταθμιστούν οι διαταραχές που οφείλονταν στις ηχογραφήσεις με διαφορετικά μικρόφωνα, αλλά φάνηκε πως ή τεχνική αυτή έχει ευεργετικά αποτελέσματα ακόμα και χωρίς να αλλάζει το κανάλι επικοινωνίας (ή καταγραφής) [74].

Γενικότερα, το CMN βοηθάει όταν στα δεδομένα υπάρχει συνελικτικός θόρυβος, στον οποίο περιλαμβάνονται διαταραχές που οφείλονται στα ιδιαίτερα χαρακτηριστικά του μικροφώ-

νου, αλλά επίσης διαταραχές λόγω αντήχησης, όπου το καθαρό ηχητικό σήμα συνελίσσεται με την κρουστική απόκριση του μικροφώνου ή του δωματίου αντίστοιχα. Λαμβάνοντας υπόψιν τη συνελικτική ιδιότητα του Fourier Μετασχηματισμού, καθώς και την ιδιότητα του λογαρίθμου $\log(x \cdot y) = \log x + \log y$, προκύπτει πως συνέλιξη στο πεδίο του χρόνου ισοδυναμεί με πρόσθεση στο πεδίο της ανασυχνότητας (quefrency), δηλαδή στο πεδίο που οδηγεί η εκτίμηση του αναφάσματος. Φορμαλιστικά, αν $y(n) = x(n) * h(n)$, τότε $Y(q) = X(q) + H(q)$, όπου με q συμβολίζεται η ανασυχνότητα. Θεωρώντας πως η κρουστική απόκριση $h(n)$ παραμένει σταθερή καθόλη τη διάρκεια της υπό μελέτη εκφοράς, για καθένα πλαίσιο i από τα F θα είναι $Y_i(q) = X_i(q) + H(q)$. Αφαιρώντας τον αριθμητικό μέσο όλων των πλαισίων προκύπτει

$$\begin{aligned} Y_i'(q) &= Y_i(q) - \frac{1}{F}\sum_{i=1}^{F} Y_i(q) \\ &= X_i(q) - \frac{1}{F}\sum_{i=1}^{F} X_i(q)\,, \end{aligned}$$

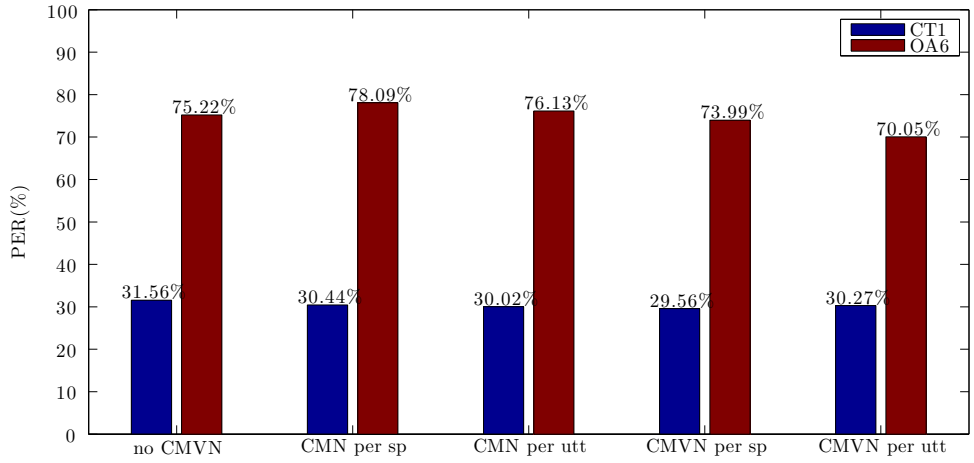
οπότε φαίνεται πως έχει εξαλειφθεί η επίδραση του $h(n)$. Εάν εκτός από την αφαίρεση του μέσου, κανονικοποιήσουμε τους αναφασματικούς συντελεστές και ως προς την τυπική απόκλιση, τότε ο προκύπτων μετασχηματισμός ονομάζεται Αναφασματική Κανονικοποίηση Μέσου και Διακύμανσης (Cepstral Mean and Variance Normalization - CMVN).

Για εφαρμογές αναγνώρισης με πολλαπλούς ομιλητές, προτείνεται ενίοτε το CMVN να γίνεται ανά ομιλητή και όχι ανά εκφορά, δηλαδή τα στατιστικά στοιχεία που χρειάζονται για την κανονικοποίηση να συλλέγονται από όλα τα πλαίσια όλων των εκφορών που ειπώθηκαν από ένα συγκεκριμένο ομιλητή (προφανώς, αυτό προϋποθέτει ότι υπάρχει η πληροφορία του ομιλητή). Κατ' αυτόν τον τρόπο, εξαλείφονται πιθανές διαφορές στα αναφασματικά χαρακτηριστικά που οφείλονται στα ιδιαίτερα χαρακτηριστικά της φωνητικής οδού του κάθε ομιλητή.

Σύμφωνα με τα αποτελέσματα του Σχήματος 5.5, τα καλύτερα αποτελέσματα για θορυβώδη δεδομένα λαμβάνονται μέσω CMVN ανά εκφορά, ενώ για καθαρά δεδομένα για CMVN ανά ομιλητή. Καθώς οι ηχογραφήσεις λαμβάνονται υπό διαφορετικές συνθήκες για τις διάφορες εκφορές και με διαφορετικές θέσεις των ομιλητών, είναι σφάλμα να θεωρηθεί ότι η επίδραση του συνελικτικού θορύβου παραμάνει σταθερή. Για τα καθαρά δεδομένα, ωστόσο, όπου δεν υπάρχουν τα παραπάνω προβλήματα, φαίνεται πως πράγματι η λήψη στατιστικών στοιχείων ανά ομιλητή είναι η καλύτερη επιλογή. Τα συγκεκριμένα πειράματα έγιναν με MFCC+Δ+ΔΔ, με κανονικοποιημένα δείγματα στο πεδίο του χρόνου, ενώ για τις παραγωγίσεις έγινε χρήση παραθύρου μήκους 7 πλαισίων, που από το Σχήμα 5.4 φαίνεται να είναι μια συνετή επιλογή, τόσο για καθαρά, όσο και για θορυβώδη δεδομένα. Σημειώνεται ότι σε πειράματα που έγιναν χωρίς να έχει προηγηθεί κανονικοποίηση στο πεδίο του χρόνου (δε φαίνονται εδώ), είχαν ευεργετικά αποτελέσματα και οι 4 παραλλαγές του CM(V)N, αλλά και πάλι το CMVN ανά εκφορά έδινε τα καλύτερα αποτελέσματα για το μικρόφωνο OA6.

5.5 Επισπεύδοντας την Παραγώγιση: Delta-Spectral Cepstral Coefficients

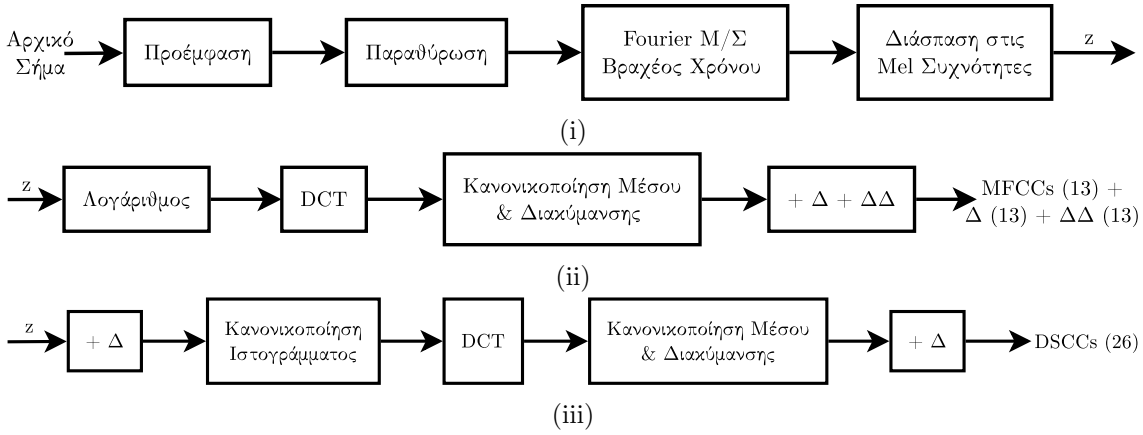
Μία ιδέα που έχει προταθεί για εύρωστη αναγνώριση αφορά στον υπολογισμό των δυναμικών χαρακτηριστικών όχι στο αναφασματικό, αλλά στα φασματικό πεδίο, με τα προκύπτοντα χαρακτηριστικά να καλούνται Δέλτα-Φασματικοί Αναφασματικοί Συντελεστές (Delta-Spectral Cepstral Coefficients - DSCCs) [75].



Σχήμα 5.5: Επίδραση του Cepstral Mean (& Variance) Normalization στην αναγνώριση. Εξετάζονται από αριστερά προς τα δεξιά οι περιπτώσεις χωρίς CMVN, με CMN ανά ομιλητή, με CMN ανά εκφορά, με CMVN ανά ομιλητή και με CMVN ανά εκφορά.

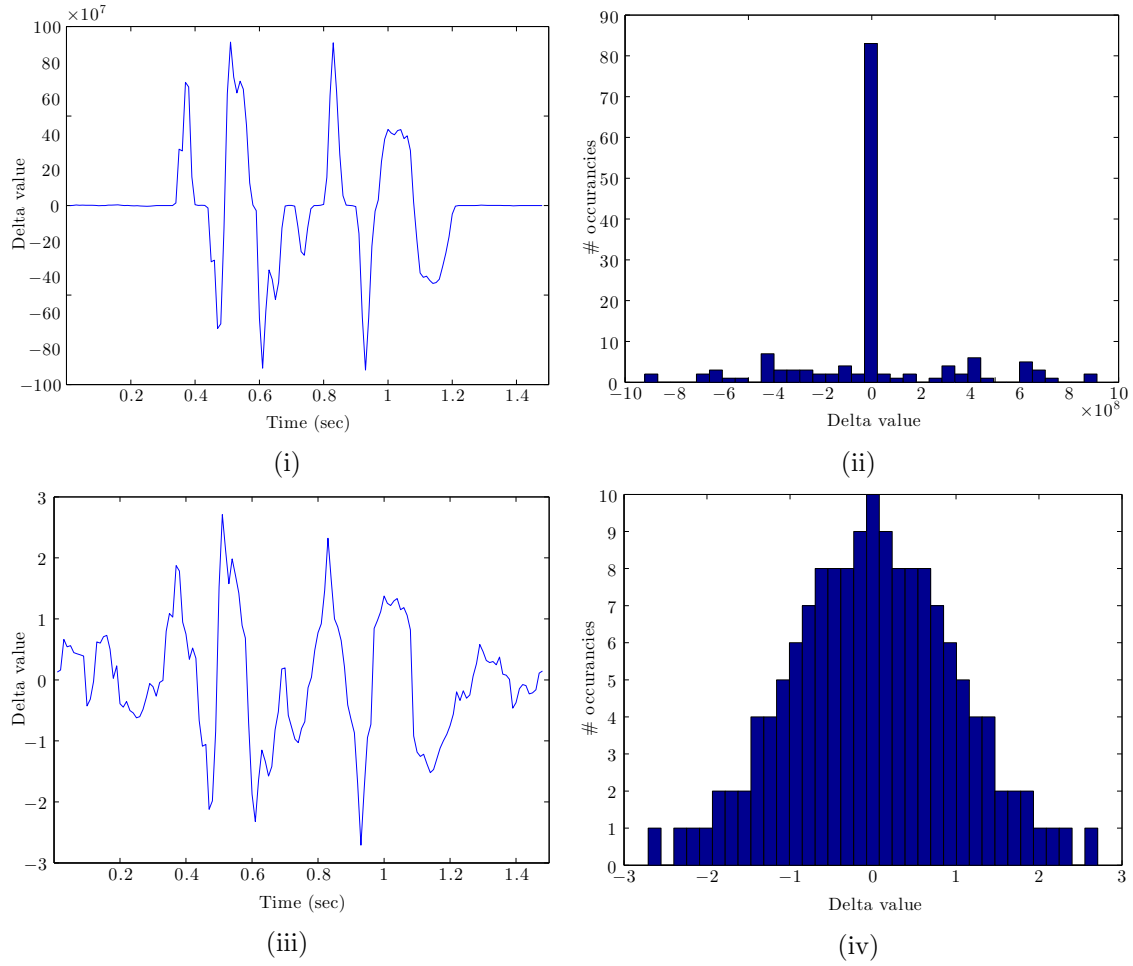
Η ιδέα στηρίζεται στο γεγονός ότι τα φασματικά χαρακτηριστικά της φωνής μεταβάλλονται πολύ πιο γρήγορα από τα φασματικά χαρακτηριστικά του θορυβώδους υποβάθρου. Η διαφορετική αυτή δυναμική μεταξύ φωνής και υποβάθρου υποστηρίζεται ότι μπορεί να αντικατοπτριστεί καλύτερα απευθείας στο φασματικό πεδίο, παρά στο αναφασματικό, δηλαδή προτού λάβει χώρα η μη-γραμμικότητα του λογαρίθμου και ο DCT. Όπως και η ίδια η διαδικασία της παραγώγισης όπως την έχουμε δει, έτσι και η παραγώγιση στο φασματικό πεδίο, είναι μια μέθοδος γενική, αλλά εδώ θα τη δούμε σε συνδυασμό με τα MFCCs.

Στο Σχήμα 5.6 παρουσιάζεται η ροή εργασίας για την εξαγωγή των DSCCs, σε αντιπαράθεση της εξαγωγής των απλών Δ και ΔΔ συντελεστών που αποτελούν το τελευταίο στάδιο της διαδικασίας. Τα DSCCs χρησιμοποιούνται σε συνδυασμό με τα MFCCs, δημιουργώντας, έτσι, ένα διάνυσμα 39 χαρακτηριστικών, ως συνήθως.



Σχήμα 5.6: Διαγραμματική αναπαράσταση της ροής εργασίας για την εξαγωγή των MFCCs και των DSCCs. (i) Πρώτα στάδια της διαδικασίας. (ii) Εξαγωγή των MFCCs και των κλασικών Δ και ΔΔ συντελεστών. (iii) Εξαγωγή των DSCCs. [*εικόνα προσαρμοσμένη από [75]*]

Η κανονικοποίηση ιστογράμματος λαμβάνει χώρα διότι η παραγώγιση στο φασματικό πεδίο έχει ως αποτέλεσμα τα δυναμικά χαρακτηριστικά να έχουν ένα πολύ μεγάλο εύρος τιμών, γεγονός που από μόνο του είναι μη επιθυμητό για τους σκοπούς της αναγνώρισης φωνής, με την πλειοψηφία, όμως, των συντελεστών να λαμβάνουν τιμές κοντά στη γειτονιά του μηδέν. Μετά την εν λόγω κανονικοποίηση, που εφαρμόζεται ανά εκφορά, οι τιμές των συντελεστών διαμορφώνονται έτσι ώστε να ακολουθούν κανονική κατανομή με μηδενική μέση τιμή και μοναδιαία απόκλιση. Η επίδραση της γκαουσιανής αυτής κανονικοποίησης παρουσιάζεται εποπτικά στο Σχήμα 5.7.



Σχήμα 5.7: Επίδραση της κανονικοποίησης ιστογράμματος κατά την εξαγωγή τψν DSCCs. (i) Τροχιά των DSCCs για το 10ο φίλτρο της συστοιχίας όπως προκύπτουν πριν την κανονικοποίηση ιστογράμματος. (ii) Ιστόγραμμα των τιμών του (i). (iii) Τροχιά των DSCCs για το 10ο φίλτρο της συστοιχίας όπως προκύπτουν μετά την κανονικοποίηση ιστογράμματος. (iv) Ιστόγραμμα των τιμών του (iii). Πρόκειται παντού για τους Δ συντελεστές πρώτης τάξης για την εκφορά της φράσης "*Έχω επικαλεστεί κάθε είδους επιχειρήματα για να το αποδείξουμε*" από το μικρόφωνο CT1.

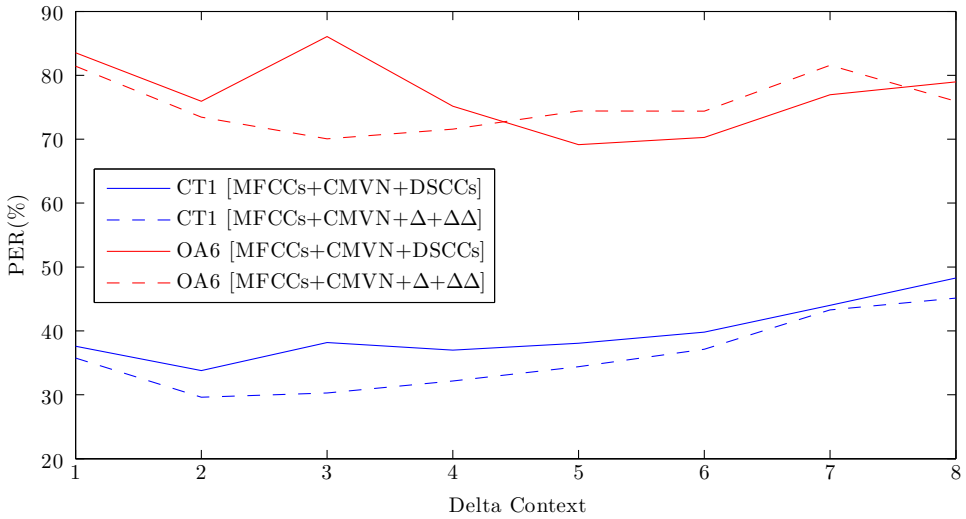
Για την υλοποίηση της μεθόδου χρησιμοποιούμε, όπως προηγουμένως, τον κώδικα που παρέχεται στο [73] για την εξαγωγή των MFCCs, με χρήση των σχετικών συναρτήσεων του κώδικα που παρέχεται στο [75][1]. Παρατηρήθηκε και πάλι ότι η τελική απόδοση είναι βελτιωμένη στην περίπτωση που χρησιμοποιούνται κανονικοποιημένα σήματα, οπότε προηγείται η

[1]Σημειώνεται πως απενεργοποιήθηκε η επιπλέον κανονικοποίηση που προτείνεται μετά τον υπολογισμό των

κανονικοποίησή τους στο πεδίο του χρόνου ώστε να έχουν μηδενική μέση τιμή και μοναδιαία απόκλιση. Τονίζεται ότι αντί της σχέσης (5.14), για την εξαγωγή των Δ συντελεστών κατά τη ροή εργασίας των DSCCs, προτείνεται η τροποποιημένη σχέση

$$\Delta x_i(k) = x_{i+M} - x_{i-M}\,. \tag{5.15}$$

Σχετικά αποτελέσματα, για διάφορες τιμές της παραμέτρου M παρουσιάζονται στο Σχήμα 5.8.

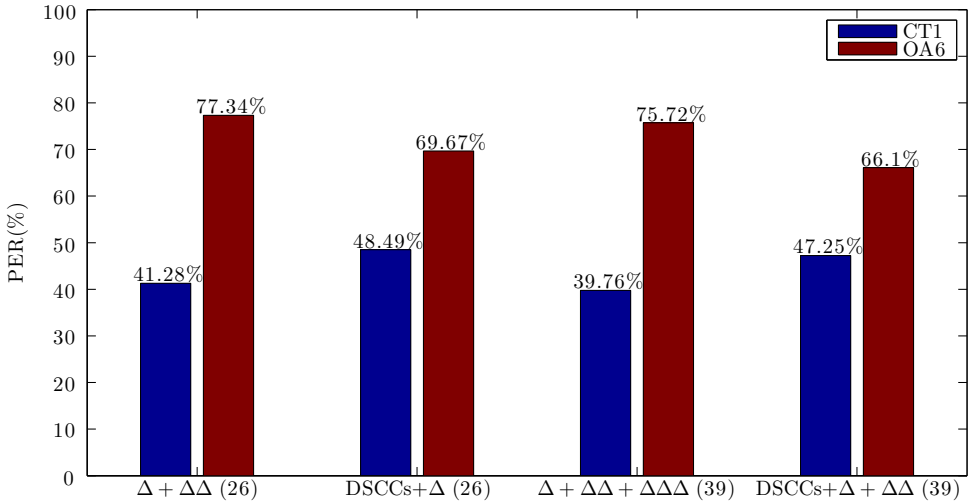


Σχήμα 5.8: PER (%) καθώς αυξάνεται το χρονικό παράθυρο που λαμβάνεται υπόψιν κατά τον υπολογισμό των δυναμικών χαρακτηριστικών, όταν αυτά ταυτίζονται με τα DSCCs ή με τους κλασικούς Δ και ΔΔ συντελεστές. Ως Delta Context συμβολίζεται η σταθερά M της σχέσης (5.14) για τους κλασικούς Δ και ΔΔ συντελεστές και της σχέσης (5.15) για τα DSCCs, ενώ σε κάθε περίπτωση λαμβάνει χώρα CMVN ανά εκφορά.

Σε καθαρές συνθήκες, είναι εμφανές από το Σχήμα ότι η χρήση των κλασικών δυναμικών συντελεστών, συνδυασμένων με το στατικό διάνυσμα των MFCCs, είναι προτιμότερη από τα DSCCs. Ωστόσο, υπό συνθήκες θορύβου, όπως στις συνθήκες του μικροφώνου OA6, και κατόπιν προσεκτικής αναζήτησης της κατάλληλης παραμέτρου M, η χρήση των DSCCs δίνει ελαφρώς μικρότερο PER (69.16% έναντι 70.05% στην περίπτωση των κλασικών Δ και ΔΔ), με τα ποσοστά, βέβαια, να είναι συγκρίσιμα. Αυτό που παρουσιάζει ιδιαίτερο ενδιαφέρον είναι το γεγονός ότι τα DSCCs επιφέρουν χαμηλά ποσοστά σφάλματος χωρίς να είναι συνδυασμένα με τα MFCCs (από τα οποία προφανώς υπολογίζονται), όπως φαίνεται στο Σχήμα 5.9.

Συγκεκριμένα, υπολογίζουμε τα DSCCs πρώτης τάξης (13 συντελεστές), σύμφωνα με το Σχήμα 5.6iii και με τη σχέση (5.15) για $M = 5$, καθώς και τους κλασικούς Δ συντελεστές πρώτης τάξης (13 συντελεστές), σύμφωνα με το Σχήμα 5.6ii και με τη σχέση (5.14) για $M = 3$. Επί των 13 αυτών συντελεστών επανυπολογίζονται τα δυναμικά χαρακτηριστικά, σύμφωνα με τη σχέση (5.14) για $M = 5$ και $M = 3$ αντίστοιχα, ώστε να προκύψουν σύνολα 26 συντελεστών. Με επαναχρησιμοποίηση της ίδιας σχέσης και με τις ίδιες τιμές για την παράμετρο M, προκύπτουν σύνολα 39 συντελεστών, δηλαδή όσο το de facto μήκος του διανύσματος χαρακτηριστικών για αναγνώριση φωνής, που ουσιαστικά περιέχουν μόνο δυναμικούς

DSCCs, εφόσον μάλιστα στις περισσότερες περιπτώσεις οδηγούσε σε ελαφρώς χειρότερα αποτελέσματα, ώστε να είμαστε συνεπείς με τη ροή του Σχήματος 5.6.



Σχήμα 5.9: PER (%) όταν γίνεται χρήση μόνο δυναμικών χαρακτηριστικών (είτε κλασικών Δ συντελεστών, είτε DSCCs), χωρίς αυτά να συνδυάζονται με τα στατικά MFCCs.

συντελεστές μέχρι και τρίτης τάξης. Όπως, λοιπόν, παρουσιάζεται στο Σχήμα 5.9, αλλά και συγκρίνοντας με το Σχήμα 5.8, τα DSCCs+Δ+ΔΔ όχι μόνο παρουσιάζουν μια απόλυτη μείωση στο PER της τάξης του 10% σε σύγκριση με τα Δ+ΔΔ+ΔΔΔ, αλλά δίνουν σημαντικά χαμηλότερο σφάλμα τόσο από τα MFCCs+Δ+ΔΔ, όσο και από τα MFCCs+DSCCs. Αξιωσημείωτο, ακόμα, είναι το γεγονός ότι με τους 26 μόνο συντελεστές των DSCCs+Δ επιτυγχάνεται χαμηλότερο PER σε σχέση με τους 39 κλασικούς Δ+ΔΔ+ΔΔΔ συντελεστές, αλλά και με τους 39 MFCCs+Δ+ΔΔ. Όλα αυτά, σαφώς, αφορούν μόνο τις συνθήκες του μικροφώνου OA6, καθώς στην περίπτωση των καθαρών συνθηκών του CT1, τα στατικά MFCCs φαίνεται να είναι απαραίτητα.

Κεφάλαιο 6

Perceptual Linear Predictive (PLP) Ανάλυση και Παραλλαγές

6.1 PLP Ανάλυση του Σήματος Φωνής και Εξαγωγή Χαρακτηριστικών

Ένα ακόμα σύνολο χαρακτηριστικών με αρκετά ευρεία χρήση που ανήκει στην κατηγορία των short-term χαρακτηριστικών είναι εκείνο που προκύπτει από Γραμμική Πρόβλεψη βασισμένη στην Αντίληψη (Perceptual Linear Prediction - PLP), με τα προκύπτοντα χαρακτηριστικά να ονομάζονται PLPs [76]. Τα PLPs βασίζονται στις ίδιες θεμελιώδεις αρχές που βασίζονται και τα MFCCs, αλλά στην PLP ανάλυση γίνεται προσπάθεια ακριβέστερης προσέγγισης των χαρακτηριστικών της ανθρώπινης ακοής.

Αφετηρία της PLP ανάλυσης υπήρξε η Γραμμική Πρόβλεψη που χρησιμοποιούταν τα πρώτα χρόνια της Αναγνώρισης Φωνής. Μέσω της Γραμμικής Πρόβλεψης μπορεί να γίνει μια ικανοποιητική προσέγγιση των πόλων της συνάρτησης μεταφοράς που περιγράφει το σύστημα παραγωγής ανθρώπινης φωνής, οι οποίοι αντιστοιχούν στους λεγόμενους φωνοσυντονισμούς (formants), δηλαδή στους χαρακτηριστικούς συντονισμούς της φωνητικής οδού [31]. Τα formants μεταβάλλονται αναλόγως του σχηματισμού που λαμβάνει η φωνητική οδός για την παραγωγή των διαφόρων ήχων· οπότε, εκτιμώντας τα, ουσιαστικά μπορούμε να εκτιμήσουμε τον ήχο προς αναγνώριση. Ένα από τα κύρια μειονεκτήματα, ωστόσο, της Γραμμικής Πρόβλεψης είναι ότι το προκύπτον μοντέλο προσεγγίζει το φάσμα του σήματος φωνής με την ίδια ακρίβεια σε όλες τις συχνότητες. Αντιθέτως, η ευαισθησία της ανθρώπινης ακοής μειώνεται με την αύξηση της συχνότητας, ιδίως πάνω από τα $800Hz$.

Η PLP ανάλυση, λοιπόν, προσπαθεί να αντισταθμίσει το παραπάνω μειονέκτημα. Μάλιστα, σύμφωνα με το [76], η Γραμμική Πρόβλεψη θα δώσει διαφορετικά αποτελέσματα για δύο εκφορές με διαφορετικά φάσματα, αλλά με την ίδια γλωσσική πληροφορία, ενώ η PLP ανάλυση θα δώσει παρόμοια αποτελέσματα, γεγονός που έχει προφανείς θετικές συνέπειες στην αναγνώριση φωνής.

Όπως και κατά την εξαγωγή των MFCCs, το σήμα πρώτα παραθυρώνεται σε επικαλυπτόμενα πλαίσια με παράθυρο Hamming και εν συνεχεία λαμβάνεται το φάσμα ισχύος του μέσω του DFT. Χωρίς να περάσουμε στο λογαριθμικό πεδίο, καταλήγουμε στη σχέση (6.1) κατ'

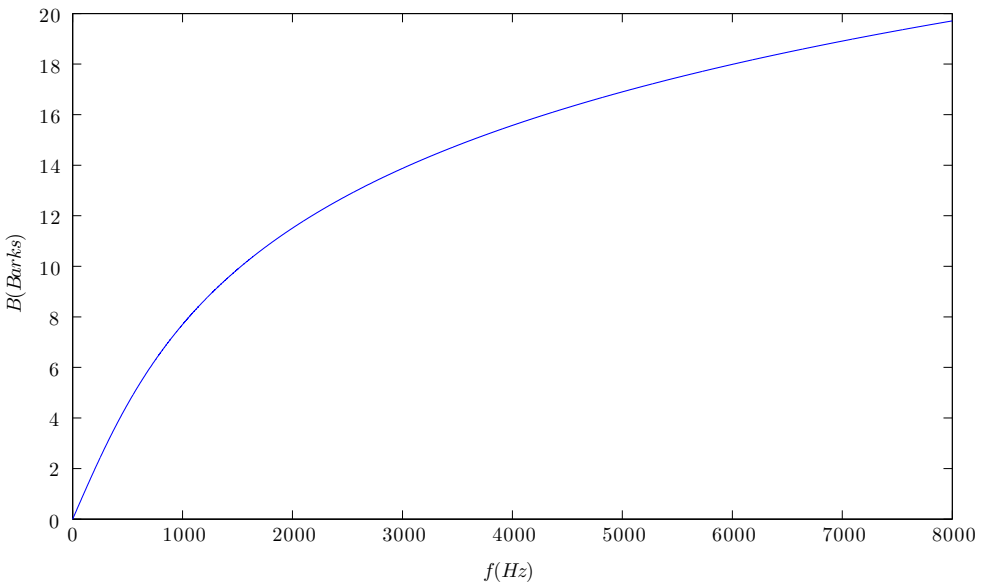
αντιστοιχία της σχέσης (5.12).

$$G_i(j) = \sum_{k=0}^{N/2} \left\{ |S_i[k]|^2 \cdot |H^j[k]| \right\} \tag{6.1}$$

Στην περίπτωση των PLPs, όμως, η μη-γραμμική ψυχοακουστική κλίμακα που χρησιμοποιείται είναι η κλίμακα $Bark$ και όχι η mel. Η αναλυτική φόρμουλα που συνδέει μια συχνότητα εκφρασμένη σε $Hertz$ με την αντίστοιχη εκφρασμένη σε $Barks$ δεν είναι μοναδική, αλλά αυτή που προτείνεται στο [76] είναι η (6.2)-(6.3), το αποτέλεσμα της οποίας απεικονίζεται γραφικά στο Σχήμα 6.1.

$$B = 6 \operatorname{arcsinh}\left(\frac{f}{600}\right) = 6 \log\left\{\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right\} \tag{6.2}$$

$$f = 600 \sinh\left(\frac{B}{6}\right) \tag{6.3}$$



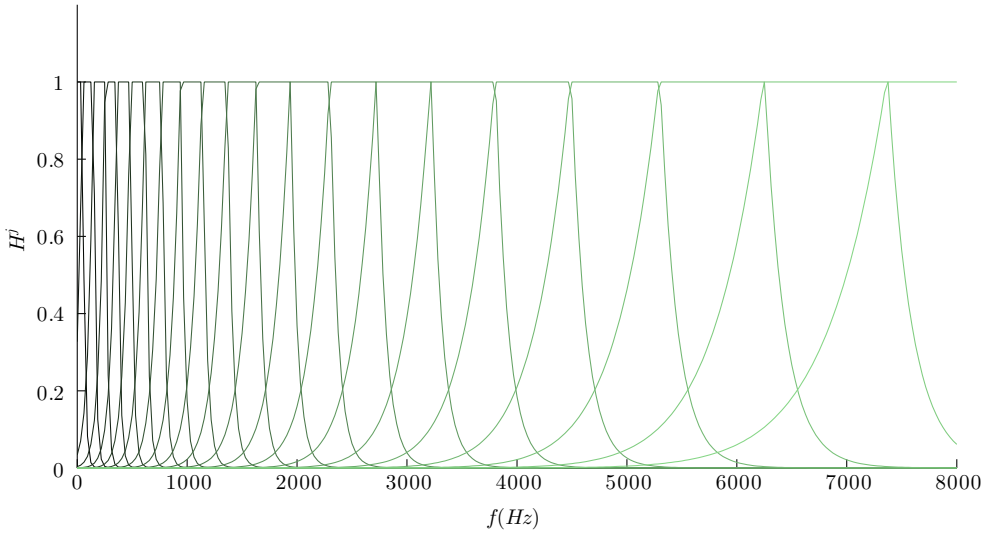
Σχήμα 6.1: Μετατροπή των συχνοτήτων από την κλίμακα $Hertz$ στην κλίμακα $Bark$.

Οι κεντρικές συχνότητες των φίλτρων της συστοιχίας που δημιουργείται είναι ομοιόμορφα κατανεμημένες στην κλίμακα $Bark$ με την κεντρική συχνότητα του πρώτου φίλτρου να είναι στα $0Hz$ και κάθε φίλτρο να απέχει από τα γειτονικά του κατά $1Bark$. Έτσι, εάν για παράδειγμα η συχνότητα δειγματοληψίας είναι $16kHz$, θα χρειαστούν 21 φίλτρα, το τελευταίο από τα οποία θα έχει κεντρική συχνότητα $8398Hz$, δηλαδή λίγο παραπάνω από τη συχνότητα Nyquist. Όλα τα φίλτρα έχουν το ίδιο σχήμα στην κλίμακα $Bark$, το οποίο περιγράφεται από

την εξίσωση (6.4), εάν θεωρηθεί η κεντρική του συχνότητα του φίλτρου j ίση με B_c^j.

$$H^j(B) = \begin{cases} 0 & , B - B_c^j < -1.3 \\ 10^{2.5(B+0.5)} & , -1.3 \leq B - B_c^j \leq -0.5 \\ 1 & , -0.5 \leq B - B_c^j \leq 0.5 \\ 10^{-(B-0.5)} & , 0.5 \leq B - B_c^j \leq 2.5 \\ 0 & , B - B_c^j > 2.5 \end{cases} \tag{6.4}$$

Τα φίλτρα, λοιπόν, προκύπτουν να έχουν ένα τραπεζοειδές σχήμα, όπου η κλίση της καμπύλης του τραπεζοειδούς προς τις χαμηλότερες συχνότητες ($10dB/Bark$) είναι αρκετά πιο ομαλή από την κλίση προς τις υψηλότερες συχνότητες ($25db/Bark$) [36]. Εποπτικά, η εν λόγω συστοιχία, για συχνότητα δειγματοληψίας $16kHz$, παρουσιάζεται στο Σχήμα 6.2.

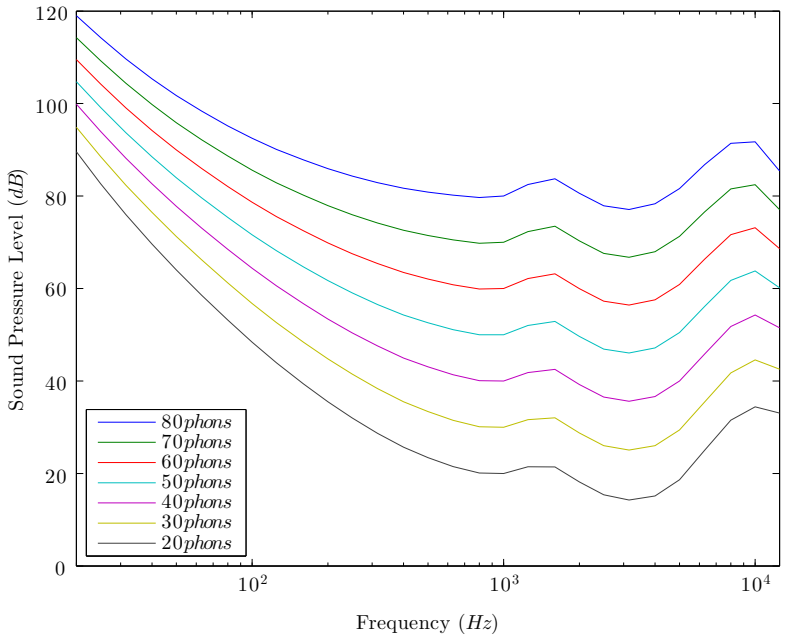


Σχήμα 6.2: Συστοιχία φίλτρων για την εξαγωγή των PLPs. Θεωρείται συχνότητα δειγματοληψίας $16kHz$, και έτσι η συστοιχία προκύπτει να αποτελείται από 21 φίλτρα.

Μέσω μιας τέτοιας συστοιχίας φίλτρων, λέμε ότι το φάσμα ισχύος του σήματος $s_i(n)$ διασπάται σε κρίσιμες ζώνες συχνοτήτων (critical bands). Για ντετερμινιστικά σήματα, το φάσμα ισχύος προκύπτει από το τετράγωνο του Fourier Μετασχηματισμού του σήματος [77], όπως άλλωστε χρησιμοποιείται στη σχέση (6.1). Στην ακουστική, ως κρίσιμη ζώνη συχνοτήτων θεωρείται ένα εύρος συχνοτήτων που εκλαμβάνονται ουσιαστικά ως ίδιες από το ανθρώπινο αυτί, διότι ενεργοποιούν την ίδια περιοχή της βασικής μεμβράνης [78]. Υπό αυτή την έννοια, λέγεται ότι και ο κοχλίας του αυτιού λειτουργεί βάσει μιας συστοιχίας φίλτρων, που ονομάζονται ακουστικά φίλτρα (auditory filters).

Το επόμενο στάδιο της διαδικασίας είναι το φιλτράρισμα του αποτελέσματος της σχέσης (6.1) με ένα φίλτρο προέμφασης που στόχο έχει την προσομοίωση της μη ομοιόμορφης αντίληψης της έντασης του ήχου από τον άνθρωπο σε διαφορετικές συχνότητες, όπως αυτή σκιαγραφείται από τις καμπύλες ίσου επιπέδου έντασης (equal-loudness-level contours). Μία καμπύλη ίσου επιπέδου έντασης [79] αποτελείται από σημεία ακουστικής πίεσης που ακούγονται εξίσου έντονα υπό διαφορετικές συχνότητες από έναν άνθρωπο-ακροατή. Η μονάδα μέτρησης του επιπέδου της έντασης είναι το *phon* και, εξ ορισμού, δύο ημιτονικά κύματα

ίσων *phons* ακούγονται εξίσου έντονα. Στο Σχήμα 6.3 απεικονίζονται καμπύλες ίσου επιπέδου έντασης για διαφορετικές τιμές έντασης, σύμφωνα με το πρότυπο ISO 226:2003 [80].



Σχήμα 6.3: Καμπύλες ίσου επιπέδου έντασης.

Η εισαγωγή της έννοιας των καμπυλών ίσου επιπέδου έντασης κατά την εξαγωγή ακουστικών χαρακτηριστικών είναι μεγάλης σημασίας στην προσπάθεια εύρεσης χαρακτηριστικών που προσομοιώνουν τον ανθρώπινο μηχανισμό ακοής, εφόσον θεωρείται ότι οι καμπύλες αυτές αποκαλύπτουν τα συχνοτικά χαρακτηριστικά του ακουστικού συστήματος του ανθρώπου [79]. Το φίλτρο που χρησιμοποιείται προσομοιώνει την ευαισθησία της ανθρώπινης ακοής περίπου στο επίπεδο των $40dB$ και περιγράφεται, στο πεδίο της συχνότητας, από τη σχέση

$$E(\omega) = \begin{cases} \dfrac{\omega^4\left(\omega^2 + 56.8\cdot 10^6\right)}{\left(\omega^2 + 6.3\cdot 10^6\right)^2\left(\omega^2 + 0.38\cdot 10^9\right)} & , Fs \leq 10kHz \\ \dfrac{\omega^4\left(\omega^2 + 56.8\cdot 10^6\right)}{\left(\omega^2 + 6.3\cdot 10^6\right)^2\left(\omega^2 + 0.38\cdot 10^9\right)\left(\omega^6 + 9.58\cdot 10^{26}\right)} & , Fs > 10kHz \end{cases}, \tag{6.5}$$

όπου $\omega = 2\pi f$. Τα βάρη που εισάγονται από το παραπάνω φίλτρο προέμφασης μπορούν να ενσωματωθούν στα βάρη που εισάγει η συστοιχία φίλτρων. Διαισθητικά, το αποτέλεσμα είναι μία νέα συστοιχία φίλτρων όπου το σχήμα του καθενός είναι όπως φαίνεται στο Σχήμα 6.2, αλλά το ύψος του καθενός αυξάνεται με τη συχνότητα. Ακόμα, οι τιμές που προκύπτουν από το πρώτο και το τελευταίο φίλτρο τίθενται ίσες με τις γειτονικές τους. Έτσι, η σχέση (6.1) μετατρέπεται στη σχέση (6.6).

$$\tilde{G}_i(j) = \begin{cases} \tilde{G}_i(2) & , j = 1 \\ \displaystyle\sum_{k=0}^{N/2}\left\{|S_i[k]|^2 \cdot |E[k]H^j[k]|\right\} & , 2 \leq j \leq Q-1 \\ \tilde{G}_i(Q-1) & , j = Q \end{cases} \tag{6.6}$$

Πέραν της διαφορετικής ευαισθησίας του ανθρώπινου αυτιού αναλόγως της συχνότητας, υπάρχουν ψυχοφυσικοί νόμοι που συνδέουν την πραγματική ένταση μιας ηχητικής πηγής με την ένταση που γίνεται αντιληπτή από τον άνθρωπο. Οι νόμοι αυτοί εντάσσονται σε μία γενικότερη κατηγορία εμπειρικών νόμων που συνδέουν μια διέγερση με την ψυχολογική ένταση που προκαλεί [81]. Όσον αφορά στην ένταση του ήχου, ο αντίστοιχος ψυχοφυσικός νόμος, όπως και σε πολλά άλλα ερεθίσματα, λαμβάνει μία εκθετική μορφή, όπου συγκεκριμένα η ψυχολογική ένταση είναι ανάλογη της πραγματικής έντασης υψωμένης περίπου στο 0.3. Οι μονάδες όπου μετράται η αντιληπτή αυτή ένταση του ήχου είναι τα *sones*. Για να προσεγγιστεί ο νόμος αυτός κατά την εξαγωγή των PLPs, τα $\tilde{G}_i(j)$ μετασχηματίζονται στα $\Phi_i(j)$:

$$\Phi_i(j) = \left(\tilde{G}_i(j)\right)^{0.33} \tag{6.7}$$

Τέλος, βρίσκονται οι συντελεστές γραμμικής πρόβλεψης που προσεγγίζουν το Φ_i. Για το σκοπό αυτό, πρακτικά χρησιμοποιείται η μέθοδος αυτοσυσχέτισης και συγκεκριμένα ο αλγόριθμος Levinson-Durbin [31]. Οπότε, χρειάζεται να εκτιμηθεί η αυτοσυσχέτιση του αντίστοιχου σήματος.

Το Φ_i υπενθυμίζεται ότι προκύπτει ως το φάσμα ισχύος του σήματος $s_i(n)$ αφότου έχει διασπαστεί σε κρίσιμες ζώνες συχνοτήτων και έχει υποστεί περαιτέρω επεξεργασία με σκοπό την προσομοίωση κάποιων χαρακτηριστικών της ανθρώπινης ακοής. Εφόσον, όμως, η συστοιχία φίλτρων τοποθετείται μέχρι τη συχνότητα Nyquist, και λόγω της συμμετρίας του DFT, ολόκληρο το τελικό φάσμα ισχύος $\tilde{\Phi}_i$ (μετά την όποια επεξεργασία) προκύπτει από την προσάρτηση στο διάνυσμα Φ_i του αντικατοπτρισμού του:

$$\tilde{\Phi}_i = [\Phi_i(1), \Phi_i(2), \cdots, \Phi_i(Q-1), \Phi_i(Q), \Phi_i(Q-1), \cdots, \Phi_i(2)]$$

Λόγω της σχέσης (6.6), ισχύει ότι $\Phi_i(2) = \Phi_i(1)$ και ότι $\Phi_i(Q) = \Phi_i(Q-1)$, γι' αυτό και οι συγκεκριμένοι 2 συντελεστές δεν επαναλαμβάνονται στον αντικατοπτρισμό του Φ_i.

Σύμφωνα, τώρα, με το θεώρημα Wiener-Khinchin [82], το φάσμα ισχύος μιας συνάρτησης είναι ο Fourier Μετασχηματισμός της αυτοσυσχέτισης της συνάρτησης. Συνεπώς, για να υπολογίσουμε τη ζητούμενη αυτοσυσχέτιση, αρκεί να πάρουμε τον αντίστροφο μετασχηματισμό (IDFT) της $\tilde{\Phi}_i$.

Έστω ότι η μοντελοποίηση που γίνεται μέσω της γραμμικής πρόβλεψης είναι τάξης p. Φορμαλιστικά, αυτό σημαίνει ότι το σήμα φωνής $s_i(n)$, όπως το αντιλαμβάνεται ο άνθρωπος, παράγεται από την εξίσωση διαφορών (6.8), όπου $u_i(n)$ η θεωρούμενη πηγή και όπου οι συντελεστές γραμμικής πρόβλεψης $a_{i,k}$ και το κέρδος G_i υπολογίζονται μέσω του αλγορίθμου Levinson-Durbin.

$$s_i(n) = \sum_{k=1}^{p} a_{i,k} s_i(n-k) + G_i u_i(n) \tag{6.8}$$

Οι συντελεστές γραμμικής πρόβλεψης μπορούν εύκολα να μετασχηματιστούν σε ένα σύνολο αναφασματικών παραμέτρων $\{\hat{\phi}_i(j)\}, j = 0, 1, \cdots, p$ που αποτελεί το σύνολο των PLPs για το σήμα s_i. Για το σκοπό αυτό χρησιμοποιείται η αναδρομική σχέση (6.9) [31].

$$\hat{\phi}_i(j) = \begin{cases} \log G_i & , j = 0 \\ a_{i,j} + \sum_{k=1}^{j-1} \left(\frac{k}{j}\right) \hat{\phi}_i(k) a_{i,j-k} & , 1 \le j \le p \end{cases} \tag{6.9}$$

6.2 Εύρωστα Χαρακτηριστικά με Χρήση RASTA Ανάλυσης

Όπως έχουμε δει, η μελέτη των φασματικών χαρακτηριστικών ενός ηχητικού σήματος μπορεί να οδηγήσει στην επιτυχή αναγνώριση της γλωσσικής πληροφορίας διότι τα φασματικά αυτά χαρακτηριστικά αντικατοπτρίζουν τις μεταβολές της φωνητικής οδού και άρα τους ήχους που παράγονται από αυτήν. Για να κατευθυνθούμε, οπότε, προς την περιοχή της εύρωστης αναγνώρισης φωνής, θα πρέπει να δούμε τι διαφορές παρουσιάζουν στο χρόνο τα φασματικά χαρακτηριστικά του ήχου γενικότερα (π.χ. θόρυβος) από τα φασματικά χαρακτηριστικά της φωνής, ώστε να δοθεί έμφαση μόνο στα τελευταία. Προς αυτήν ακριβώς την κατεύθυνση κινείται η ιδέα της Σχετικής Φασματικής (RelAtive SpecTrAl - RASTA) ανάλυσης [83].

Η RASTA ανάλυση είναι πρόδρομος των χαρακτηριστικών που βασίζονται στη Συχνότητα Διαμόρφωσης (Modulation Frequency - MF) και στο Φάσμα Διαμόρφωσης (Modulation Spectrum), ιδέες που θα μας απασχολήσουν και στο Κεφάλαιο 8. Η MF είναι η συχνότητα με την οποία μεταβάλλονται τα φασματικά χαρακτηριστικά του σήματος στο χρόνο, έννοια σημαντική για την αναγνώριση, καθώς η αντίληψη της ομιλίας εξαρτάται από τις φασματικές μεταβολές, δηλαδή τη διαφορά των χαρακτηριστικών ενός ήχου από τον προηγούμενό του [83]. Η ανθρώπινη ακοή είναι πιο ευαίσθητη σε μεταβολές της τάξης των περίπου $4Hz$, ενώ συχνότητες διαμόρφωσης άνω των $16Hz$ έχουν σχεδόν αμελητέα επίδραση στη δυνατότητα κατανόησης της φωνής [84]. Τα παραπάνω, σε συνδυασμό με το γεγονός ότι οι αρθρωτές κινούνται με ρυθμούς που δεν ξεφεύγουν από το εύρος των $[1Hz, 13Hz]$ [85], δίνουν σαφείς ενδείξεις για το εύρος συχνοτήτων διαμόρφωσης στις οποίες θα πρέπει να δοθεί έμφαση.

Η ιδέα της RASTA ανάλυσης είναι το φιλτράρισμα του σήματος σε ένα κατάλληλο πεδίο ώστε να περνάνε μόνο τα τμήματα εκείνα που φαίνεται να είναι σημαντικά για την αναγνώριση και να αποκόπτονται τα τμήματα που μεταβάλλονται πιο αργά ή πιο γρήγορα από τις τυπικές μεταβολές της φωνής. Όταν η RASTA ανάλυση συνδυάζεται με τα PLP χαρακτηριστικά, όπως προτείνεται στο [83], τα προκύπτοντα χαρακτηριστικά καλούνται RASTA-PLP. Η RASTA ανάλυση, λοιπόν, λαμβάνει χώρα αφότου έχουν παραχθεί οι φασματικοί συντελεστές του σήματος στις κρίσιμες ζώνες συχνοτήτων στην κλιμακα Bark, σύμφωνα με τη σχέση (6.1) και πριν την εφαρμογή του φίλτρου προέμφασης για την προσομοίωση των καμπυλών ίσου επιπέδου έντασης. Αποτελείται από τρία στάδια: το μη-γραμμικό μετασχηματισμό των φασματικών συντελεστών σε ένα νέο πεδίο (έστω μετασχηματισμός T), το φιλτράρισμα τους ώστε να αποκοπούν οι ζώνες συχνοτήτων διαμόρφωσης χαμηλού ενδιαφέροντος και τον εκ νέου μετασχηματισμό του αποτελέσματος μέσω του αντιστρόφου T^{-1} (ή ενός παρόμοιου μετασχηματισμού). Τονίζεται ότι η RASTA ανάλυση είναι γενική και δεν εξαρτάται αυστηρά από τα συγκεκριμένα χαρακτηριστικά εν χρήση. Για παράδειγμα, το RASTA φιλτράρισμα μπορεί να εφαρμοστεί κατά την εξαγωγή των MFCC χαρακτηριστικών, προτού περάσουμε στο αναφασματικό πεδίο, δηλαδή μετά τη σχέση (5.12).

Θα θεωρήσουμε στην πορεία πως κατά την παραθύρωση του σήματος φωνής λαμβάνονται διαδοχικά παράθυρα ανά $10msec$, οπότε ο ρυθμός παραθύρωσης είναι ${}^{1}/_{10msec} = 100Hz$. Συνεπώς, εξετάζονται συχνότητες διαμόρφωσης στο εύρος $[0Hz, 50Hz]$. Το φίλτρο που προτείνεται στο [83] έχει συνάρτηση μεταφοράς

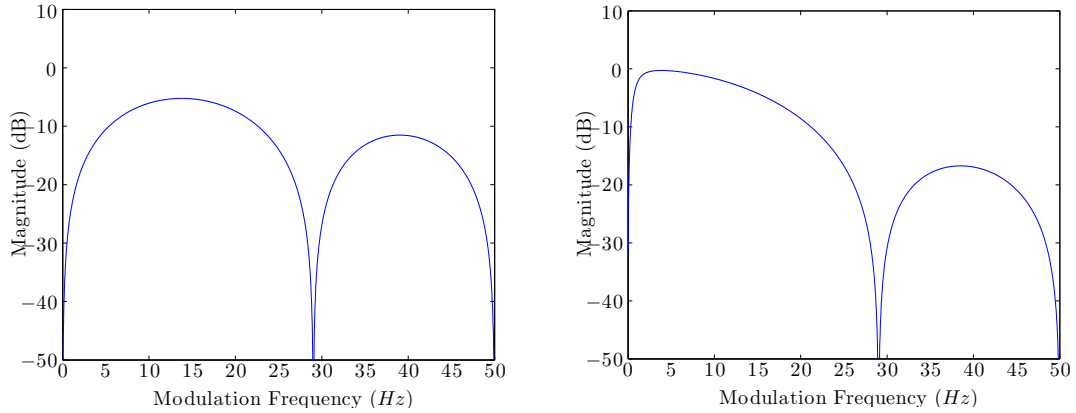
$$H(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.94z^{-1}}. \tag{6.10}$$

Έχει ενδιαφέρον να παρατηρηθεί η έντονη σχέση μεταξύ της RASTA ανάλυσης και του υπολογισμού των Δ συντελεστών, εφόσον ο αριθμητής της (6.10) είναι ουσιαστικά η συνάρτηση

μεταφοράς του φίλτρου που υπολογίζει τους Δ συντελεστές. Για την ακρίβεια, αν $M = 2$, από τη σχέση (5.14) παίρνουμε

$$\begin{aligned} &\Delta x(k) = \frac{1}{10}[-2x(k-2) - x(k-1) + x(k+1) + 2x(k+2)] \\ \Rightarrow &\Delta X(z) = 0.1[-2z^{-2}X(z) - zX(z) + zX(z) + 2z^2X(z)] \\ \Rightarrow &\tilde{H}(z) = \frac{\Delta X(z)}{X(z)} = 0.1z^2[2 + z^{-1} - z^{-3} - 2z^{-4}]\,. \end{aligned} \tag{6.11}$$

Στην πραγματικότητα, κίνητρο της RASTA ανάλυσης υπήρξαν όντως οι Δ συντελεστές και η ιδιότητά τους να εξουδετερώνουν εν μέρει τις αρνητικές επιπτώσεις των συνελικτικών διαταραχών [83]. Ωστόσο, οι Δ συντελεστές δεν μπορούν με επιτυχία να χρησιμοποιηθούν μόνοι τους και γι' αυτό χρησιμοποιούνται συνδυαστικά μαζί με τους αρχικούς, στατικούς συντελεστές, όπως έχουμε δει, οι οποίοι, όμως, είναι επιρρεπείς στις διαταραχές. Στο Σχήμα 6.4 απεικονίζονται οι αποκρίσεις συχνότητας των $H(z)$ και $\tilde{H}(z)$.



(i) (ii)

Σχήμα 6.4: (i)Απόκριση συχνότητας για το φίλτρο $\tilde{H}(z)$ του υπολογισμού των Δ συντελεστών. (ii)Απόκριση συχνότητας για το φίλτρο $H(z)$ της RASTA ανάλυσης.

Παρατηρείται ότι το φίλτρο της RASTA ανάλυσης έχει μία σχετικά σταθερή απόκριση στο εύρος $[1Hz, 10Hz]$, με το μέγιστο κοντά στα $4Hz$, σύμφωνη με τα στοιχεία της ακουστικής θεωρίας που αναλύθηκαν παραπάνω. Αντιθέτως, κατά την εξαγωγή των Δ χαρακτηριστικών, φαίνεται πως ευνοούνται κάποιες λίγες συχνότητες διαμόρφωσης, οι οποίες, μάλιστα, δεν είναι κοντά σε αυτές που χαρακτηρίζουν την ανθρώπινη ομιλία, με αποτέλεσμα να αλλοιώνεται το γλωσσικό περιεχόμενο της πληροφορίας.

Το ερώτημα που πρέπει να απαντηθεί, τώρα, είναι ποιος αποτελεί έναν κατάλληλο μετασχηματισμό T. Εάν ο θόρυβος είναι συνελικτικός, όπως συμβαίνει με την αντήχηση ή τις διαταραχές που προκαλούνται λόγω χρήσης διαφορετικών μικροφώνων, τότε μία συνετή επιλογή είναι το φιλτράρισμα να γίνει στο λογαριθμικό πεδίο. Εκεί, οι εν λόγω διαταραχές εμφανίζονται σαν προσθετικές σταθερές, οπότε είναι εύκολο να εξαλειφθεί η δράση τους. Συνεπώς, κάθε συντελεστής του φάσματος ισχύος λογαριθμίζεται, εφαρμόζεται το RASTA φιλτράρισμα και εν συνεχεία εφαρμόζεται η αντίστροφη συνάρτηση του λογαρίθμου, δηλαδή η εκθετική $\exp(\cdot)$.

Εάν, ωστόσο, υπάρχει και προσθετικός θόρυβος, τότε το λογαριθμικό πεδίο δεν είναι

κατάλληλο. Οπότε, προτείνεται η χρήση του μετασχηματισμού

$$y = \log(1 + Jx)\,, \tag{6.12}$$

όπου J κατάλληλη σταθερά. Το πεδίο αυτό συμπεριφέρεται σαν γραμμικό όταν $J \ll 1$ και σαν λογαριθμικό όταν $J \gg 1$. Η βέλτιστη επιλογή του J είναι αυτή που τοποθετεί το μεγαλύτερο μέρος του καθαρού σήματος στο λογαριθμικό κομμάτι της μη-γραμμικότητας και το μεγαλύτερο μέρος του θορύβου στο γραμμικό κομμάτι. Η αντίστροφη της (6.12) είναι η

$$x = \frac{e^y - 1}{J}\,, \tag{6.13}$$

για την οποία, όμως, δεν υπάρχει εγγύηση ότι είναι πάντα θετική. Οπότε, στην πράξη, ο αντίστροφος μετασχηματισμός προσεγγίζεται ως

$$x = \frac{e^y}{J}\,. \tag{6.14}$$

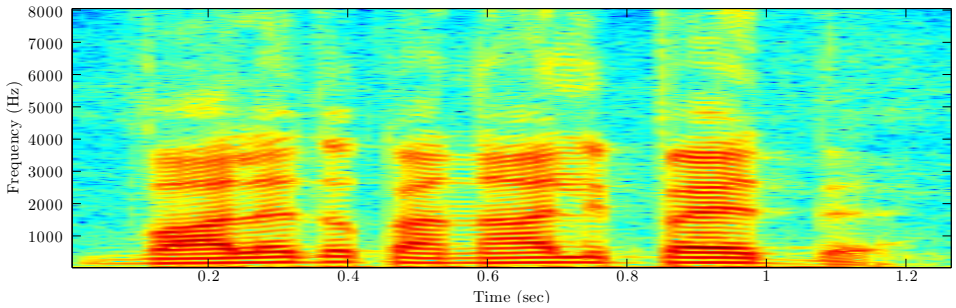
Στην περίπτωση που για το RASTA φιλτράρισμα χρησιμοποιούνται οι σχέσεις (6.12), (6.14) τότε η διαδικασία καλείται J-RASTA ή Lin-Log RASTA.

Η επίδραση της PLP και της RASTA-PLP ανάλυσης γίνεται εμφανής στο Σχήμα 6.5. Εκεί, παρουσιάζεται το αρχικό φασματογράφημα του σήματος, καθώς και τα φασματογραφήματα μετά την PLP και τη RASTA-PLP ανάλυση (προφανώς, πριν τη μετατροπή των φασματικών στα αναφασματικά χαρακτηριστικά). Κατ' αρχήν, φαίνεται άμεσα πως περνώντας από την κλίμακα *Hertz* στην κλίμακα *Bark*, αφιερώνεται "περισσότερος χώρος" στις μικρές συχνότητες. Ακόμα, φαίνεται πως μετά το RASTA φιλτράρισμα η χρονική ανάλυση είναι πολύ χαμηλότερη και με λιγότερες λεπτομέρειες. Η πληροφορία που έχει απομείνει μετά το φιλτράρισμα αναμένεται να είναι η απαραίτητη για την αναγνώριση του γλωσσικού περιεχομένου.

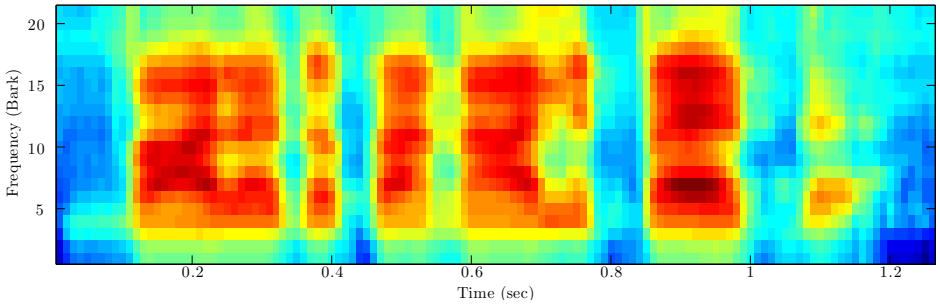
6.3 Πειραματικά Αποτελέσματα

Αρχικά, εξετάζεται η απόδοση του συστήματος αναγνώρισης όταν χρησιμοποιούνται τα χαρακτηριστικά που αναλύθηκαν στο Κεφάλαιο αυτό (PLP και RASTA-PLP) στην απλή τους μορφή. Σε όλα τα πειράματα χρησιμοποιείται μοντέλο 12ου βαθμού για το στάδιο της γραμμικής πρόβλεψης, γεγονός που οδηγεί σε διάνυσμα χαρακτηριστικών μήκους 13, ενώ η παραθύρωση των σημάτων γίνεται με παράθυρα Hamming μήκους $32msec$ με κίνηση ανά $10msec$. Για την παραγωγή των χαρακτηριστικών χρησιμοποιήθηκε το πακέτο ανοιχτού κώδικα που παρέχεται στο [73] για MATLAB. Τα σχετικά αποτελέσματα παρουσιάζονται στο Σχήμα 6.6. Για σύγκριση, παρατίθενται τα αποτελέσματα όταν χρησιμοποιούνται MFCCs, αλλά και όταν χρησιμοποιούνται μόνοι τους οι Δ συντελεστές των PLPs (Δ-PLP).

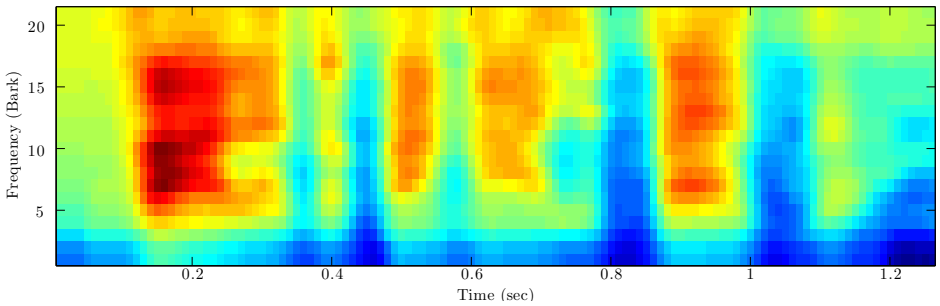
Όπως φαίνεται, τα PLPs για τα συγκεκριμένα δεδομένα λειτουργούν χειρότερα από τα MFCCs, τόσο για τις συνθήκες του μικροφώνου CT1, όσο και για τις συνθήκες του μικροφώνου OA6. Η ευεργετική επίδραση της RASTA ανάλυσης, όμως, είναι προφανής, ιδίως στην περίπτωση του CT1, δηλαδή όταν ουσιαστικά οι μόνες διαφορές μεταξύ εκπαίδευσης και ελέγχου είναι οι διαφορετικοί ομιλητές και τα διαφορετικά μικρόφωνα, δηλαδή συνελικτικές "αλλοιώσεις". Φαίνεται, ακόμα, ότι παρόλο που η χρήση των RASTA-PLP χαρακτηριστικών δίνει μεγάλες βελτιώσεις σε σχέση με τα PLPs και παρόλο που το RASTA φιλτράρισμα έχει μεγάλες ομοιότητες όπως είδαμε με την εξαγωγή των Δ συντελεστών, οι τελευταίοι, όταν



(i) Αρχικό φασματογράφημα του σήματος.



(ii) Φασματογράφημα μετά την PLP ανάλυση.



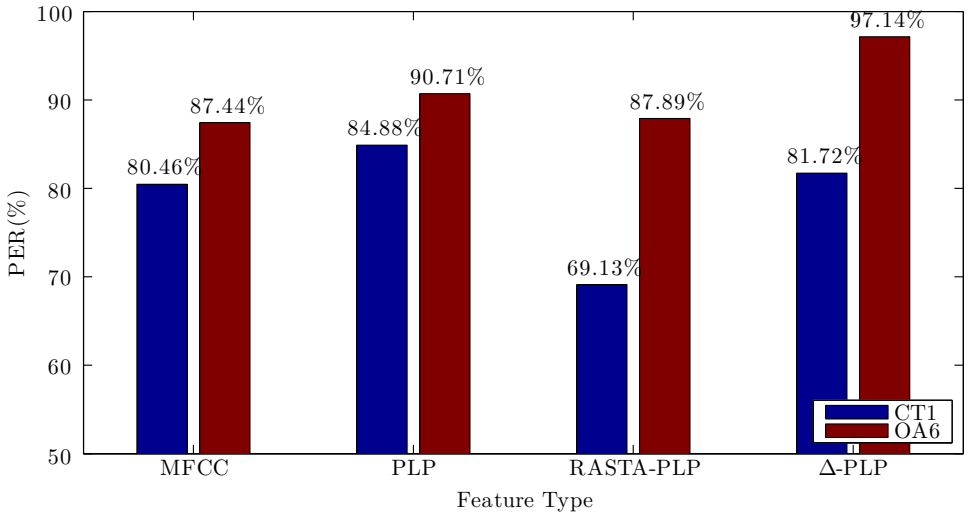
(iii) Φασματογράφημα μετά τη RASTA-PLP ανάλυση.

Σχήμα 6.5: Φασματογράφημα του σήματος φωνής με τη φράση "*Βάλε το κανάλι πέντε*" πριν και μετά την PLP και τη RASTA-PLP ανάλυση.

χρησιμοποιούνται μόνοι τους, δε δίνουν κάποιο καλύτερο αποτέλεσμα, αλλά αντίθετα μπορούν να χειροτερέψουν αρκετά την ποιότητα αναγνώρισης.

Στη συνέχεια, θέλουμε να δούμε την επίδραση στο τελικό αποτέλεσμα που έχει η χρήση του J-RASTA φιλτραρίσματος και να τη συγκρίνουμε με το απλό RASTA φιλτράρισμα στο λογαριθμικό πεδίο. Για το λόγο αυτό, δοκιμάζονται διαφορετικά J σε μία ευρεία κλίμακα τιμών και τα αποτελέσματα παρουσιάζονται στον Πίνακα 6.1. Φαίνεται πως στην καλύτερη περίπτωση τα αποτελέσματα είναι συγκρίσιμα με τα αντίστοιχα της απλής RASTA ανάλυσης, οπότε, λαμβάνοντας υπόψιν την επιπλέον δυσκολία που εισάγεται για την επιλογή της κατάλληλης παραμέτρου J, συνάγεται πως το J-RASTA φιλτράρισμα δεν αποτελεί προτιμητέα επιλογή, τουλάχιστον για τα δεδομένα που έχουμε στη διάθεσή μας.

Στην πράξη, ούτε και τα χαρακτηριστικά που προκύπτουν από τη RASTA ανάλυση επαρκούν για αξιόλογα ποσοστά αναγνώρισης και γι' αυτό το προκύπτον διάνυσμα επαυξάνεται με



Σχήμα 6.6: PER(%) όταν χρησιμοποιούνται raw PLPs και RASTA-PLPs (διανύσματα χαρακτηριστικών μήκους 13). Για σύγκριση, δίνονται τα αντίστοιχα αποτελέσματα με χρήση MFCCs και Δ συντελεστών των PLPs.

	10^{-9}	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
CT1	88.58	84.94	80.63	69.22	68.87	75.72	68.51	78.12	69.10
OA6	91.08	89.51	87.87	88.46	87.82	88.79	88.40	89.09	87.45

	10^{1}	10^{2}	10^{3}	10^{4}
CT1	70.19	**68.45**	69.64	84.67
OA6	**87.39**	87.86	87.56	90.04

Πίνακας 6.1: PER(%) όταν χρησιμοποιούνται J-RASTA-PLPs, για διαφορετικές τιμές της παραμέτρου J.

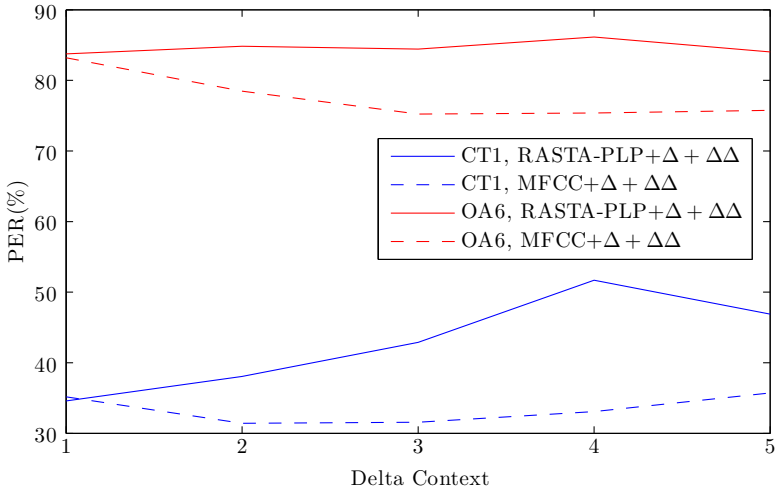
τους συντελεστές ταχύτητας και επιτάχυνσης. Στον Πίνακα 6.2 παρουσιάζονται τα σφάλματα αναγνώρισης όταν το διάνυσμα χαρακτηριστικών επαυξάνεται με τους Δ και ΔΔ συντελεστές, με χρήση των διαφορετικών βασικών χαρακτηριστικών που έχουν παρουσιαστεί. Το χρονικό παράθυρο που χρησιμοποιείται για την εξαγωγή των δυναμικών χαρακτηριστικών ισούται με $2 \cdot 3 + 1 = 7$ πλαίσια.

	MFCC+Δ+ΔΔ	PLP+Δ+ΔΔ	RASTA-PLP+Δ+ΔΔ
CT1	**31.56**	43.86	42.90
OA6	**75.22**	86.44	86.14

Πίνακας 6.2: PER(%) όταν χρησιμοποιούνται PLPs και RASTA-PLPs (και MFCCs), μετά την επαύξηση του διανύσματος χαρακτηριστικών με τους συντελεστές ταχύτητας και επιτάχυνσης. Οι δυναμικοί συντελεστές λαμβάνονται βάσει των 6 (3+3) γειτονικών πλαισίων.

Παρόλο που τα δυναμικά χαρακτηριστικά σε κάθε περίπτωση βελτιώνουν το αποτέλεσμα, τα MFCCs φανερά υπερτερούν. Ωστόσο, η χρήση 6 γειτονικών παραθύρων (context=3) για τον υπολογισμό των δυναμικών χαρακτηριστικών είναι ίσως μεροληπτική υπέρ των MFCCs, καθώς για τα τελευταία είναι σχεδόν βέλτιστη επιλογή, όπως προκύπτει από το Σχήμα 5.4. Για το λόγο αυτό, δοκιμάζουμε διαφορετικά contexts για τον εν λόγω υπολογισμό, όταν

αυτός αφορά τα RASTA-PLPs, με τα σχετικά αποτελέσματα να παρουσιάζονται στο Σχήμα 6.7. Για λόγους σύγκρισης, παρουσιάζονται και τα αντίστοιχα αποτελέσματα από το Σχήμα 5.4.



Σχήμα 6.7: PER(%) καθώς μεταβάλλεται το χρονικό παράθυρο που λαμβάνεται υπόψιν κατά τον υπολογισμό των Δ και ΔΔ συντελεστών όταν χρησιμοποιούνται RASTA-PLPs (και MFCCs).

Προκύπτει ότι η βέλτιστη επιλογή για το συνδυασμό των RASTA-PLPs με Δ και ΔΔ συντελεστές είναι η χρήση του ελάχιστου δυνατού context για τον υπολογισμό των τελευταίων, δηλαδή η αξιοποίηση μόνο των άμεσα γειτονικών πλαισίων, δίνοντας PER 34.58% για το μικρόφωνο CT1 και 83.76% για το OA6. Σε κάθε περίπτωση, όμως, η σύγκριση μεταξύ των βέλτιστων αποτελεσμάτων με MFCCs+Δ+ΔΔ και με RASTA-PLPs+Δ+ΔΔ ευνοεί τα πρώτα.

Κεφάλαιο 7

Σύνδεση Διαδοχικών Πλαισίων και Τεχνικές Μείωσης της Διαστασιμότητας

7.1 Η Ιδέα της Σύνδεσης Διαδοχικών Πλαισίων Χαρακτηριστικών

Ήδη από την Ενότητα 5.2 είδαμε πως τα λίγα στατικά χαρακτηριστικά που προκύπτουν από την επεξεργασία ενός μόνο πλαισίου του σήματος δεν αρκούν για να χαρακτηρίσουν πλήρως το εκάστοτε πλαίσιο. Αντιθέτως, προτάθηκε πως το διάνυσμα χαρακτηριστικών θα πρέπει να επαυξηθεί με δυναμικά χαρακτηριστικά, κάτι που επαληθεύτηκε και πειραματικά ότι επιφέρει καλύτερα αποτελέσματα στην τελική αναγνώριση (π.χ. Σχήμα 5.3).

Ωστόσο, η επιλογή των δυναμικών χαρακτηριστικών που επιλέχθηκαν και που υπολογίστηκαν μέσω πολυωνυμικής παρεμβολής είναι ευριστική και δεν οδηγείται ούτε από κάποια ιδιαίτερα χαρακτηριστικά της ομιλίας, ούτε από τον τελικό σκοπό της εφαρμογής, που είναι η κατά το δυνατό καλύτερη διαχωρισιμότητα και κατηγοριοποίηση των φωνημάτων προς αναγνώριση. Μία πολύ γενικότερη ιδέα είναι κάθε διάνυσμα χαρακτηριστικών να επαυξηθεί από τα γειτονικά του, έστω M από αριστερά (παρελθόν) και M από δεξιά (μέλλον), σαν να συνδέουμε τα γειτονικά πλαίσια μεταξύ τους. Έτσι, εάν $\mathbf{x}_i$ είναι το στατικό διάνυσμα χαρακτηριστικών που έχει προκύψει για το πλαίσιο i, το επαυξημένο διάνυσμα θα είναι

$$\mathbf{x}_{i,συνδ} = \left[\mathbf{x}_{i-M}^T, \mathbf{x}_{i-M+1}^T, \cdots, \mathbf{x}_i^T, \cdots, \mathbf{x}_{i+M-1}^T, \mathbf{x}_{i+M}^T\right]^T . \tag{7.1}$$

Βεβαίως, το διάνυσμα που προκύπτει από τη σύνδεση της μορφής της σχέσης (7.1) είναι εξαιρετικά μεγάλης διάστασης, γεγονός που αυξάνει πολύ τις απαιτήσεις μνήμης, αλλά κυρίως αδυνατεί να αντικατοπτρίσει τη ζητούμενη δυναμική του σήματος. Απαιτείται, λοιπόν, η προβολή του διανύσματος αυτού σε κάποιο νέο χώρο μικρότερων διαστάσεων, ο οποίος, όμως, θα είναι βέλτιστος σύμφωνα με κάποιο κριτήριο. Υπό αυτό το πρίσμα, αναζητείται η κατάλληλη μήτρα μετασχηματισμού $\mathbf{T}$, ωστε το νέο διάνυσμα χαρακτηριστικών $\tilde{\mathbf{x}}_i$ για το πλαίσιο i να δίνεται ως [34]

$$\tilde{\mathbf{x}}_i = \mathbf{T}\left[\mathbf{x}_{i-M}^T, \mathbf{x}_{i-M+1}^T, \cdots, \mathbf{x}_i^T, \cdots, \mathbf{x}_{i+M-1}^T, \mathbf{x}_{i+M}^T\right]^T . \tag{7.2}$$

Παρατηρούμε πως με τη νέα αυτή θεώρηση, η σχέση (5.14) είναι μια ειδική περίπτωση της σχέσης (7.2), με κατάλληλη επιλογή του $\mathbf{T}$. Ωστόσο, τώρα, η σχέση που περιγράφει

τη δυναμική του σήματος δεν ορίζεται εκ των προτέρων αλλά σύμφωνα με κάποιο βέλτιστο κριτήριο που θα οδηγήσει στον κατάλληλο πίνακα $\mathbf{T}$. Για το σκοπό αυτό, χρησιμοποιείται κάποια τεχνική μείωσης της διαστασιμότητας, όπως είναι η PCA, η LDA ή η HLDA, οι οποίες θα αναλυθούν στις επόμενες Ενότητες.

7.2 Principal Component Analysis (PCA)

Ίσως η πιο γνωστή μέθοδος μείωσης της διαστασιμότητας είναι η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA) [86]. Η PCA λειτουργεί στο πλαίσιο της μη επιβλεπόμενης μάθησης. Στόχος της είναι η αναπαράσταση των αρχικών χαρακτηριστικών σε ένα νέο χώρο, χαμηλότερης διάστασης, χωρίς τη χρήση επισημειωμένων παραδειγμάτων, δηλαδή χωρίς την εκ των προτέρων γνώση της κατηγορίας όπου ανήκει το κάθε δεδομένο (π.χ. σε ποιο φώνημα αντιστοιχεί).

Συμβολίζοντας για ευκολία ως $\mathbf{x} \in \mathbb{R}^m$ ένα διάνυσμα χαρακτηριστικών, όπως αυτό της σχέσης (7.1), ζητούμενο είναι να βρεθεί κατάλληλος πίνακας μετασχηματισμού $\mathbf{T}$, ώστε τα $l < m$ πρώτα χαρακτηριστικά του μετασχηματισμένου διανύσματος $\mathbf{Tx}$ να διατηρούν το μεγαλύτερο μέρος της πληροφορίας του $\mathbf{x}$. Η PCA (ή KLT) μεγιστοποιεί το ρυθμό μείωσης της διακύμανσης και άρα αποτελεί τη βέλτιστη επιλογή υπό την έννοια του ελάχιστου μέσου τετραγωνικού σφάλματος.

Η εφαρμογή της PCA στο διάνυσμα $\mathbf{x}$ υποθέτει ότι το $\mathbf{x}$ είναι μηδενικού μέσου. Για το λόγο αυτό, πρώτο βήμα πριν προχωρήσουμε στην ανάλυση είναι η αφαίρεση του μέσου από το διάνυσμα. Για απλότητα, εδώ, θεωρούμε $\mathbb{E}[\mathbf{x}] = 0$. Έστω $\mathbf{q} \in \mathbb{R}^m$ ένα μοναδιαίο διάνυσμα επί του οποίου θα προβληθεί το $\mathbf{x}$, όπου η προβολή A δίνεται ως

$$A = \mathbf{x}^T\mathbf{q} = \mathbf{q}^T\mathbf{x}\,. \tag{7.3}$$

Η προβολή αυτή έχει προφανώς μηδενική μέση τιμή, ενώ για τη διακύμανσή της ισχύει

$$\begin{aligned}\mathbb{E}[A^2] &= \mathbb{E}[(\mathbf{q}^T\mathbf{x})(\mathbf{x}^T\mathbf{q})] \\ &= \mathbf{q}^T\mathbb{E}[\mathbf{x}\mathbf{x}^T]\mathbf{q} \\ &\triangleq \mathbf{q}^T\mathbf{R}\mathbf{q} \triangleq \psi(\mathbf{q})\,,\end{aligned} \tag{7.4}$$

όπου $\mathbf{R}$ ο πίνακας συσχέτισης του $\mathbf{x}$.

Συνεπώς, εφόσον εν τέλει ζητούμενο είναι η προβολή σε ένα χώρο όπου όσο το δυνατό λιγότερες συνιστώσες συγκεντρώνουν όσο το δυνατό μεγαλύτερο ποσοστό της συνολικής διακύμανσης, ενδιαφερόμαστε να βρούμε διανύσματα $\mathbf{q}$ για τα οποία η $\psi(\mathbf{q})$ έχει ακρότατα, υπό τον περιορισμό πάντα ότι $||\mathbf{q}|| = 1$. Αποδεικνύεται [86] ότι τα διανύσματα αυτά πρόκεινται για τα ιδιοδιανύσματα της $\mathbf{R}$.

Σύμφωνα με τα παραπάνω, οι συνιστώσες του μετασχηματισμένου διανύσματος χαρακτηριστικών $\tilde{\mathbf{x}}$ θα είναι οι προβολές του $\mathbf{x}$ πάνω στα ιδιοδιανύσματα $\mathbf{q}_i$, $i = 1, 2, \cdots, m$, οι οποίες ονομάζονται κύριες συνιστώσες του $\mathbf{x}$. Δηλαδή, σύμφωνα με τη σχέση (7.3),

$$\tilde{\mathbf{x}} = [\mathbf{x}^T\mathbf{q}_1, \mathbf{x}^T\mathbf{q}_1, \mathbf{x}^T\mathbf{q}_m]^T = \mathbf{Q}^T\mathbf{x}\,. \tag{7.5}$$

Οπότε, ο ζητούμενος πίνακας μετασχηματισμού $\mathbf{T}$ ισούται με τον πίνακα $\mathbf{Q}^T$, όπου οι στήλες του $\mathbf{Q}$ είναι τα $\mathbf{q}_i$, τοποθετημένα κατά φθίνουσα σειρά των ιδιοδιανυσμάτων στα οποία αντιστοιχούν για λόγους που θα γίνουν εμφανείς στη συνέχεια.

Η διακύμανση της κύριας συνιστώσας που προκύπτει από την προβολή στο ιδιοδιάνυσμα $\mathbf{q}_i$ ισούται, σύμφωνα με τη σχέση (7.4), με

$$\sigma_i^2 = \mathbf{q}_i^T \mathbf{R} \mathbf{q}_i = \lambda_i \,, \tag{7.6}$$

όπου λ_i η αντίστοιχη ιδιοτιμή. Συνεπώς, η συνολική διακύμανση δίνεται ως

$$\sum_{i=1}^{m} \sigma_i^2 = \sum_{i=1}^{m} \lambda_i \,. \tag{7.7}$$

Εάν, λοιπόν, οι κύριες συνιστώσες υπολογισθούν έτσι ώστε η πρώτη κύρια συνιστώσα να αντιστοιχεί στη μεγαλύτερη ιδιοτιμή, η δεύτερη στην αμέσως μικρότερη κ.ο.κ., όπως προτάθηκε παραπάνω, τότε επιλέγοντας τις p πρώτες κύριες συνιστώσες, επιλέγονται εκείνες οι συνιστώσες που συγκεντρώνουν τη μεγαλύτερη συνολικά διακύμανση. Συνοπτικά, η διαδικασία της PCA έγκειται στον υπολογισμό των ιδιοτιμών και των ιδιοδιανυσμάτων του πίνακα συσχέτισης των δεδομένων και στην προβολή τους στον υποχώρο όπου εκτείνονται τα ιδιοδιανύσματα που αντιστοιχούν στις p μεγαλύτερες ιδιοτιμές. Με την αναπαράσταση αυτή, γνωστή και ως ανάλυση υποχώρων (subspace decomposition), το τελικό σφάλμα αναπαράστασης είναι

$$J = \sum_{i=p+1}^{m} \sigma_i^2 = \sum_{i=p+1}^{m} \lambda_i \,. \tag{7.8}$$

Στην παραπάνω ανάλυση, και σύμφωνα με το συμβολισμό που χρησιμοποιήθηκε, θεωρήθηκε ότι το $\mathbf{x}$ είναι τυχαίο διάνυσμα, δηλαδή αποτελείται από τυχαίες μεταβλητές. Σαφώς, στην πράξη έχουμε ντετερμινιστικά σήματα και παρατηρήσεις. Έτσι, ο πίνακας συσχέτισης προκύπτει από τα δείγματα που έχουμε στη διάθεσή μας. Για παράδειγμα, εάν τα σήματα $s_1, s_2, \cdots, s_k$ χωριστούν σε $L_1, L_2, \cdots, L_k$ πλαίσια, αντίστοιχα, σε καθένα από τα οποία υπολογιστεί ένα διάνυσμα m χαρακτηριστικών, τότε θεωρούμε ότι έχουμε m μεταβλητές $x_1, x_2, \cdots, x_m$, που συναποτελούν το "τυχαίο" διάνυσμα $\mathbf{x}$, για καθεμιά από τις οποίες υπάρχουν $L_1 + L_2 + \cdots + L_k = n$ παρατηρήσεις. Τα στοιχεία r_{yz} του πίνακα $\mathbf{R}$, όπου $y = x_i$ και $z = x_j$ για κάποια $i, j \in \{1, 2, \cdots m\}$ δίνονται ως

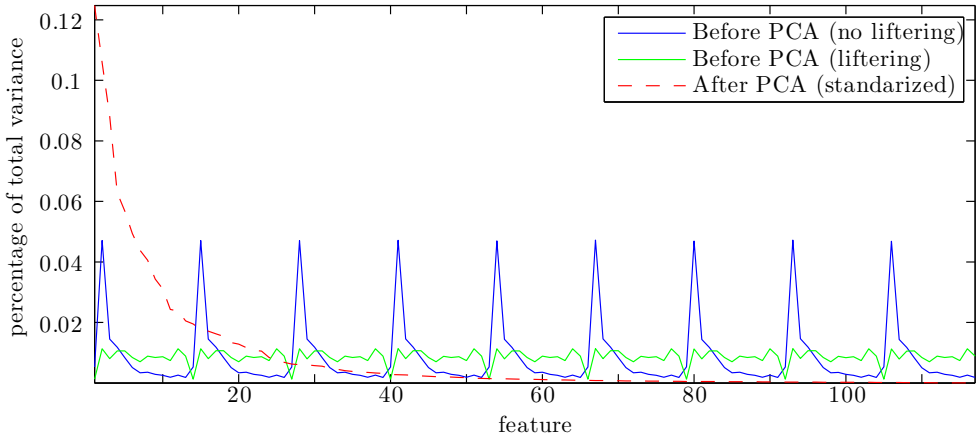
$$\begin{aligned} r_{yz} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(z_i - \bar{z})^2}} \\ &= \frac{n\sum_{i=1}^{n} y_i z_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} z_i}{\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}\sqrt{n\sum_{i=1}^{n} z_i^2 - \left(\sum_{i=1}^{n} z_i\right)^2}} \,, \end{aligned} \tag{7.9}$$

όπου με $\bar{y}$, $\bar{z}$ συμβολίζονται οι μέσες τιμές των αντίστοιχων μεταβλητών.

Η επίδραση της PCA φαίνεται στο Σχήμα 7.1, όπου απεικονίζεται η συνεισφορά στη συνολική διακύμανση καθενός από τα $13 \cdot 9 = 117$ χαρακτηριστικά που χαρακτηρίζουν κάθε πλαίσιο και που προκύπτουν από την σύνδεση 9 διαδοχικών πλαισίων. Αρχικά, έχουν υπολογιστεί για κάθε πλαίσιο τα 13 MFCCs, όπως έχει περιγραφεί στο Κεφάλαιο 5, για τις

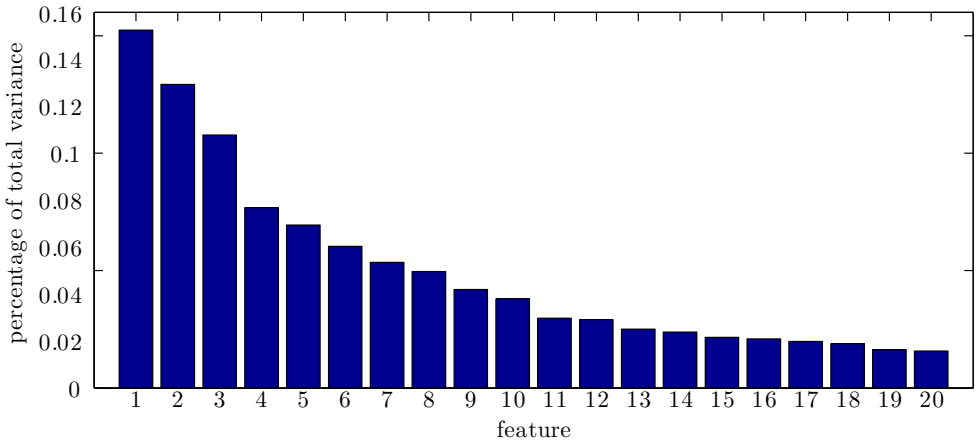
190 εκφορές που αποτελούν το σύνολο ελέγχου του βασικού μας συστήματος αναγνώρισης. Όπως φαίνεται, μετά την PCA τα πρώτα χαρακτηριστικά είναι αυτά που συγκεντρώνουν το μεγαλύτερο ποσοστό της συνολικής διακύμανσης, με τα περίπου 30 πρώτα να συνεισφέρουν συνολικά κατά το 90% και τα 60 πρώτα κατά το 97%.

Συνήθως, για την εφαρμογή της PCA τα χαρακτηριστικά κανονικοποιούνται όχι μόνο ως προς τη μέση τιμή, αλλά και ως προς την τυπική απόκλιση. Κάτι τέτοιο είναι συχνά απαραίτητο ώστε όλα τα χαρακτηριστικά να αντιμετωπίζονται από την PCA ισότιμα και να μη δίνεται ιδιαίτερη έμφαση στα χαρακτηριστικά εκείνα με μεγάλη αρχική διακύμανση. Χωρίς την εν λόγω κανονικοποίηση, η προβολή στο νέο χώρο θα γινόταν με τρόπο ώστε πολύ λίγες συνιστώσες να συγκεντρώνουν το μεγαλύτερο ποσοστό της συνολικής διακύμανσης, όπως φαίνεται στο Σχήμα 7.2. Επίσης, το ποια χαρακτηριστικά έχουν μεγάλη διακύμανση στον αρχικό χώρο δεν αποτελεί αξιόπιστο μέτρο της σημαντικότητας των εν λόγω χαρακτηριστικών. Στο παράδειγμα που εξετάζουμε, η αρχική διακύμανση των χαρακτηριστικών εξαρτάται από τον ακριβή τρόπο που εξήχθησαν τα χαρακτηριστικά, όπως το εάν έχει γίνει ή όχι χρήση liftering, κάτι που, όπως έχει εξηγηθεί στην Ενότητα 5.1 δε θα πρέπει να επηρεάζει την αναγνώριση.

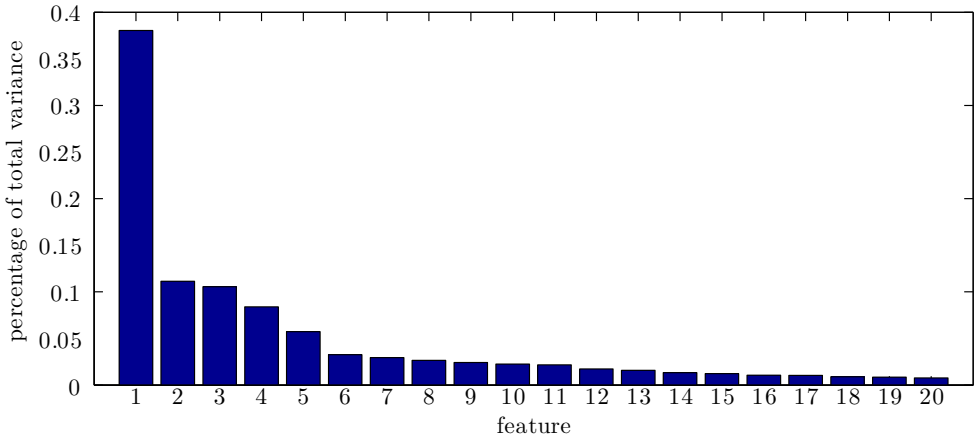


Σχήμα 7.1: Συνεισφορά των διαφόρων χαρακτηριστικών στη συνολική διακύμανση μετά τη σύνδεση 9 διαδοχικών πλαισίων πριν και μετά την PCA. Έχει γίνει χρήση 13 MFCCs με και χωρίς liftering που έχουν εξαχθεί από το σύνολο των 190 εκφορών του test set. Για την εφαρμογή της PCA, τα χαρακτηριστικά κανονικοποιούνται ως προς μέση τιμή και τυπική απόκλιση.

Παρά τη διαδεδομένη χρήση της σε μια ευρεία περιοχή εφαρμογών, η χρήση PCA στο πλαίσιο της εφαρμογής που συζητάμε δεν είναι συνετή. Μετά την κανονικοποίηση, η PCA αντιμετωπίζει όπως είπαμε όλα τα χαρακτηριστικά ισότιμα. Ως εκ τούτου, μετά τη σύνδεση έστω 9 διαδοχικών πλαισίων, για ένα πλαίσιο p η σημαντικότητα των 13 MFCCs που έχουν εξαχθεί από το p είναι ίδια με αυτή των 13 MFCCs που έχουν εξαχθεί από τα πλαίσια $p\pm1, p\pm 2, p\pm 3, p\pm 4$, κάτι που διαισθητικά δεν έχει ιδιαίτερο νόημα. Μία πρόταση θα ήταν η χρήση βαρών ώστε η σημαντικότητα των χαρακτηριστικών να μειώνεται όσο απομακρυνόμαστε από το κεντρικό πλαίσιο. Ωστόσο, η επιλογή κατάλληλων βαρών δε θα ήταν εύκολη υπόθεση και η όλη διαδικασία θα ξέφευγε από το πλαίσιο της PCA.



(i) Με κανονικοποίηση.



(ii) Χωρίς κανονικοποίηση.

Σχήμα 7.2: Συνεισφορά των 20 πρώτων χαρακτηριστικών στη συνολική διακύμανση μετά την PCA όταν έχει γίνει κανονικοποίηση των αρχικών χαρακτηριστικών ως προς την τυπική απόκλιση και όταν όχι. Τα χαρακτηριστικά εξάγονται όπως και στο Σχήμα 7.1, χωρίς liftering.

7.3 Linear Discriminant Analysis (LDA)

Η PCA, όπως είδαμε στην προηγούμενη Ενότητα, παρόλο που είναι βέλτιστη μέθοδος για εφαρμογές συμπίεσης και κωδικοποίησης, δε δίνει κάποια εγγύηση βελτιστότητας για εφαρμογές κατηγοριοποίησης, όπως είναι η αναγνώριση φωνημάτων. Για τέτοιου είδους εφαρμογές, καταλληλότερη είναι η Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis - LDA) [87], η οποία πράγματι έχει αποδειχθεί πως δίνει αξιόλογα αποτελέσματα σε εφαρμογές αναγνώρισης φωνής [88].

Θα μελετήσουμε την LDA πρώτα για την περίπτωση δύο κλάσεων και στη συνέχεια θα δούμε την άμεση γενίκευσή της στην περίπτωση πολλών κλάσεων διαχωρισμού. Στην περίπτωση των δύο κλάσεων, λοιπόν, έστω ω_1, ω_2, το ζητούμενο είναι η προβολή των m-διάστατων χαρακτηριστικών $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ σε μία διάσταση, με τρόπο ώστε να μεγιστοποιείται η διαχωρισιμότητα μεταξύ των ω_1, ω_2. Οπότε, αναζητείται το βέλτιστο m-διάστατο διάνυσμα $\mathbf{w}$ ώστε οι προβολές $y_i = \mathbf{w}^T \mathbf{x}_i$ να διαχωρίζονται κατά το βέλτιστο τρόπο.

Η LDA είναι μέθοδος επιβλεπόμενης μάθησης, οπότε θεωρούμε ότι υπάρχει ένα σύνολο εκπαίδευσης για το οποίο είναι γνωστή η πληροφορία της κλάσης όπου ανήκει το κάθε δεδομένο. Έστω, οπότε, ένα υποσύνολο $\mathcal{D}_1$ n_1 δεδομένων που ανήκουν στην ω_1 και ένα υποσύνολο $\mathcal{D}_2$ n_2 δεδομένων που ανήκουν στην ω_2. Κατ' αντιστοιχία, δημιουργείται ένα υποσύνολο $\mathcal{Y}_1$ n_1 προβολών που ανήκουν στην ω_1 και ένα υποσύνολο $\mathcal{Y}_2$ n_2 προβολών που ανήκουν στην ω_2. Ορίζονται αρχικά οι αντίστοιχοι δειγματικοί μέσοι ως

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x} \quad , i = 1, 2 \,, \tag{7.10}$$

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in \mathcal{Y}_i} y = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{w}^T \mathbf{x} = \mathbf{w}^T \mathbf{m}_i \quad , i = 1, 2 \,. \tag{7.11}$$

Ορίζονται, εν συνεχεία, οι πίνακες διασποράς

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \quad , i = 1, 2 \tag{7.12}$$

και ο πίνακας ενδοταξικής διασποράς

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 \,. \tag{7.13}$$

Βάσει των πινάκων διασποράς, μπορούν να υπολογιστούν οι διασπορές των προβολών ως

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2 = \sum_{x \in \mathcal{D}_i} (\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mathbf{m}_i)^2 = \mathbf{w}^T \mathbf{S}_i \mathbf{w} \quad , i = 1, 2 \,. \tag{7.14}$$

Ακόμη, ορίζεται ο πίνακας διαταξικής διασποράς που δείχνει πόσο απομακρυσμένα είναι μεταξύ τους τα κέντρα των δύο κλάσεων

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \tag{7.15}$$

και η διαταξική διασπορά των προβολών

$$(\tilde{m}_1 - \tilde{m}_2)^2 = (\mathbf{w}^T \mathbf{m}_1 - \mathbf{w}^T \mathbf{m}_2)^2 = \mathbf{w}^T \mathbf{S}_B \mathbf{w} \,. \tag{7.16}$$

Η LDA ψάχνει το $\hat{\mathbf{w}}$ εκείνο που οδηγεί στη μεγιστοποίηση των διαταξικών αποστάσεων και την ελαχιστοποίηση των ενδοταξικών αποστάσεων των προβολών. Φορμαλιστικά

$$\begin{aligned} \hat{\mathbf{w}} &= \underset{\mathbf{w}}{\operatorname{argmax}} \, J(\mathbf{w}) \\ &\triangleq \underset{\mathbf{w}}{\operatorname{argmax}} \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \underset{(7.13),(7.14)}{\overset{(7.16)}{=}} \underset{\mathbf{w}}{\operatorname{argmax}} \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \,. \end{aligned} \tag{7.17}$$

Η έκφραση προς μεγιστοποίηση είναι ένα γενικευμένο πηλίκο Rayleigh και η επίλυση του ζητούμενου προβλήματος ανάγεται στην επίλυση του προβλήματος γενικευμένων ιδιοτιμών

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w} \,. \tag{7.18}$$

Το ζητούμενο διάνυσμα $\hat{\mathbf{w}}$, λοιπόν, είναι το γενικευμένο ιδιοδιάνυσμα $\mathbf{w}$ που αντιστοιχεί στη μεγαλύτερη γενικευμένη ιδιοτιμή λ, σύμφωνα με τη σχέση (7.18).

Στην περίπτωση περισσότερων, έστω p κλάσεων, αναζητούμε τη βέλτιστη προβολή των m-διάστατων δεδομένων σε ένα χώρο $p-1$ διαστάσεων. Οπότε, πλέον ζητούμενο δεν είναι ένα διάνυσμα $\mathbf{w}$, αλλά ένας πίνακας μετασχηματισμού, έστω $\mathbf{W}$, ο οποίος δεν είναι παρά ο πίνακας $\mathbf{T}^T$, όπως συμβολίστηκε στην Ενότητα 7.1. Γενικεύοντας τη σχέση (7.18), τώρα έχουμε

$$\begin{aligned}\hat{\mathbf{W}} &= \underset{\mathbf{W}}{\operatorname{argmax}}\, J(\mathbf{W}) \\ &\triangleq \underset{\mathbf{W}}{\operatorname{argmax}}\, \frac{|\mathbf{W}^T\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|}\,.\end{aligned} \tag{7.19}$$

Ο πίνακας ενδοταξικής διασποράς $\mathbf{S}_W$ δίνεται από τη σχέση[1]

$$\mathbf{S}_W = \frac{1}{n}\sum_{i=1}^{p} n_i\mathbf{S}_i\,, \tag{7.20}$$

όπου

$$\mathbf{S}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T \quad ,i = 1, 2, \cdots, p\,. \tag{7.21}$$

Ο πίνακας διαταξικής διασποράς $\mathbf{S}_B$ δίνεται από τη σχέση

$$\mathbf{S}_B = \frac{1}{n}\sum_{i=1}^{p} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T\,, \tag{7.22}$$

, όπου οι δειγματικοί μέσοι δίνονται από τη σχέση (7.10), με το δείκτη i να παίρνει τις τιμές $1, 2, \cdots, p$ και ο συνολικός δειγματικός μέσος $\mathbf{m}$ ορίζεται ως

$$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}}\mathbf{x} = \frac{1}{n}\sum_{i=1}^{p} n_i\mathbf{m}_i\,. \tag{7.23}$$

Στον παραπάνω συμβολισμό, n είναι ο συνολικός αριθμός δεδομένων εκπαίδευσης και n_i είναι ο αριθμός δεδομένων που ανήκουν στην i-οστή κλάση, έστω ω_i. Προφανώς, $\sum_{i=1}^{p} n_i = n$. Ο πίνακας διαταξικής διασποράς επιλέγεται ώστε η άθροισή του με τον πίνακα ενδοταξικής διασποράς να δίνει τον πίνακα ολικής διασποράς $\mathbf{S}_T$:

$$\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W = \frac{1}{n}\sum_{\mathbf{x}}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T \tag{7.24}$$

Σημειώνεται ότι βάσει του πίνακα ολικής διασποράς, η σχέση (7.19) μπορεί να γραφεί ισοδύναμα ως

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, \frac{|\mathbf{W}^T\mathbf{S}_T\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|}\,. \tag{7.25}$$

[1] Συνήθως στη βιβλιογραφία, όλοι οι πίνακες διασποράς που χρειάζονται για την LDA υπολογίζονται χωρίς την κανονικοποίηση με το πλήθος των δεδομένων, όπως παρουσιάστηκε στην περίπτωση των δύο κλάσεων, καθώς στο τελικό κριτήριο οι όροι κανονικοποίησης απαλοίφονται και δεν παίζουν ρόλο. Ωστόσο, εδώ οι πίνακες διασποράς παρουσιάζονται κανονικοποιημένοι για λόγους συνέπειας της σημειογραφίας με την HLDA που παρουσιάζεται στην επόμενη Ενότητα.

Αποδεικνύεται ότι και πάλι το πρόβλημα ανάγεται σε ένα πρόβλημα γενικευμένων ιδιοτιμών, καθώς οι στήλες του $\mathbf{W}$ είναι τα γενικευμένα ιδιοδιανύσματα $\mathbf{w}_i$ που αντιστοιχούν στις γενικευμένες ιδιοτιμές λ_i στη σχέση

$$\mathbf{S}_B\mathbf{w}_i = \lambda_i\mathbf{S}_W\mathbf{w}_i\,. \tag{7.26}$$

Οι ιδιοτιμές μπορούν να υπολογιστούν ως οι λύσεις της χαρακτηριστικής εξίσωσης

$$|\mathbf{S}_B - \lambda_i\mathbf{S}_W| = 0 \tag{7.27}$$

και το ιδιοδιάνυσμα που αντιστοιχεί στην ιδιοτιμή λ_i υπολογίζεται μέσω της σχέσης

$$(\mathbf{S}_B - \lambda_i\mathbf{S}_W)\mathbf{w}_i = 0\,. \tag{7.28}$$

Ο πίνακας $\mathbf{S}_B$ είναι βαθμού το πολύ $p-1$, οπότε θα προκύψουν το πολύ $p-1$ μη-μηδενικές ιδιοτιμές. Επομένως, τα αντίσοιχα ιδιοδιανύσματα μπορούν να δώσουν τις ζητούμενες προβολές στον $(p-1)$-διάστατο χώρο.

Σημειώνεται πως παρόλο που η ανάλυση έγινε με γνώμονα την εύρεση ενός βέλτιστου, από την άποψη της διαχωρισιμότητας των δεδομένων, χώρου διάστασης $p-1$, δεν υπάρχει κάποιος περιορισμός ως προς τη διάσταση του προκύπτοντος διανύσματος χαρακτηριστικών μετά την εφαρμογή της LDA. Εάν, λοιπόν, θέλουμε να μειώσουμε τη διαστασιμότητα από m που είναι αρχικά σε έστω $l < m$, τότε αρκεί να βρεθούν τα ιδιοδιανύσματα που αντιστοιχούν στις l μεγαλύτερες γενικευμένες ιδιοτιμές, όπως υπολογίζονται από τη σχέση (7.27).

Όπως έχει αναφερθεί, η LDA είναι μία μέθοδος επιβλεπόμενης μάθησης. Στα σύνολα δεδομένων, όμως, που είναι διαθέσιμα για αναγνώριση φωνής δεν υπάρχει η εκ των προτέρων γνώση του φωνήματος, δηλαδή της κλάσης, στο οποίο αντιστοιχεί το κάθε διάνυσμα χαρακτηριστικών. Αυτό που μπορεί να γίνει στην πράξη είναι να εκπαιδευτεί ένα απλό μοντέλο του συστήματος αναγνώρισης, πιθανώς χρησιμοποιώντας ένα υποσύνολο των δεδομένων εκπαίδευσης, με βάση το οποίο θα γίνει εξαναγκασμένη ευθυγράμμιση των δεδομένων. Τα ευθυγραμμισμένα δεδομένα θα χρησιμοποιηθούν στην πορεία για την επανεκπαίδευση του συστήματος με αξιοποίηση της LDA.

Βασικό μειονέκτημα της LDA είναι η εξάρτησή της από τα δεδομένα, με αποτέλεσμα συνήθως να μην αποτελεί αξιόπιστη επιλογή όταν ζητούμενο είναι η ευρωστία σε εφαρμογές με μεγάλη αναντιστοιχία μεταξύ δεδομένων εκπαίδευσης και δεδομένων ελέγχου [34].

7.4 Heteroscedastic Linear Discriminant Analysis (HLDA) και Maximum Likelihood Linear Transform (MLLT)

Εάν θεωρήσουμε ότι γίνεται χρήση ενός Μπεϊζιανού ταξινομητή, η LDA θα δώσει τις βέλτιστες προβολές ως προς τη δυνατότητα διαχωρισμού των κλάσεων εάν τα δεδομένα όλων των κλάσεων ακολουθούν κανονική κατανομή με ίσους πίνακες συμμεταβλητότητας [34]. Άλλωστε, η LDA είναι ισοδύναμη με την εύρεση των ML παραμέτρων ενός γκαουσιανού μοντέλου που θεωρεί ότι όλη η χρήσιμη ως προς την ταξινόμηση πληροφορία βρίσκεται σε έναν p-διάστατο υπόχωρο του αρχικού m-διάστατου χώρου των χαρακτηριστικών και ότι οι ενδοταξικές διακυμάνσεις είναι ίδιες για όλες τις κλάσεις [89]. Στο [90] η ιδέα της LDA γενικεύεται ώστε να παρέχει μία βέλτιστη λύση στην περίπτωση που οι κλάσεις μοντελοποιούνται από GMMs. Ωστόσο, η υπόθεση ενός κοινού πίνακα συμμεταβλητότητας παραμένει.

Στην αναγνώριση φωνής πράγματι οι διαφορετικές κλάσεις μοντελοποιούνται όπως έχουμε δει από GMMs, αλλά η παραπάνω υπόθεση δεν ισχύει, οπότε η λύση που δίνει η LDA δεν

είναι η βέλτιστη. Για να καλυφθεί η περίπτωση των ετεροσκεδαστικών[2] δεδομένων, η LDA επεκτείνεται στο [91], όπου προτείνεται η τεχνική της Ετεροσκεδαστικής Γραμμικής Διακριτικής Ανάλυσης (Heteroscedastic Linear Discriminant Analysis - HLDA). Ακολουθεί μία ανάλυση της HLDA υπό τη σκοπιά της ML και στη συνέχεια δίνεται ένας αποδοτικός αλγόριθμος υπολογισμού που χρησιμοποιείται πλέον στην πράξη.

Όπως αναλύθηκε στην προηγούμενη Ενότητα, στόχος της LDA είναι η βέλτιστη προβολή των αρχικών m-διάστατων δεδομένων σε έναν μικρότερο χώρο έστω l διαστάσεων. Εφόσον απορρίπτονται οι υπόλοιπες $m-l$ διαστάσεις, γίνεται η υπόθεση ότι οι τελευταίες δεν εμπεριέχουν κάποια χρήσιμη ως προς την ταξινόμηση πληροφορία. Θεωρώντας γκαουσιανά μοντέλα για τα αρχικά δεδομένα $\mathbf{x}$, άρα και για τα μετασχηματισμένα $\mathbf{y}$, εφόσον με την LDA εφαρμόζεται ένας γραμμικός μετασχηματισμός, αυτό σημαίνει ότι οι μέσες τιμές και οι διακυμάνσεις για αυτές τις $m-l$ διαστάσεις είναι οι ίδιες για όλες τις κλάσεις. Βάσει της παρατήρησης αυτής, το διάνυσμα μέσων τιμών $\boldsymbol{\mu}_j$ και ο πίνακας συμμεταβλητότητας $\boldsymbol{\Sigma}_j$ που χαρακτηρίζουν την κλάση j μπορούν να γραφούν ως εξής:

$$\boldsymbol{\mu}_j = \begin{bmatrix} \boldsymbol{\mu}_j^l \\ \boldsymbol{\mu}_0 \end{bmatrix}, \tag{7.29}$$

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} \boldsymbol{\Sigma}_j^l & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_0 \end{bmatrix}, \tag{7.30}$$

όπου οι παράμετροι $\boldsymbol{\mu}_0 \in \mathbb{R}^{m-l}$ και $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{(m-l)\times(m-l)}$ είναι κοινές για όλες τις κλάσεις, ενώ οι παράμετροι $\boldsymbol{\mu}_j^l \in \mathbb{R}^l$ και $\boldsymbol{\Sigma}_j^l \in \mathbb{R}^{l\times l}$ μπορούν να διαφέρουν για κάθε κλάση.

Εάν είναι γνωστό ότι το δεδομένο $\mathbf{x}_i$ ανήκει στην κλάση $\mathcal{D}_i$, η αντίστοιχη πυκνότητα πιθανότητας για το μετασχηματισμένο δεδομένο $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$ (όπου δεν έχει ακόμα γίνει μείωση της διαστασιμότητας) θα είναι

$$P(\mathbf{y}_i) = \frac{1}{(2\pi)^{m/2}\left|\boldsymbol{\Sigma}_{\mathcal{D}_i}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_i - \boldsymbol{\mu}_{\mathcal{D}_i}\right)^T \boldsymbol{\Sigma}_{\mathcal{D}_i}^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}_{\mathcal{D}_i}\right)\right\}. \tag{7.31}$$

Οπότε, για το $\mathbf{x}_i$ θα είναι

$$P(\mathbf{x}_i) = \frac{|\mathbf{W}|}{(2\pi)^{m/2}\left|\boldsymbol{\Sigma}_{\mathcal{D}_i}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{W}^T\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{D}_i}\right)^T \boldsymbol{\Sigma}_{\mathcal{D}_i}^{-1}\left(\mathbf{W}^T\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{D}_i}\right)\right\}. \tag{7.32}$$

Η λογαριθμική πιθανοφάνεια $\mathcal{L}$ του συνόλου των n δεδομένων θα είναι, άρα,

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{n} \log P(\mathbf{x}_i) \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left\{\left(\mathbf{W}^T\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{D}_i}\right)^T \boldsymbol{\Sigma}_{\mathcal{D}_i}^{-1}\left(\mathbf{W}^T\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{D}_i}\right) + \log\left((2\pi)^m\left|\boldsymbol{\Sigma}_{\mathcal{D}_i}\right|\right)\right\} + n\log|\mathbf{W}|.\end{aligned} \tag{7.33}$$

Για την εύρεση των ML παραμέτρων $\hat{\mathbf{W}}$, $\hat{\boldsymbol{\Sigma}}_{\mathcal{D}_i}$ και $\hat{\boldsymbol{\mu}}_{\mathcal{D}_i}$, θεωρούμε πρώτα ένα σταθερό πίνακα μετασχηματισμού $\mathbf{W}$ και υπολογίζουμε τις ML παραμέτρους $\hat{\boldsymbol{\Sigma}}_{\mathcal{D}_i}$ και $\hat{\boldsymbol{\mu}}_{\mathcal{D}_i}$. Εν συνεχεία, βάσει

[2] Μία συλλογή μεταβλητών καλείται ετεροσκεδαστική εάν υπάρχουν υποπληθυσμοί με διαφορετικές "μεταβλητότητες" από άλλους. Η έννοια της "μεταβλητότητας" ορίζεται με κάποιο στατιστικό μέτρο διασποράς, όπως η διακύμανση.

αυτών, υπολογίζεται ο βέλτιστος $\hat{\mathbf{W}}$. Για την περίπτωση διαγώνιων πινάκων συμμεταβλητότητας, όπως χρησιμοποιούνται στην πράξη στην αναγνώριση φωνής, και επιβάλλοντας τους περιορισμούς (7.29) και (7.30), προκύπτει τελικά [91]

$$\begin{matrix} \hat{\boldsymbol{\mu}}_i^l = \mathbf{W}_l^T \mathbf{m}_i & , i = 1, 2, \cdots, p \\ \hat{\boldsymbol{\mu}}_0 = \mathbf{W}_{m-l}^T \mathbf{m} \end{matrix} , \tag{7.34}$$

$$\begin{matrix} \hat{\boldsymbol{\Sigma}}_i^l = \mathrm{diag}\left(\mathbf{W}_l^T \mathbf{S}_i \mathbf{W}_l\right) & , i = 1, 2, \cdots, p \\ \hat{\boldsymbol{\Sigma}}_0 = \mathrm{diag}\left(\mathbf{W}_{m-l}^T \mathbf{S}_T \mathbf{W}_{m-l}\right) \end{matrix} , \tag{7.35}$$

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\mathrm{argmax}} \left\{ -\frac{n}{2} \log \left| \mathrm{diag}\left(\mathbf{W}_{m-l}^T \mathbf{S}_T \mathbf{W}_{m-l}\right) \right| - \sum_{i=1}^{p} \frac{n_i}{2} \log \left| \mathrm{diag}\left(\mathbf{W}_l^T \mathbf{S}_i \mathbf{W}_l\right) \right| + n \log |\mathbf{W}| \right\} , \tag{7.36}$$

όπου ο $\mathbf{S}_T$ και ο $\mathbf{m}$ δίνονται από τις σχέσεις (7.24) και (7.23), αντίστοιχα, οι $\mathbf{S}_i$ και οι $\mathbf{m}_i$ από τις σχέσεις (7.21) και (7.10), με το δείκτη i, βέβαια, να παίρνει τις τιμές $1, 2, \cdots, p$, ενώ ο $\mathbf{W}$ διαχωρίζεται ως $\mathbf{W} = [\mathbf{W}_l \mathbf{W}_{m-l}]$, με $\mathbf{W}_l \in \mathbb{R}^{m \times l}$ και $\mathbf{W}_{m-l} \in \mathbb{R}^{m \times (m-l)}$.

Αξίζει να αναφερθεί ότι για την περίπτωση κοινών πινάκων συμμεταβλητότητας $\boldsymbol{\Sigma}_{\mathcal{D}_i} = \boldsymbol{\Sigma}$ $\forall i$, η σχέση (7.36) μετασχηματίζεται στη σχέση

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\mathrm{argmax}} \left\{ -\frac{n}{2} \log \left| \mathrm{diag}\left(\mathbf{W}_{m-l}^T \mathbf{S}_T \mathbf{W}_{m-l}\right) \right| - \frac{n}{2} \log \left| \mathrm{diag}\left(\mathbf{W}_l^T \mathbf{S}_W \mathbf{W}_l\right) \right| + n \log |\mathbf{W}| \right\} , \tag{7.37}$$

όπου ο $\mathbf{S}_W$ δίνεται από τη σχέση (7.20). Αποδεικνύεται [89] ότι οι λύσεις της (7.37) ταυτίζονται με τις λύσεις της (7.19).

Ωστόσο, η (7.36) δεν έχει αναλυτική λύση, πράγμα που σημαίνει ότι πρέπει να χρησιμοποιηθεί μία αριθμητική επαναληπτική μέθοδος για τον υπολογισμό του $\hat{\mathbf{W}}$, όπου σαν αρχική τιμή μπορεί να θεωρηθεί η λύση που προκύπτει από την LDA. Όταν η HLDA εφαρμόζεται για αναγνώριση φωνής, ο υπολογισμός του $\hat{\mathbf{W}}$ μπορεί να ενσωματωθεί στον αλγόριθμο EM για την εκπαίδευση των HMMs, ο οποίος έχει παρουσιαστεί στην Υποενότητα 2.2.3. Οι "κλάσεις", λοιπόν, τώρα είναι οι διάφορες καταστάσεις του HMM. Χρησιμοποιώντας το συμβολισμό του Κεφαλαίου 2, θεωρούμε την ακολουθία παρατηρήσεων $O = \mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T$, το σύνολο καταστάσεων $Q = \{q_1, q_2, \cdots, q_N\}$ και ορίζουμε την πιθανότητα $\gamma_t(i)$ το HMM να βρίσκεται στην κατάσταση i τη χρονική στιγμή t, σύμφωνα με τη σχέση (2.22).

Τότε, οι δειγματικοί μέσοι υπολογίζονται ως

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{t=1}^{T} \gamma_t(i) \mathbf{o}_t \quad , i = 1, 2, \cdots, N \,, \tag{7.38}$$

$$\mathbf{m} = \frac{1}{n} \sum_{t=1}^{T} \mathbf{o}_t \,, \tag{7.39}$$

όπου

$$n_i = \sum_{t=1}^{T} \gamma_t(i) \,, \tag{7.40}$$

$$n = \sum_{i=1}^{N} n_i = \sum_{t=1}^{T} \left\{ \sum_{i=1}^{N} \gamma_t(i) \right\} = \sum_{t=1}^{T} 1 = T \,. \tag{7.41}$$

Οι πίνακες διασποράς και ο πίνακας ολικής διασποράς υπολογίζονται αντίστοιχα ως

$$\mathbf{S}_i = \frac{1}{n_i}\sum_{t=1}^{T}\gamma_t(i)(\mathbf{o}_t - \mathbf{m}_i)(\mathbf{o}_t - \mathbf{m}_i)^T \quad , i = 1, 2, \cdots, N\,, \tag{7.42}$$

$$\mathbf{S}_T = \frac{1}{n}\sum_{t=1}^{T}(\mathbf{o}_t - \mathbf{m})(\mathbf{o}_t - \mathbf{m})^T\,. \tag{7.43}$$

Μέσω των παραπάνω σχέσεων, μπορούν να υπολογιστούν οι ML παράμετροι, όπως στις (7.34) - (7.36). Παρόλο που ανάλυση έγινε για HMMs όπου η πιθανότητα μιας παρατήρησης σε κάποια κατάσταση μοντελοποιείται από μία γκαουσιανή, η HLDA μπορεί εύκολα να χρησιμοποιηθεί και στην περίπτωση χρησης GMMs.

Ένας αποδοτικός αλγόριθμος υπολογισμού του πίνακα μετασχηματισμού για την HLDA προκύπτει από τον αλγόριθμο που προτείνεται για τη μέθοδο των Ημι-Δεμένων πινάκων Συμμεταβλητότητας (Semi-Tied Covariance matrices - STC) ή του Γραμμικού Μετασχηματισμού Μέγιστης Πιθανοφάνειας (Maximum Likelihood Linear Transform - MLLT), όπως είναι αλλιώς γνωστή η ίδια μέθοδος, στην περίπτωση ενός ημι-δεμένου πίνακα [92, 93].

Στόχος της STC μεθόδου είναι να δημιουργήσει μία γέφυρα ανάμεσα στην επιλογή διαγώνιων πινάκων συμμεταβλητότητας που επιφέρει μείωση των ποσοστών αναγνώρισης και την επιλογή των πλήρων πινάκων συμμεταβλητότητας που επιφέρει σημαντική αύξηση της πολυπλοκότητας του συστήματος. Αυτό που προτείνεται, λοιπόν, είναι κάθε κατάσταση i του HMM[3] να χαρακτηρίζεται από έναν διαγώνιο πίνακα συμμεταβλητότητας, έστω $\mathbf{\Sigma}_i$, με κάποιους επιπλέον μη-διαγώνιους πίνακες που διαμοιράζονται μεταξύ διαφορετικών καταστάσεων. Έτσι, εάν στην κατάσταση i αντιστοιχεί ο μη-διαγώνιος πίνακας $\mathbf{H}_r$, τότε ο πίνακας συμμεταβλητότητας $\mathbf{C}_i$ για αυτήν την κατάσταση είναι

$$\mathbf{C}_i = \mathbf{H}_r\mathbf{\Sigma}_i\mathbf{H}_r^T\,. \tag{7.44}$$

Εάν θεωρήσουμε ένα μοναδικό ημι-δεμένο πίνακα $\mathbf{H}$ για κάθε σύνολο καταστάσεων r, τότε προκύπτει η global STC μέθοδος ή MLLT. Ως πίνακα μετασχηματισμού για τον MLLT θα θεωρήσουμε τον πίνακα $\mathbf{A} = (\mathbf{H}^{-1})^T$. Ουσιαστικά, τότε στόχος είναι η εύρεση του βέλτιστου πίνακα μετασχηματισμού για την ελαχιστοποίηση της διαφοράς στην πιθανοφάνεια μεταξύ της μοντελοποίησης με πλήρεις και με διαγώνιους πίνακες συμμεταβλητότητας [94]:

$$\hat{\mathbf{A}} = \underset{\mathbf{A}}{\operatorname{argmax}}\left\{\sum_{i=1}^{p} -\frac{n_i}{2}\left(\log\left|\operatorname{diag}\left(\mathbf{A}^T\hat{\mathbf{\Sigma}}_i\mathbf{A}\right)\right| - \log\left|\left(\mathbf{A}^T\hat{\mathbf{\Sigma}}_i\mathbf{A}\right)\right|\right)\right\} \tag{7.45}$$

Ο υπολογισμός του πίνακα $\mathbf{A}$ γίνεται επαναληπτικά κατά στήλες. Συμβολίζοντας, λοιπόν, ως $\hat{\mathbf{a}}_i$ τη βέλτιστη λύση για την i-οστή στήλη του $\mathbf{A}$, ο αλγόριθμος βασίζεται στη σχέση

$$\hat{\mathbf{a}}_i = \mathbf{G}_i^{-1}\mathbf{c}_i\sqrt{\frac{n}{\mathbf{c}_i^T\mathbf{G}_i^{-1}\mathbf{c}_i}}\,, \tag{7.46}$$

όπου

$$\mathbf{G}_i = \sum_{j=1}^{N}\frac{n_j}{\hat{\sigma}_{j,i}^2}\mathbf{S}_j\,. \tag{7.47}$$

[3] Μπορούμε αντί για καταστάσεις του HMM να σκεφτόμαστε γενικότερα γκαουσιανές κατανομές, εφόσον σε μία κατάσταση αντιστοιχούν στη γενική περίπτωση περισσότερες από μία γκαουσιανές.

Με $\hat{\sigma}^2_{j,i}$ συμβολίζεται το i-οστό στοιχείο της διαγωνίου του $\hat{\mathbf{\Sigma}}_j$, ενώ με $\mathbf{c}_i$ συμβολίζεται το διάνυσμα που περιέχει τα αλγεβρικά συμπληρώματα[4] των στοιχείων του $\hat{\mathbf{a}}_i$.

Παρόλο που οι δύο μέθοδοι (HLDA και MLLT) έχουν προταθεί για να αντιμετωπίσουν διαφορετικά προβλήματα, οι προκύπτουσες συναρτήσεις προς βελτιστοποίηση υπό την ML σκοπιά είναι παρόμοιες. Αν και στο [93] προτείνεται μία παραλλαγή του αλγορίθμου του MLLT που επαναληπτικά δίνει την ακριβή λύση της HLDA, θα χρησιμοποιηθεί εδώ μία προσέγγιση που χρησιμοποιείται στην πράξη, σύμφωνα με την οποία μετά τη σύνδεση διαδοχικών πλαισίων, τα χαρακτηριστικά μετασχηματίζονται με LDA και εν συνεχεία εφαρμόζεται MLLT. Εξάλλου, στο [94] φαίνεται πως η HLDA μπορεί να δώσει χειρότερα αποτελέσματα από την LDA εάν δε συνδυαστεί με τον MLLT.

Για την ακρίβεια, λοιπόν, υπολογίζεται πρώτα ο πίνακας μετασχηματισμού $\mathbf{W}$, όπως προκύπτει από την LDA. Εν συνεχεία, βάσει των μετασχηματισμένων χαρακτηριστικών, εκπαιδεύεται ένα (τριφωνικό) μοντέλο και σε συγκεκριμένες επαναλήψεις της εκπαίδευσης λαμβάνει χώρα η ανανέωση του πίνακα μετασχηματισμού $\mathbf{A}$ του MLLT. Σε κάθε τέτοια επανάληψη ανανεώνονται όλες οι μέσες τιμές $\boldsymbol{\mu}_j$ βάσει του $\mathbf{A}$ και ανανεώνεται ο ολικός πίνακας μετασχηματισμού σύμφωνα με τη σχέση $\mathbf{W} = \mathbf{AW}$.

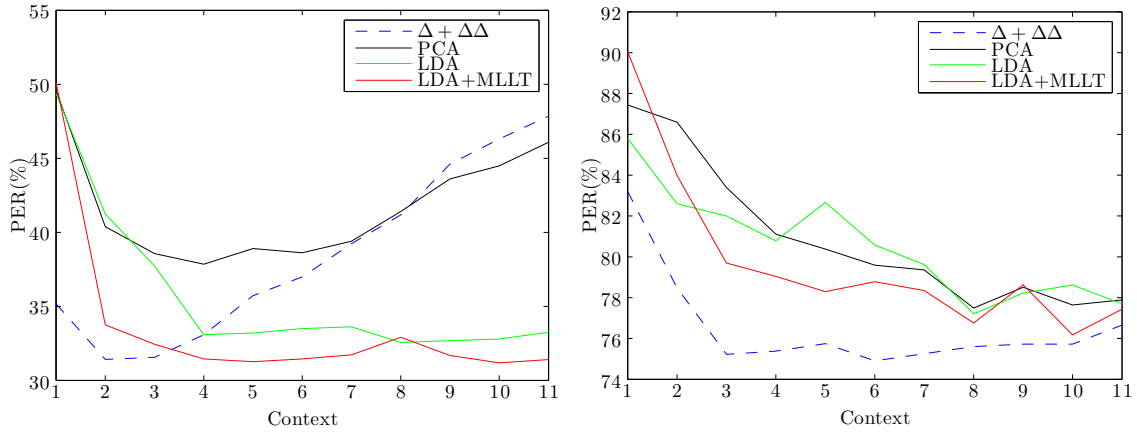
7.5 Πειραματικά Αποτελέσματα

Όπως έχουμε πει, ο MLLT βασίζεται σε έναν επαναληπτικό αλγόριθμο. Οπότε, ένα ερώτημα είναι ο αριθμός των επαναλήψεων που απαιτούνται. Μετά από πειραματικές δοκιμές είδαμε ότι από την πρώτη ήδη επανάληψη τα ποσοστά σφάλματος μειώνονται σχεδόν στην ελάχιστη δυνατή τιμή τους, ενώ μετά την πέμπτη επανάληψη δεν παρατηρείται κάποια μείωση των εν λόγω ποσοστών. Συνεπώς, σε όλα τα πειράματα που θα ακολουθήσουν ο υπολογισμός του MLLT γίνεται με 5 επαναλήψεις.

Στο Σχήμα 7.3 παρουσιάζονται συγκριτικά τα αποτελέσματα όταν χρησιμοποιείται PCA, LDA και MLLT (πάνω στα αποτελέσματα της LDA), καθώς μεταβάλλεται το χρονικό παράθυρο (σε πλαίσια) που λαμβάνεται υπόψιν για την εξαγωγή των εκάστοτε χαρακτηριστικών. Ως context ορίζεται ο αριθμός πλαισίων αριστερά και δεξιά του κεντρικού πλαισίου, τα οποία συνδέονται μεταξύ τους. Σε κάθε περίπτωση, το τελικό διάνυσμα χαρακτηριστικών αποτελείται από 39 χαρακτηριστικά, όσα, δηλαδή και όταν χρησιμοποιούνται MFCCs+Δ+ΔΔ. Για αναφορά, δίνονται και τα αποτελέσματα με χρήση MFCCs+Δ+ΔΔ, όπως παρουσιάζονται και στο Σχήμα 5.4 (αφού έχει γίνει κανονικοποίηση του σήματος). Στην τελευταία περίπτωση, όλα τα στάδια της εκπαίδευσης (μονοφωνικό και δύο περάσματα του τριφωνικού μοντέλου) γίνονται με MFCCs+Δ+ΔΔ, όπου οι Δ και ΔΔ συντελεστές λαμβάνονται με το context που φαίνεται στο Σχήμα. Στις υπόλοιπες περιπτώσεις, η εκπαίδευση του μονοφωνικού και του πρώτου περάσματος του τριφωνικού μοντέλου γίνονται με χρήση MFCCs+Δ+ΔΔ, όπου, όμως, οι Δ και ΔΔ συντελεστές λαμβάνονται με σταθερό context=3. Τα χαρακτηριστικά που προκύπτουν από PCA, LDA ή LDA+MLLT χρησιμοποιούνται μόνο κατά το τελευταίο πέρασμα του τριφωνικού μοντέλου, με το context που φαίνεται στο Σχήμα.

Ανάμεσα στις τρεις τεχνικές μείωσης της διαστασιμότητας που συγκρίνονται, η LDA μαζί με τον MLLT φαίνεται πως πετυχαίνει το μικρότερο σφάλμα αναγνώρισης, ανεξαρτήτως του context που επιλέγεται. Όσον αφορά στις συνθήκες του μικροφώνου CT1, η PCA φαίνεται καθαρά πως αποτυγχάνει να αντικατοπτρίσει τα δυναμικά χαρακτηριστικά του σήμα-

[4] Το αλγεβρικό συμπλήρωμα του στοιχείου a_{ij} του πίνακα $\mathbf{A}$ ισούται με $(-1)^{i+j}D_{ij}$, όπου D_{ij} η ορίζουσα που προκύπτει αν από τον $\mathbf{A}$ διαγραφεί η i-γραμμή και η j-στήλη.



(i) Μικρόφωνο CT1 (ii) Μικρόφωνο OA6

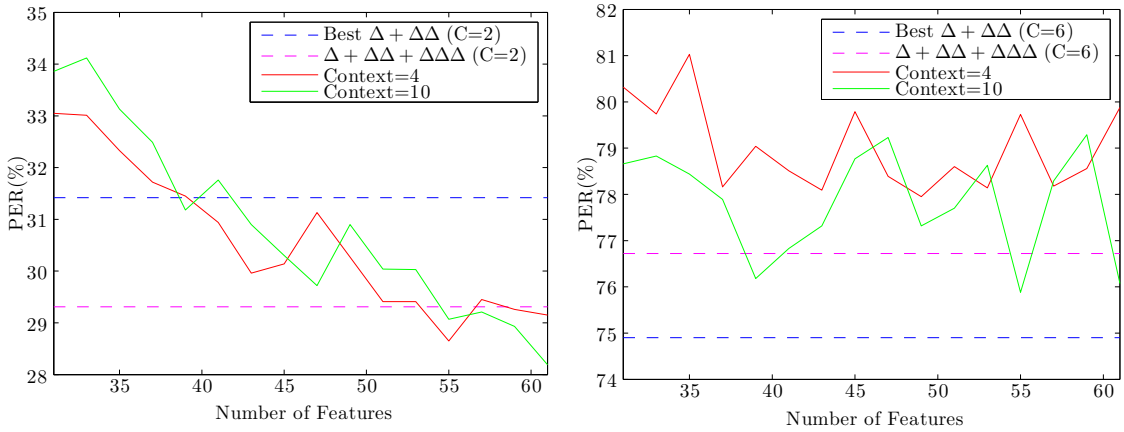
Σχήμα 7.3: PER(%) καθώς μεταβάλλεται ο αριθμός διαδοχικών πλαισίων που συνδέονται προτού λάβει χώρα η PCA, η LDA ή η LDA με MLLT. Σε κάθε περίπτωση, συνδέονται 2·Context+1 πλαίσια και κρατιούνται 39 χαρακτηριστικά. Για αναφορά, δίνεται και το PER(%) καθώς μεταβάλλεται το χρονικό παράθυρο που λαμβάνεται υπόψιν για την εξαγωγή των δυναμικών χαρακτηριστικών, όταν γίνεται χρήση Δ και ΔΔ συντελεστών.

τος. Η LDA και η LDA+MLLT, από την άλλη, παρουσιάζουν πολύ καλύτερα αποτελέσματα, χωρίς, ωστόσο, να καταφέρνουν να ξεπεράσουν σημαντικά την απόδοση των Δ+ΔΔ συντελεστών. Παρόλα αυτά, η επιλογή της LDA, σε συνδυασμό με τον MLLT, φαίνεται να είναι πιο "σίγουρη", καθώς παρουσιάζει πολύ μεγαλύτερη ευρωστία, όσον αφορά στην επιλογή του κατάλληλου context.

Όσον αφορά στις συνθήκες του μικροφώνου OA6, δηλαδή σε πραγματικές συνθήκες αναγνώρισης από απόσταση, και οι τρεις τεχνικές φαίνεται να επωφελούνται από την αύξηση του context, με το σφάλμα να παρουσιάζει μία συνεχή πτωτική τάση. Καμία, όμως, δεν καταφέρνει να ξεπεράσει την απόδοση των Δ+ΔΔ συντελεστών, οι οποίοι, μάλλον απρόσμενα, δίνουν σταθερά μικρότερο σφάλμα.

Στη συνέχεια, εξετάζουμε την επίδραση του αριθμού των χαρακτηριστικών που αποτελούν το τελικό διάνυσμα χαρακτηριστικών μετά την εφαρμογή LDA και MLLT. Για καθένα από τα δύο μικρόφωνα εξετάζονται δύο περιπτώσεις, όταν το context ισούται με 4 πλαίσια, που είναι ίσως η πιο συνηθισμένη περίπτωση, και όταν το context ισούται με 10 πλαίσια, επιλογή που έδωσε τα καλύτερα αποτελέσματα στα προηγούμενα πειράματα και για τα δύο μικρόφωνα. Τα αποτελέσματα παρουσιάζονται στο Σχήμα 7.4. Για λόγους σύγκρισης, δίνεται ως αναφορά και το PER όταν γίνεται χρήση Δ+ΔΔ συντελεστών και Δ+ΔΔ+ΔΔΔ συντελεστών, όπου το διάνυσμα χαρακτηριστικών είναι μήκους 39 και 52, αντίστοιχα. Στις περιπτώσεις αυτές, το context λήφθηκε ίσο με τη βέλτιστη επιλογή, όπως προκύπτει από τα αποτελέσματα των πειραμάτων που συνοψίζονται στο Σχήμα 5.4.

Η ανάλυση των αποτελεσμάτων που αφορούν το μικρόφωνο CT1 είναι άμεση. Με οποιαδήποτε από τις δύο επιλογές context, το σφάλμα ακολουθεί πτωτική τάση καθώς αυξάνονται τα χαρακτηριστικά, αλλά φαίνεται πως η επιλογή του μικρότερου context είναι προτιμητέα κατά μέσο όρο. Όσον αφορά στη σύγκριση με τους Δ συντελεστές, φαίνεται πως δεν υπάρχει κάποιος ιδιαίτερος λόγος να προτιμηθεί η LDA+MLLT, εάν γίνει χρήση ίσου ή μικρότερου διανύσματος χαρακτηριστικών με αυτό που προκύπτει από τη χρήση Δ+ΔΔ(+ΔΔΔ) συντελεστών. Εάν, όμως, αυξήσουμε το διάνυσμα χαρακτηριστικών, τότε με την LDA+MLLT



(i) Μικρόφωνο CT1 (ii) Μικρόφωνο OA6

Σχήμα 7.4: PER(%) καθώς μεταβάλλεται το μήκος του τελικού διανύσματος χαρακτηριστικών μετά την LDA και τον MLLT. Για αναφορά, δίνεται και το PER(%) όταν γίνεται χρήση Δ και ΔΔ συντελεστών (39 χαρακτηριστικά), καθώς και Δ, ΔΔ και ΔΔΔ συντελεστών (52 χαρακτηριστικά). Στις τελευταίες περιπτώσεις, χρησιμοποιήθηκε ως χρονικό παράθυρο αυτό που προέκυψε ως βέλτιστο κατά τη χρήση Δ και ΔΔ συντελεστών, σύμφωνα με το Σχήμα 5.4.

προκύπτει σαφής βελτίωση των αποτελεσμάτων.

Το τοπίο, ωστόσο, αλλάζει εάν δούμε τα αποτελέσματα που αφορούν το μικρόφωνο OA6. Εδώ, φαίνεται να υπάρχει μία σαφής προτίμηση προς το μεγαλύτερο context, αλλά αύξηση του αριθμού των χαρακτηριστικών δεν επιφέρει απαραίτητα μικρότερα σφάλματα αναγνώρισης. Ακόμα, η σύγκριση της LDA+MLLT με τους Δ+ΔΔ συντελεστές ευνοεί τους τελευταίους, όπως και στο Σχήμα 7.3ii. Τέλος, παρόλο που με συγκεκριμένες παραμετροποιήσεις, η απόδοση που προκύπτει από χρήση LDA+MLLT φαίνεται να μπορεί να ξεπεράσει την απόδοση που προκύπτει από χρήση Δ+ΔΔ+ΔΔΔ συντελεστών, η χρήση των τελευταίων (ΔΔΔ) δεν αποτελεί συνετή επιλογή, εφόσον όχι μόνο δε βελτιώνουν, αλλά χειροτερεύουν την απόδοση του συστήματος αναγνώρισης. Η παρατήρηση αυτή αποτελεί άλλη μία ένδειξη υπέρ της χρήσης συντελεστών παραγώγισης μόνο έως δευτέρου βαθμού και όχι μεγαλύτερου.

Κεφάλαιο 8

Τελεστής Teager Ενέργειας και AM-FM Χαρακτηριστικά

8.1 Τελεστής Teager Ενέργειας

Όπως ήδη έχουμε πει από το Κεφάλαιο 5, το ενεργειακό περιεχόμενο του σήματος αποτελεί το βασικό στοιχείο που λαμβάνεται υπόψιν κατά την εξαγωγή χαρακτηριστικών για Αυτόματη Αναγνώριση Φωνής, τουλάχιστον όσον αφορά στις κλασικές και πλέον αποδεκτές μεθόδους. Ο πιο συνήθης τρόπος εκτίμησης της στιγμιαίας ενέργειας ενός σήματος $x(t)$ είναι μέσω του τετραγώνου του σήματος αυτού ως

$$S_c[x(t)] \triangleq x^2(t)\,, \tag{8.1}$$

ενώ όμοια για ένα σήμα διακριτού χρόνου $s[n]$ έχουμε

$$S_d[s[n]] \triangleq s^2[n]\,. \tag{8.2}$$

Καλούμε τον τελεστή $S_c[\cdot]$ (ή τον $S_d[\cdot]$) Τελεστή Τετραγωνικής Ενέργειας (Squared Energy Operator - SEO) [95].

Ένας εναλλακτικός τρόπος υπολογισμού της ενέργειας προτάθηκε από τον Teager με το σχετικό τελεστή να εισάγεται φορμαλιστικά και να μελετάται συστηματικά από τον Kaiser [96, 97], γι' αυτό και τον καλούμε Τελεστή Teager Ενέργειας (Teager Energy Operator - TEO) ή και Τελεστή Teager-Kaiser Ενέργειας:

$$\Psi_c[x(t)] \triangleq \left(\frac{dx(t)}{dt}\right)^2 - x(t)\frac{d^2x(t)}{dt} = \dot{x}^2(t) - x(t)\ddot{x}(t)\,, \tag{8.3}$$

$$\Psi_d[s[n]] \triangleq s^2[n] - s[n-1]s[n+1]\,. \tag{8.4}$$

Ο χαρακτηρισμός του TEO ως "ενεργειακού τελεστή" (στη συνεχή περίπτωση) γίνεται εμφανής εάν θεωρήσουμε ένα σύστημα που εκτελεί ελεύθερη ταλάντωση (μη-αποσβενύμενη), το οποίο αποτελείται από σώμα μάζας m και ελατήριο σταθεράς $k > 0$ [98]. Η θέση του σώματος δίνεται από την εξίσωση κίνησης $m\ddot{x} + kx = 0$, η οποία επιδέχεται γενική λύση της μορφής

$$x(t) = A\cos\left(\omega_0 t + \theta\right), \quad \omega_0 = \sqrt{\frac{k}{m}}\,, \tag{8.5}$$

όπου A το πλάτος, ω_0 η συχνότητα και θ η αρχική φάση της ταλάντωσης. Όπως είναι γνωστό, η συνολική ενέργεια του συστήματος δίνεται ως το άθροισμα κινητικής και δυναμικής ενέργειας και είναι σταθερή στο χρόνο:

$$E_0 = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}kx^2 = \frac{m}{2}A^2\omega_0^2 \,. \tag{8.6}$$

Εφαρμόζοντας, τώρα, τον TEO, έχουμε:

$$\begin{aligned}\Psi_c[x(t)] &= \Psi_c[A\cos(\omega_0 t + \theta)] \\ &= (-A\omega_0 \sin(\omega_0 t + \theta))^2 - A\cos(\omega_0 t + \theta)\left(-A\omega_0^2\cos(\omega_0 t + \theta)\right) \\ &= A^2\omega_0^2[\sin^2(\omega_0 t + \theta) + \cos^2(\omega_0 t + \theta)] \\ &= A^2\omega_0^2 = \frac{E_0}{(m/2)} \,. \end{aligned} \tag{8.7}$$

Συνεπώς, πράγματι ο TEO εκφράζει τη συνολική ενέργεια του ταλαντωτή (ανά ημιμονάδα μάζας).

Υπό συγκεκριμένες συνθήκες, παρόμοια αποτελέσματα αποδεικνύεται ότι ισχύουν στη γενική περίπτωση ενός σήματος που έχει υποστεί τόσο Διαμόρφωση Πλάτους (Amplitude Modulation - AM) όσο και Διαμόρφωση Συχνότητας (Frequency Modulation - FM). Οι έννοιες των AM-FM σημάτων αναπτύχθηκαν αρχικά στον τομέα των Τηλεπικοινωνιών [99], αλλά τα τελευταία χρόνια έχουν βρει εφαρμογή στην Επεξεργασία Φωνής, μετά την εισαγωγή του αντίστοιχου μοντέλου για τα σήματα φωνής. Συγκεκριμένα, όπως παρουσιάζεται στο [98], ένα σήμα φωνής $s(t)$ μπορεί να μοντελοποιηθεί ως το άθροισμα R AM-FM ζωνοπεριορισμένων σημάτων, όπου καθένα αντιστοιχεί σε ένα formant, καθώς κάθε formant μπορεί να μεταβάλλεται γρήγορα, τόσο ως προς το πλάτος, όσο και ως προς τη συχνότητα:

$$s(t) = \sum_{r=1}^{R} a_r(t)\cos(\phi_r(t)) \,. \tag{8.8}$$

Θεωρώντας τη συχνότητα του φέροντος ίση με $\omega_{r,c}$, η γωνία $\phi_r(t)$ μπορεί να εκφραστεί ως

$$\phi_r(t) = \omega_{r,c}t + \omega_{r,m}\int_0^t q_r(\tau)d\tau + \theta_r \,, \tag{8.9}$$

ενώ ορίζοντας τη στιγμιαία συχνότητα $\omega_r(t) = \dot{\phi}_r(t)$, λαμβάνουμε την πιο συμπαγή έκφραση

$$\phi_r(t) = \int_0^t \omega_r(\tau)d\tau + \theta_r \,. \tag{8.10}$$

Με την υπόθεση ότι τα σήματα πληροφορίας $a_r(t)$ και $\omega_r(t)$ δε μεταβάλλονται με πολύ μεγάλη ταχύτητα, αλλά ούτε και σε πολύ μεγάλο βαθμό σε σύγκριση με τη συχνότητα φέροντος, τότε αποδεικνύεται [98] ότι

$$\Psi_c\left[a_r(t)\cos\left(\int_0^t \omega_r(\tau)d\tau + \theta_r\right)\right] \approx a_r^2(t)\omega_r^2(t) \,. \tag{8.11}$$

8.2 TEO στο Πεδίο Συχνότητας

Ο TEO, σύμφωνα και με την ανάλυση της Ενότητας 8.1, φαίνεται να αντικατοπτρίζει ακριβέστερα από το SEO το ενεργειακό περιεχόμενο ενός σήματος ή, πιο αυστηρά, την απαιτούμενη ενέργεια που χρειάζεται η πηγή για την παραγωγή του εκάστοτε σήματος. Για το λόγο αυτό, έχουν γίνει προσπάθειες αξιοποίησής του στο πεδίο της αναγνώρισης φωνής με υποσχόμενα αποτελέσματα.

Στο [100] το διάνυσμα των 12 MFCCs προσαυξάνεται με την Teager ενέργεια όπως υπολογίζεται από τον TEO ή από τον TEO στη Mel κλίμακα. Στο [101] εισάγονται οι βασισμένοι στον TEO Αναφασματικοί (TEO based CEPstral - TEOCEP) παράμετροι χαρακτηριστικών, όπου σε κάθε μπάντα συχνοτήτων και για κάθε παραθυρωμένο σήμα υπολογίζεται η μέση Teager ενέργεια και οι οποίοι εξάγεται ότι παρουσιάζουν μεγαλύτερη ευρωστία στην ύπαρξη θορύβου αυτοκινήτου. Μία παρόμοια προσέγγιση γίνεται στο [102], όπου με τη βοήθεια του TEO αναλύεται η έννοια του Φάσματος Ενέργειας (Energy Spectrum) [103], κατ' αντιστοιχία του Φάσματος Ισχύος (Power Spectrum), ενώ τα πειραματικά αποτελέσματα δείχνουν ότι η χρήση των δύο αυτών φασμάτων για εξαγωγή short-term χαρακτηριστικών οδηγεί σε συστήματα με συγκρίσιμες αποδόσεις. Στο [104] προτείνονται οι ακουστικοί Συντελεστές Αναφάσματος με Teager Ενέργειες (auditory Teager Energy Cepstrum Coefficients - TECCs), όπου χρησιμοποιείται συστοιχία gammatone φίλτρων για το διαχωρισμό του σήματος σε κρίσιμες μπάντες συχνοτήτων, σε κάθε μία από τις οποίες υπολογίζεται η μέση ενέργεια βραχέος χρόνου με τη βοήθεια του TEO. Τα πειραματικά αποτελέσματα δείχνουν πως τα TECCs παρουσιάζουν συγκρίσιμη απόδοση με τα MFCCs σε καθαρές συνθήκες, ενώ δίνουν σημαντικά βελτιωμένα αποτελέσματα σε διάφορα είδη προσθετικού θορύβου.

Ωστόσο, σε όλες τις παραπάνω προσεγγίσεις, ο TEO χρησιμοποιείται στο πεδίο του χρόνου, όπως στις σχέσεις (8.3) και (8.4). Αυτό σημαίνει πως για να συνδυαστεί ο TEO με οποιαδήποτε συστοιχία φίλτρων θα πρέπει να λάβουν χώρα οι κατάλληλες συνελίξεις με τις αποκρίσεις συχνότητας των εκάστοτε φίλτρων. Κάτι τέτοιο αποτελεί σαφές μειονέκτημα, από υπολογιστικής άποψης, σε σύγκριση με τις συνήθεις μεθόδους εξαγωγής χαρακτηριστικών, όπου οι διάφορες λειτουργίες γίνονται στο πεδίο συχνότητας, με την πράξη της συνέλιξης να αντικαθίσταται από την απλή πράξη του πολλαπλασιασμού, όπως στις σχέσεις (5.11) και (5.12).

Ο TEO, όμως, έχει χρησιμοποιηθεί στην πράξη και στο πεδίο συχνότητας, οπότε και αντιμετωπίζεται το παραπάνω υπολογιστικό θέμα. Στο [105] προτείνονται οι Συντελεστές Ισχύος στις Log Συχνότητες (Log Frequency Power Coefficients - LFPCs), οι οποίοι συνδυάζονται με τον TEO, είτε στο πεδίο του χρόνου είτε στο πεδίο της συχνότητας, και χρησιμοποιούνται για ταξινόμηση του stress σε σήματα φωνής. Από τα πειραματικά αποτελέσματα, μάλιστα, φαίνεται πως όταν ο TEO χρησιμοποιείται στο πεδίο συχνότητας δίνει καλύτερα αποτελέσματα. Ο TEO στο πεδίο της συχνότητας για συστήματα εύρωστης αναγνώρισης φωνής έχει χρησιμοποιηθεί στα [106] και [107], όπου συνδυάζεται αντίστοιχα με τη Διαφορά του Φάσματος Ισχύος (Power Spectrum Difference - PSD) και με τη CMN, με τα προκύπτοντα χαρακτηριστικά στην τελευταία περίπτωση να καλούνται Κανονικοποιημένα ως προς την Teager Ενέργεια Αναφασματικά (Normalized Teager Energy Cepstral - NTEC) χαρακτηριστικά.

Στο πλαίσιο αυτό, λέγοντας πως ο TEO χρησιμοποιείται στο πεδίο συχνότητας, εννοείται πως λαμβάνεται ο DFT του τυχόντος σήματος $s[n]$, έστω $S[k]$, και ο τελεστής $\Psi[\cdot]$[1] λειτουργεί

[1] Για λόγους απλότητας, στη συνέχεια απαλείφουμε από το συμβολισμό των SEO και TEO τους δείκτες c και d, καθώς θα γίνεται εμφανές εάν γίνεται αναφορά στο συνεχή ή στο διακριτό χρόνο.

ακριβώς πάνω σε αυτό το διάνυσμα:

$$\Psi[S[k]] = S^2[k] - S[k-1]S[k+1]\,. \tag{8.12}$$

Σαφώς, από την παραπάνω διαδικασία προκύπτουν γενικά φανταστικές τιμές, οπότε, για να ληφθεί η ενέργεια κατά τρόπο αντίστοιχο της σχέσης (5.10), θα πρέπει να χρησιμοποιηθεί το μέτρο των αντίστοιχων τιμών, όπως στη σχέση

$$\frac{1}{N}\sum_{k=0}^{N-1}|\Psi[S[k]]|\,. \tag{8.13}$$

Μία ελαφρώς διαφορετική προσέγγιση ακολουθείται στο [108], όπου χρησιμοποιείται ο ορισμός του TEO για μιγαδικά σήματα [109]:

$$\Phi[S[k]] \triangleq \Psi[\mathrm{Re}\{S[k]\}] + \Psi[\mathrm{Im}\{S[k]\}]\,. \tag{8.14}$$

Τα προκύπτοντα χαρακτηριστικά καλούνται Βασισμένα στην Teager Ενέργεια MFCCs (Teager Energy based MFCCs - TEMFCCs) και φαίνεται να παρουσιάζουν αυξημένη ευρωστία κατά την αναγνώριση συναισθήματος μέσω της φωνής.

Ενώ ο TEO στο πεδίο του χρόνου προσδιορίζει, όπως έχουμε δει, την πραγματική ενέργεια ενός σήματος, η χρήση του στο πεδίο της συχνότητας σύμφωνα είτε με τη σχέση (8.12) είτε με τη σχέση (8.14) είναι καθαρά ευριστική και δεν έχει κάποια διαισθητική ερμηνεία. Θέλοντας να κρατήσουμε τα υπολογιστικά πλεονεκτήματα του TEO όταν αυτός χρησιμοποιείται στο πεδίο συχνότητας, αλλά παράλληλα αυτός να διατηρεί την πρωταρχική του φύση ως ενεργειακού τελεστή, προτείνουμε ένα νέο πλαίσιο εργασίας για χρήση του TEO.

Αντίστοιχα με τη σχέση (5.10), θεωρώντας ότι δουλεύουμε με πραγματικά σήματα, υπολογίζουμε την ενέργεια βραχέος χρόνου με χρήση του TEO αντί του SEO και με τη βοήθεια των θεωρημάτων Parseval και Plancherel έχουμε

$$\begin{aligned}
\sum_{n=0}^{N-1}\Psi[s[n]] &= \sum_{n=0}^{N-1}s^2[n] - \sum_{n=0}^{N-1}s[n-1]s[n+1] \\
&= \frac{1}{N}\sum_{k=0}^{N-1}|S[k]|^2 - \frac{1}{N}\sum_{k=0}^{N-1}S_-[k]S_+^*[k] \\
&= \frac{1}{N}\sum_{k=0}^{N-1}\left\{|S[k]|^2 - S_-[k]S_+^*[k]\right\} \\
&\triangleq \frac{1}{N}\sum_{k=0}^{N-1}S_{(t)}[k]\,,
\end{aligned} \tag{8.15}$$

όπου συμβολίζουμε με $S_-[k]$ τον DFT του $s[n-1]$ και με $S_+^*[k]$ το μιγαδικό συζυγή του DFT του $s[n+1]$. Εφόσον δουλεύουμε προφανώς με παραθυρωμένα σήματα, για να χρησιμοποιήσουμε την ιδιότητα της χρονικής μετατόπισης του DFT θα έπρεπε να υποθέσουμε ότι εκτός του εκάστοτε παραθύρου κάθε παραθυρωμένο σήμα λαμβάνει μηδενικές τιμές. Αντ' αυτού, δουλεύουμε ως εξής. Έστω το σήμα φωνής $s[n]$ N' δειγμάτων

$$s = \{s[1], s[2], s[3], \cdots, s[N'-2], s[N'-1], s[N']\}\,.$$

Ορίζουμε τα παρακάτω σήματα:

$$\begin{aligned}\tilde{s} &\triangleq \{s[2], s[3], \cdots, s[N'-2], s[N'-1]\}\,,\\ \tilde{s}_- &\triangleq \{s[1], s[2], s[3], \cdots, s[N'-2]\}\,,\\ \tilde{s}_+ &\triangleq \{s[3], \cdots, s[N'-2], s[N'-1], s[N']\}\,.\end{aligned}$$

Παραθυρώνουμε και τα τρία αυτά σήματα ξεχωριστά, έστω με παράθυρο Hamming, και προκύπτει ένας ίδιος αριθμός παραθύρων για τα τρία σήματα, εφόσον και τα τρία έχουν ίδιο αριθμό δειγμάτων (ίσο με $N'-2$). Για το i-οστό παράθυρο υπολογίζουμε το φάσμα $S_{(t)i}[k]$, σύμφωνα με τη σχέση (8.15), το οποίο ονομάζουμε Φάσμα Teager Ισχύος (Teager Power Spectrum - TPS):

$$S_{(t)i}[k] = |\tilde{S}_i[k]|^2 - \tilde{S}_{i-}[k]\tilde{S}^*_{i+}[k]\,, \tag{8.16}$$

όπου $\tilde{S}_i[k]$ ο DFT N σημείων του $\tilde{s}_i[n]$ και όμοια για τα άλλα δύο σήματα. Το μέτρο του TPS παρουσιάζεται οπτικά σε μορφή φασματογραφήματος στο Σχήμα 8.1, σε αντιπαράθεση με το κλασικό φασματογράφημα για ένα τυχαίο σήμα φωνής.

Το TPS μπορεί να χρησιμοποιηθεί για την εξαγωγή των ακουστικών χαρακτηριστικών αλλάζοντας ελαφρώς τον αλγόριθμο των γνωστών μεθόδων που λειτουργούν στο πεδίο συχνότητας με ένα μικρό επιπλέον υπολογιστικό κόστος, λόγω των παραπάνω DFTs που απαιτούνται. Για παράδειγμα, η σχέση (5.11) μετασχηματίζεται στη σχέση

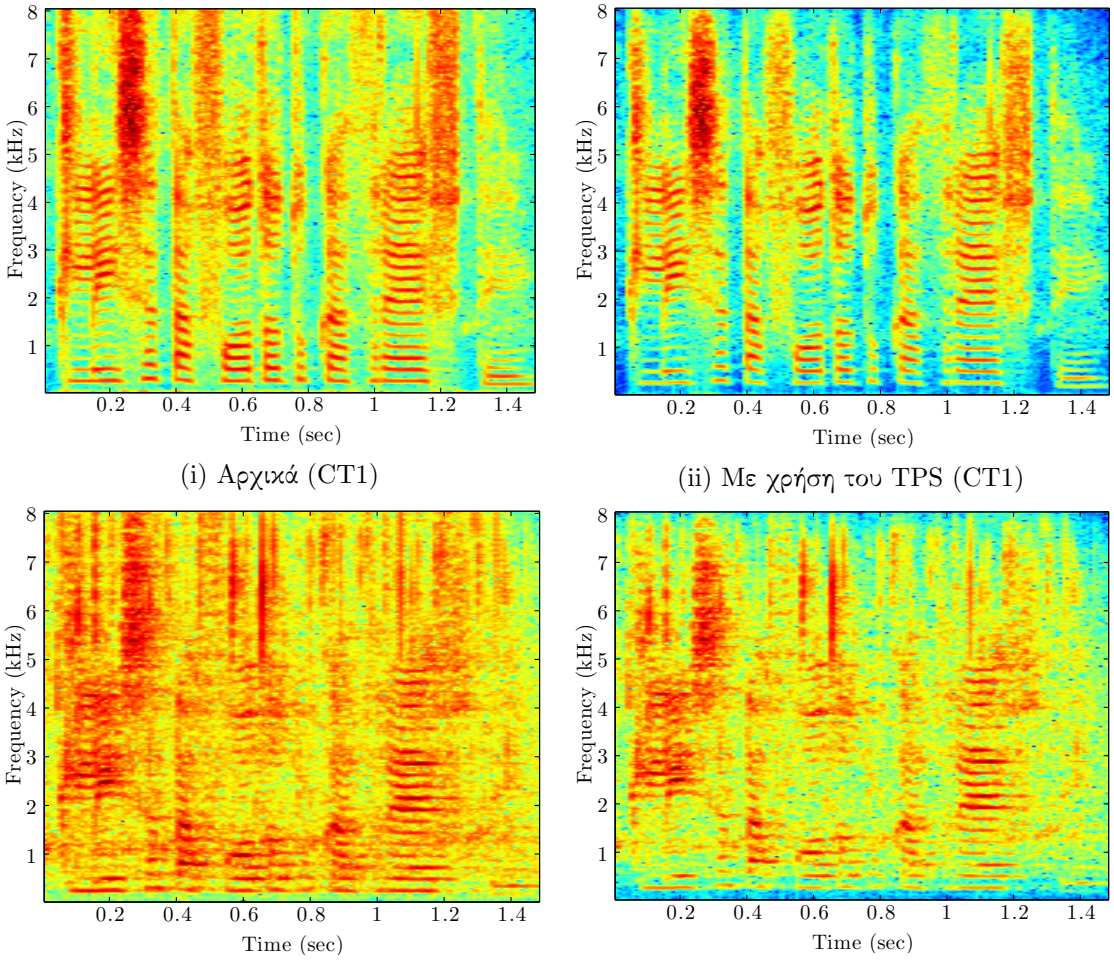
$$G_i(j) = \log\left\{\frac{2}{N}\sum_{k=0}^{N/2}\left|\sqrt{S_{(t)i}[k]}\cdot H^j[k]\right|^2\right\}\,. \tag{8.17}$$

Όμοια, η σχέση (5.12) μετασχηματίζεται στη σχέση

$$G_i(j) = \log\left\{\sum_{k=0}^{N/2}\left|S_{(t)i}[k]\cdot H^j[k]\right|\right\}\,. \tag{8.18}$$

Όπως έχουμε πει, η χρήση του ΤΕΟ αντί του SEO προτείνεται γιατί αυτός θεωρείται πως δίνει με μεγαλύτερη ακρίβεια και ευρωστία την ενέργεια του σήματος. Ωστόσο, όπως αναλύεται στο [95], όσον αφορά στο διακριτό χρόνο που ουσιαστικά μας ενδιαφέρει εδώ, ο ΤΕΟ είναι πιο αξιόπιστος για χαμηλές συχνότητες, ενώ όσο οι συχνότητες πλησιάζουν τη συχνότητα Nyquist, η χρήση του SEO είναι προτιμότερη. Συγκεκριμένα, τα εν λόγω αποτελέσματα προκύπτουν από πειράματα που γίνονται σε σήματα φωνής, αφού αυτά έχουν παραθυρωθεί σε frames των $30msec$ και έχουν διαχωριστεί σε συχνοτικές μπάντες από μια συστοιχία φίλτρων.

Σε θεωρητικό επίπεδο, η χρήση ενός αρμονικού μοντέλου για την προσέγγιση του προσθετικού θορύβου αναδεικνύει ότι οι μεταβατικοί όροι σε τέτοιας διάρκειας παράθυρα είναι πιο έντονοι όταν χρησιμοποιείται ο SEO σε σύγκριση με τον ΤΕΟ, καθώς επίσης ότι η επίδρασή τους είναι αντιστρόφως ανάλογη της συχνότητας. Ακόμα, σημειώνεται πως ο ΤΕΟ στο διακριτό χρόνο προκύπτει από μια διαδικασία διακριτοποίησης όπου η παραγώγιση προσεγγίζεται από πρώτες διαφορές, γεγονός που φυσιολογικά επιφέρει κάποιο σφάλμα προσέγγισης, το οποίο, όμως, γίνεται μεγαλύτερο στις υψηλές συχνότητες. Συνεπώς, για τα πρώτα φίλτρα της συστοιχίας, ο ΤΕΟ δίνει μια πιο εύρωστη προσέγγιση της μέσης ενέργειας του σήματος, ενώ για τα τελευταία φίλτρα υπερτερεί ο SEO. Τα εν λόγω σφάλματα εκτίμησης της ενέργειας



(i) Αρχικά (CT1)

(ii) Με χρήση του TPS (CT1)

(iii) Αρχικά (OA6)

(iv) Με χρήση του TPS (OA6)

Σχήμα 8.1: Αρχικό φασματογράφημα ((i), (iii)) και φασματογράφημα με χρήση του μέτρου του TPS ((ii), (iv)) για την ηχογράφηση της φράσης "*Κλείσε τα φώτα του καθρέφτη*" από τα μικρόφωνα CT1 ((i), (iii)) και OA6 ((ii), (iv))), η οποία είναι αλλοιωμένη λόγω θορύβου ανεμιστήρα στο υπόβαθρο (στην περίπτωση του OA6).

με χρήση του SEO ή του TEO διαθλώνται και στο αναφασματικό πεδίο [110], με αρνητικές συνέπειες για την αναγνώριση.

Βάσει των παραπάνω παρατηρήσεων, προτείνουμε να γίνεται χρήση του SEO - μέσω του φάσματος ισχύος - για τα πρώτα έστω M φίλτρα και του TEO - μέσω του TPS - για τα υπόλοιπα. Με το σκεπτικό αυτό, μπορούν, για παράδειγμα, οι σχέσεις (5.12) και (8.18) να συνδυαστούν ως εξής:

$$G_i(j) = \begin{cases} \log\left\{\sum_{k=0}^{N/2} \left|S_{(t)i}[k] \cdot H^j[k]\right|\right\}, & j \leq M \\ \log\left\{\sum_{k=0}^{N/2} \left|S_i^2[k] \cdot H^j[k]\right|\right\}, & j > M \end{cases} . \tag{8.19}$$

Βεβαίως, το πλαίσιο εργασίας που εισάγεται μέσω της παραπάνω ανάλυσης είναι γενικό

και δεν αφορά μία μόνο συγκεκριμένη μέθοδο εξαγωγής χαρακτηριστικών. Έτσι, λοιπόν, για να εξετάσουμε την αποτελεσματικότητά του, το δοκιμάζουμε σε μια ποικιλία διαφορετικών διατάξεων συστοιχιών φίλτρων, και άρα διαφορετικών χαρακτηριστικών. Για την ακρίβεια, το εισάγουμε στη ροή εργασίας των MFCCs και των PLPs, των πλέον χρησιμοποιούμενων συνόλων χαρακτηριστικών, όπως έχουν αναλυθεί στα Κεφάλαια 5 και 6, αντίστοιχα, καθώς επίσης και στη ροή εργασίας των Απλών Συντελεστών Αναφάσματος Κανονικοποιημένων ως προς την Ισχύ (Simple Power Normalized Cepstrum Coefficients - SPNCCs)[111, 112, 113]. Τα PNCCs αποτελούν ένα state-of-the-art σύνολο χαρακτηριστικών για εύρωστη Αυτόματη Αναγνώριση Φωνής, ενώ μία σύνοψη της μεθόδου εξαγωγής τους με κάποια σχετικά πειραματικά αποτελέσματα για τα δεδομένα όπου εργαζόμαστε δίνεται στο Παράρτημα I.

Η συστοιχία που χρησιμοποιείται αποτελείται από 40 τριγωνικά φίλτρα κατανεμημένα στην κλίμακα *mel* για την εξαγωγή των MFCCs, από 21 τραπεζοειδή φίλτρα κατανεμημένα στην κλίμακα $Bark$ για την εξαγωγή των PLPs και από 40 gammatone φίλτρα κατανεμημένα στην κλίμακα ERB για την εξαγωγή των SPNCCs. Οι εν λόγω συστοιχίες απεικονίζονται σχηματικά στα Σχήματα 5.2, 6.2 και I.1, αντίστοιχα (με διαφορετικό αριθμό φίλτρων). Εισάγουμε, λοιπόν, στη ροή εργασίας της εξαγωγής των τριών αυτών συνόλων χαρακτηριστικών τη μέθοδο που περιγράψαμε, δοκιμάζοντας όλες τις δυνατές τιμές της παραμέτρου M, από 0 έως Q, όπου Q ο αριθμός φίλτρων της συστοιχίας. Για $M = 0$ χρησιμοποιείται μόνο το φάσμα ισχύος, ενώ για $M = Q$ χρησιμοποιείται μόνο το TPS. Ως συνήθως, εκτελούμε το σύνολο των πειραμάτων τόσο για το μικρόφωνο OA6, όσο και για το μικρόφωνο CT1. Τα σχετικά αποτελέσματα παρουσιάζονται στον Πίνακα 8.1[2]. Για να γίνει πιο απτή η βελτιωτική επίδραση της μεθόδου, δίνεται κάθε φορά και και το σχετικό ποσοστό βελτίωσης της ακρίβειας φωνημάτων (Phone Accuracy - PACC), που ορίζεται, κατ' αντιστοιχία του WACC, ως 100%−PER. Τονίζεται πως, για τα συγκεκριμένα πειράματα, στην περίπτωση των MFCCs, μετά τον DCT διατηρούνται οι 13 πρώτοι συντελεστές, περιλαμβανομένου του μηδενικού, χωρίς αυτός να αντικαθίσταται από την ολική ενέργεια, όπως στο Κεφάλαιο 5.

	CT1			OA6		
	$M = 0$	Βέλτιστο M	$M = Q$	$M = 0$	Βέλτιστο M	$M = Q$
MFCCs	31.56	30.17 ($M = 27$)	31.09	77.67	75.71 ($M = 23$)	77.25
PACC_IMP	–	*2.03%*	*0.69%*	–	*8.78%*	*1.88%*
PLPs	45.64	35.43 ($M = 13$)	43.70	90.03	80.97 ($M = 16$)	87.28
PACC_IMP	–	*18.80%*	*3.57%*	–	*90.87%*	*27.58%*
SPNCCs	30.76	29.46 ($M = 11$)	30.03	78.60	75.75 ($M = 23$)	78.02
PACC_IMP	–	*1.88%*	*1.05%*	–	*13.32%*	*2.71%*

Πίνακας 8.1: PER(%) όταν χρησιμοποιείται το Φάσμα Ισχύος ($M = 0$), το Φάσμα Teager Ισχύος ($M = Q$) και ο "βέλτιστος" συνδυασμός τους (Βέλτιστο M) για την εξαγωγή των MFCCs, των PLPs και των SPNCCs. Στις δύο τελευταίες περιπτώσεις δίνεται, ακόμα, η σχετική βελτίωση του PACC (PACC_IMP) σε σχέση με την πρώτη περίπτωση. Το βέλτιστο M (της σχέσης (8.19)) είναι αυτό που πειραματικά δίνει το μικρότερο PER στο διάστημα $[0, Q]$. Σε κάθε περίπτωση, τα διανύσματα χαρακτηριστικών είναι επαυξημένα με τους Δ και ΔΔ συντελεστές, που λαμβάνονται βάσει των 6 (3+3) γειτονικών πλαισίων.

Παρατηρούμε ότι η προτεινόμενη μέθοδος επιφέρει σε κάθε περίπτωση βελτιωμένα αποτε-

[2]Όπως έχει γίνει αντιληπτό από τη διαδικασία υπολογισμού, το TPS μπορεί να εκτιμηθεί αξιόπιστα για όλα τα δείγματα ενός σήματος πλην του πρώτου και του τελευταίου. Για το λόγο αυτό, όλα τα αποτελέσματα στον Πίνακα 8.1, ακόμα και όσα αφορούν στους υπολογισμούς με το κλασικό Φάσμα Ισχύος, αναφέρονται ακριβώς σε αυτά τα "ενδιάμεσα" δείγματα, δηλαδή στο εκάστοτε σήμα $\tilde{s}$ της σελίδας 127.

λέσματα. Η χρήση του TPS μειώνει σε κάποιο βαθμό το PER, τόσο σε θορυβώδεις (OA6), όσο και σε καθαρές (CT1) συνθήκες. Η μείωση αυτή γίνεται σημαντικά μεγαλύτερη όταν το TPS συνδυάζεται με το Φάσμα Ισχύος (δηλαδή πρακτικά συνδυάζονται ο TEO και ο SEO), κατόπιν της αναζήτησης του βέλτιστου M της σχέσης (8.19). Ιδιαίτερα αξιοσημείωτη είναι η επίδραση της μεθόδου στο πλαίσιο εργασίας των PLPs, όπου παρατηρείται μία απόλυτη μείωση του PER της τάξης του 10%, τόσο σε καθαρές, όσο και σε θορυβώδεις συνθήκες.

8.3 Αποδιαμόρφωση AM-FM Σημάτων

Ο TEO, εκτός από τη χρησιμότητά του για τον υπολογισμό της πραγματικής ενέργειας ενός σήματος και την ευεργετική του επίδραση, όταν χρησιμοποιείται αντί του SEO, κατά την εξαγωγή χαρακτηριστικών για αναγνώριση φωνής, έχει αποτελεσματικά βρει εφαρμογή στην αποδιαμόρφωση AM-FM σημάτων. Μία εκτενής στατιστική ανάλυση που αναδεικνύει τη σχέση του TEO με τα AM-FM σήματα και τις δυνατότητες που προκύπτουν παρουσιάζεται στο [114]. Όπως αποδεικνύεται, ο TEO μπορεί να προσεγγίσει ικανοποιητικά την περιβάλλουσα πλάτους AM σημάτων, τη στιγμιαία συχνότητα FM σημάτων ή και το γινόμενο της περιβάλλουσας πλάτους και της στιγμιαίας συχνότητας, στην περίπτωση σημάτων διαμορφωμένων και ως προς τη συχνότητα και ως προς το πλάτος, τόσο για συνεχή, όσο και για διακριτά σήματα. Από την εν λόγω αποδιαμόρφωση, λοιπόν, και λόγω της δυνατότητας μοντελοποίησης των σημάτων φωνής ως το άθροισμα πεπερασμένου αριθμού ζωνοπερατών AM-FM σημάτων όπως αναλύθηκε στην Ενότητα 8.1, μπορεί να προκύψει μία ποικιλία ακουστικών χαρακτηριστικών που δυνητικά μπορούν να βελτιώσουν την ποιότητα της αναγνώρισης.

Η πλέον κλασική μέθοδος αποδιαμόρφωσης στηρίζεται στο Μετασχηματισμό Hilbert και στο αναλυτικό σήμα που προκύπτει από αυτόν [115]. Συγκεκριμένα, το αναλυτικό σήμα $z(t)$, δοθέντος ενός σήματος $s(t)$, ορίζεται ως

$$z(t) = s(t) + j\hat{s}(t) = r(t)e^{j\theta(t)}\,, \tag{8.20}$$

όπου $\hat{s}(t)$ ο Μετασχηματισμός Hilbert του $s(t)$, ο οποίος ορίζεται ως

$$\hat{s}(t) = s(t) * \frac{1}{\pi t}\,. \tag{8.21}$$

Επειδή η συνάρτηση $\frac{1}{\pi t}$ δεν είναι ολοκληρώσιμη και συνεπώς το αντίστοιχο ολοκλήρωμα που ορίζει τη συνέλιξη δε συγκλίνει, είναι διαισθητικά προτιμότερο να σκεφτόμαστε το $\hat{s}(t)$ ως το σήμα που προκύπτει με ολίσθηση κάθε αρνητικής φασματικής συνιστώσας του $s(t)$ κατά $\pi/2$ και κάθε θετικής φασματικής συνιστώσας κατά $-\pi/2$. Φορμαλιστικά, στο πεδίο της συχνότητας έχουμε

$$\hat{S}(\omega) = \begin{cases} e^{j\frac{\pi}{2}}S(\omega), & \omega < 0 \\ e^{-j\frac{\pi}{2}}S(\omega), & \omega > 0 \end{cases} = \begin{cases} jS(\omega), & \omega < 0 \\ -jS(\omega), & \omega > 0 \end{cases} = -j\mathrm{sgn}(\omega)S(\omega)\,. \tag{8.22}$$

Μία εκτίμηση του στιγμιαίου πλάτους $|a(t)|$ και της στιγμιαίας συχνότητας $\omega(t)$ δίνεται, λοιπόν, άμεσα ως

$$|a(t)| \approx r(t) = \sqrt{s^2(t) + \hat{s}^2(t)}\,, \tag{8.23}$$

$$\omega(t) \approx \dot{\theta}(t) = \frac{d}{dt}\left(\arctan\frac{\hat{s}(t)}{s(t)}\right)\,. \tag{8.24}$$

Ένας εναλλακτικός αλγόριθμος αποδιαμόρφωσης [98] στηρίζεται στον TEO και καλείται Αλγόριθμος Διαχωρισμού Ενέργειας (Energy Separation Algorithm - ESA). Σύμφωνα με τον ESA, το στιγμιαίο πλάτος και η στιγμιαία συχνότητα εκτιμώνται ως

$$|a(t)| \approx \frac{\Psi[s(t)]}{\sqrt{\Psi[\dot{s}(t)]}}\,, \tag{8.25}$$

$$\omega(t) \approx \sqrt{\frac{\Psi[\dot{s}(t)]}{\Psi[s(t)]}}\,. \tag{8.26}$$

Παρόμοια, για τα σήματα διακριτού χρόνου έχει προταθεί ο Διακριτός ESA (Discrete ESA - DESA), ο οποίος μπορεί να πάρει διαφορετικές μορφές αναλόγως της διακριτοποίησης που λαμβάνει χώρα, αλλά αυτός που φαίνεται να είναι ο πιο αξιόπιστος (με τα λιγότερα σφάλματα εκτίμησης) [98] είναι ο εξής:

$$|a[n]| \approx \sqrt{\frac{\Psi[s[n]]}{1-\left(1-\dfrac{\Psi[y[n]]+\Psi[y[n+1]]}{4\Psi[s[n]]}\right)^2}}\,, \tag{8.27}$$

$$\Omega[n] \approx \arccos\left(1-\frac{\Psi[y[n]]+\Psi[y[n+1]]}{4\Psi[s[n]]}\right)\,, \tag{8.28}$$

όπου $y[n] = x[n] - x[n-1]$.

Σύμφωνα με το [115], οι δύο μέθοδοι αποδιαμόρφωσης (με χρήση του αναλυτικού σήματος και με χρήση του ESA) δίνουν παρόμοια αποτελέσματα, με τη δεύτερη, όμως, να είναι αισθητά βελτιωμένη ως προς την υπολογιστική της πολυπλοκότητα.

Για να μπορεί να γίνει ασφαλής χρήση οποιασδήποτε από τις δύο μεθόδους αποδιαμόρφωσης, θα πρέπει το εκάστοτε σήμα να είναι ζωνοπεριορισμένο [115]. Ειδικά όσον αφορά στον ESA, η παρουσία προσθετικού θορύβου αλλοιώνει σε μεγάλο βαθμό το τελικό αποτέλεσμα, κάτι που αντιμετωπίζεται ικανοποιητικά με την εισαγωγή της έννοιας του πολυζωνικού ESA [116], όπου προηγείται η διάσπαση του σήματος σε ζώνες συχνοτήτων μέσω ζωνοπερατών φίλτρων. Στο πεδίο της Αυτόματης Αναγνώρισης Φωνής, ωστόσο, η εν λόγω διάσπαση του σήματος προτού την όποια περαιτέρω επεξεργασία του, είναι συνήθως έτσι κι αλλιώς ενσωματωμένη στη ροή εργασίας της εξαγωγής χαρακτηριστικών, οπότε δε χρειάζεται κάποια ειδική μέριμνα. Μάλιστα, η διαδικασία του ζωνοπερατού φιλτραρίσματος μπορεί να συνδυαστεί αποτελεσματικά με τον TEO [117], οδηγώντας σε έναν αλγόριθμο που αποφεύγει τα σφάλματα προσέγγισης που εισάγει ο DESA.

Για την υλοποίηση της εν λόγω μεθόδου γίνεται χρήση συστοιχίας Gabor φίλτρων, τα οποία επιλέγονται λόγω της βελτιστότητάς τους ως προς τη χρονο-συχνοτική αβεβαιότητα, αλλά και λόγω της αποφυγής μεγάλων πλευρικών λοβών, ως αποτέλεσμα της γκαουσιανής τους μορφής, που θα μπορούσαν να προκαλέσουν ασυνέπειες στην έξοδο του TEO [98]. Η απόκριση συχνότητας $g(t)$ ενός Gabor φίλτρου δίνεται στη γενική περίπτωση ως

$$g(t) = e^{-a^2t^2}\cos(2\pi f_c t + \phi)\,, \tag{8.29}$$

όπου f_c η κεντρική συχνότητα του φίλτρου και ϕ η αρχική φάση, η οποία, χωρίς βλάβη της γενικότητας, μπορεί να θεωρηθεί ίση με 0. Η παράμετρος a καθορίζει το εύρος ζώνης του φίλτρου.

Θεωρώντας προς στιγμήν πως το σήμα $s[n]$ επεκτείνεται στο συνεχή χρόνο δίνοντας το $s(t)$, η εφαρμογή του TEO στο ζωνοπεριορισμένο σήμα $s(t) * g(t)$ δίνει

$$\begin{aligned}\Psi[s(t) * g(t)] &= \left(\frac{d}{dt}[s(t) * g(t)]\right)^2 - (s(t) * g(t))\left(\frac{d^2}{dt^2}[s(t) * g(t)]\right) \\ &= \left(s(t) * \frac{dg(t)}{dt}\right)^2 - (s(t) * g(t))\left(s(t) * \frac{d^2g(t)}{dt^2}\right) ,\end{aligned} \tag{8.30}$$

όπου

$$\frac{dg(t)}{dt} = [-2a^2 t\cos(2\pi f_c t) - 2\pi f_c \sin(2\pi f_c t)]e^{-a^2t^2} , \tag{8.31}$$

$$\frac{d^2g(t)}{dt^2} = [8a^2\pi f_c t\sin(2\pi f_c t) + (4a^4t^2 - 2a^2 - 4\pi^2 f_c^2)\cos(2\pi f_c t)]e^{-a^2t^2} . \tag{8.32}$$

Όμοια,

$$\Psi\left[\frac{d}{dt}[s(t) * g(t)]\right] = \left(s(t) * \frac{d^2g(t)}{dt^2}\right)^2 - \left(s(t) * \frac{dg(t)}{dt}\right)\left(s(t) * \frac{d^3g(t)}{dt^3}\right) , \tag{8.33}$$

όπου

$$\begin{aligned}\frac{d^3g(t)}{dt^3} =&[(12a^4t - 8a^6t^3 + 24a^2\pi^2 f_c^2 t)\cos(2\pi f_c t) + \\ &+ (12a^2\pi f_c - 24a^4\pi f_c t^2 + 8\pi^3 f_c^3)\sin(2\pi f_c t)]e^{-a^2t^2} .\end{aligned} \tag{8.34}$$

Προφανώς, ασχολούμαστε με διακριτά σήματα, οπότε έχουμε διακριτή συνέλιξη του σήματος $s[n]$ με τη δειγματοληπτημένη $g[n] = g(t)|_{t=nT}$, όπου T η περίοδος δειγματοληψίας και αντίστοιχα για την πρώτη, δεύτερη και τρίτη παράγωγο της $g(t)$. Ο Gabor TEO, όπως υπολογίζεται από τις σχέσεις (8.30) και (8.33), ενσωματώνεται στον ESA, όπως εκφράζεται από τις σχέσεις (8.25) και (8.26). Η όλη διαδικασία καλείται Gabor ESA. Οι ιδέες αυτές έχουν επεκταθεί και σε περισσότερες διαστάσεις, με τους συγγραφείς στο [118] να προτείνουν τον 2D Gabor ESA για αποδιαμόρφωση 2-διάστατων σημάτων (εικόνων) με πεδίο εφαρμογής την ανάλυση υφής. Για μείωση της υπολογιστικής πολυπλοκότητας, μάλιστα, προτείνεται η εφαρμογή του 2D Gabor ESA στο πεδίο της συχνότητας, κάνοντας χρήση των απαραίτητων θεωρημάτων του Fourier Μετασχηματισμού.

Στο σημείο αυτό, αξίζει να σημειωθεί η προσέγγιση που ακολουθείται στο [119], όπου προτείνεται ένας εναλλακτικός αλγόριθμος αποδιαμόρφωσης που βρίσκει εφαρμογή σε φασματογραφήματα. Για την ακρίβεια, σε κάθε σημείο (i, j) του φασματογραφήματος ενός σήματος ορίζεται ένα επιμέρους κομμάτι $P_{ij}(f, t)$ διαστάσεων $df \times dt$, το οποίο θεωρείται συνάρτηση δύο μεταβλητών που έχει υποστεί AM-FM διαμόρφωση:

$$P_{ij}(f, t) = A_{ij}(f, t)\cos\left(\phi_{ij}(f, t)\right) . \tag{8.35}$$

Όπως υποστηρίζεται στο [119], οι περισσότερες από αυτές τις επιμέρους συναρτήσεις παρουσιάζουν φασματική δομή παρόμοια με ένα Gabor φίλτρο. Έτσι, η φάση $\phi_{ij}(f, t)$ εκτιμάται ως η φάση του "βέλτιστου" φίλτρου (υπό την έννοια του βέλτιστου ταιριάσματος) από μία οικογένεια 2D Gabor φίλτρων, με μια διαδικασία που καλείται Max-Gabor Ανάλυση. Στη συνέχεια, μέσω τεχνικών παρεμβολής εκτιμάται και η περιβάλλουσα πλάτους A_{ij}. Συνδυάζοντας τις επιμέρους ϕ_{ij} και A_{ij} με τη μέθοδο Overlap-Add λαμβάνονται οι τελικώς ζητούμενες συναρτήσεις A και ϕ που αφορούν ολόκληρο το φασματογράφημα.

Η γενικότερη ιδέα της χρήσης 2D Gabor φίλτρων που εφαρμόζονται πάνω σε μια φασματο-χρονική αναπαράσταση του σήματος, όπως είναι το φασματογράφημα, έχει χρησιμοποιηθεί και παλαιότερα με υποσχόμενα αποτελέσματα για την Αυτόματη Αναγνώριση Φωνής, όπως, για παράδειγμα, στο [120] με τα Τοπικά Φασματο-Χρονικά Χαρακτηριστικά (Localized Spectro-Temporal Features - LSTFs), αλλά και πιο πρόσφατα, όπως στα [121, 122], με την εισαγωγή των Χαρακτηριστικών Gabor Συστοιχιών Φίλτρων (GaBor FilterBank features - GBFBs) και των Διαχωρισμένων GBFBs (Separated GBFBs - SGBFBs). Η ιδέα στηρίζεται σε νευροφυσιολογικές ενδείξεις που υποστηρίζουν ότι παρόμοιες διαδικασίες ακολουθούνται στο ακουστικο σύστημα των θηλαστικών. Μάλιστα, υπάρχουν νευρώνες με Φασματο-Χρονικά Δεκτικά Πεδία (Spectro-Temporal Receptive Fields - STRFs) που είναι ευαίσθητοι σε συγκεκριμένα φασματο-χρονικά μοτίβα του σήματος που εισέρχεται στο αυτί [123, 124].

8.4 Εξαγωγή και Χρήση AM-FM Χαρακτηριστικών

8.4.1 Προηγούμενες Προσπάθειες και Αποτελέσματα

Πολλά σύνολα χαρακτηριστικών έχουν προταθεί κατά καιρούς στη βιβλιογραφία που στηρίζονται στο στιγμιαίο πλάτος και τη στιγμιαία συχνότητα των σημάτων φωνής, όπως προκύπτουν από κάποια μέθοδο αποδιαμόρφωσης, αφού αυτά φαίνεται πως μπορούν να δράσουν ευεργετικά στην αναγνώριση φωνής. Για παράδειγμα, η στατιστική ανάλυση που παρουσιάζεται στο [125] δείχνει πως οι διαμορφώσεις της περιβάλλουσας του πλάτους εξαρτώνται κυρίως από το φύλο, την ταυτότητα του ομιλητή, καθώς και την ταυτότητα του εκάστοτε φωνήματος. Συνεπώς, τα ανάλογα χαρακτηριστικά μπορούν να παίξουν σημαντικό ρόλο κατά την αναγνώριση τόσο της φωνής, όσο και του ομιλητή.

Στο [126] χρησιμοποιείται ως αφετηρία το αναλυτικό σήμα και κατόπιν της διάσπασης του σήματος φωνής σε συχνοτικές ζώνες από μία συστοιχία τραπεζοειδών φίλτρων (που διαφέρει από τη συστοιχία που χρησιμοποιείται για την εξαγωγή των PLPs), εξάγεται, μέσω κατάλληλων βαθυπερατών φίλτρων, η Μέση Λογαριθμική Περιβάλλουσα (Average Log-Envelope - ALE) και η Μέση Στιγμιαία Συχνότητα (Average Instantaneous Frequency - AIF). Τα ALEs/AIFs, παρόλο που δίνουν χειρότερα αποτελέσματα από τα MFCCs σε καθαρές συνθήκες, παρουσιάζουν μεγάλη ευρωστία στα περισσότερα ήδη θορύβου. Το αναλυτικό σήμα χρησιμοποιείται και στο [127], όπου, μέσω της αποδιαμόρφωσης όπως αναλύθηκε στην Ενότητα 8.3, εξάγεται η στιγμιαία συχνότητα. Τα 10 χαρακτηριστικά που προκύπτουν με αυτόν τον τρόπο, με χρήση μιας συστοιχίας 10 ζωνοπερατών φίλτρων κατανεμημένων στην κλίμακα *mel*, συναποτελούν τις Στιγμιαίες Συχνότητες στις Mel-Συχνότητες (Mel-Frequency Instantaneous Frequencies - MFIFs), που χρησιμοποιούνται σε μια απλή εργασία αναγνώρισης φωνηέντων. Ένα ακόμα σύνολο χαρακτηριστικών που βασίζεται στο αναλυτικό σήμα είναι το Fepstrum [128]. Για την εξαγωγή του Fepstrum λαμβάνεται το υποδειγματοληπτημένο λογαριθμικό στιγμιαίο πλάτος σε καθένα από τα γραμμικά κατανεμημένα ζωνοπερατά φίλτρα και οι χαμηλότεροι DCT συντελεστές όλων αυτών συνενώνονται σε ένα μεγάλο διάνυσμα χαρακτηριστικών, η διάσταση του οποίου μειώνεται μέσω KLT.

Στο [129] προτείνεται μία ελαφρώς παραλλαγμένη έκδοση του DESA, μέσω της οποίας εξάγεται μόνο το στιγμιαίο πλάτος του σήματος. Εάν συμβολιστεί ως $a_{i,j}[n]$ το στιγμιαίο πλάτος για το i-οστό παράθυρο του σήματος και το j-οστό φίλτρο της εν χρήση συστοιχίας gammatone φίλτρων, τότε ορίζεται η αντίστοιχη AM ισχύς, την οποία θα συμβολίζουμε εδώ AMP (Amplitude Modulation Power), ως

$$G_i(j) = a_{i,j}^T a_{i,j} \, . \tag{8.36}$$

Τα χαρακτηριστικά $G_i(j)$ περνούν από περαιτέρω στάδια επεξεργασίας, τα οποία ταυτίζονται με τα αντίστοιχα στάδια για τα PNCCs, όπως αυτά αναλύονται στο Παράρτημα I, για να παραχθούν εν τέλει οι Συντελεστές Αναφάσματος Κανονικοιημένης Διαμόρφωσης (Normalized Modulation Cepstral Coefficients - NMCCs). Παρόμοιες ιδέες χρησιμοποιούνται στο [130], όπου προτείνεται η χρήση μεγαλύτερων χρονικών παραθύρων (διάρκειας $52msec$), η αποφυγή της κανονικοποίησης ως προς το 95ο εκατοστημόριο και της PBS (βλ. Παράρτημα I), καθώς και ένα επιπλέον βαθυπερατό φιλτράρισμα των σημάτων $a_{i,j}[n]$ για τη διατήρηση της πληροφορίας μόνο στο εύρος $[5Hz, 350Hz]$. Το προκύπτον σύνολο χαρακτηριστικών καλείται Διαμόρφωση του Πλάτους Φωνής Μέσης Διάρκειας (Modulation of Medium Duration Speech Amplitude - MMeDuSA).

Στα [131, 117] γίνεται χρήση του Gabor ESA και μέσω του εκτιμώμενου στιγμιαίου πλάτους και της εκτιμώμενης στιγμιαίας συχνότητας εξάγονται διαφορετικά σύνολα χαρακτηριστικών, τα οποία συνδυάζονται με τα MFCCs για να αυξήσουν την ευρωστία του συστήματος αναγνώρισης σε προσθετικό θόρυβο. Τα χαρακτηριστικά που προτείνονται είναι το Μέσο Στιγμιαίο Πλάτος (Mean Instantaneous Amplitude - MIA), η Μέση Στιγμιαία Συχνότητα (Mean Instantaneous Frequency - MIF) και το Ποσοστό Διαμόρφωσης Συχνότητας (Frequency Modulation Percentage - FMP), που για το i-οστό πλαίσιο του σήματος και το j-οστό φίλτρο της συστοιχίας, θεωρώντας πως το κάθε πλαίσιο αποτελείται από K δείγματα, ορίζονται ως εξής:

$$MIA_i(j) = \frac{1}{K} \sum_{n=iK+1}^{(i+1)K} |a_j[n]|\,, \tag{8.37}$$

$$MIF_i(j) = \frac{1}{K} \sum_{n=iK+1}^{(i+1)K} f_j[n]\,, \tag{8.38}$$

$$FMP_i(j) = \frac{B_i(j)}{F_i(j)}\,, \tag{8.39}$$

όπου

$$B_i^2(j) = \frac{\sum_{n=iK+1}^{(i+1)K} (\dot{a}_j^2[n] + (f_j[n] - F_i(j))^2 |a_j[n]|^2}{\sum_{n=iK+1}^{(i+1)K} |a_j[n]|^2}\,, \tag{8.40}$$

$$F_i(j) = \frac{\sum_{n=iK+1}^{(i+1)K} f_j[n]|a_j[n]|^2}{\sum_{n=iK+1}^{(i+1)K} |a_j[n]|^2}\,. \tag{8.41}$$

Τα $F_i(j)$ και $B_i^2(j)$ [132], που ουσιαστικά προκύπτουν ως η πρώτη και δεύτερη σταθμισμένη ροπή της στιγμιαίας συχνότητας $f_j[n]$ με βάρος το τετραγωνικό πλάτος $|a_j[n]|^2$, αποτελούν εύρωστες εκτιμήσεις της συχνότητας και του εύρους ζώνης, αντίστοιχα, βραχέος χρόνου. Ο όρος $\dot{a}_j^2[n]$ στον ορισμό του $B_i^2(j)$ περιγράφει τη συνεισφορά στο εύρος ζώνης της διαμόρφωσης πλάτους και σχετίζεται με το ρυθμό απόσβεσης της περιβάλλουσας του πλάτους [132, 133].

Τα εν λόγω χαρακτηριστικά έχουν χρησιμοποιηθεί επιτυχώς και για αναγνώριση συναισθήματος μέσω της φωνής [134]. Εκεί δοκιμάζεται με επιτυχία, εκτός της μέσης τιμής, και η Τυπική Απόκλιση του Στιγμιαίου Πλάτους (Standard Deviation of Instantaneous Amplitude - SDIA), που ορίζεται ως

$$SDIA_i(j) = \sqrt{\frac{\sum_{n=iK+1}^{(i+1)K} (|a_j[n]| - MIA_i(j))^2}{K-1}}. \tag{8.42}$$

Όμοια, μπορεί να οριστεί και η Τυπική Απόκλιση της Στιγμιαίας Συχνότητας (Standard Deviation of Instantaneous Frequency - SDIF) ως

$$SDIF_i(j) = \sqrt{\frac{\sum_{n=iK+1}^{(i+1)K} (f_j[n] - MIF_i(j))^2}{K-1}}. \tag{8.43}$$

Σημειώνεται, βέβαια, πως στο [134] υπολογίζονται εν τέλει στατιστικές μετρικές (μέσος και τυπική απόκλιση) των διαφόρων χαρακτηριστικών που αφορούν όλη την πρόταση, κάτι που δεν έχει νόημα για τους σκοπούς της αναγνώρισης φωνημάτων.

Οι συγγραφείς των [131, 117], επηρεασμένοι από τη μοντελοποίηση της σχέσης (8.8), προτείνουν τη χρήση μιας συστοιχίας 6 Gabor φίλτρων ισοκατανεμημένων στην κλίμακα mel, με επικάλυψη 50%, που καλύπτουν το σύνολο του φάσματος $[0, Fs/2]$, όπου θεωρείται πως το κάθε φίλτρο απομονώνει το σήμα γύρω από μία συχνότητα συντονισμού (formant). Στο [133], ωστόσο, φαίνεται πως η αύξηση του αριθμού των φίλτρων επιδρά βοηθητικά στο τελικό αποτέλεσμα, με τα χαρακτηριστικά που μελετώνται να χρησιμοποιούνται αυτόνομα και να οδηγούν σε καλύτερα αποτελέσματα σε σύγκριση με τα MFCCs. Ωστόσο, από ένα σημείο και έπειτα, αύξηση του πλήθους των φίλτρων οδηγεί σε χειρότερα αποτελέσματα, γεγονός που αποδίδεται στη μείωση του εύρους ζώνης των φίλτρων, ιδίως στις χαμηλές συχνότητες. Για το λόγο αυτό, προτείνεται αύξηση του ποσοστού επικάλυψης μεταξύ διαδοχικών φίλτρων από το 50% σε ένα ποσοστό της τάξης του 70%. Ακόμα, αποδεικνύεται ότι κάποια μέθοδος αποσυσχέτισης των συντελεστών, όπως για παράδειγμα ο DCT, όχι μόνο δεν είναι απαραίτητη στην περίπτωση των AM-FM χαρακτηριστικών, αλλά οδηγεί σε αντίθετα από τα επιθυμητά αποτελέσματα, αυξάνοντας τη συσχέτιση μεταξύ τους.

Σε παρόμοιο πνεύμα, στο [135] προτείνονται τα χαρακτηριστικά Φασματικών Ροπών Επαυξημένα με Αναφασματικούς συντελεστές χαμηλής τάξης (Spectral Moment features Augmented by low order Cepstral coefficients - SMACs). Για σήματα με συχνότητα δειγματοληψίας $16kHz$ τα SMACs λαμβάνονται από μια συστοιχία 16 Gabor φίλτρων με σταθερό εύρος ζώνης ίσο με 236 $mels$ ως οι κανονικοποιημένες κεντρικές φασματικές ροπές πρώτης τάξης, που σχετίζονται στενά με τους συντελεστές $F_i(j)$ και $B_i^2(j)$ των σχέσεων (8.41) και (8.40), αντίστοιχα [133]. Το αξιοσημείωτο της μεθόδου είναι πως το διάνυσμα χαρακτηριστικών επαυξάνεται με τα πρώτα δύο MFCCs (και έπειτα με τους Δ και $\Delta\Delta$ συντελεστές όλου του διανύσματος), γεγονός που πειραματικά φαίνεται να επιδρά ευεργετικά στην τελική αναγνώριση. Μάλιστα, η προσθήκη επιπλέον MFCCs δεν επιφέρει επιπρόσθετες βελτιώσεις, ενώ σε ορισμένες περιπτώσεις αυξάνει τα ποσοστά σφάλματος.

Για να είναι αποδοτική η χρήση AM-FM χαρακτηριστικών, είναι προφανώς πολύ σημαντικό ο αλγόριθμος αποδιαμόρφωσης που χρησιμοποιείται να δίνει εύρωστες εκτιμήσεις του στιγμιαίου πλάτους και της στιγμιαίας συχνότητας. Ωστόσο, πολλές φορές τα αποτελέσματα

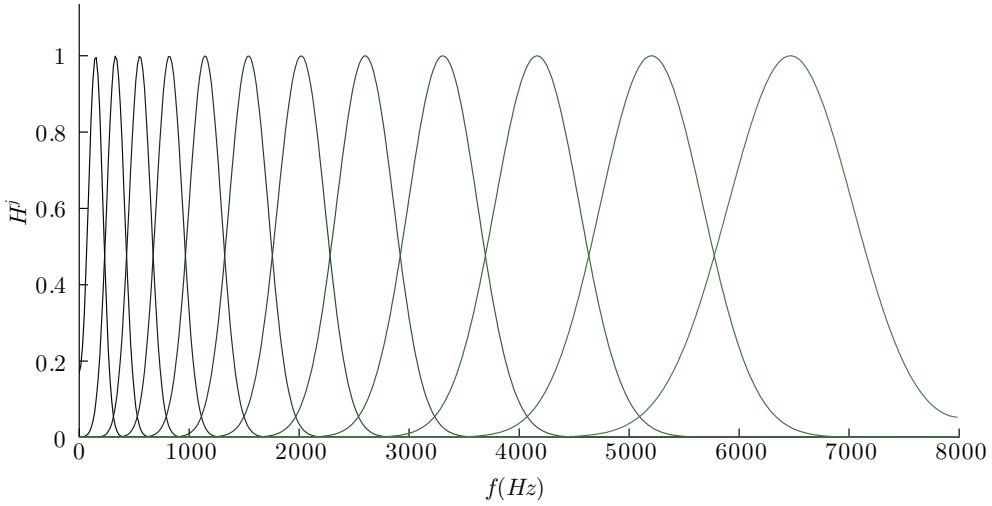
των αλγορίθμων που παρουσιάστηκαν εμφανίζουν κάποια σημεία ανωμαλίας που θα πρέπει να εξαλειφθούν [131]. Για παράδειγμα, όπως έχει ήδη αναφερθεί στην Ενότητα 8.2, το σφάλμα προσέγγισης που υπεισέρχεται από τη χρήση του TEO στο διακριτό χρόνο είναι μη αμελητέο, κυρίως στις υψηλές συχνότητες. Για το λόγο αυτό, στο [131] πριν την εφαρμογή των σχέσεων (8.25) και (8.26), τα σχετικά αποτελέσματα του TEO περνάν από βαθυπερατό φιλτράρισμα μέσω διωνυμικού φίλτρου. Ακόμα, τα αποτελέσματα της αποδιαμόρφωσης του Gabor ESA περνούν από ένα στάδιο εξομάλυνσης μέσω φίλτρου median. Για την αντιμετώπιση των διαφόρων ανωμαλιών κατά τον υπολογισμό της στιγμιαίας συχνότητας, μία εναλακτική μέθοδος είναι η χρήση συστοιχιών φίλτρων διαφορετικών κεντρικών συχνοτήτων, ευρών ζώνης και σχημάτων, τα οποία δίνουν ένα σύνολο αποτελεσμάτων που συνδυάζονται (π.χ. λαμβάνοντας τη μέση τιμή τους) παράγοντας μία πιο εύρωστη τελική εκτίμηση. Η διαδικασία αυτή καλείται Πολυζωνική Ανάλυση Αποδιαμόρφωσης (Multiband Demodulation Analysis - MDA) [136].

8.4.2 Μέθοδος και Πειραματικά Αποτελέσματα στα Πραγματικά Δεδομένα

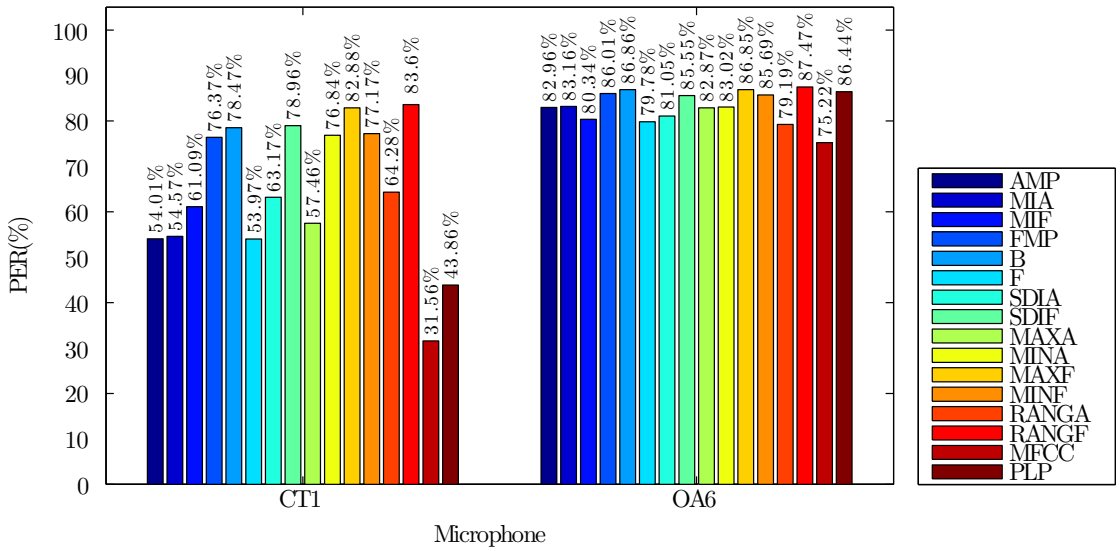
Εφορμόμενοι από τις παραπάνω ιδέες, προχωρούμε σε ορισμένα πειράματα πάνω στα δεδομένα που έχουμε στη διάθεσή μας. Συγκεκριμένα, εξάγονται τα AMP, MIA, MIF, FMP, B, F, SDIA, SDIF, όπως ορίζονται από τις σχέσεις (8.36)-(8.43), καθώς επίσης και οι μέγιστες και ελάχιστες τιμές του στιγμιαίου πλάτους και της στιγμιαίας συχνότητας σε κάθε παράθυρο και για κάθε φίλτρο (MAXA, MINA, MAXF, MINF) και τα αντίστοιχα εύρη RANGA=MAXA−MINA και RANGF=MAXF−MINF. Το στιγμιαίο πλάτος και η στιγμιαία συχνότητα εκτιμώνται βάσει του αλγορίθμου Gabor ESA, τα αποτελέσματα του οποίου εξομαλύνονται μέσω median φίλτρου μεγέθους 7. Ακόμα, τα αποτελέσματα του TEO (πριν τα τελικά στάδια του Gabor ESA) φιλτράρονται με το διωνυμικό φίλτρο τέταρτης τάξης $\frac{1}{16}[1\ 4\ 6\ 4\ 1]$. Χρησιμοποιείται, αρχικά, μία συστοιχία 12 Gabor φίλτρων H^j, $j = 1 \cdots 12$ ισοκατανεμημένων στην κλίμακα *mel* με 50% μεταξύ τους επικάλυψη, όπως φαίνεται στο Σχήμα 8.2, με τα εξαγόμενα χαρακτηριστικά να χρησιμοποιούνται αυτόνομα (χωρίς τη χρήση MFCCs), ενώ το τελικό διάνυσμα χαρακτηριστικών επαυξάνεται από τους Δ και $\Delta\Delta$ συντελεστές που προκύπτουν λαμβάνοντας υπόψιν χρονικό παράθυρο 7 πλαισίων. Η χρήση του συγκεκριμένου χρονικού παράθυρου αποτελεί μια ασφαλή επιλογή, όπως φαίνεται και στο Σχήμα 8.4. Όλα τα χαρακτηριστικά που αφορούν άμεσα το πλάτος (AMP, MIA, SDIA, MAXA, MINA, RANGA) περνούν στο λογαριθμικό πεδίο, ενώ τόσο αυτά όσο και τα υπόλοιπα κανονικοποιούνται ώστε να έχουν μηδενική μέση τιμή και μοναδιαία απόκλιση. Προηγείται μία όμοια κανονικοποίηση του σήματος στο πεδίο του χρόνου, όπως έχει εξηγηθεί στο Κεφάλαιο 5. Τα σχετικά αποτελέσματα παρουσιάζονται στο Σχήμα 8.3.

Αυτό που έχει ενδιαφέρον στο σημείο αυτό να παρατηρήσουμε είναι πως το FMP δε φαίνεται να αποτελεί αξιόπιστο χαρακτηριστικό και αντ' αυτού προτείνεται η χρήση του F, χωρίς, δηλαδή, να κανονικοποιείται από το B (που επίσης δεν αποτελεί αξιόπιστο χαρακτηριστικό). Ακόμα, AM-FM χαρακτηριστικά που δεν έχουν χρησιμοποιηθεί στο παρελθόν για αναγνώριση φωνής, όπως τα AMP, SDIA, MAXA και RANGA, δίνουν υποσχόμενα αποτελέσματα. Βέβαια, τα ποσοστά σφάλματος που έχουν εξαχθεί έως τώρα είναι αρκετά υψηλότερα σε σύγκριση με τα αντίστοιχα ποσοστά με χρήση MFCCs, ιδίως όσον αφορά στο μικρόφωνο CT1, με τις "καθαρές" ηχογραφήσεις. Ωστόσο, αξίζει να σημειωθεί ότι σχεδόν όλα τα AM-FM χαρακτηριστικά δίνουν καλύτερα αποτελέσματα από τα PLPs για τα δεδομένα του μικροφώνου OA6.

Βασει των αποτελεσμάτων του Σχήματος 8.3, εξετάζουμε περαιτέρω, με διαφορετικές



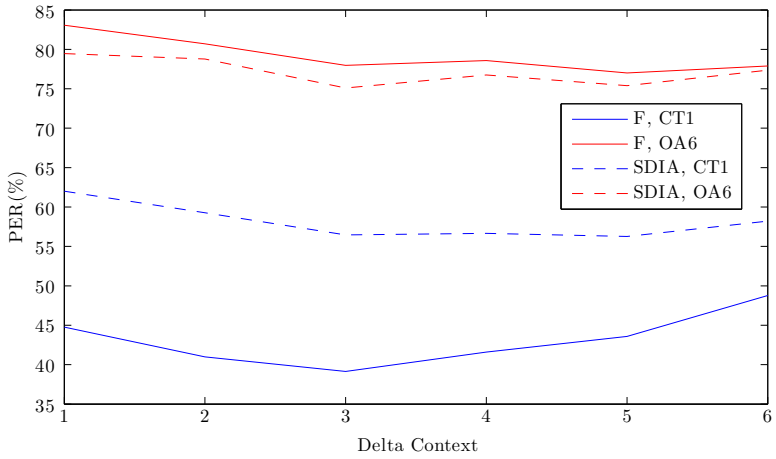
Σχήμα 8.2: Συστοιχία Gabor φίλτρων για την εξαγωγή των AM-FM χαρακτηριστικών. Θεωρείται συχνότητα δειγματοληψίας $16kHz$, ενώ η συστοιχία αποτελείται από 12 φίλτρα με 50% επικάλυψη.



Σχήμα 8.3: PER(%) όταν χρησιμοποιούνται διαφορετικά σύνολα AM-FM χαρακτηριστικών για τα μικρόφωνα CT1 και OA6. Το διάνυσμα χαρακτηριστικών αποτελείται κάθε φορά από 36 στοιχεία, τους 12 συντελεστές όπως εξάγονται από τον Gabor ESA με μια συστοιχία 12 φίλτρων και τους αντίστοιχους Δ και $\Delta\Delta$ συντελεστές. Για σύγκριση, δίνονται και τα αντίστοιχα αποτελέσματα με χρήση MFCCs και PLPs.

παραμετροποιήσεις, τα πλέον υποσχόμενα χαρακτηριστικά, κυρίως όπως αυτά προκύπτουν από τα ποσοστά σφάλματος του μικροφώνου CT1. Συγκεκριμένα, εξετάζουμε πώς μεταβάλλεται το PER καθώς μεταβάλλεται ο αριθμός φίλτρων της συστοιχίας (άρα και το μέγεθος του εκάστοτε διανύσματος χαρακτηριστικών), καθώς και το ποσοστό επικάλυψης μεταξύ των διαδοχικών φίλτρων. Χρησιμοποιούμε, λοιπόν, συστοιχίες 6, 12 και 18 φίλτρων με επικάλυψη 50% και 70%, με τα σχετικά αποτελέσματα να παρουσιάζονται στο Σχήμα 8.5.

Παρατηρούμε πως στις περισσότερες περιπτώσεις, όταν ο αριθμός των φίλτρων διατηρείται σταθερός, αύξηση του ποσοστού επικάλυψης οδηγεί σε μείωση του σφάλματος αναγνώρισης,



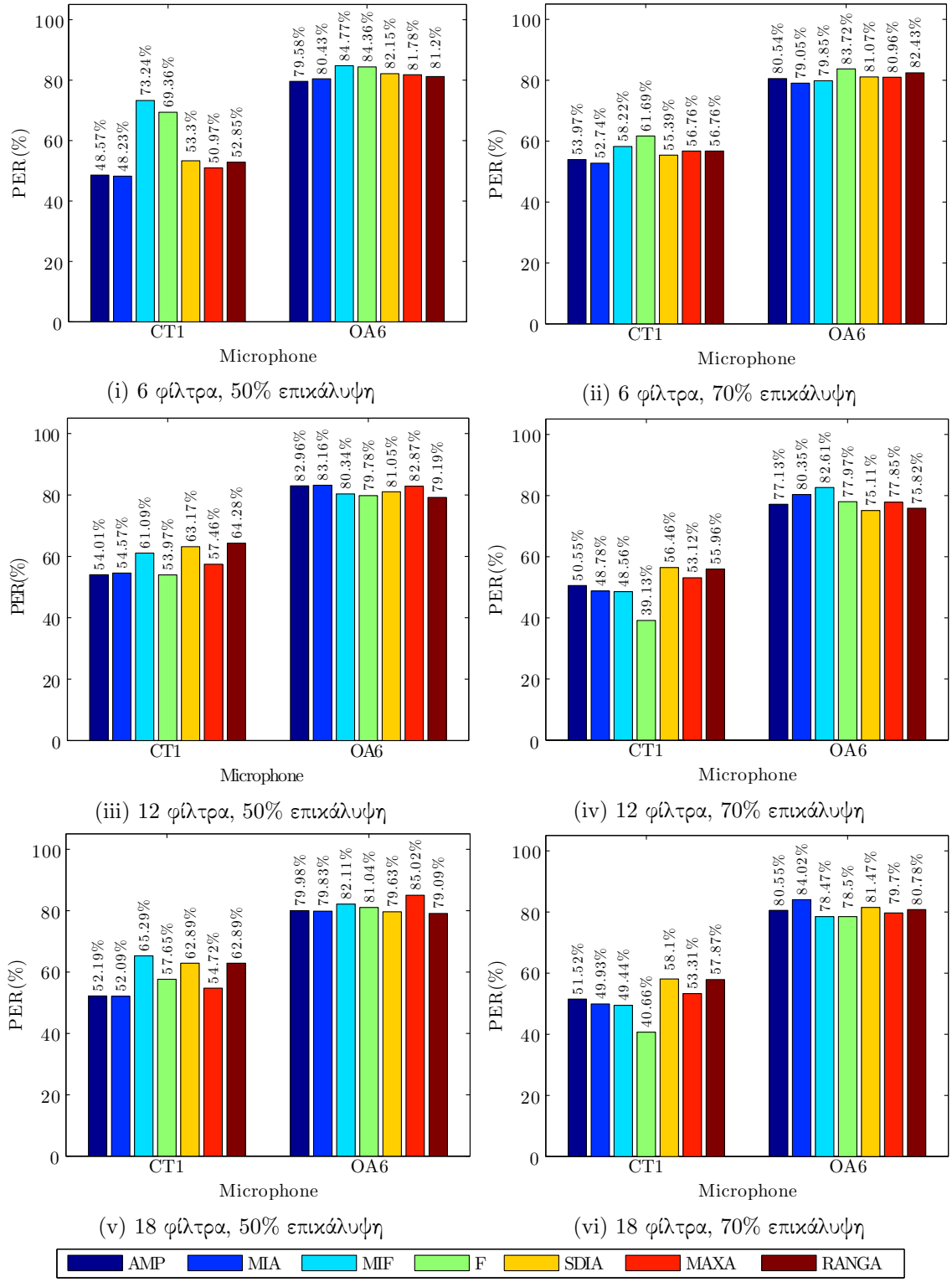
Σχήμα 8.4: PER(%) καθώς μεταβάλλεται το χρονικό παράθυρο που λαμβάνεται υπόψιν κατά τον υπολογισμό των Δ και ΔΔ συντελεστών όταν χρησιμοποιούνται AM-FM χαρακτηριστικά.

με τις εξαιρέσεις να αφορούν κυρίως το μικρόφωνο CT1 όταν γίνεται χρήση συστοιχίας 6 φίλτρων. Η αύξηση του αριθμού των φίλτρων, από την άλλη, δεν έχει πάντα ευεργετικά αποτελέσματα. Μάλιστα, με το ποσοστό επικάλυψης σταθερό στο 50%, τα αποτελέσματα είναι για τα περισσότερα χαρακτηριστικά, τόσο για το μικρόφωνο CT1, όσο και για το OA6, συγκρίσιμα ή και καλύτερα όταν γίνεται χρήση 6 φίλτρων παρά 18. Κάτι τέτοιο είναι πολύ σημαντικό λόγω του υπολογιστικού πλεονεκτήματος της χρήσης μιας μικρής συστοιχίας φίλτρων, εφόσον είναι αντίστοιχα μικρό το μέγεθος του προκύπτοντος διανύσματος χαρακτηριστικών.

Αξίζει να σημειωθεί πως έγινε προσπάθεια διατήρησης λίγων μόνο (13) συντελεστών μέσω χρήσης DCT, τόσο όταν χρησιμοποιήθηκε συστοιχία 18 φίλτρων, όσο και με μεγαλύτερες συστοιχίες. Ωστόσο, τα ποσοστά της αναγνώρισης προέκυπταν κατ' αυτόν τον τρόπο αρκετά χειρότερα, επιβεβαιώνοντας, έτσι, τα αποτελέσματα του [133].

Γενικά, τα πιο χαμηλά σφάλματα αναγνώρισης προκύπτουν στη μέση περίπτωση για συστοιχία 12 φίλτρων με το ποσοστό επικάλυψης μεταξύ διαδοχικών φίλτρων ίσο με 70%. Με τη συγκεκριμένη παραμετροποίηση, το πλέον υποσχόμενο AM-FM χαρακτηριστικό, από άποψη ευρωστίας, φαίνεται να είναι το F. Τα σφάλματα μειώνονται αισθητά, όπως φαίνεται στον Πίνακα 8.2 για το F, εάν το διάνυσμα των εν λόγω χαρακτηριστικών προσαυξηθεί με το λογάριθμο την ενέργειας, που ταυτίζεται με το πρώτο στοιχείο του διανύσματος των MFCCs. Παρατηρείται πως η απόδοση του συστήματος σε καθαρές συνθήκες είναι συγκρίσιμη με την απόδοση που προκύπτει από τη χρήση MFCCs, ενώ βελτιώνεται στις συνθήκες του μικροφώνου OA6.

Εάν δεχτούμε το επιπλέον υπολογιστικό κόστος που εισάγεται από την αύξηση του διανύσματος χαρακτηριστικών (αλλά και από την εκτίμηση των MFCCs), το σφάλμα μπορεί να μειωθεί περαιτέρω και στις δύο συνθήκες από το συνδυασμό των συντελεστών F με τα MFCCs ή με έναν περιορισμένο αριθμό από τους πρώτους μόνο συντελεστές. Τα αποτελέσματα με χρήση του υβριδικού αυτού, επαυξημένου διανύσματος χαρακτηριστικών παρουσιάζονται στο Σχήμα 8.6. Τα πειράματα διεξάγονται τόσο όταν χρησιμοποιείται CMVN για την εξαγωγή των MFCCs, όσο και όταν δε χρησιμοποιείται. Με χρήση και των 13 MFCCs, δηλαδή με ένα διάνυσμα συνολικά $(12 + 13) \cdot 3 = 75$ χαρακτηριστικών επιτυγχάνεται PER 28.16% και 70.50% για τα μικρόφωνα CT1 και OA6, αντίστοιχα, στην περίπτωση που δεν εφαρμόζεται CMVN. Όταν εφαρμόζεται CMVN, τα αντίστοιχα ποσοστά σφάλματος πέφτουν



(i) 6 φίλτρα, 50% επικάλυψη

(ii) 6 φίλτρα, 70% επικάλυψη

(iii) 12 φίλτρα, 50% επικάλυψη

(iv) 12 φίλτρα, 70% επικάλυψη

(v) 18 φίλτρα, 50% επικάλυψη

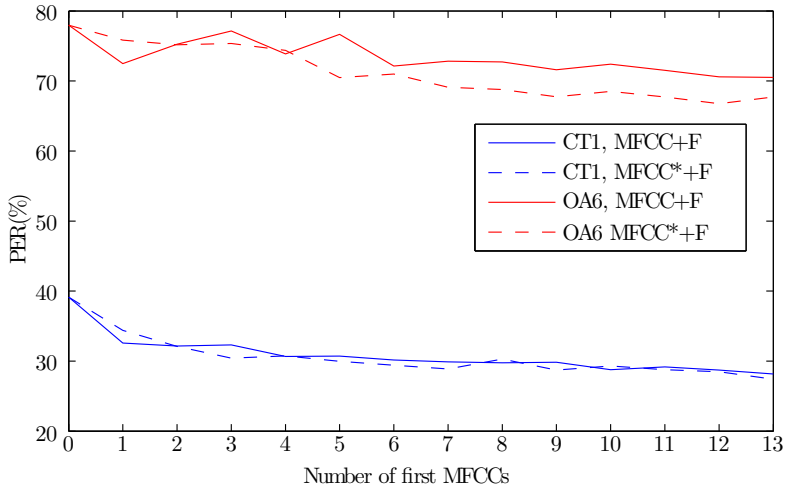
(vi) 18 φίλτρα, 70% επικάλυψη

Σχήμα 8.5: PER(%) όταν χρησιμοποιούνται διαφορετικά σύνολα AM-FM χαρακτηριστικών για τα μικρόφωνα CT1 και OA6, με διαφορετικές τιμές του πλήθους Gabor φίλτρων της συστοιχίας και της μεταξύ τους επικάλυψης.

σε 27.41% και 67.72%.

	F (36)	F+logE (39)	MFCC (39)
CT1	39.13	32.59	**31.56**
OA6	77.97	**72.49**	75.22

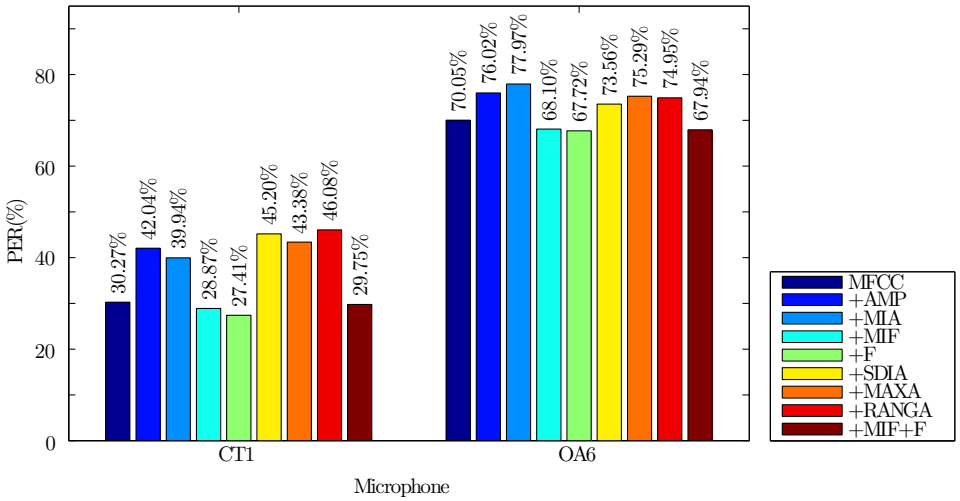
Πίνακας 8.2: PER(%) όταν χρησιμοποιείται το διάνυσμα χαρακτηριστικών F είτε μόνο του (1η στήλη), είτε προσαυξημένο με το λογάριθμο της τετραγωνικής ενέργειας (2η στήλη). Για σύγκριση δίνεται το PER και με χρήση των MFCCs (3η στήλη). Όλα τα διανύσματα χαρακτηριστικών προσαυξάνονται με τους αντίστοιχους Δ και ΔΔ συντελεστές. Για την εξαγωγή των F χρησιμοποιείται συστοιχία φίλτρων με παραμετροποίηση όπως στο Σχήμα 8.5iv. Σε παρενθέσεις αναγράφεται το μήκος του εκάστοτε διανύσματος χαρακτηριστικών.



Σχήμα 8.6: PER(%) όταν χρησιμοποιείται το διάνυσμα χαρακτηριστικών F, προσαυξημένο με τα πρώτα MFCCs. Με αστερίσκο (*) σημειώνονται τα MFCCs όταν έχει λάβει χώρα CMVN, ενώ και τα AM-FM χαρακτηριστικά είναι σε κάθε περίπτωση κανονικοποιημένα ως προς τη μέση τιμή και την τυπική απόκλιση. Το τελικό διάνυσμα χαρακτηριστικών προσαυξάνεται σε κάθε περίπτωση με τους Δ και ΔΔ συντελεστές. Ως πρώτο MFCC έχει χρησιμοποιηθεί ο λογάριθμος της τετραγωνικής ενέργειας του σήματος. Για την εξαγωγή των F χρησιμοποιείται συστοιχία φίλτρων με παραμετροποίηση όπως στο Σχήμα 8.5iv.

Όμοια με το F, δοκιμάζουμε την απόδοση του συστήματος όταν χρησιμοποιούνται, σε συνδυασμό με τα MFCCs, όλα τα σύνολα χαρακτηριστικών που εμφανίζονται στο Σχήμα 8.5, με τα σχετικά αποτελέσματα να παρουσιάζονται στο Σχήμα 8.7. Όπως παρατηρείται, τα αποτελέσματα βελτιώνονται σε σύγκριση με τα απλά MFCCs μόνο στην περίπτωση των MIF και F, δηλαδή αυτών των AM-FM χαρακτηριστικών που σχετίζονται με τη στιγμιαία συχνότητα, γεγονός που καταδεικνύει πως τα εν λόγω χαρακτηριστικά εμπεριέχουν συμπληρωματική ως προς τα MFCCs πληροφορία. Αξιοσημείωτη είναι η σύγκριση με το Σχήμα 8.5iv, όπου φαίνεται πως τα MIFs από μόνα τους δε δίνουν καλά αποτελέσματα αναγνώρισης, γεγονός που αντιστρέφεται στην περίπτωση του συνδυασμού τους με τα MFCCs. Σημειώνεται πως ενώ ο συνδυασμός των F ή των MIF με τα MFCCs αυτοτελώς δίνει βελτιωμένα αποτελέσματα, ο συνδυασμός και των τριών συνόλων χαρακτηριστικών σε ένα ενιαίο διάνυσμα χαρακτηριστικών δε μειώνει ακόμα περισσότερο το σφάλμα (Σχήμα 8.7). Παρόμοια συμπεράσματα λήφθηκαν και από το συνδυασμό άλλων τύπων AM-FM χαρακτηριστικών μεταξύ

τους.

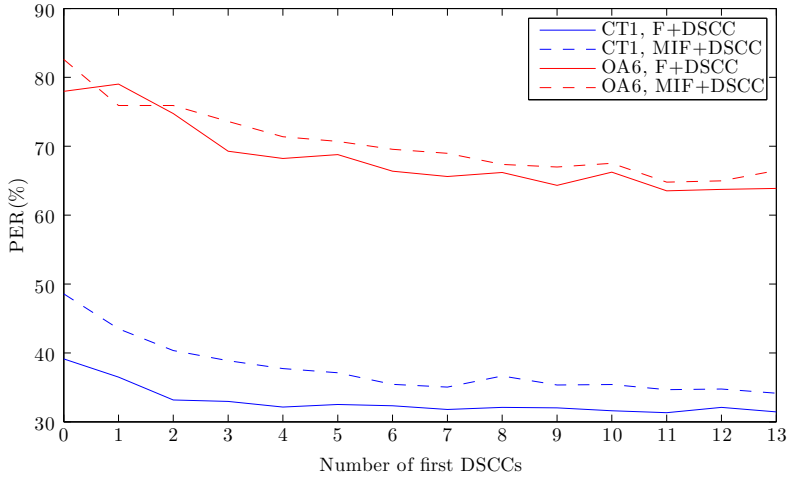


Σχήμα 8.7: PER(%) όταν χρησιμοποιούνται διαφορετικά σύνολα AM-FM χαρακτηριστικών για τα μικρόφωνα CT1 και OA6, σε συνδυασμό με τα 13 MFCCs. Το τελικό διάνυσμα χαρακτηριστικών προσαυξάνεται κάθε φορά με τους Δ και $\Delta\Delta$ συντελεστές και αποτελείται, έτσι, από $(12 + 13) \cdot 3 = 75$ στοιχεία, εκτός από την περίπτωση των σκέτων MFCCs (39 στοιχεία) και το συνδυασμό MFCC+MIF+F (111 στοιχεία). Για όλα τα χαρακτηριστικά έχει λάβει χώρα MVN. Για την εξαγωγή των AM-FM χαρακτηριστικών χρησιμοποιείται συστοιχία φίλτρων με παραμετροποίηση όπως στο Σχήμα 8.5iv.

Ακόμα καλύτερα αποτελέσματα μπορούν να επιτευχθούν όταν τα AM-FM χαρακτηριστικά συνδυάζονται, αντί των MFCCs, με τα DSCCs που παρουσιάστηκαν στην Ενότητα 5.5. Υπενθυμίζεται πως τα DSCCs προκύπτουν ως τα δυναμικά χαρακτηριστικά των MFCCs, αλλά με την παραγώγιση να λαμβάνει χώρα στο πεδίο του φάσματος και όχι του αναφάσματος. Μάλιστα, όπως είχε φανεί από τα πειράματα της σχετικής Ενότητας, η αυτοτελής χρήση των DSCCs, μαζί με τους Δ και $\Delta\Delta$ συντελεστές που υπολογίζονται επί των DSCCs, δίνει καλύτερα αποτελέσματα για τις συνθήκες του μικροφώνου OA6, σε σύγκριση με το συνδυασμό των DSCCs με τα MFCCs. Σημειώνεται, ακόμα, πως τα DSCCs, όπως και τα Δ και $\Delta\Delta$ χαρακτηριστικά επί των DSCCs, υπολογίζονται λαμβάνοντας υπόψιν χρονικό παράθυρο 5+5=10 γειτονικών πλαισίων (αντί των 3+3=6 που χρησιμοποιούμε συνήθως στην εργασία). Τα σχετικά αποτελέσματα, λοιπόν, όταν τα F και τα MIF συνδυάζονται με τα DSCCs, παρουσιάζονται στο Σχήμα 8.8. Όπως φαίνεται, η προσθήκη των DSCCs οδηγεί σε παρόμοια πτωτική συμπεριφορά του PER όπως και στο Σχήμα 8.6. Ο συνδυασμός των 12 F και των 13 DSCCs οδηγεί σε PER ίσο με 31.43% για το CT1 και 63.87% για το OA6, τη στιγμή που τα αντίστοιχα ποσοστά με χρήση μόνο των DSCCs είναι (Σχήμα 5.9) 47.25% και 66.1%. Έτσι, όχι μόνο μειώνεται το σφάλμα στην περίπτωση συνθηκών αναγνώρισης από απόσταση, αλλά βελτιώνεται σημαντικά το σύστημα και από άποψη ευρωστίας.

8.4.3 Πειραματικά Αποτελέσματα σε Συνθετικά Δεδομένα

Για μια πιο σχολαστική μελέτη της επίδρασης των υπό ανάλυση ακουστικών χαρακτηριστικών σε συνθήκες αναγνώρισης από απόσταση, εξετάζουμε την απόδοση του αναγνωριστή σε συνθετικά σήματα, όπου έχουμε πλήρη γνώση των αιτιών και του μεγέθους παραμόρφωσης των διαφόρων ηχογραφήσεων. Συγκεκριμένα, μελετάμε τα MFCCs, τα DSCCs, τα



Σχήμα 8.8: PER(%) όταν χρησιμοποιούνται τα διανύσματα χαρακτηριστικών F και MIF, προσαυξημένα με τα πρώτα DSCCs. Όλα τα χαρακτηριστικά είναι κανονικοποιημένα ως προς τη μέση τιμή και την τυπική απόκλιση. Το τελικό διάνυσμα χαρακτηριστικών προσαυξάνεται σε κάθε περίπτωση με τους Δ και $\Delta\Delta$ συντελεστές, οι οποίοι για τα DSCCs προκύπτουν από ένα χρονικό παράθυρο 11 πλαισίων, ενώ για τα F και MIF από παράθυρο 7 πλαισίων. Για την εξαγωγή των F και MIF χρησιμοποιείται συστοιχία φίλτρων με παραμετροποίηση όπως στο Σχήμα 8.5iv.

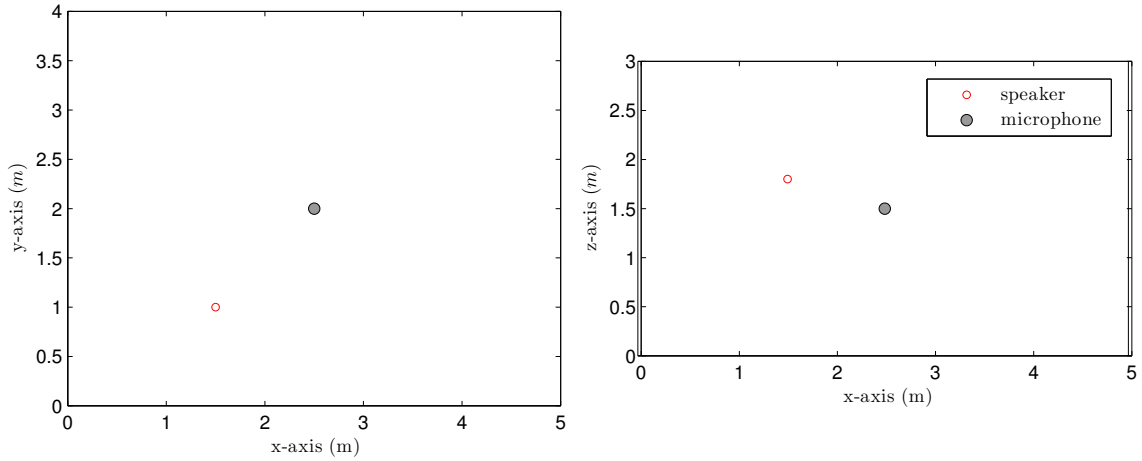
χαρακτηριστικά διαμόρφωσης F και MIF, καθώς και τους συνδυασμούς τους F+MFCCs, MIF+MFCCs, F+DSCCs, MIF+DSCCs. Όλα τα χαρακτηριστικά είναι κανονικοποιημένα ως προς μέση τιμή και τυπική απόκλιση, τα F και MIF εξάγονται από μια συστοιχία φίλτρων με παραμετροποίηση όπως στο Σχήμα 8.5iv, ενώ το τελικό διάνυσμα χαρακτηριστικών προσαυξάνεται με τους Δ και $\Delta\Delta$ συντελεστές, όπως έχει εξηγηθεί στα πειράματα της προηγούμενης Υποενότητας.

Η εκπαίδευση του συστήματος γίνεται με τα καθαρά δεδομένα της βάσης Logotypografia, όπως και σε όλα τα μέχρι τώρα πειράματα. Για τον έλεγχο, όμως, χρησιμοποιούνται συνθετικά δεδομένα που προέρχονται από τις ηχογραφήσεις της βάσης ATHENA με το μικρόφωνο CT1.

Η κυρίαρχη και αναπόφευκτη αιτία παραμόρφωσης στο πεδίο της Αναγνώρισης Φωνής από Απόσταση είναι η αντήχηση. Πρώτα απ' όλα, λοιπόν, εξετάζουμε την απόδοση του συστήματος αναγνώρισης σε ελεγχόμενες συνθήκες αντήχησης. Για το σκοπό αυτό, χρησιμοποιούμε το εργαλείο ανοιχτού κώδικα[3] που βασίζεται στη μέθοδο που περιγράφεται στο [137] για την υλοποίηση της Μεθόδου των Φανταστικών Πηγών (Image-Source Method - ISM) [138] για την προσομοίωση της ακουστικής μικρών δωματίων. Μία σύνοψη κάποιων γενικών εννοιών που σχετίζονται με την αντήχηση, αλλά και της ISM δίνεται στο Παράρτημα II. Θεωρούμε, λοιπόν, ένα ορθογωνικό δωμάτιο διαστάσεων $5m \times 4m \times 3m$, στο κέντρο του οποίου βρίσκεται το υποθετικό μικρόφωνο ηχογράφησης. Η ανακλαστικότητα των 6 επιφανειών του δωματίου θεωρείται ίδια. Ο ομιλητής θεωρείται αρχικά ακίνητος και σε απόσταση $1.45m$ από το μικρόφωνο, όπως φαίνεται στο Σχήμα 8.9. Τα σχετικά αποτελέσματα για διαφορετικούς χρόνους αντήχησης T_{60} παρουσιάζονται στο Σχήμα 8.10. Σημειώνεται ότι για $T_{60} = 0sec$ προσομοιώνεται ανηχοϊκός θάλαμος.

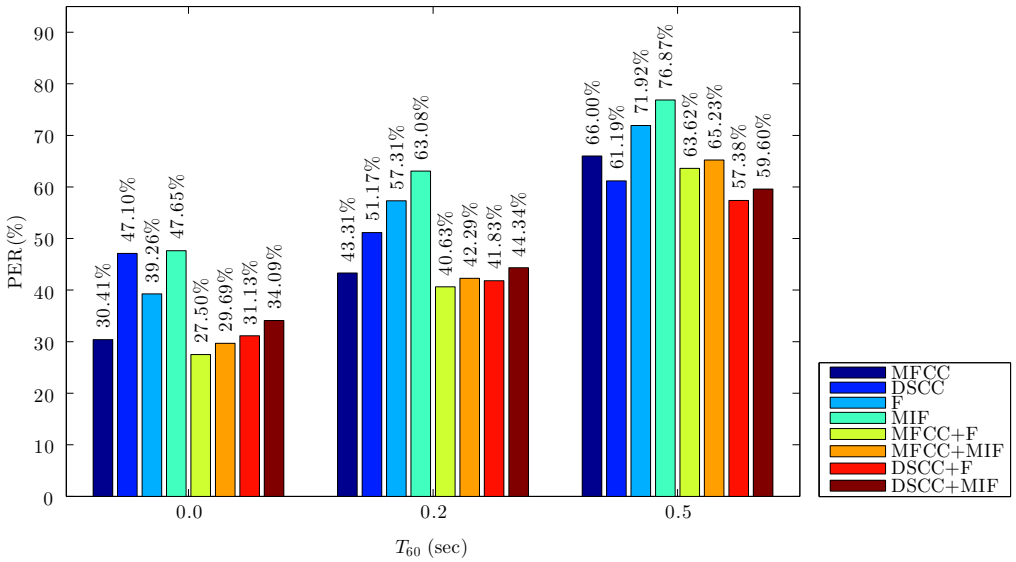
Παρατηρούμε πως σε κάθε περίπτωση η χρήση των F είναι προτιμότερη σε σύγκριση με τα MIF, ενώ η χρήση τους έχει νόημα για την παροχή συμπληρωματικής πληροφορίας προς τα MFCCs ή DSCCs και όχι αυτοτελώς. Και στις τρεις συνθήκες αντήχησης που εξετάζονται

[3] http://www.eric-lehmann.com/ism_code.html



(i) $x - y$ επίπεδο (ii) $x - z$ επίπεδο

Σχήμα 8.9: Θέση μικροφώνου και ακίνητου ομιλητή για την προσομοίωση της αντήχησης σε ορθογωνικό δωμάτιο. Οι συντεταγμένες του μικροφώνου είναι $(2.5, 2, 1.5)$ και του ομιλητή $(1.5, 1, 1.8)$.

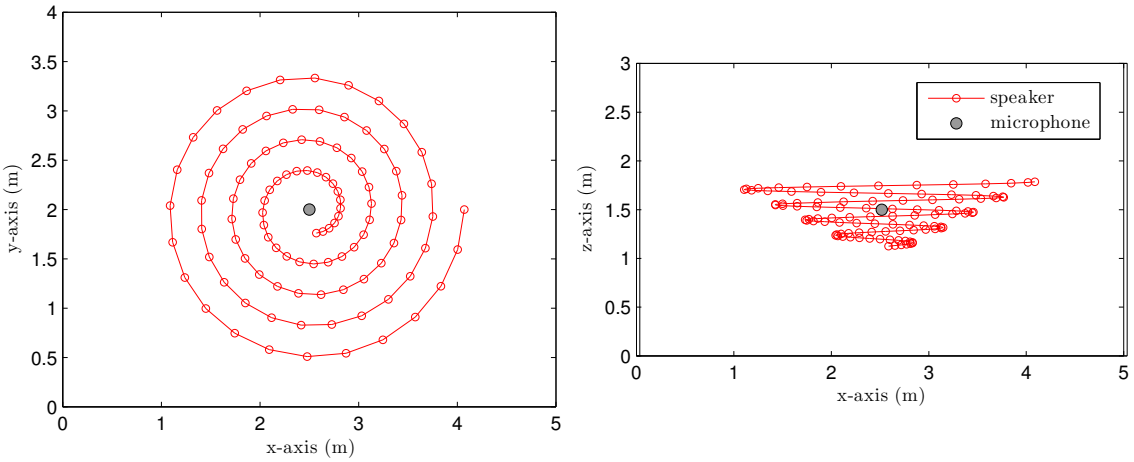


Σχήμα 8.10: PER(%) όταν χρησιμοποιούνται τα διανύσματα χαρακτηριστικών MFCC, DSCC, F, MIF και συνδυασμοί τους. Για τον έλεγχο γίνεται χρήση συνθετικών δεδομένων παραμορφωμένων με συνελικτικό θόρυβο λόγω αντήχησης σε δωμάτιο με ακίνητο ομιλητή. Όλα τα χαρακτηριστικά είναι κανονικοποιημένα ως προς τη μέση τιμή και την τυπική απόκλιση. Το τελικό διάνυσμα χαρακτηριστικών προσαυξάνεται σε κάθε περίπτωση με τους Δ και $\Delta\Delta$ συντελεστές.

υπάρχει σημαντική μείωση των ποσοστών σφάλματος που επιφέρουν τόσο τα MFCCs, όσο και τα DSCCs. Ως προς τη σύγκριση μεταξύ MFCCs και DSCCs, σε καθαρές συνθήκες υπερισχύουν σαφώς τα MFCCs, με την κατάσταση να αντιστρέφεται όσο η συνελικτική παραμόρφωση αυξάνεται. Κατ' επέκταση, τα καλύτερα αποτελέσματα για συνθήκες ανηχοϊκού θαλάμου ($T_{60} = 0sec$) λαμβάνονται με χρήση του συνδυασμού MFCC+F, ενώ για έντονη αντήχηση ($T_{60} = 0.5sec$) με το συνδυαμό DSCC+F.

Εν συνεχεία, εξετάζεται ένα πιο δύσκολο σενάριο, όπου ο ομιλητής (ή γενικά η ακου-

στική πηγή) κινείται στο χώρο ηχογράφησης, εφόσον οι τυχαίες κινήσεις των ομιλητών στο χώρο είναι ένας παράγοντας που θα πρέπει να λαμβάνεται υπόψιν σε εφαρμογές DSR. Συγκεκριμένα, θεωρούμε πως ο ομιλητής διαγράφει μια σπιροειδή τροχιά, όπως απεικονίζεται στο Σχήμα 8.11, η οποία προσομοιώνεται σε κάθε εκφορά, ενώ τα χαρακτηριστικά του δωματίου παραμένουν ίδια. Τα σχετικά αποτελέσματα, όπως παρουσιάζονται στο Σχήμα 8.12, επιβεβαιώνουν τα αποτελέσματα του Σχήματος 8.10.

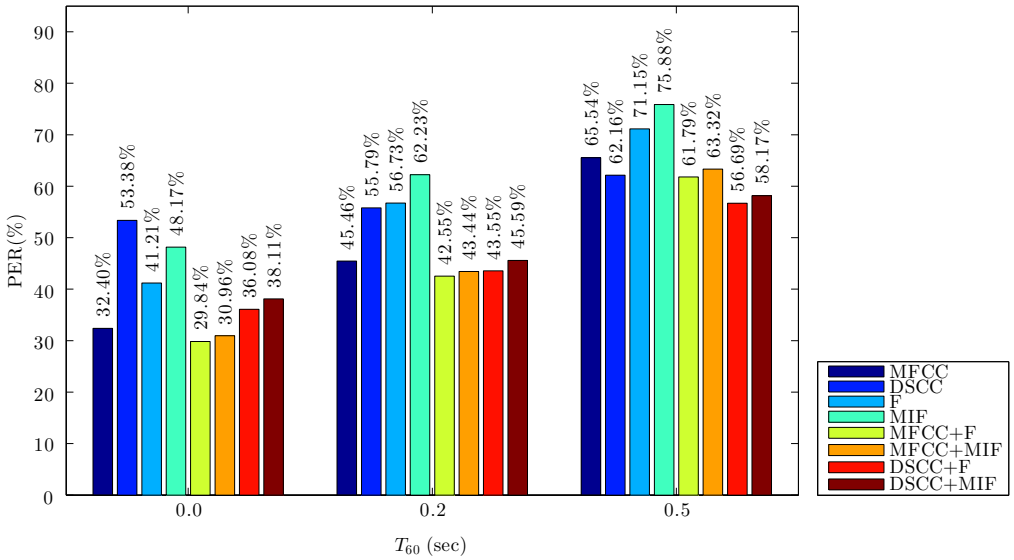


(i) $x-y$ επίπεδο (ii) $x-z$ επίπεδο

Σχήμα 8.11: Θέση μικροφώνου και κινούμενου ομιλητή για την προσομοίωση της αντήχησης σε ορθογωνικό δωμάτιο. Οι συντεταγμένες του μικροφώνου είναι $(2.5, 2, 1.5)$, ενώ ο ομιλητής ακολουθεί σπιροειδή τροχιά που σε παραμετρική μορφή δίνεται από τις εξισώσεις $x = \frac{t}{20}\cos(t)+2.5$, $y = \frac{t}{20}\sin(t)+2$ και $z = \frac{t}{40}+1$, με $t \in [5, 10\pi]$.

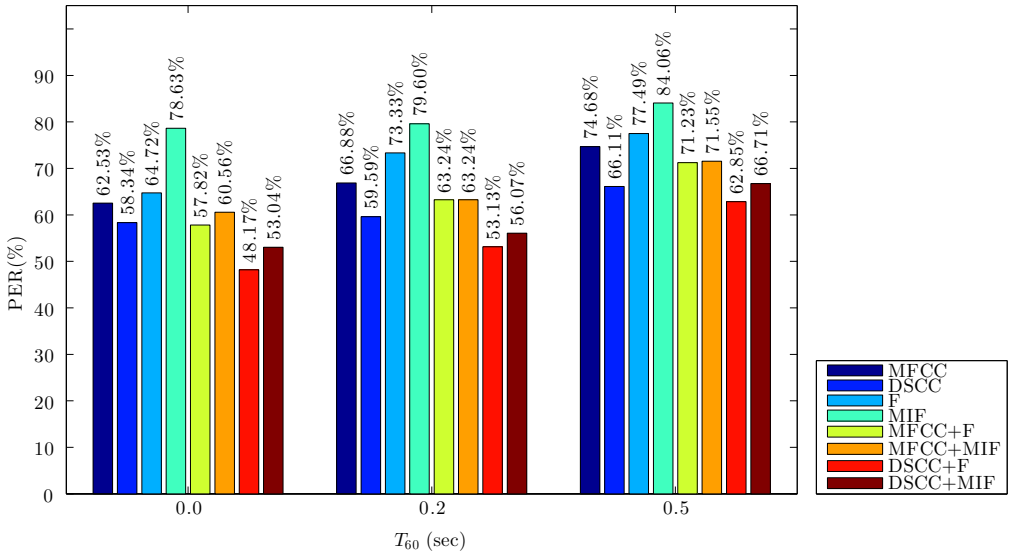
Τέλος, εκτός από το συνελικτικό θόρυβο, προσθέτουμε στα δεδομένα ελέγχου και λευκό γκαουσιανό θόρυβο. Για το σκοπό αυτό, χρησιμοποιείται και πάλι η διάταξη του Σχήματος 8.9, ενώ μελετώνται δύο διακριτές συνθήκες θορύβου, όταν ο θόρυβος είναι σχετικά ήπιος ($SNR = 15dB$) και όταν είναι πολύ έντονος ($SNR = 5dB$). Τα σχετικά αποτελέσματα παρουσιάζονται στο Σχήμα 8.13. Όπως και στην περίπτωση του σκέτου συνελικτικού θορύβου, η χρήση των F είναι προτιμότερη έναντι των MIF και προτείνεται σε συνδυασμό με τα MFCCs ή τα DSCCs. Μάλιστα, όταν τα F συνδυάζονται με τα MFCCs, προκύπτει πάντα μείωση των ποσοστών σφάλματος σε σύγκριση με χρήση μόνο των τελευταίων. Ωστόσο, σε συνθήκες έντονου προσθετικού θορύβου, και ανεξαρτήτως του μεγέθους συνελικτικής παραμόρφωσης, φαίνεται πως τα DSCCs με τον τρόπο που χρησιμοποιούνται δίνουν καλύτερα αποτελέσματα αυτοτελώς. Έτσι, ενώ στην περίπτωση όπου $SNR = 15dB$, με ή χωρίς αντήχηση, προτείνεται ο συνδυασμός DSCC+F, στην περίπτωση όπου $SNR = 5dB$ προτείνεται η χρήση μόνο των δυναμικών χαρακτηριστικών που αποτελούν τα DSCCs. Σημειώνεται, βέβαια, πως ούτως ή άλλως, σε τόσο έντονο θόρυβο, τα ποσοστά σφάλματος ανεβαίνουν σε πολύ υψηλά επίπεδα.

Καταληκτικά, προτείνουμε το συνδυασμό των DSCCs με τη σταθμισμένη εκτίμηση F της στιγμιαίας συχνότητας σε ένα κοινό διάνυσμα χαρακτηριστικών για περιπτώσεις συνελικτικού θορύβου, ο οποίος είναι σχεδόν αναπόφευκτος σε Αναγνώριση Φωνής από Απόσταση. Όταν ενυπάρχει και προσθετικός θόρυβος, η επιλογή εξαρτάται από το πόσο έντονη είναι (ή αναμένεται να είναι) η παραμόρφωση που αυτός εισάγει. Για σχετικά ήπια παραμόρφωση, που αποτελεί και την πιο ρεαλιστική κατάσταση όπου αναμένεται ένας αναγνωριστής να λειτουργήσει, προτείνεται και πάλι η συνδυασμένη χρήση DSCC και F. Με πολύ έντονο προσθετικό

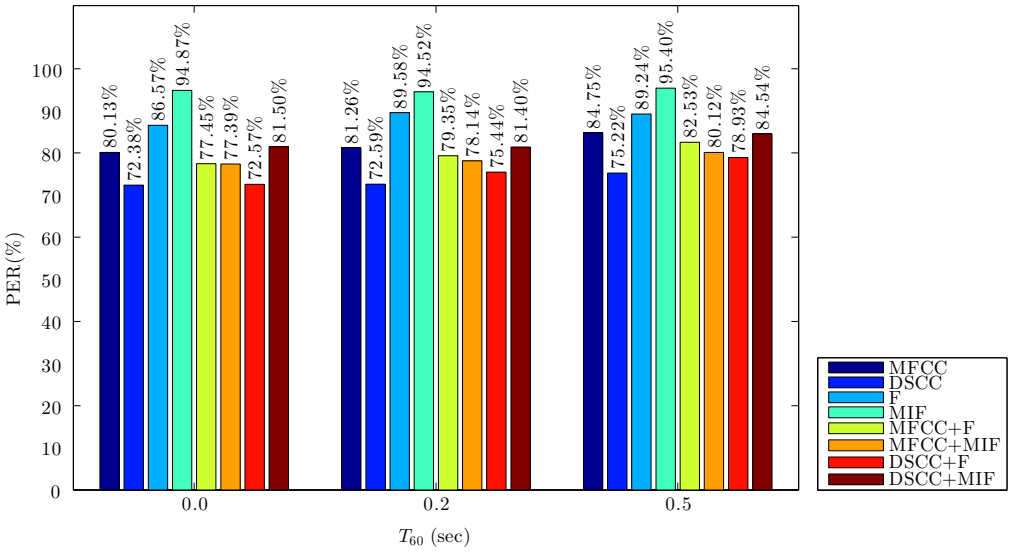


Σχήμα 8.12: PER(%) όταν χρησιμοποιούνται τα διανύσματα χαρακτηριστικών MFCC, DSCC, F, MIF και συνδυασμοί τους. Για τον έλεγχο γίνεται χρήση συνθετικών δεδομένων παραμορφωμένων με συνελικτικό θόρυβο λόγω αντήχησης σε δωμάτιο με ομιλητή σε σπειροειδή τροχιά. Όλα τα χαρακτηριστικά είναι κανονικοποιημένα ως προς τη μέση τιμή και την τυπική απόκλιση. Το τελικό διάνυσμα χαρακτηριστικών προσαυξάνεται σε κάθε περίπτωση με τους Δ και $\Delta\Delta$ συντελεστές.

θόρυβο, η χρήση μόνο των DSCCs φαίνεται να είναι η καλύτερη επιλογή. Σε κάθε περίπτωση, τα DSCCs υπενθυμίζεται πως προτείνεται να χρησιμοποιούνται αυτοτελώς, δηλαδή να αποτελούνται μόνο από δυναμικά χαρακτηριστικά, και όχι σε συνδυασμό με τα MFCCs.



(i) $SNR = 15dB$



(ii) $SNR = 5dB$

Σχήμα 8.13: PER(%) όταν χρησιμοποιούνται τα διανύσματα χαρακτηριστικών MFCC, DSCC, F, MIF και συνδυασμοί τους. Για τον έλεγχο γίνεται χρήση συνθετικών δεδομένων παραμορφωμένων με συνελικτικό θόρυβο λόγω αντήχησης σε δωμάτιο με ακίνητο ομιλητή και με προσθετικό λευκό θόρυβο. Όλα τα χαρακτηριστικά είναι κανονικοποιημένα ως προς τη μέση τιμή και την τυπική απόκλιση. Το τελικό διάνυσμα χαρακτηριστικών προσαυξάνεται σε κάθε περίπτωση με τους Δ και ΔΔ συντελεστές.

Κεφάλαιο 9

Συμπεράσματα

9.1 Σύνοψη των Αποτελεσμάτων και Συμβολή της Εργασίας

Στην παρούσα εργασία έγινε συστηματική προσπάθεια σύγκρισης διαφορετικών συνόλων χαρακτηριστικών και παραμετροποιήσεών τους όταν αυτά χρησιμοποιούνται για Αυτόματη Αναγνώριση Φωνής και ιδιαίτερα για Αναγνώριση Φωνής από Απόσταση με χρήση ενός μικροφώνου. Θέλοντας να εξετάσουμε μόνο την επίδραση των χαρακτηριστικών, χωρίς να υπεισέρχονται άλλοι παράγοντες, η πλειοψηφία των πειραμάτων αφορούσε αναγνώριση φωνημάτων χωρίς τη χρήση κάποιου γλωσσικού μοντέλου. Για την εκπαίδευση του συστήματος έγινε χρήση καθαρών δεδομένων της βάσης Logotypografia, ενώ για τον έλεγχο έγινε χρήση δεδομένων της βάσης ATHENA. Το σύστημα αναγνώρισης που αναπτύχθηκε στηρίχθηκε στο εργαλείο Kaldi.

Συγκεκριμένα, αφότου έγινε μια ανάλυση της περιοχής της Αυτόματης Αναγνώρισης Φωνής και των προκλήσεων που πρέπει να αντιμετωπιστούν (Κεφάλαιο 1) και περιγράφηκαν τα διαδοχικά στάδια ενός σύγχρονου συστήματος αναγνώρισης (Κεφάλαια 2, 3), αναλύθηκε η κατασκευή του συστήματος που χρησιμοποιήθηκε στα πειράματα της εργασίας (Κεφάλαιο 4) και μελετήθηκαν διεξοδικά, σε θεωρητικό και πειραματικό επίπεδο, τα MFCCs, τα DSCCs (Κεφάλαιο 5), τα PLPs, τα RASTA- και J-RASTA-PLPs (Κεφάλαιο 6), αλλά και τα PNCCs (Παράρτημα I). Αναλύθηκαν οι πλέον συχνά χρησιμοποιούμενες μέθοδοι μείωσης της διαστασιμότητας, η PCA, η LDA και η HLDA, καθώς και η τεχνική MLLT που συνήθως χρησιμοποιείται σε συνδυασμό με την LDA, οι οποίες, κατόπιν της σύνδεσης διαδοχικών πλαισίων, χρησιμοποιούνται συχνά για να αποτυπώσουν καλύτερα τη δυναμική του σήματος, αντί των συνηθισμένων συντελεστών ταχύτητας και επιτάχυνσης (Κεφάλαιο 7). Μελετήθηκε, ακόμα, ο TEO και η χρήση του τόσο ως αυτούσιος τελεστής ενέργειας, όσο και ως εργαλείο για την αποδιαμόρφωση AM-FM σημάτων, μέσω του ESA, και για την εξαγωγή σχετικών συνόλων χαρακτηριστικών (Κεφάλαιο 8). Στη συνέχεια, συνοψίζονται τα κύρια συμπεράσματα που εξάγονται από την παρούσα εργασία.

- Όσον αφορά στα πειράματα που έγιναν με τα MFCCs, φαίνεται πως:
  - Μια συστοιχία που καλύπτει όλο το φάσμα συχνοτήτων (από 0 έως τη συχνότητα Nyquist) είναι προτιμότερη από το να αγνοήσουμε τις πολύ χαμηλές ή πολύ υψηλές συχνότητες.

  - Η επιλογή του μήκους του παραθύρου δεν έχει ιδιαίτερη σημασία όταν αυτό ανήκει στο σύνηθες διάστημα $[15msec, 32msec]$.
  - Η προσθήκη των δυναμικών χαρακτηριστικών μέχρι και δευτέρου βαθμού βελτιώνει σημαντικά τα ποσοστά αναγνώρισης, χωρίς αξιοσημείωτες μεταβολές μετά την προσθήκη χαρακτηριστικών μεγαλύτερου βαθμού.
  - Η επιλογή του κατάλληλου παραθύρου για την εξαγωγή των δυναμικών χαρακτηριστικών, μία παράμετρος που συχνά δε λαμβάνεται σοβαρά υπόψιν, παίζει σημαντικό ρόλο. Γενικά, φαίνεται πως ένα σχετικά μικρό παράθυρο (της τάξης των 4 (2+2) γειτονικών πλαισίων), είναι προτιμότερο για καθαρές συνθήκες, ενώ μπορεί να χρειαστεί να μεγαλώσει αρκετά για να έχουμε το βέλτιστο δυνατό αποτέλεσμα σε θορυβώδεις συνθήκες.
  - Η εφαρμογή CMVN έχει εν γένει ευεργετικά αποτελέσματα. Προτείνεται, ωστόσο, για εφαρμογές αναγνώρισης από απόσταση, να γίνεται χρήση CMVN ανά εκφορά και όχι ανά ομιλητή, όπως π.χ. προτείνεται από προεπιλογή στο Kaldi, εφόσον η επίδραση του συνελικτικού θορύβου μεταβάλλεται μεταξύ διαφορετικών εκφορών, ακόμα και με τον ίδιο ομιλητή, καθώς ο ομιλητής μπορεί να βρίσκεται σε διαφορετικά σημεία του δωματίου κάθε φορά.
  - Εκτός από το CMVN, δηλαδή την τελική κανονικοποίηση στο αναφασματικό πεδίο, ιδιαίτερα βοηθητική είναι και η κανονικοποίηση των σημάτων στο πεδίο του χρόνου, κάτι που δεν αποτελεί συνήθη πρακτική.
  - Αρκετά υποσχόμενα παρουσιάζεται να είναι τα DSCCs, δηλαδή ο υπολογισμός των δυναμικών χαρακτηριστικών στο πεδίο του φάσματος αντί του αναφάσματος, πετυχαίνοντας, μάλιστα, ακόμα καλύτερα αποτελέσματα όταν αυτά δε συνδυάζονται με τα στατικά MFCCs, αλλά χρησιμοποιούνται ως ανεξάρτητο σύνολο χαρακτηριστικών.

- Η χρήση των PLPs με τα δεδομένα που έχουμε στη διάθεσή μας οδήγησε στα εξής συμπεράσματα:
  - Τα PLPs οδηγούν σε κάθε περίπτωση σε χειρότερα αποτελέσματα σε σύγκριση με τα MFCCs.
  - Η RASTA ανάλυση βελτιώνει σημαντικά την απόδοση των PLPs. Ωστόσο, όταν τα RASTA-PLPs συνδυάζονται με τους Δ και $\Delta\Delta$ συντελεστές, συνεχίζουν να δίνουν σημαντικά μεγαλύτερο σφάλμα αναγνώρισης συγκριτικά με τα MFCCs+ Δ+$\Delta\Delta$.
  - Όταν θέλουμε να συνδυάσουμε τα RASTA-PLPs με τους Δ και $\Delta\Delta$ συντελεστές, προτείνεται η χρήση του μικρότερου δυνατού παραθύρου για τον υπολογισμό των τελευταίων, δηλαδή η χρήση των δύο άμεσα γειτονικών πλαισίων.
  - Αν και τα διαθέσιμα δεδομένα είναι έντονα αλλοιωμένα και με προσθετικό θόρυβο, η J-RASTA ανάλυση δεν παρουσιάζει κάποιο πλεονέκτημα σε σχέση με την απλή RASTA ανάλυση.

- Όσον αφορά στις τεχνικές μείωσης της διαστασιμότητας που συγκρίνονται:
  - Η PCA δεν καταφέρνει να αντικατοπτρίσει ικανοποιητικά τη δυναμική των σημάτων, δίνοντας σχεδόν σταθερά τα χειρότερα αποτελέσματα.

  - Η LDA οδηγεί σε αρκετά καλύτερα αποτελέσματα, ενώ ο MLLT πάνω στα αποτελέσματα της LDA επιφέρει επιπλέον βελτιώσεις.
  - Σε καθαρές συνθήκες:
    * Εάν αναζητηθεί το βέλτιστο μήκος του παραθύρου για τον υπολογισμό των δυναμικών χαρακτηριστικών ή για τη σύνδεση των διαδοχικών πλαισίων, η LDA+MLLT δε δίνει σημαντικά βελτιωμένα αποτελέσματα σε σύγκριση με τη χρήση των Δ+ΔΔ συντελεστών.
    * Το πλεονέκτημα της LDA(+MLLT) έναντι των Δ+ΔΔ συντελεστών είναι πως εάν χρησιμοποιηθεί ένα εύλογα μεγάλο παράθυρο (μεγαλύτερο των 6 πλαισίων) η απόδοση είναι σχεδόν σταθερή, οπότε δεν απαιτείται αναλυτική αναζήτηση.
    * Το σφάλμα αναγνώρισης ακολουθεί γενικά πτωτική τάση καθώς αυξάνεται το μήκος του διανύσματος χαρακτηριστικών μετά την LDA+MLLT.
  - Στην περίπτωση αναγνώρισης από απόσταση:
    * Η αύξηση του αριθμού πλαισίων που συνδέονται έχει γενικά βελτιωτική επίδραση, παρατήρηση που συμφωνεί με την ανάλογη αύξηση του παραθύρου για τον υπολογισμό των Δ συντελεστών.
    * Ωστόσο, η σύνδεση διαδοχικών πλαισίων και η μετέπειτα μείωση της διαστασιμότητας με οποιαδήποτε μέθοδο επιφέρει αρκετά χειρότερα αποτελέσματα από την απλή χρήση των Δ+ΔΔ συντελεστών, οπότε προτείνεται η χρήση των τελευταίων.
    * Η αύξηση του αριθμού συντελεστών που συναποτελούν το τελικό διάνυσμα χαρακτηριστικών (μετά την LDA+MLLT), δεν οδηγεί απαραίτητα σε καλύτερα αποτελέσματα.

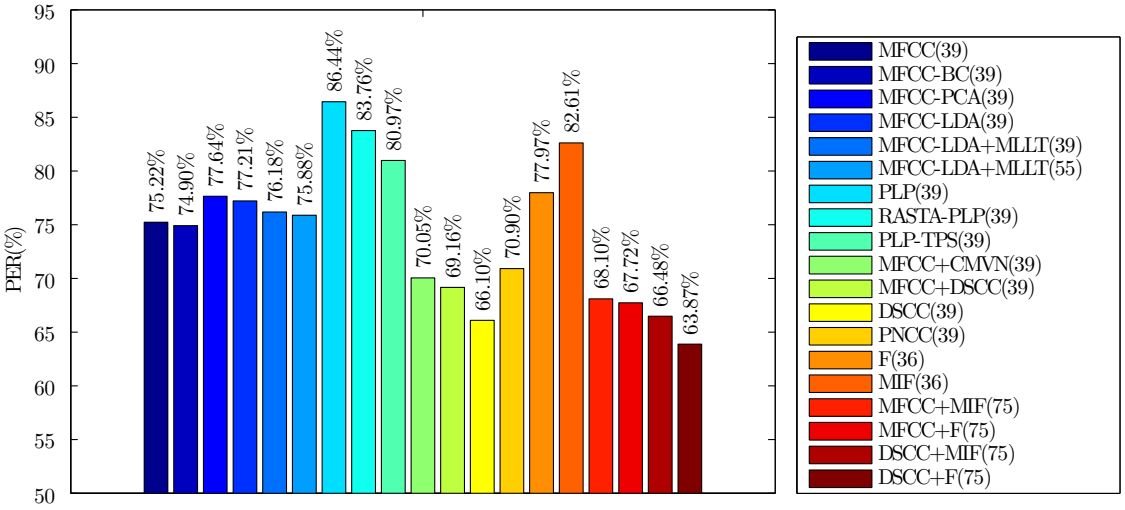
- Για τον TEO προτείνουμε ένα εναλλακτικό πλαίσιο χρήσης του, εισάγοντας την έννοια του TPS:
  - Με τη χρήση του TPS διατηρείται η φυσική ερμηνεία του TEO ως ενεργειακός τελεστής, αλλά αξιοποιούνται, παράλληλα, τα υπολογιστικά πλεονεκτήματα της εργασίας στο πεδίο της συχνότητας. Ουσιαστικά, ο ορισμός του TPS προκύπτει άμεσα από τον κλασικό ορισμό του TEO, εκμεταλλευόμενοι τα θεωρήματα του Parseval και του Plancherel.
  - Το TPS μπορεί να χρησιμοποιηθεί στη ροή εργασίας γνωστών μεθόδων εξαγωγής χαρακτηριστικών και για την ακρίβεια δοκιμάζεται με τα MFCCs, τα PLPs και τα SPNCCs, βελτιώνοντας σε κάθε περίπτωση τα αποτελέσματα, σε σύγκριση με τις κλασικές μεθόδους εξαγωγής των εν λόγω χαρακτηριστικών, που βασίζονται στην τετραγωνική ενέργεια.
  - Προτείνεται μία απλή μέθοδος συνδυασμού του TPS με το "κλασικό" φάσμα ισχύος, που στην ουσία πρόκειται για συνδυασμό του TEO με τον SEO, όπου για τα πρώτα φίλτρα μιας οποιασδήποτε συστοιχίας γίνεται χρήση του πρώτου και για τα υπόλοιπα του δεύτερου, βασιζόμενοι στην παρατήρηση ότι ο TEO είναι πιο αξιόπιστος ενεργειακός τελεστής για τις χαμηλές συχνότητες. Η εν λόγω μέθοδος δίνει σημαντικά βελτιωτικά αποτελέσματα με μία σχετική βελτίωση του PACC από 9% στην περίπτωση των MFCCs, μέχρι και 91% στην περίπτωση των PLPs.

- Αναφορικά με τις πειραματικές δοκιμές που έγιναν με χρήση AM-FM χαρακτηριστικών, όπως εξήχθησαν από τον Gabor ESA:
  - Εξήχθησε και εξετάστηκε μία ποικιλία διαφορετικών χαρακτηριστικών, και συγκεκριμένα τα AMP, MIA, MIF, FMP, B, F, SDIA, SDIF, MAXA, MINA, MAXF, MINF, RANGA και RANGF, τα περισσότερα εκ των οποίων δεν έχουν ξαναχρησιμοποιηθεί αυτόνομα στη βιβλιογραφία για αναγνώριση φωνής. Εξ' αυτών, υποσχόμενα φαίνεται να είναι τα AMP, MIA, MIF, F, SDIA, MAXA και RANGA.
  - Για τη συστοιχία Gabor φίλτρων που χρησιμοποιείται για την εξαγωγή του στιγμιαίου πλάτους και της στιγμιαίας συχνότητας, μέσω των οποίων υπολογίζονται τα AM-FM χαρακτηριστικά, προτείνεται η χρήση 12 φίλτρων κατανεμημένων στην κλίμακα *mel* με ποσοστό επικάλυψης μεταξύ των διαδοχικών φίλτρων ίσο με 70%.
  - Τα καλύτερα αποτελέσματα, από άποψη ευρωστίας, λαμβάνονται με χρήση του F, που προκύπτει ως η πρώτη σταθμισμένη ροπή της στιγμιαίας συχνότητας με βάρος το τετραγωνικό στιγμιαίο πλάτος. Τα παραπάνω αποτελέσματα βελτιώνονται σημαντικά εάν το διάνυσμα χαρακτηριστικών επαυξηθεί με το λογάριθμο της τετραγωνικής ενέργειας, οπότε τα αποτελέσματα γίνονται συγκρίσιμα ή και καλύτερα από αυτά που προκύπτουν με χρήση των MFCCs.
  - Τα AM-FM χαρακτηριστικά εξετάστηκαν, επιπλέον, σε συνδυασμό με τα MFCCs, για τη δημιουργία ενός υβριδικού διανύσματος χαρακτηριστικών. Όταν τα MFCCs συνδυάζονται με τα F ή τα MIF, τα ποσοστά σφάλματος μειώνονται τόσο σε καθαρές, όσο και σε θορυβώδεις συνθήκες. Μεταξύ των F και MIF προτείνεται η χρήση των πρώτων.
  - Επιπλέον βελτιώσεις μπορούν να προκύψουν από το συνδυασμό των F (ή των MIF) με τα DSCCs, όταν τα τελευταία χρησιμοποιούνται αυτόνομα, μαζί με τους Δ και $\Delta\Delta$ συντελεστές τους, ως σύνολο δυναμικών συντελεστών. Σύμφωνα με τα πειράματα που έγιναν στα πραγματικά, αλλά και σε συνθετικά δεδομένα, σε καθαρές συνθήκες ή συνθήκες συνελικτικού θορύβου (αντήχησης), προτείνεται ο συνδυασμός DSCC+F. Όταν ενυπάρχει και μικρή ή μέτρια παραμόρφωση λόγω προσθετικού θορύβου, τότε ο συνδυασμός DSCCs και F συνεχίζει να είναι προτιμητέος. Όταν, όμως, η παραμόρφωση αυτή γίνεται πολύ έντονη, τότε προτείνεται η αυτόνομη χρήση των DSCCs.

Για μια άμεση σύγκριση των πιο σημαντικών αποτελεσμάτων μεταξύ διαφορετικών συνόλων χαρακτηριστικών που μελετήθηκαν στην παρούσα εργασία για αναγνώριση από απόσταση, παρατίθεται το Σχήμα 9.1, όπου παρουσιάζονται τα σφάλματα αναγνώρισης για τα δεδομένα του μικροφώνου OA6. Τα διάφορα σύνολα χαρακτηριστικών που εξετάζονται στο Σχήμα αυτό περιγράφονται στον Πίνακα 9.1.

9.2 Κατευθύνσεις για Μελλοντική Έρευνα

Καθώς η παρούσα εργασία φτάνει στο τέλος της, ελπίζω ότι κατάφερε να δώσει απαντήσεις σε ερωτήματα που απασχολούν την περιοχή της Αναγνώρισης Φωνής από Απόσταση, αλλά και ότι επίσης έδωσε εναύσματα και άνοιξε το δρόμο για περαιτέρω ερευνητική μελέτη στα θέματα όπου επικεντρώθηκε. Ορισμένες κατευθύνσεις όπου θα μπορούσε να στραφεί η μελλοντική έρευνα είναι:



Σχήμα 9.1: PER(%) όταν χρησιμοποιούνται οι καλύτερες παραμετροποιήσεις διαφόρων συνόλων χαρακτηριστικών, όπως αυτές προέκυψαν από ορισμένα εκ των πειραμάτων της εργασίας, σύμφωνα με τα αποτελέσματα για το μικρόφωνο OA6. Σε παρενθέσεις αναγράφεται σε μήκος του εκάστοτε διανύσματος χαρακτηριστικών. Τα χαρακτηριστικά που εξετάζονται εδώ περιγράφονται στον Πίνακα 9.1.

Συμβολισμός	Περιγραφή
MFCC(39)	Τα 13 MFCCs μαζί με τους 13 Δ και 13 ΔΔ συντελεστές, οι οποίοι εξάγονται λαμβάνοντας υπόψιν $3+3=6$ γειτονικά πλαίσια. Ως πρώτο MFCC θεωρείται η λογαριθμική τετραγωνική ενέργεια του σήματος. Το συγκεκριμένο σύνολο αποτέλεσε το βασικό σημείο αναφοράς για συγκριτικές μελέτες στα περισσότερα πειράματα της εργασίας. [Σχήμα 5.4]
MFCC-BC(39)	Όμοια με το σύνολο MFCC(39), με τη διαφορά ότι οι δυναμικοί συντελεστές λαμβάνονται βάσει των $6+6=12$ γειτονικών πλαισίων, επιλογή που έδωσε τα καλύτερα αποτελέσματα για τις εκφορές του OA6. [Σχήμα 5.4]
MFCC-PCA(39)	Σύνδεση $2\cdot10+1=21$ γειτονικών πλαισίων, σε καθένα από τα οποία έχουν υπολογιστεί τα 13 MFCCs και μετέπειτα μείωση της διαστασιμότητας σε 39 χαρακτηριστικά με PCA. [Σχήμα 7.3ii]
MFCC-LDA(39)	Σύνδεση $2\cdot8+1=17$ γειτονικών πλαισίων, σε καθένα από τα οποία έχουν υπολογιστεί τα 13 MFCCs και μετέπειτα μείωση της διαστασιμότητας σε 39 χαρακτηριστικά με LDA. [Σχήμα 7.3ii]
MFCC-LDA+MLLT(39)	Σύνδεση $2\cdot10+1=21$ γειτονικών πλαισίων, σε καθένα από τα οποία έχουν υπολογιστεί τα 13 MFCCs, μετέπειτα μείωση της διαστασιμότητας σε 39 χαρακτηριστικά με LDA και εφαρμογή MLLT με 5 επαναλήψεις. [Σχήμα 7.3ii]

Πίνακας 9.1: Περιγραφή των συνόλων χαρακτηριστικών που εξετάζονται στο Σχήμα 9.1.

Συμβολισμός	Περιγραφή
MFCC-LDA+MLLT(55)	Σύνδεση $2{\cdot}10{+}1 = 21$ γειτονικών πλαισίων, σε καθένα από τα οποία έχουν υπολογιστεί τα 13 MFCCs, μετέπειτα μείωση της διαστασιμότητας σε 55 χαρακτηριστικά με LDA και εφαρμογή MLLT με 5 επαναλήψεις. [Σχήμα 7.4ii]
PLP(39)	Τα 13 PLPs μαζί με τους 13 Δ και 13 $\Delta\Delta$ συντελεστές, οι οποίοι εξάγονται λαμβάνοντας υπόψιν $3+3=6$ γειτονικά πλαίσια. [Πίνακας 6.2]
RASTA-PLP(39)	Τα 13 PLPs, όπως υπολογίζονται μετά από RASTA, μαζί με τους 13 Δ και 13 $\Delta\Delta$ συντελεστές, οι οποίοι εξάγονται λαμβάνοντας υπόψιν $1+1=2$ γειτονικά πλαίσια. [Σχήμα 6.7]
PLP-TPS(39)	Όμοια με το σύνολο PLP(39), με τη διαφορά ότι στη ροή εργασίας εξαγωγής των PLPs έχει εισαχθεί η έννοια του TPS. Για τα πρώτα 16 φίλτρα της συστοιχίας χρησιμοποιείται το TPS, ενώ για τα υπόλοιπα $21-16=5$ φίλτρα χρησιμοποιείται το κλασικό φάσμα ισχύος. [Πίνακας 8.1]
MFCC+CMVN(39)	Όμοια με το σύνολο MFCC(39), με τη διαφορά ότι λαμβάνει χώρα και κανονικοποίηση των αναφασματικών συντελεστών ως προς μέση τιμή και τυπική απόκλιση, ανά εκφορά. [Σχήμα 5.5]
MFCC+DSCC(39)	Τα 13 MFCCs, αφού έχει λάβει χώρα CMVN, προσαυξημένα με τα $13+13=26$ DSCCs, τα οποία εξάγονται με παράμετρο $M=5$ στη σχέση (5.15). [Σχήμα 5.8]
DSCC(39)	Τα 13 DSCCs, όπως προκύπτουν βάσει των MFCCs, μαζί με τους 13 Δ και 13 $\Delta\Delta$ συντελεστές, οι οποίοι υπολογίζονται επί των DSCCs, λαμβάνοντας υπόψιν $5+5=10$ γειτονικά πλαίσια. [Σχήμα 5.9]
PNCC(39)	Τα 13 PNCCs, όπως υπολογίζονται με παραμέτρους $C=2$ και $L=5$ στις σχέσεις (I.5) και (I.5), αντίστοιχα, μαζί με τους 13 Δ και 13 $\Delta\Delta$ συντελεστές, οι οποίοι εξάγονται λαμβάνοντας υπόψιν $3+3=6$ γειτονικά πλαίσια. [Σχήμα I.3ii]
F(36)	Οι 12 συντελεστές F, όπως υπολογίζονται από συστοιχία 12 Gabor φίλτρων με επικάλυψη 70%, μαζί με τους 12 Δ και 12 $\Delta\Delta$ συντελεστές, οι οποίοι εξάγονται λαμβάνοντας υπόψιν $3+3=6$ γειτονικά πλαίσια. Τα F κανονικοποιούνται ως προς μέση τιμή και τυπική απόκλιση. [Σχήμα 8.5iv]
MIF(36)	Οι 12 συντελεστές MIF, όπως υπολογίζονται από συστοιχία 12 Gabor φίλτρων με επικάλυψη 70%, μαζί με τους 12 Δ και 12 $\Delta\Delta$ συντελεστές, οι οποίοι εξάγονται λαμβάνοντας υπόψιν $3+3=6$ γειτονικά πλαίσια. Τα MIF κανονικοποιούνται ως προς μέση τιμή και τυπική απόκλιση. [Σχήμα 8.5iv]
MFCC+MIF(75)	Υβριδικό διάνυσμα χαρακτηριστικών που αποτελείται από τα σύνολα MFCC+CMVN(39) και MIF(36). [Σχήμα 8.7]

Πίνακας 9.1: Περιγραφή των συνόλων χαρακτηριστικών που εξετάζονται στο Σχήμα 9.1 (συνέχεια).

Συμβολισμός	Περιγραφή
MFCC+F(75)	Υβριδικό διάνυσμα χαρακτηριστικών που αποτελείται από τα σύνολα MFCC+CMVN(39) και F(36). [Σχήμα 8.7]
DSCC+MIF(75)	Υβριδικό διάνυσμα χαρακτηριστικών που αποτελείται από τα σύνολα DSCC(39) και MIF(36). [Σχήμα 8.8]
DSCC+F(75)	Υβριδικό διάνυσμα χαρακτηριστικών που αποτελείται από τα σύνολα DSCC(39) και F(36). [Σχήμα 8.8]

Πίνακας 9.1: Περιγραφή των συνόλων χαρακτηριστικών που εξετάζονται στο Σχήμα 9.1 (συνέχεια).

- Μελέτη νέων μεθόδων που μπορούν να αντικατοπτρίσουν καλύτερα τις χρονικές μεταβολές και τα δυναμικά χαρακτηριστικά των σημάτων φωνής, πέρα από τους συνήθεις Δ και ΔΔ συντελεστές και τις διάφορες τεχνικές συνένωσης διαδοχικών πλαισίων και μετέπειτα μείωσης της διαστασιμότητας, που όπως είδαμε δεν είναι πάντα βοηθητικές σε περιπτώσεις αναγνώρισης από απόσταση. Όπως φάνηκε από τη χρήση των DSCCs, η χρήση των δυναμικών χαρακτηριστικών όχι μόνο είναι απαραίτητη για τη λήψη ικανοποιητικών αποτελεσμάτων αναγνώρισης, αλλά σε θορυβώδεις συνθήκες πολλές φορές ενδείκνυται η χρήση τους αυτοτελώς, χωρίς το συνδυασμό με στατική πληροφορία.

- Προσπάθεια για επιπρόσθετους συνδυασμούς ακουστικών χαρακτηριστικών που φέρουν συμπληρωματική πληροφορία μεταξύ τους ή/και επηρεάζονται διαφορετικά από τα διάφορα είδη θορύβου και μπορούν, έτσι, να οδηγήσουν από κοινού σε μια πιο εύρωστη αναγνώριση.

- Αξιοποίηση πολλαπλών μικροφώνων σε συνδυασμό με την εύρεση κατάλληλων ακουστικών χαρακτηριστικών. Η παρούσα εργασία ασχολήθηκε αποκλειστικά με τη χρήση ενός μοναδικού μικροφώνου καταγραφής και είδαμε πως διαφορετικά σύνολα χαρακτηριστικών είναι καταλληλότερα απο άλλα αναλόγως των συνθηκών περιβάλλοντος. Παράλληλα, οι υπάρχουσες προσπάθειες αξιοποίησης συστοιχιών μικροφώνων, θεωρούν ότι τα ίδια σύνολα χαρακτηριστικών εξάγονται από κάθε ηχογράφηση (π.χ. MFCCs), τα οποία σε επόμενο στάδιο συνδυάζονται με κάποιο τρόπο. Αντιθέτως, πιθανώς να δρούσε ευεργετικά η εξαγωγή διαφορετικών συνόλων χαρακτηριστικών από τις καταγραφές κάθε μικροφώνου, αναλόγως της θέσης τους στο δωμάτιο και της αναμενόμενης επίδρασης του θορύβου.

- Συστηματική προσπάθεια συνδυασμού των δύο ενεργειακών τελεστών που δοκιμάστηκαν, του TEO και του SEO, με βάση εναλλακτικά κριτήρια αντί της ιδέας για χρήση του TEO για τα πρώτα φίλτρα (χαμηλές συχνότητες) μιας συστοιχίας και του SEO για τα υπόλοιπα, ώστε να αξιοποιηθούν στο μέγιστο τα θετικά στοιχεία του κάθε τελεστή και να απομονωθούν τα αρνητικά τους.

Παράρτημα I

Power Normalized Cepstral Coefficients (PNCCs)

Μία από τις πολυάριθμες ιδέες που έχουν προταθεί για εύρωστη αναγνώριση φωνής έχει οδηγήσει σε ένα σύνολο χαρακτηριστικών που καλούνται Συντελεστές Αναφάσματος Κανονικοποιημένοι ως προς την Ισχύ (Power Normalized Cepstral Coefficients - PNCCs) [111]. Η πλέον πρόσφατη υλοποίηση της συγκεκριμένης μεθόδου εξαγωγής χαρακτηριστικών προτείνεται στο [113], αλλά τα πειραματικά αποτελέσματα (που αναλύονται στην πορεία) ήταν αρκετά καλύτερα για τη μέθοδο που αναλύεται στο [112], γι' αυτό και εδώ θα ακολουθήσουμε την τελευταία.

Σε γενικές γραμμές, η εξαγωγή των PNCCs ακολουθεί παρόμοια βήματα με αυτή των MFCCs, παρουσιάζοντας, όμως, κάποιες θεμελιώδεις διαφορές, οι οποίες θα γίνουν σαφείς στην πορεία. Αρχικά, λοιπόν, το εκάστοτε σήμα περνάει από σύστημα προέμφασης με συνάρτηση μεταφοράς που δίνεται από τη σχέση (5.1), όπου $\tilde{a} = 0.97$, και χωρίζεται σε επικαλυπτόμενα πλαίσια μέσω Hamming παραθύρωσης. Σε κάθε παράθυρο, έστω $s_i[n]$, λαμβάνεται, μέσω του DFT N σημείων, το φάσμα του $S_i[k]$, το οποίο αναλύεται από μια συστοιχία φίλτρων $H^j[k],\ j = 1, \cdots, Q$, για να καταλήξουμε στους συντελεστές $G_i(j)$:

$$G_i(j) = \sum_{k=0}^{N/2} \left\{ \left| S_i[k] \cdot H^j[k] \right|^2 \right\} \tag{I.1}$$

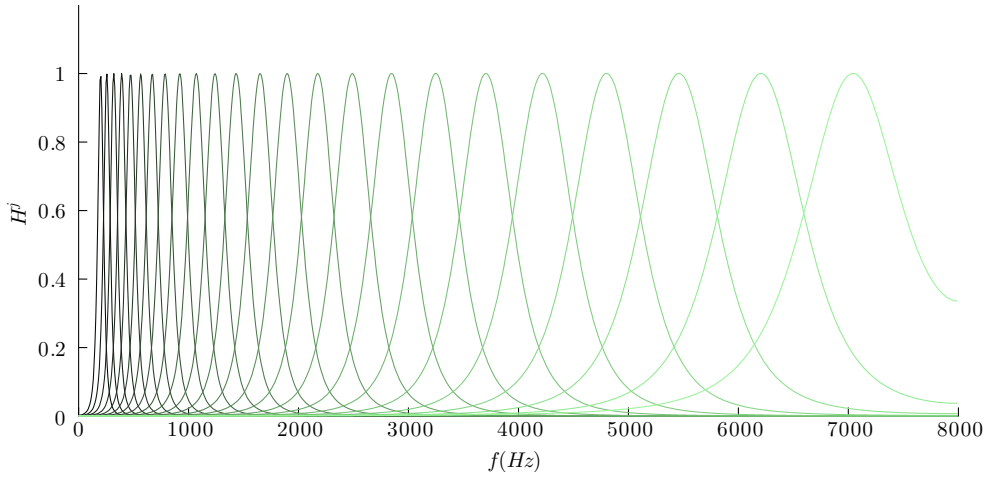
Αντί των τριγωνικών φίλτρων των MFCCs ή των τραπεζοειδών των PLPs, εδώ χρησιμοποιείται συστοιχία gammatone φίλτρων, η οποία προτάθηκε στο [139] για τη μοντελοποίηση του κοχλία του ανθρώπινου αυτιού. Η απόκριση συχνότητας $g(t)$ ενός gammatone φίλτρου δίνεται στη γενική περίπτωση ως

$$g(t) = at^{n-1} e^{-2\pi bt} \cos\left(2\pi f_c t + \phi\right), \tag{I.2}$$

όπου f_c η κεντρική συχνότητα του φίλτρου. Η παράμετρος b καθορίζει το μήκος της κρουστικής απόκρισης του φίλτρου, άρα και το εύρος ζώνης του, ενώ η παράμετρος n είναι η τάξη του φίλτρου και καθορίζει την κλίση των "ουρών" του. Για την υπό συζήτηση συστοιχία, ορίζεται $n = 4$ και $b = 1.019 ERB(f_c)$. Το ERB (Equivalent Rectangular Bandwidth - Ισοδύναμο Ορθογωνικό Εύρος Ζώνης) έχει εισαχθεί για την εκτίμηση του εύρους ζώνης ενός μη-συμμετρικού IIR φίλτρου και προσεγγίζεται εδώ ως

$$ERB(f_c) = 24.7 \frac{4.37 f_c}{1000} + 1\,. \tag{I.3}$$

Η συστοιχία που δημιουργείται, όπου τα φίλτρα είναι γραμμικά κατανεμημένα στην κλίμακα ERB, παρουσιάζεται στο Σχήμα I.1, για 24 φίλτρα και συχνότητα δειγματοληψίας $F_s = 16kHz$, όταν τα φίλτρα είναι κανονικοποιημένα, ώστε να έχουν μοναδιαίο ύψος. Στην πράξη, η συστοιχία που θα χρησιμοποιηθεί αποτελείται από 40 φίλτρα.



Σχήμα I.1: Συστοιχία gammatone φίλτρων για την εξαγωγή των PNCCs. Θεωρείται συχνότητα δειγματοληψίας $16kHz$, ενώ η συστοιχία αποτελείται από 24 φίλτρα. Η κεντρική συχνότητα του πρώτου φίλτρου είναι $200Hz$.

Οι συντελεστές $G_i(j)$ κανονικοποιούνται, στη συνέχεια, ως προς το G_{peak}, που ισούται με το 95ο εκατοστημόριο (percentile) των στοιχείων $\left\{\sum_j G_i(j)\right\}_i$:

$$\tilde{G}_i(j) = g_0 \frac{G_i(j)}{G_{peak}}\,, \tag{I.4}$$

όπου g_0 μία σταθερά άνευ σημασίας που χρησιμοποιείται για να έρθουν οι συντελεστές σε ένα εύλογο εύρος τιμών.

Τα επόμενα στάδια συναποτελούν μία διαδικασία που καλείται Αφαίρεση Μεροληψίας Ισχύος (Power Bias Subtraction - PBS). Στόχος είναι να μεγιστοποιηθεί η οξύτητα του φάσματος ισχύος στις διάφορες συχνότητες (αφού έχει περάσει από τη συστοιχία φίλτρων), καθώς η ανθρώπινη ακοή είναι πιο ευαίσθητη στις μεταβολές του φάσματος σε σύγκριση με το σταθερό φασματικό υπόβαθρο. Εάν δε λάβει χώρα η PBS, προκύπτουν τα Απλά PNCCs (Simple PNCCs - SPNCCs) [113]. Το πρώτο βήμα της PBS, λοιπόν, είναι η εύρεση της ισχύος μέσης διάρκειας μέσω του τρέχοντος μέσου όρου των $\tilde{G}_i(j)$:

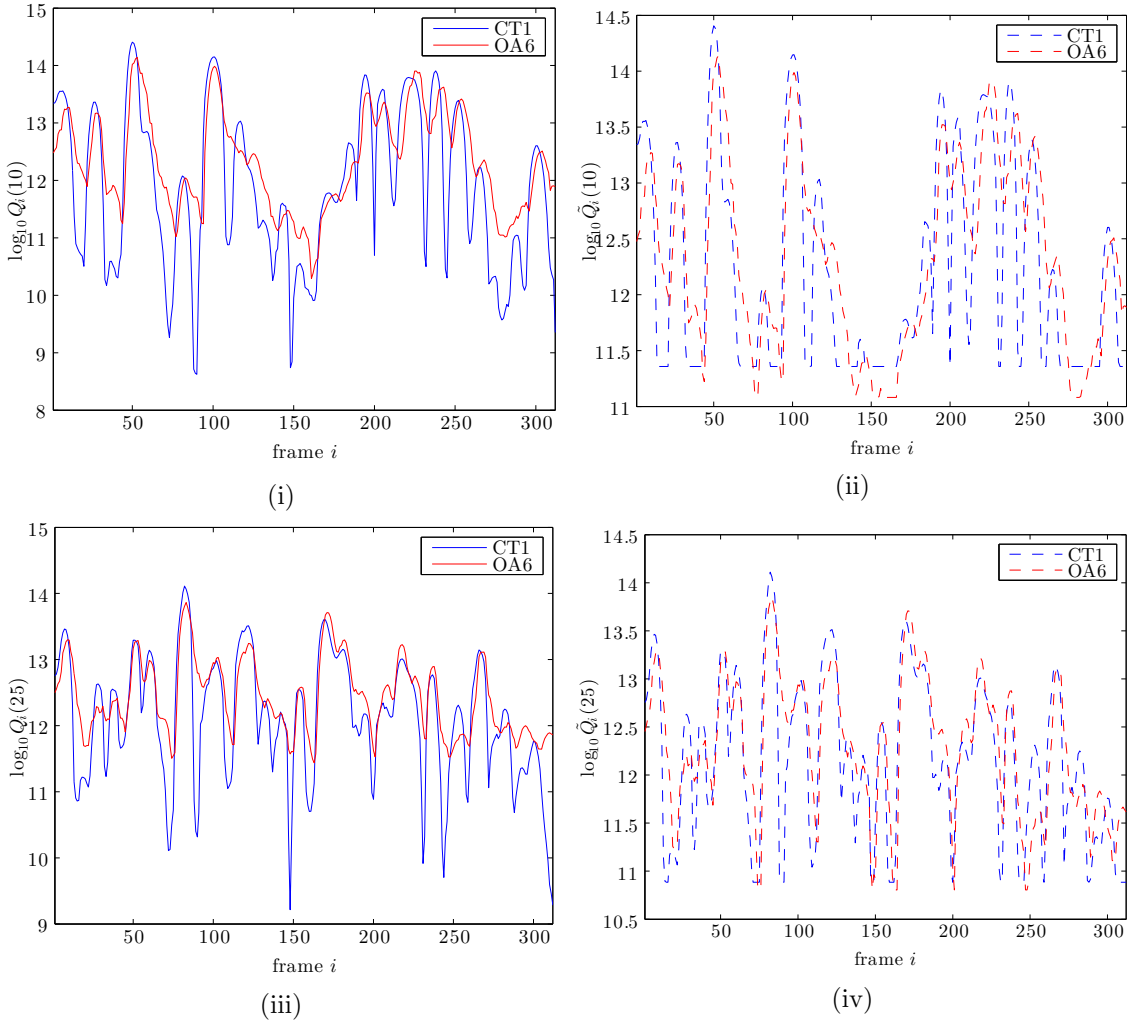
$$Q_i(j) = \frac{1}{2C+1} \sum_{j'=j-C}^{j+C} \tilde{G}_i(j') \tag{I.5}$$

Η μείωση της επίδρασης του φασματικού υποβάθρου γίνεται με την αφαίρεση από τους προκύπτοντες συντελεστές μίας τιμής q_0. Για την αποφυγή, ωστόσο, αδικαιολόγητα μικρών τιμών, υπάρχει ένα κατώφλι q_f. Προκύπτουν, έτσι, οι συντελεστές

$$\tilde{Q}_i(j) = \max\left\{Q_i(j) - q_0, q_f\right\}\,. \tag{I.6}$$

Οι τιμές των q_0, q_f προκύπτουν αναλυτικά μέσω της αναζήτησης του προτεινόμενου αλγορίθμου στο [112], που στηρίζεται στο λόγο του αριθμητικού μέσου προς το γεωμετρικό μέσο των συντελεστών ισχύος $Q_i(j)$. Η σημασία του συγκεκριμένου σταδίου γίνεται εμφανής στο Σχήμα I.2, όπου φαίνεται ότι οι $\tilde{Q}_i(j)$ για τα θορυβώδη δεδομένα (μικρόφωνο OA6) ακολουθούν σε πολύ μεγαλύτερο βαθμό τους αντίστοιχους συντελεστές για τα δεδομένα του μικροφώνου CT1, σε σύγκριση με το τι συμβαίνει με τους συντελεστές $Q_i(j)$ πριν επενεργήσει η σχέση (I.6). Ακολουθεί μία τελική ομαλοποίηση των φασματικών τιμών που έχουν προκύψει αρχικά, μέσω μιας νέας πράξης μέσου:

$$P_i(j) = \left(\frac{1}{2L+1} \sum_{j'=j-L}^{j+L} \frac{\tilde{Q}_i(j')}{Q_i(j')} \right) \tilde{G}_i(j) \tag{I.7}$$

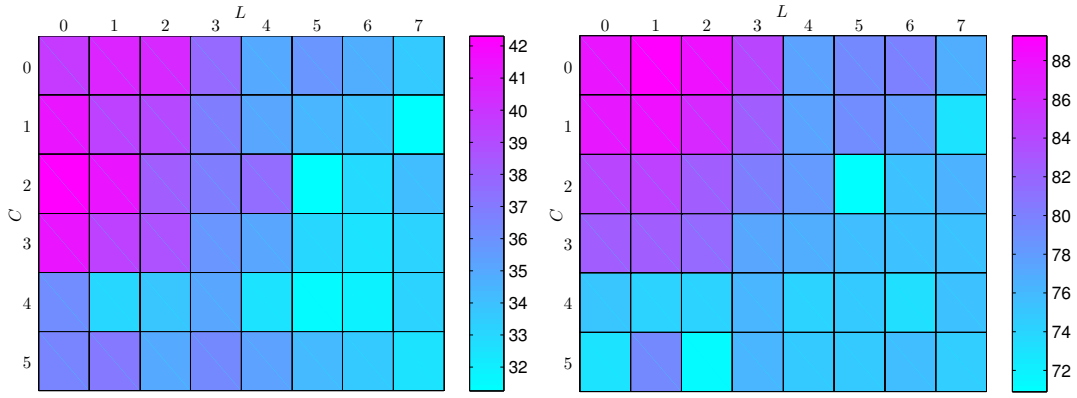


Σχήμα I.2: Ισχύς μέσης διάρκειας πριν ((i), (iii)) και μετά ((ii), (iv)) τη μείωση της επίδρασης του φασματικού υποβάθρου, όπως αυτή λαμβάνει χώρα μέσω του αλγορίθμου PBS, για το 10ο ((i), (ii)) και το 25ο ((iii), (iv)) φίλτρο της συστοιχίας για την ηχογράφηση της φράσης "*Έχω επικαλεστεί κάθε είδους επιχειρήματα για να το αποδείξουμε*" από τα μικρόφωνα CT1 και OA6, η οποία είναι αλλοιωμένη λόγω μουσικής στο υπόβαθρο (στην περίπτωση του OA6).

Τέλος, αφότου τα $P_i(j)$ (που στην περίπτωση των SPNCCs ταυτίζονται με τα $\tilde{G}_i(j)$) περάσουν από μία μη-γραμμική συνάρτηση, εφαρμόζεται DCT και διατηρούνται οι πρώτοι N_c συντελεστές, όπου πρακτικά $N_c = 13$, όπως και στην περίπτωση των MFCCs. Η μη-γραμμικότητα που εφαρμόζεται δεν είναι ο λογάριθμος, όπως στα MFCCs, αλλά η ύψωση σε έναν εκθέτη $0 < a < 1$. Ο εκθέτης αυτός, που στην περίπτωση των PLPs υπενθυμίζεται ότι ισούται με 0.33, εδώ ορίζεται ίσος με $^1/_{15}$.

Ο αλγόριθμος PBS κρίθηκε πως θα πρέπει να απενεργοποιηθεί για τα πειράματα της Ενότητας 8.2, καθώς θεωρήθηκε πως δε θα είχε κάποιο φυσικό νόημα η συνάρτηση του τρέχοντος μέσου όρου της σχεσης (I.5) για εξομάλυνση μεταξύ καναλιών όπου η ενέργεια έχει υπολογιστεί με διαφορετικούς τελεστές. Τονίζεται ότι σε όλα τα πειράματα χρησιμοποιείται ο κώδικας που είναι διαθέσιμος από τους συγγραφείς, με μόνες διαφορές την απενεργοποίηση του προτεινόμενου CMN και τη μεγέθυνση του παραθύρου από τα $25.6msec$ που προτείνονται στα $32msec$, που έχουμε χρησιμοποιήσει καθ' όλη τη διάρκεια της εργασίας. Το διάνυσμα χαρακτηριστικών των PNCCs προσαυξάνεται με τους Δ και $\Delta\Delta$ συντελεστές, όπως προκύπτουν, λαμβάνοντας υπόψιν παράθυρο μήκους 7 πλαισίων.

Στο σημείο αυτό, για να φανεί στην πράξη η επίδραση της PBS στην αναγνώριση, γίνονται κάποια πειράματα με χρήση των PNCCs. Δύο ελεύθερες παράμετροι που έχουν αρκετά μεγάλη σημασία για τα τελικά αποτελέσματα της μεθόδου είναι η C (σχέση (I.5)) και η L (σχέση (I.7)). Τα αποτελέσματα των εν λόγω πειραμάτων για διαφορετικές τιμές των παραμέτρων αυτών δίνονται στο Σχήμα I.3. Οι όποιες παράμετροι εμπλέκονται για τον υπολογισμό των q_0 και q_f της σχέσης (I.6) είναι αυτές που προτείνονται στο [112].



(i) Μικρόφωνο CT1 (ii) Μικρόφωνο OA6

Σχήμα I.3: PER(%) όταν χρησιμοποιούνται PNCCs με διαφορετικές τιμές των παραμέτρων C (σχέση (I.5)) και L (σχέση (I.7)).

Όπως φαίνεται, πολύ μικρές τιμές των παραμέτρων C και L, δηλαδή ομαλοποίηση των φασματικών παραμέτρων που βασίζεται μόνο στα πολύ κοντινά γειτονικά πλαίσια, οδηγούν σε μεγάλα σφάλματα αναγνώρισης, σε σύγκριση, πάντα, με τα MFCCs, που, όπως είδαμε στο Κεφάλαιο 5,οδηγούν σε PER 31.56% και 75.22% για τα μικρόφωνα CT1 και OA6, αντίστοιχα. Όπως είναι αναμενόμενο, βέβαια, αύξηση των τιμών αυτών δε συνεπάγεται μονότονη μεταβολή των σφαλμάτων αναγνώρισης. Ωστόσο, με μία προσεκτική επιλογή των εν λόγω τιμών, μπορεί να επιτευχθεί μη αμελητέα βελτίωση του PER, ιδίως όσον αφορά σε συνθήκες θορύβου. Συγκεκριμένα, για τα διαθέσιμα δεδομένα, ο βέλτιστος συνδυασμός C και L προκύπτει να είναι 2 και 5, αντίστοιχα, οδηγώντας σε PER ίσο με 31.29% για το μικρόφωνο CT1 και 70.90% για το μικρόφωνο OA6.

Παράρτημα ΙΙ

Μέθοδος των Φανταστικών Πηγών για Προσομοίωση της Αντήχησης

Τα ακουστικά κύματα που φτάνουν σε ένα μέσο καταγραφής μπορούν να διαχωριστούν σε τρεις κατηγορίες [34]:

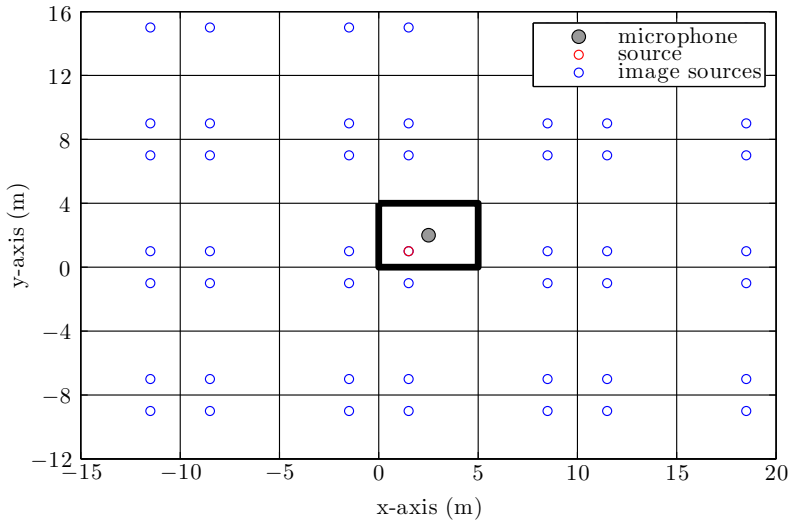
- το άμεσο κύμα,
- τις πρώιμες ανακλάσεις και
- τις αργές ανακλάσεις.

Το άμεσο κύμα είναι αυτό που μεταφέρει τη ζητούμενη πληροφορία μέσω ενός άμεσου μονοπατιού, ενώ οι ανακλάσεις οφείλονται στα διάφορα εμπόδια, όπως τοίχοι και αντικείμενα, που υπάρχουν στο χώρο καταγραφής. Οι αργές ανακλάσεις δεν είναι διακρίσιμες μεταξύ τους και έχουν ως αποτέλεσμα τη δημιουργία ενός διάχυτου θορυβώδους πεδίου. Το φαινόμενο όπου πολλά ακουστικά κύματα που μεταφέρουν την ίδια πληροφορία φτάνουν στο μέσο καταγραφής σχεδόν ταυτόχρονα, με αποτέλεσμα να μη γίνονται αντιληπτά ως διακριτές οντότητες, ονομάζεται αντήχηση (reverberation).

Η σχέση μεταξύ του δωματίου καταγραφής και της δημιουργηθείσας αντήχησης περιγράφεται μαθηματικά από την Κρουστική Απόκριση του Δωματίου (Room Impulse Response - RIR). Το τελικό σήμα που καταγράφεται δίνεται από τη συνέλιξη της RIR με το άμεσο κύμα. Μια συχνά χρησιμοποιούμενη ποσότητα που δείχνει πόσο ισχυρή είναι η συνελικτική παραμόρφωση που εισάγει η αντήχηση σε ένα δωμάτιο είναι ο λεγόμενος χρόνος αντήχησης T_{60}. Ο χρόνος αντήχησης ορίζεται ως ο χρόνος που απαιτείται για την εξασθένηση ενός σήματος $60dB$ χαμηλότερα από το αρχικό του επίπεδο ακουστικής πίεσης. Στην πραγματικότητα, αυτό καθορίζεται από το συντελεστή απορρόφησης των διαφόρων επιφανειών.

Λόγω της μεγάλης σημασίας που έχει η μελέτη του φαινομένου της αντήχησης, έχουν αναπτυχθεί υπολογιστικές μέθοδοι που προσομοιώνουν την ακουστική μικρών δωματίων, μέσω της εκτίμησης της RIR. Μία από αυτές είναι η Μέθοδος των Φανταστικών Πηγών (Image-Source Method - ISM) [138]. Όπως αποδεικνύεται, στην περίπτωση ενός τοίχου με μηδενικό συντελεστή απορρόφησης, η συνεισφορά του ανακλώμενου κύματος είναι ίδια με τη συνεισφορά μιας δεύτερης, υποθετικής ακουστικής πηγής που βρίσκεται κατοπτρικά απέναντι της πρώτης, στην άλλη πλευρά του τοίχου. Στην περίπτωση ορθογωνικού δωματίου

με 6 τοίχους, θα πρέπει να ληφθούν υπόψιν πολλαπλές ανακλάσεις, κάτι που μπορεί να γίνει με τον εκ νέου κατοπτρισμό κάθε φανταστικής πηγής σε νέες, συμμετρικά τοποθετημένες, (φανταστικές) πηγές. Δημιουργείται κατ' αυτόν τον τρόπο ένα άπειρο ορθογωνικό πλέγμα στις τρεις διαστάσεις. Ένα τμήμα αυτού του πλέγματος στις δύο διαστάσεις απεικονίζεται στο Σχήμα II.1 για τη διάταξη του Σχήματος 8.9. Στη γενικότερη, τώρα, περίπτωση τοίχων με μη μηδενικούς συντελεστές απορρόφησης, το σήμα που φτάνει στο μέσο καταγραφής από κάθε φανταστική πηγή είναι το ίδιο το σήμα της πραγματικής πηγής, με μία χρονική καθυστέρηση που εξαρτάται από την απόσταση, αλλά και με μία εξασθένηση που εξαρτάται από τους εν λόγω συντελεστές των "τοίχων" που παρεμβάλλονται. Σημειώνεται πως για λόγους απλότητας, οι συντελεστές απορρόφησης θεωρούνται ανεξάρτητοι της συχνότητας και της γωνίας πρόσπτωσης, ενώ η ακουστική πηγή θεωρείται σημειακή. Ακόμα, το ορθογωνικό πλέγμα είναι προφανώς πεπερασμένο, με τις διαστάσεις να αυξάνονται όσο μεγαλύτερος είναι ο επιθυμητός αριθμός των διαδοχικών ανακλάσεων που θα ληφθούν υπόψιν και κατ' επέκταση η επιθυμητή ακρίβεια της υπολογιστικής προσομοίωσης.



Σχήμα II.1: Οπτικοποίηση του ορθογωνικού πλέγματος με τις φανταστικές πηγές που χρησιμοποιείται στην ISM για την προσομοίωση της ακουστικής μικρών δωματίων. Παρουσιάζεται μία "τομή" στις δύο διαστάσεις του πλέγματος που δημιουργείται για τη διάταξη του Σχήματος 8.9.

Ένας αλγόριθμος που αξιοποιεί τις παραπάνω ιδέες για την εκτίμηση RIRs και ο οποίος έχει χρησιμοποιηθεί για τα πειράματα της Υποενότητας 8.4.3 αναλύεται στο [137]. Ο συγκεκριμένος αλγόριθμος φαίνεται να δίνει καλύτερα αποτελέσματα από τον αρχικό αλγόριθμο που είχε προταθεί για την υλοποίηση της ISM [138]. Σημειώνεται πως δε χρειάζεται η εκ των προτέρων γνώση των συντελεστών απορρόφησης των τοίχων, παρά μόνο η επιθυμητή σχετική απορρόφηση μεταξύ των διαφορετικών τοίχων, καθώς και ο επιθυμητός χρόνος αντήχησης T_{60}.

Κατάλογος Ακρωνυμίων[1]

AIF	Average Instantaneous Frequency	Μέση Στιγμιαία Συχνότητα
ALE	Average Log-Envelope	Μέση Λογαριθμική Περιβάλλουσσα
AM	Amplitude Modulation	Διαμόρφωση Πλάτους
AMP	Amplitude Modulation Power	Ισχύς Διαμόρφωσης Πλάτους
ASR	Automatic Speech Recognition	Αυτόματη Αναγνώριση Φωνής
CMN	Cepstral Mean Normalization	Αναφασματική Κανονικοποίηση Μέσου
CMS	Cepstral Mean Subtraction	Αναφασματική Αφαίρεση Μέσου
CMU	Carnegie-Mellon University	Πανεπιστήμιο Carnegie-Mellon
CMVN	Cepstral Mean & Variance Normalization	Αναφασματική Κανονικοποίηση Μέσου & Διακύμανσης
DARPA	Defense Advanced Research Projects Agency	Οργανισμός Προηγμένων Ερευνητικών Έργων Άμυνας
DCT	Discrete Cosine Transform	Διακριτός Μετασχηματισμός Συνημιτόνου
DESA	Discrete Energy Separation Algorithm	Διακριτός Αλγόριθμος Διαχωρισμού Ενέργειας
DFA	Deterministic Finite Automaton	Ντετερμινιστικό Πεπερασμένο Αυτόματο
DFT	Discrete Fourier Transform	Διακριτός Μετασχηματισμός Fourier
DSCC	Delta-Spectral Cepstral Coefficient	Δέλτα-Φασματικός Αναφασματικός Συντελεστής
DSR	Distant Speech Recognition	Αναγνώριση Φωνής από Απόσταση
DTW	Dynamic Time Warping	Δυναμική Χρονική Στρέβλωση
EM	Expectation - Maximization	Πρόβλεψη - Μεγιστοποίηση
ERB	Equivalent Rectangular Bandwidth	Ισοδύναμο Ορθογωνικό Εύρος Ζώνης
ESA	Energy Separation Algorithm	Αλγόριθμος Διαχωρισμού Ενέργειας
FA	Finite Automaton	Πεπερασμένο Αυτόματο
FFT	Fast Fourier Transform	Ταχύς Μετασχηματισμός Fourier
FM	Frequency Modulation	Διαμόρφωση Συχνότητας
FMP	Frequency Modulation Percentage	Ποσοστό Διαμόρφωσης Συχνότητας
FSA	Finite State Acceptor	Αποδοχέας Πεπερασμένης Κατάστασης

FSG	Finite State Grammar	Γραμματική Πεπερασμένης Κατάστασης
FST	Finite State Transducer	Μετατροπέας Πεπερασμένης Κατάστασης
GMM	Gaussian Mixture Model	Μοντέλο Μειγμάτων Γκαουσιανών
HLDA	Heteroscedastic Linear Discriminant Analysis	Ετεροσκεδαστική Γραμμική Διακριτική Ανάλυση
HMM	Hidden Markov Model	Κρυφό Μαρκοβιανό Μοντέλο
HSR	Human Speech Recognition	Αναγνώριση Φωνής από τον Άνθρωπο
HTK	Hidden Markov Model Toolkit	Εργαλειοθήκη Κρυφών Μαρκοβιανών Μοντέλων
IIR	Infinite Impulse Response	Άπειρη Κρουστική Απόκριση
ISM	Image-Source Method	Μέθοδος των Φανταστικών Πηγών
KLT	Karhunen–Loève Transform	Karhunen–Loève Μετασχηματισμός
KWS	Keyword Spotting	Στόχευση Λέξεων-Κλειδιών
LDA	Linear Discriminant Analysis	Γραμμική Διακριτική Ανάλυση
LFPC	Log Frequency Power Coefficient	Συντελεστής Ισχύος στις Log Συχνότητες
LID	Language Identification	Ταυτοποίηση Γλώσσας
LMSF	Language Model Scaling Factor	Συντελεστής Κλίμακας του Γλωσσικού Μοντέλου
LVCSR	Large-Vocabulary Continuous-Speech Recognition	Αναγνώριση Συνεχούς Λόγου Μεγάλου Λεξιλογίου
MDA	Multiband Demodulation Analysis	Πολυζωνική Ανάλυση Αποδιαμόρφωσης
MEMS	Microelectromechanical System	Μικροηλεκτρομηχανικό Σύστημα
MF	Modulation Frequency	Συχνότητα Διαμόρφωσης
MFCC	Mel-Frequency Cepstrum Coefficient	Συντελεστής Αναφάσματος στις Mel-Συχνότητες
MFIF	Mel-Frequency Instantaneous Frequency	Στιγμιαία Συχνότητα στις Mel-Συχνότητες
MIA	Mean Instantaneous Amplitude	Μέσο Στιγμιαίο Πλάτος
MIF	Mean Instantaneous Frequency	Μέση Στιγμιαία Συχνότητα
ML	Maximum Likelihood	Μέγιστη Πιθανοφάνεια
MLE	Maximum Likelihood Estimation	Εκτίμηση Μέγιστης Πιθανοφάνειας
MLLT	Maximum Likelihood Linear Transform	Γραμμικός Μετασχηματισμός Μέγιστης Πιθανοφάνειας
MMeDuSA	Modulation of Medium Duration Speech Amplitude	Διαμόρφωση του Πλάτους Φωνής Μέσης Διάρκειας
NFA	Non-deterministic Finite Automaton	Μη-ντετερμινιστικό Πεπερασμένο Αυτόματο
NMCC	Normalized Modulation Cepstral Coefficient	Συντελεστή Αναφάσματος Κανονικοποιημένης Διαμόρφωσης

NTEC	Normalized Teager Energy Cepstral [features]	Κανονικοποιημένα ως προς την Teager Ενέργεια Αναφασματικά [χαρακτηριστικά]
OOV	Out-Of-Vocabulary	Εκτός Λεξιλογίου
PACC	Phone Accuracy	Ακρίβεια Φωνημάτων
PBS	Power Bias Subtraction	Αφαίρεση Μεροληψίας Ισχύος
PCA	Principal Component Analysis	Ανάλυση Κύριων Συνιστωσών
PER	Phone Error Rate	Ποσοστό Σφάλματος Φωνημάτων
PIP	Phoneme Insertion Penalty	Κόστος Εισαγωγής Φωνήματος
PLP	Perceptual Linear Prediction	Γραμμική Πρόβλεψη βασισμένη στην Αντίληψη
PLU	Phonelike Unit	Μονάδα που μοιάζει με φθόγγο
PNCC	Power Normalized Cepstrum Coefficient	Συντελεσής Αναφάσματος Κανονικοποιημένος ως προς την Ισχύ
PSD	Power Spectrum Difference	Διαφορά του Φάσματος Ισχύος
RASR	RWTH Automatic Speech Recognition	Αυτόματη Αναγνώριση Φωνής του RWTH
RASTA	RelAtive SpecTrAl [analysis]	Σχετική Φασματική [ανάλυση]
RIR	Room Impulse Response	Κρουστική Απόκριση Δωματίου
RWTH	Rheinisch-Westfälische Technische Hochschule	Ρηχανικό-Βεστφαλικό Πολυτεχνείο
SDIA	Standard Deviation of Instantaneous Amplitude	Τυπική Απόκλιση του Στιγμιαίου Πλάτους
SDIF	Standard Deviation of Instantaneous Frequency	Τυπική Απόκλιση της Στιγμιαίας Συχνότητας
SEO	Squared Energy Operator	Τελεστής Τετραγωνικής Ενέργειας
SMAC	Spectral Moment feature Augmented by low order Cepstral coefficients	χαρακτηριστικό Φασματικών Ροπών Επαυξημένο με Αναφασματικούς συντελεστές χαμηλής τάξης
SNR	Signal to Noise Ratio	Σηματοθορυβικός Λόγος
SPNCC	Simple Power Normalized Cepstrum Coefficient	Απλός Συντελεσής Αναφάσματος Κανονικοποιημένος ως προς την Ισχύ
SSNRA	Arithmetic Segmental Signal to Noise Ratio	Αριθμητικός Τμηματικός Σηματοθορυβικός Λόγος
STC	Semi-Tied Covariance	Ημι-Δεμένη Συμμεταβλητότητα
STRF	Spectro-Temporal Receptive Field	Φασματο-Χρονικό Δεκτικό Πεδίο
SU	Subword Unit	Υπολεκτική Μονάδα
SUR	Speech Understanding Research	Έρευνα Κατανόησης της Ομιλίας
TECC	Teager Energy Cepstrum Coefficient	Συντελεστής Αναφάσματος με Teager Ενέργειες
TEMFCC	Teager Energy based MFCC	Βασισμένος στην Teager Ενέργεια MFCC

TEO	Teager Energy Operator	Τελεστής Teager Ενέργειας
TEOCEP	TEO based CEPstral [feature parameters]	Βασισμένοι στον TEO [παράμετροι χαρακτηριστικών]
TPS	Teager Power Spectrum	Φάσμα Teager Ισχύος
VAD	Voice Activity Detection	Εντοπισμός Δραστηριότητας Φωνής
WACC	Word Accuracy	Ακρίβεια Λέξεων
WER	Word Error Rate	Ποσοστό Σφάλματος Λέξεων
WFA	Weighted Finite Automaton	Πεπερασμένο Αυτόματο με Βάρη
WFSA	Weighted Finite State Acceptor	Αποδοχέας Πεπερασμένης Κατάστασης με Βάρη
WFST	Weighted Finite State Transducer	Μετατροπέας Πεπερασμένης Κατάστασης με Βάρη
WIP	Word Insertion Penalty	Κόστος Εισαγωγής Λέξης

[1]Πολλές από τις μεταφράσεις τεχνικών όρων που εμφανίζονται τόσο εδώ, όσο και στον κύριο κορμό του κειμένου, βασίστηκαν στο Λεξικό Ορολογίας του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών http://speech.di.uoa.gr/com/Lexiko/Lexiko.htm.

Ο όρος ανάφασμα (cepstrum) έχει καθιερωθεί από τον ΕΛ.Ο.Τ.

Αναφορές

[1] Αριστοτέλης, *Άπαντα, Τόμος 1, Πολιτικά Α'-Β'*. Κάκτος, 1993.

[2] T. Hobbes, *Leviathan: Or, The Matter, Form, and Power of a Common-Wealth Ecclesiastical and Civil*. 1651.

[3] V. S. Ramachandran, *The Tell-Tale Brain: a neuroscientist's quest for what makes us human*. Norton, 2012.

[4] K. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952.

[5] D. Fry, "Theoretical aspects of mechanical speech recognition," *Journal of the British Institution of Radio Engineers*, vol. 19, no. 4, pp. 211–218, 1959.

[6] P. Denes, "The design and operation of the mechanical speech recognizer at university college london," *Journal of the British Institution of Radio Engineers*, vol. 19, no. 4, pp. 219–229, 1959.

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

[8] T. K. Vintsyuk, "Speech discrimination by dynamic programming," *Kibernetika*, vol. 4, no. 1, pp. 81–88, 1968.

[9] J. R. Pierce, "Whither speech recognition?," *The Journal of the Acoustical Society of America*, vol. 46, no. 4B, pp. 1049–1051, 1969.

[10] Committee on Innovations in Computing and Communications: Lessons from History, Computer Science and Telecommunications Board, Commission on Physical Sciences, Mathematics, and Applications National Research Council, *Funding a Revolution: Government Support for Computing Research*. National Academy Press, 1999.

[11] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker-independent recognition of isolated words using clustering techniques," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 4, pp. 336–349, 1979.

[12] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Encyclopedia of Language and Linguistics*, pp. 1–24, 2005.

[13] T. Hori and A. Nakamura, *Speech Recognition Algorithms Using Weighted Finite-State Transducers*. Morgan & Claypool, 2013.

[14] F. Pereira, M. Riley, and R. Sproat, "Weighted rational transductions and their application to human language processing," in *Proceedings of the workshop on Human Language Technology*, pp. 262–267, Association for Computational Linguistics, 1994.

[15] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[16] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," pp. 1694–1694, 2001.

[17] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, "The CMU SPHINX-4 speech recognition system," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, vol. 1, pp. 2–5, Citeseer, 2003.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*. Cambridge University Engineering Department, 2006.

[19] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, "RASR-the RWTH Aachen University open source speech recognition toolkit," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," 2011.

[21] R. I. Damper, "Voice-input aids for the physically disabled," *International Journal of Man-Machine Studies*, vol. 21, no. 6, pp. 541 – 553, 1984.

[22] M. Wald, "Hearing disability and technology," in *Access All Areas: disability, technology and learning* (L. Phipps, A. Sutherland, and J. Seale, eds.), pp. 19–23, JISC Techdis Service with ALT, 2002.

[23] M. Wald, "Captioning for deaf and hard of hearing people by editing automatic speech recognition in real time," in *Computers Helping People with Special Needs*, pp. 683–690, Springer, 2006.

[24] D. Freitas and G. Kouroupetroglou, "Speech technologies for blind and low vision persons," *Technology and Disability*, vol. 20, no. 2, p. 135, 2008.

[25] E. L. Higgins and M. H. Raskind, "Compensatory effectiveness of speech recognition on the written composition performance of postsecondary students with learning disabilities," *Learning Disability Quarterly*, vol. 18, no. 2, pp. 159–174, 1995.

[26] O. Tsimhoni, D. Smith, and P. Green, "Address entry while driving: Speech recognition versus a touch-screen keyboard," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 4, pp. 600–610, 2004.

[27] A. Vizzini, "Method and system for controlling manned and unmanned aircraft using speech recognition tools," Mar. 13 2008. US Patent App. 11/688,045.

[28] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homes—present state and future challenges," *Computer methods and programs in biomedicine*, vol. 91, no. 1, pp. 55–81, 2008.

[29] W. Wahlster, *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media, 2013.

[30] C. T. Ishi, S. Matsuda, T. Kanda, T. Jitsuhiro, H. Ishiguro, S. Nakamura, and N. Hagita, "Robust speech recognition system for communication robots in real environments," in *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, pp. 340–345, IEEE, 2006.

[31] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[32] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[33] D. Jurafsky and J. Martin, *Speech and Language Processing*. Prentice Hall, 2008.

[34] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[35] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[36] H. Hermansky, "Should recognizers have ears?," *Speech communication*, vol. 25, no. 1, pp. 3–27, 1998.

[37] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "Openfst: A general and efficient weighted finite-state transducer library," in *Implementation and Application of Automata*, pp. 11–23, Springer, 2007.

[38] F. Rumsey and T. McCormick, *Sound and Recording*. Focal Press, 2009.

[39] G. A. Fink, *Markov models for pattern recognition: from theory to applications*. Springer, 2008.

[40] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. M. Doss, "Exploiting contextual information for improved phoneme recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4449–4452, IEEE, 2008.

[41] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[42] S. J. Young, "The general use of tying in phoneme-based HMM speech recognisers," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, pp. 569–572, IEEE, 1992.

[43] X. D. Huang and M. A. Jack, "Semi-continuous hidden markov models for speech signals," *Computer Speech & Language*, vol. 3, no. 3, pp. 239–251, 1989.

[44] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*, pp. 307–312, Association for Computational Linguistics, 1994.

[45] J.-P. Hosom, *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. PhD thesis, Oregon Graduate Institute of Science and Technology, 2000.

[46] A. Katsamanis, I. Rodomagoulakis, G. Potamianos, P. Maragos, and A. Tsiami, "Robust far-field spoken command recognition for home automation combining adaptation and multichannel processing," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 5547–5551, IEEE, 2014.

[47] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pp. 310–318, Association for Computational Linguistics, 1996.

[48] I. H. Witten and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *Information Theory, IEEE Transactions on*, vol. 37, no. 4, pp. 1085–1094, 1991.

[49] G. Donaj and Z. Kačič, "The use of several language models and its impact on word insertion penalty in lvcsr," in *Speech and Computer*, pp. 354–361, Springer, 2013.

[50] M. Sipser, *Introduction to the Theory of Computation*. Thomson Course Technology, 2006.

[51] W. Kuich and A. Salomaa, *Semirings, automata, languages*. Springer Verlag, 1986.

[52] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[53] M. Mohri, F. Pereira, and M. Riley, "The design principles of a weighted finite-state transducer library," *Theoretical Computer Science*, vol. 231, no. 1, pp. 17–32, 2000.

[54] M. Mohri, "Weighted automata algorithms," in *Handbook of weighted automata*, pp. 213–254, Springer, 2009.

[55] M. Riley, F. Pereira, and M. Mohri, "Transducer composition for context-dependent network expansion.," in *EUROSPEECH*, pp. 1427–1430, 1997.

[56] C. Chelba, D. Bikel, M. Shugrina, P. Nguyen, and S. Kumar, "Large scale language modeling in automatic speech recognition," *arXiv preprint arXiv:1210.8440*, 2012.

[57] J. B. Allen, "How do humans process and recognize speech?," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 567–577, 1994.

[58] A. Juneja, "A comparison of automatic and human speech recognition in null grammar," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. EL256–EL261, 2012.

[59] P. Schwarz, *Phoneme recognition based on long temporal context*. PhD thesis, 2009.

[60] A. Tsiami, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis, G. Potamianos, and P. Maragos, "Athena: A greek multi-sensory database for home automation control," *Cough*, vol. 96, pp. 1–9, 2014.

[61] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V. Diakoloukas, "Large vocabulary continuous speech recognition in greek: corpus and an automatic dictation system.," in *INTERSPEECH*, 2003.

[62] I. Rodomagoulakis, P. Giannoulis, Z.-I. Skordilis, P. Maragos, and G. Potamianos, "Experiments on far-field multichannel speech processing in smart homes," in *Digital Signal Processing (DSP), 2013 18th International Conference on*, pp. 1–6, IEEE, 2013.

[63] M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer, "Hidden markov model training with contaminated speech material for distant-talking speech recognition," *Computer Speech & Language*, vol. 16, no. 2, pp. 205–223, 2002.

[64] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments.," in *INTERSPEECH*, pp. 806–809, 2000.

[65] M. Vondrášek and P. Pollak, "Methods for speech snr estimation: Evaluation tool and analysis of vad dependency," *Radioengineering*, vol. 14, no. 1, pp. 6–11, 2005.

[66] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an open source toolkit for handling large scale language models.," in *Interspeech*, pp. 1618–1621, 2008.

[67] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[68] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various mfcc implementations on the speaker verification task," in *Proceedings of the SPECOM*, vol. 1, pp. 191–194, 2005.

[69] M. Slaney, "Auditory toolbox version 2. interval research corporation," *Indiana: Purdue University*, vol. 2010, pp. 1998–010, 1998.

[70] M. Unser, "On the approximation of the discrete Karhunen-Loève transform for stationary processes," *Signal Processing*, vol. 7, no. 3, pp. 231–249, 1984.

[71] K. K. Paliwal, "Decorrelated and liftered filter-bank energies for robust speech recognition.," in *Eurospeech*, vol. 99, pp. 85–88, 1999.

[72] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 1, pp. 52–59, 1986.

[73] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005. online web resource: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/.

[74] R. Schwartz, T. Anastasakos, F. Kubala, J. Makhoul, L. Nguyen, and G. Zavaliagkos, "Comparative experiments on large vocabulary speech recognition," in *Proceedings of the workshop on Human Language Technology*, pp. 75–80, Association for Computational Linguistics, 1993.

[75] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 4784–4787, IEEE, 2011.

[76] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[77] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Pearson/Prentice Hall Upper Saddle River, NJ, 2005.

[78] S. A. Gelfand, *Hearing: An introduction to psychological and physiological acoustics*. CRC Press, 2009.

[79] Y. Suzuki and H. Takeshima, "Equal-loudness-level contours for pure tones," *The Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 918–933, 2004.

[80] ISO, "226: 2003: Acoustics–normal equal-loudness-level contours," *International Organization for Standardization*, 2003.

[81] S. S. Stevens, "On the psychophysical law.," *Psychological review*, vol. 64, no. 3, p. 153, 1957.

[82] D. W. Ricker, *Echo signal processing*, vol. 725. Springer Science & Business Media, 2012.

[83] H. Hermansky and N. Morgan, "Rasta processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, 1994.

[84] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[85] C. L. Smith, C. P. Browman, R. S. McGowan, and B. Kay, "Extracting dynamic parameters from speech movement data," *The Journal of the Acoustical Society of America*, vol. 93, no. 3, pp. 1580–1588, 1993.

[86] S. S. Haykin, *Neural networks and learning machines*, vol. 3. Pearson Education Upper Saddle River, 2009.

[87] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[88] P. F. Brown, *The acoustic-modeling problem in automatic speech recognition*. PhD thesis, Computer Science Department, Carnegie-Mellon University, 1987.

[89] N. Kumar and A. G. Andreou, *Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition*. PhD thesis, Johns Hopkins University, 1997.

[90] T. Hastie and R. Tibshirani, "Discriminant analysis by gaussian mixtures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 155–176, 1996.

[91] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition," *Speech communication*, vol. 26, no. 4, pp. 283–297, 1998.

[92] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[93] M. J. Gales, "Semi-tied covariance matrices for hidden markov models," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 272–281, 1999.

[94] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 2, pp. II1129–II1132, IEEE, 2000.

[95] D. Dimitriadis, A. Potamianos, and P. Maragos, "A comparison of the squared energy and teager-kaiser operators for short-term energy estimation in additive noise," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2569–2581, 2009.

[96] J. F. Kaiser, "On a simple algorithm to calculate theenergy'of a signal," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 381–384, 1990.

[97] J. F. Kaiser, "On teager's energy algorithm and its generalization to continuous signals," in *Proc. 4th IEEE digital signal processing workshop*, 1990.

[98] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *Signal Processing, IEEE Transactions on*, vol. 41, no. 10, pp. 3024–3051, 1993.

[99] S. Haykin and M. Moher, *Communication systems*. John Wiley & Sons, 2009.

[100] H. Tolba and D. O'Shaughnessy, "Automatic speech recognition based on cepstral coefficients and a mel-based discrete energy operator," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, pp. 973–976, IEEE, 1998.

[101] F. Jabloun, E. Erzin, *et al.*, "Teager energy based feature parameters for speech recognition in car noise," *Signal Processing Letters, IEEE*, vol. 6, no. 10, pp. 259–261, 1999.

[102] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 196–200, 2001.

[103] A. Potamianos and P. Maragos, "Applications of speech processing using an am–fm modulation model and energy operators," in *Proc. Eur. Signal Processing Conf*, pp. 1669–1672, Citeseer, 1994.

[104] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory teager energy cepstrum coefficients for robust speech recognition.," in *INTERSPEECH*, pp. 3013–3016, 2005.

[105] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Classification of stress in speech using linear and nonlinear features," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2, pp. II–9–II–12, IEEE, 2003.

[106] N. Nehe and R. Holambe, "Power spectrum difference teager energy features for speech recognition in noisy environment," in *Industrial and Information Systems, 2008. ICIIS 2008. IEEE Region 10 and the Third international Conference on*, pp. 1–5, IEEE, 2008.

[107] N. S. Nehe and R. S. Holambe, "Isolated word recognition using normalized teager energy cepstral features," in *Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on*, pp. 106–110, IEEE, 2009.

[108] A. Georgogiannis and V. Digalakis, "Speech emotion recognition using non-linear teager energy based features in noisy environments," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pp. 2045–2049, IEEE, 2012.

[109] P. Maragos and A. C. Bovik, "Image demodulation using multidimensional energy separation," *JOSA A*, vol. 12, no. 9, pp. 1867–1876, 1995.

[110] D. Dimitriadis, P. Maragos, and A. Potamianos, "On the effects of filterbank design and energy computation on robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1504–1516, 2011.

[111] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction.," in *INTERSPEECH*, pp. 28–31, 2009.

[112] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4574–4577, IEEE, 2010.

[113] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients for robust speech recognition," *IEEE Transactions on Speech , Audio, and Language Processing*, 2015. (to appear).

[114] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *Signal Processing, IEEE Transactions on*, vol. 41, no. 4, pp. 1532–1550, 1993.

[115] A. Potamianos and P. Maragos, "A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, no. 1, pp. 95–120, 1994.

[116] A. C. Bovik, P. Maragos, and T. F. Quatieri, "Am-fm energy detection and separation in noise using multiband energy operators," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3245–3265, 1993.

[117] D. Dimitriadis and P. Maragos, "Continuous energy demodulation methods and application to speech analysis," *Speech communication*, vol. 48, no. 7, pp. 819–837, 2006.

[118] I. Kokkinos, G. Evangelopoulos, and P. Maragos, "Texture analysis and segmentation using modulation features, generative models, and weighted curve evolution," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 142–157, 2009.

[119] T. Ezzat, J. Bouvrie, and T. Poggio, "Am-fm demodulation of spectrograms using localized 2d max-gabor analysis," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–1061, IEEE, 2007.

[120] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition.," in *INTERSPEECH*, Citeseer, 2003.

[121] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.

[122] M. R. Schädler and B. Kollmeier, "Separable spectro-temporal gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2047–2059, 2015.

[123] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of neurophysiology*, vol. 85, no. 3, pp. 1220–1234, 2001.

[124] A. Qiu, C. E. Schreiner, and M. A. Escabí, "Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition," *Journal of Neurophysiology*, vol. 90, no. 1, pp. 456–476, 2003.

[125] P. Tsiakoulis and A. Potamianos, "Statistical analysis of amplitude modulation in speech signals using an am-fm model," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3981–3984, IEEE, 2009.

[126] Y. Wang, J. Hansen, G. K. Allu, and R. Kumaresan, "Average instantaneous frequency (aif) and average log-envelopes (ale) for asr with the aurora 2 database.," in *INTERSPEECH*, 2003.

[127] K. Kuldip and S. Bishnu, "Frequency-related representation of speech," EUROSPEECH Seminar, 2003.

[128] V. Tyagi and C. Wellekens, "Fepstrum representation of speech signal," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pp. 11–16, IEEE, 2005.

[129] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4117–4120, IEEE, 2012.

[130] V. Mitra, H. Franco, M. Graciarena, and D. Vergyri, "Medium-duration modulation cepstral feature for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 1749–1753, IEEE, 2014.

[131] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust am-fm features for speech recognition," *Signal Processing Letters, IEEE*, vol. 12, no. 9, pp. 621–624, 2005.

[132] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3795–3806, 1996.

[133] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Short-time instantaneous frequency and bandwidth features for speech recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pp. 103–106, IEEE, 2009.

[134] T. Chaspari, D. Dimitriadis, and P. Maragos, "Emotion classification of speech using modulation features," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pp. 1552–1556, IEEE, 2014.

[135] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust asr," *Signal Processing Letters, IEEE*, vol. 17, no. 6, pp. 551–554, 2010.

[136] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Instantaneous frequency and bandwidth estimation using filterbank arrays," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8032–8036, IEEE, 2013.

[137] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[138] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[139] R. D. Patterson, K. Robinson, I. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception: Proceedings of the 9th International Symposium on Hearing Held in Carcens, France on 9-14 June 1991*, no. 83, p. 429, Pergamon, 1992.