ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
Μεταπτυχιακό Πρόγραμμα Σπουδών

Διονυσίου Λιναρδάτου (Α.Μ. 613)

Εύρωστη Αναγνώριση Ομιλητή

Εργασία στο μάθημα: Επικοινωνία με Ομιλία
Διδάσκων: Γεώργιος Κουρουπέτρογλου

Αθήνα 1999

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Ενα σύστημα αναγνώρισης ομιλητή έχει σκοπό να αναγνωρίσει έναν ομιλητή από τη φωνή του, ενώ ένα σύστημα αναγνώρισης ομιλίας έχει σκοπό να κατανοήσει το τί λέγεται. Στην παρούσα εργασία επικεντρώνουμε το ενδιαφέρον μας στην αναγνώριση ομιλητή. Η κατασκευή ενός συστήματος αναγνώρισης ομιλητή περιλαμβάνει τρία στάδια: το στάδιο εκμάθησης, το στάδιο ελέγχου και το στάδιο υλοποίησης. Απαιτεί ιδιαίτερη επιμέλεια ώστε να ελαχιστοποιηθεί η σημαντική μείωση της απόδοσής του, φαινόμενο που συναντάται συχνά στην πράξη όταν το σύστημα λειτουργεί σε πραγματικές συνθήκες. Η μείωση της απόδοσης οφείλεται στο ότι οι συνθήκες εκμάθησης δεν ταιριάζουν με τις συνθήκες ελέγχου. Για το σκοπό αυτό κατά την κατασκευή του συστήματος εφαρμόζουμε εύρωστες στρατηγικές που αντιμετωπίζουν τα προβλήματα παραμορφώσεων της ομιλίας λόγω της επίδρασης του καναλιού και του θορύβου από διαφορετικά περιβάλλοντα.

Η κατασκευή ενός συστήματος αναγνώρισης ομιλητή ξεκινά με την εφαρμογή ανάλυσης γραμμικής πρόβλεψης της ομιλίας, παραγωγή των συντελεστών πρόβλεψης και μετασχηματισμό των συντελεστών αυτών σε διανύσματα χαρακτηριστικών μεγεθών. Το γραμμικό μοντέλο παραγωγής της ομιλίας θεωρεί ότι ο γλωττιδικός παλμός, η φωνητική οδός και η ακτινοβόληση του αέρα από τα χείλη μοντελοποιούνται το καθένα ξεχωριστά ως γραμμικά φίλτρα. Ο συνδυασμός των φίλτρων αυτών οδηγεί στην πρόβλεψη κάθε δείγματος της ομιλίας από το γραμμικό συνδυασμό προηγούμενων δειγμάτων της. Οι συντελεστές του γραμμικού συνδυασμού υπολογίζονται συνήθως με τη μέθοδο της αυτοσυσχέτισης. Η απλούστερη επιλογή διανύσματος χαρ/κών μεγεθών είναι τα σταθμισμένα πρώτα στοιχεία του cepstrum της γραμμικής συνάρτησης μεταφοράς ή η χρήση της πρώτης παραγώγου του. Η επίδραση του καναλιού αντισταθμίζεται με αφαίρεση της μέσης τιμής του cepstrum. Περισσότερο εύρωστη αντιστάθμιση επιτυγχάνουμε εάν ακολουθήσουμε τη διαδικασία φιλτραρίσματος των πόλων της γραμμικής πρόβλεψης ή τη μέθοδο της προσαρμοστικής στάθμισης ή τη μέθδο του μεταφιλτραρισμένου cepstrum.

Εξετάζοντας το θέμα της παραμόρφωσης της ομιλίας από μαθηματική άποψη διαπιστώνουμε ότι οι συντελεστές του προβλέπτη (κατά συνέπεια και τα cepstral διανύσματα) περιστρέφονται, μεταφέρονται και αλλάζουν κλίμακα, όταν το σήμα ομιλίας παραμορφώνεται από το κανάλι μετάδοσης ή αλλοιώνεται από το θόρυβο. Συγκεκριμένα η πρόσθεση λευκού θορύβου διατηρεί το γενικό προσανατολισμό του διανύσματος των συντελεστών του προβλέπτη, αλλά το συρρικνώνει μετακινώντας το πλησιέστερα στην αρχή των αξόνων. Η διέλευση της ομιλίας από ένα συγκεραστικό γραμμικό κανάλι ισοδυναμεί με γραμμικό μετασχηματισμό του διανύσματος των συντελεστών του προβλέπτη χωρίς καμμία μεταφορά. Η παρεμβολή ενός άλλου ομιλητή στο ίδιο κανάλι πάλι οδηγεί σε έναν affine μετασχηματισμό των συντελεστών του προβλέπτη. Η θεώρηση του affine μετασχηματισμού στην αντιμετώπιση των παραμορφώσεων επιτυγχάνει σχετικά καλύτερη απόδοση των συστημάτων αναγνώρισης ομιλητή σε χαμηλά SNRs, κάτι που ενθαρρύνει τη συνέχιση της έρευνας στην κατεύθυνση αυτή.

1. ΕΙΣΑΓΩΓΗ

Η μελλοντική εμπορική εκμετάλλευση της τεχνολογίας αναγνώρισης ομιλητή και αναγνώρισης ομιλίας παρεμποδίζεται από τη μεγάλη υποβάθμιση της απόδοσης ενός συστήματος αναγνώρισης λόγω των διαφορών μεταξύ των συνθηκών εκμάθησης και των συνθηκών ελέγχου. Αυτές οι διαφορές είναι γνωστές και ως *μη ταιριαστές συνθήκες*. Πολλά από τα σύγχρονα συστήματα [11] επιτυγχάνουν καλή απόδοση αναγνώρισης όταν οι συνθήκες κατά τη διάρκεια της εκμάθησης είναι παρόμοιες με τις αντίστοιχες κατά τη διάρκεια της λειτουργίας τους. Ομως συχνά, οι μη ταιριαστές συνθήκες είναι γεγονός που αναπόφευκτα συναντάται στην πράξη και υπό αυτές τις συνθήκες η απόδοση ενός συστήματος αναγνώρισης υποβαθμίζεται δραματικά. Ένα τυπικό παράδειγμα έχουμε όταν η εκμάθηση γίνεται σε καθαρή ομιλία και ο έλεγχος εφαρμόζεται σε ομιλία αλλοιωμένη με θόρυβο ή παραμορφωμένη από το κανάλι εκπομπής. Οι εύρωστες τεχνικές ομιλίας [2] προσπαθούν να διατηρήσουν την απόδοση ενός συστήματος επεξεργασίας ομιλίας κάτω από τέτοιες συνθήκες λειτουργίας που διαφέρουν από τις συνθήκες εκμάθησης.

Η παρούσα εργασία βασιζόμενη στη [42] παρουσιάζει μια ανασκόπηση των υπαρχόντων συστημάτων αναγνώρισης ομιλίας καθώς και των προβλημάτων που αυτά αντιμετωπίζουν κατά τη λειτουργία τους. Εστιάζεται στις μεθόδους βελτίωσης της απόδοσής τους, οι οποίες συνίστανται στην εξαγωγή των χαρακτηριστικών μεγεθών της ομιλίας. Πραγματεύεται την ανάλυση γραμμικής πρόβλεψης, το πρώτο βήμα εξαγωγής των χαρακτηριστικών μεγεθών και περιγράφει διάφορα εύρωστα cepstral χαρακτηριστικά που απορρέουν από τους συντελεστές της γραμμικής πρόβλεψης. Τέλος παρουσιάζεται ο affine μετασχηματισμός, ο οποίος περιγράφει την επίδραση των μη ταιριαστών συνθηκών που οφείλονται ταυτόχρονα και στην παραμόρφωση από το κανάλι και στην αλλοίωση από το θόρυβο.

2. ΑΝΑΣΚΟΠΗΣΗ

2.1 Αναγνώριση Ομιλίας και Αναγνώριση Ομιλητή

Ένα *σύστημα αναγνώρισης ομιλητή* προσπαθεί να αναγνωρίσει έναν ομιλητή από τη φωνή του. Το σύστημα μπορεί να είναι είτε εξαρτώμενο από κείμενο (όπου υπάρχει περιορισμός στο τί ομιλείται) είτε ανεξάρτητο από κείμενο (όπου δεν υπάρχει κανένας περιορισμός στο τί ομιλείται). Η φιλοσοφία του συστήματος έγκειται στην αναγνώριση των εγγενών διαφορών στα όργανα άρθρωσης, (για παράδειγμα στη δομή της φωνητικής οδού, στο μέγεθος της ρινικής κοιλότητας και στα χαρακτηριστικά των φωνητικών χορδών) καθώς και στον τρόπο της ομιλίας.

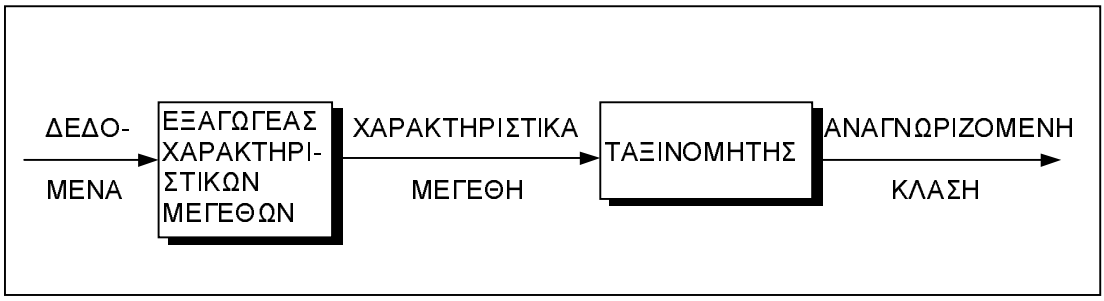
Αντίθετα, η *αναγνώριση ομιλίας* είναι η διαδικασία κατανόησης του τί λέγεται παρά του ποιος μιλάει. Πρώτα πρέπει να αναγνωριστεί η ροή των ήχων που περικλείουν την εισερχόμενη ομιλία. Στη συνέχεια ένα μοντέλο γλώσσας εφαρμόζεται στην ακολουθία των αναγνωρισμένων ήχων ώστε να βελτιωθεί η απόδοση με τη χρήση πληροφορίας από τα συμφραζόμενα.

2.2 Αναγνώριση Προτύπων

Η αναγνώριση ομιλητή και η αναγνώριση ομιλίας είναι υποσύνολα μιας ευρύτερης περιοχής γνωστής ως *αναγνώριση προτύπων*. Εχοντας δεδομένα τα χαρακτηριστικά μεγέθη που περιγράφουν τις ιδιότητες ενός αντικειμένου, ένα σύστημα αναγνώρισης προτύπων έχει ως σκοπό να αναγνωρίσει το αντικείμενο με βάση την προηγούμενη γνώση του για αυτό. Η κατασκευή ενός συστήματος αναγνώρισης προτύπων γενικά περιλαμβάνει τρία στάδια: *το στάδιο της εκμάθησης, το στάδιο του ελέγχου και το στάδιο της υλοποίησης*. Κατά το στάδιο της εκμάθησης υπολογίζεται ένα σύνολο από τις παραμέτρους ενός μοντέλου, ώστε κατά κάποιο τρόπο το μοντέλο να μαθαίνει την αντιστοιχία μεταξύ των χαρακτηριστικών μεγεθών και των κλάσεων των αντικειμένων. Ενα τέτοιο κριτήριο εκμάθησης είναι η ελαχιστοποίηση του ολικού σφάλματος εκτίμησης. Κατά το στάδιο του ελέγχου, οι παράμετροι του μοντέλου τροποποιούνται χρησιμοποιώντας ένα σύνολο από δεδομένα διασταύρωσης ώστε να επιτευχθεί μια καλή γενίκευση της απόδοσης του συστήματος. Τα δεδομένα διασταύρωσης συνήθως αποτελούνται από ένα σύνολο από χαρακτηριστικά μεγέθη και κλάσεις που είναι διαφορετικά από τα δεδομένα εκμάθησης. Η διαδικασία της αναγνώρισης εκτελείται στο στάδιο της υλοποίησης. Χαρακτηριστικά μεγέθη που ανήκουν σε άγνωστη κλάση διέρχονται από το σύστημα και αυτό αποφασίζει σε ποια κλάση ανήκουν.

Ενα σύστημα αναγνώρισης προτύπων αποτελείται από έναν εξαγωγέα χαρακτηριστικών μεγεθών και έναν ταξινομητή. Η βασική του δομή φαίνεται στο σχήμα 1. Ο εξαγωγέας χαρακτηριστικών κανονικοποιεί τα δεδομένα και τα μετασχηματίζει στο χώρο των χαρακτηριστικών μεγεθών. Στο χώρο των χαρακτηριστικών μεγεθών τα δεδομένα συμπιέζονται και αναπαρίστανται με τέτοιο αποτελεσματικό τρόπο, ώστε τα αντικείμενα της ίδιας κλάσης να συμπεριφέρονται παρόμοια και τα αντικείμενα από διαφορετικές κλάσεις να

διακρίνονται ξεκάθαρα μεταξύ τους. Ο ταξινομητής λαμβάνει τα χαρακτηριστικά μεγέθη που υπολογίστηκαν από τον εξαγωγέα χαρακτηριστικών και με βάση την ακολουθούμενη μεθοδολογία εφαρμόζει είτε ταίριασμα ιχνών είτε υπολογισμούς πιθανοφάνειας στα χαρακτηριστικά.



Σχήμα 1: Η δομή του συστήματος αναγνώρισης προτύπων

Ομως, πριν μπορέσει να χρησιμοποιηθεί για ταξινόμηση, ο ταξινομητής πρέπει να εκπαιδευθεί έτσι ώστε να επιτυγχάνει την αντιστοιχία των χαρακτηριστικών μεγεθών στην επισημείωση μιας συγκεκριμένης κλάσης. Η λειτουργία του ταξινομητή στηρίζεται στην υπόθεση ότι οι συνθήκες εκμάθησης και ελέγχου είναι συγκρίσιμες. Προβλήματα αναδεικνύονται όταν υπάρχει αναντιστοιχία μεταξύ των συνθηκών εκμάθησης και ελέγχου, η οποία γενικά παρατηρείται στις περισσότερες εφαρμογές. Για παράδειγμα, ας υποθέσουμε ότι η αναγνώριση ομιλητή εκτελείται στο τηλεφωνικό δίκτυο. Είναι πολύ πιθανό ότι ο ομιλητής ελέγχου καλεί από τηλέφωνο διαφορετικό από αυτό που χρησιμοποιούσε κατά τη διάρκεια της εκμάθησης. Εχουμε δηλαδή στην περίπτωση αυτή μη ταιριαστά κανάλια επικοινωνίας. Θα πρέπει να χρησιμοποιήσουμε εύρωστες τεχνικές κατά το σχεδιασμό του συστήματος ώστε να ξεπεράσουμε τα προβλήματα που δημιουργούν αυτές οι μη ταιριαστές συνθήκες.

2.3 Εύρωστες Τεχνικές Ομιλίας

Τα προβλήματα που δημιουργούνται από την επίδραση των καναλιών και του θορύβου αντιμετωπίζονται με δύο στρατηγικές εύρωστων τεχνικών ομιλίας. Η πρώτη στρατηγική έχει πεδίο εφαρμογής τον εξαγωγέα χαρακτηριστικών μεγεθών, δηλαδή πριν τα διανύσματα χαρακτηριστικών περάσουν στον ταξινομητή για σύγκριση και επισημείωση. Χαρακτηριστική μέθοδος της στρατηγικής αυτής είναι η ανάδειξη της ομιλίας με αφαίρεση φασμάτων [3]. Με τη μέθοδο αυτή τα χαρακτηριστικά μεγέθη είναι περισσότερο αντιπροσωπευτικά της καθαρής ομιλίας, αφού η επίδραση του θορύβου καταστέλλεται. Επίσης, εάν τα χαρακτηριστικά μεγέθη είναι τα cepstral διανύσματα, η αφαίρεση της μέσης τιμής [4] προσπαθεί να αφαιρέσει τις επιδράσεις του καναλιού διάδοσης. Ακόμη στην κατεύθυνση αυτή γίνονται προσπάθειες εύρεσης καινούριων χαρακτηριστικών μεγεθών που είναι εύρωστα στο θόρυβο και στις επιδράσεις του καναλιού (ένα παράδειγμα είναι η βραχέως χρόνου τροποποιημένη συμφωνία [5]).

Η δεύτερη στρατηγική στοχεύει στη δημιουργία ενός πιο εύρωστου ταξινομητή αντισταθμίζοντας τις παραμορφώσεις στο στάδιο ταξινόμησης.

Συνήθως για την περιγραφή των χαρακτηριστικών μεγεθών υιοθετούνται στατιστικές προσεγγίσεις με μοντέλα πιθανοτήτων. Αυτές καταλήγουν στη δημιουργία εύρωστης αντιστοιχίας από τα δεδομένα ελέγχου στα δεδομένα εκμάθησης. Μέθοδοι όπως το βέλτιστο πιθανοτικό φιλτράρισμα (Probabilistic Optimum Filtering) [6], το Γκαουσιανό Μοντέλο Ανάμιξης (GMM, Gaussian Mixture Model) [7], η προσαρμογή με κρυμμένα μοντέλα Markov (HMM: Hidden Markov Model) [8] ανήκουν σε αυτή την κατηγορία. Στο μέλλον εύρωστες μετρικές απόστασης, όπως το Itakura μέτρο φασματικής διασποράς [9, 10] και το μέτρο προβολής [11] αναμένεται να οδηγήσουν σε περισσότερο ακριβή επισημείωση των δεδομένων ελέγχου.

3. ΓΡΑΜΜΙΚΗ ΠΡΟΒΛΕΨΗ ΤΗΣ ΟΜΙΛΙΑΣ

Η διαδικασία εξαγωγής των χαρακτηριστικών μεγεθών διαιρείται σε δύο μέρη. Στο πρώτο εφαρμόζεται *ανάλυση γραμμικής πρόβλεψης* (ΓΠ) της ομιλίας και παράγεται ένα σύνολο συντελεστών πρόβλεψης. Στο δεύτερο μέρος οι συντελεστές πρόβλεψης μετασχηματίζονται σε διανύσματα χαρακτηριστικών μεγεθών. Στο κεφάλαιο αυτό παρουσιάζεται η λογική σύμφωνα με την οποία χρησιμοποιείται η ανάλυση ΓΠ, δίνονται μερικές ερμηνείες και παρατίθενται τα υπολογιστικά θέματα της ανάλυσης ΓΠ.

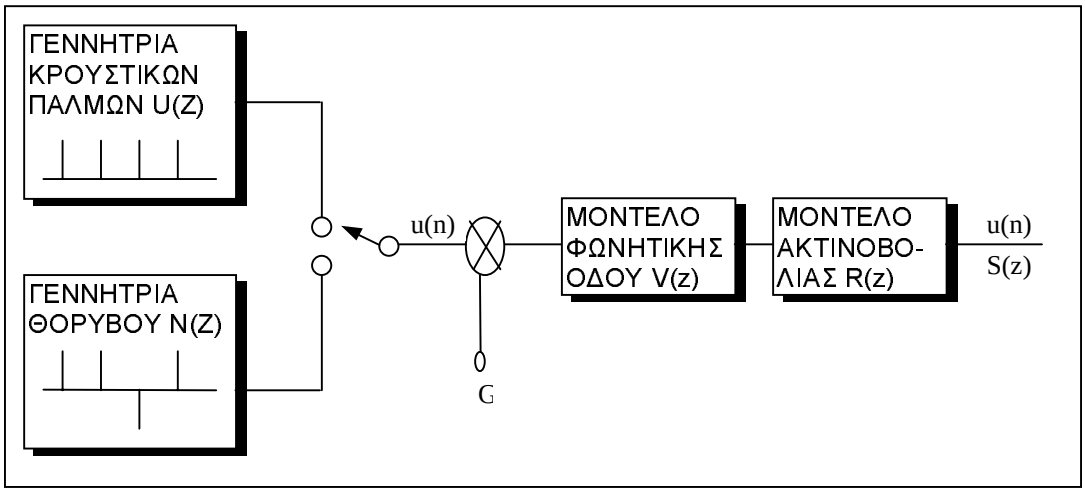
3.1 Αυτοπαλινδρούμενο Μοντέλο

Οι ήχοι ομιλίας μπορούν να ταξινομηθούν σε τρεις διακριτές κατηγορίες: στους έμφωνους ήχους, στους τυρβώδεις ή άφωνους ήχους και στους κλειστούς μη έρρινους ήχους. Η κυματομορφή ομιλίας είναι ένα ακουστικό κύμα πίεσης που προέρχεται από εκούσιες φυσιολογικές κινήσεις ανατομικών δομών, όπως οι φωνητικές χορδές, η φωνητική οδός, η ρινική κοιλότητα, η γλώσσα και τα χείλη [12, 13]. Η φωνητική οδός συνήθως μοντελοποείται ως η συρραφή ανομοιόμορφων, χωρίς απώλειες και μεταβλητής διατομής σωλήνων, που αρχίζει από τις φωνητικές χορδές και τελειώνει στα χείλη [13]. Το άνοιγμα των φωνητικών χορδών καλείται γλωττίδα. Εμφωνοι ήχοι όπως /i/ και /e/ παράγονται πιέζοντας τον αέρα διαμέσου της γλωττίδας. Η τάση των φωνητικών χορδών τροποποιείται έτσι ώστε να πάλλονται με σταθερή ταλάντωση και επομένως να διεγείρουν τη φωνητική οδό με ψευδοπεριοδικούς παλμούς αέρα. Οσο μεγαλύτερη είναι η τάση, τόσο πιο υψηλός ο μουσικός ήχος ή η βασική συχνότητα της φωνής. Οι άφωνοι ήχοι παράγονται κρατώντας εκούσια τις φωνητικές χορδές ανοικτές, σχηματίζοντας σύσφιξη με τη χρήση του αρθρωτή και πιέζοντας τον αέρα να διέλθει από τη σύσφιξη με αρκετά μεγάλη ταχύτητα ώστε να παραχθεί τύρβη. Η φωνητική οδός διεγείρεται από μία ευρείας ζώνης πηγή θορύβου κατά την παραγωγή των άφωνων ήχων. Οι κλειστοί μη έρρινοι ήχοι προκύπτουν από αύξηση της πίεσης του αέρα στο στόμα και απότομη απελευθέρωσή της.

Ενα γραμμικό μοντέλο παραγωγής της ομιλίας αναπτύχθηκε από τον Fant στα τέλη της δεκαετίας του 50 [14]. Σύμφωνα με αυτό ο γλωττιδικός παλμός, η φωνητική οδός και η ακτινοβόληση μοντελοποιούνται το καθένα ξεχωριστά ως γραμμικά φίλτρα. Ενα πλήρες μοντέλο παραγωγής ομιλίας στο χώρο του μετασχηματισμού z παρουσιάζεται στο σχήμα 2. Η πηγή είναι είτε ψευδοπεριοδική ακολουθία κρουστικών παλμών (για την περίπτωση των έμφωνων ήχων) ή μια τυχαία ακολουθία θορύβου (για την περίπτωση των άφωνων ήχων), με έναν παράγοντα απολαβής G για τον έλεγχο της έντασης της διέγερσης. Η συνάρτηση μεταφοράς V(z) για τη φωνητική οδό συσχετίζει την ταχύτητα όγκου στην πηγή με την ταχύτητα όγκου στα χείλη. Είναι γενικά ένα μοντέλο που έχει μόνο πόλους για τους περισσότερους ήχους της ομιλίας. Κάθε πόλος του V(z) αντιστοιχεί σε ένα φωνοσυντονισμό. Για τους έρρινους και τους τυρβώδεις ήχους που απαιτούν και συντονισμούς και αντισυντονισμούς (δηλαδή και πόλους και μηδενικά), επίσης προτιμούμε ένα μοντέλο που έχει μόνο πόλους επειδή η επίδραση ενός μηδενικού στη

συνάρτηση μεταφοράς μπορεί να επιτευχθεί συμπεριλαμβάνοντας περισσότερους πόλους [15]. Το μοντέλο ακτινοβόλησης R(z) περιγράφει την πίεση του αέρα στα χείλη, η οποία προσεγγίζεται από μία πρώτης τάξης οπισθοδιαφορά. Ο συνδυασμός του γλωττιδικού παλμού, της φωνητικής οδού και της ακτινοβόλησης δίνει μία μόνο συνάρτηση μεταφοράς με μόνο πόλους [13,14] που δίνεται από τη σχέση

$$H(z) = G(z)V(z)R(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \qquad (1)$$



Σχήμα 2: Το γραμμικό μοντέλο φωνητικής οδού για την παραγωγή ομιλίας

Με αυτή τη συνάρτηση μεταφοράς παίρνουμε μια εξίσωση διαφορών για τη σύνθεση των δειγμάτων ομιλίας της μορφής

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \qquad (2)$$

Δηλαδή η πρόβλεψη για το S(n) προσδιορίζεται από το γραμμικό συνδυασμό των προηγούμενων p δειγμάτων. Ετσι το μοντέλο παραγωγής ομιλίας αναφέρεται και ως *μοντέλο γραμμικής πρόβλεψης* (ΓΠ) ή ως αυτοπαλινδρούμενο μοντέλο.

3.2 Αναπαράσταση της Ομιλίας με Μοντέλο Διαμόρφωσης

Η συνάρτηση μεταφοράς στην εξίσωση (1) μπορεί να γραφεί στη μορφή

$$H(z) = \frac{1}{A(z)} = \prod_{i=1}^{p} \frac{1}{1 - z_i z^{-i}} = \sum_{i=1}^{p} \frac{r_i}{1 - z_i z^{-i}} \qquad (3)$$

όπου τα r_i αντιπροσωπεύουν τα υπόλοιπα και τα z_i τους πόλους της H(z). Οι πόλοι αναλύονται ως

$$z_i = \sigma_i e^{j\omega_i}, \quad i = 1,2,..,p \qquad (4)$$

όπου το ω_i αντιστοιχεί στην i-οστή κεντρική συχνότητα. Τα σ_i αναπαριστούν τα πλάτη των πόλων και η τιμή τους ανήκει στο διάστημα (0,1). Το εύρος ζώνης του i-οστού πόλου ορίζεται ως [16]

$$B_i = \frac{1}{\pi} \ln\left(\frac{1}{|z_i|}\right) = \frac{1}{\pi} \ln\left(\frac{1}{\sigma_i}\right) \qquad (5)$$

Επομένως το μοντέλο της φωνητικής οδού αντιστοιχεί στην αιτιατή κρουστική που δίνεται από τη σχέση

$$h(n) = \sum_{i=1}^{p} r_i z_i^n = \sum r_i \sigma_i^n e^{j\omega_i n} \qquad (6)$$

η οποία με τη σειρά της είναι ο τύπος της ομογενούς λύσης της εξίσωσης (2).

Συνεπώς το σήμα ομιλίας s(n) είναι ένα σήμα που αποτελείται από πολλές συνιστώσες. Εκφράζεται ως ο γραμμικός συνδυασμός εκθετικών σημάτων διαμορφομένων κατά πλάτος και κατά φάση, που προσδιορίζονται από το αυτοπαλινδρούμενο μοντέλο. Αυτή είναι μια ειδική περίπτωση του γενικού *μοντέλου διαμόρφωσης* για την ομιλία, το οποίο παρουσιάζεται στη [16]. Για κάθε συνιστώσα του σήματος ομιλίας υπάρχουν τρεις παράμετροι, τα r_i, σ_i και ω_i. Οι παράμετροι r_i και σ_i ορίζουν το κατά πλάτος διαμορφωμένο τμήμα, ενώ τα ω_i είναι οι παράμετροι για τη διαμόρφωση κατά φάση. Οταν ένας πόλος που χαρακτηρίζεται από την ω_i βρίσκεται πλησίον του μοναδιαίου κύκλου δηλώνει ένα φωνοσυντονισμό στην ω_i με ένα σχετικά μικρό εύρος ζώνης.

3.3 Υπολογιστικά Θέματα

Οι συντελεστές του προβλέπτη $\{a_i\}$ που περιγράφουν το αυτοπαλινδρούμενο μοντέλο πρέπει να υπολογιστούν από το σήμα ομιλίας. Ομως η ομιλία είναι χρονικά μεταβαλλόμενο σήμα, αφού η μορφή της φωνητικής οδού αλλάζει με το χρόνο. Ετσι στην πράξη ένα ακριβές σύνολο των συντελεστών του προβλέπτη καθορίζεται προσαρμοστικά για μικρά χρονικά διαστήματα (τυπικές τιμές 10ms ως 30ms), τα οποία ονομάζονται *πλαίσια*. Κατά τη διάρκεια ενός πλαισίου υποθέτουμε ότι δεν έχουμε χρονική μεταβολή. Η απολαβή G συνήθως αγνοείται για να πετύχουμε παραμετροποίηση ανεξάρτητη της έντασης του σήματος.

Η *μέθοδος της αυτοσυσχέτισης* και η *μέθοδος της διασποράς* είναι οι δύο τυποποιημένες μέθοδοι επίλυσης του προβλήματος υπολογισμού των συντελεστών του προβλέπτη [12, 17]. Και οι δύο προσεγγίσεις βασίζονται στην ελαχιστοποίηση της μέσης τετραγωνικής τιμής του σφάλματος e(n) που δίνεται από τη σχέση

$$e(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \qquad (7)$$

Οι μέθοδοι διαφέρουν ως προς τις λεπτομέρειες της υλοποίησης. Η μέθοδος της αυτοσυσχέτισης είναι υπολογιστικά πιο απλή από τη μέθοδο διασποράς και αντίθετα από αυτή θεωρεί ότι όλοι οι πόλοι της H(z) βρίσκονται εντός του μοναδιαίου κύκλου.

Στη συνέχεια εξετάξουμε τη μέθοδο αυτοσυσχέτισης. Ας θεωρήσουμε ότι το μέσο τετραγωνικό σφάλμα ελαχιστοποιείται σε ένα πλαίσιο N δειγμάτων και επιπλέον ότι τα δείγματα ομιλίας είναι εκ ταυτότητος μηδέν εκτός του πλαισίου που μας ενδιαφέρει. Η αυτοσυσχέτιση του σήματος s(n) ορίζεται ως

$$r_i(k) = \sum_{n-0}^{N-1-k} s(n)s(n+k) \qquad (8)$$

Αποδεικνύεται τότε ότι οι συντελεστές a_i του προβλέπτη προκύπτουν από την επίλυση των παρακάτω εξισώσεων

$$\begin{pmatrix} r_s(0) & r_s(1) & \cdots & r_s(p-1) \\ r_s(1) & r_s(0) & \cdots & r_s(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_s(p-1) & r_s(p-2) & \cdots & r_s(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} r_s(1) \\ r_s(2) \\ \vdots \\ r_s(p) \end{pmatrix} \qquad (9)$$

Ας ονομάσουμε R_S τον pxp Toeplitz πίνακα αυτοσυσχέτισης στο αριστερό μέλος της (9), α το διάνυσμα των συντελεστών του προβλέπτη και r_S το διάνυσμα των συντελεστών αυτοσυσχέτισης στο δεξιό μέλος της (9). Τότε η (9) γράφεται

$$R_S a = r_S \qquad (10)$$

και αν ο R_S είναι αντιστρέψιμος προκύπτει η λύση της

$$a = R_S^{-1} r_S \qquad (11)$$

Επειδή ο πίνακας R_S είναι Toeplitz, η επίλυση του συστήματος μπορεί να γίνει με τον αποδοτικό υπολογιστικά αλγόριθμο των Levinson - Durbin [13]. Στη συνέχεια αφού προσδιορίσουμε τη συνάρτηση μεταφοράς H(z), το πλάτος της απόκρισης $|H(e^{j\omega})|$ αναπαριστά τη φασματική περιβάλλουσα της ομιλίας.

Η λύση που προκύπτει για την πληροφορία της φωνητικής οδού που περιέχεται στην H(z) πρέπει να είναι εύρωστη είτε η ομιλία είναι καθαρή είτε είναι αλλοιωμένη από το θόρυβο ή/και τις επιδράσεις του καναλιού. Πρέπει δηλαδή οι συντελεστές του προβλέπτη να παραμένουν είτε αμετάβλητοι είτε να παρουσιάζουν πολύ μικρή μεταβολή όταν η ομιλία είναι αλλοιωμένη. Κατά συνέπεια τα χαρακτηριστικά μεγέθη θα είναι εύρωστα. Στη [18] παρουσιάζεται μια συγκριτική μελέτη των διαφόρων μεθοδολογιών για τον προσδιορισμό των συντελεστών του προβλέπτη που βασίζεται στην ελαχιστοποίηση διαφόρων αντικειμενικών συναρτήσεων. Πέρα των τυποποιημένων μεθόδων Γ.Π. της αυτοσυσχέτισης και της διασποράς εξετάζονται μέθοδοι που βασίζονται στο κριτήριο ελάχιστης απόλυτης τιμής, στο κριτήριο αναδρομικών βαρών ελαχίστων τετραγώνων και το κριτήριο ολικών ελαχίστων τετραγώνων. Βρέθηκε ότι η καλύτερη μέθοδος εξαρτάται από το είδος του υπάρχοντος θορύβου και επομένως μια γενικά αποδεκτή λύση δεν είναι γνωστή [18].

Μια άλλη προσπάθεια αναπαράστασης του φάσματος ομιλίας συνίσταται στην απόδοση μεγαλύτερης έμφασης σε εκείνες τις συχνότητες που παρουσιάζουν μεγαλύτερη ακουστική ευαισθησία. Αυτή η θεώρηση είναι γνωστή ως Αισθητή Γραμμική Πρόβλεψη (ΑΓΠ) [19]. Το πραγματικό φάσμα της ομιλίας (που λαμβάνεται από το DFT των δειγμάτων της ομιλίας) τροποποιείται με βάση τις αρχές της κρίσιμης ζώνης ακουστικής απόκρυψης και της άνισης ευαισθησίας της ανθρώπινης ακοής στις διαφορετικές συχνότητες [19]. Το τροποποιημένο φάσμα προσεγγίζεται από ένα αυτοπαλινδρούμενο μοντέλο για να υπολογιστεί το H(z). Οι τιμές αυτοσυσχέτισης λαμβάνονται από τον αντίστροφο DFT του τροποποιημένου φάσματος. Οι συντελεστές του προβλέπτη υπολογίζονται από την επίλυση της εξίσωσης (10). Εχει δειχθεί πρόσφατα ότι η τεχνική της ΑΓΠ είναι περισσότερο εύρωστη σε μερικά μη ταιριαστά περιβάλλοντα από ότι η τυποποιημένη μέθοδος για αναγνώριση ομιλίας μεγάλου λεξιλογίου. Στην παρούσα εργασία, χρησιμοποιούμε την τυποποιημένη μέθοδο αυτοσυσχέτισης ΓΠ για τον προσδιορισμό των συντελεστών του προβλέπτη.

4. ΕΥΡΩΣΤΗ CEPSTRAL ΑΝΑΛΥΣΗ

Το επόμενο βήμα είναι η μετατροπή των συντελεστών του προβλέπτη σε διανύσματα χαρακτηριστικών μεγεθών. Παραδείγματα τέτοιων διανυσμάτων περιλαμβάνουν [15] αυτούς τους ίδιους τους συντελεστές του προβλέπτη, τους cepstral συντελεστές και τις παραγώγους τους, ζεύγη φασματικών γραμμών (LSP), τους λογάριθμους λόγων περιοχών (LAR), τις συναρτήσεις περιοχών της φωνητικής οδού και την κρουστική απόκριση h(n) του φίλτρου H(z). Τα περισσότερα από τα παραπάνω διανύσματα χαρακτηριστικών έχουν κατά καιρούς μελετηθεί διεξοδικά. Εχει βρεθεί ότι οι cepstral συντελεστές δίνουν τα καλύτερα αποτελέσματα [21]. Επίσης οι παράγωγοι των cepstral συντελεστών περιέχουν την πληροφορία της ομιλίας που είναι βασική για εφαρμογές εξαρτώμενες από κείμενο. Πρόσφατα η χρησιμοποίηση των ζευγών φασματικών γραμμών άφησε υποσχέσεις [22]. Στην παρούσα εργασία η έμφαση δίνεται στα χαρακτηριστικά μεγέθη που σχετίζονται με το cepstrum. Ολα τα χαρακτηριστικά μεγέθη που σχετίζονται με cepstrum λαμβάνονται μετά την ανάλυση Γραμμικής Πρόβλεψης. Εξαίρεση αποτελεί το mel-warped cepstrum που λαμβάνεται από ανάλυση μιας τράπεζας φίλτρων [17], και παρουσιάζεται στην 4.8.

4.1 Cepstrum

Ας θεωρήσουμε ένα (όχι απαραίτητα αιτιατό) σήμα χ(n) του οποίου ο z-μετασχηματισμός X(z) υπάρχει και έχει περιοχή σύγκλισης στην οποία συμπεριλαμβάνεται ο μοναδιαίος κύκλος. Ας υποθέσουμε ότι ο $C(z) = \log X(z)$ αναπτύσσεται σε συγκλίνουσα δυναμοσειρά της οποίας επίσης η περιοχή σύγκλισης περιλαμβάνει το μοναδιαίο κύκλο. Το *cepstrum* ορίζεται ως ο αντίστροφος z-μετασχηματισμός του C(z) [23], δηλαδή

$$C(z) = \sum_n c(n) z^{-n} \qquad (12)$$

Ας σημειωθεί ότι το c(n) επίσης δεν είναι απαραίτητα αιτιατό. Συνεχίζουμε υποθέτοντας ότι ο X(z) είναι μια ρητή συνάρτηση του z, δηλαδή περιγράφεται πλήρως από τους πόλους, τα μηδενικά και την απολαβή. Τότε το cepstrum C(z) έχει τις ακόλουθες ιδιότητες:

1. Το δείγμα c(0) είναι ο φυσικός λογάριθμος της απολαβής.
2. Οι πόλοι και τα μηδενικά του X(z) εντός του μοναδιαίου κύκλου συνεισφέρουν μόνο στο αιτιατό μέρος του c(n) που αρχίζει από n=1.
3. Οι πόλοι και τα μηδενικά του X(z) εκτός του μοναδιαίου κύκλου συνεισφέρουν μόνο στο μη αιτιατό μέρος του c(n).
4. Το cepstrum είναι αιτιατό αν και μόνο αν ο X(z) είναι ελάχιστης φάσης.
5. Το cepstrum είναι μη αιτιατό αν και μόνο αν ο X(z) είναι μέγιστης φάσης.
6. Το cepstrum c(n) αποσβαίνει με ρυθμό $1/|n|$ όταν το n τείνει στο ∞ ή $-\infty$.
7. Το cepstrum έχει άπειρη διάρκεια είτε το χ(n) είναι πεπερασμένης είτε είναι άπειρης διάρκειας.

8. Αν το x(n) είναι πραγματικός αριθμός, τότε και το c(n) είναι πραγματικός αριθμός.

Ως ειδική περίπτωση του πιο γενικού X(z) ας θεωρήσουμε το φίλτρο ΓΠ H(z) που είναι ελάχιστης φάσης, έχει δηλαδή μόνο πόλους και προκύπτει από τη μέθοδο αυτοσυσχέτισης. Δεδομένου ότι όλοι οι πόλοι $z = z_i$ είναι εντός του μοναδιαίου κύκλου και η απολαβή είναι 1, το αιτιατό cepstrum ΓΠ $c_{lp}(n)$ της H(z) είναι [17, 23, 24]

$$c_{lp}(n) = \begin{cases} \frac{1}{n}\sum_{i=1}^{p} z_i^n & ,n > 0 \\ 0 & ,n \le 0 \end{cases} \qquad (13)$$

Μια αναδρομική σχέση μεταξύ του cepstrum ΓΠ και των συντελεστών του προβλέπτη δίνεται από τη σχέση [17]

$$c_{lp}(n) = a_n + \sum_{i=1}^{n-1} (\frac{i}{n}) c_{lp}(i) a_{n-i} \qquad (14)$$

Η χρήση αυτής της αναδρομικής σχέσης επιτρέπει τον αποδοτικό υπολογισμό του $c_{lp}(n)$ χωρίς τη χρήση κανενός είδους πολυωνυμικής παραγοντοποίησης. Εφόσον το $c_{lp}(n)$ είναι άπειρης διάρκειας, το διάνυσμα χαρακτηριστικών μεγεθών διάστασης p αποτελείται από τα στοιχεία $c_{lp}(1)$ μέχρι $c_{lp}(p)$. Οι p αυτές τιμές είναι οι πιο σημαντικές λόγω της απόσβεσης της ακολουθίας με την αύξηση του n. Ετσι παρόλο που δε λαμβάνουμε υπόψη στους υπολογισμούς τα επιπλέον των p δείγματα, η μέση τετραγωνική διαφορά μεταξύ δύο cepstral διανυσμάτων ΓΠ είναι προσεγγιστικά ίση με τη μέση τετραγωνική διαφορά μεταξύ του λογαριθμικού φάσματος των αντίστοιχων φίλτρων ΓΠ που έχουν μόνο πόλους [17]. Συνεπώς, επιτυγχάνουμε ένα καλό μέτρο της διαφοράς στην περιβάλλουσα του φάσματος των πλαισίων ομιλίας από τα οποία εξήχθησαν τα cepstral διανύσματα.

4.2 Cepstral Παράγωγοι

Το cepstrum ΓΠ αντιπροσωπεύει τις τοπικές φασματικές ιδιότητες ενός δεδομένου πλαισίου ομιλίας. Ομως δεν χαρακτηρίζει τις χρονικές ή μεταβατικές πληροφορίες σε μια ακολουθία πλαισίων ομιλίας. Σε εφαρμογές σχετικές με κείμενο, όπως η αναγνώριση ομιλίας και η εξαρτώμενη από κείμενο αναγνώριση ομιλητή, έχει διαπιστωθεί βελτιωμένη απόδοση με την εισαγωγή των cepstral παραγώγων στο χώρο των χαρακτηριστικών μεγεθών. Κι αυτό γιατί οι cepstral παράγωγοι συλλαμβάνουν τις μεταβατικές πληροφορίες της ομιλίας. Η πρώτη παράγωγος του ceptrsum (γνωστή και ως *δέλτα ceptrsum*) ορίζεται ως [17]

$$\frac{\partial c_{lp}(n,t)}{\partial t} = \Delta c_{lp}(n,t) \approx \mu \sum_{k=-K}^{K} k c_{lp}(n,t+k) \qquad (15)$$

όπου τα $c_{lp}(n,t)$ συμβολίζουν τους n-οστούς cepstral συντελεστές ΓΠ τη χρονική στιγμή t, μ είναι μια κατάλληλη σταθερά κανονικοποίησης και (2K+1) είναι το πλήθος των πλαισίων για τα οποία εκτελούνται οι υπολογισμοί. Το ceptrsum ΓΠ και το δέλτα ceptrsum μαζί έχουν χρησιμοποιηθεί για τη βελτίωση της απόδοσης της αναγνώρισης ομιλητή [25].

4.3 Cepstral Στάθμιση

Η βασική ιδέα στην οποία βασίζεται η *cepstral στάθμιση* είναι η επιμέτρηση της ευαισθησίας των χαμηλότερης τάξης cepstral συντελεστών στην ολική φασματική κλίση καθώς και της ευαισθησίας των υψηλότερης τάξης cepstral συντελεστών στο θόρυβο [17]. Η στάθμιση υλοποιείται πολλαπλασιάζοντας τα $c_{lp}(n)$ με ένα παράθυρο w(n) και χρησιμοποιώντας τα σταθμισμένα cepstrum ως το διάνυσμα των χαρακτηριστικών μεγεθών. Αυτή η διαδικασία απόδοσης βαρών είναι επίσης γνωστή και ως *λείανση*. Η πρώτη συνέπεια της λείανσης είναι η εξαγωγή ενός διανύσματος χαρακτηριστικών μεγεθών πεπερασμένης διάστασης από το $c_{lp}(n)$ που είναι άπειρης διάρκειας. Επίσης η προσεκτική επιλογή του w(n) οδηγεί σε εύρωστα αποτελέσματα.

Υπάρχουν διάφορα σχήματα στάθμισης που διαφέρουν ως προς τον τύπο του χρησιμοποιούμενου cepstral παραθύρου. Το απλούστερο είναι το ορθογώνιο παράθυρο που δίνεται από την σχέση

$$w(n) = \begin{cases} 1, & n = 1,2,\ldots,L \\ 0, & \alpha\lambda\lambda ου \end{cases} \qquad (16)$$

όπου L είναι το μέγεθος του παραθύρου. Το ορθογώνιο παράθυρο διατηρεί τα πρώτα L δείγματα που είναι τα πιο σημαντικά λόγω της φθίνουσας ιδιότητας. Αλλες μορφές του w(n) είναι η γραμμική στάθμιση (ή αλλιώς quefrency λείανση), όπου

$$w(n) = \begin{cases} n, & n = 1,2,\ldots,L \\ 0, & \alpha\lambda\lambda ου \end{cases} \qquad (17)$$

και η ζωνοπερατή λείανση (ΖΠΛ) [17,26] όπου

$$w(n) = \begin{cases} 1 + \frac{L}{2}\sin(\frac{n\pi}{L}) & n = 1,2,\ldots,L \\ 0 & \alpha\lambda\lambda ου \end{cases} \qquad (18)$$

Η γραμμική στάθμιση αποδίδει ως βάρος σε κάθε ξεχωριστό cepstral όρο το δείκτη του. Επομένως υποβαθμίζει το ρόλο των όρων χαμηλότερης τάξης. Η ΖΠΛ σταθμίζει μια cepstral ακολουθία με μια ημιτονική συνάρτηση,

έτσι ώστε να υποβαθμίζεται ο ρόλος των όρων χαμηλότερης και υψηλότερης τάξης. Ας σημειωθεί ότι τα σχήματα στάθμισης που περιγράψαμε είναι σταθερά με την έννοια ότι τα βάρη είναι συνάρτηση μόνο του cepstral δείκτη. Δεν έχουν καμμία ρητή εξάρτηση από τις στιγμιαίες μεταβολές του cepstrum, οι οποίες εισάγονται από διάφορες συνθήκες του περιβάλλοντος (όπως ο θόρυβος και η επίδραση του καναλιού).

4.4 Αφαίρεση της Cepstral Μέσης Τιμής

Ενα σήμα ομιλίας που εκπέμπεται μέσω ενός τηλεφωνικού δικτύου υφίσταται γραμμική παραμόρφωση που οφείλεται στην επίδραση του καναλιού. Αυτό απλά εκφράζεται ως T(z)=S(z)G(z), όπου ο όρος S(z) αντιστοιχεί στην αρχική καθαρή ομιλία, ο όρος G(z) αντιστοιχεί στο τηλεφωνικό κανάλι και ο όρος T(z) αντιστοιχεί στη φιλτραρισμένη ομιλία. Λογαριθμίζοντας παίρνουμε

$$\log T(z) = \log S(z) + \log G(z) \qquad (19)$$

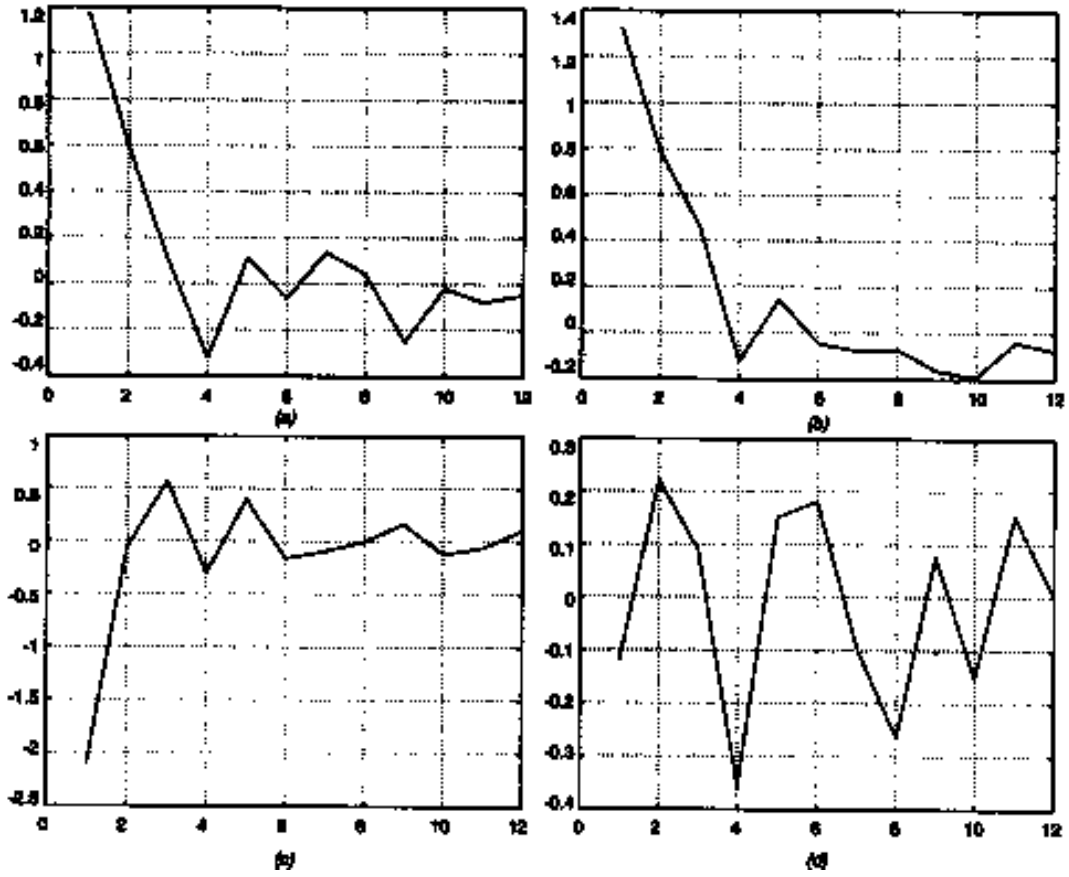
Ας υποθέσουμε ότι το φάσμα της ομιλίας και του καναλιού προσεγγίζονται καλά από μοντέλο ΓΠ που έχει μόνο πόλους. Τότε η επίδραση του καναλιού στην ομιλία προκαλεί έναν προσθετικό όρο στο cepstrum ΓΠ της καθαρής ομιλίας. Αν υποθέσουμε ότι το cepstrum ΓΠ της καθαρής ομιλίας είναι μηδέν, τότε η εκτίμηση του cepstrum του καναλιού είναι απλά και μόνο η μέση τιμή cepstrum ΓΠ της φιλτραρισμένης ομιλίας T(z). Ετσι για να αντισταθμίσουμε την επίδραση του καναλιού αρκεί να απομακρύνουμε την εκτίμησή του αφαιρώντας την cepstral μέση τιμή (CMS) [4, 21, 27]. Τότε το διάνυσμα χαρακτηριστικών μεγεθών γράφεται:

$$c_{cms}(n) = c_{lp}(n) - E[c_{lp}(n)] \qquad (20)$$

όπου η μέση τιμή λαμβάνεται σε ένα πλήθος πλαισίων ομιλίας που έχουν υποστεί την επίδραση του καναλιού. Ας σημειωθεί ότι έχουμε σιωπηρά υιοθετήσει την ορθογώνια στάθμιση (σχέση 16).

Η αφαίρεση της μέσης τιμής βελτιώνει σημαντικά την απόδοση ενός συστήματος στο οποίο η εκμάθηση γίνεται υπό ορισμένες συνθήκες για το κανάλι, ενώ ο έλεγχος κάτω από άλλες διαφορετικές συνθήκες. Ωστόσο η ακρίβεια της αναγνώρισης υφίσταται σημαντικές απώλειες όταν η CMS χρησιμοποιείται για την αναγνώριση ομιλητή στην περίπτωση που η εκμάθηση και ο έλεγχος γίνονται στο ίδιο κανάλι. Αυτό οφείλεται στην υπόθεση ότι η ασυμπτωτική cepstral μέση τιμή της καθαρής ομιλίας είναι μηδενική. Η υπόθεση έγινε ώστε το φάσμα του καναλιού να αναπαρίσταται από την ασυμπτωτική cepstral μέση τιμή της ομιλίας που εξήλθε από το κανάλι. Η υπόθεση αυτή αληθεύει μόνο όταν το τμήμα της ομιλίας φωνητικά ισορροπεί, δηλαδή συμπεριλαμβάνει την ίδια ποσότητα εμφώνων, αφώνων και κλειστών μη έρρινων ήχων. Αυτό γιατί οι τροχιές των cepstral συντελεστών για διαφορετικούς ήχους αποκλίνουν μεταξύ τους σημαντικά, ωστόσο συμπεριφέρονται παρόμοια με ήχους της ίδιας κατηγορίας. Το φαινόμενο παρατηρείται στο Σχήμα 3. Πάντως μπορούμε να υπολογίσουμε

ακριβέστερα το cepstrum του καναλιού, εάν η μέση τιμή του cepstrum η οποία οφείλεται αποκλειστικά στην καθαρή ομιλία, μπορούσε κατά μέγιστο ποσό να μειωθεί πριν από το συγκερασμό της ομιλίας με το κανάλι.

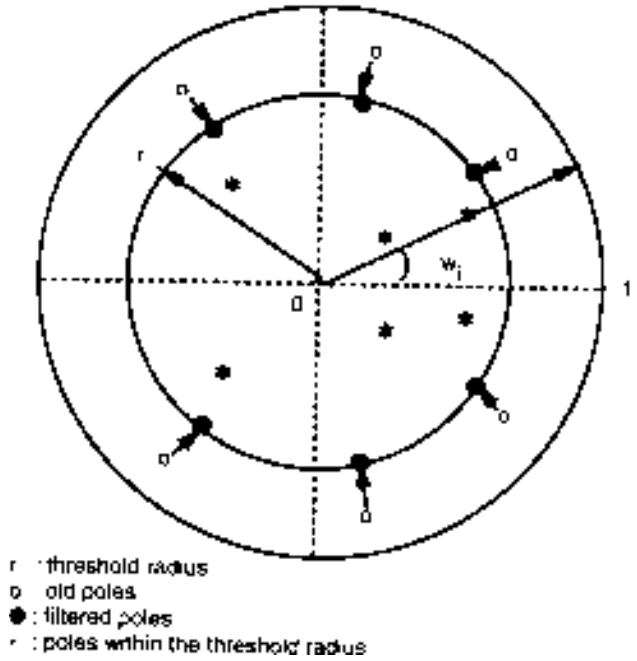


Σχήμα 3: Το cepstral διάνυσμα (α) του έμφωνου ήχου /α/ (β) του έμφωνου ήχου /u/ (γ) του άφωνου ήχου /sh/ και (δ) του κλειστού μη έρρινου ήχου /p/

4.5 Αφαίρεση της Cepstral Μέσης Τιμής Φιλτραρισμένων Πόλων

Η ιδέα της αφαίρεσης της cepstral μέσης τιμής βασίζεται στην υπόθεση ότι μια εκτίμηση του καναλιού δίνεται από τη μέση τιμή $E[c_{lp}(n)]$. Οι πόλοι της ΓΠ με στενό εύρος ζώνης που βρίσκονται πλησίον του μοναδιαίου κύκλου συνήθως αναπαριστούν τους φωνοσυντονισνούς και είναι λιγότερο ευαίσθητοι στις επιδράσεις καναλιού και θορύβου. Επομένως αυτοί οι πόλοι δε συνεισφέρουν στην εκτίμηση του καναλιού επειδή περιέχουν περισσότερη πληροφορία ομιλίας. Αντίθετα οι μεγάλου εύρους ζώνης πόλοι μοντελοποιούν τη φασματική κάμψη, την υπογλωττιδική διακύμανση και τις επιδράσεις του καναλιού. Αυτοί οι πόλοι προσφέρουν μια καλύτερη εκτίμηση του καναλιού. Μια καινούρια ιδέα γνωστή ως φιλτράρισμα πόλων τροποποιεί τους πόλους της ΓΠ έτσι ώστε να διευρύνει το εύρος ζώνης των πόλων των φωνοσυντονισμών [28]. Η διεύρυνση του εύρους ζώνης επιτυγχάνεται με την

ακτινική μεταφορά των πόλων των φωνοσυντονισμών μακριά από το μοναδιαίο κύκλο. Η συχνότητα των πόλων αφήνεται άθιγκτη. Το σχήμα 4 παρουσιάζει την ιδέα του φιλτραρίσματος των πόλων. Το cepstrum που



Σχήμα 4: Η ιδέα του φιλτραρίσματος πόλων

αντιστοιχεί σε αυτούς τους φιλτραρισμένους (ή τροποποιημένους) πόλους (συμβολίζεται ως $c_{mlp}(n)$) έχει λιγότερη πληροφορία ομιλίας και περισσότερη πληροφορία για το κανάλι από ότι έχει το $c_{lp}(n)$, ακριβώς λόγω της υποβάθμισης των πόλων των φωνοσυντονισμών. Η εκτίμηση του καναλιού δίνεται από την $E[c_{mlp}(n)]$ οπότε το διάνυσμα χαρακτηριστικών μεγεθών είναι

$$c_{pflcms}(n) = c_{lp}(n) - E[c_{mlp}(n)] \qquad (21)$$

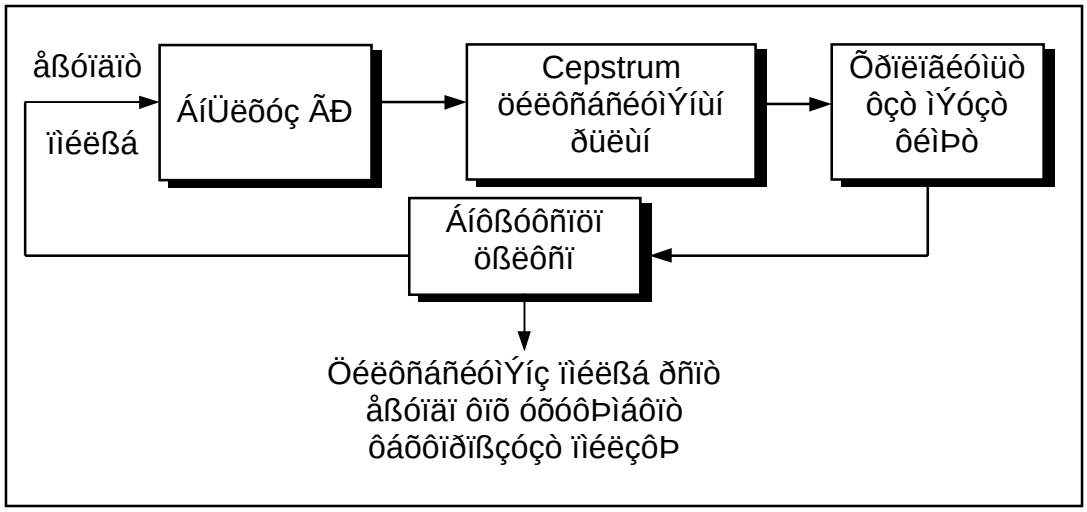
Ας σημειωθεί ότι έχουμε σιωπηρά υιοθετήσει την υπόθεση της ορθογώνιας στάθμισης (σχέση (16)).

Η παραπάνω τεχνική σχηματισμού του διανύσματος χαρακτηριστικών μεγεθών είναι γνωστή *ως αφαίρεση της cepstral μέσης τιμής φιλτραρισμένων πόλων* (PFCMS). Η υλοποίησή της περιλαμβάνει τα ακόλουθα βήματα:

- Επιλογή της ακτίνας κατωφλίου r_{th}
- Για κάθε πλαίσιο ομιλίας
  - Υπολογισμός πόλων ΓΠ $z_i, \forall i = 1,2,\ldots,p$
  - Για κάθε πόλο z_i:
    Εάν $|z_i| > r_{th}$, τροποποίηση z_i έτσι ώστε το πλάτος του να γίνει r_{th} και η φάση του να παραμείνει αμετάβλητη.
  - Υπολογισμός $c_{mlp}(n)$ βασισμένων στους τροποποιημένους ή φιλτραρισμένους πόλους.
- Υπολογισμός της εκτίμησης καναλιού $E[c_{mlp}(n)]$ για όλα τα πλαίσια ομιλίας.

- Υπολογισμός $c_{plcms}(n)$

Εχει δειχθεί ότι η PFCMS υπερισχύει στα πειράματα ταυτοποίησης ομιλητή [28]. Ας σημειωθεί ότι αφού η $E[c_{mlp}(n)]$ είναι μια καλή εκτίμηση του καναλιού, μπορεί να μετατραπεί σε ένα φίλτρο με μόνο πόλους που αναπαριστά το κανάλι. Επομένως μπορεί να μετατραπεί σε ένα αντίστροφο φίλτρο πεπερασμένης κρουστικής απόκρισης (FIR). Οταν αυτό το FIR φίλτρο εφαρμόζεται στην ομιλία, η ομιλία αναδεικνύεται καθώς η επίδραση του καναλιού απαλύνεται. Αυτή η ανάδειξη της ομιλίας πριν από την εξαγωγή των χαρακτηριστικών μεγεθών βελτιώνει την απόδοση της ταυτοποίησης του ομιλητή [29]. Το σχήμα 5 παρουσιάζει το μπλοκ διάγραμμα της διαδικασίας.



Σχήμα 5: Αντίστροφο φιλτράρισμα της επίδρασης καναλιού

4.6 Προσαρμοστική Στάθμιση των Cepstral Ορων

Οι τεχνικές της CMS και PFCMS είναι παραδείγματα τεχνικών επεξεργασίας μεταξύ πλαισίων με την έννοια ότι για τον προσδιορισμό του διανύσματος χαρακτηριστικών μεγεθών χρησιμοποιείται πληροφορία από περισσότερα του ενός πλαίσια. Σε αυτήν και την επόμενη ενότητα περιγράφονται δύο ενδοπλαισιακές μέθοδοι, δηλαδή μέθοδοι στις οποίες η πληροφορία λαμβάνεται μόνο από το τρέχον πλαίσιο. Η πρώτη είναι γνωστή ως *προσαρμοστική στάθμιση όρων* (ACW) [30].

Ας θεωρήσουμε ότι η συνάρτηση μεταφοράς ΓΠ H(z) παραμετροποιείται από τα υπόλοιπα r_k και τους πόλους z_k, οι οποίοι με τη σειρά τους περιγράφονται από τα σ_k και ω_k (σχέσεις (3) και (4)). Ετσι μπορούμε να γράψουμε την κρουστική απόκριση h(n) ως

$$\begin{pmatrix} h(1) \\ h(2) \\ \vdots \\ h(p) \end{pmatrix} = \begin{pmatrix} \sigma_1 e^{j\omega_1} & \sigma_2 e^{j\omega_2} & \cdots & \sigma_p e^{j\omega_p} \\ \sigma_1^2 e^{j2\omega_1} & \sigma_2^2 e^{j2\omega_2} & \cdots & \sigma_p^2 e^{j2\omega_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^p e^{jp\omega_1} & \sigma_2^p e^{jp\omega_2} & \cdots & \sigma_p^p e^{jp\omega_p} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{pmatrix} \qquad (22)$$

Οι φωνοσυντονισμοί του σήματος ομιλίας σταθμίζονται από τα υπόλοιπα r_i. Εχει παρατηρηθεί [30] ότι τα υπόλοιπα παρουσιάζουν αξιόλογη διακύμανση όταν η ομιλία διέρχεται από το κανάλι. Αυτό ισοδύναμα σημαίνει ότι τα πλάτη r_k των διακριτών τρόπων του μοντέλου διαμόρφωσης (εξίσωση (6)) διαταράσσονται από το κανάλι περισσότερο από ό,τι διαταράσσονται οι άλλες δύο παράμετροι σ_k και ω_k. Το ACW cepstrum απομακρύνει τις διακυμάνσεις που προκαλούνται από τη μεταβλητότητα του καναλιού κανονικοποιώντας τα υπόλοιπα. Ετσι δίνεται έμφαση στους όρους στενής ζώνης που αντιστοιχούν στους φωνοσυντονισμούς και καταστέλλονται οι όροι ευρείας ζώνης. Συνεπώς καταλήγουμε σε μια συνάρτηση μεταφοράς της μορφής

$$H_{acw}(z) = \frac{N(z)}{A(z)} = \sum_{k=1}^{p} \frac{1}{1 - z_k z^{-1}} \qquad (23)$$

όπου

$$N(z) = \sum_{k=1}^{p} \prod_{i=1 \neq k}^{p} (1 - z_i z^{-1}) \qquad (24)$$

Ο αριθμητής N(z) μπορεί να γραφεί και ως

$$N(z) = p(1 - \sum_{k=1}^{p-1} b_k z^{-1}) \qquad (25)$$

Αποδεικνύεται ότι το N(z) είναι ελάχιστης φάσης [31]. Επομένως το ACW ceptrum είναι αιτιατό και δίνεται από το $c_{acw}(0) = \log p$ και τη σχέση

$$c_{acw}(n) = c_{lp}(n) - c_{nn}(n) \quad \gamma\iota\alpha \quad n > 0 \qquad (26)$$

όπου οι όροι $c_{nn}(n)$ μπορούν να βρεθούν από μια αναδρομική σχέση που συμπεριλαμβάνει τους συντελεστές b_k (ίδιου τύπου αναδρομική σχέση όπως η εξίσωση 14). Επιπλέον, αφού τα b_k απλά εκφράζονται ως [31]

$$b_k = \frac{p-k}{p} a_k \quad \gamma\iota\alpha \quad 1 \leq k \leq p-1 \qquad (27)$$

ο υπολογισμός του ACW spectrum είναι πολύ απλός. Ο όρος $c_{nn}(n)$ αντιστοιχεί στην εκτίμηση του καναλιού. Αντίθετα από τα CMS και PFCMS, η μέθοδος είναι προσαρμοστική σε πλαίσιο-προς-πλαίσιο βάση. Στην πράξη, εφαρμόζεται η ορθογώνια στάθμιση έτσι ώστε το διάνυσμα χαρακτηριστικών μεγεθών να αποτελείται από τους όρους του $c_{acw}(n)$ για n=1 έως p. Η απόδοση των συστημάτων ταυτοποίησης ομιλητή είναι οπωσδήποτε καλύτερη αν χρησιμοποιείται το ACW cepstrum αντί για την απλή περίπτωση του cepstrum ΓΠ [30, 32].

4.7 Μεταφιλτραρισμένο Cepstrum

Η ιδέα ενός μεταφίλτρου (postfilter) εισήχθει στη [33] για την ανάδειξη της ομιλίας όταν είναι αλλοιωμένη με θόρυβο. Η φιλοσοφία της ανάπτυξης ενός μεταφίλτρου βασίζεται στο γεγονός ότι ο θόρυβος μπορεί να είναι περισσότερος ανεκτός από τις ανθρώπινες αισθήσεις όταν βρίσκεται στις περιοχές των φωνοσυντονισμών (φασματικές κορυφές) και λιγότερο ανεκτός όταν βρίσκεται στις φασματικές κοιλάδες. Το μεταφίλτρο λαμβάνεται από το A(z) και η συνάρτηση μεταφοράς του δίνεται από τη σχέση

$$H_{pfl}(z) = \frac{A(z/\beta)}{A(z/\alpha)} \quad o < \beta < \alpha \le 1 \qquad (28)$$

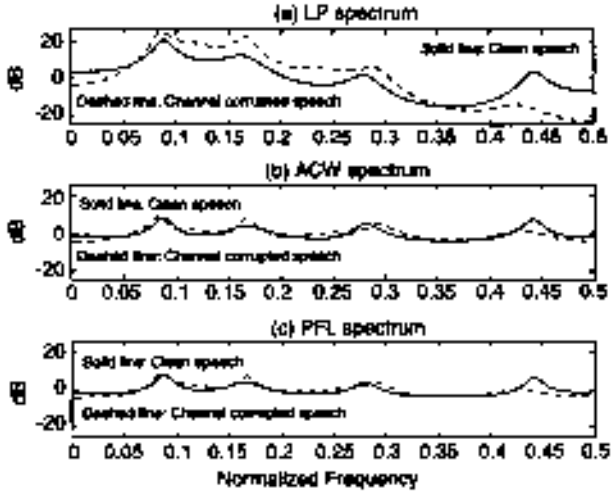
Εάν το A(z) είναι ελάχιστης φάσης, τότε το $H_{pfl}(z)$ είναι ελάχιστης φάσης. Επομένως το *μεταφιλτραρισμένο cepstrum* (που αναφέρεται σε συντομία και ως PFL cepstrum) [32] είναι αιτιατό και δίνεται από το $c_{pfl}(0) = 0$ και τη σχέση

$$c_{pfl}(n) = c_{lp}(n)[\alpha^n - \beta^n] \quad \gamma\iota\alpha \quad n > 0 \qquad (29)$$

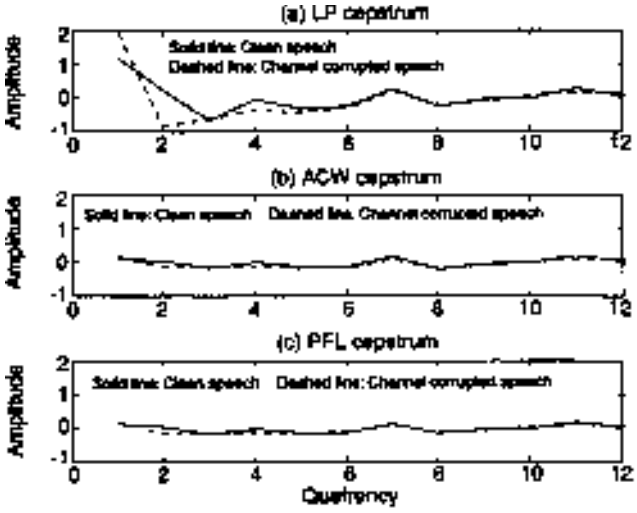
Το PFL cepstrum είναι απλά μια στάθμιση ή λείανση του cepstrum ΓΠ και είναι πολύ εύρωστο στις επιδράσεις καναλιού και θορύβου [32]. Οπως και στις άλλες μεθόδους λείανσης του cepstrum ΓΠ, δηλαδή της λείανσης ζώνης [26], της quefreny λείανσης [34] και της αντίστροφης λείανσης διακυμάνσεων [35], οι cepstral συντελεστές με τους μικρότερους δείκτες υποβαθμίζονται. Στην περίπτωση που α=1, τότε $c_{pfl}(n) = c_{lp}(n) - \beta^n c_{lp}(n)$. Υπάρχει δηλαδή ένας αφαιρετικός όρος που εξυπηρετεί ως εκτίμηση καναλιού και που είναι προσαρμοστικός σε βάση πλαίσιο-προς-πλαίσιο. Στην πράξη εφαρμόζεται ορθογώνια στάθμιση έτσι ώστε το διάνυσμα χαρακτηριστικών μεγεθών αποτελείται από τους όρους του $c_{pfl}(n)$ από n=1 έως p.

Το σχήμα 6 παρουσιάζει τις αποκρίσεις πλάτους των H(z), $H_{acw}(z)$ και $H_{pfl}(z)$ για ένα πλαίσιο καθαρής ομιλίας και για το ίδιο πλαίσιο ομιλίας αλλοιωμένο από το Continent Mid Voice (CMV) κανάλι [36]. Το CMV κανάλι είναι ένα τυπικό ζωνοπερατό κανάλι που συναντάται στα τηλεφωνικά δίκτυα. Λόγω της επίδρασης του φιλτραρίσματος από το κανάλι, υπάρχει σημαντική διαφορά στο φάσμα του 1/A(z), όπως αποκαλύπτει το σχήμα 6(α). Αυτή η διαφοροποίηση απαλύνεται σημαντικά με την εισαγωγή των $H_{acw}(z)$ και $H_{pfl}(z)$. Οπως φαίνεται και από τα σχήματα (6β) και (6γ), η διαφορά στο πλάτος του φάσματος για τις μεθόδους ACW και PFL μειώνεται ως προς αυτό του 1/A(z). Τα ACW και PFL φάσματα είναι παρόμοια αφού και τα δύο δίνουν έμφαση στις κορυφές των φωνοσυντονισμών που είναι πιο κρίσιμες για την ταυτοποίηση ομιλητή. Επίσης δεν υπάρχει φανερή φασματική κλίση. Ωστόσο το PFL φάσμα είναι ευαίσθητο στις μεταβολές των παραμέτρων α και β. Μείωση του α προκαλεί διεύρυνση του φωνοσυντονισμού, ενώ οι μεταβολές

στο β επιδρούν στη φασματική κλίση. Καθώς το β μειώνεται, η φασματική κλίση γίνεται περισσότερο φανερή.



Σχήμα 6: Πλάτη διαφόρων φασμάτων ομιλίας αλλοιωμένης από το CMV κανάλι (καθαρή ομιλία=συνεχόμενη γραφή, ομιλία αλλοιωμένη από το κανάλι=εστιγμένη γραμμή) (α) Απόκριση πλάτους του 1/A(z), (β) Απόκριση πλάτους του $H_{acw}(z)$ (γ) Απόκριση πλάτους του $H_{pfl}(z)$ (α=1, β=0.9)



Σχήμα 7: Διάφορα cepstral φάσματα ομιλίας αλλοιωμένης από το CMV κανάλι (καθαρή ομιλία=συνεχόμενη γραμμή, ομιλία αλλοιωμένη από το κανάλι=εστιγμένη γραμμή). (α) cepstrum ΓΠ $c_{lp}(n)$ (β) ACW cepstrum $c_{acw}(n)$, (γ) PFL cepstrum $c_{pfl}(n)$ (α=1, β=0.9)

Το σχήμα 7 δείχνει τους αντίστροφους cepstral συντελεστές των $c_{lp}(n)$, $c_{acw}(n)$ και $c_{pfl}(n)$ για ένα πλαίσιο καθαρής ομιλίας και για το ίδιο πλαίσιο ομιλίας αλλοιωμένο από το CMV κανάλι [36]. Υπάρχει πολύ λιγότερο μη ταίριασμα στα $c_{acw}(n)$ και $c_{pfl}(n)$ σε σύγκριση με το $c_{lp}(n)$.

4.8 Mel-warped Cepstrum

Το mel-warped cepstrum διαφέρει από το cepstrum ΓΠ στο ότι το mel-warped cepstrum υπολογίζεται με τη χρήση τράπεζας φίλτρων στην οποία το σύνολο των φίλτρων έχει το ίδιο εύρος ζώνης ως προς τη mel κλίμακα συχνοτήτων.

Η ανθρώπινη αντίληψη των συχνοτήτων των ήχων δεν ακολουθεί γραμμική κλίμακα. Αντίθετα, είναι προσεγγιστικά γραμμική με λογαριθμική συχνότητα πάνω από τα 1000Hz. Συγκεκριμένα η κρίσιμη ζώνη είναι σταθερή όταν η λογαριθμική συχνότητα είναι κάτω από τα 1000Hz και γραμμική ως προς τη λογαριθμική συχνότητα πέρα των 1000Hz. Η κρίσιμη ζώνη αναφέρεται εντός του εύρους ζώνης στο οποίο οι υποκειμενικές αποκρίσεις, όπως η ακουστότητα, παραμένουν σταθερές μέχρι το εύρος ζώνης του θορύβου να υπερβεί το εύρος της κρίσιμης ζώνης. Η κλίμακα mel ορίζεται κατά τέτοιο τρόπο ώστε τα 1000Hz στο χώρο της γραμμικής συχνότητας να είναι 1000mels. Οι υπόλοιπες τιμές λαμβάνονται με την τροποποίηση της συχνότητας ενός τόνου έτσι ώστε η ανθρώπινα αντιληπτή συχνότητα να είναι η μισή ή η διπλάσια της αντιληπτής συχνότητας ενός σημείου αναφοράς με γνωστή συχνότητα mel. Το φάσμα της κλίμακας mel προσομοιώνεται με τη χρήση τράπεζας φίλτρων ομοιόμορφα κατανεμημένων στην κλίμακα mel, όπου η ενέργεια εξόδου από κάθε ζώνη φίλτρου προσεγγίζει το τροποποιημένο φάσμα. Εάν συμβολίσουμε την ενέργεια εξόδου του κ-στού φίλτρου με $\tilde{S}_k$, το mel-warped cepstrum $c_{mel}(n)$ λαμβάνεται από τον ολισθημένο διακριτό μετασχηματισμό συνημιτόνου (DCT) του φάσματος της κλίμακας mel:

$$c_{mel}(n) = \sum_{k=1}^{K} \log(\tilde{S}_k) \cos(n(k-0.5)\frac{\pi}{K}) \qquad (30)$$

4.9 Αλλες Εύρωστες Cepstral Τεχνικές

Τα τελευταία χρόνια έχουν προταθεί μέθοδοι όπως τα μικρόφωνα ακύρωσης θορύβου, οι προεπεξεργαστές καταστολής θορύβου και η εσωτερική τροποποίηση των αλγορίθμων επεξεργασίας ώστε ρητά να αντισταθμίζουν την αλλοίωση του σήματος και να μειώνουν το θόρυβο στην ακουστική περιοχή.

Η καταστολή του θορύβου [30] με τη χρήση φασματικής κανονικοποίησης αναδεικνύει την ακουστική ποιότητα της ομιλίας. Ωστόσο, σπάνια βελτιώνει την απόδοση ενός συστήματος αναγνώρισης. Το ARMA μοντέλο που χρησιμοποιείται για την εύρωστη εκτίμηση του γραμμικού προβλέπτη λαμβάνει υπόψη την επίδραση ασυσχέτιστου προσθετικού θορύβου και προσθέτει μηδενικά στο μοντέλο της φωνητικής οδού. Ομως δεν είναι υλοποιήσιμο γιατί η υπολογιστική πολυπλοκότητα του είναι μεγάλη και η λύση δεν είναι εγγυημένο ότι συγκλίνει στο ολικό ελάχιστο της ισχυρά μη γραμμικής συνάρτησης κόστους.

Η τεχνική σχετικού φάσματος (RASTA) [37] πλεονεκτεί εξαιτίας του γεγονότος ότι ο ρυθμός μεταβολής των μη γλωσσικών όρων στην ομιλία

συχνά βρίσκεται εκτός του τυπικού ρυθμού μεταβολής του σχήματος της φωνητικής οδού. Επομένως, καταστέλει τους φασματικούς όρους που μεταβάλλονται περισσότερο αργά ή περισσότερο γρήγορα από τον τυπικό ρυθμό μεταβολής της ομιλίας. Η RASTA προσέγγιση μπορεί να συνδυαστεί με την μέθοδο της Αισθητής Γραμμικής Πρόβλεψης για να δώσουν τη συνάρτηση μεταφοράς ΓΠ H(z) [37]. Το φάσμα της κρίσιμης ζώνης της ομιλίας βρίσκεται ακριβώς όπως στη μέθοδο Αισθητής Γραμμικής Πρόβλεψης. Οι τροχιές στο χρόνο των φασματικών όρων φιλτράρονται ώστε να κατασταλούν οι μη γλωσσικοί όροι στο φάσμα. Το φιλτραρισμένο φάσμα προσεγγίζεται από ένα αυτοπαλινδρούμενο μοντέλο. Εχει δειχθεί ότι το δέλτα cepstrum, το οποίο επίσης αντανακλά τις φασματικές μεταβολές, είναι μια ειδική περίπτωση της RASTA επεξεργασίας [37]. Επίσης, αντίθετα από την αφαίρεση της cepstral μέσης τιμής η οποία απομακρύνει τη συνεχή συνιστώσα του λογαριθμικού φάσματος, η επεξεργασία RASTA επιδρά στο φάσμα της ομιλίας με ένα πιο σύνθετο τρόπο και δίνει έμφαση στις φασματικές μεταβάσεις. Η χρήση της επεξεργασίας RASTA βελτιώνει την απόδοση της αναγνώρισης ομιλίας σε μη ταιριαστές συνθήκες [37].

Εναλλακτικά, η επεξεργασία RASTA μπορεί να εφαρμοστεί κατευθείαν στους cepstral συντελεστές της συνάρτησης μεταφοράς ΓΠ που υπολογίζεται με τη συμβατική μέθοδο αυτοσυσχέτισης. Ενα ζωνοπερατό φίλτρο της μορφής [38]

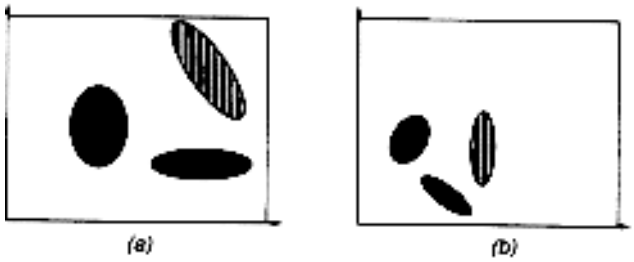
$$B(z) = \frac{a_0 + a_1 z^{-1} + a_3 z^{-3} + a_4 z^{-4}}{(1 - b_1 z^{-1}) z^{-4}} \qquad (31)$$

εφαρμόζεται στους cepstral συντελεστές. Αυτό το ζωνοπερατό φίλτρο, συνδυαζόμενο με BPL φιλτράρισμα, έχει δειχθεί ότι βελτιώνει την απόδοση αναγνώρισης ομιλητή κάτω από μη ταιριαστές συνθήκες [38].

Στις [6, 39, 40] μπορούν να βρεθούν και άλλες τεχνικές που επιχειρούν την εσωτερική τροποποίηση των αλγορίθμων για να προσαρμόσουν το μοντέλο τους από ένα περιβάλλον χωρίς θόρυβο στις ανάγκες ενός περιβάλλοντος με θόρυβο.

5. ΔΙΟΡΘΩΣΗ ΠΑΡΑΜΟΡΦΩΣΗΣ ΜΕ ΤΗ ΧΡΗΣΗ AFFINE ΜΕΤΑΣΧΗΜΑΤΙΣΜΩΝ

Οταν το σήμα ομιλίας παραμορφώνεται από το κανάλι μετάδοσης ή αλλοιώνεται από το θόρυβο, αποδεικνύεται ότι τα cepstral διανύσματα περιστρέφονται, μεταφέρονται και/ή αλλάζουν κλίμακα. Αυτό φαίνεται στο σχήμα 8.



Σχήμα 8: Η αλλαγή κλίμακας, η περιστροφή και η μεταφορά των cepstral διανυσμάτων λόγω των επιδράσεων του θορύβου και του καναλιού: (α) Η κατανομή στο χώρο των καθαρών cepstral διανυσμάτων (β) Η κατανομή στο χώρο των cepstral διανυσμάτων της ομιλίας με θόρυβο και/ ή παραμόρφωση.

Εχει βρεθεί ότι στην πράξη, η αφαίρεση της cepstral μέσης τιμής, η cepstral λείανση και οι άλλες τεχνικές που αναλύσαμε στο προηγούμενο κεφάλαιο προσφέρουν ευρωστία στα συστήματα αναγνώρισης ομιλητή. Συγκεκριμένα, εάν θεωρήσουμε την εφαρμογή της CMS για την κανονικοποίηση του καναλιού και της cepstral λείανσης για τη μείωση της επίδρασης του θορύβου, τότε οι διορθωμένοι cepstral συντελεστές $\hat{c}_i$ μπορούν να αναπαρασταθούν ως

$$\hat{c}_1 = w_1 c_1' - \bar{c}_1 \qquad (32)$$

όπου τα c_i' είναι οι cepstral συντελεστές της καθαρής ομιλίας. Η γενίκευση της εξίσωσης (32) αποτελεί ένα συσχετισμένο (affine) μετασχηματισμό που δίνεται από τη σχέση

$$c' = Ac + b \qquad (33)$$

όπου c' είναι το cepstrum της υποβαθμισμένης ομιλίας και c είναι το cepstrum της αρχικής καθαρής ομιλίας. Οταν ο πίνακας A είναι διαγώνιος και το διάνυσμα b μηδενικό, η σχέση (33) αποτελεί ένα μετασχηματισμό ομοιότητας.

Η ιδέα της χρήσης ενός affine μετασχηματισμού για τη διόρθωση των παραμορφώσεων των cepstral συντελεστών που οφείλονται στις επιδράσεις του καναλιού και του θορύβου προτείνεται στην [41]. Η βασική ιδέα είναι ότι οι συντελεστές του προβλέπτη υφίστανται συσχετισμένο (affine) μετασχηματισμό, όταν το σήμα ομιλίας αλλοιώνεται από περιβαλλοντολογικές αλλαγές. Αποτέλεσμα του μετασχηματισμού αυτού είναι ο συσχετισμένος

(affine) μετασχηματισμός των cepstral συντελεστών. Οι μετασχηματισμοί εξαρτώνται από τις φασματικές ιδιότητες των ήχων. Ετσι ένα φασματικά παρόμοιο σύνολο cepstral διανυσμάτων υφίσταται τον ίδιο μετασχηματισμό.

Στη συνέχεια παρουσιάζουμε σε συντομία την ανάλυση της μεθόδου και εξετάζουμε τις επιδράσεις του προσθετικού θορύβου και του γραμμικού καναλιού ξεχωριστά. Για την ανάλυση χρησιμοποιούμε τη λύση των συντελεστών του προβλέπτη που προκύπτει από την εφαρμογή της μεθόδου της αυτοσυσχέτισης και δίνεται από τη σχέση $a = R_S^{-1} r_S$.

5.1 Προσθετικός θόρυβος

Ο τυχαίος θόρυβος που προέρχεται από το υπόβαθρο και οι διακυμάνσεις του καναλιού μετάδοσης υποθέτουμε γενικά ότι ανήκουν στην κατηγορία του προσθετικού λευκού θορύβου (AWN). Ετσι η παρατήρηση του αρχικού σήματος ομιλίας που έχει αλλοιωθεί από το θόρυβο δίνεται από τη σχέση

$$s'(n) = s(n) + q(n) \qquad (34)$$

Ο θόρυβος q(n) είναι τέτοιος ώστε

$$E[q(n)] = 0 \quad \text{και} \quad E[q^2(n)] = \sigma^2 \qquad (35)$$

Αποδεικνύεται [41] ότι οι συντελεστές του προβλέπτη της ομιλίας που έχει αλλοιωθεί από θόρυβο, σύμφωνα με την εξίσωση (11) είναι

$$a' = R_{S'}^{-1} r_{S'} = (R_S + \sigma^2 I)^{-1} r_S = (R_S + \sigma^2 I)^{-1} R_S a \qquad (36)$$

Συγκρίνοντας με την εξίσωση (11) συμπεραίνουμε ότι η πρόσθεση του λευκού θορύβου στην ομιλία είναι ισοδύναμη με το γραμμικό μετασχηματισμό των συντελεστών του προβλέπτη. Ο γραμμικός μετασχηματισμός εξαρτάται από την αυτοσυσχέτιση της ομιλίας. Επομένως σε ένα φασματικό μοντέλο, όλοι οι φασματικά παρόμοιοι προβλέπτες αντιστοιχούν σε ένα παρόμοιο γραμμικό μετασχηματισμό.

Το επόμενο βήμα είναι η SVD ανάλυση του μετασχηματισμού της εξίσωσης (36). Η ανάλυση αυτή μας επιτρέπει να εμβαθύνουμε και να καταλάβουμε καλύτερα την αλληλεπίδραση του θορύβου και των συντελεστών του προβλέπτη. Ας υποθέσουμε ότι ο Toeplitz πίνακας αυτοσυσχέτισης του αρχικού σήματος ομιλίας R_S αναλύεται με SVD στη μορφή

$$R_S = U \Lambda U^{\mathrm{T}} \qquad (37)$$

όπου U είναι ένας μοναδιαίος πίνακας και Λ είναι ένας διαγώνιος πίνακας του οποίου τα διαγώνια στοιχεία είναι οι ιδιοτιμές του πίνακα R_S. Αποδεινύεται [41] τότε ότι η εξίσωση (36) γράφεται στη μορφή

$$a' = [U(\Lambda + \sigma^2 I)U^r]^{-1}(U\Lambda U^r)a = U[(\Lambda + \sigma^2 I)U^r a, \qquad (38)$$

$$= U \begin{pmatrix} \frac{\lambda_1^2}{\lambda_1^2 + \sigma^2} & & & \\ & \frac{\lambda_2^2}{\lambda_2^2 + \sigma^2} & & \\ & & \ddots & \\ & & & \frac{\lambda_n^2}{\lambda_n^2 + \sigma^2} \end{pmatrix} U^r a$$

Από την εξίσωση αυτή διαπιστώνουμε ότι η νόρμα των συντελεστών του προβλέπτη μειώνεται όταν η ομιλία διαταράσσεται από λευκό θόρυβο. Συνεπώς όταν η ομιλία αλλοιώνεται από προσθετικό λευκό θόρυβο, το διάνυσμα συντελεστών του προβλέπτη διατηρεί το γενικό του προσανατολισμό αλλά υφίσταται συρρίκνωση που το μετακινεί πλησιέστερα στην αρχή των αξόνων.

5.2 Γραμμικό Κανάλι

Οταν μια ακολουθία δειγμάτων S(n) διέρχεται από ένα συγκεραστικό κανάλι με κρουστική απόκριση p(n), το σήμα που λαμβάνεται στην έξοδο του καναλιού είναι ο συγκερασμός

$$s'(n) = p(n) \otimes s(n) \qquad (39)$$

Αν το φάσμα ισχύος των σημάτων s(n) και s'(n) συμβολίζεται με $S_S(\omega)$ και $S_{S'}(\omega)$ αντίστοιχα, τότε ισχύει

$$S_{S'}(\omega) = |P(\omega)|^2 S_S(\omega) \qquad (40)$$

Μετατρέποντας την προηγούμενη εξίσωση στο πεδίο του χρόνου έχουμε

$$r_{S'}(k) = [p(n) \otimes p(-n)] \otimes r_S(k) = r_p(k) \otimes r_S(k) \qquad (41)$$

όπου $r_S(k)$ και $r_{S'}(k)$ είναι η αυτοσυσχέτιση του σήματος εισόδου και του σήματος εξόδου, αντίστοιχα και $\otimes$ ο συγκεραστικός τελεστής. Με τη χρήση αυτών των σχέσεων, αποδεικνύεται ότι οι συντελεστές του προβλέπτη του σήματος εξόδου s'(n) δίνοντα [41] από τη σχέση

$$a_{S'} = Aa \qquad (42)$$

Επομένως οι συντελεστές του προβλέπτη του σήματος ομιλίας που φιλτράρεται από ένα συγκεραστικό κανάλι λαμβάνονται από ένα γραμμικό μετασχηματισμό των συντελεστών του προβλέπτη του σήματος ομιλίας εισόδου. Παρατηρούμε ότι δεν προκύπτει καμμία μεταφορά των συντελεστών

του προβλέπτη. Τέλος σημειώνουμε ότι ο μετασχηματισμός στην εξίσωση (42) εξαρτάται από την ομιλία, καθώς οι εκτιμήσεις των πινάκων αυτοσυσχέτισης υποθέτουν στατικότητα.

5.3 Ενδοκαναλική Παρεμβολή

Η ενδοκαναλική παρεμβολή που οφείλεται σε ένα δεύτερο ομιλητή μπορεί επίσης να ερμηνευθεί ως ένας συσχετισμένος μετασχηματισμός. Στην περίπτωση αυτή της παρεμβολής ενός άλλου ομιλητή που μιλά στο ίδιο κανάλι, το παρατηρούμενο σήμα s_T είναι

$$s_T = s_1 + s_2 \qquad (43)$$

όπου s_1, s_2 τα δύο σήματα ομιλίας. Κατά συνέπεια

$$R_{s_T} = R_{s_1} + 2R_{s_1 s_2} + R_{s_2} \qquad (44)$$

και

$$r_{s_T} = r_{s_1} + 2r_{s_1 s_2} + r_{s_2} \qquad (45)$$

Επομένως οι συντελεστές γραμμικής πρόβλεψης για το σήμα s_T είναι

$$\begin{aligned} a &= R_{s_T}^{-1} r_{s_T} \\ &= (R_{s_T} + 2R_{s_1 s_2} + R_{s_2})^{-1} R_{s_1} R_{s_1}^{-1} (r_{s_T} + 2r_{s_1 s_2} + r_{s_2}) \\ &= Aa_1 + b \end{aligned} \qquad (46)$$

όπου

$$A = (R_{s_1} + 2R_{s_1 s_2} + R_{s_2})^{-1} R_{s_1} \qquad (47)$$

και

$$b = (R_{s_1} + 2R_{s_1 s_2} + R_{s_2})^{-1} R_{s_1} (2r_{s_1 s_2} + r_{s_2}) \qquad (48)$$

Πάλι, η ενδοκαναλική παρεμβολή οδηγεί σε συσχετισμένο μετασχηματισμό των συντελεστών του προβλέπτη.

Τα παραπάνω αποτελέσματα δείχνουν ότι οι μη ταιριαστές συνθήκες που οφείλονται σε αλλοίωση από προσθετικό θόρυβο, σε διέλευση από γραμμικό κανάλι και σε ενδοκαναλική παρεμβολή είναι το καθένα είτε ένας γραμμικός είτε ένας συσχετισμένος μετασχηματισμός των συντελεστών του γραμμικού προβλέπτη. Γενικά, λόγω του μεταβατικού χαρακτήρα του συσχετισμένου μετασχηματισμού, μια ακολουθία παραμορφώσεων που προκύπτουν από τις επιδράσεις του θορύβου και του καναλιού είναι επίσης ισοδύναμη με ένα συσχετισμένο μετασχηματισμό της μορφής

$$a' = Aa + b \qquad (49)$$

5.4 Affine Μετασχηματισμός του Cepstrum

Εμπειρικά έχει βρεθεί ότι τα cepstral χαρακτηριστικά μεγέθη είναι τα πιο εύρωστα στις διάφορες πηγές υποβάθμισης. Σε αυτήν την παράγραφο θα δούμε για το cepstrum ΓΠ ότι η επίδραση του καναλιού και του θορύβου μπορεί επίσης να μοντελοποιηθεί ως ένας συσχετισμένος μετασχηματισμός.

Το cepstrum ΓΠ από τον ορισμό του ισούται με

$$c_{lp}(n) = Z^{-1}[\log(H(z))] = Z^{-1}[\log(\frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}})] \qquad (50)$$

Η πρώτη μερική παράγωγος του $c_{lp}(n)$ ως προς το a_i δίνεται από τη σχέση

$$\begin{aligned} \frac{\partial c_{lp}(n)}{\partial a_i} &= \frac{\partial Z^{-1}[\log(\frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}})]}{\partial a_i} \qquad (51) \\ &= Z^{-1}[\frac{\log(\frac{1}{1-\sum_{k=1}^{p} a_i z^{-k}})}{\partial a_i} \\ &= h(n-i) \end{aligned}$$

όπου h(n) είναι η αιτιατή και ευσταθής κρουστική απόκριση που σχετίζεται με τη συνάρτηση μεταφοράς ΓΠ H(z). Επομένως αν c είναι το διάνυσμα των p πρώτων cepstral συντελεστών ΓΠ, τότε

$$dc = Hda \qquad (52)$$

όπου

$$H = \begin{pmatrix} h(0) & 0 & \cdots & 0 \\ h(1) & h(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h(p-1) & h(p-2) & \cdots & h(0) \end{pmatrix} \qquad (53)$$

Ο πίνακας κρουστικής απόκρισης H θα είναι ο ίδιος για μια ομάδα φασματικά παρόμοιων cepstral διανυσμάτων. Η σχέση μεταξύ ενός διαφορικού cepstral διανύσματος ΓΠ και ενός διαφορικού διανύσματος συντελεστών προβλέπτη δίνεται από την εξίσωση (52).

Στη συνέχεια ας υποθέσουμε ότι το σήμα ομιλίας υφίσταται παραμόρφωση από ένα κανάλι και/ή θόρυβο. Τότε έχουμε μια νέα συνάρτηση μεταφοράς ΓΠ H'(z). Το αντίστοιχο διάνυσμα συντελεστών του προβλέπτη δίνεται από το α' και το διάνυσμα του cepstral ΓΠ δίνεται από το c'. Για ένα

σύνολο φασματικά παρόμοιων cepstral διανυσμάτων, ο ίδιος μετασχηματισμός μπορεί να γραφεί ως

$$dc' = H' da' \qquad (54)$$

Αφού, όταν υπάρχει παρεμβολή οι συντελεστές του προβλέπτη υφίστανται συσχετισμένο μετασχηματισμό υπό την έννοια ότι $a' = Aa + b$, διαφορίζοντας και τα δύο μέλη παίρνουμε

$$da' = Ada \qquad (55)$$

Τότε έχουμε

$$\frac{dc'}{dc} = (\frac{dc'}{da'})(\frac{da'}{da})(\frac{da}{dc}) = H' AH^{-1} \qquad (56)$$

Ολοκληρώνοντας βλέπουμε ότι το cepstrum ΓΠ που αντιστοιχεί στην παραμορφωμένη ομιλία δίνεται από τη σχέση

$$c' = H' AH^{-1} c + b \qquad (57)$$

που αποτελεί έναν συσχετισμένο μετασχηματισμό. Συμπεραίνουμε λοιπόν ότι οι cepstral συντελεστές ΓΠ υφίστανται συσχετισμένο μετασχηματισμό όταν έχουμε μη ταιριαστές συνθήκες θορύβου και καναλιού. Οι παράμετροι του συσχετισμένου μετασχηματισμού εξαρτώνται από τα φάσματα.

5.5 Υπολογισμός των Παραμέτρων των Affine Μετασχηματισμών

Ας υποθέσουμε ότι η αντιστοιχία μεταξύ των cepstral διανυσμάτων των συνθηκών εκμάθησης $c_i = \begin{bmatrix} c_{i1} & c_{i2} & \cdots & c_{ip} \end{bmatrix}^T$ και των cepstral διανυσμάτων των συνθηκών ελέγχου $c_i' = \begin{bmatrix} c_{i1}' & c_{i2}' & \cdots & c_{ip}' \end{bmatrix}^T$ είναι γνωστή για i=1 έως N, όπου N είναι το πλήθος των cepstral διανυσμάτων. Ο συσχετισμένος μετασχηματισμός που συσχετίζει τα διανύσματα c_i και c_i' δίνεται από τη σχέση

$$c_i' = Ac_i + b \qquad (58)$$

Η σχέση (58) αναπτύσσεται στην

$$\begin{pmatrix} c_{i1}' \\ \vdots \\ c_{ip}' \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix} \begin{pmatrix} c_{i1} \\ \vdots \\ c_{ip} \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} \quad \gamma\iota\alpha \quad i = 1 \quad \varepsilon\omega\varsigma \quad N \qquad (59)$$

Κάθε γραμμή του πίνακα A (τα στοιχεία a_{jk} για k=1 έως p) και το αντίστοιχο στοιχείο του διανύσματος b (δηλαδή το b_j) καθορίζονται ξεχωριστά. Για τον καθορισμό της j-οστής γραμμής του A και του b_j συλλέγουμε το j-οστό όρο καθενός από τα cepstral διανύσματα των συνθηκών ελέγχου και καταλήγουμε στο ακόλουθο σύστημα εξισώσεων

$$\begin{pmatrix} c'_{ij} \\ \vdots \\ c'_{Nj} \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1p}\,1 \\ \vdots & \ddots & \vdots\,\vdots \\ c_{N1} & \cdots & c_{Np}\,1 \end{pmatrix} \begin{pmatrix} a_{j1} \\ \vdots \\ a_{jp} \\ b_j \end{pmatrix} = \begin{pmatrix} c_1\,1 \\ \vdots\,\vdots \\ c_N\,1 \end{pmatrix} \begin{pmatrix} a_{j1} \\ \vdots \\ a_{jp} \\ b_j \end{pmatrix} \quad \gamma\iota\alpha \quad i=1 \quad \varepsilon\omega\varsigma \quad p \qquad (60)$$

Αυτό είναι σχεδόν πάντα ένα υπερκαθορισμένο σύστημα εξισώσεων και επομένως λαμβάνουμε μια λύση ελαχίστων τετραγώνων [42] της μορφής

$$\begin{pmatrix} a_{j1} \\ \vdots \\ a_{jp} \\ b_j \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} c_i c_i^T & \sum_{i=1}^{N} c_i \\ (\sum_{i=1}^{N} c_i)^T & N \end{pmatrix} (c_1 \cdots c_N \quad 1) \begin{pmatrix} c'_{1j} \\ \vdots \\ c'_{Nj} \end{pmatrix} \quad \gamma\iota\alpha \quad i=1 \quad \varepsilon\omega\varsigma \quad N \qquad (61)$$

Με τον συσχετισμένο μετασχηματισμό, όπως παρατηρούμε παραπάνω, τα διανύσματα των συνθηκών εκμάθησης μπορούν να αντιστοιχηθούν στο χώρο των διανυσμάτων των συνθηκών ελέγχου. Η ανάστροφη αντιστοίχηση είναι επίσης δυνατή αν επιλύσουμε ως προς τα διανύσματα των συνθηκών εκμάθησης αντί για τα διανύσματα των συνθηκών ελέγχου.

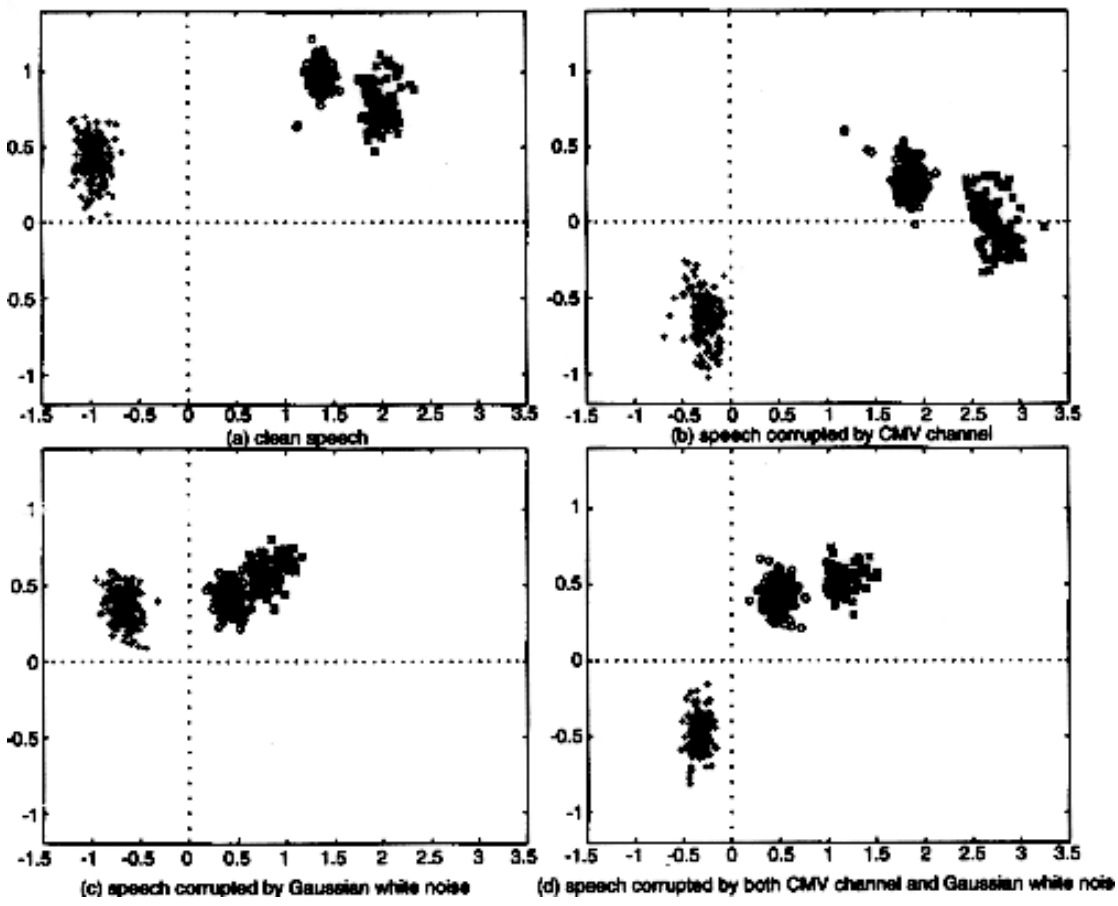
5.6 Γεωμετρική Ερμηνεία των Affine Μετασχηματισμών

Ο συσχετισμένος μετασχηματισμός μπορεί να συμπεριλάβει μια μεγάλη ποικιλία μη ταιριαστών συνθηκών και να περιγράψει ως υποπεριπτώσεις του άλλες εύρωστες τεχνικές. Ας μελετήσουμε τα επόμενα παραδείγματα.

1. Εάν οι συνθήκες εκμάθησης ταιριάζουν με τις συνθήκες ελέγχου, με την έννοια ότι τα αντίστοιχα cepstral διανύσματα είναι ταυτόσημα, οι παράμετροι του συσχετισμένου μετασχηματισμού είναι $A = I$ και $b = 0$.
2. Η παραμόρφωση από το κανάλι οδηγεί σε πίνακα A πολύ κοντά στο μοναδιαίο πίνακα και το διάνυσμα b να παριστάνει μεταφορά που είναι ισοδύναμη με τις τεχνικές της CMS και της PFCMS. Οι τεχνικές αυτές αναδεικνύονται ως υποπεριπτώσεις του συσχετισμένου μετασχηματισμού.
3. Η λείανση, η οποία προσφέρει ευρωστία [17], είναι και αυτή υποπερίπτωση, αφού ο A είναι διαγώνιος και $b = 0$.
4. Ο προσθετικός θόρυβος προκαλεί τη μείωση του μεγέθους των cepsrtal διανυσμάτων χωρίς να αλλάζει σημαντικά τον προσανατολισμό τους [11]. Αυτός ο τύπος παραμόρφωσης μπορεί να περιγραφεί από ένα διαγώνιο A του οποίου τα διαγώνια στοιχεία είναι γενικά διάφορα μεταξύ τους και έχουν

μέγεθος μικρότερο από τη μονάδα. Επίσης $b = 0$. Ας σημειωθεί ότι αυτή είναι επίσης μια ειδική περίπτωση λείανσης.

5. Σύνθετες επιδράσεις από το κανάλι και το θόρυβο επίσης μοντελοποιούνται ως ένας συσχετισμένος μετασχηματισμός με τον A κύρια υπεύθυνο για την παραμόρφωση από το θόρυβο και το b κύρια υπεύθυνο για την παραμόρφωση από το κανάλι. Το σχήμα 9 δείχνει την αλλαγή της ταξινόμησης στο χώρο των cepstral συντελεστών που οφείλεται στις επιδράσεις του γραμμικού καναλιού, του λευκού θορύβου και του συνδυασμού και των δύο.



Σχήμα 9: Η χωρική κατανομή των cepstral συντελεστών σε διάφορες συνθήκες, '∗' για το φωνήεν /α/, 'ο' για τον ένρινο ήχο /n/, και '+' για τον ήχο /sh/ (α) Cepstrum καθαρής ομιλίας (β) Cepstrum των σημάτων φιλτραρισμένων από το ηπειρωτικό κανάλι φωνής CMV των ΗΠΑ (γ) Cepstrum των σημάτων με 15dB SNR. Ο θόρυβος είναι προσθετικός λευκός Gaussian (AWG); (δ) Cepstrum της ομιλίας αλλοιωμένο από το CMV κανάλι και τον AWG θόρυβο των 15dB SNR.

Τελειώνοντας επισημαίνουμε ότι η χρήση του συσχετισμένου μετασχηματισμού αποδεικνύεται ότι βελτιώνει δραστικά την απόδοση συστημάτων αναγνώρισης ομιλητή εξαρτωμένων από κείμενο [41].

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα εργασία παρουσίασε μια ανασκόπηση μερικών από τις χρησιμοποιούμενες τεχνικές για την εύρωστη αναγνώριση ομιλητή με έμφαση στα θέματα ανάδειξης και εξαγωγής των χαρακτηριστικών μεγεθών. Τα περισσότερα από τα χαρακτηριστικά μεγέθη που περιγράφονται βασίζονται στο μοντέλο γραμμικής πρόβλεψης της ομιλίας. Η κλασσική μέθοδος αυτοσυσχέτισης για τον υπολογισμό των συντελεστών γραμμικής πρόβλεψης δεν είναι από μόνη της πολύ εύρωστη σε μια πολύ ευρεία γκάμα διαφορετικών περιβαλλοντικών συνθηκών. Ετσι είναι απαραίτητο να αναζητηθεί στο μέλλον ένα καλύτερο μοντέλο που να είναι εύρωστο και υπολογιστικά υλοποιήσιμο.

Οι συντελεστές γραμμικής πρόβλεψης μετατρέπονται σε διάφορους τύπους cepstral χαρακτηριστικών συντελεστών. Ειδικά οι μέθοδοι της προσαρμοστικής στάθμισης και του μεταφιλτραρισμένου cepstrum είναι αρκετά εύρωστες στις επιδράσεις του καναλιού και του θορύβου. Ομως χρειάζεται περισσότερη προσπάθεια για την εύρεση χαρακτηριστικών μεγεθών που επιτυγχάνουν πολύ υψηλή απόδοση αναγνώρισης, ειδικά στις περιπτώσεις που έχουμε ισχυρές παραμορφώσεις από το κανάλι και πολύ χαμηλά SNRs.

Ο συσχετισμένος (affine) μετασχηματισμός είναι μια πολύ πρόσφατη και πολλά υποσχόμενη τεχνική. Αντιστοιχίζει στο χώρο των χαρακτηριστικών μεγεθών μια περιοχή σε μιαν άλλη, με σκοπό τη διόρθωση των αποκλίσεων που προκαλούνται από την αλλοίωση του σήματος ομιλίας από το κανάλι και το θόρυβο. Με το συσχετισμένο μετασχηματισμό επιτυγχάνεται σχετικά καλύτερη απόδοση σε χαμηλά SNRs κάτι που ενθαρρύνει την περαιτέρω έρευνα στην κατεύθνση αυτή.

ΒΙΒΛΙΟΓΡΑΦΙΑ - ΑΝΑΦΟΡΕΣ

1. A.E. Rosenberg, F.K. Soong. ‘’Recent Research in Automatic Speaker Recognition.’’ In S. Furui and M. M. Sondhi, editors, Advances in Speech Signal Processing, 701-738, Marcel Dekker, 1991.
2. J.H.L. Hansen, R.J. Mammone, S. Young, editors, IEEE Transactions on Speech and Audio Processing, October, 1994
3. S.F. Boll. ‘’Suppression of Acoustic Noise in Speech Using Spectral Subtraction,’’ IEEE Trans. Acoust., Speech, Signal Processing, 27:113-120, April, 1979.
4. S. Furui. ‘’Cepstral Analysis Techniques for Automatic Speaker Verification,’’ IEEE Trans. Acoust. Speech, Signal Processing, 29:254-272, April, 1981.
5. D. Mansour and B.H. Juang. ‘’The short-time modified coherence representation and noisy speech recognition,’’ IEEE Trans. Acoust. Speech, Signal Processing, 37: 795-804, June, 1989.
6. L. Neweyer and M. Weintraub. ‘’Probabilistic Optimum Filtering for Robust Speech Recognition,’’ Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, 1:417-420, 1994.
7. D.A. Reynolds and R.C. Rose. ‘’Robust Text-indepedent Speaker Identification Using Gaussian Mixture Speaker Models,’’ Trans. Speech, Audio Processing, January, 1995.
8. J.A. Nolazco Flores, S.J. Young. ‘’Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation,’’ Proc IEEE Int. Conf. Acoust. Speech Signal Processing, 1 409:412, 1994.
9. F. Itakura, Minimum Prediction residual principle applied to speech redognition, IEEE Trans Acoust., Speech, Signal Processing, 23:67-72, Feb., 1975.
10. F.K. Soong, M.M. Sondhi. «A Frequency-weighted Itakura Spectral Distortion Measure and its Application to Speech Recognition in Noise,’’ IEEE Trans. Acoust. Speech, Signal Processing, 36: 41-48, Jan, 1988.
11. D. Mansour, B.H. Juang. ‘’A Family of Distortion Measures Based Upon Preojection Operation for Robust Speech Recognition,’’ IEEE Trans. Acoust. Speech, Signal Processing, 37: 1659-1671, November, 1989.
12. J.D. Markel, A.H. Gray Jr. Linear Prediction of Speech, Springer-Verlag, Berlin Heidelberg New York, 1976.
13. L.R. Rabiner, R.W. Schafer. Digital Processing Of Speech Signals, Prentice Hall, Englewood Cliffs, NJ, 1978.
14. G. Fant, Acoustic Theory of Speech Production, Mouton and Co., Gravenhage, The Netherland, 1960.
15. J.R Deller, J.G. Proakis, J.H. Hansen, Discrete-time Processing of Speech Signals, Macmillan, New York NY, 1993.
16. K.T. Assaleh, R. J. Mammome, M.G. Rahim, J.L. Flanagan. ‘’Speech Recognition Using The Modulation Model.’’ Proc. IEEE Int. Conf. Acoust. Speech, Signal processing, 2, 664-667, April, 1993.
17. L.R. Rabiner, B.H. Juang. Fundamentals of Speech recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.
18. R.P. Ramachandran, M.S. Zilovic, R.J. Mammone. ‘’A Comparative Study Of Robust Linear Predictive Analysi+s Methods with Applications to

Speaker Identification,'' IEEE Trans. Speech, Audio Processing, 3: 117-125, March, 1995.
19. H. Hermansky. ''Perceptual Linear Predictive (PLP) Analysis of Speech,'' Journal of the Acoust. Society of Amer., 87:1738-1752, April, 1990.
20. P.C. Woodland, M.J.F. Gales, D. Pye, V. Valtchev. ''The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task.'' ARPA Speech Recognition Workshop, Februaray, 1996.
21. B.S. Atal. ''Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification,'' J. Acoust. Soc. Am., 55:1304-1312, 1974.
22. C.-S. Liu, M.-T. Lin, W. -J. Wang, H.-C. Wang. ''Study of Line Spectrum Pair Frequencies for Speaker Recognition,'' Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 277-280, 1990.
23. A.V. Oppenheim, R.W. Schafer, Discrete-Time Signal Processing, Prentice Hall, Englewood Cliffs, NJ, 1989.
24. M.R. Schroeder. ''Direct (Nonrecursive) Relations Between Cepstrum and Predictor Coefficients,'' IEEE Trans. Acoust., Speech, Signal Processing, 29:297-301, April, 1981.
25. F.K. Soong, A.E. Rosenberg ''On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition,'' IEEE Trans. Acoust., Speech, Signal Processing, 36:871-879, June 1988.
26. B.H. Juang, L.R. Rabiner. J.G. Wilpon, ''On the Use of Bandpass Liftering in Speech Recognition,'' IEEE Trans. Acoust., Speech, Signal Processing, 35:947-954, July, 1987.
27. B.S. Atal. ''Automatic Recognition of Speakers from their Voices,'' Proc. IEEE, 64:460-475, April, 1976.
28. D.Naik. ''Pole-filtered Cepstral Mean Subtraction,'' Proc. IEEE Int Conf. Acoust., Speech, Signal Processing, 1:157-160, 1995.
29. D. Naik, K.T. Assaleh, R.J. Mammone. ''Robust Speaker Identification Using Pole Filtering,'' ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1994.
30. K.T. Assaleh and R.J. Mammone, New LP-derived features for speaker identification, IEEE Trans. Speech, Audio Processing, 2:630-638, October, 1994.
31. M.S. Zilovic, R.P. Ramachandran, R.J. Mammone. ''A Fast Algorithm for Finding the Adaptive Component Weighted Cepstrum for Speaker Recognition,'' Submitted to IEEE Trans. Speech, Audio Processing, March, 1995.
32. M.S. Zilovic, R.P. Ramachandran, R.J. Mammone, '' Speaker Identification Based on the Use of Robust Cepstral Features Obtained from Pole-zero Transfer Functions,'' Submitted to IEEE Trans. Speech, Audio Processing, March, 1995.
33. V. Ramamoorthy, N.S. Jayant, R.V. Cox, M.M. Sondhi. ''Enchancement of ADPCM Speech Coding with Backward Adaptive Algorithms for Post-filtering and Noise Feeddback,'' IEEE Jour, on Select. Areas in Commun, 6:364-382, February, 1988.
34. K.K. Paliwal. ''On the Performance of the Frequency-weighted Cepstral Coefficients in Vowel Recognition,'' Speech Communication, 1:151-154, May, 1982.

35. Y. Tohkura. ‘’A Weighted Cepstrtal Distance Measure for Speech Recognition,’’ IEEE Trans. Acoust., Speech, Signal Processing, 35:1414-1422, Oct., 1987.
36. J. Kupin. ‘’A Wireless Simulator (Software),’’ CCR-P, April, 1993.
37. H. Hermansky, N. Morgan. ‘’RASTA Processing of Speech,’’ IEEE Trans. Speech, Audio Processing, 2:578-589, October, 1994.
38. J.S. Baras, P.K. Rajasekaran. ‘’Robustness Study of Free-text Speaker Identification and Verification,’’ Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 2:379-382, 1993.
39. A. Nadas, D. Nahamoo, M.A. Picheny, ‘’Adaptive Labeling: Normalization of Speech by Adaptive Transformation Based on Vector Quantization,’’ ’’ Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 521-524, 1988.
40. H. Gish, K. Ng and J.R. Rohlicek. ‘’Robust Mapping of Noisy Speech Parameters for HMM Word Spotting,’’ ’’ Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, 2:109-112, 1992.
41. Xiaoyu Zhang, R.J. Mammone. ‘’The Affine Transformed Cepstrum for Robust Speaker Identification,’’ Submitted to IEEE Trans. Speech, Audio Processing, May, 1995.
42. Rj. Mammone, Xiaoyu Zhang, R.P. Ramachandran, “Robust Speaker Recognition: A Feature - based Approach’’, IEEE Signal Processing Magazine, Sept. 1996, pp. 58-71

ΠΑΡΑΡΤΗΜΑ

ΠΙΝΑΚΑΣ ΑΓΓΛΙΚΩΝ ΟΡΩΝ

adaptive	προσαρμοστικός	
affine transform	συσχετισμένος μετασχηματισμός	
autoregressive model	αυτοπαλινδρούμενο μοντέλο	βλπ. AR
classifier	ταξινομητής	
coherence	συμφωνία	
contextual	συμφραζόμενα	
covariance	συνδιασπορά, συνδιακύμανση, συμμεταβλητότητα	
feature	χαρακτηριστικά μεγέθη, χαρακτηριστικές ποσότητες	
feature extractor	εξαγωγέας χαρακτηριστικών μεγεθών	
frame	πλαίσιο	
gain	απολαβή, κέρδος	
identification	ταυτοποίηση	
implementation stage	στάδιο υλοποίησης	βλπ. pattern recognition
impulse response	κρουστική απόκριση	
interference	παρεμβολή	
likelihood	πιθανοφάνεια	
mean-square error	μέσο τετραγωνικό σφάλμα	
minimum phase	ελάχιστης φάσης	Εχει όλους τους πόλους και τα μηδενικά εντός του μοναδιαίου κύκλου στο z-επίπεδο.
mismatched conditions	μη ταιριαστές συνθήκες	
modulation	διαμόρφωση	πχ. AM και FM
pattern recognition	αναγνώριση προτύπων, αναγνώριση μορφών	Περιλαμάνει 3 στάδια: το στάδιο εκμάθησης, το στάδιο ελέγχου και το στάδιο υλοποίησης
pole filtering	φιλτράρισμα πόλων	
postfilter	μεταφίλτρο	
radiation	ακτινοβόληση, ακτινοβολία	
residue	υπόλοιπο	
spectrum	φάσμα	
speaker identification	ταυτοποίηση ομιλητή	
speech frame	πλαίσιο ομιλίας	
testing stage	στάδιο ελέγχου	βλπ. pattern recognition
time-varying	χρονικά μεταβαλλόμενο	
training stage	στάδιο εκμάθησης	βλπ. pattern recognition
transfer function	συνάρτηση μεταφοράς	
vocabulary	λεξιλόγιο	
weighting	στάθμιση	

ΠΙΝΑΚΑΣ ΣΥΝΤΜΗΣΕΩΝ

ACW	Adaptive Component Weighting	
AR	AutoRegressive model	$s(n) = \sum_{i=1}^{p} a_i s(n-i) + e(n)$
ARMA	AutoRegressive Moving Average model	
AWG	Additive White Gaussian	
AWN	Additive White Noise	
BPL	BandPass Liftering	
CMS	Cepstral Mean Subtraction	
CMV	Continental Mid Voice	
DCT	Discrete Cosine Transform	
FIR	Finite Impulse Response	
GMM	Gaussian Mixture Model	
LAR	Log Area Ratios	
LP	Linear Prediction	βλπ. AR
LSP	Line Spectra Pairs	
PFCMS	Pole-Filtered CMS	
PFL	PostFiLter	
PLP	Perceptual Linear Prediction	
SVD	Singular Value Decomposition	