

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολο-
γιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Ζωγράφου 157 73, Αθήνα



Κατάτμησης Φωνής με Αυτόματες μεθόδους

Ιωάννης Ατσόνιος

Επιβλέπων Καθηγητής: Γεώργιος Καραγιάννης

Διπλωματική Εργασία

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολο-
γιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Ζωγράφου 157 73, Αθήνα



Copyright © Ιωάννης Ατσόνιος

Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ' ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν την χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

ΕΜΠ
Σελίδες 97

Ιωάννης Ατσόνιος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών ΕΜΠ.

Ευχαριστίες:

Ευχαριστώ θερμά τον καθηγητή Γεώργιο Καραγιάννη που μου εμπιστεύτηκε αυτό το θέμα της διπλωματικής και που για εμένα είναι πρότυπο επιστήμονα και ανθρώπου με ιδιαίτερο ήθος που εκκλίνει στις μέρες μας, συνδιάζει το πρότυπο ενός αναγεννησιακού και πολυσχιδούς ανθρώπου ως προς την επιστημονική κατάρτιση και παράλληλα το αρχαίο ελληνικό πρότυπο του καλού και αγαθού που πάντοτε υπερασπίζει το κοινό συμφέρον και προκοπή. Οι συμβουλές του και κατ'ιδίαν συζητήσεις θα με συντροφεύουν σε όλη μου την ζωή.

Παράλληλα ευχαριστώ και το οικογενειακό και φιλικό μου περιβάλλον με κορωνίδα τους γονείς μου, Ιάκωβο Ατσόνιο και Ειρήνη Ατσόνιου (γένος Χρυσανθοπούλου) που μου συμπαράσταθηκαν όλα αυτά τα χρόνια με την αμέριστη τους αγάπη, εμπιστοσύνη και υποστήριξη τόσο υλικά όσο και ηθικά και δημιούργησαν τις κατάλληλες συνθήκες να ευοδωθεί και να πραγματοποιηθεί η όποια μου αναζήτηση και με εμπύχωσαν σε στιγμές δύσκολες και μοιράστηκαν αντίστοιχα την χαρά μου

.Τους είμαι βαθιά ευγνώμων, ελπίζω και δεσμεύομαι κάποτε να τους το ανταποδώσω. Παράλληλα και τον αδερφό μου Λευτέρη του οποίου το χαοτικό ταμπεραμέντο ελπίζω να τον βοηθήσει να εκπληρωθεί κάθε του προσδοκία.

Θα ήταν βαθια παράλειψη μου να μην ευχαριστήσω και τους μηχανικούς και ερευνητές του ΙΕΛ (τμήμα τεχνολογίας φωνής) των οποίων η εμπειρία και κατάρτιση υπήρξε πολύτιμη σε περιόδους “εγκλωβισμού” σε “τοπικά” ελάχιστα και για την φιλικότητα που τους διακρίνει και ήταν μια μορφή ακαδημαϊκής εστίας.

Τέλος είμαι ευγνώμων και σε όλους τους συμφοιτητές/τριες και φίλους/ες για το πιο ανελαστικό μέγεθος σε αυτή την ζωή τόσο ως προς τη ζήτηση όσο και τη προσφορά, για **τον χρόνο** τους.

Γιαννης Ατσόνιος

Ιταλία, Μπολώνια-Μπερτινόρο, Οκτώβρης 2006

Περίληψη:

Σκοπός της συγκεκριμένης διπλωματικής είναι η μελέτη του προβλήματος της κατάτμησης της φωνής με αυτόματες μεθόδους.

Ιδιαίτερη έμφαση δίδεται σε μεθόδους που έχουν προταθεί διεθνώς και έχουν μελετηθεί σε βάθος. Μάλιστα αναλύεται διεξοδικά το θεωρητικό υπόβαθρο και δίδεται έμφαση στον μαθηματικό φορμαλισμό των ιδεών, συνακόλουθα δίδεται η διαισθητική όψη που αποτέλεσε το εναρκτήριο λάκτισμα για την ανάπτυξη αυτών των μεθόδων. Εν παραλλήλω η μελέτη γίνεται με τρόπο που προσπαθεί να συγκεράσει τον τρόπο της ανάπτυξης των ιδεών σε σχέση με το χρόνο, αλλά και το πώς αυτές εμπλουτίστηκαν, διευρύνθηκαν και ωρίμασαν από διαφόρους επιστήμονες και ερευνητικές ομάδες ανά τον κόσμο. Σε επίπεδο υλοποίησης αναφέρονται τα πειράματα που πραγματοποιήθηκαν και τα αποτελέσματα που προέκυψαν με τις μεθοδολογίες καθώς και αναδεικνύονται όψεις του προβλήματος και πως αυτές μπορούν να αντιμετωπιστούν.

Το πρόβλημα είναι εγγενώς πολυδιάστατο και επαφίεται στην τομή πολλών πεδίων της πληροφορικής, της επεξεργασίας σημάτων, των βάσεων δεδομένων, των μη-γραμμικών και δυναμικών συστημάτων και της μηχανικής μάθησης.

Τέλος ολοκληρώνουμε την διπλωματική παρατίθοντας μία πλειάδα ανοικτών προβλημάτων, αλλά και θέματα προς διερεύνηση και μελέτη που αποτελούν αιχμή της υπάρχουσας βιβλιογραφίας και έρευνας εν έτει 2006.

Λέξεις Κλειδιά:

Κατάτμηση φωνής, αναγνώριση φωνής, εκθέτες Lyapunov, Kolmogorov Complexity, μηχανική μάθηση, συμπίεση, γραφήματα, μη γραμμικές μέθοδοι επεξεργασίας, υψηλότερης τάξης ροπές, fractals

Abstract:

The main purpose of this thesis is the study of the speech segmentation problem via automatic techniques.

Special treatment is dedicated to methods that are proposed globally and have been studied thoroughly and in depth. More specifically is scrutinized the theoretical background and is given emphasis in the mathematical formulation and nomenclature of these ideas and techniques. Consequently we give the intuitive notion and facet that gave the initial impetus for the development of these methods and strategies. In parallel the study is performed in such a way that tries to combine the development of the ideas during the last decades, and how these flourished, been enriched become fruitful and approximate a stable level of consensus among the global research society. Practically we propose the algorithms that developed and tested, we survey our results that gained via these methodologies and we shed light facets and difficulties of the problem and give insight how we can surpass these.

The problem is inherently multi dimensional and lies on the intersection of several fields of active research, namely of Computer Science, Signal Processing, Database Theory, non linear and dynamic systems and machine learning.

Finally we summarize this diploma thesis by proposing a plethora of open problems and direction who need further research and study and enjoy the fact that are in the state of the art of the existing and known literature in the year of 2006.

Key words:

speech segmentation, speech recognition, Lyapunov eigenvalues, Kolmogorov Complexity, machine learning, data compression, graphs, non-linear signal processing, higher order statistics, fractals

Περιεχόμενα:

Ευχαριστίες	5
Περίληψη	6
Abstract	7
Περιεχόμενα	8
1.Εισαγωγή	
1.1Κίνητρα	10
1.2Οργάνωση Διπλωματικής	11
2. Η φωνή ως σήμα και το υπόβαθρο της	
2.1 Εισαγωγή	14
2.2Επικοινωνία μέσω φωνής	15
3. Μερικές σημειώσεις για της μη- γραμμικότητες της φωνής	
3.1_Μη-γραμμικότητες στην παραγωγή της φωνής	18
3.2 Μη γραμμικότητες στην ακουστική της Φωνής	20
3.3 Μη γραμμικότητες στην αντιληπτικότητα της Φώνης.	
3.3.1 Μηχανική συμπεριφορά του αυτιού.	23
3.3.2 Νευρολογική ακουστική επεξεργασία	26
3.4 Ακουστικά μοντέλα	
3.4.1Εισαγωγή	28
3.4.2Φυσιολογία Ακοής	28
4.Δυναμικά συστήματα και Μη-γραμμικές Μέθοδοι	
4.1 Εισαγωγή	29
4.2 Σύγκριση γραμμικών και μη γραμμικών τρόπων επεξεργασίας	31
4.3Μη γραμμική ανάλυση σήματος	32
4.4Υψηλότερης τάξης στατιστικής	32
4.5Μη γραμμική πρόβλεψη	34
4.6Ειδικότερα περί δυναμικών συστημάτων και φωνής	35
4.7 Φωνή και Χάος	41
4.7.1Δυναμικές αναλύσεις της φωνής	41
4.7.2Αλγόριθμος εκτίμησης εκθέτη Lyapunov	42
5.Αναπαράσταση Φωνής	
5.1 MEL FREQUENCY CEPSTRUM COEFFICENTS	46
5.2 RASTA-PLP PROCESSING	50
5.3 AM-FM Models	60
5.4 Χρονικά δυναμικά χαρακτηριστικά της φωνής	62
5.5Εξαγωγή χαρακτηριστικών στο πεδίο του χρόνου	63
5.6Έμφωνα/Αφωνα πλαίσια	66
6. Η φύση του πολυδιάστατου χώρου	
6.1 Εισαγωγή	69
6.2Επιπτώσεις των χαρακτηριστικών υψηλής διάστασης στην επιβλεπόμενη μάθηση	76
6.3 Συμπεράσματα	78
7. Αλγόριθμοι Κατάτμησης Φωνής	
7.1 Εισαγωγή	79
7.2Γενική επισκόπηση των τεχνικών για κατάτμηση φωνής	80
7.3Προτεινόμενος Αλγόριθμος	83
7.3.1Εκτίμηση απόδοσης	85
7.3.2Πειράματα κατάτμησης	85

7.3.3Αποτελέσματα	87
7.3.4Πηγές σφαλμάτων	89
7.3.5Συμπεράσματα	90
8.Μελλοντική Έρευνα	
8.1 Μελλοντική Εργασία	91
8.2 Προτάσεις για περαιτέρω έρευνα	91
9.Βιβλιογραφία	95

Κεφάλαιο 1

Εισαγωγή

Σκοπός της συγκεκριμένης διπλωματικής είναι η μελέτη του προβλήματος της κατάτμησης της φωνής με αυτόματες μεθόδους. Ιδιαίτερη έμφαση δίδεται σε μεθόδους που έχουν προταθεί διεθνώς και έχουν μελετηθεί σε βάθος. Μάλιστα αναλύεται διεξοδικά το θεωρητικό υπόβαθρο και δίδεται έμφαση στον μαθηματικό φορμαλισμό των ιδεών, συνακόλουθα δίδεται η διαισθητική όψη που αποτέλεσε το εναρκτήριο λάκτισμα για την ανάπτυξη αυτών των μεθόδων. Εν παραλλήλω η μελέτη γίνεται με τρόπο που προσπαθεί να συγκεράσει τον τρόπο της ανάπτυξης των ιδεών σε σχέση με το χρόνο, αλλά και το πώς αυτές εμπλουτίστηκαν, διευρύνθηκαν και ωρίμασαν από διάφορους επιστήμονες και ερευνητικές ομάδες ανά τον κόσμο. Σε επίπεδο υλοποίησης αναφέρονται τα πειράματα που πραγματοποιήθηκαν και τα αποτελέσματα που προέκυψαν με τις μεθοδολογίες καθώς και αναδεικνύονται όψεις του προβλήματος και πώς αυτές μπορούν να αντιμετωπιστούν.

1.1 Κίνητρα

Αποτελεί αδιαφιλονίκητο γεγονός ότι στην εποχή μας, η οποία από πολλούς αποκαλείται και εποχή της Γνώσης και Πληροφορίας, συνιστά μία νέα εποχή όπου οι υπολογιστές και εν γένει η τεχνολογία αποτελεί σημείο αιχμής και έχει καταλάβει σχεδόν ολοκληρωτικά κάθε κλάδο της ανθρώπινης δραστηριότητας και μάλιστα οι επιπτώσεις που είχε είναι άμεσα παρατηρησιμοί και αναγνώσιμοι ακόμη και στην καθημερινότητα μας.

Όπως αναφέρθηκε οι υπολογιστές αποτελούν αναπόσπαστο κομμάτι της καθημερινότητας και η ανάγκη για ανάπτυξη ενός εύρωστου τρόπου επικοινωνίας μεταξύ ανθρώπου και υπολογιστή έχει τραβήξει το ενδιαφέρον πολλών επιστημόνων, μηχανικών κτλ. Μάλιστα αυτή η διεπαφή ανθρώπου και υπολογιστή κατευθύνεται προς την υιοθέτηση τρόπων που είναι πιο κοντά στον φυσικό τρόπο μεταξύ ανθρώπων. Το κατεξοχήν μέσο επικοινωνίας των ανθρώπων αποτελεί ο λόγος, καθώς αυτός διακρίνεται σε γραπτό και προφορικό. Ιδιαίτερα ο προφορικός είναι ακόμη πιο φυσικός και γίνεται μέσω της ικανότητας του ανθρώπου για ομιλία που τον διακρίνει από τα υπόλοιπα όντα, ομιλία και εν γένει η φωνή είναι μία διαδικασία η οποία στον άνθρωπο μπορεί να γίνεται, αλλά η διαδικασία τόσο παραγωγής λόγου, όσο δε και η κατανόηση του από άλλο άνθρωπο είναι ιδιαιτέρως ακανθώδης και εξαιρετικής πολυπλοκότητας. Βέβαια ο άνθρωπος είναι ικανός μέσω της φύσης για την επίλυση αυτού της διαδικασίας με πολύπλοκα συστήματα όπως π.χ αυτί, φωνητικό σωλήνα και πάνω από όλα αυτά τα συστήματα τον εγκέφαλο, του οποίου η λειτουργία είναι ακόμη ανεξερεύνητη και άγνωστη, παρά τις επίπονες και ειλικρινείς προσπάθειες των επιστημόνων. Σε αυτή την προσπάθεια έχουν προταθεί διάφορες μέθοδοι για την μοντελοποίηση του συστήματος οι οποίες διακρίνονται για την πολυπλοκότητα τους η οποία βαθμιαία αυξάνεται λόγω του ότι αρχίζουμε να κατανοούμε καλύτερα το πρόβλημα, αλλά έχουμε και στην διάθεση μας μαθηματικά εργαλεία τα οποία συνεχώς επεκτείνονται, παράλληλα δίδεται η ικανότητα λόγω των υπολογιστών οι

οποίοι έχουν ολοένα και μεγαλύτερη ισχύ συνδυασμένη με τον σχεδιασμό όλο και πιο έξυπνων αλγοριθμικών τρόπων να ευρεθεί λύση στο πρόβλημα.

Κεντρικό ρόλο σε αυτή την προσπάθεια αποτελεί και η αναγνώριση φωνής, η οποία για πολλούς επιστήμονες αποτελεί και το Άγιο Δισκοπότηρο, το οποίο αναζητείται. Η φωνή είναι ένα σήμα το οποίο ενσωματώνει νόημα, συναισθηματική κατάσταση από τον ομιλητή και κυρίως διέπεται από πολλή μεγάλη μεταβλητότητα (variability). Η μεγάλη μεταβλητότητα είναι που κάνει το πρόβλημα ιδιαίτερος δύσκολο. Γενικά οι μεθοδολογίες που έχουν χρησιμοποιηθεί προσπαθούν να συγκεράσουν να αντιμετωπίσουν τους πολλούς βαθμούς ελευθερίας που έχει εγγενώς η φωνή αν την αντιμετωπίσουμε συστημικά. Η φωνή εγγενώς είναι ένα πλούσιο σήμα το οποίο ενσωματώνει πολλή πληροφορία σε πολλά επίπεδα και γι' αυτό όλες οι μέθοδοι γενικά αστοχούν να δημιουργήσουν ένα αποδοτικό και ορθό πλαίσιο εργασίας (framework).

Η έρευνα και η βιβλιογραφία στην αναγνώριση φωνής είναι ογκώδης, ζώσα και έχει ιστορικά τα θεμέλια της γύρω στο 1960, όπου οι υπολογιστές βρίσκονταν ακόμη σε εμβρυακό στάδιο και η ισχύς τους ήταν τάξεις μεγέθους χαμηλότερη από αυτή που διαθέτουμε σήμερα. Ειδικά στις μέρες μας που η ενσωμάτωση και η ολοκλήρωση υπολογιστικών συστημάτων είναι καθολική, υπάρχει ανάγκη για αποτελεσματική αναγνώριση φωνής και μάλιστα για ευρύ λεξιλόγιο, για συνεχή λόγο, ανεξαρτήτως παρουσίας θορύβου, για πολλούς ομιλητές και για πολλές γλώσσες. Οι τελευταίες απαιτήσεις κάνουν το πρόβλημα ακόμη πιο δύσκολο και ακανθώδες.

Η συγκεκριμένη διπλωματική εργασία κάνει επισκόπηση των μεθόδων που έχουν αναπτυχθεί και δοκιμαστεί και πως αυτές μπορούν να λύσουν το πρόβλημα της αυτόματης κατάτμησης που σχετίζεται άμεσα με την αναγνώριση φωνής. Λόγω της σπονδυλωτής αρχιτεκτονικής ενός τέτοιου συστήματος π.χ Επεξεργασία Σήματος, Αναγνώριση Προτύπων (Φωνημάτων), Ταυτοποίηση, Λεκτικά Μοντέλα κτλ., κάθε τμήμα (module) αναλύεται διεξοδικά και εφαρμόζεται μία πληθώρα μεθοδολογιών. Η προτεινόμενη μέθοδος της κατάτμησης της φωνής (ιδιαίτερα σε φωνήματα-phonemes-) είναι αρκετά χρήσιμη και γενικά έχει εφαρμογή για την εκπαίδευση (training process) ενός ολοκληρωμένου αυτόματου συστήματος αναγνώρισης φωνής-ASR (Automatic Speech Recognition). Αυτή η διαδικασία γίνεται ως επί το πλείστον χειροκίνητα από ανθρώπους και είναι ιδιαίτερος επίπονη και κοπιαστική εργασία.

1.2 Οργάνωση της Διπλωματικής

Με εντελώς αδρό τρόπο η διάσχιση των κεφαλαίων ακολουθεί ένα επαγωγικό τρόπο, που στηρίζεται από την αναγωγή του γενικού μέρους στο ειδικό και συνακόλουθα ανάλυση με όσο το δυνατό πιο ενδελεχή τρόπο. Προσπάθεια καταβλήθηκε έτσι ώστε να γίνεται ομαλή η μετάβαση του αναγνώστη από τις πιο απλές (primitive) σε πιο σύνθετες έννοιες και μεθοδολογίες οι οποίες συνιστούν ένα σύστημα αναγνώρισης φωνής. Ταυτόχρονα γίνεται κριτική σύγκριση των μεθόδων που έχουν αναπτυχθεί και υπογραμμίζονται κάποια χαρακτηριστικά αυτών όπως:

- Ομοιότητες, τόσο σε θεωρητικό υπόβαθρο, όσο και σε επίπεδο υλοποίησης.
- Διαφορές, οι οποίες κρίνονται ως προς την αποτελεσματικότητά τους.
- Δυνατότητα επεκτασιμότητας, όπου αυτό είναι εφικτό.

Συνεπώς αρχικά μελετώνται η φωνή τόσο σε επίπεδο παραγωγής από τον άνθρωπο, όσο και κάποια χαρακτηριστικά της τα οποία την διακρίνουν σαν σήμα από τα υπόλοιπα είδη και μάλιστα με θεμελιώδη τρόπο. Λόγω αυτής της ιδιαιτερότητας της φωνής σαν σήμα που φέρει πληροφορία μόνο σε μαθηματικό επίπεδο που θα το

εξηγήσουμε και θα εμβαθύνουμε στην συνέχεια της διπλωματικής ,αλλά παράλληλα και σε ένα πιο αφαιρετικό και ανώτερο επίπεδο που συνθέτει το λόγο και κατ' επέκταση μία γλώσσα, το κατ' εξοχή μέσο επικοινωνίας των ανθρώπων. Στην συνέχεια της εργασίας γίνεται επισκόπηση στις μεθόδους επεξεργασίας σημάτων και γίνεται μία πρώιμη αναφορά στην εξαγωγή χαρακτηριστικών(feature extraction)ειδικά αυτών που είναι πιο εύρωστες .Ιδιαίτερο βάθος δίνεται στο υπόβαθρο που διέπει αυτές τις μεθόδους και πως αυτές προέκυψαν χρονικά .Επιγραμματικά γίνεται διάκριση των μεθόδων επεξεργασίας που κυρίως διακρίνονται στις ακόλουθες περιοχές:

- Στατιστικές που έχουν κομψό και θεμελιωμένο υπόβαθρο από την περιοχή της Στατιστικής ,των Πιθανοτήτων ,αλλά και της Θεωρίας της Πληροφορίας, ειδικά η έννοια της εντροπίας .Μάλιστα μελετάται και η έννοια της εντροπίας κατά Kolmogorov,που αποτελεί ένα συναφή τρόπο, αλλά δομικά προσανατολισμένο σε άλλη κατεύθυνση σε σχέση με τον κλασσικό ορισμό του Shannon.Η σύγκριση αυτών των δύο εννοιών που προσπαθούν να χαρακτηρίσουν την τυχαιότητα ενός σήματος ,μάλιστα αποδεικνύονται ενδιαφέρουσες ιδιότητες .
- Ενεργειακές που ενσωματώνουν χρήσιμη πληροφορία για το σήμα. Χαρακτηριστικός εκπρόσωπος τέτοιας μεθόδου είναι η μέθοδος AM-FM .
- Μέθοδοι που προσπαθούν να ενσωματώσουν ακουστική(auditory) πληροφορία στο στάδιο επεξεργασίας. Ενδεικτικά αναφέρουμε τις μεθόδους για την εξαγωγή MFCC συντελεστών,RASTA,PLP κτλ. Μάλιστα ιδιαίτερη μνεία θα γίνει και στο ακουστικό σύστημα του ανθρώπου που προσπαθούν να προσομοιώσουν αυτά τα μοντέλα.
- Διάκριση μεταξύ γραμμικών και μη γραμμικών μεθόδων. Ιδιαίτερη προσοχή δίδεται στη μαθηματική σκοπιά αυτών των μεθόδων.Πρώιμα πρέπει να αναφερθεί ,ότι στη βιβλιογραφία οι γραμμικές μέθοδοι έχουν μελετηθεί ευρύτατα και λόγω απλότητας(σε σχέση πάντα με τις μη γραμμικές),αλλά παράλληλα έχουν εξαντληθεί και σε επίπεδο υλοποίησης. Βέβαια έχουν αναδειχθεί οι αδυναμίες που έχουν, αλλά και διάφορες ελκυστικές ιδιότητες τις οποίες διαθέτουν. Αντίστοιχα συγκρίνονται με τις μη γραμμικές ,οι οποίες όλο και εφαρμόζονται πιο συχνά, ειδικά στις μέρες μας ,λόγω του ότι αναδύονται θετικά αποτελέσματα, αλλά και της σύγχρονης ωρίμανσης σε μαθηματικό και θεωρητικό υπόβαθρο .
- Αναφορά σε σύγχρονες μεθόδους οι οποίες ακόμα παραμένουν σε πειραματικό στάδιο, αλλά έχουν στιβαρό μαθηματικό υπόβαθρο και πλέον στις μέρες λόγω της (σχετικά) υψηλής υπολογιστικής ισχύος σε σχέση με παλαιότερα ,υπάρχει η δυνατότητα για υλοποίηση σε ένα υπολογιστικό σύστημα. Χαρακτηριστικές αντιπρόσωποι τέτοιων μεθόδων είναι οι ακόλουθες:
 - ✓ Χαοτικές, που σχετίζονται με την μοντελοποίηση της φωνής σαν ένα δυναμικό σύστημα, του οποίου εξάγονται διάφορα χαρακτηριστικά όπως οι εκθέτες Lyapunov που χαρακτηρίζουν την ευστάθεια του μοντέλου. Μάλιστα η μοντελοποίηση με βάση τα δυναμικά συστήματα, προσπαθεί να αντιμετωπίσει ένα ανοιχτό πρόβλημα γενικά στον χώρο της αναγνώρισης της φωνής που σχετίζεται με την μοντελοποίηση του φωνητικού σωλήνα και έχει τις ρίζες της στην κατασκευή αεροστροβίλων(jet engines) και γίνεται μία αναγωγή των εξισώσεων.

- ✓ Μέθοδοι βασισμένοι στα fractal και συγκεκριμένα την εξαγωγή της fractal διάστασης ενός σήματος. Μάλιστα αυτές οι μέθοδοι βοηθούν διότι δίνουν μία πρώτη μορφή προσέγγισης και χαρακτηρισμού ήχων όπως των fricative ήχων, κρουστικών ήχων (π.χ /s/ ,/t/). Αξίζει να σημειωθεί ότι δεν υπάρχει πλήρως ανεπτυγμένο μαθηματικό μοντέλο και υπόβαθρο ακόμα και στις μέρες μας.

Εν συνεχεία γίνεται μία επισκόπηση των μεθόδων αναγνώρισης προτύπων και μηχανικής μάθησης. Ειδικότερα θα γίνει μία αδρή αναφορά σε παραμετρικές και μη παραμετρικές μεθόδους που έχουν αναπτυχθεί. Εν συνεχεία γίνεται εμβάθυνση στις μεθόδους που έχουν χρησιμοποιηθεί με επιτυχία στο πρόβλημα της αναγνώρισης φωνής. Ενδεικτικά θα γίνει λεπτομερή μνεία στα HMM(Hidden Markov Models),στα ANN(Artificial Neural Networks),σε υβριδικές μεθόδους(hybrid methods)που προσπαθούν να συγκεράσουν τα πλεονεκτήματα των δύο μεθόδων και να υπερκεράσουν τις εγγενείς αδυναμίες των 2 παραπάνω μεθόδων. Συνακόλουθα γίνεται αναφορά στα SVM(support vector machines),οι οποίες είναι σχετικά πρόσφατες και έχουν ιδιαίτερα ενθαρρυντικά αποτελέσματα.

Στο επόμενο κεφάλαιο γίνεται ανάλυση σε μοντέλα γλώσσας τα οποία βοηθούν σε ανώτερα επίπεδα ανάλυσης της αναγνώρισης λέξεων ,φωνημάτων κτλ. Μάλιστα έχουν αποδειχθεί διάφορες ενδιαφέρουσες ιδιότητες οι οποίες έχουν πρακτική αξία, αλλά παράλληλα έχουν αποδειχθεί και ιδιότητες που έχουν αρνητικό αποτέλεσμα, όχι ως προς την αδυναμία αυτών των μοντέλων, αλλά σε σχέση με την πολυπλοκότητα τους και συγκεκριμένα την υπολογιστική.

Επιπρόσθετα αναλύεται η μέθοδος που εφαρμόζεται σε αυτή την διπλωματική προκειμένου να επιλυθεί το πρόβλημα. Αναλύεται το θεωρητικό υπόβαθρο και η υλοποίηση της και γίνεται μία πρώτη συγκομιδή των αποτελεσμάτων που προκύπτουν. Παράλληλα θέματα όπως η πολυπλοκότητα ,η ορθότητα και η ευρωστία του αλγορίθμου επιλύονται αναλυτικά με φορμαλιστικό τρόπο όπως επιτάσσει η θεωρία των αλγορίθμων και η θεωρία υπολογισμού.

Συνακόλουθα γίνεται ανάλυση των αποτελεσμάτων και διαπιστώνονται επιτυχίες και αστοχίες της μεθοδολογίας. Παράλληλα τονίζονται πιθανοί τρόποι αντιμετώπισης αυτών των αστοχιών και την εγγενή τους φύση όπου αυτή υπάρχει,π.χ αδυναμία υπολογιστική ως προς την αριθμητική των αλγορίθμων κατά την υλοποίησή τους, αλλά και εγγενείς αδυναμίες που οφείλονται σε υποθέσεις κατά την μοντελοποίηση κτλ.

Λίγο πριν το επίλογο της εργασίας γίνεται επισκόπηση των πιο σημαντικών προβλημάτων και πως αυτά επιλύονται και πως οι τεχνικές που έχουν εφαρμοστεί αντιπαραβάλλονται με αυτή που υλοποιήθηκε και γίνεται μία κριτική σύγκριση και σύνθεση των επιμέρους ιδεών.

Εν κατακλείδι στον επίλογο της εργασίας γίνονται μία προσπάθεια για αναφορά προτάσεων για μελλοντική έρευνα προς την προσπάθεια επίλυσης του προβλήματος. Γνώμονας των προτάσεων αποτελεί η ολική αντιμετώπιση της φωνής σαν ένα σήμα το οποίο περιέχει χαρακτηριστικά και πληροφορία σε πολλά επίπεδα και διαστάσεις και εν γένει η γνώση της φυσιολογίας του φωνητικού συστήματος και ακουστικού το οποίο έχει ο άνθρωπος και πως μπορούν οι ερευνητές να εμπνευστούν από την φύση και να προσπαθούν να την προσεγγίσουν, ενσωματώνοντας γνώση από πολλούς τομείς που τέμνονται σε κάποιο σημείο και η συνάρθρωση τους συνιστά την πολυποίκιλη φύση του προβλήματος.

Ως επίλογος της διπλωματικής αποτελεί η πλήρης αναφορά της βιβλιογραφίας που συνίσταται από τα βιβλία ,τα επιστημονικά άρθρα ,εργασίες κυρίως σε επίπεδο

διδασκαλίας καθώς και έτοιμο λογισμικό το οποίο χρησιμοποιήθηκε σε επικουρικό επίπεδο από τον συγγραφέα για την ολοκλήρωση της εργασίας.

Κεφάλαιο 2

Η φωνή ως σήμα και το υπόβαθρο της

2.1 Εισαγωγή

Οι φωνητικοί ήχοι είναι η απήχηση των ταλαντώσεων της πίεσης του αέρα που εκπνέεται από τους πνεύμονες και διαμορφώνονται (Modulated) και προσαρμόζονται (shaped) σε σχήμα από τις ταλαντώσεις των φωνητικών χορδών (glottal chords) καθώς και την αντήχηση ή συντονισμό (resonance) του φωνητικού σωλήνα καθώς ο αέρας πιέζεται μέσω των χειλιών και της μύτης.

Η φωνή είναι πολύ πλούσιο σήμα σε πληροφορία το οποίο εκμεταλλεύεται την διαμόρφωση σε συχνότητα (Frequency-modulated), διαμόρφωση σε πλάτος (amplitude-modulated) και σε φέρων διαμορφωμένο στο χρόνο (and time-modulated carriers) π.χ αρμονικές και θόρυβος, διάρκεια ισχύος, κυματισμός του pitch, μετακινήσεις της αντήχησης κτλ, για να μεταφέρει πληροφορία σχετικά με λέξεις, την ταυτότητα του ομιλητή, προφορά, συναισθηματική κατάσταση, αλλά και κατάσταση υγείας του. Από μία τηλεπικοινωνιακή σκοπιά όλη αυτή η πληροφορία μπορεί να εκμαιευθεί από το παραδοσιακό τηλέφωνο με εύρος 4KHz, η ενέργεια της φωνής πάνω από αυτό το εύρος μπορεί να μεταφέρει πληροφορία για την ποιότητα του ήχου, αλλά και την αίσθηση (sensation).

Σε αυτό το κεφάλαιο θα μελετήσουμε τις θεμελιώδεις αρχές της παραγωγής και αντίληψης (perception) φωνής. Θα μελετηθούν οι μηχανισμοί που παράγουν και μεταφέρουν φωνητικούς ήχους ομιλίας και θα εξεταστούν οι ακουστικοί συσχετισμοί. Η φωνή όπως αναφέρθηκε είναι ένα σύνθετο σήμα το οποίο παράγεται σαν αποτέλεσμα πολλών μετασχηματισμών που συμβαίνουν σε διάφορα επίπεδα και επεκτείνοντας την προηγούμενη αναφορά, όπως:

- ο Σημασιολογικό (semantic) και έχει σχέση με το νόημα που εξαγάγεται απ' την ομιλία / λόγο κάποιου ομιλητή.
- ο Γλωσσολογικό (linguistic), που σχετίζεται με την γλώσσα αυτή καθ' αυτή (per se) και πιο συγκεκριμένα και με αδρό τρόπο σε σχέση με το συντακτικό, γραμματική κτλ και γενικά με κανόνες που διέπουν κάποια γλώσσα. Είναι πασιφανές ότι οι γλώσσες γενικά διαφέρουν ω προς συντακτικό τους και γι' αυτό π.χ η μετάφραση από μία γλώσσα σε κάποια άλλη δεν είναι πάντοτε εφικτή και αποτελεσματική, ακόμη και αν την αναλαμβάνει αυτή τη διεργασία ένας ειδήμων.
- ο Εναρθρωτικό (articulatory), και σχετίζεται με την άρθρωση και τον τρόπο που εκφέρει κάποιος τις λέξεις. Μάλιστα χαρακτηριστική είναι και η περίπτωση ομιλίας κάποιου μίας γλώσσας η οποία δεν είναι η μητρική του, π.χ να μιλήσει ένας Έλληνας αγγλικά και το αντίστοιχο κάποιος Άγγλος ελληνικά.
- ο Ακουστικό (acoustic), το οποίο συσχετίζεται με το αυτί και γενικά με τις διεργασίες που συμβαίνουν σε αυτό (auditory process). Μάλιστα πολλές μεθοδολογίες και ιδιαίτερα οι πιο σύγχρονες προσπαθούν να

μοντελοποιήσουν αυτή τη βιολογική διαδικασία. Ιδιαίτερα έχουν ανακλύψει νέοι μέθοδοι επεξεργασίας της φωνής, αλλά ακόμα και αναπαράσταση της. Διαφορές σε αυτά τους μετασχηματισμούς εμφανίζονται ως διαφορές στις ακουστικές ιδιότητες του σήματος της φωνής και έτσι εξηγείται η μεγάλη διαφορετικότητα που παρατηρούμε στην φωνή. Οι συσχετισμένες με ομιλητή διαφορές είναι αποτέλεσμα του συνδυασμού των ανατομικών διαφορών που είναι εγγενείς στον φωνητικό σωλήνα (vocal tract) καθώς και των ομιλητικών συνηθειών των διαφόρων ατόμων. Στην αναγνώριση φωνής και σε πιο γενικό στην αναγνώριση ομιλητών μπορούν να χρησιμοποιηθούν για να επιτύχουμε την επιθυμητή διαφοροποίηση

2.2 Επικοινωνία μέσω φωνής

Η φωνή είναι από τους πιο φυσικούς τρόπους επικοινωνίας των ανθρώπων. Πρόκειται από τα πιο σημαντικά σήματα που ενσωματώνουν πληροφορία, οι ήχοι φωνής έχουν πλούσια και πολύεπίπεδη φασματική και χωρική (spectra-temporal) μεταβολή που εκφράζει λέξεις, φύλο, ηλικία, τύπο ομιλίας, και κατάσταση υγείας και ψυχισμού. Οι ήχοι παράγονται από ταλαντώσεις της πίεσης του αέρα που πιέζονται από αέρα που έχει γίνει εισπνοή από τους πνεύμονες από τις ταλαντωμένες φωνητικές χορδές και του φωνητικού σωλήνα και έξω από τους αεραγωγούς των χειλίων και της μύτης. Ο αέρας διαμορφώνεται από τις ταλαντώσεις των φωνητικών χορδών, την αντήχηση του φωνητικού σωλήνα και των ρινικών κοιλοτήτων, την θέση της γλώσσας και του ανοίγματος και κλεισίματος του στόματος.

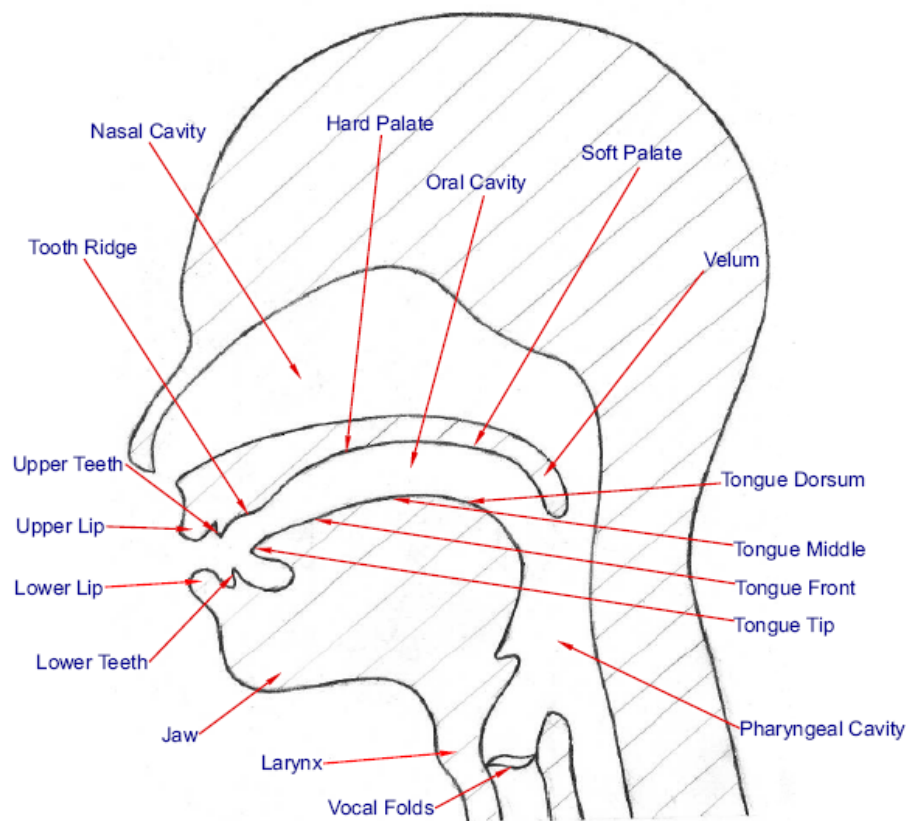
Κατά αντιστοιχία με την γραμμένη μορφή μίας γλώσσας που είναι σαν ακολουθία ενός στοιχειώδους αλφαβήτου, για την φωνή είναι επίσης μία ακολουθία από στοιχειώδεις ακουστικούς ήχους ή σύμβολα τα οποία είναι γνωστά ως φωνήματα και μεταφέρουν την φωνητική μορφή μίας γλώσσας. Για τα ελληνικά υπάρχουν 32 τέτοιοι ήχοι, αντίστοιχα για τα αγγλικά υπάρχουν 40-60 τέτοια φωνήματα. Από αυτό το στοιχειώδες σύνολο μπορεί να δημιουργηθεί ένας πολύ μεγάλος αριθμός ακουστικών λέξεων. Πρέπει όμως να τονίσουμε ότι στην πράξη ότι η παραγωγή κάθε φωνητικού ήχου επηρεάζεται άλλοτε σε μεγάλο βαθμό και άλλοτε σε μικρότερο από το περιεχόμενο το γειτονικών φωνημάτων.

Τα σήματα φωνής μεταφέρουν πολύ περισσότερο νόημα απ' ό,τι κάποιες λέξεις που εκφωνούνται. Η πληροφορία που μεταφέρεται από την φωνή όπως ειπώθηκε είναι πολυεπίπεδη και περιέχει διαμόρφωση συχνότητας και χρόνου του φέροντος πληροφορίας όπως τα formants και του κυματισμού του pitch. Με εντελώς αδρό μπορούμε να πούμε ότι τα formants είναι οι αντηχήσεις του φωνητικού σωλήνα και το pitch περιγράφει την θεμελιώδη συχνότητα του ανοίγματος και του κλεισίματος των γλωττικών πτυχώσεων/ τοιχωμάτων (glottal folds). Η πληροφορία που παρέχεται μέσω φωνής περιέχει τα ακόλουθα:

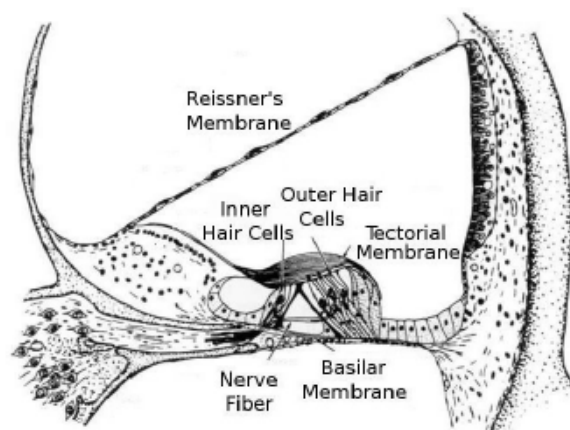
1. Ακουστικά φωνητικά σύμβολα. Πρόκειται για τις στοιχειώδεις ακουστικές μονάδες από τις οποίες μεγαλύτερες ηχητικές μονάδες ομιλίας όπως λέξεις και συλλαβές προκύπτουν. Μερικές λέξεις έχουν μόνο 2 φωνήματα όπως οι λέξεις 'me', 'you'.
2. Προσωδία. Είναι μία έννοια συνυφασμένη με τον ρυθμό της φωνής, τα πρότυπα τονισμού (stress patterns). Η προσωδία βοηθά να εξαγάγουμε πληροφορία σχετικά με τα όρια μεταξύ τμημάτων φωνής, να υπογραμμίσουμε υποφράσεις και να ξεκαθαρίσουμε την πρόθεση και να αναιρέσουμε ασάφειες όπως αν η φράση είναι κατάφαση ή ερώτηση. Η προσωδία παίζει σπουδαίο ρόλο στην αντίληψη της ομιλίας και επιπρόσθετα μεταφέρει πληροφορία για τον ομιλητή. Εν γένει η μοντελοποίηση της προσωδίας είναι χρήσιμη για την

σύνθεση φωνής(speech synthesis), αλλά πρόκειται για δύσκολο πρόβλημα και για μία προβληματική διαδικασία , ιδιαίτερα στην περίπτωση μεγάλου λεξιλογίου. Συσχετίζεται με το χρονισμό , την βασική συχνότητα και με τη ροή της φωνής.

3. Πληροφορία για το φύλο. Το φύλο μπορεί να καθοριστεί από το pitch(που σχετίζεται με τη θεμελιώδη συχνότητα των έμφωνων ήχων)και το μέγεθος και τα φυσικά χαρακτηριστικά του φωνητικού σωλήνα. Λόγω των διαφορών στην φωνητική ανατομία ,η γυναικεία φωνή έχει υψηλότερες συχνότητες αντήχησης, αλλά και υψηλότερο pitch.
4. Ηλικία. Αυτό οφείλεται στο μέγεθος, αλλά και την ελαστικότητα των φωνητικών χορδών, σωλήνα και του pitch. Το pitch των παιδιών μπορεί να είναι παραπάνω από 300HZ.
5. Προφορά, η οποία μπορεί να διαγνωστεί από :
 - Μεταβολές στο λεξιλόγιο προφοράς με τη μορφή υποκατάστασης, διαγραφής και εισαγωγής φωνημάτων στην τυπική εγγραφή των λέξεων
 - Τις συστηματικές μεταβολές στις συχνότητες αντήχησης(formants),στον κυματισμό του pitch, διάρκεια, στην έμφαση και τον τονισμό.
6. Η ταυτότητα του ομιλητή προκύπτει από τα φυσικά χαρακτηριστικά των πτυχώσεων του φωνητικού σωλήνα ,το Pitch και το στυλ ομιλίας.
7. Ψυχισμός και υγεία, μέσω του φάσματος του φωνητικού σωλήνα και της δυναμικής του pitch,της διάρκειας και της αντήχησης του φωνητικού σωλήνα.



Σχήμα 1 Ανατομία των κυριότερων οργάνων παραγωγής ομιλίας



Κεφάλαιο 3

Μερικές σημειώσεις για της μη- γραμμικότητες της φωνής

3.1 Μη-γραμμικότητες στην παραγωγή της φωνής:

Η ομιλία είναι η κινητήριος ικανότητα η οποία αποτελείται από ελεγχόμενες και συντονισμένες κινήσεις οι οποίες πραγματοποιούνται κυρίως από τα όργανα του φωνητικού σωλήνα που ενεργούν πάνω στον αέρα στις αναπνευστικές οδούς και παράγουν φωνητικούς ήχους. Τα φωνητικά όργανα παράγουν τοπικές διαταραχές στα μόρια του αέρα σε διάφορα σημεία της φωνητικής οδού και δημιουργούν τις πηγές για την παραγωγή φωνητικών ήχων. Οι πιο κοινοί φωνητικοί παραγωγοί είναι 1) η ημι-περιοδική ταλάντωση των φωνητικών χορδών -ηχητική πηγή, 2) ο τυρβώδης θόρυβος που δημιουργείται από την μετάβαση του αέρα μέσα από μία ημι-περιορισμένη στενωπό στην φωνητική κοιλότητα-τυρβώδης πηγή, 3) ο εκρηκτικός θόρυβος που ακολουθεί την αποδέσμευση του αέρα πίσω από την ολική εμπέδηση της φωνητικής κοιλότητας-μεταβατική πηγή. Παρ'όλα αυτά η πολύπλοκη δομή των φωνητικών ήχων δεν οφείλεται μόνο στα χαρακτηριστικά των πηγών που την παράγουν, αλλά κυρίως από τα χαρακτηριστικά της απόκρισης του φωνητικού σωλήνα που βασίζεται κυρίως στην μορφή του και την κατάσταση του. Η κατάσταση του φωνητικού σωλήνα αλλάζει σύμφωνα με την κίνηση των φωνητικών οργάνων που διαμορφώνουν κυρίως το μήκος του, την χαρακτηριστική του απόκριση και τις περιοχές που τέμνονται. Η δομή των φωνητικών ήχων δημιουργείται από την συνδιασμένη δράση των ηχητικών πηγών και των χαρακτηριστικών του φωνητικού σωλήνα.

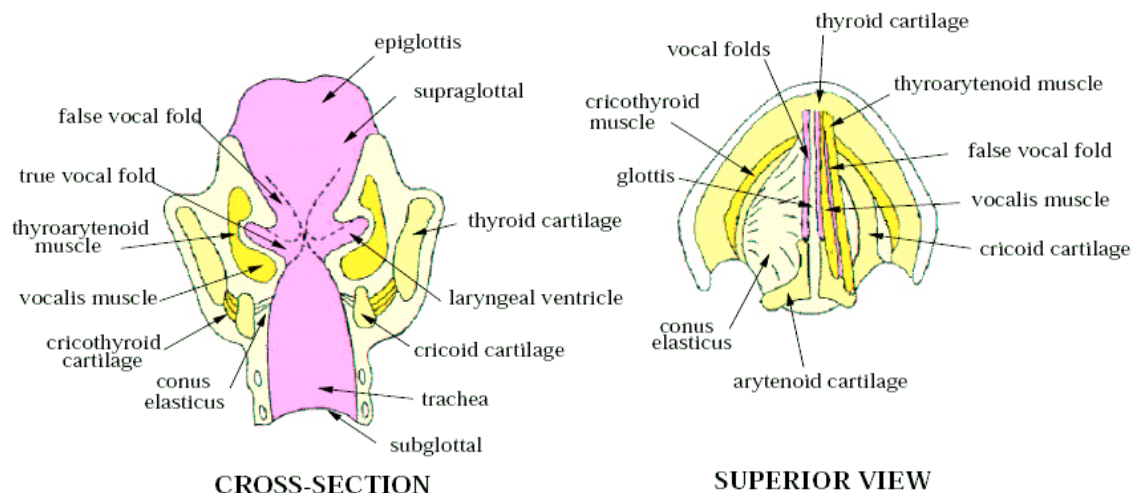
Η απουσία των ήχων, ο φωνητικών σωλήνας μπορεί να μοντελοποιηθεί σαν ένας απλός σωλήνας και τα μόρια του αέρα μπορούν να μοντελοποιηθούν σαν ένας γραμμικός ταλαντωτής ο οποίος ανταποκρίνεται στις διαταραχές με μικρές μεταβολές από την υπόλοιπη θέση. Οι συνθήκες είναι πολύ πιο πολύπλοκες όταν οι φωνητικοί ήχοι παράγονται, διότι η κίνηση των φωνητικών οργάνων μεταβάλλει το σχήμα του φωνητικού σωλήνα. Ένα αρκετά αδρό μοντέλο του φωνητικού σωλήνα σε αυτές τις συνθήκες είναι ένα σύνολο επικαλυπτόμενων σωληνών διαφορετικού μήκους και τεμνόμενων περιοχών. Λόγω αυτών των χαρακτηριστικών, ο κάθε σωλήνας υπόκειται σε διαφορετική πίεση αέρα, η οποία με την σειρά της δημιουργεί διαφορετικές δυνάμεις που ενεργούν πάνω στα μόρια του αέρα και προκαλούν αρκετά πολύπλοκη κίνηση. Σε αυτή την περίπτωση μία γραμμική περιγραφή αποτυγχάνει να περιγράψει την πολύπλοκη δυναμική και είναι επιτακτικό να χρησιμοποιηθούν μη γραμμικές τεχνικές. Παρ'όλα αυτά οι μη γραμμικότητες εισαγάγουν αβέβαια και πολλαπλές λύσεις στα διαφορετικά προβλήματα φωνής. Σαν ένα παράδειγμα θα αναφέρουμε το πρόβλημα γνωστό ως “ακουστική προς άρθρωση απεικόνιση” (acoustic to articulatory mapping). Κυρίως συνίσταται στην αναγνώριση ενός φωνητικού σωλήνα και του μοντέλου του, το οποίο αποτελείται από επικαλυπτόμενους σωλήνες, ένα σύνολο παραμέτρων τέτοιο ώστε οι ηχητικοί συντονισμοί (resonance) να αντιστοιχούν στα αντίστοιχα formants που μπορούν να παρατηρηθούν σε ένα παραγόμενο ήχο φωνής. Έχει παρατηρηθεί ότι συγκεκριμένα σχήματα φωνητικού σωλήνα αντιστοιχούν και παράγουν συγκεκριμένο σύνολο formants συχνοτήτων, οπότε ένα συγκεκριμένο σύνολο formants δε μπορεί να αναγνωριστεί μόνο ηχητικά από το σχήμα του φωνητικού σωλήνα που τις παρήγαγε (αντιστροφο πρόβλημα-inverse problem). Εν γένει υπάρχουν άπειρες λύσεις για ένα αντιστροφο πρόβλημα. Ακόμα και αν εισαγάγουμε

συναρτησιακούς περιορισμούς (όπως ελάχιστη ανακατασκευή ή ελαχιστο ρυθμό ανακατασκευής του φωνητικού σωλήνα ως προς ένα δεδομένο σχήμα-minimal deformation or minimal deformation rate of the vocal tract about reference shape), η αντίστροφη απεικόνιση δεν διορθώνει πάντοτε το μοντέλο, δηλαδή να μας δώσει την κατάσταση/σχήμα του φωνητικού σωλήνα που παρήγαγε τον ήχο. Από μελέτες που έχουν γίνει κυρίως στην αμερικανική αγγλική γλώσσα για την μέτρηση των συχνοτήτων formants σε σχέση με την ακουστική-αρθρωτική απεικόνιση Schoentgen, επεδείξε κάποια προβλήματα στην υλοποίηση αυτής της ιδέας και με μεθοδολογίας.

Έτσι κυρίως έγινε λαμβάνοντας υπόψιν τα ακόλουθα: 1) η ακρίβεια του σχήματος του φωνητικού σωλήνα για να απεικονιστούν οι συχνότητες των formants και η σύγκριση με πραγματικά σχήματα του φωνητικού σωλήνα, 2) τα μοντέλα σχεδίασης του φωνητικού σωλήνα, 3) η αριθμητική ευστάθεια και ευρωστία. Τα αποτελέσματα αποδεικνύουν ότι μια καλή προσέγγιση μπορεί να επιτευχθεί μόνο για φωνητικούς ήχους που παράγονται με ένα απλή μοντελοποίηση του φωνητικού σωλήνα ως "απλή κοιλοτητα, απλούς περιορισμούς, κυρτή γλώσσα και επίσης περιορισμένη στο λάρυγγα-φάρυγγα και στα χείλη", τα μοντέλα βασίζονται κυρίως σε μικρό αριθμό κωνικών σωληνιδίων τα οποία συνεχώς αλλάζουν και εκτιμώνται σε σχέση με τα κυλινδρικά ή εκθετικά σωληνίδια, η σύγκλιση ως προς το επιθυμητό formant μπορεί να επιτευχθεί με ακρίβεια μικρότερου του 1Hz ακόμα και αν τα εκτιμώμενα σχήματα του φωνητικού σωλήνα διαφέρουν και ποιοτικά, αλλά και ποσοτικά από αυτά που προκύπτουν οι παρατηρούμενες formant συχνότητες.

Ένα ακόμη ανοικτό πρόβλημα είναι αυτό της προσέγγισης του γλωττικού κύκλου κυματομορφής (glottal cycle waveform), δηλαδή της μορφής της ροής του αέρα από την γλωττίδα, πριν το σήμα διαμορφωθεί από την επίδραση της κατάστασης του φωνητικού σωλήνα. Η κυματομορφή που προκύπτει από την γλωττίδα είναι πολύ σημαντική διότι διαμορφώνει την ποιότητα της προκύπτουσας φωνής και ορίζεται ως "η χαρακτηριστική ακουστική χρωματικότητα της φωνής ενός ατόμου και προκύπτει από μια ποικιλία χαρακτηριστικών του λάρυγγα και ρέει συνεχώς μέσω της φωνής κάποιου ατόμου". Ένας αποτελεσματικός συνθέτης ήχου πρέπει να είναι ικανός να ελέγχει ικανοποιητικά την ποιότητα της φωνής τέτοιος ώστε η φωνή που προκύπτει να έχει φυσικό και διακριτικό τόνο. Αρκετές μεθοδολογίες έχουν γίνει για να επιτευχθεί η μοντελοποίηση και κυρίως ώστε να βρουν χαρακτηριστικά του γλωττικού κύκλου που να είναι συμβατά με την μεταβλητή ποιότητα της φωνής. Η πιο ευθεία και αυτοματοποιημένη τεχνική που δεν βασίζεται σε τέτοια χαρακτηριστικά είναι μέσω της αντίστροφης τεχνικής φίλτρων (inverse filtering technique), η οποία απαιτεί εγγραφή του γλωττικού αεραγωγού στο στόμα και γίνεται επεξεργασία με ένα σύστημα φίλτρων που διαχωρίζει τα χαρακτηριστικά της φωνητικής οδού από αυτά της γλωττικής πηγής. Παρα ταύτα λόγω των μη γραμμικοτήτων που προκύπτουν στην συνάρτηση μεταφοράς του φωνητικού σωλήνα, αυτός ο διαχωρισμός δεν είναι εύκολος και ευθύς και τα προκύπτοντα μοντέλα δεν μπορούν να διαχειριστούν διαφορα χαρακτηριστικά ποιότητας της φωνής. Οι μη γραμμικότητες οφείλονται σε διάφορους λόγους, μεταξύ αυτών είναι και η διάδραση της φωνητικής οδού με την ρινική όδο (κυρίως για ένρινους ήχους), και επίσης η παρουσία διαφόρων πηγών ακουστικής ενέργειας μέσα στον φωνητικό σωλήνα (όπως εμπόδια σε συγκεκριμένα σημεία), όπως οι διαφορές των καταστάσεων στην διαδικασία ανόγματος και κλεισίματος στον κύκλο ταλάντωσης, η παρουσία ακανονικοτήτων σε αναπνευστικούς ήχους και πολλά άλλα. Οι παραπάνω μη γραμμικότητες μπορούν να προκαλέσουν αρκετά σφάλματα στην προσέγγιση της γλωττικής κυματομορφής κυρίως λόγω της μη επάρκειας των μαθηματικών μοντέλων τα οποία δεν λαμβάνουν

υπόψην όλες τις μη γραμμικότητες που προαναφέραμε και γιατί δεν μπορούν να προσεγγίσουν επακριβώς τα πραγματικά δεδομένα. Πολλές απόπειρες έχουν γίνει για να χρησιμοποιηθούν αυτές οι μέθοδοι σε διάφορα στάδια επεξεργασίας της φωνής και κυρίως να μην απαιτούν την ανθρώπινη παρέμβαση, διότι αυτό τις κάνει πιο εύρωστες για μια ευρεία γκάμα εφαρμογών.



3.2. Μη γραμμικότητες στην ακουστική της Φωνής

Όπως όλα τα περισσότερα πραγματικά σήματα, οι ήχοι της φωνής είναι αρκετά σύνθετοι στις μορφές τους, τόσο και στην εφαρμοσμένη όσο και στην βασική έρευνα, υπάρχει έντονη κινητικότητα για να βρεθεί η κατάλληλη ψηφιακή καταγραφή της φωνής. Από την πρακτική σκοπιά, μία κατάλληλη ψηφιακή απεικόνιση της φωνής θα μπορούσε να λύσει τα προβλήματα σύνθεσης φωνής και αναγνώρισης και θα επιτρέψει την εύκολη διάδραση ανθρώπου και Η/Υ. Περισσότερα, στην βασική έρευνα, θα βοηθήσει στην κατανόηση των διαδικασιών της φωνής και είναι εύφορο έδαφος για την έρευνα των επιστημόνων που ασχολούνται με την επικοινωνία μέσω φωνής (speech communication science).

Η πιο απλή ψηφιακή αναπαράσταση της φωνής σχετίζεται με την ηχητική κυματομορφή και βασίζεται στο θεώρημα του Nyquist το οποίο δηλώνει ότι οποιοδήποτε σήμα περιορισμένο σε ένα συγκεκριμένο εύρος, το οποίο μπορεί να ανασκευαστεί επακριβώς - με το να λαμβάνουμε περιοδικά δείγματα και ο ρυθμός είναι ο διπλάσιος από την υψηλότερη συχνότητα του σήματος - . Βασισμένα σε αυτή την ιδέα υπάρχουν τεχνικές όπως PCM (Pulse Code Modulation), DM (Delta Modulation) & DPCM (Differential Pulse Code Modulation). Παρ' όλα αυτά σε πολλές εφαρμογές φωνής το ενδιαφέρον μας δεν συνίσταται στην ανασκευή του ήχου όσο στο να αναπαραστήσουμε σαν ένα σύνολο χαρακτηριστικών τα οποία με κομψό, συμπαγή και εύρωστο τρόπο εκφράζουν τις ιδιότητες του σήματος. Όλες οι παραμετρικές αναπαραστάσεις της φωνής βασίζονται κυρίως στην ιδέα της ανάλυσης μικρού-χρόνου (short time analysis), π.χ σε ένα μικρό διάστημα της τάξεως των 10-30ms, θεωρούμε ότι το σήμα είναι αναλλοίωτο (invariant), όμως με την έννοια σε κάποιο χαρακτηριστικό του, π.χ να είναι γκαουσιανό, να είναι στάσιμο κτλ. Με αυτό

τον τρόπο ουσιαστικά λαμβάνουμε τα χαρακτηριστικά του σήματος σε μικρά διαστήματα όπου θεωρούμε ότι το σήμα δεν επηρεάζεται από μεταβολές.Συνακόλουθα αυτά τα χαρακτηριστικά χρησιμοποιούνται από ένα αλγόριθμο για περαιτέρω ανάλυση κτλ.Γενικά όπως θα δούμε πιο επισταμένα στην συνέχεια της εργασίας .

Η ανάλυση βραχέως χρόνου επιτρέπει να αναγνωριστούν χρήσιμες φωνητικές παραμέτρους από απλά χαρακτηριστικά της κυματομορφής τα οποία διακρίνονται στο πεδίο του χρόνου και στο πεδίο της συχνότητας:

Πεδιο Χρόνου:

α)Κορυφές,

β)Ενέργεια σήματος

γ)Ρυθμός όπου το σήμα μηδενίζεται(zero crossing rate)

δ)Ανάλυση βασισμένη στην αυτοσυσχέτιση (autocorrelation analysis).

Στην συνέχεια της εργασίας θα περιγράψουμε αναλυτικότερα αυτά τα χαρακτηριστικά.Αδρά αξίζει να δούμε ότι η διαίσθηση για τον τρόπο λειτουργίας τους βασίζεται κυρίως στο γεγονός ότι μετρούν απλά χαρακτηριστικά με φυσικό και στατιστικό ενδιαφέρον,γι'αυτό το λόγο χρησιμοποιούνται συχνότατα,διότι είναι πολύ απλές,αλλά παράλληλα έχουν και πολλά μειονεκτήματα.

Πεδίο συχνότητας:

Κυρίως αυτά τα χαρακτηριστικά βασίζονται πάνω στον μετασχηματισμό Fourier και χρησιμοποιούνται σε μία συστοιχία βαθυπερατών φίλτρων ή κάνοντας χρήση του ταχύ μετασχηματισμού Fourier(FFT) ο οποίος έχει πολυπλοκότητα $O(n \log n)$ και ανακαλυφθηκε απο τον Tucker et al.Το προκύπτον φάσμα φραχέως χρόνου είναι το πρώτο βήμα πριν χρησιμοποιηθούν μετέπειτα για πιο σύνθετους υπολογισμούς στους αλγορίθμους.Βασικά αυτή η μεθοδολογία βασίζεται στην ομομορφική επεξεργασία φωνής(homomorphic speech processing) σύμφωνα με την οποία μία διαδικασία που βασίζεται στην χρονική μεταβολή της παραγωγ'ης φωνής μπορεί να αντιμετωπιστεί στη βραχέως χρόνου ανάλυση σαν η συνέλιξη(convolution) της πηγής (σαν τυχαίος θόρυβος για τους τυρβώδεις και μεταβατικούς ήχους και ημιπεριοδικό συρμός παλμών για μία φωνητική πηγή) με την κρουστική απόκριση του φωνητικού σωλήνα.Το αποτέλεσμα αυτής της ανάλυσης καλείται cepstrum και βασίζεται στην αθροιστική συνδιάση του φωνητικού σωλήνα και της πηγής ήχου τα οποία δεν επικαλύπτονται και μπορούν εύκολα να διαχωριστού λόγω του ότι τα τμήματα μικρού χρόνου περιέχουνβ πληροφορία σχετικά με την συνάρτηση μεταφοράς του φωνητικού σωλήνα και τα δείγματα μεγαλύτερου χρόνου παρέχουν πληροφορία σχετικά με την πηγή.

Άλλες χρήσιμες μέθοδοι για την ανάλυση και κωδικοποίηση της φωνής βασίζονται στην ιδέα της γραμμικής πρόβλεψης(Linear prediction).Η ιδέα πίσω από αυτή την τεχνική είναι ότι θεωρούμε τα φωνητικά δείγματα σαν την έξοδο ενός ψηφιακού φίλτρου το οποίο αλλάζει χρονικά και του οποίου η συνάρτηση μεταφοράς προσεγγίζει την φασματικές ιδιότητες και του φωνητικού σωλήνα ,αλλά και του γλωττικού παλμού.

Η ευρεία ποικιλία της αναπαράστασης της φωνής μπορεί να χρησιμοποιηθεί για διάφορες εφαρμογές επεξεργασίας φωνής όπως αναγνώριση φωνής,σύνθεση φωνής,προσαρμογή σε ομιλητή(speaker adaptation).Παρά το γεγονός ότι οι αυτές οι μέθοδοι προκάλεσαν πρόοδο σε αυτήν την περιοχή τα τελευταία 30 χρόνια,δυστυχώς παραμελούν(ίσως και εσκεμμένα)την μηγραμμική φύση των σημάτων φωνής,βέβαια ο λόγος είναι οτι και τα μηγραμμικά μαθηματικά δεν έχουν “ωριμάσει” ακόμα αρκετά ακόμα για να εφαρμόζονται με την ίδια ευκολία που εφαρμόζονται οι γραμμικές μέθοδοι.Για πρακτικές εφαρμόγες είναι πιο εύκολο να γίνει αυτή η υπόθεση οτι το

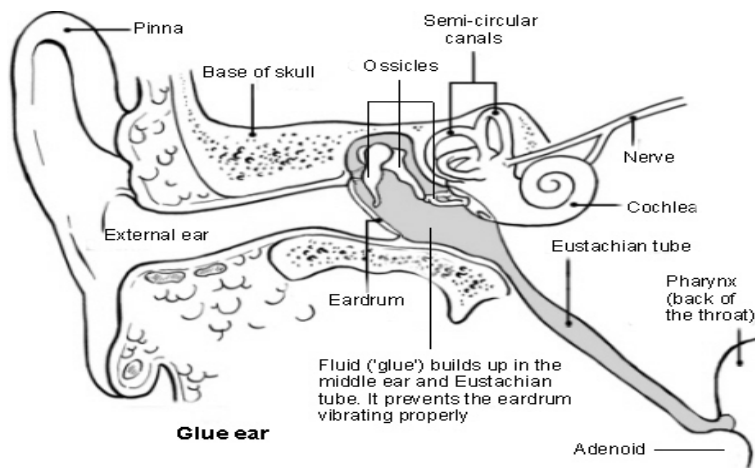
σήμα είναι γραμμικό ,αλλά έχει το κόστος στις εφαρμογές μιας και λαμβάνουμε μικρότερους ρυθμούς συμπίεσης,ένα λιγότερο φυσικό συνθετικό σήμα στην σύνθεση,μικρότερη επιτυχία στην αναγνώριση φωνής κτλ.

Οι μηγραμμικότητες στην φωνή είναι φανερές τουλάχιστον σε 2 πεδία: στο φωνητικό και στο υπερ-τμηματικό(supra segmental).Στο φωνητικό,όπως η ταχεία δυναμική των μεταβατικών και τυρβωδών ήχων(όπως τα σύμφωνα),η ταχεία μεταβολή της συχνότητας των formants(που παρατηρείται κυρίως στα φωνήεντα λόγω του φωνητικού περιεχομένου),συγχετικοί ήχοι και φαινόμενα κατάτμησης τα οποία δεν μπορούν να θεωρηθούν ως στάσιμα και γραμμικά.Αντίστοιχα στο υπερ-τμηματικό,υπάρχουν ταχείς μεταβολές στον τρόπο που κάποιος μιλάει,στην ενέργεια της φωνής,επηρεασμός από την συναισθηματική κατάσταση και υγεία,και επίσης μεγάλη ποικιλία λόγω της ποικιλίας στην ακουστική έκφραση διαφόρων ομιλητών,ομόφωνες λέξεις και εν γένει μη γραμμικότητες στον τρόπο που παράγεται η φωνή και δεν μπορούν να εξαχθούν από αλγορίθμους που κανουν εξαγωγή μη γραμμικών χαρακτηριστικών.Για τον λόγο αυτό έχουν προταθεί διάφορες μη γραμμικές τεχνικές όπως μη γραμμικές παραμετρικές μέθοδοι και μη παραμετρικές αυτοσυσχετιστικές,μέθοδοι που βασίζονται σε ρευστομηχανική της φωνής(Navier Stokes εξισώσεις),διαμόρφωση και φράκταλ μεθόδους.Στην συνέχεια θα αναφερθούμε σε αυτές τις μεθόδους πιο επισταμένα.Τα αποτελέσματα αυτών των ερευνών είναι ιδιαίτερα ενθαρρυντικά και προσφέρουν μια καλή εναλλακτική στις κλασσικές και παραδοσιακές γραμμικές τεχνικές.

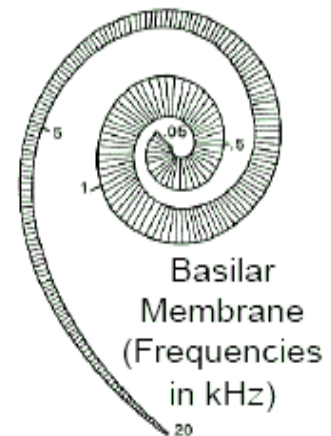
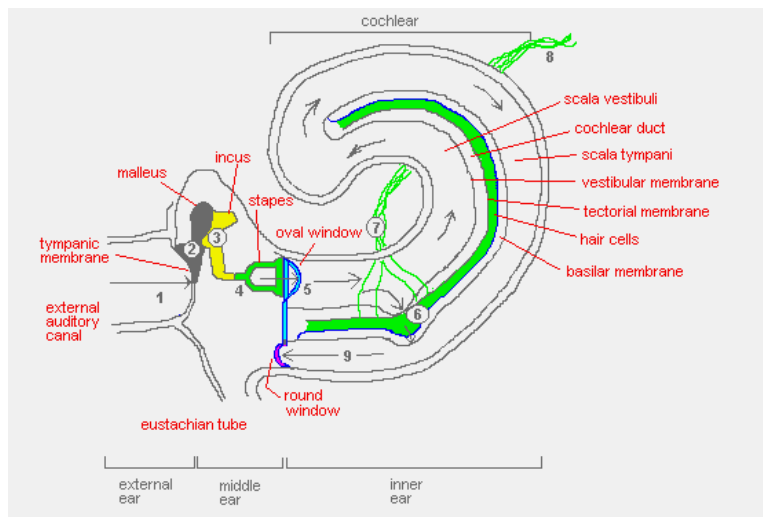
3.3.Μη γραμμικότητες στην αντιληπτικότητα της Φώνης.

Το ανθρώπινο ακουστικό σύστημα είναι το βέλτιστο σύστημα για την αναγνώριση φωνής ακόμα και σε εχθρικές συνθήκες π.χ με ύπαρξη θορύβου.Αν και η επικοινωνία προκύπτει από ταυτόχρονη επεξεργασία διαφόρων σημάτων όπως η φωνή ,οι χειρονομίες,αλλά ακόμη και το περιβάλλον που βρίσκονται.Είναι προφανές ότι το ακουστικό σύστημα επεξεργασίας προσφέρει σημαντικά στην κατανόηση της φωνής.Το αντί είναι μία “συσκευή” για να μετράει ήχους ,επιδεικνύοντας αξιοθεύμαστες λειτουργίες όπως ακρίβεια κατωφλίου(threshold sensitivity),δυναμικό εύρος ταυτοποίησης,διακριτική ικανότητα συχνότητων,ενίσχυση και μηχανισμοί καταστολής θορύβου.Οι μηχανισμοί που υλοποιούν αυτά τα χαρακτηριστικά είναι τόσο μηχανικοί όσο και νευρολογικοί-που προσομοιάζουν την υπολογιστική υφή του ηχητικού συστήματος.

Υπάρχουν ενδείξεις ότι το ανθρώπινο ακουστικό σύστημα υλοποιεί μία μη γραμμική επεξεργασία σήματος που είναι εντελώς διαφορετική από αυτή που γίνεται στην κλασική βραχέος χρόνου ανάλυση.Υπάρχει ενδιαφέρον στην ανάπτυξη τέτοιων αλγορίθμων που προσομοιώνουν το ανθρώπινο σύστημα διότι βοηθάει και αυξάνει την επιτυχία της αναγνώρισης της φωνής και της κατάτμησης.Τα διάφορα μοντέλα ουσιαστικά κάνουν χρήση της μέχρι τώρα γνώσης που υπάρχει στο πως λειτουργεί το περιφερειακό ακουστικό σύστημα καθώς και το πως ο εγκέφαλος χρησιμοποιεί την πληροφορία που πυροδοτείται στα νευρικά όργανα έτσι ώστε να αναγνωρίζουν τους ήχους.Παρακάτω στην εικόνα υπάρχει μία όψη της τομής του αυτιού,που περιέχει όλα τα τμήματα και θα περιγράψουμε το πως συνιστά το κάθε μέρος στην ανάλυση της φωνητικής πληροφορίας.



Τα μαθηματικά μοντέλα του ακουστικού μηχανισμού γίνεται κυρίως μέσω 4 σταδίων όπου το καθένα πραγματοποιεί μια συγκεκριμένη συμπεριφορά. Επι παραδείγματι το πρώτο και το δεύτερο στάδιο προσομοιώνει την μηχανική συμπεριφορά του αυτιού, το οποίο περιέχει την επεξεργασία του κοχλίου. Στο τρίτο στάδιο μοντελοποιεί κάποιες μη γραμμικές διαδικασίες όπως προσαρμογή μικρού χρόνου (short term adaptation), μείωση υψηλής συχνότητας



συγχρονισμού, το οποίο λαμβάνει μέρος στη βασική μεμβράνη και μεταφράζεται σε δραστηριότητα των νευρικών ινών. Τελικά στο τελευταίο στάδιο ολοκληρώνει την εξαγωγή χαρακτηριστικών των σημάτων και η ερμηνεία τους που γίνεται σε υψηλότερες αντιληπτικές ικανότητες.

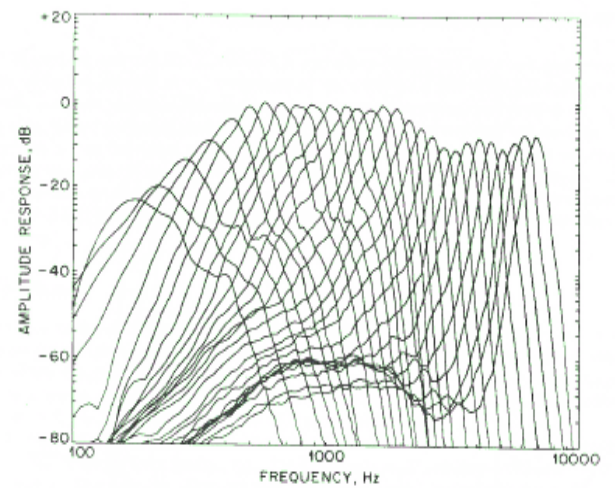
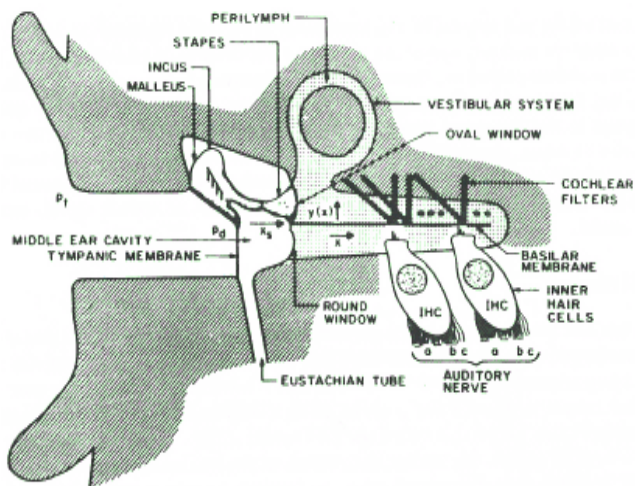
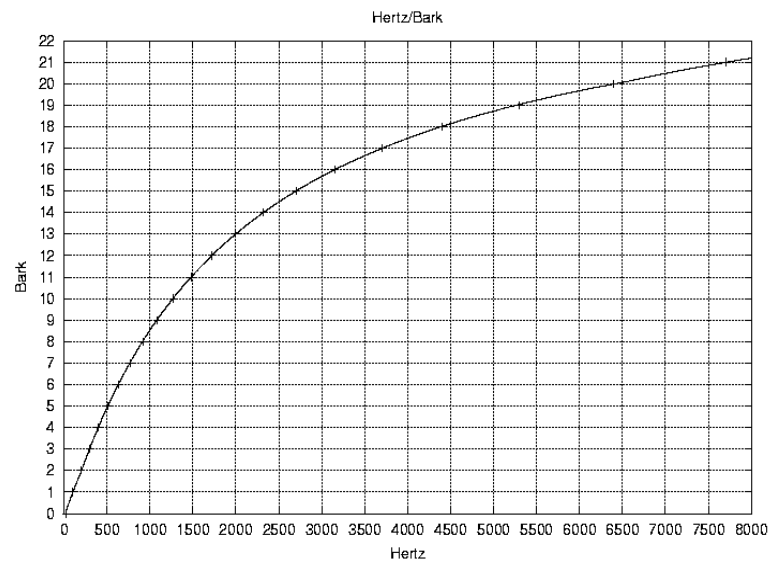
4.1) Μηχανική συμπεριφορά του αυτιού.

Τα φυσιολογικά δεδομένα σχετικά με την απόδοση του αυτιού είναι πράγματι εντυπωσιακά. Η ελάχιστη ακουστική ηχητική ένταση την οποία μπορεί να αντιληφθεί το αυτί είναι 10^{-6} W/cm^2 , η οποία είναι μόλις λίγο πιο πάνω από τον θερμικό θόρυβο που υπάρχει στο ακουστικό εύρος. Η ακουστική αντιληπτική πυκνότητα ποικίλει από το 0 έως 120 dB, που καλύπτει ένα εύρος 12 τάξεων μεγέθους και προτείνει αποτελεσματικούς μηχανισμούς για την μείωση, όπως και την ενίσχυση για υψηλά και χαμηλά επίπεδα σημάτων. Επιπλέον σαν φασματικός αναλυτής, το αυτί έχει απίστευτη διακριτική ικανότητα αφού παρουσιάζει πολύ απαιτητική διακριτική ικανότητα ($\Delta f = 1 \text{ Hz}$) στο χαμηλό εύρος συχνοτήτων ($< 500 \text{ Hz}$) και μία σχετική

σταθερή διακριτική ικανότητα ($\Delta f/f=0.0035$) στο υψηλό εύρος συχνοτήτων ($>1\text{kHz}$). Αυτά τα χαρακτηριστικά υλοποιούνται στο μηχανικό μέρος του αυτιού. Ένα προαπαιτούμενο για το τόσο υψηλό κατώφλι ευαισθησίας είναι η υπόθεση ότι δεν πρέπει να χαθεί ενέργεια από το εισερχόμενο σήμα για να μπορέσει να επεξεργαστεί στα επόμενα στάδια. Αυτό απαιτεί μία εμπέδηση σε μια διαδικασία η οποία γίνεται στο μεσαίο αυτί και η οποία προσπαθεί να δεσμεύσει το ηχητικό κύμα με τα ηχητικά κύτταρα αίσθησης. Πολλές βιολογικές έρευνες δείχνουν ότι η τυμπανική μεμβράνη στο εισερχόμενο σήμα κάνει μία συγκεκριμένο μετασχηματισμό ο οποίος ενισχύει ένα σήμα χαμηλής έντασης και αντίστοιχα μειώνει την ένταση ενός υψηλού σήματος. Συνήθως, ένα μαθηματικό μοντέλο υλοποιείται από ένα γραμμικό φίλτρο, το οποίο περιλαμβάνει ένα εύρος συχνοτήτων από 100Hz μέχρι τα 7000Hz , δηλαδή ένα φίλτρο το οποίο περιστρέφεται γύρω από αυτές τις συχνότητες. Παρα ταύτα το φίλτρο από μόνο του δεν εξηγεί την ακουστική ευαισθησία του κοχλίου, διότι άλλες διαδικασίες γίνονται στο κοχλίο που πρέπει να εξηγηθεί.

Ο κοχλίας, είναι ένα 3.5 εκατοστών σχήμα σε μορφή σαλλιγκαριού. Έχει μια πολύ λεπτά τοιχώματα τα οποία είναι γεμάτα με υγρό ασυμπίεστο. Ο κοχλίας χωρίζεται σε 2 πολύ ελαστικές μεμβράνες. Όταν εισέρχεται ένα σήμα, τα ηχητικά κύματα διέρχονται από το κανάλι του αυτιού, χτυπώντας και ταλαντώνοντας το τύμπανο, το οποίο με την σειρά του ταλάντώνει μία συστοιχία μερών του αυτιού. Μέσα στον κοχλίο, η συμπίεση στο σήμα οφείλεται στο υγρό του το οποίο μετατρέπεται σε ένα κύμα επιφανείας σε όλη την διάσταση της βασιλικής μεμβράνης. Λόγω της μεταβλητής πυκνότητας της βασιλικής μεμβράνης, τα κύματα ταξιδεύουν ταχύτερα στο βασιλικό τμήμα και εν συνεχεία ελαττώνονται στην περιοχή του apical. Λόγω του ότι η ταχύτητα μειώνεται, το πλάτος του κύματος αυξάνει λόγω της αρχής διατήρησης της ενέργειας. Όμως το περιβάλλον του κοχλίου έχει την τάση να αποσβένει και έτσι προστατεύει από το να μην αυξηθεί απεριόριστα από κορυφές του κύματος, τα οποία λαμβάνουν ένα μέγιστο και έπειτα εξαφανίζονται. Ο εντοπισμός του μέγιστου εξαρτάται από την κύρια συχνότητα, η απόκριση των υψηλών συχνοτήτων γίνεται στο μέρος της βασιλικής μεμβράνης διότι είναι και πιο σκληρό, ενώ αντίστοιχα οι πιο χαμηλές συχνότητες εντοπίζονται στο μέρος του apex. Με αυτό το τρόπο η βασιλική μεμβράνη κάνει μία ανάλυση συχνοτήτων στο εισερχόμενο σήμα και μετατρέποντας τις συχνότητες που το συνιστούν στις θέσεις των μέγιστων ταλαντώσεων στη βασιλική μεμβράνη. Αυτός ο τύπος ανάλυσης απεικονίζει την γραμμική κλίμακα συχνοτήτων σε μία λογαριθμική κλίμακα λόγω της υψηλής συχνότητας κανονικοποίησης κλίσης ($\Delta f/f=0.0035$) όπως αναφέρθηκε προτύτερα.

Η μηχανική συμπεριφορά της ανάλυσης των ηχητικών κυμάτων είναι άμεσα συνδεδεμένη με την ανάλυση φουριερ και αναπαριστά μία γραμμική και παθητική αντιμετώπιση της ανάλυσης σημάτων. Όμως τα ρεαλιστικά μοντέλα της ακουστικής “συσκευής” υλοποιούν την συμπεριφορά του εσωτερικού αυτιού με την έννοια μιας συστοιχίας φίλτρων. Η απόκριση της συστοιχίας φίλτρων είναι σχετικά συμβατό με αυτό που γίνονται στην βασιλική μεμβράνη επιτρέποντας στα εύρη του φίλτρου να επικαλύπτονται. Επιπλέον, λόγω της μη γραμμικής κλιμάκωσης στις αντιδράσεις της βασιλικής μεμβράνης, ο υπολογισμός του εύρους των φίλτρων γίνεται χρησιμοποιώντας ένα βιολογικό μέτρο, την κλίμακα Bark, η οποία εικονίζεται παρακάτω:



Παρά την ανάλυση, η συμπεριφορά ενός πλήρους μηχανικού φίλτρου δεν είναι αρκετή να δικαιολογήσει την διακριτικότητα των συχνοτήτων του αυτιού, και γι' αυτό υποτίθεται ότι η συσχέτιση με άλλους μηχανισμούς με την βασιλική μεμβράνη και ενισχύει και οξύνει τα οδεύοντα κύματα. Γι' αυτό το λόγο θα δείξουμε παρακάτω τα εξωτερικά τριχοειδή κύτταρα του αυτιού ότι συμβάλλουν στην υψηλής απόδοσης λειτουργία του.

3.2) Νευρολογική ακουστική επεξεργασία:

Επισυναπτόμενη στην βασιλική μεμβράνη είναι το Corti όργανο, ένα πολύπλοκο όργανο το οποίο τα εσωτερικά και εξωτερικά τριχοειδή αγγεία, οι απολείξεις των κυττάρων (οι οποίες συντίθενται από 2 έως 4 γραμμές από τρίχες, κάποιες δομές οι οποίες αποκαλούνται stereocilia) και βρίσκονται μεταξύ της βασιλικής και tectorial μεμβράνης, και είναι οι βάσεις όπου κυρίως συνδέονται οι νευρικές ίνες του ακουστικού νεύρου. Τα εσωτερικά τριχοειδή αγγεία κυρίως συνδέονται μέσω των afferent νευρώνων, τα οποία μεταφέρουν σήματα στον εγκέφαλο. Τα εσωτερικά τριχοειδή αγγεία λειτουργούν σαν μηχανοηλεκτρολογικοί υπολίστρες, μετατρέποντας μετακινήσεις στις τρίχες σε δυναμικό δράσης. Αντίστοιχα τα εξωτερικά τριχοειδή αγγεία λειτουργούν σαν ηλεκτρομηχανικοί μετατροπείς μετατρέποντας τις τάσεις κατά μήκος του κυττάρου σε μεταβολές στο μήκος. Οι ακουστικοί νευρώνες σχετίζονται με τα τριχοειδή αγγεία τα οποία τροφοδοτούν τον 8η κρανιακό ακουστικό νεύρο και τερματίζουν στον πυρήνα του κοχλία, με αυτό τον τρόπο διασφαλίζουν την διατήρηση του σήματος μέσω αυτής της γεωμετρικής κατανομής της αναπαράστασης του σήματος. Πειραματικά δεδομένα δείχνουν ότι δοθέντος ενός ήχου, οι αποκρίσεις του στο ακουστικό νεύρο είναι μία συνάρτηση της συχνότητας του, π.χ. οι νευρικές αποκρίσεις είναι επιλεκτικές κάθε νευρώνας είναι έχει μέγιστη απόκριση (ένα υψηλό ρυθμό πυροδότησης) σε μία συγκεκριμένη συχνότητα, η οποία καλείται Χαρακτηριστική συχνότητα (characteristic frequency). Η χαρακτηριστική συχνότητα οφείλεται στους μηχανισμούς του κοχλία, και ο κοχλίας από μόνος του παρέχει ανάλυση συχνοτήτων της ακουστικής διέγερσης, διότι κάθε τμήμα της βασιλικής μεμβράνης, όταν διεγείρεται από την χαρακτηριστική συχνότητα, δημιουργεί ένα υψηλό ρυθμό πυροδότησης στους συγκεκριμένους νευρώνες.

Οι δύο τύποι τριχοειδών αγγείων συνισφέρουν διαφορετικά στην επεξεργασία του εισερχόμενου σήματος. Πράγματι, ενώ τα εσωτερικά αγγεία μεταφέρουν τις παραμέτρους του σήματος στον εγκέφαλο, τα εξωτερικά φαίνεται να υλοποιούν ένα είδος ανατροφοδότησης η οποία οξύνει το πρότυπο της διέγερσης στις νευρικές ίνες, έτσι ώστε 2 αντιληπτικά κοντινοί τόνοι να μπορούν να διαχωριστούν παρόλο τον δυνατό περισπασμό λόγω αυτών που γίνεται στο μετασχηματισμό συχνότητας-θέσης στην βασιλική μεμβράνη.

Ένας τρόπος που τα εσωτερικά τριχοειδή αγγεία διεγείρουν τα ακουστικά νεύρα είναι υψηλά μη-γραμμικός, διότι οι νευρώνες μπορούν να μεταφέρουν αποκρίσεις όλα ή τίποτα (all or nothing) φύσεως. Έτσι διάφορα μη γραμμικά φαινόμενα προκύπτουν όταν οι κινήσεις της βασιλικής μεμβράνης παράγουν τα πρότυπα της πυροδότησης στα ακουστικά νεύρα.

Μεταξύ των διαφόρων μη γραμμικών μηχανισμών που συνιστούν την λειτουργία του ακουστικού νεύρου είναι η δυνατότητα “καταπίεσης” (suppression) στον ρυθμό πυροδότησης κατά τη διάρκεια αρνητικών τμημάτων του εισερχόμενου σήματος. Ο μηχανισμός είναι γνωστός σαν διόρθωση μισής κυματομορφής (half wave

rectification) και προϋπεί διότι τα τριχοειδή αγγεία παρουσιάζουν υψηλό επίπεδο κατευθυντικής ευαισθησίας μιας και τείνουν να αποκρίνονται σε μία κατεύθυνση της μετατόπισης της διεγερτικής κυματομορφής. Το επακριβές σχήμα του μισού κύματος μη γραμμικότητας δεν είναι προφανές, τα μαθηματικά μοντέλα στα ακουστικά συστήματα υλοποιούν αυτό το χαρακτηριστικό μέσω του μισού κύματος διόρθωσης που βασίζεται σε κορεσμένη μη γραμμικότητα (saturating nonlinearity). Φυσιολογικά πειραματικά δεδομένα δείχνουν ότι κορυφές (spikes) των ακουστικών νευρικών διαρκούν για 0.2ms και μπορούν να συμβούν και χωρίς ακουστική διέγερση. Λόγω των ακουστικών διεγέρσεων ο ρυθμός μπορεί να πιάσει μέχρι και 1000 spikes/sec. Υψηλότεροι ρυθμοί πυροδότησης δεν είναι δυνατοί λόγω της σκληρής (refractory) περιόδου των νευρώνων. Για συνεχείς ακουστικές διεγέρσεις ο στιγμιαίος ρυθμός αποφόρτισης του ακουστικού νεύρου αποσβέννεται μέχρι να πιάσει κάποιο ρυθμό σταθερού επιπέδου. Τέτοια κατάπτωση προκαλείται αμέσως μετά την διέγερση η οποία καλείται μικρού ρυθμού μετατροπής. Άλλο ένα χαρακτηριστικό της νευρικής κωδικοποίησης είναι στην φάση έρευνας της χαρακτηριστικής των νευρικών κορυφών. Βασίζεται στο γεγονός ότι η νευρική απόκριση, κάτω από μικρής συχνότητας περιοδικές διεγέρσεις είναι κλειδωμένες ως προς την φάση (phase locked), δηλαδή των διαστημάτων εύρεσης μεταξύ πυροδότησης τείνων να συσταδοποιούν κοντά σε πολλαπλάσια της περιόδου διεγέρσεως. Για παράδειγμα, ένα κύμα 500Hz έχει περίοδο 12ms και έτσι τα διαστήματα μεταξύ διαδοχικών νευρικών πυροδοτήσεων θα είναι περίπου 2, 4, 6ms κτλ. Παρα ταύτα το κλειδωμα φάσεων δεν συμβαίνει για συχνότητες υψηλότερες των 4-5 KHz λόγω της απώλειας ακρίβειας μεταξύ της έναρξης της νευρικής πυροδότησης και της φάσης εισαγωγής της εισερχόμενης διέγερσης. Υπάρχει, πράγματι, μία μεταβλητότητα στην επακριβή έναρξη της νευρικής πυροδότησης. Στις υψηλές συχνότητες η μεταβλητότητα αυτή γίνεται συγκρίσιμη με την εισερχόμενη περίοδο, και γιαυτό πάνω από μία συγκεκριμένη συχνότητα οι πυροδοτήσεις θα κατανεμηθούν σε όλη την περίοδο διέγερσης. Αυτές οι διαδικασίες που μόλις περιγράφηκαν έχουν εξαχθεί έπειτα από μελέτες που έχουν γίνει και σχετίζονται με τις αναπαραστάσεις των ακουστικών σημάτων και είναι δομικά διαφορετικές από τις διαδικασίες τις ηχητικές και τις φασματικές που γίνονται με βάση τις κλασσικές τεχνικές επεξεργασίας σήματος. Οι ηχητικές παραμέτροι της φωνής μπορούν να κωδικοποιηθούν καλύτερα χρησιμοποιώντας χαρακτηριστικά που προκύπτουν από μη γραμμικά πρότυπα στους ηχητικούς νευρώνες και χρησιμοποιώντας μηχανισμούς που επιτρέπουν στις νευρικές ίνες να είναι επιλεκτικές ως προς τον χρόνο, χώρο και συχνότητα. Αυτά τα χαρακτηριστικά είναι ικανά να ξεπεράσουν περιορισμούς της ανάλυσης βραχέως χρόνου και καθολικά να εξαγάγουν πληροφορία για την παραγωγή σήματος και την αντίληψη της φωνής με μη γραμμικές δυναμικές. Αποδεικτικές της αληθοφάνειας αυτών των μεθόδων και λεπτομέρειες της επεξεργασίας και των τεχνικών μοντελοποίησης βασίζονται σε μηχανισμούς της ακουστικής επεξεργασίας και υλοποιούνται σαν ανάλυση της φωνής και της αναγνώρισης σε εφαρμογές και μπορούν να βρεθούν σε κατάλληλη βιβλιογραφία.



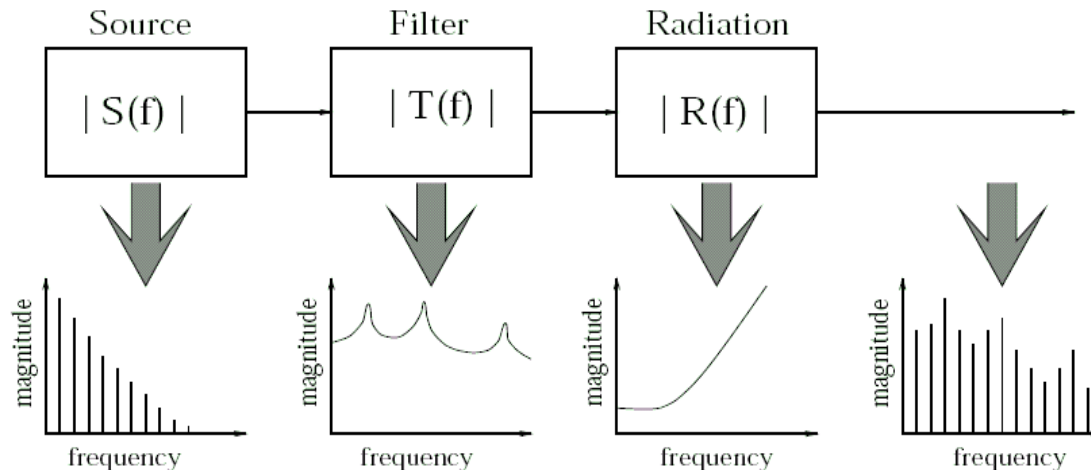
3.4 Ακουστικά μοντέλα:

1)Εισαγωγή:

Το τελευταίο μισό αιώνα ,το ακουστικό σύστημα είναι αντικείμενο έρευνας .Η γνώση της φυσιολογίας ,της ψυχολογίας και των επιστημών μηχανικού έδωσε την δυνατότητα για την δημιουργία μοντέλου τα οποία μέχρι ένα συγκεκριμένο σημείο προσπαθούν να περιγράψουν και να μιμηθούν τους μηχανισμούς ακοής.Τα μαθηματικά και υπολογιστικά μοντέλα χρησιμοποιούνται για να κατασκευάσουν ποσοτικές προσομοιώσεις για να περιγράψουν ένα σύστημα και διελευκάνουν το σύστημα αυτο καθαυτό και την συμπεριφορά του υπό συγκεκριμένες προυποθέσεις.Τα ακουστικά μοντέλα μπορούν να χρησιμοποιηθούν για διάφορες λειτουργίες,όπως η βοήθεια σε προβλήματα φωνής και ακοής.Επίσης μπορεί να είναι πολύ χρήσιμα για την μουσική έρευνα καθώς προσομοιώνουν τον μηχανισμό την αντίληψη του ήχου,έτσι να είναι δυνατό στην έρευνα του ηχοχρώματος.

2)Φυσιολογία ακοής:

Βασικά το περιφερειακό ακουστικό σύστημα έχει τρία μέρη:το εξωτερικό,το μεσαίο και το εσωτερικό αυτί.Το εξωτερικό αυτί έχει ένα οπτικό τμήμα στο αυτί.Περιλαμβάνει την πένα,το κανάλι του αυτιού και το τύμπανο του αυτιού.Η πέννα συλλέγει ήχους και κατευθύνεται κάτω στο κανάλι του αυτιού.Το σχήμα του βοηθά να εξαγάγει απευθείας πληροφορία από τους ήχους.Το κανάλι του αυτιού είναι ένας σωλήνας 2.5εκ και λειτουργεί σαν αντηχητής.Έτσι ενισχύει τις συχνότητες γύρω από τα 3400 Ηζ,οι περιοχές της μέγιστης ευαισθησίας της ακουστικής συχνότητας.Αυτή η



περιοχή μπορεί εύκολα να παρατηρηθεί από τις διάσημες καμπύλες της ίσης έντασης. Προσέγγιση του μοντέλου φωνής/φίλτρου(θα την αναλύσουμε περαιτέρω σε επόμενο κεφάλαιο)

Κεφάλαιο 4

Δυναμικά συστήματα και Μη-γραμμικές μέθοδοι

1)Εισαγωγή:

Για πολλά χρόνια τα μηγραμμικά συστήματα δεν θεωρούνταν η πρώτη επιλογή για έρευνα από τους διάφορες επιστήμονες και η βιβλιογραφία κυρίως εδραζόταν στην χρήση των πιο προσπελάσιμων γραμμικών συστημάτων. Σήμερα οι ρόλοι γίνονται σχεδόν ίσοι λόγω της έλευσης και σταδιακής ωρίμανσης της θεωρίας των μηγραμμικών δυναμικών συστημάτων και της συναφούς περιοχής, της θεωρίας του χάους. Η θεωρία του χάους είναι μία επιστήμη αναγκαστικά σε εμβρυακό επίπεδο και σχετίζεται με πολλά πεδία ερεύνης, όπως η έρευνα για το σύμπαν κτλ. Δυστυχώς παρ'όλο τον ενθουσιασμό των πολλών θεωριών οι περισσότερες έχουν εφαρμοστεί και κατανοηθεί με λανθασμένο τρόπο. Σε αυτό το κεφάλαιο θα περιγράψουμε τη φιλοσοφία και των δομικών τμημάτων που συνθέτουν την θεωρία του χάους και ιδανικά θα αναδείξουμε τους περιορισμούς αυτής της μεθοδολογίας και της εφαρμοσιμότητας τους σε πραγματικά δεδομένα και εν προκειμένω στην επεξεργασία και κατανόηση της φωνής.

Ιστορία:

Οι απαρχές των όσων γνωρίζουμε σαν θεωρία του χάους είναι πολύ δύσκολο να οριστούν στον χρόνο και παράλληλα να τις αποδώσουμε στην εργασία ενός και μόνο επιστήμονα. Είναι ορθότερο να αναφέρουμε πως ότι γνωρίζουμε σαν θεωρία του χάους δημιουργήθηκε από μία πλειάδα μη σχετιζομένων θεωριών και ανακαλύψεων από ένα εύρος ασυσχέτιστων περιοχών. Οι αρχικές εργασίες του πεδίου προέκυψαν από την εργασία του ρώσου μαθηματικού Lyapunov που διερεύνησε την μηγραμμική δυναμική συμπεριφορά και παρήγαγε ένα ορισμό για να ποσοτικοποιήσει κάποια αμετάβλητα χαρακτηριστικά, όπως την ευαισθησία στις αρχικές συνθήκες. Η εργασία του ήταν ο ακρογωνιαίος λίθος και μία πλήρης ανάλυση χάους και έγινε γραφηματικά πιο εμφανής με τις εργασίες που θα ακολουθούσαν. Ο Poincare χρησιμοποίησε την τοπολογία για να δείξει ότι κάποια φαινομενικά πολύπλοκα συστήματα στην πραγματικότητα παρουσίαζαν πολύ κανονική συμπεριφορά. Οι Julia και Fatou βρήκαν ότι απλές εξισώσεις μπορούν να παράγουν απείρως πολύπλοκα γεωμετρικά σχήματα, η ιδέα που αργότερα ο Benoit Mandelbrot θα ξεκαθαρίσει και θα ορίσει την λέξη fractal για να περιγράψει την γεωμετρική δομή. Ο Edward Lorenz αν και προσπάθησε να παράξει ένα μοντελο για να περιγράψει τον καιρό παρουσιάστηκε σαν το πιο διάσημο σύμβολο του χάους το ελκυστή Lόρεντζ (Lorenz attractor). Παρά αυτές τις έρευνες το χάος ήλθε στο ευρύτερο κοινό μέσω των εργασιών των Feigenbaum, May και Yorke για να συνδέσουν όλες αυτές τις ιδέες και να δώσουν τις απαραίτητες αρχές για μία εφαρμόσιμη θεωρία, πράγματι ο Li και Yorke ήταν οι πρώτοι που χρησιμοποίησαν τον όρο χάος για αυτή την περιοχή. Αυτές οι αρχές έχουν ερευνηθεί από πολλούς επιστήμονες. Οι Ruelle και Takens για να ονομάσουμε 2, δημιούργησαν την θεωρία του χάους.

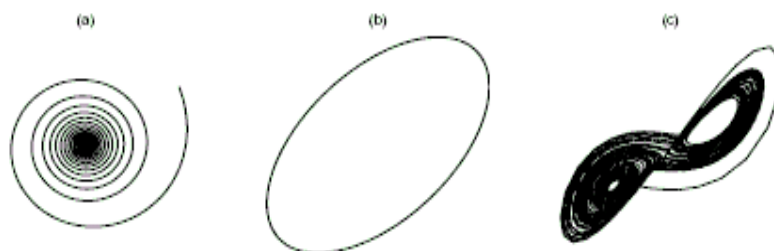
Έχουμε πολυποίκιλες εφαρμογές σε πολλές περιοχές όπως η κωδικοποίηση fractal, διαχωρισμό σημάτων (signal separation), υπογραφές radar, υποβρύχια συνεχή σήματα κυμάτων, χαοτική πρόβλεψη (chaotic prediction) και hidden Markov models (HMM). Στις επόμενες σελίδες του κεφαλαίου θα δώσουμε μια γενική επισκόπηση των στοιχείων αυτής της θεωρίας και τις εφαρμογές τους στην φωνή.

Η θεωρία των δυναμικών συστημάτων είναι ένα πολύ σημαντικό μέρος της μη γραμμικής ανάλυσης σημάτων. Ένα δυναμικό σύστημα είναι ένα σύστημα του οποίου η κατάσταση αλλάζει με τον χρόνο. Στον συνεχή χρόνο, το σύστημα περιγράφεται με διαφορικές εξισώσεις, ενώ στον διακριτό χρόνο από επαναληπτικές προβολές-συναρτήσεις. Σε αυτό το κεφάλαιο θα γίνει μία μικρή επισκόπηση της θεωρίας και αυτών των εξισώσεων. Αξίζει να σημειωθεί, ότι στα σήματα και ιδιαίτερα τα ψηφιακά ασχολούμαστε στον διακριτό χρόνο, οπότε θα επικεντρωθεί η μελέτη κυρίως σε αυτό το χωρίο.

Η δυναμική ενός διακριτού χρόνου συστήματος καθορίζεται κυρίως από τις πιθανές καταστάσεις σε ένα πολυδιάστατο χώρο, ο οποίος καλείται χώρος καταστάσεων ή χώρος φάσεων. Οι μεταβάσεις μεταξύ των καταστάσεων μπορεί να περιγραφεί με διανύσματα, και αυτά τα διανύσματα περιγράφουν κυρίως την τροχιά που περιγράφει την χρονική εξέλιξη του συστήματος με βάση την εξίσωση:

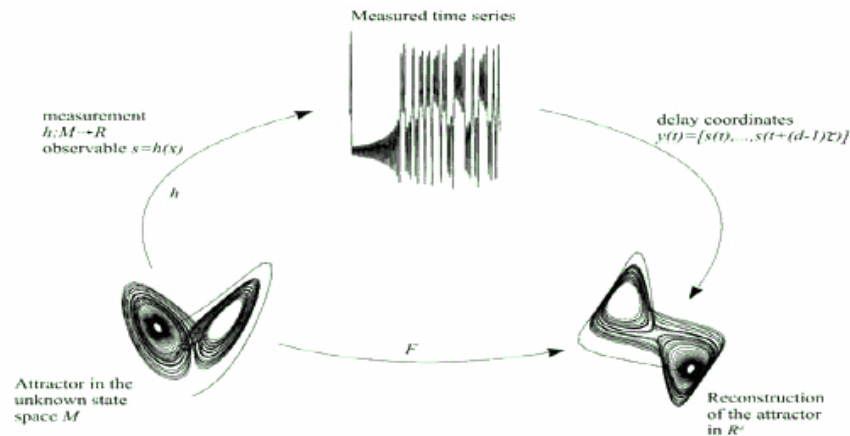
$$x(t+1) = \varphi(x(t))$$

,όπου $\chi(t)$ εκφράζει την κατάσταση του συστήματος και t είναι η χρονική στιγμή και είναι μια συνάρτηση απεικόνισης $\varphi: M \rightarrow M$, όπου M είναι ο πραγματικός χώρος καταστάσεων. Παρακάτω δίδουμε μία γεωμετρική απεικόνιση μίας τέτοιας τροχιάς.



Παραπάνω εικονίζονται ελκυστές σταθερού σημείου, πιο συγκεκριμένα έχουμε, α) έναν οριακό κύκλο, β) έναν παράξενο ελκυστή από ένα σύστημα Λόρεντζ, γ) Ένα φυσικό παράδειγμα ενός ελκυστή σταθερού σημείου είναι ένα εκρεμμές, όπου όλες οι αρχικές καταστάσεις συγκλίνουν σε ένα συγκεκριμένο σημείο. Παραλάσσοντας το παράδειγμα, έτσι ώστε το εκρεμμές να έχει μία δύναμη, τότε δημιουργείται απλή ταλάντωση, και τότε λαμβάνουμε ένα περιοδικό ελκυστή. Τα χαοτικά συστήματα όπως τα συστήματα Λόρεντζ χρησιμοποιούνται για να περιγράψουν τον καιρό και προκύπτουν παράξενοι ελκυστές των οποίων οι τροχιές δεν τέμνονται ποτέ και δεν εφάπτονται ποτέ.

Ο πραγματικός χώρος καταστάσεων περιλαμβάνει τις πραγματικές καταστάσεις χ , του οποίου η χρονική εξέλιξη περιγράφεται από την απεικόνιση $\varphi, \chi(t) = \varphi_t(\chi(0))$. Ας υποθέσουμε ότι η μόνη πληροφορία διαθέσιμη σχετικά με το σύστημα είναι αυτή η οποία μπορούμε να βρούμε με βαθμωτή μορφή $s(t) = h(x(t))$, όπου $h: M \rightarrow \mathbb{R}$. Εάν $s(t)$ είναι η προβολή από τον πολυδιάστατο χώρο καταστάσεων M , τότε μπορούμε να αντιστρέψουμε την διαδικασία όπως εικονίζεται στο παρακάτω σχήμα:



Αυτό προστάζει ότι δοθέντος ενός σήματος ,μπορούμε να δημιουργήσουμε μία απεικόνιση από μία άγνωστη κατάσταση στο M στο αντίστοιχο σημείο $Y(\tau)$ στο ανασκευασμένο χώρο καταστάσεων στον $\mathbb{R}^d$. Όλα αυτά βασίζονται στο θεώρημα του Τακενς:

$$F: M \rightarrow \mathbb{R}^d$$

$$x(t) \rightarrow y(t) = F(x(t)) = [s(t), s(t+\tau), \dots, s(t+(d-1)\tau)]$$

όπου τ είναι η παράμετρος καθυστέρησης, d είναι η διάσταση εμπεντινγκ και F είναι η απεικόνιση από τον πραγματικό χώρο καταστάσεων στον ανασκευασμένο χώρο καταστάσεων. Ιδιαίτερης σημασίας είναι η επιλογή του τ και του d , διότι επηρεάζει την ακρίβεια ανασκευάσεως του χώρου καταστάσεων. Παρά την σημαντικότητα τους, δυστυχώς δεν υπάρχει πάντοτε προκαθορισμένος τρόπος επιλογής. Μπορούμε όμως να χρησιμοποιήσουμε την μέθοδο που βασίζεται στην μέση αμοιβαία πληροφορία για το τ και d την μέθοδο του Cao. Το θεώρημα του Τάκενς ουσιαστικά αποδεικνύει ότι ο χώρος καταστάσεων που ανασκευάζεται είναι δυναμικά και τοπολογικά ισοδύναμος με τον M . Διότι η δυναμική του ανασκευασμένου χώρου περιέχει την ίδια τοπολογική πληροφορία όπως ο αρχικός χώρος καταστάσεων, ο χαρακτηρισμός και η πρόβλεψη βασίζονται στον ανασκευασμένο χώρο είναι αληθείς σαν να είχαν γίνει από τον πραγματικό χώρο καταστάσεων..

2) Συγκριση γραμμικών και μη γραμμικών τρόπων επεξεργασίας σημάτων:

Γραμμικές Μέθοδοι

Προπαρασκευή σήματος
Διαχωρίζουμε το σήμα από το θόρυβο
χρησιμοποιώντας φασματικά χαρακτηριστικά
μέσω της χρήσης φίλτρων στο πεδίο συχνοτήτων.
καταστάσεων, είτε
ελκυστή

Βρίσκοντας τον χώρο:
Χρησιμοποιώντας μέθοδο Fourier, όπου
μεταβλητών
ουσιαστικά μετατρέπουμε τις εξισώσεις
διαφορών σε αλγεβρικές μορφές.

Μη γραμμικές μέθοδοι

Προπαρασκευή σήματος
Διαχωρίζουμε το σήμα από το θόρυβο
χρησιμοποιώντας την ντετερμινιστική
φύση του σήματος, μέσω του να κάνουμε
αποσυνθεση του χώρου
μέσω της χρήσης στατιστικών του
Βρίσκοντας τον χώρο:
Μέσω χρονικά καθυστερημένων
που συνθέτουν τις συντεταγμένες του
ανασκευασμένου χώρου σε d - διαστάσεις

$s(t)$ το σήμα,

$$S(f) = \sum s(t) e^{i2\pi ft}$$

Ταξινόμηση του σήματος:

- 1)Οξείες φασματικές κορυφές
- 2)Συχνότητες συντονισμού του συστήματος point)

Πρόβλεψη:

$$s(t+1) = \sum \alpha_k s(t-k)$$

Βρίσκουμε τους συντελεστές με βάση τις φασματικές κορυφές.

$$y(t) = [s(t), s(t+\tau), \dots, s(t+(d-1)\tau)]$$

όπου τ, d είναι οι μεταβλητές από το

θεώρημα του Τάκενς
Ταξινόμηση του σήματος:

- 1)Εκθέτες Lyapunov
- 2)Μετρική της φράκταλ διάστασης
- 3)Ασταθή σημεία ισορροπίας(fixed point)
- 4)Χρήση αναδρομικών σχέσεων
- 5)Στατιστική κατανομή στον ελκυστή

Πρόβλεψη:

$$y(t) \rightarrow y(t+1)$$

$$y(t+1) = F[y(t), a_1, a_2, \dots, a_p]$$

Βρίσκουμε τις παραμέτρους με βάση τους συντελεστές Lyapunov και τις φράκταλ διαστάσεις.

3)Μη γραμμική ανάλυση σήματος:

Με βάση τα παραπάνω είναι εύληπτο ότι η μη γραμμική ανάλυση έχει εντελώς διαφορετικά εργαλεία για την ανάλυση και δίδει νέα μεθοδολογία αρκετά βοηθητική για την μελέτη των σημάτων.

α)Μη-ακέραιες διαστάσεις:

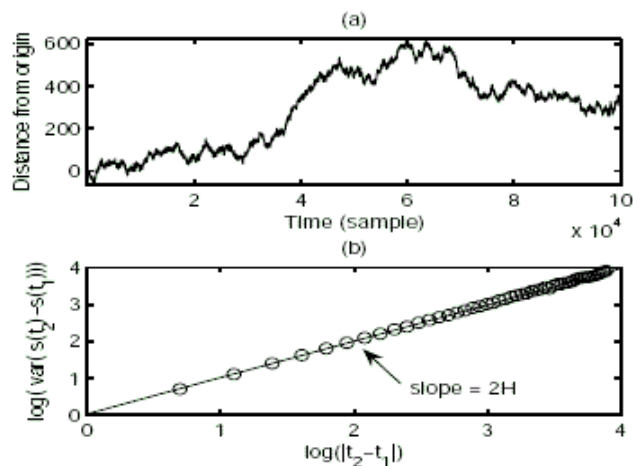
Όπως αναφέραμε σε άλλο κεφάλαιο σχετικά με τα φράκταλ, τα φράκταλ χαρακτηρίζονται από το γεγονός ότι δεν έχουν ακέραια τιμή. Επίσης μερικοί παράξενοι ελκυστές(strange attractors) έχουν επίσης φράκταλ διάσταση και μάλιστα αυτή η διάσταση είναι μικρότερη από την διάσταση του χώρου όπου υπάρχουν. Υπάρχουν δύο τύποι προσέγγισης για να υπολογίσουμε την φράκταλ διάσταση, ιδιαίτερα για να εφαρμοστούν επακριβώς πάνω στην κυματομορφή και εκείνων που εφαρμόζονται πάνω στον ανασκευασμένο χώρο καταστάσεων. Αξίζει να σημειωθεί ότι η διάσταση του ελκυστή είναι κανονικά διαφορετική από αυτή της κυματομορφής και της φράκταλ διάστασης -που κυμένεται μεταξύ της διάστασης 1 έως 2-. Υπάρχουν ένα πλήθος προβλημάτων όταν προσπαθούμε να εκτιμήσουμε την φράκταλ διάσταση του ελκυστή, ένας λόγος είναι το υπολογιστικό φορτίο- πολυπλοκότητα- των αλγορίθμων. Για αυτό το λόγο, επικεντρωνόμαστε στην μελέτη της φράκταλ διάστασης της κυματομορφής. Θεωρούμε ότι η κυματομορφή είναι στο επίπεδο, όπου η κυματομορφή είναι μία γεωμετρική φιγούρα. Αν και υπάρχουν πολλοί τρόποι για να εκτιμηθεί η φράκταλ διάσταση, εδώ θα επικεντρωθούμε στην διακύμανση της φράκταλ διάστασης(VFD), λόγω του ότι είναι αρκετά εύρωστη μέθοδος ως προς το θόρυβο. Οι υπολογισμοί βασίζονται στην κατανομή power-law.

$$Var(s(t_2) - s(t_1)) \sim |t_2 - t_1|^{2H}$$

όπου H είναι ο δείκτης Hurst, και αποτελεί μία μετρική της ομαλότητας του σήματος. Ισοδύναμα έχουμε λογαριθμίζοντας:

$$\log(Var(s(t_2) - s(t_1))) = 2H \cdot \log(|t_2 - t_1|)$$

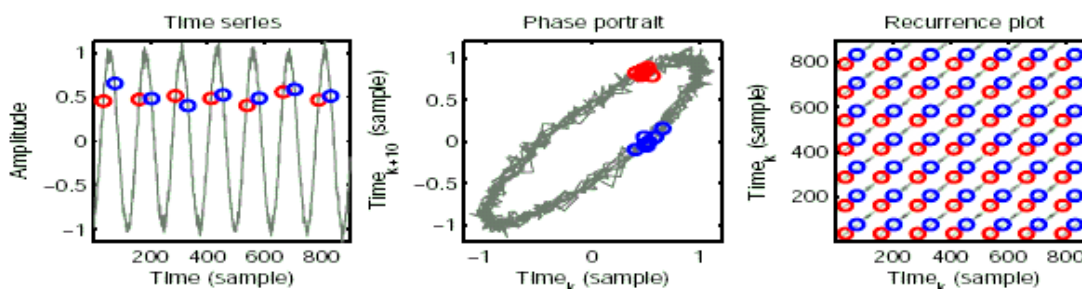
Παρακάτω έχουμε μία λογαριθμική απεικόνιση, ένα απλό παράδειγμα:



Η φράκταλ διαστάση διακύμανσης ορίζεται ως προς το H , ως $\Phi\Delta\Delta=(\text{ευκλείδια διάσταση})+1-H$, για χρονοσειρές η ευκλείδια διασταση είναι 1. Η επιλογή των χρονικών αυξήσεων αντικατοπτρίζεται στην τροχιά της διάστασης, όπου το σήμα υπάρχει. Εάν οι χρονικές αυξήσεις επιλέγονται ως δυνάμεις του 2 (1,2,4,8,...), οι διαφορές μεταξύ διαφόρων συστατικών του σήματος αναδεικνύονται, ενώ αν είναι μοναδιαία η αύξηση (1,2,3,...), η κατάτμηση του σήματος από το θόρυβο αναδεικνύονται πιο εύκολα. Συνήθως χρησιμοποιούμε την τεχνική του ολισθένοντος παραθύρου για να μειώσουμε την πολυπλοκότητα του αλγορίθμου.

β) Αναδρομική υπολογιστική ανάλυση:

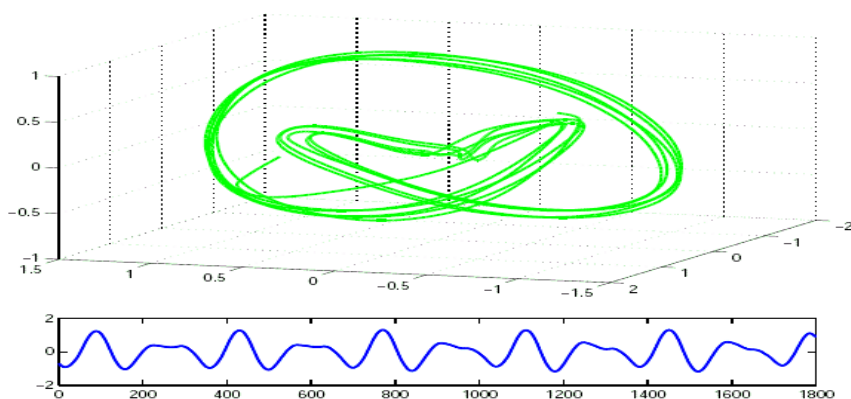
Ο χώρος καταστάσεων ενός συστήματος είναι συχνά υψηλής διάστασης, ειδικά όταν ανασκευάζεται από πειραματικά δεδομένα όπου ο θόρυβος τείνει να αυξάνει την διάσταση. Το πορτραίτο της φάσης μπορεί να οπτικοποιηθεί με προβολές σε 2 ή 3 διαστάσεις. Αυτή η λειτουργία τείνει να “διπλώνει” τον ελκυστή και κάνοντας αυτό καταστρέφει την δομή του. Μια αναδρομική απεικόνιση (recurrence plot) είναι ένας τρόπος να ερευνήσουμε οπτικά την δ-διάστατη τροχιά του χώρου καταστάσεων μέσω μίας 2-διάστατης απεικόνισης. Συνήθως πρέπει να τα χρησιμοποιούμε για μικρής διάρκειας χρονοσειρές και περιγράφουν την αναδρομή των καταστάσεων. Σε σχέση με άλλες μεθόδους όπως Φουριερ, wavelets, οι αναδρομές είναι απλές σχέσεις που μπορούν να χρησιμοποιηθούν για γραμμικά και μη γραμμικά δεδομένα. Μία τέτοια απεικόνιση βασίζεται σε ένα πίνακα $N \times N$, όπου κάθε καταχώρηση (i,j) εκφράζει την απόσταση $y(i)$ και $y(j)$.



Εχουμε ένα παράδειγμα τέτοιας απεικόνισης:

Υπάρχουν κυρίως 7 παράμετροι που επηρεάζουν το αποτέλεσμα μιας τέτοιας μελέτης, όπως η διάσταση d , η χρονική καθυστέρηση τ , το μήκος της χρονοσειράς, η νόρμα για τις αποστάσεις. Η τελευταία παράμετρος χρησιμοποιείται για την ποσοτικοποίηση της μελέτης και όχι για την δημιουργία της τεχνικής. Μετρικές όπως η RQA (Recurrence quantification analysis) κυρίως βασίζονται στις διαγώνιες

δομές,κάθετες δομές και τα στατιστικά του χρόνου.Απομονωμένα σημεία αναδρομής προκύπτουν όταν οι καταστάσεις είναι σπάνιες,εάν δεν επιμένουν σε κάθε χρόνο και δεν έχουν μεγάλη διακύμανση.Οι διαγώνιες γραμμές προκύπτουν όταν ένα τμήμα της τροχιάς τρέχει παράλληλα με κάποιο άλλο τμήμα,π.χ όπως το παρακάτω σχήμα.



4)Υψηλότερης τάξης στατιστικές:

Οι παραδοσιακές τεχνικές στην επεξεργασία σήματος βασίζονται στις 2-ου βαθμού στατιστικές,αλλά οι δευτερόν βαθμού μετρικές δεν περιέχουν πληροφορία σχετικά με την φάση.Σαν συνέπεια,τα μη ελάχιστης φάσης σηματα και συγκεκριμένοι τύποι σύζευξης των φάσεων(που σχετίζονται κυρίως με τις μηγραμμικότητες) δεν μπορούν κατάλληλα να αναγνωριστούν.Εν αντιθέσει με τις 2ου βαθμού στατιστικές,οι υψηλότερης τάξης στατιστικές βασίζονται σε μέσους όρους πάνω από γινόμενα 3 ή και περισσοτέρων δειγμάτων του σήματος,έτσι οι μη γραμμικές εξαρτήσεις σε πολλαπλά δείγματα των σημάτων μπορούν να εκτιμηθούν.Θεωρώντας ότι το σήμα έχει μηδενικό μέσο όρο ,και περιορίζοντας την μελέτη μέχρι 4-ου βαθμού τότε οι αντίστοιχες αθροιστικές συναρτήσεις είναι οι ακόλουθες:

$$m_s^{(1)} = c_s^{(1)} = E \{ s(t) \} = 0$$

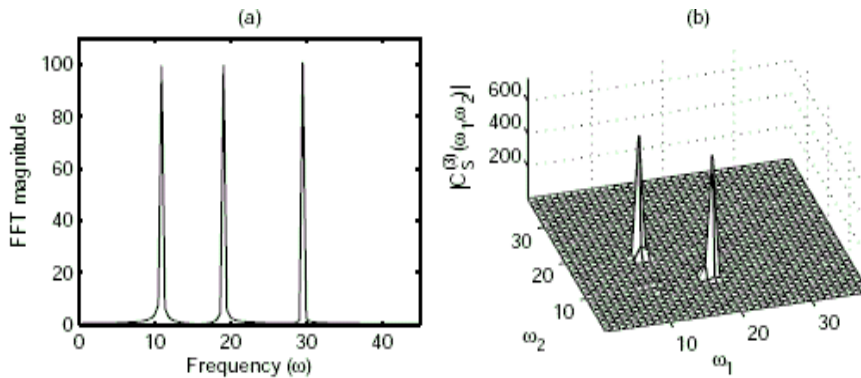
$$m_s^{(2)}(\tau) = c_s^{(2)}(\tau) = E \{ s(t) s(t+\tau) \}$$

$$m_s^{(3)}(\tau_1, \tau_2) = c_s^{(3)}(\tau_1, \tau_2) = E \{ s(t) s(t+\tau_1) s(t+\tau_2) \}$$

$$m_s^{(4)}(\tau_1, \tau_2, \tau_3) = E \{ s(t) s(t+\tau) s(t+\tau_2) s(t+\tau_3) \}$$

$$c_s^{(4)}(\tau_1, \tau_2, \tau_3) = E \{ s(t) s(t+\tau) s(t+\tau_2) s(t+\tau_3) \} - 3 \left[E \{ s(t) s(t+\tau) \} \right]^2$$

όπου το E παρουσιάζει την μέση τιμή.Ιδιαίτερης σημασίας έχουν οι διακυμάνσεις , κυρτώσεις ,η τραχύτητα που αντιστοιχούν στους βαθμούς 2,3 και 4 αντίστοιχα.Οι μετασχηματισμοί Fourier αυτών καλούνται bispectra(διφάσματα).Ένα μικρό παράδειγμα τέτοιου μετασχηματισμού(3ου βαθμού)εικονίζεται παρακάτω:



Σχήμα: Ένα παράδειγμα σύζευξης φάσεων. Το φάσμα συχνοτήτων συνίσταται από 3 ημίτονα με συχνότητες ω_1, ω_2 και $\omega_3 = \omega_1 + \omega_2$. Το αντίστοιχο διάγραμμα bispectrum εικονίζεται ακριβώς δίπλα. Λόγω της σύζευξης που προαναφέραμε βλέπουμε μία κορυφή στα αντίστοιχα σημεία του φάσματος

Ορισμός των φασμάτων:

$$C_s^{(2)}(\omega) = FT\{c_s^{(2)}(\tau)\} \quad \text{Power spectrum}$$

$$C_s^{(3)}(\omega_1, \omega_2) = FT\{c_s^{(3)}(\tau_1, \tau_2)\} \quad \text{Bispectrum}$$

$$C_s^{(4)}(\omega_1, \omega_2, \omega_3) = FT\{c_s^{(4)}(\tau_1, \tau_2, \tau_3)\} \quad \text{Trispectrum}$$

Ολοκληρωτικός χαρακτηρισμός της στοχαστικής διαδικασίας απαιτεί γνώση όλων των ροπών του σήματος. Γενικά μιλώντας αξίζει να πούμε ότι οι ροπές αντιστοιχούν στις συσχετίσεις και τις αθροιστικές στις διακυμάνσεις. Αν και τα 2 αυτά που αναφέραμε περιέχουν την ίδια στατιστική πληροφορία, στην πράξη χρησιμοποιούμε της αθροιστικές λόγω των παρακάτω:

- 1) Το αθροιστικό φάσμα για τάξη μεγάλη του 2 είναι μηδενικό για Γκαουσιανά σήματα και έτσι τα πολυφασματα παρέχουν μία μετρική για την έκταση της μη-Γκαουσιανής ιδιότητας.
- 2) Η συνάρτηση συδιακύμανσης ενός λευκού θορύβου είναι ένα παλμός όπου το φάσμα του είναι επίπεδο. Ομοίως οι αθροιστικές λευκού θορύβου σε πολυδιάστατους παλμούς με πολυδιάστατα επίπεδα πολυφάσματα.
- 3) Η αθροιστική 2 ανεξαρτήτων τυχαίων διαδικασιών ισούται με το άθροισμα αυτών των τυχαίων μεταβλητών.

Οι υψηλότερης τάξης αθροιστικές δίδουν μία μετρική σχετικά με το πόσο το τυχαίο διάνυσμα αποκλείει από το γκαουσιανό τυχαίο διάνυσμα με μέσο όρο 1 και πίνακα συνδιακύμανσης. Αυτή η ιδιότητα μπορεί να χρησιμοποιηθεί για την εξαγωγή του μη-γκαουσιανού τμήματος του σήματος (μία εφαρμογή είναι για την αφαίρεση θορύβου). Άλλες ενδιαφέρουσες ιδιότητες είναι ότι το διφάσμα είναι μηδεν ενός γκαουσιανού σήματος και το κανονικοποιημένο διφάσμα (δυσυνοχή-bicoherence) είναι σταθερό για ένα γραμμικό σήμα.

5) Μη γραμμική πρόβλεψη:

Υπάρχουν διάφορες πηγές προβλεπτικότητας στις χρονο-σειρές. Εάν ένα σήμα περιέχει γραμμικές συσχετίσεις ως προς το χρόνο, τότε τα γραμμικά μοντέλα είναι κατάλληλα (μετακινούμενου μέσου όρου μοντέλα-MA-moving average models, μοντέλα αυτοσυσχέτισης-autoregressive models-AR, και μοντέλα αυτοσυσχέτισης μεταβλητού μέσου-autoregressive moving averages-ARMA). Τα MA μοντέλα μπορούν να χρησιμοποιηθούν αν το φάσμα ενός σήματος συμπεριφέρεται

σαν χρωματισμένος θόρυβος, ενός τα AR είναι προτιμώμενα εάν το φάσμα κυριαρχείται από κάποιες διακεκριμένες κορυφές. Το ARMA είναι φυσική εξέλιξη αυτών των 2:

$$s(t) = \sum_{k=1}^N \alpha_k s(t-k) + e(t)$$

όπου ακ είναι τα γραμμικά βάρη. Για την πρόβλεψη, μόνο οι βεβαρυμένοι σταθμιστές είναι σημαντικοί και η πρόβλεψη επιτυγχάνεται αγνοώντας τον όρο e(t). Αυτό το μοντέλο μπορεί να επεκταθεί για να επιτρέπει μη

$$\begin{aligned} \hat{s}(t|\theta) &= \sum_{k=1}^N \alpha_k g_k(\varphi) \\ \theta &= [\alpha_1, \alpha_2, \dots, \alpha_n]^T \\ g_k(\varphi) &= \kappa(\beta_k(\varphi - \gamma_k)) \\ \varphi &= [s(t-k), \dots, s(t-1)] \end{aligned}$$

γραμμικές εξαρτήσεις μεταξύ των προηγούμενων εξόδων s(t-k). Ειδικότερα έχουμε:

όπου όλα τα G_k σχηματίζονται από επεκτάσεις και μετατρεπόμενες εκδόσεις της μητρικής βάσης της συνάρτησης K. Θ είναι ένα διάνυσμα είναι ένα διάνυσμα βαρών και φ είναι ένα διάνυσμα γνωστών δειγμάτων του σήματος, α είναι οι συντεταγμένες των βαρών και β είναι η κλιμακα ή επέκταση των βαρών και γ είναι η τοποθεσία ή η μετατροπή των παραμέτρων. Μερικά παραδείγματα σε ποιο πλαίσιο μοντέλου μπορούν να χρησιμοποιηθούν και είναι τα εξής:

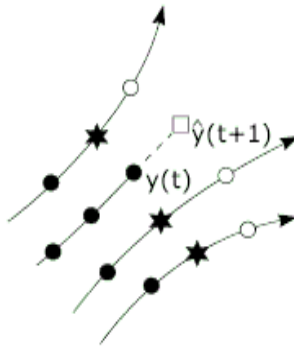
Αυτοσυσχετιστικά μοντέλα: όπου ορίζουν τις περισσότερες μεταβλητές ίσες με μονάδα.

Σιγμοειδή νευρωνικά δίκτυα: είναι η κλασσική κατασκευή με σιγμοειδείς συναρτήσεις.

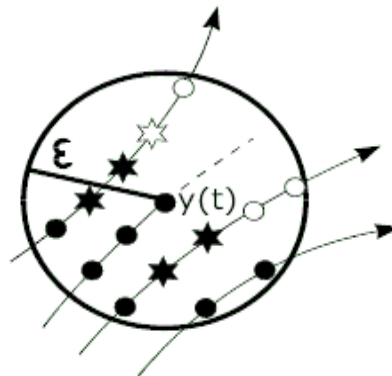
Ακτινικής βάσης δίκτυα: είναι ακτινικές κατασκευές τέτοιες όπως αυτή της γκαουσιανής καμπύλης.

Αν επιστρέψουμε στην μεθοδολογία του ανασκευασμένου χώρου καταστάσεων θα δούμε αρκετές ομοιότητες. Όπως η παράμετρος καθυστέρησης τ, αν θέσουμε τ=1, τότε το φ γίνεται γ. Ένας τρόπος να το δούμε είναι σαν ένα μοντέλο που περιγράφει έναν ελκυστή. Συνήθως όλες οι παράμετροι είναι σχεδιασμένες παράμετροι, είτε μπορούν να ποικίλουν σε ένα προκαθορισμένο τρόπο ή να είναι σταθερές. Αν εισαγάγουμε μία συνάρτηση κόστους, τότε οδηγούμαστε σε γραμμικές εξισώσεις όταν αποτιμούμε το α:s, αλλά απλοποιεί κατά πολύ τον ορισμό. Μιας και το μοντέλο προσπαθεί να μοντελοποιήσει όλο τον ελκυστή καλείται καθολικό μοντέλο.

Αυτό λέγεται για καθολικά μοντέλα τα οποία θα εγκαταλείψουν μεταξύ τους και θα συγκεντρωθούν σε στις τοπικές μεθόδους για την ανασκευή όλου του ελκυστή. Παράλληλα οι τροχιές σε ένα χώρο φάσεων μοιράζονται τα ίδια χαρακτηριστικά κυματομορφής με το πεδίο του χρόνου, και είναι μία μέθοδος να προβλέψεις το μέλλον μέσω του να μιμηθείς την διαδικασία εξέλιξης στις γειτονικές τροχιές, όπως φαίνεται και στο παρακάτω σχήμα. Αν τα δεδομένα δειγματοληπτηθούν με μεγάλη συχνότητα, τα περισσότερα κοντινού γείτονα σημεία θα είναι δείγματα γειτονικά και στις χρονο-σειρές. Μπορούμε να δούμε θεαματική αλλαγή αν αντί χρήσης των κοντινών γειτόνων με την χρήση των κοντινών τροχιών.



Τρεις τροχιές που τέμνουν και προβλέπουν την τροχιά στον 2διάστατο χώρο φάσης. Ο μέσος όρος αλλάζει μεταξύ των κοντινότερων γειτονικών σημείων (μαύρα αστέρια) και των διαδόχων (λευκοί κύκλοι) που χρησιμοποιούνται για την πρόβλεψη των επόμενων σημείων (λευκό τετράγωνο).



Πολλοί από τους κοντινούς γείτονες του $y(t)$, αστέρια, είναι τυπικά παράσιτα λόγω της υψηλής δειγματοληψίας. Χρησιμοποιώντας έναν αλγόριθμο κοντινής τροχιάς απ'ότι ενός κοντινού γείτονα είναι ία λύση.

Υπάρχουν 2 τρόποι να προβλέψουμε p βήματα μπροστά, είτε μέσω της επαναληπτικής πρόβλεψης, είτε της ευθείας πρόβλεψης. Εάν η πρόβλεψη επαναλαμβάνεται, ο αλγόριθμος προβλέπει ένα βήμα μπροστά p φορές (οι προβλεπόμενες τιμές θα χρησιμοποιηθούν σαν σημεία εκκίνησης στις επόμενες επαναλήψεις). Στην ευθεία πρόβλεψη οι εξελίξεις των κοντινότερων γειτόνων μοντελοποιούνται σαν μία συνάρτηση που απεικονίζει p βήματα προς το μέλλον. Έχει εμπειρικά δειχθεί ότι η επαναληπτική πρόβλεψη είναι καλύτερη στις προβλέψεις βραχέως χρόνου για μία ευρεία γκάμα μη γραμμικών μοντέλων. Παρ'όλα αυτά οι επαναληπτικές προβλέψεις δεν λαμβάνουν τα σφάλματα σαν είσοδο με αποτέλεσμα να αυξάνουν εκθετικά. Πολλές επιλογές είναι διαθέσιμες. Στην πρόβλεψη μέσου όρου, ο μέσος όρος των γειτονικών διαδόχων (λευκοί κύκλοι) επιλέγονται με βάση την προβλεπόμενη τιμή η οποία είναι πρόβλεψη του επομένου σημείου συν την μεσαία μεταβολή των γειτόνων. Εάν αυτή η τροχιά είναι λανθασμένη τότε ο μέσος όρος των γειτονικών σημείων θα φεύγει.

Για να συνοψίζουμε, τα τοπικά μοντέλα δίνουν εκπληκτική πρόβλεψη βραχέως χρόνου, επειδή είναι απλά στην σύλληψη, αλλά απαιτούν πολύ υπολογιστικό φορτίο λόγω της εξάρτησης των υπολογισμών στους κοντινούς γείτονες.

6)Ειδικότερα περι δυναμικών συστημάτων και φωνής:

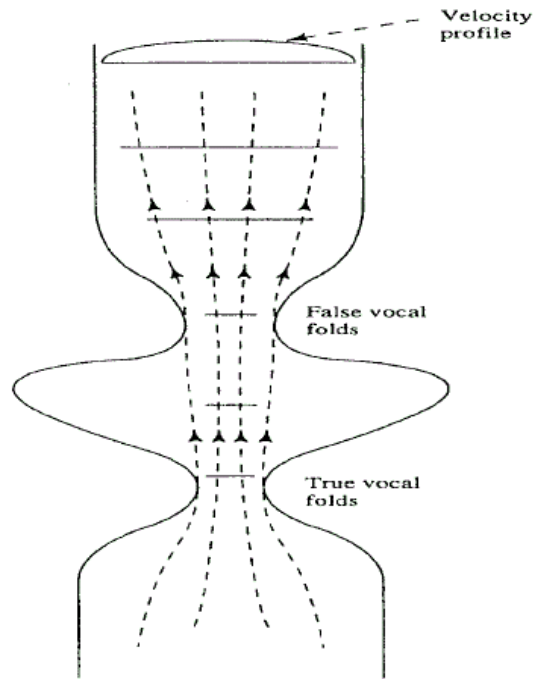
Η παραγωγή φωνής κλασσικά έχει μοντελοποιηθεί σαν γραμμικό σύστημα ,το οποίο παρέχει ένα απλό και εύκολο μαθηματικό φορμαλισμό.Παρά ταύτα,ένας αριθμός από μη γραμμικές επιδράσεις είναι παρούσες σε κάθε φυσική διαδικασία με που φτάνει στα όρια την αποτελεσματικότητα του γραμμικού μοντέλου.Μία πιο προχωρημένη στρατηγική είναι να μοντελοποιήσουμε την φωνή σαν ένα μηγραμμικό δυναμικό σύστημα.

α)Παραγωγή Φωνής

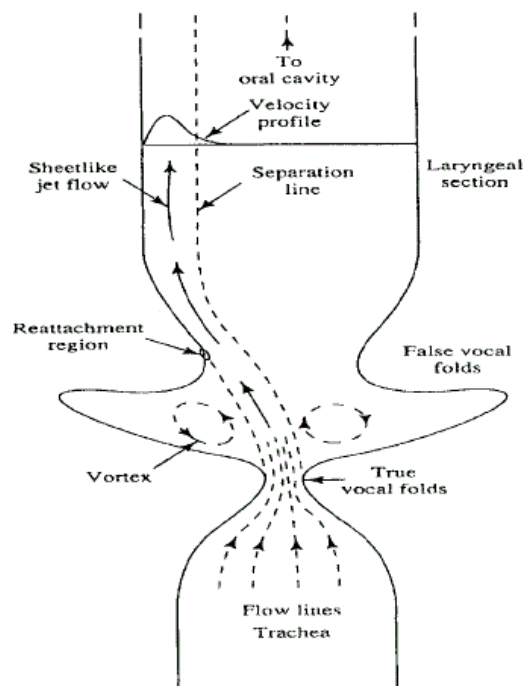
Η παραγωγή φωνής είναι μία πολύ περίπλοκη διαδικασία.Μπορεί να περιγραφεί με πολύ απλές λέξεις ως εξής.Μία πίεση αέρα παράγεται από τους πνεύμονες και το κλυμα πίεσης διαμορφώνεται καθώς αυτό ρέει μέσω του λάρυγγα.Απαραίτητα,ο λάρυγγας έχει φτιαχτεί από 2 συμμετρικές μάζες γνωστές σαν φωνητικές κοιλότητες που είναι ικανές να κλείνουν ολοκληρωτικά μεταξύ τους ,καθώς μετακινούνται μεταξύ τους,δημιουργώντας ένα τριγωνικό άνοιγμα το οποίο καλείται γλωττίδα.Κατά την διάρκεια της κανονικής αναπνοής και και παραγωγής των λεγομένων άφωνων ήχων ,ο αέρας ρέει ελεύθερα μέσω της γλωττίδας.Όταν οι φωνητικές κοιλότητες ταλαντώνονται σε ένα ημί-περιοδικό τρόπο τότε οι φωνητικοί ήχοι παράγονται.Η συχνότητα της διέγερσης είναι γνωστή σαν η κύρια συχνότητα.Η γλωττική κυματομορφή που προκύπτει και διεγείρει τον φωνητικό σωλήνα του οποίου η περιοχή εκτείνεται από τον λάρυγγα μέχρι τα χείλη.Διάφορες καταστάσεις του φωνητικού σωλήνα καταλήγουν να δημιουργούν διαφορετικές διαμορφώσεις της γλωττικής κυματομορφής και παράγουν διαφόρους συγκεκριμένους ήχους.

β)Μη γραμμικότητες στη φωνή:

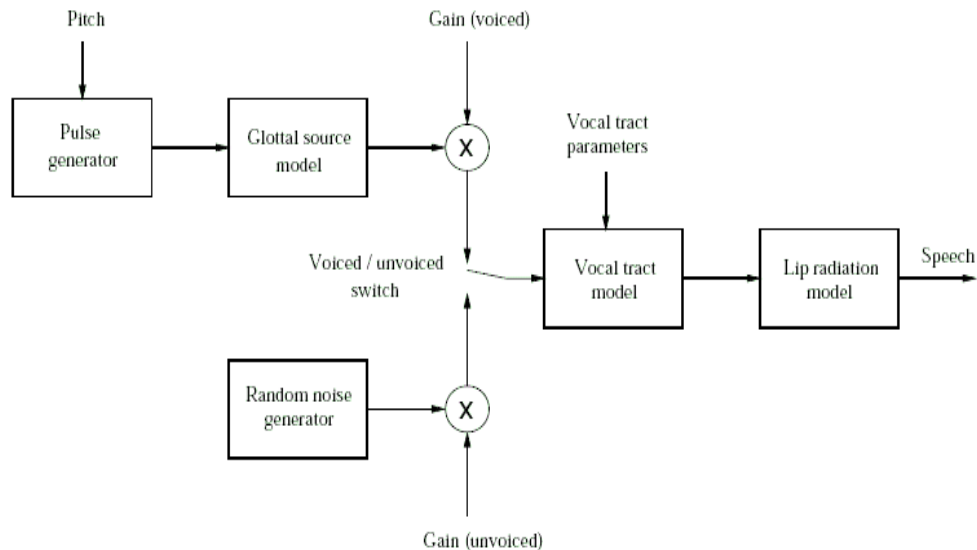
Υπάρχουν γνωστές μη γραμμικές επιδράσεις από την διαδικασία της παραγωγής φωνής.Πρώτα,είχε αποδεχθεί για αρκετό χρόνο ότι ο φωνητικός σωλήνας και οι φωνητικές κοιλότητες δεν λειτουργούν ανεξάρτητα η μία από την άλλη,αλλά υπάρχουν κάποια στοιχεία ότι κάποια μορφή σύζευξης υπάρχει μεταξύ τους όταν η γλωττίδα είναι ανοιχτή.Αυτό μπορεί να προκαλέσει σημαντικές αλλαγές στα χαρακτηριστικά των formants όταν είναι ανοιχτοί ή κλειστοί οι γλωττικοί κύκλοι.Πιο αντιφατικά οι Teager and Teager ισχυρίζονται(βασίζόμενοι σε φυσικές μετρήσεις) ότι οι ακουστικοί ήχοι μπορούν να χαρακτηριστούν από πολύπλοκες ροές αέρα στον ηχητικό σωλήνα που σχετίζεται με jets,παρά από μία καλά ορισμένη ροή .Η τυρβώδης ροή της φύσης είναι επίσης αποδεκτή ότι συμβαίνει και κατά τον άφωνο ήχο,όπου η παραγωγή ήχου οφείλεται σε κάποια συνθλιβή σε κάποιο σημείο του φωνητικού σωλήνα.Επιπλέον οι φωνητικές κοιλότητες απο μόνες τους είναι υπεύθυνες για ακόμη πιο μη γραμμική συμπεριφορά,μιας και οι μύες θα εξαναγκάσουν τον φάρυγγα να έχει μηγραμμικές εφελκυστικές ιδιότητες.Τέτοιες μηγραμμικότητες συμπεριλαμβάνονται απλά σε κάθε απόπειρα μοντελοποίησης της φυσικής διαδικασίας στην ταλάντωση του φωνητικού σωλήνα ,η οποία έχει επικεντρωθεί σαν μοντελοποίηση 2 ή περισσοτέρων μαζών ,όπου η κίνηση των ηχητικών κοιλοτήτων μπορεί να μοντελοποιηθεί σαν μάζες που συνδέονται με ελατήρια με μηγραμμική σύζευξη.Παρατηρήσεις της γλωττικής κυματομορφής ενισχύουν αυτή την ένδειξη ,όπου έχει δείχθει ότι η κυματομορφή μπορεί να αλλάζει σχήμα για διάφορα πλάτη.Τέτοια αλλαγή δεν θα ήταν εφικη με ένα αυστηρά γραμμικό μοντέλο όπου το σχήμα της κυματομορφής δεν είναι ανεπηρρέαστο από αλλαγές του πλάτους.



(a)



Αυτή η υψηλά πολύπλοκη διαδικασία μπορεί γενικά να μοντελοποιηθεί σαν μία γραμμική πηγή -και μοντέλο φίλτρου φωνής.



Όπως φαίνεται και στο παραπάνω σχήμα υπάρχει αυστηρή διακοπή μεταξύ των άφωνων και έμφωνων ήχων. Η γλωττική κυματομορφή των έμφωνων ήχων παράγονται με ένα περιοδικό σήμα, του οποίου η περίοδος είναι ίση με την περίοδο του pitch, όπου η παραγωγή άφωνων ήχων μπορεί να μοντελοποιηθεί σαν λευκός θόρυβος. Αυτά τα σήματα όταν εφαρμοστούν σε ένα αργώς μεταβαλλόμενο γραμμικό φίλτρο του οποίου η συνάρτηση μεταφοράς αναπαριστά την συμβολή του φωνητικού σωλήνα ακολουθούμενου από άλλο ένα φίλτρο το οποίο εκπροσωπεί την ακτινοβολία μέσω των χειλιών.

Έτσι ώστε να φτάσουμε σε αυτό το απλοποιημένο μοντέλο, ένας αριθμός μεγάλων υποθέσεων πρέπει να γίνει. Μερικές από τις υποθέσεις είναι οι ακόλουθες:

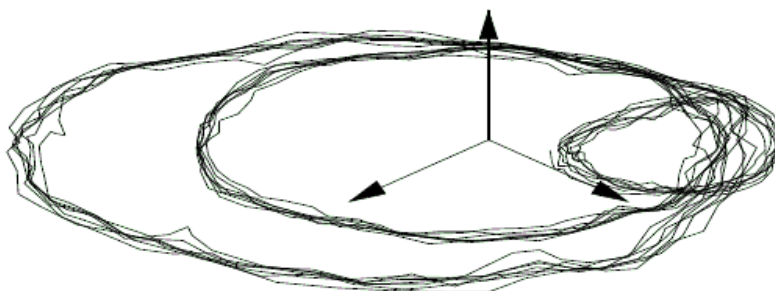
- ο φωνητικός σωλήνας και η φωνητική πηγή είναι ασύζευκτη (επιτρέποντας έτσι τον διαχωρισμό πηγής-φίλτρου)
- η ροή του αέρα μέσω του φωνητικού σωλήνα είναι laminar
- οι φωνητικές κοιλότητες ταλαντώνονται σε έναν επακριβώς περιοδικό τρόπο κατά την διάρκεια της παραγωγής έμφωνου ήχου.
- η κατάσταση του φωνητικού σωλήνα θα αλλάζει πολύ αργά.

Συνεκτιμώντας τα πειστήρια που δείξαμε για τις παραπάνω μη γραμμικότητες της φωνής, το μοντέλο αυτό προφανώς έχει αρκετές αδυναμίες. Αυτές οι υποθέσεις απλά προκαλούν απώλεια πληροφορίας, ενώ την καθιστούν μαθηματικά πιο εύκολη, που σημαίνει ότι με τα γραμμικά μοντέλα δεν μπορεί να συλληφθεί πλήρως η δυναμική του φωνητικού σήματος. Αυτές οι αδυναμίες μπορούν να ειπωθούν στην πράξη π.χ στην παραγωγή φωνής όπου η φωνή έχει αδυναμία φυσικότητας. Αυτό είναι προφανές ακόμη και σε τεχνικές παράταξης όπου αντιγράφουμε και αλλάζουμε τα πραγματικά φωνητικά τμήματα, αλλά είναι ακόμη πιο προφανή όταν επειχθούμε παραγωγή φωνής με γραμμικό μοντέλο. Αυτό προκύπτει διότι το γραμμικό μοντέλο δεν μοντελοποιεί πλήρως το φωνητικό σήμα και δεν μπορεί να αναπαραγάγει ποιότητες που το κάνουν πιο ανθρώπινο, όπως η μεταβλητότητα του pitch.

γ) Η μη γραμμική επιλογή:

Αντί του να επιλέξουμε ένα γραμμικό μοντέλο μπορούμε να επιλέξουμε το μη γραμμικό τότε δεν είναι απαραίτητο να κάνουμε πολλές υποθέσεις σχετικά με τις αντίστοιχες του γραμμικού μοντέλου. Η φωνή μοντελοποιείται σαν η έξοδος ενός αυτόνομου δυναμικού συστήματος, το οποίο έχει μικρό αριθμό από βαθμούς ελευθερίας. Μέσω μίας διαδικασίας χρονικής υστέρησης μπορούμε να κάνουμε

προβολή ενός στάσιμου ήχου ενός φωνήεντος .Στο παρακάτω σχήμα βλέπουμε πως είναι στον 3διάστατο χώρο αυτό.



Όπου το τ επιλέχθηκε στα 0.45msec χρησιμοποιώντας την τεχνική της αμοιβαίας πληροφορίας.Εξετάζοντας την ανακατασκευή του χώρου καταστάσεων από όλες τις πλευρές στον 3-ατο χώρο βλέπουμε πως δεν υπάρχουν τομές στην επιφάνεια και έτσι έχουμε αρκετή προβολή σύμφωνα με την θεωρία των μη τομών.Έτσι τα φωνήεντα μπορούν να ανασκευαστούν με κάποιας μορφής ελκυστή μικρής διάστασης,και αυτό είναι δυνατό λόγω της μη γραμμικής δυναμικής τεχνικής:

$$x_{i+1} = \mathcal{F}\{x_i\}$$

όπου μία μη γραμμική απεικόνιση μεταξύ των προηγούμενων δειγμάτων και του επομένου γίνεται.Αυτή η μικρής διάστασης χαρακτηριστική είναι κοινή στους έμφωνους ήχους,αλλά δεν είναι αληθής για άφωνους ήχους όπως θα δουμε στη συνέχεια.Από το θεώρημα του Τάκενς υπάρχει 1-1 απεικόνιση μεταξύ του F και του μηγραμμικού συστήματος.Έτσι έχουμε ακριβής μοντελοποίησης που σημαίνει ότι αρκετά χαρακτηριστικά της δυναμικής της φωνής έχουν συλληφθεί.

7)Φωνή και Χάος:

Στις προηγούμενες παραγράφους δώσαμε φυσικές αποδείξεις ότι η φωνή είναι μία μηγραμμική διαδικασία σε σχέση με την παραγωγή φωνής .Αυτό έδωσε αρκετή πληροφορία σχετικά με τον θεμελιώδη μηχανισμό της φωνής ότι έχει χαοτικές ιδιότητες.

7.1)Δυναμικές αναλύσεις της φωνής:

Μία από τις πρώτες εργασίες σε αυτό το θέμα έγιναν από τον Tishby,ο οποίος κατασκεύασε ένα μηγραμμικό προβλεπτή βασιζόμενο σε δυναμικά συστήματα.Το πιο κύριο του εύρημα ήταν ότι στο επίπεδο των δυναμικών συστημάτων έδωσε μία προσέγγιση της διάστασης συσχέτισης.Είχε βρεθεί ότι είναι μεταξύ 3 και 5 για έμφωνο ήχο,όπου ο άφωνος ήχος έδωσε υψηλότερη διάσταση μεταξύ 5 και 8.Το τελευταίο ήταν μη αξιόπιστο διότι βασίστηκε σε ελλιπή δεδομένα.Επίσης φαίνεται ότι τα σήματα που αναλύθηκαν ήταν μη στάσιμα,τα οποία μπορούν να δώσουν παραπλανητικά αποτελέσματα.Μία επακόλουθη εργασία υποστήριξε αυτά τα ευρήματα.Ο Townshend έδωσε μία διάσταση συσχέτισης για τα φωνήεντα στο 2.9 και 3.3 και δώσανε μία υλοποίηση ενός μηγραμμικού προβλέπτη που μοντελοποιούσε τις τοπικές δυναμικές.Αυτή η μέθοδος του Casdagli ότι ένα σύστημα είναι υψηλής ή χαμηλής διάστασης και υποδεικνύει ότι η φωνή παρουσιάζεται να είναι μικρής διάστασης.Και πάλι όμως παρουσιάστηκε ότι συνεχές και μη στάσιμο έμφωνο ηχητικό σήμα ότι χρησιμοποιήθηκε για αυτές τις αναλύσεις.Όμως οι McLaughlin&Lowty ανέλυσαν στάσιμους ήχους φωνηέντων και κατέληξαν στα ίδια και βρήκαν μία διάσταση συσχέτισης μεταξύ 3 και 5.Χρησιμοποιώντας λανθασμένες τεχνικές κοντινότερου γειτονα εξαγάγανε ότι η διάσταση είναι 7.Επίσης ο μεγαλύτερος εκθέτης Λιαπούνοφ είναι 0 που σημαίνει την παντελή έλλειψη χάους.

Τεχνικές βασισμένες στην θεωρία της πληροφορίας εκτιμούν την βέλτιστη παράμετρο καθυστέρησης και μία διάσταση η οποία τείνει να εισαγάγει την ανακατασκευή του χώρου φάσης. Τα αποτελέσματα δείχνουν ότι η διάσταση του 3 είναι αρκετή, και δείχνουν ότι τα εκτεταμένα φωνήεντα είναι χαοτικά και έχουν μια μή ακεραία διάσταση συσχέτισης (μεταξύ 1 και 2) και τον υπολογισμό θετικού ρυθμού εντροπίας. Αυτό συμφωνεί με μελέτη του Kumar and Mullick, που βρίσκουν ότι η μεγαλύτερη ιδιοτιμή Λιαπούνοφ και θετική εντροπία, σε μεγάλο πλήθος φωνημάτων και φωνηέντων.

Ο Tokuda et al έχει μία μελέτη για τα ιαπωνέζικα φωνήεντα, και είδαν ότι τα φωνήεντα έχουν χαοτική συμπεριφορά. Κάνουν ευρύτατη εφαρμογή της μεθόδου ανάλυσης σε κύριες συνιστώσες και βρήκαν ότι η κύρια διάσταση είναι 3. Ισχυρίζονται ότι έχουν παρατηρήσει τάση, δίπλωμα και συμπίεση του χώρου καταστάσεων με μελέτη του τομών Πουανκαρέ, που υποδεικνύει χάος. Αυτό επεκτείνεται μέσω της εύρεσης ενός θετικού-μηδενικού-αρνητικού φάσματος Λιαπούνοφ. Τελικά υπολογίζουν ότι η φράκταλ διάσταση της φωνής είναι μεταξύ 2 και 3. Παρα ταύτα επιδεικνύουν ότι τα δεδομένα του με Λιαπούνοφ θέλουν ιδιαίτερη προσοχή, λόγω των πολλών δυσκολιών που ανακύπτουν στους υπολογισμούς. Εν προκειμένο ο αλγόριθμος των Sano and Sawada μπορεί να δώσει παραπλανητικά αποτελέσματα σε πραγματικά δεδομένα, εάν υπάρχει θόρυβος. Επιπλέον εργασία έχει δείξει ότι τα περισσότερα φωνήεντα δείχνουν υψηλό βαθμό μη γραμμικότητας, αλλά όχι ύπαρξη χάους. Έχουν επίσης μελετήσει τις μεταβολές από pitch-to-pitch σε φωνήεντα. Αυτή η μεταβολή jitter μπορεί να θεωρηθεί φράκταλ.

Μία ανεξάρτητη μελέτη δείχνει ότι οι ήχοι των συμφώνων έχουν μεγαλύτερη διάσταση και ότι το φάσμα Λιαπούνοφ έδειξε την ύπαρξη χάους. Επίσης σε σχέση με τα φωνήεντα με αυτή την μελέτη έδειξαν την απουσία του χάους.

7.2) Αλγόριθμος εκτίμησης εκθέτη Λιαπούνοφ (Lyapunov exponent estimation algorithm)

Με βάση τα παραπάνω βλέπουμε ότι δεν υπάρχει πλήρης κατανόηση του φαινομένου του χάους σε σχέση με τους ήχους των φωνηέντων. Μια έρευνα του Banbrook επιχειρεί να λύσει αυτό το θέμα, μέσω ενός εύρωστου αλγορίθμου για τους εκθέτες και έπειτα εφαρμόζοντας το σε μία βάση στάσιμων ήχων φωνηέντων.

Εν γένει υπάρχουν πολλοί αλγόριθμοι που βασίζονται στην χρήση Ιακωβιανών πινάκων για τοπικές δυναμικές γύρω από τον χώρο ανακατασκευής των φάσεων, αλλά είναι επιρρεπείς στην παρουσία θορύβου. Μια μεγάλη πρόοδος έχει γίνει ανάγοντας το πρόβλημα σε αποσύνθεση μοναδιαίας τιμής (singular value decomposition) για την παρουσία θορύβου. Αυτή η μέθοδος αναπτύχθηκε στο πανεπιστήμιο του Εδιμβούργου από τους Banbrook, Ushaw και McLaughlin, χρησιμοποιώντας σαν βάση τους Darbyshire and Broomhead.

Αλγόριθμος:

Σε αυτό το χωρίο, ο αλγόριθμος που προτάθηκε από τους Banbrook, Ushaw and McLaughlin περιγράφεται με κάποια λεπτομέρεια. Όπως συμβαίνει με όλους τους αλγορίθμους υπολογισμού των εκθετών Lyapunov ο αλγόριθμος είναι πολύπλοκος. Για να παράγει σωστά αποτελέσματα βασίζεται σε ένα αριθμό από παραμέτρους οι οποίες μπορούν να καθοριστούν πειραματικά. Παρακάτω εικονίζεται μία επισκόπηση του αλγορίθμου και της λειτουργίας. Όπως μπορεί να ειπωθεί η διαδικασία μπορεί να κατανεμηθεί σε ένα αριθμό από διακριτά βήματα τα οποία περιγράφονται πιο κάτω:

1) Διαδικασία προβολής (Embedding)

Το φωνητικό σήμα μπορεί να προβληθεί και να εγκλωβιστεί σε ένα d-διάστατο χώρο καταστάσεων με τη μέθοδο SVD.

2)Επιλογή γειτονιάς.Η πρακτική μέθοδος για τον υπολογισμό των εκθετών Lyapunov από πειραματικά δεδομένα σχετίζεται με την κατασκευή ενός d-διάστατου χώρου υπερ-σφαίρας που περιέχουν μερικά από τα σημεία του χώρου καταστάσεων και έπειτα την παρατήρηση της εξέλιξης στην υπερ-σφαίρα.Η υπερ-σφαίρα κατασκευάζεται από την εύρεση ενός ικανοποιητικού αριθμού σημείων μέσα σε ένα γραμμικό χώρο γύρω από ένα επιλεγμένο σημείο, $\chi(i)$ και σχηματίζουμε το διανυσματικό σύνολο της γειτονιάς $\Gamma(i)$.Ένας έλεγχος των σημείων γίνεται μέσα στο $\Gamma(i)$ που δεν είναι λανθασμένοι γείτονες,εξασφαλίζοντας ότι θα εξελίσσονται σε κατάλληλες τροχιές.Τώρα μπορούμε με την μορφή πίνακα να το γράψουμε ως εξής,που είναι ο πίνακας γειτνίασης:

$$\mathbf{B}_i = \begin{pmatrix} \gamma_1 - \mathbf{x}_i \\ \gamma_2 - \mathbf{x}_i \\ \vdots \\ \gamma_M - \mathbf{x}_i \end{pmatrix}$$

όπου $\gamma(v)$ είναι η v-οστη συνιστώσα του $\Gamma(i)$.Είναι απαραίτητο να μετασχηματίσουμε τον πίνακα έπειτα από ένα αριθμό βημάτων, a , μιας και τα σημεία γίνονται μέχρι στιγμής όλο και πιο διαχωρίσιμα και έτσι καταρρέει η υπόθεση της τοπικής γραμμικότητας.Η επιλογή αυτής της παραμέτρου,καλείται αριθμός των βημάτων εξέλιξης ,καθώς και ο αριθμός των σημείων στην υπερ-σφαίρα,που πρέπει να γίνει πειραματικά.

3)Υπολογισμός του πίνακα εφαπτομένης.

Εαν μεταξύ ενός βήματος $B(i)$ και του πίνακα εξελισσόμενης γειτνίασης $B(i+a)$,τότε είναι μικρό αρκετά ώστε να χαρακτηριστεί τοπικά γραμμικός,έπειτα οι ιδιοτιμές της απεικόνισης του πίνακα εφαπτομένης μεταξύ των 2 πινάκων θα μας δώσει τους εκθέτες Lyapunov.Η απεικόνιση εφαπτομένης $T(i)$ δημιουργεί μία γραμμική σχέση μεταξύ $B(i)$ και $B(i+a)$ ως εξής:

$$\mathbf{B}_{i+a}^T = \mathbf{T}_i \mathbf{B}_i^T$$

όπου αναγεται σε μία άλλη μορφή:

$$\mathbf{B}_i^+ \mathbf{B}_{i+a} = \mathbf{T}_i^T$$

που λόγω της ύπαρξης του ψευδοαντίστροφου τότε λύνεται με SVD.

4)Μέσοι εκθέτες:

Η απεικόνιση εφαπτομένης $T(i)$ δίνει μία περιγραφή της δυναμικής του τοπικού συστήματος.Για να είμαστε ικανοί να υπολογίσουμε την καθολική δυναμική,μια σειρά από απεικονίσεις μπορούν να γίνουν σε ολόκληρο τον χώρο φάσης και πρέπει να συνδιαστούν μεταξύ τους.Αυτό γίνεται μέσω της παραγοντοποίησης QR,που επιτρέπει ο κάθε πίνακας να γραφεί σαν:

$$\mathbf{A} = \mathbf{Q}\mathbf{R}$$

όπου $\mathbf{Q}$ έχει ορθογώνιες στήλες και $\mathbf{R}$ έχει ένα άνω-δεξιά τριγωνικό πίνακα με θετικές τιμές στην διαγώνιο.Αυτό επιτρέπει οι καθολικές ιδιοτιμές Lyapunov να υπολογιστούν ως εξής:

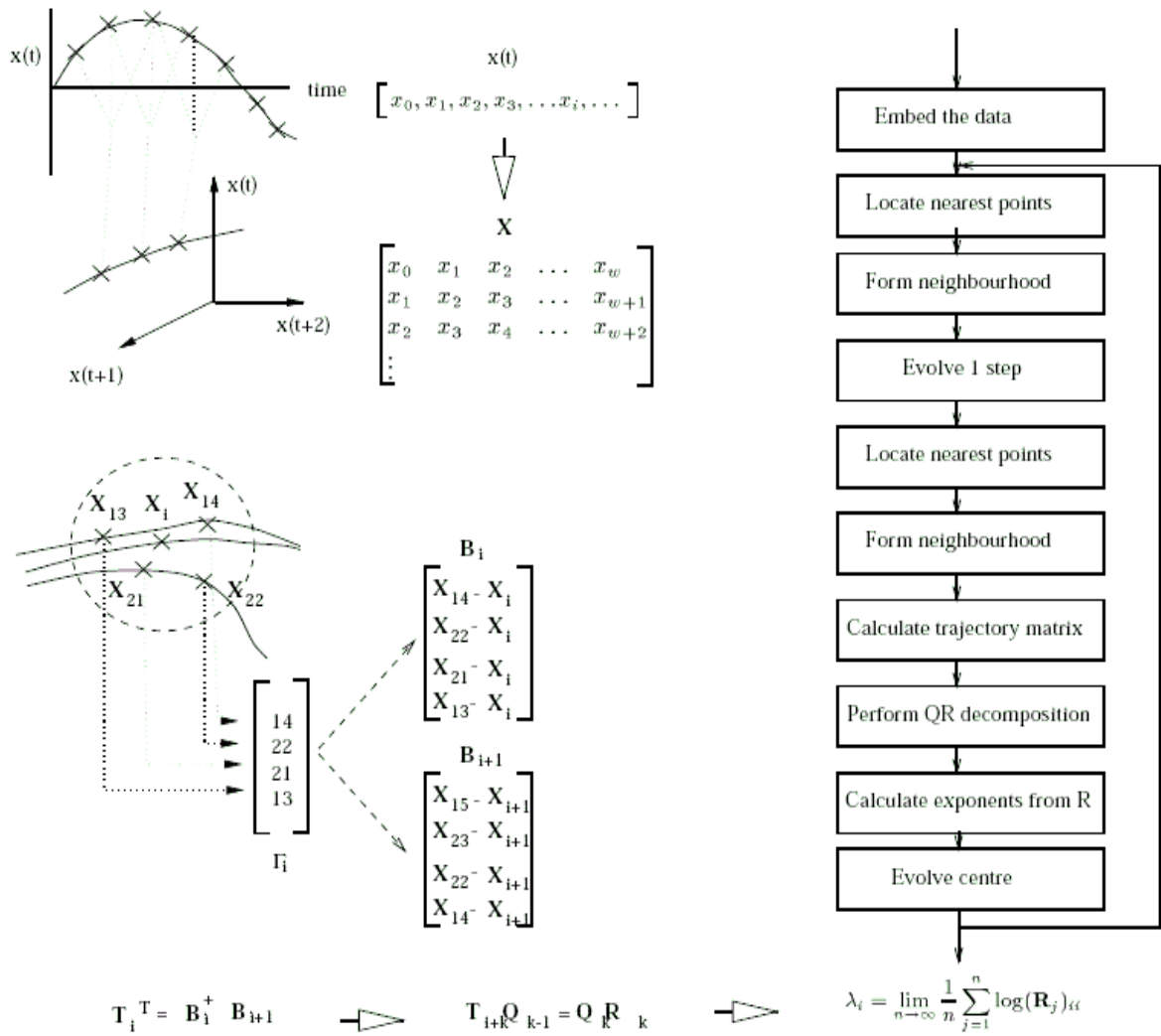
$$\lambda_i = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \log(\mathbf{R}_j)_{ii} \quad 1 \leq i \leq d$$

Όπως αναφέρθηκε πιο πάνω,υπάρχουν ένας αριθμός από παραμέτρους που μπορούν να επιλεγούν ώστε να μας δώσουν σωστά αποτελέσματα από τον αλγόριθμο.

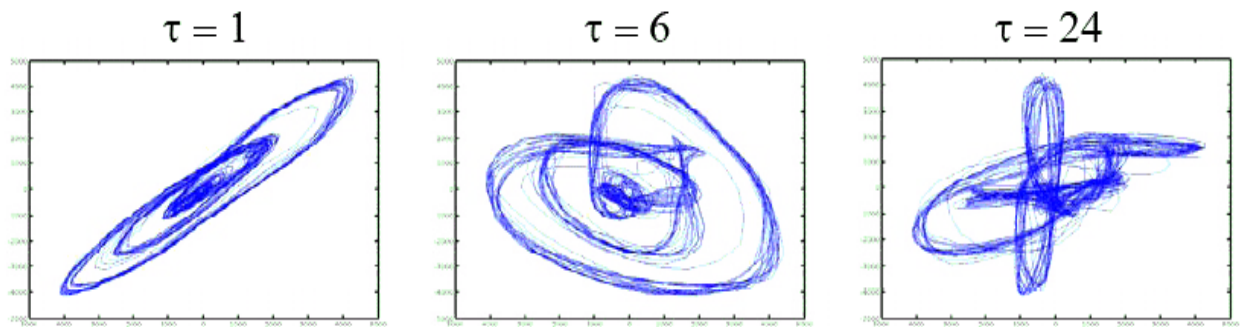
Μερικά από αυτά είναι:

- Ο αριθμός των σημείων στο υπερ-επίπεδο, τα οποία πρέπει να είναι όσο μεγαλύτερα γίνεται χωρίς όμως να χαλάμε την υπόθεση της τοπικής γραμμικότητας.
- Ο αριθμός της διάστασης προβολής(embedding dimension). Το οποίο μπορεί να υπολογιστεί από την διάσταση συσχέτισης και πρέπει να πειραματιστούμε με αυτό.
- το μέγεθος του παραθύρου SVD έτσι ώστε να είμαστε ικανοί να κάνουμε μείωση και αφαίρεση του θορύβου.
- Η παράμετρος καθυστέρησης. Υπάρχει μία διελκυστίνδα μεταξύ του μεγάλου αριθμού βημάτων που δείχνουν μεγαλύτερη αλλαγή, ενάντια στο δεδομένο ότι η γειτονία θα διασκεδαστεί αρκετά και τότε δεν θα είναι πια αρκετά τοπικά γραμμική.
- συνολικός αριθμός απεικονίσεων εφαιπτομένης που σχετίζονται τουλάχιστον με την εξέλιξη ενός βήματος κατά την διαδικασία ανακατασκευής του χώρου καταστάσεων.
- Ο αριθμός των δεδομένων , που μπορούν να είναι αρκετά μεγάλα έτσι ώστε α μπορούν να χωρούν αρκετά στις περιοχές του χώρου φάσης κατά την ανακατασκευή.

Ο αλγόριθμος έχει και ικανότητα μείωσης του θορύβου . Αυτό σχετίζεται με μία διαφορετική αντιμετώπιση και παρουσίαση του πινάκου, που επιτρέπει την λειτουργία μέσου όρου(averaging operation) να γίνει. Πειράματα ακόμη και σε θορυβώδη δεδομένα Lorenz δείχνουν σοβαρή βελτίωση στην εκτίμηση της θετικής τιμής του εκθέτη Lyapunov με αυτή την τεχνική.



Παραδείγματα διαφόρων τιμών καθυστέρησης τ



Κεφάλαιο 5

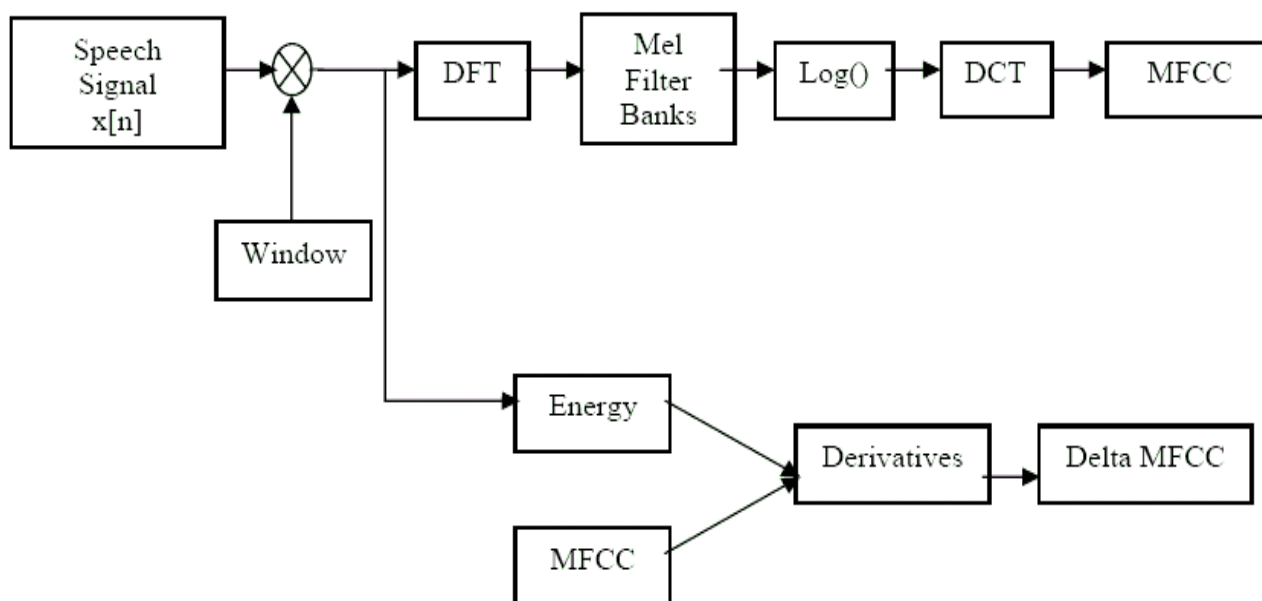
Αναπαράσταση φωνής

1) MEL FREQUENCY CEPSTRUM COEFFICIENTS

Οι πιο ευρύτατα διαδεδομένοι τρόποι εξαγωγής χαρακτηριστικών για αυτόματη αναγνώριση είναι οι συντελεστές της mel συχνότητας. Εκτός από το μοντέλο της γραμμικής πρόβλεψης που κυρίως βασίζεται στην μοντελοποίηση του φωνητικού σωλήνα, η κλίμακα του μελ για τους συντελεστές στο φάσμα της ισχύος της φωνής έτσι ώστε να μιμηθεί το ανθρώπινο ακουστικό μηχανισμό.

1) Εισαγωγή:

Μία τυπική MFCC επεξεργασία, στο πρώτο βήμα γίνεται παραθυροποίηση του σήματος έτσι ώστε να τεμαχιστεί σε πλαίσια. Μιας και τα υψηλής συχνότητας formants έχουν μικρότερο εύρος από αυτά με χαμηλή συχνότητα, τότε τα υψηλής συχνότητας μπορούν να ενισχυθούν ώστε να έχουν όλα παρόμοιο πλάτος με όλα τα formants. Μετά την παραθύρωση κάνουμε μετασχηματισμό FFT για να βρούμε το φάσμα ισχύος για κάθε πλαίσιο. Ομως έπειτα έχουμε μια συστοιχία φίλτρων για επεξεργασία του φάσματος ισχύος, το οποίο χρησιμοποιεί την κλίμακα μελ. Έπειτα ο διακριτός μετασχηματισμός συνημιτόνου χρησιμοποιείται για να μετατραπεί στο φάσμα ισχύος σε λογαριθμική κλίμακα για να υπολογιστούν οι συντελεστές MFCC, όπως εικονίζεται παρακάτω.

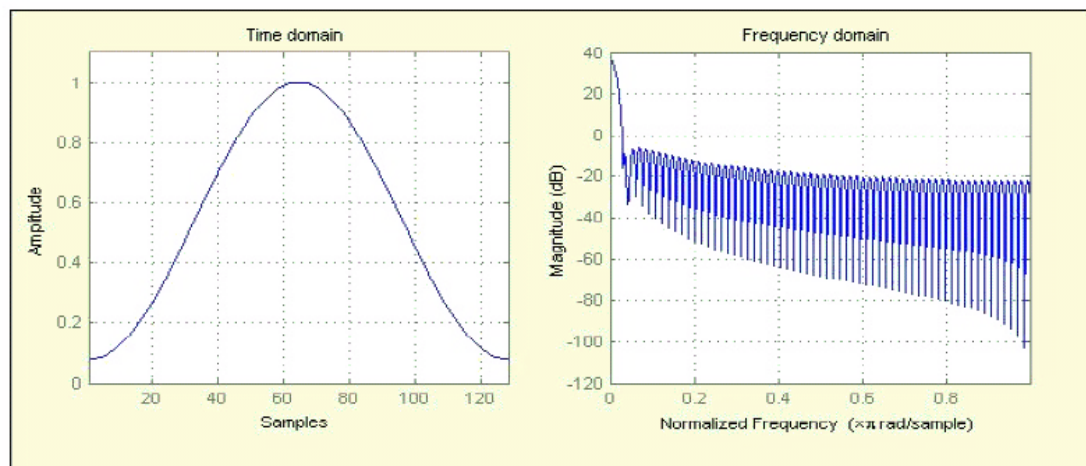
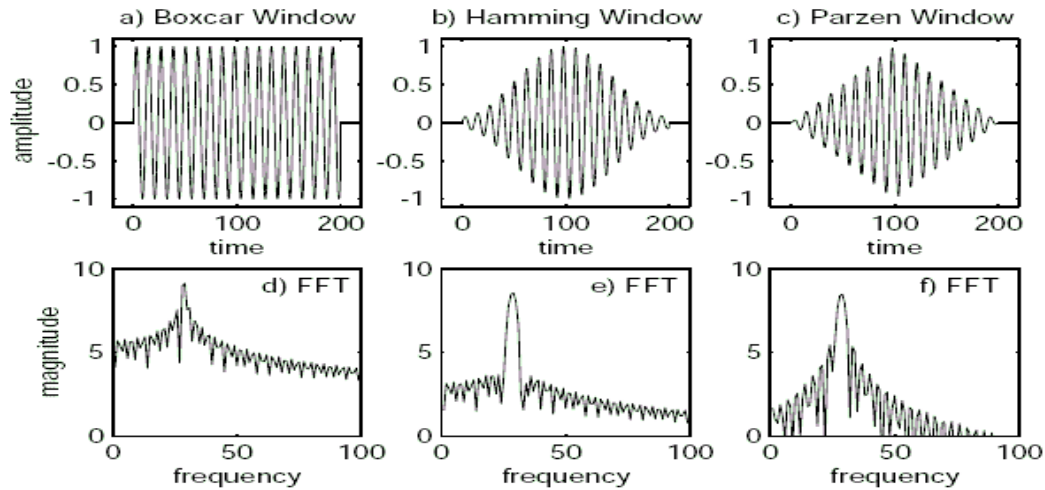


2) Φασματική ανάλυση:

Ο φασματικός υπολογισμός των σημάτων είναι αξιόπιστος στην περίπτωση που το σήμα που επεξεργαζόμαστε θεωρούμε πως είναι στάσιμο. Αυτό υποδηλοί ότι τα χαρακτηριστικά της φωνής δεν αλλάζουν αρκετά μέσα σε λίγο χρόνο έτσι εφαρμόζουμε το παράθυρο στο σήμα. Έτσι το μέγεθος του παραθύρου και το αποτέλεσμα του σχήματος συμβάλλουν στον να βρεθούν οι συντελεστές των χαρακτηριστικών. Τυπικά παράθυρα στην επεξεργασία σήματος είναι τα ορθογώνια

,τα παράθυρα Hamming και Hanning.Κυρίως χρησιμοποιείται το Hamming διότι εγκλωβίζει την απώλεια στους λοβούς της ενέργειας του σήματος και είναι πιο βέλτιστα από άλλα.Τυπικά χρησιμοποιούνται παράθυρα 10-25 msec,ο τύπος του παραθύρου είναι:

$$W(n) = 0.54 + 0.46\cos(2\pi n (N-1))$$



Ο FFT χρησιμοποιείται για να μετατρέπει ένα σήμα στο πεδίο των συχνοτήτων.Η φάση του σήματος (ειδικά για τα ακουστικά σήματα μας είναι αδιάφορη) διότι δεν παρέχει πληροφορία από ακουστική σκοπιά.

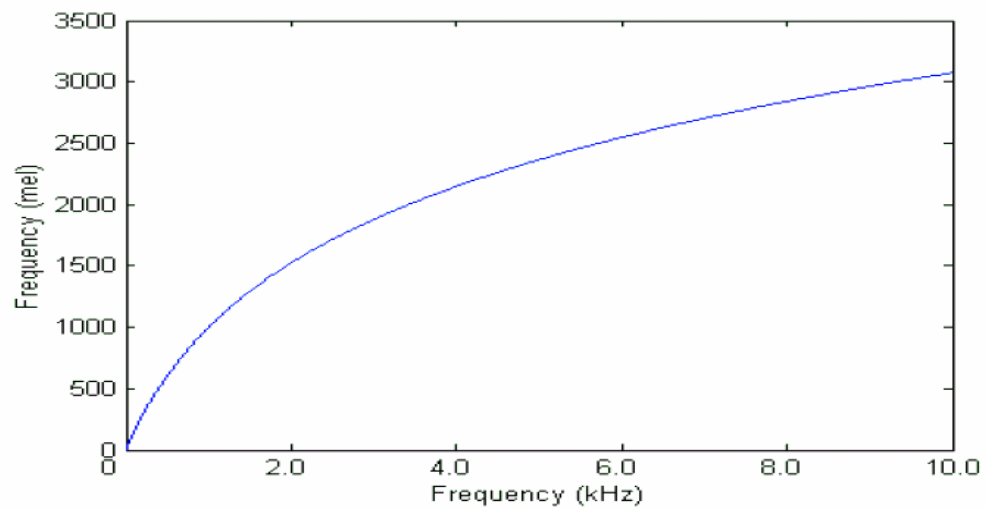
3)Επεξεργασία από συστοιχία φίλτρων(Filter Bank Processing)

Τα φίλτρα χρησιμοποιούνται για να δώσουμε έμφαση σε συγκεκριμένες συχνότητες για το φάσμα ισχύος της φωνής όπως θεωρείται πως κάνει αδρά μια στάθμιση αντιληπτική.Τα περισσότερα επεξεργάζονται το φάσμα κάτω από το 1KHZ μια και η φωνή περιέχει το πιο χρήσιμο τμήμα της πληροφορίας της εκεί,όπως το πρώτο

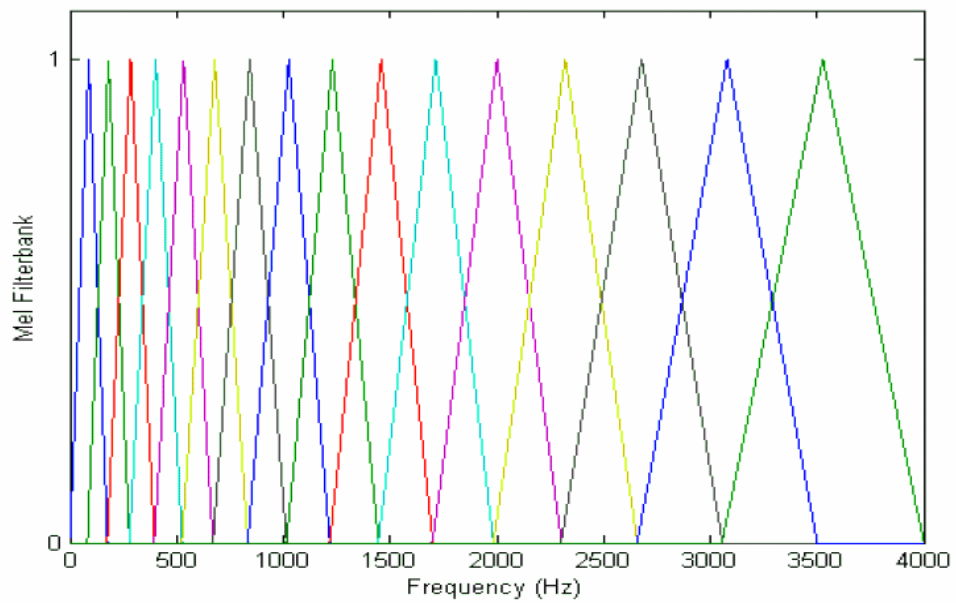
$$mel(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

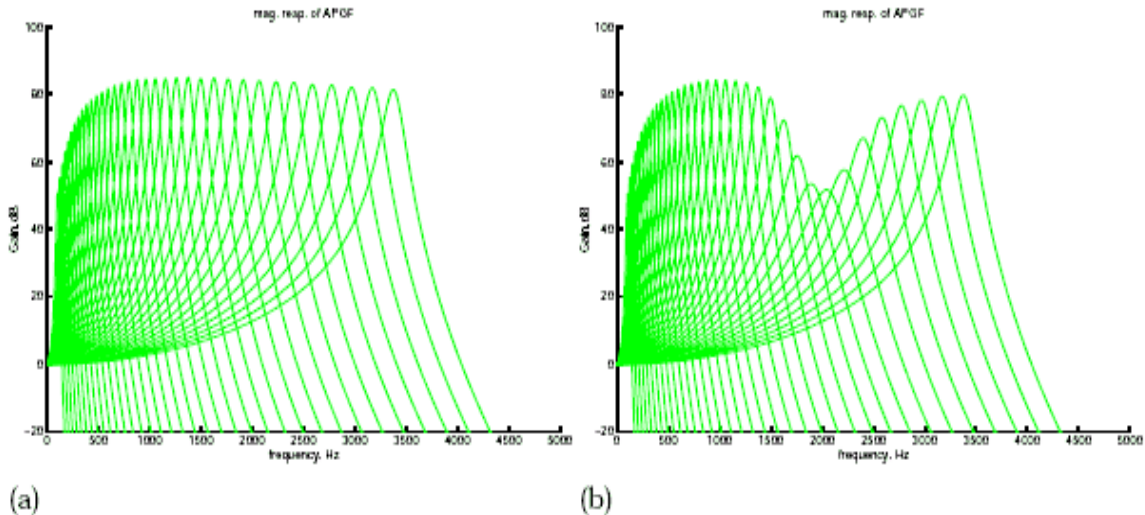
formant που είναι στις χαμηλότερες συχνότητες.

Η αντίστοιχη κλίμακα Μελ:



και η αντίστοιχη συστοιχία φίλτρων:





δ)Υπολογισμός των φασματικών παραμέτρων:

Μετά από την διάσχιση του φάσματος από την συστοιχία των φίλτρων,τότε βρίσκεται ο λογάριθμος των παραμέτρων.Το ανθρώπινο αυτί επίσης ομαλοποιεί το φάσμα και χρησιμοποιεί λογαριθμική κλίμακα.Η λογαριθμική κλίμακα είναι χρήσιμη για την έμφαση στα formants σε θορυβώδη φωνή.Ο αντίστροφος DFT γίνεται πάνω στο λογάριθμο του μεγέθους στην συστοιχία των φίλτρων και την έξοδο.Μιας και το λογαριθμικό φάσμα ισχύος είναι συμμετρικό και αληθινό,ο αντίστροφος DFT μπορεί να αναχθεί στο διακριτό μετασχηματισμό συνημιτόνου.Αυτός ο μετασχηματισμός απομπλέκει τα χαρακτηριστικά,το οποίο οδηγεί σε διαγώνιους πίνακες μεταβλητότητας απ'ότι πλήρων πινάκων καθώς η μοντελοποίηση των παραμέτρων των χαρακτηριστικών μπορεί να γίνει με γραμμικό συνδιασμό γκαουσιανών συναρτήσεων.Με αυτό τον τρόπο ελαχιστοποιείται η χρονική πολυπλοκότητα και το υπολογιστικό φορτίο αυτών.Αυτό είναι ιδιαιτέρως χρήσιμο για μεγάλα λεξικά.Μιας και ο μετασχηματισμός συμπυκνώνει την πληροφορία στους χαμηλοτερης τάξης συντελεστές.Τυπικά ο αριθμός των συντελεστών είναι μεταξύ 8 και 13.

Ο λογάριθμος της ενέργειας του πλαισίου(ο μηδενικός όρος) μπορεί να προστεθεί στην κλιμακα μελ διότι διαφορετικά φωνήματα έχουν διαφορετική ενέργεια.Για να συλλάβουμε και τις δυναμικές ιδιότητες της φωνής ,τότε χρησιμοποιούμε της δελτα και δελτα δελτα παραμέτρους.Πολλά συστήματα αναγνώρισης φωνής χρησιμοποιούν αυτούς τους συντελεστές για να αυξήσουν την απόδοση.

ε)Ευρωστία στην αναγνώριση φωνής.

Η ευρωστία κυρίως αναφέρεται στην διατήρηση υψηλής απόδοσης για όλα τα δυνατά δεδομένα που αντιμετωπίζουμε στην πραγματική ζωή.Ο θόρυβος,η γραμμική παράλλαξη στα κανάλια επικοινωνίας και στην μεταβλητότητα των ομιλητών και συνήθως μειώνουν την απόδοση της αναγνώρισης.Έτσι για να παρακάμψουμε αυτά τα προβλήματα πολλοί αλγόριθμοι έχουν προταθεί.Επίσης υπάρχουν και μοντέλα παραμέτρων που μπορούν να προσαρμοστούν άνετα στο περιβάλλον μέσω καλής εκπαίδευσης και έτσι πετυχαίνεται υψηλός ρυθμός αναγνώρισης.Αργότερα θα αναφέρουμε την επεξεργασία RASTA η οποία χρησιμοποιείται για να αποφύγουμε την γραμμική παράλλαξη.

στ)Προσθετικός θόρυβος και μοντέλο γραμμικής παράλλαξης

Το παρακάτω σχήμα παρουσιάζει 2 κύριες πηγές χειροτέρευσης, τον αθροιστικό θόρυβο και την γραμμική διατάρραξη. Θα χρησιμοποιήσουμε τον συνελκτικό θόρυβο και την γραμμική διατάρραξη που συνελκίζεται με το καθαρό σήμα. Θεωρούμε την καθαρή φωνή $u(t)$, ο αθροιστικός θόρυβος $n(t)$ και η γραμμική διατάρραξη $d(t)$ και το συνολικό σήμα είναι $y(t)$. Και δίδεται από τον παρακάτω τύπο:

$$y(t) = [u(t) + n(t)] * d(t) = x(t) * d(t)$$

Εάν θεωρήσουμε τον αθροιστικό θόρυβο που είναι στάσιμος και ασυσχέτιστος με την

$$P_y(f) = P_x(f) \cdot |D(f)|^2 = [P_u(f) + P_n(f)] \cdot |D(f)|^2$$

καθαρή φωνή τότε το φάσμα ισχύος του $y(t)$ είναι:

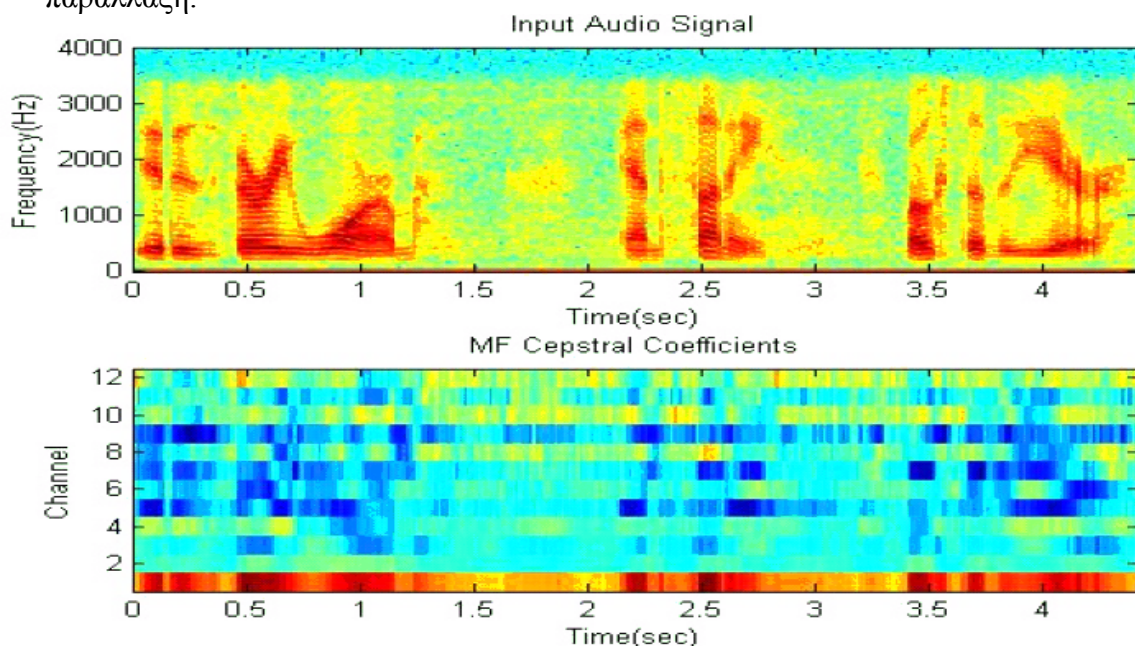
Λογαριθμίζοντας έχουμε:

$$\log[P_y(f)] = \log[P_x(f)] + \log|D(f)|^2$$

Με διαφορετική μορφή έχουμε:

$$\log[P_y(f)] = \left\{ \log[P_u(f)] + \log\left[1 + \frac{1}{SNR(f)}\right] \right\} + \log|D(f)|^2$$

όπου $SNR(f)$ υποδεικνύει τον σηματοθορυβικό λόγο. Από αυτή την εξίσωση βλέπουμε για υψηλές τιμές λόγου τότε η χειροτέρευση βασίζεται στην γραμμική παράλλαξη.



2) RASTA-PLP PROCESSING

Εισαγωγή:

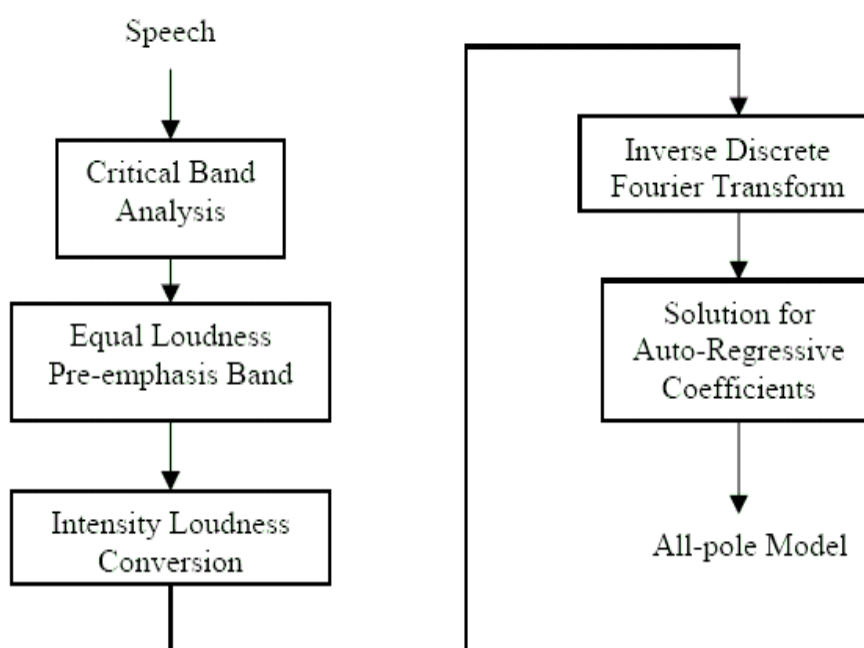
Η ανάλυση με την γραμμική πρόβλεψη έχει ευρύτατα χρησιμοποιηθεί και χρησιμοποιείται για ανάλυση της φωνής και αναγνώριση της φωνής. Σε αυτό το κεφάλαιο θα παρουσιάσουμε μια διαφορετική τεχνική για την ανάλυση της φωνής, η

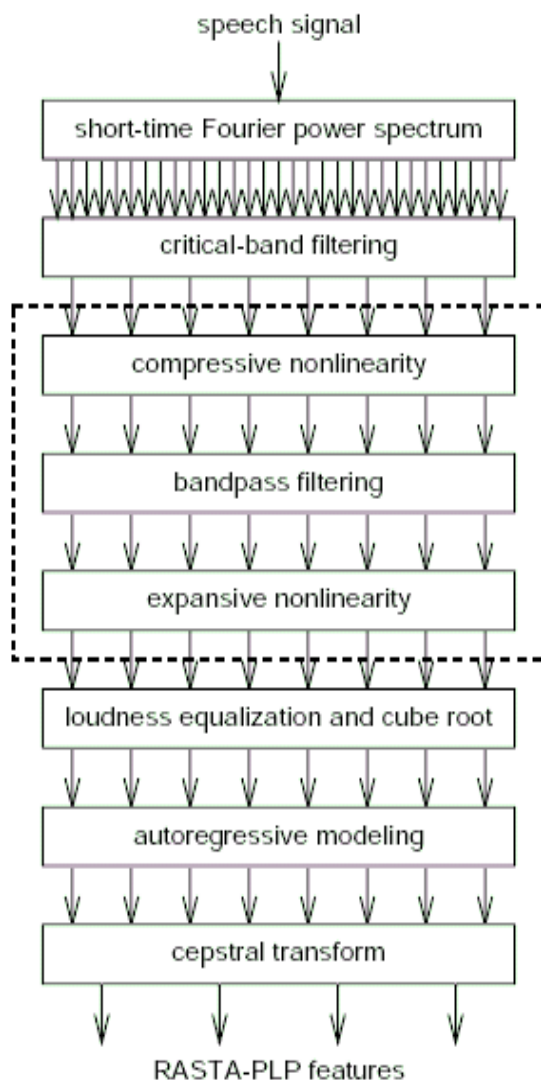
οποία είναι η αντιληπτική γραμμική προβλεπτική ανάλυση(perceptual predictive linear analysis).Χρησιμοποιεί ιδέες από την ψυχοακουστική της ακοής και παρουσιάζει μία χαμηλότερη διάσταση αναπαράστασης της φωνής,η οποία είναι αρκετά χρήσιμη για αναγνώριση φωνής ανεξαρτήτως ομιλητή.Μαζί με την PLP υπάρχει και η μέθοδος του σχετικού φάσματος(relative spectra -RASTA) που μπορεί να χρησιμοποιηθεί συναθροιστικά στους αναγνωριστές.Η RASTA χρησιμοποιεί φιλτράρισμα στο λογαριθμικό πεδίο του φάσματος ισχύος για να παράξει εύρωστες αναπαραστάσεις της φωνής.

α)Συντελεστές αντιληπτικής γραμμικής πρόβλεψης:

Το μοντέλο της γραμμικής πρόβλεψης χρησιμοποιεί την συνάρτηση μεταφοράς όλων των πόλων του φωνητικού σωλήνα και χρησιμοποιείται για να εκτιμήσει το φάσμα ισχύος της φωνής.Εάν η τάξη του μοντέλου είναι καλά επιλεγμένη τότε προσεγγίζει καλά τις περιοχές του φάσματος μεγάλης ενέργειας και ομαλοποιεί την αρμονική δομή και άλλες άσχετες λεπτομέρειες του φάσματος.

Αυτές οι ιδιότητες έχουν οδηγήσει το μοντέλο LP να χρησιμοποιηθεί για πολλές περιοχές της έρευνας της φωνής.Το μειονέκτημα αυτό του μοντέλου είναι ότι αυτό αναλύει το σήμα της φωνής βασισμένο στον μηχανισμό της άρθρωσης.Το μοντέλο αυτό προσεγγίζει την φωνή σε όλες τις συχνότητες με τον ίδιο τρόπο.Αυτό δεν είναι σωστή προσέγγιση για την ακοή επειδή είναι πιο ευαίσθητη στις συχνότητες μεσαίας περιοχής.Για να προσεγγίσουμε αυτή την ιδιότητα υπάρχει η τεχνική διαμόρφωσης του φάσματος(spectral warping) όπως βασισμένη στην κλίμακα μελ.Η PLP διαμορφώνει το φάσμα ισχύος της φωνής πριν την προσέγγιση από το μοντέλο της αυτοπαλινδρόμησης.Η PLP χρησιμοποιεί χαμηλής τάξης μοντέλο όλων των πόλων ,έτσι για να γίνεται πιο αποδοτικός υπολογισμός.Είναι συνεπές με σχετικά φαινόμενα που παρατηρούνται κατά την διαδικασία της αντίληψης της φωνής με βάση την θεωρία του δευτέρου formant και την θεωρία φασματικής ολοκλήρωσης του Μπαρκ.Η τεχνική της PLP σε μπλοκ διάγραμμα είναι η παρακάτω για να βρει τους συντελεστές.





β)1)Φασματική ανάλυση:

Το σήμα της φωνής διαιρείται σε πλαίσια και σταθμίζεται με ένα παράθυρο, το οποίο είναι συνήθως το παράθυρο Hamming. Μετά την παραθύρωση, το DTFT χρησιμοποιείται για να μετατρέψει το πλαίσιο της φωνής στο πεδίο της συχνότητας. Στην βραχέως χρόνου ανάλυση το φάσμα ισχύος είναι το εξής:

$$P(\omega) = \text{Re}(S(\omega))^2 + \text{Im}(S(\omega))^2$$

β)2)Ανάλυση του κριτικού εύρους:

Χρησιμοποιώντας τα κριτικά εύρη για την ανάλυση του σήματος της φωνής είναι σημαντικό και από τον μηχανισμό ακοής. Το κριτικό εύρος για μία δεδομένη κεντρική συχνότητα ορίζεται σαν το μικρότερο εύρος συχνοτήτων γύρω από αυτό το οποίο ενεργοποιεί το ίδιο μέρος της βασιλικής μεμβράνης του αυτιού. Συνεχείς τόνοι εμπίπτουν στην ίδια κριτική μπάντα συχνοτήτων που δε αυξάνουν την αντιλαμβανόμενη ένταση πάνω από ένα συγκεκριμένο τόνο σαν να είχαν όλοι την ίδια ηχητική πίεση. Έτσι το κριτικό εύρος μπορεί να αναπαρασταθεί σαν την ισχύ του αυτιού για ταυτόχρονους τόνους. Όμοια, διαφορικό κατώφλι μπορεί να ορισθεί σαν η ελάχιστη αντιληπτική διαφορά για μία μοναδική συχνότητα.

Έπειτα από την εύρεση του φάσματος ισχύος $P(\omega)$ τότε αυτό συνελίσσεται γύρω από τον άξονα συχνοτήτων ω στην Μπαρκ συχνότητα, έχουμε:

$$\Omega(\omega) = 6 \ln \left[\frac{\omega}{1200\pi} + \left(\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right)^{0.5} \right]$$

Αν κάνουμε συνέλιξη του φάσματος ισχύος warping με βάση την παρακάτω συνάρτηση:

$$\Psi(\Omega) = \begin{cases} 0 & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & -1.3 \leq \Omega \leq -0.5 \\ 1 & \text{for } -0.5 \leq \Omega \leq 0.5 \\ 10^{-1.0(\Omega-0.5)} & 0.5 \leq \Omega \leq 2.5 \\ 0 & \Omega > 2.5 \end{cases}$$

Αυτή η κρίσιμου εύρους μάσκα καμπύλη είναι μία προσέγγιση στην ασύμμετρη καμπύλη μάσκας του Schroeder. Χρησιμοποιεί την πρόταση του Zwicker, ότι το κάθε σχήμα των ακουστικών φίλτρων είναι σχεδόν σταθερό στην κλίμακα του Μπαρκ. Αυτή η συνέλιξη μειώνει την φασματική ανάλυση. Η κλίμακα του Μπαρκ υποδειγματολειπτήται με επαναδειγματοληψία κάθε Μπάρκ που καλύπτει από 0 έως 16.9 Μπαρκ (π.χ 0-5 KHz).

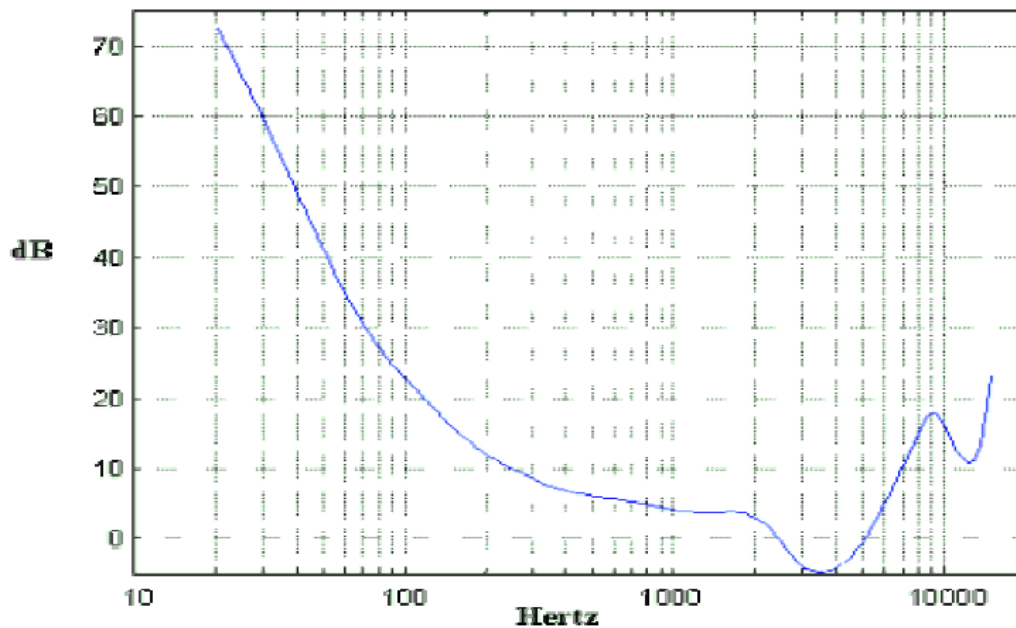
β)γ) Πριέμφαση ίσης ακουστικότητας.

Μετά την συνέλιξη και την επαναδειγματοληψία, το αποτέλεσμα επιδέχεται πριέμφαση με το να προσομοιώνει την καμπύλη της ίσης έντασης. Αυτή η συνάρτηση προσομοιώνει την ευαισθησία του αυτιού στην ακοή στα 40-dB και εφαρμόστηκε με προσαρμογή από τους Makhoul and Cosell:

$$E(\omega) = (\omega^2 + 56.8 \times 10^6) \omega^4 ((\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9))$$

Αυτή η συνάρτηση μεταφοράς του φίλτρου με ασυμπτωτικές των 12dB/oct μεταξύ 0 και 400Hz, 0dB/oct μεταξύ 40 και 1200Hz, 6 dB/oct μεταξύ 1200 και 3100 Hz και 0dB/oct από 3100 μέχρι την συχνότητα του Nyquist.

Αυτή η παραπάνω εξίσωση χρησιμοποιείται για να προσομοιώνει την ακουστική ανάλυση ισχύος. Οι μεταβολές της ισχύος από άτομο σε άτομο εικονίζονται στην παρακάτω εικόνα.



Με αυτή τα πρώτα δείγματα και τα τελευταία δεν είναι καλά ορισμένα και γίνονται ίσα με τις τιμές των πιο κοντινών γειτόνων τους. Τα πρώτα βήματα αυτής της τεχνικής μοιάζουν αρκετά με τον υπολογισμό των συντελεστών των συχνοτήτων μελ. Πραγματικά χρησιμοποιούμε PLP συστοιχία φίλτρων εκτός από την χρήση της συστοιχίας μελ. Η συνέλιξη και η πριέμφαση γίνονται για κάθε δείγμα του E στο $P(\omega)$ πεδίο με μία σταθμισμένη φασματική άθροιση. Το φασματικό δείγμα δίδεται από την εξής εξίσωση:

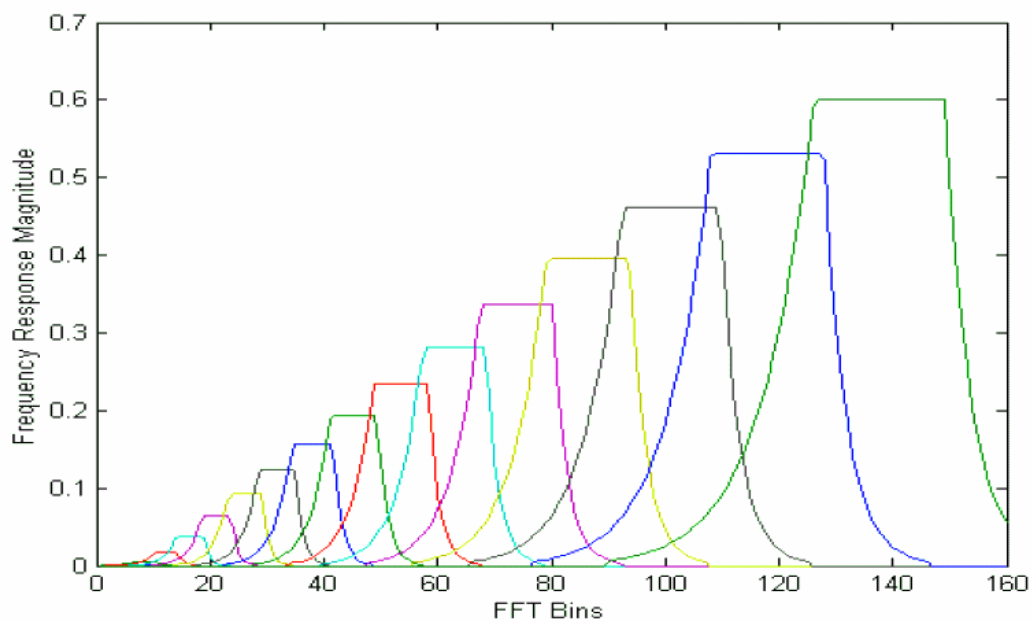
$$E(\Omega(\omega_i)) = \sum_{\omega=\omega_{ik}}^{\omega_{lm}} w_i(\omega) P(\omega)$$

όπου i είναι το νούμερο των σταθμισμένων συναρτήσεων, όπου k και m είναι ο πρώτος και ο τελευταίος δείκτης του ω στο πεδίο της ισχύος του φάσματος.

Η αντίστροφη δίδεται:

$$w = 1200\pi \sinh(\Omega / 6)$$

Οι συνάρτησεις στάθμισης υπολογίζονται για δεδομένη συχνότητα δειγματοληψίας και το μέγεθος του FFT. Μπορεί ναδειχθεί ότι το φασματικό διάστημα ολοκλήρωσης αυξάνει με την συχνότητα. Η χαμηλής συχνότητας κλίση είναι λιγότερο απότομη απ' ό,τι για την κλίση υψηλής συχνότητας. Επίσης το πλάτος αυξάνει όσο αυξάνει η συχνότητα.



β)δ)Μοντελοποίηση αυτοσυσχέτισης:

Το τελικό βήμα είναι να προσεγγίσει την $\Phi(\Omega)$ μέσω του φάσματος ενός μοντέλου ολικών πόλων χρησιμοποιώντας την μέθοδο αυτοσυσχέτισης για την φασματική μοντελοποίηση όλων των πόλων. Αντίστροφος DFT εφαρμόζεται στο $\Phi(\Omega)$ για να έχουμε μία συνάρτησης αυτοσυσχέτισης και τις πρώτες $M+1$ τιμές της αυτοσυσχέτισης χρησιμοποιούνται για να λύνονται οι εξισώσεις Yule-Walker. Οι τελικο συντελεστές αυτοσυσχέτισης μπορούν να σε άλλες παραμέτρους που μας ενδιαφέρουν.

Η ανάλυση PLP είναι όμοια με την ανάλυση συχνοτήτων μελ. Οι διαφορές είναι η χρήση της κλίμακας μελ και χρησιμοποιώντας το μοντέλο ολικών πόλων αντί του διακριτού μετασχηματισμού συνημιτόνων για να βρεθούν οι συντελεστές.

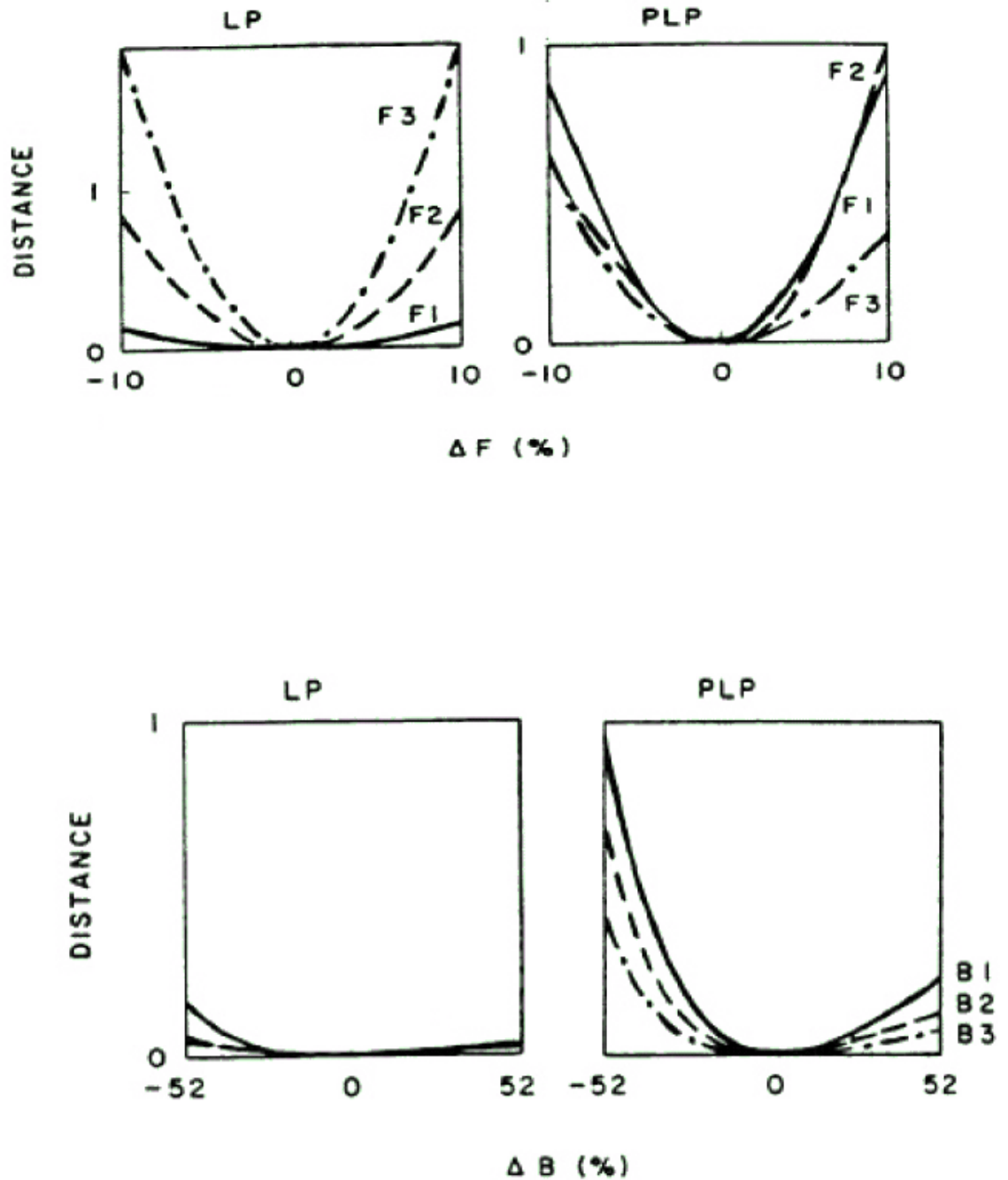
β)ε)PLP και η ανθρώπινη ακοή:

Η ανάλυση LP είναι σε αντίθεση με κάποια ήδη υπάρχουσα γνώση που έχουμε σε σχέση με την κατανόηση της φωνής.

Σύμφωνα με τα πειράματα του Flanagan ,μόλις ελάχιστη κατανοητή διαφορά στην αντίληψη των 3 συχνοτήτων formants είναι σταθερές σε μια σχετική συχνότητα. Παρα ταύτα η ανάλυση γραμμικής πρόβλεψης εναντιώνεται στα ευρύματα του Flanagan ,ενώ τα PLP δεν έχουν αυτή την ανικανότητα.

Από τα πειράματα, έχει βρεθεί ότι το μοντέλο γραμμικής πρόβλεψης δεν είναι το ίδιο ευαίσθητες σε όλα τα formants. Ωστόσο η PLP είναι ευαίσθητες ακόμη και σε μικρές μεταβολές στο εύρος των formants. Αυτό είναι ιδιαίτερα χρήσιμο ακόμη και για φωνητικές διαταράξεις. Αυτές εικονίζονται παρακάτω στα σχήματα με βάση τις μελέτες του Hermansky.

Η PLP μέθοδος συμπλέει με τα ευρύματα πολλών ερευνητών σχετικά με την ακοή. Σε σχέση με την ανάλυση γραμμικής πρόβλεψης δίνει καλύτερα αποτελέσματα σε διαδικασίες ανεξαρτήτως ομιλητή. και είναι μία εναλλακτική για άλλες περιοχές έρευνας.



3) Σχετική φασματική επεξεργασία φωνής(RASTA)

Επιδράσεις του περιβάλλοντος δυσχεραίνουν την απόδοση πολλών αναγνωριστών φωνής. Μία από τις μεθόδους να παρακάμψουμε τους παράγοντες του περιβάλλοντος, είναι τα σχετικά φάσματα που αναπτύχθηκε από τους Hermansky and Morgan. Αυτή είναι μία μέθοδος να φιλτράρουμε παραλλάξεις και τους θορύβους που προκαλείται από αναπάντεχους περιβαλλοντικούς παράγοντες έτσι ώστε να αυξηθεί η απόδοση της αναγνώρισης.

Οι ιδιότητες της φωνής μπορούν εύκολα να αλλάξουν χρησιμοποιώντας διάφορα μικρόφωνα μέσω της εγγραφής ή με χαρακτηριστικά συχνοτήτων από τα κανάλια

επικοινωνίας και τηλεφωνικά δίκτυα. Αυτές οι γραμμικές παραλλάξεις μπορούν να θεωρηθούν σαν συνελεκτικό θόρυβο επειδή εμφανίζονται σαν συνελεκτικά τμήματα των σημάτων. Οι αναγνωριστές, που χρησιμοποιούν βραχέως χρόνου λογαριθμικό φάσμα και γι' αυτό επηρεάζονται από αθροιστικό θόρυβο, διότι γίνονται αθροιστικές στο πεδίο του λογαρίθμου. Οι μεγαλύτερας κλιμακας ισχύ μπορεί να αφαιρεθεί από το cepstrum της φωνής για να αποφύγουμε αυτό το αποτέλεσμα το οποίο χρησιμοποιείται στην μέση φασματική αφαίρεση ή στις μεθόδους τυφλής αποσυνέλιξης (blind deconvolution methods). Η RASTA χρησιμοποιεί μεθόδους φιλτραρίσματος στο λογαριθμικό πεδίο που συνεκτιμούν τα χαρακτηριστικά συχνοτήτων των καναλιών επικοινωνίας συχνά αλλάζουν πολύ αργά στο χρόνο. Στην περίπτωση που το σήμα της φωνής παραλλάσσεται από αθροιστικό θόρυβο, παρουσιάζεται να είναι ένα αθροιστικό στοιχείο στο φάσμα ισχύος που λαμβάνεται από τμήματα που δεν έχουν φωνή και μπορούν να αφαιρεθούν από το φάσμα ισχύος για να αποφύγουμε αυτή την χειρότερηση.

Ο Hirsch χρησιμοποιεί ένα υπερυπερατό φίλτρο στο πεδίο του φάσματος ισχύος ειδικά για να συμπίεσει τον αθροιστικό θόρυβο.

Το φίλτρο RASTA μπορεί να εφαρμοστεί στην φωνή εκτός από το πλήρως λογαριθμικό πεδίο ενός φάσματος ισχύος αλλά εισαγάγει έναν παράγοντα που μετατρέπει το φάσμα ισχύος σε ένα γραμμικό λογαριθμικό πεδίο.

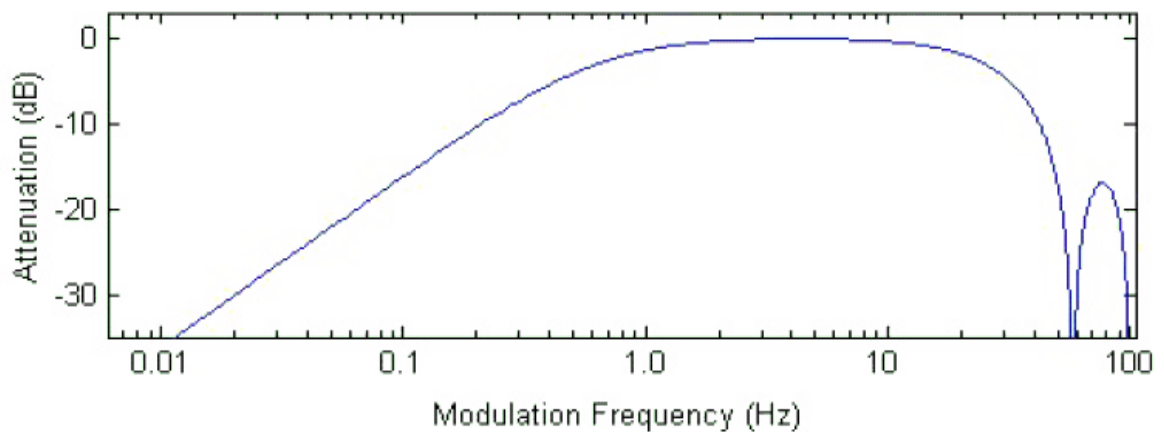
Η αντίληψη των ανθρώπων επηρεάζεται από σχετικές τιμές της φωνής και όχι με απόλυτες τιμές. Ο Zwicker, Green δείχνει ότι υπάρχει μία μεγαλύτερη ευαισθησία της ανθρώπινης ακοής στις συχνότητες διαμόρφωσης γύρω από τα 4Hz σε σχέση με μικρότερες ή υψηλότερες συχνότητες διαμόρφωσης. Η RASTA χρησιμοποιεί βαθυπερατά φίλτρα για να κάνει την ανάλυση της φωνής λιγότερο ευαίσθητη στους αργά μεταβαλλόμενους ή σταθερούς παράγοντες της φωνής. Αυτή η μέθοδος φιλτραρίσματος χρησιμοποιείται συνεπικουρικά με την μέθοδο PLP χρησιμοποιώντας αθροιστικά βήματα που εξηγούνται παρακάτω. Μετά τον υπολογισμό του κρίσιμου εύρους συχνοτήτων του φάσματος ισχύος, έπειτα το μετατρέπουμε συμπίεζοντας το με μία μη γραμμική συνάρτηση όπως ο λογάριθμος. Κάθε μετατρεπόμενο φασματικό πλάτος μετέπειτα φιλτράρεται με ένα φίλτρο RASTA. Αυτό έχει ως αποτέλεσμα να επεκταθεί χρησιμοποιώντας μη γραμμικούς μετασχηματισμούς όπως η αντίστροφη συνάρτηση που χρησιμοποιούμε όταν κάνουμε συμπίεση.

Η τεχνική PLP υπολογίζει το μοντέλο ολικών πόλων και έχει σαν αποτέλεσμα ένα φάσμα μετά από την εφαρμογή καμπύλης ίσης έντασης και τον λόγο του ισχυρού της ακοής (power law of hearing).

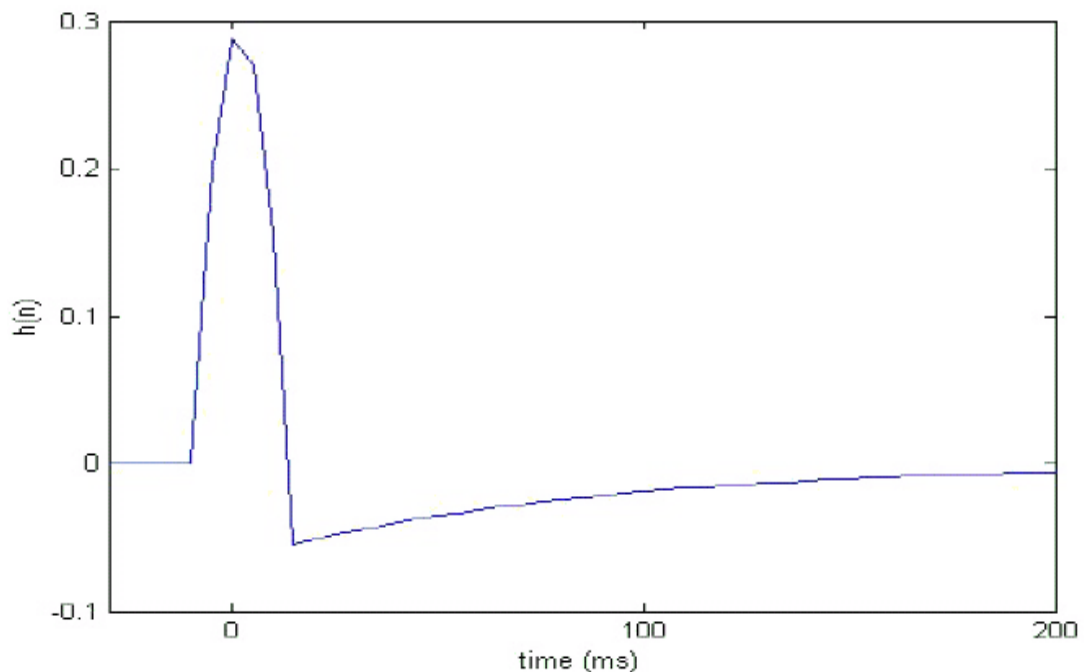
$$H(z) = 0.1 z^4 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}$$

Ο σκοπός αυτού του φίλτρου για να συμπίεσει σε σταθερούς παράγοντες και σε αργά μεταβαλλόμενη συνιστώσα με το να εισαχθεί ένα οξυ φασματικό μηδενικό στην μηδενική συχνότητα και υπακούει στα ευρήματα του Green σχετικά με την ευαισθησία της ακοής στις συχνότητες διαμόρφωσης.

Το βαθυπερατό φίλτρο ομαλοποιεί τις αλλαγές από πλαίσιο σε πλαίσιο σε βραχέως χρόνου φασματικές αναλύσεις. Το υπερυπερατό τμήμα υποτίθεται ότι αναφέρεται στον συνελεκτικό θόρυβο. Η χαμηλή συχνότητα αποκοπής είναι 0.26 Hz και η κλίση του φίλτρου πέφτει από 6dB/oct από τα 12.8 Hz με οξεία μηδενικά στα 28.9 και 50 Hz. Το φίλτρο έχει σταθερό χρόνο γύρω στα 160msec, διότι είναι αρκετά λογική για τις περισσότερες αναπαραστάσεις της φωνής. Τα πειράματα δείχνουν ότι η διέγερση του ήχου δεν χρειάζεται να είναι πάνω από 150msec για την καλύτερη διάκριση συχνοτήτων. Παρακάτω εικονίζεται οι γραφικές όσων προαναφέραμε:



Το RASTA μεσοπερατό φίλτρο μπορεί να μην είναι αποδοτικό να απομακρύνει



ασυσχέτιστο αθροιστικές συνιστώσες του σήματος επειδή γίνονται εξαρτώμενες από το σήμα μετά την λογαριθμική διαδικασία. Ένα σήμα θετικό σχετίζεται από μία θετική σταθερά, J παράγοντα, μπορεί να χρησιμοποιηθεί για να μετατρέψει το φάσμα ισχύος σε γραμμικό-λογαριθμικό πεδίο χρησιμοποιώντας μία εξίσωση:

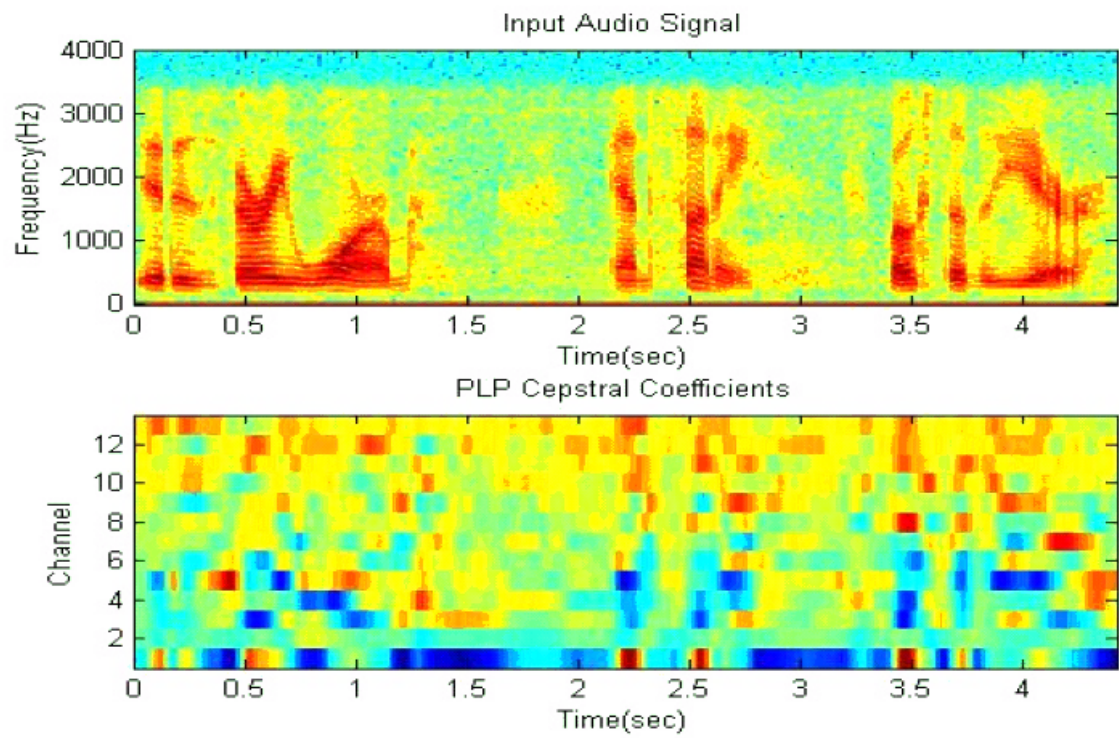
$$y = \ln(1 + Jx)$$

Αν το J είναι μεγάλο αρκετό τότε θα είναι τύπου λογαριθμικού αυτή ο μετασχηματισμός, αν πάλι είναι μικρό τότε γίνεται γραμμικό. Για να αποφύγουμε αρνητικές τιμές, χρησιμοποιούμε μία άλλη αντίστροφη:

$$x = \frac{e^{y/J}}{J}$$

Υπάρχει μία βέλτιστη τιμή για κάθε επίπεδο θορύβου. Για λόγους ευκολίας μία σταθερή τιμή χρησιμοποιείται για διαφορετικούς σηματοθορυβικούς λόγους στους

αναγνωριστές φωνής ή η εκπαίδευση μπορεί να γίνει για μία εκτιμώμενη τιμή για διάφορα επίπεδα θορύβου.

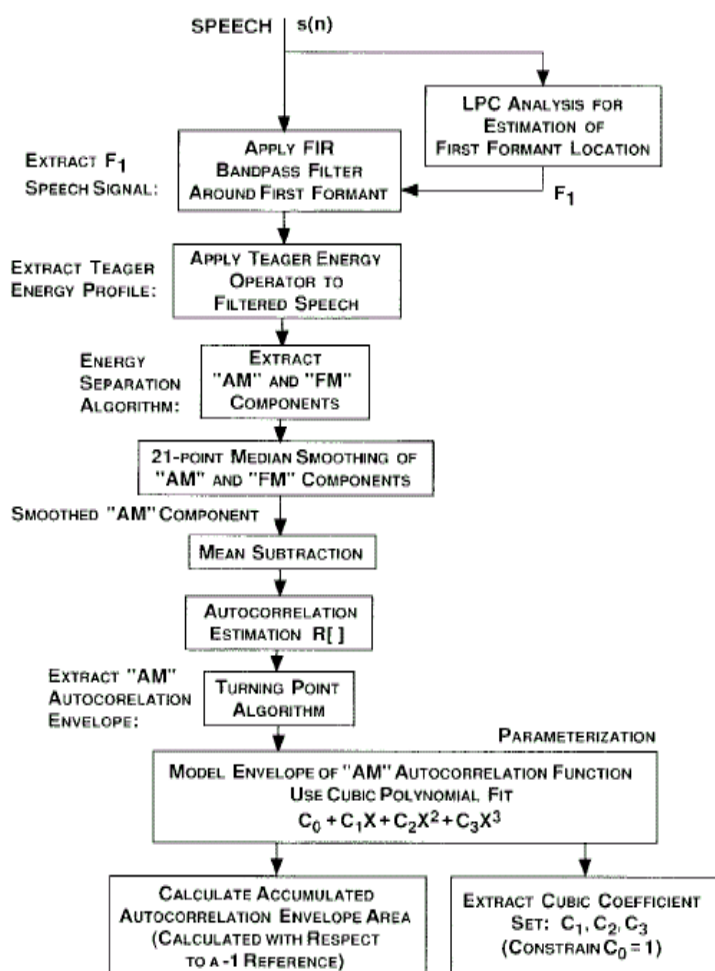


3)AM-FM Models

Το γραμμικό μοντέλο της φωνής βασίζεται στην υπόθεση ότι οι συντονισμοί(δηλαδή τα κέντρα των συχνοτήτων των formants) παραμένουν σταθεροί για σχετικά μικρά χρονικά διαστήματα.Ένα μηγραμμικό μοντέλο προτείνει ότι αυτοί οι συντονισμοί δεν είναι σταθεροί και διακυμαίνονται γύρω από την κεντρική συχνότητα και μπορεί να μοντελοποιηθεί σαν άθροισμα σημάτων AM-FM.Τα χαρακτηριστικά της διαμόρφωσης μικρού χρόνου επιχειρούν να ποσοτικοποιήσουν αυτές τις μεταβολές και να συλλάβουν τη χρονική και δυναμική φύση των φωνητικών συντονισμών.Η βελτίωση των ρυθμών αναγνώρισης υποστηρίζουν την ακρίβεια ενός τέτοιου μηγραμμικού μοντέλου.Αυτά τα χαρακτηριστικά χρησιμοποιούνται για να ενισχύσουν τα κλασσικά φασματικά χαρακτηριστικά σαν ένα επαυξημένο σύνολο χαρακτηριστικών για εφαρμογές ASR και δείχνουν ευρωστία σε θορυβώδη φωνή λόγω της χρήσης μιας συστοιχίας φίλτρων και τον αλγόριθμο διαχωρισμού ενέργειας(Energy Separation Algorithm (ESA)).

Εναλλακτικά,άλλοι αλγόριθμοι έχουν προταθεί για να λάβουν το στιγμιαίο πλάτος και συχνότητα των σημάτων για μεσοπερατά σήματα φωνής.Τέτοιες τεχνικές χρησιμοποιούν το φίλτρο Κάλμαν και τον μετασχηματισμό Χίλμπερτ.Η πρώτη τεχνική έχει υψηλή υπολογιστική πολυπλοκότητα και δεν μπορεί να χρησιμοποιηθεί για εφαρμογές πραγματικού χρόνου.Η τελευταία έχει πολύ χαμηλή ανάλυση στο πεδίο του χρόνου και οι οξείες μεταβολές ομαλοποιούνται.

Τελικά,μερικά πειράματα που έχουν γίνει μεσω της συγχώνευσης του μη γραμμικού μοντέλου για τους συντονισμούς και την εκτίμηση των χαρακτηριστικών τύπου MFCC.Πιο συγκεκριμένα το τετράγωνο του πλάτους ενός σήματος φωνής έχει εναλλαχθεί με την χρήση του μη γραμμικού τελεστή Teager στον κλασσικό αλγόριθμο υπολογισμού των χαρακτηριστικών συνόλου MFCC.Η σωστός ρυθμός αναγνώρισης φωνημάτων έχει δείξει τις περιθώριες διαφορές μεταξύ καθαρής φωνής συγκρινόμενη με τους ρυθμούς που λαμβάνουμε χρησιμοποιώντας MFCC.Αυτό



οφείλεται στην επεξεργασία ομαλοποίησης του MFCC(λόγω της αποκοπής των φασματικών παραμέτρων υψηλότερης τάξης).Παρα ταύτα τα χαρακτηριστικά που βασίζονται στον τελεστή ενέργειας δείχνουν να είναι πιο ομαλά και πιο εύρωστα για θορυβώδη σήματα και δίνουν καλύτερα χαρακτηριστικά.

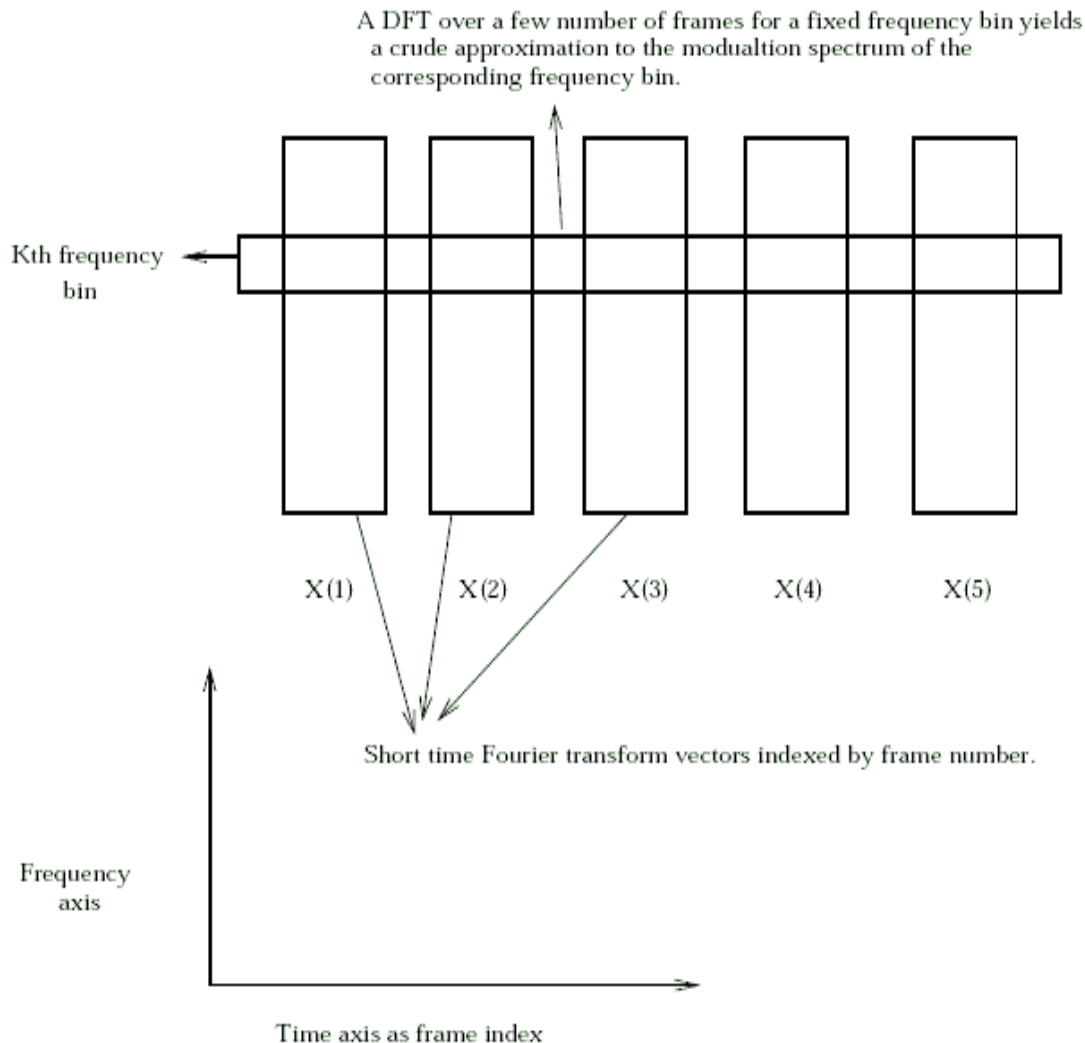
Άλλα σύνολα μη γραμμικών χαρακτηριστικών έχουν προταθεί όπου λαμβάνουν το διαμορφωμένου πλάτους μέρος του μη γραμμικού μοντέλου φωνής.Ο αλγόριθμος που περιγράφηκε πιο πάνω(η συστοιχία φίλτρων και ο αλγόριθμος αποδιαμόρφωσης) έχουν χρησιμοποιηθεί για την εκτίμηση του στιγμιαίου πλάτους(απόλυτο φάκελο) των μεσοπερατών σημάτων φωνής.Αυτοί οι φάκελοι διαμορφώνονται από βαθυπερατά σήματα και περιέχουν γλωσσολογική πληροφορία.Το προτεινόμενο σύνολο χαρακτηριστικών παραμετροποιεί τα διαμορφωμένα σήματα και λαμβάνουν τα στατιστικά(την πρώτη και δεύτερη τάξη των φασματικών ροπών).Αυτό το σύνολο χαρακτηριστικών δείχνει μία στατιστική σημαντική αύξηση συγκρινόμενη με των ακριβειών του συστήματος αναφοράς με βάση τα MFCC.Πρόσφατα πειράματα δείχνουν ότι τα χαρακτηριστικά έχουν μεγαλύτερη ευρωστία στο θόρυβο,κυρίως λόγω της βαθυπερατής φύσης τους.Ονομαστικά αυτά τα πλάτη είναι βαθυπερατά σήματα και κατ'επέκταση πιο εύρωστα στο θόρυβο.Οι εκτιμήσεις τους φαίνονται να είναι πολύ ομαλές,σε σχέση με κορυφές και ασυνέχειες,ακόμα και σε μικρούς σηματοθορυβικούς λόγους.Έχειδειχθεί ότι περιέχουν σημαντική ποσότητα πληροφορίας σε σχέση με τον ομιλητή και το γλωσσικό περιεχόμενο των σημάτων φωνής.

Τα στιγμιαίου πλάτους σήματα και τα αντίστοιχα σήματα διαμόρφωσης,έχουν πολυ μικρή χρονική εξέλιξη.Αυτή η ιδιότητα εκμεταλεύεται από μία άλλη όψη της έρευνας για διαμόρφωση μεγάλης κλίμακας,π.χ το φασματόμετρο διαμόρφωσης.Οι διαμορφώσεις βραχέως και μακροπρόθεσμου χρόνου έχουν μελετηθεί σε παράθυρα χρόνου για 10-30ms έτσι ώστε να λαμβάνουν και τις μικρολεπτομέρειες(πολύ γρήγορες μεταβολές) των σημάτων φωνής.Στον αντίποδα,οι μακροπρόθεσμες διαμορφώσεις εξετάζουν την ενέργεια της εξέλιξης της φωνής και των αντιστοιχων παραθύρων στο εύρος των 200-500ms.

$$f(n) \approx \frac{1}{2\pi T} \arccos \left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]} \right),$$

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

$$\begin{aligned} \Psi_e[x(t)] &= \left(\frac{d}{dt} x(t) \right)^2 - x(t) \left(\frac{d^2}{dt^2} x(t) \right) \\ &= [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \end{aligned}$$



4)Χρονικά δυναμικά χαρακτηριστικά της φωνής:

Παρόλα αυτά για ένα τμήμα βραχέως χρόνου 20ms των παραμέτρων της φωνής και υποτίθενται ότι είναι ανεξάρτητα ως προς το χρόνο, όμως σε ένα μεγαλύτερο τμήμα η φωνή μεταβάλλεται ως προς τον χρόνο. Επιπλέον οι μεταβολές της φωνής στον χρόνο μας δίδουν χρήσιμη γλωσσολογική, προσοδική και χαρακτηριστική του ομιλητή. Έτσι παράμετροι της φωνής, όπως συντελεστές του γραμμικού προβλεπτικού μοντέλου, το κέρδος της διέγερσης, pitch, και μεταβλητές που αλλάζουν ως προς τον χρόνο. Χαρακτηριστικά χρονικών διαφορών είναι απλοί και αποτελεσματικοί τρόποι περιγραφής των παραμέτρων φωνής ως προς τον χρόνο. Η μοντελοποίηση των χαρακτηριστικών διαφοράς είναι απαραίτητη για να βελτιωθεί η ακρίβεια της αναγνώρισης φωνής ή ακόμα και την αντιληπτική ποιότητα για την σύνθεση της φωνής. Για την αναγνώριση φωνής σχετικά απλά χαρακτηριστικά είναι αυτά που ορίζονται ως εξής:

$$\partial c(m) = c(m+1) - c(m-1)$$

$$\partial \partial c(m) = \partial c(m+1) - \partial c(m-1)$$

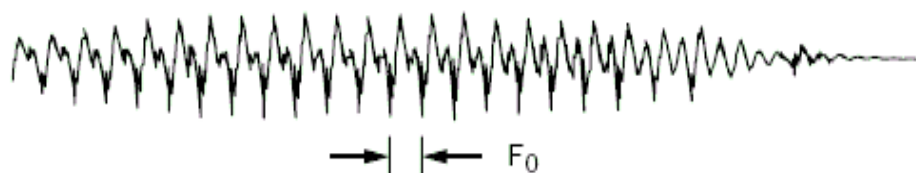
όπου $\partial c(m)$ είναι η πρώτης τάξης χρονική διαφορά των φασματικών χαρακτηριστικών, στην επεξεργασία φωνής αντιστοιχίζεται σαν χαρακτηριστικά μιας ταχύτητας. Οι δευτέρας τάξης χρονικές διαφορές των φασματικών χαρακτηριστικών $\partial \partial c(m)$ και σχετίζεται στην επιτάχυνση. Παράλληλα μπορούμε να πάρουμε την

ελάχιστη μέση τετραγωνική κλίση των 3 φασματικών δειγμάτων $c(m-1)$, $c(m)$ και $c(m+1)$. Τα δυναμικά χαρακτηριστικά είναι αποτελεσματικά για την βελτίωση της αναγνώρισης φωνής. Η χρήση των διαφορικών χαρακτηριστικών είναι χρήσιμα για την βελτίωση της ποιότητας της κωδικοποίησης της φωνής, σύνθεσης και εφαρμογών ενίσχυσης.

5)Εξαγωγή χαρακτηριστικών στο πεδίο του χρόνου(time domain features):

Αντίστοιχα με το πεδίο συχνοτήτων έχουμε και το πεδίο του χρόνου, όπου ουσιαστικά εργαζόμαστε πάνω στην δοθείσα κυματομορφή. Αξίζει να σημειωθεί ότι κατά πολλούς ερευνητές η χρήση του πεδίου χρόνου είναι κυρίως επικουρική, διότι αν και δίνει αρκετά χρήσιμα χαρακτηριστικά, εντούτοις είναι επιρρεπής ως προς τον θόρυβο και συγκριτικά είναι λιγότερο εύρωστο από το πεδίο της συχνότητας.

Οι τεχνικές του χρόνου αντιμετωπίζουν ευθέως την κυματομορφή του ήχου



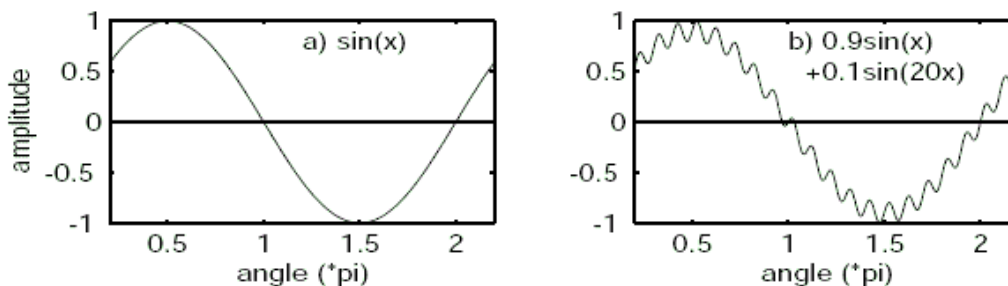
Ρυθμός Zero crossing:

Ένα από τα πρώτα πράγματα που οι ερευνητές χρησιμοποίησαν το ZCR ήταν για την εκτίμηση f_0 . Η σκέψη ήταν ότι ZCR μπορεί να εφαρμοστεί ευθέως και να σχετίζεται με τον αριθμό των φορών που επαναλαμβάνεται η κυματομορφή ανά μονάδα χρόνου. Όμως φάνηκε ξεκάθαρο ότι υπήρχε πρόβλημα με αυτή την μέτρηση f_0 . Εάν το φάσμα ισχύος της κυματομορφής περιστρέφεται γύρω από το f_0 , τότε θα διασχίσει την γραμμή του μηδενός 2 φορές ανά κύκλο. Όμως εάν το σήμα περιέχει υψηλότερες συχνότητες (όπως εικονίζεται στο σχήμα), τότε μπορεί να διασχίσει την γραμμή περισσότερες από 2 φορές.

Ένας ανιχνευτής ZCR f_0 μπορεί να χρησιμοποιηθεί με αρχικό φιλτράρισμα για να αφαιρέσουμε τις υψηλότερες συχνότητες που “μολύνουν” την μέτρηση και αφαιρούν όσο υψηλότερης συχνότητας στοιχεία. Άλλη μία πιθανότητα είναι να χρησιμοποιηθεί ο ZCR f_0 ανιχνευτής για να διαγνώσει πρότυπα στο zero-crossings και να υποθέσουν μία τιμή για το f_0 βασισμένο σε αυτά τα πρότυπα. Πρέπει να σημειωθεί ότι ο ZCR είναι ένα χαρακτηριστικό πληροφορίας για αυτό και άσχετο με το πόσο καλά βρίσκει το f_0 . Πολλοί ερευνητές έχουν ελέγξει στατιστικά τις ιδιότητες του ZCR. Επί παραδείγματι το ZCR χρησιμοποιείται για να συσχετίσει το φασματικό κέντρο ή σημείο ισορροπίας της κυματομορφής, εκτός αν το φάσμα είναι διτροπο (bimodal), και είναι συχνά η περιοχή με την υψηλότερη συχνότητα στην κυματομορφή. Εάν το φασματικό κέντρο είναι σχετικά υψηλή συχνότητα, μπορεί να σημαίνει ότι το σήμα είναι σύμφωνο ή ένα ανθρώπινο φώνημα άηχο. Ορίζεται ως ακολούθως:

$$F_0 = \frac{ZF_s}{2} \quad Z = \sum_{n=0}^{N-1} |\text{sgn}[x(n)] - \text{sgn}[x(n-1)]|$$

Με βάση τον ορισμό αποτελεί μία αδρή μετρική για την θεμελιώδη συχνότητα του σήματος.



Μία πλήρως στατιστική χρήση του ZCR έχει βρεθεί. Ο John Saunders συνέλεξε δεδομένα σχετικά με το πώς αλλάζει ο ZCR ανά χρόνο και καλείται ZCR ισόπεδη (contour). Έχει βρεθεί ότι ZCR contour της φωνής είναι σημαντικά διαφορετική από την μουσική και έχει χρησιμοποιηθεί για να διαφοροποιήσουν και τα 2. Μια όμοια χρήση του ZCR είναι η βραχέως χρόνου μέση ZCR. Αυτές οι μετρικές δείχνουν μια υπεροχή σε σχέση με αντιληπτικά χαρακτηριστικά τα οποία μπορεί να διαθετούν τα μουσικά σήματα και να μην είναι ευθέως αντιλαμβανόμενα από το ανθρώπινο αυτί.

Μία από τις πιο ελκυστικές του ZCR και των σχετικών χαρακτηριστικά είναι ότι αυτά τα χαρακτηριστικά μπορούν να υπολογιστούν αρκετά γρήγορα. Ο ZCR είναι ένα χαρακτηριστικό του πεδίου του χρόνου το οποίο μπορεί να υπολογιστεί και σε πραγματικό χρόνο (on the fly). Μάλιστα μπορεί να χρησιμοποιηθεί και σε αναλογικά σήματα (απλά ο έλεγχος μόνο της πολικότητας) και να γίνεται μέτρηση κάθε φορά που αλλάζει η πολικότητα.

RMS Ισχύς

Όπως είναι ένα από τα πιο φυσικά χαρακτηριστικά, ισχύς είναι ένα μετρο της κυματομορφής οποιαδήποτε στιγμή. Η ισχύς ορίζεται σαν η ενέργεια ανά μοναδιαίο χρόνο και για την περιοχή των ακουστικών σημάτων, η ισχύς σχετίζεται με το πλάτος της κυματομορφής. Όσο πιο δυνατή ένταση έχει το σήμα, τόσο περισσότερη ισχύ έχει. Οι μετρικές της ισχύς χρησιμοποιούνται για να βρουν την παύση στη κυματομορφή, όπως επίσης και το δυναμικό εύρος. Η ισχύς της κυματομορφής χρησιμοποιείται τυπικά για τον υπολογισμό σε βάση βραχέως χρόνου, αυτό γίνεται με την παραθύρωση όπως στο FFT, τετραγωνίζοντας τους όρους και λαμβάνοντας τον μέσο όρο αυτών. Η τετραγωνική ρίζα είναι μία ποσότητα στην μηχανική γνωστή σαν μέση τετραγωνική τιμή και η οποία χρησιμοποιείται ευρύτατα. Η μέση κανονικοποιημένη ισχύς, για ένα ψηφιακό σήμα ορίζεται σαν κανονικοποιημένη ενέργεια ανά δείγμα. Παρακάτω δίνουμε τους τύπους υπολογισμού:

$$\begin{aligned}
W_{rms} &= \sqrt{|w^2(t)|} \\
p(t) &= w^2(t) \\
P &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} w^2(t) dt \\
E &= \lim_{T \rightarrow \infty} \int_{-T/2}^{T/2} w^2(t) dt
\end{aligned}$$

Τα χαρακτηριστικά σχετίζονται με την ισχύ που μεταβάλλεται στο χρόνο. Ισχύς σε συγκεκριμένες μπάστες, και συγκεκριμένα η διακύμανση της ισχύς σε ορισμένα τμήματα χρησιμοποιείται για την εύρεση της σιγής. Το επιχείρημα τους είναι η εφαρμογή σαν ανυπαρκτό κατώφλι ισχύος για να γίνεται διάγνωση ισχύος, αλλά αστοχεί στην περίπτωση όπου έχουμε πλαίσια με σιγή είτε πλαίσια με μικρή ισχύ, όπως είναι στην αρχή ή στο τέλος του σήματος.

Μάλιστα η κατανομή της ισχύος χρησιμοποιείται και για να ξεχωριστεί η φωνή από την μουσική. Η φωνή τείνει να συνίσταται από περιόδους υψηλής ενέργειας (έμφωνα φωνήματα), ενώ η μουσική τείνει να έχει πιο συνεπή κατανομή ισχύος. Επίσης υπάρχει και η μετρική της διαμόρφωσης της ισχύς όπου πολλοί ερευνητές αναφέρονται ότι η φωνή έχει ρυθμό διαμόρφωσης στην ισχύ (power modulation rate) γύρω από τα 4Hz. Η συνέλιξη μεταξύ 2 κυματομορφών είναι ένα μέτρο της ομοιότητας τους. Οι κυματομορφές συγκρίνονται σε διαφορετικά χρονικά διαστήματα και η ομοιότητα τους υπολογίζεται σε κάποιο χρονικό διάστημα. Το αποτέλεσμα της συσχέτισης είναι μία μέτρηση της ομοιότητας σαν μία συνάρτηση της χρονικής διαφοράς μεταξύ των ενάρξεων αυτών των 2 κυματομορφών. Η συνάρτηση αυτοσυσχέτισης είναι μία συσχέτιση μιας κυματομορφής με την ίδια την κυματομορφή. Θα ήταν αποδεκτό ότι η ομοιότητα με χρονική καθυστέρηση μηδέν, με αυξανόμενη την μη-ομοιότητα καθώς η χρονική καθυστέρηση αυξάνεται. Ο μαθηματικός ορισμός δίδεται από την εξίσωση: (Η πρώτη εξίσωση έχουν άπειρο μέγεθος και στην δεύτερη πεπερασμένο μέγεθος μεγέθους N)

$$\begin{aligned}
R_x(\nu) &= \sum_{n=-\infty}^{\infty} x[n]x[n+\nu] \\
R_{x'}(\nu) &= \sum_{n=0}^{N-1-\nu} x'[n]x'[n+\nu]
\end{aligned}$$

Η συσχέτιση μεταξύ 2 συναρτήσεων μεταξύ $x[n]$ και $y[n]$ και υπολογίζεται ως εξής:

$$R_{xy}(\nu) = \sum_{n=-\infty}^{\infty} x[n]y[n+\nu].$$

Η τιμή του $R_{xy}(0)$ δίνει μία μέτρηση της ομοιότητας μεταξύ 2 διαφορετικών συναρτήσεων. Αυτή είναι μια μετρική που χρησιμοποιείται αρκετά.

Οι περιοδικές κυματομορφές έχουν μία ενδιαφέρουσα χαρακτηριστική αυτοσυσχέτισης: η συνάρτηση αυτοσυσχέτισης είναι περιοδική. Καθώς η χρονική καθυστέρηση μεγαλώνει μέχρι το μέσο της περιόδου μιας κυματομορφής, η συσχέτιση μειώνεται μέχρι το ελάχιστο. Αυτό είναι επειδή η κυματομορφή είναι σε φάση με ένα αντίγραφο με χρονική καθυστέρηση. Η πρώτη κορυφή δείχνει την περίοδο της κυματομορφής.

Προβλήματα με αυτή τη μέθοδο προκύπτουν όταν μία αρμονική πολύπλοκη, ψευδο-περιοδική κυματομορφή ληφθεί. Η πρώτη κορυφή δεν θα είναι η περίοδος της κυματομορφής, αλλά η αρμονική που θα έχει προστεθεί. Η πρώτη μεγάλη κορυφή πράγματι θα δείξει την θεμελιώδη περίοδο της κυματομορφής, αλλά μειώνει την ευρωστία και αυξάνει την υπολογιστική πολυπλοκότητα του αλγορίθμου για να διακρίνει “μεγάλες” και “μικρές” κορυφές.

$$\Psi(j) = \frac{\sum_{n=0}^{N-1} x(n) x(n-j)}{\sqrt{\left(\sum_{n=0}^{N-1} x(n)^2\right) \left(\sum_{n=0}^{N-1} x(n-j)^2\right)}}$$

Θεμελιώδης συχνότητα: (Fundamental Frequency (f_0))

Μόνο τα περιοδικά ή ψευδο-περιοδικά σήματα έχουν αληθή f_0 . Αντιληπτικά, τα περιοδικά ή ψευδο-περιοδικά σήματα έχουν pitch. Τα περιοδικά σήματα έχουν την ακόλουθη περιγραφή ($w(t+\tau) = w(t)$) with a period of τ and $f_0 = 1/\tau$ για την μεγαλύτερη τιμή του τ . Τα

ψευδο-περιοδικά σήματα σχεδόν επαναλαμβάνονται σαν ($w(t+\tau) = w(t) + \varepsilon$).

Υπάρχει μια μικρή διαφορά στην κυματομορφή από περίοδο σε περίοδο, αλλά πάλι μπορούμε να ισχυριστούμε ότι $f_0 = 1/\tau$, που αντιστοιχεί στην μεγαλύτερη περίοδο στην οποία η κυματομορφή επαναλαμβάνεται με μια κάποια ανοχή.

Είναι προφανές ότι η εξαγωγή του f_0 από ένα σήμα έχει νόημα μόνο αν η κυματομορφή είναι περιοδική. Οι f_0 ανιχνευτές συνήθως εξυπηρετούν ένα διττό σκοπό σε αυτή την περίπτωση — αν το διεγνωσμένο f_0 έχει νόημα και για το υπόλοιπο σήμα, τότε το σήμα θεωρείται περιοδικό. Αν το f_0

εμφανίζεται τυχαία ή εάν ο ανιχνευτής παρέχει ένα αδύνατο ή αναληθές αποτέλεσμα, το σήμα είναι απεριοδικό.

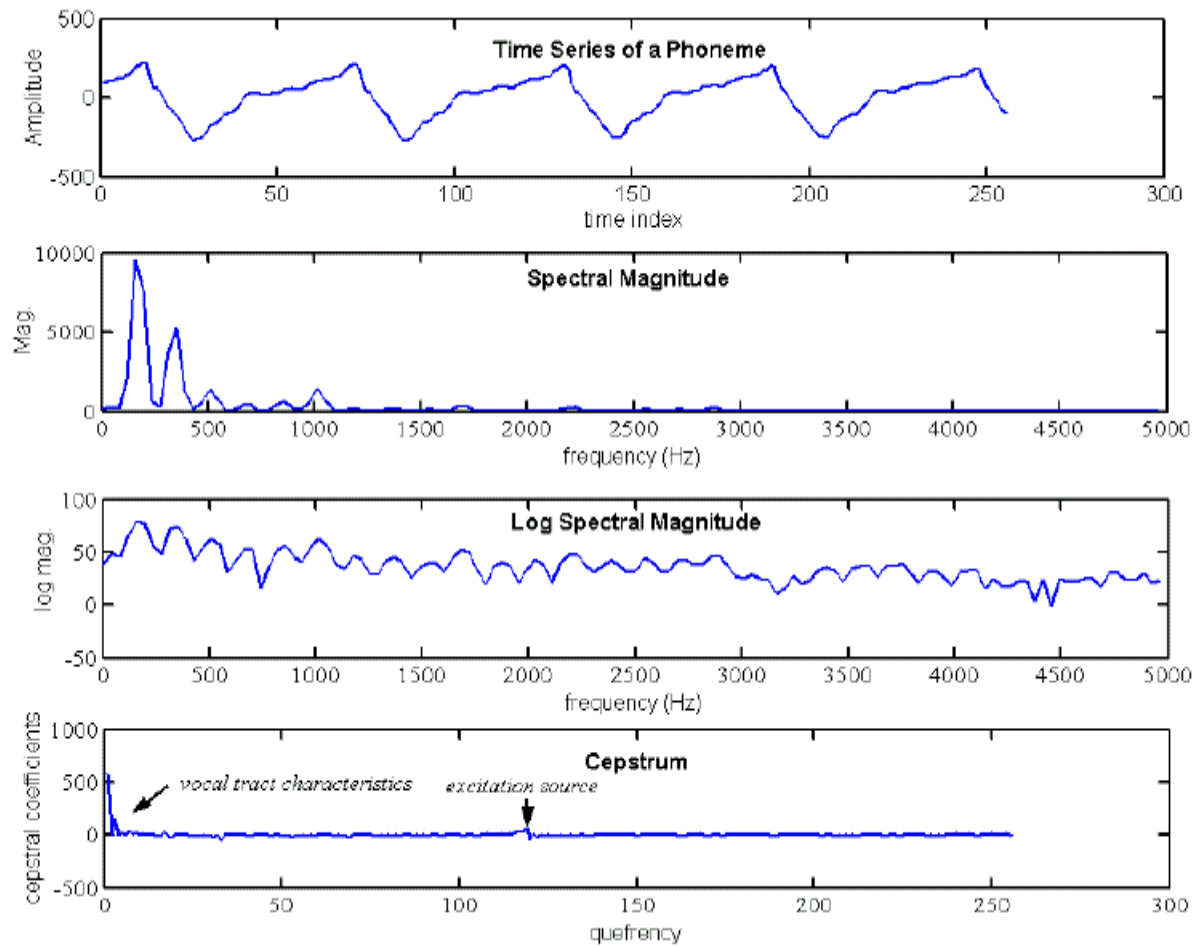
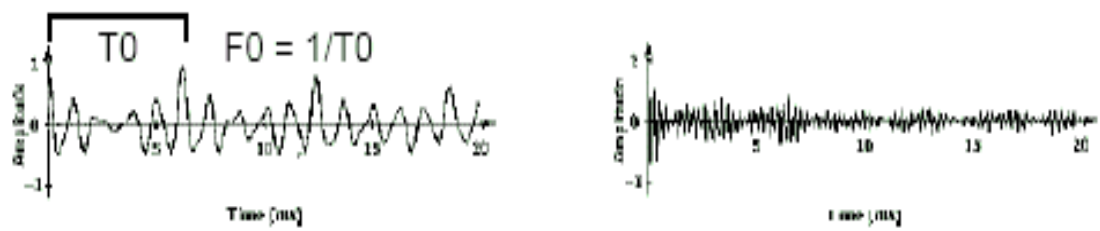
Τα όρια των λέξεων συχνά ευρισκονται με την χρήση του f_0 . Η ιδέα είναι ότι μεγάλη μεταβολή στο f_0 είναι απίθανες να συμβούν στο μέσο μίας λέξης. Αυτές είναι πιο πιθανές να συμβούν στο τέλος των λέξεων. Μάλιστα έχει δοκιμαστεί με επιτυχία σε διάφορες γλώσσες της Ινδίας (Hindi, Bengali, Marathi and Telugu) όπως και στα γερμανικά.

Σε πολλά από αυτά τα συστήματα, δεν υπάρχει διαφορά μεταξύ του pitch & f_0 , και η διαφορά είναι καλά κατανοημένη και εύκολα μοντελοποιήσιμη, είναι σημαντικό να θυμόμαστε ότι πολλά από αυτά τα συστήματα δεν περιέχουν αντιληπτικά μοντέλα για την ανίχνευση του pitch.

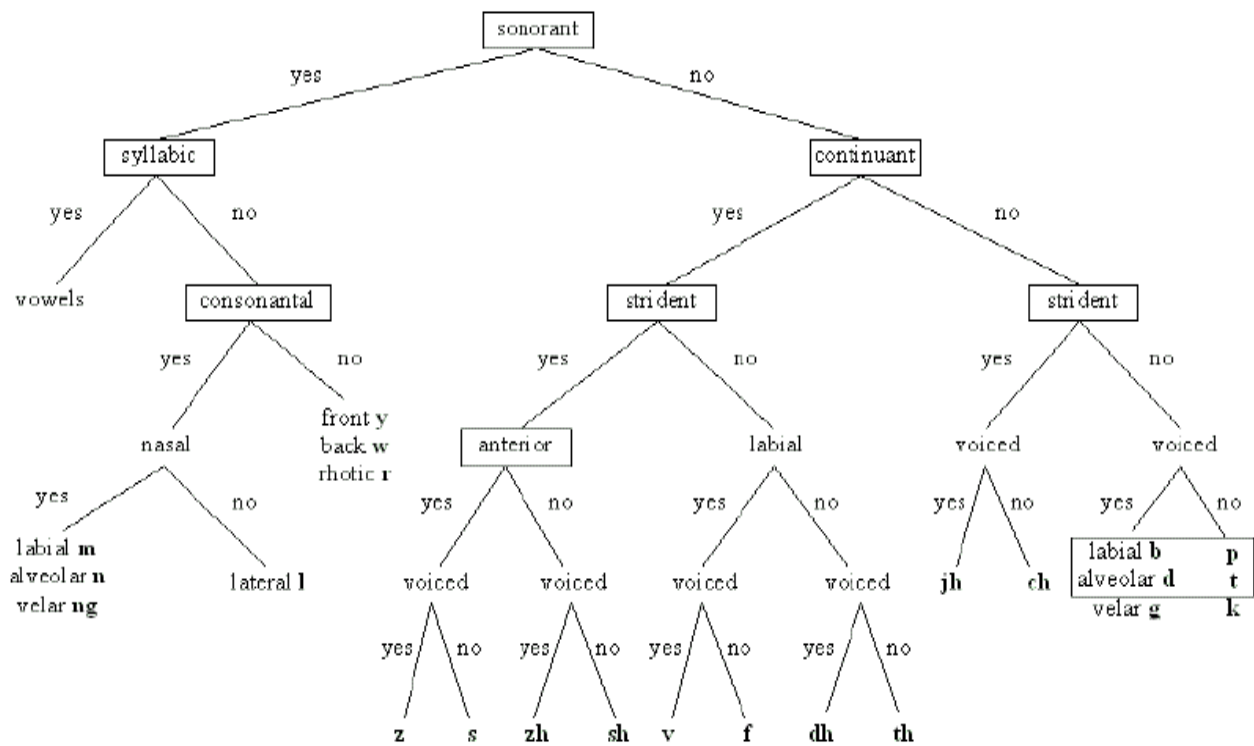
6) Έμφωνα/Άφωνα πλαίσια

Ένα από τα πρώτα θεμελιώδη βήματα σε κάθε σύστημα αναγνώρισης φωνής είναι η ταξινόμηση των πλαισίων σε έμφωνα/άφωνα. Στην πρώτη επισκόπηση, φαίνεται σαν ένα απλό φυσικό μέτρο διότι: τα έμφωνα πλαίσια τείνουν να είναι αρμονικά, και έχουν λιγότερο φασματικό κέντρο απ' ό,τι τα άφωνα που τείνουν να είναι μη αρμονικά.

Η ηχητικότητα ενός ηχητικού τμήματος μπορεί να χρησιμοποιηθεί σαν μέτρο ταξινόμησης και συχνά γίνεται με την χρήση ενός f_0 εκτιμητή. Αυτή η υπόθεση βασίζεται στο ότι η είσοδος είναι σημά φωνής και έτσι υπάρχει ισχύς σε ένα πλαίσιο αλλά αν όχι κατανοητό pitch στο κανονικό εύρος φωνής, το πλαίσιο μπορεί αξιόπιστα να ταξινομηθεί σαν άνηχο. Το σύστημα ταξινομείται στην βάση των pitches σε ένα συγκεκριμένο εύρος, αλλά στο πεδίο της αναγνώρισης της φωνής σαν ταξινόμηση έμφωνου/άφωνου.



Πίνακας άφωνων/έμφωνων πλαισίων



Κεφάλαιο 6

Η φύση του πολυδιάστατου χώρου

1)Εισαγωγή:

Η πολυπλοκότητα της διάστασης είναι γνωστή εδώ και 3 και παραπάνω δεκαετίες και έχει επιπτώσεις από το ένα πεδίο μέχρι και το άλλο. Στην συνδυαστική βελτιστοποίηση (combinatorial optimization) σε πολλές διαστάσεις, έχει ειπωθεί σαν μία εκθετική αύξηση του υπολογιστικού φόρτου σε σχέση με τον βαθμό των διαστάσεων. Στην στατιστική, εκφράζει ότι είναι ένα πρόβλημα με παραμέτρους ή εκτίμηση πυκνότητας λόγω της ανεπάρκειας των δεδομένων. Η αρνητική επίδραση αυτής της ανεπάρκειας προκύπτει από κάποιες γεωμετρικές, στατιστικές και ασυμπτωτικές ιδιότητες του πολυδιάστατου χώρου χαρακτηριστικών. Αυτά τα χαρακτηριστικά παρουσιάζουν μία εντυπωσιακή συμπεριφορά των δεδομένων στις υψηλότερες διαστάσεις.

Υπάρχουν πολλές υποθέσεις που κάνουμε σχετικά με τα χαρακτηριστικά των χώρων χαμηλής διάστασης κυρίως βασίζοντας την εμπειρία μας από τον 3διαστατο ευκλείδιο χώρο. Υπάρχει ένα αντιληπτικό εμπόδιο που το κάνει δύσκολο να έχουμε την κατάλληλη διαίσθηση για τον χώρο υψηλών διαστάσεων και τις επιπτώσεις που έχει στην συμπεριφορά των πολυδιάστατων δεδομένων. Οι περισσότερες από τις υποθέσεις μας που είναι απαραίτητες για στατιστικές ιδιότητες βασίζονται στην διαίσθηση μας για τον 3διαστατο χώρο, όπως για παράδειγμα, όπως η συγκέντρωση του όγκου σε τέτοια σχήματα όπως κύβοι, σφαίρες και ελλειψοειδή ή που τα δεδομένα και η συγκέντρωση τους είναι σχετική και γνωστή από μία οικογένεια συναρτήσεων πυκνότητας όπως οι κανονικές και οι ομοιόμορφες. Άλλες σημαντικές αντιλήψεις που είναι σχετικές με την στατιστική ανάλυση, για παράδειγμα, πως οι διαγώνιοι σχετίζονται με τις συντεταγμένες, ο αριθμός που δειγμάτων που απαιτείται για επιβλεπόμενη μάθηση, η υπόθεση της κανονικότητας των δεδομένων και η σημαντικότητα του μέσου και της διακύμανσης στην διαδικασία διαχωροποίησης μεταξύ διαφόρων στατιστικών τάξεων. Στη συνέχεια θα δώσουμε κάποια χαρακτηριστικά του πολυδιάστατου χώρου και το πως επηρεάζει την επιβλεπόμενη ταξινόμηση της ανάλυσης των δεδομένων και θα συζητηθεί. Οι περισσότερες από αυτές τις ιδιότητες δεν ταιριάζουν με τις ιδιότητες του 3διαστατου χώρου που προαναφέραμε.

2)Γεωμετρικές, Στατιστικές και ασυμπτωτικές ιδιότητες.

Σε αυτό το τμήμα θα απεικονίσουμε κάποιες ασυνήθιστες και αναπάντεχες χαρακτηριστικές ιδιότητες του υπερχώρου και θα παρατάξουμε κάποιες αποδείξεις και αποτελέσματα (επεξήγηση). Αυτά αποδεικνύουν ότι ο πολυδιάστατος χώρος είναι αρκετά διαφορετικός από τον 3διάστατο χώρο με τον οποίο είμαστε εξοικειωμένοι.

Καθώς η διάσταση αυξάνει:

Α) Ο όγκος του υπερκύβου συγκεντρώνεται στις γωνίες.

Έχει δειχθεί ότι ο όγκος μιας υπερσφαίρας ακτίνας ρ και διάστασης d δίδεται από την εξίσωση:

$$V_s(r) = \text{volume of a hypersphere} = \frac{2r^d}{d} \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$$

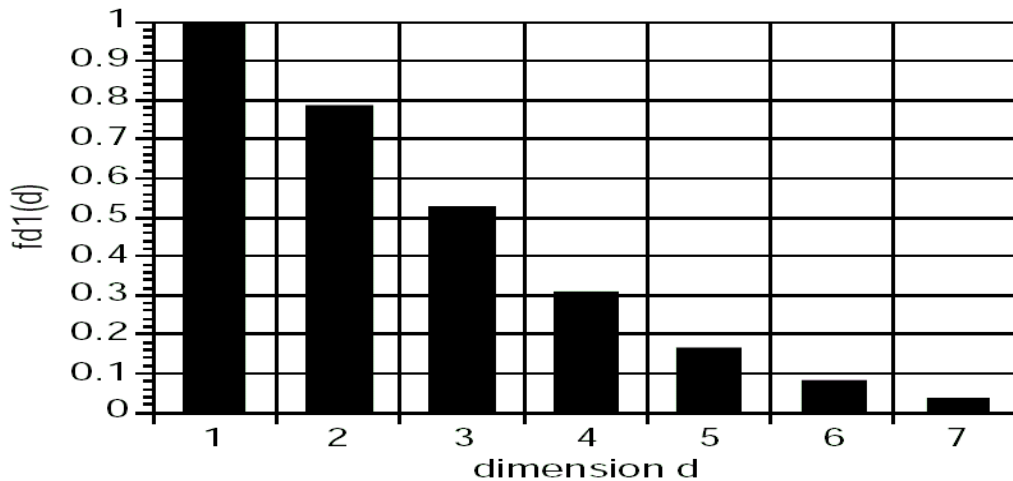
όπου ο όγκος του υπερκύβου δίδεται από την αντίστοιχη εξίσωση:

$$V_c(r) = \text{volume of a hypercube} = (2r)^d$$

Αν λάβουμε τον λόγο του όγκου της υπερσφαίρας που περιέχεται στον υπερκύβο δίδεται από :

$$f_{d1} = \frac{V_s(r)}{V_c(r)} = \frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)}$$

όπου d η διάσταση του χώρου. Κάνοντας την παρακάτω γραφική έχουμε:



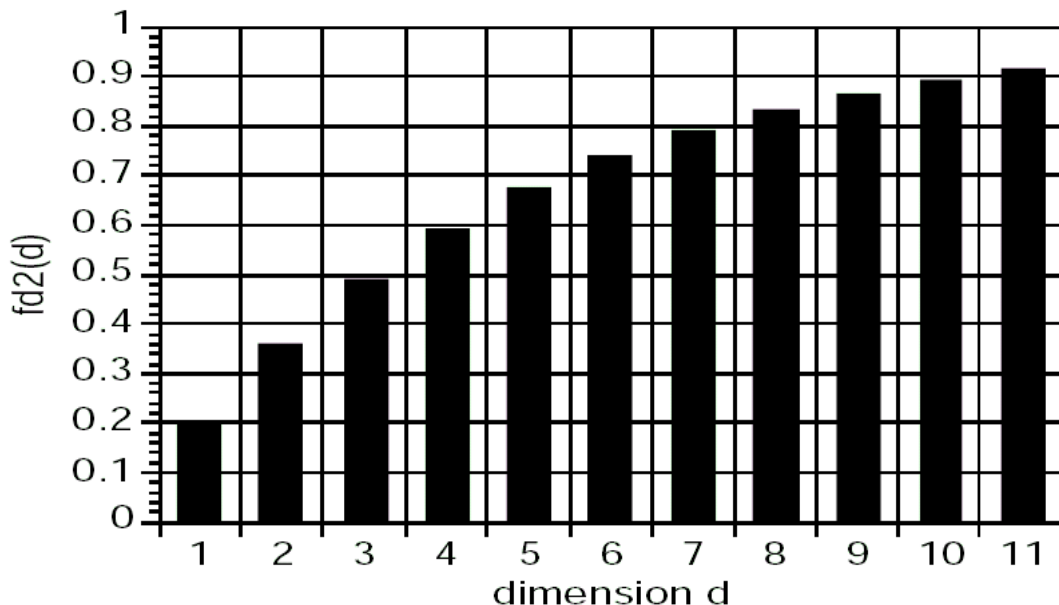
Παρατηρούμε πως καθώς η διάσταση αυξάνεται ο λόγος τείνει στο 0 που σημαίνει ότι ο όγκος συγκεντρώνεται στις γωνίες.

B) Ο όγκος της υπερσφαίρας συγκεντρώνεται στον εξωτερικό φλοιό:

Ο λόγος του όγκου που περιέχεται στον φλοιό ορίζεται σαν μια σφαίρα ακτίνας r:

$$f_{d2} = \frac{V_d(r) - V_d(r - \varepsilon)}{V_d(r)} = \frac{r^d - (r - \varepsilon)^d}{r^d} = 1 - \left(1 - \frac{\varepsilon}{r}\right)^d$$

Κάνοντας την αντίστοιχη γραφική παράσταση έχουμε:



Παρατηρούμε πως το όριο τείνει στο 1, αυτό σημαίνει ότι ο περισσότερος όγκος μιας υπερσφαίρας συγκεντρώνεται στον εξωτερικό της φλοιό.

Γ)Ο όγκος ενός υπερελλειψοειδούς συγκεντρώνεται στονεξωτερικό φλοιό.
Εδώ τα προηγούμενα αποτελέσματα θα γενικευτούν σε υπερελλειψοειδή.Έτσι η εξίσωση του είναι:

$$\frac{X_1^2}{\lambda_1^2} + \frac{X_2^2}{\lambda_2^2} + \dots + \frac{X_d^2}{\lambda_d^2} = 1$$

Και ο όγκος του είναι:

$$V_e(\lambda_i) = \frac{2 \prod_{i=1}^d \lambda_i}{d} \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$$

Ο όγκος του μπορεί να θεωρηθεί ως:

$$\frac{X_1^2}{(\lambda_1 - \delta_1)^2} + \frac{X_2^2}{(\lambda_2 - \delta_2)^2} + \dots + \frac{X_d^2}{(\lambda_d - \delta_d)^2} = 1$$

Έτσι για τον φλοιό έχουμε:

$$V_e(\lambda_i - \delta_i) = \frac{2 \prod_{i=1}^d (\lambda_i - \delta_i)}{d} \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}$$

οπότε κάνοντας τις ίδιες πράξεις με πιο πριν έχουμε το λόγο:

$$f_{d3} = \frac{\prod_{i=1}^d (\lambda_i - \delta_i)}{\prod_{i=1}^d \lambda_i} = \prod_{i=1}^d \left(1 - \frac{\delta_i}{\lambda_i}\right)$$

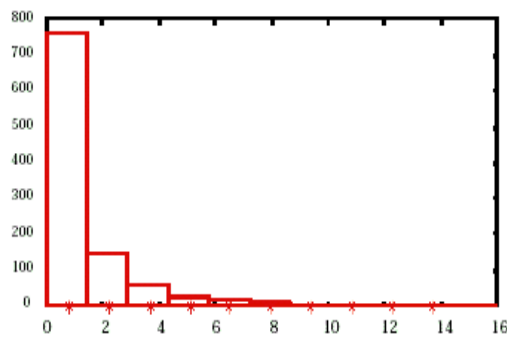
Παρατηρούμε πως το όριο συγκλίνει και αυτό στο μηδέν.

Τα χαρακτηριστικά που μόλις προαναφέρθηκαν έχουν κυρίως 2 συνέπειες για πολυδιάστατα δεδομένα που εμφανίζονται αμέσως. Το πρώτο είναι ότι ο πολυδιάστατος χώρος είναι κυρίως άδειος και αυτό σημαίνει ότι πολυδιάστατα δεδομένα στο ν-διάστατο ευκλείδιο χώρο συνήθως έχουν δομή μικρότερης διάστασης. Σαν συνέπεια αυτού είναι ότι τα πολυδιάστατα δεδομένα μπορούν να προβληθούν σε χώρους με μικρότερη διάσταση χωρίς να χάνουμε χρήσιμη πληροφορία σε όρους διαχωρισμότητας μεταξύ των διαφόρων κλάσεων. Η δεύτερη συνέπεια είναι πως τα κανονικά κατανομημένα δεδομένα τείνουν να συγκεντρώνονται στις ουρές, ομοια τα δεδομένα από ομοιόμορφη κατανομή να συγκεντρώνονται στις γωνίες, κάνοντας έτσι την εκτίμηση της πυκνότητας πιθανότητας αρκετά δύσκολη. Οι τοπικές γειτονιές είναι σχεδόν σίγουρα αραιές και χάνουν την αποτελεσματικότητα να περιγραφούν με ακρίβεια καθώς απαιτούν μεγάλο εύρος. Υποστήριξη για αυτή την τάση μπορεί να βρεθεί από την στατιστική συμπεριφορά των κανονικών και ομοιόμορφων πολυδιάστατων δεδομένων στην υψηλή διάσταση. Είναι αναμενόμενο ότι η αν η διάσταση αυξάνει τότε τα δεδομένα θα συγκεντρωθούν στον εξωτερικό φλοιό. Καθως μάλιστα αυξάνει ο αριθμός αυξάνει και η απόσταση από την αρχή επίσης.

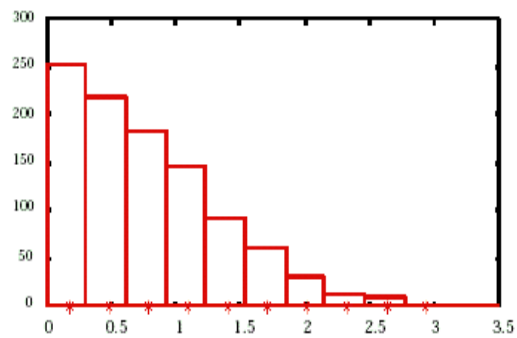
Για να δείξουμε αυτή την κατανομή των δεδομένων πρέπει να εφεύρουμε ένα πείραμα. Πολυδιάστατα δεδομένα από κανονικές και ομοιόμορφες κατανομές πρέπει να γεννηθούν. Οι κανονικές και οι ομοιόμορφες μεταβλητές είναι ανεξάρτητες και τα δείγματα προκύπτουν από τις $N(0,1)$ και $U(-1,1)$.

Στα παρακάτω σχεδιαγράμματα το δείχνουμε αυτό μέσω της απόστασης από το O ,του τετραγώνου των κατανομών για διάφορες διαστάσεις.

Normal, dimensions = 1

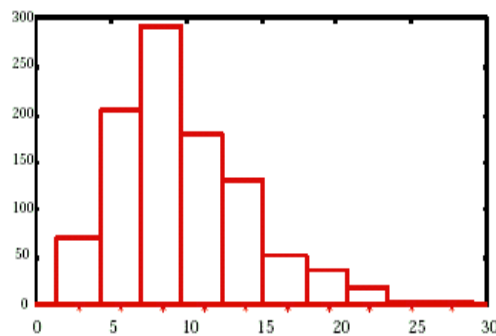


$$\sum_i x_i^2, \mu = 1.0168, \sigma = 1.4017$$

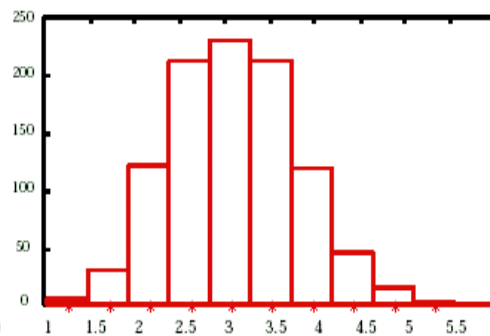


$$\sqrt{\sum_i x_i^2}, \mu = .7736, \sigma = .5737$$

Normal, dimensions = 10

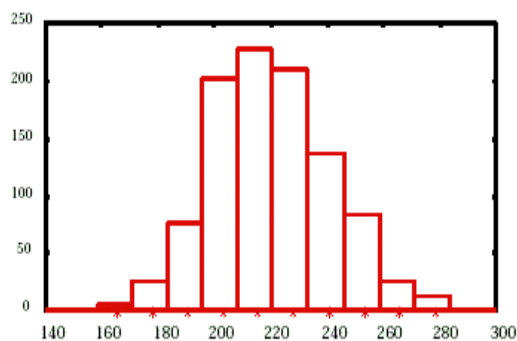


$$\sum_i x_i^2, \mu = 9.8697, \sigma = 4.5328$$

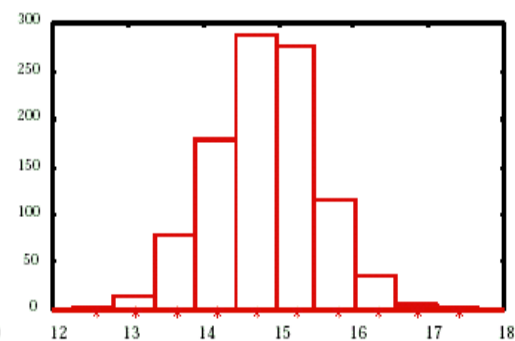


$$\sqrt{\sum_i x_i^2}, \mu = 3.1026, \sigma = 0.7042$$

Normal, dimensions = 220

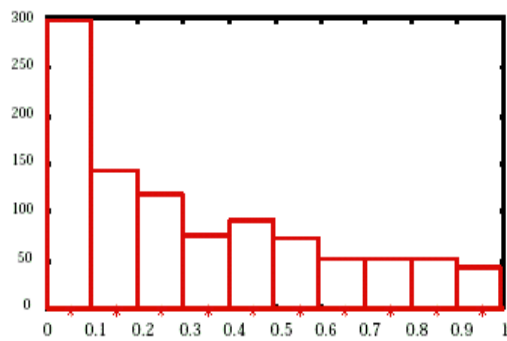


$$\sum_i x_i^2, \mu = 220.3732, \sigma = 20.7862$$

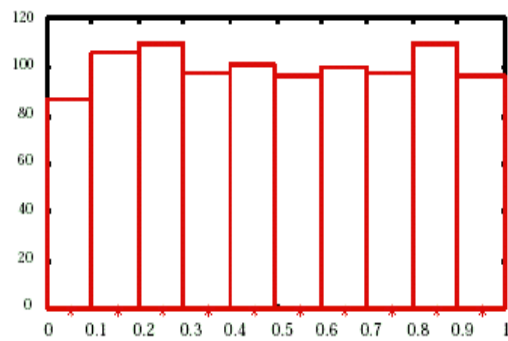


$$\sqrt{\sum_i x_i^2}, \mu = 14.8497, \sigma = 0.7129$$

Uniform, dimensions = 1

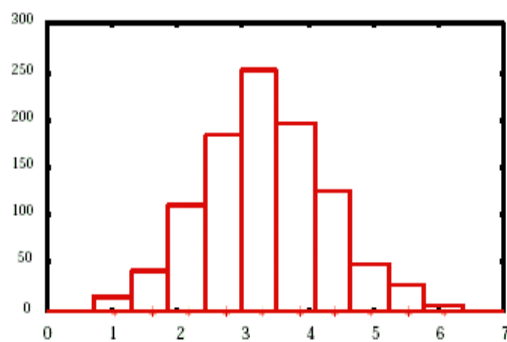


$$\sum_1 x_i^2, \mu = 0.3277, \sigma = 0.2883$$

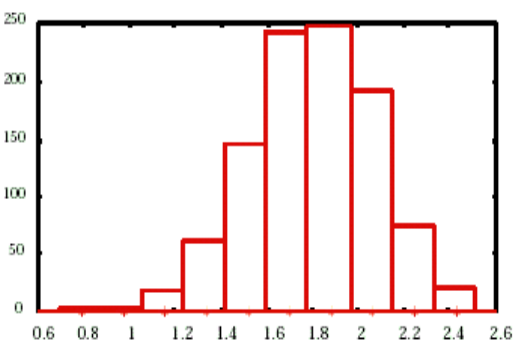


$$\sqrt{\sum_1 x_i^2}, \mu = 0.5041, \sigma = 0.2887$$

Uniform, dimensions = 10

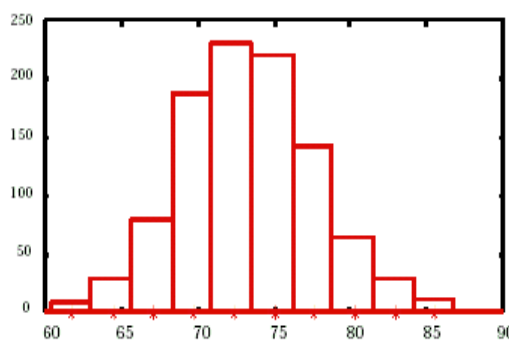


$$\sum_1 x_i^2, \mu = 3.3444, \sigma = 0.9390$$

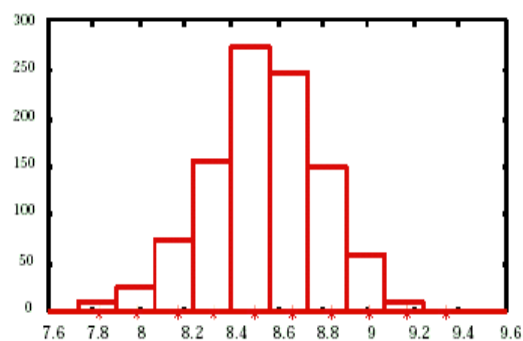


$$\sqrt{\sum_1 x_i^2}, \mu = 1.8010, \sigma = 0.2678$$

Uniform, dimensions = 220



$$\sum_1 x_i^2, \mu = 73.3698, \sigma = 4.3854$$



$$\sqrt{\sum_1 x_i^2}, \mu = 8.5488, \sigma = 0.2505$$

Αυτά τα παραδείγματα δείχνουν ότι ο μέσος όρος και η απόκλιση σαν συναρτήσεις

της διάστασης. Καθώς η διάσταση αυξάνει τότε τα δεδομένα συγκεντρώνονται στον εξωτερικό φλοιό. Ο μέσος όρος και η διακύμανση είναι:

$$r = \sqrt{\sum_{i=1}^d x_i^2} \quad \text{and} \quad R = \sum_{i=1}^d x_i^2$$

και υπολογίζονται. Αυτές οι μεταβλητές είναι και οι αποστάσεις στον ευκλείδειο χώρο. Αυτές οι τιμές των παραμέτρων και τα ιστογράμματα των τυχαίων μεταβλητών φαίνονται στα δεδομένα.

Καθώς η διάσταση αυξάνει, αυξάνει και η απόσταση από το κέντρο και τα δεδομένα συγκεντρώνονται στον φλοιό.

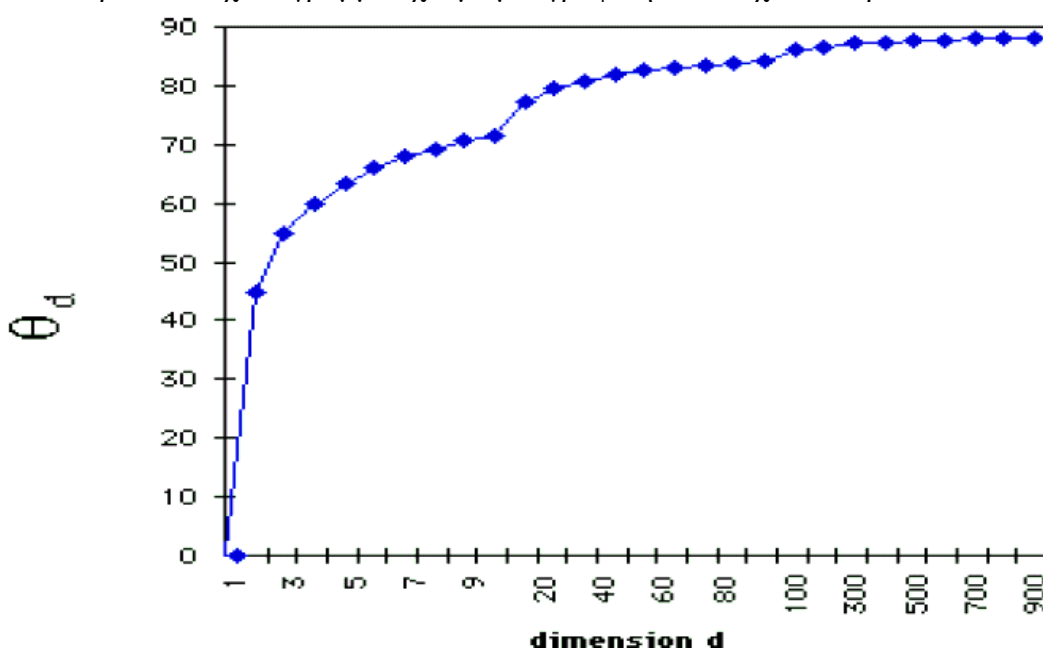
Υπό αυτές τις συνθήκες είναι δύσκολο να υλοποιήσουμε μια διαδικασία εκτίμησης της πυκνότητας και να λάβουμε ακριβή δεδομένα. Γενικά οι μη παραμετρικές τεχνικές θα έχουν ακόμη μεγαλύτερα προβλήματα με τις πολυδιάστατες απεικονίσεις και δεδομένα.

δ) Οι διαγώνιοι είναι σχεδόν κάθετοι στους άξονες συντεταγμένων:

Το συνημίτονο της γωνίας μεταξύ του διαγωνίου διανύσματος και του ευκλείδειου άξονα συντεταγμένων είναι:

$$\cos(\theta_d) = \pm \frac{1}{\sqrt{d}}$$

Στο παρακάτω σχεδιάγραμμα έχουμε μία γραφική που δείχνει το όριο:



Το αποτέλεσμα είναι σημαντικό διότι κάθε προβολή μίας συστάδας σε μία διαγώνιο, μέσω π.χ του μέσου όρου των χαρακτηριστικών μπορούν να χαθούν πληροφορίες που περιέχονται στα πολυδιάστατα δεδομένα. Για να το εξηγήσουμε αυτό, θα λάβουμε μία οποιαδήποτε διαγώνιο στο d -διάστατο χώρο. Έστω ότι είναι η συντεταγμένη σε αυτό τον χώρο. Έτσι κάθε σημείο μπορεί να περιγραφεί ως:

$$\mathbf{P} = \sum_{i=1}^d \alpha_i \mathbf{a}_i$$

οπότε η προβολή γίνεται σαν:

$$\mathbf{P}_{\text{diag}} = (\mathbf{P}^T \mathbf{a}_{\text{diag}}) \mathbf{a}_{\text{diag}} = \sum_{i=1}^d \alpha_i (\mathbf{a} \mathbf{c}_i^T \mathbf{a}_d) \mathbf{a}_d$$

όμως καθώς η διάσταση αυξάνεται τότε μηδενίζονται κάποιοι όροι ειδικά οι όροι μέσα στην παρένθεση, με αυτόν το τρόπο χάνουμε πληροφορία μέσα στο δ-διάστατο χώρο για αυτό το σημείο.

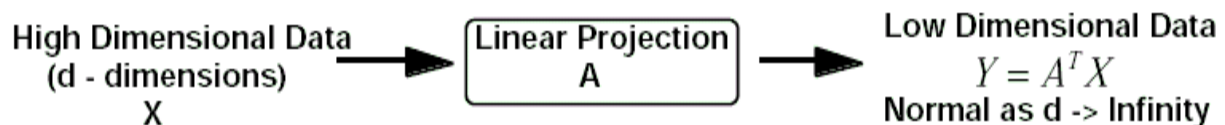
ε) Ο απαιτούμενος αριθμός των δειγμάτων για μία επιβλεπόμενη μάθηση και διαδικασία ταξινόμησης αυξάνει σαν αριθμός της διάστασης.

Ο Fukunaga αποδεικνύει ότι ο απαιτούμενος αριθμός δεδομένων εκπαίδευσης είναι γραμμικά εξαρτημένος από την διάσταση του γραμμικού ταξινομητή και το τετράγωνο της διάστασης για ένα τετραγωνικό ταξινομητή. Αυτό το γεγονός είναι πολύ σχετικό, μιας και τα πειράματα έχουν δείξει ότι υπάρχουν καταστάσεις όπου οι στατιστικές 2ου βαθμού έχουν αποδείξει πως είναι πιο σχετικά απ' ότι του πρώτου βαθμού για να ξεχωριστούν οι κλάσεις σε πολυδιάστατα δεδομένα. Στην περίπτωση των μη παραμετρικών ταξινομητών που η κατάσταση είναι ακόμη πιο ακραία. Έχει εκτιμηθεί ότι καθώς ο αριθμός των διαστάσεων αυξάνει, το μέγεθος πρέπει να αυξηθεί εκθετικά έτσι ώστε να έχουμε αποτελεσματική εκτίμηση για πολυδιάστατες πυκνότητες.

Έτσι είναι λογικό να περιμένουμε ότι τα πολυδιάστατα δεδομένα περιέχουν αρκετή πληροφορία σε σχέση με την ικανότητα για αναγνώριση πολλών κλάσεων με αρκετή ακρίβεια. Συγχρόνως τα παραπάνω χαρακτηριστικά μας επιδεικνύουν ότι με τις σύγχρονες τεχνικές που βασίζονται σε υπολογισμούς σε όλες τις διαστάσεις ίσως αυτό να μην δίνει πλεονεκτήματα αρκετά. Αυτό βασίζεται σε μία εργασία που αποδεικνύει πως ο ελάχιστος αριθμός δειγμάτων υπάρχει μία τιμωρία ως προς την ακρίβεια τιμωρίας καθώς ο αριθμός των χαρακτηριστικών ξεπερνά ένα συγκεκριμένο κατώφλι.

στ) Για τα περισσότερα υψηλής διάστασης δεδομένα, οι γραμμικές προβολές χαμηλής τάξης που έχουν την τάση να είναι κανονικά ή ένας συνδιασμός κανονικών κατανομών καθώς η διάσταση αυξάνεται.

Ένα σημαντικό χαρακτηριστικό των δεδομένων υψηλής τάξης είναι σχετικά σύμφωνο με την ανάλυση τους. Έχει αποδειχθεί ότι καθώς η διάσταση πάει προς το άπειρο, μικρότερες διαστάσεις γραμμικών προβολών μπορούν να προσεγγίσουν ένα κανονικό μοντέλο με πιθανότητα που τείνει στην μονάδα. Όπως εικονίζεται στο



σχεδιάγραμμα:

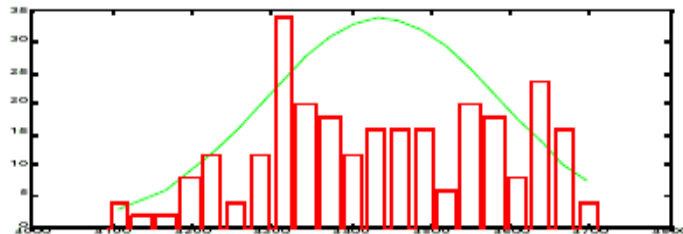
Η κανονικότητα σε αυτή την περίπτωση υποδεικνύει κανονική ή μείγμα κανονικών κατανομών.

Διάφορα πειράματα ότι αυτό μπορεί να αποδειχθεί τόσο με πραγματικά ή με προσωμειομένα δεδομένα. Αυτή η διαδικασία σε αυτά τα πειράματα είναι να προβάλλουν τα δεδομένα από ένα χώρο υψηλής διάστασης χώρο προς ένα μονοδιάστατο υποχώρο. Εξετάζουμε την συμπεριφορά των προβαλλόμενων δεδομένων και ένας αριθμός διαστάσεων στο αρχικό υψηλής διάστασης χώρο από 1 στο 10 και τελικά στο 100. Αυτή η μέθοδος προβολής των δεδομένων είναι να το πολλαπλασιάζει με ένα κανονικό διάνυσμα με τυχαίες γωνίες από τις συντεταγμένες. Ένα ιστόγραμμα μπορεί να χρησιμοποιηθεί για να παρατηρηθεί η

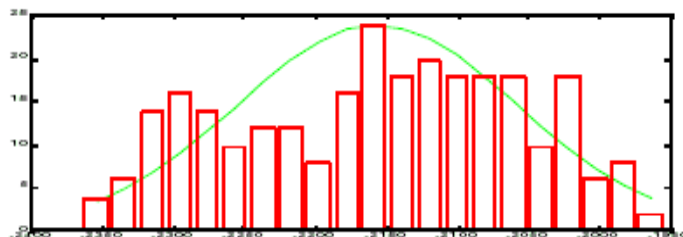
κατανομή των δεδομένων. Μια κανονική συνάρτηση πυκνότητας σκιαγραφείται με ένα ιστόγραμμα για να συγκριθούν τα αποτελέσματα.
Για την περίπτωση των ομοιόμορφα κατανεμημένων δεδομένων:

Από πραγματικά πολυδιάστατα δεδομένα έχουμε:

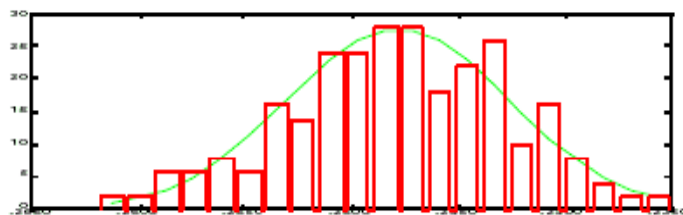
$d = 1$



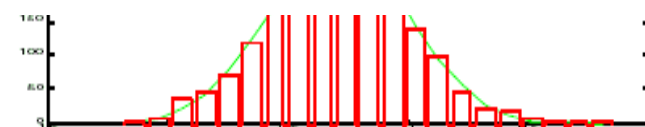
$d = 10$



$d = 100$



$d = 100$



High dimensional characteristics implications for supervised classification

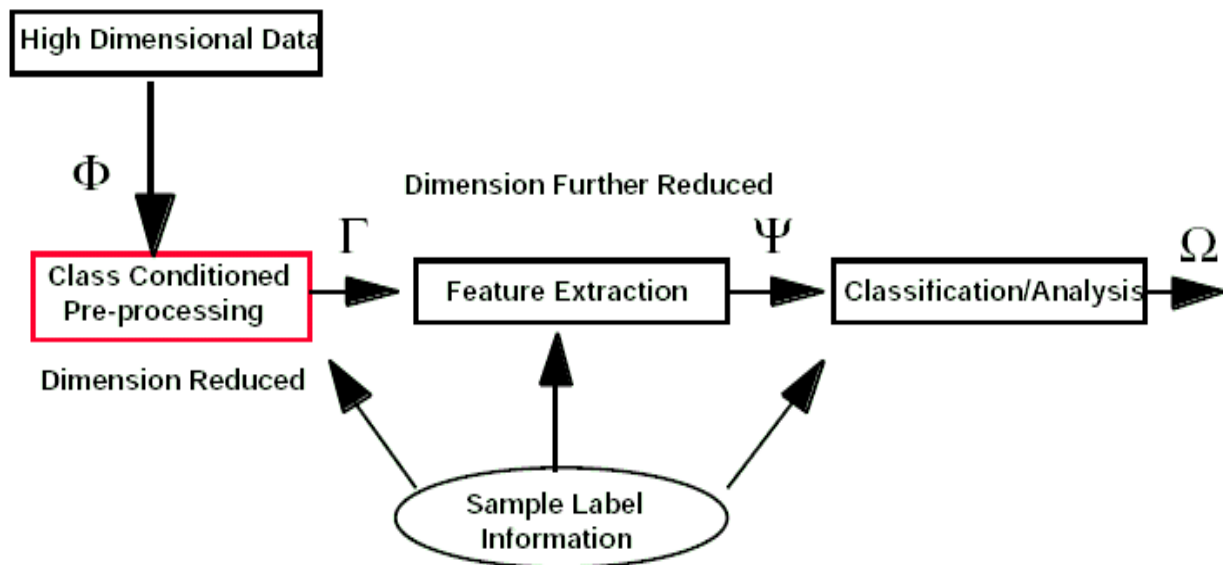
6.2 Επιπτώσεις των χαρακτηριστικών υψηλής διάστασης στην επιβλεπόμενη μάθηση:

Βασισμένο στα χαρακτηριστικά των υψηλών διαστάσεων δεδομένων ότι ο όγκος των υπερκύβων έχουν την τάση να συγκεντρώνονται στα άκρα και σε ένα υπερελλειψοειδές στον εξωτερικό φλοιό, και είναι προφανές ότι ο χώρος υψηλών διαστάσεων είναι κυρίως άδειος και τα πολυδιάστατα δεδομένα είναι συνήθως σε μια δομή χαμηλότερης διάστασης. Σαν συνέπεια αυτού είναι ότι είναι δυνατό να μειωθεί η διάσταση χωρίς να χαθεί σημαντική πληροφορία και η διαχωρισιμότητα. Λόγω των δυσκολιών της εκτίμησης πυκνότητας σε μη παραμετρικές μεθόδους, μία παραμετρική έκδοση της ανάλυσης δεδομένων των αλγορίθμων συνήθως θεωρείται ότι έχουν καλύτερη απόδοση όταν υπάρχουν ένας περιορισμένος αριθμός δειγμάτων και είναι διαθέσιμα για να παρέχουν την απαραίτητη πρωθύστερη πληροφορία. Ο αυξανόμενος αριθμός των δειγμάτων που απαιτείται για επιβλεπόμενη ταξινόμηση καθώς η διάσταση αυξάνεται παρουσιάζει ένα πρόβλημα στους τρέχοντες

αλγορίθμους εξαγωγής χαρακτηριστικών όπου ο υπολογισμός γίνεται στην πλήρη διάσταση, π.χ. Principal Components, Discriminant Analysis and Decision Boundary Feature Extraction. Μία νέα μέθοδος απαιτείται, αντί να κάνει υπολογισμούς στον πλήρη χώρο, να υπολογίζει ένα υποχώρο μικρότερης διάστασης. Κάνοντας τον υπολογισμό σε υποχώρο μικρότερης διάστασης και είναι αποτέλεσμα της γραμμικής προβολής από τον αρχικό πολυδιάστατο χώρο που κάνει την υπόθεση της κανονικότητας που είναι προσαρμοσμένος στην πραγματικότητα, δίνοντας έτσι μια καλύτερη εκτίμηση παραμέτρων και μια καλύτερη ακρίβεια ταξινόμησης.

Μία μέθοδος προεπεξεργασίας των δεδομένων υψηλής διάστασης βασίζεται σε αυτά τα χαρακτηριστικά που βασίζονται στις προβολές Projection Pursuit. Αυτή ονομάζεται Parametric

Projection Pursuit. Η Parametric Projection Pursuit μειώνει την διάσταση των δεδομένων και διατηρεί όσο το δυνατόν περισσότερη πληροφορία μέσω της προβολής και είναι μία μετρική της διαχωρισιμότητας. Η προβολή βασίζεται στην ελάχιστη απόσταση Bhattacharyya μεταξύ των κλάσεων λαμβάνοντας υπόψη τα χαρακτηριστικά πρώτης ή δεύτερης τάξης. Ο υπολογισμός βασίζεται σε ένα υποχώρο μικρής διάστασης όπου τα δεδομένα προβάλλονται. Τέτοιο είναι η προεπεξεργασία που γίνεται πριν την χρήση κάποιου αλγορίθμου εξαγωγής χαρακτηριστικών και την διαδικασία ταξινόμησης. Στο σχήμα εικονίζεται αυτό:



Ασυμπτωτικές πρώτης και δεύτερης τάξης στατιστικές ιδιότητες:

Οι Lee and Landgrebe [13] επιχείρησαν ένα πείραμα όπου ταξινόμησαν μερικά από τα υψηλής διάστασης δεδομένα έτσι ώστε να δούμε τον σχετικό ρόλο που παίζουν τα στατιστικά πρώτης και δεύτερης τάξης. Αυτό θα βοηθήσει στην κατανόηση των συνθηκών που απαιτούνται τόσο από τα πρώτης ή δεύτερης τάξης στατιστικά μεγέθη στην διάκριση μεταξύ των στατιστικών κλάσεων του χώρου υψηλής διάστασης. Είναι λογικό να υποθέσουμε, ότι καθώς ο αριθμός των χαρακτηριστικών αυξάνει, η δυναμική πληροφορία στα πολυφασματικά δεδομένα αυξάνει το ίδιο. Στην επιβλεπόμενη ταξινόμηση η αύξηση της πληροφορίας μεταφράζεται στον αριθμό των κλάσεων και την διαχωρισιμότητα τους. Προβάλλει ένα όριο στην ακρίβεια της ταξινόμησης λαμβάνοντας τα στατιστικά πρώτης και δεύτερης τάξης. Θα χρησιμοποιηθεί η απόσταση Bhattacharyya σαν μετρική της διαχωρισιμότητας. Η

Bhattacharyya απόσταση είναι το άθροισμα 2 στοιχείων, το ένα βασίζεται κυρίως στις διαφορές των μέσων και η άλλη βασίζεται στις διαφορές των διακυμάνσεων.

Η απόσταση Bhattacharyya υπό την προϋπόθεση της κανονικότητας υπολογίζεται ως εξής:

$$\mu = \frac{1}{8} (M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1} (M_2 - M_1) + \frac{1}{2} \ln \left[\frac{\left| \frac{\Sigma_1 + \Sigma_2}{2} \right|}{\sqrt{|\Sigma_1| |\Sigma_2|}} \right]$$

Θα θεωρήσουμε $\mu = \mu_M + \mu_C$ where μ_M , όπου ο πρώτος όρος αναφέρεται στις διαφορές μέσων και ο δεύτερος όρος στις διαφορές διακυμάνσεων.

3) Συμπεράσματα:

Δείξαμε ότι τα χαρακτηριστικά του πολυδιάστατου χώρου είναι πολύ διαφορετικά από τον 3-διάστατο χώρο που έχουμε συνηθίσει. Αυτό έχει σημαντικές επιπτώσεις στις τεχνικές επιβλεπόμενης ταξινόμησης. Υπό τους όρους της εκτίμησης παραμέτρων των τάξεων, ένας μεγάλος αριθμός από δείγματα απαιτείται για να γίνει επακριβής εκτίμηση, και το πρόβλημα μεγαλώνει καθώς η διάσταση αυξάνει. Στις μη παραμετρικές μεθόδους ο αριθμός των δειγμάτων που απαιτείται να έχει ακόμη μεγαλύτερο αριθμό δειγμάτων για την εκτίμηση την πυκνότητα των τάξεων. Αυτά τα 2 είδη εκτιμήσεων αντιμετωπίζονται όταν έχουμε ένα πρόβλημα με χαρακτηριστικά υψηλής διάστασης.

Οι παρούσες εργασίες υπό αυτές τις παρούσες πολυφασματικές, τα σύνολα εκπαίδευσης για να εκτιμηθούν τα στατιστικά της τάξης είναι αρκετά μικρά. Σαν συνέπεια είναι επιθυμητό να προβάλλουμε τα δεδομένα σε χώρο μικρότερης διάστασης όπου οι παρενέργειες των χαρακτηριστικών της γεωμετρίας της υψηλής διάστασης και το φαινόμενο του Hughes μπορεί να μειωθεί. Κοινές μέθοδοι που χρησιμοποιούνται είναι η ανάλυση σε κύριες συνιστώσες (Principal Components, Discriminant Analysis) και εξαγωγή χαρακτηριστικών στα όρια αποφάσεων (Decision Boundary Feature Extraction) έχουν όμως το μειονέκτημα ότι απαιτούν υπολογισμούς στην πλήρη διάσταση και έτσι ο αριθμός των απαιτούμενων δειγμάτων είναι πολύ μεγάλος. Οι μέθοδοι που χρησιμοποιούνται εκτιμούν στατιστικές οι οποίες δεν είναι απαραίτητα ακριβείς, εντούτοις καταλήγουν σε μικρότερης ακρίβειας ταξινόμηση. Άλλο πρόβλημα είναι η υπόθεση της κανονικότητας. Τίποτα δεν εξασφαλίζει ότι η μεγάλη διάσταση, το μοντέλο έχει πιστότητα μεγάλη. Έχειδειχθεί ότι οι χώροι υψηλής διάστασης είναι κυρίως άδαιοι, υποδεικνύοντας ότι η δομή δεδομένων υπάρχει κυρίως σε ένα υποχώρο. Το πρόβλημα του ποιον υποχώρο μπορεί να βρεθεί και είναι σχετιζόμενο με την κάθε περίπτωση. Ο στόχος να μειωθεί η διάσταση των δεδομένων στον κατάλληλο υποχώρο χωρίς να χαθεί η πληροφορία της διαχωρισιμότητας.

Κεφάλαιο 7

Αλγόριθμοι Κατάτμησης Φωνής-Algorithms for speech segmentation

1)Εισαγωγή

Η αυτόματη κατάτμηση φωνής είναι μία δύσκολη εργασία στην οποία τα σημερινής γενιάς συστήματα φωνητικής αναγνώρισης αποφεύγουν να ασχοληθούν ευθέως,αντίθετα σε τέτοιες εφαρμογές χρησιμοποιούν στατιστική μοντελοποίηση των μεταβάσεων μεταξύ των φωνητικών μονάδων μοιάζει να είναι πιο ελκυστικό από μία καθολική ,a priori συγκεκριμένη κατάτμηση.

Παρα ταύτα η αυτόματη κατάτμηση ανεξαρτήτου γλώσσας είναι μία σημαντική πρόκληση για την έρευνα φωνής η οποία παίζει σημαντικό ρόλο για την σύγχρονη τεχνολογία φωνής για εφαρμογές αυτόματης αναγνώρισης φωνής ,κωδικοποίησης,κειμένου σε φωνή σύνθεση .

Η αξιόπιστη αυτόματη κατάτμηση είναι θεμελιώδους σημασίας για την αυτόματη αναγνώριση φωνής.Με κάποιον τρόπο,η ακολουθία ηχητικών πλαισίων προέρχεται από την βραχέως χρόνου ανάλυση η οποία μπορεί να οργανωθεί σε ομογενή τμήματα τα οποία μπορούν να συσχετιστούν με ένα σύνολο συμβόλων που αναπαριστούν φωνήματα,λέξεις,συλλαβές και άλλες συγκεκριμένες ακουστικές μονάδες.Από αυτήν την σκοπιά ,η αυτόματη κατάτμηση φωνής μπορεί να είναι πολύ

χρήσιμη,επιτρέποντας το πρώτο τμήμα της φωνής του ήχου να μετατραπεί σε ένα σύνολο συμβόλων και υλοποιούν αλγορίθμους για την αναγνώριση.Όμως ,μία τεχνική λύση για το πρόβλημα κοινώς αποδεκτή δεν έχει ακόμη βρεθεί.Αυτός είναι ένας από τους κύριους λόγους για τον οποίο τα σύγχρονα συστήματα αναγνώρισης φωνής δεν μπορούν να λειτουργήσουν με ένα τρόπο από κάτω προς τα πάνω,π.χ ξεκινώντας από μια ολοκληρωμένη κατάτμηση.Αντ'αυτού καθορίζουν ένας γενικευτικός ,από πάνω προς τα κάτω αντιμετώπιση,εκτιμώντας την πιθανοφάνεια των πιο κορυφαίων γλωσσικών υποθέσεων και μοντελοποιώντας σαν στοχαστικές διαδικασίες,έχοντας ως βάση τη παρατηρούμενη ακολουθία των ηχητικών πλαισίων.Τα Hidden Markov Models(HMMs) είναι από τις πιο κοινές υλοποιήσεις σε μία αναπαραγωγική κατεύθυνση,και χρησιμοποιούνται ευρύτατα στην έρευνα και σε αρκετές εμπορικές εφαρμογές σε συστήματα ASR.Παρα ταύτα,η προαναφερθείσα τεχνική τείνει να πριμοδοτεί λεκτικούς περιορισμούς πάνω στην φωνητική πραγματικότητα όπου γνωστές φωνητικές περιγραφές χρησιμοποιούνται να περιορίζουν την αναγνώριση του σήματος ομιλίας σε μία ευθυγράμμιση κατευθυνόμενη(forced alignment).Μια πιο ισορροπημένη τεχνική μπορεί να υλοποιηθεί περιλαμβάνοντας φωνητική πληροφορία σε μία γενικευμένη αποκωδικοποιητική διαδικασία.Ένα πιθανόν βήμα σε αυτή την κατεύθυνση είναι να χρησιμοποιηθεί μία πιθανότητα μεταβίβασης μεταξύ των φωνητικών μονάδων σε μία καθολική κατάτμηση και ακουστική-φωνητική ουρά.

Ο σκοπός της κωδικοποίησης της φωνής είναι έτσι ώστε να παραχθεί μία συμπιεσμένη αναπαράσταση του ηχητικού σήματος για την αποδοτική μετάδοση και αποθήκευση.Η ψηφιακή φωνητική κωδικοποίηση χρησιμοποιούνται ευρύτατα σε καθημερινές εφαρμογές και σε μερικές από αυτές όπως κινητή τηλεφωνία και είναι νευραλγικής σημασίας.

Οι μέθοδοι ανάλυσης συχνά ανήκουν μεταξύ της ASR και των περιοχών κωδικοποίησης.Παρα ταύτα η προεπεξεργασία της φωνής και της μοντελοποίησης για λόγους κωδικοποίησης συχνά ικανοποιούν συγκεκριμένες ανάγκες οι οποίες διαφέρουν από αυτές που απαιτούνται στην ASR.Πράγματι,το σήμα της φωνής μπορεί να αναδομηθεί από κωδικοποιημένη πληροφορία και όλη η μεταβλητότητα

του και πλούσια δομή του έτσι ώστε να διατηρήσει όχι μόνο το γλωσσικό μήνυμα, αλλά και άλλες φυσικές ιδιότητες του σήματος (που μπορεί ο ακροατής να αντιληφθεί)

Διάφοροι κωδικοποιητές της φωνής συχνά βασίζονται σε ιδέες κατάτμησης όπως η στιγμιαία αποσύνθεση. Επιπλέον γλωσσικώς ανεξάρτητες τεχνικές κατάτμησης είναι απαραίτητες για να επιτευχθεί πολύ χαμηλούς ρυθμούς κωδικοποίησης φωνής μέσω της διαδικασίας ευρετηρίου στην μνήμη των φωνητικών τμημάτων. Για αυτούς τους λόγους είναι απαραίτητη η κατάτμηση των σημάτων της φωνής σε φωνητικές μονάδες, διότι η αναληθής ανίχνευση των φωνητικών οτίων μπορεί σημαντικά να μειώσει την απόδοση όλου του συστήματος. Γι' αυτό είναι επιτακτική η ανάπτυξη πιο αποτελεσματικών τεχνικών κατάτμησης φωνής που μπορούν να βοηθήσουν στην κωδικοποίηση της φωνής.

Στην έρευνα της φωνής, η διαθεσιμότητα μεγάλου και χαρακτηρισμένου ηχητικού λεξικού είναι δομικό θέμα. Τέτοιο λεξικό είναι απαραίτητο για την αυτόματη αναγνώριση φωνής και τεράστια δεδομένα εκπαίδευσης συνήθως χρησιμοποιούνται για να λάβουμε μεγαλύτερη απόδοση. Τεράστιες ποσότητες ταυτοποιημένων ηχητικών μονάδων χρειάζονται για να λάβουμε υψηλής πιστότητας για σύνθεση φωνής από κείμενο σε φωνή (TTS-text-to Speech) και εφαρμογές κωδικοποίησης. Η συλλογή και η κατάτμηση του λεξικού με χειρωνακτικό κόπο σε λεκτικό ή φωνητικό επίπεδο είναι διαδικασία πολύπλοκη και ακριβή, ειδικά για ηχογραφήσεις αυθόρμητες. Σαν συνέπεια αυτού, αξιόπιστες βάσεις για πειράματα υπάρχουν σε πολύ λίγες γλώσσες. Για να αποφύγουμε αυτές τις δυσκολίες, υπάρχουν διάφορες τεχνικές που κυρίως βασίζονται σε μια διαδικασία τύπου top-down ASR, όπως Viterbi alignment με μοντελοποίησης της των διαφόρων τρόπων προφοράς. Παρά ταύτα οι παραπάνω μεθοδολογίες έχουν καλή απόδοση μόνο εάν ένα επακριβές μοντέλο το οποίο συλλαμβάνει τα φωνητικά δεδομένα όπως διαλεκτικές μεταβολές, όταν έχουμε παραπάνω λέξεις από μία σε ένα σήμα κλτ, τα οποία είναι αρκετά συχνά στον αυθόρμητο λόγο.

Δισταγμοί, λανθασμένες ενάρξεις και άλλες δυσκολίες είναι άλλη μία πηγή του προβλήματος για αυτές τις μεθόδους. Επιπρόσθετα, η ακρίβεια της κατάτμησης για εφαρμογές TTS είναι υψηλότερες απ' ό,τι αυτές του ASR μιας και τα συστήματα ASR επικεντρώνονται στην σωστή αναγνώριση της φωνητικής ακολουθίας και δεν χρειάζονται όπως τα συστήματα TTS, ακριβείς τοποθετήσεις των φωνητικών ορίων. Η κατάτμηση χωρίς κείμενο είναι χρήσιμη για την επίλυση αυτών των προβλημάτων. Επιπλέον η έλλειψη γραπτού κειμένου ή όταν η επακριβής φωνητική μοντελοποίηση δεν είναι διαθέσιμη (λόγω χρημάτων και χρόνου), μία bottom up στρατηγική για γλωσσικώς-ανεξάρτητη ταυτοποίηση φωνημάτων, που βασίζεται στην αυτόματη φωνητική κατάτμηση είναι αρκετά ελκυστική ιδέα.

Επιπλέον πέρα από τις εφαρμογές που συζητήσαμε, υπάρχουν περιοχές της έρευνας στη φωνή που θα επωφελούνταν από μία τέτοια τεχνική. Για παράδειγμα η διαμόρφωση σε κλίμακα χρόνου (time scale modification) για να παράγει διάφορα φωνήματα με μεταβολές, είναι ένα ερευνητικό εργαλείο χρήσιμο για μελέτες στην γλωσσολογία και την φωνολογία.

2) Γενική επισκόπηση των τεχνικών για κατάτμηση φωνής

Αυτόματες μέθοδοι κατάτμησης της φωνής έχουν αντιμετωπιστεί με διάφορες στρατηγικές που βασίζονται σε διάφορα χαρακτηριστικά της φωνής, μερικά από αυτά έχουν και γλωσσική πληροφορία, άλλα πάλι βασίζονται σε διάφορες ταμπλέτες αναφοράς. Γλωσσολογικά περιορισμένη κατάτμηση βασίζεται στην χρήση μιας έγγραφης μορφής μίας πρότασης και προσπαθούμε να βρούμε τα όρια των

φωνημάτων. Η επιγραφή της πρότασης μπορεί να είναι τόσο χειρωνακτική όσο και να προέλθει από ορθογραφική επιγραφή. Η γλωσσικά περιορισμένη φωνητική κατάτμηση βασίζεται στην αξιόπιστη κατάτμηση που λαμβάνει φωνή και πληροφορία σχετικά με το περιεχόμενο (είτε μέσω της φωνητικής επιγραφής, ή ενός πιθανοκρατικού μοντέλου που προέρχεται μέσω μιας διαδικασίας μάθησης που βασίζεται σε ένα σύνολο χειρωνακτικώς κατετμημένων δεδομένων εκπαίδευσης). Πολλές διαδικασίες έχουν βασιστεί πάνω σε αυτά που προαναφέραμε όπως Dynamic Time warping (DTW), αλγόριθμοι Forward-Backwards συζευγμένα με Hidden Markov Models και τεχνητά νευρωνικά δίκτυα.

Η βασική στρατηγική σε αυτές τις μεθόδους που μόλις προαναφέρθηκαν είναι να υπολογίσει όλα τα όρια των φωνητικών τμημάτων που μας ενδιαφέρουν (μιας και η κατάτμηση μπορεί να γίνει σε επίπεδο φωνημάτων, συλλαβών, λεκτικών μερών, λέξεων καθώς και άλλων ακουστικών μονάδων) και να επιλέξουμε το καλύτερο που βασίζεται σε ένα αποδοτικό αλγόριθμο ανάλυσης συντακτικού (parsing algorithms). Παρόλα αυτά, αυτή η στρατηγική προσπαθεί να αποδόσει μόνο για επακριβής αντιστοιχία (exact matching) παραβλέποντας όμως το πρόβλημα της εύρεσης της μεταβίβασης (detection of transition) είτε διότι μερικές δεν μπορούν να ευρεθούν και να παρατηρηθούν με βάση το λεξικό που έγινε η εκπαίδευση ή επειδή είναι σχετικά σπάνια να προκύψουν και έτσι το μοντέλο εκτίμησης είναι σχετικά φτωχό στην διακριση και την περιγραφή. Σαν συνέπεια αυτής της φτωχής μοντελοποίησης προκύπτουν πολλά σφάλματα εισαγωγής που παράγονται και επηρεάζουν έτσι το επόμενο επεξεργαστικό τμήμα στο οποίο η αναγνώριση ή η σύνθεση των φωνητικών τμημάτων διαγνώστηκε. Για να ξεπεράσουμε αυτή τη δυσκολία, πολλές και διάφορες τεχνικές μοντελοποίησης του περιεχομένου (context modelling), που έγιναν κυρίως για εφαρμογές αναγνώρισης και σύνθεσης φωνής, έχουν προταθεί και βασίζονται σε αιτιοκρατικούς κανόνες ή στατιστικούς αλγόριθμους. Αυτές οι ιδέες δείχνουν να δουλεύουν αρκετά καλά και αυξάνουν παραπάνω από 10% την απόδοση σε μεθόδους ανεξάρτητως περιεχομένου (context independence). Τα μειονέκτηματα είναι το υψηλότερο υπολογιστικό φορτίο. Παρά ταύτα της μη επαρκούς μοντελοποίησης των περιοχών των μεταβάσεων, παράγει πολλά σφάλματα εισαγωγής (insertion error), το οποίο παραμένει. Παρ' όλα αυτά μερικοί ερευνητές έχουν προβάλλει ένα τμήμα αναγνώρισης των μεταβάσεων στο εμπρόσθιο τμήμα των εφαρμογών κατάτμησης φωνής. (Προσθεση κειμένου, σελ 264) Οι μέθοδοι ανεξαρτήτως κειμένου είναι ιδιαίτερες χρήσιμες σε εφαρμογές που απαιτούν ή μπορούν να επωφεληθούν από καθολική κατάτμηση φωνής, όταν γραπτό κείμενο δεν είναι διαθέσιμο ή ανακριβές όπως στην διαδικασία εύρεσης φωνής (speech annotation). Μάλιστα τέτοιες μέθοδοι είναι ιδιαίτερες χρήσιμες για πολυγλωσσικές εφαρμογές, ακόμα αν και η ανεξαρτησία από κείμενο δεν σημαίνει απαραίτητα και ανεξαρτησία από φωνή, μιας και συγκεκριμένες γλωσσικές ιδιότητες μπορούν να χρησιμοποιηθούν για την κατάτμηση.

Πολλές τεχνικές σε αυτό το μοτίβο υπάρχουν στην βιβλιογραφία. Μάλιστα μπορούν να διαχωριστούν σε 2 κατηγορίες: τις βασισμένες σε μοντέλο και τις ανεξαρτήτως μοντέλου τεχνικές. Και οι 2 μέθοδοι αναγκαστικά βασίζονται σε μετρικές αποστάσεων σε φασματικές αλλαγές μεταξύ 2 συνεχών τμημάτων φωνής. Η κύρια διαφορά αυτών των μεθόδων συνίσταται κυρίως στο εμπρόσθιο τμήμα της ανάλυσης της φωνής και προηγείται του υπολογισμού των μετρικών αποστάσεων.

Οι μέθοδοι βασισμένοι σε μοντέλα εξαρτάται κυρίως στην γραμμική πρόβλεψη (Linear Predictive Coding-LPC) και έχουν χρησιμοποιηθεί αρκετά σε μεθόδους κατάτμησης συνεχών τμημάτων από την ακολουθιακή ανάλυση των φασματικών μεταβολών. Λόγω της απλότητας και της υπολογιστικής

απλότητας, κυρίως μπορούν να χρησιμοποιηθούν για εφαρμογές καθαρής φωνής, παρ'όλα αυτά δεν είναι χρήσιμες για την μοντελοποίηση ήχων συμφώνων όπως τα ρηνικά, σύμφωνα και παύσεις, λόγω του ότι η LPC υπολογίζει φασματικά και είναι πάντοτε περιστρεφόμενη από την F0 αρμονική.

Η γενική ιδέα πίσω από αυτές τις τεχνικές είναι κυρίως να πετύχουν την κατάτμηση μέσω του να κάνουν ακολουθιακές διαγνώσεις των αλλαγών του μοντέλου. Δεδομένου του σήματος, τότε υπάρχουν 2 υποθέσεις προς μελέτη, πρώτα ότι δεν έχει διαγνωστεί αλλαγή στο δοθέν τμήμα και η άλλη ότι ήδη έχει γίνει μία αλλαγή. Όταν χρησιμοποιείται το μοντέλο που αναφέραμε ουσιαστικά η γκαουσιανή συνάρτηση πιθανοφάνειας υπολογίζεται κάτω από αυτές τις 2 υποθέσεις και ο λόγος πιθανοφάνειας υπολογίζεται για κάθε δεδομένο n.

Τα μοντέλα LPC συχνά χρησιμοποιούνται όταν είναι να εκτιμηθεί το φάσμα των φωνητικών σημάτων όταν άλλες μέθοδοι για την μέτρηση παραλλαγής των φασμάτων (όπως συναρτήσεις αλλαγής και συναρτήσεις διακύμανσης των φασμάτων). Δύο από αυτές τις συναρτήσεις μεταβολών είναι η Δελτα Φασματική συνάρτηση (Delta Cepstral Function-DCF) που εισήχθη από τον Mitchell καθώς και η Φασματικής μεταβολής συνάρτηση (Spectral Variation function-SVF). Ο μαθηματικός ορισμός τους είναι ο ακόλουθος:

$$DCF_k(n) = C_k(n+1) - C_k(n-1), k=1, 2, \dots, Q$$

όπου $C_k(n)$ είναι η κ-οστή φασματική παράμετρος για το τμήμα n και Q είναι ο αριθμός των φασματικών παραμέτρων που εξαγάγονται από αυτό το δείγμα. Έπειτα χρησιμοποιείται το $DCF_k(n)$ για να υπολογιστεί η συνάρτηση κόστους που θα μας καταστήσει δυνατό την ανίχνευση των φασματικών μεταβολών σχετικά με την μεταβίβαση των φωνημάτων.

$$SVF(N) = C_{\sim}(n-1)C(n-1) / (||C_{\sim}(n-1)|| * ||C(n-1)||)$$

όπου $C_{\sim}(n-1)$ είναι η διαφορά του ν-οστού φασματικού διανύσματος από τον αριθμητικό μέσο όρο που είναι εντός του παραθύρου γύρω από το n.

Οι μέθοδοι ανεξαρτήτως μοντέλου δίδει ένα μέσο να ξεπεράσει τα προβλήματα που έχουν διάφοροι αλγόριθμοι όπως στο παραμετρικό φιλτράρισμα. Αυτές οι μέθοδοι προσπαθούν να ταυτοποιήσουν τις φασματικές μεταβολές στο σήμα της φωνής, μέσω του να αντιστοιχούν το φωνητικό φάσμα με διάφορες τεχνικές μέτρησης διαφορών. Αυτό σημαίνει ότι εάν $S1(\omega)$ και $S2(\omega)$ είναι οι φασματικές συναρτήσεις κατανομών που εκτιμάται μεταξύ 2 γειτονικών τμημάτων του σήματος, μία μεταβολή μπορεί να βρεθεί εάν η απόσταση υπερβαίνει ένα συγκεκριμένο κατώφλι. Ανάμεσα σε αυτές τις γραμμές, διάφορες φασματικές μετρικές έχουν προταθεί μεταξύ αυτών και η φασματική διαφορά Kullback-Leibler και ορίζεται ως εξής:

$$D_{KL}(P||Q) = - \int_X \log \frac{dQ}{dP} dP$$

Το μειονέκτημα αυτών των μεθόδων ανεξαρτήτως μοντέλων μετράει αν αυτές είναι αυστηρά συνδεδεμένες σε σχέση με τις φασματικές τους πυκνότητες. Όταν μεικτού τύπου φάσματα είναι παρόντα σε σήματα, τότε μπορεί να ληφθούν μεγάλες διαφορές ενώ απαραίτητα δεν σημαίνει ότι τα σήματα απέχουν τόσο πολύ.

Όταν μία μεταβολή έχει διαγνωστεί τότε η διαδικασία κατάτμησης μπορεί να υλοποιηθεί με πολλούς τρόπους. {μεθοδοι περιγραφικά} Πέρα από τις μεθόδους που προαναφέρθηκαν υπάρχουν και άλλες που παρέχουν σχήματα καθολικής κατάτμησης περιλαμβάνοντας αθροιστικές μεθόδους και τοπική ομαλότητα για τις μετρικές μεταβολής ειδικά στην εκτίμηση των ορίων, κατάτμηση σε πολλαπλές κλίμακες

μεταβάλλοντας το εύρος,μετρικές λογαριθμικής μεταβολής στο πεδίο της συχνότητας,μετασχηματισμοί wavelet,εκτιμητές του ρυθμού της εντροπίας,μέθοδοι ελάχιστης φάσης,κυρτός χώρος(convex hull),χρονική αποσύνθεση ,μερικές μεταβολές των παραμετρικών μεθόδων φιλτραρίσματος και πολλές άλλες.Επίσης έχει εισαχθεί μία νέα μετρική φασματικής διαφοράς η οποία καλείται DAPCOSH η οποία προτάθηκε από τους Wei Gibson,που επίσης συγκρίνους το σφάλμα σταθερής κατάστασης για διάφορες μετρικές φασμάτων.

Ένα μειονέκτημα κοινό τόσο στις μεθόδους που βασίζονται σε μοντέλα και σε αυτές που δεν βασίζονται είναι ότι μαζί με τα σημεία που είναι σωστά για την κατάτμηση εισαγώνται και πάρα πολλά άλλα σημεία τέτοια ώστε δεν είναι σωστά.Έτσι υπάρχουν 2 προβλήματα τα οποία μπορούν να λυθούν όταν ο στόχος μας είναι να λύσουμε μία αυτόματη κατάτμηση φωνής χωρίς την χρήση πρότερης γλωσσολογικής πληροφορίας, 1)τι είδους ακουστικές μετρικές θα χρησιμοποιήσουμε για να είμαστε ικανοί να έχουν εύρωστη πληροφορίας για τις ακουστικές μεταβάσεις από το ένα φώνημα στο άλλο και 2) να μειώσουμε τον αριθμό των λανθασμένων εισαγωγών έτσι ώστε το ποσοστό των σωστών διαγνώσεων να μην επηρεαστεί από αυτό των λανθασμένων.Για να λύσουμε αυτό το πρόβλημα κάνουμε την διαδικασία της κατάτμησης σε 2 βήματα: 1)χρησιμοποιούμε διάφορες μεθόδους κωδικοποίησης και δοκιμάζουμε του νέου αλγόριθμου κατάτμησης σε σχέση με αυτών που ήδη προαναφέραμε.Το κύριο κριτήριο είναι να χρησιμοποιήσουμε αυτή την σύγκριση για να έχουμε την ικανότητα για κατάτμηση σε υψηλή ποιότητα π.χ με το να την συγκρίνουμε με μία χειρωνακτική κατάτμηση από πεπειραμένους χρήστες.Άλλο σημαντικό κριτήριο είναι η υπολογιστική απόδοση και σαφώς η ικανότητα να χρησιμοποιηθεί σε πιο πολυπλοκές φωνητικές εφαρμογές.

3)Προτεινόμενος Αλγόριθμος

Ο προτεινόμενος αλγόριθμος επιχειρεί να κάνει κατάτμηση σε έναν αριθμό από χρονικά μεταβαλλόμενα χαρακτηριστικά που λαμβάνονται από επεξεργασία βραχέως χρόνου.Η αυτόματη διαδικασία επιχειρεί να συλλάβει τις οξείες μεταβολές στην εξέλιξη αυτών των παραμέτρων .Θεμελιώδες είναι ότι ο αλγόριθμος συνεκτικά διάφορα γεγονότα οξέων μεταβολών που μπορούν να διαγνωστούν από συγκεκριμένα χαρακτηριστικά σε ένα συγκεκριμένο όριο φωνήματος,μιας και έχει παρατηρηθεί ότι οξείες μεταβολές δεν συμβαίνουν για κάθε παράμετρο,παρα ταύτα σε ένα μικρό χρονικό διάστημα μπορούν να παρατηρηθούν.Για τον λόγο αυτό ο αλγόριθμος χρησιμοποιεί μια fitting procedure,η οποία τοποθετεί το όριο του φωνήματος σε αυτό που καλείται βαρύκεντρο της κάθε ομάδας των ημι-ταυτόχρονων οξέων μεταβολών και συνδιάζει τις οξείες μεταβάσεις που έχουν διαγνωστεί και γειτονεύουν με το ν-οστό πλαίσιο σε ένα μοναδικό μαρκάρισμα του ορίου του φωνήματος.Ο αλγόριθμος ρυθμίζεται από 3 παραμέτρους.Οι 2 πρώτοι παράμετροι a,b σχετίζονται με την αναγνώριση και εύρεση των απότομων μεταβολών που συμβαίνουν στις τιμές των ακουστικών χαρακτηριστικών που κωδικοποιούν το σήμα στην χρονική του εξέλιξη. Η παράμετρος a ρυθμίζει πόσα συνεχόμενα πλαίσια χρησιμοποιούνται για να εκτιμήσουν την ένταση(ύψος) μίας απότομης μεταβολής.Όταν το ύψος περνάει ένα συγκεκριμένο κατώφλι b,τότε μία πιθανή μεταβολή από το ένα φώνημα σε κάποιο άλλο θα υπολογιστεί από την συνάρτηση fitting procedure.Η παράμετρος c χρησιμοποιείται από την διαδικασία και ορίζει το εύρος της γειτονιάς όπου το βαρύκεντρο απομονώνεται.

Ο αλγόριθμος σε ψευδοκώδικα είναι ο ακόλουθος:

Αλγόριθμος:(θεωρούμε είσοδο ένα σύνολο κ time-sequence $x_i[n]$ $i=1, 2, \dots, N, N=\text{\#αριθμός frames στην πρόταση}$)

Begin

{Για κάθε πρόταση στο σύνολο δεδομένων;

{Για κάθε χρονική ακολουθία $x_i[n]$ στην πρόταση;

Υπολογίζουμε την συνάρτηση $J_i[n]$;

}

Ανίχνευση peaks στο $J_i[n]$ σύμφωνα με το κατώφλι b ;

Αποθήκευση των κορυφών σε ένα πίνακα $S(i,n)$;

%Fitting Procedure

{Για κάθε frame $n=a$ to $N-a-c$

Υπολογίζουμε την συνάρτηση $f[n]$ και το ελάχιστο στο διάστημα $[n, n+c]$;

$nw=n'$, όπου $f[n'] = \min f[n]$;

$acc[nw] = acc[nw] + 1$;

$n=n+1$;

}

Ορίζουμε το όριο στο frame n σύμφωνα με το peak της $acc[n]$;

}

}

Στην συνέχεια ορίζουμε τις συναρτήσεις που ορίσαμε πιο πάνω

$J_i[n] = |\text{sum}(x_i[m], n-a, n-1) - \text{sum}(x_i[m], n+1, n+a)|$

$F[n'] = \min f[n], p < n < q$

$F(n) = \text{sum}(\text{sum}(S(m,i)|n-m|, 1, k), p, q)$ για το διάστημα $[p, q]$ (έχουμε διπλό άθροισμα)

$$J_i^a[n] = \left| \sum_{m=n-a}^{n-1} \frac{x_i[m]}{a} - \sum_{m=n+1}^{n+a} \frac{x_i[m]}{a} \right|$$

$$f[n] = \sum_{m=p}^q \sum_{i=1}^k S(m,i) |n-m| \quad \forall n \in [p, q]$$

$$n_{win} = n' : f[n'] = \min_{n \in [p, q]} f[n]$$

- Το a αναπαριστά το εύρος διαστήματος των υπολογισμών
- Το b σχετικό κατώφλι για την εύρεση των peaks
- Το c είναι το εύρος της fitting procedure να βρει κατάλληλα όρια κατάτμησης.
- S είναι δυαδικός πίνακας όπου $S(n,i)=1$ αν για ορθή μετάβαση στο frame n στην χρονική ακολουθία i .
- Η fitting procedure ουσιαστικά χρησιμοποιεί τα peaks και τοποθετεί το όριο του φωνήματος στο μέσο και κατ'ουσίαν κάνει μία αδρή συσταδοποίηση των ημι-ταυτόχρονων peaks που γίνονται σε παράθυρο μήκους c .
- Η συνάρτηση συνάθροισης (accumulation function), υλοποιεί την διαδικασία συσταδοποίησης αναγνωρίζοντας σε κάθε frame τα όρια φωνημάτων και τα συσταδοποιεί, δηλαδή αναγνωρίζει το υψηλότερο αριθμό peaks των κ coefficients που κωδικοποιούν το speech frame.
- Η accumulation function παράγει μια σειρά από frames Μεταβολής. Ο επακριβής χρονισμός (exact timing) γίνεται μέσω της ανάθεσης του

- segmentation point στο κεντρικό δείγμα του κάθε transition frame..Εδώ θα μπορούσαμε να χρησιμοποιήσουμε πιο σοφιστικέ εργαλεία για να το επιτύχουμε αυτό.
- Είναι σημαντική η επιλογή του τρόπου αναπαράστασης πληροφορίας(π.χLPC, RASTA,MFCCκτλ)
- Εξίσου σημαντική και ίσως και μεγαλύτερη είναι η επιλογή των a,b,c(εδώ ουσιαστικά με δοκιμές βρήκαμε κάποιες τιμές που είχαν καλύτερα αποτελέσματα).Όμως θα μπορούσε να επεκταθεί και να υπολογίζονται αυτοί οι παράμετροι με εντελώς αυτόματο π.χ ANN,όμως ο λόγος που δεν εφαρμόστηκε εδώ είναι ότι θέλαμε να δώσουμε περισσότερο κίνητρο στην ανάδειξη του αλγορίθμου,απ'ότι να εφαρμόσουμε μία τεχνική μηχανικής μάθησης.

3.1)Εκτίμηση απόδοσης

Το αποτέλεσμα του προτεινόμενου αλγορίθμου βασίζεται κυρίως σε 3 παραμέτρους a,b,c.Ωστόσο και η επιλογή κατάλληλου σχήματος αναπαράστασης της φωνής βοηθά ιδιαίτερα,επίσης και η διαδικασία κατωφλιοποίησης επηρεάζει τον αλγόριθμο.Αξίζει να σημειωθεί ότι η κατωφλιοποίηση είναι μία μη γραμμική διαδικασία.Έτσι μια συνεκτική ανάλυση θα πρέπει να παρέχει τα βέλτιστα σε αυτά που αναφέραμε για να εκτιμηθεί η απόδοση του αλγορίθμου.

Γι'αυτό το λόγο δοκιμάσαμε τον αλγόριθμο μας σε μέρος της βάσεως TIMIT καθώς και σε δείγματα από το IEL Corpus που είχε άριστη ποιότητα εγγραφής.Επίσης τα δοκιμάσαμε με τα παραπάνω διότι μας παρέχεται και η ανθρώπινη κατάτμηση που μπορούμε να θεωρήσουμε ότι είναι ιδανική.

Η πρώτη μετρική βασίζεται στην απόδοση,δηλαδή στο ποσοστό επιτυχών ανιχνεύσεων των ορίων των φωνημάτων.Στην δική μας εργασία θεωρήσαμε και μία ανοχή της τάξης 20ms για το όριο των φωνημάτων(αυτό είναι αρκετά κοινό στην βιβλιογραφία)

Έτσι ορίζεται ως $P_c = 100 S_c / S_t$,όμως από μόνο του αυτό το ποσοστό δεν είναι αρκετό για να μετρήσει την απόδοση.Γι'αυτό ορίζουμε και την μετρική που δηλώνει το πρόβλημα της υπερ-κατάτμησης(over-segmentation)που ορίζεται από την διαφορά S_{ins} που ορίζεται σαν όλος ο αριθμός των σημείων που διαγνώστηκαν και τον αριθμό των πρότυπων κατατμήσεων S_t

$$S_{ins} = S_d - S_t$$

Έτσι ορίζεται ο ρυθμός over-segmentation(D)

$$D = 100(S_d/S_t - 1).$$

Βέβαια μετρικές για την εκτίμηση της απόδοσης υπάρχουν και πιο σοφιστικέ και ακριβείς μαθηματικά και ορίζονται από την θεωρία της signal detectability.

3.2)Πειράματα κατάτμησης

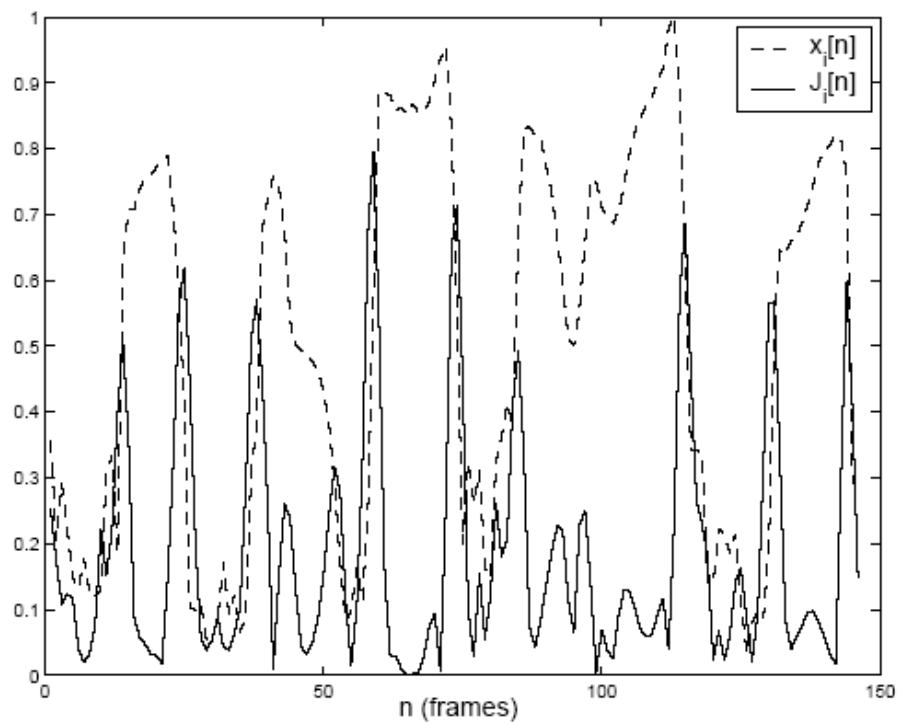
Τα πειράματα μας γίνανε ως ακολούθως.

- Δοκιμές στην ελληνική με ILSP corpus(1 γυναίκα,20 προτάσεις),κυρίως χρησιμοποιήθηκε για μελέτη του αλγορίθμου κατά πόσο μπορεί να γενικευτεί(πολύ καλή ποιότητα ήχου,ηχογράφηση ποιότητας studio 44,1 Khz)
- Δοκιμή σε TIMIT database(american english)σε τμήμα αυτής με ομιλητές(16αντρες,16 γυναικές,300 προτάσεις)

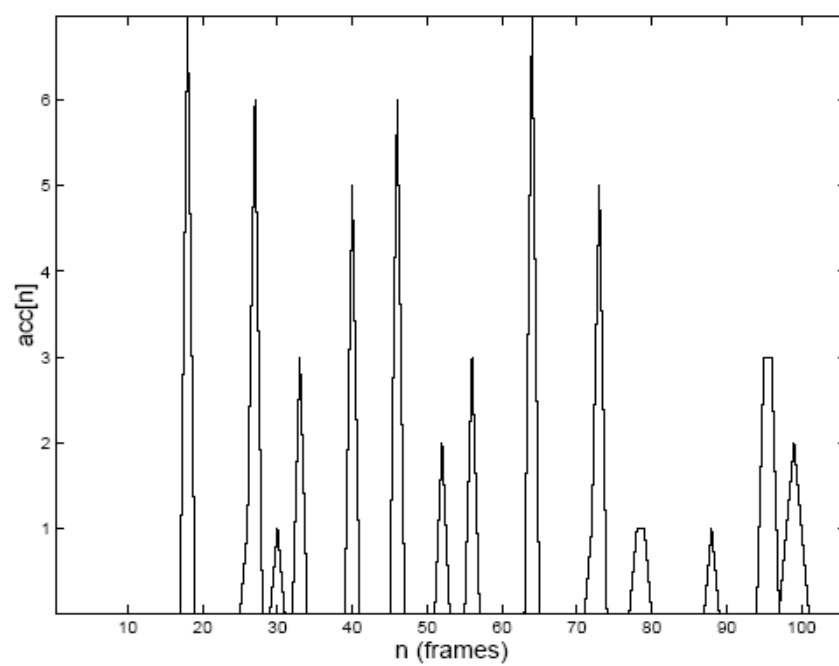
Θεωρούμε ότι οι δοθείσες βάσεις έχουν τέλεια ιδανική κατάτμηση και ως προς αυτές θα συγκριθεί η μέθοδος.

Αξίζει να σημειωθεί ότι χρησιμοποιήθηκε η ελληνική μιας και ο αλγόριθμος μας θέλαμε να δούμε πόσο εύκολα μπορεί να επεκταθεί σε πολυγλωσσικό περιβάλλον.

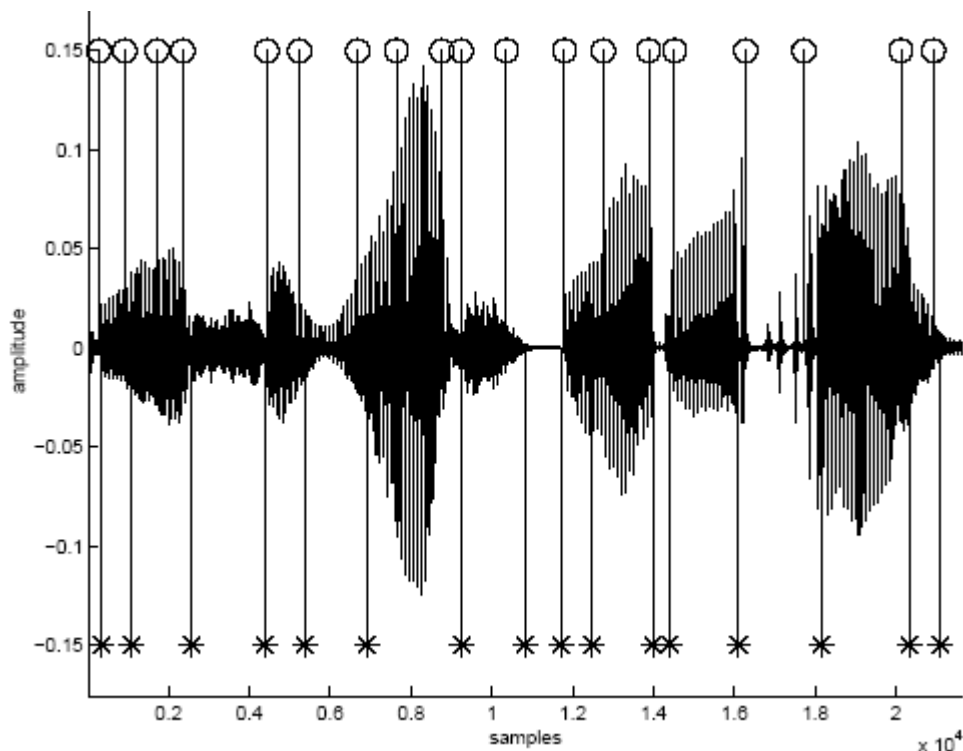
Παρακάτω δείχνουμε ένα στιγμιότυπο υπολογισμού από τον αλγόριθμο:



Η αντίστοιχη αθροιστική συνάρτηση:



Και η τελική κατάτμηση:



Αξίζει να σημειωθεί ότι η υλοποίηση έγινε με κώδικα σε C/Matlab και κάποια χρήση του HTK του κλασσικού εργαλείου για HMM. Καθώς και του προγράμματος Pratt. Το μεγαλύτερο μέρος του κώδικα υλοποιήθηκε από εμάς, αλλά παράλληλα δοκιμάσαμε και άλλα πακέτα που κυκλοφορούν στο διαδίκτυο για π.χ RASTA, PCA, Lyapunov eigenevalues, Higher order statistics κυρίως για εκπαιδευτικούς σκοπούς και όχι για την συλλογή μετρήσεων.

3.3) Αποτελέσματα

Κυρίως εμβαθύνουμε στο να αναδείξουμε τόσο τα πλεονεκτήματα όσο και τα μειονεκτήματα του αλγορίθμου. Γι' αυτό παρακάτω παραθέτουμε τους συνολικούς πίνακες και στην συνέχεια θα κάνουμε κριτική επ' αυτών.

α) Ο αλγόριθμος έχει αρκετά καλή επίδοση στην ανίχνευση της συχνότητας των μεταβολών από φώνημα σε φώνημα (σε κάθε επιτρεπόμενη κλάση, στην συνέχεια θα δείξουμε τι γίνεται αν συγκεκριμένες κλάσεις φωνημάτων). Δηλαδή τα πιο συχνά φωνήματα έχουν καλύτερη ανίχνευση από τα πιο αραιά εμφανιζόμενα φωνήματα (ίσως να μην εδράζεται αυτό το αποτέλεσμα μόνο στην εκπαίδευση των συστημάτων, αλλά ίσως να φέρει κάποια πιο χρήσιμα συμπεράσματα για την επεξεργασία γλώσσας ή ότι η κατανομή των φωνημάτων στις βάσεις θα πρέπει να είναι πιο ομοιόμορφα κατανεμημένη)

MFCC, $a=2, c=7, b=0.3$

Συχνότητα εμφάνισης	Επιτυχία
$100 < n < 500$	86%
$50 < n < 99$	78,3%
$10 < n < 50$	77.3%
$3 < n < 10$	78%
$n=3$	80%
$n=2$	75%
$N=1$	66%
Total	78%

β) Μεταβάσεις μεταξύ διαφόρων κλάσεων φωνημάτων (εδώ θα αναφέρουμε τις αδύναμες επιδόσεις <50%) με βάση όμως την άρθρωση (articulation)

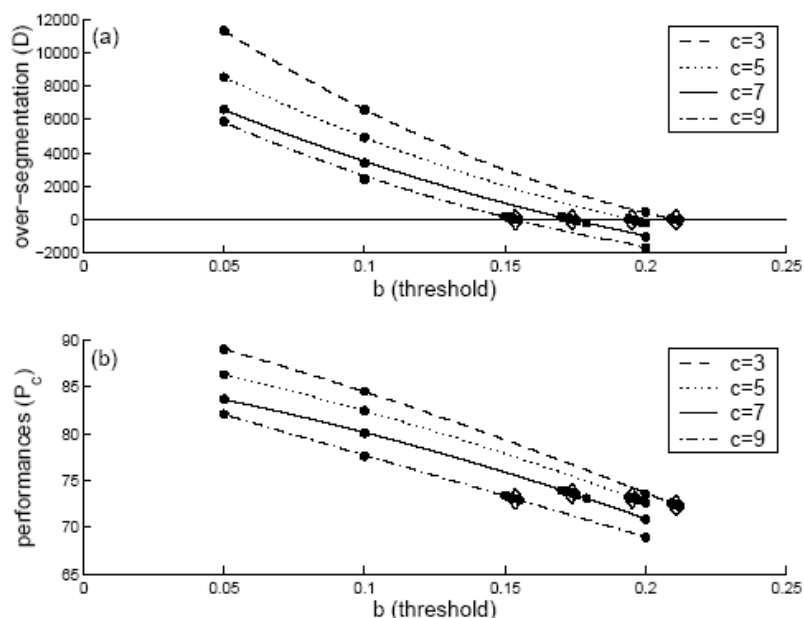
Κατηγορία Φωνήματος1	Κατηγορία Φωνήματος2	Επιτυχία(Αγγλικά/Ελληνικά)
Nasal	semi-vowel	48.1%/48.4%
Semi-vowel	vowel	47.9%/48.32%
Vowel	semi-vowel	46.51%/47.5%
Nasal	silence	40.4%/38.2%
Stop	fricative	34.92%/36%
Stop	silence	31.4%/30%
Vowel	vowel	29.2%/30.2%
Stop(καταληκτικό)	silence	17.8%/18.4%
Nasal	nasal	16.1%/17.4%
Total		42.78%/43.1%

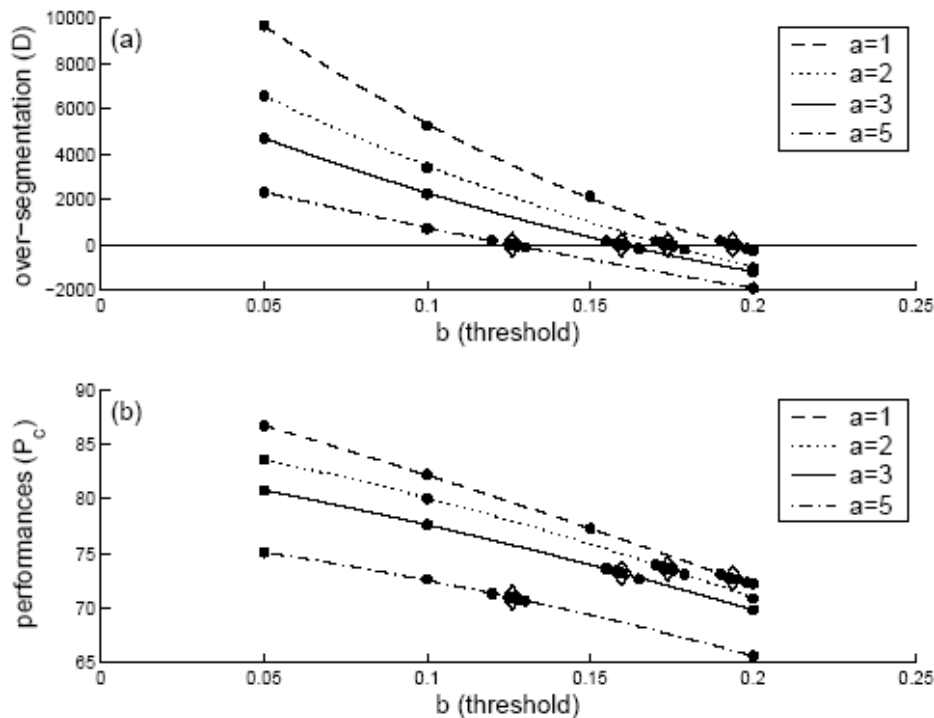
Τα ελληνικά απλά χρησιμοποιήθηκαν για μία αδρή σύγκριση(είναι μικρό δείγμα σε σχέση με τα αγγλικά)

Τα αποτελέσματα αναδεικνύουν τις αδυναμίες του αλγορίθμου για ταξινόμηση σε σχέση με κλάσεις άρθρωσης για ειδικές περιπτώσεις. Αυτό όμως είναι αναμενόμενο μιας και ο αλγόριθμος βασίζεται ιδιαίτερα στην ανίχνευση μεταβολών. Π.χ από vowel σε vowel(που είναι αρκετά ομαλοί και περιοδικοί ήχοι) είναι αναμενόμενο να αστοχούσε. Βέβαια πρέπει να διερευνηθεί πως σε πολλές από τις παραπάνω μεταβάσεις, ακόμα και αρκετοί expert έχουν πρόβλημα, αυτό ίσως θα έπρεπε να μας κάνει να σκεφτούμε την υφή της κατάτμησης αν όντως υπάρχουν αρκετές ασυνέχειες για να κάνουν εφικτή την εύρεση

γ) Παράλληλα με την παραπάνω μελέτη εκπονήσαμε και την συλογή κάποιων μετρήσεων για την απόδοση του αλγορίθμου για το πως εξαρτάται από τις κεντρικές παραμέτρους a, b, c:

Ενδεικτικά κάποιοι πίνακες





Δυστυχώς είναι πολύ κρίσιμη η βέλτιστη επιλογή των μεταβλητών μιας και το σύστημα εδράζεται συνολικά πάνω σε αυτές. Ιδιαίτερη μνεία πρέπει να γίνει για την δημιουργία μιάς αυτόματης μεθοδολογίας, μία ιδέα μπορεί να βασίζεται σε ANN, SVM κτλ

3.4) Πηγές σφαλμάτων

Ο αλγόριθμος έχει κάποιες εγγενείς αδυναμίες και τις οποίες προσπαθούμε να αναδείξουμε και παράλληλα αν μπορούμε σε μελλοντική εργασία να τις επεκτείνουμε

- Όπως αναφέραμε θεωρούμε ότι η ανθρώπινη κατάτμηση είναι ιδανική.. Όμως έρευνες έχουν δείξει ότι μεταξύ ειδικών οι κατατμήσεις διαφέρουν (πάλι με ανοχή 20ms) μέχρι και 4-6%...
- Έχουμε μεγάλη εξάρτηση της γειτονιάς των φωνημάτων (αυτό δεν συμβαίνει μόνο αλγοριθμικά, αλλά και στους ανθρώπους) π.χ είναι πιο εύκολη η διάγνωση ενός ρινικού φωνήματος σε σχέση με ένα φωνήεν, απ' ότι π.χ η παύση από ένα σύμφωνο. Αυτό όμως δεν είναι παράξενο διότι διαισθητικά μιας και η φωνή φέρει λεκτική πληροφορία και δεν είναι και στάσιμο σήμα είναι απολύτως φυσιολογικό.
- Ο αλγόριθμος εδώ δείχνει πως έχει μεγάλη εξάρτηση από την αναπαράσταση του σήματος (encoding scheme). Δηλαδή αν χρησιμοποιήσουμε πιο καλή αναπαράσταση (π.χ να χρησιμοποιήσουμε δυναμικά χαρακτηριστικά στην φωνή, χρήση φρακταλς, AM-FM models κτλ)
- Στην περίπτωση θορύβου, ο αλγόριθμος θα αστοχεί αρκετά συχνά διότι θα γεμίσει το σήμα με αρκετά peaks (false alarm-detection), όμως αν έχουμε εύρωστη αναπαράσταση ίσως αυτό το πρόβλημα να αποφευχθεί.
- Ο λόγος που αστοχεί ο αλγόριθμος είναι στην απλή υπόθεση που κάναμε ότι κατά την μεταβολή από φώνημα σε φώνημα θεωρούμε ότι υπάρχουν αρκετές ασυνέχειες (οξείες αλλαγές) για να είναι εφικτή η διάγνωση (δεν είναι πάντα αληθές)
- Αυτή η υπόθεση δεν συνεκτιμά ότι η άρθρωση μεταβάλλεται συνεχώς και ομαλά κατά την διάρκεια παραγωγής της φωνής και ότι οι αλλαγές του

φωνητικού σωλήνα από την μία κατάσταση γίνεται συνεχώς και όχι σε συγκεκριμένα χρονικά στιγμιότυπα.

- Αυτές όμως οι δυσκολίες αντιμετωπίζονται και από άλλες μεθοδολογίες που βασίζονται σε αυτό και ο μόνος τρόπος είναι ο αλγόριθμος απλά να αναγνωρίζει αυτές τις συνεχόμενες μεταβολές(π.χ μέσω χρήσης στατιστικών μοντέλων υψηλότερης τάξης,εντροπίας)

3.5)Συμπεράσματα

- Η κατάτμηση φωνής είναι ένα ανοικτό πρόβλημα με εφαρμογές στην κωδικοποίηση φωνής,αναγνώριση και σύνθεση.
- Μία πρακτική είναι αυτή που προτείναμε εδώ(bottom-up)όπου χρησιμοποιούμε ακουστικές ιδιότητες του σήματος,ή σαν ένα παραπροϊόν σε άλλες διαδικασίες τύπου top-down όπως στα HMM.
- Είναι αρκετά πρωτότυπος και απλός αλγόριθμος που δεν έχει πλήρως εκμεταλλευτεί και μπορεί κάλλιστα να βελτιωθεί με διάφορες προσθήκες,ειδικότερα για τις απαλές μεταβολές μεταξύ κάποιων κλάσεων φωνημάτων.
- Υπερτερεί στο γεγονός ότι είναι απλός υπολογιστικά και φθηνός σε πόρους(χαμηλή πολυπλοκότητα) .
- Ουσιαστικά ο αλγόριθμος προσπαθεί να διαγνώσει φασματική αστάθεια σε πολλαπλά εύρη συχνοτήτων
- Είναι ανεξάρτητος της λεκτικής πληροφορίας και της γλώσσας και πιο εύρωστος σε περιπτώσεις αστοχίας εκπαίδευσης και δοκιμής όπως συμβαίνει στα HMM,όπου απαιτούν υψηλή ποιότητα δεδομένων.
- Ο αλγόριθμος έχει υψηλή εξάρτηση με το τι μοντέλο εξαγωγής χαρακτηριστικών θα χρησιμοποιηθούν.
- Δεν χρησιμοποιεί ιδιαίτερες τεχνικές εκπαίδευσης..αν και δομικά επηρεάζεται από μεθοδολογίες των τεχνικών εκπαίδευσης(π.χ η ποιότητα των δεδομένων και η κατανομή τους)
- Πετυχαίνει ποσοστό επιτυχίας 78%(αγγλικά),75%(ελληνικά),80%(σε εγκεφαλογραφήματα),μπορεί να επεκταθεί και σε σεισμικά σήματα.

Κεφάλαιο 8

1)Μελλοντική Εργασία:

Αυτό το κεφάλαιο συνοψίζει τα αποτελέσματα της έρευνας μας στο πεδίο της κατάτμησης της φωνής και κατ'επέκταση της αναγνώρισης της φωνής. Προτείνουμε μια σειρά από ανοικτά ερωτήματα προς έρευνα. Το προσωπικό ενδιαφέρον του συγγραφέα εδράζεται κατά κύριο λόγο στην περιοχή της μηχανικής μάθησης και της σχεδίασης αλγορίθμων. Ωστόσο το πρόβλημα της αναγνώρισης φωνής (καθώς και το συναφές πρόβλημα της κατάτμησης) είναι ένα πολυπλευρο πρόβλημα, το οποίο είναι δεκτικό σε πολλές λύσεις, όμως ελάχιστες από αυτές οδηγούν σε μία καθολική λύση, διότι το πρόβλημα είναι εγγενώς πολύ δύσκολο και απαιτεί την ωρίμανση αρκετών τομέων και πεδίων. Επι παραδείγματι, τα μη γραμμικά μαθηματικά δίνουν μία άλλη διαίσθηση και οπτική για την επίλυση, αλλά είναι ακόμη σε αρχικό επίπεδο. Παράλληλα και η θεωρία των αλγορίθμων αν και έχει κάνει άλματα τα τελευταία 40 χρόνια, ωστόσο ακόμη αναπτύσσεται και θα δώσει στο μέλλον καλύτερη ποιότητα και τρόπο για την σχεδίαση αλγορίθμων για το πρόβλημα καθώς και άλλη κλάση τέτοιων προβλημάτων.

2)Προτάσεις για περαιτέρω έρευνα:

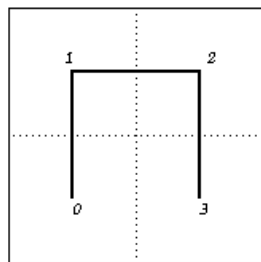
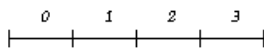
Το προτεινόμενο σύστημα κατάτμησης μπορεί να ενισχυθεί σε πολλές περιοχές. Μερικές προτάσεις για μελλοντική έρευνα περιλαμβάνουν τα ακόλουθα πεδία:

- Ανάπτυξη νέου μοντέλου της φωνής που θα λαμβανεί υπόψη καινούργια ευρήματα-κυρίως ποσοτικά- για την φύση των φωνητικών σημάτων.
- Μοντέλα που θα βοηθούν καλύτερα στην αναπαράσταση της φωνής (τόσο στο φασματικό ή στο πεδίο του χρόνου) που θα είναι αρκετά εύρωστη ώστε να ενθυλακώνει με κομψό τρόπο την μεγάλη μεταβλητότητα που υπάρχει στην ομιλία ειδικά στην ομιλία τόσο ενός ομιλητού όσο και συνόλου ομιλητών. Με αυτό τον τρόπο θα ενισχυθεί η διαδικασία της κατάτμησης.
- Πρόσφατα αναπτύχθηκε μία νέα θεωρία που ονομάζεται Compressive Sampling και ουσιαστικά διατυπώθηκε ένα νέο μη-γραμμικό θεώρημα που είναι πιο κομψό και φειδωλό από το αντίστοιχο του Shannon και ουσιαστικά μεταθέτει την επεξεργασία του σήματος όχι πια σε σχέση με το εύρος του σήματος, όσο σε σχέση με την δομή του (και βασίζεται στο γεγονός ότι τα σήματα τείνουν να είναι αραιά). Διάφορες εφαρμογές δείχνουν αρκετά καλή ποιότητα αποτελεσμάτων
- Επιλογή νέων μετρικών αποστάσεων που θα έχουν την ικανότητα να παράγουν μια συνάρτηση ομαλής φασματικής μεταβολής. Γενικά οι μετρικές θα πρέπει να είναι τόσο δομικά εύρωστες, όσο και υπολογιστικά αποτελεσματικές. Επι παραδείγματι καθώς αυξάνεται η διάσταση του χώρου τότε οι αλγόριθμοι που έχουμε προς το παρόν είναι αργοί, π.χ. αν επιθυμούμε να κάνουμε αναζήτηση σε μία διάσταση από μία συλλογή n στοιχείων τότε αυτό μπορεί να γίνει σε χρόνο $O(\log n)$, αν όμως αρχίζουμε να “ανεβαίνουμε” σε διαστάσεις, τότε οι καλύτεροι υπάρχοντες αλγόριθμοι είναι γραμμικοί, δηλαδή για την εύρεση ενός στοιχείου, θα πρέπει να διαβάσουμε

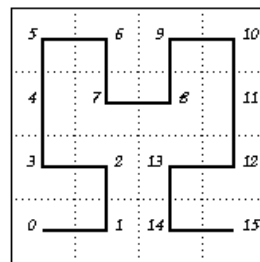
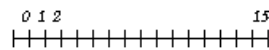
όλο το χώρο.Ο λόγος που γίνεται αυτό δεν οφείλεται μόνο στις ιδιότητες του χώρου που περιγράψαμε σε προηγούμενο κεφάλαιο,αλλά και στο γεγονός ότι πέρα από την διάσταση-1 χάνουμε την ιδιότητα της διάταξης .Παρα ταύτα η κοινότητα των ερευνητών βάσεων δεδομένων έχει προτάξει διάφορες μεθοδολογίες για να αντιμετωπίσει αυτό το πρόβλημα το οποίο είναι κεφαλαιώδους σημασίας.Μια ιδέα είναι με τις καμπύλες κάλυψης του χώρου(space filling curves).Μία μεθοδολογία είναι με τις καμπύλες Hilbert(εικονίζονται παρακάτω)

The Hilbert Curve

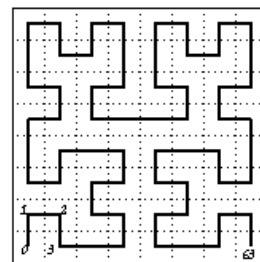
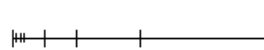
First Order



Second Order

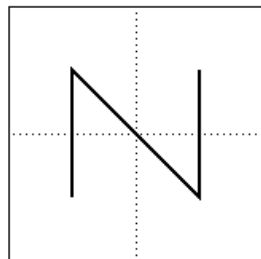


Third Order

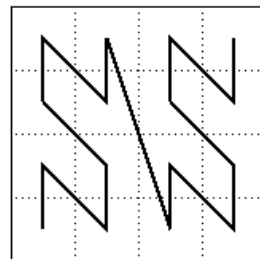


The Z-Order Curve

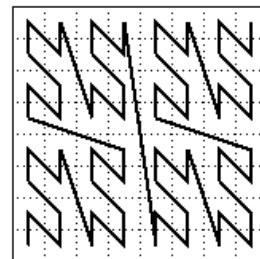
First Order

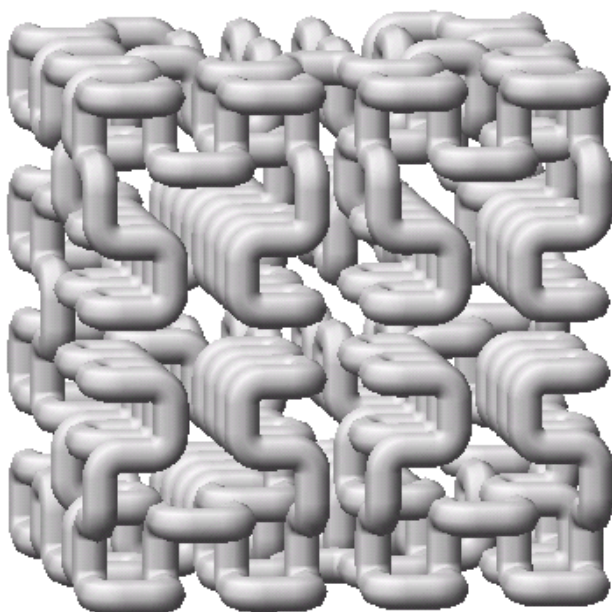


Second Order

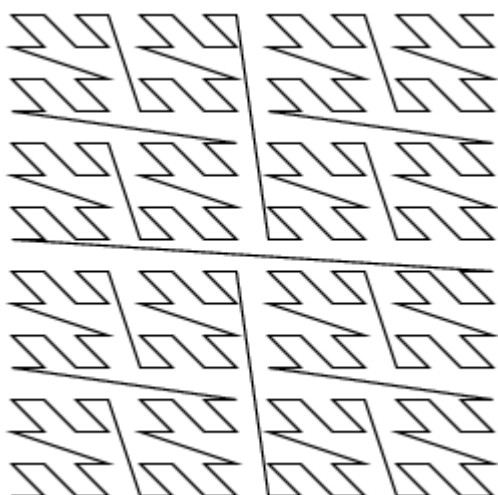


Third Order





Παρά το γεγονός ότι είναι μία τεχνική αρκετά έξυπνη, δυστυχώς δεν συμπεριφέρεται καλά πέρα από κάποιες διαστάσεις (διάσταση $> 7,8$) και οι αλγόριθμοι πάλι τείνουν να είναι γραμμικοί. Βέβαια αυτό οφείλεται όπως αναλύσαμε σε προηγούμενο κεφάλαιο στην υφή του πολυδιάστατου χώρου



- Περί του νόμου του Zipf: Ο νόμος του Zipf δεν συναντάται μόνο στην γλωσσολογία, αλλά ανήκει σε μία γενικότερη κλάση κατανομών που ονομάζονται power law. Συναντώνται μέχρι και στην οικονομία π.χ μέσω της αρχής του Παρέτο. Όμως μήπως είναι βαθύτερη η σχέση με τον τρόπο που κάνουν οι άνθρωποι (σε σύνολο) βελτιστοποίηση με πολλαπλά κριτήρια (multi-objective optimization).
- Ορμώμενοι από τον νόμο του Zipf και το γεγονός ότι η φωνή πρόκειται για μέσο συνεννόησης, αλλά παράλληλα κάνουν βελτιστοποίηση στο ότι καταβάλλουν την ελάχιστη δυνατή προσπάθεια για επικοινωνία. Μήπως θα πρέπει να χρησιμοποιηθεί άλλη μαθηματική κατεύθυνση και εργαλεία. Μία από αυτές ίσως να μπορούσε να έχει σχέση με την θεωρία παιγνίων, όπου παίκτες άνθρωποι και μηχανές προσπαθούν να επικοινωνήσουν και να αλληλοκατανοηθούν. Ο παραπάνω νόμος δίνει την ευκαιρία να κάνουμε αυτή τη σκέψη διότι ο νόμος μπορεί να θεωρηθεί σαν μία μορφή ισορροπίας κατά Nash μιας και οι παίκτες (άνθρωποι) όταν συννενοούνται δεν έχουν κίνητρο να

αποκλίνουν από την καταβολή ελάχιστης προσπάθειας(που είναι η ισορροπία).Είναι θέμα ανοικτό προς διερεύνηση.

- Χρήση νέων τεχνικών μάθησης πέρα από τα ANN,HMM,SVM και χρήση πιο εξελιγμένων αλγορίθμων εκπαίδευσης.
- Χρήση της εντροπίας κατά Κολμογκόροφ,η οποία καλείται και περιγραφική πολυπλοκότητα και σχετίζεται με την MDL(minimum description length)και μπορεί να οδηγήσει σε πιο συμπαγή και κομψή περιγραφή της πληροφορίας σε ένα σήμα καθώς και στην διαδικασία μάθησης.

Σίγουρα τα πεδία της έρευνας είναι πάρα πολλά και άλλωτε προχωράει με γρήγορους ρυθμούς και άλλωτε με πιο αργούς.Σίγουρα όμως απαιτείται κάποιο είδος ωρίμανσης μιας και βρισκόμαστε στις απαρχές πολλών τομέων.Η έρευνα στην φωνή και σε σχέση με την επεξεργασία της είναι μία πρόκληση η οποία τραβά την προσοχή πολλών επιστημόνων και λόγω φύσεως του προβλήματος είναι καθαρά διεπιστημονική.

Κεφάλαιο 9

Βιβλιογραφία

- [1] Καραγιάννης Γ. Και Ραγκούση Μ., “Ψηφιακή Επεξεργασία Σήματος”, Εκδόσεις Συμμεών, 1989.
- [2] Rabiner L. and Schafer R., “Digital Processing Of Speech Signals”, Englewood Cliffs, Prentice-Hall, 1978.
- [3] Rabiner L. and Juang B.H., “Fundamentals of Speech Recognition”, Englewood Cliffs, Prentice-Hall, 1993.
- [4] Lawrence R. Rabiner, Fellow, IEEE , “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, 1989
- [5] Biologically-Inspired Engineering, Hynek Hermansky, Tutorial
- [6] Hermansky, H. and N. Morgan, "RASTA Processing of Speech", IEEE Transactions on Speech and Audio Processing, Vol. 2, no. 4, October 1994.
- [7] Oppenheim, A.V. and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 1989.
- [8] Hermansky, H., “Perceptual linear predictive (PLP) analysis of speech,” *J. Acoust. Soc. Am.* **87**(4), 1738-1752, 1990.
- [9] Feller, “An Introduction to Probability Theory and Its Applications”, Vol 1, Wiley 1957
- [10] Li M. and Vitanyi P. “An Introduction to Kolmogorov Complexity and Its Application”, Springer-Verlag, 1997
- [11] Krylov N.V., “Introduction to the Theory of Random Processes”, American Mathematical Society, 2002
- [12] Shannon C.E., Weaver W., *The Mathematical Theory of Communication*, Univ. Of Illinois Press, Urbana and Chicago, 1949
- [13] Cover, T. M., and Thomas, J. A. Elements of Information Theory. John Wiley & Sons, Inc., New York, 1991.
- [14] Robert M. Gray, “Entropy and Information Theory”, Springer-Verlag, 1990
- [15] Thomas L. Griffiths
“Probability, algorithmic complexity, and subjective randomness”,
- [16] Christoph Bandt and Bernd Pompe, “Permutation entropy — a natural complexity measure for time series”, *Phys. Rev. Lett.*, 2002
- [17] Khurram Waheed, Kim Weaver and Fathi M. Salam, “A ROBUST ALGORITHM FOR DETECTING SPEECH SEGMENTS USING AN ENTROPIC CONTRAST”

- [18]Pattern Recognition,Sergios Theodoridis,Konstantinos Koutroumbas,Elsevier
- [19]Randomized Algorithms,Motwani,Raghavan,Cambridge Press
- [20]Nonlinear Speech Modeling and Applications,Advanced Lectures,Springer
- [21]How to Solve It:Modern Heuristics,Michalewicz,Fogel,Springer Verlag
- [22]The Simple Genetic Algorithm:Foundations and Theory,Michael D.Vose,Prentice Hall of India.
- [23]Speech and Language Processing:An Introduction to Natural Language Processing,Computational Linguistics and Speech Processing,Jurafsky & Martin,Prentice Hall
- [24]Genetic Algorithms in Search,Optimization and Machine Learning,David Goldberg,Addison-Wesley
- [25]An introduction to genetic algorithms,Mitchell,MIT Press
- [26]Graphs,Algorithms,and Optimization,Kocay & Kreher,Chapman&Hall/CRC
- [27]An information-theoretic approach to Neural Computing,Deco&Obradovic,Springer Verlag
- [28]An introduction to Probability:Theory and its Applications,Volume 1,Wiley
- [29]Πιθανότητες,Τυχαίες Μεταβλητές και Στοχαστικές Διαδικασίες,3η έκδοση,Α.Παπουλής,Εκδόσεις Τζιόλα
- [30] V. Pitsikalis and P. Maragos, "Speech analysis and feature extraction using chaotic models," in Proc.ICASSP , 2002.
- [31] Andrew Lindgren, Michael Johnson, and Richard Povinelli, "Speech recognition using reconstructed phase space features," in Proc. ICASSP , 2003.
- [32] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum,"IEEE Trans ASSP , vol. 34, no. 1, pp. 52–59, 1986.
- [33] H. AKAIKE, A New Look at the Statistical Model Identification, IEEE Trans. on Automatic Control, Vol.19, pp. 716-723, December 1974.
- [34] S. Kullback. Information Theory and Statistics . Wiley, New York, 1959.
- [35] ICASSP (1991). Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing , Toronto. Institute of Electrical and Electronic Engineers.
- [36]ICASSP (1992). Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing , San Francisco. Institute of Electrical and Electronic Engineers.
- [37]ICASSP (1993). Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing , Minneapolis, Minnesota. Institute of Electrical and Electronic Engineers.
- [38]ICASSP (1994). Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing , Adelaide, Australia. Institute of Electrical and Electronic Engineers.
- [39]ICSLP (1994). Proceedings of the 1994 International Conference on Spoken Language Processing , Yokohama,

Japan.

[40]Maragos, P., Kaiser, J., and Quatieri, T. (1992). On separating amplitude from frequency modulations using energy operators.

[41]CARNEGIE MELLON UNIVERSITY NONSTATIONARY GAUSSIAN PROCESSES FOR REGRESSION AND SPATIAL MODELLING,Christopher Joseph Paciorek

[42]Nonlinear Order Statistic Filter Design: Methodologies and Challenges ,Moncef Gabbouj and Jaakko Astola

[43]Poincare Sections for Pitch Mark Determination, Gernot Kubin

[44]An Investigation of Nonlinear Speech Synthesis and Pitch Modification Techniques,Iain Mann edinburg

[45]A nonlinear based speech feature analysis method with application to vocal fold pathology assessmentL. Gavidia-Ceballos , John H.L. Hansen, and James KaiserIEEE Transactions on Biomedical Engineering

[46]Segmentation, Structure Detection And Summarization of Multimedia Sequences Hari Sundaram Columbia University

[47]IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 1, JANUARY 2006
Automatic Segmentation and Identification of fixed-Language Speech Using Delta-BIC and LSA-Based GMMs Chung-Hsien Wu, Senior Member, IEEE, Yu-Hsien Chiu, Chi-Jiun Shia,

[48]Introduction to Machine Learning Draft Version 1.1.5 Thomas Hofmann

[49]Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and SpeakerVerification

[50]Dynamical behaviour of Lotka-Volterra systems,Jean-Luc Gouzé