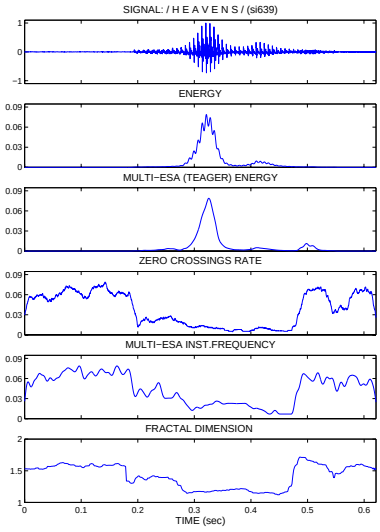
Σύγχρονες Μέθοδοι Ανάλυσης Σημάτων με Εφαρμογή στην Ανίχνευση Σιωπής – Φωνής – Θορύβου

Γεώργιος Ευαγγελόπουλος

Επιβλέπων: Καθηγητής Πέτρος Μαραγκός



Διπλωματική Εργασία

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Οκτώβριος 2001

θα ήθελα να ευχαριστήσω τον επιβλέποντα Καθηγητή μου κ. Π.Μαραγκό για την υποστήριξη, την καθοδήγηση και την προθυμία του να με βοηθήσει σε οποιοδήποτε πρόβλημα παρουσιάστηκε.

Επίσης τα παιδιά του CVS Lab (Βασίλη,Νατάσσα και Δημήτρη) για τις συμβουλές και τη βοήθεια.

Τους δικούς μου (που με αφήνουν να επιλέγω και με στηρίζουν)
Τα παιδιά του ΕΜΠ (για ότι περάσαμε...)
Τα παιδιά από 'πάνω'(δε χανόμαστε ποτέ...)
Τον αδερφό μου (για την υπομονή...)
Τη *Φωτεινή* (γιατί ήταν εδώ...)

Αθήνα
24 Οκτωβρίου 2001

Γιώργος Ευαγγελόπουλος

Περιεχόμενα

1 **Εισαγωγή** 3

2 **Η Κλασσική Μέθοδος Ανίχνευσης Φωνής (Αλγόριθμος Rabiner-Sambur)** 9

2.1 **Επεξεργασία Σημάτων Φωνής στο Πεδίο του Χρόνου** 10

2.1.1 Ενέργεια Βραχέως Χρόνου (Short-Time Energy) 12

2.1.2 Μέσο Πλάτος (Short-Time Average Magnitude) 12

2.1.3 Η Επίδραση του 'Παραθύρου' Ανάλυσης 13

2.1.4 Ενέργεια ή Πλάτος ; 15

2.1.5 Μέσος Ρυθμός 'Μεταβάσεων' από το Μηδέν (Short-Time Average Zero-Crossings Rate) 17

2.2 **Ένας Αλγόριθμος Προσδιορισμού των Αρχικών και Τελικών Σημείων Φωνής** 19

2.2.1 Προβλήματα Εντοπισμού των Endpoints 21

2.2.2 Ο Αλγόριθμος των Rabiner&Sambur - Μια υλοποίηση . . . 23

2.2.3 Παραδείγματα Χρήσης του Αλγορίθμου 26

2.2.4 Γενικά Συμπεράσματα και Παρατηρήσεις πάνω στην Απόδοση του Αλγορίθμου-Προοπτικές 27

2.3 **Ανακεφαλαίωση** 30

3 **Σύγχρονες Μη-Γραμμικές Τεχνικές Επεξεργασίας Σημάτων Φωνής** 32

3.1 **Ανίχνευση Ενέργειας και Αποδιαμόρφωση AM-FM σημάτων** 33

3.1.1 Ο Ενεργειακός Τελεστής Ψ (Teager-Kaiser Energy Operator). 33

3.1.2 Ανίχνευση Ενέργειας AM-FM σημάτων 34

3.1.3 Ένας Αλγόριθμος Διαχωρισμού της Ενέργειας (ESA) 37

3.2 **Μοντελοποίηση και Αποδιαμόρφωση Φωνητικών Σημάτων με βάση AM-FM μοντέλα** 42

3.2.1 Γραμμικά και Μη-Γραμμικά μοντέλα Παραγωγής Φωνής . 42

3.2.2 Ανάλυση και Αποδιαμόρφωση σε Πολλαπλές Ζώνες (MDA) 45

3.2.3 Μια MDA διαδικασία για την αποδιαμόρφωση σημάτων φωνής . . . 46
3.3 **Οι Νέες Απεικονίσεις στο Πεδίο του Χρόνου** . . . 53
3.3.1 Ενέργεια, Πλάτος Περιβάλλουσας και Στιγμιαία Συχνότητα για το συνολικό Σήμα . . . 53
3.3.2 Απεικονίσεις Βραχέως Χρόνου . . . 55
3.4 **Λίγα λόγια για την *Fractal* (κλασματική) διάσταση των Σημάτων** 58
3.5 **Ανακεφαλαίωση** . . . 60

4 Μια μέθοδος ανίχνευσης φωνής από σιωπή με βάση μη-γραμμικά εργαλεία 61
4.1 **Ποιοτική Σύγκριση Κλασσικών και Νέων Απεικονίσεων στο Χρόνο** . . . 62
4.2 **Ανίχνευση Φωνής από Σιωπή - Ένας αλγόριθμος βασισμένος στις μη-γραμμικές απεικονίσεις** . . . 72
4.2.1 Παραδείγματα εφαρμογής του αλγορίθμου σε απομονωμένες διαταραχές . . . 76
4.2.2 Συγκρίσεις με τον Κλασσικό Αλγόριθμο . . . 78
4.2.3 Εναλλακτικές υλοποιήσεις για ανίχνευση φωνής . . . 83
4.2.4 Γενικά Συμπεράσματα-Αποτελέσματα . . . 86
4.3 **Ανίχνευση Φωνής και Θόρυβος** . . . 87
4.3.1 Αντοχή της διαδικασίας ανίχνευσης σε θόρυβο . . . 88
4.3.2 Διάκριση Σιωπής, Φωνής, Θορύβου 'σε σειρά' . . . 90
4.4 **Ανακεφαλαίωση** . . . 93

5 Επίλογος 95

Κεφάλαιο 1

Εισαγωγή

'Η σιωπή είναι χρυσός'

Όχι πάντα ...

Σε σύγχρονα υπολογιστικά συστήματα όπου η φωνή χρησιμοποιείται ως εργαλείο επικοινωνίας ανάμεσα στον άνθρωπο και σε μια μηχανή, η παρουσία σιωπής πριν και μετά από φωνητικά σήματα θεωρείται ανασταλτικός παράγοντας. Η ανάγκη να χωριστεί ένα σύνθετο σήμα σε επιμέρους τμήματα και να απομονωθούν μόνο τα χρήσιμα από άποψη πληροφορίας διαστήματα του, ωθεί την ανάπτυξη και υλοποίηση μεθόδων *διάκρισης φωνής* (speech detection).

Η έννοια της σιωπής σε συστήματα επεξεργασίας φωνής και τηλεπικοινωνιακά ζητήματα έχει να κάνει με το περιβάλλον στο οποίο ηχογραφείται, παράγεται ή μεταφέρεται ένα φωνητικό σήμα. Το μικρής εντάσεως και θορυβώδους φύσεως σήμα που βρίσκεται στο 'φόντο' ενός τέτοιου περιβάλλοντος θεωρείται σιωπή (background noise). Εκτός από το ακουστικό αυτό περιβάλλον σιωπής διάφορα άλλα γεγονότα μη-φωνητικής φύσεως μπορεί να είναι παρόντα. Γεγονότα που έχουν να κάνουν με τον ομιλητή και τον τρόπο παραγωγής της φωνής όπως 'χτυπήματα' των χειλιών ή βαριά αναπνοή, καθώς και με το σύστημα μέσω του οποίου είναι πιθανόν να μεταφέρεται ένα φωνητικό σήμα όπως διαφωνία σε ενσύρματα δίκτυα, προϊόντα ενδοδιαμόρφωσης και τονικές παρεμβολές διαφόρων τύπων [9].

Ο εντοπισμός των χρονικών στιγμών στις οποίες αρχίζουν και τελειώνουν τα ακουστικά γεγονότα σημασίας ή διαφορετικά τα άκρα της φωνής (speech endpoints), σε μεμονωμένες λέξεις ή φράσεις, μπορεί να μειώσει την ποσότητα του υπό επεξεργασία σήματος αλλά και να απλοποιήσει εφαρμογές που απαιτούνε την επεξεργασία χρήσιμης πληροφορίας μόνο. Η ανάγκη για διάκριση φωνής παρουσιάζεται κυρίως σε συστήματα επεξεργασίας - αναγνώρισης φωνής αλλά και σε τηλεπικοινωνιακές εφαρμογές.

- Για αυτόματη αναγνώριση φωνής, η διάκριση είναι απαραίτητη έτσι

ώστε να απομονωθεί η φωνή και να δημιουργηθεί ένα φωνητικό πρότυπο (speech pattern or template). Πειραματικά δεδομένα αποδεικνύουνε τη σημασία του ακριβούς εντοπισμού, έτσι ώστε να δημιουργηθούν τα πιο κατάλληλα πρότυπα για αναγνώριση, καθώς και τη μείωση της απόδοσης που ενδεχόμενη λάθος τοποθέτηση των άκρων της φωνής μπορεί να προκαλέσει. Συγκεκριμένα για εφαρμογή αναγνώρισης ψηφίων από πολλαπλούς ομιλητές [13], απομάκρυνση από τα άκρα τα οποία εντοπίστηκαν μετά από παρατηρήσεις οδήγησε σε ομοιόμορφη μείωση της ακρίβειας αναγνώρισης. Για αναγνώριση σε ποσοστό 93%, λάθος και στα δύο άκρα άνω των 80ms οδηγεί σε μείωση πάνω από 20% στην απόδοση.

- Στις τηλεπικοινωνίες η ανάγκη για αποδοτική χρήση τόσο του εύρους ενός καναλιού όσο και του χρόνου που αυτό είναι κατειλημμένο απαιτεί τη γνώση του που και πότε υπάρχει ομιλία, δηλαδή χρήσιμη πληροφορία σε ένα μέσο. Για παράδειγμα , σε αναλογικά, πολυκαναλικά συστήματα μετάδοσης η τεχνική TASI (time assignment speech interpolation) εκμεταλλεύεται το χρόνο σιγής ενός καναλιού για να εξυπηρετήσει περισσότερους χρήστες, παραχωρώντας 'χρόνο' στο μέσο μόνο όταν ανιχνεύεται σιωπή. Επιτυγχάνεται έτσι μια αύξηση του αριθμού των χρηστών κατά έναν παράγοντα 2.5 (235 χρήστες σε 96 κανάλια φωνής).

Το πρόβλημα της ανίχνευσης φωνής είναι σχετικά απλό όταν έχουμε να κάνουμε με ιδανικές συνθήκες παραγωγής και σύλληψης της. Κάτι τέτοιο θα προϋπόθεται σταθερή και 'καθαρή' προφορά από τη μεριά του ομιλητή, ένα περιβάλλον με χαμηλό επίπεδο στατικού θορύβου (π.χ. ηχογράφηση σε ανηχοϊκό θάλαμο ή δωμάτιο με ηχομονωτικά τοιχώματα), όπως επίσης και ένα 'καθαρό' μέσο μεταφοράς της φωνής (π.χ. τηλεφωνικά σήματα πάνω από τοπικά PBXs). Για τέτοια περιβάλλοντα ο σηματοθορυβικός λόγος των σημάτων είναι μεγάλος και υπερβαίνει τα 30db rms. Για συστήματα μεταφοράς, ελεύθερα από παρεμβολές που συνήθως δημιουργούνε στατικό θόρυβο ο SNR κορυφής λαμβάνει τιμές (35-50)db.

Κάτω από πραγματικές όμως συνθήκες, στις οποίες λαμβάνουν χώρα οι περισσότερες εφαρμογές τα πράγματα δυστυχώς δεν είναι τόσο απλά. Τα πραγματικά δεδομένα αναιρούν τις προηγούμενες συνθήκες και σύνθετοι παράγοντες δυσκολεύουν τη διάκριση φωνής από ένα μη φωνητικό ακουστικό περιβάλλον. Τέτοιοι παράγοντες σχετίζονται και ευθύνονται για τα μη-φωνητικά γεγονότα που παρουσιάζονται μαζί με τη σιωπή και έχουν να κάνουν με τον ομιλητή καθώς και με το πιθανό σύστημα μέσω του οποίου μεταφέρεται η φωνή(όπως π.χ. διάφοροι θόρυβοι και ήχοι που παρουσιάζονται σε τυπικές τηλεφωνικές γραμμές).

Ο κυριότερος ίσως παράγοντας που δυσκολεύει τη διαδικασία ανίχνευσης φωνής σε ένα πρακτικό σύστημα είναι οι απρόβλεπτες συνθήκες τού περιβάλλοντος στο οποίο παράγεται. Διάφορες πηγές μπορούν να παράγουν ακουστικό θόρυβο και σήματα που παρεμβάλλουν. Έτσι για να ανιχνεύσει κανείς φωνή σε ένα τέτοιο περιβάλλον θα πρέπει να λάβει υπόψιν του τέτοιες πηγές θορύβου (μηχανές που δουλεύουν, θόρυβο πλήθους), μη στατικές διαταραχές (θόρυβος από αυτοκινητόδρομους) καθώς και παρεμβολές φωνής (από ραδιόφωνο, τηλεόραση, συζητήσεις στο παρασκήνιο). Αποδοτική και σθεναρή διάκριση φωνής από σιωπή κάτω από συνθήκες θορύβου απαιτείται σε εφαρμογές κινητών τηλεπικοινωνιών, πλοήγησης και αναγνώρισης εντολών σε αεροπλάνα και αυτοκίνητα κ.α.

Το πεδίο έχει βρει ιδιαίτερη ανταπόκριση ως ξεχωριστός κλάδος του τομέα της αναγνώρισης φωνής. Η ερευνητική δραστηριότητα είναι έντονη και διάφοροι αλγόριθμοι ανίχνευσης αρχής και τέλους της φωνής έχουν προταθεί. Οι μέθοδοι διάκρισης φωνής για χρήση σε συστήματα αναγνώρισης μπορούνε να χωριστούνε σε τρεις κατηγορίες ανάλογα με το βαθμό αλληλεπίδρασης τους με τη διαδικασία αναγνώρισης [9]:

1. *Η άμεση ή αποκλειστική,* στην οποία η ανίχνευση της φωνής γίνεται ανεξάρτητα από οποιαδήποτε διαδικασία αναγνώρισης. Η μέθοδος δίνει καλή ακρίβεια για φωνή με υψηλό σηματοθορυβικό λόγο αλλά αποτυγχάνει σε θορυβώδη περιβάλλοντα.

2. *Η έμμεση* , αντιμετωπίζει ταυτόχρονα τις διαδικασίες ανίχνευσης και αναγνώρισης. Το πρότυπο αναφοράς περιλαμβάνει και στοιχεία για τη σιωπή και η τελική απόφαση δίνει τόσο την πιο κατάλληλη λέξη όσο και τα αντίστοιχα άκρα της. Η μέθοδος αυτή αν και απαιτεί μεγαλύτερο υπολογιστικό φόρτο παρέχει μεγαλύτερη ακρίβεια από την άμεση προσέγγιση.

3. *Η υβριδική τεχνική,* χρησιμοποιεί την άμεση προσέγγιση για να εντοπίσει ένα σύνολο από πιθανές στιγμές και την έμμεση για να διαλέξει την πιθανή λέξη και τα αντίστοιχα άκρα της. Συνδυάζει έτσι την πολυπλοκότητα της πρώτης και την ακρίβεια της δεύτερης τεχνικής.

Ο βασικός αλγόριθμος διάκρισης φωνής προτάθηκε από τους Rabiner και Sambur [10]. Περιλαμβάνει επεξεργασία μιας διαταραχής (συνήθως απομονωμένης λέξης σε ένα ακουστικό περιβάλλον σιωπής) στο πεδίο του χρόνου και εξαγωγή χαρακτηριστικών μετρήσεων στις οποίες στη συνέχεια βασίζεται η απόφαση για τα άκρα της φωνής. Συγκεκριμένα χρησιμοποιούνται η ενέργεια βραχέως χρόνου(short-time energy) καθώς ο αριθμός περασμάτων του σήματος από το μηδέν στη μονάδα του χρόνου (average zero-crossings rate). Ο αλγόριθμος

είναι κατά κάποιο τρόπο αυτοπροσαρμοζόμενος στο ακουστικό περιβάλλον της φωνής αφού αποκτά τα σχετικά κατώφλια για τα κριτήρια απόφασης από μετρήσεις που γίνονται κατευθείαν στο πραγματικό διάστημα ηχογράφησης. Είναι σχεδιασμένος για να αποδίδει καλά σε περιβάλλον με σηματοθορυβικό λόγο μεγαλύτερο από 30db και να ελαχιστοποιεί τα μεγάλα σφάλματα ανίχνευσης (άνω των 50ms).

Προσπάθειες γίνανε στη συνέχεια για αλγόριθμους που θα αποδίδανε καλά τόσο σε 'καθαρό' όσο και σε απρόβλεπτα θορυβώδες περιβάλλον. Οι αλγόριθμοι αυτοί ανήκουνε στην κατηγορία υβριδικών μεθόδων. Χρησιμοποιώντας ξανά τη μέτρηση της ενέργειας βραχέως χρόνου και εξισορροπώντας τη σε σχέση με το επίπεδο σιωπής,αναζητώνται ενεργειακοί παλμοί που μπορεί να αποτελούν ένδειξη φωνής, στο διάστημα της λέξης ή φράσης. Οι υποψήφιοι αυτοί παλμοί κατατάσσονται με βάση κανόνες λογικής για να εκτιμηθούν τα πιθανότερα ζευγάρια αρχής και τέλους.

Σε αυτό το πνεύμα η Lamel κ.α. [4] πρότεινε έναν αλγόριθμο που αναζητεί πιθανά endpoints βασιζόμενος στην ενέργεια φωνής που ανέρχεται πάνω από ένα καθορισμένο επίπεδο από την ενέργεια της σιωπής(background noise). Η ενέργεια (ή καλύτερα ο λογάριθμος της) ελέγχεται αφού αφαιρεθεί αρχικά το μέσο επίπεδο θορύβου που υπολογίζεται μετά από στατιστική επεξεργασία του ενεργειακού σήματος. Κανόνες λογικής και περιορισμοί ως προς τη διάρκεια των εντοπιζόμενων παλμών, χρησιμοποιούνται για την ταξινόμηση και την τελική επιλογή των ορίων της φωνής. Η προσέγγιση αυτή καλείται "από κάτω προς τα πάνω" προσέγγιση (bottom-up approach) και ο αλγόριθμος χρησιμοποιήθηκε και χρησιμοποιείται για εφαρμογές πραγματικού χρόνου [13].

Ο Wilpon κ.α. [13] παρουσίασε μια "από πάνω προς τα κάτω" προσέγγιση (top-down approach), με έναυσμα τη δημιουργία ενός βελτιωμένου αλγορίθμου ανίχνευσης φωνής, μεγάλης ακριβείας σε περιβάλλον έντονα μεταβλητού θορύβου τόσο σε ένταση όσο και σε φασματικό περιεχόμενο. Τέτοιες συνθήκες συναντώνται συνήθως σε δημόσια τηλεφωνικά δίκτυα λόγω των χαρακτηριστικών των γραμμών μεταφοράς, τονικοτήτων κλπ. Ο αλγόριθμος μοιάζει με τον προηγούμενο στο ότι υπολογίζει την κανονικοποιημένη ενέργεια, εντοπίζει τους παλμούς και τους συνδυάζει για την επιλογή των τελικών σημείων. Διαφέρει στο ότι ανιχνεύει το μέγιστο ενός ενεργειακού παλμού και στη συνέχεια αναζητεί πιθανά endpoints γύρω από αυτή την κορυφή. Πρόσθετες παράμετροι όπως η κλίση των παλμών χρησιμοποιούνται, ενώ για να προκύψουν τα τελικά σημεία εμπλέκονται συντακτικοί(λεξιλόγιο) αλλά και σημαντικοί(σχετικά με την εφαρμογή) περιορισμοί. Πειραματικά αποτελέσματα για εφαρμογή ανίχνευσης ψηφίων, κάτω από τηλεφωνικές συνθήκες έδειξαν ακρίβεια αναγνώρισης κοντά στο 90% .

Οι αλγόριθμοι που χρησιμοποιήθηκαν εντόπισαν το πρόβλημα της διάκρισης της φωνής από ένα ακουστικό περιβάλλον, στην επιλογή χαρακτηριστικών μετρήσεων που διαφέρουν για τα σήματα φωνής και σιωπής. Προσπάθειες για να αυξηθεί η ευαισθησία τόσο σε ομαλά όσο και σε απρόβλεπτα περιβάλλοντα γίνανε και γίνονται, είτε εισάγοντας περισσότερες παραμέτρους (π.χ. συντελεστές πρόβλεψης) είτε αυξάνοντας την πολυπλοκότητα της διαδικασίας επιλογής των πιθανών endpoints (π.χ. εξειδικεύοντας ανάλογα με την εφαρμογή).

Ένα σημαντικό γεγονός που φαίνεται να 'προσπερνάται' είναι το ότι στην ουσία τα σήματα φωνής, σιωπής αλλά και πιθανού θορύβου είναι από τη φύση τους διαφορετικά. Επιπλέον σύγχρονες έρευνες αποδεικνύουν την ύπαρξη μη γραμμικών φαινομένων και διαμορφώσεων κατά τη διαδικασία παραγωγής της φωνής.

Ο Teager [12], σε μια προσπάθεια για να μοντελοποιήσει την παραγωγή φωνής , ανέπτυξε έναν *ενεργειακό τελεστή* ο οποίος παρουσιάστηκε από τον Kaiser [3] ως ένας νέος τρόπος μέτρησης της ενέργειας της πηγής που παράγει το σήμα. Για απλά ημιτονικά σήματα αυτή η ενέργεια είναι το γινόμενο των τετραγώνων του πλάτους και της συχνότητας.

$$\Psi[A\cos(\omega t+\theta)] = (A\omega)^2$$

Εξετάζοντας πλέον την ανάλυση σημάτων από την πλευρά της ενέργειας που χρειάζεται για να δημιουργηθούνε, ο Maragos κ.α. [6] χρησιμοποίησαν αυτόν τον μη-γραμμικό τελεστή της ενέργειας Ψ και ανέπτυξαν έναν αλγόριθμο διαχωρισμού της ενέργειας AM-FM διαμορφωμένων σημάτων , σε στιγμιαίο πλάτος και συχνότητα (Energy Seperation Algorithm). Επιπρόσθετα με βάση ενδείξεις για την ύπαρξη διαμορφώσεων στα σήματα φωνής (όπως ασταθής και διαχωρίσιμη ροή του αέρα στη φωνητική οδό, στρόβιλοι που διαμορφώνουν την ενέργεια του αέρα κ.α.) [6], ένα μικρό τμήμα φωνής μπορεί να μοντελοποιηθεί σαν υπέρθεση AM-FM σημάτων (speech resonances).

Σκοπός αυτής της εργασίας είναι η εισαγωγή αυτών των νέων εργαλείων που βρίσκονται πλέον στη διάθεση μας, στην αναζήτηση νέων τεχνικών διάκρισης φωνής. Οι ως τώρα τεχνικές βασίζονται σε γραμμική μοντελοποίηση της φωνής και στην ενέργεια που 'κουβαλάει' το ίδιο το σήμα. Στόχος είναι να εξετάσουμε πόσο διαφορετικά αποτελέσματα δίνουν αυτά τα νέα εργαλεία και πόση πληροφορία σχετικά με τη φύση της φωνής, της σιωπής και του θορύβου.

Αρχίζοντας από χαμηλά, προτείνεται μία νέα μέθοδος διάκρισης, η οποία χρησιμοποιεί τα νέα αυτά μη - γραμμικά εργαλεία (ενεργειακό τελεστή, στιγμιαίο αποδιαμορφωμένο πλάτος - συχνότητα), βασίζεται όμως στη λογική των στατιστικών κατωφλιών του βασικού αλγορίθμου των Rabiner-Sambur. Συγκεκριμένα, αφού το υπό εξέταση τμήμα φιλτράρεται από μια τράπεζα ζωνοπερατών

Gabor φίλτρων έτσι ώστε να απομονωθούν οι πιο ισχυρές συνιστώσες του, εφαρμόζεται ο ESA σε κάθε έξοδο της τράπεζας (Multiband Demodulation Analysis). Η ενεργειακά πιο ισχυρή δίνει ανά χρονική στιγμή τη μέτρηση της ενέργειας, του στιγμιαίου πλάτους και της στιγμιαίας συχνότητας. Έτσι λαμβάνονται τρεις νέες μετρήσεις, οι οποίες μετά από μια διαδικασία παραθυροποίησης χρησιμοποιούνται αντί για την ενέργεια βραχέως χρόνου και τον μέσο ρυθμό zero-crossings. Οι διακυμάνσεις τους για τα σήματα φωνής και σιωπής καθορίζουν τη διάκριση μεταξύ τους.

Για να εκτιμηθεί η συμβολή των νέων εργαλείων απαραίτητη ήταν η σύγκριση με τα ήδη υπάρχοντα. Έτσι ο βασικός αλγόριθμος χρησιμοποιήθηκε ως μέτρο σύγκρισης στο κατά πόσο βελτιώνεται η ικανότητα διάκρισης της φωνής. Αναμφίβολα ο καλύτερος έλεγχος μπορεί να γίνει κάτω από συνθήκες αναγνώρισης των εντοπισμένων λέξεων. Κάτι τέτοιο όμως ξεφεύγει από τα πλαίσια της παρούσας εργασίας. Έτσι η σύγκριση διατηρείται μόνο σε επίπεδο εκτίμησης των χρονικών στιγμών αρχής και τέλους των λέξεων, με βάση τα εκ των προτέρων γνωστά πραγματικά σημεία. Χρησιμοποιείται γι'αυτό το σκοπό ένα σύνολο φράσεων από την βάση TIMIT.

Σύγκριση και έλεγχος γίνεται επίσης και για σήματα στα οποία έχει προστεθεί θόρυβος για διάφορα επίπεδα σηματοθορυβικού λόγου. Επίσης με βάση τα νέα εργαλεία, επιχειρείται και μια διάκριση θορύβου 'σε σειρά' με τα σήματα φωνής και σιωπής. Για το σκοπό αυτό άλλο ένα μη-γραμμικό εργαλείο χρησιμοποιείται. Η fractal διάσταση των σημάτων, που εκφράζει τη γεωμετρική πολυπλοκότητα των κυματομορφών τους, εκτιμάται για να χαρακτηρίσει ένα θορυβώδες σήμα.

Η διάρθρωση του κειμένου που ακολουθεί έχει ως εξής:

Στο **Κεφάλαιο 2** περιγράφονται τα εργαλεία που χρησιμοποιούνται συνήθως και που αφορούνε επεξεργασία του σήματος στο πεδίο του χρόνου. Γίνεται αναφορά στις περιπτώσεις που δημιουργούν πρόβλημα στη διαδικασία διάκρισης. Περιγράφεται ο αλγόριθμος των Rabiner&Sambur [10] και δίνονται παραδείγματα εφαρμογής του στην ανίχνευση μεμονωμένων λέξεων.

Στο **Κεφάλαιο 3** περιγράφονται αναλυτικά τα νέα μη-γραμμικά εργαλεία, ο ενεργειακός τελεστής και ο ESA, η διαδικασία αποδιαμόρφωσης σε πολλαπλές μπάντες (MDA) καθώς και τα νέα προς μέτρησιν μεγέθη που προκύπτουν μετά το φιλτράρισμα του σήματος.

Στο **Κεφάλαιο 4** εισάγεται ο νέος αλγόριθμος και δοκιμάζεται σε μεμονωμένες λέξεις. Συγκρίνεται με τον αλγόριθμο του Rabiner για ένα σύνολο λέξεων, ενώ δοκιμάζονται διάφοροι συνδυασμοί των νέων εργαλείων που θα δώσουν μεγαλύτερη ακρίβεια.Τέλος μελετάται η ευστάθεια των δύο αλγορίθμων κάτω από συνθήκες θορύβου αλλά και γίνεται μια προσπάθεια διάκρισης τριών διαφορετικών ειδών σημάτων (φωνής - σιωπής - θορύβου) με χρήση

συνδυασμού γραμμικών και μη – γραμμικών εργαλείων.

Στο **Κεφάλαιο 5** δίνονται συμπεράσματα της όλης έρευνας και γίνεται μια αναφορά σε προοπτικές για περαιτέρω ανάπτυξη και μελέτη των νέων τεχνικών στο πεδίο της διάκρισης φωνής.

Κεφάλαιο 2

Η Κλασσική Μέθοδος Ανίχνευσης Φωνής (Αλγόριθμος Rabiner-Sambur)

Ένα από τα πρώιμα βήματα στο πεδίο της ανίχνευσης φωνής έγινε από τους Rabiner-Sambur στο [10] με την ενεργειακή προσέγγιση που εφάρμοσαν στο πρόβλημα. Έκτοτε η συγκεκριμένη προσέγγιση χρησιμοποιήθηκε σαν αναφορά σε οποιαδήποτε προσπάθεια βελτίωσης ενώ τα εργαλεία και η λογική της μεθόδου παραμένουν στο προσκήνιο αφού εμφανίζονται σε κάθε μοντέρνο σύστημα ανίχνευσης.

Ο αλγόριθμος σχεδιάστηκε για ανίχνευση μεμονωμένων λέξεων που βρίσκονται σε ένα ακουστικό περιβάλλον. Κάθε τέτοιο σήμα ονομάζεται συμβατικά 'διαταραχή'. Βασίστηκε σε δύο μετρήσεις, την ενέργεια βραχέως χρόνου(short-time energy) και τον μέσο ρυθμό περασμάτων από το μηδέν(average zero-crossings rate) και σε στατιστικά κατώφλια σιωπής παρμένα από ένα τμήμα από την αρχή του σήματος το οποία θεωρείται σιωπή. Η λογική είναι η αναζήτηση σημείων στα οποία η ενέργεια του σήματος ανέρχεται πάνω από ένα καθορισμένο επίπεδο και τα οποία χαρακτηρίζουν την ύπαρξη φωνής. Η έναρξη και λήξη της λέξης ελέγχονται ακόμη περισσότερο με τη μέτρηση του ρυθμού zero-crossings για τις περιπτώσεις συριστικών φωνημάτων ή παύσεων στην αρχή ή στο τέλος της.

Οι μετρήσεις αυτές προϋποθέτουν επεξεργασία του σήματος στο πεδίο του χρόνου(time-domain processing) και αποτελούν γραμμικά εργαλεία υπό την έννοια ότι η εφαρμογή τους στο σήμα ισοδυναμεί με την έξοδο ενός γραμμικού συστήματος με την κατάλληλη χρονική απόκριση. Πρόκειται για χρήσιμες απεικονίσεις, εύκολα υλοποιήσιμες που παρέχουν πληροφορία για σημαντικά χαρακτηριστικά του σήματος φωνής.

Το κεφάλαιο ξεκινάει με μια ανασκόπηση της επεξεργασίας φωνητικών σημάτων στο πεδίο του χρόνου και των γραμμικών εργαλείων που χρησιμοποιούνται στον βασικό αλγόριθμο. Στη συνέχεια γίνεται αναφορά στα προβλήματα

που σχεδιάστηκε για να λύσει ο αλγόριθμος, και αφορούνε φωνήματα που δεν εντοπίζονται στην αρχή και στο τέλος λέξεων. Περιγράφεται ο αλγόριθμος και παρουσιάζονται αποτελέσματα προσομοίωσης από τη χρήση του σε ανίχνευση μεμονωμένων λέξεων. Τέλος γίνεται μια γενική εκτίμηση της απόδοσης του και επισημαίνονται τα μειονεκτήματα αλλά και τα μονοπάτια που άνοιξε.

2.1 Επεξεργασία Σημάτων Φωνής στο Πεδίο του Χρόνου

Η επεξεργασία ενός φωνητικού σήματος γίνεται με σκοπό να αποκτηθεί μια πιο χρήσιμη και βολική απεικόνιση η οποία να τονίζει ή να απορρίπτει κάποια χαρακτηριστικά του σήματος. Ανάλογα με την εφαρμογή οι απεικονίσεις αυτές επιλέγονται έτσι ώστε να κρατάνε σχετική μόνο πληροφορία, κάνοντας τα επιθυμητά χαρακτηριστικά εμφανή και εκμεταλλεύσιμα. Στην περίπτωση μας ο σκοπός της επεξεργασίας είναι να εξαχθεί από την κυματομορφή ενός σήματος πληροφορία σχετικά με τη φύση του (δηλ. φωνή, σιωπή ή ακόμη και θόρυβος).

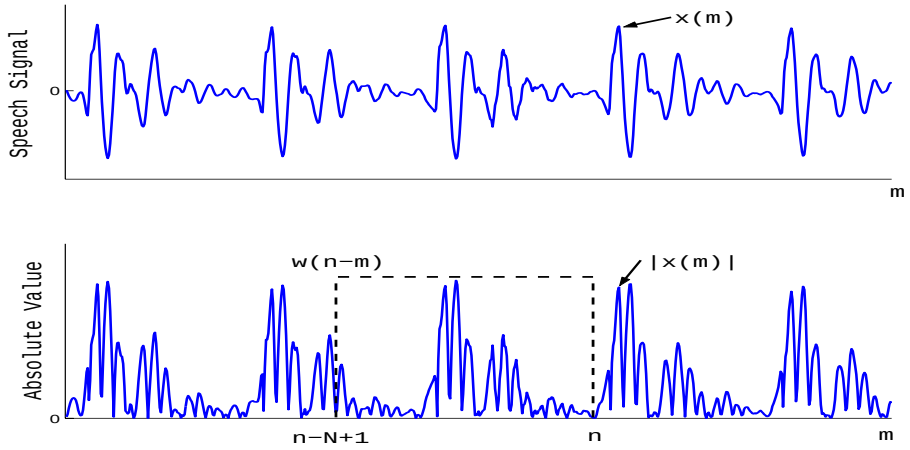
Οι τεχνικές που περιλαμβάνουν επεξεργασία της κυματομορφής του σήματος απευθείας ονομάζονται *τεχνικές στο πεδίο του χρόνου*. Για ψηφιακή επεξεργασία η κυματομορφή είναι στην ουσία ένας αριθμός από διαδοχικά δείγματα του αναλογικού σήματος.[1] Σε ένα φωνητικό σήμα, η κυματομορφή μεταβάλλεται με το χρόνο ως προς το πλάτος των κορυφών της, η διέγερση διαφέρει ανάμεσα στα έμφωνα και άφωνα τμήματα καθώς και η θεμελιώδης συχνότητα (pitch) μέσα σε ένα έμφωνο τμήμα. Με απλές τεχνικές αυτές οι μεταβολές μπορούν να δώσουν χρήσιμες απεικονίσεις ορισμένων χαρακτηριστικών του σήματος.

Μια κατηγορία τέτοιων τεχνικών είναι οι *μέθοδοι επεξεργασίας 'βραχέως χρόνου'* (short-time methods) που βασίζονται στην παρατήρηση ότι οι ιδιότητες της φωνής μεταβάλλονται σχετικά αργά με το χρόνο. Μικρά τμήματα του σήματος, που συνήθως ονομάζονται *πλαίσια ή παράθυρα ανάλυσης* (analysis frames) απομονώνονται και επεξεργάζονται ξεχωριστά. Η διαδικασία επαναλαμβάνεται για όλο το σήμα, με περιοδικά συνήθως 'άλματα' των πλαισίων έτσι ώστε να καλυφθεί ολόκληρο. Τελικά παράγεται μια νέα χρονική ακολουθία αριθμών,που προκύπτουν από την επεξεργασία του κάθε πλαισίου, η οποία αποτελεί μια νέα απεικόνιση στο χρόνο του σήματος φωνής.

Η λειτουργία των short-time τεχνικών περιγράφεται μαθηματικά από τη σχέση

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]w(n \quad m) \tag{2.1}$$

[1] Για σήματα φωνής που διαθέτουν εύρος συχνοτήτων 4kHz, το θεώρημα δειγματοληψίας του Nyquist καθορίζει συχνότητα δειγματοληψίας τουλάχιστον 8kHz (8000 samples/sec). Σε υπολογιστικές εφαρμογές χρησιμοποιείται συχνότητα 16kHz για μεγαλύτερη ακρίβεια.



Σχήμα 2.1: Υπολογισμός μιας short-time απεικόνισης

η οποία ερμηνεύεται ως εξής: Το σήμα της φωνής μετασχηματίζεται μέσω του τελεστή $T[\ ]$, ο οποίος μπορεί να είναι γραμμικός ή μη γραμμικός και η ακολουθία που προκύπτει πολλαπλασιάζεται με ένα παράθυρο $w(\ )$, συνήθως πεπερασμένης διάρκειας τοποθετημένο στο δείγμα με δείκτη n. Το γινόμενο αθροίζεται στη συνέχεια για όλες τις τιμές του σήματος. Οι τιμές του Q_n αποτελούν στην ουσία ένα σύνολο τοπικών μέσων όρων του $T[x(m)]$.

Έστω για παράδειγμα ο μετασχηματισμός $T[x] = |x|$. Η short-time απεικόνιση της απόλυτης τιμής του σήματος ή το μέσο πλάτος του δίνεται από τη σχέση

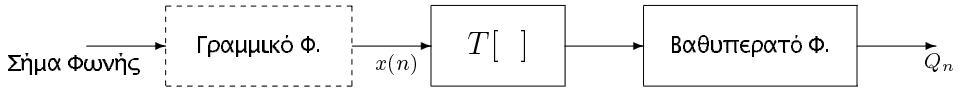
$$A_n = \sum_{m=n\ \ N+1}^{n} |x(m)|$$

δηλαδή η τιμή της στο n είναι το άθροισμα των απολύτων τιμών των δειγμάτων $n\ \ N+1$ έως n. Σύμφωνα με την γενική έκφραση της Εξ. (2.1) το παράθυρο ανάλυσης είναι

$$\begin{aligned} w(n) &= 1 \qquad 0 \le n \le N\ \ \ 1 \\ &= 0 \qquad \textit{αλλού} \end{aligned}$$

Το Σχήμα 2.1 απεικονίζει τον παραπάνω υπολογισμό και την κίνηση του παραθύρου ώστε να επιλεγεί το ανάλογο τμήμα του σήματος.

Μια σημαντική ιδιότητα των απεικονίσεων βραχέως χρόνου είναι η ομοιότητα που παρουσιάζουν με την απόκριση γραμμικών συστημάτων. Πράγματι, η Εξ. (2.1) παριστάνει τη συνέλιξη του $T[x(n)]$ με ένα παράθυρο $w(n)$, δηλαδή την έξοδο ενός γραμμικού χρονικά-αμετάβλητου συστήματος με κρουστική απόκριση $h(n) = w(n)$. Για ένα σήμα φωνής πιθανώς να προηγείται και



Σχήμα 2.2: Αρχή των τεχνικών *'short-time'* απεικονίσεων

ένα στάδιο γραμμικού φιλτραρίσματος ώστε να απομονωθεί ένα επιθυμητό εύρος συχνοτήτων Αυτή η άποψη των διαδικασιών αυτών φαίνεται στο block διάγραμμα του Σχήματος (2.2).

Στη συνέχεια θα αναφερθούμε σε τρεις απεικονίσεις βραχέως χρόνου οι οποίες χρησιμοποιούνται ευρύτατα ως εργαλεία ανίχνευσης φωνής. Πρόκειται για την *ενέργεια*, το *πλάτος* και το *μέσο ρυθμό zero-crossings*.

2.1.1 Ενέργεια Βραχέως Χρόνου (Short-Time Energy)

Οι μεταβολές του πλάτους σε ένα σήμα φωνής, ειδικά κατά την μετάβαση από τα άφωνα στα έμφωνα τμήματα του απεικονίζονται αισθητά με τη βοήθεια της μέτρησης της ενέργειας. Με τον όρο ενέργεια εννοούμε συμβατικά την ενέργεια που μεταφέρει το σήμα με το χρόνο. Η ενέργεια ενός διακριτού σήματος ορίζεται ως $E_n = \sum_{m=\ \infty}^{\infty} x^2(m)$, αλλά η πληροφορία που δίνει για τις χρονικές μεταβολές του δεν είναι βολική.
Έτσι ορίζεται η ενέργεια βραχέως χρόνου δηλ.

$$E_n = \sum_{m=\ \infty}^{\infty} [x(m)w(n \quad m)]^2 \tag{2.2}$$

Γράφοντας

$$h(n) = w^2(n)$$

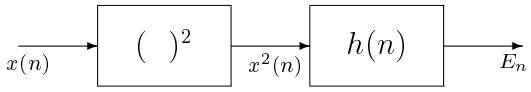
η σχέση (2.2) έρχεται στη γενική μορφή της Εξ. (2.1)

$$E_n = \sum_{m=\ \infty}^{\infty} x^2(m)h(n \quad m) \tag{2.3}$$

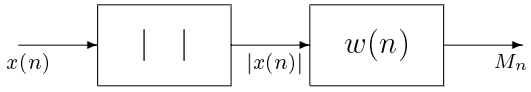
Το $h(n)$ είναι δηλαδή η κρουστική απόκριση του φίλτρου που φαίνεται στο Σχήμα (2.3). Η επιλογή του είδους και του μήκους του παραθύρου $h(n)$ καθορίζουν και τη φύση της απεικόνισης. Μπορεί να είναι οποιοδήποτε βαθυπερατό φίλτρο, αρκεί να λειαίνει επαρκώς το επιλεγμένο τμήμα ενώ το κατάλληλο μήκος καθορίζεται με βάση πρακτικές απαιτήσεις.

2.1.2 Μέσο Πλάτος (Short-Time Average Magnitude)

Μια εναλλακτική απεικόνιση που 'αιχμαλωτίζει' πληροφορία σχετικά με τις χρονικές μεταβολές του πλάτους ή της έντασης του σήματος φωνής επιτυγχά-



Σχήμα 2.3: Block διάγραμμα της Short-Time Ενέργειας



Σχήμα 2.4: Block διάγραμμα του Short-Time Average Magnitude

νεται με τη μέτρηση του μέσου πλάτους. Η συνάρτηση υπολογισμού του ορίζεται ως εξής

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)| w(n - m) \tag{2.4}$$

Η εφαρμογή της Εξ. (2.4) στο σήμα και η ισχύς του γενικού κανόνα υλοποίησης της ως γραμμικό φιλτράρισμα παρουσιάζεται στο Σχήμα (2.4).

Η επίδραση του παραθύρου $w(n)$ στον υπολογισμό εξαρτάται, όπως και για την ενέργεια, από το μήκος και το είδος του. Τα φαινόμενα αυτά θα συζητηθούν σύντομα στη συνέχεια.

2.1.3 Η Επίδραση του 'Παραθύρου' Ανάλυσης

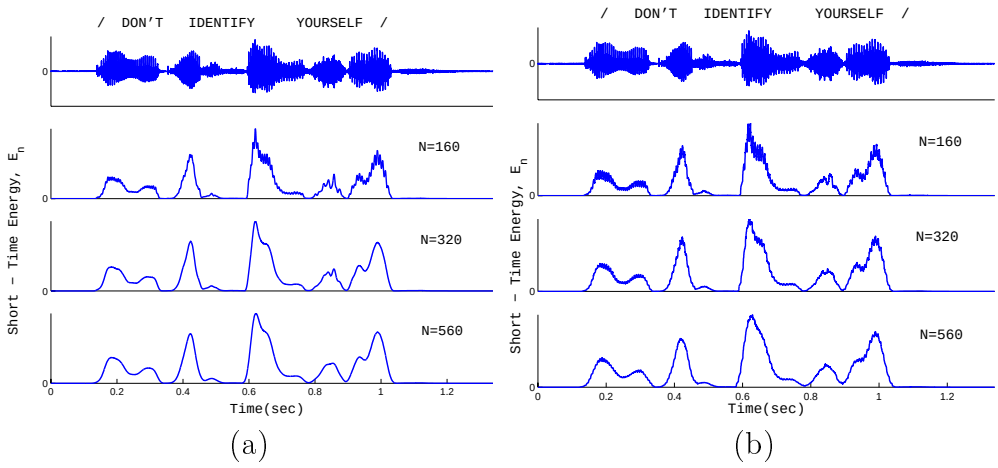
Το παράθυρο, $w(n)$ ή $h(n)$, που χρησιμοποιείται για την ανάλυση και τον υπολογισμό των short-time μεγεθών χρειάζεται να είναι μεν σύντομο σε διάρκεια ώστε να ανταποκρίνεται σε απότομες αλλαγές του πλάτους του σήματος αλλά να έχει και αρκετό μήκος ώστε να παρέχει επαρκή τοπική εξισορρόπηση των τιμών για ομαλές απεικονίσεις.

Δύο αντιπροσωπευτικά παράθυρα που χρησιμοποιούνται συχνά στη ψηφιακή επεξεργασία σημάτων και σε μεθόδους φασματικής ανάλυσης είναι το τετραγωνικό

$$\begin{aligned} h(n) &= 1 \qquad 0 \le n \le N-1 \\ &= 0 \qquad \textit{αλλού} \end{aligned} \tag{2.5}$$

που εφαρμόζει το ίδιο βάρος στα δείγματα που εμπλέκονται στον υπολογισμό, όπως είδαμε και στο Σχήμα (2.1),
και το παράθυρο Hamming

$$\begin{aligned} h(n) = 0.54 - 0.46\cos(2\pi n/(N-1)) &\qquad 0 \le n \le N-1 \\ = 0 &\qquad \textit{αλλού} \end{aligned} \tag{2.6}$$



Σχήμα 2.5: Short-Time Ενέργεια για παράθυρα ανάλυσης διαφορετικού μήκους και τύπου (a) Hamming (b) Τετραγωνικό

που λειαίνει το σήμα και ελαχιστοποιεί τα φαινόμενα άκρων στο τμήμα που επιλέγεται.

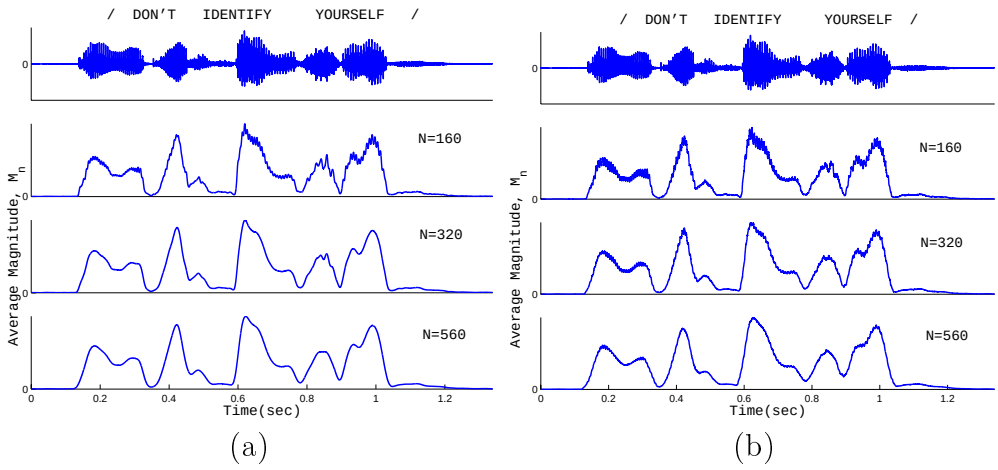
Αποτελούν στην ουσία βαθυπερατά φίλτρα που αποσβένουν το σήμα έξω από μια ζώνη συχνοτήτων. Το εύρος του παραθύρου Hamming είναι σχεδόν διπλάσιο ενός τετραγωνικού παραθύρου και η απόσβεση που δίνει έξω από τη βαθυπερατή ζώνη είναι μεγαλύτερη. Αυξάνοντας το μήκος N ενός παραθύρου μειώνεται το εύρος ζώνης του. Αν το N είναι μικρό π.χ. της τάξεως μίας περιόδου pitch ή λιγότερο, η ενέργεια ή το πλάτος θα μεταβάλλεται ταχύτατα ακολουθώντας τις λεπτομέρειες της κυματομορφής του σήματος. Αν το N είναι μεγάλο, π.χ. αρκετών περιόδων pitch, οι short-time απεικονίσεις θα αλλάζουν πολύ αργά και δεν θα καταγράφουν τις ιδιότητες του σήματος.

Στο Σχήμα (2.5) φαίνεται η ενέργεια για τα δύο παράθυρα ανάλυσης με τρία διαφορετικά μήκη. Έτσι για συχνότητα δειγματοληψίας 16kHz[2], παρουσιάζονται παράθυρα των 10ms ($N = 160$), 20ms ($N = 320$) και 35ms ($N = 560$), στη διαταραχή / don't identify yourself / από άντρα ομιλητή.

Παρόμοια στο Σχήμα (2.6) φαίνονται τα διαφορετικά μήκη για τα δύο παράθυρα, αλλά για την απεικόνιση του μέσου πλάτους.

Η τιμή του παραθύρου που τελικά επιλέγεται προκύπτει από παρατήρηση και πρακτικούς περιορισμούς. Επειδή η διάρκεια του pitch μεταβάλλεται από 2ms (στα 16kHz συχνότητα δειγματοληψίας) για παιδική ή γυναικεία φωνή με υψηλό pitch , έως 25ms για αντρική φωνή χαμηλού pitch, μια καλή επιλογή που δίνει και καλά πειραματικά αποτελέσματα είναι η επιλογή παραθύρου

[2]Ως συχνότητα δειγματοληψίας για όλα τα πειράματα και παραδείγματα χρησιμοποιήθηκε η F_s =16kHz. Έτσι από δω και πέρα θα θεωρείται καθορισμένη στα 16kHz, όπου αναφέρεται.



Σχήμα 2.6: Μέσο Πλάτος για παράθυρα ανάλυσης διαφορετικού μήκους και τύπου (a) Hamming (b) Τετραγωνικό

μήκους 10-20ms (160-320 δείγματα).

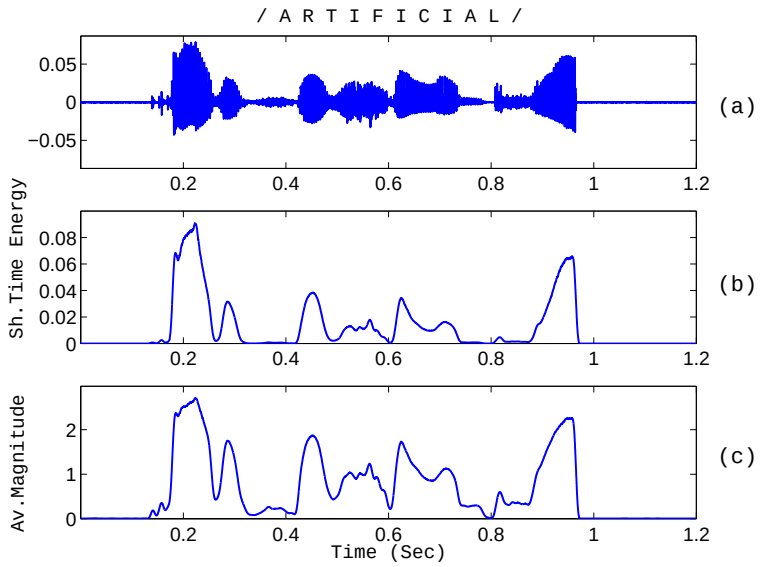
Να σημειωθεί τέλος ότι για short-time απεικονίσεις δε χρειάζεται να περιοριζόμαστε σ'αυτούς τους δύο τύπους παραθύρων μόνο. Οποιοδήποτε παράθυρο που μπορεί να δράσει ως βαθυπερατό φίλτρο μπορεί να χρησιμοποιηθεί(π.χ. Hanning, Kaiser κ.α.), όπως επίσης και παράθυρα μη-πεπερασμένου μήκους, αρκεί η απόκριση συχνότητας τους να έχει την επιθυμητή βαθυπερατή ιδιότητα.

2.1.4 Ενέργεια ή Πλάτος ;

Η βασική ιδιότητα της ενεργειακής απεικόνισης είναι η ικανότητα της να διακρίνει τα έμφωνα από τα άφωνα τμήματα του λόγου. Από το Σχ. (2.5) φαίνεται ότι η τιμή της ενέργειας είναι πολύ μικρότερη για τα άφωνα τμήματα. Έτσι μπορεί να χρησιμοποιηθεί για τον εντοπισμό των χρονικών στιγμών που υπάρχει μετάβαση από τον ένα στον άλλο ηχητικό τύπο.

Η ευαισθησία της ενέργειας σε μεγάλα επίπεδα του σήματος, μπορεί να δημιουργήσει προβλήματα και υπερεκτιμήσεις μεγεθών αφού μεγάλες μεταβολές από δείγμα σε δείγμα ενισχύονται με τη χρήση του τετραγώνου στην Εξ. (2.3). Αντίθετα χρησιμοποιώντας το μέσο πλάτος, όπου οι υπολογισμοί γίνονται με την απόλυτη τιμή του σήματος τα προβλήματα υπερευαισθησίας περιορίζονται ενώ απλοποιείται υπολογιστικά και η διαδικασία.

Στο Σχ. (2.6) φαίνονται οι διαφορές σε σχέση με την ενέργεια και ειδικά η μείωση της απόστασης ανάμεσα στα έμφωνα και στα άφωνα τμήματα.



Σχήμα 2.7: Απεικονίσεις Βραχέως Χρόνου με παράθυρο Hamming 15ms στα 16kHz (a) Κυματομορφή της λέξης /artificial/ (b) Ενέργεια (c) Μέσο Πλάτος

Ισχύει προσεγγιστικά ότι

$$\frac{max(M_n)}{min(M_n)} \quad \sqrt{\frac{max(E_n)}{min(E_n)}}$$

και επομένως η διαφορά επιπέδων τιμών είναι μεγαλύτερη για την ενέργεια όσον αφορά τη διάκριση άφωνων και έμφωνων γεγονότων. Στο Σχ. (2.7) φαίνονται οι δύο απεικονίσεις για τη λέξη / artificial / από γυναικεία φωνητικά. Για τους υπολογισμούς χρησιμοποιήθκε παράθυρο Hamming, μήκους 240 δειγμάτων (15 ms) με συχνότητα δειγματοληψίας 16kHz.

Μια χρήσιμη παρατήρηση, όσον αφορά τη διαδικασία υπολογισμού των δύο μεγεθών έχει να κάνει με την ποσότητα πληροφορίας που απαιτούνε. Συγκεκριμένα ο ρυθμός δειγματοληψίας στην έξοδο των συστημάτων στα Σχ. (2.3), (2.4) δεν χρειάζεται να είναι όσο αυτός της εισόδου. Το εύρος των δύο μετρήσεων είναι αυτό του βαθυπερατού παραθύρου που χρησιμοποιείται για τον υπολογισμό τους και έτσι η συχνότητα δειγματοληψίας τους μπορεί να είναι πολύ μικρότερη από αυτή του σήματος. Για παράθυρο μήκους N δειγμάτων αρκεί $f_s \quad 2F_s/N$. Αυτό επιτυγχάνεται είτε υποδειγματοληπτώντας τις εξόδους των παραπάνω συστημάτων, είτε ισάξια, μετακινώντας το παράθυρο περισσότερο από ένα δείγμα κάθε φορά. Μ'αυτόν τον τρόπο αρκετή πληροφορία απορρίπτεται, ενώ αυτή που αφορά τις μεταβολές του μεγέθους του σήματος διατηρείται σε μια βολική μορφή.

2.1.5 Μέσος Ρυθμός 'Μεταβάσεων' από το Μηδέν (Short-Time Average Zero-Crossings Rate)

Ένα πέρασμα από το μηδέν ή ένα zero-crossing συμβαίνει σε διακριτά σήματα όταν δύο διαδοχικά δείγματα του σήματος έχουν διαφορετικό πρόσημο. Ο ρυθμός με τον οποίο συμβαίνουνε zero-crossings στο σήμα δίνει μια καλή εκτίμηση του φασματικού περιεχομένου του. Για σήματα στενής ζώνης ο ρυθμός αυτός παρέχει έναν άμεσο τρόπο υπολογισμού συχνότητας. Για σήματα ευρείας ζώνης όπως είναι τα σήματα φωνής, χρησιμοποιείται ένας short-time μέσος ρυθμός zero-crossings, για να αποκαλυφθούν χαρακτηριστικές φασματικές ιδιότητες όπως τα επίπεδα συχνοτήτων ή ο βαθμός διέγερσης.

Ένας κατάλληλος ορισμός του ρυθμού αυτού είναι

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] \quad sgn[x(m \quad 1)]| \, w(n \quad m) \tag{2.7}$$

με $sgn[\ ]$ τη συνάρτηση προσήμου

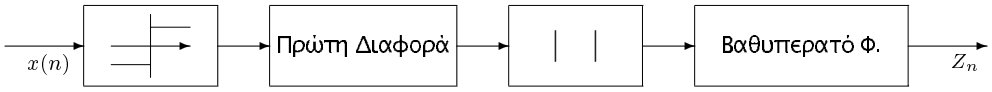
$$\begin{aligned} sgn[x(n)] &= 1 \qquad x(n) \quad 0 \\ &= \quad 1 \qquad x(n) < 0 \end{aligned} \tag{2.8}$$

και το παράθυρο ανάλυσης

$$\begin{aligned} w(n) &= \frac{1}{2N} \qquad 0 \le n \le N \quad 1 \\ &= 0 \qquad \textit{αλλού} \end{aligned} \tag{2.9}$$

Στην ουσία για τον υπολογισμό του Z_n, τα δείγματα ελέγχονται ανά δύο για ενδεχόμενη αλλαγή προσήμου και στη συνέχεια υπολογίζεται ο μέσος όρος για N συνεχόμενα δείγματα. Οι διαδικασίες υπολογισμού φαίνονται στο block του Σχ. (2.8). Επειδή ο μέσος ρυθμός εκφράζεται συνήθως σε αριθμό zero-crossings ανά παράθυρο ανάλυσης η διαίρεση με το N μπορεί να αποφευχθεί. Επειδή ισχύουν οι γενικές ιδιότητες των short-time απεικονίσεων, το παράθυρο εκτός από τετραγωνικό μπορεί να είναι π.χ. και Hamming, ενώ η συχνότητα δειγματοληψίας της εξόδου του Σχ. (2.8) μπορεί να είναι πολύ μικρότερη από αυτή της εισόδου (δηλ. το παράθυρο να μετακινείται κατά πολύ περισσότερα δείγματα ανά μέτρηση).

Η σχέση της μέτρησης του μέσου ρυθμού zero-crossings με τις συχνότητες ενός σήματος φωνής, μπορεί να επεκταθεί και να δώσει μια εκτίμηση του κατά πόσο ένα τμήμα φωνής μπορεί να είναι έμφωνο ή άφωνο. Συγκεκριμένα μεγάλο zero-crossing rate σημαίνει ψηλές συχνότητες στο εύρος του σήματος, ενώ χαμηλό zero-crossing rate σημαίνει χαμηλές συχνότητες. Επιπλέον είναι



Σχήμα 2.8: Block διάγραμμα του μέσου ρυθμού zero-crossings

γνωστό ότι για έμφωνο λόγο, οι συχνότητες με το πιο ισχυρό ενεργειακό περιεχόμενο εντοπίζονται στη ζώνη κάτω από τα 3kHz, ενώ για άφωνο (π.χ. άφωνες παύσεις και συριστικά σύμφωνα) οι ισχυρές ενεργειακά συχνότητες είναι υψηλότερες. Έτσι μια γενικευμένη προσέγγιση ίσως είναι ότι τα άφωνα τμήματα ενός σήματος χαρακτηρίζονται από μεγάλο zero-crossing rate, ενώ τα έμφωνα από μικρό. Φυσικά κάτι τέτοιο δεν μπορεί να είναι απόλυτο κριτήριο για το διαχωρισμό τους, αλλά αποτελεί ένα πολύ χρήσιμο εργαλείο προς αυτή την κατεύθυνση.

Ομαλοποίηση του average zero-crossings rate

Κατά τον υπολογισμό της zero-crossings απεικόνισης είναι πιθανόν να εμφανιστούν θορυβώδη προϊόντα στο σήμα εξόδου είτε ως συνέπεια σφαλμάτων στην επεξεργασία του αναλογικού σήματος πριν τη δειγματοληψία (π.χ. dc offset, 60Hz speech hum), είτε λόγο μικρού χρονικού παραθύρου του μέσου όρου. Για την πρώτη κατηγορία ιδιαίτερη φροντίδα πρέπει να ληφθεί για την ελαχιστοποίηση τους κατά τη φάση της μετατροπής του αναλογικού σε ψηφιακό σήμα [11]. Εξαιτίας αυτών των περιορισμών διάφορες άλλες παρόμοιες απεικονίσεις βραχέως χρόνου έχουν προταθεί στη βιβλιογραφία[11].

Για τη δεύτερη κατηγορία, καθώς και για οποιαδήποτε σφάλματα εκτίμησης που δημιουργούν μεγάλες ασυνέχειες στην απεικόνιση, εφαρμόζεται συνήθως μια διαδικασία ομαλοποίησης (*smoothing*) της απεικόνισης. Αυτές οι ασυνέχειες εκδηλώνονται ως απότομες αλλαγές τιμών ανάμεσα σε δύο διαδοχικά δείγματα ('spikes') και δυσκολεύουν την επεξεργασία και ερμηνεία των μετρήσεων. Η χρήση ενός απλού γραμμικού βαθυπερατού φίλτρου ομαλοποίησης μπορεί να διαταράξει την απεικόνιση και να εξαφανίσει και ασυνέχειες χρήσιμες για κάποια εφαρμογή(π.χ. κάποια μετάβαση από άφωνο σε έμφωνο ήχο). Για την εξάλειψη τέτοιων θορυβωδών συντελεστών ένα είδος *μη-γραμμικής* ομαλοποίησης αποδεικνύεται ότι περιορίζει τα μεγάλα σφάλματα ενώ διατηρεί τις 'χρήσιμες' ασυνέχειες.

Συγκεκριμένα ένας συνδυασμός από *median* και γραμμικά φίλτρα δίνει τα επιθυμητά αποτελέσματα[11]. Ένα κινούμενο median φίλτρο l βημάτων, $m_l[x(n)]$ είναι ο μέσος όρος των τιμών

$$x(n), \ldots, x(n \quad l+1)$$

Ως γραμμικό μπορεί να χρησιμοποιηθεί οποιοδήποτε βαθυπερατό συμμετρικό φίλτρο π.χ.ένα Hanning. Η λογική είναι το γραμμικό να 'κόβει' τις θορυβώδεις ασυνέχειες ενώ το median να διατηρεί τις μεταβολές που χρειάζονται στο σήμα ομαλοποιώντας ταυτόχρονα σε μικρότερο βαθμό. Αποδεικνύεται ότι ο συνδυασμός median-median-γραμμικό, με δύο μη γραμμικά περάσματα είναι ο κατάλληλος[11]. Έτσι χρησιμοποιείται ένα median 5 σημείων, ακολουθούμενο από ένα median 3 σημείων και ένα Hanning φίλτρο 3 σημείων με κρουστική απόκριση

$$\begin{aligned} h(n) &= 1/4 \quad n = 0 \\ &= 1/2 \quad n = 1 \\ &= 1/4 \quad n = 2 \end{aligned}$$

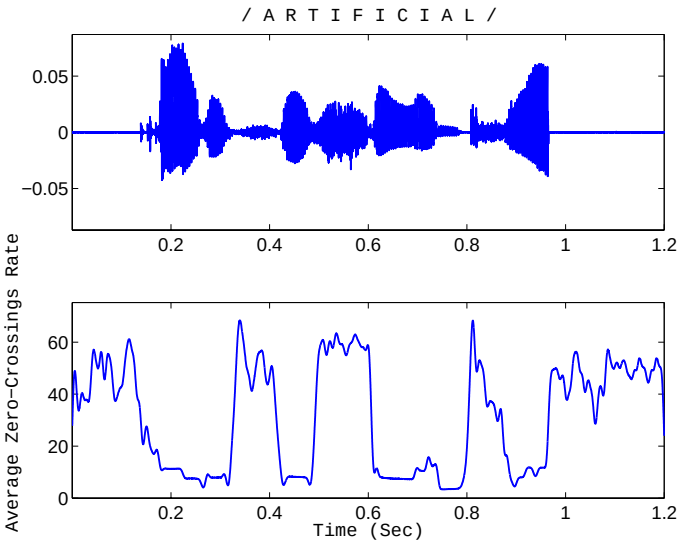
Στη συνέχεια ακολουθούνε δύο παραδείγματα υπολογισμού του short-time average zero crossings rate. Πρόκειται για τις δύο διαταραχές /artificial/, στο Σχ. 2.9(a) και /don't identify yourself/, στο Σχ. 2.9(b) που παρουσιάστηκαν προηγουμένως. Για τον υπολογισμό χρησιμοποιήθηκε παράθυρο Hamming 15ms (στα 16kHz) και το συνδυασμένο ομαλό φιλτράρισμα που περιγράφηκε. Φαίνεται καθαρά η διαφορά επιπέδων ανάμεσα στα έμφωνα και άφωνα τμήματα φωνής, ενώ ακόμη και οπτικά μπορεί κανείς να εξάγει από την απεικόνιση συμπεράσματα σχετικά με την ηχητική φύση των διαφόρων τμημάτων.

2.2 Ένας Αλγόριθμος Προσδιορισμού των Αρχικών και Τελικών Σημείων Φωνής

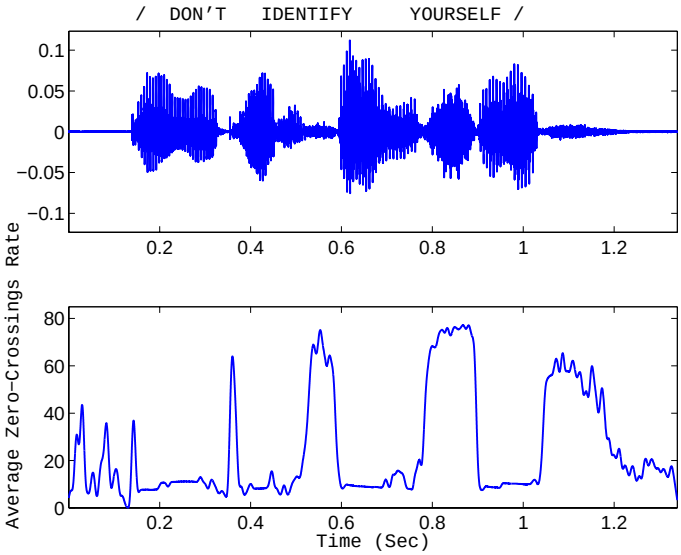
Ένας απλός τρόπος υπολογισμού των αρχικών και τελικών σημείων της φωνής (*endpoints*) προτάθηκε από τους L.Rabiner & M.Sambur [10] και εμπλέκει δύο απεικονίσεις βραχέως χρόνου. Οδηγούμενοι από την ανάγκη για έναν έμπιστο και γρήγορο αλγόριθμο διαχωρισμού της φωνής από ένα ακουστικό περιβάλλον, χρησιμοποίησαν τη μέτρηση του μέσου πλάτους και του μέσου ρυθμού zero-crossings σε μια μέθοδο με κατώφλια και συγκρίσεις.

Η μέθοδος ανήκει στην κατηγορία των αποκλειστικών τεχνικών (*explicit approach*) όπου η ανίχνευση γίνεται ανεξάρτητα από την αναγνώριση, η οποία βασίζεται σε κάποιον αλγόριθμο σύμπτωσης ιχνών (*template matching algorithm*). Πρόκειται για εφαρμογή ανίχνευσης μεμονωμένων λέξεων στην οποία ο ομιλητής προφέρει μια λέξη και όλο το διάστημα ηχογράφησης αποθηκεύεται και δειγματοληπτείται. Σκοπός είναι να βρεθεί η αρχή και το τέλος της λέξης έτσι ώστε μετέπειτα επεξεργασία και σύγκριση προτύπων να αγνοήσει τα διαστήματα σιωπής.

Υποτίθεται ότι κάπου μέσα στο διάστημα ηχογράφησης υπάρχει μια φωνητική διαταραχή και ότι θα μπορεί εύκολα να απομονωθεί. Για κάτι τέτοιο, σε



(a)



(b)

Σχήμα 2.9: Μέσος Ρυθμός Zero-Crossings με παράθυρο Hamming 15ms στα 16kHz για τις διαταραχές (α) /artificial/ (b) /don't identify yourself/. Στις απεικονίσεις εφαρμόστηκε συνδυασμός median-γραμμικού smoothing.

ένα περιβάλλον με πολύ υψηλό SNR (π.χ. ηχογράφηση σε ανηχοϊκό θάλαμο) μια απλή μέτρηση της ενέργειας του σήματος θα αρκούσε αφού η ενέργεια των πιο αδύναμων ήχων (π.χ.ασθενή συριστικά φωνήματα)ξεπερνάει την ενέργεια του ακουστικού περιβάλλοντος. Επειδή όμως οι συνθήκες δεν είναι πάντα ιδανικές μόνο μια τέτοια μέτρηση δεν αρκεί. Ο αλγόριθμος ξεχωρίζει αρχικά μια ευρεία περιοχή στην οποία εντοπίζεται το σήμα φωνής και θέτει κατώφλια για την ενέργεια της για να περιορίσει το υποψήφιο διάστημα. Χρησιμοποιεί και τη μέτρηση του μέσου ρυθμού zero-crossings για να δώσει μια καλύτερη εκτίμηση των άκρων της, ελέγχοντας την ύπαρξη άφωνων γεγονότων πριν και μετά από το «ενεργειακό» διάστημα.

Αποδείχθηκε ότι αποδίδει καλά σε οποιοδήποτε περιβάλλον όπου ο λόγος σήματος προς σιωπή (background noise) είναι τουλάχιστον 30db, ενώ μπορεί να προσαρμοστεί στις συνθήκες της κάθε εφαρμογής αφού τα κατώφλια λαμβάνονται με στατιστική επεξεργασία απευθείας στο διάστημα της διαταραχής. Κάτι τέτοιο βέβαια τον καθιστά ευάλωτο σε μη-στατικά περιβάλλοντα αφού υποθέτει παρόμοια συμπεριφορά του σήματος σιωπής καθ'όλη τη διάρκεια του διαστήματος ηχογράφησης.

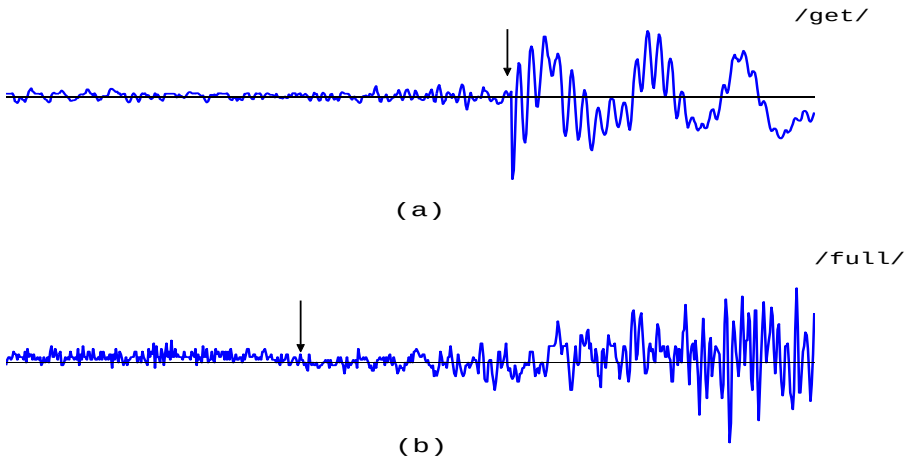
Πριν προχωρήσουμε στην περιγραφή του αλγορίθμου θα ήταν καλό να εξετάσουμε ποια χαρακτηριστικά των φωνητικών σημάτων δυσκολεύουν μια διαδικασία ανίχνευσης. Συγκεκριμένα, φωνήματα τα οποία όταν ανήκουν στην αρχή ή στο τέλος μιας λέξης είναι εύκολο να γίνουν αντιληπτά ως τμήματα σιωπής.

2.2.1 Προβλήματα Εντοπισμού των Endpoints

Το πρόβλημα εντοπισμού φωνής σε ένα περιβάλλον σιωπής είναι στη βάση του ένα πρόβλημα αναγνώρισης προτύπων. Αν θελήσει κάποιος να αναγνωρίσει με το μάτι τα όρια της φωνής σε ένα σύνθετο σήμα, θα πρέπει να συνηθίσει τη μορφή ή το 'πρότυπο' της σιωπής και στη συνέχεια να παρατηρήσει αλλαγές στην κυματομορφή που θα δηλώνουν ένα διαφορετικό 'πρότυπο' φωνής. Αρκετές φορές αυτό μπορεί να γίνει εύκολα, όταν παρατηρούνται μεγάλες μεταβολές στη στάθμη του σήματος καθώς και έντονες αλλαγές στη διακύμανση του.

Στο Σχήμα(2.10) φαίνονται 40ms, από τα τμήματα 2 σημάτων[3] στα οποία συμβαίνει η μετάβαση από φωνή σε σιωπή. Στο Σχ. 2.10 (α) φαίνεται η αρχή του /get/. Η διαφορά στην ενέργεια του σήματος ανάμεσα στη σιωπή και στο φωνητικό τμήμα είναι μεγάλη, οπότε εύκολα διακρίνεται το σημείο έναρξης. Έτσι για έμφωνη έναρξη ή λήξη της λέξης η διάκριση είναι απλή. Επίσης

[3] Να σημειωθεί ότι οι λέξεις προέρχονται από φράσεις της βάσης TIMIT οπότε οι συνθήκες ηχογράφησης είναι προκαθορισμένες.



Σχήμα 2.10: Κυματομορφές 40ms από την έναρξη των λέξεων (α) /get/ (b) /full/. Με τα βέλη σημαδεύονται τα πραγματικά σημεία έναρξης των διαταραχών.

εύκολα γίνεται η διάκριση σε περιπτώσεις που, ενώ η ενέργεια του σήματος δεν αυξάνεται αισθητά, το φασματικό περιεχόμενο αλλάζει έντονα σε σχέση με το περιβάλλον. Κάτι τέτοιο σημαίνει αύξηση στο ρυθμό zero-crossings του σήματος και συμβαίνει σε λέξεις που ξεκινάνε ή καταλήγουν με έντονα συριστικά όπως είναι το /s/.

Σε άλλες περιπτώσεις τα πράγματα δεν είναι τόσο απλά. Για παράδειγμα στο Σχ. 2.10 (b) το πραγματικό σημείο έναρξης εύκολα θα μπορούσε να προσπεραστεί με παρατήρηση και να χαθεί έτσι το αρχικό /f/. Γενικά ασθενή συριστικά είναι δύσκολο να εντοπισθούν με το μάτι. Το ίδιο συμβαίνει και με ένρινα σύμφωνα στο τέλος λέξεων. Τέλος στην περίπτωση άφωνων παύσεων (π.χ. /p/) στα σύνορα μιας λέξης, ένα μεγάλο διάστημα ηρεμίας ή μια ασθενής απελευθέρωση μπορούν να χαθούν τελείως.

Τα προβλήματα που σχετίζονται με την ανίχνευση των αρχικών και τελικών στιγμών φωνής, μπορούν να χωριστούν στις εξής κατηγορίες:

(i) Ασθενή συριστικά (/ f ,th ,h /) στην αρχή ή στο τέλος μιας διαταραχής

(ii) Ασθενείς παύσεις (/p ,t ,k /)

(iii) Τελικά ένρινα (/n ,ng /)

(iv) Έμφωνα συριστικά στο τέλος λέξεων που γίνονται άφωνα

(v) Ορισμένοι έμφωνοι ήχοι που καταλήγουν άφωνοι π.χ. το /i/ στη λέξη "binary" (/b-al-n-e-r-i/)

Για να αποδώσει ένας αλγόριθμος ανίχνευσης φωνής καλά θα πρέπει να λαμβάνει υπόψιν του τις παραπάνω περιπτώσεις 'προβληματικών' φωνημάτων στην αρχή και στο τέλος των λέξεων.

Ένα λογικό ερώτημα που προκύπτει είναι πότε θεωρείται ότι ένας αλγόριθμος αποδίδει 'καλά'. Για αυτόματη ανίχνευση φωνής ο ορισμός του 'καλά' είναι πραγματολογικός δηλ. καλύτερη απόδοση είναι αυτή που δίνει τα καλύτερα αποτελέσματα αναγνώρισης των εντοπιζόμενων λέξεων. Ο αλγόριθμος που εξετάζεται εδώ σχεδιάστηκε με την προοπτική να απομονώνεται αρκετό διάστημα από μια λέξη ώστε η ανάλυση του σήματος που προκύπτει να είναι αρκετή για την αναγνώριση της. Έτσι δεν είναι απαραίτητο να εντοπιστεί το *ακριβές* σημείο έναρξης ή λήξης μιας λέξης, αλλά είναι σημαντικό να συμπεριληφθούν όλα τα σημαντικά ακουστικά γεγονότα.

Για παράδειγμα για τη λέξη /full/ είναι σημαντικό να συμπεριληφθεί το αρχικό /f/ στα όρια της λέξης, αλλά δεν είναι απαραίτητο να εντοπιστεί ολόκληρη η διάρκεια του. Η εμπειρία λέει ότι 30 με 50 ms άφωνου διαστήματος αρκούνε για εφαρμογές αναγνώρισης φωνής [10]. Αυτή η παρατήρηση είναι πολύ σημαντική για οποιονδήποτε αλγόριθμο γιατί επιτρέπει τη χρήση συντηρητικών κατωφλιών απόφασης, πράγμα που σημαίνει λιγότερα σφάλματα και υπερεκτιμήσεις των ορίων των λέξεων.

Αξίζει να αναφερθεί εδώ ότι μετέπειτα έρευνες αποδεικνύουνε τη βελτίωση που μπορεί να επιτευχθεί σε εφαρμογές αναγνώρισης με την ακριβή τοποθέτηση των ορίων της φωνής [13]. Παρ'όλα αυτά ο αλγόριθμος που εξετάζεται στη συνέχεια σχεδιάστηκε με βάση την παραπάνω παρατήρηση και με αυτή τη λογική κινούμαστε από δω και πέρα.

2.2.2 Ο Αλγόριθμος των Rabiner&Sambur - Μια υλοποίηση

Ο σχεδιασμός του αλγορίθμου είναι τέτοιος ώστε να περιλαμβάνει απλή και γρήγορη επεξεργασία, να εντοπίζει τα σημαντικά ηχητικά γεγονότα σε μια διαταραχή και να προσαρμόζεται σε διαφορετικά περιβάλλοντα σιωπής.

Η δομή και η λογική του, διατηρήθηκαν όπως παρουσιάστηκαν αρχικά. Όπως είναι φυσικό όμως πολλά από τα στοιχεία και οι λεπτομέρειες αλλάξανε έτσι ώστε να προσαρμοστούν στα δικά μας δεδομένα. Έτσι διαφορές π.χ. στη συχνότητα δειγματοληψίας, στην προεπεξεργασία του αναλογικού σήματος κ.α. οδηγούν σε διαφορετικά παράθυρα ανάλυσης ή χρονικούς περιορισμούς στη διαδικασία κ.λ.π. Θα αναφέρονται λοιπόν τα στοιχεία της δικιάς μας υλοποίησης και για την πληρότητα όπου χρειάζεται τα αυθεντικά στοιχεία.

Ο αλγόριθμος βασίζεται, όπως έχει αναφερθεί, στις δύο μετρήσεις βραχέως χρόνου, το *μέσο πλάτος* που δίνεται από τη σχέση 2.4 και το *μέσο ρυθμό zero-crossings*, που δίνεται από τη σχέση 2.7. Και οι δύο είναι απλές και γρήγορες απεικονίσεις και συνδυασμένα μπορούν να δώσουν μια καλή εκτίμηση για

την παρουσία ή την απουσία φωνής. Η χρήση πλάτους αντί της ενέργειας προτιμήθηκε για αύξηση της ταχύτητας των υπολογισμών (υπολογίζοντας απόλυτη τιμή αντί για τετράγωνο) αλλά και για περιορισμό του φαινομένου ενίσχυσης των μεγάλων διαφορών πλάτους που προκαλεί η χρήση της ενεργειακής μέτρησης.

Το σύνθετο σήμα, σε ψηφιακή μορφή έχει ένα ρυθμό δειγματοληψίας 16κHz[4]. Τα παράθυρα ανάλυσης και για τις δύο απεικονίσεις έχουν μήκος 15ms ή 240 δείγματα[5].

Μια σημαντική υπόθεση του αλγορίθμου είναι ότι τα πρώτα 100ms του επεξεργαζόμενου σήματος είναι οπωσδήποτε σιωπή. Αυτή η υπόθεση είναι σημαντική αφού σ'αυτή στηρίζονται τα μέτρα σύγκρισης των δύο μετρήσεων. Κατά τη διάρκεια αυτού του διαστήματος υπολογίζονται τα στατιστικά χαρακτηριστικά της σιωπής και πιο συγκεκριμένα η *μέση τιμή* και η *τυπική απόκλιση* του ρυθμού zero-crossings και του μέσου πλάτους. Χρησιμοποιώντας αυτά τα μεγέθη καθώς και τη μέγιστη τιμή του πλάτους στη σιωπή αλλά και σε όλο το διάστημα, υπολογίζονται κατώφλια για το πλάτος και το zero-crossings.

Το κατώφλι, $IZCT$, του zero-crossings για άφωνο λόγο υπολογίζεται ως το ελάχιστο ενός καθορισμένου κατωφλιού, IF (60 crossings per 15ms), και το άθροισμα της μέσης τιμής zero-crossings κατά τη διάρκεια της σιωπής, $\overline{IZC}$, συν δύο φορές την τυπική απόκλιση του zero-crossings της σιωπής δηλ.

$$IZCT = min(IF, \overline{IZC} + 2\sigma_{IZC}) \tag{2.10}$$

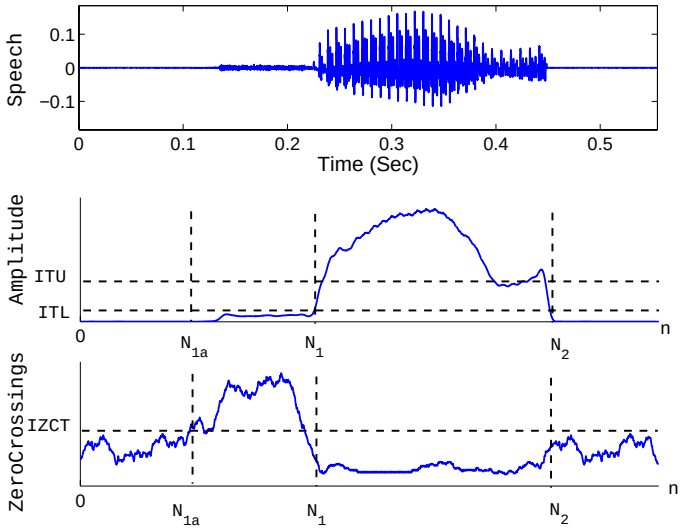
Από το μέγιστο πλάτος για όλο το διάστημα, IMX , και το μέγιστο πλάτος στο διάστημα της σιωπής των 100ms, IMN υπολογίζονται τα δύο κατώφλια πλάτους σύμφωνα με τους κανόνες:

$$I_1 = 0.03(IMX \quad IMN) + IMN \tag{2.11}$$
$$I_2 = 4 \quad IMN \tag{2.12}$$
$$ITL = min(I_1, I_2) \tag{2.13}$$
$$ITU = 5 \quad ITL \tag{2.14}$$

Στο Σχήμα(2.11) φαίνονται οι απεικονίσεις του μέσου πλάτους και του μέσου ρυθμού zero-crossings (ανά παράθυρο 15ms) για τη λέξη /full/. Παρουσιάζονται γραφικά τα κατώφλια των υπολογισμών και με βάση αυτό το σχήμα θα περιγραφεί η λειτουργία του αλγορίθμου.

[4]στο [10] χρησιμοποιείται F_s =10kHz, αφού φιλτραριστεί το σήμα από ζωνοπερατό φίλτρο 100-4000Hz.

[5]Σε αυτό το μήκος παραθύρου (N=240 στα 16kHz) καταλήξαμε μετά από δοκιμές και αυτό θα χρησιμοποιείται ως βάση για όλα τα παράθυρα ανάλυσης από δω και πέρα. Στο [10] τα παράθυρα είναι 10ms και οι απεικονίσεις υπολογίζονται με ρυθμό 100φορές/sec δηλ. με συχνότητα f_s =100Hz).



Σχήμα 2.11: Απεικόνιση της λειτουργίας του αλγορίθμου για τη λέξη /full/. Παρουσιάζονται η κυματομορφή, το μέσο πλάτος και ο ρυθμός zero-crossings ενώ διακρίνονται τα κατώφλια για την κάθε μέτρηση και τα υποψήφια σημεία.

Αρχικά ελέγχεται το πλάτος και αναζητείται το διάστημα που αυτό υπερβαίνει το συντηρητικό ψηλό κατώφλι ITU. Θεωρείται ότι τα όρια της φωνής εντοπίζονται έξω από αυτό το διάστημα. Στη συνέχεια ελέγχεται η απεικόνιση από το σημείο που ξεπερνιέται για πρώτη φορά το ITU προς τα πίσω και εντοπίζεται το σημείο που το πλάτος πέφτει κάτω από το χαμηλό κατώφλι ITL. Το σημείο αυτό ονομάζεται $N1$ στο Σχ. (2.11) και αποτελεί μια πρώτη εκτίμηση του αρχικού σημείου της λέξης. Αυτός ο έλεγχος με διπλό κατώφλι εξασφαλίζει ότι απότομές πτώσεις του πλάτους δεν θα σηματοδοτήσουν εσφαλμένο σημείο αρχής. Παρόμοια διαδικασία ακολουθείται και στο άλλο άκρο και εντοπίζεται το σημείο $N2$ ως υποψήφιο τέρμα της λέξης.

Μέχρι αυτό το σημείο χρησιμοποιούνται τα συντηρητικά κατώφλια του πλάτους και μπορούμε επομένως να πούμε με βεβαιότητα ότι μπορεί μεν τμήμα της λέξης να υπάρχει εκτός του διαστήματος $(N1, N2)$ αλλά τα σημεία αρχής και τέλους δεν βρίσκονται εντός του.

Στη συνέχεια ακολουθεί ο έλεγχος της απεικόνισης zero-crossings για τον εντοπισμό άφωνων αρχικών ή τελικών φωνημάτων. Εξετάζεται το διάστημα από $N1$ έως $N1 \quad 1200$ δηλ. ένα διάστημα 75ms από το $N1$,[6]και υπολογίζεται ο αριθμός των στιγμών που η zero-crossings μέτρηση ξεπερνάει το κατώφλι

[6]Στο [10] ελέγχονται τα προηγούμενα 25 frames, δηλ.(αφού η μέτρηση γίνεται με ρυθμό $F_s/100$) τα προηγούμενα 250ms. Στη δική μας υλοποίηση, για πρακτικούς κυρίως λόγους (επειδή τα σήματα είναι λίγο μικρότερα από 1sec(16K δείγματα)) επιλέγονται τα προηγούμενα 75ms αλλά και αυτό το διάστημα είναι αρκετό για τον έλεγχο.

$IZCT$. Αν ο αριθμός αυτός είναι πάνω από 145 (το προτεινόμενο όριο στο [10] είναι 3), τότε το πιθανό σημείο αρχής μετατοπίζεται στο πρώτο χρονικά σημείο που ξεπεράστηκε το κατώφλι $IZCT$. Στο Σχήμα το σημείο αυτό είναι το $N1a$. Διαφορετικά διατηρείται το σημείο N_1 ως εκτίμηση του σημείου έναρξης της φωνής. Παρόμοια διαδικασία ακολουθείται και στο τέλος της λέξης, και ελέγχεται η ύπαρξη άφωνης ενέργειας στο διάστημα από $N2$ έως $N2 + 1200$.

Η λογική πίσω από αυτό το δεύτερο έλεγχο είναι ότι η υπέρβαση του zero-crossings πάνω από ένα συντηρητικό κατώφλι, είναι οπωσδήποτε ισχυρή ένδειξη για την παρουσία άφωνης ενέργειας στο εξεταζόμενο διάστημα. Ο έλεγχος περιορίζεται σ'αυτό το διάστημα αφού αν προηγείται κάποιο φώνημα, 75ms διάρκειας είναι μεν υπέρ αρκετά ενώ αν η διάρκεια του είναι μεγαλύτερη αυτά τουλάχιστον θα συμπεριληφθούν στη λέξη. Φυσικά αυξάνοντας αυτόν τον χρονικό περιορισμό (και επομένως και την πολυπλοκότητα του αλγορίθμου) μπορεί ίσως να βελτιωθεί και η ακρίβεια του.

Με αναφορά στο Σχ. (2.11), η λέξη /full/ ξεκινάει με ένα δυνατό συριστικό σύμφωνο το /f/. Ο έλεγχος του πλάτους εντοπίζει αρχικά το διάστημα $(N1, N2)$ ως μια πρώτη εκτίμηση της λέξης, με $N1$ το υποψήφιο σημείο αρχής (*beginning point*) και $N2$ το υποψήφιο σημείο λήξης (*ending point*). Στη συνέχεια ο έλεγχος zero-crossings στην αρχή αποκαλύπτει ένα μεγάλο αριθμό στιγμών πάνω από το κατώφλι, και έτσι το σημείο αρχής μετατοπίζεται στο πρώτο χρονικά σημείο που συμβαίνει αυτή η υπέρβαση, το σημείο $N1a$. Ο έλεγχος στο τέλος δεν δίνει ένδειξη άφωνης ενέργειας και έτσι το σημείο λήξης παραμένει στο $N2$.

2.2.3 Παραδείγματα Χρήσης του Αλγορίθμου

Ο αλγόριθμος των Rabiner&Sambur που περιγράφηκε στην προηγούμενη ενότητα προσομοιώθηκε με χρήση του υπολογιστικού πακέτου MATLAB. Γίνανε διάφορες δοκιμές για μεμονωμένες λέξεις παρμένες από φράσεις της βάσης δεδομένων TIMIT[7]. Σε κάθε παράδειγμα που ακολουθεί οι ονομασίες δίνονται στη μορφή /"λέξη"/ ("TIMIT φράση από την οποία προέρχεται") για λόγους αναφοράς (π.χ. /full/ (si2211)). Να σημειωθεί εδώ ότι οι λέξεις που εξετάσθηκαν προσομοιώνουν καλές συνθήκες ηχογράφησης αφού ο λόγος σήματος προς σιωπή(background noise) είναι 30-40db.

Στο Σχήμα (2.12) φαίνονται έξι παραδείγματα του αλγορίθμου ανίχνευσης φωνής, για διάφορες κατηγορίες 'προβληματικών' φωνημάτων στην αρχή ή

[7]Οι φράσεις "τεμαχίστηκαν" σε λέξεις και τμήματα σιωπής απομονώθηκαν από τη αρχή και τέλος της κάθε φράσης έτσι ώστε να προσομοιωθούν όσο το δυνατόν καλύτερα πραγματικές συνθήκες ηχογράφησης. Φυσικά με πραγματικές ηχογραφήσεις μεμονωμένων λέξεων τα αποτελέσματα του αλγορίθμου μπορούν να είναι πολύ καλύτερα.

στο τέλος λέξεων. Στο Σχ. (2.12a) απεικονίζεται ο εντοπισμός των άκρων στη λέξη /harsh/. Εδώ τα κατώφλια του πλάτους είναι αρκετά για να εντοπίσουν τα κατάλληλα σημεία έναρξης και λήξης της φωνής, αφού το αρχικό /h/ και το τελικό /sh/ είναι αρκετά ισχυρά. Στο Σχ. (2.12b) η μέτρηση zero-crossings χρησιμοποιήθηκε για να εντοπιστεί το όριο του αρχικού /th/. Παρ'όλο που η ενέργεια (το πλάτος) του είναι σημαντική, τα κατώφλια ενέργειας είναι σχετικά συντηρητικά για να το εντοπίσουν.

Στα Σχ. (2.12 c , d) φαίνονται οι περιπτώσεις παύσεων σε αρχή και τέλος. Στο (c) η λέξη /coincided/ αρχίζει με την άφωνη παύση /c/ και τελειώνει με την έμφωνη /d/. Το τέλος εντοπίζεται με τη μέτρηση του πλάτους ενώ η αρχή εντοπίζεται με τη βοήθεια του zero-crossings. Στο (d),στη διαταραχή /quick/ από γυναικεία φωνητικά έχουμε δύο άφωνες παύσεις /κ/ σε αρχή και τέλος. Ο αλγόριθμος χρησιμοποιεί αποδοτικά τις εξάρσεις του zero-croosings κατά την έναρξη και λήξη της προφοράς της λέξης και κάνει μια πολύ καλή εκτίμηση των δύο σημείων. Είναι ενδεικτικό ότι παρά το χαμηλό πλάτος στην αρχή και το μεγάλο διάστημα πριν την απελευθέρωση στο τέλος, ο αλγόριθμος κατάφερε με επιτυχία να συμπεριλάβει τις 'εκρήξεις' του zero-crossings

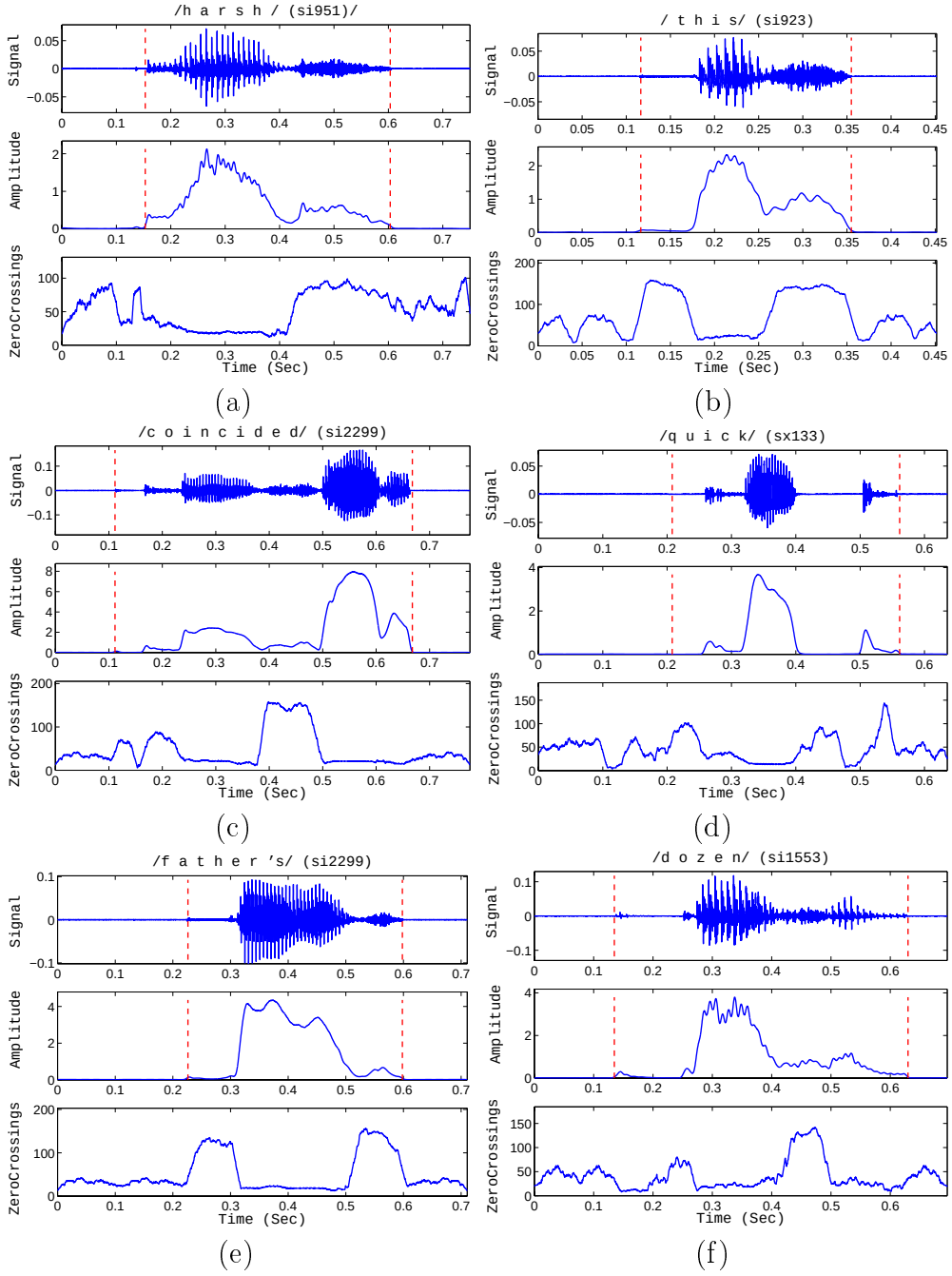
Στο Σχ. (2.12e) δοκιμάζεται η λέξη /father's/ με συριστικά σύμφωνα σε αρχή και τέλος. Το τελικό, ισχυρό ενεργειακά /s/ εντοπίζεται από το κατώφλι πλάτους,ενώ για το αρχικό ασθενές /f/, το μεγάλο zero-crossings rate δίνει το σημείο έναρξης της λέξης. Τέλος για το /dozen/ στο Σχ. (2.12f), η μέτρηση του πλάτους καταφέρνει να εντοπίσει το έμφωνο /d/ στην αρχή αλλά και το /n/ στο τέλος.

Φυσικά υπάρχουν και οι περιπτώσεις που ο αλγόριθμος χάνει τμήμα της φωνητικής πληροφορίας είτε λόγο χαμηλού zero-crossings rate είτε λόγο μικρού πλάτους. Δύο τέτοιες περιπτώσεις φαίνονται στο Σχ. (2.13). Η πληροφορία που αποκομίζεται τελικά άλλοτε είναι αρκετή για αναγνώριση (όπως στο τελικό /κ/ του /back/) και άλλοτε το ποσό που χάνεται είναι κρίσιμο και ο αλγόριθμος αποτυγχάνει(όπως στο τελικό /t/ που χάνεται κατά το μεγαλύτερο μέρος του /get/).

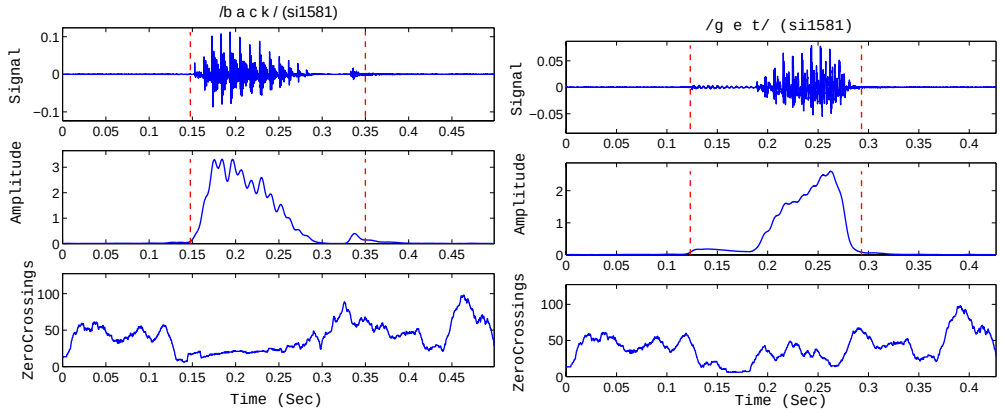
2.2.4 Γενικά Συμπεράσματα και Παρατηρήσεις πάνω στην Απόδοση του Αλγορίθμου-Προοπτικές

Ο αλγόριθμος δοκιμάστηκε σε ένα σύνολο διαταραχών όπως αυτές που παρουσιάστηκαν προηγούμενα με σκοπό να βρεθούν τα πλεονεκτήματα και τα αδύναμα σημεία του. Ορισμένες γενικές παρατηρήσεις:

1. Είναι απλός και γρήγορος υπολογιστικά,αφού χρησιμοποιεί βασικές απεικονίσεις βραχέως χρόνου οι οποίες μπορούν να θεωρηθούν και ως απο-



Σχήμα 2.12: Παραδείγματα εφαρμογής του αλγορίθμου των Rabiner&Sambur για διάφορες λέξεις, που δείχνουν την απόδοση του σε διαφορετικές κατηγορίες προβλημάτων.



Σχήμα 2.13: Παραδείγματα δοκιμών που ο αλγόριθμος (a) εντοπίζει αρκετή άφωνη διάρκεια (b) αποτυγχάνει

τέλεσμα γραμμικών συστημάτων στο σήμα.

2. Προσαρμόζεται εύκολα σε διαφορετικά περιβάλλοντα σιωπής (έντασης ή φασματικού περιεχομένου) αφού η πληροφορία για τις όποιες αποφάσεις λαμβάνεται από το αντίστοιχο διάστημα ενδιαφέροντος.

3. Είναι, θα μπορούσαμε να πούμε, ανεξάρτητος από την ίδια την εφαρμογή για την οποία προορίζονται οι λέξεις που ανιχνεύονται. Δεν εμπλέκονται γλωσσικοί ή λεξιλογικοί περιορισμοί, παρά μόνο εξετάζεται η φύση των φωνημάτων σε οποιαδήποτε διαταραχή.

4. Γενικά ανιχνεύει χωρίς μεγάλα σφάλματα, τα σημεία έναρξης και λήξης των λέξεων. Σε αρκετές περιπτώσεις προβληματικών φωνημάτων συμπεριλαμβάνει αρκετή άφωνη ενέργεια(30-50ms) έτσι ώστε να είναι δυνατή η έμπιστη χρήση τους.

5. Είναι δε σχεδιασμένος για να ελαχιστοποιεί τα μεγάλα σφάλματα (άνω των 50ms) και χρησιμοποιεί συντηρητικά κατώφλια οπότε περιορίζει τις εσφαλμένες εκτιμήσεις.

6. Παρουσιάζεται όμως και ένας αριθμός λαθών, άλλοτε μικρά άλλοτε μεγάλα, όπως ασθενή άφωνα σύμφωνα ή έμφωνα που καταλήγουν άφωνα και παύσεις , τα οποία χάνονται αρκετά ή τελείως.

7. Η απόδοση του εξαρτάται από το επίπεδο της σιωπής ή θορύβου περιβάλλοντος. Όπως αναφέρθηκε θεωρείται δεδομένος ένας υψηλός λόγος σήματος προς ακουστικό θόρυβο (>30db). Σε ακόλουθο κεφάλαιο εξετάζεται και η κάμψη της απόδοσης με την προσθήκη θορύβου στο σήμα.

8. Παρατηρήθηκε ότι αντιδρά απρόβλεπτα σε μη-στατικές συνθήκες του περιβάλλοντος. Αν η σιωπή μεταβάλλεται αισθητά κατά το διάστημα της διαταραχής, τότε οι εκτιμήσεις 'πέφτουν έξω' και αυτό γιατί το 'πρότυπο' της σιωπής που δημιουργείται στην αρχή θεωρείται αντιπροσωπευτικό και για το υπόλοιπο διάστημα. Αποτέλεσμα είναι να περιλαμβάνει και μεγάλα τμήματα σιωπής στο εκτιμώμενο διάστημα της λέξης.

Οι προοπτικές που δημιούργησε ο αλγόριθμος των Rabiner&Sambur για το πρόβλημα εντοπισμού των άκρων φωνής είναι σημαντικές. Απέδειξε την πρακτική σημασία χρονικών απεικονίσεων πλούσιων σε πληροφορία, όπως η ενέργεια και ο ρυθμός zero-crossings και προσέγγισε το πρόβλημα χρησιμοποιώντας απλή λογική. Έτσι οποιοδήποτε σχόλιο εκφράζεται περισσότερο με μια αναφορική ματιά.

Για να λειτουργήσει αποδοτικότερα ο αλγόριθμος οι προσπάθειες μπορούν να επικεντρωθούν σε δύο κατηγορίες. Η μία είναι να αυξηθεί η διακριτικότητα και επομένως και η ακρίβεια του αλγορίθμου. Αυτό σημαίνει να ενταχθούν κι άλλες παράμετροι του σήματος στη διαδικασία όπως π.χ.οι συντελεστές αυτοσυσχέτισης, ή οι συντελεστές γραμμικής πρόβλεψης. Στο ίδιο πλαίσιο, διατηρώντας την απλή λογική του, νέες μοντέρνες απεικονίσεις μπορούν ίσως να λάβουν τη θέση των παλιών με την προϋπόθεση να 'περικλείουν' περισσότερη πληροφορία σχετικά με τη διαφορά φωνής και σιωπής. Κάτι τέτοιο προτείνεται στα επόμενα κεφάλαια.

Μια δεύτερη κατηγορία, μπορεί να είναι η βελτίωση της συμπεριφοράς του σε θορυβώδεις συνθήκες. Ίσως όμως αυτό να σημαίνει εγκατάλειψη της απλής λογικής και μια τελείως διαφορετική δομή(π.χ. μια υβριδική μέθοδο αντί για αποκλειστική).

2.3 Ανακεφαλαίωση

Παρουσιάστηκε η μέθοδος, που αποτέλεσε τη βάση για πολλά από τα σύγχρονα συστήματα ανίχνευσης φωνής. Ο αλγόριθμος εντοπισμού των endpoints της φωνής που προτάθηκε από τους Rabiner&Sambur, βασίζεται σε απεικονίσεις του σήματος στο πεδίο του χρόνου, και συγκεκριμένα απεικονίσεις βραχέως χρόνου (short-time) όπως είναι η ενέργεια, το πλάτος αλλά και ο μέσος ρυθμός zero-crossings. Οι μετρήσεις αυτές δίνουν μια σαφή εικόνα για τις αλλαγές ανάμεσα σε έμφωνα τμήματα, άφωνα τμήματα και σιωπή με απλό τρόπο. Οι δοκιμές του αλγορίθμου απέδειξαν ότι παρ'όλο που υπάρχουν αποτυχίες ελαχιστοποιεί τα μεγάλα λάθη. Τέλος να σημειωθεί ότι ο συγκεκριμένος αλγόριθμος επιλέχτηκε ως πρωτοπόρα μέθοδος και απλή έκφραση των τρόπων και των εργαλείων που χρησιμοποιήθηκαν και χρησιμοποιούνται σήμερα. Θα

αποτελέσει αναφορά και μέτρο σύγκρισης για τα νέα,σύγχρονα εργαλεία που θα παρουσιαστούν στη συνέχεια.

Κεφάλαιο 3

Σύγχρονες Μη-Γραμμικές Τεχνικές Επεξεργασίας Σημάτων Φωνής

Πολλές σύγχρονες έρευνες στο πεδίο της παραγωγής και επεξεργασίας σημάτων φωνής στρέφονται σε μη-γραμμικές μεθόδους, αναζητώντας εναλλακτικούς τρόπους προσέγγισης. Άλλες φορές πάλι προβάλλουν ως μια αναπόφευκτη ανάγκη σε ένα σύστημα τόσο άγνωστο και περίπλοκό από τη φύση του όπως είναι το φωνητικό. Παράλληλα πολλές απ'αυτές επεκτείνονται από τα δισδιάστατα σήματα φωνής και στην επεξεργασία εικόνων.

Ακολουθώντας ενδείξεις από διάφορα μη-γραμμικά και χρονικά μεταβαλλόμενα φαινόμενα που λαμβάνουν χώρα κατά την παραγωγή της φωνής, ένα φωνητικό σήμα μπορεί να μοντελοποιηθεί από ένα άθροισμα ταλαντωτών διαμορφωμένων κατά AM-FM. Μ'αυτό τον τρόπο οι χαρακτηριστικοί ήχοι του σήματος ή αλλιώς τα 'formants' της φωνής θεωρούνται ημιτονικά σήματα με χρονικά μεταβαλλόμενο πλάτος και συχνότητα. Οι διαμορφώσεις αυτές είναι δυνατόν να εντοπιστούν, και τα στιγμιαία πλάτη και συχνότητες να εκτιμηθούν, προσφέροντας έτσι πληροφορία για τις μεταβολές ολόκληρου του σήματος.

Προς αυτή την κατεύθυνση ένας *αλγόριθμος αποδιαμόρφωσης* προτάθηκε από το Marago κ.α.[6] που βασίζεται σε ένα μη-γραμμικό 'ενεργειακό' τελεστή $\Psi[\ ]$. Η έκφραση ενεργειακός οφείλεται στο γεγονός ότι υπολογίζει την ενέργεια της πηγής που παράγει το σήμα, και μάλιστα ως συνάρτηση τόσο του πλάτους όσο και της συχνότητας του σήματος κάθε χρονική στιγμή. Ο τελεστής αναπτύχθηκε από τον Teager[12], ενώ για σκοπούς ενεργειακής εκτίμησης χρησιμοποιήθηκε πρώτη φορά από τον Kaiser[3].

Ο Αλγόριθμος διαχωρισμού της ενέργειας(Energy Separation Algorithm) ή ESA, εκτιμάει τα σήματα διαμόρφωσης πλάτους και συχνότητας για κάθε formant. Για να γίνει όμως αυτό και να έχει ρεαλιστικό νόημα, το σήμα φιλτράρεται πρώτα από ένα πλέγμα ζωνοπερατών φίλτρων έτσι ώστε να

απομονωθούν οι ισχυροί συντελεστές του σε κάθε ζώνη (δηλ. τα formants). Αυτή η διαδικασία, συστηματικά παρουσιασμένη από τον Bovik[1], είναι γνωστή ως Αποδιαμόρφωση σε πολλαπλές 'μπάντες' (Multiband Demodulation Analysis) ή MDA.

Με αυτές τις νέες μεθόδους αντιμετώπισης των φωνητικών σημάτων διαθέσιμες, υπήρξε το κίνητρο για να διερευνηθεί η χρησιμότητα τους στην ανάπτυξη νέων μεθόδων διάκρισης φωνής από σιωπή, ή ακόμη και θορύβου. Ένα πρώτο βήμα ήταν να βρεθούν πως αντιδράνε αυτές οι διαφορετικές κλάσεις σημάτων στον ενεργειακό τελεστή αλλά και στον ESA, μέσα από την διαδικασία της MDA ανάλυσης. Ακριβώς επειδή η έννοια της ενέργειας πλέον σχετίζεται με την ενέργεια της πηγής παραγωγής του σήματος,έγινε μια προσπάθεια εμπλοκής των σχετικών μεγεθών σε απεικονίσεις στο πεδίο του χρόνου. Προτείνεται μια μέθοδος συνδυασμού των επιμέρους καναλιών του MDA σε μία μέτρηση καθώς και η αντίστοιχη απεικόνιση βραχέως χρόνου, με σκοπό να αναδείξουν διαφορές ανάμεσα σε σιωπή, άφωνους και έμφωνους ήχους.

Το κεφάλαιο ξεκινάει με την περιγραφή και παρουσίαση των μη-γραμμικών μεθόδων επεξεργασίας που αναφέρθηκαν. Στη συνέχεια εξετάζεται η εφαρμογή τους σε σήματα φωνής, σιωπής και θορύβου καθώς και οι διαφορές που τονίζουν. Παρουσιάζεται η μέθοδος υπολογισμού της συνολικής ενέργειας, αποδιαμορφωμένου πλάτους και στιγμιαίας συχνότητας ενός σήματος (και όχι των επιμέρους ζωνοπερατών συντελεστών του) ενώ ορίζονται και τα αντίστοιχα 'μέσα' μεγέθη, παρόμοια με τις απεικονίσεις βραχέως χρόνου. Τέλος γίνονται συγκρίσεις με την short-time ενέργεια, το μέσο πλάτος και τον μέσο ρυθμό zero-crossings. Ως ένα επιπλέον μη-γραμμικό εργαλείο περιγράφεται και η μέτρηση της short-time fractal διάστασης, μια απεικόνιση που προκύπτει από μορφολογική επεξεργασία του σήματος και δίνει πληροφορία για την πολυπλοκότητα και τη λεπτομέρεια της κυματομορφής του.

3.1 Ανίχνευση Ενέργειας και Αποδιαμόρφωση AM-FM σημάτων

3.1.1 Ο Ενεργειακός Τελεστής Ψ (Teager-Kaiser Energy Operator).

Η χρήση μοντέλων διαμορφώσεων πλάτους(AM) και συχνότητας(FM) είναι έντονη σε συστήματα μετάδοσης πληροφορίας. Η πρωτοπόρα δουλειά του Teager[12] στο πεδίο της μη-γραμμικής μοντελοποίησης της φωνής τόνισε την κυριαρχική παρουσία τέτοιων διαμορφώσεων και κατά τη διαδικασία παραγωγής της φωνής. Ανάμεσα στα εργαλεία που πρότεινε για τη μη-

γραμμική επεξεργασία σημάτων είναι και ο μη-γραμμικός *ενεργειακός τελεστής*:

$$\Psi_c[s(t)] \triangleq [\dot{s}(t)]^2 \quad s(t)\ddot{s}(t) \tag{3.1}$$

για αναλογικά σήματα, με $\dot{s} = ds/dt$ και

$$\Psi_d[s(n)] \triangleq [s(n)]^2 \quad s(n \quad 1)s(n+1) \tag{3.2}$$

για σήματα σε διακριτή μορφή $s(n), n = 0, \pm 1, \pm 2, \quad$.

Ο τελεστής χρησιμοποιήθηκε συστηματικά για πρώτη φορά από τον Kaiser[3] για τον εντοπισμό της ενέργειας απλών αρμονικών ταλαντωτών. Συγκεκριμένα ας θεωρήσουμε την ταλάντωση ενός συστήματος μάζας m και ελατηρίου σταθεράς k, η μετατόπιση του οποίου $x(t)$ δίνεται από την εξίσωση κίνησης $m\ddot{x} + kx = 0$. Η λύση της διαφορικής εξίσωσης είναι ένα συνημίτονο της μορφής $x(t) = A\cos(\omega_0 t + \theta)$, με $\omega_0 = \sqrt{k/m}$. Η ολική ενέργεια του ταλαντωτή (κινητική και δυναμική) είναι σταθερή και ανάλογη με τα τετράγωνα πλάτους και συχνότητας, δηλ.:

$$E = \frac{1}{2}(m\dot{x}^2 + kx^2) = \frac{m}{2}(A\omega_0)^2$$

Η εφαρμογή του τελεστή (3.1) στο $x(t)$ δίνει:

$$\Psi_c[A\cos(\omega_0 t + \theta)] = A^2\omega_0^2 = \frac{E}{(m/2)}$$

Επομένως το αποτέλεσμα του τελεστή στο σήμα ταλάντωσης είναι η ενέργεια (ανά μισή μονάδα μάζας) της πηγής που παράγει το σήμα.

Τα σημαντικά νέα στοιχεία που προκύπτουν από τη θεώρηση του ενεργειακού τελεστή των Teager-Kaiser είναι α) η προοπτική ανάλυσης σημάτων από την πλευρά της ενέργειας της πηγής που τα παράγει και β) η παρουσία της έννοιας της συχνότητας στην ενεργειακή ποσότητα.

3.1.2 Ανίχνευση Ενέργειας AM-FM σημάτων

Πέρα από τα στατικά, σε πλάτος και συχνότητα, σήματα η ανίχνευση της ενέργειας μέσω των ενεργειακών τελεστών μπορεί να επεκταθεί και σε σήματα AM-FM δομής, που συναντώνται πολύ συχνά σε συστήματα επικοινωνιών. Ένα τέτοιο σήμα πραγματικών τιμών είναι της μορφής

$$s(t) = a(t)\cos\left(\omega_c t + \omega_m \int_0^t q(\tau)d\tau + \theta\right) \tag{3.3}$$

Το $s(t)$ μπορεί να θεωρηθεί είτε ως ένα FM σήμα του οποίου το πλάτος μεταβάλλεται σύμφωνα με ένα AM σήμα, είτε ως ένα AM σήμα του οποίου η

συχνότητα δεν είναι σταθερή, αλλά μεταβάλλεται σύμφωνα με ένα FM σήμα διαμόρφωσης $q(t)$. Πρόκειται στην ουσία για ένα συνημίτονο με φέρουσα συχνότητα ω_c, χρονικά μεταβαλλόμενο πλάτος $a(t)$ και φάση

$$\phi(t) = \omega_c t + \omega_m \int_0^t q(\tau) d\tau + \theta \tag{3.4}$$

Η γωνιακή συχνότητα είναι επίσης χρονικά μεταβαλλόμενη και δίνεται από την πρώτη παράγωγο της φάσης δηλ.

$$\omega_i(t) = \frac{d}{dt}\phi(t) = \omega_c + \omega_m q(t) \tag{3.5}$$

, όπου $q(t)$ το σήμα διαμόρφωσης συχνότητας με $|q(t)| \leq 1$, ω_m η μέγιστη απόκλιση από τη φέρουσα ω_c, και $\theta = \phi(0)$ η αρχική τιμή της φάσης. Τέτοια σήματα αναφέρονται ως *AM–FM σήματα* και μεταφέρουν πληροφορία τόσο στο πλάτος $a(t)$ όσο και στη συχνότητα $\omega_i(t)$.

Η εφαρμογή των τελεστών (3.1) και (3.2) στην ευρεία τάξη των AM–FM σημάτων, ανεξάρτητα από την πηγή ή τη φύση τους, μελετήθηκε από το Marago κ.α [5], και αποδείχτηκε πολύ χρήσιμη για την ανίχνευση της πληροφορίας διαμόρφωσης.

Συγκεκριμένα στο [5] αποδεικνύεται ότι ο ενεργειακός τελεστής μπορεί να προσφέρει μια εκτίμηση του πλάτους AM σημάτων, της συχνότητας FM σημάτων και του γινομένου τους για την περίπτωση AM–FM σημάτων (τόσο στη συνεχή όσο και στη διακριτή μορφή τους), με πολύ μικρό σφάλμα προσέγγισης κάτω από γενικευμένες ρεαλιστικές συνθήκες.

- Για AM σήματα της μορφής $s_{am}(t) = a(t)\cos(\omega_c t + \theta)$, με την υπόθεση ενός ζωνο-περιορισμένου $a(t)$, δηλαδή πολύ αργά μεταβαλλόμενου σε σχέση με τη φέρουσα ω_c, ο αναλογικός τελεστής δίνει:

  $$\Psi[s_{am}(t)] \quad a^2(t)\omega_c^2 \tag{3.6}$$

  ενώ για το αντίστοιχο διακριτό σήμα $d_{am} = a(n)\cos(\Omega_c n + \theta)$, με Ω_c την αντίστοιχη διακριτή φέρουσα, υποθέτοντας ξανά αργή μεταβολή:

  $$\Psi[d_{am}(t)] \quad a^2(n)\sin^2(\Omega_c) \tag{3.7}$$

  Έτσι ο τελεστής $\sqrt{\Psi}$ μπορεί να εκτιμήσει την περιβάλλουσα των AM σημάτων, με πολύ μικρό σφάλμα πάντα με την υπόθεση της μικρής μεταβολής της, σε σχέση με τη φέρουσα.

- Για απλά FM σήματα της μορφής $s_{fm}(t) = \cos[\phi(t)]$, όπου η φάση $\phi(t)$ είναι της μορφής (3.4), η εκτίμηση του τελεστή είναι:

  $$\Psi[s_{fm}(t)] \quad \omega_i^2(t) \tag{3.8}$$

Η προϋπόθεση για να ισχύει η παραπάνω με πολύ μικρό σφάλμα είναι η στιγμιαία συχνότητα ω_i να μην αλλάζει πολύ (δηλ. μικρό ω_m) ούτε γρήγορα σε σχέση με τη φέρουσα ω_c. Ανάλογες προϋποθέσεις ισχύουν και για το διακριτό FM σήμα

$$d_{fm}(n) = \cos[\phi(n)] = \cos\left[\Omega_c n + \Omega_m \int_0^n q(m)dm + \theta\right]$$

για το οποίο ο τελεστής εκτιμάει το τετράγωνο του ημιτόνου της στιγμιαίας διακριτής συχνότητας δηλ.:

$$\Psi[d_{fm}(n)] \quad \sin^2(\Omega_i(n)) \tag{3.9}$$

Να σημειωθεί ότι στο [5] δίνονται τα όρια των σφαλμάτων των προσεγγίσεων των τελεστών τα οποία τηρουμένων των γενικών συνθηκών που οριστήκαν είναι αρκετά μικρά.

Η περίπτωση των AM–FM σημάτων, που πρόκειται να μας απασχολήσει περισσότερο, παρουσιάζει και το μεγαλύτερο ενδιαφέρον από πρακτικής πλευράς. Ένα αναλογικό AM–FM σήμα δίνεται από τη σχέση (3.3) και εφαρμόζοντας τον ενεργειακό τελεστή προκύπτει (όπου με a εννοείται $a(t)$ και με ϕ το $\phi(t)$):

$$\Psi[a\cos(\phi)] = (a\dot{\phi})^2 + \underbrace{a^2\ddot{\phi}\sin(2\phi)/2 + \cos^2(\phi)\Psi(a)}_{E(t)} \tag{3.10}$$

Οι δύο τελευταίοι όροι συντελούν το σφάλμα προσέγγισης $E(t)$ και γίνονται πολύ μικροί υπό τις ρεαλιστικές συνθήκες ότι τα σήματα πληροφορίας $a(t)$ και $q(t)$ είναι ζωνοπεριορισμένα και μάλιστα $\omega_a, \omega_m << \omega_c$. Έτσι ο ενεργειακός τελεστής δίνει με πολύ μικρό σφάλμα μια εκτίμηση του τετραγωνικού γινομένου πλάτους και στιγμιαίας συχνότητας (δηλ. έμμεσα και της ενέργειας) του σύνθετου AM–FM σήματος:

$$\Psi[a(t)\cos(\phi(t))] \quad a^2(t)\omega_i^2(t) \tag{3.11}$$

υποθέτοντας ότι *τα σήματα* $a(t)$ *και* $\omega_i(t)$ *δε μεταβάλλονται πολύ γρήγορα ή σε μεγάλο βαθμό σε σχέση με τη φέρουσα συχνότητα* ω_c. Αντίστοιχα για τη διακριτή μορφή AM–FM η έξοδος του τελεστή είναι:

$$\Psi[a(n)\cos(\phi(n))] \quad a^2(n)\sin^2(\Omega_i(n)) \tag{3.12}$$

,δηλαδή ο διακριτός τελεστής $\sqrt{\Psi}$ μπορεί με πολύ καλή προσέγγιση να εκτιμήσει το γινόμενο της AM περιβάλλουσας και του ημιτόνου της FM στιγμιαίας συχνότητας. Πειράματα στο [5] αποδεικνύουν ότι η προσέγγιση είναι πολύ καλή ακόμη και για μεγάλα ποσά AM ή FM διαμόρφωσης.

Τέλος αξίζει να σημειωθεί ότι τα αποτελέσματα για συνημιτονικά σήματα μπορούν να συμπεριλάβουν οποιοδήποτε σταθερό πλάτος $A \neq 1$ ή/και εκθετικό παράγοντα e^{rt} (r^n για τη διακριτή περίπτωση), αφού $\Psi_c[Ae^{rt}s(t)] = A^{2r}e^{2rt}\Psi_c[s(t)]$ (και αντίστοιχα για διακριτά σήματα
$\Psi_d[Ar^n s(n)] = A^2 r^{2n}\Psi_d[s(n)]$) [3].

3.1.3 Ένας Αλγόριθμος Διαχωρισμού της Ενέργειας (ESA)

Πέρα από την ανίχνευση της ενέργειας είναι επιθυμητό οι εκτιμήσεις (3.11) και (3.12) να μπορέσουν να διαχωριστούν σε πλάτος περιβάλλουσας και στιγμιαία συχνότητα. Κάτι τέτοιο είναι πολύ σημαντικό σε πολλές εφαρμογές επικοινωνιών αφού η αποδιαμόρφωση σημαίνει απόκτηση της πληροφορίας που κουβαλάει ένα σήμα. Όπως θα φανεί στη συνέχεια, για την ανάλυση φωνητικών σημάτων, μοντελοποιημένων με τη βοήθεια AM-FM διαμορφώσεων, ο διαχωρισμός της ενέργειας, που εκτιμάει ο ενεργειακός τελεστής δίνει το πλάτος $|a(t)|$ και τη στιγμιαία συχνότητα $f(t)$ των συντονισμών (*speech resonanses*) της φωνής.

Προς την επίλυση αυτού του προβλήματος προτάθηκε από το Marago κ.α. [6] μια μέθοδος που χρησιμοποιεί μη-γραμμικούς συνδυασμούς της στιγμιαίας εξόδου του ενεργειακού τελεστή για να μπορέσει να διαχωρίσει την ενέργεια στους συντελεστές AM και FM διαμόρφωσης. Ο αλγόριθμος ονομάστηκε *Energy Separation Algorithm* (ESA) αφενός λόγο της εξάρτησης της ενέργειας ενός ταλαντωτή από το γινόμενο του πλάτους και της συχνότητας και αφετέρου λόγο της χρήσης των ενεργειακών τελεστών.

Διαχωρισμός Ενέργειας για σήματα Συνεχούς Χρόνου (CESA)

Αν θεωρήσουμε ένα απλό μονοχρωματικό σήμα(σταθερό πλάτος και συχνότητα) $s(t) = A\cos(\omega_c t + \theta)$ και το διαφορικό πρώτης τάξης του $\dot{s}(t) = \quad A\omega_c \sin(\omega_c t + \theta)$, τότε το αποτέλεσμα του τελεστή στο σήμα είναι ακριβώς

$$\Psi[s(t)] = A^2\omega_c^2 \tag{3.13}$$

ενώ με την ίδια λογική:

$$\Psi[\dot{s}(t)] = (\quad A\omega_c)^2\omega_c^2 = A^2\omega_c^4 \tag{3.14}$$

Από τις (3.13) και (3.14) προκύπτει ότι η σταθερή συχνότητα και το απόλυτο πλάτος μπορούν να υπολογιστούν από τις εξισώσεις:

$$\omega_c = \sqrt{\frac{\Psi[\dot{s}(t)]}{\Psi[s(t)]}} \tag{3.15}$$

$$|A| = \frac{\Psi[s(t)]}{\sqrt{\Psi[\dot{s}(t)]}} \tag{3.16}$$

Ας θεωρήσουμε την γενική περίπτωση AM-FM σημάτων της μορφή (3.3) δηλ. $s(t) = a(t)\cos[\phi(t)]$. Υπό την προϋπόθεση ότι τα σήματα πλάτους και συχνότητας δε μεταβάλλονται πολύ σε μέγεθος ή πολύ γρήγορα σε σχέση με το φέρον, η έξοδος του ενεργειακού τελεστή, με πολύ μικρό σφάλμα προσέγγισης δίνεται από τη σχέση (3.11) η οποία επαναλαμβάνεται εδώ για λόγους αναφοράς[1]:

$$\Psi[s(t)] \quad a^2(t)\omega_i^2(t) \tag{3.17}$$

Για διαχωρισμό πλάτους-συχνότητας χρειάζεται και η εφαρμογή του τελεστή Ψ στο διαφορικό πρώτης τάξης του σήματος το οποίο είναι:

$$\dot{s}(t) = \dot{a}(t)\cos[\phi(t)] \quad a(t)\dot{\phi}(t)\sin[\phi(t)]$$

Πράγματι αποδεικνύεται στο [6] ότι για παρόμοιες συνθήκες ισχύει με αμελητέο σφάλμα προσέγγισης:

$$\Psi[\dot{s}(t)] \quad a^2\omega_i^4(t) \tag{3.18}$$

Ο συνδυασμός των σχέσεων (3.17) και (3.18) δίνει το διαχωρισμό των σημάτων στιγμιαίας συχνότητας $\omega_i(t)$ και του πλάτους περιβάλλουσας $|a(t)|$:

$$\sqrt{\frac{\Psi[\dot{s}(t)]}{\Psi[s(t)]}} \quad \omega_i(t) \tag{3.19}$$

$$\frac{\Psi[s(t)]}{\sqrt{\Psi[\dot{s}(t)]}} \quad |a(t)| \tag{3.20}$$

Οι παραπάνω σχέσεις αποτελούν τον αλγόριθμο διαχωρισμού ενέργειας για σήματα διακριτού χρόνου (*Continuous Energy Separation Algorithm*) ή αλλιώς $CESA$. Παρατηρούμε ότι οι (3.19),(3.20) για σταθερό πλάτος και συχνότητα καταλήγουν ακριβώς στη μορφή των (3.15) και (3.16) και δίνουν τη λύση $\omega_i(t) = \omega_c, a(t) = |A|$.

[1] Γενικά ο συμβολισμός όπου αναφέρεται σημαίνει

$$\Psi(s) = D + E \quad D \Leftrightarrow \frac{E_{max}}{D_{max}} << 1$$

,όπου E το σφάλμα προσέγγισης και D η επιθυμητή τιμή που εκτιμάει ο τελεστής.

Διαχωρισμός Ενέργειας για σήματα Διακριτού Χρόνου (DESA)

Η εφαρμογή του ενεργειακού διαχωρισμού για διακριτά σήματα είναι λίγο πιο περίπλοκη. Συγκεκριμένα υπάρχουν τρεις διαφορετικές μέθοδοι εκτίμησης του πλάτους και της στιγμιαίας συχνότητας, ανάλογα με τον τρόπο που εκφράζεται ο διακριτός ενεργειακός τελεστής Ψ_d. Στο [5] αναφέρονται τρεις τρόποι διακριτοποίησης του αναλογικού τελεστή (3.1) ανάλογα με το πως προσεγγίζεται το διαφορικό πρώτης τάξης $\dot{s}(n)$ (σε σχέση με κάποια διαφορά προηγούμενων, τρεχόντων και επόμενων τιμών του διακριτού $s(n)$).

Παρακάτω θα αναφερθούν οι τρεις διαφορετικοί $DESA$ μαζί με τους αντίστοιχους τελεστές που χρησιμοποιούν. Για όλες τις περιπτώσεις κάτω από τις γενικές συνθήκες μικρών και αργών μεταβολών σε σχέση με το φέρον, τα σφάλματα προσέγγισης είναι σχεδόν αμελητέα. Οι λεπτομέρειες των υπολογισμών καθώς και η αιτιολόγηση των προσεγγίσεων, για τα πολύ μικρά σφάλματα δίνονται αναλυτικά στο [6].

Καταρχήν ας θεωρήσουμε ξανά το γενικευμένο διακριτό AM–FM σήμα:

$$s(n) = a(n)\cos[\phi(n)] = a(n)\cos[\Omega_i(n)n + \theta]$$

όπου $\Omega_i(n)$ (σε rad/δείγμα) η στιγμιαία συχνότητα

$$\Omega_i(n) = \frac{d}{dn}\phi(n) = \Omega_c + \Omega_m q(n)$$

που μεταβάλλεται σύμφωνα με το σήμα πληροφορίας $q(n)$, και λαμβάνει τιμές στο διάστημα $(0, \pi)$. Το πέρασμα του τελεστή (όπως και αν οριστεί αυτός) από το $s(n)$ δίνεται από τη σχέση (3.12),δηλαδή:

$$\Psi[s(n)] \quad a^2(n)\sin^2[\Omega_i(n)] \tag{3.21}$$

Προσεγγίζοντας το διαφορικό $\dot{s}$ με τη διαφορά 2–δειγμάτων προς τα πίσω (ή προς τα μπρος) δηλ. αντικαθιστώντας στην (3.1) το t με n, το $s(t)$ με $s(nT)$, και το $\dot{s}(t)$ με $[s(n) \quad s(n \quad 1)]/T$ (ή με $[s(n+1) \quad s(n)]/T$) ,ο διακριτός τελεστής είναι όπως στη σχέση (3.2) δηλ.[2]:

$$\Psi[s(n)] = s^2(n) \quad s(n+1)s(n \quad 1)$$

Η εφαρμογή αυτού του τελεστή στην πρώτη παράγωγο $\dot{s}(n)$, όπως προσεγγίστηκε προηγούμενα, δίνει:

$$\Psi[\dot{s}(n)] = \Psi[s(n) \quad s(n \quad 1)] \quad 4a^2(n)\sin^2[\Omega_i(n)/2] \quad sin^2[\Omega_i(n)] \tag{3.22}$$

[2]Το T είναι η συνάρτηση δειγματοληψίας δηλ. ένας παράγοντας κλίμακας που συμπεριλαμβάνεται στην Ω_i. Έτσι μπορούμε, για απλοποίηση να υποθέσουμε $= 1$ και στο τέλος να ανάγουμε τα αποτελέσματα σύμφωνα με τη σχέση $f_i = \Omega_i/2\pi T$, όπου f_i η συχνότητα σε Hz.

Ο συνδυασμός των σχέσεων (3.21) και (3.22) δίνει το βασικό αλγόριθμο διαχωρισμού ενέργειας διακριτών σημάτων (*Discrete Energy Separation Algorithm*):

$$\arccos\left(1 - \frac{\Psi[s(n) - s(n-1)]}{2\Psi[s(n)]}\right) \quad \Omega_i(n) \tag{3.23}$$

$$\sqrt{\frac{\Psi[s(n)]}{1 - \left(1 - \frac{\Psi[s(n) - s(n-1)]}{2\Psi[s(n)]}\right)^2}} \quad |a(n)| \tag{3.24}$$

Αυτός ο αλγόριθμος ονομάστηκε *DESA-1a*, όπου το "1" συμβολίζει την προσέγγιση με απλή διαφορά του $\dot{s}$ σε διακριτή μορφή, και το "a" την ασύμμετρη διαφορά.

Αν τώρα χρησιμοποιήσουμε και την διαφορά 'προς τα εμπρός' δηλ. το $[s(n+1) - s(n)]$ και χρησιμοποιήσουμε το μέσο όρο των δύο εξόδων από τον τελεστή, η εφαρμογή του διακριτού τελεστή στο $\dot{s}(n)$ λαμβάνεται ως:

$$\Psi[\dot{s}(n)] = \frac{\Psi[s(n) - s(n-1)] + \Psi[s(n+1) - s(n)]}{2} \\ 4a^2(n)\sin^2[\Omega_i(n)/2] \quad sin^2[\Omega_i(n)] \tag{3.25}$$

Έτσι από το 'συμμετρικό' πλέον ορισμό του αποτελέσματος του τελεστή, ο συνδυασμός των σχέσεων (3.21) και (3.25) δίνει έναν άλλο τρόπο υπολογισμού πλάτους και στιγμιαίας συχνότητας, αρκετά βελτιωμένο[6], που περιγράφεται από τις σχέσεις:

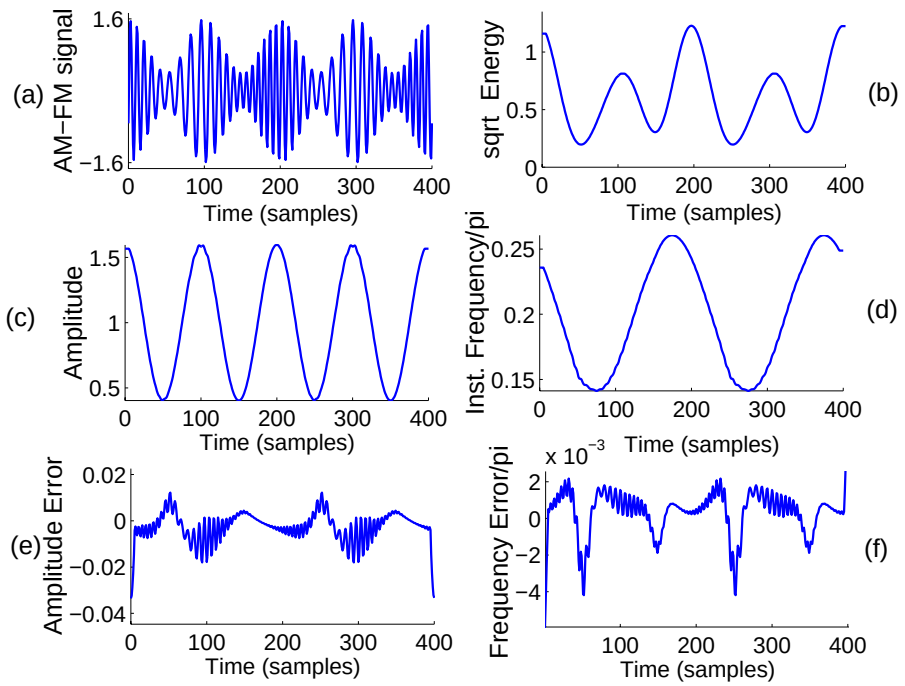
$$s(n) - s(n-1) = c(n)$$

$$\arccos\left(1 - \frac{\Psi[c(n)] + \Psi[c(n+1)]}{4\Psi[s(n)]}\right) \quad \Omega_i(n) \tag{3.26}$$

$$\sqrt{\frac{\Psi[c(n)]}{1 - \left(1 - \frac{\Psi[c(n)] + \Psi[c(n+1)]}{4\Psi[s(n)]}\right)^2}} \quad |a(n)| \tag{3.27}$$

Η παραπάνω διαδικασία αποτελεί τον αλγόριθμο *DESA-1*. Επειδή η σχέση(3.26) ισχύει για $0 < \Omega_i(n) < \pi$, ο *DESA-2* μπορεί να υπολογίσει στιγμιαίες συχνότητες $\leq 1/2$ της συχνότητας δειγματοληψίας.

Ο *DESA-1* αποδίδει παρόμοια με τον *DESA-1a*, με τον τελευταίο πάντως να δίνει λίγο μεγαλύτερα σφάλματα προσέγγισης. Ένα παράδειγμα της εκτίμησης πλάτους και συχνότητας, που επιτελεί ο *DESA-1* για ένα AM-FM σήμα φαίνεται στο Σχήμα (3.1). Στο σχήμα αποδιαμορφώνεται ένα σήμα AM-FM/WC (With Carrier) διακριτού χρόνου, για το οποίο το πλάτος $a(n)$ λαμβάνει μόνο θετικές



Σχήμα 3.1: Παράδειγμα του αλγορίθμου DESA-1 (α) AM-FM σήμα με 60% AM και 30% FM διαμόρφωση. (b) Η ρίζα της εκτιμώμενης Ενέργειας δηλ. $\sqrt{\Psi}$. (c) Το εκτιμώμενο πλάτος περιβάλλουσας δηλ.$|\hat{a}(n)|$. (d) Η εκτιμώμενη στιγμιαία συχνότητα/π, δηλ. $\hat{\Omega}_i(n)/\pi$. (e) Το σφάλμα εκτίμησης πλάτους, $|\hat{a}|$ $|a|$. (f) Το σφάλμα εκτίμησης της στιγμιαίας συχνότητας , $(\hat{\Omega}_i$ $\Omega_i)/\pi$

τιμές. Επίσης η συχνότητα μεταβάλλεται σύμφωνα με ένα συνημιτονικό σήμα διαμόρφωσης $q(n)$.

$$a(n)\cos[\phi(n)] = [1 + (0.6cos(\pi\frac{n}{50})][\cos(\pi\frac{n}{5}) + 6\sin(\pi\frac{n}{100} + \frac{\pi}{4})]$$

Το σήμα παρουσιάζει μεγάλα ποσοστά διαμόρφωσης 30% FM και 60% AM. Παρά το γεγονός αυτό βλέπουμε από το σχήμα (3.1) ότι η εκτίμηση ενέργειας και πλάτους λειτουργεί πολύ καλά για τον *DESA1*.

Τέλος υπάρχει και μια τελευταία μέθοδος διακριτής αποδιαμόρφωσης, για την οποία το διαφορικό πρώτης τάξης προσεγγίζεται με μια συμμετρική διαφορά 3 τιμών του σήματος, και συγκεκριμένα:

$$\begin{aligned} \dot{s}(n) &= [(s(n+1) \quad s(n)) + (s(n) \quad s(n \quad 1))]/2 \\ &= [s(n+1) \quad s(n \quad 1)]/2 \end{aligned} \tag{3.28}$$

,τότε η εφαρμογή του τελεστή στο $\dot{s}(n)$ δίνει προσεγγιστικά και με μικρό

σφάλμα

$$\Psi[\dot{s}(n)] = \Psi[s(n+1) \quad s(n \quad 1)]/2 \quad a^2(n)sin^4[\Omega_i(n)] \tag{3.29}$$

Η επεξεργασία των σχέσεων(3.21) και (3.29) οδηγούν σε δύο διαφορετικές εξισώσεις για τον υπολογισμό του πλάτους και της στιγμιαίας συχνότητας AM–FM σημάτων:

$$\frac{1}{2}\arccos\left[1 \quad \frac{\Psi[s(n+1) \quad s(n \quad 1)]}{2\Psi[s(n)]}\right] \quad \Omega_i(n) \tag{3.30}$$

$$\frac{2\Psi[s(n)]}{\sqrt{\Psi[s(n+1) \quad s(n \quad 1)]}} \quad |a(n)| \tag{3.31}$$

Οι δύο προηγούμενες σχέσεις συντελούν τον αλγόριθμο *DESA–2* (το "2" οφείλεται στην προσέγγιση του διαφορικού με τιμές του σήματος που διαφέρουν κατά 2 χρονικές στιγμές,σχ. (3.28)). Επειδή η σχέση(3.30) προϋποθέτει $0 < \Omega_i(n) \leq \pi/2$, ο *DESA–2* μπορεί να χρησιμοποιηθεί για να υπολογιστούν συχνότητες $\leq 1/4$ της συχνότητας δειγματοληψίας.

Στο [6] πραγματοποιήθηκαν συγκρίσεις ανάμεσα στους τρεις DESA, τόσο ως προς την απόδοση όσο και προς την πολυπλοκότητα της κάθε έκφρασης του ίδιου αλγορίθμου. Έτσι αναφορικά, και οι τρεις εντοπίζουν το πλάτος και τη στιγμιαία συχνότητα των διαμορφώσεων με λάθη μικρότερα του 1%. Οι *DESA–1* και *DESA–2* αποδίδουν σαφώς ανώτερα από τον *DESA–1a* με τον πρώτο να δίνει λίγο καλύτερα αποτελέσματα (της τάξης του 0.01% – 0.1%). Από άποψη πολυπλοκότητας και οι τρεις είναι πολύ απλοί στην υλοποίηση τους και αντιδράνε στο σήμα σχεδόν στιγμιαία. Αυτό γιατί τα παράθυρα που χρησιμοποιούνε για τον υπολογισμό μιας χρονικής στιγμής είναι μερικών δειγμάτων μόνο(4 και 5) και όχι της τάξης αρκετών ms όπως συμβαίνει με τις 'συμβατικές' μετρήσεις ενέργειας. Παρ'όλο που οι διαφορές είναι πολύ μικρές ο *DESA–2* είναι ο πιο γρήγορος από άποψη υπολογισμών και ο *DESA–1* ο πιο αργός.

3.2 Μοντελοποίηση και Αποδιαμόρφωση Φωνητικών Σημάτων με βάση AM–FM μοντέλα

3.2.1 Γραμμικά και Μη–Γραμμικά μοντέλα Παραγωγής Φωνής

Σύμφωνα με στοιχεία από την ακουστική θεωρία, [11], [2], η φωνητική οδός έχει μοντελοποιηθεί σαν ένα χρονικά μεταβαλλόμενο φίλτρο που διεγείρεται από μία ή περισσότερες πηγές, και στην έξοδο του δίνει το φωνητικό σήμα. Ο ήχος δημιουργείται με τρεις τρόπους (έμφωνο, τυρβώδη, εκρηκτικό) και κάθε

τρόπος οδηγεί σε μια ιδιαίτερη διέγερση. Παράλληλα η φωνητική οδός, με κατάλληλους σχηματισμούς από τοπικές κοιλότητες, επιβάλλει συντονισμούς στη διέγερση ενισχύοντας ορισμένες συχνότητες και περιορίζοντας κάποιες άλλες, προσδίδοντας έτσι μια ιδιαιτερότητα στον ήχο που τελικά παράγεται. Οι συντονισμοί αυτοί είναι γνωστοί ως *speech resonanses* και πρόκειται για συστήματα ταλαντωτών που μεταβάλλονται γύρω από αυτές τις συχνότητες.

Η συνηθισμένη προσέγγιση για τη μοντελοποίηση της φωνής υποθέτει ένα γραμμικό μοντέλο με συνάρτηση μεταφοράς της μορφής [11]:

$$U(z) = \frac{G}{1 \quad \sum_{k=1}^{N} a_k z^{\ k}} \tag{3.32}$$

όπου οι συντελεστές a_k και ο όρος G είναι παράμετροι σχετικοί με τη φωνητική οδό. Για αυτό το μοντέλο τα resonanses που συνήθως αποκαλούνται και *formants* της φωνής, αντιστοιχούν στους πόλους της παραπάνω συνάρτησης μεταφοράς. Για τους περισσότερους ήχους ένα μοντέλο με πόλους μόνο αρκεί, αν και για ορισμένα φωνήματα όπως ένρινα ή συριστικά είναι πιθανόν να χρειάζεται να συμπεριληφθούν και μηδενικά στη συνάρτηση μεταφοράς.

Οι ρίζες του πολυωνύμου στον παρονομαστή της (3.32) είναι είτε πραγματικές είτε ζεύγη συζυγών μιγαδικών. Ένα τυπικό formant (ή συχνότητα συντονισμού) της φωνητικής οδού μπορεί να έχει τη μορφή:

$$s_k, s_k^* = \quad \sigma_k \pm j2\pi F_k$$

,κεντρικής συχνότητας $\omega_k = 2\pi F_k$, και εύρους $2\sigma_k$ [11], ενώ οι αντίστοιχοι συζυγείς πόλοι της διακριτής μορφής θα είναι:

$$\begin{aligned} z_k, z_k^* &= e^{\ \sigma_k} e^{\pm j\omega_k} \\ &= e^{\ \sigma_k} \cos(\omega_k) \pm j e^{\ \sigma_k} \sin(\omega_k) \end{aligned}$$

Έτσι για N resonanses η s - μορφή της συνάρτησης μεταφοράς θα μπορεί να αναπτυχθεί σε άθροισμα του οποίου ο κάθε όρος θα αποτελεί τη συμβολή του κάθε ζεύγους πόλων, δηλ:

$$U(s) = \sum_{k=1}^{N} \frac{g_k(s)}{(s \quad s_k)(s \quad s_k^*)}$$

Τελικά καταλήγουμε σε μια απεικόνιση του κάθε ζεύγους, επομένως και του αντίστοιχου resonanse της φωνής:

$$R_{lin}(t) = Ae^{\ \sigma t} \cos(\omega_c t + \theta) \tag{3.33}$$

όπου ω_c η συχνότητα του formant και σ η παράμετρος που ελέγχει το εύρος του, και επομένως το σήμα της φωνής μοντελοποιείται τελικά ως το άθροισμα Ν τέτοιων όρων, όσα και τα formants του αντίστοιχου ήχου. Η βασική υπόθεση του γραμμικού μοντέλου είναι η *τοπική στασιμότητα* στο σήμα. Αυτό σημαίνει ότι οι γενικές ιδιότητες της διέγερσης και της φωνητικής οδού παραμένουν αμετάβλητες για περιόδους 10-30ms ή 1-3 περιόδους του pitch. Η κλασσική προσέγγιση λοιπόν εμπλέκει ένα αργά μεταβαλλόμενο με το χρόνο γραμμικό μοντέλο οι συντελεστές του οποίου και επομένως και τα formants της φωνής θεωρούνται σταθερά για ένα μικρό χρονικό διάστημα.

Οι σύγχρονες τάσεις και έρευνες μοντέλων παραγωγής φωνής έρχονται να καταρρίψουν την υπόθεση της τοπικής στασιμότητας. Πειραματικά και θεωρητικά αποτελέσματα [12] ,[6] αποτέλεσαν ενδείξεις για μια νέα θεώρηση σύμφωνα με την οποία τα resonanses της φωνής μπορούν να μεταβάλλονται ταχύτατα τόσο σε πλάτος όσο και σε συχνότητα ακόμη και σε πολύ μικρές κλίμακες. Η παρουσία τέτοιων διαμορφώσεων AM και FM σε φωνητικά σήματα οδήγησαν το Marago κ.α.[5] στην πρόταση ενός μοντέλου που αποτελείται από την υπέρθεση AM-FM σημάτων για ένα σήμα φωνής.

Επιγραμματικά θα αναφέρουμε αιτίες και ενδείξεις για την παρουσία διαμορφώσεων στα σήματα φωνής, που σχετίζονται με μη-γραμμικότητες και ιδέες από τη δυναμική ρευστών παρούσες κατά τη διαδικασία παραγωγής της φωνής:

1. Η *ασταθής και διαχωρίσιμη ροή αέρα* στη φωνητική οδό κατά την παραγωγή της φωνής, η οποία μεταβάλλεται ταχύτατα και ταλαντώνεται ανάμεσα στα τοιχώματα, αλλάζει τα χαρακτηριστικά που καθορίζουν τη συχνότητα των κοιλοτήτων συντονισμού. Το γεγονός αυτό προκαλεί διαμόρφωση της στιγμιαίας συχνότητας των formants σε κλίμακες πολύ μικρότερες από τη διάρκεια της περιόδου pitch[12].

2. Κατά την παραγωγή της φωνής σχηματίζονται *στρόβιλοι* που περιβάλλουν τη ροή αέρα και δρουν έτσι ως διαμορφωτές της ενέργειας της ροής. Πειραματικά έχουν εντοπισθεί από τον Teager [12].

3. *Αργές μεταβολές των στοιχείων ενός ταλαντωτή* μπορούν να οδηγήσουν σε διαμορφώσεις πλάτους ή συχνότητας. Έτσι μεταβολή των αερίων μαζών και των 'ενεργών' επιφανειών (δηλ. των επιφανειών που σχηματίζουν τις 'ηχητικές' κοιλότητες στη φωνητική οδό), που μπορεί να οφείλονται στη διαχωρίσιμη ροή προκαλούν διαμορφώσεις στους ταλαντωτές συντονισμού (δηλ.στα formants) [5].

4. Ένδειξη της παρουσίας διαμορφώσεων είναι και οι *'ενεργειακοί' παλμοί*, όπως τους βάφτισε ο Teager και παρουσιάστηκαν στην έξοδο του ενερ-

γειακού τελεστή από ένα σήμα φωνής ζωνοπερατά φιλτραρισμένο. Η παρουσία αρκετών τέτοιων παλμών και μάλιστα μέσα σε μια περίοδο pitch, έρχεται σε αντίθεση με ένα μοντέλο όπως το κλασσικό, σχέση (3.33) και προϊδεάζει για την ύπαρξη κάποιας μορφής διαμορφώσεων στα formants. Τέτοιοι παλμοί βρέθηκαν πειραματικά και στο [5].

Το νέο αυτό μη-γραμμικό μοντέλο για το σήμα της φωνής θεωρεί κάθε 'resonance' της φωνής ως AM-FM διαμορφωμένο σήμα:

$$\begin{aligned} R_{nonlin}(t) &= a(t)\cos[\phi(t)] \\ &= e^{\ \sigma t} A(t)\cos\left[\omega_c t + \omega_m \int_0^t q(\tau)d\tau + \theta\right] \end{aligned} \tag{3.34}$$

και αντίστοιχα σε διακριτή μορφή:

$$\begin{aligned} R(n) &= a(n)\cos[\phi(n)] \\ &= r^n A(n)\cos\left[\Omega_c n + \Omega_m \int_0^n q(k)dk + \theta\right] \end{aligned} \tag{3.35}$$

όπου κατά τα γνωστά Ω_c είναι η κεντρική συχνότητα του formant, $\Omega_i(n) = \Omega_c + \Omega_m q(n)$ η στιγμιαία συχνότητα, $A(n)$ το μεταβαλλόμενο πλάτος, ενώ το r έχει να κάνει με το ρυθμό εξασθένισης (ανάλογα με το $e^{\ \sigma}$).

Το μοντέλο συμπληρώνεται από την έκφραση για το συνολικό σήμα της φωνής, που αν αποτελείται από N formants είναι:

$$S(n) = \sum_{k=1}^{N} R_k(n) = \sum_{k=1}^{N} a_k(n)\cos[\phi_k(n)] \tag{3.36}$$

Σύμφωνα με την παραπάνω AM-FM μοντελοποίηση είναι δυνατή η αποδιαμόρφωση ενός formant της φωνής σε πλάτος $|a(n)|$ και στιγμιαία συχνότητα $\Omega_i(n)$, χρησιμοποιώντας κάποιον από τους αλγορίθμους ESA που περιγράφηκαν στην προηγούμενη ενότητα. Για να γίνει όμως αυτό θα πρέπει να απομονωθεί ένα formant , δηλ ένα AM-FM σήμα, από το φάσμα του σήματος με ζωνοπερατό φιλτράρισμα ώστε στη συνέχεια να εφαρμοστεί ο ενεργειακός τελεστής Ψ.

3.2.2 Ανάλυση και Αποδιαμόρφωση σε Πολλαπλές Ζώνες (MDA)

Για να μπορέσει να εφαρμοστεί ο ESA στο σήμα φωνής, στο [5] για πειραματικούς καθαρά σκοπούς, τα ζωνοπερατά φίλτρα τοποθετούνται 'με το χέρι' στην περιοχή του formant ενδιαφέροντος. Κάτι τέτοιο για αυτόματες εφαρμογές δεν αρκεί. Ένα εναλλακτικό σχέδιο είναι η αυτόματη προσαρμογή του φίλτρου

έτσι ώστε να συμπίπτει με την κεντρική συχνότητα του formant, τοποθετώντας το αρχικά στην ευρύτερη περιοχή του formant, εφαρμόζοντας τον DESA και στη συνέχεια προσαρμόζοντας την κεντρική συχνότητα του φίλτρου σύμφωνα με τη μέση εκτιμώμενη στιγμιαία συχνότητα. Μετά από κάποιες επαναλήψεις η συχνότητα του φίλτρου συγκλίνει στην κεντρική συχνότητα του formant.

Με αφορμή την ανάγκη απομόνωσης ενός AM-FM σήματος που δημιουργείται σε ένα θορυβώδες περιβάλλον ο Bovik κ.α. [1] πρότεινε μια μέθοδο βαθυπερατού φιλτραρίσματος σε πολλαπλές ζώνες. Σύμφωνα με τη τεχνική αυτή το φάσμα του σήματος φιλτράρεται 'πυκνά' από ένα "πλέγμα" βαθυπερατών φίλτρων (filter -bank) με συγκεκριμένες ιδιότητες και επιλέγεται η πιο ισχυρή έξοδος κάθε χρονική στιγμή ενώ απορρίπτονται ταυτόχρονα τα προϊόντα θορύβου που υπάρχουν στις άλλες ζώνες. Η τεχνική αυτή αποδείχτηκε ότι βελτιώνει κατά πολύ την απόδοση του ESA, η οποία για AM-FM σήματα σε θόρυβο μπορεί να είναι απρόβλεπτη και μη ερμηνεύσιμη[1]. Η όλη διαδικασία ονομάστηκε Αποδιαμόρφωση σε Πολλαπλές Ζώνες (*Multiband Demodulation Analysis*) ή *MDA*. Η MDA ανάλυση έχει καθιερωθεί σαν ένα πολύ ισχυρό εργαλείο ανάλυσης σημάτων που παρέχει πληροφορία τόσο στο φάσμα όσο και στο χρόνο. Χρησιμοποιήθηκε στο [8] για ανίχνευση της κεντρικής συχνότητας και του εύρους των formants φωνής.

3.2.3 Μια MDA διαδικασία για την αποδιαμόρφωση σημάτων φωνής

Με στόχο την ανάπτυξη νέων εργαλείων που θα τονίζουν και θα διαχωρίζουν σήματα διαφορετικής φύσης όπως η φωνή από τη σιωπή ή/και το θόρυβο, χρησιμοποιήθηκε μια διαδικασία MDA και στη συνέχεια αποδιαμόρφωση μέσω του DESA, έτσι ώστε να αποκαλυφθεί πληροφορία σε σχέση με το πλάτος και τη στιγμιαία συχνότητα που κρύβουν οι διαμορφώσεις αυτών των σημάτων. Προτείνεται ένας τρόπος που αποδιαμορφώνει την πιο ισχυρή ενεργειακή συνιστώσα του σήματος κάθε χρονική στιγμή, κάτι το οποίο είναι πολύ σημαντικό αφού έτσι απομονώνονται οι κυρίαρχες AM-FM διαμορφώσεις, δηλ. ο τρόπος που μεταβάλλεται το σήμα με το χρόνο.

Στη συνέχεια περιγράφεται αρχικά ο τρόπος με τον οποίο σχεδιάζεται και λειτουργεί η προτεινόμενη MDA ανάλυση. Βασικό στοιχείο για το σχεδιασμό και την απόδοση της διαδικασίας είναι το φίλτρο, πολλαπλά μετατοπισμένα στη συχνότητα αντίγραφα του οποίου θα αποτελέσουν το πλέγμα που θα σαρώνει το φάσμα του κάθε χρονική στιγμή. Στατιστική ανάλυση στο [1] παρουσιάζει ως βέλτιστη λύση, φίλτρα *ελάχιστης αβεβαιότητας*.

Ο Gabor ανακάλυψε μια τάξη σημάτων που επιτυγχάνουν την ελάχιστη τιμή στην στην *αρχή της αβεβαιότητας του Fourier*, που διατυπώθηκε από τον ίδιο και σύμφωνα με την οποία $(\Delta t)(\Delta\omega) \quad 1/2$, δηλ. το γινόμενο της

διάρκειας [3] ενός φίλτρου επί το εύρος του δεν μπορούν να είναι μικρότερα από $1/2$. Τα **Gabor** φίλτρα που εισήγαγε πετυχαίνουν αυτή την ελάχιστη τιμή και για το λόγο αυτό καλούνται φίλτρα ελάχιστης αβεβαιότητας. Πρόκειται για ημίτονα τα πλάτη των οποίων είναι διαμορφωμένα από συναρτήσεις Gauss. Η κρουστική απόκριση και το φάσμα τους δίνονται από τις σχέσεις:

$$h(t) = exp[\ \ a^2t^2]cos(\omega_c t) \tag{3.37}$$

$$H(\omega) = \frac{\sqrt{\pi}}{2a}\left(exp\left[\frac{(\omega\ \ \omega_c)^2}{4a^2}\right] + exp\left[\frac{(\omega+\omega_c)^2}{4a^2}\right]\right) \tag{3.38}$$

με a μια παράμετρο της περιβάλλουσας και ω_c τη συχνότητα του διαμορφωμένου ημιτόνου του. Οι rms τιμές για τη διάρκεια ($\Delta t = 1/2a$) και για το εύρος τους($\Delta\omega = a$) δίνουν γινόμενο ακριβώς 1/2. Έτσι τα Gabor φίλτρα προκύπτουν ως βέλτιστη εκλογή για το ζωνοπερατό φιλτράρισμα αφ'ενός λόγο αυτής της ιδανικής σχέσης ανάμεσα σε συχνότητα και χρόνο και αφετέρου επειδή η μορφή του $H(\omega)$ αποφεύγει τη δημιουργία πλευρικών λοβών, που μπορεί να οδηγήσουν σε λανθασμένους παλμούς στην έξοδο του Ψ.

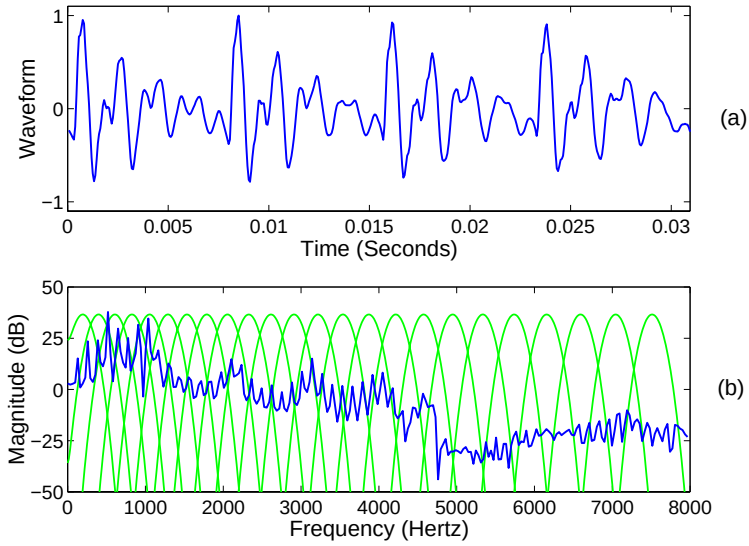
Για να δημιουργηθεί το πλέγμα των φίλτρων (δηλ. η τοποθέτηση τους στο φάσμα), πέρα από το είδος του φίλτρου χρειάζεται και μια διαδικασία επιλογής του αριθμού τους, του τρόπου διάταξης αλλά και του εύρους του κάθε φίλτρου. Μια σημαντική παρατήρηση για κάθε filter-bank είναι ότι κάθε τέτοια υλοποίηση πρέπει να προσομοιώνει με κάποιο τρόπο το ακουστικό σύστημα[2]. Για το σκοπό αυτό χρειάζεται η σχετική απόσταση μεταξύ δύο συνεχόμενων στο φάσμα φίλτρων να αυξάνεται σχεδόν *γραμμικά* για συχνότητες κάτω από 1 kHz και σχεδόν *λογαριθμικά* για συχνότητες πάνω από αυτή την τιμή. Να σημειωθεί ότι επομένως , σύμφωνα με τη σχέση (3.38), αυξάνει και το εύρος τους κατά αυτόν τον τρόπο, αφού αυξάνεται η συχνότητα του διαμορφωμένου συνημιτόνου. Υπάρχουν διάφορες τεχνικές που βασίζονται σε μια κλίμακα συχνοτήτων η οποία ικανοποιεί την προηγούμενη απαίτηση.

Η *mel-κλίμακα*, η οποία χρησιμοποιείται και στην συγκεκριμένη υλοποίηση τοποθετεί τα φίλτρα ανά διάστημα του mel-άξονα, που προκύπτει από τον άξονα συχνοτήτων με ανάθεση (frequency warping) σύμφωνα με τη σχέση:

$$f_{mel} = 2595\ \ \log(1 + f/5000)$$

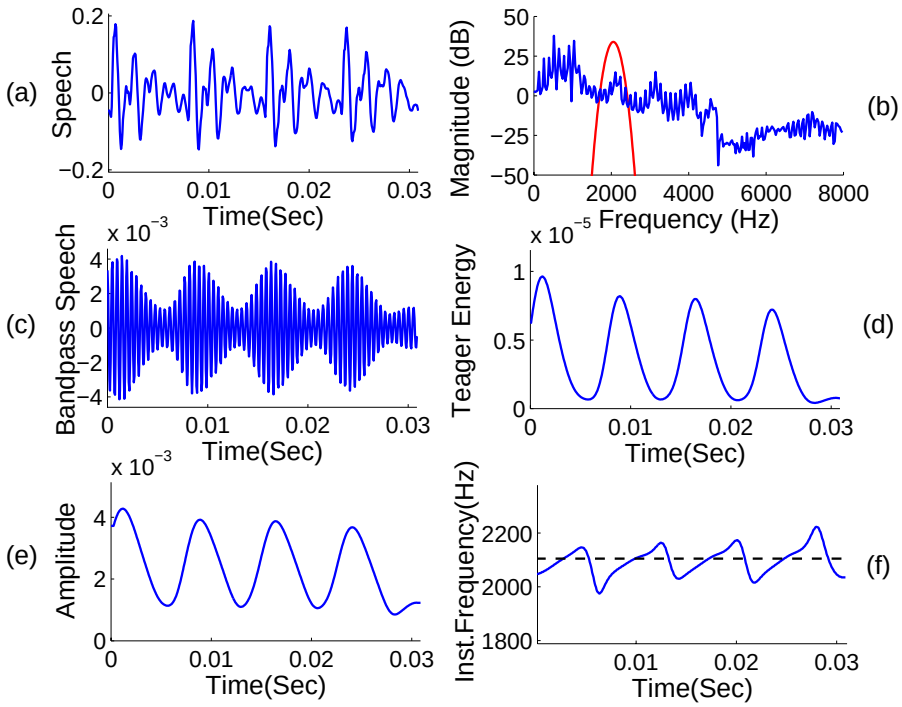
Το φάσμα του σήματος, σε ένα εύρος 0-8kHz ($F_s/2$) καλύπτεται από 24 φίλτρα Gabor (ο αριθμός έχει προκύψει πειραματικά), με τυπικό rms ενεργό εύρος περίπου 400-950Hz. Ένα παράδειγμα του πως τα φίλτρα κατανέμονται πάνω στο φάσμα των 8kHz φαίνεται στο Σχήμα (3.2) σε ένα μικρό τμήμα από ένα τυπικό φώνημα.

[3] Οι τιμές $\Delta(\ )$ είναι οι μέσες τετραγωνικές τιμές (rms) των ορισμάτων τους.

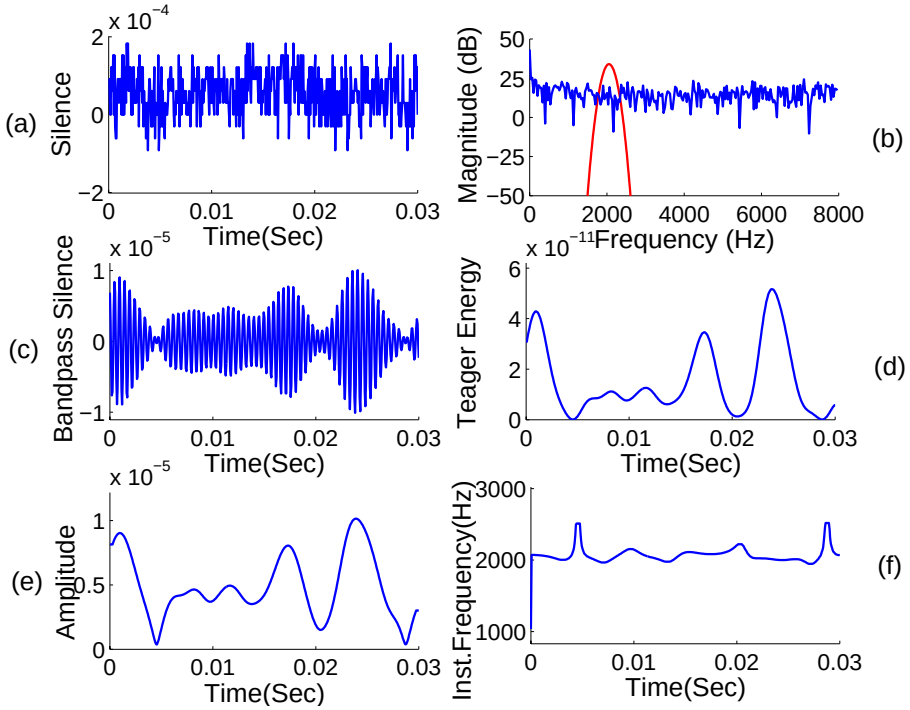


Σχήμα 3.2: Πλέγμα 24 Φίλτρων Gabor στο φάσμα του /ow/ από τη λέξη /how/ (a) Κυματομορφή του σήματος 30ms (4 περίοδοι pitch). (b) Τα φίλτρα, τοποθετημένα με τη mel-κλίμακα καλύπτουν πυκνά και επαρκώς το φάσμα των 8kHz. Το rms εύρος του πρώτου είναι περίπου 400Hz. Για τη διακριτή μορφή της σχ.(3.37) είναι $a = 400/F_s$.

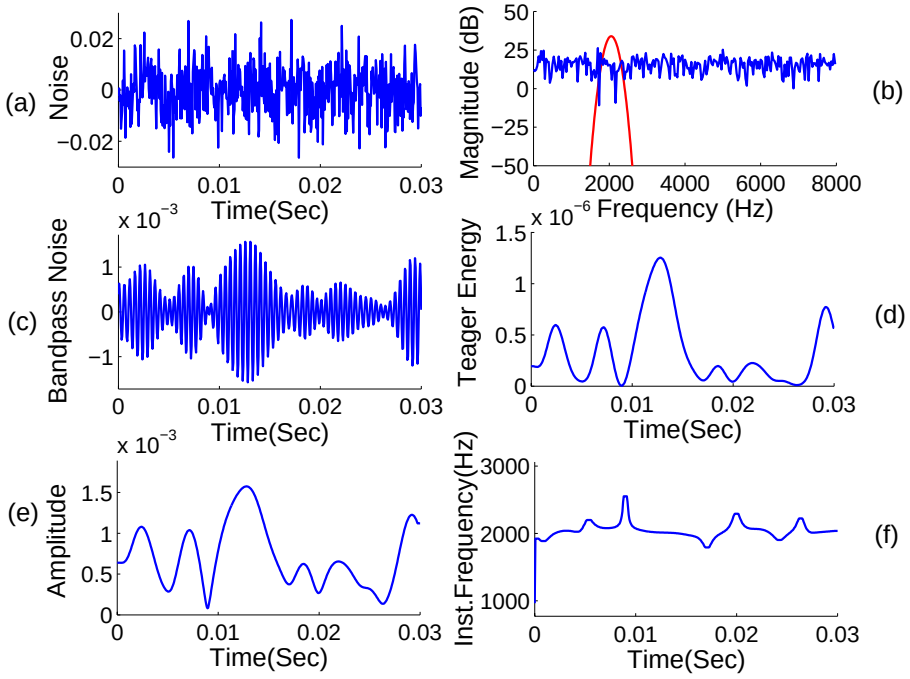
Το ζωνοπερατό φιλτράρισμα που επιτελείται από τα Gabor φίλτρα, έχει ως αποτέλεσμα την απομόνωση των ισχυρών συνιστωσών του σήματος σε κάθε ζώνη. Εφαρμογή του ενεργειακού τελεστή Ψ στην έξοδο του κάθε φίλτρου αλλά και αποδιαμόρφωση με τον αλγόριθμο DESA, μπορεί να εντοπίσει την AM και FM συνιστώσα που επικρατεί σε κάθε ζώνη, σε όλο το χρονικό διάστημα του σήματος. Για να γίνει κατανοητό αυτό, το Σχήμα (3.3) παρουσιάζει την έξοδο ενός φίλτρου του πλέγματος το οποίο είναι τοποθετημένο γύρω από το τρίτο formant του φωνήματος που παρουσιάστηκε στο Σχ. (3.2). Το (c) παρουσιάζει τη μορφή του φιλτραρισμένου σήματος από το 9ο φίλτρο του πλέγματος. Παρατηρούμε ότι η μορφή του μοιάζει με AM-FM σήμα, ενώ η έξοδος του ενεργειακού τελεστή στο (d) παρουσιάζει τους ενεργειακούς παλμούς που αποτελούν ένδειξη διαμορφώσεων [6],[12] και δικαιολογούν το μη-γραμμικό μοντέλο. Η αποδιαμόρφωση μέσω του DESA-1 αποκαλύπτει έναν ισχυρό AM παράγοντα στο (e), αφού ακολουθεί την ενεργειακή μεταβολή, αλλά και FM διαμόρφωση στο (f) αφού η στιγμιαία συχνότητα μεταβάλλεται γύρω από την κεντρική τιμή του formant (2100 kHz). Για άλλα φωνήματα, με μεγαλύτερα ποσά διαμόρφωσης οι ενεργειακοί παλμοί είναι περισσότεροι ανά περίοδο pitch ενώ για τυρβώδεις ήχους οι παλμοί που παρουσιάζονται είναι ακανόνιστοι σε σχήμα και αριθμό ανά περίοδο. Επειδή απώτερος σκοπός



Σχήμα 3.3: (a) Σήμα $s(n)$ από το φώνημα /ow/ στα 16kHz. (b) Φάσμα του σήματος με το 9ο Gabor φίλτρο του πλέγματος, γύρω από τρίτο formant ($f_3 = 2100$ Hz). (c) Το σήμα που προκύπτει από το ζωνοπερατό φιλτράρισμα στη συγκεκριμένη ζώνη. (d) $\Psi[s(n)]$. (e) Εκτιμώμενο πλάτος με αποδιαμόρφωση μέσω του DESA-1. (f) Εκτιμώμενη στιγμιαία συχνότητα $F_i = \Omega_i/2\pi T_s$, μετά από φιλτράρισμα με median 12 σημείων. Η διακεκομμένη γραμμή δείχνει τη συχνότητα του formant.



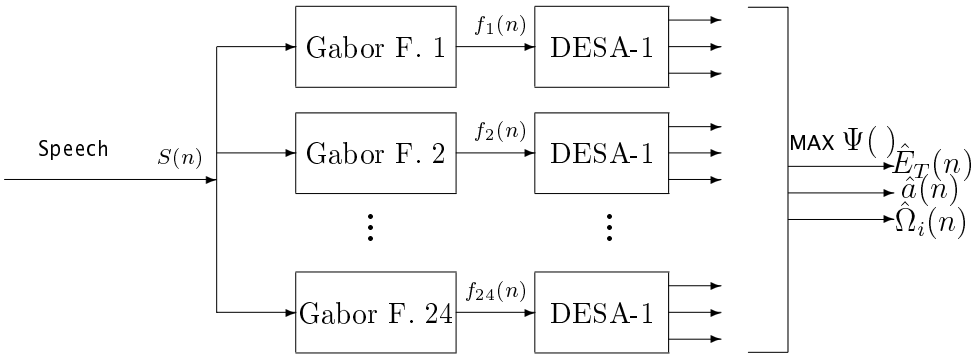
Σχήμα 3.4: (a) Σήμα $s(n)$ σιωπής (30ms) στα 16kHz. (b) Φάσμα της σιωπής με το 9ο Gabor φίλτρο του πλέγματος. (c) Το σήμα που προκύπτει από το ζωνοπερατό φιλτράρισμα στη συγκεκριμένη ζώνη. (d) $\Psi[s(n)]$. (e) Εκτιμώμενο πλάτος με αποδιαμόρφωση μέσω του DESA-1. (f) Εκτιμώμενη στιγμιαία συχνότητα $F_i = \Omega_i/2\pi T_s$, μετά από φιλτράρισμα με median 12 σημείων.



Σχήμα 3.5: (a) Θόρυβος WGN $N(n)$ (30ms), τυχαία δημιουργημένος, στα 16kHz. (b) Φάσμα θορύβου με το 9ο Gabor φίλτρο του πλέγματος. (c) Το σήμα θορύβου που προκύπτει από το ζωνοπερατό φιλτράρισμα. (d) $\Psi[N(n)]$. (e) Εκτιμώμενο πλάτος με αποδιαμόρφωση μέσω του DESA-1. (f) Εκτιμώμενη στιγμιαία συχνότητα $F_i = \Omega_i/2\pi T_s$, μετά από φιλτράρισμα με median 12 σημείων.

μας είναι η διάκριση φωνής από σιωπή σε θόρυβο ή και από θόρυβο, αξίζει να σταθούμε για λίγο στο πως αντιδράνε στη διαδικασία φιλτραρίσματος και αποδιαμόρφωσης αυτά τα σήματα. Στο Σχήμα (3.4) φαίνεται η ίδια διαδικασία για την ίδια ζώνη, όπως για τη φωνή προηγούμενα. Οι ενεργειακοί παλμοί είναι τυχαίοι, σε ακανόνιστες θέσεις κάτι που είναι υπαρκτό σε όλες τις ζώνες που φιλτράρεται το σήμα. Αυτό πιθανότατα να οφείλεται στις τυχαίες ενεργειακές εξάρσεις που παρουσιάζει η σιωπή (στην ουσία ο 'θόρυβος' που υπάρχει στο περιβάλλον ενός σήματος). Η στιγμιαία συχνότητα μεταβάλλεται αργά, με μερικά απότομα 'πετάγματα', γύρω από την κεντρική συχνότητα του φίλτρου. Παρόμοια συμπεράσματα βγαίνουν και για ένα σήμα θορύβου. Στο Σχήμα (3.5) φαίνεται λευκός θόρυβος Gauss (WGN) και η αποδιαμόρφωση του στην έξοδο του 9ου φίλτρου. Τυχαίοι παλμοί, πλάτος που ακολουθεί την ενέργεια, και στιγμιαία συχνότητα με μικρούς κυματισμούς (0,5-1,5 db).

Σαφής εικόνα δεν υπάρχει ακόμη για τις διαμορφώσεις που κρύβονται γύρω από τέτοιου είδους σήματα, αλλά μια καλή υπόθεση είναι η μοντελοποίηση τους με ένα AM-FM μοντέλο όπου τα σήματα διαμόρφωσης είναι τυχαία μεταβαλλόμενα.



Σχήμα 3.6: Αναπαράσταση της MDA διαδικασίας για την εκτίμηση των συνολικών μεγεθών. Στην έξοδο επιλέγεται η ζώνη που δίνει τη max τιμή του ενεργειακού τελεστή κάθε χρονική στιγμή (*energy tracking*), και λαμβάνεται το πλάτος $\hat{a}$ και η στιγμιαία συχνότητα $\hat{\Omega}_i$ που αντιστοιχεί στη ζώνη αυτή.

Επιστρέφοντας στην διαδικασία φιλτραρίσματος σε πολλαπλές μπάντες, τα φίλτρα που καλύπτουν πυκνά το φάσμα του σήματος αντιδρούν και δίνουν στην έξοδο τους σήματα όπως αυτά που περιγράφηκαν προηγούμενα. Αν δε, στην κάθε έξοδο εφαρμοστεί ο ενεργειακός τελεστής και στη συνέχεια αποδιαμόρφωση μέσω του ESA, προκύπτουν για κάθε ζώνη η ενέργεια, το πλάτος και η στιγμιαία συχνότητα των κυρίαρχων διαμορφώσεων. Η προσέγγιση που προτείνεται για να αναχθούμε από τις διάφορες ζώνες σε απεικονίσεις αυτών των μεγεθών για ολόκληρο το σήμα βασίζεται στην MDA ανάλυση του Bovik κ.α.[1] και περιγράφεται με βάση το Σχήμα (3.6).

Το σήμα φωνής $S(n)$ φιλτράρεται από το πλέγμα των Gabor φίλτρων και στην έξοδο του κάθε καναλιού $f_k(n), k = 1, 2 \quad 24$ εφαρμόζεται ο αλγόριθμος αποδιαμόρφωσης DESA-1. Στη συνέχεια ακολουθεί μια διαδικασία επιλογής του καναλιού που θα δώσει τη μέτρηση ανά χρονική στιγμή, σύμφωνα με κάποιο κριτήριο μεγιστοποίησης. Για να περιοριστούν τα σφάλματα στη διαδικασία αλλά και για να έχουμε πιο εύρωστα αποτελέσματα το κριτήριο αυτό είναι η έξοδος του ενεργειακού τελεστή Ψ. Το πλάτος και η συχνότητα προέρχονται από επεξεργασία των μετρήσεων του τελεστή, οπότε οποιοδήποτε σφάλμα σε εκτίμηση της ενέργειας θα αυξάνεται μέσα από τους υπολογισμούς.

Έτσι η λογική είναι η εξής: *Κάθε χρονική στιγμή n, επιλέγεται το κανάλι k με τη μέγιστη ενέργεια $E_T(n) = max(\Psi[f_k(n)])$, και δίνει το πλάτος της περιβάλλουσας $a(n)$ και τη στιγμιαία συχνότητα $\Omega_i(n)$, του αποδιαμορφωμένου σήματος*. Με το συμβολισμό E_T θα θεωρούμε από δω και στο εξής την Teager ενέργεια, δηλ. την έξοδο του ενεργειακού τελεστή:

$$E_T(n) = \Psi[f(n)] \quad a^2(n)\Omega_i^2(n)$$

Στην ουσία η μέθοδος αυτή εντοπίζει ανά χρονική στιγμή το πιο ισχυρό ενεργειακά κανάλι (*maximum energy tracking*). Βέβαια σε ένα κανάλι μπορεί να είναι

παρούσες περισσότερες από μια συνιστώσες του σήματος, δηλ. περισσότερα από ένα AM-FM σήματα, από τα οποία το ένα υπερτερεί των υπολοίπων. Η διαδικασία εντοπίζει αυτή τη μια ισχυρή συνιστώσα με μικρό σφάλμα. Επειδή ενδιαφερόμαστε στην πραγματικότητα για τη διαφορά που παρουσιάζουν η ενέργεια, το πλάτος και η συχνότητα ανάμεσα σε διαφορετικά ήδη σημάτων, τα όποια σφάλματα παίζουν μικρό ρόλο στο τελικό αποτέλεσμα. Να σημειωθεί τέλος ότι για εφαρμογές πραγματικού χρόνου η διαδικασία μπορεί να απλοποιηθεί υπολογιστικά εφαρμόζοντας τον ενεργειακό τελεστή σε κάθε κανάλι, και όχι τον ESA και αποδιαμορφώνοντας το πιο ισχυρό κάθε χρονική στιγμή.

3.3 Οι Νέες Απεικονίσεις στο Πεδίο του Χρόνου

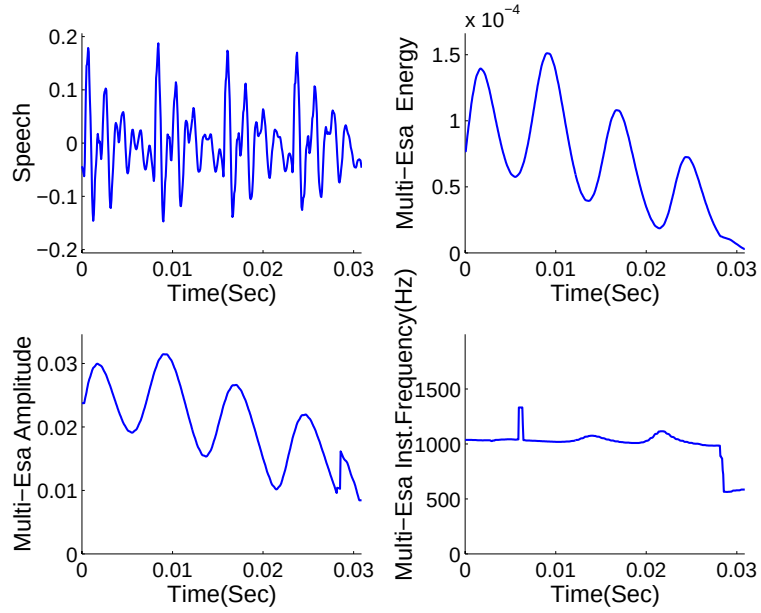
3.3.1 Ενέργεια, Πλάτος Περιβάλλουσας και Στιγμιαία Συχνότητα για το συνολικό Σήμα

Τα νέα μετρήσιμα μεγέθη φαίνονται στο Σχήμα (3.7), όπου απεικονίζονται τα αποτελέσματα της επιλογής των τιμών του καναλιού με τη μέγιστη ενέργεια, μετά από ζωνοπερατό φιλτράρισμα σε πολλαπλές ζώνες για δύο ήχους (έναν έμφωνο και έναν τυρβώδη). Οι ονομασίες που δόθηκαν είναι συμβολικές, για να εκφράζουν τον τρόπο δημιουργίας τους:

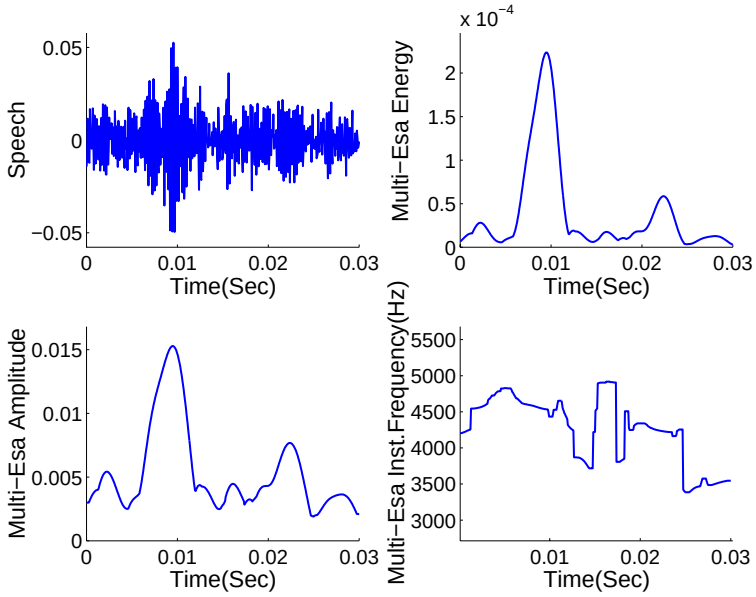
- *Multi-Esa Energy*: Η Teager ενέργεια, της πηγής που παράγει το σήμα.
- *Multi-Esa Amplitude*: Το αποδιαμορφωμένο πλάτος, από το πιο 'ισχυρό' κανάλι κάθε χρονική στιγμή, για το συνολικό σήμα.
- *Multi-Esa Inst. Frequency*: Η αποδιαμορφωμένη στιγμιαία συχνότητα για το συνολικό σήμα.

Από τα παραδείγματα του Σχ.(3.7) παρατηρούμε πόσο καλά φαίνεται να εντοπίζονται τα μεγέθη αυτά για το συνολικό σήμα. Για το /ow/ στο (a), η ενέργεια ακολουθεί τις μεταβολές της περιβάλλουσας του σήματος. Το πλάτος ακολουθεί την ενέργεια και αποκαλύπτει έτσι μια ισχυρή AM δομή που επικρατεί στην ενεργειακή μέτρηση. Όσον αφορά τη συχνότητα αυτή μεταβάλλεται με αργή διακύμανση, πρώτα γύρω από το δεύτερο formant του /ow/ ($f_2 = 1036$Hz) και στη συνέχεια γύρω από το πρώτο($f_1 = 520$Hz). Κάτι τέτοιο μοιάζει πολύ ρεαλιστικό, αφού εξετάζεται ένας έμφωνος ήχος με δύο ισχυρά πρώτα formant να κυριαρχούν στις μετρήσεις. Η διαδικασία εντόπισε δηλαδή την ενέργεια κυρίως γύρω από τα δύο πρώτα formant.

Για το /z/ στο (b) τα πράγματα είναι λίγο πιο πολύπλοκα. Η ενέργεια και το πλάτος ακολουθούνε την περιβάλλουσα του σήματος, με τις τυχαίες



(a)



(b)

Σχήμα 3.7: Τα νέα μεγέθη που προκύπτουν από την MDA ανάλυση και διαδικασία επιλογής καναλιού, για δύο διαφορετικά σήματα. (a) Το έμφωνο /ow/ από το /how/. (b) Το τυρβώδες /z/ από το /was/. Και στα δύο φαίνονται η πολυκαναλική ενέργεια, το πλάτος και η στιγμιαία συχνότητα.

ενεργειακές εξάρσεις. Η συχνότητα για το συνολικό σήμα παρουσιάζεται ακανόνιστη , με τιμές από 4kHz-5kHz. Αν σκεφτεί κανείς ότι το ισχυρό formant του είναι περίπου 4.5kHz φαίνεται ότι η διαδικασία εντόπισε την επικρατούσα συνιστώσα και αποκάλυψε, για το συνολικό πλέον σήμα τυχαία AM και FM δομή.

Μια τελευταία παρατήρηση είναι ότι οι ενεργειακοί παλμοί που εμφανίζονται για το συνολικό σήμα μπορούν να χρησιμεύσουν σε μεθόδους ανίχνευσης της μικρής μεταβολής του pitch σε έμφωνους ήχους(*pitch jitter*).

3.3.2 Απεικονίσεις Βραχέως Χρόνου

Οι προηγούμενες μετρήσεις για το συνολικό σήμα είναι χρήσιμες για την παρατήρηση και την εξαγωγή συμπερασμάτων για το σήμα σε μικρές κλίμακες (μερικών περιόδων του pitch). Για εφαρμογές όμως που απαιτούνε την επεξεργασία μεγαλύτερων τμημάτων σήματος, π.χ. ολόκληρες λέξεις ή φράσεις, όπως είναι μια εφαρμογή ανίχνευσης φωνής οι απεικονίσεις αυτές περικλείουν πολύ περισσότερη πληροφορία απ'ότι χρειάζεται. Επιπλέον είναι δύσκολο να χειριστούν και να δώσουν πρακτικά αποτελέσματα.

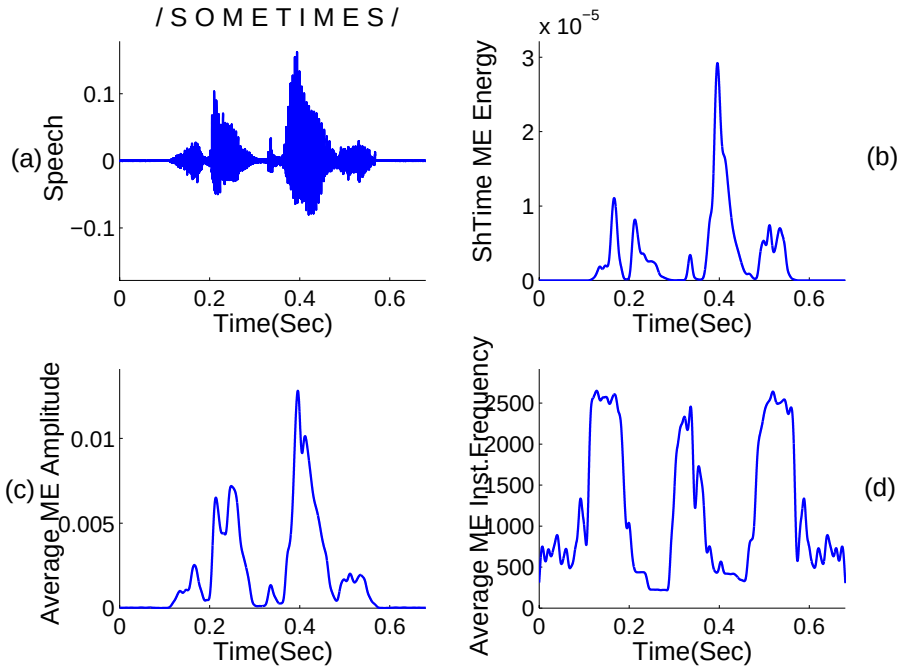
Για το σκοπό αυτό φαίνεται καλή ιδέα να ληφθούν κάποιες τοπικές μέσες τιμές των μεγεθών αυτών. Αυτό μπορεί να γίνει με επεξεργασία στο πεδίο του χρόνου σύμφωνα με τις μεθόδους βραχέως χρόνου (short-time methods) που περιγράφηκαν στην ενότητα 2.1. Έτσι από τα σήματα ενέργειας, πλάτους και στιγμιαίας συχνότητας που λαμβάνονται από τη διαδικασία MDA και την επιλογή του καναλιού με τη μέγιστη ενέργεια κάθε στιγμή (MDA-maxE), προκύπτουν με τη βοήθεια τοπικών παραθύρων ανάλυσης οι απεικονίσεις βραχέως χρόνου. Όπως θα φανεί αυτές οι απεικονίσεις είναι ικανές να διακρίνουν τα διαφορετικά είδη φωνής αλλά και τη σιωπή.

Μπορούμε να θεωρήσουμε ξανά την έκφραση για τις αναπαραστάσεις αυτές, στις οποίες λαμβάνεται ο μέσος όρος των τιμών της μέτρησης, σε ένα μετακινούμενο παράθυρο N δειγμάτων:

$$\bar{P}(n) = \sum_{m=-\infty}^{\infty} P(m) \quad w(n \quad m) \tag{3.39}$$

όπου $P(n)$ οποιοδήποτε από τα τρία μεγέθη που προέρχεται από τη διαδικασία MDA-maxE και με τον συμβολισμό $\bar{P}$ θα εννοούμε από δω και πέρα τις τιμές βραχέως χρόνου των μεγεθών αυτών. Το $w(n)$ είναι κατά τα γνωστά το παράθυρο με τη βαθυπερατή ιδιότητα, π.χ. hamming ή τετραγωνικό, που αποτελεί και το πλαίσιο ανάλυσης. Τα νέα εξισορρόπημένα μεγέθη για το συνολικό σήμα είναι:

- *ShortTime Multi-Esa Energy*, $\bar{E}_T(n)$: Η Teager ενέργεια βραχέως χρόνου μετά το πλέγμα των φίλτρων και την επιλογή του ισχυρού καναλιού.



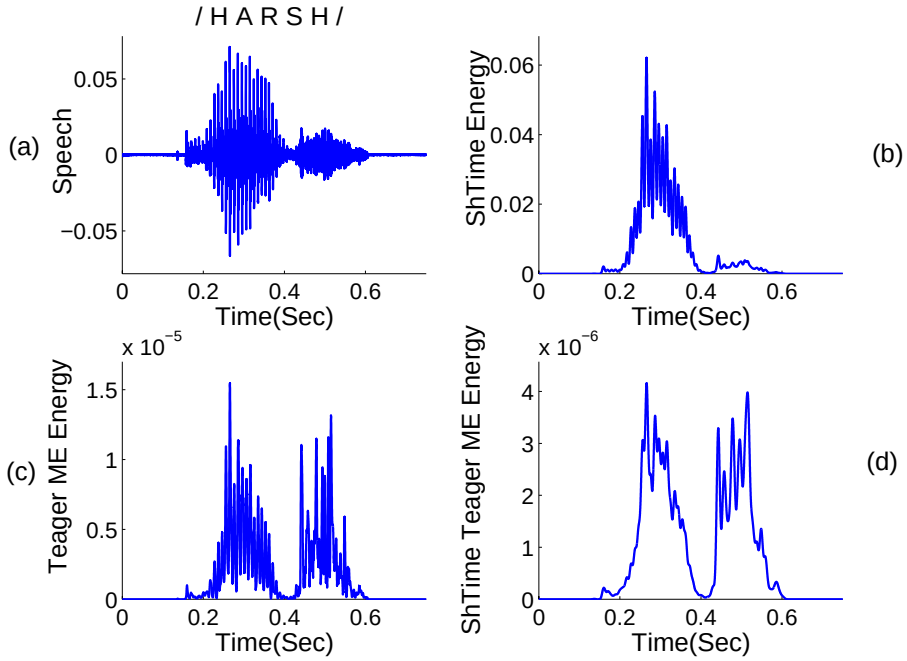
Σχήμα 3.8: Τα νέα μεγέθη που προκύπτουν μετά την επεξεργασία της διαταραχής /sometimes/ στα 16kHz, μέσω MDA-maxE και εκτιμήσεις βραχέως χρόνου των προϊόντων της διαδικασίας για το συνολικό σήμα. (a) Η λέξη /sometimes/ με σιωπή σε αρχή και τέλος. (b) ShortTime Multi-Esa Energy. (c) Average Multi-Esa Amplitude. (d) Average Multi-Esa Inst. Frequency. Με το ME εννοείται Multi-Esa.

- *Average Multi-Esa Amplitude*, $\bar{a}(n)$: Το μέσο αποδιαμορφωμένο πλάτος από το πιο 'ισχυρό' κανάλι.
- *Average Multi-Esa Inst.Frequency*, $\bar{\Omega}_i(n)$: Η μέση μέτρηση της αποδιαμορφωμένης στιγμιαίας συχνότητας.

Στο Σχήμα (3.8) φαίνονται οι μεταβολές των μεγεθών αυτών για μια διαταραχή (0.7ms στα 16kHz). Πρόκειται για τη λέξη /sometimes/ με δύο ισχυρά τυρβώδεις ήχους σε αρχή και τέλος. Παρατηρούμε ότι προκύπτουν αρκετά ομαλές αναπαραστάσεις που παρουσιάζουν με απλό τρόπο τη χρονική μεταβολή του σήματος ανάμεσα στα διαφορετικά τμήματα του. Η ενέργεια[4] και το πλάτος είναι ισχυρά για έμφωνους ήχους ενώ η συχνότητα είναι αρκετά μεγάλη για άφωνους-τυρβώδεις ήχους, όπως είναι το /s/ και το /t/.

Περισσότερα παραδείγματα για αυτές τις νέες απεικονίσεις θα δοθούν

[4]Όταν αναφερόμαστε σε ενέργεια, πλάτος και συχνότητα εννοούνται τα νέα μεγέθη, που υπολογίζονται με τη διαδικασία που περιγράφηκε, εκτός και αν δηλώνεται διαφορετικά στο κείμενο.



Σχήμα 3.9: Σύγκριση των ενεργειακών μετρήσεων με τον κλασσικό και το νέο τρόπο για τη λέξη /harsh/. (a) Η διαταραχή σε περιβάλλον σιωπής. (b) Ενέργεια βραχέως χρόνου (c) Multi-Esa Teager Energy. (d) ShortTime Multi-Esa Teager Energy.

σε επόμενο κεφάλαιο. Για να φανούν οι προοπτικές που ανοίγουν οι νέες μετρήσεις, στο Σχήμα (3.9) φαίνεται μια σύγκριση των τριών ενεργειακών μετρήσεων. Της κλασσικής ενέργειας βραχέως χρόνου[5], της ενέργειας Teager με την επιλογή του ισχυρότερου καναλιού μετά το πλέγμα των φίλτρων και της Teager ενέργειας βραχέως χρόνου. Για τις δύο τελευταίες χρησιμοποιούμε τον όρο 'Teager' αφενός για να τις διακρίνουμε και αφετέρου λόγο του ενεργειακού τελεστή Ψ του Teager που χρησιμοποιούν. Όπως φαίνεται και από το (d) η νέα ενεργειακή μέτρηση 'τονίζει' περισσότερο τους άφωνους ήχους. Μια σημαντική παρατήρηση είναι ότι επειδή σύμφωνα με τον ορισμό της, εντοπίζει την ενέργεια της πηγής που παράγει το σήμα, αποδίδει περίπου παρόμοιο επίπεδο ενέργειας στον άφωνο τελικό ήχο με το έμφωνο τμήμα της λέξης.

Τέλος πρέπει να σημειωθεί ότι αναφορά αλλά και δοκιμή σε μια μέθοδο μέτρησης της Teager ενέργειας με τη βοήθεια παραθύρων ανάλυσης και τοπικών μέσων, γίνεται και από τον Ying κ.α. στο [14]. Η διαφορά είναι ότι εκεί χρησιμοποιείται ο τελεστής Ψ με τον ορισμό του στο σήμα και οι έξοδοι του αθροίζονται για τα δείγματα ενός παραθύρου. Η αναφορά αυτή αποτέλεσε

[5] Βλ. Κεφ.2 , Ενότητα 2.1.1

σίγουρα έμπνευση για την διαδικασία που προτείνεται εδώ.

3.4 Λίγα λόγια για την Fractal (κλασματική) διάσταση των Σημάτων

Στα πλαίσια των μη-γραμμικών τεχνικών επεξεργασίας σημάτων που έρχονται στο προσκήνιο, θα γίνει μια αναφορά στη *fractal* διάσταση και σε έναν τρόπο υπολογισμού μιας μέτρησης της βραχέως χρόνου. Η χρήση της fractal διάστασης ως εργαλείο ποσοτικοποίησης της πολυπλοκότητας της κυματομορφής ενός σήματος έγινε από τους Marago και Potamiano [7].

Ανάμεσα στα μη-γραμμικά φαινόμενα που συνοδεύουν τη δημιουργία και την διάδοση του ηχητικού κύματος στη φωνητική οδό κατατάσσεται και ένα ποσοστό αστάθειας ή διαταραχής (*turbulence*) της ροής του αέρα. Αιτίες και ενδείξεις για αυτή την ασταθή ροή δίνονται στο [7] και έχουν να κάνουν με την αεροδυναμική της φωνής. Τα διάφορα ποσοστά αστάθειας σε ένα σήμα καθώς και γενικά ο μεγάλος βαθμός κατάτμησης και γεωμετρικής πολυπλοκότητας της κυματομορφής τους, ειδικά σε τυρβώδεις και άφωνους ήχους μπορούν να ποσοτικοποιηθούν με τη βοήθεια των fractals.

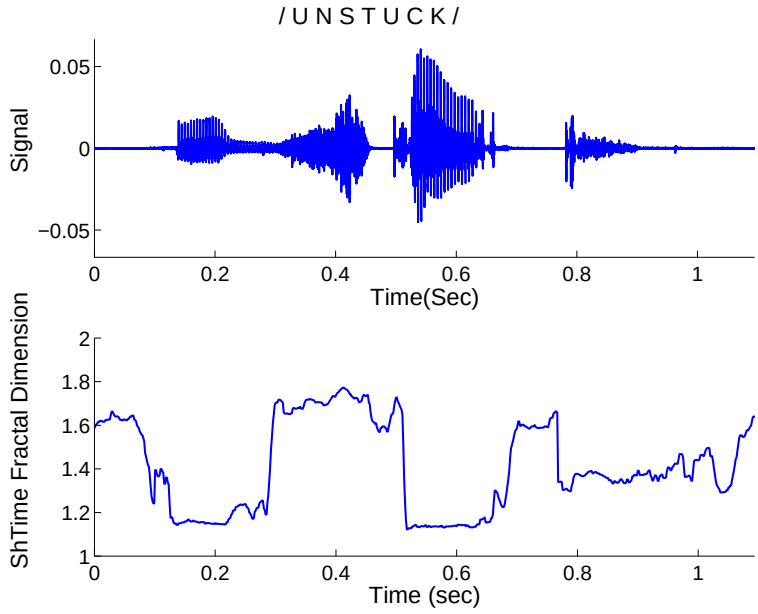
Σύμφωνα με τον **Mandelbrot**, που εισήγαγε την θεωρία αυτών των μαθηματικών συνόλων η fractal διάσταση είναι μια *"διάσταση ανάμεσα στις διαστάσεις"*. Για ηχητικά σήματα 1-D (δηλαδή με τοπολογική διάσταση 1), η διάσταση αυτή της κυματομορφής τους λαμβάνει τιμές $1 \leq F_d \leq 2$, και αν δε $1 < F_d$ το σήμα είναι fractal. Ο υπολογισμός της fractal διάστασης απαιτεί μορφολογική επεξεργασία της γραφικής παράστασης ενός σήματος, που σημαίνει αντιμετώπιση του ως δυαδική εικόνα και τεχνικές μαθηματικών συνόλων.

Στο [7] προτείνεται ένας αλγόριθμος υπολογισμού της *fractal διάστασης σε πολλαπλές κλίμακες*, όπου για μια κλίμακα ε η F_d δίνεται από τη σχέση:

$$F_d = 2 \quad \lim_{\varepsilon \to 0} \frac{\log[A_B(\varepsilon)]}{\log(\varepsilon)} \tag{3.40}$$

Η $A_B(\varepsilon)$ είναι η 'περιοχή' που προκύπτει από επαναληπτικές διαστολές (dilations) της γραφικής του σήματος σε κλίμακα ε [7]. Έτσι η διάσταση μπορεί να μετρηθεί με τη βοήθεια της κλίσης της $(\log \varepsilon, \log[A_B(\varepsilon)])$.Για τη διάσταση σε πολλαπλές κλίμακες ο αλγόριθμος περιλαμβάνει ένα κινούμενο παράθυρο για τη μέτρηση αυτής της κλίσης. Περισσότερες λεπτομέρειες αλλά και η υλοποίηση του αλγορίθμου για ψηφιακά σήματα δίνονται στο [7].

Αυτό που μας ενδιαφέρει άμεσα εδώ είναι μια μέτρηση της fractal διάστασης σαν απεικόνιση βραχέως χρόνου (*short-time fractal dimension*) με σκοπό να



Σχήμα 3.10: Κυματομορφή της λέξης /unstuck/ στα 16kHz και μέτρηση της fractal διάστασης βραχέως χρόνου. Για τη μέτρηση χρησιμοποιήθηκε παράθυρο 15ms (240 δειγμάτων) κάθε 1.5 ms. (Ο υπολογισμός σε κάθε παράθυρο γίνεται σε πολλαπλές κλίμακες με $\varepsilon = 1, w = 5$). Ακολουθεί median-3 σημείων που φιλτράρει την απεικόνιση.

χρησιμοποιηθεί ως μέσο διάκρισης διαφορετικών φωνητικών ή απλά ακουστικών σημάτων. Ανάλογα με το ποσό αστάθειας που παρουσιάζουν οι φωνητικοί ήχοι η short-time fractal διάσταση είναι: (i) μικρή για φωνήεντα, (ii) μεσαία για έμφωνα συριστικά, π.χ. /v/,/th/ και (iii) μεγάλη για παύσεις, άφωνα συριστικά και κάποια έμφωνα. Επιπλέον εκτός από φωνή η μέτρηση μπορεί να επεκταθεί και σε σιωπή ή και θόρυβο, αφού και αυτά τα σήματα παρουσιάζουν υψηλό βαθμό γεωμετρικής πολυπλοκότητας.

Ένα παράδειγμα υπολογισμού της για μια τυπική λέξη με σιωπή σε αρχή και τέλος φαίνεται στο σχήμα (3.10). Πρόκειται για τη διαταραχή /unstuck/ με δείγματα στα 16kHz. Η διαδικασία υπολογισμού είναι η εξής: Το σήμα χωρίζεται σε παράθυρα 15ms και σε κάθε τέτοιο τμήμα υπολογίζεται η fractal διάσταση σε πολλαπλές κλίμακες. Υπολογίζεται δηλαδή η κλίση της $(\log \varepsilon, \log[A_B(\varepsilon)])$ σε κλίμακες $\varepsilon, \ldots, \varepsilon+w$, όπου w το μέγιστο παράθυρο. Το παράθυρο ανάλυσης ανανεώνεται με βήμα 1.5ms και αυτό έχει ως αποτέλεσμα μια υποδειγματοληψία σε σχέση με το σήμα $f_s = F_s/24$.

Περισσότερες συγκρίσεις και διαφορές σε σχέση με τις άλλες απεικονίσεις που έχουν αναφερθεί, γραμμικές και μη-γραμμικές, θα γίνουν σε επόμενο κεφάλαιο. Εδώ αρκεί να πούμε, ότι η μέτρηση της short-time fractal dimen-

sion έχει χρησιμοποιηθεί ήδη ως πρόσθετο χαρακτηριστικό σε συστήματα αυτόματης αναγνώρισης φωνής [7].

3.5 Ανακεφαλαίωση

Με κεντρικό σημείο αναφοράς τη μη-γραμμική φύση τους παρουσιάστηκαν κάποιες σύγχρονες τεχνικές επεξεργασίας σημάτων. Ο σκοπός ήταν να οδηγηθούμε μέσα από τη χρήση τέτοιων τεχνικών σε νέες απεικονίσεις στο πεδίο του χρόνου, ικανές να διαχωρίσουν τα φωνητικά από άλλα σήματα. Ξεκινώντας με μια διαφορετική θεώρηση της ενέργειας σε σχέση με την πηγή που παράγει το σήμα και μοντελοποιώντας τα βασικά συστατικά του φωνητικού σήματος (formants) με τη βοήθεια AM-FM σημάτων, παρουσιάστηκε ο αλγόριθμος διαχωρισμού της ενέργειας (ESA) αλλά και μια διαδικασία αποδιαμόρφωσης σε πολλαπλές ζώνες (MDA), με τη βοήθεια ενός πλέγματος ζωνοπερατών φίλτρων (filter bank). Προτάθηκε μια μέθοδος εξαγωγής συνολικών αναπαραστάσεων για το σήμα μετά το φιλτράρισμα η οποία έδωσε αρκετά ενδιαφέροντα αποτελέσματα και τελικά κάποιες νέες απεικονίσεις βραχέως χρόνου για τη μεταβολή της περιβάλλουσας αλλά και το φασματικό περιεχόμενο του σήματος. Μένει τώρα να δούμε πως θα εμπλακούν αποδοτικά αυτά τα νέα εργαλεία σε μια νέα μέθοδο ανίχνευσης φωνής.

Κεφάλαιο 4

Μια μέθοδος ανίχνευσης φωνής από σιωπή με βάση μη-γραμμικά εργαλεία

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν νέα εργαλεία τα οποία προέκυψαν από μη γραμμική μοντελοποίηση και επεξεργασία των σημάτων φωνής. Οι εφαρμογές τους μπορεί να είναι πολλές και ακολουθώντας τις σύγχρονες τάσεις προς τη μη-γραμμικότητα , να δοκιμαστούν στη θέση κλασσικών εργαλείων επεξεργασίας σημάτων. Εδώ προσπαθήσαμε να εξετάσουμε τη συμβολή τους στο ζήτημα της ανίχνευσης φωνής σε ένα ακουστικό περιβάλλον σιωπής.

Οι δοκιμές απέδειξαν ότι οι νέες απεικονίσεις λειτουγούν με τρόπο παρόμοιο με αυτόν που λειτουργούν οι μετρήσεις ενέργειας βραχέως χρόνου και μέσου ρυθμού zero-crossings, που περιγράφηκαν στο 2ο κεφάλαιο. Έτσι το πλάτος περιβάλλουσας ή η ενέργεια Teager που προκύπτουν από την ανάλυση σε πολλαπλές μπάντες ενός σήματος μπορούν να αποτελέσουν ένδειξη έμφωνων ήχων και η στιγμιαία συχνότητα να αποτελέσει κάποια μορφή επιβεβαίωσης άφωνης αλλά έντονης φασματικά διέγερσης. Σε ορισμένες περιπτώσεις ακόμη και η ενέργεια μόνη της, με τη νέα θεώρηση της που περιλαμβάνει και την έννοια της στιγμιαίας συχνότητας μπορεί να είναι ικανή να διακρίνει ένα τμήμα φωνής.

Για το σκοπό αυτό οι απεικονίσεις αυτές εμπλέκονται σε μια μέθοδο που βασίζεται στις ίδιες αρχές με την κλασσική προσέγγιση ανίχνευσης των ορίων της φωνής που παρουσιάσανε οι Rabiner και Sambur [10]. Όπως παρουσιάστηκε αναλυτικά σε προηγούμενο κεφάλαιο, ο κλασσικός αυτός αλγόριθμος πάσχει στην ανίχνευση ασθενών φωνημάτων ή παύσεων μεγάλης διάρκειας στην αρχή και στο τέλος λέξεων καθώς και από ισχυρή παρουσία θορύβου. Η προσπάθεια επικεντρώθηκε στο κατά πόσο η νέα μέθοδος, ή καλύτερα τα νέα εργαλεία, δίνουν καλύτερα αποτελέσματα στην ανίχνευση των ορίων μεμονωμένων λέξεων.

Τα αποτελέσματα είναι πολύ ενθαρρυντικά καθώς η μέθοδος παρουσίασε

μια βελτιωμένη απόδοση σε αντάλλαγμα με μεγαλύτερη υπολογιστική πολυπλοκότητα. Προχωρώντας περισσότερο ,δοκιμάστηκε διάκριση *από* θόρυβο αλλά και διάκριση *σε* θόρυβο, με αρκετά χρήσιμα συμπεράσματα και πρακτική σημασία.

Στη συνέχεια ακολουθεί μια παρουσίαση πειραμάτων και αποτελεσμάτων σχετικά με τη νέα μέθοδο. Αρχικά γίνεται μια σύγκριση με τα κλασσικά εργαλεία και, κυρίως ποιοτική με βάση διαγράμματα διαφορετικών χρονικών απεικονίσεων μικρών διαταραχών. Ακολουθεί η περιγραφή της μεθόδου και δοκιμάζονται διαφορετικοί συνδυασμοί των νέων εργαλείων που θα δώσουν το καλύτερο αποτέλεσμα. Γίνονται συγκρίσεις με τον αλγόριθμο των Rabiner-Sambur και παρουσιάζονται και κάποια ποσοτικά αποτελέσματα. Τέλος εμπλέκεται και η έννοια του θορύβου από δύο οπτικές πλευρές. Αφ'ενός οι δύο αλγόριθμοι δοκιμάζονται όσον αφορά την αντοχή τους στο θόρυβο και αφ'ετέρου προτείνεται ένας τρόπος διάκρισης τριών ειδών σημάτων, σιωπής-φωνής-θορύβου, με χρήση σύγχρονων και κλασσικών μετρήσεων.

4.1 Ποιοτική Σύγκριση Κλασσικών και Νέων Απεικονίσεων στο Χρόνο

Οι νέες απεικονίσεις στο πεδίο του χρόνου που προέκυψαν με τη χρήση μηγραμμικών, σύγχρονων τεχνικών παρέχουν πληροφορία τόσο για τη μεταβολή της έντασης ενός σήματος με το χρόνο όσο και για το φασματικό περιεχόμενο του. Με τον όρο φασματικό περιεχόμενο εννοούμε είτε το ρυθμό μεταβολής του σήματος, είτε τα επίπεδα συχνοτήτων του είτε το βαθμό διέγερσης του.

Οι μετρήσεις της περιβάλλουσας του σήματος αποκαλύπτουν μεταβάσεις από άφωνα σε έμφωνα τμήματα, με την απότομη αύξηση του μεγέθους αλλά και γενικά μετάβαση από ένα φώνημα σε ένα άλλο διαφορετικής φύσεως (π.χ. από φωνήεν σε έμφωνο σύμφωνο). Για το σκοπό αυτό το κλασσικό *Μέσο Πλάτος*, M, δίνεται από την Εξ. (2.4) ενώ το *Μέσο Multi-Esa Πλάτος*, $\bar{a}$, υπολογίζεται με βάση το αποδιαμορφωμένο μέσω ESA πλάτος του καναλιού του πλέγματος των φίλτρων με τη μέγιστη απόκριση στον ενεργειακό τελεστή. Αυτά τα μεγέθη εκφράζουν παρόμοια χαρακτηριστικά του σήματος σε αντίθεση με τις δύο μετρήσεις για την ενέργεια. Η κλασσική *Ενέργεια Βραχέως Χρόνου*, E, δίνεται από την Εξ. (2.3) και εκφράζει την ενέργεια που 'κουβαλάει' το σήμα ενώ η *Multi-Esa Teager Ενέργεια Βραχέως Χρόνου*, E_T, υπολογίζεται βασισμένη στη θεώρηση του ενεργειακού τελεστή Ψ που εντοπίζει την ενέργεια της πηγής που δημιουργεί το σήμα. Για σήματα φωνής αυτή η πηγή είναι ολόκληρο το φωνητικό σύστημα.

Από την άλλη, οι απεικονίσεις που εκφράζουν τι φασματικές ιδιότητες του σήματος μπορούν να αποτελέσουν ένδειξη άφωνης διαταραχής. Ο *Μέσος*

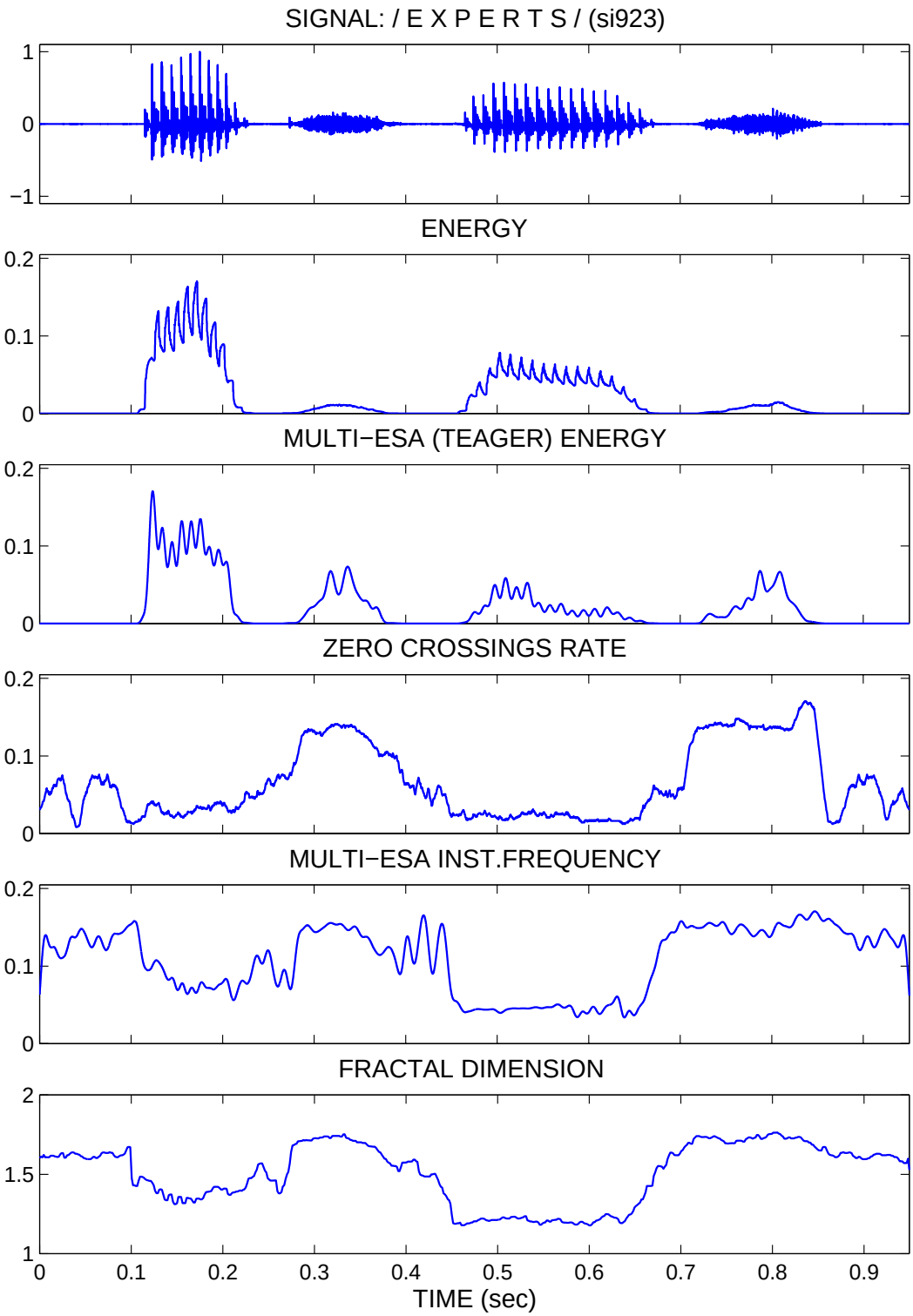
ρυθμός Zero Crossings, Z, δίνεται από την Εξ. (2.7) και όπως έχει αναφερθεί είναι μεγάλος για άφωνα τμήματα. Τέτοια τμήματα χαρακτηρίζονται και από μεγάλες συχνότητες, οπότε και η *Μέση Multi-Esa Συχνότητα*, $\bar{\Omega}_i$, που υπολογίζεται με τη διαδικασία αποδιαμόρφωσης και επιλογής του ισχυρότερου καναλιού είναι μεγάλη. Τέλος η *Κλασματική (Fractal) Διάσταση Βραχέως Χρόνου*, , D_f, που λαμβάνει τιμές στο διάστημα $(1, 2]$, είναι μεγαλύτερη για σήματα με υψηλό βαθμό διαταραχής στην κυματομορφή τους, όπως είναι οι άφωνοι ήχοι και κάποια έμφωνα συριστικά απ'ότι για έμφωνους ήχους.

Ακολουθούνε κάποια διαγράμματα αυτών των μετρήσεων για να φανεί αν οι νέες μετρήσεις τονίζουν καλύτερα τις διαφορές ανάμεσα στα διαφορετικά είδη φωνής αλλά κυρίως αν διαχωρίζουν πιο έντονα τη φωνή από την περιβάλλουσα σιωπή. Οι εξεταζόμενες διαταραχές είναι μεμονωμένες λέξεις με σιωπή στην αρχή και στο τέλος. Προέρχονται από φράσεις της βάσης σημάτων TIMIT, με αντρικά και γυναικεία φωνητικά. Ο ρυθμός δειγματοληψίας είναι 16kHz για όλες τις διαταραχές, ενώ όλα τα σχετικά μεγέθη , που είναι απεικονίσεις βραχέως χρόνου, δημιουργούνται με παράθυρα των 240 δειγμάτων (15ms). Επειδή για θέματα ανίχνευσης μας ενδιαφέρουν όχι τόσο οι απόλυτες τιμές τους αλλά η διαφορά των επιπέδων των τιμών ανάμεσα στα διαφορετικά ηχητικά τμήματα, η κυματομορφή του σήματος κανονικοποιείται ως προς 1, ενώ οι διαφορετικές χρονικές απεικονίσεις του φαίνονται κανονικοποιημένες ως προς το λόγο $max(P)/max(s)$ [1], με εξαίρεση τη μέτρηση της fractal διάστασης.

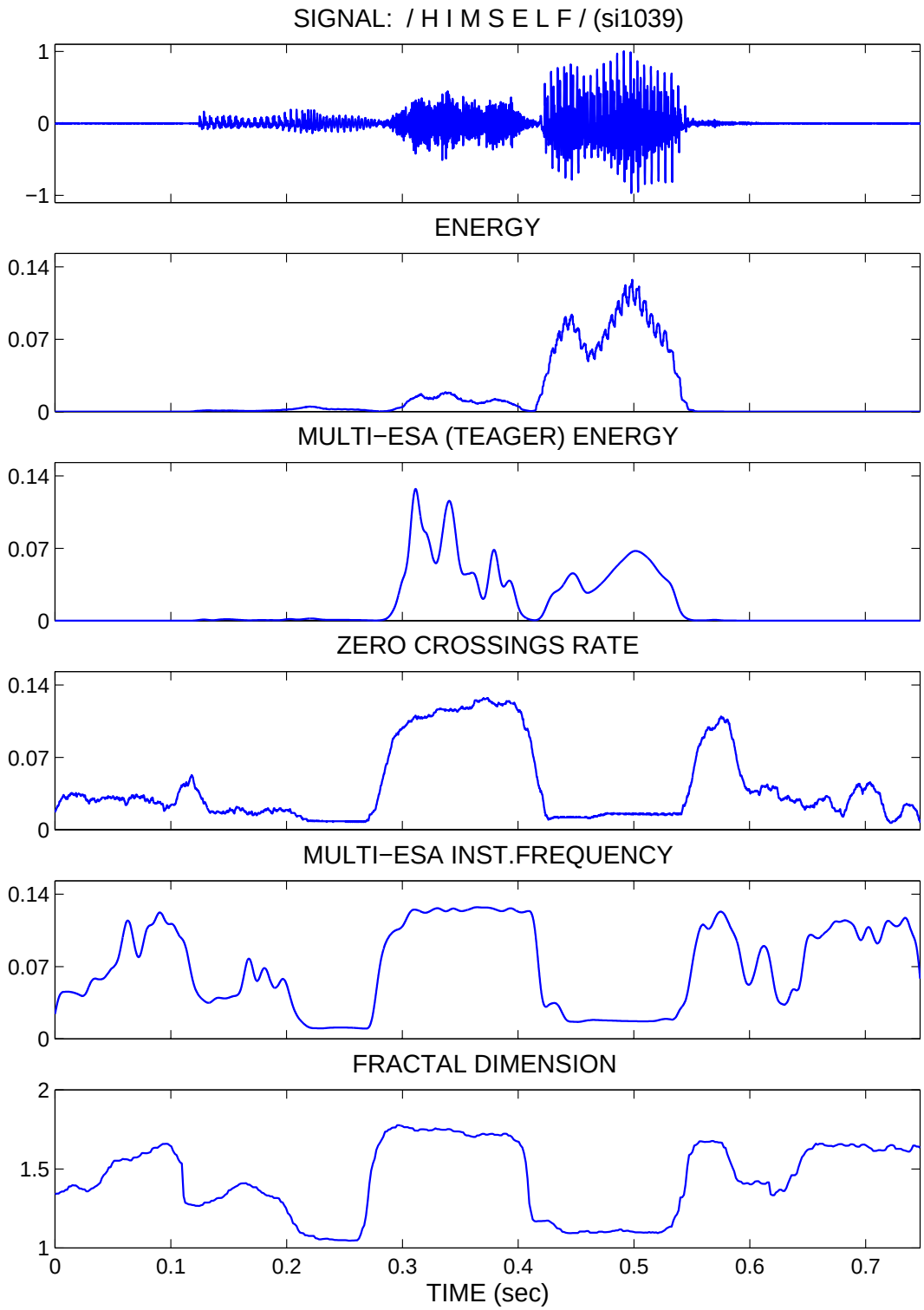
Τα τέσσερα πρώτα σχήματα είναι με μετρήσεις της ενέργειας ενώ τα επόμενα τρία είναι με μετρήσεις πλάτους, με τον κλασσικό αλλά και με τον νέο τρόπο. Γενικά η νέα μέτρηση της ενέργειας τονίζει περισσότερο από την κλασσική, έντονα τυρβώδεις ήχους όπως τα /s/, στη λέξη /experts/ του Σχ. (4.1), στη λέξη /himself/ του Σχ. (4.2) αλλά και στην αρχή /ceramic/ του Σχ. (4.5). Επίσης με τον ίδιο τρόπο αυξάνει τη διαφορά ανάμεσα σε μια αρχική άφωνη παύση και στη σιωπή, όπως συμβαίνει για το /t/ στην αρχή του /trying/, στο Σχ. (4.4). Αυτό οφείλεται στο γεγονός ότι η Teager ενέργεια λαμβάνει υπόψιν της και τον παράγοντα της συχνότητας στην ενεργειακή μέτρηση, με αποτέλεσμα άφωνα μεν φωνήματα με έντονη όμως διέγερση να εμφανίζονται με μεγάλη ενέργεια.

Οι νέες εκφράσεις για το πλάτος άλλοτε συμπεριφέρονται παρόμοια με την κλασσική έκφραση του μέσου πλάτους, και άλλοτε παρουσιάζονται πιο σύνθετα είτε λόγω σφαλμάτων στην αποδιαμόρφωση είτε διότι ίσως αποκαλύπτουν κάποια πιο περίπλοκη δομή από μια απλή καταγραφή της περιβάλλουσας του σήματος. Παρατηρούμε ότι φωνήματα που καταγράφουν και υψηλά επίπεδα συχνοτήτων παρουσιάζονται με μειωμένο πλάτος σε σχέση με την κλασσική μέτρηση, όπως το /v/ στο /convenient/, του Σχ. (4.8). Όμως

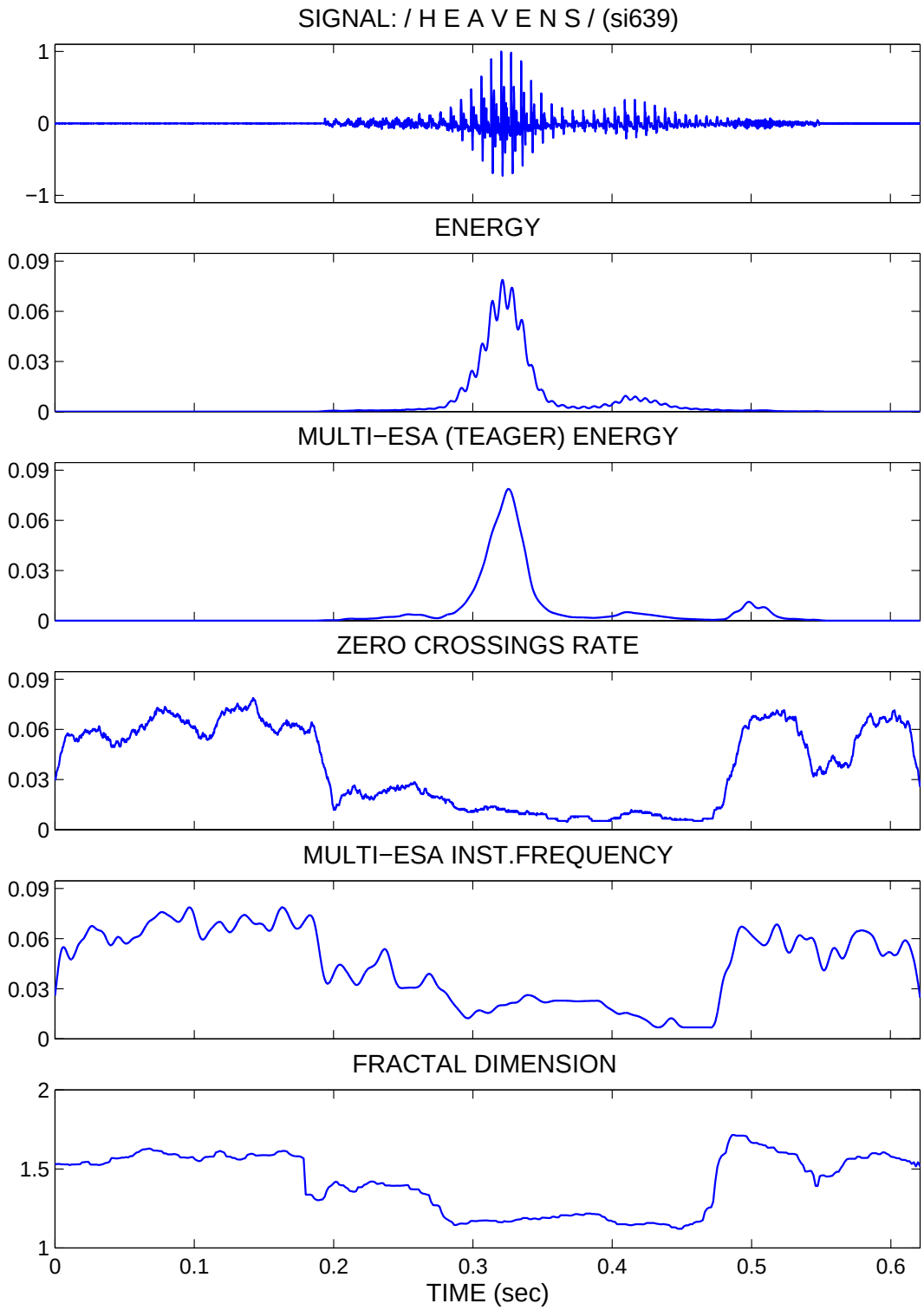
[1] όπου $max(P)$ η μέγιστη τιμή της μέτρησης και $max(s)$ η μέγιστη τιμή του σήματος



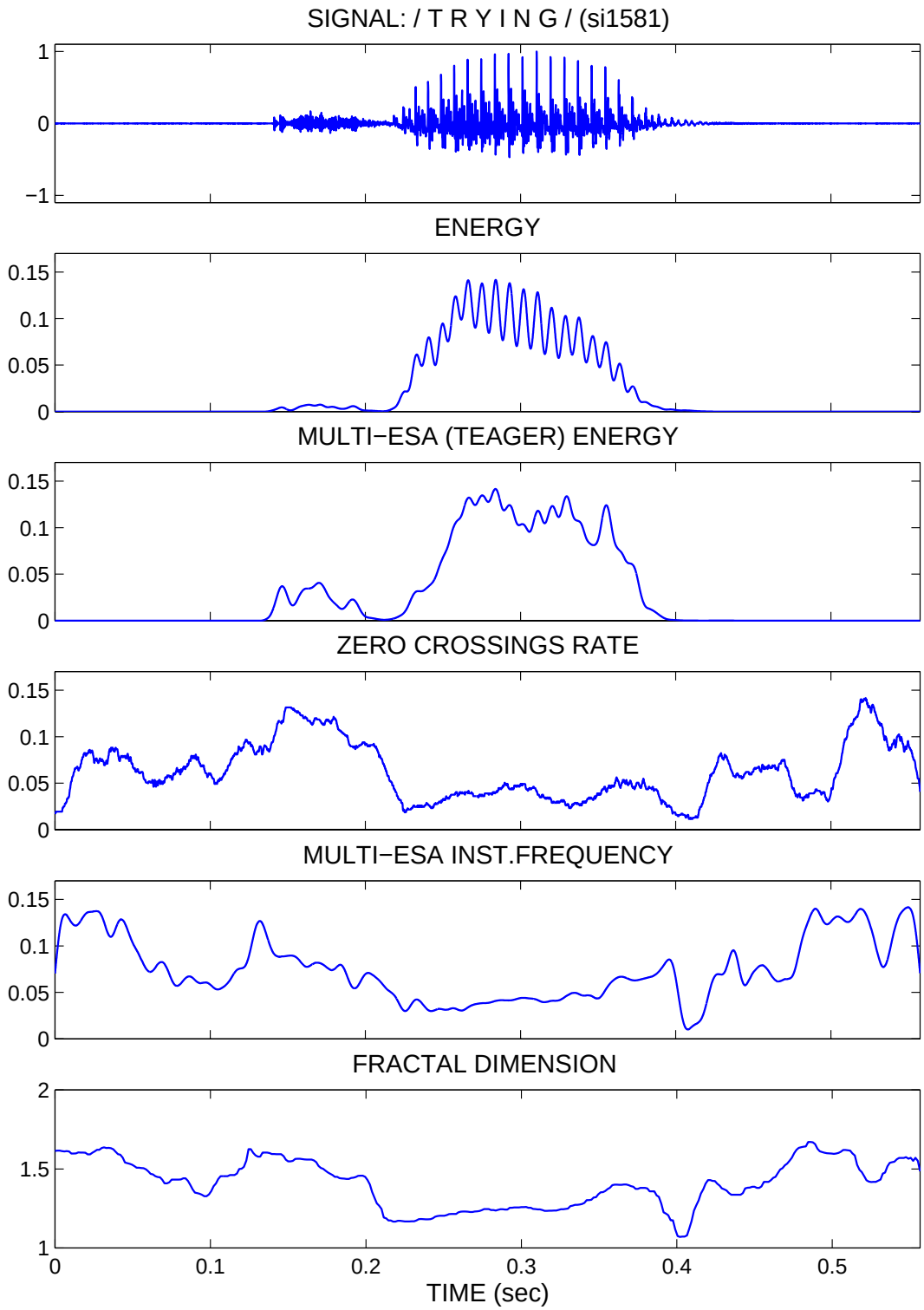
Σχήμα 4.1: Η λέξη /experts/, στα 16kHz, και μετρήσεις ενέργειας, συχνότητας, zero-crossings και fractal διάστασης, με παράθυρα 15ms.



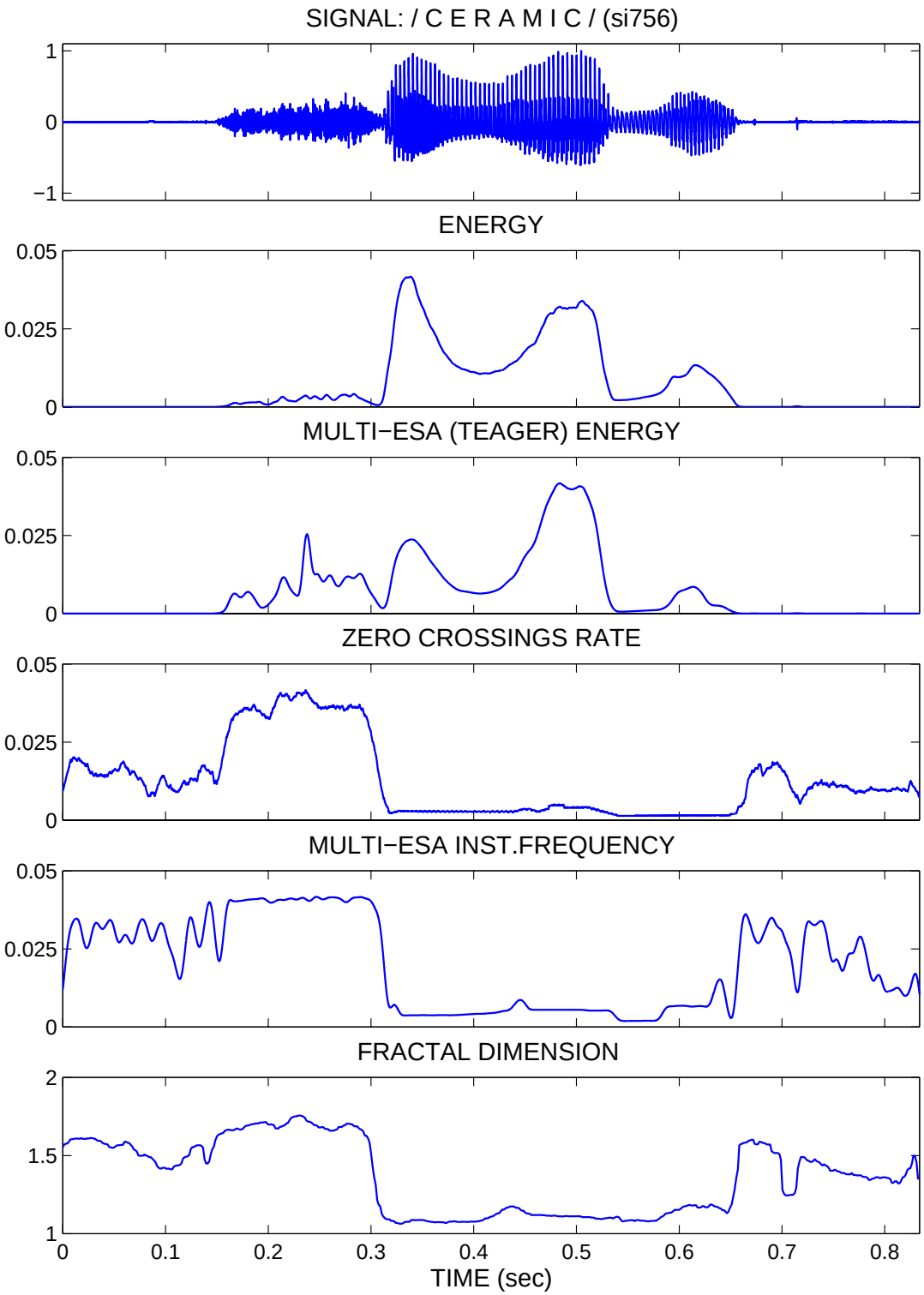
Σχήμα 4.2: Η λέξη /himself/, στα 16kHz, και μετρήσεις ενέργειας, συχνότητας, zero-crossings και fractal διάστασης, με παράθυρα 15ms.



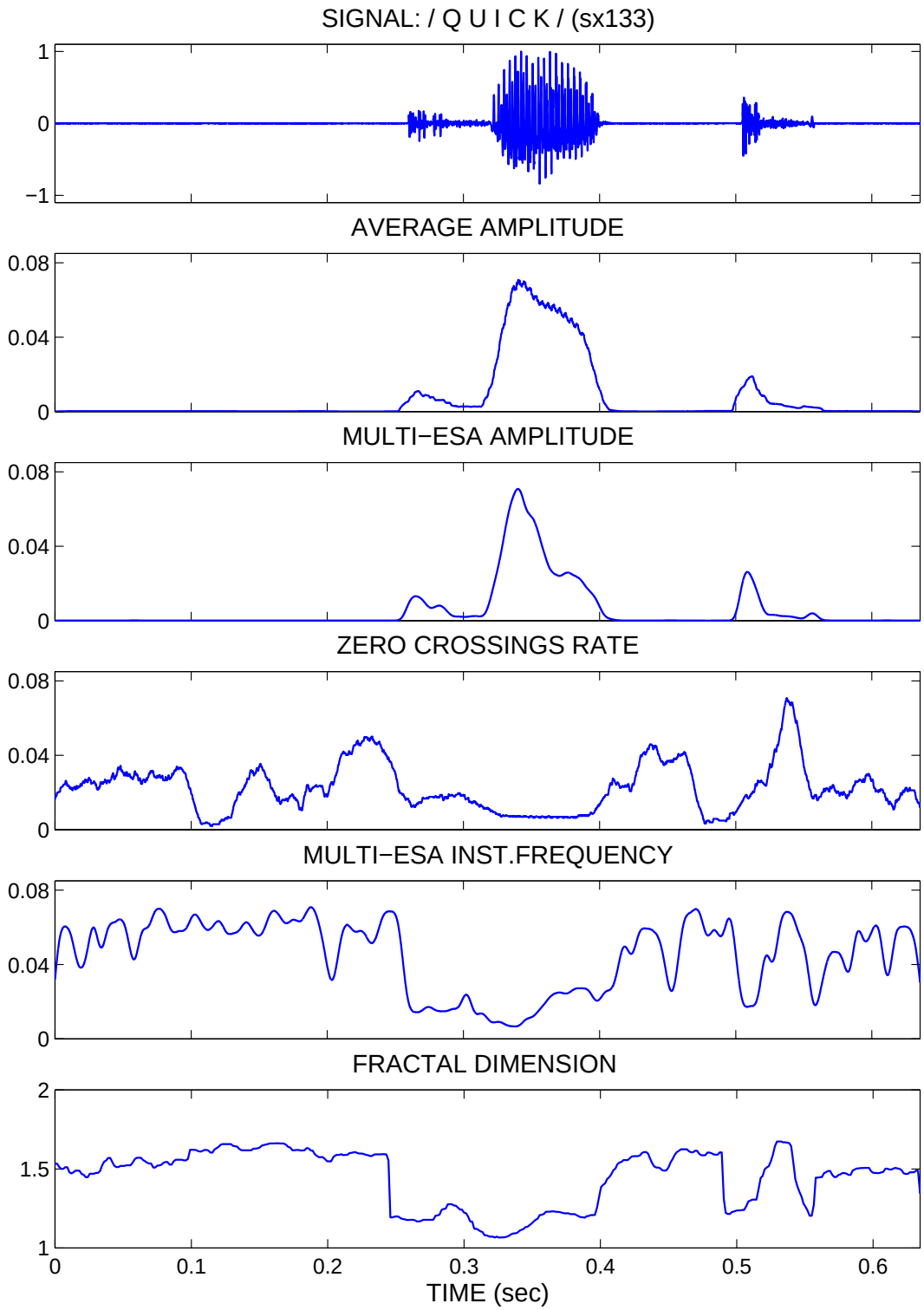
Σχήμα 4.3: Η λέξη /heavens/, στα 16kHz, και μετρήσεις ενέργειας, συχνότητας, zero-crossings και fractal διάστασης, με παράθυρα 15ms.



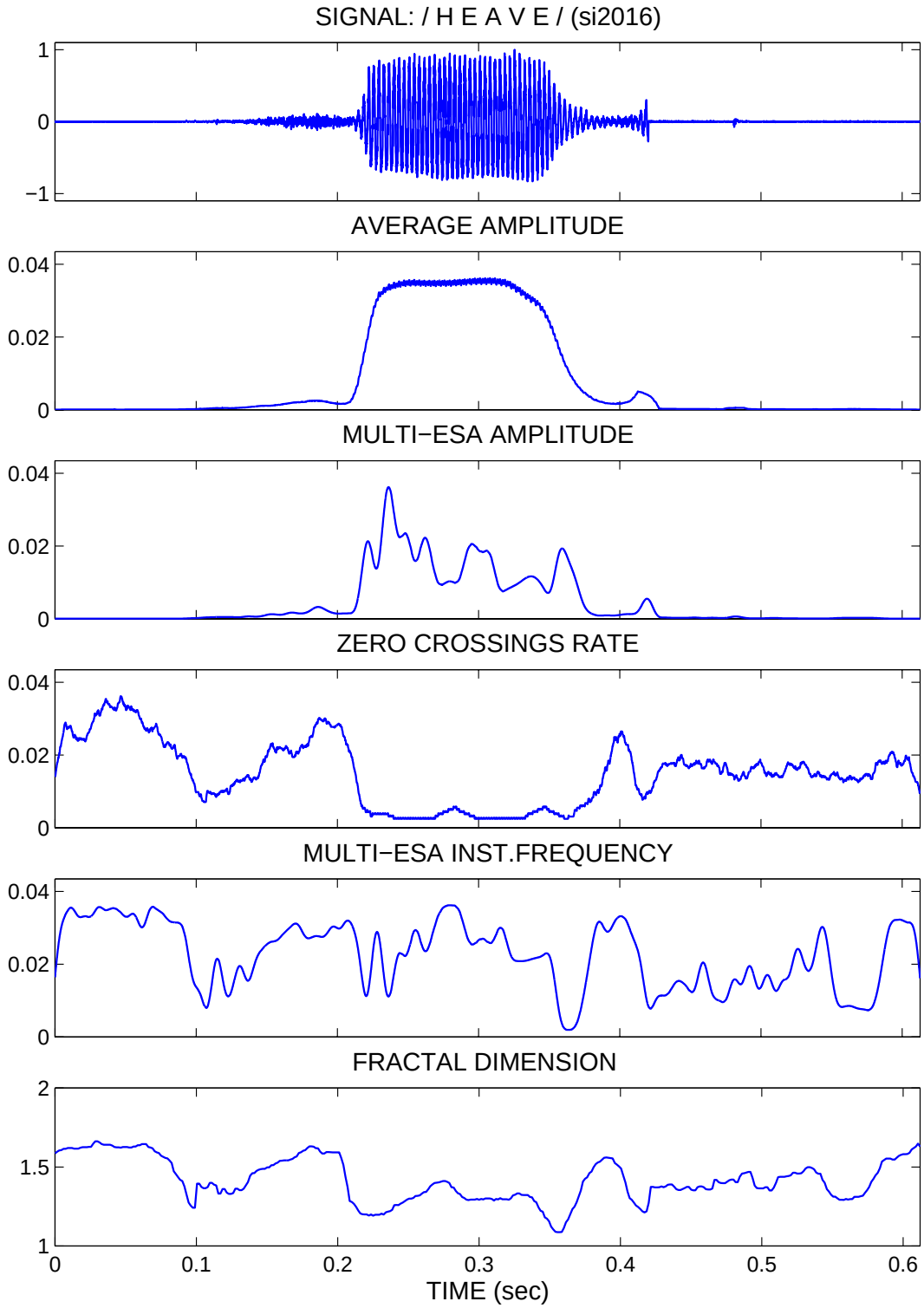
Σχήμα 4.4: Η λέξη /trying/, στα 16kHz, και μετρήσεις ενέργειας, συχνότητας, zero-crossings και fractal διάστασης, με παράθυρα 15ms.



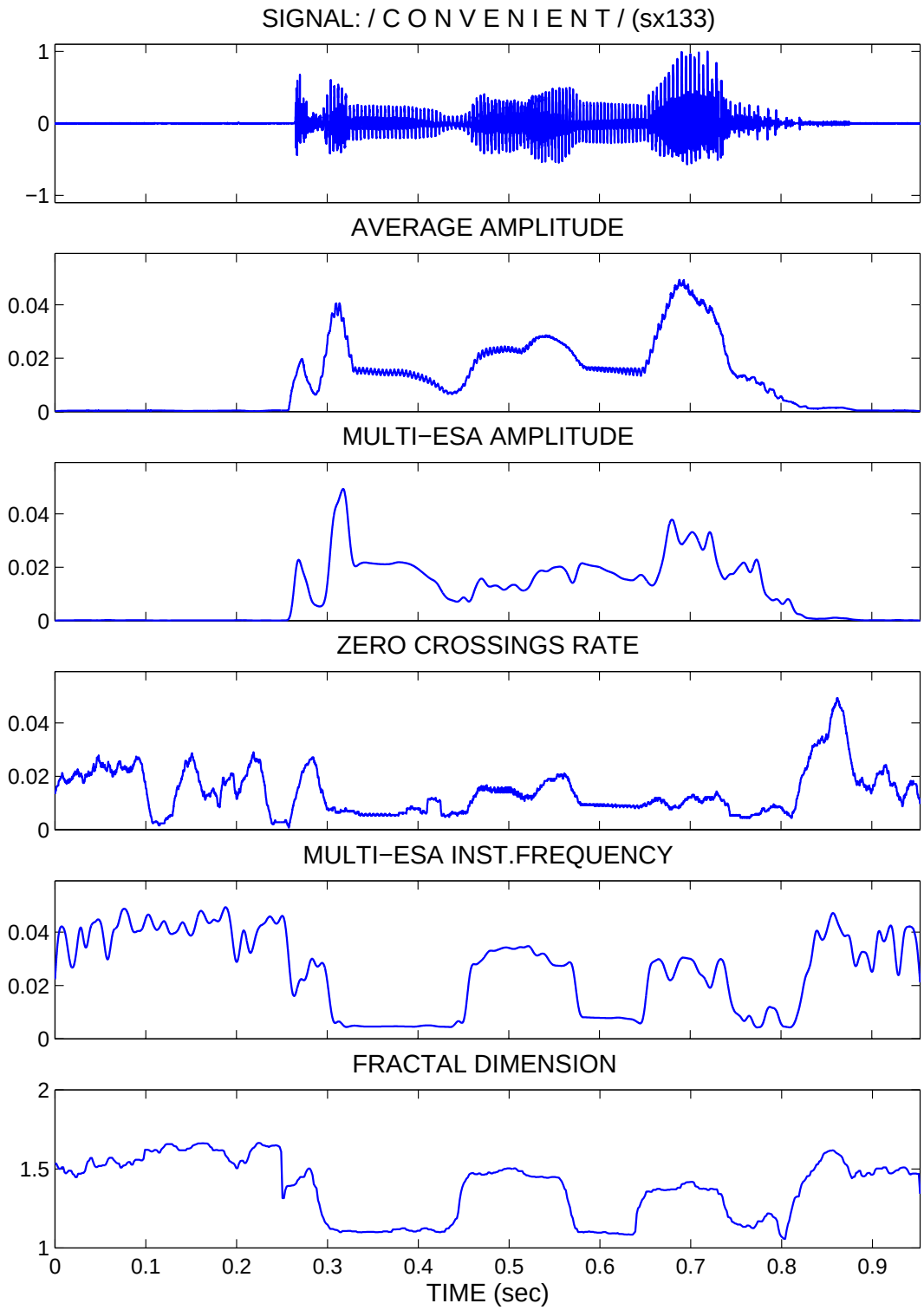
Σχήμα 4.5: Η λέξη /ceramic/, με γυναικεία φωνητικά στα 16kHz, και μετρήσεις ενέργειας, συχνότητας, zero-crossings και fractal διάστασης, με παράθυρα 15ms.



Σχήμα 4.6: Η λέξη /quick/, με γυναικεία φωνητικά στα 16kHz, και μετρήσεις πλάτους, συχνότητας, zero-crossings και fractal διάστασης, με παράθυρα 15ms.



Σχήμα 4.7: Η λέξη /heave/, με γυναικεία φωνητικά στα 16kHz, και μετρήσεις πλάτους, συχνότητας, zero-crossings και fractal διάστασης, με παράθυρα 15ms.



Σχήμα 4.8: Η λέξη /convenient/, με γυναικεία φωνητικά στα 16kHz, και μετρήσεις πλάτους, συχνότητας, zero-crossings και fractal διάστασης, με παράθυρα 15ms.

κάτι τέτοιο σημαίνει ότι και τα τμήματα σιωπής, που χαρακτηρίζονται από έντονη φασματική δραστηριότητα, θα παρουσιάζονται με αρκετά μειωμένο πλάτος, δηλ. η διαφορά επιπέδων τιμών σε σχέση με κάποια αρχικά ή τελικά φωνήματα θα είναι μεγαλύτερη.

Για τις νέες εκφράσεις των φασματικών ιδιοτήτων του σήματος, όπως φαίνεται από όλα τα σχήματα λειτουργούν και καταγράφουν μεταβολές με τον ίδιο τρόπο σε γενικές γραμμές. Η νέα μέτρηση της συχνότητας μοιάζει περισσότερο με τη μέτρηση της fractal διάστασης. Και οι δύο καταγράφουν έμφωνα τυρβώδη, όπως το /v/ στο Σχ. (4.8) με μεγαλύτερες τιμές απ'ότι τα φωνήεντα, κάποια έμφωνα που καταλήγουν σε άφωνα όπως το /v/(/f/) στο τέλος του /heave/ στο Σχ. (4.7), καθώς και δίνουν μεγαλύτερα διαστήματα σιγής πριν από την 'έξαρση' κάποιας παύσης, όπως στο /p/ του Σχ. (4.1) και το τελικό /k/ του Σχ. (4.6). Παρ' όλα αυτά πρέπει να σημειώσουμε ότι ο ρυθμός zero-crossings δίνει γενικά χαμηλότερα επίπεδα σιωπής, σε αρκετές περιπτώσεις, απ'ότι οι άλλες δύο μετρήσεις. Αυτό οφείλεται στο γεγονός ότι η κυματομορφή της σιωπής μπορεί να παρουσιάζει έντονες μεταβολές χωρίς να μεταφράζονται αυτές σε ταλαντώσεις γύρω από το μηδέν. Έτσι οι μεταβολές αυτές καταγράφονται σαν μεγάλα επίπεδα στιγμιαίας συχνότητας αλλά και σαν έντονη κατάτμηση της κυματομορφής μέσω της fractal μέτρησης, ενώ ο zero-crossings ρυθμός τους είναι πιο χαμηλός.

Όλα τα προηγούμενα αποτελούν ενδείξεις, ότι τα νέα εργαλεία που προτάθηκαν μπορούν να χρησιμοποιηθούν σε κάποια μέθοδο διάκρισης φωνής από ένα ακουστικό περιβάλλον σιωπής. Εξαιτίας της διαφορετικότητας τους αλλά κυρίως εξαιτίας της διαφορετικής αντίληψης που εισάγουνε ως προς την ενέργεια ενός σήματος, μπορούν να δώσουν στοιχεία για σήματα διαφορετικής φύσης όπως είναι η φωνή και η σιωπή.

4.2 Ανίχνευση Φωνής από Σιωπή - Ένας αλγόριθμος βασισμένος στις μη-γραμμικές απεικονίσεις

Το κίνητρό για την εφαρμογή των προηγούμενων μετρήσεων σε μια μέθοδο διάκρισης φωνής από σιωπή ήταν η ομοιότητα στον τρόπο με τον οποίο καταγράφουν τις μεταβολές του σήματος στο χρόνο, σε σχέση με τις κλασσικές μετρήσεις ενέργειας και μέσου ρυθμού zero-crossings. Ο σκοπός δεν ήταν να προταθεί μια ριζικά νέα μέθοδος διάκρισης ή ένας σύνθετος τρόπος που θα ενέπλεκε στη διαδικασία και άλλους σύνθετους περιορισμούς εκτός από συγκρίσεις. Η ουσία είναι να παρουσιαστεί η βελτίωση που μπορεί να προκύψει σε εφαρμογές διάκρισης από τη χρήση μη-γραμμικών εργαλείων.

Για το σκοπό αυτό χρησιμοποιήθηκε ο κλασσικός αλγόριθμος των Rabiner-

Sambur [2], ως βάση αναφοράς και διατηρήθηκε η απλή συνδυαστική λογική του για την υλοποίηση μιας μεθόδου που βασίζεται στις δύο νέες μετρήσεις. Την νέα ενεργειακή μέτρηση (*ShortTime Multi-Esa (Teager) Energy*) και τη μέτρηση της στιγμιαίας συχνότητας από πολλαπλά κανάλια (*Avergage Multi-Esa Inst. Frequency*).

Τα βασικά στοιχεία του αλγορίθμου διατηρούνται με κάποιες τροποποιήσεις. Χρησιμοποιούνται ξανά στατιστικά κατώφλια από τα πρώτα 100ms του σήματος που θεωρούνται σιωπή και συγκρίνονται με τις δύο μετρήσεις που υπολογίζονται για όλο το σήμα. Η απόφαση για τα όρια της φωνής στηρίζεται στο κατά πόσο και σε ποια σημεία οι μετρήσεις αυτές υπερβαίνουν τα προϋπολογισμένα κατώφλια σιωπής. Πρώτα γίνεται έλεγχος της απεικόνισης της ενέργειας και τα επιλεγμένα σημεία εξετάζονται περισσότερο σχετικά με τη μέση στιγμιαία συχνότητα τμημάτων σήματος πριν και μετά για την ύπαρξη άφωνης αλλά και χαμηλής ενέργειας. Να σημειωθεί ότι χρησιμοποιείται η μέτρηση της ενέργειας αντί για αυτή του αποδιαμορφωμένου πλάτους γιατί αποδείχτηκε πειραματικά καλύτερη όπως θα παρουσιαστεί στη συνέχεια.

Καταρχήν υποθέτουμε διαταραχές που αποτελούνται από μικρές απομονωμένες λέξεις σε ένα ακουστικό περιβάλλον σιωπής. Όλο το διάστημα της διαταραχής υπόκειται σε επεξεργασία για να υπολογιστούν οι μετρήσεις βραχέως χρόνου της ενέργειας(ή του πλάτους) και της στιγμιαίας συχνότητας. Αυτό περιλαμβάνει φιλτράρισμα από το πλέγμα των φίλτρων (MDA), αποδιαμόρφωση (ESA) και επιλογή του φίλτρου ή καναλιού κάθε χρονική στιγμή που παρουσιάζει τη μέγιστη απόκριση του ενεργειακού τελεστή ($maxE_T$). Για εφαρμογές πραγματικού χρόνου η διαδικασία μπορεί να απλοποιηθεί υπολογιστικά πολύ, με παράλληλη επεξεργασία των εξόδων των φίλτρων, εφαρμογή του ενεργειακού τελεστή και τελικά αποδιαμόρφωση της πιο ισχυρής εξόδου κάθε χρονική στιγμή (οπότε χρειάζεται και η αποθήκευση δύο μόνο σημάτων, διάρκειας όσο το αρχικό). Από τις απεικονίσεις που προκύπτουν λαμβάνονται οι αντίστοιχες βραχέως χρόνου ως βαθυπερατό φιλτράρισμα με κάποιο 'παράθυρο' ανάλυσης. Για την συγκεκριμένη υλοποίηση χρησιμοποιούνται παράθυρα των 15ms, με 16kHz τη συχνότητα δειγματοληψίας.

Από τα πρώτα 100ms του σήματος, που θεωρούνται σιωπή, λαμβάνονται στατιστικές μετρήσεις που χαρακτηρίζουν το σήματα της σιωπής. Συγκεκριμένα υπολογίζεται η *μέση τιμή*, και η *τυπική απόκλιση* για τη στιγμιαία συχνότητα, καθώς επίσης και η μέγιστη τιμή της ενέργειας του σήματος σιωπής αλλά και της ενέργειας όλου του διαστήματος. Από τις μετρήσεις αυτές επιλέγονται τα κατώφλια για τη στιγμιαία συχνότητας και ενέργειας.

Το κατώφλι της στιγμιαίας συχνότητας, THF, λαμβάνεται ως το άθροισμα

[2] Κεφάλαιο 2

της μέσης τιμής του διαστήματος σιωπής,$\overline{F}$, και της τυπικής απόκλισης, σ_F:

$$THF = \overline{F} + \sigma_F \tag{4.1}$$

Λόγο του περιορισμού που αναφέρθηκε προηγούμενα, σχετικά με το ψηλό επίπεδο συχνοτήτων της σιωπής, το κατώφλι αυτό δυστυχώς δεν μπορεί να τεθεί περισσότερο συντηρητικά. Από τη μέγιστη τιμή της ενέργειας της σιωπής, SMX, και τη μέγιστη ενέργεια του διαστήματος, που ανήκει σε φωνή, EMX, λαμβάνονται ένα μεγάλο και ένα μικρότερο και πιο αυστηρό κατώφλι ενέργειας σύμφωνα με τους εξής κανόνες:

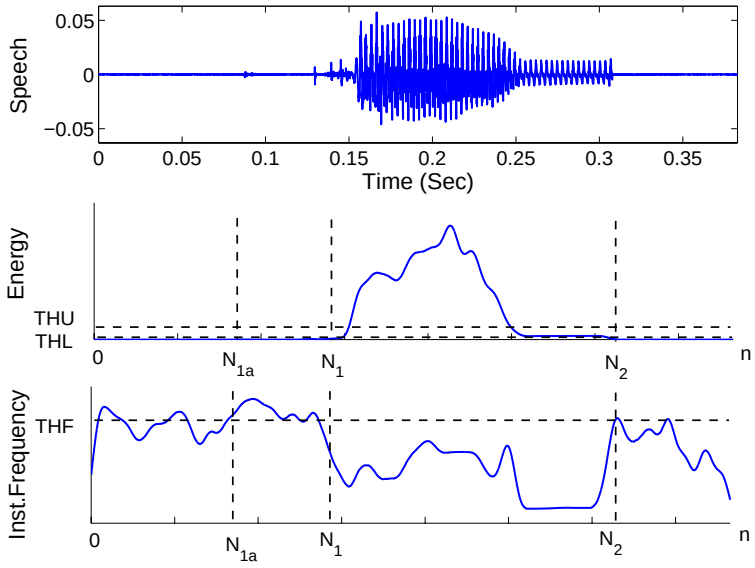
$$\begin{aligned} T_1 &= 0.02 \quad (EMX \quad SMX) + SMX && (4.2)\\ T_2 &= 3 \quad IMN && (4.3)\\ THL &= min(T_1, T_2) && (4.4)\\ THU &= 5 \quad ITL && (4.5) \end{aligned}$$

Το διπλό ενεργειακό κατώφλι εξασφαλίζει προστασία από απότομες πτώσεις της ενέργειας που μπορεί να σηματοδοτήσουν πρόωρα εξασθένιση της ενέργειας σε επίπεδα σιωπής.

Θα περιγραφεί σύντομα, διότι η λογική της είναι παρόμοια με του κλασσικού αλγορίθμου, η διαδικασία προσδιορισμού του αρχικού και τελικού σημείου φωνής με αναφορά το Σχήμα (4.9), στο οποίο φαίνεται η διαταραχή /thin/, οι δύο μετρήσεις, της (Teager) ενέργειας και της στιγμιαίας συχνότητας καθώς και τα κατώφλια στα οποία στηρίζεται η απόφαση του αλγορίθμου.

Αρχικά αναζητείται μια ευρεία περιοχή, έντονη σε ενεργειακή δραστηριότητα, εντός της οποίας αποκλείεται να υπάρχουν τα σημεία αρχής και τέλους της λέξης. Για το σκοπό αυτό ελέγχεται η ενεργειακή μέτρηση και εντοπίζεται το διάστημα που αυτή είναι ξεπερνάει το μεγάλο κατώφλι, THU. Στη συνέχεια σε ένα δεύτερο έλεγχο, αναζητείται η στιγμή πριν από το σημείο που ξεπερνιέται για πρώτη φορά το THU, που η ενέργεια γίνεται μικρότερη από ένα πιο αυστηρό χαμηλό κατώφλι, THL. Το σημείο αυτό είναι το $N1$ στο Σχ. (4.9). Παρόμοιος διπλός έλεγχος γίνεται και στο τέλος της λέξης και οδηγεί στο σημείο $N2$. Τα σημεία $N1$ και $N2$ είναι τα πρώτα υποψήφια σημεία αρχής και λήξης αντίστοιχα.

Σε αρκετές περιπτώσεις η νέα αυτή ενεργειακή μέτρηση, εξαιτίας της φύσης της και της διαφοράς που αποκαλύπτει ανάμεσα στα σήματα φωνής και σιωπής είναι αρκετή από μόνη της να εντοπίσει με μεγάλη ακρίβεια τα όρια τη φωνής. Παρ'όλα αυτά για μεγαλύτερη ασφάλεια ελέγχεται και η περίπτωση ύπαρξης φωνή έξω από αυτό το αρχικά εκτιμώμενο διάστημα $(N1, N2)$ με τη βοήθεια της απεικόνισης της στιγμιαίας συχνότητας. Η ύπαρξη ενός διαστήματος όπου η μέτρηση της στιγμιαίας συχνότητας έχει τιμή πάνω



Σχήμα 4.9: Παρουσίαση γραφικά του τρόπου που λειτουργούν τα κατώφλια για τον αλγόριθμο με τις νέες απεικονίσεις. Διακρίνονται η κυματομορφή της λέξης /thin/, στα 16kHz, η νέα μέτρηση της Teager ενέργειας (ShortTime Multi-Esa Energy) και η μέτρηση της στιγμιαίας συχνότητας του σήματος (Average Multi-Esa Inst. Frequency). Τα σημεία που εντοπίζονται τελικά τα όρια της φωνής είναι τα N_{1a} και N_2.

από τις τιμές στις οποίες κυμαίνεται το διάστημα της σιωπής αποδεικνύεται ικανή ένδειξη άφωνης διαταραχής.

Εξετάζεται το διάστημα από $N1$ έως $N1 \quad 1000$, δηλ. τα προηγούμενα 60ms περίπου και υπολογίζονται τα δείγματα που έχουν στιγμιαία συχνότητα πάνω από την τιμή του κατωφλιού THF. Αν οι χρονικές στιγμές με τέτοια υψηλή συχνότητα είναι πάνω από 120, τότε το υποψήφιο σημείο έναρξης μεταφέρεται πίσω στην πρώτη στιγμή που ξεπεράστηκε το κατώφλι (το σημείο $N1a$, στο σχήμα) . Διαφορετικά διατηρείται η επιλογή του σημείου $N1$. Με τον ίδιο τρόπο ελέγχονται και τα δείγματα $N2$ έως $N2+1000$, για την ύπαρξη άφωνων ή πολύ ασθενών έμφωνων ήχων.

Για το Σχήμα (4.9) και τη λέξη /thin/, ο αλγόριθμος εντοπίζει αρχικά μέσω της ενέργειας το διάστημα $(N1, N2)$. Στη συνέχεια ο έλεγχος της απεικόνισης της συχνότητας αποκαλύπτει ένα μεγάλο διάστημα χρονικών στιγμών με 'ψηλές' συχνότητες , το αρχικό άφωνο /th/, με αποτέλεσμα ως σημείο έναρξης της λέξης να επιλεγεί το σημείο $N1a$ όπου ξεπερνιέται για πρώτη φορά το κατώφλι. Ο ανάλογος έλεγχος στο τέλος δε δίνει κάποια ένδειξη, οπότε η ενέργεια και μόνο έχει εντοπίσει το σημείο λήξης της λέξης $N2$.

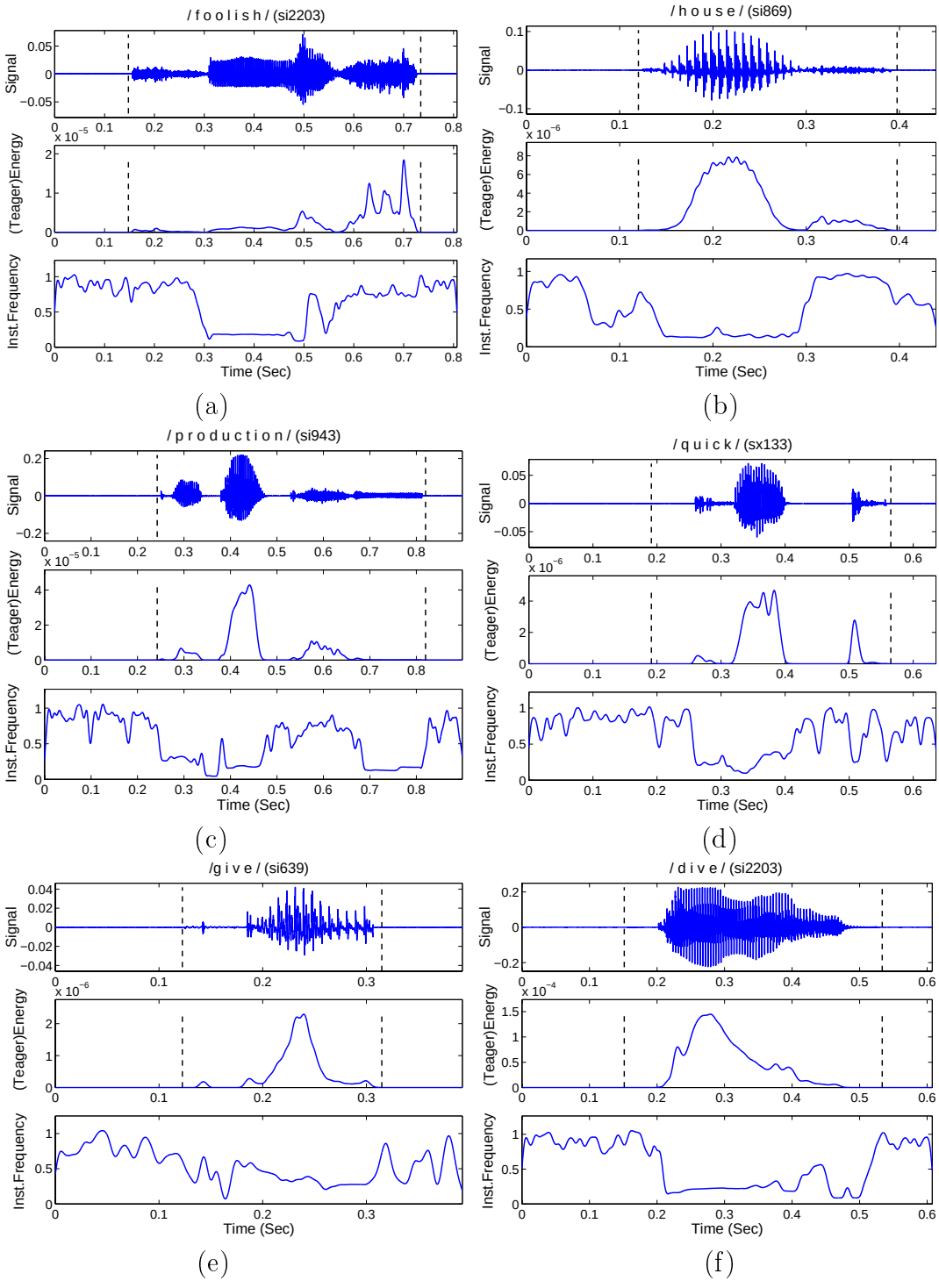
4.2.1 Παραδείγματα εφαρμογής του αλγορίθμου σε απομονωμένες διαταραχές

Η προσομοίωση του αλγορίθμου που χρησιμοποιεί τις νέες απεικονίσεις έγινε με τη βοήθεια του υπολογιστικού πακέτου MATLAB, έτσι ώστε να εξακριβωθούν και πειραματικά τα προηγούμενα συμπεράσματα και η βελτίωση που επιφέρει η χρήση των νέων εργαλείων που παρουσιάστηκαν. Η υλοποίηση βασίστηκε στα στοιχεία που περιγράφηκαν στα προηγούμενα κεφάλαια. Έτσι χρησιμοποιήθηκε ένα πλέγμα από 24 *Gabor* φίλτρα, για να καλύπτουν το φάσμα του σήματος και η έξοδος κάθε φίλτρου αποδιαμορφώθηκε μέσω του αλγορίθμου *DESA-1*. Οι απεικονίσεις της συχνότητας από κάθε φίλτρο, ομαλοποιούνται πριν από οποιαδήποτε άλλη επεξεργασία με ένα *median* φίλτρο 13 σημείων έτσι ώστε να απορριφθούν απότομα "καρφιά" που μπορεί να οφείλονται σε σφάλματα της διαδικασίας .Μετά την επιλογή του καναλιού με τη μέγιστη ενέργεια, λαμβάνονται οι *short-time* μετρήσεις με τη βοήθεια παραθύρων 15ms, στα 16kHz.

Ο αλγόριθμος εφαρμόστηκε σε ένα πλήθος λέξεων, με σιωπή στην αρχή και στο τέλος τους, από φράσεις της βάσης TIMIT. Προσομοιώθηκαν όσο το δυνατόν καλύτερες συνθήκες πραγματικών ηχογραφήσεων με τμήματα σιωπής από την αρχή και το τέλος των φράσεων στις οποίες ανήκουν οι λέξεις (και οι οποίες θα αναφέρονται δίπλα στη λέξη μέσα σε παρενθέσεις). Στο Σχήμα (4.10) φαίνονται έξι χαρακτηριστικά παραδείγματα με τις διεκεκομμένες γραμμές να ορίζουν τα σημεία που εντοπίζει τα άκρα της φωνής ο αλγόριθμος.

Στο Σχ. (4.10a) φαίνεται η λέξη /foolish/, από γυναίκα ομιλητή, η οποία ξεκινάει με ένα άφωνο αλλά ισχυρά τυρβώδες /f/ και καταλήγει με ένα έντονο /sh/, κάτι το οποίο απεικονίζεται με υψηλά επίπεδα στο διάγραμμα της ενέργειας. Ο αλγόριθμος εντοπίζει πολύ καλά την αρχή και το τέλος της λέξης βασιζόμενος μονάχα στην ενεργειακή μέτρηση. Παρόμοια στο (b), για τη λέξη /house/, η ενέργεια εντοπίζει το τελικό /s/ αλλά και τον σχετικά αδύναμο 'ψίθυρο', /h/, στην αρχή της λέξης.

Στα σχήματα (c) και (d) έχουμε τις περιπτώσεις άφωνων παύσεων ως έναρξη ή λήξη, οι οποίες όπως έχει αναφερθεί αρκετές φορές μπορεί να δημιουργήσουν πρόβλημα. Στη λέξη /production/ του (c), η ενεργειακή μέτρηση εντοπίζει τόσο το αρχικό /p/ όσο και το τελικό ένρινο /n/, παρ'όλο που δεν φαίνεται απόλυτα στο διάγραμμα. Τα επίπεδα (Teager)ενέργειας και των δύο είναι πολύ χαμηλά αλλά της σιωπής είναι πολύ χαμηλότερα. Στο (d) , η λέξη /quick/ ξεκινάει και τελειώνει με την παύση /k/. Τα όρια της εντοπίζονται με μεγάλη ακρίβεια. Το τέλος μόνο με την ενέργεια ενώ η αρχή χρησιμοποιεί σε μικρό βαθμό και τη μέτρηση της συχνότητας. Εδώ να αναφέρουμε ότι για τη συγκεκριμένη λέξη ο κλασσικός αλγόριθμος χρειαζόταν και τη μέτρηση του



Σχήμα 4.10: Παραδείγματα εφαρμογής του αλγορίθμου με τις νέες μετρήσεις, ενέργειας και στιγμιαίας συχνότητας για χαρακτηριστικές περιπτώσεις λέξεων.

zero-crossings ρυθμού για να εντοπίσει την έναρξη της.

Τέλος στα σχήματα (e),(f) του (4.10) φαίνονται έμφωνες παύσεις στην αρχή και έμφωνα τυρβώδη στο τέλος των δύο λέξεων. Στο (e), στη λέξη /give/ εντοπίζεται όλο το διάστημα ηρεμίας της παύσης /g/ στην αρχή ενώ το τελικό /v/ είναι καθ'όλη τη διάρκεια του έμφωνο οπότε η ενέργεια του είναι σε πολύ μεγαλύτερα επίπεδα από αυτά της σιωπής. Στο σχήμα (f), φαίνεται η λέξη /dive/ με γυναικεία φωνητικά, η οποία παρουσιάζει το εξής σύνηθες φαινόμενο. Το τελικό /v/ γίνεται άφωνο /f/ από κάποιο σημείο και μετά. Ο αλγόριθμος χρησιμοποιώντας και τη μέτρηση της συχνότητας εντοπίζει όλη τη διάρκεια του. Για την έναρξη, η χρήση της ενεργειακής μέτρησης είναι αρκετή για να δώσει όλο το διάστημα του αρχικού /d/.

Όπως είναι φανερό η διάκριση βασίζεται κυρίως στη μέτρηση της νέας (Teager) ενέργειας και ανάλογα με την περίπτωση χρησιμοποιείται και πληροφορία από τη μέτρηση της στιγμιαίας συχνότητας. Πάντως οι δοκιμές δείξανε ότι μη-χρήση της συχνότητας χειροτερεύει την γενική απόδοση της διαδικασίας.

4.2.2 Συγκρίσεις με τον Κλασσικό Αλγόριθμο

Για να διαπιστώσουμε πόσο βελτιώνεται η διαδικασία ανίχνευσης του κλασσικού αλγορίθμου των Rabiner-Sambur, με τη χρήση των νέων μετρήσεων της Multi-Esa Teager Energy και της Multi-Esa Inst.Frequency εφαρμόσαμε και τις δύο διαδικασίες σε ένα σύνολο από μεμονωμένες διαταραχές. Φυσικά η πιο αξιόπιστη σύγκριση των δύο που μπορεί να γίνει είναι στα πλαίσια αναγνώρισης των λέξεων που ανιχνεύονται, οπότε έτσι να φανεί ποιος από τους δύο δίνει τα μεγαλύτερα ποσοστά αναγνώρισης αφού αυτό είναι που μετράει στην ουσία. Κάτι τέτοιο ξεφεύγει από το σκοπό του παρόντος και αφήνεται ως προοπτική μελλοντικού ενδιαφέροντος.

Το μέτρο σύγκρισης εδώ είναι πόσο κοντά βρίσκονται τα εντοπισμένα από τους δύο αλγορίθμους σημεία στα σημεία που δίνονται από τα στοιχεία της TIMIT για τα όρια των λέξεων. Οι διαδικασίες είναι σχεδιασμένες έτσι ώστε αφενός να ελαχιστοποιούν τα μεγάλα σφάλματα (>40ms) και αφετέρου να περικλείουν αρκετή πληροφορία από κάποιο άφωνο αρχικό ή τελικό φώνημα έτσι ώστε να μπορεί να γίνει επιτυχής αναγνώριση της ανιχνευόμενης λέξης. Αυτό σημαίνει τουλάχιστον 30-50ms του φωνήματος να συμπεριλαμβάνονται στο διάστημα της λέξης [10]. Επομένως σαν σφάλμα στη διαδικασία θεωρούμε:

i) την περίπτωση που το εκτιμώμενο σημείο απέχει περισσότερο από 40ms από το γνωστό σημείο αρχής ή τέλους (*υπερεκτίμηση της σιωπής*).

ii) την περίπτωση που στα όρια μιας ανιχνευόμενης λέξης περιλαμβάνεται καθόλου ή μικρή διάρκεια (<30ms) από ένα αρχικό ή τελικό φώνημα (*'χαμένο' φώνημα*).

Ο νέος αλγόριθμος [3] παρουσίασε γενικά πολύ βελτιωμένη συμπεριφορά σε σχέση με τον κλασσικό. Η διαφορά φαίνεται σε περιπτώσεις χαμένων φωνημάτων που παρουσιάστηκαν με τον κλασσικό αλγόριθμο και τα οποία εντοπίστηκαν από το νέο. Κάποια τέτοια χαρακτηριστικά παραδείγματα παρουσιάζονται στα Σχήματα (4.11) και (4.13).

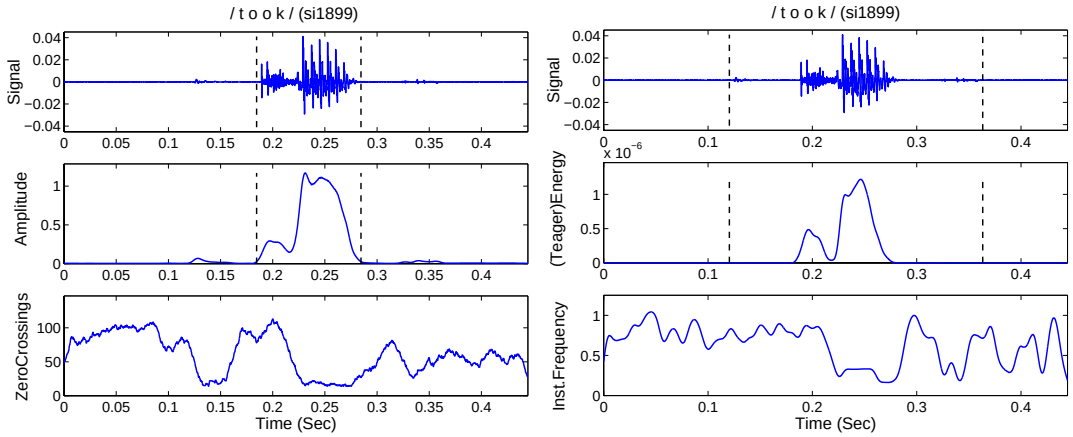
Στο Σχ. (4.11a) εξετάζεται η λέξη /took/ , από άντρα ομιλητή και η εφαρμογή των δύο μεθόδων. Όπως φαίνεται και από τις διακεκομμένες γραμμές ο κλασσικός αλγόριθμος με τις μετρήσεις πλάτους και zero-crossings χάνει τόσο το αρχικό /t/, όσο και το τελικό /k/. Τα κατώφλια του πλάτους δεν καταφέρνουν να εντοπίσουν τα χαμηλά επίπεδα αυτών των φωνημάτων, ενώ επειδή η διέγερση τους είναι 'χαλαρή' το zero-crossings είναι μικρό για να μετατοπίσει τα αρχικά υποψήφια σημεία. Αντίθετα βλέπουμε ότι η νέα ενεργειακή μέτρηση τονίζει αρκετά, ώστε να εντοπιστούν, τη διαφορά ανάμεσα στα φωνήματα αυτά και στη σιωπή πρίν και μετά. Η διαφορά με τα πραγματικά σημεία έναρξης και λήξης είναι5ms και για τα δύο, σε αντίθεση με τα 57 και 72ms διαφοράς που δίνει η κλασσική μέθοδος.

Ανάλογα αποτελέσματα έχουμε και για την περίπτωση του /back/, από άντρα ομιλητή, στο Σχ. (4.11b). Η άφωνη παύση /k/ στο τέλος, παρ'όλο που εμφανίζει ένα σχετικά μεγαλύτερο ρυθμό zero-crossings από τη σιωπή, χάνεται (56ms) όπως και η αρχική έμφωνη αλλά ασθενής παύση /b/(115ms). Η νέα προσέγγιση εντόπισε και τα δύο, δίνοντας απόκλιση από τις πραγματικές τιμές 8 και 4ms για τα δύο σημεία.

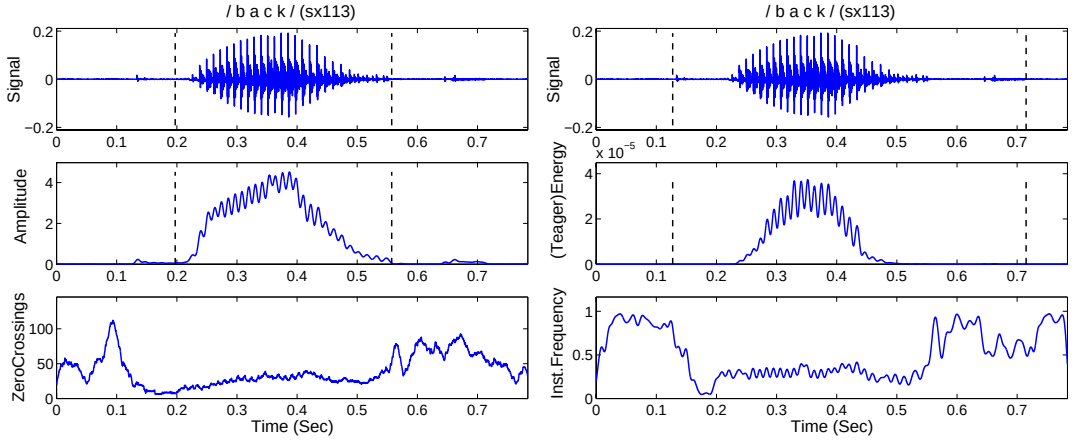
Στο Σχ. (4.11c) φαίνεται η λέξη /up/ με μια πολύ ασθενή τελική παύση. Το άφωνο /p/ χάνεται τόσο από το πλάτος όσο και από το ρυθμό zero-crossings με την κλασσική μέθοδο(37ms απόκλιση). Αντίθετα με την νέα διαδικασία, η ενέργεια αρκεί για να το εντοπίσει (1ms απόκλιση). Για το αρχικό /u/ η υπόθεση είναι εύκολη και για τους δύο αλγορίθμους (7 και 10 ms αντίστοιχα ,από τα πραγματικά). Για να φανεί καλύτερα η διαφορά που ενισχύει η νέα ενεργειακή μέτρηση για την περίπτωση αυτού του ασθενούς τελικού /p/ παρουσιάζονται στο Σχήμα (4.12), σε db τιμές τα διαγράμματα της νέας ενεργειακής μέτρησης, του νέου πλάτους αλλά και της κλασσικής μέτρησης του πλάτους. Βλέπουμε πόσο παραπάνω από το επίπεδο της σιωπής "ανεβαίνει" το /p/ για την νέα ενέργεια απ'ότι για την κλασσική μέτρηση του πλάτους.

Στο Σχήμα (4.13) παρουσιάζονται περισσότερα παραδείγματα. Στο (a) έχουμε τη λέξη /novelty/, που καταλήγει με τη συλλαβή /-ty/. Το /t/ είναι άφωνο και το τελικό /i/ είναι πολύ ασθενές και καταλήγει άφωνο. Αυτό οδηγεί τον κλασσικό αλγόριθμο να το αγνοήσει τελείως. Χρησιμοποιεί και

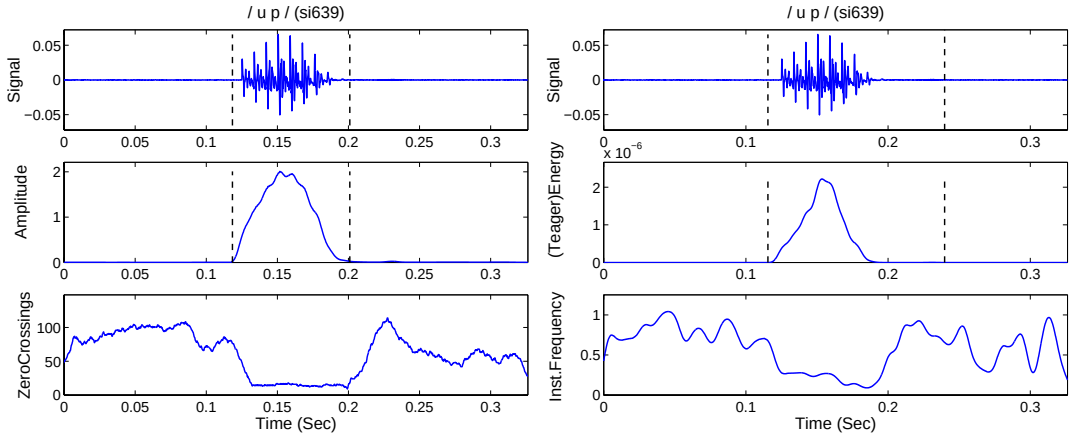
[3] Με την έκφραση νέος εννοούμε τον αλγόριθμο όπως περιγράφηκε σ'αυτό το κεφάλαιο με τις νέες μετρήσεις ενέργειας και στιγμιαίας συχνότητας, ενώ ως κλασσικό θεωρούμε τον αλγόριθμο που περιγράφηκε στο 2ο Κεφάλαιο και χρησιμοποιεί τις κλασσικές μετρήσεις πλάτους (ή ενέργειας) και ρυθμού zero-crossings.



(a)

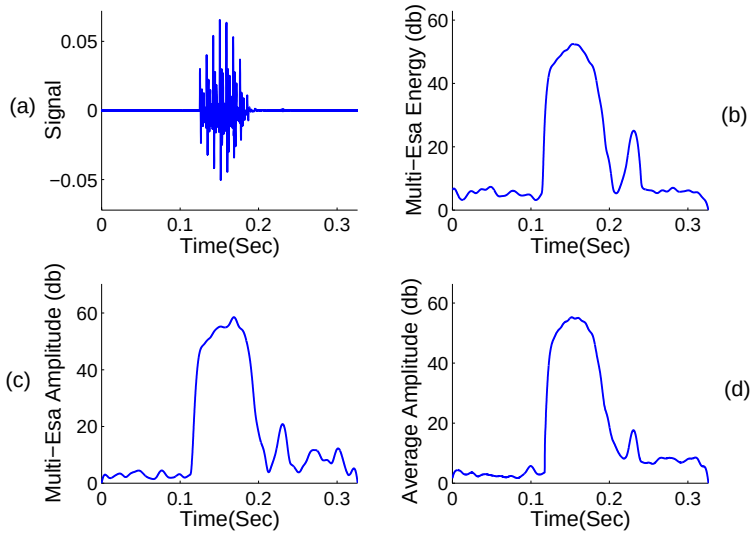


(b)



(c)

Σχήμα 4.11: Διαφορές στην απόδοση των δύο μεθόδων. Σε κάθε διάγραμμα, στην αριστερή πλευρά παρουσιάζεται η κλασσική προσέγγιση ενώ στη δεξιά η νέα προσέγγιση με τα μη-γραμμικά εργαλεία. Το σήμα δειγματοληπτείται στα 16kHz και για όλες τις απεικονίσεις χρησιμοποιούνται παράθυρα των 15ms.

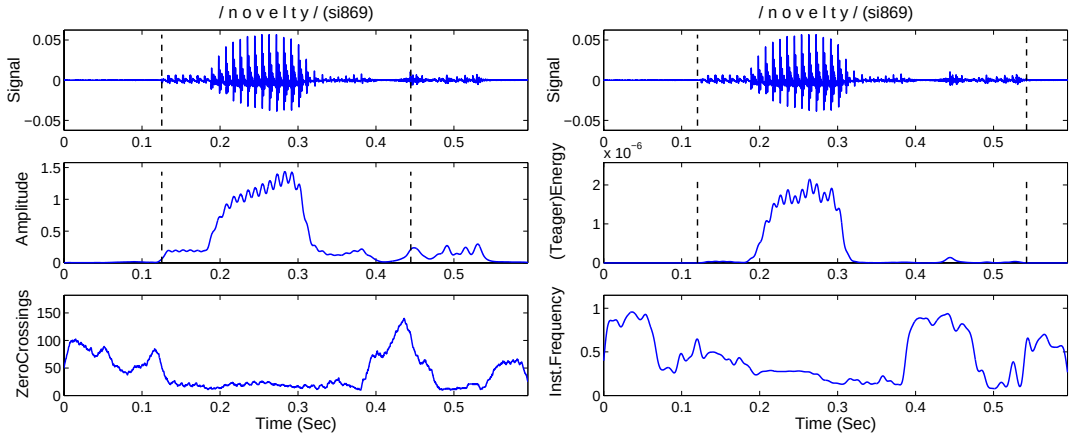


Σχήμα 4.12: Διαφορές ανάμεσα στα επίπεδα φωνής και σιωπής για τη λέξη /up/. (a) Η κυματομορφή του σήματος στα 16kHz. (b) Νέα ενεργειακή μη-γραμμική μέτρηση (Multi-Esa (Teager) Energy). (c) Νέο αποδιαμορφωμένο πλάτος (Multi-Esa Amplitude). (d) Κλασσική μέτρηση πλάτους (Average Amplitude). Όλες οι μετρήσεις είναι βραχέως χρόνου.

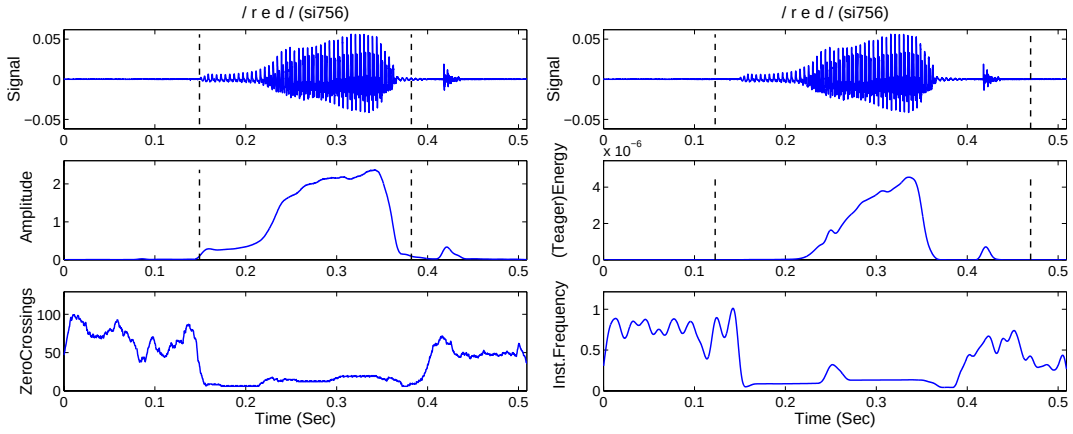
τη μέτρηση zero-crossings αλλά αποτυγχάνει να συμπεριλάβει ένα μεγάλο τελικό τμήμα στα εκτιμώμενα όρια της λέξης (100ms απόκλιση). Από την άλλη ο αλγόριθμος με τα νέα εργαλεία εντοπίζει ολόκληρη την τελική άφωνη ενέργεια με μεγάλη επιτυχία (2ms από τα πραγματικά).

Στο (4.13b) συγκρίνεται η λέξη /red/. Παρ'όλο που η τελική έμφωνη παύση /d/ έχει πλάτος που διακρίνεται από τη σιωπή η κλασσική προσέγγιση τη χάνει τελείως ενώ η νέα την εντοπίζει. Αυτό όμως συνοδεύεται και από μια υπερεκτίμηση της σιωπής στην αρχή και στο τέλος (περίπου 30ms), η οποία σε τέτοιο μικρό βαθμό δεν μπορεί να αποτελεί πρόβλημα. Τέτοιες υπερεκτιμήσεις σιωπής, σε μικρό ή μεγάλο βαθμό, συναντήσαμε αρκετές φορές κατά τη διάρκεια των δοκιμών και για τους δύο αλγορίθμους. Πιθανότατα οφείλονται στον τρόπο που δημιουργήθηκαν οι υπό εξέταση διαταραχές (με απότομη συνένωση σιωπής- φωνής-σιωπής) με αποτέλεσμα ο ρυθμός zero-crossings ή η συχνότητα να πιάνει τέτοιες ψηλές συνιστώσες στα σημεία μετάβασης. Για πραγματικά ηχογραφημένες διαταραχές αυτά τα φαινόμενα θα εκλείπουν αφού η μετάβαση σιωπής-φωνής θα είναι πιο ομαλή.

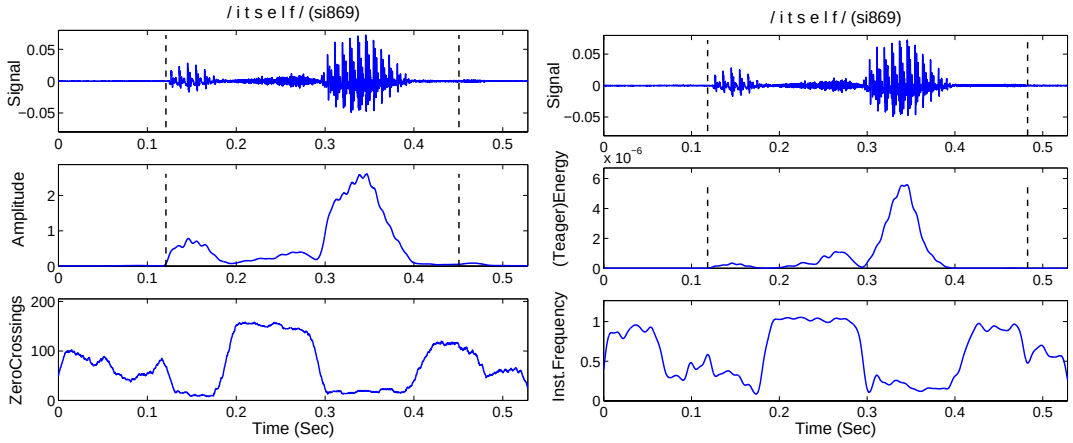
Τέλος στο (4.13c) παρουσιάζεται η περίπτωση που ο νέος αλγόριθμος αποκαλύπτει μεγαλύτερο διάστημα απ΄ότι ο κλασσικός. Πρόκειται για τη διαταραχή /itself/ για την οποία ο οποίος εντοπίζει πλάτος και zero-crossings εντοπίζουν ένα μέρος του τελικού /f/ αλλά όχι ολόκληρο (πιθανότατα αρκετό



(a)

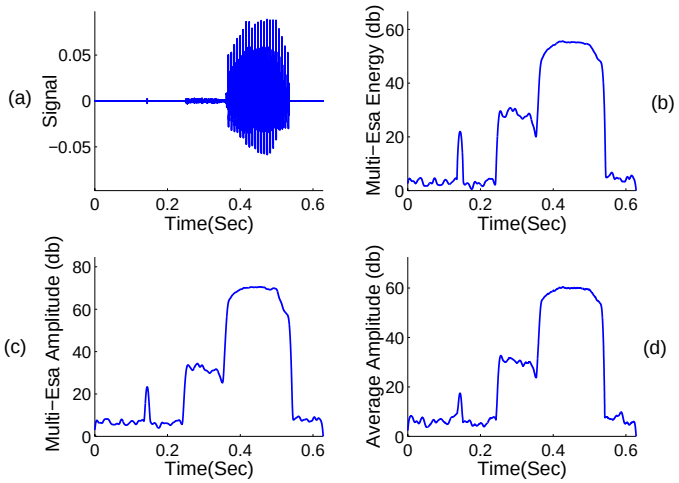


(b)



(c)

Σχήμα 4.13: Περισσότερα παραδείγματα ανίχνευσης λέξεων που παρουσιάζουν τη βελτίωση που επιφέρει η χρήση των νέων απεικονίσεων και ειδικά της νέας ενεργειακής μέτρησης. Σε κάθε περίπτωση, αριστερά φαίνεται το αποτέλεσμα του κλασσικού αλγορίθμου ενώ δεξιά η νέα προσέγγιση.



Σχήμα 4.14: Παράδειγμα της ευαισθησίας της νέας ενεργειακής μέτρησης σε πιθανά αρχικά 'mouth clicks'. Η διαταραχή είναι η λέξη /file/. Οι μετρήσεις είναι σε db τιμές. Το 'καρφί' στην αρχή της νέας ενεργειακής μέτρησης στο (d) είναι ένα click. Τα πλάτη, όπως φαίνεται, δεν παρουσιάζουν την ίδια ευαισθησία.

για να γίνει η αναγνώριση της λέξης). Τα νέα πάντως εργαλεία εντοπίζουν όλη τη διάρκεια του τελικού φωνήματος.

Όπως είναι φυσικό υπάρχουνε και οι περιπτώσεις που και ο νέος αλγόριθμος αποτυγχάνει είτε χάνοντας κάποιο ολόκληρο φώνημα είτε εντοπίζοντας μόνο ένα μέρος του. Αυτό συμβαίνει σε περιπτώσεις αρχικών και τελικών εξαιρετικά αδύναμων φωνημάτων. Επίσης ένα άλλο θέμα το οποίο δεν θίξαμε μέχρι τώρα είναι η περίπτωση άλλων ήχων, μη-φωνητικής φύσεως που μπορεί να είναι παρόντα μαζί με τη σιωπή. Τέτοια γεγονότα, γνωστά ως *'mouth clicks'* οφείλονται στον τρόπο προφοράς του ομιλητή (περισσότερες λεπτομέρειες δόθηκαν στο Κεφ.1). Επειδή η νέα μέτρηση της (Teager) ενέργειας ενισχύει τη διαφορά γεγονότων σε σχέση με τη σιωπή, όπως είναι φυσικό μεγενθύνει και τα 'clicks' που μπορεί να είναι παρόντα στην αρχή των λέξεων. Έτσι η μέτρηση αποκτά υπερευαισθησία και σχεδόν πάντα ανιχνεύει τέτοια 'clicks' στην αρχή. Ένα παράδειγμα φαίνεται στο Σχήμα (4.14). Η ενέργεια ανιχνεύει το αρχικό αυτό 'click' ως φωνητικό γεγονός με αποτέλεσμα να δίνει μια απόκλιση 115ms από το πραγματικό σημείο έναρξης της λέξης.

4.2.3 Εναλλακτικές υλοποιήσεις για ανίχνευση φωνής

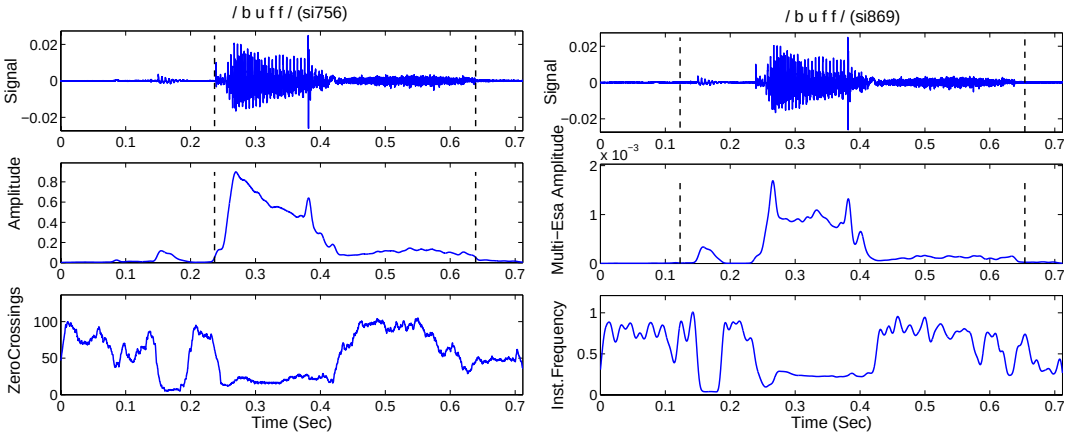
Με αφορμή την πληθώρα απεικονίσεων, νέων και κλασσικών, που διαθέτουμε στα χέρια μας πλέον δοκιμάστηκαν διάφοροι συνδυασμοί προς αναζήτηση της βέλτιστης από πλευράς ανίχνευσης λύσης, σε σχέση με τον κλασσικό

αλγόριθμο. Θα αναφέρουμε σύντομα τους εναλλακτικούς αυτούς τρόπους με ένα παράδειγμα-σύγκριση σε σχέση με τον κλασσικό αλγόριθμο για την κάθε περίπτωση.

Νέο Πλάτος και Στιγμιαία Συχνότητα

Υλοποιώντας τη διαδικασία όπως περιγράφηκε στην Ενότητα 4.1, με τη διαφορά για την ανίχνευση έμφωνης ενέργειας να χρησιμοποιηθεί το νέο πλάτος αποδιαμόρφωσης αντί για την νέα ενέργεια. Ως γνωστό, αυτό το πλάτος προκύπτει από τη διαδικασία επιλογής του αποδιαμορφωμένου πλάτους στην έξοδο κάθε φίλτρου, που αντιστοιχεί στο κανάλι με τη μέγιστη έξοδο του ενεργειακού τελεστή (*Average Multi-Esa Amplitude*). Για τον εντοπισμό άφωνης ενέργειας χρησιμοποιείται όπως και πριν η στιγμιαία συχνότητα (*Average Multi-Esa Inst.Frequency*).

Η συγκεκριμένη υλοποίηση βελτιώνει τα ποσοστά ανίχνευσης του κλασσικού αλγορίθμου αφού ανιχνεύει φωνήματα τα οποία χάνονται. Επίσης υπερτερεί του αλγορίθμου με τη νέα ενέργεια αφού δεν παρουσιάζει την ίδια ευαισθησία στα 'clicks' που αναφέρθηκαν προηγούμενα και μειώνει την υπερεκτίμηση σιωπής που συμβαίνει σε ορισμένες περιπτώσεις. Όμως λόγο της μικρότερης ευαισθησίας στις διαφορές του πλάτους τα ποσοστά ανίχνευσης του είναι μικρότερα και αρκετά φωνήματα χάνονται. Στο Σχ. (4.15) φαίνεται ένα παράδειγμα όπου αυτή η διαδικασία ανιχνεύει τη λέξη /buff/. Για τη συγκεκριμένη διαταραχή ο αλγόριθμος με τη νέα (Teager) ενέργεια απέτυχε να ανιχνεύσει



Σχήμα 4.15: Παράδειγμα της νέας υλοποίησης με χρήση του νέου πλάτους αντί της ενέργειας. Η εξεταζόμενη διαταραχή είναι η λέξη /buff/, στα 16kHz. (a) Η εφαρμογή του κλασσικού αλγορίθμου ανίχνευσης. (b) Η εφαρμογή της νέας υλοποίησης.

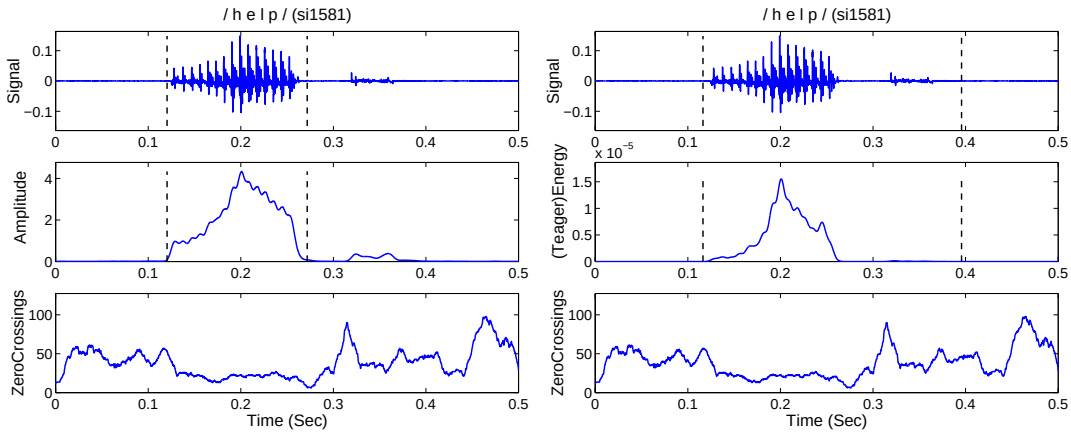
το αρχικό /b/ και υπερεκτίμησε κατά 40ms τη σιωπή στο τέλος.

Νέα (Teager) Ενέργεια και Zero-Crossings

Αντί για την απεικόνιση της στιγμιαίας συχνότητας χρησιμοποιήθηκε το 'γραμμικό' εργαλείο του μέσου ρυθμού zero-crossings. Τα αποτελέσματα είναι παρόμοια με τη βασική υλοποίηση του νέου αλγορίθμου και φυσικά καλύτερα από αυτά του κλασσικού αφού εμπλέκεται η νέα ενεργειακή μέτρηση. Σε αρκετές περιπτώσεις πάντως μειώνεται η υπερεκτίμηση σιωπής αλλά και υπάρχουν και χαμένα φωνήματα. Στο Σχήμα (4.16) φαίνεται η επιτυχία με την οποία εντοπίζει το τελικό /p/, πράγμα το οποίο ο κλασσικός αλγόριθμος αποτυγχάνει να κάνει.

Αλλες υλοποιήσεις

Εκτός από τα προηγούμενα εξετάστηκαν και δύο ακόμη περιπτώσεις. Εξαιτίας του κυρίαρχου ρόλου που επιτελεί η νέα ενέργεια σε αυτή τη νέα προσέγγιση που προτάθηκε, ελέγχθηκε η περίπτωση χρήσης της *ενέργειας μόνο (Multi-Esa (Teager) Energy)* για ανίχνευση των άκρων της φωνής, έτσι ώστε να μειωθεί και η υπερεκτίμηση που μπορεί να οφείλεται στον έλεγχο της στιγμιαίας συχνότητας. Αυτή η υλοποίηση έδωσε μικρότερα ποσοστά ανίχνευσης από τη μέθοδο με τη συχνότητα αλλά βελτίωση σε σχέση με την κλασσική. Εξάλλου



(a) (b)

Σχήμα 4.16: Παράδειγμα της νέας υλοποίησης με χρήση του μέσου ρυθμού zero-crossings αντί για τη στιγμιαία συχνότητα. Η εξεταζόμενη διαταραχή είναι η λέξη /help/, στα 16kHz. (a) Η εφαρμογή του κλασσικού αλγορίθμου ανίχνευσης. (b) Η εφαρμογή της νέας υλοποίησης.

η νέα έννοια της ενέργειας περιλαμβάνει και κάποια πληροφορία για τη συχνότητα και επομένως για τα άφωνα τμήματα υψηλής διέγερσης.

Τέλος ελέγχθηκε και η περίπτωση χρησιμοποίησης της *fractal διάστασης βραχέως χρόνου* ως εργαλείο πληροφορίας μαζί με τη νέα ενεργειακή μέτρηση. Η περίπτωση αυτή θέλει λίγο περισσότερο ψάξιμο ως προς την εφαρμογή της για ανίχνευση φωνής, αφού τα αποτελέσματα σε άλλες περιπτώσεις είχαν νόημα και σε άλλες όχι.

4.2.4 Γενικά Συμπεράσματα-Αποτελέσματα

Μια πρώτη σύγκριση σε επίπεδο απόδοσης έγινε και σε προηγούμενη ενότητα ανάμεσα στον κλασσικό αλγόριθμο και στον αλγόριθμο που χρησιμοποιεί τα νέα εργαλεία που προέκυψαν από τη μη-γραμμική επεξεργασία του σήματος. Οι δύο διαδικασίες, καθώς και ο εναλλακτικές υλοποιήσεις δοκιμάστηκαν σε ένα πλήθος 130 λέξεων από τη βάση TIMIT με σιωπή στην αρχή και στο τέλος του, από έξι διαφορετικούς ομιλητές, τέσσερις άντρες και δύο γυναίκες. Κάποια στατιστικά ποσοστά ανίχνευσης για αυτό το σύνολο, σε σχέση πάντα με τα όρια των λέξεων όπως δίνονται φαίνονται στον Πίνακα (4.1) .

Θα πρέπει εδώ να σημειώσουμε ότι για να εξαχθούν τα προηγούμενα ποσοστά ως λάθη θεωρήθηκαν χαμένα φωνήματα στην αρχή ή/και στο τέλος των λέξεων καθώς και υπερεκτιμήσεις σιωπής μεγαλύτερες από 50ms. Επίσης επαναλαμβανόμενες υπερεκτιμήσεις σιωπής για λέξεις προερχόμενες από την ίδια φράση θεωρήθηκαν σφάλματα προερχόμενα από τη δημιουργία των εξεταζόμενων διαταραχών και αγνοήθηκαν. Τέλος τα επίπεδα φωνής ως προς τη σιωπή σε όλα τα σήματα ήτανε μεγαλύτερα από 30db, πράγμα που προσομοιώνει καλές συνθήκες ηχογράφησης.

Από τον Πίνακα (4.1) βλέπουμε ότι το μεγαλύτερο ποσοστό ανίχνευσης το δίνει η νέα προσέγγιση, όπως προτάθηκε στο κεφάλαιο αυτό με χρήση της νέας ενεργειακής μέτρησης και τη στιγμιαίας συχνότητας αποδιαμόρφωσης των ζωνοπερατών συνιστωσών του σήματος. Αυτό το ποσοστό είναι 11% πάνω από το ποσοστό που δίνει η κλασσική μέθοδος όπως προτάθηκε αρχικά από τους Rabiner-Sambur. Όμως θα πρέπει να σημειώσουμε την ευαισθησία της σε αρχικά 'clicks' που μπορεί να εμφανιστούν στη φωνή, καθώς και κάποια

Μέθοδος Ανίχνευσης	Κλασσική M, Z	Νέα $E_T, \overline{\Omega}_i$	$\overline{a}, \overline{\Omega}_i$	E_T, Z	E_T
Ποσοστό $\%$	83	94	90	93	92

Πίνακας 4.1: Αποτελέσματα ανίχνευσης μεμονωμένων λέξεων από δοκιμές σε 130 διαταραχές, του κλασσικού αλγορίθμου, της νέας προσέγγισης και των εναλλακτικών μεθόδων.

υπερεκτίμηση της σιωπής($30ms$) που μπορεί να προκύψει ανάλογα με την περίπτωση λόγο της χρήσης της συχνότητας και του όχι και τόσο αυστηρού κατωφλιού που χρησιμοποιεί.

Για να μειώσουμε την υπερεκτίμηση μπορούμε να χρησιμοποιήσουμε την υλοποίηση που χρησιμοποιεί μόνο την ενέργεια αλλά τότε το ποσοστό ανίχνευσης είναι μικρότερο κατά 2% αλλά παραμένει ανώτερο από την κλασσική μέθοδο. Για να περιορίσουμε την ανίχνευση αρχικών μη-φωνητικών γεγονότων μαζί με τη λέξη η καλύτερη περίπτωση είναι η μέθοδος που χρησιμοποιεί το νέο αποδιαμορφωμένο πλάτος και τη συχνότητα αλλά το ποσοστό είναι μικρότερο κατά 5% της βασικής νέας μεθόδου. Τέλος η προσέγγιση με τη χρήση του ρυθμού zero-crossings αντί της συχνότητας δίνει σχεδόν το ίδιο ποσοστό και υπερεκτίμηση αυτή που μπορεί να δίνει και η κλασσική μέθοδος.

Μια τελευταία παρατήρηση είναι ότι η μέθοδος με τα νέα εργαλεία είναι σαφώς πιο πολύπλοκη υπολογιστικά και πιο αργή από τον κλασσικό αλγόριθμο αφού περιλαμβάνει στην ουσία τρία στάδια επεξεργασίας, και μάλιστα το ένα από αυτά σε πολλαπλά κανάλια ζωνοπερατών συνιστωσών, αντί για ένα που χρησιμοποιεί ο κλασσικός αλγόριθμος. Παρ'όλα αυτά η μέθοδος είναι σχετικά απλή στην υλοποίηση της, δεν εμπλέκει κάποια σύνθετη λογική απόφασης και ,με τα σύγχρονα μέσα και παράλληλη επεξεργασία των καναλιών του σήματος μπορεί να χρησιμοποιηθεί και για εφαρμογές πραγματικού χρόνου.

4.3 Ανίχνευση Φωνής και Θόρυβος

Μέχρι στιγμής με την έννοια του θορύβου ασχοληθήκαμε λίγο. Θεωρήσαμε την περίπτωση σημάτων της μορφής σιωπή-φωνή-σιωπή και μάλιστα σε καλές συνθήκες ηχογράφησης (SNR>30db). Ο έλεγχος και η σύγκριση για την κλασσική και τη νέα προσέγγιση έγινε κάτω από τέτοιες συνθήκες. Τι συμβαίνει όμως στις περιπτώσεις όπου παράλληλα με το σήμα ενδιαφέροντος υπάρχει και επιπρόσθετος θόρυβος; Τέτοιες περιπτώσεις συναντώνται πολύ συχνά σε πρακτικές εφαρμογές ανίχνευσης και αναγνώρισης φωνής όπως για παράδειγμα στην κινητή τηλεφωνία, στη μετάδοση φωνής πάνω από τηλεφωνικές γραμμές κ.α. Έρευνες γίνανε και γίνονται πάνω στο θέμα της ανίχνευσης φωνής σε θορυβώδη περιβάλλοντα (βλ.[4],[13]). Εδώ δοκιμάσαμε μια σύγκριση των δύο μεθόδων κάτω από διάφορες στάθμες προσθετικού θορύβου για να εξετάσουμε ποια επηρεάζεται περισσότερο.

Επίσης έγινε και μια προσπάθεια ανίχνευσης, με όλα τα διαθέσιμα εργαλεία, γραμμικά και μη γραμμικά, ανίχνευσης τριών ειδών σημάτων που μπορεί να βρίσκονται 'σε σειρά'. Κάτι τέτοιο είναι χρήσιμο σε περιπτώσεις όπου ένα θορυβώδες σήμα, τυχαίας φύσεως μπορεί να βρίσκεται πριν ή μετά το σήμα της φωνής (π.χ. λόγο υψίσυχνων μικροφωνισμών) και να αντιμετωπιστεί ως

φωνή από κάποια διαδικασία αναγνώρισης ή ανίχνευσης.

4.3.1 Αντοχή της διαδικασίας ανίχνευσης σε θόρυβο

Ο κλασσικός αλγόριθμος είναι σχεδιασμένος να δίνει καλά αποτελέσματα σε συνθήκες ηχογράφησης με υψηλό σηματοθορυβικό λόγο φωνής προς σιωπή (η οποία στην ουσία θεωρείται ακουστικός θόρυβος του περιβάλλοντος). Αντίθετα η νέα προσέγγιση που προτάθηκε εδώ και χρησιμοποιεί τα νέα μηγραμμικά εργαλεία, παρότι βασίζεται στις ίδιες γενικές αρχές με την κλασσική, προσφέρει την προοπτική μεγαλύτερης αντοχής σε θόρυβο.

Αυτή η υπόθεση βασίζεται στο γεγονός ότι για να εξαχθούν οι απεικονίσεις του σήματος που εμπλέκει στις αποφάσεις, χρησιμοποιείται ανάλυση σε πολλαπλές μπάντες (MDA). Η MDA διαδικασία αναπτύχθηκε εξαρχής με σκοπό τη βελτίωση της απόδοσης του ESA σε AM-FM σήματα που συνυπάρχουν με μεγάλα ποσοστά θορύβου. Πράγματι η απόδοση του ενεργειακού τελεστή και του ESA βελτιώνεται σημαντικά μέσα από το φιλτράρισμα του σήματος, από ένα σύνολο πυκνών στο φάσμα ζωνοπερατών φίλτρων, καθώς η επίδραση του θορύβου περιορίζεται στη ζώνη του φίλτρου που επιλέγεται και αποδιαμορφώνεται κάθε χρονική στιγμή. Οι συνιστώσες του στις υπόλοιπες ζώνες είναι αμελητέες ειδικά αν το εύρος των φίλτρων είναι μικρό [1].

Με βάση αυτή την πληροφορία, περιμένουμε ότι και στην εφαρμογή που παρουσιάσαμε εδώ για ανίχνευση φωνής και που χρησιμοποιεί μια MDA διαδικασία τα αποτελέσματα σε συνθήκες θορύβου θα είναι ποιο ευσταθή απ' ότι με την κλασσική μέθοδο. Το τυχαίο σήμα του θορύβου επιδρά σε όλο το φάσμα και ενώ η κλασσική διαδικασία χρησιμοποιεί όλο το φάσμα κάθε χρονική στιγμή, η νέα προσέγγιση χρησιμοποιεί μόνο τη ζώνη του που δίνει τη μεγαλύτερη έξοδο του ενεργειακού τελεστή. Φυσικά εδώ μιλάμε για σήματα φωνής, που μοντελοποιούνται με τη βοήθεια υπέρθεσης AM-FM σημάτων και όχι για απλά διαμορφωμένα σήματα σε συχνότητα και πλάτος, οπότε πρέπει η παραπάνω υπόθεση με πρέπει να διατηρείται με κάποια επιφύλαξη.

Για να διαπιστωθεί πειραματικά το αποτέλεσμα του θορύβου στις δύο διαδικασίες ανίχνευσης, δοκιμάστηκε η εφαρμογή τους, σε διαταραχές που δίνανε και για τους δύο τρόπους καλά αποτελέσματα εντοπισμού των αρχικών και τελικών σημείων της φωνής. Στις διαταραχές προστέθηκε λευκός θόρυβος για διαφορετικά επίπεδα SNR. Χρησιμοποιήθηκαν 30 λέξεις από την TIMIT και τέσσερις στάθμες θορύβου (40,30,20,10 db). Παρ'όλο που ο αριθμός των διαταραχών είναι μικρός για να δώσει στατιστικά συμπεράσματα μια αρχική πρώτη εκτίμηση μπορεί να γίνει. Ο αριθμός των λαθών ανίχνευσης, δηλαδή ο αριθμός των λέξεων στις οποίες 'καταρρέουν' σε κάθε στάθμη θορύβου οι δύο τρόποι φαίνεται στον Πίνακα (4.2), ενώ στον Πίνακα (4.3) φαίνεται το ποσοστό ορθής ανίχνευσης λέξεων σε κάθε στάθμη.

Σύνολο:30 λέξεις	Λάθη Ανίχνευσης	
SNR(db)	Κλασσική	Σύγχρονη
40	12	7
30	6	8
20	5	10
10	7	5

Πίνακας 4.2: Δοκιμές με τέσσερις στάθμες θορύβου για την κλασσική και τη σύγχρονη μέθοδο.Από τις 30 λέξεις φαίνεται πόσες δίνουν λάθος ανίχνευση σε κάθε στάθμη του SNR.

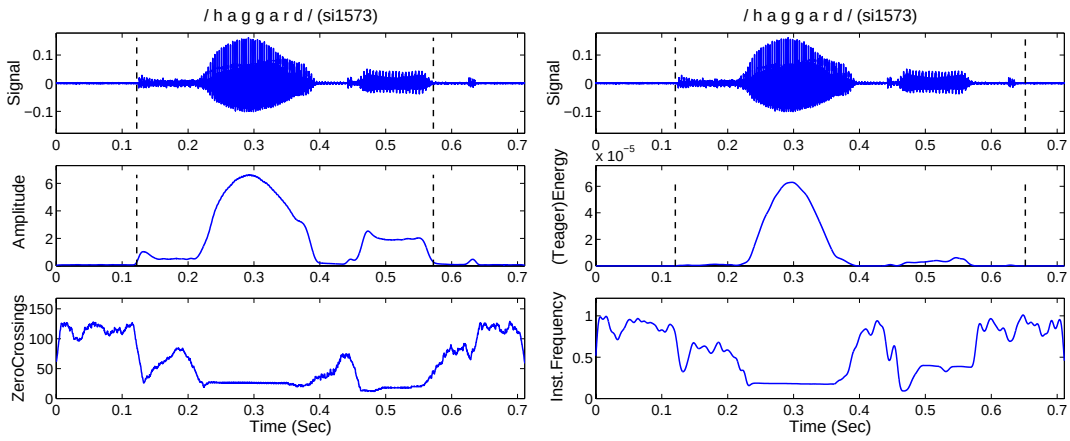
Σύνολο:30 λέξεις	Ποσοστό Ανίχνευσης%	
SNR(db)	Κλασσική	Σύγχρονη
40	60	77
30	40	50
20	23	17

Πίνακας 4.3: Ποσοστά ορθής ανίχνευσης λέξεων σε κάθε στάθμη θορύβου.

Παρατηρήθηκε ότι γενικά και οι δύο τρόποι πάσχουνε από την προσθήκη θορύβου, και αυτό εκδηλώνεται είτε ως χαμένα φωνήματα στην αρχή ή στο τέλος των λέξεων είτε ως μεγάλα σφάλματα υπερεκτίμησης της σιωπής (απόκλιση πάνω από 70ms από τα πραγματικά άκρα). Οι λέξεις που παρουσιάσανε το μεγαλύτερο σφάλμα στις μικρές στάθμες θορύβου (40,30db) ήτανε κυρίως λέξεις με έμφωνες παύσεις στην αρχή ή στο τέλος /d/,/b/ και κάποιες άφωνες παύσεις π.χ. /c/,/k/ ενώ στις μεγάλες (20db) συριστικά όπως /f/,/h/ και ένρινα τελικά όπως /n/,/ng/. Στο Σχήμα (4.17) φαίνεται η ανίχνευση για τις δύο μεθόδους, στη λέξη /haggard/ για σηματοθορυβικό λόγο 30db.

Αξίζει να αναφέρουμε ότι για κάποιες λέξεις η νέα μέθοδος παρουσίασε το εξής "περίεργο" φαινομενικά γεγονός. Ενώ στα 40db έδωσε μεγάλες αποκλίσεις, ενώ αυξήθηκε ο SNR η απόδοση όχι μόνο δεν μειώθηκε αλλά βελτιώθηκε αισθητά σε επίπεδα ακριβούς εντοπισμού των άκρων της φωνής.

Τέλος να σημειώσουμε ξανά ότι τα αποτελέσματα με πραγματικές ηχογραφήσεις μεμονωμένων λέξεων σε συνθήκες θορύβου μπορεί να είναι διαφορετικά, όπως επίσης ότι για να εξαχθούν ασφαλή στατιστικά συμπεράσματα απαιτείται ο έλεγχος ενός πολύ μεγαλύτερου δείγματος διαταραχών. Σε αυτή την περίπτωση μπορεί να αποκαλυφθεί μια ακόμη καλύτερη συμπεριφορά της μεθόδου με τις νέες απεικονίσεις σε θορυβώδες περιβάλλον.



(a) (b)

Σχήμα 4.17: Παράδειγμα ανίχνευσης σε περιβάλλον θορύβου. Η διαταραχή είναι η λέξη /haggard/ στα 16kHz, με SNR=30db. (a) Η εφαρμογή του κλασσικού αλγορίθμου ανίχνευσης. (b) Η εφαρμογή της νέας μεθόδου με τα μη-γραμμικά εργαλεία.

4.3.2 Διάκριση Σιωπής, Φωνής, Θορύβου 'σε σειρά'

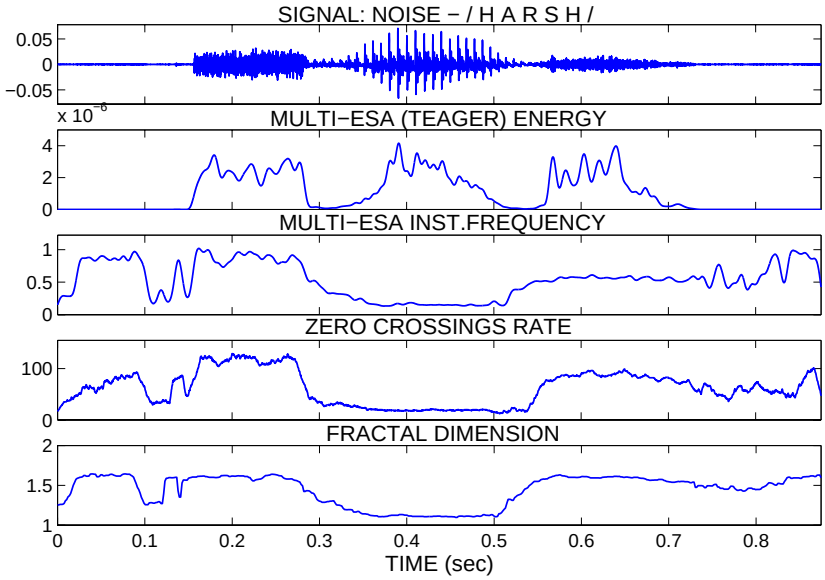
Η απομόνωση σημάτων θορύβου που βρίσκονται σε αλληλουχία με φωνή ή και σιωπή δεν είναι εύκολη διαδικασία. Το να διακριθεί δηλαδή το χρήσιμο σήμα της φωνής από τα άλλα δύο πιθανότατα απαιτεί σύνθετη αλγοριθμική λογική και τεχνικές αναγνώρισης προτύπων. Με κίνητρο την πληθώρα πλέον εργαλείων που παρουσιάστηκαν, γραμμικών και μη-γραμμικών για επεξεργασία σημάτων στο χρόνο καθώς και με βάση την τεχνική που προτάθηκε για διάκριση φωνής από σιωπή επιχειρούμε να προσεγγίσουμε μια ειδική περίπτωση του προβλήματος.

Συγκεκριμένα υποθέτουμε σήματα θορύβου σε σειρά με φωνή (concatenation) σε ένα ακουστικό περιβάλλον σιωπής, δηλαδή θόρυβο *ανάμεσα* στη φωνή και στη σιωπή. Ως θόρυβος θεωρείται ένα τυχαίο σήμα, αρκετά μεγαλύτερου μεγέθους από τη σιωπή, με μηδενική μέση τιμή. Προτείνεται ότι για να απομονωθεί το χρήσιμο πληροφοριακά σήμα της φωνής από τα άλλα δύο, η διάκριση μπορεί να βασιστεί σε τέσσερις απεικονίσεις βραχέως χρόνου:

1. **Multi-Esa (Teager) Energy**, η οποία είναι μεγάλη για φωνή και θόρυβο και μικρή για σιωπή.

2. **Multi-Esa Inst.Frequency**, η οποία μπορεί να διακρίνει ανάμεσα σε άφωνο ήχο και σιωπή.

3. **Fractal Dimension**, η οποία για θόρυβο, σιωπή και άφωνους ήχους είναι μεγάλη (1.6), ενώ για έμφωνο λόγο είναι χαμηλή (1.1, 1.2).



Σχήμα 4.18: Οι τέσσερις απεικονίσεις για τη διάκριση σιωπής-φωνής-θορύβου, σε ένα σύνθετο σήμα. Πριν από τη λέξη /harsh/, στα 16kHz, προηγείται ένα θορυβώδες σήμα μεγάλου πλάτους και διάρκειας 45ms. Οι απεικονίσεις δημιουργήθηκαν με παράθυρα 15ms.

4. **ZeroCrossings Rate**, ο οποίος είναι εξαιρετικά μεγάλος για σήματα θορύβου λόγω της τυχαίας φύσης τους.

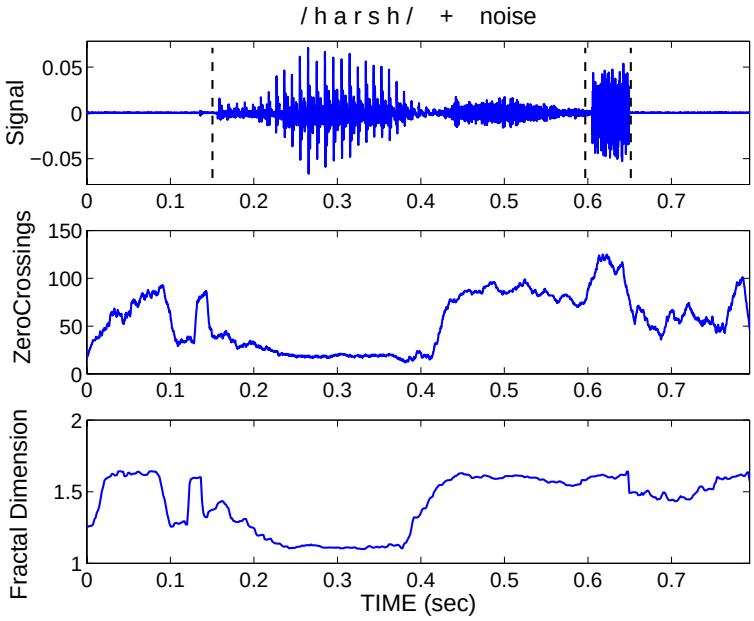
Ένα παράδειγμα με τα διαφορετικά αυτά μεγέθη για ένα σήμα σιωπής-θορύβου-φωνής-σιωπής φαίνεται στο Σχήμα (4.18). Πρόκειται για τη λέξη /harsh/ με 45ms τυχαίου θορύβου πριν την έναρξη της λέξης. Από το Σχήμα παρατηρούμε ότι ρυθμός zero-crossings του θορύβου είναι μεγαλύτερος και από τη σιωπή αλλά και από το τελικό τυρβώδες /sh/. Μάλιστα οι μεταβολές του θορύβου είναι τέτοιες ώστε οι τιμές του zero-crossings είναι της τάξεως του μισού παραθύρου ανάλυσης (N=240). Αυτό σημαίνει ότι για κάθε τέσσερα περίπου διαδοχικά δείγματα έχουμε δύο μεταβάσεις από το μηδέν.

Η διαδικασία με την οποία επιχειρήθηκε διάκριση των τριών σημάτων βασίζεται ξανά σε απόφαση με βάση πειραματικά κατώφλια. Περιληπτικά: Στο πρώτο στάδιο της διακρίνει με τη βοήθεια της ενέργειας και τη συχνότητας τη σιωπή πριν και μετά το υπόλοιπο σήμα, σύμφωνα με τη νέα μέθοδο ανίχνευσης φωνής. Κάτι τέτοιο είναι εφικτό αφού ο θόρυβος έχει πλάτος αρκετά μεγαλύτερο της σιωπής, και επομένως το σημείο όπου χωρίζονται τα δύο είναι εύκολα εντοπήσιμο. Με την προϋπόθεση ότι η διαδικασία θα δώσει αποδώσει ικανοποιητικά, έχουμε πλέον ένα σήμα φωνής-θορύβου.

Σε ένα δεύτερο στάδιο ελέγχεται η fractal διάσταση στην αρχή και στο

τέλος του σύνθετου σήματος και αναζητείται το διάστημα όπου αυτή ξεπερνάει αυστηρά ένα κατώφλι $TD_f = 1.6$. Αυτό το διάστημα με μεγάλη fractal διάσταση είναι ή άφωνος ήχος ή θόρυβος, από τη στιγμή που δεν υπάρχει σιωπή πλέον. Ελέγχεται και ο ρυθμός zero-crossings(zcr) του διαστήματος που εντοπίστηκε, ως προς το πόσες φορές ξεπερνάει ένα κατώφλι λίγο μικρότερο από το μισό μήκος παραθύρου, $TZc = 110$. Αν το κατώφλι ξεπερνιέται πάνω από 100 φορές, τότε το διάστημα είναι θόρυβος ή κάποιο εξαιρετικά τυρβώδες φώνημα, π.χ. /tch/. Αυτό ελέγχεται με ένα χρονικό περιορισμό σχετικά με το πρώτο σημείο που υπερβαίνεται το κατώφλι για πρώτη φορά. Από κει και πέρα αναζητείται το σημείο που ο zcr πέφτει κάτω από $0.6 * TZc$ και πλέον έχουμε και τα όρια του θορύβου.

Η διαδικασία για τις περισσότερες δοκιμές απέδωσε σχετικά καλά. Παρουσίασε βέβαια και προβλήματα στις περιπτώσεις κάποιων συριστικών, π.χ./s/ που παρουσίασαν υψηλό zcr και εντοπίστηκαν ως θόρυβος αλλά και στις περιπτώσεις όπου ο βασικός αλγόριθμος διάκρισης της σιωπής έδωσε μεγάλες αποκλίσεις από τα πραγματικά όρια. Επειδή αυτός είναι και ο κορμός της, σε περιπτώσεις που αρκετό τμήμα σιωπής συμπεριλήφθηκε στο σήμα θορύβου-



Σχήμα 4.19: Παράδειγμα αυτόματης διάκρισης των τριών ειδών σημάτων (Φωνής- Σιωπής- Θορύβου). Στο σχήμα φαίνονται η κυματομορφή /harsh/, με θόρυβο μετά το τέλος της λέξης με τα σημεία εντοπισμού των ορίων σιωπής (εξωτερικά), φωνής και θορύβου, καθώς και οι μετρήσεις zero-crossing και fractal διάστασης που εμπλέκονται στους υπολογισμούς.

σιωπής, ώστε να δώσει μεγάλη fractal διάσταση πριν ή μετά το θόρυβο, η διαδικασία εντοπισμού του θορύβου απέτυχε.

Στο Σχήμα (4.19), φαίνεται η διάκριση σήματος θορύβου μετά από τη λέξη /harsh/. Εκτός από τη σιωπή που ανιχνεύτηκε αρχικά (και φαίνεται με τις εξωτερικές διακεκομμένες) ανιχνεύεται και ο θόρυβος μετά το τελικό /sh/(που διαχωρίζεται με την εσωτερική διακεκομμένη γραμμή). Αν αυτό το τελικό /sh/ είχε μεγαλύτερο ρυθμό zcr, δηλαδή ισχυρότερη 'διέγερση', η διαδικασία μπορεί να μην απέδιδε.

Το σημαντικό πάντως από την προηγούμενη προσπάθεια είναι ότι με κατάλληλους συνδυασμούς κλασσικών (γραμμικών) εργαλείων αλλά και νέων (μη-γραμμικών) είναι δυνατή η κατηγοριοποίηση και η διάκριση τμημάτων σήματος διαφορετικής φύσεως.

4.4 Ανακεφαλαίωση

Τα νέα εργαλεία που προέκυψαν από τη μη-γραμμική μοντελοποίηση και επεξεργασία του φωνητικού σήματος και παρουσιάστηκαν στο προηγούμενο κεφάλαιο, εμπλέχθηκαν εδώ σε μια διαδικασία ανίχνευσης φωνής και σιωπής, ανάλογη με τον κλασσικό αλγόριθμο των Rabiner-Sambur. Οι νέες απεικονίσεις στο πεδίο του χρόνου παρουσιάστηκαν ως ισχυρά εργαλεία επεξεργασίας σημάτων που τονίζουν τις μεταβολές ανάμεσα στα διαφορετικά τμήματα φωνής, ανάλογα με τα κλασσικά και σε ορισμένες περιπτώσεις πολύ καλύτερα. Αυτό αντανακλάται στη βελτίωση που επιτυγχάνει η νέα μέθοδος με τη χρήση των νέων εργαλείων και η οποία είναι της τάξης του 11%. Διάφορες άλλες εναλλακτικές υλοποιήσεις προτάθηκαν για ανίχνευση φωνής με ειδικές παραμέτρους η κάθε μία. Τέλος ελέγχθηκε και η περίπτωση του θορύβου που δυσχεραίνει τό έργο της ανίχνευσης. Η νέα μέθοδος παρουσίασε ελαφρώς πιο ανθεκτική συμπεριφορά στο θόρυβο από την κλασσική μέθοδο και έδωσε ελπίδες για καλύτερα αποτελέσματα.

Κεφάλαιο 5

Επίλογος

Η παρούσα δουλειά ξεκίνησε ως μια προσπάθεια διερεύνησης νέων τρόπων ανίχνευσης φωνητικών σημάτων και διάκρισης τους από διάφορα άλλα σήματα ακουστικής αλλά όχι φωνητικής φύσεως. Το βασικότερο από αυτά τα γεγονότα είναι η σιωπή που μπορεί να υπάρχει πριν, μετά ή κατά τη διάρκεια συνεχούς λόγου ανάμεσα σε δύο πλευρές που ανταλλάσσουν πληροφορία. Η ανάγκη να απομονωθεί η πληροφορία, δηλαδή η φωνή, από ένα τέτοιο σύνθετο σήμα έχει οδηγήσει στη δημιουργία ενός ολόκληρου πεδίου έρευνας μεθόδων και αλγορίθμων ανίχνευσης φωνής. Η εφαρμογή τέτοιων μεθόδων μπορεί να βρεθεί σε σύγχρονα συστήματα αναγνώρισης φωνής, για τη δημιουργία των φωνητικών προτύπων και τηλεπικοινωνιών όπου το φωνητικό σήμα απαιτείται είτε για να μειωθεί ο χρόνος των διαδικασιών, είτε για κέρδος σε εύρος καναλιών και εξυπηρέτηση.

Το θέμα δεν είναι απλό και τα πράγματα δυσκολεύουν ακόμη περισσότερο από την παρουσία θορύβου επιπρόσθετα στο σήμα. Οι περισσότερες εφαρμογές πραγματικού χρόνου και ιδιαίτερα εμπορικών συστημάτων γίνονται σε περιβάλλον με πολλές πηγές θορύβου παρούσες, αρκετές φορές μη στατικές. Συστήματα εντολών και απόκρισης σε εξωτερικό περιβάλλον, τηλεφωνικά δίκτυα, ασύρματα και ενσύρματα που εισάγουν παρεμβολές, παραμορφώσεις και ζωνοπερατές συνιστώσες θορύβου κάνουν το έργο της διάκρισης της φωνής σύνθετο.

Πολλές προσπάθειες έχουνε γίνει, σχετικά με το θέμα (βλ. [4],[13]) που ως βασικό χαρακτηριστικό τους έχουν ότι προσεγγίζουν το θέμα από την πλευρά της ανίχνευσης της ενέργειας του σήματος. Η θεμελιώδης από αυτές είναι οι δουλειά των Rabiner - Sambur στο ζήτημα [10] που άνοιξε το δρόμο για σημερινές σύνθετες τεχνικές με κανόνες απόφασης, μεθοδολογίες αναγνώρισης προτύπων και περιορισμούς. Η μέθοδος που προτάθηκε χρησιμοποιεί κάποιες κλασσικές μετρήσεις και απεικονίσεις στο πεδίο του χρόνου, τις οποίες έχουν υιοθετήσει και μεταγενέστερες τεχνικές. Πρόκειται για τη μέτρηση της*ενέργειας*

του σήματος, που εκφράζει το άθροισμα των τετραγώνων των τιμών του σήματος σε τοπικά παράθυρα ανάλυσης, το *πλάτος* του και ο μέσος ρυθμός μεταβάσεων από το μηδέν (*zero-crossing rate*), μια έμμεση περιγραφή του φασματικού περιεχομένου του σήματος. Η μέθοδος αυτή δίνει πολύ καλά αποτελέσματα ανίχνευσης μεμονωμένων λέξεων για καλές όμως συνθήκες σηματοθορυβικού λόγου.

Εδώ δοκιμάστηκε μια μη-γραμμική προσέγγιση του θέματος, που όπως κάθε τι μη-γραμμικό φάνηκε ενδιαφέρον και πολλά υποσχόμενο. Βασιζόμενοι σε θεωρίες και μεθόδους που έχουν αναπτυχθεί στο πεδίο της μη-γραμμικής παραγωγής και μοντελοποίησης της φωνής (βλ.[5],[6],[3],[12]), οδηγηθήκαμε σε ανάλυση του σήματος σε πολλαπλά ζωνοπερατά κανάλια, και αποδιαμόρφωση μέσω του ESA των συνιστωσών του σήματος σε στιγμιαίο πλάτος και στιγμιαία συχνότητα, AM-FM σημάτων. Προτάθηκε μια μέθοδος, που στηρίζεται στον εντοπισμό της ισχυρότερης ενέργειας κάθε καναλιού, όπου πλέον η ενέργεια χρησιμοποιείται με αναφορά στην πηγή που δημιουργεί το σήμα. Η μέθοδος αυτή οδήγησε σε νέες και πολύ ενδιαφέρουσες απεικονίσεις για το ολικό σήμα πλέον και όχι για τα επιμέρους κανάλια του.

Οι απεικονίσεις αυτές, και μετά από μια διαδικασία τοπικής εξισορρόπησης, έδωσαν νέες εκφράσεις για το επίπεδο τιμών του σήματος (*ενέργεια, πλάτος*) αλλά και για το φασματικό του περιεχόμενο (*στιγμιαία συχνότητα*). Η υπόθεση πίσω από αυτές τις απεικονίσεις είναι βάσιμη καθώς η πιο ισχυρή ενεργειακή συνιστώσα στο φάσμα υπερτερεί των υπολοίπων και συμμετέχει με το στιγμιαίο πλάτος και τη στιγμιαία συχνότητα της. Επομένως κάτω από τις φαινομενικά ομαλές μεταβάσεις της περιβάλλουσας του σήματος ή τις μικρές διαφορές, τμημάτων διαφορετικής φύσεως μπορεί να κρύβονται ανώμαλες μεταβολές ή ακόμη μεγαλύτερες διαφορές που καλύπτονται από την υπέρθεση των συνιστωσών του σήματος. Αυτή την 'κρυμμένη' πληροφορία θελήσαμε να δεσμεύσουμε και να τη χρησιμοποιήσουμε ένα βήμα πιο πέρα. Για να δούμε πόσο θα μπορούσε να διευκολυνθεί η διάκριση όχι μόνο διαφορετικών τμημάτων ενός φωνητικού σήματος αλλά σημάτων διαφορετικής φύσεως όπως είναι η σιωπή και η φωνή.

Με γνώμονα τον πειραματισμό και τη δοκιμή αυτές οι νέες απεικονίσεις δοκιμάστηκαν σε μια μέθοδο ανίχνευσης φωνής από ένα ακουστικό περιβάλλον σιωπής. Δανειζόμενοι στοιχεία από τον κλασσικό αλγόριθμο, και ιδιαίτερα την απλή λογική του και τη στατιστική εξαγωγή πληροφορίας για τη σιωπή, αναπτύξαμε μια παρόμοια μέθοδο με κατώφλια και έλεγχο του ενεργειακού και του φασματικού επιπέδου του σήματος. Για το σκοπό αυτό η νέα ενεργειακή μέτρηση και η μέτρηση της στιγμιαίας συχνότητας χρησιμοποιήθηκαν ως βασικά εργαλεία μετά από τη βέλτιστη πειραματική απόδοση.

Η νέα μέθοδος έδωσε καλά και πολλά υποσχόμενα αποτελέσματα σε θέματα ανίχνευσης μεμονωμένων λέξεων με σιωπή πριν την αρχή και μετά το τέλος

τους. Σε σύγκριση με τον κλασσικό αλγόριθμο παρουσίασε μια βελτίωση ανίχνευσης των ορίων των λέξεων της τάξεως του 11%, ενώ ανάλογες εναλλακτικές υλοποιήσεις , με βάση πάντα τα νέα εργαλεία παρουσίασαν επίσης μια βελτιωμένη συμπεριφορά σε σχέση με την διαδικασία με τα κλασσικά εργαλεία. Φυσικά αρκετή δουλειά μπορεί να γίνει ακόμη για να περιοριστούν κάποια δυσμενή φαινόμενα όπως η ανίχνευση και άλλων γεγονότων μη-φωνητικής φύσεως μαζί με το σήμα φωνής (π.χ. 'clicks') αλλά και για τη μείωση της υπερεκτίμησης της σιωπής που μπορεί να παρατηρηθεί ανάλογα με την περίπτωση.

Ο παράγοντας θόρυβος δεν θα μπορούσε να μείνει εκτός από μια προσπάθεια για κάτι νέο. Το ερώτημα που προέκυψε ήτανε πόσο καλύτερα συμπεριφέρονται αυτοί οι νέοι τρόποι επεξεργασίας σημάτων αλλά και η νέα μέθοδος ανίχνευσης φωνής που βασίζεται σε αυτά. Μια πρώτη εκτίμηση και δοκιμές δώσανε ποιο ευσταθή συμπεριφορά για ποσοστά θορύβου της τάξεως των $40 \quad 30db$ επιπρόσθετα στο σήμα, ενώ για μεγάλα ποσά , $\leq 20db$ ο κλασσικός αλγόριθμος παρουσιάστηκε λίγο καλύτερος. Περισσότερες δοκιμές μπορούν να γίνουν για την επίδραση του θορύβου στις διαδικασίες ανίχνευσης για πιο ασφαλή συμπεράσματα και ίσως με πραγματικές ηχογραφήσεις όπου τα πράγματα είναι πιο απρόβλεπτα.

Το κόστος σε υπολογιστικό φόρτο που απαιτεί η εφαρμογή της νέας μεθόδου είναι σίγουρα μεγαλύτερο από την κλασσική προσέγγιση. Κάτι τέτοιο βέβαια δεν μπορεί να αποτελεί απαγορευτικό παράγοντα για εφαρμογές πραγματικού χρόνου, αφού επενδύουμε συνεχώς στην ταχύτητα και στην παράλληλη επεξεργασία πληροφορίας.

Προοπτικές και Έρευνα

Προοπτικές για έρευνα και επέκταση των τεχνικών που αναφέρθηκαν αλλά και των αποτελεσμάτων που προέκυψαν ανοίγονται τόσο για το θέμα της ανίχνευσης φωνής όσο και για την επεξεργασία σημάτων αλλά και την τεχνολογία φωνής γενικότερα.

Αρκετή δουλειά μπορεί να γίνει ακόμη για να διαπιστωθεί η απόδοση της νέας μεθόδου με γνώμονα την αναγνώριση των λέξεων που ανιχνεύονται. Το θέμα της αναγνώρισης δεν εξετάστηκε εδώ, και για εφαρμογές αναγνώρισης φωνής είναι αυτό που μετράει περισσότερο και όχι ο ακριβής εντοπισμός των ορίων των λέξεων. Έρευνες έχουν αποδείξει τη ρόλο του εντοπισμού των ορίων της φωνής σε θέματα αναγνώρισης (βλ.[13]). Μένει να διαπιστωθεί η αποδοτικότητα της νέας μεθόδου αλλά και το ποσοστό της πιθανής βελτίωσης που επιφέρει κάτω από θέματα αναγνώρισης.

Αλλά και στο θέμα του θορύβου είναι ακόμη ανοιχτό. Δοκιμές μπορούνε να γίνουνε με πραγματικές ηχογραφημένες διαταραχές (π.χ. λέξεις από κινητά τηλέφωνα ή μετά από ενσύρματη τηλεφωνική μετάδοση) για να διαπιστωθεί

η αξιοπιστία της νέας μεθόδου και κατά πόσο πιο ευσταθής είναι από την κλασσική. Τα πραγματικά θορυβώδη περιβάλλοντα είναι πιο απρόβλεπτα από κάθε υπολογιστική προσομοίωση θορύβου που μπορεί να επιτευχθεί. Η αποκάλυψη αντοχής σε θόρυβο, μια πεποίθηση που υπάρχει και βασίζεται στον τρόπο παραγωγής των νέων απεικονίσεων, μπορεί να βρει μεγάλη πρακτική εφαρμογή αφού οι εφαρμογές ενδιαφέροντος γίνονται σε πραγματικό χρόνο και κάτω από 'πραγματικές' συνθήκες.

Τα νέα εργαλεία εφαρμόστηκαν σε μια προσέγγιση της κλασσικής μεθόδου προσαρμοσμένη στα νέα δεδομένα. Θα μπορούσαν να δοκιμαστούν και σε ποιο σύνθετες υλοποιήσεις ανίχνευσης φωνής, π.χ. σε κάποια υβριδική μέθοδο όπου η ανίχνευση θα γίνεται παράλληλα με την ταύτιση προτύπων, με μεγαλύτερη ακρίβεια αποτελεσμάτων. Επίσης μπορούν να εμπλακούν στη διαδικασία περιορισμοί (π.χ. συντακτικοί ή σημαντικοί), ανάλογα με την εφαρμογή. Σκοπός μας δεν ήτανε η υλοποίηση κάποιας νέας καινοτομικής μεθόδου αλλά να φανεί η βελτίωση που μπορεί να επιτευχθεί σε κάποια διαδικασία ανίχνευσης μέσω των νέων απεικονίσεων του σήματος που προτάθηκαν.

Επίσης η ανίχνευση φωνής θα μπορούσε να επεκταθεί και για περιπτώσεις όχι απλά απομονωμένων λέξεων αλλά διαδοχικών λέξεων με σιωπή ανάμεσα τους. Μια τέτοια ανάπτυξη κάποιας μεθοδολογίας θα μπορούσε να είναι πολύ χρήσιμη για πραγματικές εφαρμογές, π.χ. φωνητικές εντολές με περισσότερες της μιας λέξης σε απομακρυσμένο υπολογιστικό σύστημα μέσω τηλεφωνικών γραμμών για πρόσβαση σε μια βάση δεδομένων. Για τέτοιες εφαρμογές εκτός από την ανίχνευση σιωπής πριν και μετά τη φράση , μπορεί να απαιτείται και ενδιάμεση ανίχνευση λόγο παύσης του ομιλητή.

Επιθυμία και προοπτική είναι να μπορούσε επίσης να ελεγχθεί το θέμα ανίχνευση φωνής και γενικότερα αλλά και η νέα μέθοδος που προτάθηκε για λέξεις από την ελληνική γλώσσα. Όπως σε όλα τα θέματα τεχνολογίας φωνής, οι μεθοδολογίες και οι αλγόριθμοι αναπτύσσονται με βάση την αγγλική διάλεκτο και τις ανάγκες της. Για λόγους αναφοράς και σύγκρισης αλλά και λόγο των διαθέσιμων δεδομένων η παρούσα προσπάθεια έγινε με βάση αυτή τη διάλεκτο. Η όλη δουλειά θα μπορούσε να ελεγχθεί και με βάση την ελληνική διάλεκτο, που είναι πλούσια σε λέξεις και προφορές.

Τέλος τα νέα εργαλεία που παρουσιάστηκαν εδώ μπορούν να χρησιμοποιηθούν γενικότερα στα πεδία της επεξεργασίας σημάτων και τεχνολογίας φωνής. Είναι δυνατή η χρήση τους ως χαρακτηριστικά ενός φωνητικού σήματος για θέματα αναγνώρισης φωνής, δηλ ως επιπλέον δεδομένο στα διανύσματα που επιτελούν την ταύτιση προτύπων , αφού αποκαλύπτουν πληροφορία σχετικά με το σήμα και την παραγωγή του που μπορεί να μη φαίνεται από άλλα χαρακτηριστικά (π.χ. σχετικά με τις διαμορφώσεις ή τις ισχυρές ενεργειακά συνιστώσες του). Επίσης γενικότερα στην επεξεργασία σημάτων όπου μπορεί να χρησιμοποιηθεί μοντελοποίηση με AM-FM σήματα, τα νέα εργαλεία

μπορούν να βρουν εφαρμογή π.χ. στην επεξεργασία εικόνων που μπορούν να μοντελοποιηθούν μέσω AM–FM 2Δ σημάτων ή στις τηλεπικοινωνίες για κωδικοποίηση ή αποκωδικοποίηση σημάτων.

Βιβλιογραφία

[1] A.C. Bovik, P. Maragos and T.F. Quatieri,"AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators", *IEEE Trans. on Signal Processing*, Vol.41, No. 12, pp. 3245-3265, December 1993

[2] B. Gold and N. Morgan, *Speech And Audio Signal Processing*, John Wiley And Sons Inc., 2000

[3] J.F. Kaiser, "On Teager's Energy Algorithm and it's generalization to Continuous Signals", *in Proc. 4th IEEE Digital Signal Processing Workshop*, September 1990

[4] L. F. Lamel, L.R. Rabiner, A.E. Rosenberg and J.G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-29, No. 4, pp. 777-785, August 1981

[5] P. Maragos, J.F. Kaiser and T.F. Quatieri,"On Amplitude and Frequency Demodulation Using Energy Operators", *IEEE Trans. on Signal Processing*, Vol. 41, No. 4, pp. 1532-1550, April 1993.

[6] P. Maragos, J.F. Kaiser and T.F. Quatieri,"Energy Separation in Signal Modulations with Application to Speech Analysis", *IEEE Trans. on Signal Processing*, Vol. 41, No. 10, pp. 3024-3051, October 1993.

[7] P. Maragos and A. Potamianos, "Fractal Dimensions of Speech Sounds: Computation and Application to Automatic Speech Recognition", *J.Acoust. Soc.Amer.* 105 (3), pp. 1925-1932, March 1999

[8] A. Potamianos and P. Maragos, "Speech Formant Frequency and Bandwidth Tracking using Multiband Energy Demodulation", *J.Acoust. Soc.Amer.* 99 (6), pp.3795-3806, June 1996

[9] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[10] L.R. Rabiner and M.R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", *Bell Systems Technical Journal*,Vol. 54, No. 2, pp.297-315, February 1975.

[11] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[12] H.M. Teager and S.M. Teager, "Evidence of Nonlinear Sound Production Mechanisms in the Vocal Tract", in *Speech Production and Speech Modelling*, Kluwer Academic, pp.241-261, 1990

[13] J.G. Wilpon, L.R. Rabiner and T.B. Martin, "An Improved Word-Detection Algorithm for Telephone-Quality Speech Incorporating Both Syntactic and Semantic Constraints", *AT&T Technical Journal*, Vol. 63, No. 3, pp. 479-498, March 1984

[14] G.S. Ying, C.D. Mitchell, L.H. Jamieson, "Endpoint Detection of Isolated Utterances based on a Modified Teager Energy Measurement", in *Proc. of the 1993 International Conference on Acoustics, Speech and Signal Processing*, April 1993