ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. ΚΑΘΟΡΙΣΜΟΣ ΑΚΟΥΣΤΙΚΗΣ ΠΟΙΟΤΗΤΑΣ ΒΑΣΙΣΜΕΝΟΣ ΣΕ ΤΕΧΝΙΚΕΣ ΜΕΤΡΗΣΗΣ ΑΝΤΙΛΗΨΗΣ

Περίληψη: Μία νέα, αξιοσημείωτη, προσέγγιση για τον καθορισμό της ακουστικής ποιότητας συζητάται εδώ. Η μέθοδος αυτή δεν χαρακτηρίζει το υπό δοκιμή ακουστικό σύστημα αλλά χαρακτηρίζει την αντίληψη του σήματος εξόδου του ακουστικού συστήματος. Με την σύγκριση της διαβαθμισμένης εξόδου με την ιδανική (αναφοράς), χρησιμοποιώντας ένα μοντέλο του ανθρώπινου ακουστικού συστήματος, μπορούν να γίνουν προβλέψεις για την υποκειμενικά παρατηρούμενη ακουστική ποιότητα της εξόδου του συστήματος χρησιμοποιώντας οποιοδήποτε σήμα εισόδου. Ένα αξιοσημείωτο μοντέλο χρησιμοποιείται για τον υπολογισμό των εσωτερικών αναπαραστάσεων τόσο της διαβαθμισμένης εισόδου όσο και της εισόδου αναφοράς. Ένα απλό εκπαιδευόμενο μοντέλο απεικονίζει τις διαφορές μεταξύ των εσωτερικών αναπαραστάσεων. Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί για την αξιολόγηση της ποιότητας των μουσικών κωδικοποιητών—αποκωδικοποιητών ευρείας ζώνης όπως και των κωδικοποιητών—αποκωδικοποιητών ομιλίας τηλεφωνικού εύρους ζώνης (300-3400 Hz). Η συσχέτιση μεταξύ υποκειμενικών και αντικειμενικών αποτελεσμάτων είναι μεγαλύτερη από 0.9 για μία μεγάλη ποικιλία βάσεων δεδομένων που προέρχονται από υποκειμενικές εκτιμήσεις ποιότητας κωδικοποιητών—αποκωδικοποιητών ομιλίας και μουσικής. Για την μέτρηση της ποιότητας των κωδικοποιητών—αποκωδικοποιητών ομιλίας τηλεφωνικής ζώνης δίνεται μία απλοποιημένη μέθοδος. Αυτή η μέθοδος έχει καθοριστεί από το International Telecommunications Union σαν πρόταση P.861.

2.1 ΕΙΣΑΓΩΓΗ

Με την εισαγωγή και τον καθορισμό των νέων, βασιζόμενων στην αντίληψη, ακουστικών (μουσικής και ομιλίας) κωδικοποιητών—αποκωδικοποιητών [ISO92st,1993], [ISO94st,1994], [ETSIstdR06,1992], [CCITTrecG728, 1992], [CCITTrecG729,1995], οι κλασσικές μέθοδοι για την μέτρηση της ακουστικής ποιότητας, όπως ο λόγος σήματος προς θόρυβο και η ολική αρμονική διασπορά, παραγκωνίστηκαν. Έτσι κατά την διαδικασία του καθορισμού αυτών των κωδικοποιητών—αποκωδικοποιητών η ποιότητα των διαφόρων προτάσεων εκτιμήθηκε μόνο υποκειμενικά (βλέπε πχ [Natvig, 1988], [ISO90,1990] και [ISO91, 1991]). Παρόλα αυτά οι υποκειμενικές εκτιμήσεις είναι χρονοβόρες, ακριβές και δύσκολο να αναπαραχθούν.

Μία βασική ερώτηση είναι αν οι υποκειμενικές μέθοδοι μπορούν να τροποποιηθούν έτσι ώστε να μπορούν να χρησιμοποιηθούν για πρόβλεψη της υποκειμενικής ποιότητας τέτοιων τεχνικών κωδικοποίησης της αντίληψης με ένα αξιόπιστο τρόπο. Μία διαφορά με τις κλασσικές προσεγγίσεις για την εκτίμηση της ακουστικής ποιότητας είναι ότι η περιγραφή του συστήματος δεν είναι πλέον χρήσιμη λόγω των χρονικά μεταβαλλόμενων, προσαρμοστικού σήματος (signal adaptive), τεχνικών που χρησιμοποιούνται σε αυτούς τους κωδικοποιητές—αποκωδικοποιητές. Γενικά η ποιότητα των μοντέρνων ακουστικών κωδικοποιητών—αποκωδικοποιητών εξαρτάται από το σήμα εισόδου. Η πρόσφατα ανεπτυγμένη μέθοδος λοιπόν πρέπει να είμαι ικανή να μετράει την ποιότητα των κωδικοποιητών—αποκωδικοποιητών χρησιμοποιώντας οποιοδήποτε ακουστικό σήμα, όπως ομιλία, μουσική και

δοκιμαστικά σήματα. Μέθοδοι οι οποίες βασίζονται πάνω σε δοκιμαστικά σήματα μόνο, με ή χωρίς την χρήση μοντέλων αντίληψης, δεν μπορούν να χρησιμοποιηθούν.

Αυτό το κεφάλαιο θα παρουσιάσει μία γενική μέθοδο για την μέτρηση της ποιότητας των ακουστικών συσκευών συμπεριλαμβάνοντας τους βασιζόμενους στην αντίληψη κωδικοποιητές—αποκωδικοποιητές. Η μέθοδος αυτή χρησιμοποιεί την ιδέα της εσωτερικής αναπαράστασης ήχου, της αναπαράστασης που πλησιάζει όσο το δυνατόν περισσότερο αυτή που χρησιμοποιείται για την κρίση της ποιότητας από το ανθρώπινο ακουστικό σύστημα. Η είσοδος και η έξοδος της ακουστικής συσκευής απεικονίζονται πάνω στην εσωτερική αναπαράσταση σήματος και η διαφορά σε αυτή την αναπαράσταση χρησιμοποιείται για τον καθορισμό ενός μέτρου αντιληπτής ακουστικής ποιότητας (Perceptual Audio Quality Measure—PAQM). Θα δειχθεί ότι αυτό το PAQM έχει υψηλή συσχέτιση με την υποκειμενικά αντιληπτή ακουστική ποιότητα, ειδικά όταν οι διαφορές στην εσωτερική αναπαράσταση αναλύονται, με ένα περιγραφικό εξαρτημένο τρόπο, από ένα τύπο εκμάθησης (cognitive module). Ακόμα περισσότερο μία απλοποιημένη μέθοδος, η οποία προήλθε από το PAQM, για την μέτρηση της ποιότητας των τηλεφωνικού εύρους ζώνης (300-3400 Hz) κωδικοποιητών—αποκωδικοποιητών ομιλίας παρουσιάζεται. Αυτή η μέθοδος έχει καθοριστεί από το ITU-T (International Telecommunications Union-Telecom Sector) σαν πρόταση P.861 [ITUTrecP861, 1996].

2.2 ΒΑΣΙΚΗ ΦΙΛΟΣΟΦΙΑ ΜΕΤΡΗΣΕΩΝ

Στην βιβλιογραφία για την μέτρηση της ποιότητας των ακουστικών συσκευών ως επί το πλείστον κάποιος μπορεί να βρει τεχνικές μέτρησης οι οποίες περιγράφουν τις ακουστικές συσκευές με την χρήση δοκιμών. Η περιγραφή πρέπει να χτιστεί πάνω στην γνώση της ανθρώπινης ακουστικής αντίληψης ή να εξηγηθεί με βάση την ανθρώπινη ακουστική αντίληψη.

Για γραμμικά, χρονικά αμετάβλητα συστήματα μία πλήρης περιγραφή δίνεται από την παλμική ή μιγαδική απόκριση συχνότητας (impulse or complex frequency response) [Papoulis, 1977]. Με την επεξήγηση, βάση της αντίληψης, αυτής της περιγραφής κάποιος μπορεί να καθορίσει την ακουστική ποιότητα του υπό δοκιμή συστήματος. Αν ο σχεδιαστικός στόχος του υπό δοκιμή συστήματος είναι να είναι διαπερατό (δεν υπάρχουν παρατηρήσιμες διαφορές μεταξύ εισόδου και εξόδου) τότε η αξιολόγηση της ποιότητας είναι απλή και επικεντρώνεται στην απαίτηση σταθερού πλάτους και απόκρισης φάσης (μέσα σε ένα καθορισμένο πλαίσιο) πάνω από το εύρος των ακουστών συχνοτήτων (20-20000 Hz).

Για συστήματα που είναι κατά προσέγγιση γραμμικά ή χρονικά μεταβαλλόμενα η ιδέα της παλμικής (μιγαδικής συχνότητας) απόκρισης μπορεί επίσης να εφαρμοστεί. Για ασθενή μη γραμμικά συστήματα η περιγραφή μπορεί να επεκταθεί συμπεριλαμβάνοντας μετρήσεις της μη γραμμικότητας (θόρυβος, διασπορά, clipping point). Για χρονικά μεταβαλλόμενα συστήματα η περιγραφή μπορεί να επεκταθεί συμπεριλαμβάνοντας μετρήσεις της εξάρτησης από τον χρόνο της παλμικής απόκρισης. Μερικές από τις βοηθητικές μετρήσεις συμπεριλαμβάνουν γνώση του ανθρώπινου ακουστικού συστήματος η οποία οδηγεί σε περιγραφές συστημάτων που έχουν απευθείας σύνδεση με την ακουστική αντιληπτή ποιότητα (πχ το αντιλαμβανόμενο ζυγισμένο(weighted) λόγο σήματος προς θόρυβο).

Το προτέρημα της προσέγγισης της περιγραφής του συστήματος είναι ότι αυτή είναι (ή καλύτερα πρέπει να είναι) σε μεγάλο βαθμό ανεξάρτητη από τα σήματα δοκιμής που χρησιμοποιούνται. Οι περιγραφές μπορούν έτσι να μετρούνται με καθορισμένα σήματα και διαδικασίες μέτρησης. Αν και η περιγραφή του συστήματος

είναι συνήθως ανεξάρτητη από τα σήματα η υποκειμενικά αντιλαμβανόμενη ποιότητα στις περισσότερες περιπτώσεις εξαρτάται από το ακουστικό σήμα που χρησιμοποιείται. Αν πάρουμε πχ ένα σύστημα που προσθέτει λευκό θόρυβο στο σύστημα εισόδου τότε η αντιλαμβανόμενη ακουστική ποιότητα θα είναι πολύ υψηλή αν το σήμα εισόδου είναι ευρείας ζώνης. Για ένα σήμα εισόδου ευρείας ζώνης ο θόρυβος που εισάγεται από το ακουστικό σύστημα θα αποκρύπτεται από το σήμα εισόδου. Για ένα σήμα εισόδου ευρείας ζώνης ο θόρυβος θα είναι παρατηρήσιμος σε περιοχές συχνοτήτων όπου δεν υπάρχει ενέργεια σήματος εισόδου. Οπότε σε αυτή την περίπτωση οι περιγραφές των συστημάτων δεν χαρακτηρίζουν την αντιλαμβανόμενη ποιότητα του σήματος εξόδου.

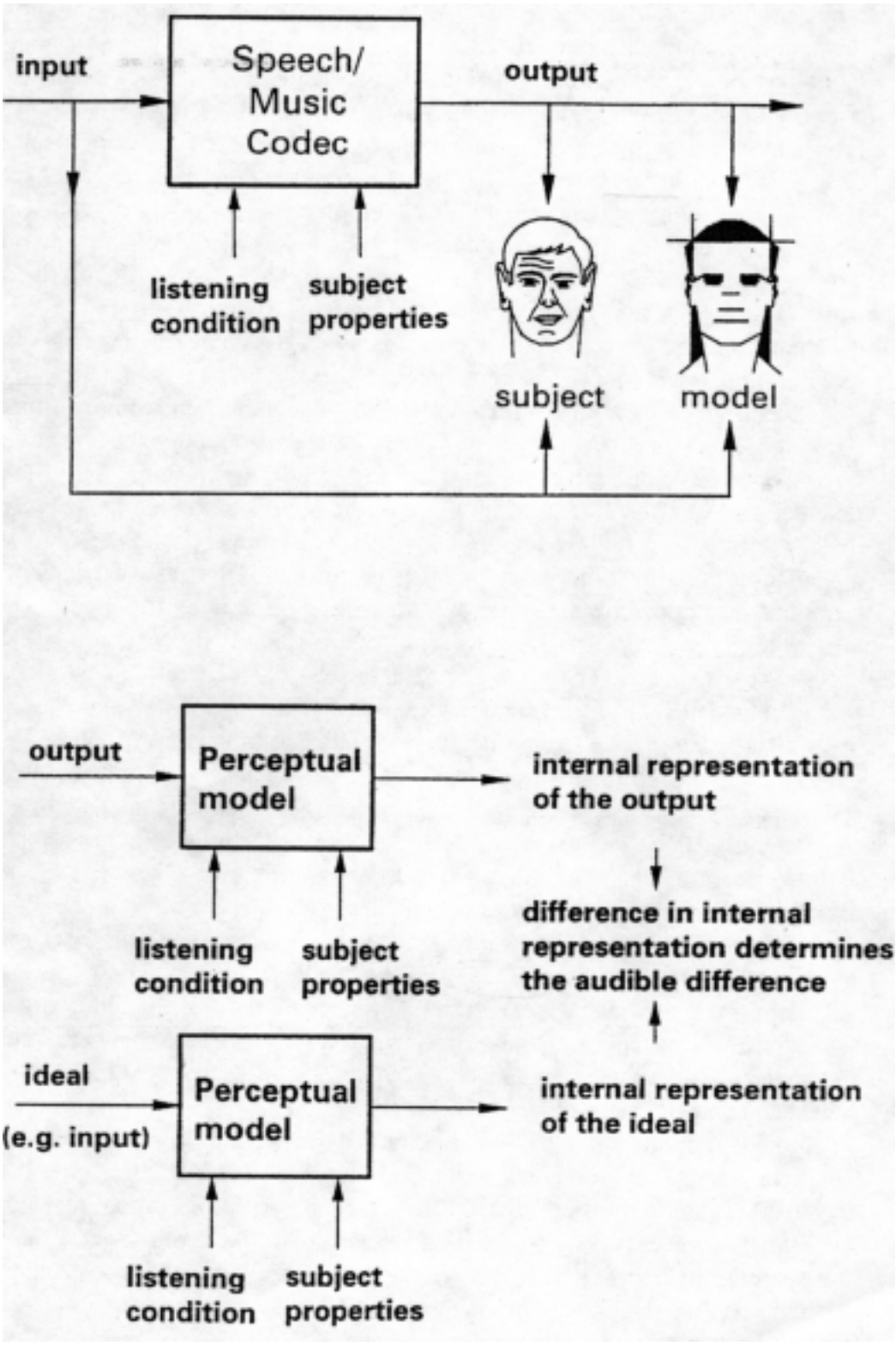
Ένα μειονέκτημα της προσέγγισης περιγραφής συστήματος είναι το γεγονός ότι αν και η περιγραφή ισχύει για μία μεγάλη ποικιλία σημάτων εισόδου μπορεί μόνο να μετρηθεί πάνω στην βάση της γνώσης του συστήματος. Αυτό οδηγεί σε περιγραφές οι οποίες εξαρτώνται από τον τύπο του συστήματος που δοκιμάζεται. Ένα σημαντικό πρόβλημα στην προσέγγιση περιγραφής συστήματος είναι το ότι είναι εξαιρετικά δύσκολο να περιγραφούν συστήματα που έχουν μη γραμμική και χρονικά μεταβαλλόμενη συμπεριφορά.

Μία εναλλακτική προσέγγιση της περιγραφής συστήματος, που ισχύει για κάθε σύστημα είναι η προσέγγιση της αντίληψης (perceptual approach). Σε αυτό το κεφάλαιο μία προσέγγιση της αντίληψης παρουσιάζεται σαν μία προσέγγιση στην οποία οι πτυχές την ανθρώπινης αντίληψης μοντελοποιούνται έτσι ώστε να κάνουμε μετρήσεις σε ακουστικά σήματα τα οποία έχουν υψηλή συσχέτιση με την υποκειμενικά αντιλαμβανόμενη ποιότητα αυτών των σημάτων και η οποία μπορεί να εφαρμοστεί σε κάθε σήμα, δηλαδή, ομιλία, μουσική και σήματα δοκιμής.

Στην προσέγγιση της αντίληψης κάποιος δεν χαρακτηρίζει το υπό δοκιμή σύστημα αλλά την ακουστική ποιότητα του σήματος εξόδου του υπό δοκιμή συστήματος. Αυτή χρησιμοποιεί το ιδανικό σήμα σαν αναφορά και ένα ακουστικό μοντέλο αντίληψης για τον καθορισμό των διαφορών που είναι δυνατό να ακουστούν μεταξύ του σήματος εξόδου και του ιδανικού. Για ακουστικά συστήματα που πρέπει να είναι διαπερατά το ιδανικό σήμα είναι το σήμα εισόδου. Μία γενική εικόνα της βασικής φιλοσοφίας που χρησιμοποιείται στις τεχνικές μέτρησης ακουστικής αντιλαμβανόμενης ποιότητας δίνεται στο σχήμα 1.

Το μεγάλο προτέρημα της προσέγγισης αντίληψης είναι το ότι είναι ανεξάρτητη του συστήματος και μπορεί να εφαρμοστεί σε κάθε σύστημα, συμπεριλαμβάνοντας συστήματα που παρουσιάζουν μη γραμμική και χρονικά μεταβαλλόμενη συμπεριφορά. Ένα μειονέκτημα είναι το ότι για την περιγραφή της ακουστικής ποιότητας ενός συστήματος κάποιος χρειάζεται ένα μεγάλο σύνολο από σχετικά δοκιμαστικά σήματα (σήματα ομιλίας και μουσικής).

Εάν η προσέγγιση αντίληψης χρησιμοποιείται για την πρόβλεψη της υποκειμενικά αντιληπτής ακουστικής ποιότητας της εξόδου ενός γραμμικού, χρονικά αμετάβλητου συστήματος τότε η προσέγγιση περιγραφής συστήματος και η προσέγγιση αντίληψης πρέπει να οδηγήσουν στο ίδιο αποτέλεσμα. Στην προσέγγιση περιγραφής συστήματος κάποιος πρέπει πρώτα να περιγράψει το σύστημα και μετά να εξηγήσει τα αποτελέσματα χρησιμοποιώντας την γνώση τόσο του ακουστικού συστήματος όσο και του σήματος εισόδου για τα οποία κάποιος θέλει να καθορίσει την ποιότητα. Στην προσέγγιση αντίληψης κάποιος πρέπει να περιγράψει την ποιότητα αντίληψης των σημάτων εξόδου που προέρχονται από τα σήματα εισόδου αναφοράς.



Σχήμα 1: Γενική παρουσίαση της βασικής φιλοσοφίας που χρησιμοποιείται στην ανάπτυξη των τεχνικών μέτρησης αντιληπτής ακουστικής ποιότητας. Ένα υπολογιστικό μοντέλο του ανθρώπου χρησιμοποιείται για να συγκρίνει την έξοδο της υπό δοκιμή συσκευής (πχ κωδικοποιητής—αποκωδικοποιητής ομιλίας ή μουσικής) με την ιδανική, χρησιμοποιώντας οποιοδήποτε ακουστικό σήμα. Αν η υπό δοκιμή συσκευή πρέπει να είναι διαπερατή τότε η ιδανική έξοδος είναι ίση με την είσοδο

Μέχρι πρόσφατα αρκετές τεχνικές μετρήσεων αντίληψης έχουν προταθεί αλλά οι περισσότερες από αυτές είναι εστιασμένες είτε πάνω στην ποιότητα των κωδικοποιητών—αποκωδικοποιητών ομιλίας [Gray and Markel, 1976], [Scroeder et al.,1979], [Gray et al., 1980], [Nocerino et al., 1985], [Quackenbush et al., 1988], [Hayasi and Kitawaki, 1992], [Halka and Heute, 1992], [Wang et al., 1992], [Ghitza, 1994], [Beerends and Stemerdink, 1994b] είτε πάνω στην ποιότητα των μουσικών κωδικοποιητών—αποκωδικοποιητών [Pailard et al., 1992], [Brandenburg and Sporer, 1992], [Beerends and Stemerdink, 1992], [Colomes et al., 1992]. Αν και κάποιος μπορεί να περιμένει ότι ένα μοντέλο για την μέτρηση της ποιότητας των μουσικών κωδικοποιητών—αποκωδικοποιητών ευρείας ζώνης μπορεί να εφαρμοστεί στους τηλεφωνικής ζώνης κωδικοποιητές—αποκωδικοποιητές ομιλίας, οι πρόσφατες έρευνες δείχνουν ότι κάτι τέτοιο είναι μάλλον δύσκολο [Beerends, 1995].

Σε αυτό το κεφάλαιο παρουσιάζεται γενικά το μέτρο της αντιληπτής ακουστικής ποιότητας (PAQM) [Beerends and Stemerdink, 1992] και θα δειχθεί ότι η προσέγγιση PAQM μπορεί να χρησιμοποιηθεί για την μέτρηση της ποιότητας κωδικοποιητών—αποκωδικοποιητών ομιλίας και μουσικής. Η μέθοδος PAQM είναι προς το παρόν υπό μελέτη στο ITU-R (International Telecommunication Union—Radio Sector) [ITURsg10con9714, 1997], [ITURsg10con9719, 1997] για μελλοντικό καθορισμό μιας βασισμένης στην αντίληψη μεθόδου μέτρησης ακουστικής ποιότητας. Μία απλοποιημένη μέθοδος, η οποία προέρχεται από την PAQM, για την μέτρηση της ποιότητας κωδικοποιητών—αποκωδικοποιητών ομιλίας τηλεφωνικής ζώνης (300-3400 Hz) έχει καθοριστεί από το ITU-T (International Telecommunication Union—Telecom Section) σαν πρόταση P.861 [ITUTrecP861, 1996], [ITUTsg12rep31.96, 1996]. Η ανεξάρτητη επαλήθευση αυτής της απλοποιημένης μεθόδου, η οποία καλείται μέτρηση αντιληπτής ποιότητας ομιλίας (Perceptual Speech Quality Measure—PSQM) έδειξε μεγάλη συσχέτιση μεταξύ υποκειμενικών και αντικειμενικών αποτελεσμάτων, όταν συγκρίθηκε με αρκετές άλλες μεθόδους [ITUTsg12con9674, 1996].

Ένα γενικό πρόβλημα στην ανάπτυξη τεχνικών μέτρησης αντίληψης είναι το ότι κάποιος χρειάζεται ακουστικά σήματα για τα οποία η υποκειμενική ποιότητα, όταν συγκρίνεται με ένα σήμα αναφοράς, είναι γνωστή. Η δημιουργία βάσεων δεδομένων ακουστικών σημάτων και των ακουστικών ποιοτήτων τους είναι εικονική και πολλά από τα προβλήματα που συναντούνται στις υποκειμενικές δοκιμές έχουν μία απευθείας συσχέτιση με τα προβλήματα στις τεχνικές μέτρησης αντίληψης.

Υψηλές συσχετίσεις μεταξύ υποκειμενικών και αντικειμενικών αποτελεσμάτων μπορούν να επιτευχθούν μόνο όταν οι υποκειμενικές και αντικειμενικές αξιολογήσεις είναι στενά συνδεδεμένες. Στην επόμενη ενότητα μερικά ενδιαφέροντα σημεία παρουσιάζονται τα οποία αφορούν την σχέση μεταξύ υποκειμενικών και αντικειμενικών δοκιμών αντίληψης.

2.3 ΥΠΟΚΕΙΜΕΝΙΚΕΣ ΑΝΤΙ ΑΝΤΙΚΕΙΜΕΝΙΚΩΝ ΔΟΚΙΜΕΣ ΑΝΤΙΛΗΨΗΣ

Κατά την ανάπτυξη τεχνικών μέτρησης αντίληψης χρειάζονται βάσεις δεδομένων με αξιόπιστες κρίσεις ποιότητας, οι οποίες είναι προτιμότερο να χρησιμοποιούν την ίδιες πειραματικές συνθήκες (setup) και την ίδια κοινή υποκειμενική κλίμακα ποιότητας.

Όλα τα υποκειμενικά αποτελέσματα που θα χρησιμοποιηθούν σε αυτό το κεφάλαιο προέρχονται από μεγάλες βάσεις δεδομένων του ITU για την δημιουργία των οποίων ζητήθηκε από άτομα να δώσουν την γνώμη τους για την ποιότητα ενός ακουστικού κομματιού χρησιμοποιώντας μία κλίμακα αξιολόγησης των πέντε

σημείων. Ο μέσος όρος των κρίσεων της ποιότητας που δίνουν οι ερωτηθέντες δίνει το αποκαλούμενο αποτέλεσμα μέσης γνώμης (Mean Opinion Score—MOS) σε αυτή την κλίμακα των πέντε σημείων. Χρησιμοποιούνται υποκειμενικά πειράματα στα οποία έχουν αξιολογηθεί η ποιότητα των κωδικοποιητών—αποκωδικοποιητών ομιλίας τηλεφωνικής ζώνης (300-3400 Hz) ή μουσικών κωδικοποιητών—αποκωδικοποιητών ευρείας ζώνης (20-20000 Hz ποιότητα compact disk). Και για τα δύο, αξιολόγηση κωδικοποιητών—αποκωδικοποιητών ομιλίας και μουσικής, χρησιμοποιείται η κλίμακα πέντε σημείων ITU MOS αλλά οι διαδικασίες που ακολουθούνται κατά την αξιολόγηση κωδικοποιητών—αποκωδικοποιητών ομιλίας [CCITTrecP80, 1994] είναι διαφορετικές από τις πειραματικές διαδικασίες που ακολουθούνται κατά την αξιολόγηση μουσικών κωδικοποιητών—αποκωδικοποιητών [CCIRrec562, 1990], [ITURrecBS1116, 1994].

Κατά τις αξιολογήσεις κωδικοποιητών—αποκωδικοποιητών ομιλίας, απόλυτη αξιολόγηση κατηγορίας (Absolute Category Rating—ACR) εφαρμόζονταν με ταμπέλες ποιότητας που κυμαίνονταν από κακή (MOS=1.0) έως άριστη (MOS=5.0) [CCITTrecP80, 1994]. Στα πειράματα ACR οι ερωτηθέντες δεν έχουν πρόσβαση στο αυθεντικό μη κωδικοποιημένο ακουστικό σήμα. Στις αξιολογήσεις μουσικών κωδικοποιητών—αποκωδικοποιητών μία κλίμακα διαβαθμισμένης αξιολόγησης κατηγορίας εφαρμόζεται (degradation category rating—DCR) με ταμπέλες ποιότητας που κυμαίνονται από την περίπτωση όπου «η διαφορά είναι δυνατόν να ακουστεί και είναι πολύ ενοχλητική» (MOS=1.0) έως την περίπτωση όπου «η διαφορά δεν είναι αντιληπτή» (MOS=5.0). Οι βάσεις δεδομένων των κωδικοποιητών—αποκωδικοποιητών μουσικής που χρησιμοποιούνται σε αυτή την δημοσίευση έχουν προέλθει όλες από πειράματα DCR όπου οι ερωτηθέντες είχαν ένα γνωστό και ένα κρυφό σήμα αναφοράς [ITURrecBS1116, 1994].

Γενικά δεν επιτρέπεται να συγκρίνουμε τιμές MOS οι οποίες προέρχονται από διαφορετικές πειραματικές συνθήκες. Ένα τμήμα ομιλίας τηλεφωνικού εύρους ζώνης μπορεί να έχει ένα MOS το οποίο να είναι πάνω από 4.0 σε κάποιες συγκεκριμένες πειραματικές συνθήκες ενώ το ίδιο τμήμα μπορεί να έχει ένα MOS το οποίο να είναι χαμηλότερο από 2.0 σε άλλες πειραματικές συνθήκες. Ακόμα και αν οι τιμές MOS λαμβάνονται μέσα από τις ίδιες πειραματικές συνθήκες αλλά μέσα σε διαφορετικό πολιτιστικό (cultural) περιβάλλον μεγάλες διαφορές στις τιμές MOS μπορούν να προκύψουν [Goodman and Nash, 1982]. Είναι έτσι αδύνατον να αναπτυχθεί μία τεχνική μέτρησης αντίληψης η οποία θα προβλέπει σωστές τιμές MOS κάτω από όλες τις συνθήκες.

Πριν κάποιος αρχίσει να προβλέπει τιμές MOS κάποια προβλήματα πρέπει πρώτα να επιλυθούν. Το πρώτο είναι το ότι διαφορετικοί άνθρωποι έχουν διαφορετικά ακουστικά συστήματα τα οποία οδηγούν σε ένα μεγάλο εύρος δυνατών μοντέλων. Εάν κάποιος θέλει να καθορίσει την ποιότητα των τηλεφωνικού εύρους ζώνης κωδικοποιητών—αποκωδικοποιητών ομιλίας (300-3400 Hz) οι διαφορές μεταξύ των ανθρώπων έχουν ελάχιστη σημασία. Κατά τον καθορισμό της ποιότητας των μουσικών κωδικοποιητών—αποκωδικοποιητών ευρείας ζώνης (ποιότητα compact disk, 20-20000 Hz) οι διαφορές μεταξύ των ανθρώπων είναι βασικό πρόβλημα ειδικά εάν οι κωδικοποιητές—αποκωδικοποιητές παρουσιάζουν περιορισμό δυναμικής ζώνης στο εύρος των 10-20 kHz. Πρέπει μία τεχνική αντικειμενικής μέτρησης αντίληψης να χρησιμοποιεί ένα ακουστικό μοντέλο το οποίο να αναπαριστάνει το καλύτερο διαθέσιμο αφτί, το οποίο μοντελοποιεί τον μέσο άνθρωπο, ή να χρησιμοποιεί ένα ξεχωριστό μοντέλο για κάθε άνθρωπο [Treurniet, 1996]. Η απάντηση εξαρτάται από την εφαρμογή. Για την πρόβλεψη των αποτελεσμάτων κοινής γνώμης (MOS) πρέπει να υιοθετηθεί το ακουστικό μοντέλο του μέσου ανθρώπου. Σε αυτό το κεφάλαιο όλες οι μετρήσεις αντίληψης

πραγματοποιήθηκαν με ένα όριο ενός μέσου ανθρώπου ηλικίας μεταξύ 20 και 30 ετών και ένα όριο μέγιστης συχνότητας στα 18 kHz. Ακριβή δεδομένα για τους ανθρώπους δεν ήταν διαθέσιμα.

Ένα άλλο πρόβλημα στις υποκειμενικές δοκιμές είναι το ότι ο τρόπος με τον οποίο το ακουστικό ερέθισμα παρουσιάζεται έχει μεγάλη επιρροή στην αντιλαμβανόμενη ακουστική ποιότητα. Είναι η παρουσίαση μέσα σε ένα ήσυχο δωμάτιο ή υπάρχει κάποιος θόρυβος στο υπόβαθρο (background noise) ο οποίος αποκρύπτει τις μικρές διαφορές; Τα ερεθίσματα παρουσιάζονται με ηχητικά συστήματα (loudspeakers) τα οποία εισάγουν αλλοιώσεις είτε με τα ηχεία αυτού καθεαυτού ή με αλληλεπίδραση με το δωμάτιο ακοής (listening room); Επιτρέπεται οι άνθρωποι να ρυθμίσουν την ένταση για κάθε ακουστικό κομμάτι; Μερικές από αυτές τις διαφορές όπως το επίπεδο της ακουστότητας και ο θόρυβος υποβάθρου, μπορούν να μοντελοποιηθούν στις μετρήσεις αντίληψης σχετικά εύκολα, αν και σε άλλες περιπτώσεις είναι σχεδόν αδύνατον. Μία μη πρακτική λύση σε αυτό το πρόβλημα είναι να πραγματοποιηθούν εγγραφές του σήματος εξόδου της υπό δοκιμής συσκευής και του σήματος αναφοράς (σήμα εισόδου) στην είσοδο του αφτιού των ανθρώπων και να χρησιμοποιηθούν αυτά τα σήματα στην αξιολόγηση αντίληψης.

Σε αυτό το κεφάλαιο όλες οι αντικειμενικές μετρήσεις αντίληψης πραγματοποιούνται απευθείας πάνω ηλεκτρικό σήμα εξόδου των κωδικοποιητών—αποκωδικοποιητών χρησιμοποιώντας ένα επίπεδο ρύθμισης (level setting) το οποίο αναπαριστά το μέσο επίπεδο ακοής στο πείραμα. Ακόμα περισσότερο ο θόρυβος υποβάθρου που είναι παρών κατά την διάρκεια των πειραμάτων ακρόασης μοντελοποιήθηκε με την χρήση θορύβου Hoth σταθερής κατάστασης [CCITTsup13, 1989]. Σε μερικά πειράματα οι συμμετέχοντες επιτρέπονταν να ρυθμίσουν μόνοι τους το επίπεδο του κάθε ακουστικού κομματιού πράγμα το οποίο οδηγεί σε συσχετίσεις οι οποίες είναι πιθανόν χαμηλότερες από αυτές που θα προέρχονταν αν το επίπεδο στο υποκειμενικό πείραμα ήταν προκαθορισμένο για όλα τα τμήματα. Σωστή ρύθμιση του επιπέδου αποδείχθηκε ότι είναι πολύ χρήσιμη για τις μετρήσεις αντίληψης.

Είναι ξεκάθαρο ότι υψηλές συσχετίσεις μεταξύ αντικειμενικών μετρήσεων και υποκειμενικών αποτελεσμάτων ακρόασης μπορούν να επιτευχθούν μόνο όταν το πειραματικό περιβάλλον είναι γνωστό και μπορεί να ληφθεί υπόψη σωστά από το μοντέλο εκμάθησης ή αντίληψης.

Το μοντέλο αντίληψης όπως αναπτύσσεται σε αυτό το κεφάλαιο χρησιμοποιείται για να απεικονίσει την είσοδο και την έξοδο της ακουστικής συσκευής πάνω σε εσωτερικές αναπαραστάσεις οι οποίες προσεγγίζουν όσο το δυνατόν περισσότερο τις εσωτερικές αναπαραστάσεις που χρησιμοποιούνται από τον άνθρωπο για την κρίση της ακουστικής συσκευής. Αποδεικνύεται ότι η διαφορά στην εσωτερική αναπαράσταση μπορεί να διαμορφώσει την βάση ενός μέτρου αντίληψης ακουστικής ποιότητας (PAQM) το οποίο έχει υψηλή συσχέτιση με την υποκειμενικά αντιληπτή ακουστική ποιότητα. Ακόμα περισσότερο αποδεικνύεται ότι με ένα απλό τύπο εκμάθησης ο οποίος επεξηγεί την διαφορά στην εσωτερική αναπαράσταση η συσχέτιση μεταξύ υποκειμενικών και αντικειμενικών αποτελεσμάτων είναι πάντα πάνω από 0.9 τόσο για τα ευρείας ζώνης μουσικά σήματα όσο και για τα τηλεφωνικής ζώνης σήματα ομιλίας. Για την μέτρηση της ποιότητας των τηλεφωνικής ζώνης κωδικοποιητών—αποκωδικοποιητών ομιλίας παρουσιάζεται μία απλοποιημένη έκδοση του PAQM, το μέτρο αντίληψης ποιότητας ομιλίας (Perceptual Speech Quality Measure—PSQM).

Πριν από την παρουσίαση της μεθόδου για τον υπολογισμό των εσωτερικών αναπαραστάσεων τα βασικά σημεία της ψυχοακουστικής του μοντέλου αντίληψης εξηγούνται στο ακόλουθο κεφάλαιο.

2.4 ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΨΥΧΟΑΚΟΥΣΤΙΚΗΣ ΓΙΑ ΤΟΝ ΥΠΟΛΟΓΙΣΜΟ ΤΗΣ ΕΣΩΤΕΡΙΚΗΣ ΑΝΑΠΑΡΑΣΤΑΣΗΣ ΗΧΟΥ

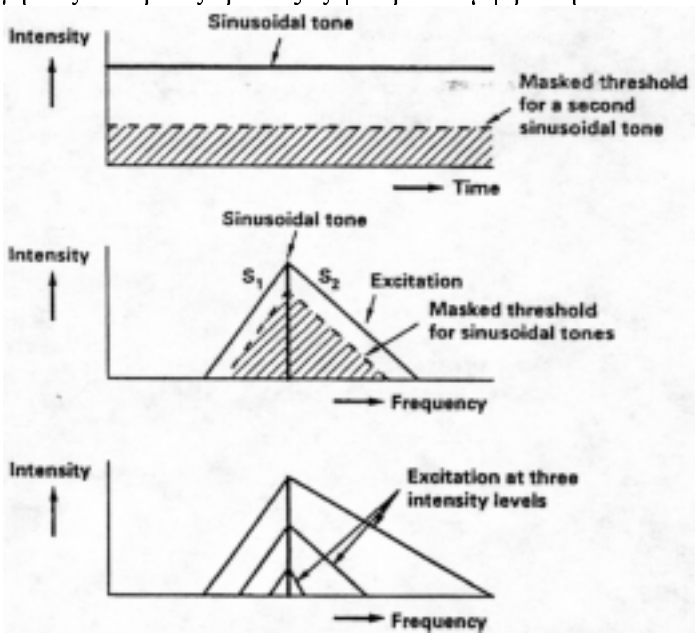
Πάνω στην σκέψη για το πως να υπολογίσουμε την εσωτερική αναπαράσταση ενός σήματος κάποιος μπορεί να ονειρεύεται μία μέθοδο στην οποία όλοι οι μετασχηματισμοί χαρακτηριστικών των ανεξαρτήτων τμημάτων του ανθρώπινου ακουστικού συστήματος θα μετρούνται και θα μοντελοποιούνται. Σε αυτή την ακριβή προσέγγιση κάποιος θα είχε την, σχεδόν απίθανη, δουλειά της μοντελοποίησης του αφτιού, του μηχανισμού μετατροπής ενέργειας και της νευρωνικής επεξεργασίας για έναν αριθμό διαφορετικών επιπέδων αφαίρεσης.

Η βιβλιογραφία παρέχει παραδείγματα της ακριβούς προσέγγισης [Kates, 1991b], [Yang et al., 1992], [Giguere and Woodland, 1994a], [Giguere and Woodland, 1994b] αλλά δεν έχουν εκδοθεί ακόμα αποτελέσματα από μεγάλα πειράματα υποκειμενικής αξιολόγησης ποιότητας. Τα πρώτα αποτελέσματα από την χρήση της ακριβούς προσέγγισης κατά την μέτρηση της ποιότητας των κωδικοποιητών—αποκωδικοποιητών ομιλίας έχουν ήδη εκδοθεί (πχ [Ghitza, 1994]) αλλά παρουσιάζουν μάλλον απογοητευτικά αποτελέσματα με την έννοια της συσχέτισης μεταξύ αντικειμενικών και υποκειμενικών μετρήσεων. Προφανώς είναι πολύ δύσκολο να υπολογιστεί η σωστή εσωτερική αναπαράσταση ήχου πάνω στην βάση του ποιος άνθρωπος κρίνει την ποιότητα του ήχου. Ακόμα περισσότερο μπορεί να μην είναι αρκετό να υπολογίσουμε απλώς τις διαφορές στις εσωτερικές αναπαραστάσεις, μιας και οι επιδράσεις της εκπαίδευσης μπορεί να συνεισφέρουν στην ποιότητα αντίληψης.

Κάποιος μπορεί να αμφιβάλλει για το αν είναι αναγκαίο να έχουμε ένα ακριβές μοντέλο των χαμηλότερων επιπέδων αφαίρεσης (abstraction level) του ακουστικού συστήματος (εξωτερικό, μέσο, εσωτερικό αφτί, μετατροπή ενέργειας). Επειδή οι μετρήσεις της ακουστικής ποιότητας είναι, στο τέλος, μία διαδικασία εκπαίδευσης μία γενική προσέγγιση της εσωτερικής αναπαράστασης ακολουθούμενη από μία γενική εκπαιδευόμενη διαδικασία επεξήγησης μπορεί να είναι περισσότερο κατάλληλη από το να έχουμε μία ακριβή εσωτερική αναπαράσταση χωρίς εκπαιδευόμενη διαδικασία επεξήγησης των διαφορών.

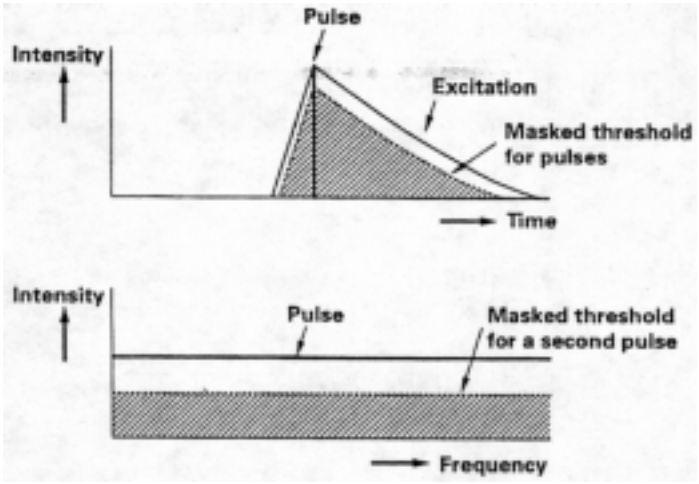
Για την εύρεση της κατάλληλης εσωτερικής αναπαράστασης μπορούν να χρησιμοποιηθούν τα αποτελέσματα των ψυχοακουστικών πειραμάτων στα οποία οι άνθρωποι κρίνουν συγκεκριμένα χαρακτηριστικά του ακουστικού σήματος με την έννοια των ψυχολογικών ποσοτήτων όπως ακουστότητα και ύψος ήχου. Αυτές οι ποσότητες ήδη εσωκλείουν ένα συγκεκριμένο επίπεδο υποκειμενικής επεξήγησης φυσικών ποσοτήτων όπως είναι η ένταση (intensity) και η συχνότητα. Αυτή η ψυχοακουστική προσέγγιση έχει οδηγήσει σε μία μεγάλη ποικιλία μοντέλων τα οποία μπορούν να προβλέψουν συγκεκριμένα χαρακτηριστικά του ήχου, πχ [Zwicker and Feldtkeller, 1967], [Zwicker, 1977], [Florentine and Buus, 1981] , [Martens, 1982], [Srulovicz and Goldstein, 1983], [Durlach et al., 1986], [Beerends, 1989], [Meddis and Hewitt, 1991]. Παρόλα αυτά εάν κάποιος θέλει να προβλέψει την υποκειμενικά αντιλαμβανόμενη ποιότητα μίας ακουστικής συσκευής ένα μεγάλο τμήμα των διαφορετικών πτυχών της αντίληψης του ήχου πρέπει να μοντελοποιηθεί. Τα πιο σημαντικά χαρακτηριστικά που πρέπει να μοντελοποιηθούν για την εσωτερική αναπαράσταση είναι η απόκρυψη, η ακουστότητα συνισταμένων χρόνου-συχνότητας που έχουν υποστεί μερική απόκρυψη και η ακουστότητα συνισταμένων χρόνου συχνότητας που δεν έχουν υποστεί απόκρυψη.

Για στατικούς ήχους (stationary) η εσωτερική αναπαράσταση περιγράφεται καλύτερα με την χρήση της φασματικής αναπαράστασης. Η εσωτερική αναπαράσταση μπορεί να μετρηθεί με την χρήση ενός δοκιμαστικού σήματος με μικρό εύρος ζώνης. Ένα σχηματικό παράδειγμα δίνεται στο σχήμα 2 για ένα μονό ημιτονοειδή τόνο (αποκρύπτης—masker) όπου το κατώφλι απόκρυψης (masked threshold) ενός τέτοιου τόνου μετριέται με ένα δεύτερο ημιτονοειδή δοκιμαστικό τόνο (στόχος—target). Το κατώφλι απόκρυψης μπορεί να εξηγηθεί σαν το αποτέλεσμα της εσωτερικής αναπαράστασης που παρουσιάζεται στο σχήμα 2 σαν ένα πρότυπο διέγερσης. Το σχήμα 2 επίσης αποτελεί μία ένδειξη του επιπέδου εξάρτησης του προτύπου διέγερσης ενός μονού ημιτονοειδούς τόνου. Αυτή η εξάρτηση επιπέδου κάνει τις επεξηγήσεις σε όρους τράπεζας φίλτρων αμφίβολη.



Σχήμα 2: Από το πρότυπο απόκρυψης μπορεί να παρατηρηθεί ότι η διέγερση που παράγεται από ένα ημιτονοειδή τόνο εξαπλώνεται στο πεδίο συχνοτήτων. Η κλίση του δεξιού τμήματος του προτύπου της διέγερσης βλέπουμε ότι μεταβάλλεται σαν συνάρτηση της έντασης του αποκρύπτη

Για μη στατικούς ήχους (non-stationary) η εσωτερική αναπαράσταση περιγράφεται καλύτερα με την χρήση μίας χρονικής αναπαράστασης (temporal representations). Η εσωτερική αναπαράσταση μπορεί να μετρηθεί με την χρήση ενός σήματος δοκιμής μικρής διάρκειας. Ένα σχηματικό παράδειγμα για ένα μονό click (αποκρύπτης—masker) δίνεται στο σχήμα 3 όπου το κατώφλι απόκρυψης (masked threshold) ενός τέτοιου click μετριέται με ένα δεύτερο click (στόχος—target). Το κατώφλι απόκρυψης μπορεί να εξηγηθεί σαν το αποτέλεσμα μίας εσωτερικής, εξαπλωμένης (smeared out), αναπαράστασης των παλμών (puls) (σχήμα 3, πρότυπο διέγερσης).



Σχήμα 3:Από το αποκρύπτον πρότυπο παρατηρείται ότι η διέγερση που παρατηρείται από ένα click εξαπλώνεται στο πεδίο συχνοτήτων.

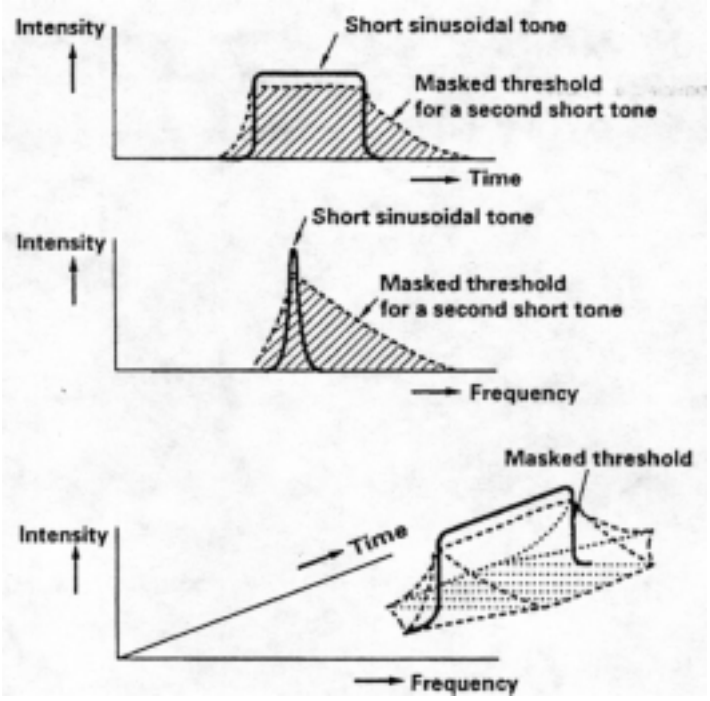
Ένα παράδειγμα συνδυασμού απόκρυψης πεδίου χρόνου και συχνοτήτων, χρησιμοποιώντας ένα απότομο τόνο (tone burst), δίνεται στο σχήμα 4.

Για τα παραδείγματα που δίνονται στα σχήματα 2-4 κάποιος μπορεί να συνειδητοποιήσει ότι το κατώφλι απόκρυψης καθορίζεται με ένα σήμα στόχου (target signal) το οποίο είναι ένα ακριβές αντίγραφο του αποκρύπτοντος σήματος (masker signal). Για σήματα στόχου τα οποία είναι διαφορετικά από το αποκρύπτον σήμα (πχ ένα ημίτονο το οποίο αποκρύπτει μία ζώνη θορύβου) το κατώφλι απόκρυψης μοιάζει διαφορετικό, καταστώντας αδύνατη την συζήτηση για το κατώφλι απόκρυψης ενός σήματος. Το κατώφλι απόκρυψης ενός σήματος εξαρτάται από τον στόχο σε αντίθεση με την εσωτερική αναπαράσταση και το πρότυπο διέγερσης τα οποία δεν εξαρτώνται από το στόχο.

Στα σχήματα 2-4κάποιος μπορεί να δει ότι οποιαδήποτε συνισταμένη χρόνου-συχνότητας του σήματος εξαπλώνεται τόσο στον άξονα των συχνοτήτων όσο και στον άξονα του χρόνου. Αυτή η εξάπλωση του σήματος έχει ως αποτέλεσμα περιορισμένη ανάλυση χρόνου-συχνότητας του ακουστικού συστήματος. Ακόμα περισσότερο είναι γνωστό ότι δύο εξαπλωμένες συνιστώσες χρόνου-συχνότητας στο πεδίο διέγερσης δεν αθροίζονται σε μία συνδυασμένη διέγερση πάνω στην βάση της πρόσθεσης ενέργειας. Για τον λόγο αυτό η εξάπλωση απαρτίζεται από δύο μέρη, ένα το οποίο περιγράφει πως η ενέργεια σε ένα σημείο του πεδίου χρόνου-συχνότητας έχει ως αποτέλεσμα την διέγερση σε ένα άλλο σημείο, και ένα τμήμα που περιγράφει πως οι διαφορετικές διεγέρσεις σε ένα συγκεκριμένο σημείο, προερχόμενες από την εξάπλωση των ανεξαρτήτων συνιστωσών χρόνου-συχνότητας, προστίθενται.

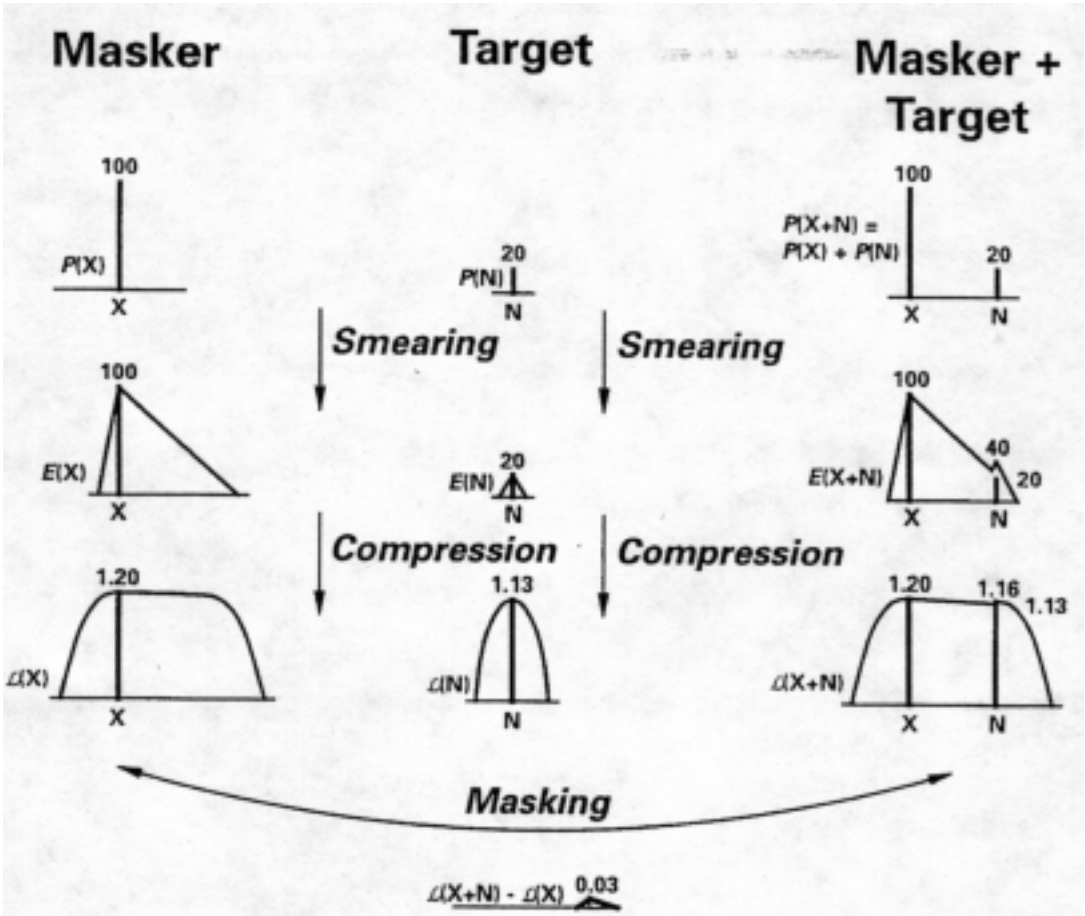
Μέχρι τώρα έχει περιγραφτεί μόνο εξάπλωση χρόνου-συχνότητας του ακουστικού σήματος από το αφτί, η οποία οδηγεί σε μία αναπαράσταση διέγερσης. Αυτή η αναπαράσταση διέγερσης συνήθως μετριέται σε dB SPL (Sound Pressure Level—Επίπεδο Πίεσης Ήχου) σαν μία συνάρτηση του χρόνου και της συχνότητας. Για την κλίμακα συχνοτήτων συνήθως δεν χρησιμοποιείται η γραμμική κλίμακα Hz αλλά η μη γραμμική κλίμακα Bark. Αυτή η κλίμακα Bark είναι μία κλίμακα ύψους ήχου που αναπαριστά το ψυχοφυσικό ισοδύναμο της συχνότητας. Αν και η εξάπλωση συσχετίζεται με μία σημαντική ιδιότητα του ανθρώπινου ακουστικού συστήματος, την απόκρυψη στο πεδίο χρόνου-συχνότητας, η αναπαράσταση που προκύπτει στην μορφή ενός προτύπου διέγερσης δεν είναι πολύ χρήσιμη ακόμα. Για να πάρουμε μία εσωτερική αναπαράσταση η οποία προσεγγίζει όσο το δυνατόν περισσότερο την

εσωτερική αναπαράσταση που χρησιμοποιείται από τους ανθρώπους για την αξιολόγηση ποιότητας πρέπει να συμπιεστεί η αναπαράσταση διέγερσης με ένα τρόπο που αντανακλά την συμπίεση που συναντάται στο εσωτερικό αφτί και στην νευρωνική επεξεργασία.



Σχήμα 4: Το πρότυπο διέγερσης για ένα μικρό εκρηκτικό τόνο. Η διέγερση που προκύπτει είναι εξαπλωμένη στα πεδία χρόνου και συχνότητας.

Η συμπίεση που χρησιμοποιείται για τον υπολογισμό της εσωτερικής αναπαράστασης απαρτίζεται από ένα κανόνα μετασχηματισμού από την πυκνότητα διέγερσης στην συμπιεσμένη πυκνότητα Sone όπως διατυπώθηκε από τον Zwicker [Zwicker and Feldtkeller, 1967]. Η εξάπλωση της ενέργειας είναι κυρίως το αποτέλεσμα των περιφερειακών διεργασιών [Viergever, 1986] ενώ η συμπίεση είναι μία πιο κεντρική διαδικασία [Pickles, 1988]. Με τις δύο απλές μαθηματικές πράξεις, την εξάπλωση και την συμπίεση, είναι δυνατό να μοντελοποιηθούν οι ιδιότητες της απόκρυψης του ακουστικού συστήματος όχι μόνο στο κατώφλι απόκρυψης, αλλά επίσης και στη μερική απόκρυψη [Scharf, 1964] πάνω από το κατώφλι απόκρυψης (βλέπε σχήμα 5).



Σχήμα 5: Παρουσίαση του τρόπου με τον οποίο μοντελοποιείται η απόκρυψη στο μοντέλο εσωτερικής αναπαράστασης. Η εξάπλωση και η συμπίεση με $L=E^{0.04}$ οδηγεί στην απόκρυψη. Η πρώτη αναπαράσταση (πάνω) είναι στην μορφή ισχύος και μπορεί να αναπαριστά clicks στο πεδίο του χρόνου ή ημίτονα στο πεδίο των συχνοτήτων. Με X συμβολίζουμε το σήμα, ή αποκρύπτη, και με N τον θόρυβο, ή στόχο. Η αριστερή μεριά δείχνει τους μετασχηματισμούς του αποκρύπτη, στα δεξιά παρουσιάζονται οι μετασχηματισμοί του σύνθετου σήματος (αποκρύπτης και στόχος). Η δεύτερη αναπαράσταση είναι σε όρους διέγερσης E και δείχνει την διέγερση σαν συνάρτηση του χρόνου και της συχνότητας. Η τρίτη αναπαράσταση είναι η εσωτερική αναπαράσταση

2.5 ΥΠΟΛΟΓΙΣΜΟΣ ΤΗΣ ΕΣΩΤΕΡΙΚΗΣ ΑΝΑΠΑΡΑΣΤΑΣΗΣ ΗΧΟΥ

Σαν αρχή στην ποσοτικοποίηση των δύο μαθηματικών πράξεων, της εξάπλωσης και της συμπίεσης, που χρησιμοποιούνται στο εσωτερικό μοντέλο αναπαράστασης μπορούν να χρησιμοποιηθούν τα αποτελέσματα των ψυχοακουστικών πειραμάτων στην απόκρυψη χρόνου-συχνότητας και στην αντίληψη της ακουστότητας. Η εξάπλωση της συχνότητας μπορεί να προέλθει από πειράματα απόκρυψης στο πεδίο συχνοτήτων όπου ένας μονός σταθερής κατάστασης στενής ζώνης αποκρύπτης και ένας μονός σταθερής κατάστασης στενής ζώνης στόχος χρησιμοποιούνται για την μέτρηση της κλίσης της συνάρτησης απόκρυψης [Scharf and Buus, 1986], [Moore, 1997]. Αυτές οι συναρτήσεις εξαρτώνται από το επίπεδο και τη συχνότητα του αποκρύπτοντος σήματος. Εάν ένα από τα σήματα είναι μία μικρή ζώνη θορύβου και το άλλο καθαρός τόνος τότε οι κλίσεις μπορούν να προσεγγιστούν από την εξίσωση (1) (βλέπε Terhardt 1979, [Terhardt, 1979]):

S_1=31dB/Bark, συχνότητα στόχου<συχνότητα αποκρύπτη

$$S_2=(22+\min(230/f, 10)-0.2L)\ dB/Bark, \quad \text{συχνότητα στόχου} > \text{συχνότητα αποκρύπτη} \tag{1}$$

όπου f είναι η συχνότητα απόκρυψης σε Hz και L το επίπεδο σε dB SPL. Ένα σχηματικό παράδειγμα αυτής της απόκρυψης στο πεδίο των συχνοτήτων φαίνεται στο σχήμα 2. Το κατώφλι απόκρυψης μπορεί να θεωρηθεί σαν το αποτέλεσμα της εξάπλωσης των σημάτων στενής ζώνης στο πεδίο των συχνοτήτων (βλέπε σχήμα 2). Οι κλίσεις όπως δίνονται από τις εξισώσεις (1) μπορούν να χρησιμοποιηθούν σαν μία προσέγγιση της εξάπλωσης της διέγερσης στο πεδίο των συχνοτήτων και στην περίπτωση αυτή το κατώφλι απόκρυψης μπορεί να θεωρηθεί σαν ένα τμήμα της διέγερσης.

Εάν περισσότεροι από ένας αποκρύπτες υπάρχουν την ίδια χρονική στιγμή το κατώφλι απόκρυψης ενέργειας του σύνθετου σήματος $M_{composite}$ δεν είναι απλώς το άθροισμα των n ανεξαρτήτων κατωφλίων απόκρυψης ενέργειας M_i, αλλά δίνεται προσεγγιστικά από την σχέση:

$$M_{composite} = \left(\sum_{i=1}^{n} M_i^a \right)^{1/a} \tag{2}$$

Αυτός ο κανόνας πρόσθεσης ισχύει για ταυτόχρονη (πεδίο συχνοτήτων) [Luffi, 1983], [Luffi, 1985] και για μη ταυτόχρονη (πεδίο χρόνου) [Penner, 1980], [Penner and Shiffrin, 1980] απόκρυψη [Humes and Jesteadt, 1989] αν και η τιμή της δύναμης συμπίεσης a μπορεί να είναι διαφορετική πάνω στους άξονες συχνότητας (a_{freq}) και χρόνου (a_{time}).

Στο ψυχοακουστικό μοντέλο το οποίο χρησιμοποιείται σε αυτό το κεφάλαιο κανένα κατώφλι απόκρυψης δεν υπολογίζεται ευθέως σε καμία μορφή. Η απόκρυψη μοντελοποιείται από ένα συνδυασμό εξάπλωσης και συμπίεσης όπως εξηγείται στο σχήμα 5. Άρα το μέγεθος της απόκρυψης εξαρτάται από τις παραμέτρους a_{freq} και a_{time} οι οποίες καθορίζουν μαζί με τις κλίσεις S_1 και S_2 το μέγεθος της εξάπλωσης.

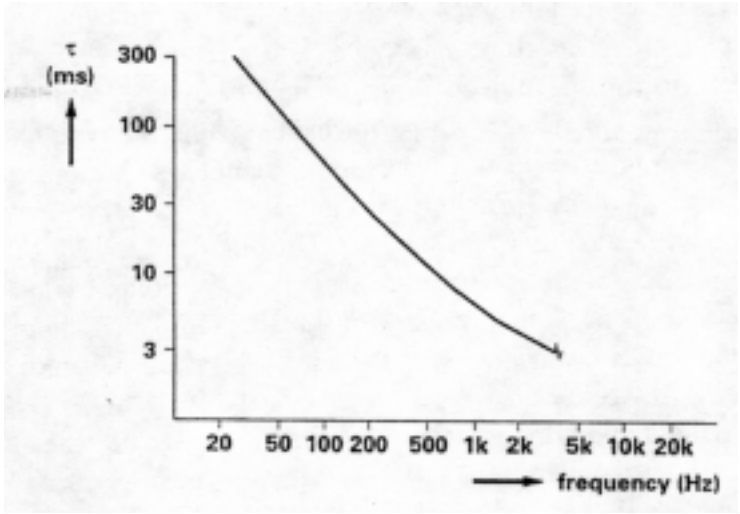
Όμως οι τιμές a_{freq} και a_{time} που βρίσκονται στην βιβλιογραφία βελτιστοποιούνται ως προς το κατώφλι απόκρυψης και έτσι δεν μπορούν να χρησιμοποιηθούν στο μοντέλο μας. Έτσι αυτά τα δύο a θα βελτιστοποιηθούν πριν από τις μετρήσεις ακουστικής ποιότητας.

Στο ψυχοακουστικό μοντέλο η φυσική αναπαράσταση χρόνου-συχνότητας υπολογίζεται με την χρήση ενός FFT με ένα κατά 50% επικαλύπτον παράθυρο Hanning ($\sin^2$) των 40 ms περίπου, το οποίο οδηγεί σε μία ανάλυση χρόνου της τάξης των 20 ms. Μέσα σε αυτό το παράθυρο οι συνιστώσες της συχνότητας εξαπλώνονται σύμφωνα με τις εξισώσεις (1) και οι διεγέρσεις προστίθενται σύμφωνα με την εξίσωση (2). Λόγω της περιορισμένης ανάλυσης χρόνου μπορεί να εφαρμοστεί μόνο μία πρόχειρη προσέγγιση της εξάπλωσης στο πεδίο συχνοτήτων.

Από τα δεδομένα απόκρυψης που βρίσκονται στην βιβλιογραφία [Jesteadt et al., 1982] έγινε μία εκτίμηση του πόση ενέργεια μένει μέσα σε ένα πλαίσιο από ένα προηγούμενο πλαίσιο χρησιμοποιώντας μία μετατόπιση μισού παραθύρου (50% επικάλυψη). Αυτό το ποσοστό μπορεί να εκφραστεί σαν μία σταθερή χρόνου τ στην έκφραση:

$$\Delta E(\Delta t) = e^{-\Delta t/\tau} \tag{3}$$

όπου Δt είναι η χρονική απόσταση μεταξύ δύο πλαισίων $=T_f$. Το ποσοστό της ενέργειας που εμφανίζεται στο επόμενο παράθυρο εξαρτάται από την συχνότητα και έτσι διαφορετικό τ χρησιμοποιείται για κάθε ζώνη συχνοτήτων. Αυτό το ποσοστό της ενέργειας εξαρτάται επίσης από το επίπεδο του αποκρύπτη [Jesteadt et al., 1982] αλλά αυτή η εξάρτηση του τ από το επίπεδο δεν οδήγησε σε κάποια βελτίωση στην συσχέτιση και έτσι παραλήφθηκε από το μοντέλο. Σε συχνότητες άνω των 2000 Hz η εξάπλωση ελέγχεται από νευρωνικές διαδικασίες και παραμένει σχεδόν η ίδια [Pickles, 1988]. Οι τιμές του τ δίνονται στο σχήμα 6 και δίνουν μία εκθετική προσέγγιση την απόκρυψης στο πεδίο του χρόνου χρησιμοποιώντας μετατοπίσεις στην γειτονιά των 20 ms.



Σχήμα 6: Η σταθερά χρόνου τ που χρησιμοποιείται στην εξάπλωση στο πεδίο του χρόνου, σαν συνάρτηση της συχνότητας. Αυτή η συνάρτηση ισχύει μόνο για μετατοπίσεις παραθύρων της τάξης των 20 ms και επιτρέπει μόνο μία γενική εκτίμηση της εξάπλωσης στο πεδίο χρόνου χρησιμοποιώντας ένα a_{time} με τιμή περίπου 0.6.

Ένα παράδειγμα της αποσύνθεσης (decomposition) ενός ημιτονοειδούς εκρηκτικού (burst) τόνου στο πεδίο χρόνου-συχνότητας δίνεται στο σχήμα 4. Πρέπει να γίνει αντιληπτό ότι αυτές οι σταθερές χρόνου τ δίνουν μόνο μία εκθετική προσέγγιση των συναρτήσεων απόκρυψης του πεδίου χρόνου, στην απόσταση του μισού μήκους παραθύρου.

Έχοντας εφαρμόσει την λειτουργία της εξάπλωσης χρόνου-συχνοτήτων λαμβάνεται μία αναπαράσταση μίας διέγερσης προτύπων του ακουστικού σήματος σε (dB_{exc}, seconds, Bark). Εν συνεχεία αυτή η αναπαράσταση μεταμορφώνεται σε μία εσωτερική αναπαράσταση χρησιμοποιώντας μία μη γραμμική συνάρτηση συμπίεσης. Η μορφή αυτής της συνάρτησης συμπίεσης μπορεί να εξαχθεί από πειράματα ακουστότητας.

Τα πειράματα κλίμακας που χρησιμοποιούν σήματα σταθερής κατάστασης έχουν δείξει ότι η ακουστότητα ενός σήματος είναι μη γραμμική συνάρτηση της έντασης. Εκτεταμένες μετρήσεις πάνω στην σχέση που υπάρχει μεταξύ της έντασης και της ακουστότητας έχουν οδηγήσει στον ορισμό του Sone. Ένα σταθερής κατάστασης ημιτονοειδές του 1 kHz σε ένα επίπεδο των 40 dB SPL έχει εξ' ορισμού

ακουστότητα ένα Sone. Η ακουστότητα άλλων θορύβων μπορεί να εκτιμηθεί σε ψυχοακουστικά πειράματα. Σε μία πρώτη προσέγγιση για τον υπολογισμό της εσωτερικής αναπαράστασης μπορεί να απεικονιστεί η φυσική αναπαράσταση που δίνεται σε dB/Bark σε μία αναπαράσταση σε Sone/Bark:

$$L = k\left(P - P_0\right)^{\gamma} \tag{4}$$

όπου το k είναι μία σταθερά κλίμακας (περίπου 0.01), P το επίπεδο του τόνου σε μPa, P_0 το απόλυτο όριο ακοής για τον τόνο σε μPa, και γ η παράμετρος συμπίεσης, η οποία στην βιβλιογραφία εκτιμάται ότι είναι περίπου 0.6 [Scharf and Houtsma, 1986]. Αυτή η συμπίεση συνδέει μία φυσική ποσότητα (ακουστική πίεση P) με μία ψυχοφυσική ποσότητα (ακουστότητα L).

Οι εξισώσεις (1), (2) και (4) περιέχουν ποσότητες που μπορούν να μετρηθούν απευθείας. Ύστερα από την εφαρμογή της εξίσωσης (1) σε οποιαδήποτε συνιστώσα χρόνου-συχνότητας και πρόσθεση όλων των ανεξαρτήτων συνεισφορών διέγερσης με την χρήση της εξίσωσης (2), το πρότυπο διέγερσης που προκύπτει μορφοποιεί την βάση της εσωτερικής αναπαράστασης. (Η ακριβής μέθοδος για τον υπολογισμό του προτύπου διέγερσης δίνεται στα παραρτήματα (appendix) A, B και C των [Beerends and Stemerdink, 1992] ενώ ένας συμπαγής αλγόριθμος δίνεται στο παράρτημα D των [Beerends and Stemerdink, 1992]).

Επειδή η εξίσωση (4) απεικονίζει το φυσικό πεδίο απευθείας σε ένα εσωτερικό πεδίο πρέπει να αντικατασταθεί από μία απεικόνιση από την διέγερση στην εσωτερική αναπαράσταση. Ο Zwicker έδωσε μία τέτοια απεικόνιση (εξίσωση 52,17 στο [Zwicker and Feldtkeller, 1967]):

$$L = k\left(\frac{E_0}{s}\right)^{\gamma}\left[\left(1-s+s\frac{E}{E_0}\right)^{\gamma}-1\right] \tag{5}$$

στην οποία το k είναι μία αυθαίρετη σταθερά κλίμακας, το E είναι το επίπεδο διέγερσης του τόνου, το E_0 είναι η διέγερση στο απόλυτο όριο ακοής για τον τόνο, το s είναι ο παράγοντας “schwel” όπως καθορίστηκε από τον Zwicker [Zwicker and Feldtkeller, 1967] και το γ είναι μία παράμετρος συμπίεσης η οποία προσαρμόστηκε στα δεδομένα ακουστότητας. Ο Zwicker βρήκε μία ιδανική τιμή για το γ γύρω στο 0.23.

Αν και το γ του 0.23 μπορεί να είναι βέλτιστο ως προς την κλίμακα ακουστότητας δεν θα είναι κατάλληλο για το υποκειμενικό μοντέλο ποιότητας το οποίο χρειάζεται μία εσωτερική αναπαράσταση η οποία είναι όσο το δυνατό πλησιέστερη στην αναπαράσταση που χρησιμοποιείται από τους ανθρώπους και πάνω στην οποία βασίζεται η κρίση τους για την ποιότητα. Έτσι το γ λαμβάνεται σαν μία παράμετρος η οποία μπορεί να προσαρμοστεί στην συμπεριφορά απόκρυψης των ανθρώπων μέσα στο πλαίσιο των μετρήσεων ακουστικής ποιότητας. Η σταθερά κλίμακας k δεν έχει καμία επίδραση στην απόδοση του μοντέλου. Η παράμετρος γ προσαρμόστηκε στην ISO/MPEG 1990 (International Standards Organization/Motion Picture Expert Group) βάση δεδομένων [ISO90, 1990] με την έννοια της μέγιστης συσχέτισης (ελάχιστης απόκλισης) μεταξύ υποκειμενικών και αντικειμενικών αποτελεσμάτων.

Η σύνθετη λειτουργία, της εξάπλωσης που ακολουθείται από συμπίεση, έχει ως αποτέλεσμα την μερική απόκρυψη (βλέπε σχήμα 5). Το πλεονέκτημα αυτής της

μεθόδου είναι το ότι το μοντέλο δίνει αυτόματα μία πρόβλεψη της συμπεριφοράς του ακουστικού συστήματος όταν οι αλλοιώσεις είναι πάνω από το κατώφλι απόκρυψης.

Ανακεφαλαιώνοντας το μοντέλο θα χρησιμοποιεί τους ακόλουθους μετασχηματισμούς (βλέπε σχήμα 7):

- Το σήμα εισόδου $x(t)$ και το σήμα εξόδου $y(t)$ μετασχηματίζονται στο πεδίο συχνοτήτων, χρησιμοποιώντας έναν μετασχηματισμό FFT και ένα παράθυρο Hanning ($\sin^2$) $\omega(t)$ των 40 ms περίπου. Αυτό οδηγεί στις φυσικές αναπαραστάσεις σήματος $P_x(t,f)$ και $P_y(t,f)$ σε (dB, seconds, Hz) με μία ανάλυση χρόνου-συχνότητας που είναι αρκετά καλή σαν σημείο έναρξης για την εξάπλωση χρόνου –συχνότητας.

- Η κλίμακα συχνότητας f (σε Hz) μετατρέπεται σε μία κλίμακα ύψους ήχου z (σε Bark) και το σήμα φιλτράρεται με την συνάρτηση μεταφοράς $\alpha_0(z)$ από το εξωτερικό στο εσωτερικό αφτί (ελεύθερο ή διασταλμένο πεδίο). Αυτό έχει ως αποτέλεσμα τις αναπαραστάσεις ισχύος-χρόνου-ύψους ήχου $p_x(t,z)$ και $p_y(t,z)$ που μετριούνται σε (dB, seconds, Bark). Μία πιο λεπτομερής περιγραφή αυτών των μετασχηματισμών δίνεται στο παράρτημα Α του [Beerends and Stemerdink, 1992].

- Οι αναπαραστάσεις ισχύος-χρόνου-ύψους ήχου $p_x(t,z)$ και $p_y(t,z)$ πολλαπλασιάζονται με ένα εξαρτώμενο από την συχνότητα κλάσμα $e^{-T_f/\tau(z)}$ χρησιμοποιώντας την εξίσωση (3) και το σχήμα 6, για την πρόσθεση με το α_{time} μέσα στο επόμενο πλαίσιο (T_f = χρονική μετατόπιση μεταξύ δύο πλαισίων ≈ 20 ms). Αυτό μοντελοποιεί την εξάπλωση πεδίου χρόνου των $x(t)$ και $y(t)$.

- Υπολογίζεται ο συγκερασμός των αναπαραστάσεων ισχύος-χρόνου-ύψους ήχου $p_x(t,z)$ και $p_y(t,z)$ με την συνάρτηση εξάπλωσης συχνότητας Λ, η οποία μπορεί να υπολογιστεί από την εξίσωση (1), ο οποίος οδηγεί σε αναπαραστάσεις διέγερσης-χρόνου-ύψους ήχου (dB_{exc}, seconds, Bark) $E_x(t,z)$ και $E_y(t,z)$ (βλέπε παραρτήματα B, C και D του [Beerends and Stemerdink, 1992]). Η μορφή της συνάρτησης εξάπλωσης συχνότητας εξαρτάται από την ένταση και την συχνότητα, και ο συγκερασμός πραγματοποιείται με ένα μη γραμμικό τρόπο χρησιμοποιώντας την εξίσωση (2) με την παράμετρο α_{freq} (βλέπε παράρτημα Α του [Beerends and Stemerdink, 1992]).

- Οι αναπαραστάσεις διέγερσης-χρόνου-ύψους ήχου $E_x(t,z)$ και $E_y(t,z)$ (dB_{exc}, seconds, Bark) μετατρέπονται σε συμπιεσμένες αναπαραστάσεις ακουστότητας-χρόνου-ύψους ήχου $L_x(t,z)$ και $L_y(t,z)$ (compressed Sone, seconds, Bark) με την χρήση της εξίσωσης (5) με παράμετρο γ (βλέπε παράρτημα Ε του [Beerends and Stemerdink, 1992]).

- Η συμπιεσμένη αναπαράσταση ακουστότητας-χρόνου-ύψους ήχου $L_y(t,z)$ της εξόδου της ακουστικής συσκευής διαβαθμίζεται ανεξάρτητα σε τρία διαφορετικά εύρη ύψους ήχου με φραγμούς στα 2 και 22 Bark. Αυτή η λειτουργία πραγματοποιεί ένα οικουμενικό ταίριασμα προτύπου ανάμεσα στις αναπαραστάσεις εισόδου και εξόδου και ήδη μοντελοποιεί κάποια από τα υψηλότερα, εκπαιδευόμενα, επίπεδα της επεξεργασίας ήχου.

Στην βιβλιογραφία της ψυχοακουστικής αρκετά πειράματα πάνω στην συμπεριφορά της απόκρυψης μπορούν να βρεθούν για τα οποία το μοντέλο εσωτερικής αναπαράστασης πρέπει, στην πράξη, να είναι ικανό να προβλέπει την

συμπεριφορά των ανθρώπων. Μία από αυτές τις επιδράσεις είναι η όξυνση του προτύπου διέγερσης ύστερα από την παύση κάποιου ακουστικού ερεθίσματος [Houtgast, 1977], το οποίο έμμεσα μοντελοποιείται μερικώς εδώ, στην μορφή της εξάρτησης της κλίσης S2 στην εξίσωση (1.1) από την ένταση. Μετά από την παύση του αποκρύπτη η αναπαράσταση στο επόμενο πλαίσιο του μοντέλου είναι μία "οξυμένη έκδοση του προηγούμενου πλαισίου".

Μία άλλη σημαντική επίδραση είναι η ασυμμετρία της απόκρυψης μεταξύ ενός τόνου που αποκρύπτει μία ζώνη και μίας ζώνης θορύβου που αποκρύπτει ένα τόνο [Hellman, 1972]. Στα μοντέλα που χρησιμοποιούν το κατώφλι απόκρυψης αυτή η επίδραση πρέπει να μοντελοποιηθεί άμεσα κάνοντας το όριο εξαρτημένο από τον τύπο του αποκρύπτη πχ με τον υπολογισμό ενός δείκτη τονικότητας όπως γίνεται στα ψυχοακουστικά μοντέλα που χρησιμοποιούνται στο ISO/MPEG standard ακουστικής κωδικοποίησης [ISO92st, 1993]. Στην προσέγγιση εσωτερικής αναπαράστασης αυτή η επίδραση υπολογίζεται από την μη γραμμική πρόσθεση των ανεξαρτήτων συνιστωσών χρόνου συχνότητας στο πεδίο διέγερσης.

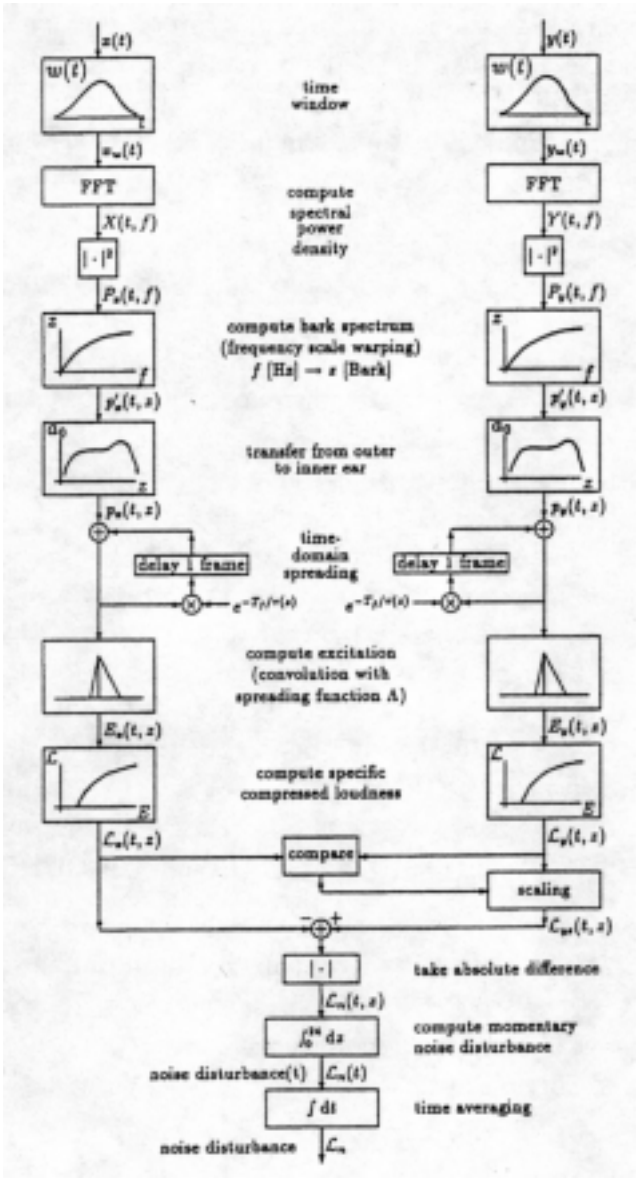
2.6 ΤΟ ΜΕΤΡΟ ΑΝΤΙΛΗΠΤΗΣ ΑΚΟΥΣΤΙΚΗΣ ΠΟΙΟΤΗΤΑΣ (PAQM)

Ύστερα από τον υπολογισμό των εσωτερικών αναπαραστάσεων ακουστότητας-χρόνου-ύψους ήχου της εισόδου και της εξόδου της ακουστικής συσκευής η αντιλαμβανόμενη ποιότητα του σήματος εξόδου μπορεί να εξαχθεί από την διαφορά μεταξύ των εσωτερικών αναπαραστάσεων. Οι συναρτήσεις πυκνότητας $L_x(t,z)$ (η πυκνότητα ακουστότητας σαν συνάρτηση του χρόνου και του ύψους ήχου για την είσοδο x) και οι κλιμακωτές $L_y(t,z)$ αφαιρούνται για να προκύψει μία συνάρτηση πυκνότητας διασποράς θορύβου (noise disturbance) $L_n(t,z)$. Αυτό το $L_n(t,z)$ ολοκληρώνεται πάνω από τις συχνότητες γεγονός που έχει σαν αποτέλεσμα μία στιγμιαία διασπορά θορύβου $L_n(t)$ (βλέπε σχήμα 7).

Από τον μέσο όρο της στιγμιαίας διακύμανσης θορύβου υπολογίζεται η διακύμανση του θορύβου L_n . Δεν θα χρησιμοποιήσουμε τον όρο ακουστότητα θορύβου γιατί η τιμή του γ λαμβάνεται να είναι τέτοια ώστε το μοντέλο υποκειμενικής ποιότητας να είναι βέλτιστο. Σε αυτή την περίπτωση το L_n δεν αναπαριστάνει απαραίτητα ακουστότητα θορύβου. Ο λογάριθμος ($\log_{10}$) της διασποράς θορύβου ορίζεται σαν μέτρο αντιληπτής ακουστικής ποιότητας (PAQM).

Η βελτιστοποίηση των α_{freq}, a_{time} και γ πραγματοποιείται με την χρήση της βάσης δεδομένων υποκειμενικής ακουστικής ποιότητας που προέκυψε από την ISO/MPEG 1990 δοκιμή ακουστικών κωδικοποιητών-αποκωδικοποιητών [ISO90, 1990]. Η βελτιστοποίηση χρησιμοποίησε το τυπικό σφάλμα (standard error) της εκτιμημένης MOS από μία τρίτης τάξης γραμμή αναδρομής προσαρμοσμένης μέσα από τα PAQM, MOS σημεία δεδομένων. Η βελτιωποίηση επιτεύχθηκε με την ελαχιστοποίηση του τυπικού σφάλματος της εκτιμημένης MOS σαν μία συνάρτηση των α_{freq}, a_{time} και γ.

Οι ιδανικές τιμές των παραμέτρων α_{freq} και a_{time} εξαρτώνται από την δειγματοληψία του πεδίου χρόνου –συχνότητας. Για τις τιμές που χρησιμοποιούνται στην εφαρμογή μας, όπου Δz=0.2 Bark και Δt=20 ms (το συνολικό μήκος του παραθύρου είναι περίπου 40 ms), οι ιδανικές τιμές των παραμέτρων του μοντέλου έχει βρεθεί ότι είναι α_{freq}=0.8, a_{time}=0.6 και γ=0.04. Η εξάρτηση της συσχέτισης από την παράμετρο απόκρυψης πεδίου χρόνου a_{time} αποδείχθηκε μικρή.



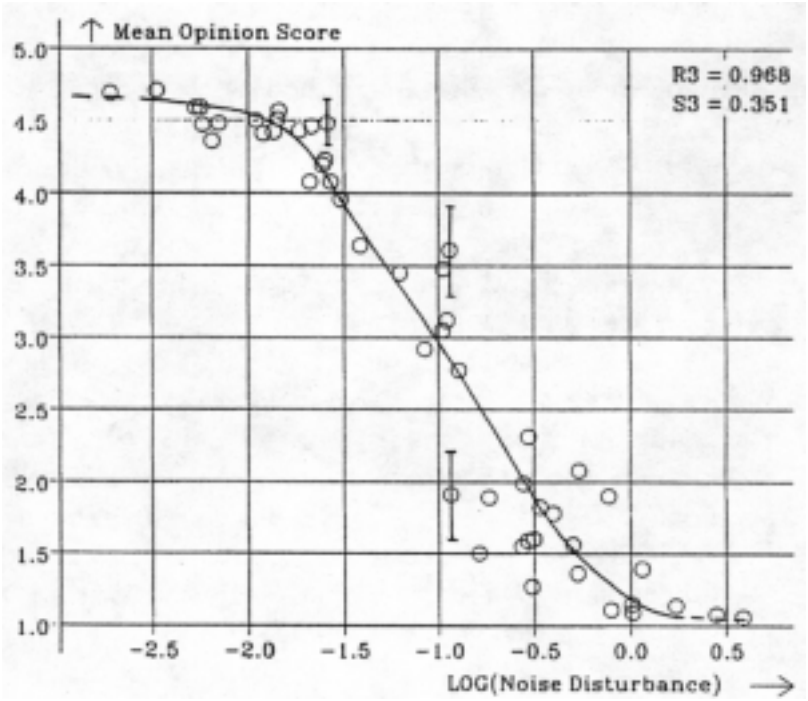
Σχήμα 7: Γενική παρουσίαση των βασικών μετασχηματισμών που χρησιμοποιούνται στην ανάπτυξη του PAQM. Τα σήματα x(t) και y(t) σαρώνονται από ένα παράθυρο w(t) και εν συνεχεία μετασχηματίζονται στο πεδίο των συχνοτήτων. Το φάσμα ισχύος σαν συνάρτηση του χρόνου και της συχνότητας $P_x(t,f)$ και $P_v(t,f)$ μετασχηματίζεται σε φάσμα ισχύος σαν συνάρτηση χρόνου και ύψους ήχου, $p_x(t,z)$ και $p_y(t,z)$ τα οποία συνελίσσονται με την συνάρτηση εξάπλωσης και δίνουν τις διεγέρσεις $E_x(t,z)$ και $E_y(t,z)$. μετά τον μετασχηματισμό με την συνάρτηση συμπίεσης λαμβάνονται οι εσωτερικές αναπαραστάσεις $L_x(t,z)$ και $L_y(t,z)$ από τις οποίες προκύπτει η μέση διακύμανση θορύβου στο ακουστικό κομμάτι.

Λόγω του μικρού γ το οποίο βρέθηκε κατά την βελτιστοποίηση η πυκνότητα που προκύπτει σαν συνάρτηση του ύψους του ήχου (σε Bark) και του χρόνου δεν αναπαριστά την πυκνότητα ακουστότητας αλλά μία συμπιεσμένη πυκνότητα ακουστότητας. Η ολοκληρωμένη διαφορά μεταξύ των συναρτήσεων πυκνότητας της εισόδου και της εξόδου συνεπώς δεν αναπαριστά την ακουστότητα του θορύβου αλλά την συμπιεσμένη ακουστότητα του θορύβου.

Η σχέση ανάμεσα στο αντικειμενικό (PAQM) και το υποκειμενικό μέτρο ποιότητας (MOS) στις ιδανικές ρυθμίσεις των α_{freq}, α_{time} και γ για την ISO/MPEG 1990 βάση δεδομένων [ISO90, 1990] δίνεται στο σχήμα 8.

Η εσωτερική αναπαράσταση κάθε ακουστικού σήματος μπορεί τώρα να υπολογιστεί με την χρήση των μετασχηματισμών που δίνονται στην προηγούμενη

ενότητα. Η ποιότητα μίας ακουστικής συσκευής μπορεί έτσι να μετρηθεί με δοκιμαστικά σήματα (ημιτονοειδή, sweeps, θόρυβο κτλ) όπως και με πραγματικά σήματα (ομιλία, μουσική). Έτσι η μέθοδος αυτή είναι παγκοσμίως εφαρμόσιμη. Γενικά οι ακουστικές συσκευές ελέγχονται για διαπερατότητα (δηλαδή η έξοδος πρέπει να πλησιάζει την είσοδο όσο το δυνατόν περισσότερο) και στην περίπτωση αυτή η είσοδος και η έξοδος απεικονίζονται και οι δύο στις εσωτερικές τους αναπαραστάσεις και η ποιότητα της ακουστικής συσκευής καθορίζεται από την διαφορά ανάμεσα σε αυτές τις εσωτερικές αναπαραστάσεις εισόδου (η αναφοράς) και εξόδου.



Σχήμα 8: Σχέση μεταξύ του MOS και του PAQM για τις 50 καταχωρίσεις της ISO/MPEG 1990 δοκιμής κωδικοποιητών-αποκωδικοποιητών[ISO90, 1990] στην παρουσίαση με loudspeaker.

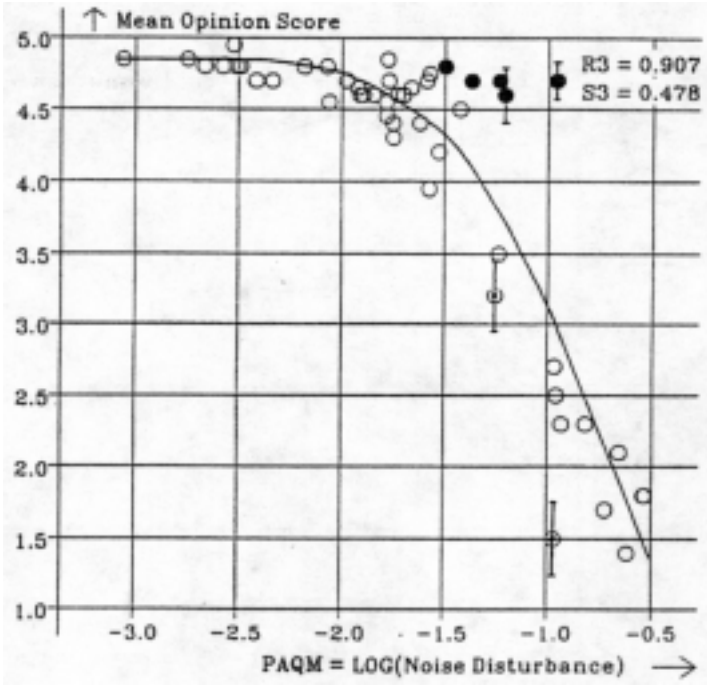
2.7 ΕΦΑΡΜΟΓΗ ΤΟΥ PAQM ΣΤΙΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΚΩΔΙΚΟΠΟΙΗΤΩΝ-ΑΠΟΚΩΔΙΚΟΠΟΙΗΤΩΝ ΟΜΙΛΙΑΣ ΚΑΙ ΜΟΥΣΙΚΗΣ

Η βελτιστοποίηση του PAQM που περιγράφεται στην προηγούμενη ενότητα οδηγεί σε ένα PAQM το οποίο παρουσιάζει μία καλή συσχέτιση ανάμεσα στα αντικειμενικά και στα υποκειμενικά αποτελέσματα. Σε αυτή την ενότητα το PAQM τίθεται σε εφαρμογή με την χρήση των αποτελεσμάτων της δεύτερης ISO/MPEG δοκιμής ακουστικών κωδικοποιητών-αποκωδικοποιητών (ISO/MPEG 1991 [ISO91, 1991]) και των αποτελεσμάτων της ITU-R TG10/2 1993 [ITURsg10cond9343,1993] δοκιμής ακουστικών κωδικοποιητών-αποκωδικοποιητών. Σε αυτή την τελευταία δοκιμή αξιολογείται υποκειμενικά πλήθος διαδοχικών συνθηκών του ISO/MPEG Layer II και III ενώ τρία διαφορετικά μοντέλα αντικειμενικής αξιολόγησης παρουσιάζουν αντικειμενικά αποτελέσματα.

Αυτή η ενότητα δίνει επίσης μία εφαρμογή του PAQM σε βάσεις δεδομένων η οποία προήλθε από τις αξιολογήσεις κωδικοποιητών-αποκωδικοποιητών ομιλίας τηλεφωνικού εύρους ζώνης (300-3400 Hz).

Το αποτέλεσμα της εφαρμογής με την χρήση της ISO/MPEG 1991 βάσης δεδομένων δίνεται στο σχήμα 9. Μία καλή συσχέτιση (R3=0.91) και ένα λογικό

χαμηλό τυπικό σφάλμα της εκτίμησης (S3=0.48) μεταξύ των αντικειμενικών PAQM και των υποκειμενικών MOS τιμών βρέθηκε.



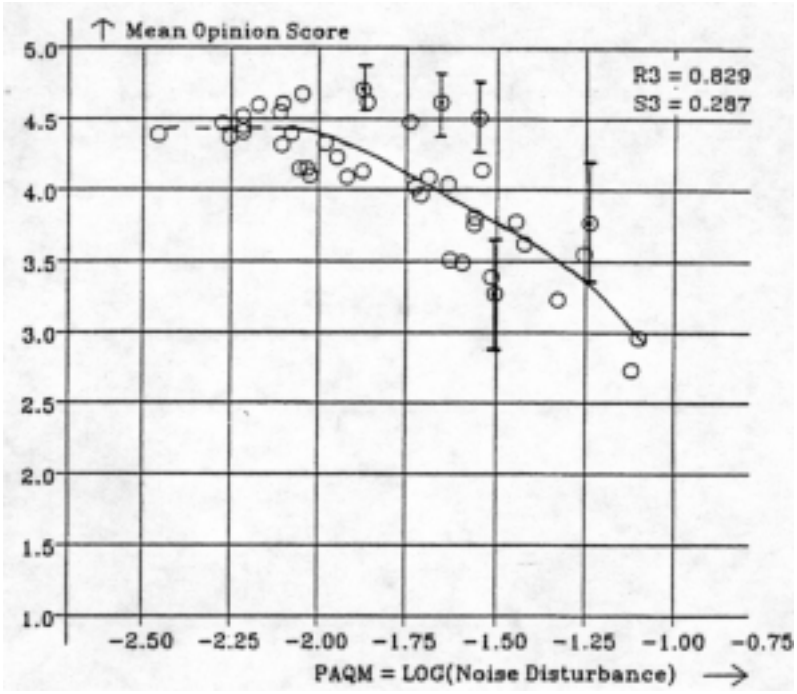
Σχήμα 9: Σχέση μεταξύ του MOS και του PAQM για τις 50 καταχωρίσεις της ISO/MPEG 1991 δοκιμής κωδικοποιητών-αποκωδικοποιητών[ISO91, 1991] στην παρουσίαση με loudspeaker. Οι μαυρισμένοι κύκλοι είναι καταχωρίσεις των οποίων η ποιότητα κρίθηκε πολύ χαμηλότερα από το μοντέλο από ότι από τους ερωτηθέντες.

Ένα ενδιαφέρον σημείο είναι το ότι για τις ίδιες PAQM τιμές μερικές φορές βρίσκονται μεγάλες αποκλίσεις στα υποκειμενικά αποτελέσματα.(βλέπε σχήμα 1.9).

Το αποτέλεσμα της εφαρμογής χρησιμοποιώντας την ITU-R TG10/2 1993 βάση δεδομένων (για την δοκιμασία Συνεισφοράς Διασποράς Εκπομπής) δίνεται στο σχήμα 10 και δείχνει καλή συσχέτιση και χαμηλό τυπικό σφάλμα εκτίμησης (R3=0.83 και S3=0.29) ανάμεσα στο αντικειμενικό PAQM και στο υποκειμενικό MOS. Αυτά τα αποτελέσματα εξακριβώθηκαν από το Swedish Broadcasting Corporation [ITURsg10cond9351, 1993] χρησιμοποιώντας ένα αντίγραφο λογισμικού το οποίο παραδόθηκε πριν η δοκιμή ITU-R TG10/2 πραγματοποιηθεί.

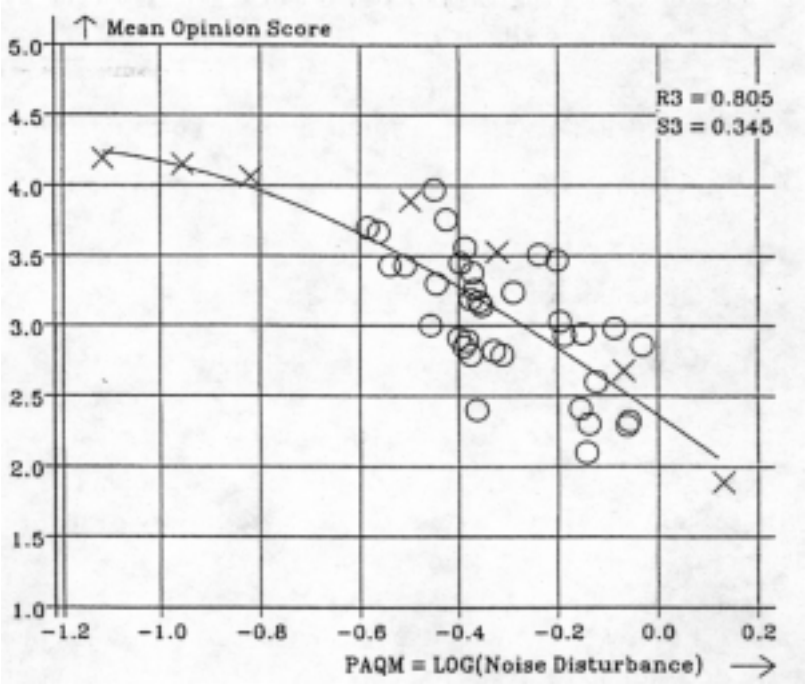
Και οι δύο εφαρμογές που πραγματοποιήθηκαν χρησιμοποιούν βάσεις δεδομένων στις οποίες αξιολογήθηκε η υποκειμενική ποιότητα των σημάτων εξόδου των μουσικών κωδικοποιητών-αποκωδικοποιητών. Αν το PAQM είναι όντως ένα παγκόσμιο μέτρο ακουστικής ποιότητας τότε πρέπει να είναι εφαρμόσιμο και στην αξιολόγηση κωδικοποιητών-αποκωδικοποιητών ομιλίας. Αν και οι κωδικοποιητές-αποκωδικοποιητές ομιλίας χρησιμοποιούν μία διαφορετική προσέγγιση για την μείωση των δεδομένων του ακουστικού bitstream απ' ότι οι μουσικοί κωδικοποιητές-αποκωδικοποιητές η κρίση της ποιότητας και των δύο πραγματοποιείται πάντα με το ίδιο ακουστικό σύστημα. Συνεπώς μία παγκόσμια αντικειμενική προσέγγιση της αντίληψης για την μέτρηση της ποιότητας των κωδικοποιητών-αποκωδικοποιητών ομιλίας και μουσικής πρέπει να γίνει πραγματοποιήσιμη. Στην βιβλιογραφία βρίσκεται μεγάλη ποσότητα πληροφοριών για το πως είναι δυνατό να μετρηθεί η ποιότητα των κωδικοποιητών-αποκωδικοποιητών ομιλίας (πχ [Gray and Markel, 1976], [Schroeder et al., 1979], [Gray et al., 1980], [Nocerino et al., 1985], [Quackenbush et al., 1988], [Hayashi and Kitawaki, 1992], [Halka and Heute, 1992], [Wang et al., 1992], [Ghitza, 1994], [Beerends and Stemerdink, 1994b]), αλλά καμία

από τις μεθόδους δεν μπορεί να εφαρμοστεί και στους δύο στους στενής ζώνης κωδικοποιητές-αποκωδικοποιητές ομιλίας και στους ευρείας ζώνης μουσικούς κωδικοποιητές-αποκωδικοποιητές.



Σχήμα 10: Σχέση μεταξύ του MOS και του PAQM για τις 43 ISO layer II συνθήκες της ITU-R TG10/2 1993 δοκιμής ακουστικών κωδικοποιητών-αποκωδικοποιητών [ITURsg10cond9343, 1993]

Για να ελέγξουμε εάν το PAQM μπορεί να εφαρμοστεί για την αξιολόγηση των κωδικοποιητών-αποκωδικοποιητών ομιλίας εγκαταστήθηκε μία εφαρμογή που χρησιμοποιεί αποτελέσματα υποκειμενικών δοκιμών στους ETSI GSM (European Telecommunications Standards Institute, Global System for Mobile communications) υποψήφιους κωδικοποιητές-αποκωδικοποιητές ομιλίας. Και οι δύο η GSM πλήρους ρυθμού (13 kbit/s, [Natvig, 1988]) και μισού ρυθμού (6 kbit/s, [ETSI91tm74, 1991]) αξιολογήσεις κωδικοποιητών-αποκωδικοποιητών ομιλίας έχουν χρησιμοποιηθεί στη εφαρμογή. Σε αυτά τα πειράματα τα σήματα ομιλίας κρίθηκαν σε ποιότητα με την χρήση μίας κλασσικής τηλεφωνικής συσκευής [CCITTrecP48, 1989]. Συνεπώς κατά την εφαρμογή της PAQM τόσο το σήμα ομιλίας εισόδου αναφοράς και το διαβαθμισμένο σήμα ομιλίας εξόδου φιλτραρίστηκαν με την χρήση της χαρακτηριστικής του φίλτρου του κλασσικού τηλεφώνου [CCITTrecP48, 1989]. Ακόμα οι αξιολογήσεις της ποιότητας ομιλίας πραγματοποιήθηκαν σε ένα ελεγχόμενο περιβάλλον θορύβου χρησιμοποιώντας θόρυβο Hoth σαν έναν αποκρύπτοντα θόρυβο υποβάθρου. Στην PAQM εφαρμογή αυτός ο θόρυβος μοντελοποιήθηκε με την πρόσθεση του σωστού φασματικού επιπέδου θορύβου Hoth [CCITTsup13, 1989] στις αναπαραστάσεις ισχύος-χρόνου-ύψους ήχου των σημάτων ομιλίας εισόδου και εξόδου.



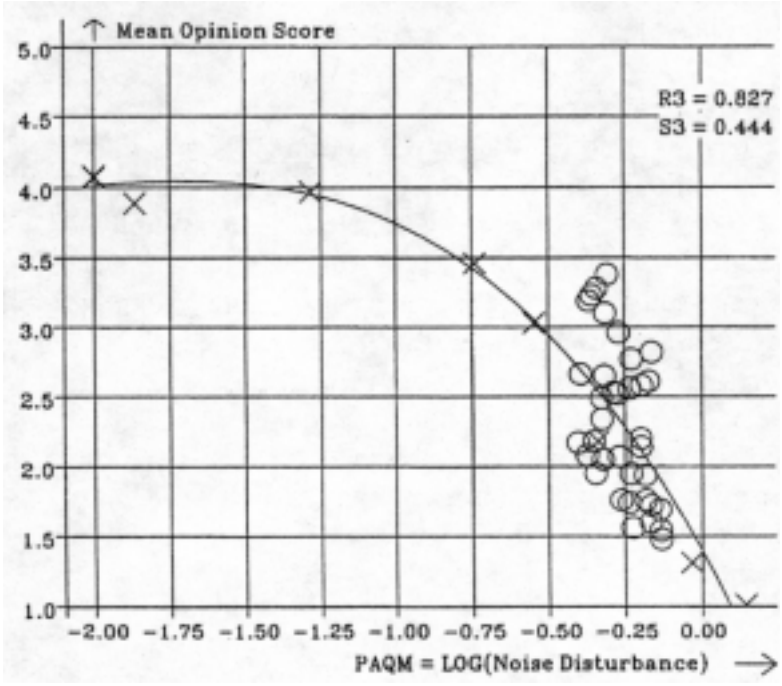
Σχήμα 11: Σχέση μεταξύ του MOS και του PAQM για την ETSI GSM πλήρους ρυθμού βάση δεδομένων ομιλίας. Οι σταυροί αναπαριστούν δεδομένα από το πείραμα που βασίζεται στην διαμορφωμένη μονάδα θορύβου αναφοράς, και οι κύκλοι αναπαριστούν δεδομένα από τους κωδικοποιητές-αποκωδικοποιητές ομιλίας.

Τα αποτελέσματα της εφαρμογής στους κωδικοποιητές-αποκωδικοποιητές ομιλίας δίνονται στα σχήματα 11 και 12. μία εμφανής διαφορά ανάμεσα σε αυτή την εφαρμογή και σε εκείνη που πραγματοποιήθηκε με την χρήση μουσικών κωδικοποιητών-αποκωδικοποιητών είναι η διασπορά των τιμών PAQM. Στους μουσικούς οι τιμές PAQM είναι όλες κάτω του –0.5 (βλέπε σχήματα 9, 10) ενώ για τους ομιλίας είναι κυρίως άνω των –0.5 (βλέπε εικόνες 11, 12). Προφανώς οι διασπορές σε αυτές τις βάσεις δεδομένων είναι σημαντικά μεγαλύτερες από αυτές στις βάσεις δεδομένων μουσικής. Ακόμα περισσότερο η συσχέτιση μεταξύ των υποκειμενικών και αντικειμενικών αποτελεσμάτων αυτής της εφαρμογής είναι χειρότερη από αυτήν της εφαρμογής που χρησιμοποιεί μουσικούς κωδικοποιητές-αποκωδικοποιητές

2.8 ΕΚΠΑΙΔΕΥΟΜΕΝΕΣ ΕΠΙΔΡΑΣΕΙΣ ΣΤΗΝ ΚΡΙΣΗ ΤΗΣ ΑΚΟΥΣΤΙΚΗΣ ΠΟΙΟΤΗΤΑΣ

Αν και τα αποτελέσματα της εφαρμογής του PAQM στις βάσεις δεδομένων κωδικοποιητών-αποκωδικοποιητών μουσικής και ομιλίας έδειξαν μία αρκετά καλή συσχέτιση ανάμεσα στα αντικειμενικά και στα υποκειμενικά αποτελέσματα, χρειάζονται ακόμα βελτιώσεις. Η αξιοπιστία των MOS προβλέψεων δεν είναι αρκετά καλή για την επιλογή των κωδικοποιητών-αποκωδικοποιητών μουσικής ή ομιλίας με την υψηλότερη ακουστική ποιότητα. Όπως δηλώθηκε στην ενότητα με τις ψυχοακουστικές αρχές της μεθόδου, μπορεί να είναι καλύτερο να έχουμε γενικά (crude) μοντέλα αντίληψης συνδυασμένα με γενικές εκπαιδευόμενες επεξηγήσεις (interpretation) από το να έχουμε ένα ακριβές μοντέλο αντίληψης. Συνεπώς η μέγιστη βελτίωση αναμένεται να έρθει από μία καλύτερη μοντελοποίηση των επιδράσεων εκμάθησης. Κατά την PAQM προσέγγιση όπως έχει παρουσιαστεί μέχρι στιγμής, η

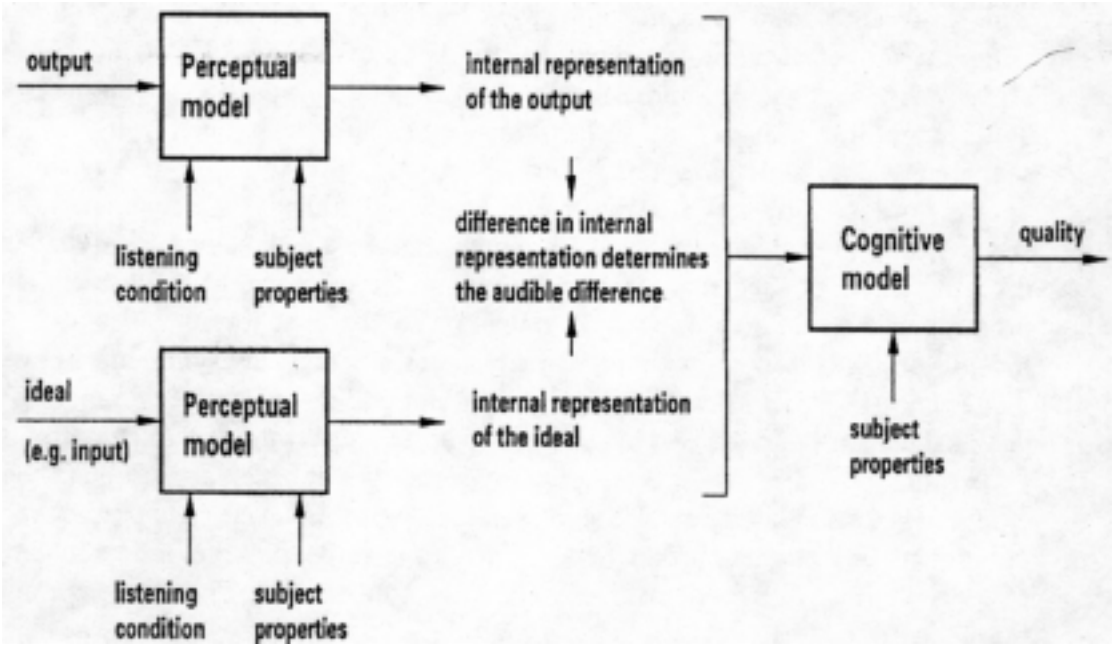
μόνη επίδραση εκμάθησης που έχει μοντελοποιηθεί είναι το ολικό ταίριασμα των χαρακτηριστικών σε τρεις διαφορετικές περιοχές συχνότητας. Αυτή η ενότητα θα εστιάσει σε βελτιώσεις στο πεδίο εκμάθησης και η βασική προσέγγιση που παρουσιάζεται στο σχήμα 1 τροποποιείται ελαφρώς (βλέπε σχήμα 13) με την ενσωμάτωση μίας κεντρικής μονάδας η οποία επεξηγεί τις διαφορές στην εσωτερική αναπαράσταση.



Σχήμα 12: Ομοίως όπως στο σχήμα 11 αλλά για την ETSI GSM μισού ρυθμού βάση δεδομένων ομιλίας.

Πιθανές κεντρικές επιδράσεις της διαδικασίας εκμάθησης οι οποίες είναι σημαντικές στον καθορισμό της υποκειμενικής ακουστικής ποιότητας είναι:

1. **Πληροφοριακή απόκρυψη,** όπου το κατώφλι απόκρυψης ενός σύνθετου στόχου που αποκρύπτεται από ένα σύνθετο αποκρύπτη μπορεί να ελαττωθεί μετά από εκπαίδευση για περισσότερα από 40 dB [Leek and Watson, 1984].
2. **Διαχωρισμός γραμμικών από μη γραμμικές αλλοιώσεις.** Γραμμικές αλλοιώσεις του σήματος εισόδου είναι λιγότερο ανεπιθύμητες από τις μη γραμμικές αλλοιώσεις.
3. **Ανάλυση ακουστικής σκηνής,** κατά την οποία λαμβάνονται αποφάσεις σύμφωνα με τις οποίες τα τμήματα ενός ακουστικού συμβάντος ολοκληρώνονται στο ένα τοις εκατό [Bregman, 1990].
4. **Φασματο-χρονικό ζύγισμα (weighting).** Μερικές φασματο-χρονικές περιοχές του ακουστικού σήματος μεταφέρουν περισσότερη πληροφορία, και έτσι μπορεί να είναι πιο σημαντικές, από άλλες. Για παράδειγμα αν κάποιος θεωρεί ότι τα διαστήματα σιγής στην ομιλία δεν μεταφέρουν πληροφορία τότε αυτά θεωρούνται λιγότερο σημαντικά.



Σχήμα 13: Η βασική προσέγγιση που χρησιμοποιείται στην ανάπτυξη του $PAQM_C$, το PAQM εκπαιδευόμενης διόρθωσης. Οι διαφορές στην εσωτερική αναπαράσταση κρίνονται από μία κεντρική διαδικασία εκμάθησης.

1) Η πληροφοριακή απόκρυψη μπορεί να μοντελοποιηθεί με τον ορισμό ενός μέτρου φασματο-χρονικής πολυπλοκότητας, παρόμοιο με την εντροπία. Το αποτέλεσμα είναι κατά πάσα πιθανότητα εξαρτώμενο από το μέγεθος της εκπαίδευσης στο οποίο εκτίθονται οι άνθρωποι πριν πραγματοποιηθεί η υποκειμενική αξιολόγηση. Γενικά, οι αξιολογήσεις ποιότητας των κωδικοποιητών-αποκωδικοποιητών ομιλίας πραγματοποιούνται από άπειρους ακροατές [CCITTrecP80, 1994],ενώ οι μουσικοί κωδικοποιητές-αποκωδικοποιητές αξιολογούνται κυρίως από πεπειραμένους ακροατές [CCITTrec562, 1990], [ITURrecBS1116, 1994].

Για μερικές βάσεις δεδομένων η επίδραση της πληροφοριακής απόκρυψης διαδραματίζει σημαντικό ρόλο και η μοντελοποίηση αυτής της επίδρασης έγινε τελικά υποχρεωτική για την απόκτηση υψηλών συσχετίσεων μεταξύ υποκειμενικών και αντικειμενικών αποτελεσμάτων [Beerends et al., 1996]. Η μοντελοποίηση πραγματοποιείται καλύτερα με τον υπολογισμό ενός αριθμού τοπικής πολυπλοκότητας πάνω από ένα παράθυρο χρόνου των 100 ms περίπου. Εάν αυτή η τοπική πολυπλοκότητα είναι υψηλή τότε οι αλλοιώσεις μέσα σε αυτό το χρονικό παράθυρα είναι πιο δύσκολο να ακουστούν από όταν η τοπική πολυπλοκότητα είναι χαμηλή [Beerends et al., 1996].

Αν και η μοντελοποίηση της πληροφοριακής απόκρυψης δίνει υψηλότερες συσχετίσεις για μερικές βάσεις δεδομένων , σε άλλες βάσεις δεδομένων μπορεί να υπάρξει ελάττωση της συσχέτισης. Καμία γενική μορφοποίηση δεν έχει βρεθεί που να μπορεί να χρησιμοποιηθεί για την μοντελοποίηση πληροφοριακής απόκρυψης με ένα ικανοποιητικό, γενικά εφαρμόσιμο, τρόπο. Συνεπώς η μοντελοποίηση αυτής της επίδρασης είναι ακόμα υπό μελέτη και δεν λαμβάνεται υπόψη εδώ.

2) Ο διαχωρισμός των γραμμικών από τις μη γραμμικές αλλοιώσεις μπορεί να πραγματοποιηθεί σχετικά εύκολα με την χρήση προσαρμοστικού αντιστρόφου φιλτραρίσματος του σήματος εξόδου. Όμως δεν έδωσε ουσιαστική βελτίωση στην συσχέτιση μεταξύ των αντικειμενικών και των υποκειμενικών αποτελεσμάτων με την

χρήση των διαθέσιμων βάσεων δεδομένων (ISO/MPEG 1990, ISO/MPEG 1991, ITU-R 1993, ETSI GSM full rate 1988, ETSI GSM half rate 1991).

Παρόλα αυτά ανεπίσημα πειράματα έδειξαν ότι αυτός ο διαχωρισμός είναι απαραίτητος όταν το σήμα εξόδου περιλαμβάνει έντονες γραμμικές αλλοιώσεις.

3) Η ανάλυση ακουστικής σκηνής είναι μία επίδραση εκμάθησης η οποία περιγράφει πως οι άνθρωποι ξεχωρίζουν διαφορετικά ακουστικά γεγονότα και τα ομαδοποιούν σε διαφορετικά αντικείμενα (objects). Αν και ένα πλήρες μοντέλο ακουστικής ανάλυσης σκηνής είναι πέρα από το επίκεντρο αυτού του κεφαλαίου η επίδραση ερευνήθηκε λεπτομερώς. Μία πραγματική προσέγγιση που δίνεται στο [Beerends and Stemerdink, 1994a] έγινε πολύ σημαντική στην ποσοτικοποίηση μίας επίδρασης ανάλυσης ακουστικής σκηνής. Η κεντρική ιδέα αυτής της προσέγγισης είναι το γεγονός ότι εάν μία συνιστώσα χρόνου-συχνότητας δεν κωδικοποιείται από έναν κωδικοποιητή-αποκωδικοποιητή, το υπολειπόμενο σήμα ακόμη μορφοποιεί μία περιεκτική (coherent) ακουστική σκηνή ενώ η εισαγωγή μίας νέας ασυσχέτιστης συνιστώσας χρόνου-συχνότητας οδηγεί σε δύο διαφορετικές παρατηρήσεις (percepts). Λόγω του διαχωρισμού σε δύο διαφορετικές παρατηρήσεις η αλλοίωση θα γίνει περισσότερο ανεπιθύμητη από ότι μπορεί να περιμένει κανείς ως προς την ακουστότητα της νεοεισερχόμενης συνιστώσας αλλοίωσης. Αυτό οδηγεί σε μία αντιληπτή ασυμμετρία ανάμεσα στην διακύμανση της αλλοίωσης που προκαλείται όταν δεν κωδικοποιείται μία συνιστώσα χρόνου-συχνότητας και στη διακύμανση που προκαλείται από την εισαγωγή μίας νέας συνιστώσας χρόνου-συχνότητας.

Για να είμαστε ικανοί να κωδικοποιήσουμε αυτή τη επίδραση της εκμάθησης ήταν αναγκαίο να ποσοτικοποιήσουμε σε κάποια έκταση την αλλοίωση, όπως βρέθηκε από το μοντέλο, που προκαλείται από την παράληψη μίας συνιστώσας χρόνου-συχνότητας ή από την εισαγωγή μίας νέας συνιστώσας χρόνου-συχνότητας στο σήμα. Ένα πρόβλημα ήταν το ότι όταν μία αλλοίωση εισάγεται στο σήμα σε ένα συγκεκριμένο σημείο χρόνου-συχνότητας γενικά θα υπάρχει ήδη ένα συγκεκριμένο επίπεδο ισχύος σε αυτό το σημείο. Άρα μία συνιστώσα χρόνου-συχνότητας δεν θα είναι ποτέ εντελώς καινούργια. Μία πρώτη προσέγγιση για την ποσοτικοποίηση της ασυμμετρίας ήταν η χρήση του πηλίκου ισχύος ανάμεσα στην έξοδο και την είσοδο σε ένα συγκεκριμένο σημείο χρόνου-συχνότητας για την ποσοτικοποίηση της “νεότητας” αυτής της συνιστώσας. Όταν το πηλίκο ισχύος μεταξύ της εξόδου y και της εισόδου x, p_y / p_x σε ένα συγκεκριμένο σημείο χρόνου-συχνότητας είναι μεγαλύτερο από 1.0 μία ακουστή αλλοίωση είναι πιο ενοχλητική από όταν αυτό το πηλίκο είναι μικρότερο από 1.0.

Στο μοντέλο εσωτερικής αναπαράστασης το επίπεδο χρόνου-συχνότητας διαιρείται σε κυψέλες (cells) με μία ανάλυση των 20ms κατά μήκος του άξονα του χρόνου (δείκτης χρόνου m) και των 0.2 Bark κατά μήκος του άξονα των συχνοτήτων (δείκτης συχνότητας l). Μία πρώτη προσέγγιση ήταν η χρήση του πηλίκου ισχύος μεταξύ της εξόδου y και της εισόδου x, p_y / p_x σε κάθε (Δf, Δt) κυψέλη (m,l) σαν ένα παράγοντα διόρθωσης για την αλλοίωση του θορύβου $L_n(m,l)$ σε αυτή την κυψέλη (η ονοματολογία διαλέγεται να είναι συμβατή με το [Beerends and Stemerdink, 1992]).

Μία καλύτερη προσέγγιση που παρουσιάστηκε είναι ο υπολογισμός του μέσου όρου του πηλίκου ισχύος p_y / p_x μεταξύ της εξόδου y και της εισόδου x για ένα αριθμό διαδοχικών χρονικών πλαισίων. Αυτό υπονοεί ότι εάν ένας κωδικοποιητής-αποκωδικοποιητής εισάγει μία νέα συνιστώσα χρόνου-συχνότητας αυτή θα είναι περισσότερο ενοχλητική εάν αυτή παρουσιάζεται πάνω από ένα αριθμό διαδοχικών πλαισίων. Η γενική μορφή της διόρθωσης εκμάθησης ορίζεται ως εξής:

$$L_{C_n}(m,l) = \begin{cases} C(m,l)^{\lambda} L_n(m,l) & \text{εάν } C(m,l) < 5 \\ 5^{\lambda} L_n(m,l) & \text{εάν } C(m,l) \geq 5 \end{cases}$$

όπου

$$C(m,l) = \sum_{i=0}^{4} \frac{p_y(m-i,l)}{p_x(m-i,l)} \qquad (6)$$

και ένας επιπρόσθετος ψαλιδισμός της αλλοίωσης θορύβου σε κάθε χρονικό παράθυρο:

$$L_{C_n}(m) = \sum_{l=1}^{l=max} L_{C_n}(m,l)$$

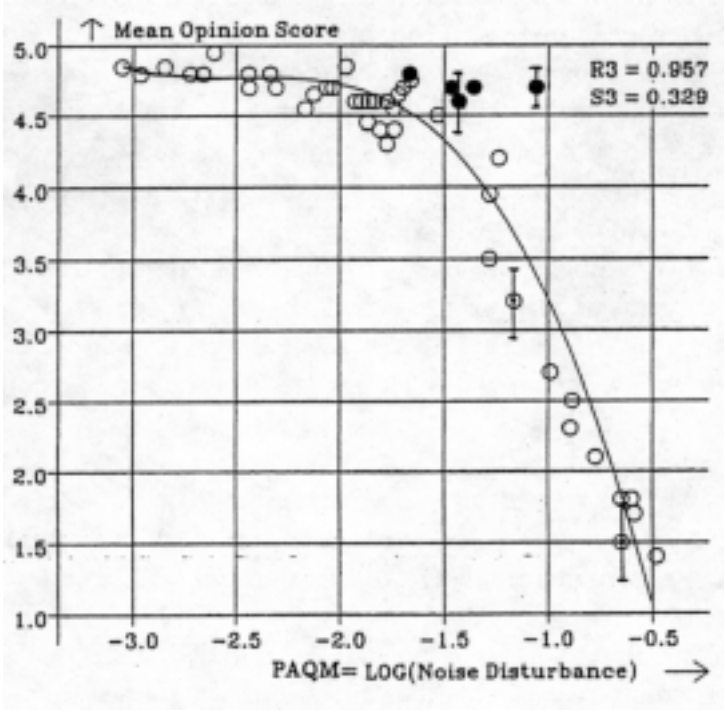
σε ένα επίπεδο των 20.

Η απλή μοντελοποίηση της ακουστικής ανάλυσης σκηνής με τον παράγοντα ασυμμετρίας *C(m,l)* προσέφεραν σημαντικές βελτιώσεις στην συσχέτιση μεταξύ υποκειμενικών και αντικειμενικών αποτελεσμάτων. Παρόλα αυτά βρέθηκε ότι για μέγιστη συσχέτιση το μέγεθος της διόρθωσης, όπως αυτό ποσοτικοποιήθηκε από τον παράγοντα λ, ήταν διαφορετικό για ομιλία και για μουσική. Όταν εφαρμόστηκε σε βάσεις δεδομένων μουσικής η βέλτιστα διορθωμένη διακύμανση θορύβου βρέθηκε για λ =1.4 ($PAQM_{C1.4}$) ενώ για βάσεις δεδομένων ομιλίας το λ ήταν περίπου 4.0 ($PAQM_{C4.0}$).

Τα αποτελέσματα αξιολόγησης κωδικοποιητών-αποκωδικοποιητών μουσικής δίνονται στα σχήματα 14 (ISO/MPEG 1991) και 15 (ITU-R TG10/2 1993) και παρουσιάζουν μία μείωση στο τυπικό σφάλμα της MOS εκτίμησης μεγαλύτερη από 25%. Για την ISO/MPEG 1990 βάση δεδομένων δεν παρουσιάστηκε καμία βελτίωση. Για την ομιλία η βελτίωση ήταν ελαφρώς μικρότερη αλλά όπως προέκυψε η τελευταία από τις ταξινομημένες επιδράσεις εκμάθησης, το φασματο-χρονικό ζύγισμα, ελέγχει τις υποκειμενικές εκτιμήσεις ποιότητας ομιλίας. Το τυπικό σφάλμα της MOS εκτίμησης στις βάσεις δεδομένων ομιλίας μπορεί να μειωθεί αρκετά περισσότερο όταν και η ασυμμετρία και το φασματο-χρονικό ζύγισμα μοντελοποιούνται ταυτόχρονα.

4) Το φασματο-χρονικό ζύγισμα βρέθηκε ότι είναι σημαντικό μόνο στις κρίσεις ποιότητας στους κωδικοποιητές-αποκωδικοποιητές ομιλίας. Πιθανώς στην μουσική όλα τα φασματο-χρονικά τμήματα του σήματος, ακόμα και οι παύσεις, μεταφέρουν πληροφορία, ενώ στην ομιλία μερικά φασματο-χρονικά τμήματα, όπως οι φωνοσυντονισμοί μεταφέρουν περισσότερη πληροφορία από άλλα, όπως οι παύσεις. Επειδή οι βάσεις δεδομένων ομιλίας που χρησιμοποιούνται σε αυτό το paper είναι όλες περιορισμένες στην τηλεφωνική ζώνη το φασματικό ζύγισμα παρουσιάζεται να είναι ελάχιστης σημασίας και μόνο το ζύγισμα στο χρόνο χρειάζεται να μοντελοποιηθεί.

Αυτή η επίδραση του ζυγίσματος πάνω από τον χρόνο μοντελοποιήθηκε με ένα πολύ απλό τρόπο, τα πλαίσια ομιλίας κατηγοριοποιήθηκαν σε δύο σύνολα, ένα σύνολο για τα ενεργά πλαίσια ομιλίας και ένα για τα πλαίσια σιγής. Με το ζύγισμα της διακύμανσης θορύβου που παρουσιάζεται στα πλαίσια σιγής με ένα παράγοντα W_{sil} με τιμές μεταξύ 0.0 (οι παύσεις δεν λαμβάνονται υπόψη) και 0.5 (οι παύσεις είναι τόσο σημαντικές όσο και η ομιλία) η επίδραση ποσοτικοποιήθηκε.



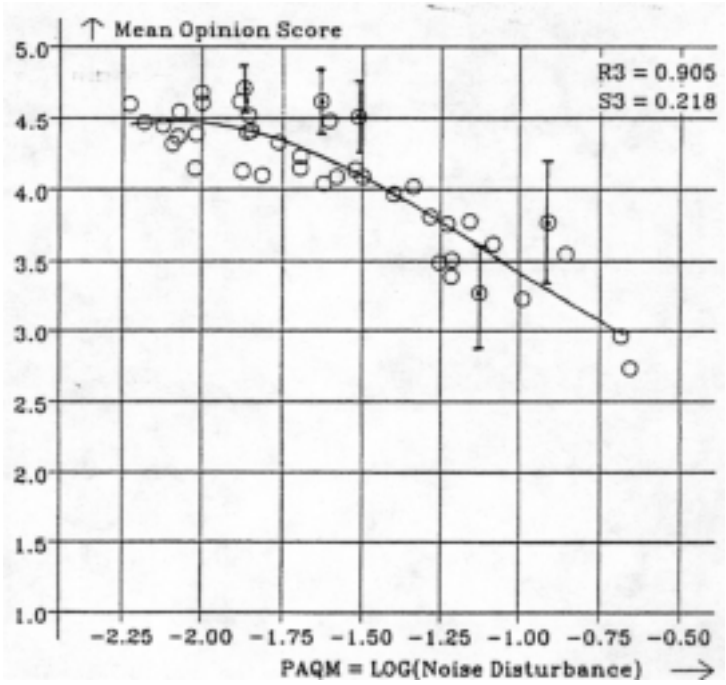
Σχήμα 14: Σχέση μεταξύ του MOS και του $PAQM_{C1.4}$ για τις 50 καταχωρίσεις της ISO/MPEG 1991 δοκιμής κωδικοποιητών-αποκωδικοποιητών σε αναπαράσταση με loudspeaker.

Ένα πρόβλημα στην ποσοτικοποίηση της συμπεριφοράς των διαστημάτων σιγής είναι το ότι η επιρροή των διαστημάτων σιγής εξαρτάται άμεσα από το μήκος αυτών των διαστημάτων. Εάν η ομιλία εισόδου δεν περιέχει διαστήματα σιγής τότε η επιρροή είναι μηδέν. Εάν η ομιλία εισόδου περιέχει ένα συγκεκριμένο ποσοστό πλαισίων σιγής η επιρροή είναι ανάλογη με αυτό το ποσοστό. Χρησιμοποιώντας ένα σύνολο οριακών συνθηκών, με L_{spn} την μέση διακύμανση θορύβου πάνω από ενεργά πλαίσια ομιλίας και L_{siln} την μέση διακύμανση θορύβου πάνω από πλαίσια σιγής μπορεί να αποδειχθεί ότι το σωστό ζύγισμα είναι:

$$L_{Wn} = \frac{W_{sp} \cdot p_{sp}}{W_{sp} \cdot p_{sp} + p_{sil}} L_{spn} + \frac{p_{sil}}{W_{sp} \cdot p_{sp} + p_{sil}} L_{si\,ln} \qquad (7)$$

όπου
LWn είναι η διορθωμένη κατά ένα παράγοντα Wsil διακύμανση θορύβου
$W_{sp} = (1\text{-}W_{sil})/W_{sil}$
p_{sil} το ποσοστό των πλαισίων σιγής

p_{sp} το ποσοστό των πλαισίων ενεργής ομιλίας ($p_{sil} + p_{sp} = 1.0$)



Σχήμα 15: Σχέση μεταξύ του MOS και του $PAQM_{C1.4}$ για τις 43 ISO layer II συνθήκες της ITU-R TG10/2 1993 δοκιμής ακουστικών κωδικοποιητών-αποκωδικοποιητών [ITURsg10cond9343, 1993].

Όταν και τα διαστήματα σιγής και τα διαστήματα ενεργής ομιλίας είναι εξίσου σημαντικά, όπως συναντάται στις δοκιμές κωδικοποιητών-αποκωδικοποιητών μουσικής, ο παράγοντας βάρους W_{sil} είναι ίσος με 0.5 και η εξίσωση (1.7) γίνεται ως εξής $L_{Wn} = p_{sp} \cdot L_{spn} + p_{sil} \cdot L_{siln}$. Και για τις δύο από τις βάσεις δεδομένων ομιλίας ο παράγοντας βάρους για θόρυβο διαστήματος σιγής για μέγιστη συσχέτιση μεταξύ αντικειμενικών και υποκειμενικών αποτελεσμάτων βρέθηκε ότι είναι 0.1 δείχνοντας ότι ο θόρυβος σε διαστήματα σιγής είναι λιγότερο ενοχλητικός από τον θόρυβο ίδιας έντασης κατά την διάρκεια ομιλίας.

Όταν η επίδραση ασυμμετρίας που προκύπτει από την ακουστική ανάλυση σκηνής και το χρονικό ζύγισμα ποσοτικοποιούνται σωστά, η συσχέτιση ανάμεσα στα αντικειμενικά και στα υποκειμενικά αποτελέσματα και για τις δύο βάσεις δεδομένων ομιλίας βελτιώνεται σημαντικά. Χρησιμοποιώντας $\lambda = 0.4$ (ασύμμετρη μοντελοποίηση) και ένα βάρος διαστημάτων σιγής 0.1 (υποδηλώνεται σαν $PAQM_{c4.0,w0.1}$) η ελάττωση στο τυπικό σφάλμα της MOS εκτίμησης είναι περίπου 40% και για τις δύο βάσεις δεδομένων την ETSI GSM full rate και την half rate .

Ένα πρόβλημα των δύο μονάδων εκμάθησης που προκύπτουν είναι ότι η πρόβλεψη της υποκειμενικά αντιληπτής ποιότητας εξαρτάται από το πειραματικό περιβάλλον. Άρα πρέπει προκαταβολικά να θεσπιστούν τιμές για την επίδραση ασυμμετρίας και για το βάρος των διαστημάτων σιγής.

2. ΓΕΝΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΤΗΣ ΕΠΙΚΟΙΝΩΝΙΑΣ ΜΕ ΟΜΙΛΙΑ

Περίληψη: Η πρόοδος στην ψηφιακή επεξεργασία ομιλίας επιτρέπει πλέον την εφαρμογή και χρήση μιας ποικιλίας τεχνολογιών ομιλίας για την επικοινωνία ανθρώπου μηχανής. Μάλιστα νέες εταιρίες δημιουργούνται ταχέως γύρω από αυτές τις τεχνολογίες. Όμως αυτές οι δυνατότητες χρησιμοποιούνται μόνο αν η κοινωνία μπορεί να τις αντέξει οικονομικά. Ευτυχώς η εκρηκτική πρόοδος στον τομέα της μικροηλεκτρονικής όσο και στην τεχνολογία των υπολογιστών τις τελευταίες δύο δεκαετίες έχει κάνει οικονομικώς προσιτά αυτά τα επιτεύγματα.

Οι προκλήσεις έρευνας όσον αφορά την επεξεργασία ομιλίας παραμένουν στους παραδοσιακά αναγνωρισμένους τομείς της αναγνώρισης, σύνθεσης και κωδικοποίησης. Αυτές οι τρεις περιοχές τυπικά έχουν δρομολογηθεί ανεξάρτητα η μία από την άλλη, ουσιαστικά με σημαντική απομόνωση μεταξύ τους. Παρόλα αυτά αποτελούν πλευρές του ίδιου θεμελιώδους θέματος—το οποίο είναι το πως θα αναπαρασταθεί και θα προσδιοριστεί η πληροφορία του σήματος ομιλίας. Αυτό προϋποθέτει βαθύτερη κατανόηση της φυσικής της παραγωγής ομιλίας, των περιορισμών που επιβάλλονται από την εφαρμογή των κανόνων της γλώσσας και των μηχανισμών επεξεργασίας πληροφορίας στο σύστημα ακοής. Οπότε στην έρευνα πλέον αναζητούμε πιο ακριβή μοντέλα παραγωγής ομιλίας, καλύτερα υπολογιστικά σχήματα παραγωγής της γλώσσας και ρεαλιστικούς οδηγούς αντίληψης για επεξεργασία ομιλίας—μαζί με τρόπους συνδυασμού των βασικών θεμάτων αναγνώρισης, σύνθεσης και κωδικοποίησης. Επιτυχής λύση θα παράγει την μακρινής αναζήτησης μηχανή υπαγόρευσης, υψηλής ποιότητας σύνθεση από κείμενο και την καλύτερη δυνατή εκπομπή ομιλίας σε χαμηλούς ρυθμούς δεδομένων. Θα ανοίξει επίσης τον δρόμο για την μετάφραση γλωσσών στην τηλεφωνία, όπου η συνθετική μετάφραση θα γίνεται με την φωνή του ομιλητή.

2.1 ΕΙΣΑΓΩΓΗ

Η ομιλία είναι ένα από τα προτιμώμενα μέσα επικοινωνίας μεταξύ των ανθρώπων. Μάλιστα έχει ξεκινήσει να γίνεται ένα επιθυμητό μέσο επικοινωνίας για την επικοινωνία ανθρώπου – μηχανής. Σταδιακά, για καλά καθορισμένες δουλειές, οι μηχανές κατάφεραν να αποκτήσουν πολλές από τις δυνατότητες της ανταλλαγής διαλόγου. Έτσι λοιπόν η ισχύς των πολύπλοκων υπολογιστών μπορεί να εκμεταλλευτεί για τις κοινωνικές ανάγκες χωρίς να επιβαρύνεται ο χρήστης με την περαιτέρω γνώση των ομιλουμένων γλωσσών.

Επειδή οι άνθρωποι είναι σχεδιασμένοι να ζούνε μέσα στην ατμόσφαιρα, ήταν αναπόφευκτο να μάθούν να ανταλλάσσουν πληροφορίες με την μορφή ηχητικών κυμάτων τα οποία δημιουργούνται με μετακινήσεις των μορίων του αέρα. Αλλά μεταξύ των χιλιάδων τύπων ακουστικών πληροφοριακών σημάτων η ομιλία αποτελεί ένα ξεχωριστό είδος. Είναι περιορισμένη κατά τους τρεις ακόλουθους τρόπους:

- Από τη φυσική που διέπει την παραγωγή ήχου στο σύστημα ομιλίας.
- Από τις ιδιότητες της ανθρώπινης ακοής και αντίληψης, και
- Από τους κανόνες της γλώσσας.

Αυτοί οι περιορισμοί έχουν γίνει καθοριστικοί στην έρευνα της ομιλίας και τυγχάνουν εξαιρετικής σημασίας ακόμα και σήμερα.

Σχήμα 1: Οι Αρχαίοι λαοί χρησιμοποιούσαν ομιλούντα αγάλματα για να εντυπωσιάζουν, να ψυχαγωγούν και να προκαλούν δέος.

Αυτή η εργασία προτείνει την μελέτη του πεδίου επεξεργασίας ομιλίας με τρεις τρόπους :

- Πρώτον, παρουσιάζοντας μια σύντομη αναδρομή της επιστήμης
- Δεύτερον, υποδεικνύοντας σημαντικές κατευθύνσεις έρευνας, και
- Τρίτον διακινδυνεύοντας κάποιες τεχνολογικές εκτιμήσεις

2.2 ΑΡΧΕΣ ΤΗΣ ΤΕΧΝΟΛΟΓΙΑΣ ΟΜΙΛΙΑΣ

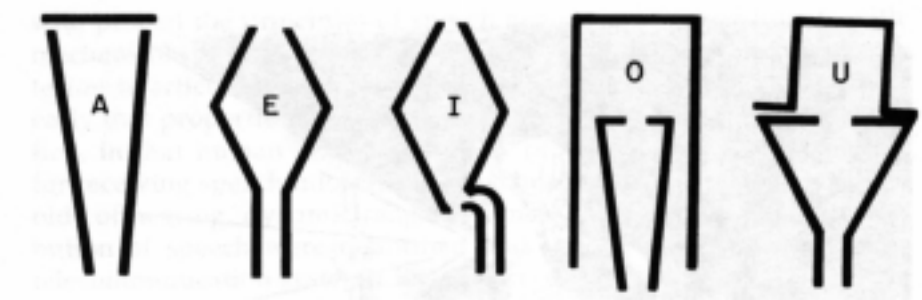
Η επεξεργασία ομιλίας , σαν επιστήμη, μπορεί να θεωρηθεί ότι έχει προέλθει από την εξέλιξη των ηλεκτρικών επικοινωνιών. Η ανακάλυψη του τηλεφώνου και η ανάπτυξη

των τηλεπικοινωνιών σαν μια υπηρεσία που εξυπηρετεί την κοινωνία, προκάλεσαν την εργασία πάνω στην θεωρία των δικτύων, την έρευνα πάνω στους μετατροπείς ενέργειας (transducers), τον σχεδιασμό φίλτρων, την φασματική ανάλυση, την ψυχοακουστική (psychoacoustics), τις μεθόδους διαμόρφωσης και τις ενσύρματες και ασύρματες τεχνικές μετάδοσης. Παλιότερα η ακουστική και η φυσιολογία της παραγωγής ομιλίας είχαν αναγνωριστεί ως σημαντικά θέματα προς κατανόηση. Το ίδιο ισχύει και σήμερα, παρόλο που έχει αποκτηθεί πολύ περισσότερη γνώση από τότε. Ο Alexander Graham Bell ήταν ανάμεσα σε αυτούς που εξερεύνησαν της αρχές της παραγωγής ομιλίας σε πειράματα με μηχανικές μηχανές ομιλίας. Ακόμα είχε αναγνωριστεί από νωρίς ότι οι ιδιότητες της ακοής και της κατανόησης χρειάζονταν να προσδιοριστούν, διότι η ανθρώπινη ακοή ουσιαστικά παρέχει ένα κριτήριο ακριβείας για την λήψη των πληροφοριών της ομιλίας. Η ψυχοακουστική συμπεριφορά για τα όρια της ακοής, το δυναμικό εύρος, την ένταση, το ύψος και την φασματική διασπορά της ομιλίας προσδιορίστηκαν και χρησιμοποιήθηκαν κατά τον σχεδιασμό των πρώτων τηλεπικοινωνιακών συστημάτων. Όμως μόνο πρόσφατα, με την χρήση της ισχύος των υπολογιστών, έχουν γίνει προσπάθειες για να συμπεριληφθούν και άλλες λεπτομέρειες της ακοής—όπως η απόκρυψη στον χρόνο και την συχνότητα—στους αλγορίθμους επεξεργασίας ομιλίας. Ακόμα, μόνο πρόσφατα δόθηκε αρκετή προσοχή στην αναλυτική μοντελοποίηση της γλώσσας και αυτό έχει γίνει ιδιαίτερα σημαντικό καθώς οι τεχνικές σύνθεσης ομιλίας από κείμενο και αυτόματης αναγνώρισης συνεχούς λόγου έχουν εξελιχθεί.

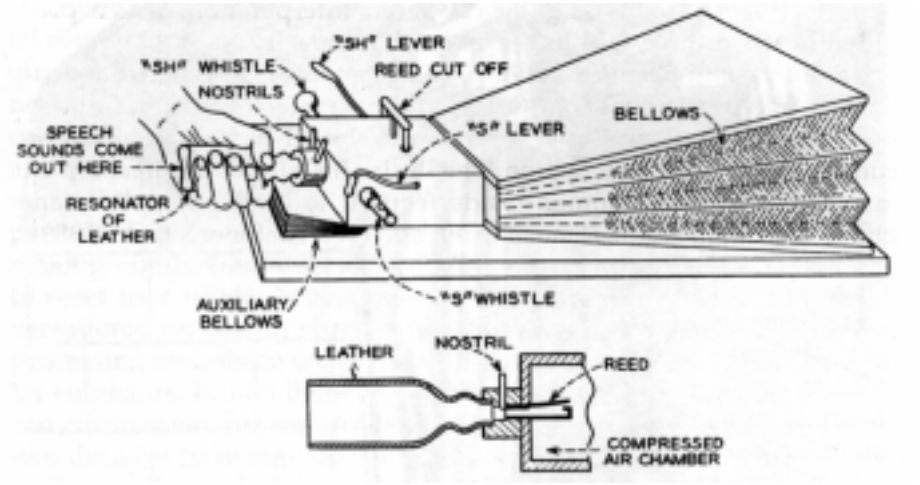
Περίπου στο μέσον αυτού του αιώνα αναπτύχθηκαν ταυτόχρονα η θεωρία δειγματοληψίας δεδομένων και οι ψηφιακοί υπολογισμοί, ανοίγοντας νέες προοπτικές για υψηλής ποιότητας και μακρινής απόστασης τηλεπικοινωνίες και για εξομοίωση του μηχανικού σχεδίου πολύπλοκων συστημάτων γρήγορα και οικονομικά. Αλλά η τεχνολογία υπολογισμών γρήγορα αναπτύχθηκε πέρα από την ταξινόμηση δεδομένων για εταιρείες και την εξομοίωση αλγορίθμων για την επιστήμη. Φθηνή αριθμητική και οικονομική αποθήκευση, μαζί με την επεκτεινόμενη γνώση για σήματα πληροφορίας, επέτρεψαν στους υπολογιστές να αναλάβουν λειτουργίες περισσότερο συσχετιζόμενες με την λήψη αποφάσεων—κατανοώντας με λεπτομέρεια τις προθέσεις του χρήστη και εφαρμόζοντας τρόπους εξυπηρέτησης των αναγκών του. Η επεξεργασία ομιλίας—που δίνει στους υπολογιστές την διαλογική ικανότητα—έχει γίνει το επίκεντρο αυτής της ανάπτυξης. Παρόμοια έμφαση δίνεται στην επεξεργασία εικόνας και, πιο πρόσφατα, στην αλληλεπίδραση επαφής. Αλλά όλες αυτές οι δυνατότητες είναι χρήσιμες μόνο αν η κοινωνία μπορεί να τις αντέξει οικονομικά. Η εκρηκτική πρόοδος στον τομέα της μικροηλεκτρονικής τις τελευταίες δύο δεκαετίες έχει εξασφαλίσει προσιτή πρόσβαση σε αυτά τα επιτεύγματα όσο και στην τεχνολογία των υπολογιστών. Όλες οι ενδείξεις μας οδηγούν στο συμπέρασμα ότι οι πρόοδοι στην τεχνολογία των υπολογισμών θα συνεχιστούν και ότι όταν θα χρειαστεί θα γίνουν οικονομικοί υπολογισμοί που θα υποστηρίξουν την τεχνολογία ομιλίας.

2.3 ΚΙΝΗΤΡΑ ΣΤΗΝ ΕΡΕΥΝΑ ΤΗΣ ΟΜΙΛΙΑΣ

Οι πειραματισμοί με την ομιλία συχνά παροτρύνονταν από την επιθυμία του ενθουσιασμού , της διασκέδασης ή του δέους. Ομιλούντα αγάλματα και θεοί προτιμήθηκαν από τους αρχαίους Έλληνες και τους Ρωμαίους. Αλλά μερικές φορές



Σχήμα 2: Η βραβευμένη κατασκευή αντιχείων που εξομοίωνε ανθρώπινους ήχους φωνηέντων του **Kratzenstein** *(1779). Τα αντιχεία ενεργοποιούνταν από παλλόμενα ελάσματα ανάλογα με τις φωνητικές χορδές. Η διαφορά με τα φυσικά σχήματα άρθρωσης υποδεικνύει την μη μοναδικότητα ανάμεσα στο φάσμα του ήχου και στο σχήμα του αντιχείου.*



Σχήμα 3: Ανακατασκευή της ομιλούσας μηχανής του **Von Kempelen** *(1791), που χρεώθηκε στον Sir Charles Wheatstone (1879). Τυπικά ο βραχίονας και η παλάμη του ενός χεριού τοποθετούνταν πάνω από το κύριο* **bellows** *και το αντιχείο αντίστοιχα για να παράγουν φωνητικούς ήχους ενώ το άλλο χέρι χειρίζονταν τα βοηθητικά* **bellows** *και θύρες για τους μη φωνητικούς ήχους.*

το κίνητρο ήταν η δημιουργική περιέργεια. Και μερικές φορές οι προσπάθειες δεν είχαν επιστημονική αξιοπιστία (η ομιλούσα μηχανή του Von Kemplen αγνοήθηκε ευρέως λόγω του αυτόματου του που έπαιζε σκάκι ενώ στην ουσία είχε έναν άνθρωπο κρυμμένο).

Τα ακουστικά κύματα εξαπλώνονται σφαιρικά και δεν διαδίδονται καλά σε μεγάλες αποστάσεις. Όμως η επικοινωνία σε μεγάλες αποστάσεις ήταν από παλιά βασική ανάγκη της ανθρώπινης κοινωνίας. Καθώς η κατανόηση των ηλεκτρικών φαινομένων εξελίχθηκε, ο ηλεκτρικός τηλέγραφος εμφανίστηκε στα μέσα του δεκάτου ενάτου αιώνα. Ακολουθώντας αυτή την επιτυχία της επικοινωνίας με τελείες και παύλες, περισσότερη προσοχή δόθηκε στην προοπτική αποστολής φωνητικών

σημάτων πάνω από ηλεκτρικά καλώδια. Έτσι η ανακάλυψη της τηλεφωνίας αποτελεί σήμερα ιστορία.

Στα πρώτα χρόνια του εικοστού αιώνα, το κίνητρο παρέμενε η φωνητική επικοινωνία σε ακόμα μεγαλύτερες αποστάσεις. Η ενίσχυση των αναλογικών σημάτων τα οποία εξασθενίζουν με την απόσταση και εμπλουτίζονται με θόρυβο χρειάζονταν. Το 1915 η διηπειρωτική τηλεφωνία επιτεύχθηκε με οριακή πιστότητα με την χρήση ηλεκτρομηχανικών επαναληπτών (“repeaters”). Τα διαατλαντικά καλώδια τηλεγραφίας δεν μπορούσαν να υποστηρίζουν το εύρος ζώνης που χρειάζονταν για την φωνή και οι προσπάθειες έρευνας στράφηκαν στους φωνοκωδικοποιητές για συμπίεση του εύρους ζώνης. Το 1927, καθώς η ηλεκτρονική τεχνολογία προόδευσε, η διαατλαντική ραδιοτηλεφωνία έγινε πραγματικότητα. Η κατανόηση της συμπίεσης του εύρους ζώνης εφαρμόστηκε τότε για την διασφάλιση του απορρήτου και για την κρυπτογράφηση. Η διαατλαντική μετάδοση φωνής πάνω από συρμάτινα καλώδια έπρεπε να περιμένει μέχρι την ανάπτυξη αξιόπιστων υποβρύχιων ενισχυτών που έλαβε χώρα το 1956. Μαζί με αυτά τα ακριβά και συγχρόνως υψηλής ποιότητας φωνητικά κυκλώματα, το ενδιαφέρον για την διατήρηση του εύρους ζώνης εμφανίστηκε πάλι και προκάλεσε νέες εξελίξεις όπως η χρονική αποστολή τμημάτων ομιλίας (Time Assignment Speech Interpolation) η οποία παρέχει σχεδόν τρεις φορές αύξηση στη χωρητικότητα του καλωδίου.

Στα μέσα του εικοστού αιώνα παρουσιάστηκε πρόοδος στις τεχνικές δειγματοληψίας δεδομένων, στις ψηφιακές επικοινωνίες και στην μικροηλεκτρονική. Λόγω αυτών των εξελίξεων ένα ισχυρό ενδιαφέρον αναπτύχθηκε στην επικοινωνία ανθρώπου—μηχανής και στην αλληλεπίδραση μεταξύ τους. Η επιθυμία για ευκολία χρήσης των πολύπλοκων μηχανών που εξυπηρετούν τις ανθρώπινες ανάγκες οδήγησε στην εστίαση του ενδιαφέροντος πάνω στην επικοινωνία με την χρήση της ομιλούμενης γλώσσας (Flanagan et al., 1970; Rabiner et al., 1989). Έτσι προέκυψε ουσιαστική πρόοδος στην αναγνώριση και σύνθεση ομιλίας. Έμφαση δόθηκε στη διατήρηση του εύρους ζώνης και στη κωδικοποίηση χαμηλού ρυθμού δεδομένων τόσο για την οικονομία αποθήκευσης (σε υπηρεσίες όπως το φωνητικό ταχυδρομείο) όσο και για την οικονομία στην χωρητικότητα εκπομπής. Οι τελευταίες εξελίξεις στις κινητές κυψελοειδείς, προσωπικές και ασύρματες τηλεπικοινωνίες έχουν ανανεώσει το ενδιαφέρον για την διατήρηση του εύρους ζώνης και ταυτόχρονα να αποτελούν ένα ενισχυμένο κίνητρο για την διαφύλαξη του απορρήτου και για την κρυπτογραφία.

Σχήμα 4α: Επίδειξη ων αρχών της μεταφράζουσας τηλεφωνίας από την NEC Corporation στην Telecom στην Γενεύη το 1983. Το σενάριο της εφαρμογής ήταν ο διάλογος μεταξύ ενός σταθμάρχη στην

Ιαπωνία και μίας βρετανής τουρίστριας που είχε χάσει τις αποσκευές της. Η συζήτηση ήταν πραγματικού χρόνου, συνδεδεμένης ομιλίας, και μεταφράζονταν από Αγγλικά σε Ιαπωνικά και αντίστροφα, και χρησιμοποιούσε ένα καθορισμένο λεξιλόγιο και γραμματική «βιβλίου φράσεων».

Καθώς πλησιάζουμε στα όρια του εικοστού πρώτου αιώνα, νέα συστήματα παρουσιάζονται για την μεταφραστική τηλεφωνία. Αυτά τα συστήματα απαιτούν αυτόματη αναγνώριση μεγάλων λεξιλογίων ακριβείας σε μια γλώσσα από μια μεγάλη ποικιλία ομιλητών, μετάδοση της υπάρχουσας πληροφορίας ομιλίας και φυσικής ποιότητας σύνθεση σε μια ξένη γλώσσα—με την ακριβή ποιότητα φωνής του ομιλητή αν αυτό είναι δυνατό. Την παρούσα στιγμή μόνο μετάφραση τύπου «βιβλίου φράσεων» («phrase book») έχει επιτευχθεί με περιορισμένη γραμματική και υποτυπώδη λεξιλόγια και η φωνή που συντίθεται δεν πλησιάζει την ποιότητα φωνής των ομιλητών. Η μεταφραστική τηλεφωνία και οι μηχανές υπαγόρευσης απαιτούν σημαντική πρόοδο των υπολογιστικών μοντέλων της γλώσσας που μπορούν να συνδυάσουν φυσική προφορική γραμματική με μεγάλα λεξιλόγια. Συστήματα αναγνώρισης που χρησιμοποιούν μοντέλα για τις ρίζες των λέξεων θα πρέπει να έχουν γλωσσικούς κανόνες που σχηματίζουν (α) αποδεκτές υποψήφιες λέξεις από τις εκτιμούμενες αλληλουχίες φωνητικών μονάδων, (β) υποψήφιες προτάσεις από τις αλληλουχίες λέξεων και (γ) υποψήφιες έννοιες από τις προτάσεις. Η απλή καθημερινή διαλογική ομιλία, με όλες τις παραλλαγές της και την χωρίς γραμματική δομή της, παρουσιάζει ξεχωριστές προκλήσεις στον σχεδιασμό ελέγξιμων γραμματικών, συντακτικών και εννοιολογικών μοντέλων.

2.4 ΚΑΤΑΣΤΑΣΗ ΤΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

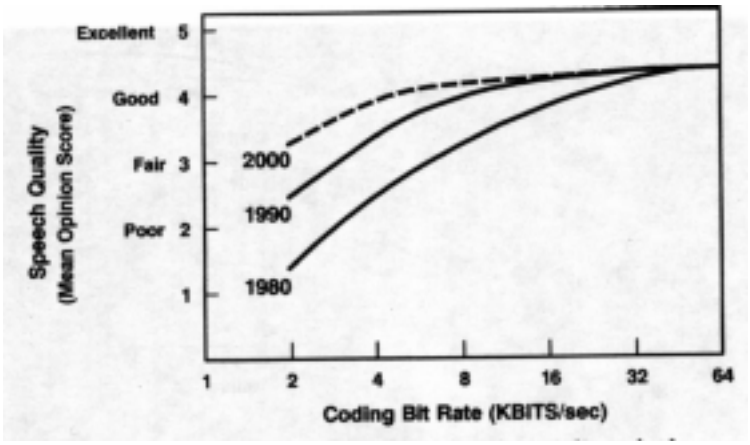
Μία λειτουργική πρόκληση στην επεξεργασία ομιλίας είναι το πως να αναπαραστήσουμε, να μετρήσουμε και να ανακτήσουμε την πληροφορία που βρίσκεται στο σήμα ομιλίας. Έχει καθιερωθεί η έρευνα να εστιάζεται πάνω στους τομείς της κωδικοποίησης, της αναγνώρισης ομιλίας και του ομιλητή, και της σύνθεσης.

Σχήμα 4β: Ένα διεθνές κοινό πείραμα πάνω στην μεταφραστική τηλεφωνία πραγματοποιήθηκε τον Ιανουάριο του 1993, και σύνδεσε τα ATR laboratories (Ιαπωνία), το Carnegie-Mellon University (ΗΠΑ), την Siemens A. G. (Γερμανία) και το Karlsruhe University (Γερμανία). Οι προτάσεις των ομιλούντων πρώτα αναγνωρίζονταν και μεταφράζονταν από ένα υπολογιστή σε γραπτή μορφή, η οποία αποστέλλονταν με ένα modem πάνω από μία τηλεφωνική γραμμή. Ένας συνθέτης φωνής που βρίσκονταν στον δέκτη έλεγε τις μεταφρασμένες λέξεις. Το σύστημα που

παρουσιάστηκε είχε περιοριστεί στην εργασία της εγγραφής συμμετεχόντων για μία διεθνή συνδιάσκεψη.

Κωδικοποίηση. Η υψηλής ποιότητας ψηφιακή κωδικοποίηση ομιλίας χρησιμοποιείται εδώ και πολλά χρόνια στις τηλεπικοινωνίες, στην μορφή της διαμόρφωσης PCM—Pulse Code Modulation, χρησιμοποιώντας ένα τυπικό ρυθμό μετάδοσης στα 64 Kbits per second. Πρόσφατα, η εκτεινόμενης χωρητικότητας ADPCM—Adaptive Differential PCM διαμόρφωση στα 32 Kbits per second έχει τοποθετηθεί στο τηλεφωνικό δίκτυο, ως επί το πλείστον για ιδιωτικές γραμμές. Οικονομικά συστήματα για φωνητικό ταχυδρομείο έχουν προέλθει από αλγορίθμους συμπίεσης για 16 Kbits per second κωδικοποίηση υποζώνης συχνοτήτων *(Sub-Band)* και μικρής καθυστέρησης γραμμική πρόβλεψη με χρήση κώδικα *(CELP—Code Excited Linear Prediction)* , και η τεχνολογία αυτή—ανεπτυγμένη για ρυθμούς της τάξης των 8 Kbits per second—δοκιμάζεται τώρα σε ψηφιακά κινητά κυψελοειδή τηλέφωνα.

Η ποιότητα του σήματος τυπικά ελαττώνεται με την αύξηση του ρυθμού κωδικοποίησης, με ένα αξιοσημείωτο ακρότατο *("Knee")* περίπου στα 8 Kbits per second. Όμως παρόλα αυτά οι ρυθμοί φωνοκωδικοποιητών στα 4 και 2 Kbits per second χρησιμοποιούνται για ψηφιακή κρυπτογράφηση σε κανάλια με εύρος ζώνης φωνής. Η πρόκληση στην κωδικοποίηση είναι η βελτίωση της ποιότητας σε χαμηλούς ρυθμούς μετάδοσης. Πρόοδος πραγματοποιείται με τον συνδυασμό υπολογίσιμων παραγόντων και με την βελτιωμένη αναπαράσταση των φασματικών παραμέτρων και των παραμέτρων διέγερσης *(excitation)* (Jayant et al., 1990).



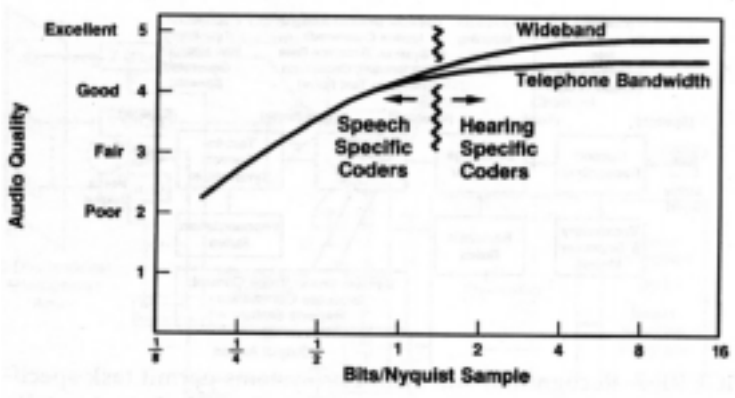
Σχήμα 5: Επίδραση του ρυθμού κωδικοποίησης στην ποιότητα της ομιλίας τηλεφωνικού εύρους ζώνης. Ολοένα και πιο σύνθετοι αλγόριθμοι χρησιμοποιούνται καθώς ο ρυθμός κωδικοποίησης ελαττώνεται. Η ερευνητική προσπάθεια επικεντρώνεται στην βελτίωση της ποιότητας και την ανθεκτικότητα στην παρεμβολή σε ρυθμούς κωδικοποίησης κάτω των 8Kbps.

Υπάρχουν πειραματικοί λόγοι οι οποίοι μας οδηγούν στο να πιστέψουμε ότι η υψηλή ποιότητα μπορεί να επιτευχθεί με ρυθμούς που φτάνουν μέχρι το εύρος των 2000 bps. Βελτιώσεις σε αυτούς τους ρυθμούς μπορούν να προέλθουν από δύο διευθύνσεις: (i) Δυναμική προσαρμογή των παρατηρήσιμων κριτηρίων στην κωδικοποίηση και (ii) Μοντελοποίηση της άρθρωσης του σήματος ομιλίας.

Στην κωδικοποίηση ακουστικών σημάτων ευρείας ζώνης η ευρεία χρήση των ακουστικών παραγόντων αντίληψης στον αλγόριθμο κωδικοποίησης (κωδικοποιητές ακουστικού προσδιορισμού—hearing specific) έχει γίνει αξιοσημείωτα επιτυχής, επιτρέποντας την αναπαράσταση σημάτων ευρείας ζώνης με μέσο ρυθμό μικρότερο των 2 bits ανά δείγμα. Από αυτό συνεπάγεται το γεγονός ότι η ποιότητα εκπομπής

FM stereo μπορεί να μεταδοθεί πάνω από τα ψηφιακά κανάλια τηλεφωνίας (public switched digital telephone channels) που παρέχονταί από τον βασικό ρυθμό της υπηρεσίας ISDN (Integrated Services Digital Network). Εναλλακτικά κάποιος μπορεί να αποθηκεύσει έως και 8 φορές περισσότερο σήμα σε μια υψηλής πιστότητας εγγραφή compact disk σε σχέση με ότι γίνεται συμβατικά.

Για στερεοφωνική κωδικοποίηση το δεξιό σήμα προστίθεται στο αριστερό και επίσης το δεξί αφαιρείται από το αριστερό και δημιουργούνται δύο νέα σήματα τα οποία διαμορφώνονται και κωδικοποιούνται ξεχωριστά (τυπικά από FFT 2048 σημείων). Για κάθε φάσμα, κάθε χρονική στιγμή, υπολογίζεται ένα όριο απόκρυψης βασισμένο στη διασπορά της φασματικής ενέργειας και στην απόκρυψη κρίσιμης ζώνης στο αυτί. Οποιεσδήποτε συνιστώσες σήματος έχουν φασματικά πλάτη μικρότερα αυτού του ορίου δεν θα ακουστούν προς στιγμή στην παρουσία ισχυρότερων γειτόνων και έτσι αυτές οι συνιστώσες δεν χρειάζεται να δεσμεύουν bits για εκπομπή. Ομοίως αν τα bits διατίθενται στις ισχυρότερες συνισταμένες, έτσι ώστε το κβαντισμένο φάσμα θορύβου να συγκεντρώνεται κάτω από αυτό το όριο απόκρυψης, ο κβαντισμένος θόρυβος δεν θα είναι δυνατό να ακουστεί. Ο υπολογισμός που θα ολοκληρώσει την κωδικοποίηση, εφόσον είναι ουσιαστικός, δεν είναι ανέφικτος με την χρήση των μέχρι στιγμής διαθέσιμων chips ψηφιακής επεξεργασίας σήματος.



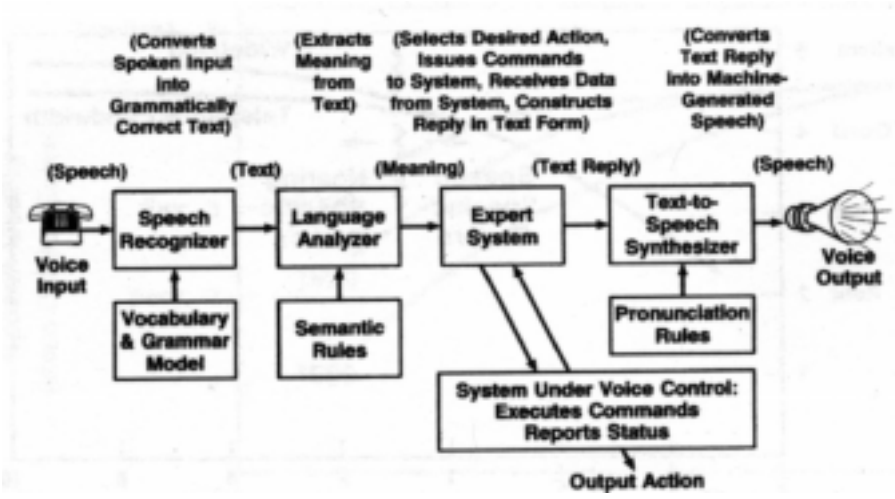
Σχήμα 6:Επίδραση της ψηφιακής αναπαράστασης στην ποιότητα ακουστικού σήματος. Ολοένα και πιο σύνθετοι αλγόριθμοι απαιτούνται καθώς τα bits την αναπαράστασης ανά δείγμα ελαττώνονται. Hearing specific κωδικοποιητές συμπεριλαμβάνουν ανθρώπινους παράγοντες αντίληψης, όπως την απόκρυψη των συχνοτήτων.

Αυτή αλλά και άλλες παρόμοιες τεχνικές επηρεάζουν ισχυρά τα διεθνή standards της κωδικοποίησης ομιλίας και της μουσικής.

Αναγνώριση και σύνθεση. Δυστυχώς η πρόοδος στην αναγνώριση και στην σύνθεση, ειδικά στην σύνθεση ομιλίας από κείμενο, δεν έχουν συνδεθεί αρκετά και δεν αλληλοϋποστηρίζονται ουσιαστικά η μια με την άλλη. Αυτό φαίνεται να οφείλεται κυρίως στο γεγονός ότι η αναγνώριση ακολούθησε μια κατεύθυνση απόκρυψης προτύπων με τα υπερβολικά πετυχημένα μοντέλα Markov (Hidden Markov Models—HMMs), ενώ η σύνθεση βασίστηκε κυρίως στην ακουστική φωνητική με μοντέλα φωνοσυντονισμού (formant) και βιβλιοθήκες βασικών συλλαβών να συνεισφέρουν στην επιτυχία. Παρόλα ταύτα οι τεχνικές αυτές προορίζονται να χρησιμοποιηθούν παράλληλα σε φωνητικά συστήματα αλληλεπίδρασης. Και οι δύο μπορούν να ωφεληθούν από τα βελτιωμένα υπολογιστικά μοντέλα των γλωσσών.

Οι παρούσες δυνατότητες για διάλογο μηχανών επιτρέπουν ευφυή και άπταιστη αλληλεπίδραση από μια μεγάλη ποικιλία ομιλητών παρότι το λεξιλόγιο

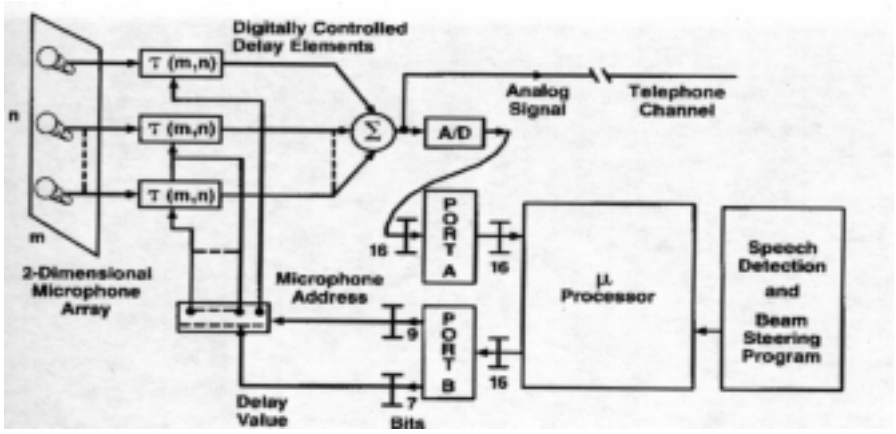
είναι περιορισμένο όπως επίσης το και το πεδίο εφαρμογών (Flanagan 1992). Τυπικά χρησιμοποιείται μια γραμματική περιορισμένων καταστάσεων για να παρέχει αρκετή κάλυψη για μια χρήσιμη διαλογική ανταλλαγή. Λεξιλόγια ενός με δύο εκατοντάδων λέξεων και μια γραμματική η οποία επιτρέπει δισεκατομμύρια προτάσεων για μια συγκεκριμένη εργασία—ας πούμε για την συλλογή πληροφοριών για αεροπορικές πτήσεις—είναι τυπικά. Η ακρίβεια της αναγνώρισης λέξεων είναι πάνω από 90 τοις εκατό για λεξιλόγια μερικών εκατοντάδων λέξεων ομιλουμένων σε συνδεδεμένη μορφή από μια ευρεία ποικιλία ομιλητών. Για μικρότερα λεξιλόγια, όπως είναι τα ψηφία , η ακρίβεια της αναγνώρισης κυμαίνεται επίσης πάνω του 90 τοις εκατό για σύνολα ψηφίων (πχ επταψήφιοι τηλεφωνικοί αριθμοί) ομιλούμενα σε συνδεδεμένη μορφή. Με τα μέχρι στιγμής διαθέσιμα chips επεξεργασίας σήματος το hardware που υποστηρίζει αναγνώριση συνδεδεμένων ψηφίων είναι σχετικά εφικτό.



Σχήμα 7: Τα συστήματα αναγνώρισης και σύνθεσης επιτρέπουν συγκεκριμένης εργασίας διαλογική αλληλεπίδραση. Επεκτάσεις του μεγέθους του λεξιλογίου, η ανεξαρτησία του ομιλητή και τα γλωσσικά μοντέλα που πλησιάζουν ικανοποιητικά την φυσικά ομιλούμενη γλώσσα, μαζί με τη σύνθεση υψηλής ποιότητας είναι ερευνητικοί στόχοι.

Πάλι ένα ουσιαστικό όριο παρουσιάζεται στην ανάπτυξη γλωσσικών υπολογιστικών μοντέλων τα οποία καλύπτουν περισσότερο φυσικά την γλώσσα και επιτρέπουν απεριόριστη αλληλεπίδραση. Η υπολογιστική γλωσσολογία μπορεί να συνεισφέρει αρκετά σε αυτόν τον τομέα.

Επαλήθευση της ταυτότητας του ομιλητή. Χρησιμοποιώντας cepstrum (το αποτέλεσμα του μετασχηματισμού Fourier του λογαριθμικού φάσματος), δέλτα cepstrum και HMM τεχνικές η δυνατότητα του ελέγχου των εγγεγραμμένων ομιλητών πάνω από ένα καθαρό κανάλι είναι σχετικά καλά καθορισμένη (Soong and Rosenberg 1988). Ο υπολογισμός που χρειάζεται υποστηρίζεται εύκολα αλλά ακόμα δεν έχει παρουσιαστεί ικανοποιητική εμπορική ανάπτυξη. Αυτό δεν προκαλείται από έλλειψη επιθυμίας για χρήση αυτής της δυνατότητας αλλά από μια μικρή σχετικά πρόθεση του καταναλωτικού κοινού να πληρώσει για αυτή. Επειδή η αναγνώριση ομιλίας και η επαλήθευση του ομιλητή χρησιμοποιούν κοινές διεργασίες είναι φυσικό να συνδυάζονται σε ένα interface. Η επένδυση στην αναγνώριση λοιπόν μπορεί να παρέχει και επαλήθευση με μια στοιχειώδη αύξηση του κόστους. Νέες υπηρεσίες αυτού του τύπου εμφανίζονται στον τραπεζικό τομέα, όπου η επαλήθευση των πελατών είναι απαραίτητη για υπηρεσίες όπως οι αυτόματες τραπεζικές μηχανές ανάληψης και κατάθεσης χρημάτων.



Σχήμα 8α: Πίνακες μικροφώνων τροποποίησης ακτίνας και αναγνώρισης σήματος επιτρέπουν φυσική επικοινωνία χωρίς την χρήση φορητών μικροφώνων.

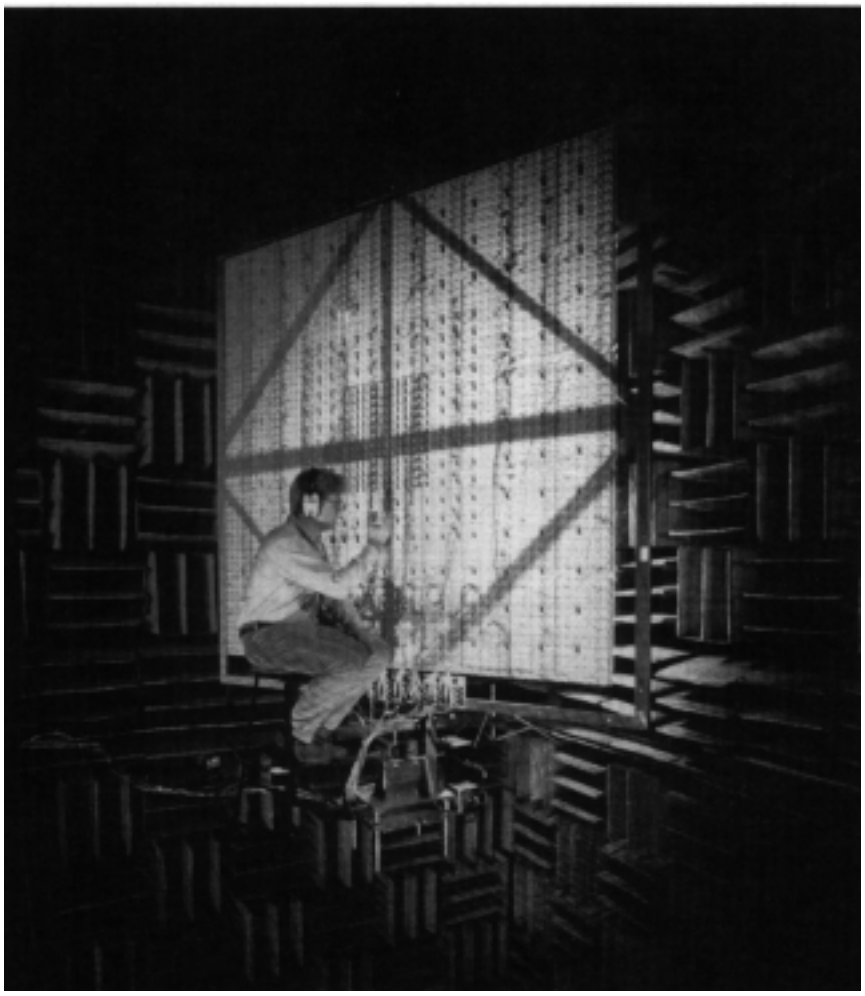
Αυτοκατευθυνόμενοι πίνακες μικροφώνων. Σε πολλά περιβάλλοντα επικοινωνίας ομιλίας, ειδικά στην τηλεδιάσκεψη και στην χρήση τερματικών φωνητικής αλληλεπίδρασης, είναι πιο φυσικό να επικοινωνούμε χωρίς μικρόφωνα που θα πρέπει να φοράμε ή να κρατάμε. Η ελευθερία κινήσεων μέσα στον χώρο εργασίας χωρίς να είμαστε δεμένοι ή να παρεμποδιζόμαστε και η δυνατότητα να μιλάμε σαν σε διάλογο πρόσωπο με πρόσωπο είναι συνήθως προτέρημα. Οι αυτοκατευθυνόμενοι πίνακες μικροφώνων, συγκεκριμένα τα συστήματα μορφοποίησης ακτινών, επιτρέπουν καλής ποιότητας σύλληψη ήχου και εξομαλύνουν τα αποτελέσματα της ηχούς που εμφανίζεται κατά την ομιλία μέσα σε δωμάτια και του παρεμβαλλόμενου ακουστικού θορύβου (Flanagan et al.,1991).

Τα υψηλής απόδοσης και χαμηλού κόστους μικρόφωνα ηλεκτρίτη, σε συνδυασμό με τους οικονομικά κατανεμημένους επεξεργαστές σήματος, κάνουν τους μεγάλους πίνακες αναζήτησης ομιλίας πρακτικούς. Κάθε αισθητήρας μπορεί να έχει ένα επεξεργαστή αφοσιωμένο σε αυτόν για να πραγματοποιήσει την μορφοποίηση της ακτίνας και την καθοδήγηση. Ένας ενσωματωμένος ελεγκτής προσφέρει κατάλληλες τιμές μορφοποίησης και εντοπισμού ακτίνας σε κάθε αισθητήρα και υποστηρίζει αλγορίθμους για εύρεση πηγής θορύβου και αναγνώριση ομιλίας /παύσης. Ο πίνακας τυπικά χρησιμοποιείται με πολλαπλές ακτίνες σε έναν track-while-scan τύπο λειτουργίας. Νέες έρευνες πάνω σε πίνακες τριών διαστάσεων και μορφοποίηση πολλαπλών ακτινών οδηγεί σε υψηλής ποιότητας καταγραφή του σήματος από εκλεγμένους χωρικούς όγκους.

2.5 ΣΗΜΑΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ ΤΗΣ ΕΡΕΥΝΑΣ ΟΜΙΛΙΑΣ

2.5.1. Φυσική της παραγωγής ομιλίας. Αρχές υδροδυναμικής

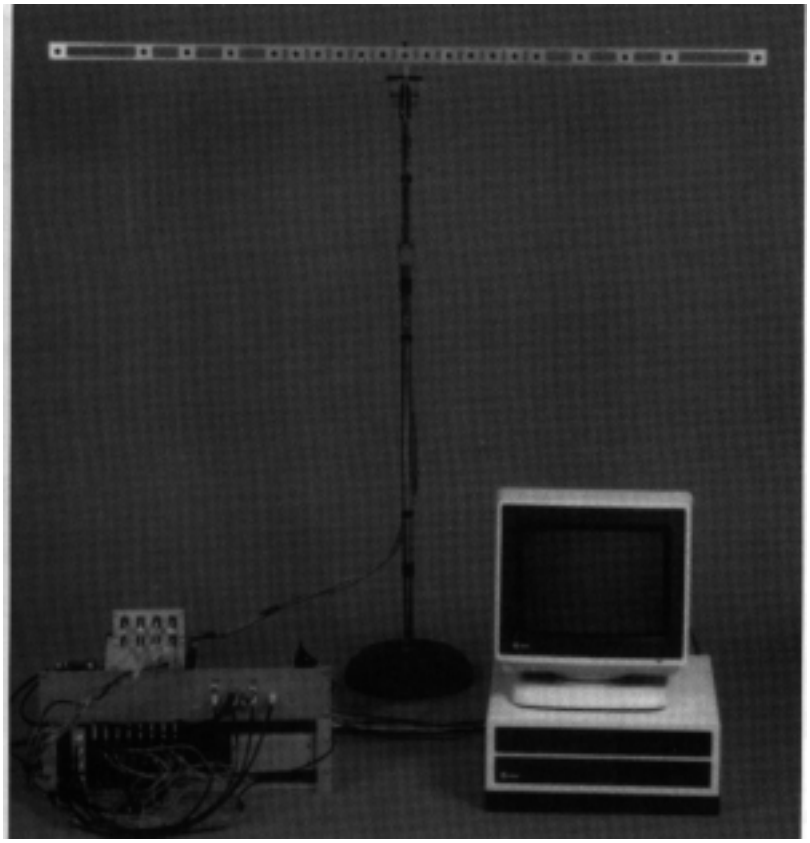
Η προαναφερθείσα έλλειψη φυσικότητας στην ομιλία που παράγεται από αυστηρούς κανόνες προέρχεται από δύο δυνατές πηγές. Η μία είναι η ακατέργαστη προσέγγιση των ακουστικών ιδιοτήτων του φωνητικού συστήματος από τον συνθέτη ομιλίας. Η άλλη είναι τα λάθη στα δεδομένα ελέγχου που δεν αντανακλούν ικανοποιητικά την φυσική άρθρωση και την προσωδία. Κάθε μία από τις δύο παραπάνω περιπτώσεις επηρεάζει την ποιότητα της ομιλίας και την δυνατότητα αναπαραγωγής συγκεκριμένων χαρακτηριστικών φωνής.



Σχήμα 8β: Μεγάλος δυσδιάστατος πίνακας αποτελούμενος από 408 μικρόφωνα ηλεκτρίτη. Κάθε μικρόφωνο έχει ένα chip αποκλειστικά δικό του για την μορφοποίηση ακτίνας.

Η παραδοσιακή σύνθεση λαμβάνει ως σημείο αναφοράς μια προσέγγιση πηγής—φίλτρου του φωνητικού συστήματος στην οποία η πηγή και το φίλτρο δεν αλληλεπιδρούν. Τυπικά η λειτουργία του φίλτρου προσεγγίζεται από ένα σωλήνα με σκληρά τοιχώματα ο οποίος επιτρέπει μονοδιάστατη διάδοση του κύματος. Όμως ούτε αυτό είναι ρεαλιστικό.

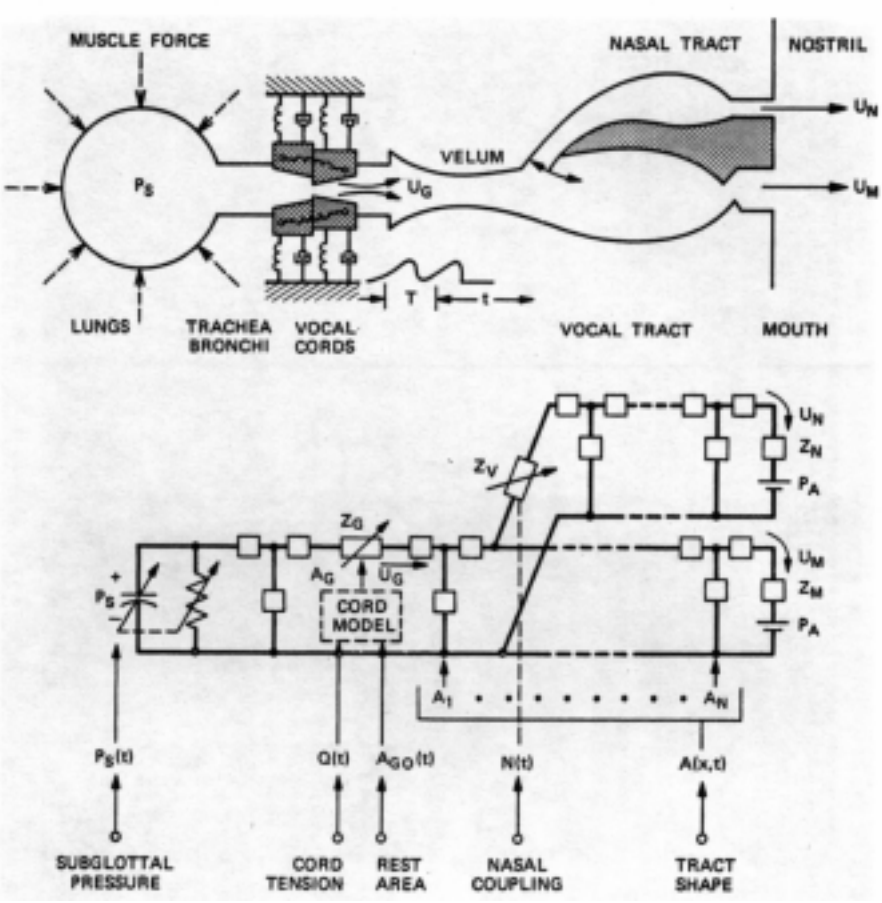
Η πρόοδος στους παράλληλους υπολογισμούς ανοίγει τον δρόμο για πραγματοποίηση σύνθεσης ομιλίας από τις βασικές αρχές της υδροδυναμικής. Δοθείσας μιας τρισδιάστατης, χρονικά μεταβαλλόμενης φωνητικής οδού με μαλακά τοιχώματα, η οποία διεγείρεται από περιοδική βαλβοειδή ροή στις φωνητικές χορδές και από τυρβώδη ροή στις συσφίγξεις, η εξίσωση Navier—Stokes μπορεί να λυθεί αριθμητικά σε ένα καθορισμένο διάγραμμα (grid) χώρου—χρόνου για να παράγει μια αξιοσημείωτη ρεαλιστική περιγραφή της εκπεμπόμενης ηχητικής πίεσης. Οι μη γραμμικότητες της διέγερσης, η παραγωγή της τύρβης, οι διασταυρωμένοι τύποι (cross-modes) του συστήματος και η ακουστική αλληλεπίδραση ανάμεσα στις πηγές και τα αντηχεία (resonators) λαμβάνονται υπόψη. Η μορφοποίηση απαιτεί πληθώρα υπολογισμών, αλλά οι τωρινές συνθήκες στην υπολογιστική υψηλής απόδοσης υπόσχονται τις απαραίτητες δυνατότητες.



Σχήμα 8γ: Μονοδιάστατος track-while-scan μορφοποιητής ακτίνας για μικρά δωμάτια συνδιασκέψεων.

2.5.2. Υπολογιστικά Μοντέλα Γλωσσών

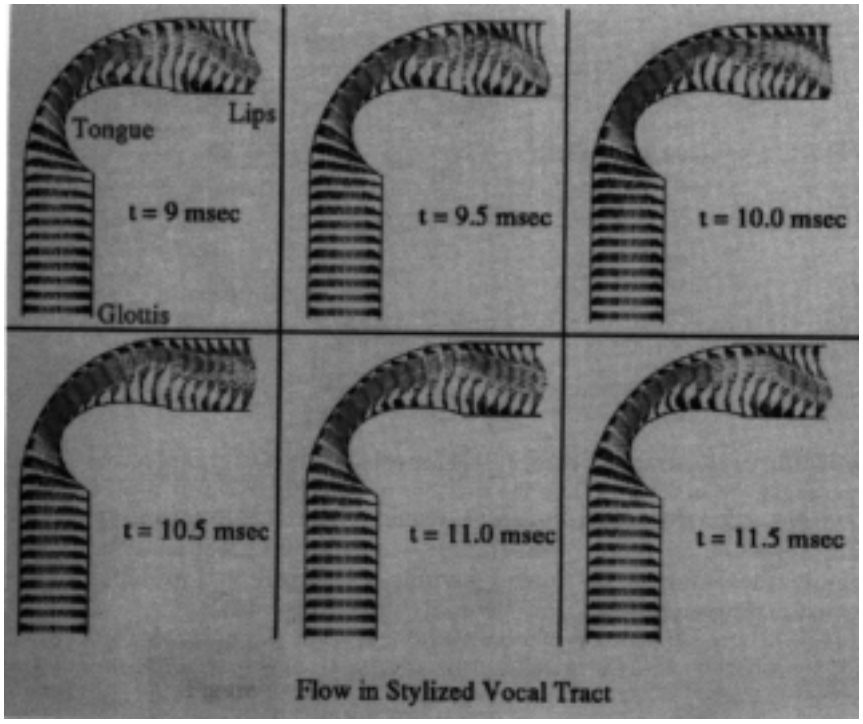
Έχει ήδη αναφερθεί η σημαντικότητα των γλωσσικών μοντέλων για την άψογη, με μεγάλα λεξιλόγια αναγνώριση ομιλίας. Ελέγξιμα μοντέλα τα οποία είναι υπεύθυνα για την γραμματική (στην ομιλούμενη γλώσσα), την συντακτική και την σημασιολογική συμπεριφορά χρειάζονται επειγόντως τόσο για την σύνθεση από κείμενο όσο και για την αναγνώριση. Οι στατιστικοί περιορισμοί στην προφορική γλώσσα είναι τόσο ισχυροί όσο και αυτοί στην γραπτή και μπορούν να χρησιμοποιηθούν για να συμπληρώσουν ικανοποιητικά τις παραδοσιακές προσεγγίσεις της ανάλυσης και καθορισμού τμημάτων της ομιλίας.



Σχήμα 9α: Παραδοσιακή αναπαράσταση της γέννησης και διάδοσης του ήχου στην φωνητική οδό. Η μονοδιάστατη προσέγγιση της διάδοσης του ήχου επιτρέπει τον υπολογισμό των διασπορών της ταχύτητας και της πίεσης κατά μήκος της οδού και στα εκπέμποντα σημεία. Η τυρβώδης διέγερση υπολογίζεται από τον αριθμό του Reynolds σε κάθε τοποθεσία κατά μήκος της οδού. Η εξομοίωση των φωνητικών χορδών επιτρέπει την αλληλεπίδραση πηγής-φίλτρου.

2.5.3. Επεξεργασία πληροφορίας στο ακουστικό σύστημα. Ακουστική συμπεριφορά

Η μηχανική και η λειτουργία του περιφερειακού αφτιού είναι σχετικά καλά κατανοητές. Η ψυχοακουστική συμπεριφορά είναι εκτενώς μετρημένη. Οι λεπτομέρειες για την νευρωνική επεξεργασία, και ο μηχανισμός για την ανάκτηση της νευρωνικής πληροφορίας δεν είναι καλά καθορισμένες. Αυτό όμως δεν εμποδίζει την εποικοδομητική χρήση αυτών των παραγόντων συμπεριφοράς στην επεξεργασία ομιλίας. Στο παρελθόν η τηλεπικοινωνιακή και η ακουστική τεχνολογία έχουν αναπτύξει κύριους τομείς της ανθρώπινης ακοής, όπως τα εύρη συχνοτήτων, τα πλάτη και οι λόγοι σήματος προς θόρυβο.

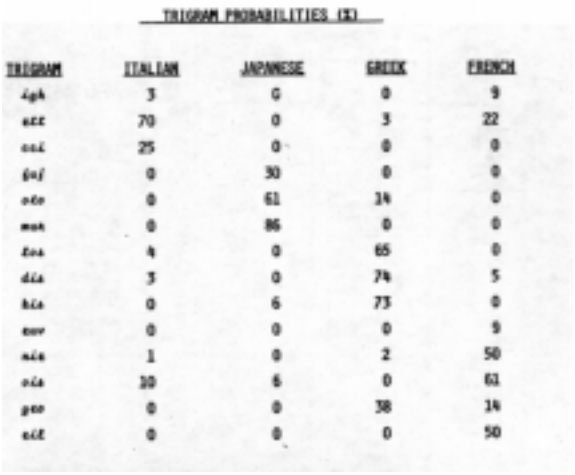


Σχήμα 9β: Η παραγωγή του ήχου στην φωνητική οδό υπολογίζεται με βάση τις αρχές της υδροδυναμικής. Το πλάτος και η διεύθυνση του διανύσματος της ταχύτητας σε κάθε σημείο στις δύο διαστάσεις, εξαιτίας ενός βήματος ακτινικής ταχύτητας στις φωνητικές χορδές, υπολογίζονται σε ένα υπερυπολογιστή. Τα θερμά χρώματα υποδεικνύουν πλάτους υψηλής ταχύτητας. Το διάγραμμα δείχνει διαχωρισμό ροής κατά την διεύθυνση (downstream) της σύσφιξης της γλώσσας και nonplanar wavefronts.

Τώρα όμως, με ανέξοδους υπολογισμούς, βοηθητικά τεχνάσματα μπορούν να ενσωματωθούν στην αναπαράσταση ακουστικών σημάτων. Ήδη η υψηλής ακριβείας ακουστική κωδικοποίηση συμπεριλαμβάνει μερικούς περιορισμούς λόγω της ταυτόχρονης απόκρυψης στο πεδίο των συχνοτήτων. Η απόκρυψη στον χρόνο είναι ένας φανερός μελλοντικός στόχος. Σχετικά ανέπαφη μέχρι στιγμής είναι η εσωτερική συμπεριφορά της binaural απελευθέρωσης των δύο αυτιών (binaural) από την απόκρυψη, ενώ η φάσης μεταξύ των δύο αυτιών (interaural) ελέγχει αξιοσημείωτα την παρατηρησιμότητα.

2.5.4. Ομαδοποίηση Κωδικοποίησης, Σύνθεσης και Αναγνώρισης Ομιλίας

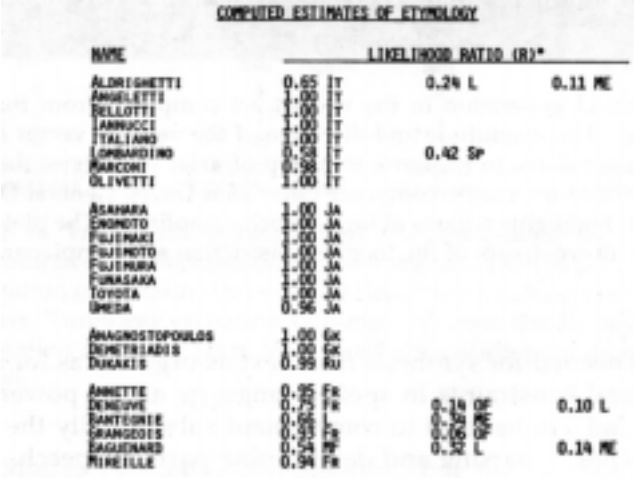
Τα θέματα της κωδικοποίησης, της αναγνώρισης και της σύνθεσης δεν είναι ασυσχέτιστα—είναι πλευρές της ίδιας βασικής διαδικασίας της ομιλίας και της ακοής. Συνεπώς μπορούμε να πασχίσουμε για μία έρευνα η οποία θα ενοποιεί τα θέματα αυτών των διαφορετικών τομέων. Ακόμα καλύτερα, μπορούμε να αναζητήσουμε μια προσέγγιση η οποία θα ομαδοποιεί αυτά τα προβλήματα σε μια κοινή θεωρία. Μία τέτοια προσπάθεια είναι η «μίμηση φωνής».



TRIGRAM PROBABILITIES (%)

TRIGRAM	ITALIAN	JAPANESE	GREEK	FRENCH
[illegible]	3	0	0	9
[illegible]	70	0	3	22
[illegible]	25	0	0	0
[illegible]	0	30	0	0
[illegible]	0	61	14	0
[illegible]	0	86	0	0
[illegible]	4	0	65	0
[illegible]	3	0	74	5
[illegible]	0	6	73	0
[illegible]	0	0	0	9
[illegible]	1	0	2	50
[illegible]	10	6	0	61
[illegible]	0	0	38	14
[illegible]	0	0	0	50

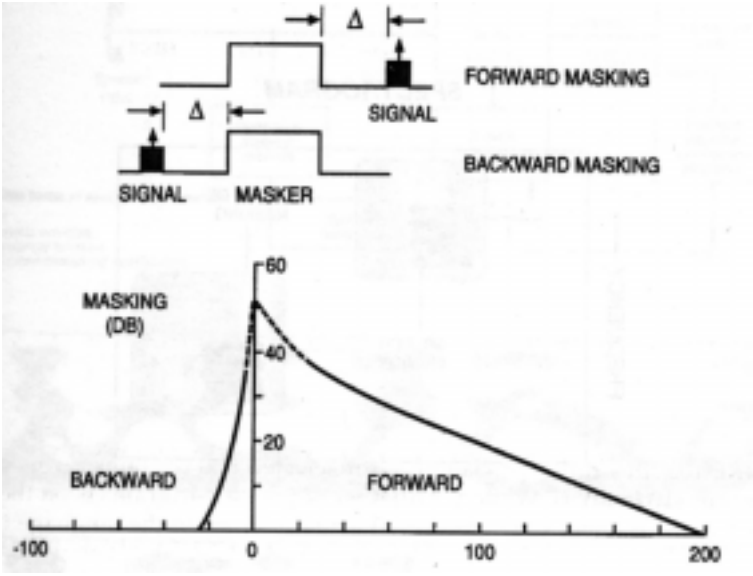
Σχήμα 10α: Απεικόνιση πιθανοτήτων για επιλεγμένα κειμένων τριών γραμμάτων (text trigrams) για αρκετές γλώσσες (συνολικά 10). Όσο ο αριθμός των πιθανών τριγραμμάτων είναι της τάξης των 20000, ο αριθμός των τριγραμμάτων που εμφανίζονται ουσιαστικά στην γλώσσα είναι τυπικά μικρότερος από μία τάξη μεγέθους—συντάσσοντας μεγάλη επιρροή στην εκτίμηση επιτρεπτών σειρών συμβόλων σε μία γλώσσα και παρέχοντας ένα εργαλείο για τη εκτίμηση της ετοιμολογίας από τις ανεξάρτητες πιθανότητες.



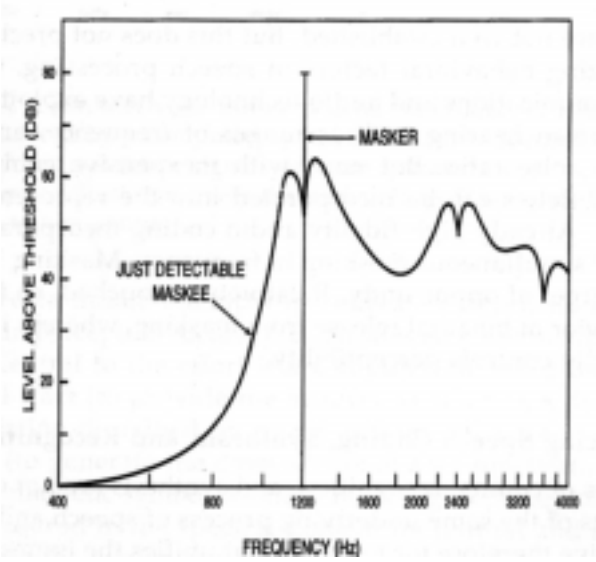
COMPUTED ESTIMATES OF ETYMOLOGY

NAME	LIKELIHOOD RATIO (R)*		
ALDRIGHETTI	0.65 IT	0.24 L	0.11 ME
ANGELETTI	1.00 IT		
BELLOTTI	1.00 IT		
IANNUCCI	1.00 IT		
ITALIANO	1.00 IT		
LOMBARDINO	0.58 IT	0.42 SP	
MARCONI	0.98 IT		
OLIVETTI	1.00 IT		
ASAHARA	1.00 JA		
ENOMOTO	1.00 JA		
FUJIMAKI	1.00 JA		
FUJIMOTO	1.00 JA		
FUJIMURA	1.00 JA		
FUNASAKA	1.00 JA		
TOYODA	1.00 JA		
UMEDA	0.96 JA		
ANAGNOSTOPOULOS	1.00 Gk		
DEMETRIADES	1.00 Gk		
DUKAKIS	0.99 Ru		
ANNETTE	0.95 Fr		
DENEUVE	0.72 Fr	0.14 OF	0.10 L
BANTEGNIE	0.66 L	0.[illegible] [illegible]	
GRANGEOIS	0.[illegible] Fr	0.[illegible] [illegible]	
BAGUENARD	0.54 MF	0.32 L	0.14 ME
MIREILLE	0.94 Fr		

Σχήμα 10β: Παραδείγματα εκτιμήσεων ετοιμολογίας για κατάλληλες ονομασίες. Η εκτίμηση βασίζεται στο λόγο πιθανοτήτων (ο λόγος της πιθανότητας το όνομα να ανήκει σε μία γλώσσα j, προς την μέση πιθανότητα του ονόματος σε όλες τις γλώσσες). Οι γλώσσες που περιλαμβάνονται είναι αγγλικά, γαλλικά, γερμανικά, Ιαπωνικά, ελληνικά, ρώσικα, σουηδικά, ισπανικά, ιταλικά και λατινικά.



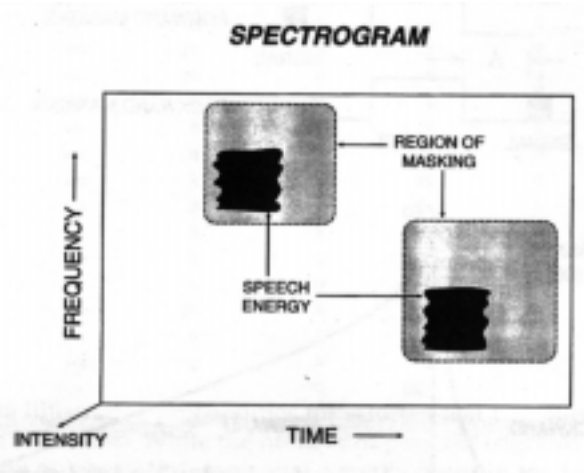
Σχήμα 11α: Απόκρυψη στον χρόνο. Ένας ισχυρός ήχος πριν και μετά από ένα ασθενέστερο μπορεί να αυξήσει το όριο ανιχνευσιμότητας του τελευταίου.



Σχήμα 11β: Απόκρυψη στη συχνότητα. Ένας ισχυρός τόνος (στα 1200 Hz εδώ) μπορεί να ανεβάσει το όριο ανιχνευσιμότητας ενός γειτονικού τόνου, ειδικά κάποιου με υψηλότερη συχνότητα.

Η μίμηση φωνής προσπαθεί να δημιουργήσει ένα συνθετικό σήμα ομιλίας το οποίο, με αξιοσημείωτη ακρίβεια, αναπαράγει μία είσοδο αυθαίρετης φυσικής ομιλίας. Κεντρικό σημείο σε αυτή την προσπάθεια είναι ένα υπολογιστικό μοντέλο των φωνητικών χορδών και της φωνητικής οδού (για να παρέχει την δυνατότητα ακουστικής σύνθεσης), ένα δυναμικό μοντέλο άρθρωσης περιγραφόμενο από σχεδόν ορθογώνιες παραμέτρους μορφής της φωνητικής οδού (για να παράγει την διατμηματικής περιοχής συνάρτηση) και, ιδανικά, μια διακριτή φωνητική απεικόνιση συμβόλου—σε—σχήμα (symbol—to—shape mapping). Ένα προφανώς ζυγισμένο λάθος (weighted error), μετρημένο στο φασματικό πεδίο για φυσικά και συνθετικά

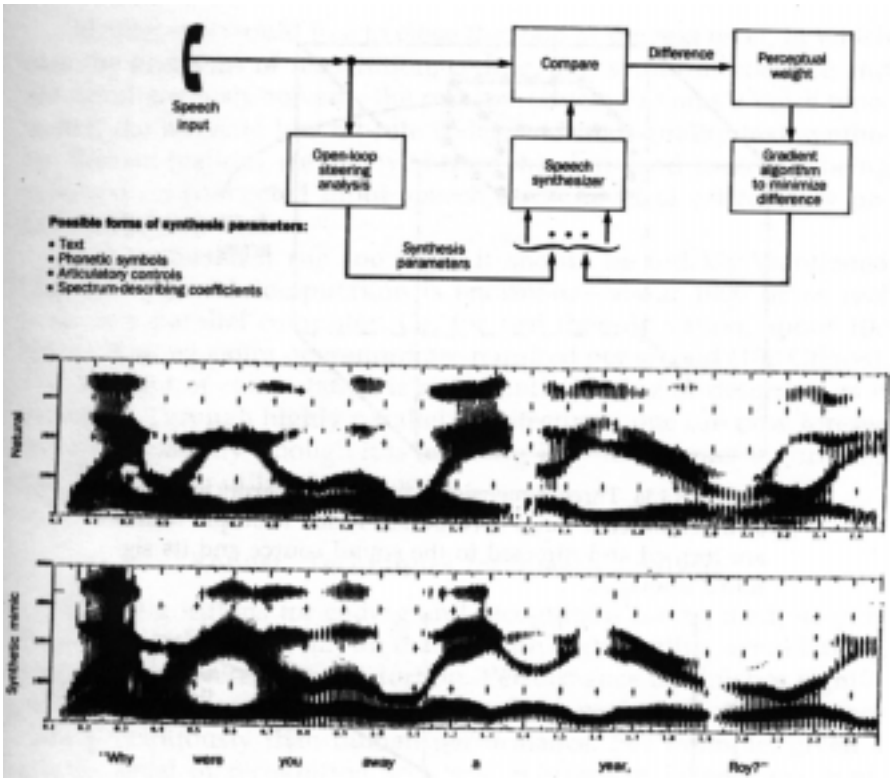
σήματα, οδηγεί τις παραμέτρους της σύνθεσης έτσι ώστε να ελαχιστοποιείται το λάθος μίμησης, από στιγμή σε στιγμή (moment by moment). Η ανάλυση ανοιχτού βρόχου της εισερχόμενης φυσικής ομιλίας είναι χρήσιμη στην καθοδήγηση της πραγματοποίησης κλειστού βρόχου.



Σχήμα 11γ: Απεικόνιση της περιοχής χρόνου-συχνότητας που περικυκλώνει έντονα σήματα στίξης όταν απόκρυψη στο χρόνο και στη συχνότητα λαμβάνουν χώρα.

Ιδανικά, κάποιος μπορεί να θέλει να κλείσει τον βρόχο στο επίπεδο κειμένου και σε αυτή την περίπτωση τα προβλήματα της αναγνώρισης, της κωδικοποίησης και της σύνθεσης ομαδοποιούνται και λύνονται εξομοιωτικά (simultaneously solved)—το αποτέλεσμα περιγράφει σαν ένα μια γραφομηχανή, τον καλύτερο κωδικοποιητή χαμηλού ρυθμού μετάδοσης δεδομένων (low bit rate) και υψηλής ποιότητας σύνθεση κειμένου. Η παρούσα κατάσταση είναι απομακρυσμένη από αυτό, αλλά καλά αποτελέσματα επιτυγχάνονται σε συνδεδεμένη είσοδο ομιλίας στο επίπεδο της ρύθμισης των παραμέτρων της άρθρωσης.

Προτού όμως αυξηθεί ο ενθουσιασμός, θα αναφερθεί συντόμως ότι οι απαραίτητοι υπολογισμοί είναι πολυπληθείς—περίπου 100 φορές πραγματικού χρόνου (100 times real time) σε ένα παράλληλο υπολογιστή. Ή αλλιώς, για μία λειτουργία πραγματικού χρόνου απαιτούνται περίπου 100 δισεκατομμύρια πράξεις κινητής υποδιαστολής ανά δευτερόλεπτο (100 Gflops). Αυτό το μέγεθος των υπολογισμών δεν είναι τόσο τρομακτικό και τόσο αποτρεπτικό όσο ήταν κάποτε. Με την χρήση αρχιτεκτονικών μεγάλου βαθμού παραλληλίας μπορεί πλέον να προβλεφθεί ικανότητα τετράκις εκατομμυρίων πράξεων ανά δευτερόλεπτο teraflops (αν και δεν είναι τόσο εμφανές πως να οργανώσουμε τους αλγορίθμους και το λογισμικό ώστε να αξιοποιηθεί αυτή η ισχύς).

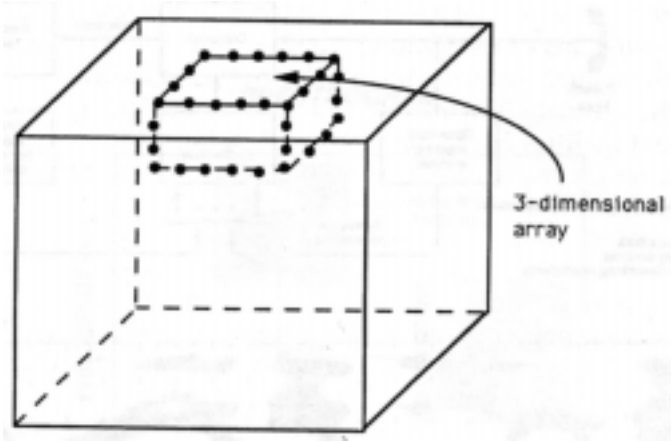


Σχήμα 12: Σύστημα μίμησης φωνής με υπολογιστή. Η φυσική συνεχόμενη είσοδος ομιλίας προσεγγίζεται από μία υπολογισμένη σύνθετη εκτίμηση. Οι φασματικές διαφορές μεταξύ του πραγματικού και του σύνθετου σήματος επισυνάπτονται με βάρη και χρησιμοποιούνται σε ένα κλειστό βρόχο για να ρυθμιστούν οι παράμετροι της σύνθεσης, κάνοντας με επαναλήψεις την διαφορά ελάχιστη.

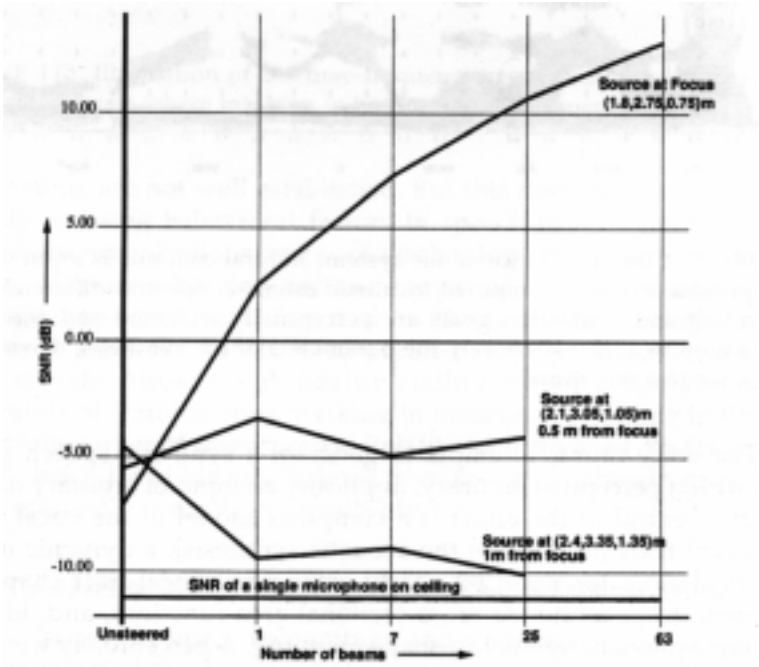
2.5.5. «Εύρωστες» Τεχνικές Ανάλυσης Ομιλίας

Οι περισσότεροι αλγόριθμοι για κωδικοποίηση και αναγνώριση μπορούν να κατασκευαστούν ώστε να λειτουργούν καλά με «καθαρή» είσοδο, δηλαδή με υψηλής ποιότητας σήμα το οποίο έχει ασήμαντη παρεμβολή ή αλλοίωση. Η απόδοση μειώνεται ουσιαστικά με υποβαθμισμένη είσοδο. Και η απόδοση της μηχανής μειώνεται πιο απότομα σε σχέση με την ανθρώπινη απόδοση. Για παράδειγμα, δοθέντος ενός συγκεκριμένου επιπέδου ακρίβειας αναγνώρισης, ο άνθρωπος—ακροατής μπορεί τυπικά να κατορθώσει να φτάσει αυτό το επίπεδο με λόγους σήματος προς θόρυβο στην είσοδο οι οποίοι είναι 10 με 15 dB μικρότεροι από ότι είναι στα τυπικά αυτόματα συστήματα.

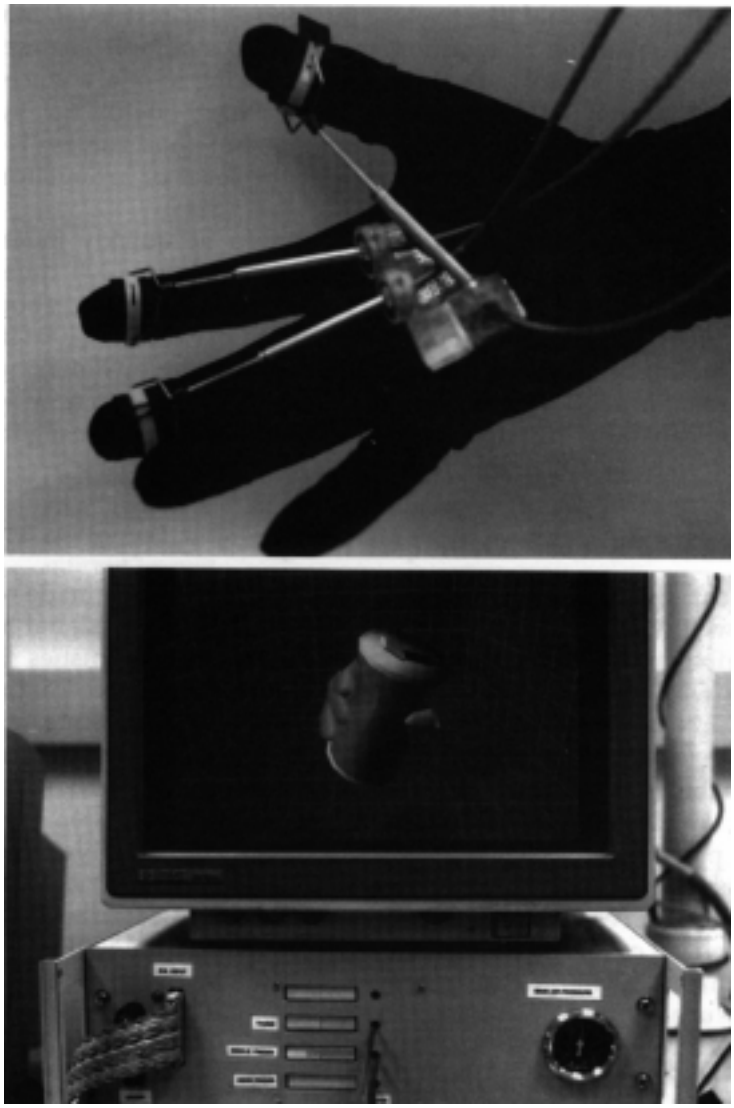
Ένα τμήμα αυτού του προβλήματος εμφανίζεται να είναι η γραμμική ανάλυση η οποία χρησιμοποιείται στις περισσότερες επεξεργασίες. Η γραμμική κωδικοποίηση πρόβλεψης, για να εκτιμηθούν φάσματα μικρής χρονικής διάρκειας, είναι αντιπροσωπευτική. Ικανοποιητικού μεγέθους διάρκειες των σημάτων συνεισφέρουν στον υπολογισμό των τιμών της συνδιακύμανσης, έτσι ώστε ένας μεγάλος αριθμός σημάτων μολυσμένων με θόρυβο να χρησιμοποιείται για τον υπολογισμό του μέσου όρου στην ανάλυση. Μία εναλλακτική διαδικασία η οποία παρουσιάζει ενδιαφέρον, προς το παρόν, είναι η απομάκρυνση των περισσότερο μολυσμένων με θόρυβο σημάτων και η επανασύνταξη των ξεσκαρταρισμένων δειγμάτων από ένα μη γραμμικό αλγόριθμο παρεμβολής. Μία άλλη διαδικασία είναι η χρήση ακουστικών μοντέλων με την χρήση basilar membrane filtering και νευρωνικής μετατροπής ενέργειας (neural transduction) για χαρακτηρισμό των χαρακτηριστικών του σήματος.



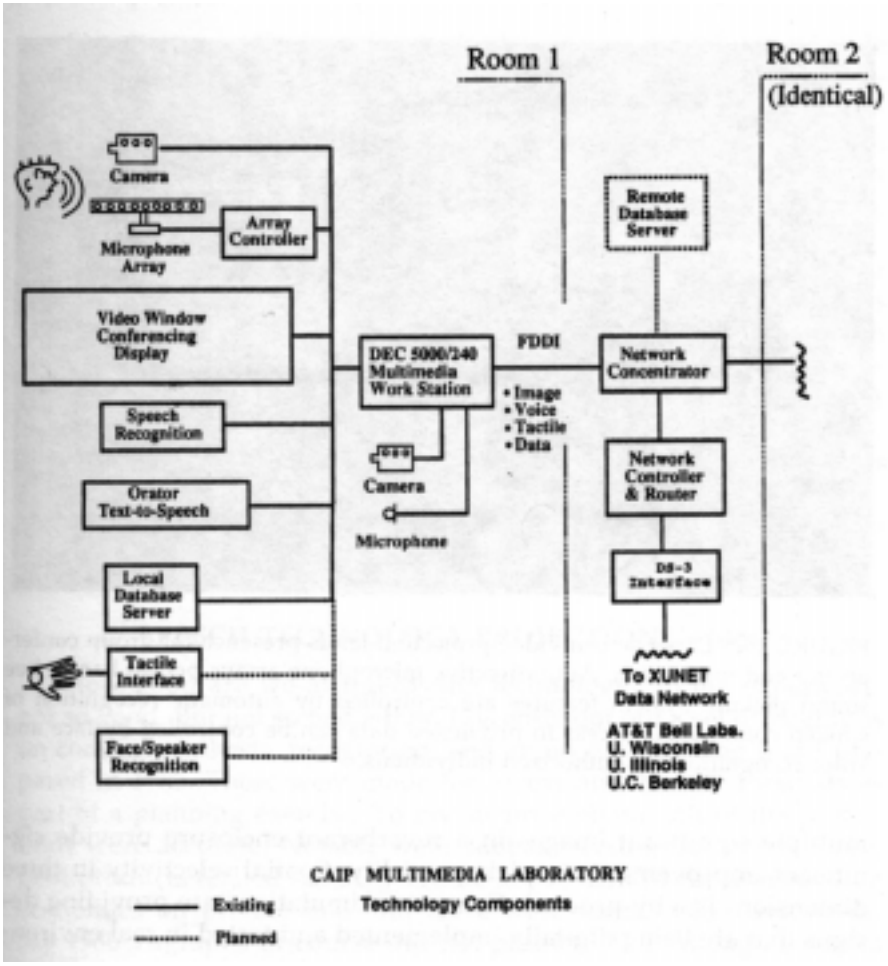
Σχήμα 13α: τρισδιάστατος πίνακας μικροφώνων τοποθετημένος σαν πολύφωτο σε ένα δωμάτιο με ηχώ. Πολλαπλές ακτίνες μορφοποιούνται και κατευθύνονται στην πηγή του ήχου και στα σημαντικότερα είδωλα του.



Σχήμα 13β: Πηλίκα σήματος προς θόρυβο για δύο οκτάβες ομιλίας για ένα 7x7x7 τετραγωνικό πίνακα μικροφώνων τοποθετημένο στο κέντρο του ταβανιού για ένα δωμάτιο με σκληρά τοιχώματα που εξομοιώνεται από υπολογιστή, διαστάσεων 7x5x3 μέτρων. Τα είδωλα της πηγής μέσω τρίτης τάξης υπολογίζονται και πολλαπλές ακτίνες κατευθύνονται στην πηγή και στα είδωλα της



Σχήμα 14: Εξαναγκασμένη ανατροφοδότηση applique για ένα VPL γάντι δεδομένων στο CAIP Ceter. Χρησιμοποιώντας το γάντι εξαναγκασμένης ανατροφοδότησης, αυτός που το φορά μπορεί να υπολογίσει ένα φανταστικό αντικείμενο, και να αισθανθεί με την αφή την σχετική θέση του αντικειμένου και την προγραμματισμένη συμβατότητα του(πάνω). Με την χρήση του γαντιού εξαναγκασμένης ανατροφοδότησης , ο χρήστης δημιουργεί και αισθάνεται την πλαστική παραμόρφωση ενός εικονικού κουτιού αναψυκτικού.



Σχήμα 15: Εξαναγκασμένη ανατροφοδότηση applique για ένα VPL γάντι δεδομένων στο CAIP Ceter. Χρησιμοποιώντας το γάντι εξαναγκασμένης ανατροφοδότησης, αυτός που το φορά μπορεί να υπολογίσει ένα φανταστικό αντικείμενο, και να αισθανθεί με την αφή την σχετική θέση του αντικειμένου και την προγραμματισμένη συμβατότητα του(πάνω). Με την χρήση του γαντιού εξαναγκασμένης ανατροφοδότησης , ο χρήστης δημιουργεί και αισθάνεται την πλαστική παραμόρφωση ενός εικονικού κουτιού αναψυκτικού.

2.5.6. Τρισδιάστατη Σύλληψη Ήχου και Προβολή

Τα υψηλής ποιότητας και χαμηλού κόστους μικρόφωνα ηλεκτρίτη καθώς και οι οικονομικοί ψηφιακοί επεξεργαστές σήματος επιτρέπουν την χρήση μεγάλων πινάκων μικροφώνων για hands—free σύλληψη ήχου σε εχθρικά ακουστικά περιβάλλοντα. Ακόμα περισσότερο τριών διαστάσεων πίνακες με καθοδήγηση ακτίνας (beam steering) στην ηχητική πηγή και πολλαπλές σημαντικές εικόνες σε μία αντηχητική περίφραξη (enclosure) παρέχουν ουσιαστικές βελτιώσεις στην ποιότητα σύλληψης (pickup quality). Η χωρική επιλεκτικότητα στις τρεις διαστάσεις είναι ένα υποπροϊόν. Οι προσομοιώσεις υπολογιστών παρέχουν σχέδια τα οποία εφαρμόζονται και δοκιμάζονται ψηφιακά σε πραγματικά περιβάλλοντα.

Ο σχεδιασμός των πινάκων—δεκτών είναι όμοιος αυτόν των πινάκων εκπομπής (ή προβολής) ήχου—αν και το κόστος των transducers για λήψη και εκπομπή διαφέρουν πολύ. Αυξημένος χωρικός ρεαλισμός στην προβολή ήχου θα προκύψει από αυτή την νέα κατανόηση.

2.5.7. Ολοκλήρωση των Τρόπων Αίσθησης για Όραση, Ήχο και Ακοή

Η ικανότητα του ανθρώπου να αφομοιώνει πληροφορίες, να τις αντιλαμβάνεται και να αντιδρά είναι τυπικά περισσότερο περιορισμένη σε αναλογία με τις ικανότητες εκπομπής οι οποίες μεταφέρουν πληροφορίες στο τερματικό του χρήστη. Η εξέλιξη της παγκόσμιας ψηφιακής μεταφοράς από άκρη σε άκρη θα μεγαλώσει αυτή την διαφορά και θα τονίσει την ανάγκη να αναζητήσουμε τους βέλτιστους τρόπους για να ταιριάξουμε τις επιδείξεις πληροφορίας με την ανθρώπινη ικανότητα επεξεργασίας.

Ταυτόχρονα επιδείξεις για πολλούς τρόπους αίσθησης παρέχει οφέλη εάν αυτά μπορούν να ενορχηστρωθούν καταλλήλως. Οι τρόποι αίσθησης άμεσου ενδιαφέροντος είναι η όραση, ο ήχος και η αφή. Η γνώση μας για τα πρώτα δύο είναι πιο προηγμένη από ότι για το τελευταίο, αλλά νέες μέθοδοι για μετατροπείς ενέργειας με εξαναγκασμένη ανάδραση (force feedback transducers) σε γάντια δεδομένων (data gloves) και οι εφαρμογές έξυπνου δέρματος (smart skin) φιλοδοξούν να προάγουν την τεχνολογία αφής (Flanagan, in press).

Η ευκολία χρήσης συνδέεται απευθείας με επιτυγχυμένη ολοκλήρωση πολλαπλών αισθητηρίων καναλιών. Από την πλευρά της τεχνολογίας ομιλίας αυτό σημαίνει ολοκλήρωση μέσα στο πληροφοριακό σύστημα των στοιχειωδών τμημάτων για αναγνώριση, σύνθεση και επαλήθευση ομιλίας, χαμηλού ρυθμού δεδομένων κωδικοποίηση και hands—free σύλληψη ήχου. Οι αρχικές προσπάθειες σε αυτή την κατεύθυνση είναι σχεδιασμένες για συνδιάσκεψη πάνω από ψηφιακά τηλεφωνικά κανάλια (Burkley and Flanagan 1990). Τα χαρακτηριστικά της ομιλίας επιτρέπουν την εγκατάσταση τηλεφωνικής κλήσης, την ανάκτηση πληροφοριών, την επαλήθευση του ομιλητή και την πραγματοποίηση συνδιασκέψεων—όλα συνολικά κάτω από ένα έλεγχο φωνής. Συμπληρωματικά χαμηλού ρυθμού δεδομένων κωδικοποίηση έγχρωμων εικόνων δίνει την δυνατότητα μετάδοσης video υψηλής ποιότητας πάνω από μέτρια χωρητικότητα.

2.6 ΠΡΟΟΠΤΙΚΕΣ ΤΕΧΝΟΛΟΓΙΑΣ ΟΜΙΛΙΑΣ—2000

Πόσο καλοί είμαστε στην πρόβλεψη των εξελίξεων της τεχνολογίας; Σύμφωνα με την εμπειρία μου όχι και τόσο καλοί. Πρόσφατα απέκτησα ένα σύνολο vugraphs πάνω στην κωδικοποίηση, σύνθεση, αναγνώριση και ακουστική συνδιάσκεψη τα οποία προετοίμασα το 1980. Αυτά κατασκευάστηκαν για 5ετείς και 10ετείς προβλέψεις σαν τμήμα μίας άσκησης σχεδιασμού. Προς έκπληξη μου περίπου οι μισές εκ των προβλέψεων ήταν ακριβείς. Αξιοσημείωτες ήταν η κωδικοποίηση υποζώνης για αρχικά προϊόντα φωνητικού ταχυδρομείου (καλείται AUDIX) και η χρήση 32 Kbps ADPCM για εκπεμπόμενες οικονομίες (transmission economies) σε ιδιωτικές γραμμές. Αλλά εκεί υπάρχουν ορισμένες αστρικές παραβλέψεις. Τα vugraphs μου του 1980 φυσικά δεν προέβλεψαν το CELP, αν και ήμουν σε στενή επαφή με την βασική δουλειά που οδήγησε σε αυτό.

Παρά τους σφοδρούς κινδύνους στην πρόβλεψη γεγονότων, πληθώρα επιτευγμάτων φαίνεται πιθανή μέχρι το έτος 2000:

- *Αναπαράσταση σήματος καλής ποιότητας αντίληψης σε<0.5 Bits/δείγμα*

Αυτό θα εξαρτηθεί από τις συνεχιζόμενες εξελίξεις στην μικροηλεκτρονική, ειδικά στην συγχώνευση των ψυχοακουστικών παραγόντων στους αλγόριθμους κωδικοποίησης.

- *Πολυγλωσσική σύνθεση κειμένου σε ομιλία με γενικές ποιότητες φωνής.*

Τα πολυγλωσσικά συστήματα προβάλλονται τώρα. Η προοπτική για αναπαραγωγή των ατομικών φωνητικών χαρακτηριστικών κατά κανόνα δεν υποστηρίζεται ακόμα από βασική γνώση. Αλλά οι γενικές ποιότητες, όπως τα χαρακτηριστικά φωνής για άντρα, γυναίκα και παιδί θα είναι δυνατά.

- *Διαλογική αλληλεπίδραση μεγάλου λεξιλογίου (100Κ λέξεις) με μηχανές με συγκεκριμένου καθήκοντος μοντέλα γλώσσας.*

Η αναγνώριση απεριορίστου λεξιλογίου, από οποιονδήποτε ομιλητή και για οποιοδήποτε θέμα, θα υπάρχει ακόμα στον μακρινό ορίζοντα. Όμως συγκεκριμένου καθήκοντος συστήματα θα λειτουργούν αξιόπιστα και θα αναπτύσσονται ευρέως. Μια ισχυρή έμφαση θα συνεχίζει να υπάρχει στα υπολογιστικά μοντέλα που προσεγγίζουν την φυσική γλώσσα.

- *Εκτεταμένη μετάφραση γλωσσών συγκεκριμένου καθήκοντος.*

Συστήματα τα οποία προχωρούν σημαντικά πάνω από την κατηγορία του “phrase book” είναι πιθανά, αλλά ακόμα με τον περιορισμό του συγκεκριμένου καθήκοντος και τις γενικές ποιότητες της σύνθεσης φωνής.

- *Αυτοματοποιημένη επαύξηση σήματος, η οποία προσεγγίζει την αισθητή οξύτητα.*

Αυτή συγκαταλέγεται μεταξύ των περισσότερο προβληματικών προβλέψεων, αλλά βελτιωμένα μοντέλα, ακοής και μη γραμμικής επεξεργασίας σήματος, για αυτόματη αναγνώριση θα μικρύνουν το χάσμα ανάμεσα στην απόδοση του ανθρώπου και των μηχανών σε θορυβώδη συστήματα. Συγκρίσιμες επιδόσεις αναγνώρισης μεταξύ ανθρώπου και μηχανής φαντάζουν πραγματοποιήσιμες για περιορισμένα λεξιλόγια και θορυβώδεις εισόδους. Οι ευαίσθητες στην παρεμβολή επικοινωνίες όπως οι αέρος προς έδαφος (air to ground) και το προσωπικό κυψελοειδές ραδιόφωνο θα επωφεληθούν.

- *Τρισδιάστατη σύλληψη και προβολή ήχου.*

Οι ανέξοδοι, υψηλής ποιότητας μετατροπείς ενέργειας ηλεκτρίτη, μαζί με τους οικονομικούς μονού chip επεξεργαστές, ανοίγουν προοπτικές για την αντιμετώπιση της πολυκαναλικής αλλοίωσης (multipath distortion) (αντήχηση δωματίου) έτσι ώστε να επιτευχθεί υψηλής ποιότητας σύλληψη ήχου από καθορισμένες χωρικές περιοχές. Ο χωρικός ρεαλισμός στην προβολή και οι φυσικές hands—free επικοινωνίες επιπρόσθετα οφέλη. Η παρούσα έρευνα προτείνει ότι αυτά τα επιτεύγματα είναι δυνατό να υποστηριχθούν.

- *Ολοκλήρωση εικόνας, θορύβου και τρόπων αίσθησης*

Αν και οι συστατικές τεχνολογίες για την όραση, τον ήχο και την αφή θα έχουν ατελείς προοπτικές για το άμεσο μέλλον, ο κατάλληλος σχεδιασμός των σεναρίων εφαρμογών θα κάνει δυνατή την παραγωγική χρήση αυτών των τρόπων αίσθησης στους αλληλεπιδραστικούς σταθμούς εργασίας. Η μηχανική των ανθρωπίνων παραγόντων είναι βασική για την επιτυχία. Εκτεταμένη χρησιμότητα των απτών απεικονίσεων εξαρτάται από τις νέες εξελίξεις στους μετατροπείς ενέργειας—για παράδειγμα ο σχεδιασμός των πινάκων των μετατροπέων ενέργειας ώστε να είναι ικανοί για λεπτομερειακή αναπαράσταση της υφής.

- *Αναγκαίος οικονομικός προγραμματισμός.*

Οι ενδείξεις είναι ότι οι πρόοδοι στην μικροηλεκτρονική θα συνεχιστούν. Προς το παρόν εφαρμόζεται σε ευρεία βάση τεχνολογία 0.9μm η οποία παρέχει υπολογισμούς

της τάξης των 50Mflops σε ένα μονό chip και κοστίζει λιγότερο από ένα δολάριο ανά Mflops. Έως το 2000 η προσδοκία είναι η ευρεία εφαρμογή τεχνολογίας 0.35μm (ή μικρότερης) με ανάλογες πυκνότητες πυλών. Υπολογισμοί της τάξης του 1Gflops θα γίνουν διαθέσιμοι σε ένα μονό chip. Αυτή η διαθεσιμότητα υπολογισμών θα προκαλεί διαρκώς τους ερευνητές της ομιλίας να κατασκευάζουν αλγόριθμους τεράστιας πολυπλοκότητας. Εάν η πρόκληση μάλιστα επιτευχθεί το έτος 2001 μπορεί πραγματικά να υλοποιηθεί μια τύπου HAL διαλογική μηχανή.

2.7 ΛΕΞΙΚΟ ΟΡΩΝ

autocorrelation	αυτοσυσχέτιση	
bandpass	ζωνοπερατό, διέλευσης ζώνης	
bandwidth	εύρος ζώνης	
baseband	βασική ζώνη	
band	ζώνη (συχνοτήτων)	περιοχή συχνοτήτων
bandpass	ζωνοπερατό, διέλευσης ζώνης	
bark		κλίμακα ύψους ήχου που αναπαριστά το ψυχοφυσικό ισοδύναμο της συχνότητας
belows	ασκοί	
binaural	και στα δύο αφτιά	
click	παλμός στο πεδίο του χρόνου	
codec	κωδικοποιητής-αποκωδικοποιητής	
cognitive module	διαδικασία εκμάθησης, ή εκπαίδευσης	
constrictions	συσφίξεις	
decoder	αποκωδικοποιητής	
electret	ηλεκτρίτης	
encoder	κωδικοποιητής	
envelope	περιβάλλουσα	
filter bank	τράπεζα φίλτρων	σύνολο ζωνοπερατών φίλτρων που καλύπτουν μια ζώνη συχνοτήτων [*αναφέρεται και ως* bank-of-filters]
formant	φωνοσυντονισμός	χαρακτηριστικοί συντονισμοί της φωνητικής οδού
identification	ταυτοποίηση	
impulse	κρουστικός παλμός	
internal representation	εσωτερική αναπαράσταση	ο τρόπος με τον οποίο απεικονίζεται στο εσωτερικό του αφτιού κάποιο ακουστικό σήμα
intensity	ένταση	
interaction	αλληλεπίδραση	
interaural	Ανάμεσα στα δυο αφτιά	
loudness	ακουστότητα	το χαρακτηριστικό γνώρισμα με το οποίο καταλαβαίνουμε αν ένας ήχος είναι ισχυρός ή ασθενής ή το μέγεθος του προκαλούμενου ακουστικού αισθήματος
mapping	απεικόνιση	
mapping function	συνάρτηση απεικόνισης	
masked threshold	κατώφλι απόκρυψης	
masking	απόκρυψη	η κατάσταση κατά την οποία η παρουσία ενός ήχου καθιστά αδύνατο το άκουσμα ενός άλλου ήχου
modal	υποθετικός	
model	μοντέλο	
neural processing	νευρωνική επεξεργασία	
normalization	κανονικοποίηση	
passband	ζώνη διέλευσης	περιοχή συχνοτήτων μέσα από την οποία μεταδίδεται ικανοποιητικά ένα σήμα
pattern	πρότυπο, μορφή	

pattern recognition	αναγνώριση προτύπων, αναγνώριση μορφών	
perceptual interpretation	επεξήγηση βάση της αντίληψης	
pitch	μουσικός τόνος, ύψος ήχου	πως ένας ακροατής αντιλαμβάνεται αν ένας ήχος είναι χαμηλός (βαθύς) ή υψηλός (οξύς) με βάση μια υποκειμενική κλίμακα ύψους, χωρίς δηλαδή να λαμβάνει υπόψη του τις φυσικές ιδιότητες του ήχου. Ακουστικά συσχετίζεται με τη *βασική συχνότητα* ή τη θεμελιώδη περίοδο της ομιλίας.
prosody	προσωδία	μαζί τα εξής χαρακτηριστικά του ήχου: μουσικός τόνος, διάρκεια και ένταση
psycoacoustics	ψυχοακουστική	
psycholinguistics	ψυχογλωσσολογία	
pulse	παλμός	
recursion	αναδρομή	
resonator	αντηχείο	
robustness speech recognition	εύρωστη αναγνώριση ομιλίας	η δυνατότητα ενός συστήματος αναγνώρισης ομιλίας να συμπεριφέρεται με ακρίβεια σε δύσκολες συνθήκες
round vowel	στρογγυλό φωνήεν	
sensory modalities	τρόποι αίσθησης	
sone		μονάδα υποκειμενικής ακουστότητας
sonorant	ηχηρός	ήχοι που παράγονται χωρίς κανένα φράξιμο της φωνητικής οδού, όπως τα υγρά, έρρινα, και όσα πλευρικά παράγονται χωρίς έμφραξη.
speech generation	δημιουργία ομιλίας	
speech processing	επεξεργασία ομιλίας	
speech synthesis	σύνθεση ομιλίας	
standard	πρότυπο	
standardization	τυποποίηση	
stationary	στατικό	Στατιστικό μέγεθος του οποίου οι παράμετροι (πχ μέση τιμή, διασπορά κτλ) δεν μεταβάλλονται με τον χρόνο
stimuli	ερεθίσματα, διεγέρσεις	
supervised learning	εκμάθηση με επίβλεψη	
system characterization	περιγραφή συστήματος	
talker verification	εξακρίβωση ομιλητή	
temporal representation	χρονική αναπαράσταση	
transducer	μετατροπέας ενέργειας	
transparent system	διαπερατό σύστημα	το σύστημα του οποίου η είσοδος είναι ίδια με την έξοδο
text-to-speech synthesis	σύνθεση ομιλίας από κείμενο	
threshold of hearing	κατώφλι ακοής	
time-frequency	χρονοσυχνοτικός	
tone	τόνος	(στη γλωσσολογία) ο ιδιαίτερος τρόπος χρησιμοποίησης του μουσικού τόνου σε μια γλώσσα
turbulent flow	τυρβώδης ροή	
valved flow	βαλβοειδής ροή	
verification	εξακρίβωση	
vocal cords	δες vocal folds	
vocoder	φωνοκωδικοποιητής	
voice mimic	μίμηση φωνής	τεχνική που χρησιμοποιείται στην επεξεργασία ομιλίας κατά την οποία μεταφράζεται η ομιλία ενός

		ανθρώπου σε μία άλλη γλώσσα αλλά διατηρούνται από τον συνθέτη ομιλίας τα χαρακτηριστικά της φωνής του
voice processing	επεξεργασία φωνής	
voiced	ηχηρό, έμφωνο	ήχοι κατά την παραγωγή των οποίων πάλλονται οι φωνητικές χορδές, όπως οι [b, d, g, v, …]
voiceless	άηχο, άφωνο	ήχοι κατά την παραγωγή των οποίων δεν πάλλονται οι φωνητικές χορδές, όπως οι [p, t, k, f, …]

2.8 ΠΙΝΑΚΑΣ ΑΚΡΟΝΥΜΙΩΝ

ACR	Absolute Category Rating
ADPCM	Adaptive PCM
CCITT	Comite Consultatif International pour le Telephone et le Telegraphe
CD	Compact Disc
CELP	Code Excited Linear Prediction
dB	deciBell
DCR	Degradation Category Rating
HMM	Hidden Markov Model
ITU	International Telecommunications Union
MOS	Mean Opinion Score
PAQM	Perceptual Audio Quality Measure
PSQM	Perceptual Speech Quality Measure
PCM	Pulse Code Modulation
SPL	Sound Pressure Level

2.9 ΒΙΒΛΙΟΓΡΑΦΙΑ

1. John G. Beerends “AUDIO QUALITY DETERMINATION BASED ONNN PERCEPTUAL MEASURMENT TECHNIQUES”
2. National Academy of Scinces “VOICE COMMUNICATIONS BETWEEN HUMANS AND MACHINES”