ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
Μεταπτυχιακό Πρόγραμμα Σπουδών

Ασημακόπουλος Γιάννης (Α.Μ. 97502)
Γρυλλάκης Αυγουστίνος (Α.Μ. 97508)

Τελευταία λέξη της τεχνολογίας στην αναγνώριση συνεχούς ομιλίας

Εργασία στο μάθημα: Επικοινωνία με Ομιλία
Διδάσκων: Γεώργιος Κουρουπέτρογλου

Αθήνα 1999

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Έχουν γίνει πολύ μεγάλες πρόοδοι στην αυτόματη αναγνώριση ομιλίας από μηχανή. Αυτή η εργασία εστιάζει στις προόδους της αναγνώρισης ομιλίας που έγιναν μέσα από καλύτερες τεχνικές μοντελοποίησης ομιλίας, κυρίως μέσω μεγαλύτερης ακρίβειας μαθηματικών μοντέλων στους ήχους της ομιλίας.

Πιο συγκεκριμένα ξεκινάει από την αναγνώριση ομιλίας και τα προβλήματά της λόγω διακυμάνσεων στα χαρακτηριστικά του σήματος ομιλίας. Προχωράει στις μονάδες ομιλίας και την εμφάνισή τους στο φασματογράφημα. Το μαθηματικό υπόβαθρο που χρησιμοποιείται είναι τα κρυμμένα μοντέλα Markov. Αυτά εξετάζονται διεξοδικά αφού πρώτα γίνει μία αναφορά στις αλυσίδες Markov. Μετά εξηγείται πως χρησιμοποιούνται τα κρυμμένα μοντέλα Markov για να μοντελοποιήσουν φωνητικά ηχητικά γεγονότα. Ύστερα αναλύεται η εκπαίδευση τους και τονίζεται η σημασία της γραμματικής στην όλη διαδικασία. Αφού οριστεί η αναγνώριση αναλύονται τα τρία προβλήματα που πρέπει να λυθούν για να επεκταθεί στην συνεχή ομιλία: το πρόβλημα εκπαίδευσης, ή πώς να εκπαιδεύσουμε HMMs για συνεχή ομιλία, το πρόβλημα μοντελοποίησης, ή πώς να εμπλουτίσουμε την ακουστική-φωνητική μοντελοποίηση για συνεχή ομιλία, και το πρόβλημα αναγνώρισης.

Στο πρόβλημα της εκπαίδευσης αναφέρονται δύο βασικοί αλγόριθμοι: ο αλγόριθμος μπρος-πίσω και ο κατατμημένος αλγόριθμος κ-μέσων. Στο πρόβλημα της μοντελοποίησης εξετάζονται τέσσερις περιοχές έρευνας που μπορεί να αυξήσουν την απόδοση του συστήματος: η επεξεργασία και απεικόνιση σήματος, η δομή κρυμμένων μοντέλων Markov, ο αλγόριθμος μάθησης, και ο κατάλογος των μονάδων υπολέξεων. Στο πρόβλημα αναγνώρισης αναλύεται η θεμελιώδης εξίσωση της αναγνώρισης ομιλίας. Αναφέρεται ένας αλγόριθμος βασικής μορφής έρευνας, η έρευνα απαρίθμησης, καθώς και τέσσερις εμπλουτίσεις που μπορεί να γίνουν στην έρευνα για καλύτερη αναγνώριση. Επίσης γίνεται αναφορά στον αλγόριθμο Viterbi, στον αλγόριθμο Viterbi δέσμης και σε κάποιες παραλλαγές του.

Ύστερα έχουμε κάποιες βελτιώσεις στην απόδοση λόγω: χρήσης κοινών ηχητικών στοιχείων γλώσσας, βελτιωμένης ακουστικής μοντελοποίησης, βελτιωμένης γλωσσικής μοντελοποίησης, και γρηγορότερου κύκλου πειραματικής έρευνας. Τονίζονται οι επιδράσεις της εκπαίδευσης και της γραμματικής στον ρυθμό σφάλματος λέξης. Μετά ορίζονται οι εξαρτώμενη-ομιλητή και ανεξάρτητη-ομιλητή αναγνώριση ομιλίας, καθώς και το σύστημα αναγνώρισης ομιλίας που να είναι ανεξάρτητο τομέα. Βλέπουμε πως είναι δυνατόν να αυξήσουμε την απόδοση ενός εξαρτώμενου-ομιλητή ή ανεξάρτητου-ομιλητή συστήματος με αυξητική προσαρμογή στην φωνή του καινούργιου ομιλητή καθώς αυτός χρησιμοποιεί το σύστημα, όπως και προσθέτοντας νέες λέξεις στο σύστημα.

Τέλος εξετάζονται συνοπτικά δύο εναλλακτικά μοντέλα: τα στοχαστικά κατατμημένα μοντέλα και τα νευρωνικά δίκτυα, τα οποία χρησιμοποιούμενα με το σύστημα HMM για την αυξάνουν την απόδοση.

ΚΕΦΑΛΑΙΑ

- Το πρόβλημα αναγνώρισης ομιλίας
  - Γενική Σύνθεση/Διαδικασία Αναγνώρισης
  - Μονάδες Ομιλίας
- Κρυμμένα μοντέλα Markov
  - Αλυσίδες Markov
  - Κρυμμένα Markov μοντέλα
  - Φωνητικά HMMs
- Εκπαίδευση και αναγνώριση
  - Εκπαίδευση
    - Φωνητικά HMMs και Λεξιλόγιο
    - Γραμματική
  - Αναγνώριση
- Πρόβλημα εκπαίδευσης
- Πρόβλημα μοντελοποίησης
  - Βελτιωμένη Απεικόνιση Σήματος
  - Βελτιωμένες HMM Δομές
  - Βελτιωμένη Μάθηση και Διάκριση
  - Βελτιωμένος Κατάλογος Υπολέξεων
- Πρόβλημα αναγνώρισης
  - Περίληψη του Προβλήματος
    - Το μοντέλο Γλώσσας
    - Το Ακουστικό Μοντέλο
    - Συνδυάζοντας Γλωσσικά και Ακουστικά Μοντέλα
    - Απεριόριστη Έρευνα
    - Εμπλουτίσεις έρευνας
  - Έρευνα Δέσμης Viterbi
    - Βέλτιστη Έρευνα Viterbi
    - Η Έρευνα Δέσμης Viterbi
    - Παραλλαγές Έρευνας Δέσμης Viterbi
- Τελευταία λέξη της τεχνολογίας
  - Βελτιώσεις στην Απόδοση
    - Κοινά Ηχητικά Στοιχεία Γλώσσας
    - Ακουστική Μοντελοποίηση
    - Μοντελοποίηση Γλώσσας
    - Κύκλος Έρευνας Πειραματισμού
  - Σχήματα Απόδοσης Δειγμάτων
  - Επιδράσεις της Εκπαίδευσης και της Γραμματικής
  - Αναγνώριση Εξαρτώμενη-Ομιλητή & Ανεξάρτητη-Ομιλητή
  - Προσαρμογή
  - Προσθέτοντας Νέες Λέξεις
- Αναγνώριση ομιλίας πραγματικού-χρόνου
- Εναλλακτικά μοντέλα
  - Κατατμημένα Μοντέλα
  - Νευρωνικά Δίκτυα

Τελευταία λέξη της τεχνολογίας στην αναγνώριση συνεχούς ομιλίας

John Makhoul and Richard Schwartz

ΠΕΡΙΛΗΨΗ

Την τελευταία δεκαετία, τεράστιες πρόοδοι έχουν γίνει στην τελευταία λέξη της τεχνολογίας αυτόματης αναγνώρισης φωνής από μηχανή. Μία μείωση στον ρυθμό λάθους λέξης περισσότερο από ένα παράγοντα του 5 και μία αύξηση στις ταχύτητες αναγνώρισης σε διάφορες τάξεις μεγέθους (οφειλόμενες σε ένα συνδυασμό γρηγορότερων αλγορίθμων αναζήτησης και πιο ισχυρών υπολογιστών), έχουν συνδυαστεί ώστε να παρέχουν υψηλή ακρίβεια, ανεξάρτητη ομιλητή, αναγνώριση συνεχούς ομιλίας για μεγάλα λεξιλόγια σε πραγματικό χρόνο, σε μη κατά παραγγελία φτιαγμένους υπολογιστές, χωρίς την βοήθεια ειδικών συσκευών. Αυτές οι πρόοδοι υπόσχονται να κάνουν την τεχνολογία αναγνώρισης ομιλίας άμεσα διαθέσιμη στο ευρύ κοινό. Αυτή η εργασία εστιάζει στις προόδους της αναγνώρισης ομιλίας που έγιναν μέσα από καλύτερες τεχνικές μοντελοποίησης ομιλίας, κυρίως μέσω μεγαλύτερης ακρίβειας μαθηματικών μοντέλων στους ήχους της ομιλίας.

ΕΙΣΑΓΩΓΗ

Όλο και περισσότερο, η τεχνολογία της αναγνώρισης ομιλίας διαγράφει μία διαδρομή από το εργαστήριο σε εφαρμογές για το ευρύ κοινό. Πρόσφατα, μια ποιοτική αλλαγή στην τελευταία λέξη της τεχνολογίας ήρθε στην επιφάνεια και υπόσχεται να φέρει τις δυνατότητες στην αναγνώριση ομιλίας κοντά σε οποιονδήποτε έχει πρόσβαση σε έναν υπολογιστή. Υψηλής ακρίβειας, πραγματικού χρόνου, ανεξάρτητη από ομιλητή, αναγνώριση συνεχούς ομιλίας για μετρίου μεγέθους λεξιλόγια (μερικών χιλιάδων λέξεων) είναι τώρα δυνατή σε λογισμικό σε μη κατά παραγγελία φτιαγμένους υπολογιστές. Οι χρήστες θα μπορούν να προσαρμόζουν δυνατότητες αναγνώρισης στις δικές τους εφαρμογές. Τέτοιες βασισμένες σε λογισμικό, πραγματικού χρόνου λύσεις προαγγέλλουν μια νέα εποχή στην ανάπτυξη και χρήση της τεχνολογίας αναγνώρισης ομιλίας.

Όπως συμβαίνει συχνά στην τεχνολογία, παραδειγματική μετακίνηση συμβαίνει όταν αρκετές εξελίξεις συγκλίνουν για να κάνουν μία καινούργια δυνατότητα εφικτή. Στην περίπτωση της αναγνώρισης συνεχούς ομιλίας, οι ακόλουθες πρόοδοι έχουν συνδυαστεί για να κάνουν τη νέα τεχνολογία δυνατή:

- Υψηλότερης ακρίβειας αναγνώριση συνεχούς ομιλίας, βασισμένη σε καλύτερες τεχνικές μοντέλων ομιλίας
- Καλύτερες στρατηγικές στην αναγνώριση ομιλίας που μειώνουν τον απαιτούμενο χρόνο για υψηλής ακρίβειας αναγνώριση και
- Αυξημένη ισχύς σε μη κατά παραγγελία φτιαγμένους υπολογιστές πολυμέσων.

Η παραδειγματική μετακίνηση συμβαίνει μεταξύ του πως βλέπουμε και χρησιμοποιούμε την αναγνώριση ομιλίας. Παρά να είναι περισσότερο μια εργαστηριακή προσπάθεια, η αναγνώριση ομιλίας γίνεται γρήγορα μια τεχνολογία που είναι διεισδυτική και θα έχει μια βαθιά επίδραση στον τρόπο που οι άνθρωποι επικοινωνούν με τις μηχανές και μεταξύ τους.

Αυτή η εργασία εστιάζει στα πλεονεκτήματα της μοντελοποίησης ομιλίας στην αναγνώριση συνεχούς ομιλίας, με μια αναφορά στα κρυφά μοντέλα Markov (HMMs), τη μαθηματική ραχοκοκαλιά πίσω από αυτές τις προόδους. Ενώ η γνώση των ιδιοτήτων του σήματος ομιλίας και της αντίληψης της ομιλίας παίζανε πάντα ένα ρόλο, πρόσφατες βελτιώσεις έχουν βασιστεί πολύ σε ακλόνητες μαθηματικές και πιθανοκρατικής μοντελοποίησης μεθόδους, ειδικά η χρήση των HMMs για μοντελοποίηση ήχων ομιλίας. Αυτές οι μέθοδοι είναι ικανές για μοντελοποίηση χρόνου και διακύμανσης φάσματος ταυτόχρονα, και οι παράμετροι του μοντέλου μπορούν να προσεγγιστούν αυτόματα από δοθέντα δεδομένα εκπαιδευτικής ομιλίας. Οι παραδοσιακές διαδικασίες κατάτμησης και ονοματολογίας των ήχων ομιλίας τώρα ενώνονται σε μια μοναδική πιθανοκρατική διαδικασία που μπορεί να βελτιστοποιήσει την ακρίβεια αναγνώρισης.

Αυτή η εργασία περιγράφει την διαδικασία αναγνώρισης ομιλίας και παρέχει τυπικά σχήματα αναγνώρισης ακριβείας από εργαστηριακά πειράματα συναρτήσει του λεξικού, της εξάρτησης ομιλητή, της πολυπλοκότητας γραμματικής, και του ποσού ομιλίας που χρησιμοποιήθηκε για να εκπαιδευτεί το σύστημα. Σαν αποτέλεσμα των προόδων της μοντελοποίησης, οι ρυθμοί σφαλμάτων αναγνώρισης έχουν πέσει πολύ. Σημαντικό σε αυτές τις βελτιώσεις είναι η διαθεσιμότητα ηχητικών στοιχείων κοινής ομιλίας για εκπαιδευτικούς και δοκιμαστικούς σκοπούς και η υιοθέτηση αναγνωρισμένων διαδικασιών ελέγχου.

Αυτή η εργασία επίσης παρουσιάζει περισσότερο πρόσφατες κατευθύνσεις έρευνας, συμπεριλαμβανομένης της χρήσης κατατμημένων μοντέλων και δικτύων τεχνητής νοημοσύνης βελτιώνοντας την απόδοση των HMMs συστημάτων. Οι ικανότητες των νευρωνικών δικτύων να μοντελοποιούν τις υψηλά μη γραμμικές συναρτήσεις μπορούν να χρησιμοποιηθούν για να αναπτύξουν νέα χαρακτηριστικά από το σήμα της ομιλίας, και η δυνατότητά τους να μοντελοποιούν μεταγενέστερες πιθανότητες μπορεί να χρησιμοποιηθεί για να βελτιώσει την ακρίβεια αναγνώρισης.

Θα αμφισβητήσουμε το γεγονός πως οι μελλοντικές πρόοδοι στην αναγνώριση ομιλίας πρέπει να συνεχίσουν να βασίζονται στην εύρεση καλύτερων τρόπων να ενσωματώσουν την γνώση μας στην ομιλία σε προοδευμένα μαθηματικά μοντέλα, με μία έμφαση σε μεθόδους που αντέχουν σε μεταβλητότητα ομιλητών, θορύβων, και άλλων ακουστικών διαταραχών.

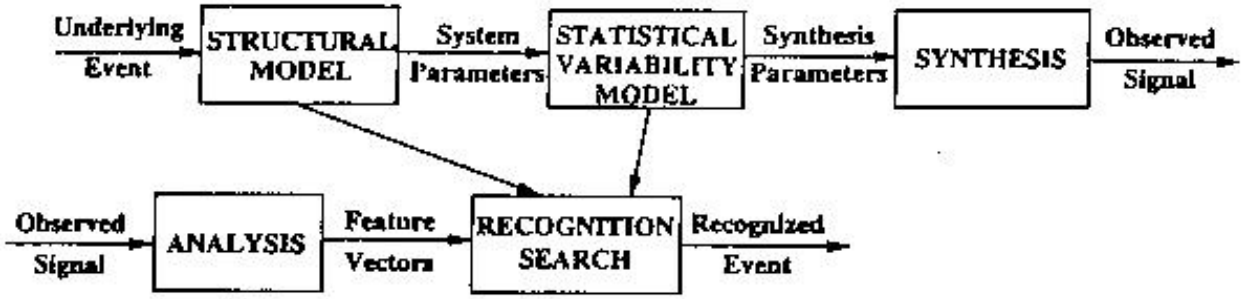
ΤΟ ΠΡΟΒΛΗΜΑ ΑΝΑΓΝΩΡΙΣΗΣ ΟΜΙΛΙΑΣ

Η αυτόματη αναγνώριση ομιλίας μπορεί να ειδωθεί σαν μία αντιστοίχηση από ένα συνεχές σήμα, το σήμα ομιλίας, σε μία ακολουθία διακριτών οντοτήτων, για παράδειγμα, φωνήματα (ή ήχοι ομιλίας), λέξεις, και προτάσεις. Το σοβαρότερο εμπόδιο για υψηλής ακρίβειας αναγνώριση είναι η μεγάλη διακύμανση στα χαρακτηριστικά του σήματος ομιλίας. Αυτή η διακύμανση οφείλεται σε τρεις κυρίως λόγους: γλωσσολογική διακύμανση, διακύμανση ομιλητών, και διακύμανση καναλιού. Η γλωσσολογική διακύμανση περιλαμβάνει τις επιδράσεις της φωνητικής, φωνολογίας, σύνταξης, σημασιολογίας, και ομιλίας στο σήμα ομιλίας. Η διακύμανση ομιλητών περιλαμβάνει ένδο- και μεταξύ ομιλητών διακύμανση, συμπεριλαμβανομένης των επιδράσεων της συνάρθρωσης, δηλαδή, των επιδράσεων γειτονικών ήχων στην ακουστική πραγματοποίηση ενός ειδικά φωνήματος, λόγω της συνέχειας και των περιορισμών στην κίνηση του ανθρώπινου αρθρωτικού μηχανισμού. Η διακύμανση καναλιού περιλαμβάνει τις επιδράσεις του θορύβου του περιβάλλοντος και του καναλιού μετάδοσης (π.χ. μικρόφωνο, τηλέφωνο, αντήχηση). Όλες αυτές οι διακυμάνσεις τείνουν να καλύψουν το μήνυμα με αβεβαιότητα, η οποία πρέπει να εξαλειφθεί από την διαδικασία της αναγνώρισης.

Γενική Σύνθεση/Διαδικασία Αναγνώρισης

Βλέπουμε την διαδικασία της αναγνώρισης σαν μέρος μιας γενικότερης διαδικασίας σύνθεσης/αναγνώρισης, όπως φαίνεται στο Σχήμα 1. Υποθέτουμε πως η διαδικασία της σύνθεσης αποτελείται από τρία μέρη: ένα μοντέλο δόμησης, ένα μοντέλο στατιστικής διακύμανσης, και τη σύνθεση του σήματος ομιλίας. Η είσοδος μπορεί να είναι μία ακολουθία λέξεων, και η έξοδος το πραγματικό σήμα ομιλίας. Το μοντέλο δόμησης συνδυάζει πολλές όψεις της γνώσης μας για ομιλία και γλώσσα, και το μοντέλο στατιστικής διακύμανσης αντιστοιχεί για τις διαφορετικές διακυμάνσεις

που συναντώνται. Η διαδικασία αναγνώρισης ξεκινά με ανάλυση του σήματος ομιλίας σε μία ακολουθία από διανύσματα χαρακτηριστικών. Αυτή η ανάλυση στοχεύει στο να μειώσει μία όψη της διακύμανσης σήματος λόγω των αλλαγών στο ύψος ήχου, κλπ. Δοθέντος της ακολουθίας των διανυσμάτων χαρακτηριστικών, η διαδικασία αναγνώρισης μειώνεται σε μία έρευνα πάνω σε όλα τα δυνατά γεγονότα (ακολουθίες λέξεων) γι' αυτό το γεγονός που έχει την υψηλότερη πιθανότητα δοθέντος της ακολουθίας των διανυσμάτων χαρακτηριστικών, βασισμένη στα μοντέλα δομικής και στατιστικής διακύμανσης χρησιμοποιούμενα στην σύνθεση.

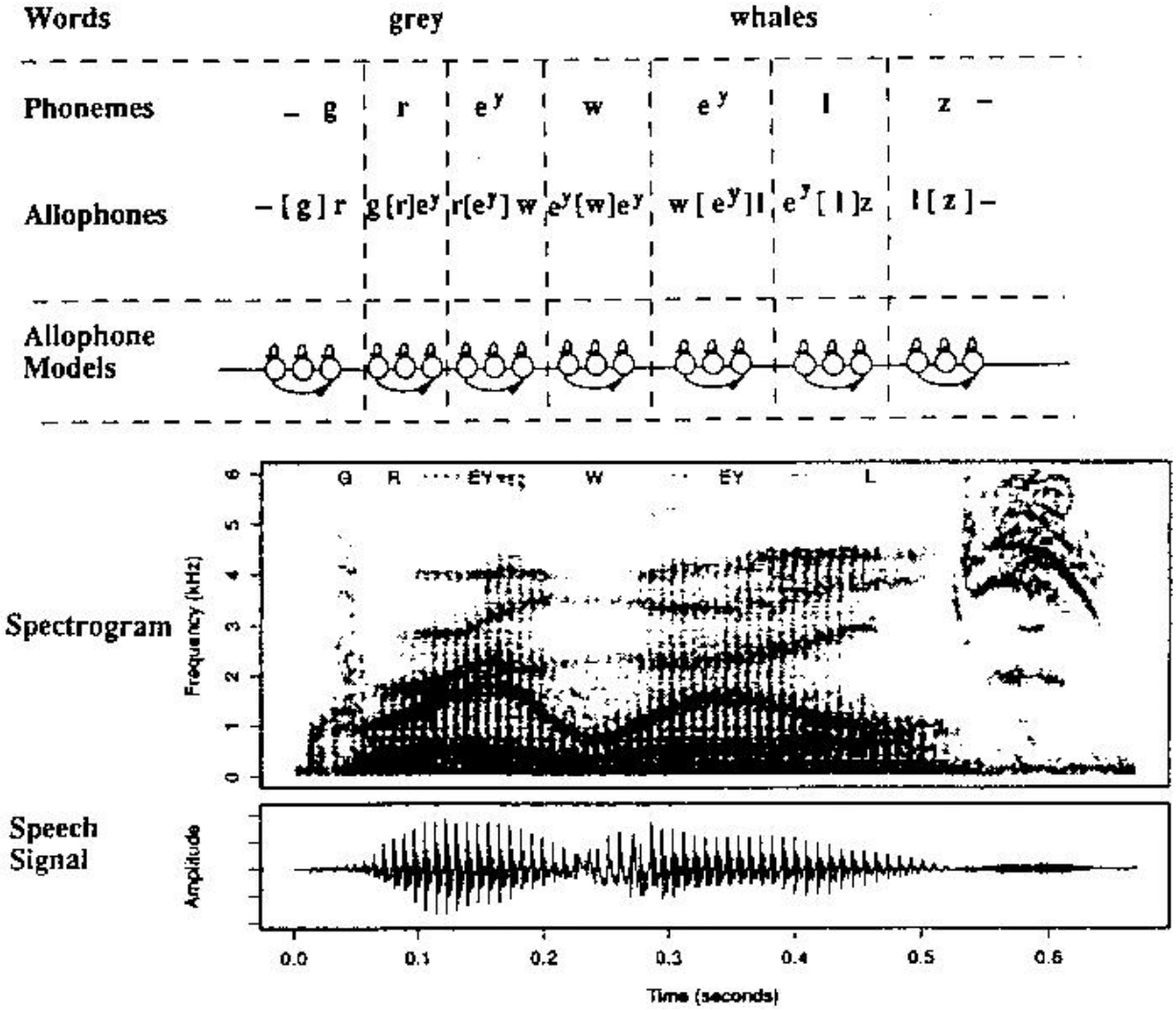


Σχήμα Error! Unknown switch argument. **Γενική διαδικασία σύνθεσης/αναγνώρισης**

Είναι σημαντικό να σημειώσουμε πως ένα σημαντικό μέρος της γνώσης ομιλίας είναι ενσωματωμένο στο μοντέλο δόμησης, συμπεριλαμβανομένης της γνώσης μας για τη δομή της γλώσσας, παραγωγή ομιλίας, και αντίληψης ομιλίας. Παραδείγματα της δομής της γλώσσας περιλαμβάνουν το γεγονός πως η συνεχής ομιλία αποτελείται από μία αλληλουχία λέξεων και πως οι λέξεις είναι μία αλληλουχία από βασικούς ήχους ομιλίας ή φωνήματα. Αυτή η γνώση της δομής της γλώσσας είναι αρκετά παλιά, τουλάχιστον 3000 χρόνια παλιά. Μία πιο σύγχρονη όψη της δομής της γλώσσας που εκτιμήθηκε αυτόν τον αιώνα είναι το γεγονός πως η ακουστική πραγματοποίηση των φωνημάτων είναι κατά πολύ εξαρτώμενη από το γειτονικό περιβάλλον. Η γνώση μας στην παραγωγή ομιλίας, σε όρους τρόπου άρθρωσης (π.χ. έμφωνο, δασύ, έρρινο) και μερών άρθρωσης (π.χ. υπερωικό, ουρανικό, οδοντικό, χειλικό), για παράδειγμα, μπορεί να χρησιμοποιηθεί για παροχή φειδωλών ομάδων από φωνητικά περιβάλλοντα. Όσο για την αντίληψη ομιλίας, πολλά είναι γνωστά για την ανάλυση ήχου στους κοχλίες, για παράδειγμα, πως η βρισκόμενη στα θεμέλια μεμβράνη πραγματοποιεί μία μορφή περίπου συχνοτικής ανάλυσης σε μία γραμμική κλίμακα συχνοτήτων, και για φαινόμενα απόκρυψης στο χρόνο και την συχνότητα. Όλη αυτή η γνώση μπορεί να ενσωματωθεί ευεργετικά στην μοντελοποίηση μας για το σήμα ομιλίας για λόγους αναγνώρισης.

Μονάδες Ομιλίας

Για να έχουμε μία εκτίμηση για το τι μοντελοποίηση απαιτείται για να πραγματοποιηθεί αναγνώριση, θα χρησιμοποιήσουμε σαν παράδειγμα τη φράση "grey whales," της οποίας το σήμα ομιλίας φαίνεται στο κάτω μέρος του Σχήματος 2 με το αντίστοιχο φασματογράφημα (ή φωνητικό αποτύπωμα) να υπάρχει ακριβώς από πάνω. Το φασματογράφημα δείχνει το αποτέλεσμα μίας ανάλυσης συχνότητας της ομιλίας, με τις σκοτεινές ζώνες να αναπαριστούν συντονισμούς των φωνητικών περιοχών. Στο πάνω μέρος του Σχήματος 2 είναι οι δύο λέξεις "grey" και "whales", οι οποίες είναι το ζητούμενο αποτέλεσμα του συστήματος αναγνώρισης. Το πρώτο πράγμα που σημειώνουμε είναι πως το σήμα ομιλίας και το φασματογράφημα δεν δείχνουν κανένα διαχωρισμό ανάμεσα στις δύο λέξεις "grey" και "whales", οι λέξεις συνδέονται η μία με την άλλη, χωρίς κανένα εμφανή διαχωρισμό. Η ανθρώπινη αντίληψη πως μία ομιλούμενη έκφραση αποτελείται από μία ακολουθία διακριτών λέξεων είναι ένα φαινόμενο αντίληψης. Η πραγματικότητα είναι πως οι λέξεις δεν διαχωρίζονται καθόλου.



Σχήμα Error! Unknown switch argument. **Μονάδες ομιλίας**

Κάτω από το επίπεδο λέξης στο Σχήμα 2 είναι το φωνητικό επίπεδο. Εδώ οι λέξεις αναπαριστούνται από όρους ενός φωνητικού αλφαβήτου που μας λέει τι είναι οι διάφοροι ήχοι στις δύο λέξεις. Σε αυτήν την περίπτωση η φωνητική εγγραφή είναι $[g\ r\ e^{y}\ y\ w\ e^{y}\ l\ z]$. Πάλι, ενώ η ακολουθία των φωνημάτων είναι διακριτή, δεν υπάρχει φυσικός διαχωρισμός μεταξύ των διαφορετικών ήχων στο σήμα ομιλίας. Πράγματι, δεν είναι ξεκάθαρο που ο ένας ήχος τελειώνει και που ο επόμενος ξεκινά. Η διακεκομμένες κάθετες γραμμές που φαίνονται στο Σχήμα 2 δίνουν μία χοντρική κατάτμηση του σήματος ομιλίας, η οποία δείχνει περίπου τις αντιστοιχίες μεταξύ των φωνημάτων και της ομιλίας.

Τώρα, το φώνημα $[e^{y}]$ προκύπτει μία φορά σε καθεμία από τις δύο λέξεις. Αν κοιτάξουμε στα τμήματα του φασματογράμματος που αντιστοιχούν στα δύο $[e^{y}]$ φωνήματα, παρατηρούμε κάποιες ομοιότητες μεταξύ των δύο μερών, αλλά επίσης και κάποιες διαφορές. Οι διαφορές οφείλονται περισσότερο λόγω του γεγονότος ότι τα δύο φωνήματα βρίσκονται σε διαφορετικά περιβάλλοντα, το πρώτο $[e^{y}]$ φώνημα προηγείται από $[r]$ και ακολουθείται από $[w]$, ενώ το δεύτερο προηγείται από $[w]$ και ακολουθείται από $[1]$. Αυτές οι συναφείς περιβαλλοντολογικές επιδράσεις είναι αποτέλεσμα αυτού που ονομάζεται συνάρθρωση, του γεγονότος πως η άρθρωση κάθε ήχου αναμειγνύεται με την άρθρωση του επόμενου ήχου. Σε πολλές περιπτώσεις, περιβαντολλογικές φωνητικές επιδράσεις συνδέουν αρκετά φωνήματα, αλλά οι κυριότερες επιδράσεις συμβαίνουν από δύο γειτονικά φωνήματα.

Για να ερμηνεύσουμε το γεγονός πως το ίδιο φώνημα έχει διαφορετικές ακουστικές πραγματοποιήσεις, εξαρτούμενες από το περιβάλλον, αναφερόμαστε σε κάθε ειδικό περιβάλλον σαν

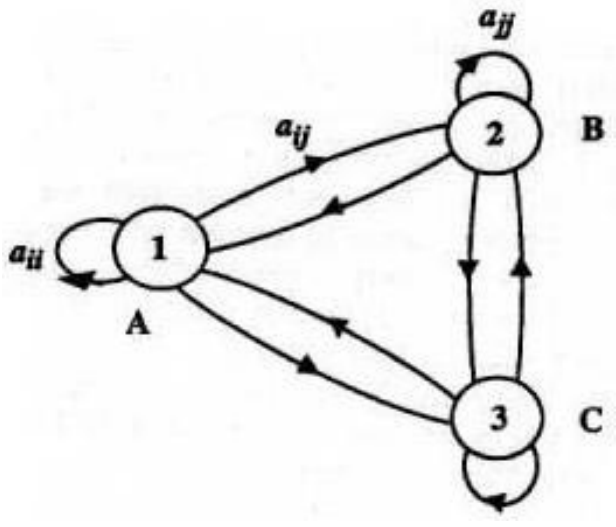
αλλόφωνο. Γι' αυτόν τον λόγο, στο Σχήμα 2, έχουμε δύο διαφορετικά αλλόφωνα του φωνήματος $[e^y]$, ένα για καθένα από τα δύο περιβάλλοντα στις δύο λέξεις. Με αυτόν τον τρόπο, είμαστε ικανοί να αντιμετωπίσουμε την φωνητική διακύμανση που είναι έμφυτη στην συνάρθρωση και που είναι φανερή στο φασματογράφημα στο Σχήμα 2.

Για να πραγματοποιήσουμε την αναγκαία αντιστοίχηση από το συνεχές σήμα ομιλίας στο διακριτό φωνητικό επίπεδο, εισάγουμε ένα μοντέλο – μία μηχανή διακριτών καταστάσεων στην περίπτωσή μας – για καθένα από τα αλλόφωνα που συναντάμε. Σημειώνουμε από το Σχήμα 2 πως οι δομές αυτών των μοντέλων είναι πανομοιότυπες, οι διαφορές θα είναι στις τιμές των διαφόρων παραμέτρων του μοντέλου. Καθένα από αυτά τα μοντέλα είναι ένα κρυμμένο μοντέλο Markov, το οποίο συζητείται παρακάτω.

ΚΡΥΜΜΕΝΑ MARKOV ΜΟΝΤΕΛΑ

Αλυσίδες Markov

Πριν εξηγήσουμε τι είναι ένα κρυμμένο μοντέλο Markov, ας δούμε τι είναι μία αλυσίδα Markov. Μία αλυσίδα Markov αποτελείται ένα αριθμό καταστάσεων, με μεταβάσεις μεταξύ των καταστάσεων. Συνδεμένη με κάθε μετάβαση είναι μία πιθανότητα και συνδεμένη με κάθε κατάσταση είναι ένα σύμβολο. Το Σχήμα 3 δείχνει μία αλυσίδα Markov τριών καταστάσεων, με πιθανότητες μετάβασης a_{ij} μεταξύ των καταστάσεων i και j. Το σύμβολο A είναι συνδεμένο με την κατάσταση 1, το σύμβολο B με την κατάσταση 2, και το σύμβολο C με την κατάσταση 3. Καθώς έχουμε μία μετάβαση από την κατάσταση 1 στην κατάσταση 2, για παράδειγμα, το σύμβολο B παράγεται σαν έξοδος. Αν η επόμενη μετάβαση είναι από την κατάσταση 2 στον εαυτό της, το σύμβολο B είναι η έξοδος πάλι, ενώ η μετάβαση στην κατάσταση 3, δίνει σαν έξοδο το σύμβολο C. Αυτά τα σύμβολα καλούνται σύμβολα εξόδου γιατί μία αλυσίδα Markov θεωρείται σαν παραγωγικό μοντέλο, έχει σαν έξοδο σύμβολα καθώς έχουμε μετάβαση από μία κατάσταση σε μία άλλη. Ας σημειώσουμε πως σε μία αλυσίδα Markov η μετάβαση από μία κατάσταση σε μία άλλη είναι πιθανοκρατική, αλλά η παραγωγή των συμβόλων εξόδου είναι ντετερμινιστική.

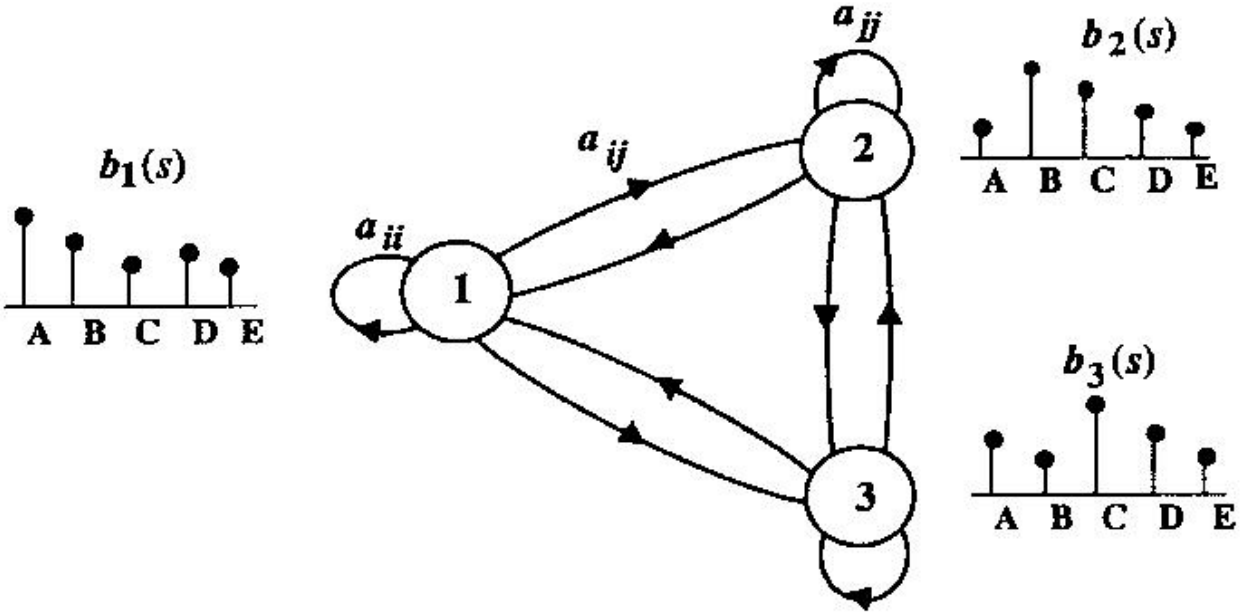


Σχήμα Error! Unknown switch argument. **Μία αλυσίδα Markov τριών καταστάσεων**

Τώρα, δοθέντος μίας ακολουθίας από σύμβολα εξόδου που παρήχθησαν από μία αλυσίδα Markov, ένας μπορεί να ανακαλύψει την αντίστοιχη ακολουθία από καταστάσεις τελείως και αναμφίβολα (αν το σύμβολο εξόδου από κάθε κατάσταση είναι μοναδικό). Για παράδειγμα, η ακολουθία συμβόλων δείγματος B A A C B B A C C C A παράγεται με μεταβάσεις από την εξής ακολουθία καταστάσεων: 2 1 1 3 2 2 1 3 3 3 1.

Κρυμμένα Markov μοντέλα

Ένα κρυμμένο μοντέλο Markov (HMM) είναι το ίδιο με μία αλυσίδα Markov, εκτός από μία σημαντική διαφορά: τα σύμβολα εξόδου σε ένα HMM είναι πιθανοκρατικά. Αντί να συνδέσουμε ένα μόνο σύμβολο εξόδου ανά κατάσταση, σε ένα HMM όλα τα σύμβολα είναι πιθανά σε μία κατάσταση, καθένα με την δική του πιθανότητα. Έτσι, συνδεμένη με κάθε κατάσταση είναι μία κατανομή πιθανότητας όλων των συμβόλων εξόδου. Ακόμα, ο αριθμός των συμβόλων εξόδου μπορεί να είναι αυθαίρετος. Οι διαφορετικές καταστάσεις μπορεί να έχουν τότε διαφορετικές κατανομές πιθανότητας ορισμένες στο σύνολο των συμβόλων εξόδου. Οι πιθανότητες οι συνδεμένες με καταστάσεις είναι γνωστές σαν πιθανότητες εξόδου. (Αν αντί να έχουμε ένα διακριτό αριθμό συμβόλων εξόδου έχουμε ένα συνεχώς εκτιμώμενο διάνυσμα, είναι πιθανό να ορίσουμε μία συνάρτηση πυκνότητας πιθανότητας πάνω σε όλες τις δυνατές τιμές του τυχαίου διανύσματος εξόδου. Αυτή την στιγμή, θα περιορίσουμε την συζήτησή μας σε διακριτά σύμβολα εξόδου.)



Σχήμα Error! Unknown switch argument. **Ένα κρυφό μοντέλο Markov τριών καταστάσεων**

Το Σχήμα 4 δείχνει ένα παράδειγμα από ένα HMM τριών καταστάσεων. Έχει τις ίδιες πιθανότητες μετάβασης όπως η αλυσίδα Markov στο Σχήμα 3. Η διαφορά βρίσκεται στο ότι έχουμε συνδέσει μία κατανομή πιθανότητας $b_i(s)$ με κάθε κατάσταση i , ορισμένη πάνω στο σύνολο των συμβόλων εξόδου s - σε αυτήν την περίπτωση έχουμε πέντε σύμβολα εξόδου-A, B, C, D, και E. Τώρα, όταν μεταβαίνουμε από την μία κατάσταση στην άλλη, το σύμβολο εξόδου επιλέγεται σύμφωνα με την κατανομή πιθανότητας την αντίστοιχη σε αυτήν την κατάσταση. Συγκρίνοντας με την αλυσίδα Markov, οι ακολουθίες εξόδου που παράγονται από ένα HMM είναι γνωστές σαν διπλά στοχαστικές, όχι μόνο η μετάβαση από μία κατάσταση σε μία άλλη είναι στοχαστική (πιθανοκρατική) αλλά έτσι είναι και το σύμβολο εξόδου που παράγεται από κάθε κατάσταση.

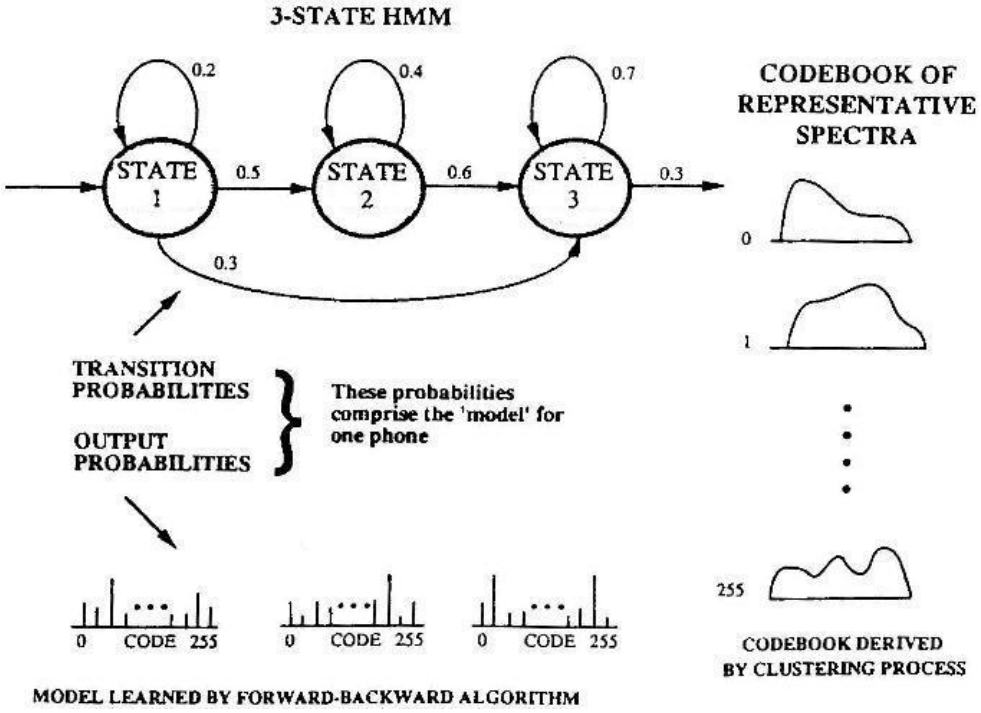
Τώρα, δοθέντος μίας ακολουθίας από σύμβολα παραγόμενα από ένα ειδικό HMM, δεν είναι δυνατόν να ξαναανακαλύψουμε την ακολουθία των καταστάσεων χωρίς αμφιβολίες. Κάθε ακολουθία καταστάσεων του ίδιου μήκους όπως η ακολουθία των συμβόλων είναι δυνατή, καθεμία με διαφορετική πιθανότητα. Δοθέντος της ακολουθίας δειγμάτων εξόδου –C D A A B E D B A C C- δεν υπάρχει τρόπος να ξέρουμε σίγουρα ποια ακολουθία καταστάσεων παρήγε αυτά τα σύμβολα εξόδου. Λέμε ότι η ακολουθία των καταστάσεων είναι κρυμμένη κατά το ότι είναι κρυμμένη από τον παρατηρητή αν ότι κάποιος βλέπει είναι η ακολουθία εξόδου, και γι' αυτό αυτά τα μοντέλα καλούνται *κρυμμένα* Markov μοντέλα.

Αν και δεν είναι δυνατόν να αποφασίσουμε σίγουρα ποια ακολουθία καταστάσεων παρήγε μία συγκεκριμένη ακολουθία συμβόλων, κάποιος μπορεί να ενδιαφερθεί για την ακολουθία των καταστάσεων που είχε την μεγαλύτερη πιθανότητα να έχει παράγει την δεδομένη ακολουθία. Το να

βρούμε μία τέτοια ακολουθία καταστάσεων απαιτείται μία διαδικασία αναζήτησης που, κατ'αρχήν, πρέπει να εξετάσει όλες τις πιθανές ακολουθίες καταστάσεων και να υπολογίσει τις αντίστοιχες πιθανότητες. Ο αριθμός των πιθανών ακολουθιών καταστάσεων αυξάνει εκθετικά με το μήκος της ακολουθίας. Πάντως, λόγω της Markov φύσης ενός HMM, δηλαδή πως ότι υπάρχει σε μία κατάσταση είναι συνάρτηση μόνο της προηγούμενης κατάστασης, υπάρχει μία αποτελεσματική διαδικασία αναζήτησης, ο Viterbi αλγόριθμος (Forney, 1973), που μπορεί να βρει την ακολουθία των καταστάσεων που είναι πιθανότερο να έχουν παράγει την δεδομένη ακολουθία συμβόλων, χωρίς να έχει ψάξει όλες τις πιθανές ακολουθίες. Αυτός ο αλγόριθμος απαιτεί υπολογισμούς που είναι ανάλογοι του αριθμού των καταστάσεων στο μοντέλο και του μήκους της ακολουθίας.

Φωνητικά HMMs

Εξηγούμε τώρα πως τα HMMs χρησιμοποιούνται για να μοντελοποιήσουν φωνητικά ηχητικά γεγονότα. Το Σχήμα 5 δείχνει ένα παράδειγμα HMM τριών καταστάσεων για ένα ανεξάρτητο φώνημα. Το πρώτο στάδιο στην συνεχή σε διακριτή αντιστοίχιση που απαιτείται για αναγνώριση πραγματοποιείται από το κουτί ανάλυσης στο Σχήμα 1. Τυπικά, η ανάλυση αποτελείται από την εκτίμηση του βραχείας διάρκειας φάσματος του σήματος ομιλίας πάνω σε ένα πλαίσιο (παράθυρο) περίπου 20ms. Ο υπολογισμός του φάσματος τότε ενημερώνεται περίπου κάθε 10ms, το οποίο αντιστοιχεί σε μία ροή πλαισίων 100 πλαίσια το δευτερόλεπτο. Αυτό συμπληρώνει την αρχική διακριτότητα στο χρόνο. Πάντως, το HMM, όπως περιγράφεται σε αυτήν την εργασία, επίσης απαιτεί τον ορισμό ενός διακριτού συνόλου "συμβόλων εξόδου". Έτσι, χρειαζόμαστε να διακριτοποιήσουμε το φάσμα σε ένα πεπερασμένο σύνολο φασμάτων. Το Σχήμα 5 περιέχει ένα σύνολο από φόρμες φασμάτων (γνωστό σαν κωδικοβιβλίο) που αναπαριστά τον χώρο των πιθανών φασμάτων ομιλίας. Δοθέντος του υπολογισμένου φάσματος για ένα πλαίσιο ομιλίας, κάποιος μπορεί να βρει την φόρμα στο κωδικοβιβλίο που είναι πιο "κοντά" σε αυτό το φάσμα, χρησιμοποιώντας μία διαδικασία γνωστή σαν κβάντιση διανύσματος (Makhoul et al., 1987). Το μέγεθος του κωδικοβιβλίου στο Σχήμα 5 είναι 256 φόρμες. Αυτό γιατί ο ελάχιστος ρυθμός δειγματοληψίας είναι ο ρυθμός Nyquist που είναι 8KHz. Οπότε $(20ms)*(8KHz)=(20*10^{-3}s)*(8000s^{-1})=160$ φόρμες. Επειδή όμως είμαστε στο δυαδικό σύστημα και 160>128 τελικά έχουμε 256 φόρμες. Αυτές οι φόρμες, ή οι δείκτες τους (από το 0 ως το 255), παίζουν το ρόλο των συμβόλων εξόδου της HMM. Βλέπουμε στο Σχήμα 5 πως συνδεμένη με κάθε κατάσταση είναι μία κατανομή πιθανότητας στο σύνολο των 256 συμβόλων. Ο ορισμός μίας φωνητικής HMM είναι τώρα πλήρης. Τώρα θα περιγράψουμε πως λειτουργεί.



Σχήμα Error! Unknown switch argument. **Βασική δομή ενός φωνητικού HMM**

Ας δούμε πρώτα πως ένα φωνητικό HMM λειτουργεί σαν ένα παραγωγικό μοντέλο (σύνθεσης). Καθώς μπαίνουμε στην κατάσταση 1 στο Σχήμα 5, κάποιο από τα 256 σύμβολα εξόδου παράγεται βάση της αντίστοιχης κατανομής πιθανότητας της κατάστασης 1. Τότε, βάση των πιθανοτήτων μετάβασης από την κατάσταση 1, μία μετάβαση γίνεται είτε πίσω στην κατάσταση 1, ή στην κατάσταση 2, ή στην κατάσταση 3, και άλλο ένα σύμβολο παράγεται βασισμένο στην αντίστοιχη κατανομή πιθανότητας της κατάστασης στην οποία έγινε η μετάβαση. Με αυτόν τον τρόπο παράγεται μία ακολουθία συμβόλων μέχρι να γίνει μία μετάβαση έξω από την κατάσταση 3. Σε αυτό το σημείο, η ακολουθία αντιστοιχεί σε ένα ανεξάρτητο φώνημα.

Το ίδιο μοντέλο μπορεί να χρησιμοποιηθεί στην μέθοδο αναγνώρισης. Σε αυτήν την μέθοδο κάθε μοντέλο μπορεί να χρησιμοποιηθεί για να υπολογίσει την πιθανότητα να έχουμε παραγωγή μίας ακολουθίας φασμάτων. Υποθέτοντας πως αρχίζουμε με την κατάσταση 1 και δοθέντος ενός φάσματος ομιλίας εισόδου που έχει κβαντιστεί σε μία από τις 256 φόρμες, κάποιος μπορεί να κάνει ένα ψάξιμο του πίνακα και να βρει την πιθανότητα αυτού του φάσματος. Αν τώρα υποθέσουμε ότι μία μετάβαση έγινε από την κατάσταση 1 στην κατάσταση 2, για παράδειγμα, η προηγούμενη πιθανότητα εξόδου πολλαπλασιάζεται από την πιθανότητα μετάβασης από την κατάσταση 1 στην κατάσταση 2 (0.5 στο Σχήμα 5). Ένα νέο φάσμα υπολογίζεται τώρα πάνω στο επόμενο πλαίσιο ομιλίας και κβαντίζεται, η αντίστοιχη πιθανότητα εξόδου τότε καθορίζεται από την αντίστοιχη κατανομή πιθανότητας εξόδου στην κατάσταση 2. Αυτή η πιθανότητα πολλαπλασιάζεται με το προηγούμενο γινόμενο, και η διαδικασία συνεχίζεται μέχρι την έξοδο του μοντέλου. Το αποτέλεσμα του πολλαπλασιασμού των ακολουθιών των εξόδων και των πιθανοτήτων μεταβάσεων μας δίνει την ολική πιθανότητα που η ακολουθία φάσματος εισόδου «παρήχθηκε» από αυτό το HMM χρησιμοποιώντας μία ειδική ακολουθία καταστάσεων. Για κάθε ακολουθία καταστάσεων, μία διαφορετική τιμή πιθανότητας προκύπτει. Για αναγνώριση, ο υπολογισμός της πιθανότητας που μόλις περιγράφτηκε πραγματοποιείται για όλα τα δυνατά φωνητικά μοντέλα και για όλες τις δυνατές ακολουθίες καταστάσεων. Αυτή η ακολουθία που έχει την μεγαλύτερη πιθανότητα είναι η αναγνωριζόμενη ακολουθία των φωνημάτων.

Σημειώνουμε στο Σχήμα 5 πως δεν επιτρέπονται όλες οι μεταβάσεις (οι μεταβάσεις που δεν εμφανίζονται έχουν πιθανότητα μηδέν). Αυτό το μοντέλο είναι γνωστό σαν «από αριστερά στα δεξιά» μοντέλο, και αναπαριστά το γεγονός ότι, στην ομιλία, ο χρόνος ρέει σε μία προς τα μπρος διεύθυνση μόνο, αυτή η προς τα μπρος διεύθυνση αναπαριστάται στο Σχήμα 5 από μία γενική από τα αριστερά στα δεξιά κίνηση. Έτσι, δεν υπάρχουν επιτρεπόμενες μεταβάσεις από τα δεξιά στα αριστερά. Μεταβάσεις από οποιαδήποτε κατάσταση πίσω στον εαυτό της εξυπηρετούν στην διακύμανση του μοντέλου στο χρόνο, η οποία είναι πολύ αναγκαία για ομιλία μιας και διαφορετικές στιγμές φωνημάτων και λέξεων προφέρονται με διαφορετικές χρονικές καταγραφές. Η μετάβαση από την κατάσταση 1 στην κατάσταση 3 εννοεί πως το μικρότερο φώνημα που μοντελοποιείται από το μοντέλο στο Σχήμα 5 είναι αυτό που είναι δύο πλαίσια μακρύ, ή 20ms. Ένα τέτοιο φώνημα θα κατέχει την κατάσταση 1 για ένα πλαίσιο και την κατάσταση 3 για ένα πλαίσιο μόνο. Μία εξήγηση για την ανάγκη τριών καταστάσεων, γενικά, είναι ότι η κατάσταση 1 αντιστοιχεί χοντρικά στο αριστερό μέρος του φωνήματος, η κατάσταση 2 στο μεσαίο μέρος, και η κατάσταση 3 στο δεξιό μέρος. (Περισσότερες καταστάσεις μπορεί να χρησιμοποιηθούν, αλλά τότε περισσότερα δεδομένα θα χρειαστούν για να προσεγγιστούν σθεναρά οι παράμετροί τους.)

Συνήθως, υπάρχει ένα HMM για καθένα από τα φωνητικά περιβάλλοντα που μας ενδιαφέρουν. Αν και τα διαφορετικά περιβάλλοντα μπορούν να έχουν διαφορετικές δομές, συνήθως όλα αυτά τα μοντέλα έχουν την ίδια δομή όπως αυτή που φαίνεται στο Σχήμα 5, αυτό που τα διαφοροποιεί είναι η μετάβαση και οι πιθανότητες εξόδου.

ΜΙΑ ΙΣΤΟΡΙΚΗ ΑΝΑΔΡΟΜΗ

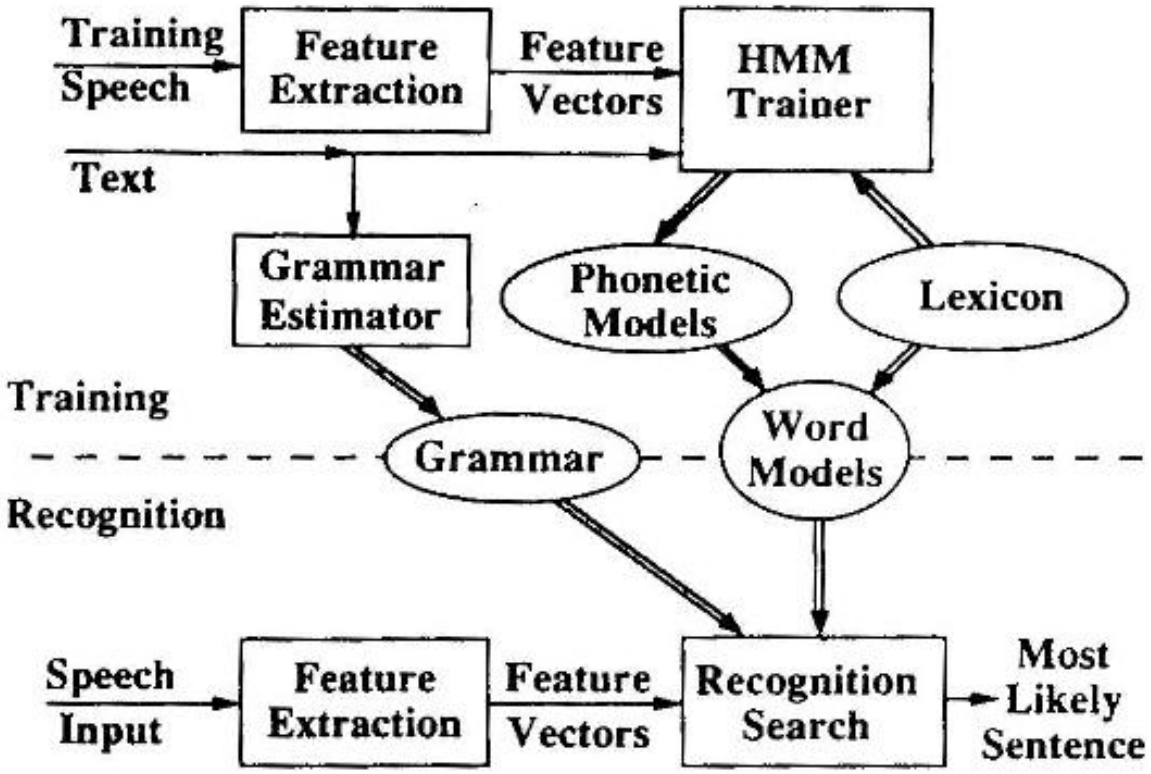
Η θεωρία HMM αναπτύχθηκε στο τέλος της δεκαετίας του 1960 από τον Baum και τους συναδέλφους του (Baum and eagon, 1967) στο ινστιτούτο για την Defense Analyses (IDA). Μια

πρώτη δουλειά χρησιμοποιώντας HMMs για αναγνώριση ομιλίας πραγματοποποιήθηκε την δεκαετία του 1970 στο IDA, IBM (Jelinek et al., 1975), και Carnegie-Mellon University (Baker, 1975). Το 1980 ένας αριθμός ερευνητών στην αναγνώριση ομιλίας στις Ηνωμένες Πολιτείες προσκαλέστηκαν σε ένα εργαστήριο όπου οι IDA ερευνητές ανασκόπησαν τις ιδιότητες των HMMs και των χρησιμοτήτων τους στην αναγνώριση ομιλίας. Αυτό το εργαστήριο υποκίνησε κάποιους οργανισμούς σαν AT&T και BBN, να ξεκινήσουν να εργάζονται με HMMs (Levinson et al., 1983, Schwartz et al., 1984). Το 1984 ένα πρόγραμμα στην αναγνώριση συνεχούς ομιλίας ξεκίνησε από το Advanced Research Projects Agency (ARPA), και γρήγορα τα HMMs έδειξαν να είναι καλύτερα από τις άλλες προσεγγίσεις (Chow et al., 1986). Μέχρι τότε, μόνο λίγοι οργανισμοί παγκοσμίως είχαν εργαστεί με HMMs. Λόγω της αποτυχίας των HMMs και λόγω της ισχυρής επίδρασης του ARPA προγράμματος, με την έμφασή του στις περιοδικές εκτιμήσεις χρησιμοποιώντας κοινά ηχητικά στοιχεία γλώσσας, η χρησιμοποίηση HMMs για αναγνώριση ομιλίας άρχισε να εξαπλώνεται παγκοσμίως. Σήμερα, η χρήση τους έχει επικρατήσει άλλων προσεγγίσεων για αναγνώριση ομιλίας σε πολλά εργαστήρια ανά την υδρόγειο. Επιπρόσθετα των εργαστηρίων που προαναφέρθηκαν, σημαντική εργασία έχει γίνει στα, για παράδειγμα, Massachusetts Institute of Technology's Lincoln Laboratory, Dragon, SRI και TI στις Ηνωμένες Πολιτείες, CRIM και BNR στον Καναδά, RSRE και Cambridge University στην Μεγάλη Βρετανία, ATR, NTT, και NEC στην Ιαπωνία, LIMSI στην Γαλλία, Philips στην Γερμανία και Βέλγιο, και CSELT στην Ιταλία. Περιεκτικές αναφορές των HMMs και της χρησιμότητάς τους στην αναγνώριση ομιλίας μπορούν να βρεθούν στους Rabiner (1989), Lee (1989), Huang et al. (1990), Rabiner and Juang (1993). Αποτελέσματα έρευνας σε αυτόν τον τομέα συνήθως αναφέρονται στις ακόλουθες εφημερίδες και αποτελέσματα συνεδρίων: IEEE transactions on Speech and Audio Processing, IEEE Transactions on Signal Processing, Speech Communication Journal, IEEE International Conderence on Acoustics, Speeech, kai Signal Processing, EuroSpeech, and the International Conference on Speech and Language Processing.

Τα HMMs αποδειχτήκανε να είναι ένα καλό μοντέλο διακύμανσης ομιλίας στο χρόνο και χώρο. Η αυτόματη εκπαίδευση των μοντέλων από δεδομένα ομιλίας επιτάχυνε την ταχύτητα έρευνας και βελτίωσε την απόδοση της αναγνώρισης. Επίσης, η πιθανοκρατική διατύπωση των HMMs μας παρέχει ένα ενωτικό πλαίσιο εργασίας για εξαγωγή υποθέσεων και για συνδυασμό διαφορετικών πηγών γνώσεων. Για παράδειγμα, η ακολουθία από ομιλούντες λέξεις μπορεί επίσης να μοντελοποιηθεί σαν την έξοδο μίας άλλης στατιστικής διαδικασίας (Bahl et al., 1983). Με αυτόν τον τρόπο είναι φυσικό να συνδυάσουμε τις HMMs για ομιλία με τα στατιστικά μοντέλα για γλώσσα.

ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΑΝΑΓΝΩΡΙΣΗ

Το σχήμα 6 δείχνει ένα block διάγραμμα ενός γενικού συστήματος για εκπαίδευση και αναγνώριση. Να σημειώσουμε πως και στην εκπαίδευση και στην αναγνώριση το πρώτο βήμα στην διαδικασία είναι να πραγματοποιήσουμε εξαγωγή χαρακτηριστικών στο σήμα ομιλίας.



Σχήμα Error! Unknown switch argument. **Γενικό σύστημα για εκπαίδευση και αναγνώριση**

Στην θεωρία πρέπει να είναι δυνατόν να αναγνωρίζουμε ομιλία άμεσα από το σήμα. Πάντως, λόγω της μεγάλης διακύμανσης του σήματος ομιλίας, είναι καλή ιδέα να πραγματοποιήσουμε κάποιας μορφής εξαγωγή χαρακτηριστικών για να μειώσουμε αυτή την διακύμανση. Ειδικότερα, υπολογίζοντας το παράθυρο του βραχυπρόθεσμου φάσματος έχουμε μεγάλη μείωση της διακύμανσης εξομαλύνοντας το λεπτομερές φάσμα, έτσι εξαλείφοντας αρκετά χαρακτηριστικά της πηγής, έτσι ώστε αν ο ήχος να είναι έμφωνος ή δασύς, και, εάν έμφωνος, αυτό εξαλείφει την επίδραση της περιοδικότητας ή του ύψους του ήχου. Η απώλεια της πληροφορίας πηγής δεν φαίνεται να επηρεάζει πολύ την απόδοση της αναγνώρισης γιατί προκύπτει ότι το παράθυρο του φάσματος είναι πολύ συσχετισμένο με την πληροφορία της πηγής.

Ένας λόγος για τον υπολογισμό του βραχυπρόθεσμου φάσματος είναι ότι ο κοχλίας του ανθρώπινου αυτιού πραγματοποιεί μία περίπου συχνοτική ανάλυση. Η ανάλυση στον κοχλία λαμβάνει μέρος σε μία μη γραμμική κλίμακα συχνοτήτων (γνωστή σαν Bark κλίμακα ή mel κλίμακα). Αυτή η κλίμακα είναι σχεδόν γραμμική μέχρι περίπου τα 1000Hz και είναι περίπου λογαριθμική από εκεί και πέρα. Έτσι, στην εξαγωγή χαρακτηριστικών, είναι πολύ σύνηθες να πραγματοποιούμε μία αναδίπλωση συχνότητας του άξονα συχνοτήτων αφού υπολογίσουμε το φάσμα.

Οι ερευνητές έχουν πειραματιστεί με πολλούς διαφορετικούς τύπους χαρακτηριστικών για χρήση με HMMs (Rabiner και Juang, 1993). Παραλλαγές στον βασικό υπολογισμό του φάσματος, όπως ο συνυπολογισμός του χρόνου και η απόκρυψη συχνότητας, έχουν δείξει ότι παρέχουν κάποια πλεονεκτήματα σε ορισμένες περιπτώσεις. Η χρήση ακουστικών μοντέλων σαν βάση για εξαγωγή χαρακτηριστικών είναι χρήσιμη σε μερικά συστήματα (Cohen, 1989), ειδικά σε θορυβώδη περιβάλλοντα (Hunt et al., 1991).

Ίσως τα πιο δημοφιλή χαρακτηριστικά που χρησιμοποιούνται για αναγνώριση ομιλίας με HMMs σήμερα είναι γνωστά σαν συντελεστές cepstrum mel-συχνότητας ή MFCCs (Davis και Mermelstein, 1980). Μετά την αναδίπλωση mel-κλίμακας του φάσματος, υπολογίζεται ο λογάριθμος του φάσματος και προκύπτει ένας αντίστροφος μετασχηματισμός Fourier στο cepstrum. Διατηρώντας την πρώτη δωδεκάδα περίπου συντελεστών του cepstrum, κάποιος θα μπορούσε να διατηρήσει την πληροφορία του φασματικού παραθύρου που επιθυμεί. Τα χαρακτηριστικά που προκύπτουν είναι τα MFCCs, τα οποία αντιμετωπίζονται σαν ένα διάνυσμα και τυπικά υπολογίζονται για κάθε πλαίσιο των 10ms. Αυτά τα διανύσματα χαρακτηριστικών σχηματίζουν την είσοδο των συστημάτων εκπαίδευσης και αναγνώρισης.

Εκπαίδευση

Εκπαίδευση είναι η διαδικασία της προσέγγισης των παραμέτρων του μοντέλου ομιλίας από πραγματικά δεδομένα ομιλίας. Για προετοιμασία εκπαίδευσης, χρειάζεται το κείμενο της ομιλίας εκπαίδευσης και ένα λεξιλόγιο όλων των λέξεων της εκπαίδευσης, μαζί με τις προφορές τους, γραμμένες σαν φωνητικές ορθογραφίες. Έτσι, μία εγγραφή της ομιλίας εκπαίδευσης γίνεται ακούγοντας την ομιλία και γράφοντας την ακολουθία των λέξεων. Όλες οι ευδιάκριτες λέξεις τότε τοποθετούνται σε ένα λεξιλόγιο και κάποιος πρέπει να παράγει μία φωνητική ορθογραφία για κάθε μία. Σε περιπτώσεις όπου μία λέξη έχει περισσότερες της μίας προφορές, για κάθε λέξη συμπεριλαμβάνονται τόσες φωνητικές ορθογραφίες όσες και οι προφορές. Αυτές οι φωνητικές ορθογραφίες μπορούν να αποκτηθούν από υπάρχοντα λεξικά ή μπορούν να γραφτούν από οποιονδήποτε με ελάχιστη εκπαίδευση στη φωνητική.

Φωνητικά HMMs και Λεξιλόγιο

Δοθέντος της ομιλίας εκπαίδευσης, του κειμένου της ομιλίας, και του λεξιλογίου των φωνητικών ορθογραφιών όλων των λέξεων, οι παράμετροι όλων των φωνητικών HMMs (πιθανότητες μετάβασης και εξόδου) υπολογίζονται αυτόματα χρησιμοποιώντας μία επαναληπτική διαδικασία γνωστή σαν Baum-Welch ή μπρος-πίσω αλγόριθμος (Baum και Eagon, 1967). Αυτός ο αλγόριθμος υπολογίζει τις παραμέτρους των HMMs έτσι ώστε να μεγιστοποιείται η πιθανότητα, η ομιλία εκπαίδευσης όντως να παράγεται από αυτά τα HMMs. Η επαναληπτική διαδικασία εγγυάται να συγκλίνει σε ένα τοπικό βέλτιστο. Τυπικά, γύρω στις πέντε επαναλήψεις μέσω των δεδομένων χρειάζονται για να πάρουμε μία λογικά καλή εκτίμηση του μοντέλου ομιλίας. (Δείτε την εργασία του Jelinek σε αυτόν τον τόμο για περισσότερες πληροφορίες για τον αλγόριθμο εκπαίδευσης HMM.)

Είναι σημαντικό να τονίσουμε το γεγονός ότι η εκπαίδευση HMM δεν απαιτεί τα δεδομένα να περιγράφονται λεπτομερώς σε όρους της θέσης των διαφόρων λέξεων και φωνημάτων, δηλαδή, δεν χρειάζεται ευθυγράμμιση χρόνου μεταξύ της ομιλίας και του κειμένου. Δοθέντος μίας λογικής αρχικής εκτιμήσεως των HMM παραμέτρων, ο Baum-Welch αλγόριθμος εκπαίδευσης πραγματοποιεί μία δυνατή ευθυγράμμιση της ακολουθίας του φάσματος εισόδου στις καταστάσεις του HMM, η οποία χρησιμοποιείται μετά για να πάρουμε μία βελτιωμένη προσέγγιση. Ότι χρειάζεται επιπλέον της ομιλίας εκπαίδευσης είναι η εγγραφή του κειμένου και το λεξιλόγιο. Αυτά είναι από τα πιο σημαντικά χαρακτηριστικά της προσέγγισης HMM για την αναγνώριση. Η εκπαίδευση απαιτεί σημαντικό μέγεθος υπολογισμών αλλά δεν απαιτεί πολλά σε ανθρώπινη εργασία.

Προετοιμάζοντας την αναγνώριση είναι σημαντικό το λεξιλόγιο να περιέχει λέξεις που μπορεί να προκύψουν σε μελλοντικά δεδομένα, αν και δεν προέκυψαν στην εκπαίδευση. Τυπικά, κλάσεις λέξεων κλειστών συνόλων συμπληρώνονται – για παράδειγμα, οι μέρες της εβδομάδας, οι μήνες του χρόνου, οι αριθμοί.

Αφού συμπληρωθεί το λεξιλόγιο, τα HMM μοντέλα λέξεων μεταγλωττίζονται από το σύνολο των φωνητικών μοντέλων χρησιμοποιώντας τις φωνητικές ορθογραφίες στο λεξιλόγιο. Αυτά τα μοντέλα λέξεων είναι απλά μία συγκέντρωση των κατάλληλων φωνητικών HMM μοντέλων. Τότε μεταγλωττίζουμε την γραμματική (η οποία ορίζει ακολουθίες φωνημάτων για κάθε λέξη) σε μία πιθανοκρατική γραμματική για την ακολουθία των φωνημάτων. Το αποτέλεσμα της αναγνώρισης είναι μία ειδική ακολουθία λέξεων, που αντιστοιχεί στην αναγνωριζόμενη ακολουθία φωνημάτων.

Γραμματική

Μία άλλη όψη της εκπαίδευσης που χρειάζεται να βοηθήσει στην αναγνώριση είναι να παράγει την γραμματική που χρησιμοποιείται στην αναγνώριση. Χωρίς γραμματική, όλες οι λέξεις θα θεωρούνται εξ ίσου ίδιες σε κάθε σημείο στην προφορά, το οποίο θα έκανε την αναγνώριση δύσκολη, ειδικά με μεγάλα λεξιλόγια. Εμείς, σαν άνθρωποι, κάνουμε μεγάλη χρήση της γνώσης για τη γλώσσα για να μας βοηθήσει να αναγνωρίσουμε τι λέει ένα πρόσωπο. Μία γραμματική τοποθετεί περιορισμούς στις ακολουθίες των λέξεων που επιτρέπονται, δίνοντας στην αναγνώριση λιγότερες επιλογές σε κάθε σημείο της προφοράς και, επομένως, βελτιώνοντας την απόδοση της αναγνώρισης.

Οι περισσότερες γραμματικές που χρησιμοποιούνται στην αναγνώριση ομιλίας αυτές τις μέρες είναι στατιστικές γραμματικές Markov που δίνουν τις πιθανότητες διαφορετικών ακολουθιών λέξεων- έτσι καλούνται n-γράμματες γραμματικές. Για παράδειγμα, διγράμματες γραμματικές δίνουν τις πιθανότητες για όλα τα ζευγάρια των λέξεων, ενώ τριγράμματες γραμματικές δίνουν τις πιθανότητες για όλες τις τριπλέττες των λέξεων στο λεξιλόγιο. Πρακτικά, τριγράμματα εμφανίζονται να αρκούν να κωδικοποιήσουν πολλούς από τους φυσικούς περιορισμούς που μπαίνουν σε ακολουθίες λέξεων σε μία γλώσσα. Σε μία n-γράμματη γραμματική Markov, η πιθανότητα μίας λέξης είναι συνάρτηση των προηγούμενων n-1 λέξεων. Ενώ αυτή η υπόθεση μπορεί να μην ισχύει γενικά, εμφανίζεται να αρκεί για να πετύχουμε καλή ακρίβεια αναγνώρισης. Επιπλέον, αυτή η υπόθεση μας επιτρέπει επαρκείς υπολογισμούς της πιθανότητας μίας ακολουθίας λέξεων.

Ένα μέτρο του πόσο περιορισμένη είναι μία γραμματική δίνεται από την *περιπλοκή* της (Bahl et al., 1983). Η περιπλοκή ορίζεται σαν το 2 υψωμένο στην δύναμη της εντροπίας Shannon της γραμματικής. Αν όλες οι λέξεις είναι εξ ίσου ίδιες σε κάθε σημείο στην πρόταση, η περιπλοκή είναι ίση με το μέγεθος του λεξιλογίου. Πρακτικά, οι ακολουθίες των λέξεων έχουν αρκετά διαφορετικές πιθανότητες, και η περιπλοκή είναι συχνά πολύ μικρότερη από το μέγεθος του λεξιλογίου, ειδικά για μεγάλα λεξιλόγια. Επειδή οι γραμματικές υπολογίζονται από ένα σύνολο δεδομένων εκπαίδευσης, συχνά είναι σημαντικό να μετράμε την περιπλοκή σε ένα ανεξάρτητο σύνολο δεδομένων, ή όπως είναι γνωστό σαν σύνολο ελέγχου περιπλοκής (Bahl et al., 1983). Η περιπλοκή συνόλου ελέγχου Q υπολογίζεται από τον τύπο:

$$Q = P\left(w_1 w_2 \ldots w_M\right)^{-1/M}$$

όπου $w_1 w_2 \ldots w_M$ είναι η ακολουθία των λέξεων που υπολογίζεται συνδέοντας αλυσιδωτά όλες τις προτάσεις ελέγχου και P είναι η πιθανότητα όλης της ακολουθίας. Λόγω του Markov χαρακτηριστικού των n-γράμματων γραμματικών, η πιθανότητα P μπορεί να υπολογιστεί σαν το γινόμενο των συνεχόμενων υποθετικών πιθανοτήτων των n-γράμματων γραμματικών.

Αναγνώριση

Όπως φαίνεται στο Σχήμα 6, η διαδικασία της αναγνώρισης ξεκινά με το στάδιο της εξαγωγής των χαρακτηριστικών, το οποίο είναι όμοιο με αυτό στην εκπαίδευση. Τότε, δεδομένων της ακολουθίας των διανυσμάτων χαρακτηριστικών, των HMM μοντέλων λέξεων, και της γραμματικής, η αναγνώριση είναι απλά μία μεγάλη έρευνα μεταξύ όλων των πιθανών ακολουθιών λέξεων για την ακολουθία λέξεων με την μεγαλύτερη πιθανότητα που έχει παράγει την υπολογισμένη ακολουθία διανυσμάτων χαρακτηριστικών. Θεωρητικά η έρευνα είναι εκθετική με τον αριθμό των λέξεων στην έκφραση. Πάντως, λόγω του Markov χαρακτηριστικού της υποθετικής ανεξαρτησίας στο HMM, είναι δυνατόν να μειώσουμε την έρευνα δραστικά χρησιμοποιώντας δυναμικό προγραμματισμό (χρησιμοποιώντας τον Viterbi αλγόριθμο). Ο Viterbi αλγόριθμος απαιτεί υπολογισμούς που είναι ανάλογοι του αριθμού των καταστάσεων στο μοντέλο και του μήκους της ακολουθίας εισόδου. Περισσότερο προσεγγιστικοί ερευνητικοί αλγόριθμοι έχουν αναπτυχθεί για να επιτρέψουν τους υπολογισμούς της έρευνας να μειωθούν περισσότερο, χωρίς σημαντικές απώλειες στην απόδοση. Η πιο πολύ χρησιμοποιούμενη τεχνική είναι η έρευνα beam (Lowerre, 1976), η οποία αποφεύγει τους υπολογισμούς για καταστάσεις που έχουν μικρή πιθανότητα.

Όπως είπαμε τα πιο επιτυχημένα συστήματα αναγνώρισης συνεχούς ομιλίας σήμερα βασίζονται σε κρυμμένα μοντέλα Markov (HMMs). Μιας και συζητήσαμε τα βασικά για τα HMMs θα προχωρήσουμε τώρα στις τεχνικές και στους αλγόριθμους που χρησιμοποιούνται στην βασισμένη

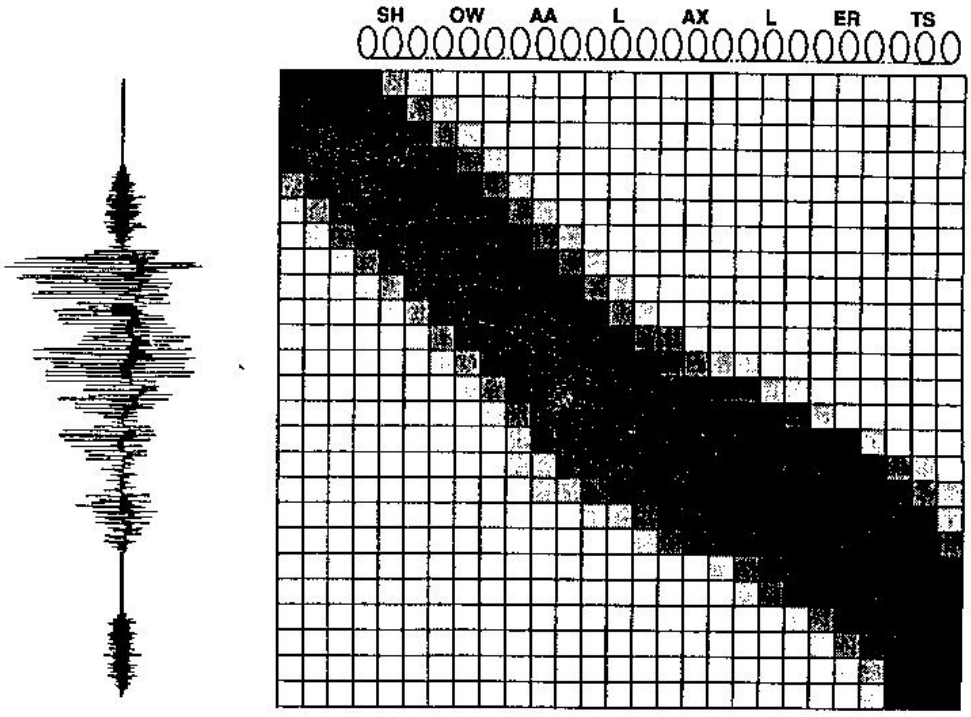
σε HMM αναγνώριση συνεχούς ομιλίας. Υπάρχουν τρία προβλήματα που πρέπει να λυθούν για να επεκτείνουμε την αναγνώριση ομιλίας στην συνεχή ομιλία: (1) το πρόβλημα εκπαίδευσης, ή πώς να εκπαιδεύσουμε HMMs για συνεχή ομιλία, (2) το πρόβλημα μοντελοποίησης, ή πώς να εμπλουτίσουμε την ακουστική-φωνητική μοντελοποίηση για συνεχή ομιλία, και (3) το πρόβλημα αναγνώρισης, ή πώς να οργανώσουμε την έρευνα και μειώσουμε την έκταση της έρευνας για να αντιμετωπίσουμε την συνδυαστική έκρηξη στην συνεχή ομιλία.

1. ΤΟ ΠΡΟΒΛΗΜΑ ΕΚΠΑΙΔΕΥΣΗΣ

Πολλές τεχνικές αναγνώρισης ομιλίας αντιμετωπίζουν προβλήματα στην εκπαίδευση για συνεχή ομιλία γιατί τα όρια των λέξεων δεν είναι αυτόματα ανιχνεύσιμα. Συχνά, χρειάζεται χειροκίνητη διάκριση, το οποίο είναι βαρετό και μπορεί να είναι σχεδόν ευνοϊκότερο. Τα HMMs, πάντως, δεν έχουν αυτό το πρόβλημα. Συγκεκριμένα, τα HMMs εκπαίδευσης στην συνεχή ομιλία δεν διαφέρουν πολύ από την εκπαίδευση σε απομονωμένες λέξεις.

Μιας και η ακολουθία καταστάσεων είναι κρυμμένη στα HMMs, *δεν πειράζει που είναι τα όρια των λέξεων*. Απλά προσποιούμαστε ότι μία πρόταση είναι μία λέξη. Για κάθε πρόταση, χρειαζόμαστε μόνο την ορθογραφική εγγραφή. Πρώτα, κάθε λέξη στιγματοποιείται με το HMM της (το οποίο μπορεί, διαδοχικά, να είναι μία αλληλουχία μοντέλων υπολέξεων). Μετά, οι λέξεις στην πρόταση συνδέονται αλυσιδωτά με επιλεκτικά μοντέλα σιωπής μεταξύ των λέξεων, όπως επίσης στην αρχή και στο τέλος της πρότασης. Αυτό το μεγάλο HMM συνδεμένης αλυσιδωτά πρότασης τότε εκπαιδεύεται σε ολόκληρη την πρόταση χρησιμοποιώντας τον αλγόριθμο μπρος-πίσω (Bahl et al., 1983). Το Σχήμα 7 δείχνει την αλληλουχία των ακουστικών μοντέλων σε ένα μοντέλο πρότασης και την ευθυγράμμιση των καταστάσεων με την ομιλία.

Εναλλακτικά, μπορεί να χρησιμοποιηθεί ο κατατμημένος αλγόριθμος K-μέσων (Rabiner et al., 1986). Αυτός ο αλγόριθμος πρώτα ευθυγραμμίζει κάθε πλαίσιο ομιλίας με μία κατάσταση και ξαναυπολογίζει όλες τις παραμέτρους δοθέντων των ευθυγραμμίσεων. Επαναληπτική εκτέλεση αυτού του αλγορίθμου βελτιώνει τις ευθυγραμμίσεις, όπως επίσης και τα όρια των λέξεων. Αποφέρει συγκρίσιμα αποτελέσματα με τον αλγόριθμο μπρος-πίσω για HMMs συνεχών πυκνοτήτων.



Σχήμα Error! Unknown switch argument. **Η ευθυγράμμιση των καταστάσεων σε σχέση με την ομιλία χρησιμοποιώντας τον μπρος-πίσω αλγόριθμο. Τα σκοτεινότερα τετράγωνα αντιστοιχούν σε μεγάλο α και αναπαριστούν περισσότερο αληθοφανείς ευθυγραμμίσεις.**

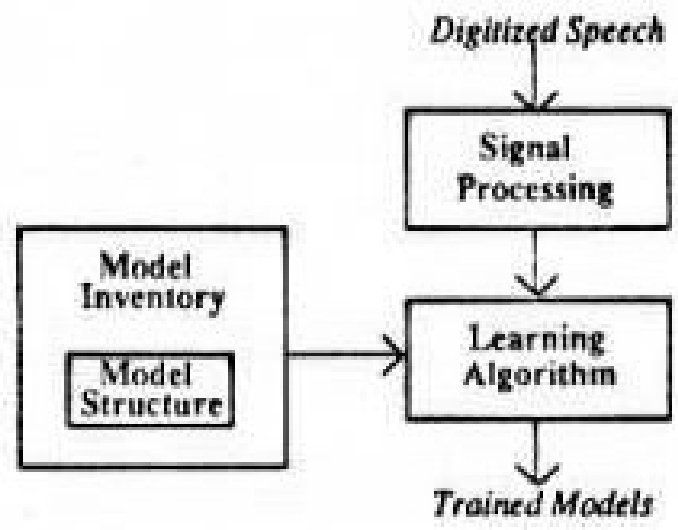
Και για τους δύο αλγόριθμους, η εκπαίδευση στην αναγνώριση ομιλίας απαιτεί σχεδόν καμία αλλαγή στον αλγόριθμο ή στον κώδικα. Η μόνη διαφορά είναι η δημιουργία ενός HMM πρότασης που μελετά πολλούς τρόπους που μία πρόταση μπορεί να λεχθεί. Αυτό είναι ένα από τα πιο σημαντικά πλεονεκτήματα στην χρήση HMMs για αναγνώριση συνεχούς ομιλίας.

2. ΤΟ ΠΡΟΒΛΗΜΑ ΜΟΝΤΕΛΟΠΟΙΗΣΗΣ

Αυτό το τμήμα αφορά εμπλουτισμένες τεχνικές κρυμμένης Markov μοντελοποίησης για υψηλής απόδοσης αναγνώριση συνεχούς ομιλίας. Το Σχήμα 3 δείχνει ένα block διάγραμμα ενός συστήματος αναγνώρισης συνεχούς ομιλίας βασισμένο σε HMMs. Αυτό το σχήμα δείχνει τέσσερις περιοχές έρευνας που μπορεί να αυξήσουν την απόδοση του συστήματος: (1) επεξεργασία και απεικόνιση σήματος, (2) την HMM δομή, (3) τον αλγόριθμο μάθησης, και (4) τον κατάλογο των μονάδων υπολέξεων. Τώρα θα συζητήσουμε καθεμία από αυτές τις περιοχές.

2.1 Βελτιωμένη Απεικόνιση Σήματος

Αν και οι HMM αλγόριθμοι μάθησης είναι ισχυροί, αν η απεικόνιση σήματος δεν παρέχει επαρκείς πληροφορίες, τα HMMs δεν μπορούν να πραγματοποιήσουν τις σωστές διακρίσεις.



Σχήμα Error! Unknown switch argument. **Μία βασισμένη σε HMM αρχιτεκτονική CSR**

Το πιο σημαντικό θέμα για να εξετάσουμε εδώ είναι καθαρά, "ποια είναι η πιο καλή απεικόνιση ομιλίας;". Δυστυχώς, δεν υπάρχει τελεσίδικη απάντηση σε αυτήν την ερώτηση, λόγω έλλειψης εργασίας στις σχετικές μελέτες. Δύο από τις εξαιρέσεις είναι οι Davis and Mermelstein (1980) και Rabiner et al., (1985). Οι Davis and Mermelstein (1980) βρήκαν ότι το cepstrum της mel-συχνότητας έδωσε την καλύτερη απόδοση, ενώ ο Rabiner et al., (1985) έδειξε ότι το LPC cepstrum υπερτέρησε τεσσάρων άλλων παραμέτρων. Και οι δύο απεικονίσεις έχουν χρησιμοποιηθεί από έναν αριθμό από συστήματα επιτυχούς αναγνώρισης ομιλίας (Rabiner et al., 1988; Lee, 1989; Chow et al., 1987; Paul and Martin, 1988). Πιο πρόσφατα, ακουστικά μοντέλα (Seneff, 1988; Cohen, 1989) βασισμένα στην ανθρώπινη ακοή έχουν προταθεί, αλλά δεν έχουν δείξει ακόμα υπερβολικής απόδοσης βελτιώσεις.

Για τα διακριτά συστήματα HMM, μία άλλη ευκαιρία να βελτιώσουμε την απεικόνιση σήματος είναι με την διαδικασία κβαντοποίησης διανύσματος. Η κβαντοποίηση διανύσματος είναι ένα στάδιο προεπεξεργασίας που αντικαθιστά κάθε πλαίσιο ομιλίας με τον δείκτη του κοντινότερου ταιριαστού κεντροειδούς διανύσματος. Τα κεντροειδή διανύσματα προϋπολογίζονται χρησιμοποιώντας έναν αλγόριθμο συσταδοποίησης. Η μετρική απόσταση που χρησιμοποιείται και στην συσταδοποίηση και κβάντιση είναι αναίσθητη στον τύπο των γεγονότων ομιλίας που συγκρίνονται. Επιπλέον, είναι αμερόληπτη σε σχέση με το HMM μάθησης, και λάθη σε αυτό το

στάδιο μπορεί να είναι επιζήμια για το HMM μάθησης αργότερα. Εν όψει αυτού του προβλήματος, έχουν προταθεί τεχνικές για να εμβάλλουν περισσότερες φωνητικές πληροφορίες στα HMMs. Μία μέθοδος είναι ο αλγόριθμος του Kohonen για κβάντιση διανύσματος μάθησης (Iwamida et al., 1990) για οργάνωση πολλαπλών κεντροειδών για κάθε φώνημα για αύξηση της διάκρισης. Μία άλλη είναι η κβάντιση επιβλέποντος διανύσματος (Bahl et al., 1989a). Η κβάντιση επιβλέποντος διανύσματος επαναληπτικά βελτιώνει τα κεντροειδή της χρησιμοποιώντας φωνητικά ονομαζόμενα πλαίσια εκπαίδευσης για βελτίωση της ακουστικής συνάφειας των κεντροειδών. Και οι δύο αλγόριθμοι κβάντισης διανύσματος σκοπεύουν στην αύξηση της φωνητικής πληροφορίας στο στάδιο κβάντισης διανύσματος.

Μία από τις μεγαλύτερες βελτιώσεις στα συστήματα αναγνώρισης συνεχούς ομιλίας είναι η ενσωμάτωση των χρονικών αλλαγών στο φάσμα. Μία από τις πρώτες προσπάθειες που ενσωμάτωσαν δυναμικές αλλαγές ήταν η χρήση από τον Furui του *δέλτα cepstrum* (Furui, 1986), ή μία επικλινής μέτρηση υπολογισμένη πάνω σε ένα παράθυρο γύρω στα πέντε πλαίσια (50ms). Πιο απλές μετρήσεις όπως διαφορικοί συντελεστές (Shikano et al., 1986) έχουν επίσης χρησιμοποιηθεί με επιτυχία. Αυτές οι διαφορικές παράμετροι όχι μόνο παρέχουν επιπρόσθετες πληροφορίες για την κίνηση του φάσματος, αλλά ευρύνουν επίσης το πεδίο δράσης ενός μοναδικού πλαισίου που είναι παρών στο HMM. Κάτι εναλλακτικό σε αυτούς τους διαφορικούς συντελεστές είναι να συνδέσουμε αλυσιδωτά πλαίσια και μετά να μειώσουμε την διάσταση (Brown, 1987). Η ενσωμάτωση μεταβατικής πληροφορίας έχει μειώσει τα σφάλματα περίπου 50% (Rabiner et al., 1988; Lee, 1989).

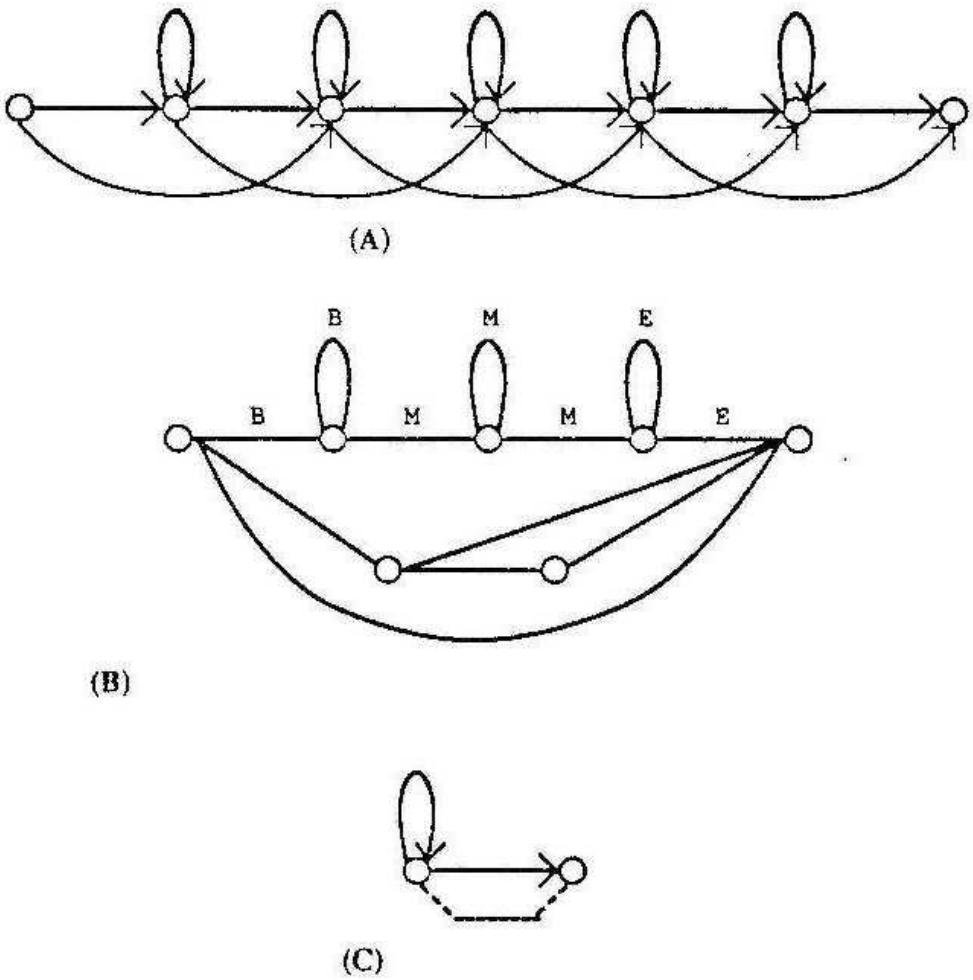
Ο συνδυασμός ετερογενών παραμέτρων έχει επίσης βρεθεί ότι βοηθάει. Ισχύς και διαφορική ισχύς (Lee 1989) έχουν χρησιμοποιηθεί επιτυχώς σε ένα LPC σύστημα, το ποίο δεν έχει επαρκές μοντέλο ισχύος. Μια περισσότερο φιλόδοξη προσπάθεια ενοποίησε πολλά διαφορετικά χαρακτηριστικά, συμπεριλαμβανομένων του επίπεδου ομιλίας, της ταχύτητας μετάβασης φάσματος, της ενέργειας πλαισίου και φίλτρου, και της ενέργειας διαφορικού πλαισίου και φίλτρου (Doddington, 1989). Αν και αυτές οι μετρήσεις είναι πιθανότατα συσχετισμένες, η χρήση τεχνικών μείωσης διάστασης, όπως της ανάλυσης κύριων συνιστωσών, τις αποσυσχετίζει αποτελεσματικά.

2.2 Βελτιωμένες HMM Δομές

Το προηγούμενο τμήμα ασχολήθηκε με το πώς να προετοιμάσομε τις παραμέτρους ομιλίας για κρυμμένη Markov μοντελοποίηση. Σε αυτό το τμήμα, θα συζητήσουμε αρκετές τεχνικές που βελτιώνουν την κρυμμένη Markov μοντελοποίηση για καλύτερη αναγνώριση συνεχούς ομιλίας. Τα "βελτιωμένα" HMM θα πρέπει να είναι ένα περισσότερο "σωστό" μοντέλο ομιλίας. Για παράδειγμα, εάν ένα φώνημα υποβάλλεται σε πολλές αλλαγές προσωρινά, πρέπει να τοποθετηθούν σε αυτό αρκετές καταστάσεις. Ή εάν ένα μοντέλο πρέπει να αφομοιώσει αρσενικές και θηλυκές φωνές, πρέπει να ανατεθούν αρκετές κωδικολέξεις (για διακριτό HMM) ή μείγματα (για συνεχές HMM) για να μοντελοποιήσουν μία διωνυμική κατανομή. Το να έχουμε το σωστό μοντέλο είναι πολύ σημαντικό, μιας και η τεχνική υπολογισμού μέγιστης πιθανότητας που χρησιμοποιείται στην μάθηση HMM βασίζεται σε σωστές υποθέσεις μοντέλου (Brown, 1987).

Δοθέντων επαρκών δεδομένων, τα κρυμμένα μοντέλα Markov θα αποδίδουν καλύτερη απόδοση με καλύτερα μοντέλα. Αλλά πως κάποιος κατασκευάζει καλύτερα μοντέλα; Ένας προφανής τρόπος είναι να αυξήσει τον αριθμό των καταστάσεων. Περισσότερες καταστάσεις μπορούν να μοντελοποιήσουν καλύτερα μέσα σε παραλλαγές μονάδων, το οποίο βελτιώνει την "ορθότητα" του μοντέλου. Ο Deng et al. (1989) βρήκε ότι προσαρμόζοντας το μήκος του μοντέλου στην μέση διάρκεια του φωνήεντος αυξάνει την απόδοση. Σε αυτήν την μελέτη, ο Deng et al. χρησιμοποίησε το μοντέλο Bakis, όπου κάθε κατάσταση έχει μία μετάβαση *σταματήματος*, μία *προοδευτική* μετάβαση, και μία μετάβαση *υπερπήδησης*. Ένα παράδειγμα του μοντέλου Bakis φαίνεται στο Σχήμα 9A. Ο Doddington (1989) επίσης χρησιμοποίησε το μοντέλο Bakis και βρήκε ότι τα καλύτερα αποτελέσματα παράγονται όταν ο αριθμός των καταστάσεων αντιστοιχούν στον μέσο αριθμό

πλαισίων για τις μονάδες που είναι να μοντελοποιηθούν.* Επιπλέον, η λεπτομερής μοντελοποίηση ενισχύθηκε επιτρέποντας μία μέγιστη διάρκεια καταστάσεων τριών πλαισίων.



Σχήμα Error! Unknown switch argument. **Τρεις διαφορετικοί τύποι τοπολογιών HMM**

Ένα πρόβλημα με το Bakis (ή οποιοδήποτε γραμμικό) μοντέλο είναι ότι δεν θα είναι ικανό να αναγνωρίζει δείγματα μικρότερα από την μικρότερη διαδρομή υπερπήδησης. Μία πιθανή λύση είναι η δομή μοντέλου που φαίνεται στο Σχήμα 9Β. Αυτή η δομή έχει σαφείς διαδρομές για δείγματα μικρότερα από τέσσερα πλαίσια και έχει δώσει καλά αποτελέσματα για φωνητική μοντελοποίηση (Lee, 1989). Εναλλακτικά, οι τροχιές υπερπήδησης, ή οι τροχιές που κάνουν δυνατή την κίνηση από την μία κατάσταση σε μία άλλη χωρίς να καταναλώνουν ένα πλαίσιο εισόδου, μπορούν να χρησιμοποιηθούν για να ενεργοποιήσουν μία πιο κατάλληλη εκκίνηση από ένα μοντέλο. Ο Bahl et al., (1988a) χρησιμοποίησε μοντέλα που μπορούσαν τελείως να υπερπηδηχτούν για να μοντελοποιήσουν fenones (πολύ μικρά ακουστικά γεγονότα). Πάντως, ενώ τα fenones έχουν επιτυχώς εφαρμοστεί στην αναγνώριση απομονωμένων λέξεων, δεν έχουν δείξει συγκρίσιμα αποτελέσματα στην αναγνώριση συνεχούς ομιλίας.

Μία άλλη περιοχή καλύτερης μοντελοποίησης είναι στο επίπεδο πυκνότητας πιθανότητας εξόδου. Για διακριτά κρυμμένα μοντέλα Markov, ένα μεγαλύτερο κωδικοβιβλίο αυξάνει την ακρίβεια στην ακουστική μοντελοποίηση. Τυπικά, γύρω στις 200-300 κωδικολέξεις χρησιμοποιούνται. Πάντως, για τις περισσότερες αποστολές, τα μοντέλα γίνονται ανεκπαίδευτα εάν αυξήσουμε σημαντικά τον αριθμό των κωδικολέξεων. Για την αντιμετώπιση αυτού του προβλήματος, *πολλαπλά κωδικοβιβλία* έχουν προταθεί (Gupta et al., 1987). Αρκετά κωδικοβιβλία δημιουργούνται, καθένα

* Να σημειώσουμε ότι πρέπει να υπάρχουν επαρκή δεδομένα για να μοντελοποιήσουν αυτό το επίπεδο λεπτομέρειας. Για την μελέτη του Doddington (1989) περισσότερες από 8000 προτάσεις χρησιμοποιήθηκαν για να εκπαιδεύσουν 29 μοντέλα.

μοντελοποιώντας ένα υποσύνολο των συντελεστών της ομιλίας. Υποθέτοντας πως τα πολλαπλά κωδικοβιβλία είναι ανεξάρτητα, ο μπρος-πίσω αλγόριθμος μπορεί να τροποποιηθεί να εκπαιδεύσει αυτές τις πολλαπλές συναρτήσεις πυκνότητας πιθανότητας (pdfs) ταυτόχρονα. Η αρχή των πολλαπλών κωδικοβιβλίων μπορεί να συνδυαστεί με μεταβατικές και άλλες παραμέτρους. Για παράδειγμα, ένα κωδικοβιβλίο μπορεί να χρησιμοποιηθεί για στιγμιαίες παραμέτρους, άλλο για διαφορικές παραμέτρους, και ακόμα ένα τρίτο για ισχύ (Lee, 1989).

Για HMMs συνδυασμένης Γκαουσιανής συνεχής πυκνότητας, ο αριθμός των συνδυασμών ανά κατάσταση μπορεί να αυξηθεί για μοντελοποίηση ακουστικής με περισσότερη λεπτομέρεια. Ο Rabiner et al. (1988) χρησιμοποίησε μέχρι εννιά Γκαουσιανούς συνδυασμούς ανά κατάσταση σε έναν αναγνωριστή συνεχών ψηφίων. Πάντως, προσθέτοντας περισσότερους συνδυασμούς ανά κατάσταση έχουμε αύξηση των παραμέτρων πολύ γρήγορα. Ένας τρόπος να ελέγξουμε την αύξηση των παραμέτρων είναι να χρησιμοποιήσουμε *ημισυνεχή κρυμμένα μοντέλα Markov* (Huang and Jack, 1989; Bellegarda and Nahamoo, 1989), έναν συμβιβασμό μεταξύ των συνεχών και διακριτών HMMs. Το ημισυνεχές κρυμμένο μοντέλο Markov είναι ένα HMM συνεχούς πυκνότητας, όπου οι συνδυασμοί μοιράζονται από όλες τις καταστάσεις. Εάν δούμε αυτούς τους συνδυασμούς σαν κωδικοβιβλίο, τότε το ημισυνεχές κρυμμένο μοντέλο Markov είναι επίσης παρόμοιο με το διακριτό HMM, εκτός του ότι χρησιμοποιούνται κωδικοβιβλία συνεχούς πυκνότητας και πολλές κωδικολέξεις μπορούν να αντιστοιχηθούν σε κάθε κατάσταση. Συγκρινόμενο με το HMM συνεχούς πυκνότητας, το ημισυνεχές κρυμμένο μοντέλο Markov μπορεί να έχει ένα καλύτερο ακουστικό μοντέλο γιατί κάθε κατάσταση μπορεί να χρησιμοποιήσει ένα μεγάλο αριθμό συνδυασμών. Την ίδια στιγμή, το ημισυνεχές κρυμμένο μοντέλο Markov είναι περισσότερο εκπαιδεύσιμο γιατί ο αριθμός των ολικών συνδυασμών πιθανών είναι μικρότερος αυτού κανονικών HMMs συνεχούς πυκνότητας εξαρτωμένων από την κατάσταση.

2.3 Βελτιωμένη Μάθηση και Διάκριση

Για μάθηση, τα περισσότερα συστήματα που βασίζονται σε HMM χρησιμοποιούν τον μπρος-πίσω αλγόριθμο (Baum, 1972q Jelinek, 1976; Bahl et al., 1983), ο οποίος προσαρμόζει τις παραμέτρους για να αποκομίσει μία προσέγγιση στις εκτιμήσεις μέγιστης πιθανότητας των HMM παραμέτρων. Οι εκτιμητές μέγιστης πιθανότητας έχουν πολλές επιθυμητές ιδιότητες, και πολλά επιτυχημένα συστήματα (Jelinek, et al., 1985; Chow et al., 1987; Lee, 1989; Rabiner et al., 1988) βασίζονται σε εκτιμητές μέγιστης πιθανότητας. Πάντως, η εκτίμηση μέγιστης πιθανότητας έχει ένα σοβαρό ελάττωμα: υποθέτει ότι τα βασικά μοντέλα είναι σωστά. Στην πραγματικότητα, πάντως, τα τυπικά HMMs κάνουν εξαιρετικά ανακριβείς υποθέσεις για την διαδικασία παραγωγής ομιλίας.

Για την αντιμετώπιση αυτού του προβλήματος, προτάθηκε ο υπολογισμός της μέγιστης κοινής πληροφορίας (Brown, 1987). Ο υπολογισμός της μέγιστης κοινής πληροφορίας ελαχιστοποιεί την πληροφορία που χρειάζεται για να ορίσει την ακολουθία των λέξεων όταν είναι γνωστή η ακουστική ακολουθία. Δοθέντος μίας ακουστικής ακολουθίας Y και της αντίστοιχης ακολουθίας λέξεων W, ο υπολογισμός της μέγιστης κοινής πληροφορίας όχι μόνο προσπαθεί να αυξήσει την Pr(Y|W) (όπως κάνει ο εκτιμητής μέγιστης πιθανότητας), αλλά επίσης προσπαθεί να μειώσει την Pr(Y|W') για λάθος λέξεις W'. Η εκπαίδευση έναντι των αρνητικών υποδειγμάτων αυξάνει την σημασία μιας σωστής αναγνώρισης.

Ένας άλλος εναλλακτικός τρόπος, ο αλγόριθμος διορθωτικής εκπαίδευσης (Bahl et al., 1988b) προσπαθεί να μεγιστοποιήσει τον ρυθμό αναγνώρισης στα δεδομένα εκπαίδευσης. Το κάνει τροποποιώντας τις HMM παραμέτρους, έτσι ώστε όχι μόνο να αυξάνει την πιθανότητα για σωστά δεδομένα εκπαίδευσης αλλά επίσης να μειώνει την πιθανότητα για κακές αναγνωρίσεις. Αν και η διορθωτική εκπαίδευση δεν έχει τις θεωρητικές δικαιολογίες του υπολογισμού μέγιστης κοινής πληροφορίας, συμπεριφέρεται καλά για αναγνώριση απομονωμένων λέξεων (Bahl et al., 1988b), όπως επίσης και για αναγνώριση συνεχούς ομιλίας (Lee and Martin, 1990).

Μία άλλη διαχωρίζουσα προσέγγιση μάθησης βασίζεται στις *γραμμικές διακρίσεις* (Brown, 1987; Doddington, 1989). Αυτή η προσέγγιση είναι παρόμοια με την διορθωτική εκπαίδευση όπου και τα σωστά και τα σχεδόν σωστά δεδομένα χρησιμοποιούνται για να ενισχύσουν την διάκριση. Υπολογίζεται ένας πίνακας μετασχηματισμού έτσι ώστε η διακύμανση για τα σωστά δεδομένα να ελαχιστοποιηθεί, ενώ η διακύμανση μεταξύ των σωστών και των σχεδόν σωστών κλάσεων να

μεγιστοποιηθεί. Ο Brown (1987) χρησιμοποίησε επιτυχώς ένα γραμμικό διαχωριστή για όλα τα πλαίσια ομιλίας, ουσιωδώς σαν μεταεπεξεργαστή. Σε μία περισσότερο περίπλοκη εφαρμογή, ο Doddington (1989) δημιούργησε γραμμικούς διαχωριστές εξαρτημένους από την κατάσταση. Πρώτα, συλλέχτηκαν σωστά και σχεδόν σωστά δείγματα για να υπολογίσουν έναν εντός-κλάσεων και μεταξύ-κλάσεων πίνακα συμμεταβολής για κάθε κατάσταση. Αυτοί οι πίνακες τότε χρησιμοποιήθηκαν για να μετατρέψουν τα δεδομένα ομιλίας ώστε να μεγιστοποιήσουν την μεταξύ-κλάσεων διακύμανση και να ελαχιστοποιήσουν την εντός-κλάσεων διακύμανση για κάθε κατάσταση. Και οι δύο μελέτες έδειξαν μεγάλες μειώσεις στο ρυθμό σφάλματος.

Αυτοί οι αλγόριθμοι προσθέτουν μία ικανότητα διάκρισης στην αναγνώριση συνεχούς ομιλίας. Να σημειώσουμε πως, γενικά, τα καλύτερα αποτελέσματα παίρνονται όχι διακρίνοντας όλες τις λέξεις ή όλα τα φωνήματα αλλά *διαλέγοντας* διακριτές λέξεις και φωνήματα που μπορούν να προκαλέσουν σύγχυση. Δοθέντος αυτού του σκοπού, η απόδοσή τους βελτιώνεται σταθερά για συγκεκριμένους σκοπούς και γραμματικές, πάντως, είναι πιθανόν να είναι λιγότερο ρωμαλέοι, κάνοντας την αλλαγή ή αύξηση των νέων λέξεων δύσκολη.

2.4 Βελτιωμένος Κατάλογος Υπολέξεων

Οι προηγούμενες βελτιώσεις υποθέτουν την ύπαρξη ενός σταθερού καταλόγου υπολέξεων, όπως φωνήματα, τρίφωνα, ή συλλαβές. Σε αυτό το τμήμα, εξετάζουμε πως ο κατάλογος αυτών των μονάδων μπορεί να εμπλουτιστεί.

Για αναγνώριση συνδεδεμένης ομιλίας μικρών λεξιλογίων, οι λέξεις είναι εφαρμόσιμα μοντέλα. Πράγματι, όταν υπάρχει επαρκής εκπαίδευση, τα μοντέλα λεπτομερών λέξεων έχουν οδηγήσει στα καλύτερα αποτελέσματα (Rabiner et al., 1988; Doddington, 1989). Πάντως, για αποστολές μεγάλων λεξιλογίων, τα μοντέλα λέξεων δεν είναι πια εφικτά λόγω του μεγάλου αριθμού παραμέτρων και της έλλειψης επαρκών δεδομένων ανά λέξη.

Τα φωνήματα είναι φυσικές μονάδες ομιλίας. Πάντως, λόγω του ότι τα φωνήματα επηρεάζονται πολύ από το περιβάλλον, τα μοντέλα φωνημάτων έχουν πολύ ευρείς κατανομές και δεν είναι επαρκή για αναγνώριση ομιλίας υψηλής απόδοσης. Η πιο σημαντική επίδραση περιβάλλοντος είναι αυτή του άμεσου φωνητικού γείτονα. Δηλαδή, η πραγματοποίηση ενός φωνήματος επηρεάζεται πολύ από τους αριστερούς και δεξιούς του γείτονες. Αυτή η επίδραση οξύνεται στην συνεχή ομιλία, όπου οι φθόγγοι είναι πιο σύντομοι. Για την αντιμετώπιση αυτού, εισαχθήκανε οι τρίφθογγοι (Schwartz et al., 1985) για να μοντελοποιήσουν κάθε φθόγγο στο περιβάλλον των αριστερών και δεξιών του γειτόνων. Ο Bahl et al., (1989a) επίσης βρήκε ότι ενώ η λεπτομερής μοντελοποίηση λέξης είναι η πιο αποτελεσματική τεχνική για αναγνώριση απομονωμένων λέξεων, για αναγνώριση συνεχούς ομιλίας η μοντελοποίηση περιβάλλοντος έδωσε τα καλύτερα αποτελέσματα. BYBLOS (Kubala et al., 1988), ένα σύστημα αναγνώρισης ομιλίας βασισμένο σε τριφθόγγους, είχε πολύ καλά αποτελέσματα. Πολλά άλλα επιτυχή συστήματα (Paul and Martin, 1988; Lee, 1989; Bernstein et al., 1989; Lee et al., 1990b; Bahl, et al., 1989a) βασίστηκαν σε τριφθόγγους ή σε μονάδες παρόμοιες με αυτούς.

Για τα συστήματα HMM, οι τρίφθογγοι υλοποιούνται αντικαθιστώντας κάθε φθόγγο με το αντίστοιχό του τρίφθογγο. Αλλά τι γίνεται με τους τριφθόγγους ορίων λέξεων; Ο πρώτος φθόγγος μίας λέξης επηρεάζεται καθαρά από την προηγούμενη λέξη, και ο τελευταίος φθόγγος μίας λέξης επηρεάζεται από την επόμενη λέξη. Υλοποίηση αυτής της επίδρασης σχετίζεται με τροποποίηση του αλγόριθμου έρευνας για να χρησιμοποιεί δυναμικά τα σωστά μοντέλα ορίων λέξεων. Μία διαδικασία για μοντελοποίηση τριφθόγγων μεταξύ λέξεων τεκμηριώθηκε στον Hwang et al. (1989).

Ένα άλλο πρόβλημα με τις μονάδες τριφθόγγων είναι ότι υπάρχουν πάρα πολλές από αυτές. Με περίπου 50 φθόγγους, υπάρχουν ενδεχομένως 125000 τρίφθογγοι. Ακόμα και αν απορρίψουμε τους αδύνατους τριφθόγγους, υπάρχουν ακόμα δεκάδες χιλιάδων από τριφθόγγους. Επιπλέον, αν επιθυμούμε να μοντελοποιήσουμε άλλες αλλοφωνικές παραλλαγές (όπως δυναμικό τόνο, όριο λέξης, όριο συλλαβής, ή τετρά-φθογγοι), ο αριθμός των μοντέλων αυξάνει γεωμετρικά. Μία προσέγγιση να ελέγξουμε αυτήν την εξάπλωση των μοντέλων είναι να συνδυάσουμε όμοιους φθόγγους (ή αλλόφωνα). Αυτό είναι φωνητικά εύλογο, διότι πολλοί φθόγγοι είναι όμοιοι. Ο Lee (1990) χρησιμοποίησε μία συσσωρευτική ομαδοποίηση για να ανάγει τους τριφθόγγους σε γενικευμένους

τριφθόγγους. Μία απόσταση μέγιστης πιθανότητας χρησιμοποιήθηκε για να παραμείνει ακριβής στο κριτήριο εκπαίδευσης. Οι γενικευμένοι τρίφθογγοι έχουν οδηγήσει επίσης σε αναγνώριση υψηλής απόδοσης (Lee, 1989).

Εάν το ποσό της εκπαίδευσης είναι τόσο περιορισμένο ακόμα και για να πραγματοποιήσει ομαδοποίηση τριφθόγγων/αλλοφώνων, μπορεί να χρησιμοποιηθεί η ανθρώπινη γνώση για να μειώσει τον αριθμό των περιβαντολλογικών παραλλαγών. Για παράδειγμα, για την αντιμετώπιση κακής αναγνώρισης εκρηκτικών ήχων, ο Derouault (1987) χρησιμοποίησε εκρηκτικά μοντέλα εξαρτώμενα από το περιβάλλον που εξαρτιόντουσαν μόνο από την θέση της άρθρωσης του γειτονικού φωνήεντος. Αυτά τα μοντέλα ουσιαστικά βελτίωσαν την απόδοση στα φωνήεντα. Σαν ένα άλλο παράδειγμα, ο Deng et al., (1989) δημιούργησε μοντέλα πολλαπλών φωνηέντων εξαρτωμένων από το περιβάλλον που μοιράζονταν τις ίδιες πιθανότητες αλλά είχαν τις δικές τους πιθανότητες μετάβασης. Αυτός ο τύπος της εξαρτώμενης από το περιβάλλον μοντελοποίησης είναι κατάλληλος όταν υπάρχουν ανεπαρκή δεδομένα εκπαίδευσης για να υποστηρίξουν εκπαίδευση όλων των παραμέτρων και όταν τα χαρακτηριστικά του φάσματος είναι αμετάβλητα επί του περιβάλλοντος ενώ τα χαρακτηριστικά της διάρκειας είναι μεταβλητά.

Μία άλλη δυσκολία με την συνεχή ομιλία είναι η φτωχή προφορά των λειτουργικών λέξεων. Για να αντιμετωπίσουμε τις λειτουργικές λέξεις, προτάθηκαν *μοντέλα λειτουργικών φθόγγων εξαρτωμένων από λέξεις* (Lee, 1990). Αυτά τα μοντέλα εστιάζουν στο πιο μπερδεμένο υπολεξιλόγιο. Ακόμα μπορούν να εκπαιδευτούν καλά διότι οι λειτουργικές λέξεις προκύπτουν συχνά. Για να εξασφαλίσουμε την ρωμαλεότητα της αναγνώρισης, τα μοντέλα λειτουργικών εξαρτωμένων από λέξεις φθόγγων (όπως άλλα μοντέλα εξαρτώμενα από το περιβάλλον φθόγγων) μπορούν επίσης να παρεμβληθούν με μοντέλα ανεξάρτητα του περιβάλλοντος φθόγγων.

Όλες οι προηγούμενες τεχνικές δημιουργούν ειδικά μοντέλα για ειδικά περιβάλλοντα ή λέξεις. Μία άλλη προσέγγιση είναι να δημιουργήσουμε πολλαπλά μοντέλα για το ίδιο γεγονός. Για παράδειγμα, ένα μοντέλο μπορεί να χρησιμοποιηθεί για αρσενικούς ομιλητές και ένα άλλο για θηλυκούς ομιλητές. Για την αναγνώριση, και τα δύο διαστήματα φύλου ερευνώνται ενώ η συνάφεια του γένους σε κάθε υποθετική πρόταση διατηρείται (Doddington, 1989). Η ομαδοποίηση φύλου είναι μία απλή κανόνων-απορρέουσα ομαδοποίηση. Εναλλακτικά, μπορούν να χρησιμοποιηθούν αυτόματες τεχνικές ομαδοποίησης για να δημιουργήσουν πολλαπλά μοντέλα στο επίπεδο λέξης ή φωνήματος. Ο Rabiner et al., (1989b) βρήκε ότι δημιουργώντας ομάδες για outliers έχουμε σημαντική βελτίωση της ακρίβειας αναγνώρισης.

Τελικά, πολλαπλές προφορές των λέξεων μπορούν να χρησιμοποιηθούν για να διευθετήσουν διάφορους τρόπους που μπορεί μία λέξη να ειπωθεί. Αυτό είναι χρήσιμο για μοντελοποίηση λέξεων με πολλαπλές προφορές (όπως το *the*), ή ακόμα λέξεις που μπορεί να προφερθούν διαφορετικά λόγω διαλέκτου ή συνήθειας (όπως *tomato*). Ο συνδυασμός των φωνολογικών κανόνων και HMMs ερευνήθηκε από τον Bernstein et al., (1989). Βρέθηκε ότι ολική φωνολογική επέκταση αύξησε υπερβολικά τον αριθμό των παραμέτρων. Ο καλύτερος συνδυασμός φωνολογικής και HMM εκπαίδευσης είναι να επιτρέψουμε ένα μικρό αριθμό παραλλαγών ομιλίας (γύρω στις 1.3 προφορές ανά λέξη).

2.5 Περίληψη

Περιγράψαμε έναν αριθμό μεθόδων που βελτιώνουν την ακουστική μοντελοποίηση χρησιμοποιώντας HMMs. Αυτές οι τεχνικές εμπνεύστηκαν από την γνώση ακουστικής-φωνητικής, όπως επίσης και κατανόησης των δυνάμεων και αδυναμιών του HMM. Αυτές οι τεχνικές εμπλουτίζουν την λεπτομέρεια που μοντελοποιείται από HMMs, ενώ κρατάμε τις παραμέτρους εκπαιδεύσιμες. Ο συνδυασμός βελτιωμένων παραμέτρων, βελτιωμένων HMMs, βελτιωμένης μάθησης, και βελτιωμένου καταλόγου υπολέξεων μπορεί να μειώσει τον ρυθμό σφάλματος ενός συστήματος αναγνώρισης συνεχούς ομιλίας κατά μία τάξη μεγέθους (Lee, 1989).

3. ΤΟ ΠΡΟΒΛΗΜΑ ΑΝΑΓΝΩΡΙΣΗΣ

3.1 Περίληψη του Προβλήματος

Ένα παράδειγμα έρευνας απασχολείται για να επιλέξουμε από έναν αριθμό εναλλακτικών μία λύση που να ικανοποιεί κάποια σκόπιμα κριτήρια. Ο απόλυτος στόχος της αναγνώρισης συνεχούς ομιλίας είναι να αποφασίσει μία *σωστή* αποκωδικοποίηση (σε όρους ακολουθίας λέξεων) μίας παρατηρούμενης ακουστικής ακολουθίας. Δυστυχώς, κατασκευάζοντας μία αποτίμηση που να μπορεί πάντα να αναγνωρίζει την σωστή ακολουθία λέξεων είναι αδύνατον. Γι'αυτό, τροποποιούμε τον στόχο του παραδείγματος έρευνας για να επιλέξουμε μία ακολουθία λέξεων που είναι η πιο πιθανή για τα ακουστικά μας μοντέλα δοθέντος της παρατηρούμενης ακουστικής ακολουθίας. Αυτό οδηγεί στον ακόλουθο ορισμό της αναγνώρισης συνεχούς ομιλίας.

Δοθέντος μίας ακουστικής ακολουθίας y, να παράγουμε μία ακολουθία λέξεων w' τέτοια ώστε $P(w'|y) = \max_w P(w|y)$.

Εφαρμόζοντας τον κανόνα Bayes σε αυτόν τον ορισμό της αναγνώρισης συνεχούς ομιλίας παράγουμε την Εξ. (1), ή την θεμελιώδη εξίσωση της αναγνώρισης ομιλίας.

$$P(w|y) = \frac{P(w) \times P(y|w)}{P(y)} \tag{1}$$

Θεμελιώδης εξίσωση της αναγνώρισης ομιλίας

Η εξίσωση (1) μας επιτρέπει να εκφράσουμε την $P(w|y)$ σε όρους που μπορούμε να υπολογίσουμε. Η $P(w)$ είναι η εκ των προτέρων πιθανότητα της ακολουθίας των λέξεων w. Η $P(w)$ υπολογίζεται από το μοντέλο γλώσσας. Η $P(y|w)$ είναι η υποθετική πιθανότητα της ακουστικής ακολουθίας y δοθέντος της ακολουθίας λέξεων w. Η $P(y|w)$ υπολογίζεται από το ακουστικό μοντέλο. Τελικά, η $P(y)$ είναι η πιθανότητα της ακουστικής ακολουθίας y. Το να αποφασίσουμε ένα λογικό μοντέλο για την $P(y)$ είναι δύσκολο και συχνά όχι αναγκαίο, μιας και η $P(y)$ είναι ισοδύναμη για όλες τις πλήρεις αποκωδικοποιήσεις του y. Έτσι, απορρίπτουμε την $P(y)$ και ξαναγράφουμε την Εξ. (1) σαν

$$P(w, y) = P(w) \times P(y|w) \tag{2}$$

3.1.1 Το μοντέλο Γλώσσας

Αν και η ανάπτυξη των μοντέλων γλώσσας καλύπτεται αλλού σε αυτόν τον τόμο, η επιλογή του μοντέλου γλώσσας (ή $P(w)$ στην Εξ. 2) έχει σημαντικό αντίκτυπο στην διαδικασία της αναγνώρισης. Συγκεκριμένα, ο περιορισμός που παρέχεται από το μοντέλο γλώσσας μπορεί ουσιαστικά να βελτιώσει την απόδοση ενός συστήματος, και το μέγεθος του χώρου έρευνας που δημιουργείται από το μοντέλο γλώσσας συχνά αποφασίζει για τον καλύτερο αλγόριθμο έρευνας. Έτσι, έχουμε μία σύντομη περίληψη εδώ.

Τα μοντέλα γλώσσας που χρησιμοποιούνται από συστήματα αναγνώρισης συνεχούς ομιλίας ανήκουν σε πολυάριθμες κατηγορίες.

- *Ομοιόμορφα* μοντέλα γλώσσας, όπου κάθε λέξη σε κάθε πρόταση είναι ισοδύναμα πιθανή. Το μέγεθος του μοντέλου γλώσσας είναι 0 αλλά έτσι είναι και ο περιορισμός.
- *Στοχαστικά* μοντέλα γλώσσας όπως μοντέλα τριγραμμάτων, διγραμμάτων, και μονογραμμάτων (Jelinek et al., 1985), όπου η πιθανότητα μίας λέξης σε μία ακολουθία υπολογίζεται από την πιθανότητα να ανήκει σε μία κλάση ισοδυναμίας ορισμένη από τις προηγούμενες λέξεις σε αυτήν την ακολουθία.
- *Πεπερασμένα* μοντέλα γλώσσας (Baker, 1975; Bahl et al., 1978) είναι απλές τεχνητές γλώσσες που μοντελοποιούν όλες τις νόμιμες προτάσεις χρησιμοποιώντας ένα απλό δίκτυο.
- *Άλλα* δυνατά μοντέλα γλώσσας περιλαμβάνουν ελεύθερα περιβάλλοντος (Ney, 1987), ενοποίησης (Hemphill and Picone, 1989), στατιστικά βασισμένα σε δέντρο (Bahl et al., 1989b), και περιπτώσεις γραμματικών πλαισίων.

Αυτά τα μοντέλα γλώσσας παρέχουν αρκετά επίπεδα περιορισμών στον αναγνωριστή. Οι γραμματικές πεπερασμένων καταστάσεων τείνουν να είναι περιοριστικές, πάντως, δεν είναι ρεαλιστικές γραμματικές για τους περισσότερους σκοπούς. Από την άλλη πλευρά, οι γραμματικές χωρίς περιβάλλοντα και οι γραμματικές ενοποίησης είναι πιο ρεαλιστικές, αλλά έχουν πάρα πολλές κατά στάσεις να απαριθμηθούν. Οι στατιστικές γραμματικές μπορεί να είναι μικρές (μονόγραμματες, δίγραμματες), αλλά καθώς ο αριθμός των προηγούμενων μοντελοποιούμενων λέξεων αυξάνεται σε τριγράμματα, ο αριθμός των καταστάσεων αυξάνει πολύ γρήγορα.

3.1.2 Το Ακουστικό Μοντέλο

Ο δεύτερος όρος της Εξ. (2), $P(y \mid w)$, υπολογίζεται από το ακουστικό μοντέλο. Σε αυτό το κεφάλαιο, αυτό υποτίθεται ότι υπολογίζεται από κρυμμένα μοντέλα Markov. Για να υπολογίσουμε την $P(y \mid w)$ για δεδομένα y και w, δημιουργείται το αντίστοιχο μοντέλο της πρότασης w συνδέοντας αλυσιδωτά μοντέλα λέξεων ή υπολέξεων. Μετά, ο μπρος αλγόριθμος δίνει την $P(y \mid w)$.

3.1.3 Συνδυάζοντας Γλωσσικά και Ακουστικά Μοντέλα

Αν και η θεωρία των πιθανοτήτων υποδηλώνει ότι οι ακουστικές και οι γλωσσικές πιθανότητες μπορούν να συνδυαστούν μέσω πολλαπλασιασμού, στην πράξη κάποια βαρύτητα είναι αναγκαία. Αυτό είναι διότι το HMM ακουστικό μοντέλο κάνει συγκεκριμένες ανακριβείς ανεξάρτητες υποθέσεις που έχουν σαν αποτέλεσμα σε μία υποτιμούμενη πιθανότητα ακουστικού μοντέλου. Συμπερασματικά, οι δύο ποσότητες έχουν πάρα πολύ διαφορετικές δυναμικές ακτίνες δράσης. Ένας τρόπος να τις ισορροπήσουμε είναι να αντικαταστήσουμε τον όρο $P(w)$ με τον όρο $P(w)^l$ στις Εξ. (1) και (2) (Bahl et al., 1980). Ο εκθέτης l αναφέρεται σαν το *βάρος γλώσσας* και αποφασίζεται πειραματικά για την βελτιστοποίηση της απόδοσης της αναγνώρισης του συστήματος αναγνώρισης συνεχούς ομιλίας. Μιας και τα ακουστικά υποτιμόνται, $l > 1$. Τυπικές τιμές του l βρίσκονται μεταξύ του 2 και 5.

3.1.4 Απεριόριστη Έρευνα

Έχοντας περιγράψει τα δύο τμήματα ενός αναγνωριστή ομιλίας και πως συνδυάζονται, μπορούμε να προχωρήσουμε στο πρόβλημα της εύρεσης της πιο πιθανής ακολουθίας λέξεων δοθέντων αυτών των τμημάτων και μίας ομιλούσας άρθρωσης.

Για να διασαφηνίσουμε τα προβλήματα της αναγνώρισης συνεχούς ομιλίας, εξετάζουμε πρώτα μία βασική μορφή έρευνας, την απλή απαρίθμηση. Στην αναγνώριση συνεχούς ομιλίας, μία έρευνα που απαριθμεί όλες τις πιθανότητες εγγυάται να βρει την βέλτιστη ακολουθία λέξεων w, δοθέντος μιας παρατηρήσιμης ακουστικής ακολουθίας y. Ο ακόλουθος αλγόριθμος περιγράφει την έρευνα απαρίθμησης.

Έρευνα Απαρίθμησης

1. Για κάθε πιθανό μήκος ακολουθίας, N,
    1. Απαρίθμησε κάθε πιθανή ακολουθία λέξεων, w, μήκους N,
        a. Εκτίμησε την $P(w \mid y)$.
        b. Επέλεξε την πιο πιθανή w.

Όπως περιγράφτηκε παραπάνω, η $P(w \mid y)$ αποφασίζεται συνδυάζοντας το HMM ακουστικό μοντέλο [δοθέντος της $P(y \mid w)$] και το μοντέλο γλώσσας [δοθέντος της $P(w)$].

3.1.5 Εμπλουτίσεις έρευνας

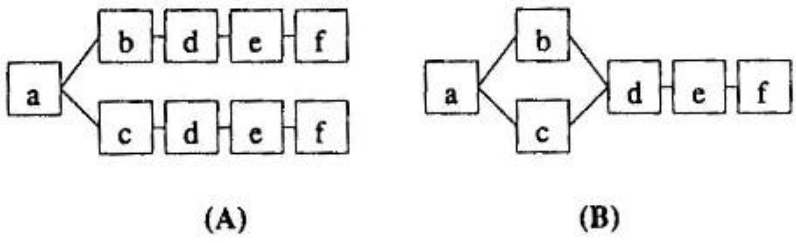
Δυστυχώς, εκτός από το μικρότερο των προβλημάτων, η έρευνα απαρίθμησης αποτυγχάνει να τερματίσει σε ένα πρακτικό ποσό χρόνου, μιας και ο αριθμός των ακολουθιών λέξεων αυξάνει εκθετικά με το μήκος της ακολουθίας λέξεων N. Παρ'όλα αυτά, όλοι οι αλγόριθμοι αναγνώρισης συνεχούς ομιλίας αναπαριστούν μία αλλαγή δομής αυτού του απλοϊκού αλγόριθμου για να βρουν τις λύσεις τους σε μία *μικρή* πεπερασμένη περίοδο χρόνου. Αυτή η αλλαγή δομής σχετίζεται με την εφαρμογή μίας εκ των ακόλουθων τεχνικών στην διαδικασία απεριόριστης έρευνας που περιγράφεται παραπάνω:

- Δημιουργία βέλτιστης υπόθεσης
- Μείωση της έκτασης του προβλήματος
- Μείωση της έρευνας
- Εφαρμογή γνώσεων

Η δημιουργία βέλτιστης υπόθεσης πραγματοποιείται συνενώνοντας κοινές μερικές υποθέσεις, το οποίο είναι η βάση για μία έρευνα δένδρου. Το σύνολο των μερικών υποθέσεων δημιουργείται αυξητικά έτσι ώστε κάθε μοναδική μερική λύση να εκτιμάται ακριβώς μία φορά. Για παράδειγμα, στο Σχ. 10Α, δύο προτάσεις δημιουργούνται σειριακά στην πληρότητά τους. Η αναγνώριση χρησιμοποιώντας αυτήν την αναπαράσταση απαιτεί δύο πλήρεις σειριακές έρευνες. Από την άλλη πλευρά, στο Σχ. 10Β, ένα δένδρο δημιουργήθηκε συνενώνοντας κοινά τμήματα των δύο προτάσεων. Χρησιμοποιώντας το δένδρο στην 10Β για αναγνώριση, η μερική λύση (ab) εκτιμάται μόνο μία φορά. Εκτιμώντας κοινές μερικές λύσεις, μπορούν να πραγματοποιηθούν σημαντικές αποταμιεύσεις στην αναγνώριση συνεχούς ομιλίας.

Σχήμα Error! Unknown switch argument. **Ενώνοντας κοινές μερικές λύσεις**

Η μείωση της έκτασης του προβλήματος σχετίζεται με την μεταμόρφωση του χώρου καταστάσεων του προβλήματος για να κάνουμε την έρευνα πιο αποτελεσματική. Αυτό έχει σχέση με την δημιουργία ενός γραφήματος από το δένδρο έρευνας τέτοιο ώστε καμιά κατάσταση του δένδρου έρευνας να μην είναι διπλότυπη. Έτσι, κάθε μοναδική κατάσταση επισκέπτεται μόνο μία φορά, με αποτέλεσμα μεγάλες μειώσεις στην έκταση της έρευνας. Για να το επεξηγήσουμε αυτό, ας θεωρήσουμε το δένδρο και το γράφημα έρευνας στο Σχ. 11. Ξανασυνδυάζοντας τα μονοπάτια στην κατάσταση d ο αριθμός των κόμβων που πρέπει να επισκεφτούν μειώνεται. Στα συστήματα αναγνώρισης συνεχούς ομιλίας, τα μονοπάτια γενικά συνενώνονται όταν αναπαριστούν ολόιδιες ακολουθίες λέξεων. Τα μονοπάτια μπορούν επίσης να συνενωθούν σε *καταστάσεις σύνδεσης*, όπως στην κατάσταση d στο Σχ. 11, στο δίκτυο γραμματικής.



Σχήμα Error! Unknown switch argument. **Ενώνοντας κοινές καταστάσεις**

Μιας και έχει αποφασιστεί ότι τα δύο μονοπάτια θα συνενωθούν, κάποιος πρέπει να αποφασίσει πώς να συνδυάσει τις τιμές τους. Εάν αυτά τα μονοπάτια αναπαριστούν ολόιδιες

ακολουθίες λέξεων, είναι σύμφωνο με την μπρος-πίσω εκπαίδευση να προσθέσουμε τις πιθανότητες. Μερικά συστήματα αναγνώρισης συνεχούς ομιλίας μπορούν να επιλέξουν να πάρουν το μέγιστο για αποδοτικότητα.* Εάν τα μονοπάτια έχουν διαφορετικά ιστορικά, ένα μέγιστο μπορεί να επιλεγεί όταν συνενώνονται. Αυτό ακόμα εγγυάται την βέλτιστη ανώτατη επιλογή, μιας και το κατώτερο μονοπάτι σε αυτό το σημείο δεν θα μπορούσε πιθανώς να ξεπεράσει το ανώτερο. Πάντως, αυτή η τεχνική δεν μπορεί παρά να παράγει τον ανώτερο υποψήφιο.

Η μείωση της έρευνας σχετίζεται με υπόθεση από επιπλέον μελέτες. Η αποκλειόμενη υπόθεση υπόκειται σε δύο κατηγορίες:

1. Πιθανώς κατώτερη υπόθεση, της οποίας η μερική εκτίμηση είναι χειρότερη από μία ήδη πλήρη εκτίμηση.
2. Πιθανώς κατώτερη υπόθεση, της οποίας η μερική εκτίμηση είναι χειρότερη από κάποιο δοκιμαστικά ευρισκόμενο κατώφλι. Αποφασισμένη από τις προσπάθειες που έγιναν, αυτή η εξάλειψη μπορεί να οδηγήσει σε σφάλματα έρευνας.+

Η εφαρμογή της γνώσης σχετίζεται με την εφαρμογή ειδικής γνώσης στο πεδίο προβλήματος της αναγνώρισης συνεχούς ομιλίας για να αυξήσουμε την αποτελεσματικότητα της έρευνας. Η ειδική γνώση είναι πολύ αποτελεσματική στην δημιουργία μίας βάσης δεδομένων που περιορίζει την έρευνα στα πεδία της ακουστικής και γλωσσικής μοντελοποίησης. Πιο σφικτοί περιορισμοί στην βάση δεδομένων μεταφράζονται άμεσα σε μικρότερους χώρους έρευνας. Μία άλλη περιοχή της εφαρμογής γνώσεων είναι στην σχεδίαση συναρτήσεων δοκιμαστικής εκτίμησης. Οι συναρτήσεις δοκιμαστικής εκτίμησης χρησιμοποιούνται στην απόφαση το ποιες υποθέσεις θα ακολουθηθούν και ποιες θα εξαλειφθούν από την μελέτη.

Στο επόμενο τμήμα θα δούμε έναν αλγόριθμο, την έρευνα Viterbi, όπου έχουν εφαρμοστεί οι τεχνικές της δημιουργίας βέλτιστης υπόθεσης, μείωσης της έκτασης του προβλήματος, και ικανής μείωσης στο πρόβλημα αναγνώρισης συνεχούς ομιλίας.

3.3 Έρευνα Δέσμης Viterbi

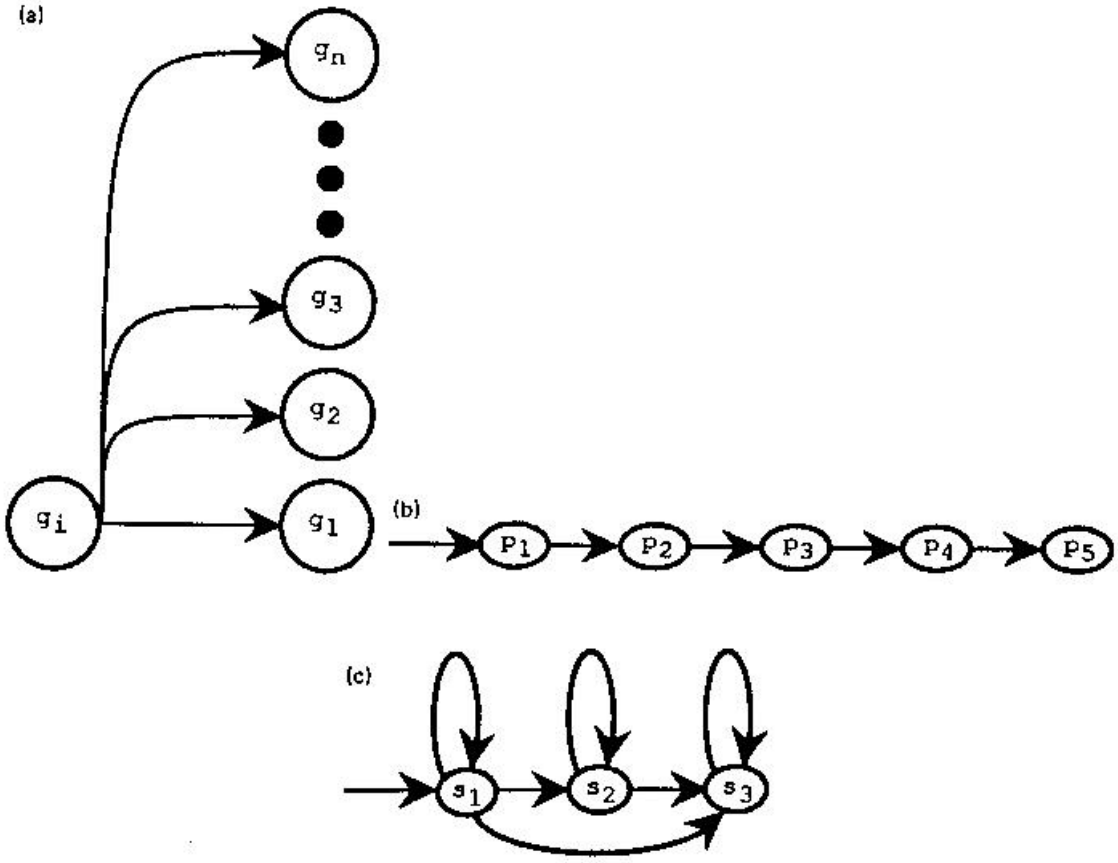
Η έρευνα δέσμης Viterbi εφαρμόστηκε επιτυχώς στα Harpy (Lowerre, 1976), Sphinx (Lee and Hon, 1988; Lu, 1988; Lee et al., 1990a), SPICOS (Ney et al., 1987), και BYBLOS (Chow et al., 1987) συστήματα αναγνώρισης συνεχούς ομιλίας. Θα περιγράψουμε τώρα τον βασικό Viterbi αλγόριθμο έρευνας και την εφαρμογή της ικανής μείωσης που αποφέρει η έρευνα δέσμης Viterbi. Θα συζητήσουμε επίσης και τα χαρακτηριστικά που κάνουν την έρευνα Viterbi κατάλληλη για αναγνώριση συνεχούς ομιλίας και θα εξετάσουμε τα προβλήματα που παρουσιάζει η έρευνα Viterbi.

3.3.1 Βέλτιστη Έρευνα Viterbi

Η έρευνα Viterbi δημιουργεί το δένδρο έρευνάς της από ένα γράφημα καταστάσεων S που ορίζεται από το δίκτυο γραμματικής και τα μοντέλα λέξεων. Τα μοντέλα λέξεων μπορεί να αναπαρασταθούν άμεσα σαν δίκτυα Markov, ή μπορεί να οριστούν παραπέρα από ένα δίκτυο του οποίου οι καταστάσεις αναπαριστούν μονάδες υπολέξεων. Το Σχήμα 12 παρουσιάζει αυτήν την διαδικασία επέκτασης.

* Τα συστήματα αναγνώρισης συνεχούς ομιλίας συχνά χρησιμοποιούν λογαριθμικές αναπαραστάσεις πιθανότητας. Ένας αλγόριθμος έρευνας που παίρνει πάντα το μέγιστο θα απαιτεί μόνο αθροίσεις των λογαρισθμικών πιθανοτήτων. Αυτό έχει το πλεονέκτημα της αποφυγής δαπανηρών πολλαπλσίων και underflows κινητής υποδιαστολής.

+ Ένα σφάλμα έρευνας σημαίνει ότι η έρευνα τερματίζει με μία μη βέλτιστη λύση.



Σχήμα Error! Unknown switch argument. **(a) Ένα μικρό τμήμα του δικτύου γραμματικής G. Κάθε κατάσταση γραμματικής g_i αντιστοιχεί σε ένα μοντέλο λέξης $W=Wm(G_i)$ όπου η συνάρτηση Wm μπορεί να είναι πολλά σε ένα. (b) Μία λέξη W επεκταμένη στο δίκτυό της των μονάδων υπο-λέξεων P. (c) Κάθε p_i επεκτείνεται στο δίκτυό του Markov.**

Βέλτιστη Έρευνα Viterbi

1. Δημιούργησε N καταστάσεων λίστες SL, όπου N είναι το μήκος της ακουστικής ακολουθίας, και όπου κάθε λίστα περιέχει μία είσοδο για κάθε κατάσταση Markov στο γράφημα καταστάσεων S.
2. Τοποθέτησε αρχικές τιμές SL[t] για t=0 θέτοντας την πιθανότητα της αρχικής κατάστασης ίση με 1.0 και όλων των άλλων 0.0.
3. Καθάρισε τις SL[t+1]
4. Για κάθε κατάσταση s στην SL[t]
    1. Για κάθε μετάβαση από την s ορισμένη από την S
        a. Υπολόγισε την πιθανότητα μετάβασης.
        b. Εάν η κατάσταση προορισμού στην SL[t+1] δεν έχει αρχική τιμή, τότε ενημέρωσέ την με κάποια τιμή και δείξε πίσω στην *κατάσταση*.
        c. Εάν η κατάσταση προορισμού στην SL[t+1] έχει ήδη αρχική τιμή, τότε έλεγξε εάν αυτή η μετάβαση είναι καλύτερη. Εάν είναι τότε ενημέρωσέ την με μία νέα τιμή και δείξε πίσω στην *κατάσταση*.
5. Εάν t=N, ξανακάνε την πορεία και επέστρεψε.
6. Αλλιώς t=t+1 και πήγαινε στο βήμα 3.

Αντίθετα προς την αποκωδικοποίηση σωρού, η έρευνα Viterbi είναι *χρονικά σύγχρονη*. Αυτό σημαίνει ότι, οποιαδήποτε στιγμή, κάθε μερική λύση w^l υπολογίζεται για το ίδιο τμήμα της ακουστικής ακολουθίας, ονομαστικά y_1^i. Έτσι, οι μερικές λύσεις μπορούν να συγκριθούν άμεσα χωρίς την επεξεργασία συναρτήσεων εκτίμησης.

Πάντως, στο βήμα 4 της έρευνας Viterbi, και οι εντός-λέξης και οι μεταξύ-λέξεων μεταβάσεις θεωρώνται σε ένα ενιαίο πλαίσιο εργασίας. Αυτό έχει το πλεονέκτημα της αποτελεσματικής εκτίμησης μερικών λύσεων λόγω των δυναμικών προγραμματιστικών ιδιοτήτων. Σαν αρνητική συνέπεια αυτού είναι ότι οι μερικές λύσεις ξανασυνδυάζονται χρησιμοποιώντας μία *max* συνάρτηση αντί μιας *sum* συνάρτησης. Αυτό είναι υπολογιστικά αποδοτικό γιατί τα μη μέγιστα μονοπάτια μπορούν απλά να απορριφθούν, και το *max* είναι καλύτερο από το *sum*. Πάντως, είναι αντιφατικό με την αθροιστική διαδικασία που χρησιμοποιείται από τον μπρος-πίσω αλγόριθμο. Για μία είσοδο y και μία υποθετική πρόταση w, οι αναφορές αποκωδικοποίησης σωρού $P(y \mid w)$, ενώ η έρευνα Viterbi αναφέρει $P(y \mid w, s)$, όπου s είναι η περισσότερο πιθανή ακολουθία.

3.3.2 Η Έρευνα Δέσμης Viterbi

Η έρευνα Viterbi εκμεταλλεύεται των τεχνικών δυναμικού προγραμματισμού που μειώνουν το μέγεθος του προβλήματος στο O(*αριθμός καταστάσεων* x *μήκος εσόδου*). Πάντως, το πρόβλημα είναι ακόμα πολύ μεγάλο όταν χρησιμοποιούνται γραμματικές πολύπλοκων πεπερασμένων καταστάσεων ή στοχαστικές γραμματικές, μιας και κάθε κατάσταση πρέπει να απαριθμηθεί. Παρατηρώντας ότι οι περισσότερες καταστάσεις στην SL[t] έχουν μηδενική ή σχεδόν μηδενική πιθανότητα μπορούμε να τροποποιήσουμε την έρευνα Viterbi ώστε να μελετά μόνο τις καταστάσεις που έχουν πιθανότητες που είναι εντός του *ε* της καλύτερης κατάστασης στην SL[t], όπου *ε* ορίζει το πλάτος της *δέσμης* στην έρευνα. Η έρευνα δέσμης Viterbi ορίζεται τότε όπως παρακάτω:

1. Δημιούργησε N σωρούς B για τις δέσμες δυναμικών καταστάσεων. N είναι το μήκος της ακουστικής ακολουθίας.
2. Πρόσθεσε την αρχική κατάσταση στην B[0] με πιθανότητα 1.0. Θέσε t=0.
3. Καθάρισε το B[t+1].
4. Για κάθε s στο B[t].
    1. Για κάθε μετάβαση από *κατάσταση* ορισμένη από την S.
        a. Υπολόγισε την πιθανότητα μετάβασης.
        b. Εάν η κατάσταση προορισμού δεν είναι στον B[t+1] τότε πρόσθεσε την στο B[t+1] με την τιμής της και δείξε πίσω στην *κατάσταση*.
        c. Εάν η κατάσταση προορισμού είναι ήδη στο B[t+1] τότε έλεγξε αν αυτή η μετάβαση είναι καλύτερη. Εάν είναι τότε ενημέρωσε την κατάσταση με την νέα τιμή και δείξε πίσω στο s.
5. Εάν t=N, ξανακάνε την πορεία και τερμάτισε.
6. Βρες την καλύτερη κατάσταση στο B[t+1] και απόρριψε τις καταστάσεις που είναι περισσότερο από *ε* χειρότερες από την καλύτερη κατάσταση.
7. t=t+1 και πήγαινε στο βήμα 3.

Το καθαρό πλεονέκτημα της έρευνας δέσμης Viterbi είναι ότι το μέγεθος του προβλήματος μειώνεται στο O(*μέσο μέγεθος δέσμης* x *μήκος εισόδου*) και η συνάρτηση συνένωσης μονοπατιού max έχει σημαντικά υπολογιστικές ευκολίες πάνω στην συνάρτηση *add* όταν εφαρμόζεται σε λογαριθμικές πιθανότητες. Αν και το καλύτερο *μέσο μέγεθος δέσμης* δεν μπορεί να καθοριστεί εκ των προτέρων, μπορεί να υπολογιστεί εμπειρικά. Το πλεονέκτημα μίας εκτίμησης δυναμικής δέσμης είναι ότι μελετά μόνο *καλές* υποθέσεις σχετικές με την καλύτερη υπόθεση. Αντιστρόφως, υπάρχει μία καθαρά καλύτερη υπόθεση μόνο μερικές εναλλακτικές χρειάζονται να διατηρηθούν από την έρευνα.

Θα εξετάσουμε τώρα μερικά θέματα υλοποίησης για την έρευνα δέσμης. Εάν θέλουμε να κρατήσουμε την πολυπλοκότητα του προβλήματος στο O(*μέσο μέγεθος δέσμης* x *μήκος εισόδου*) αντί O(*αριθμός καταστάσεων* x *μήκος εισόδου*), δεν μπορούμε να απαριθμήσουμε όλες τις καταστάσεις. Ακόμα, χωρίς απαρίθμηση τα κελιά δέσμης δεν μπορούν να προσπελαστούν τυχαία. Αυτό μπορεί να οδηγήσει σε μεγάλους σταθερούς παράγοντες στον εσώτατο βρόγχο της έρευνας. Τυπικά αυτό το πρόβλημα έχει λυθεί δημιουργώντας μία *από-καταστάσεις λίστα* και μία *σε-καταστάσεις λίστα* που απαριθμούν τις ενεργές καταστάσεις στο S. Οι σωροί B τότε γίνονται δείκτες στα ενεργά κελιά σε

αυτές τις δύο λίστες. Όταν μία μετάβαση πραγματοποιείται μπορεί εύκολα να αποφασιστεί αν ένα κελί είναι ενεργό. Εάν το κελί δεν είναι ενεργό, μπορεί να προστεθεί στον σωρό. Εάν το κελί είναι ήδη ενεργό, μπορεί να ενημερωθεί κατάλληλα. Δεν χρειάζεται να γίνει καμία αλλαγή στον σωρό. Οι δείκτες προς τα πίσω μπορούν να χειρισθούν είτε σαν μέρος του σωρού ή σαν ξεχωριστές δομές δεδομένων.

Αλλά τι γίνεται αν είναι ανέφικτο να απαριθμήσουμε το γράφημα καταστάσεων S; Μία λύση, φυσικά, θα ήταν να χρησιμοποιήσουμε μία hashing συνάρτηση χρησιμοποιώντας ένα κλειδί βασισμένο στην κατάσταση γραμματικής, λέξη και HMM κατάσταση. Μία άλλη λύση χρησιμοποιούμενη από το SPICOS (Ney et al., 1987b) σύστημα είναι να δημιουργήσουμε δυναμικά μία ιεραρχημένη δομή από τις ενεργές καταστάσεις γραμματικής, ενεργές λέξεις, και ενεργές HMM καταστάσεις. Η δυναμική εγκάρσια διάβαση γραμματικής απαιτεί πρόσθετους υπολογισμούς, αλλά έχει το πλεονέκτημα ότι σώζει πολύ χώρο.

3.3.3 Παραλλαγές Έρευνας Δέσμης Viterbi

Δύο παραλλαγές της έρευνας δέσμης Viterbi έχουν υλοποιηθεί στο BBN. Η πρώτη παραλλαγή (Chow et al., 1987) διαφέρει από την έρευνα Viterbi στο ότι *προσθέτει* πιθανότητες όταν τα μονοπάτια συνενώνονται εντός των λέξεων και γίνεται *μέγιστη* όταν τα μονοπάτια συνενώνονται στα όρια των λέξεων. Τα κύρια πλεονεκτήματα είναι ότι αυτός ο αλγόριθμος είναι περισσότερο αποδοτικός από την αποκωδικοποίηση σωρού και φέρνει την διαδικασία αποκωδικοποίησης *πιο κοντά* στην υπόθεση της εκπαίδευσης. Ο αλγόριθμος επίσης διατηρεί το χαρακτηριστικό ότι όλα τα μονοπάτια έχουν το ίδιο μήκος για να κρατήσουν την μείωση απλή. Τα προβλήματα με αυτήν την προσέγγιση είναι ότι (1) είναι πιο αργή από την έρευνα Viterbi, (2) ακόμα μελετά μόνο ένα όριο στις εξόδους λέξεων, και (3) τα μονοπάτια είναι συνδυασμένα (και οι πιθανότητές τους προστίθενται) ακόμα και όταν έχουν διαφορετικά ιστορικά.

Η δεύτερη παραλλαγή της έρευνας δέσμης Viterbi είναι η N-Καλύτερη έρευνα (Schwartz and Chow, 1990), η οποία παρέχει τις N καλύτερες αντιστοιχούμενες καταστάσεις. Η N-Καλύτερη έρευνα διατηρεί μία ξεχωριστή δέσμη από κάθε κατάσταση. (Αυτό είναι αντίθετο με την έρευνα Viterbi, η οποία χρησιμοποιεί μία γενική δέσμη με ένα μονοπάτι ανά ενεργή κατάσταση.) όταν τα μονοπάτια συνενώνονται σε μία κατάσταση s τα N καλύτερα εναλλακτικά μονοπάτια στο s κρατιούνται στην τοπική δέσμη. Σαν αποτέλεσμα, αυτό επιτρέπει στην έρευνα να αναφέρεται στα N ανώτατες αντιστοιχούμενες προτάσεις. Να σημειώσουμε ότι αυτή η έρευνα είναι πραγματικά υβριδική και είναι παρόμοια με και την έρευνα δέσμης Viterbi και την αποκωδικοποίηση σωρού.

3.4 Περίληψη

Η έρευνα δέσμης Viterbi θυσιάζει την καταλληλότερη λύση για την αποδοτικότητα και την ευκολία της εκτίμησης. Από την άλλη μεριά, η έρευνα δέσμης Viterbi μπορεί να υλοποιηθεί εύκολα και αποδοτικά αν το γράφημα καταστάσεων μπορεί να επεκταθεί πλήρως σε Markov καταστάσεις. Αντίθετα, πλήρης επέκταση είναι αδύνατη για μεγάλα μοντέλα γλώσσας.

Τελευταία λέξη της τεχνολογίας

Σε αυτό το τμήμα βλέπουμε την τελευταία λέξη της τεχνολογίας στην αναγνώριση συνεχούς ομιλίας. Παρουσιάζουμε μερικούς από τους κυριότερους παράγοντες που οδηγούν σε σχετικά μεγάλες βελτιώσεις στην απόδοση και δίνουν επιδόσεις δειγμάτων κάτω από διαφορετικές συνθήκες. Μετά βλέπουμε αρκετά από τα θέματα που επηρεάζουν την απόδοση, συμπεριλαμβανομένων των επιδράσεων της εκπαίδευσης και της γραμματικής, της εξαρτώμενης από ομιλητή σε σχέση με την ανεξάρτητη από ομιλητή αναγνώριση, της προσαρμογής ομιλητή, των μη ιθαγενών ομιλητών, και της συμπερίληψης νέων λέξεων στο λεξιλόγιο. Τα περισσότερα από τα αποτελέσματα και παραδείγματα παρακάτω έχουν παρθεί από το πρόγραμμα ARPA, το οποίο έχει υποστηρίξει την συλλογή και

διασπορά ηχητικών στοιχείων γλώσσας για συγκριτική εκτίμηση, με συγκεκριμένα παραδείγματα που έχουν παρθεί από εργασίες γνωστές στους συγγραφείς.

Βελτιώσεις στην Απόδοση

Οι βελτιώσεις στην απόδοση της αναγνώρισης ομιλίας είναι τόσο μεγάλες που στο πρόγραμμα ARPA το σφάλμα έχει πέσει κατά ένα παράγοντα του 5 σε 5 χρόνια! Αυτή η χωρίς προηγούμενο πρόοδος στην τελευταία λέξη της τεχνολογίας οφείλεται σε τέσσερις παράγοντες: χρήση κοινών ηχητικών στοιχείων γλώσσας, βελτιωμένη ακουστική μοντελοποίηση, βελτιωμένη γλωσσική μοντελοποίηση, και γρηγορότερος κύκλος πειραματικής έρευνας.

Κοινά Ηχητικά Στοιχεία Γλώσσας

Στο πρόγραμμα ARPA πρέπει να αποδοθεί το ξεκίνημα και η διατήρηση ενός μεγάλου προγράμματος για μεγάλο λεξιλόγιο, ανεξάρτητο ομιλητή, αναγνώριση συνεχούς ομιλίας. Ένας από τους θεμέλιους λίθους του προγράμματος ARPA ήταν η συλλογή και χρήση κοινών ηχητικών στοιχείων ομιλίας για ανάπτυξη συστημάτων και δοκιμές. (Τα διάφορα κοινά ηχητικά στοιχεία ομιλίας που συλλέχτηκαν από αυτό το πρόγραμμα είναι διαθέσιμα από το Linguistic Data Consortium, με γραφεία στο Πανεπιστήμιο της Πενσυλβάνιας.) Η πρώτη συλλογή κοινών στοιχείων ομιλίας ήταν η Resource Management (RM) συλλογή κοινών στοιχείων ομιλίας (Price et al., 1988), η οποία ήταν μία συλλογή από διαβασμένες προτάσεις από ένα λεξιλόγιο 1000 λέξεων στο πεδίο της διαχείρισης ναυτικών πηγών πληροφοριών. Χρησιμοποιώντας αυτή την συλλογή κοινών στοιχείων ομιλίας σαν βάση για την εργασία μας, οι διάφορες θέσεις του προγράμματος υποβλήθηκαν σε μία σειρά από τεστ ανταγωνιζόμενες αλγόριθμους αναγνώρισης κάθε 6 με 9 μήνες. Οι διάφοροι αλγόριθμοι που αναπτύχθηκαν διαμοιράστηκαν με τους άλλους συμμετέχοντες μετά από κάθε εκτίμηση, και οι επιτυχόντες γρήγορα ενσωματώθηκαν από τα διάφορα μέρη. Επιπρόσθετα από τους αλγόριθμους που αναπτύχθηκαν από διάφορα μέρη του προγράμματος , άλλοι αλγόριθμοι επίσης ενσωματώθηκαν από όλον τον κόσμο, κυρίως από την Ευρώπη και την Ιαπωνία. Αυτός ο κύκλος της ανάπτυξης αλγορίθμων, εκτίμησης, και διαμοιρασμού των λεπτομερών τεχνικών πληροφοριών οδήγησε σε απίστευτη μείωση στο ρυθμό σφάλματος όπως σημειώθηκε παραπάνω.

Ακουστική Μοντελοποίηση

Ένας αριθμός ιδεών στην ακουστική μοντελοποίηση έχει οδηγήσει σε σημαντικές βελτιώσεις στην απόδοση. Αναπτύσσοντας φωνητικά μοντέλα HMM που εξαρτώνται από το περιβάλλον, δηλαδή, από τα αριστερά και δεξιά φωνήματα, έχουν δείξει μείωση του ρυθμού σφάλματος λέξης κατά ένα παράγοντα του 2 πάνω σε ανεξάρτητα περιβάλλοντος μοντέλα, (Chow et al., 1986). Φυσικά, με μοντέλα εξαρτημένα περιβάλλοντος, ο αριθμός των μοντέλων αυξάνεται σημαντικά. Θεωρητικά, αν υπάρχουν 40 φωνήματα στο σύστημα, ο αριθμός των πιθανών τρίφωνων μοντέλων είναι $40^3 = 64000$. Πάντως, πρακτικά, μόνο λίγες χιλιάδες από τα τρίφωνα μπορούν πραγματικά να προκύψουν. Έτσι, μόνο μοντέλα των τριφώνων που προκύπτουν από εκπαιδευτικά δεδομένα υπολογίζονται πραγματικά. Εάν ειδικά τρίφωνα στους ελέγχους δεν προκύπτουν στην εκπαίδευση, τα αλλόφωνα μοντέλα που χρησιμοποιήθηκαν μπορεί να είναι τα δίφωνα ή ακόμα τα ανεξάρτητα περιβάλλοντος μοντέλα. Ένα από τα χαρακτηριστικά των HMMs είναι πως διαφορετικά μοντέλα (όπως εξαρτημένα περιβάλλοντος, δίφωνα, και τρίφωνα μοντέλα) μπορούν να παρεμβληθούν με τέτοιο τρόπο ώστε να κάνουν την καλύτερη δυνατή χρήση των δεδομένων εκπαίδευση, έτσι αυξάνοντας την ρωμαλεότητα του συστήματος.

Λόγω του ότι τα περισσότερα συστήματα υλοποιούνται σαν συστήματα αναγνώρισης λέξης (περισσότερο από συστήματα αναγνώρισης φωνήματος), δεν είναι συνήθως μέρος του συστήματος αναγνώρισης να ασχοληθούμε με επιδράσεις οφειλόμενες στο περιεχόμενο μεταξύ λέξεων και, συνεπώς, συμπεριλαμβανομένων αυτών των επιδράσεων στην αναγνώριση μπορεί να αυξήσει ουσιαστικά το υπολογιστικό φορτίο. Η μοντελοποίηση των επιδράσεων οφειλομένων στο περιεχόμενο μεταξύ λέξεων είναι περισσότερο σημαντική για μικρές λέξεις, ειδικά για λειτουργικές

λέξεις (όπου συμβαίνουν αρκετά σφάλματα), και μπορεί να μειώσει τον συνολικό ρυθμό σφαλμάτων λέξης περίπου κατά 20%.

Επιπρόσθετα της χρήσης των διανυσμάτων χαρακτηριστικών, όπως τα MFCCs, έχει βρεθεί ότι συμπεριλαμβάνοντας τα δέλτα χαρακτηριστικά-την αλλαγή του διανύσματος χαρακτηριστικού με τον χρόνο-μπορεί να μειώσει τον ρυθμό σφάλματος κατά ένα παράγοντα περίπου του 2 (Furui, 1986). Τα δέλτα χαρακτηριστικά συμπεριφέρονται σαν επιπρόσθετο διάνυσμα χαρακτηριστικού του οποίου η κατανομή πιθανότητας πρέπει επίσης να εκτιμηθεί από τα δεδομένα εκπαίδευσης. Ακόμα και αν και το πρωταρχικό διάνυσμα χαρακτηριστικού περιέχει όλη την πληροφορία που μπορεί να χρησιμοποιηθεί από την αναγνώριση, φαίνεται πως η HMM δεν εκμεταλλεύεται πλήρως την εξέλιξη χρόνου των διανυσμάτων χαρακτηριστικών. Υπολογίζοντας τις δέλτα παραμέτρους είναι ένας τρόπος να εξάγουμε την πληροφορία χρόνου και να την δώσουμε άμεσα στην HMM (Gupta et al., 1987).

Η σωστή εκτίμηση των HMM παραμέτρων-η μετάβαση και οι πιθανότητες εξόδου-από τα δεδομένα εκπαίδευσης είναι πολύ σημαντική. Λόγω του ότι μόνο ένας μικρός αριθμός από τις δυνατές τιμές διανυσμάτων χαρακτηριστικών θα προκύψουν στο σύνολο εκπαίδευσης, είναι σημαντικό να χρησιμοποιήσουμε τεχνικές εκτίμησης πιθανότητας και ομαλοποίησης που όχι μόνο θα μοντελοποιήσουν τα δεδομένα εκπαίδευσης αλλά επίσης θα μοντελοποιήσουν και όλα τα άλλα περιστατικά σε μελλοντικά δεδομένα. Ένας αριθμός τεχνικών εκτίμησης πιθανότητας και ομαλοποίησης έχει αναπτυχθεί και πετυχαίνει ένα καλό συμβιβασμό μεταξύ του υπολογισμού, ρωμαλεότητας, και ακρίβειας αναγνώρισης και έχει σαν αποτέλεσμα μειώσεις του ρυθμού σφάλματος κατά περίπου 20% συγκρινόμενα με τις διακριτές HMMs που παρουσιάστηκαν στο τμήμα “Hidden Marcov Models” (Bellegarda and Nahamoo, 1989; Gauvain and Lee, 1992; Huang et al., 1990; Schwartz et al., 1989).

Μοντελοποίηση Γλώσσας

Όπως αναφέρθηκε παραπάνω, οι στατιστικές n-γράμματες γραμματικές, ειδικά τα τριγράμματα λέξεων, είναι πολύ επιτυχημένες στην μοντελοποίηση των πολύ πιθανών ακολουθιών λέξεων σε πραγματικά δεδομένα ομιλίας. Για να πετύχουμε ένα καλό μοντέλο γλώσσας, είναι σημαντικό να χρησιμοποιήσουμε όσο μεγάλη συλλογή ηχητικών στοιχείων γλώσσας όσο είναι δυνατό έτσι ώστε όλα τριγράμματα που φαίνονται σε ερευνητική ύλη να βλέπονται στην εκπαίδευση με την ίδια πιθανότητα. Να σημειώσουμε πως μόνο το κείμενο χρειάζεται για εκπαίδευση του μοντέλου γλώσσας, όχι η πραγματική ομιλία. Τυπικά, εκατομμύρια λέξεων κειμένου χρησιμοποιούνται για να αναπτύξουν καλά μοντέλα γλώσσας. Ένας αριθμός μεθόδων έχουν αναπτυχθεί και παρέχουν ρωμαλέα εκτίμηση των πιθανοτήτων τριγραμμάτων (Katz, 1987; Placeway et al., 1993).

Για ένα σύστημα μεγάλου λεξιλογίου, υπάρχει μία μικρή αμφιβολία πως η πληρότητα, ακρίβεια, και ρωμαλεότητα του συστήματος γλώσσας μπορούν να παίξουν σημαντικό ρόλο στην απόδοση αναγνώρισης του συστήματος. Μιας και κάποιος δεν μπορεί πάντα να προβλέψει τι καινούργιο υλικό είναι δυνατό σε περιοχές μεγάλου λεξιλογίου, είναι σημαντικό να αναπτύξουμε μοντέλα γλώσσας που μπορούν να αλλάξουν σημαντικά καθώς τα δεδομένα εισόδου αλλάζουν (Della Pietra et al., 1992).

Κύκλος Έρευνας Πειραματισμού

Έχουμε δώσει έμφαση παραπάνω τις βελτιώσεις αναγνώρισης που είναι δυνατές με καινοτομίες στην ανάπτυξη του αλγορίθμου. Πάντως, αυτές οι βελτιώσεις δεν θα ήταν δυνατές χωρίς τα σωστά υπολογιστικά εργαλεία που έχουν επιτρέψει στον ερευνητή να συντομέψει τον κύκλο έρευνας πειραματισμού. Γρηγορότεροι αλγόριθμοι έρευνας, όπως επίσης γρηγορότεροι υπολογιστές, έχουν κάνει δυνατό να τρέχουμε ένα μεγάλο πείραμα σε λίγο χρόνο, τυπικά σε μία νύχτα, έτσι ώστε ο ερευνητής μπορεί να κάνει τις αναγκαίες αλλαγές την επόμενη μέρα και να εκτελεί ένα άλλο πείραμα. Οι συνδυασμένες αυξήσεις στην ταχύτητα με την καλύτερη έρευνα και γρηγορότερες μηχανές είναι αρκετές τάξεις μεγέθους.

Σχήματα Απόδοσης Δειγμάτων

Το Σχήμα 13 δείχνει αντιπροσωπευτική δειγματοληψία της απόδοσης στην τελευταία λέξη της τεχνολογίας αυτόματης αναγνώρισης φωνής. Η απόδοση φαίνεται με όρους ρυθμού σφάλματος λέξης, ο οποίος ορίζεται σαν το άθροισμα των αντικαταστάσεων, διαγραφών, και εισαγωγών λέξεων, σαν ποσοστό του πραγματικού αριθμού λέξεων στο τεστ. Όλοι οι ομιλητές της εκπαίδευσης και των ελέγχου ήταν ιθαγενείς ομιλητές των Αμερικάνικων Αγγλικών. Οι ρυθμοί σφαλμάτων είναι για αναγνώριση ανεξάρτητη ομιλητή, δηλαδή, οι ομιλητές των ελέγχων ήταν διαφορετικοί από τους ομιλητές της εκπαίδευσης. Όλα τα αποτελέσματα στο Σχήμα 13 είναι για εργαστηριακά συστήματα, πάρθηκαν από τις ακόλουθες αναφορές (Bates et al., 1993, Haeb-Umbach et al., 1993, Huang et al., 1991, Pallett et al., 1993).

Corpus	Training Data		Vocabulary		Test Data		Word Error Rate
	Type	Amount	Size	Open/ Closed	Type	Perplexity	
TI Digits	Read	4 hrs	10	Closed	Read	11	0.3%
ARPA Resource Management	Read	4 hrs	1000	Closed	Read	60	4%
ARPA Airline Travel	Spontaneous	13 hrs	1800	Open	Spontaneous	12	4%
ARPA Wall Street Journal Dictation	Read	12 hrs	5000	Closed	Read	45	5%
	Read	12 hrs	20,000	Open	Read	200	13%
	Read	12 hrs	20,000	Open	Spontaneous	255	26%

Σχήμα Error! Unknown switch argument. **Τελευταία λέξη της τεχνολογίας σε ανεξάρτητη-ομιλητή αναγνώριση συνεχούς ομιλίας**

Τα αποτελέσματα για τέσσερις συλλογές ηχητικών στοιχείων γλώσσας είναι: η ψηφιο-συνδεμένη TI συλλογή ηχητικών στοιχείων γλώσσας (Leonard, 1984), η ARPA Resource Management συλλογή ηχητικών στοιχείων γλώσσας, η ARPA Airline Travel Information Service (ATIS) συλλογή ηχητικών στοιχείων γλώσσας (MADCOW, 1992), και η ARPA Wall Street Journal (WSJ) συλλογή ηχητικών στοιχείων γλώσσας (Paul, 1992). Οι δύο πρώτες συλλογές ηχητικών στοιχείων γλώσσας συλλέκτηκαν σε πολύ ήσυχα δωμάτια στο TI, ενώ οι δύο τελευταίες συλλέκτηκαν σε περιβάλλοντα γραφείου σε διάφορες τοποθεσίες. Η ATIS συλλογή ηχητικών στοιχείων γλώσσας συλλέκτηκε από άτομα που προσπαθούσαν να πάρουν πληροφορίες για αεροπορικές γραμμές χρησιμοποιώντας κανονικά Αγγλικά ερωτήματα, είναι η μόνη συλλογή ηχητικών στοιχείων γλώσσας από τις τέσσερις που παρουσιάζονται εδώ για την οποία η ομιλία της εκπαίδευσης και ελέγχου είναι αυθόρμητη αντί να είναι διαβαζόμενες προτάσεις. Η WSJ συλλογή ηχητικών στοιχείων γλώσσας αποτελείται κυρίως από διαβαζόμενες προτάσεις από την Wall Street Journal, με κάποιες αυθόρμητες προτάσεις χρησιμοποιούμενες για ελέγχους. Στο Σχήμα 13 φαίνονται το μέγεθος του λεξιλογίου για κάθε συλλογή ηχητικών στοιχείων γλώσσας και το αν το λεξιλόγιο είναι κλειστό ή ανοικτό. Το λεξιλόγιο είναι κλειστό όταν όλες οι λέξεις στον έλεγχο είναι εγγυημένα στο λεξικό του συστήματος, ενώ στην ανοικτή περίπτωση ο έλεγχος μπορεί να περιέχει λέξεις που δεν είναι στο λεξικό του συστήματος και, επομένως, θα προκαλέσει σφάλματα στην αναγνώριση. Η περιπλοκή είναι η περιπλοκή ελέγχου-συνόλου που περιγράφτηκε παραπάνω. Μιλώντας αυστηρά, η περιπλοκή δεν ορίζεται στην περίπτωση του ανοικτού λεξιλογίου, έτσι η τιμή της περιπλοκής που φαίνεται βγήκε

κάνοντας μερικές απλές υποθέσεις για την πιθανότητα των n-γραμμάτων που περιέχουν τις άγνωστες λέξεις.

Τα αποτελέσματα που φαίνονται στο Σχήμα 13 είναι τα μέσα αποτελέσματα πάνω σε έναν αριθμό ομιλητών ελέγχου. Οι ρυθμοί σφάλματος για ξεχωριστούς ομιλητές διαφέρουν κατά ένα μεγάλο σχετικά εύρος και μπορούν να είναι αρκετές φορές χαμηλότεροι ή ψηλότεροι από τις μέσες τιμές που φαίνονται. Μιας και πολλά από τα δεδομένα συλλέχτηκαν σε σχετικά ήπιες συνθήκες, κάποιος μπορεί να περιμένει την απόδοση να μειώνεται στην παρουσία θορύβου και διαστροφής καναλιού. Είναι ξεκάθαρο από το Σχήμα 13 πως υψηλότερη περιπλοκή, ανοικτό λεξιλόγιο, και αυθόρμητη ομιλία τείνουν να αυξήσουν τον ρυθμό σφάλματος λέξης. Θα ποσοτικοποιήσουμε κάποιες από αυτές τις επιδράσεις μετά και θα συζητήσουμε μερικά σημαντικά θέματα που επηρεάζουν την απόδοση.

Επιδράσεις της Εκπαίδευσης και της Γραμματικής

Είναι δεκτό ότι αυξάνοντας το ποσό των δεδομένων εκπαίδευσης γενικά μειώνεται ο ρυθμός σφάλματος λέξης. Πάντως, είναι σημαντικό η αυξημένη εκπαίδευση να είναι αντιπροσωπευτική των τύπων των δεδομένων του ελέγχου. Αλλιώς, η αυξημένη εκπαίδευση μπορεί να μην βοηθήσει.

Με την RM συλλογή ηχητικών στοιχείων γλώσσας, βρέθηκε πως ο ρυθμός σφάλματος είναι αντιστρόφως ανάλογος της τετραγωνικής ρίζας του ποσού των δεδομένων εκπαίδευσης, έτσι που τετραπλασιάζοντας τα δεδομένα εκπαίδευσης έχουμε σαν αποτέλεσμα το κόψιμο του ρυθμού σφάλματος λέξης κατά ένα παράγοντα του 2. Αυτή η μεγάλη μείωση στον ρυθμό σφάλματος αυξάνοντας τα δεδομένα εκπαίδευσης μπορεί να είναι το αποτέλεσμα ενός τεχνουργήματος της RM συλλογής ηχητικών στοιχείων γλώσσας, δηλαδή, πως τα πρότυπα προτάσεων των δεδομένων ελέγχου ήταν τα ίδια με αυτά της εκπαίδευσης. Σε μία ρεαλιστική συλλογή ηχητικών στοιχείων γλώσσας, όπου τα πρότυπα προτάσεων του ελέγχου μπορούν συχνά να είναι αρκετά διαφορετικά από της εκπαίδευσης, τέτοιες βελτιώσεις μπορεί να είναι δραματικές. Για παράδειγμα, πρόσφατες έρευνες με WSJ συλλογή ηχητικών στοιχείων γλώσσας απέτυχαν να δείξουν σημαντική μείωση στο ρυθμό σφάλματος διπλασιάζοντας το ποσό της εκπαίδευσης. Πάντως, είναι δυνατό αυξάνοντας την πολυπλοκότητα των μοντέλων καθώς αυξάνονται τα δεδομένα εκπαίδευσης, να έχουμε σαν αποτέλεσμα μεγαλύτερη μείωση στο ρυθμό σφάλματος. Αυτό είναι ακόμα θέμα ερευνών.

Οι ρυθμοί σφάλματος λέξης γενικά αυξάνονται με αύξηση στη περιπλοκή γραμματικής. Ένας γενικός κανόνας είναι πως ο ρυθμός σφάλματος αυξάνεται με την τετραγωνική ρίζα της περιπλοκής, με καθετί άλλο να παραμένει ίδιο. Αυτός ο κανόνας μπορεί να μην είναι γενικά ένας καλός προγνώστης της απόδοσης, αλλά είναι μία λογική προσέγγιση. Να σημειώσουμε πως το μέγεθος του λεξικού δεν είναι ο πρωταρχικός παράγοντας της απόδοσης αναγνώρισης αλλά μάλλον η ελευθερία με την οποία οι λέξεις τοποθετούνται μαζί, η οποία αναπαρίσταται στην γραμματική. Μία λιγότερο περιορισμένη γραμματική, όπως της WSJ συλλογής ηχητικών στοιχείων γλώσσας, έχει σαν αποτέλεσμα υψηλότερους ρυθμούς σφάλματος.

Αναγνώριση Εξαρτώμενη-Ομιλητή & Ανεξάρτητη-Ομιλητή

Οι όροι εξαρτώμενη-ομιλητή (SD) και ανεξάρτητη-ομιλητή (SI) αναγνώριση χρησιμοποιούνται συχνά για να περιγράψουν διαφορετικούς τρόπους χρήσης ενός συστήματος αναγνώρισης ομιλίας. Η SD αναγνώριση αναφέρεται στην περίπτωση όπου ένας μόνο ομιλητής χρησιμοποιείται για να εκπαιδεύσει το σύστημα και ο ίδιος ομιλητής χρησιμοποιείται και για τον έλεγχο. Η SI αναγνώριση αναφέρεται στην περίπτωση όπου ο ομιλητής ελέγχου δεν συμπεριλαμβάνεται στην εκπαίδευση. Τα συστήματα που βασίζονται σε HMM μπορούν να λειτουργήσουν είτε σαν SD είτε σαν SI, αναλόγως των δεδομένων εκπαίδευσης που χρησιμοποιήθηκαν. Στον SD τρόπο λειτουργίας η ομιλία εκπαίδευσης συλλέγεται από ένα μόνο ομιλητή, ενώ στον SI τρόπο λειτουργίας η ομιλία εκπαίδευσης συλλέγεται από πολλούς ομιλητές.

Οι SD και SI τρόποι αναγνώρισης μπορούν να συγκριθούν σε όρους του ρυθμού σφάλματος λέξης για ένα δεδομένο ποσό εκπαίδευσης. Ένας γενικός κανόνας είναι πως, εάν το ολικό ποσό της ομιλίας εκπαίδευσης είναι αμετάβλητο μέχρι κάποιο σημείο, οι ρυθμοί σφάλματος SI είναι περίπου τετραπλάσιοι των SD ρυθμών σφάλματος. Ένας άλλος τρόπος να εκφράσουμε αυτόν τον κανόνα είναι πως, για να έχει η SI αναγνώριση την ίδια απόδοση με την SD αναγνώριση, απαιτεί περίπου 15 φορές

το ποσό των δεδομένων εκπαίδευσης της δεύτερης (Schwartz et al., 1993). Αυτά τα αποτελέσματα προκύπτουν μετά από μίας ώρας ομιλίας για τον υπολογισμό των SD μοντέλων. Πάντως, οριακά, καθώς το ποσό των δεδομένων εκπαίδευσης των SD και SI μοντέλων μεγαλώνει όλο και περισσότερο, δεν είναι ξεκάθαρο ποιο ποσό δεδομένων εκπαίδευσης θα επιτρέψει την απόδοση SI να πλησιάσει την απόδοση SD.

Το σκεπτικό πίσω από την αναγνώριση SI είναι πως η εκπαίδευση γίνεται μία φορά, μετά την οποία κάθε ομιλητής μπορεί να χρησιμοποιήσει το σύστημα με καλή απόδοση. Η αναγνώριση SD αντιμετωπίζεται αρνητικά για μελλοντικούς χρήστες. Πάντως, κάποιος πρέπει να έχει υπόψιν πως η εκπαίδευση SI πρέπει να γίνει για κάθε νέα χρήση ενός συστήματος σε ένα διαφορετικό τομέα. Εάν το σύστημα χρησιμοποιείται κάπου όπου δεν έχει εκπαιδευτεί, τότε η απόδοση μειώνεται. Είναι διαχρονικό αξίωμα ότι, για την βέλτιστη απόδοση SI, είναι καλύτερο να συλλέγουμε δεδομένα εκπαίδευσης από όσους περισσότερους χρήστες μπορούμε σε κάθε τομέα. Για παράδειγμα, αντί να συλλέγουμε 100 αρθρώσεις από καθένα από 100 ομιλητές, πιστεύεται ότι είναι πολύ καλύτερο να συλλέγουμε, ας πούμε, 10 αρθρώσεις από καθένα από 1000 ομιλητές. Πρόσφατες έρευνες έδειξαν ότι, για μερικές εφαρμογές, συλλέγοντας ομιλία από μόνο μία δωδεκάδα ομιλητές μπορεί να είναι αρκετό για καλή απόδοση SI. Σε ένα πείραμα με WSJ συλλογή ηχητικών στοιχείων γλώσσας, για δεδομένο ποσό δεδομένων εκπαίδευσης, φάνηκε πως η εκπαίδευση με 12 ομιλητές έδωσε βασικά την ίδια απόδοση SI όπως η εκπαίδευση με 84 ομιλητές (Schwartz et al., 1993). Αυτό είναι ένα ευπρόσδεκτο αποτέλεσμα, κάνει ευκολότερο να αναπτύξουμε μοντέλα SI σε έναν καινούργιο τομέα μιας και συλλέγοντας δεδομένα από λίγους ομιλητές είναι φθηνό και περισσότερο βολικό.

Ο τελικός στόχος των ερευνών αναγνώρισης ομιλίας είναι να έχουμε ένα σύστημα που να είναι ανεξάρτητο τομέα (DI), δηλαδή, ένα σύστημα που εκπαιδεύεται μία για πάντα έτσι ώστε να μπορεί να χρησιμοποιηθεί σε κάθε τομέα και από οποιονδήποτε ομιλητή χωρίς επανεκπαίδευση. Πρόσφατα, η μόνη μέθοδος που χρησιμοποιείται για αναγνώριση DI είναι να εκπαιδεύσουμε το σύστημα σε ένα πολύ μεγάλο αριθμό δεδομένων από διαφορετικούς τομείς. Πάντως, προκαταρκτικοί έλεγχοι έχουν δείξει πως η αναγνώριση DI σε ένα καινούργιο τομέα που δεν συμπεριλαμβάνεται στην εκπαίδευση μπορεί να αυξήσει το ρυθμό σφάλματος κατά ένα παράγοντα από 1.5 έως 2 πάνω από την αναγνώριση SI όταν εκπαιδεύεται σε ένα νέο τομέα, υποθέτοντας ότι η γραμματική έρχεται από το νέο τομέα (Hon, 1992). Εάν μία καλή γραμματική δεν είναι διαθέσιμη από τον καινούργιο τομέα, η απόδοση μπορεί να είναι αρκετά χειρότερη.

Προσαρμογή

Είναι δυνατό να αυξήσουμε την απόδοση ενός συστήματος SI ή DI με αυξητική προσαρμογή στην φωνή του καινούργιου ομιλητή καθώς αυτός χρησιμοποιεί το σύστημα. Αυτό χρειάζεται ειδικά για ασυνήθιστους ομιλητές με υψηλούς ρυθμούς σφάλματος που μπορεί αλλιώς να βρούνε το σύστημα ανέθιστο. Τέτοιες κατηγορίες ομιλητές συμπεριλαμβάνουν ομιλητές με ασυνήθιστες διαλέκτους και αυτούς για τους οποίους τα μοντέλα SI απλά δεν είναι καλά μοντέλα της ομιλίας τους. Πάντως, η αυξητική προσαρμογή μπορεί να απαιτήσει ώρες χρήσης και πολύ υπομονή από ένα νέο χρήστη πριν η απόδοση γίνει ικανοποιητική.

Μία καλή λύση στο πρόβλημα ασυνήθιστου ομιλητή είναι να χρησιμοποιήσουμε μία μέθοδο γνωστή σαν γρήγορη προσαρμογή ομιλητή. Σε αυτήν την μέθοδο μόνο ένα μικρό ποσό ομιλίας (γύρω στα δύο λεπτά) συλλέγεται από τον νέο ομιλητή πριν χρησιμοποιήσει το σύστημα. Έχοντας συλλέξει τις ίδιες αρθρώσεις πιο πριν από ένα ή περισσότερους πρωτότυπους ομιλητές, έχουν αναπτυχθεί μέθοδοι για εξαγωγή ενός μοντέλου ομιλίας από το νέο ομιλητή μέσω μίας απλής μετατροπής του μοντέλου ομιλίας των πρωτότυπων ομιλητών (Furui, 1989; Kubala and Schwartz, 1990; Nakamura and Shikano, 1989). Είναι δυνατό με αυτές τις μεθόδους να πετύχουμε μία μέση απόδοση SI για ομιλητές που αλλιώς θα είχε αρκετά μεγαλύτερο ρυθμό σφάλματος.

Ένα αξιοπρόσεκτο παράδειγμα που ασυνήθιστοι ομιλητές είναι μη ιθαγενείς ομιλητές, δοθέντος πως το σύστημα SI εκπαιδεύτηκε σε μόνο ιθαγενείς ομιλητές. Σε ένα δοκιμαστικό πείραμα όπου τέσσερις μη ιθαγενείς ομιλητές ελέγθησαν στο RM τομέα με SI τρόπο, υπήρχε μία οχταπλάσια αύξηση στο ρυθμό σφάλματος λέξης πάνω από τους ιθαγενείς ομιλητές! Οι τέσσερις ομιλητές ήταν ιθαγενείς ομιλητές των Αραβικών, Εβραϊκών, Κινέζικων και Βρετανικών Αγγλικών. Συλλέγοντας δύο

λεπτά ομιλίας από καθένα από αυτούς τους ομιλητές και χρησιμοποιώντας γρήγορη προσαρμογή ομιλητή, ο μέσος ρυθμός σφάλματος λέξης για τους τέσσερις ομιλητές μειώθηκε πέντε φορές.

Προσθέτοντας Νέες Λέξεις

Οι έξω από τα λεξιλόγια λέξεις προκαλούν σφάλματα αναγνώρισης και μειώνουν την απόδοση. Έχουν γίνει λίγες προσπάθειες στην αυτόματη αναγνώριση της παρουσίας νέων λέξεων, περιορισμένη επιτυχία (Asadi et al., 1990). Τα περισσότερα συστήματα απλά δεν κάνουν τίποτα ιδιαίτερο για αντιμετωπίσουν την παρουσία τέτοιων λέξεων.

Αφού ο χρήστης συνειδητοποιήσει ότι μερικά από τα σφάλματα προκαλούνται από νέες λέξεις και αποφασίσει τι είναι αυτές, είναι δυνατόν να τις προσθέσει στο λεξιλόγιο του συστήματος. Στην βασισμένη σε λέξεις αναγνώριση, όπου ολόκληρες λέξεις μοντελοποιούνται χωρίς να έχει παρεμβληθεί ένα ενδιάμεσο φωνητικό στάδιο, η πρόσθεση νέων λέξεων στο λεξιλόγιο απαιτεί ειδική εκπαίδευση του συστήματος στις νέες λέξεις (Bahl et al., 1988). Πάντως, στην φωνητικά βασισμένη αναγνώριση, όπως των φωνητικών HMM προσέγγιση που παρουσιάζεται σε αυτήν την εργασία, η πρόσθεση νέων λέξεων στο λεξιλόγιο μπορεί να επιτευχθεί συμπεριλαμβάνοντας τις φωνητικές τους προφορές στο λεξικό του συστήματος. Αν οι νέες λέξεις δεν είναι μέσα στο λεξικό, μία φωνητική προφορά μπορεί να παραχθεί από έναν συνδυασμό ηχογράφησης και μίας πραγματικής προφοράς της λέξης από τον ομιλητή (Bahl et al., 1990a). Τα HMMs για τις νέες λέξεις τότε αυτόματα μεταγλωττίζονται από τα προϋπάρχοντα φωνητικά μοντέλα, όπως φαίνεται στο Σχήμα 6. Οι νέες λέξεις πρέπει επίσης να προστεθούν στην γραμματική με τον κατάλληλο τρόπο.

Πειράματα έδειξαν ότι, χωρίς επιπρόσθετη εκπαίδευση για τις νέες λέξεις, ο SI ρυθμός σφάλματος είναι περίπου δύο φορές αυτού που η εκπαίδευση περιέλαβε τις νέες λέξεις. Επομένως, τα ορισμένα από τον χρήστη λεξιλόγιο και γραμματική μπορούν εύκολα να ενσωματωθούν σε ένα σύστημα αναγνώρισης με μία μέτρια αύξηση του ρυθμού σφάλματος για τις νέες λέξεις.

ΑΝΑΓΝΩΡΙΣΗ ΟΜΙΛΙΑΣ ΠΡΑΓΜΑΤΙΚΟΥ-ΧΡΟΝΟΥ

Μέχρι πρόσφατα, θεωρείτο ότι για να επιτευχθεί μεγάλης ακρίβειας, πραγματικού χρόνου, αναγνώριση συνεχούς ομιλίας για μεγάλα λεξιλόγια θα απαιτούσε είτε ειδικού τύπου VLSI υλικό ή ένα πολυεπεξεργαστή. Πάντως, νέες ανακαλύψεις στους αλγόριθμους έρευνας έχουν επιταχύνει τους υπολογισμούς αναγνώρισης κατά τουλάχιστον δύο τάξεις μεγέθους, με λίγη ή καθόλου απώλεια ακρίβειας αναγνώρισης (Austin et al., 1991; Bahl et al., 1990b; Ney, 1992; Schwartz and Austin, 1991; Soong and Huang, 1991). Επιπρόσθετα, υπολογιστικοί πρόοδοι πέτυχαν αύξηση δύο τάξεις μεγέθους στις ταχύτητες των υπολογιστών την τελευταία δεκαετία. Αυτές οι δύο πρόοδοι έκαναν την βασισμένη σε λογισμικό, πραγματικού χρόνου, αναγνώριση συνεχούς ομιλίας μία πραγματικότητα. Η μόνη απαίτηση είναι πως ο υπολογιστής πρέπει να έχει ένα A/D μετατροπέα για να ψηφιοποιήσει την ομιλία. Ολόκληρη η επεξεργασία σήματος, εξαγωγή χαρακτηριστικών, και έρευνα αναγνώρισης πραγματοποιείται τότε με λογισμικό σε πραγματικό χρόνο σε ένα υπολογιστή ενός επεξεργαστή.

Για παράδειγμα, είναι δυνατόν να πραγματοποιήσουμε μία 2000 λέξεων ATIS εργασία σε πραγματικό χρόνο σε υπολογιστές όπως ο Silicon Graphics Indigo R3000 ή ο Sun SparcStation 2. Πιο πρόσφατα, μία εργασία 20000 λέξεων WSJ συνεχούς υπαγόρευσης επιδείχτηκε σε πραγματικό χρόνο (Nguyen et al., 1993) σε έναν υπολογιστή Hewlett-Packard 735, ο οποίος έχει περίπου τρεις φορές την ισχύ ενός SGI R3000. Έτσι, οι υπολογισμοί αυξάνουν πιο αργά παρά γραμμικά με το μέγεθος του λεξικού.

Τα κατορθώματα πραγματικού χρόνου που μόλις περιγράφτηκαν έχουν επιτευχθεί με ένα σχετικά μικρό κόστος στην ακρίβεια λέξης. Τυπικά, οι ρυθμοί σφάλματος λέξης είναι λιγότερο από δύο φορές αυτών των καλύτερων συστημάτων έρευνας.

Η πιο προοδευμένη από αυτές τις πραγματικού χρόνου επιδείξεις δεν έχει ακόμα γίνει στην αγορά. Πάντως, είναι δυνατόν σήμερα να αγοράσουμε προϊόντα που πραγματοποιούν ανεξάρτητη ομιλητή, αναγνώριση συνεχούς ομιλίας για λεξιλόγια λίγων χιλιάδων λέξεων. Η υπαγόρευση μεγάλων λεξιλογίων περίπου 30000 λέξεων είναι δυνατή εμπορικά, αλλά ο ομιλητής πρέπει να

σταματά πολύ λίγο μεταξύ των λέξεων και το σύστημα προσαρμόζεται στην φωνή του χρήστη για να αυξήσει την απόδοση. Για τα περισσότερα από τα διαθέσιμα προϊόντα και τις εφαρμογές τους, ο αναγνώστης μπορεί να αναφερθεί σε άλλες εργασίες αυτού του τόμου.

ΕΝΑΛΛΑΚΤΙΚΑ ΜΟΝΤΕΛΑ

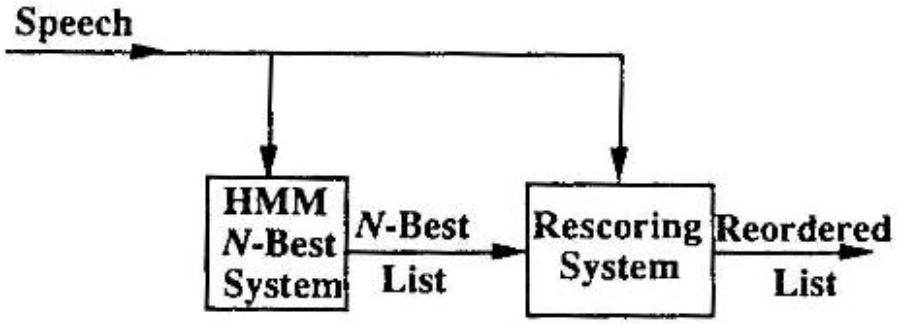
Οι HMMs έχουν αποδειχτεί να είναι πολύ καλές για διακύμανση μοντελοποίησης στο χρόνο και χώρο χαρακτηριστικών και αποτελεσματικές σε μεγάλες προόδους στην αναγνώριση συνεχούς ομιλίας. Πάντως, μερικές από τις υποθέσεις που κάνουν οι HMMs δεν είναι αυστηρά αληθινές για την ομιλία-για παράδειγμα, οι γενόμενες υπό όρους υποθέσεις ανεξαρτησίας με τις οποίες η πιθανότητα να βρίσκεται σε μία κατάσταση εξαρτάται μόνο από την προηγούμενη κατάσταση και η πιθανότητα εξόδου σε μία κατάσταση εξαρτάται μόνο από αυτήν την κατάσταση και όχι από την προηγούμενη καταστάσεις ή τις προηγούμενες εξόδους. Έχουν γίνει προσπάθειες βελτίωσης των επιδράσεων αυτών των υποθέσεων αναπτύσσοντας εναλλακτικά μοντέλα ομιλίας. Πιο κάτω, περιγράφουμε εν συντομία μερικές από αυτές τις προσπάθειες, συμπεριλαμβανομένης της χρήσης κατατμημένων μοντέλων και νευρωνικών δικτύων. Σε όλες αυτές τις προσπάθειες, πάντως, σημαντικοί υπολογιστικοί περιορισμοί έχουν καθυστερήσει την εξερεύνηση αυτών των μεθόδων και έχουν σαν αποτέλεσμα μόνο μικρές βελτιώσεις στην απόδοση μέχρι τώρα.

Κατατμημένα Μοντέλα

Τα φωνητικά κατατμημένα μοντέλα σχηματίζουν ένα μοντέλο ενός ολόκληρου φωνητικού τμήματος, μάλλον από το να μοντελοποιήσουν την ακολουθία των πλαισίων όπως με μια HMM. Τα κατατμημένα μοντέλα δεν περιορίζονται από την γενόμενη υπό όρους υπόθεση των HMMs γιατί, μοντελοποιούν εξαρτήσεις μεταξύ όλων των πλαισίων ενός τμήματος άμεσα. Επιπρόσθετα, τα κατατμημένα μοντέλα μπορούν να ενσωματώσουν διάφορα κατατμημένα χαρακτηριστικά με έναν άμεσο τρόπο, ενώ είναι αδέξια να συμπεριλάβουν κατατμημένα χαρακτηριστικά σε μία HMM. Τα κατατμημένα χαρακτηριστικά περιλαμβάνουν οποιεσδήποτε μετρήσεις έχουν γίνει σε όλο το τμήμα ή σε μέρη ενός τμήματος, όπως η διάρκεια του τμήματος.

Έχουν προταθεί λίγα κατατμημένα μοντέλα, μεταξύ των οποίων στοχαστικά κατατμημένα μοντέλα και κατατμημένα νευρωνικά δίκτυα (περιγράφονται στο επόμενο τμήμα). Τα στοχαστικά κατατμημένα μοντέλα βλέπουν την ακολουθία των διανυσμάτων χαρακτηριστικών σε ένα φωνητικό τμήμα σαν ένα μόνο μακρύ διάνυσμα χαρακτηριστικών (Ostendorf et al., 1992). Το κύριο έργο είναι τότε να εκτιμήσουμε την από κοινού πυκνότητα πιθανότητας των στοιχείων στο διάνυσμα χαρακτηριστικών. Πάντως, λόγω της μεταβλητότητας του αριθμού των πλαισίων σε ένα τμήμα, είναι σημαντικό πρώτα να κανονικοποιήσουμε το τμήμα σε ένα δεδομένο αριθμό πλαισίων. Χρησιμοποιώντας κάποιας μορφής παρεμβολή, τυπικά περίπου γραμμική, γεννιέται ένας δεδομένος αριθμός πλαισίων που μαζί σχηματίζουν το μοναδικό διάνυσμα χαρακτηριστικών του οποίου η κατανομή πιθανότητας είναι να υπολογιστεί. Τυπικά, η κατανομή υποτίθεται ότι είναι πολυδιαστατική Γκαουσιανή και οι παράμετροί της υπολογίζονται από τα δεδομένα εκπαίδευσης. Πάντως, λόγω του μεγάλου μεγέθους του διανύσματος χαρακτηριστικών και του πάντα περιορισμένου διαθέσιμου ποσού δεδομένων εκπαίδευσης, δεν είναι δυνατόν να πάρουμε καλές προσεγγίσεις όλων των παραμέτρων της κατανομής πιθανότητας. Έτσι, γίνονται υποθέσεις για να μειώσουν τον αριθμό των παραμέτρων που πρέπει να υπολογιστούν.

Λόγω του ότι τα κατατμημένα μοντέλα τμηματοποιούν άμεσα ένα μοντέλο, μία κατάτμηση της ομιλίας σε φωνητικά τμήματα μπορεί να επιτευχθεί πριν την μοντελοποίηση και την αναγνώριση. Θεωρητικά κάποιος μπορεί να προσπαθήσει όλες τις δυνατές κατατμήσεις, υπολογίσει την πιθανότητα κάθε κατάτμησης, και διαλέξει αυτήν που έχει την μεγαλύτερη πιθανότητα δοθέντος της ομιλίας εισόδου. Πάντως, το να προσπαθήσουμε όλες τις δυνατές κατατμήσεις είναι απαγορευτικό υπολογιστικά σε κανονικούς υπολογιστές. Μία λύση ήταν να χρησιμοποιήσουμε ένα βασισμένο σε HMM σύστημα για να παράγουμε κατάλληλους υποψήφιους για κατατμήσεις, οι οποίοι μετά ξαναδιαβαθμίζονται με τα κατατμημένα μοντέλα.



Σχήμα Error! Unknown switch argument. **Ν-καλύτερο παράδειγμα για συνδυασμό πηγών γνώσεων**

Το Σχήμα 14 δείχνει την βασική ιδέα αυτού που είναι γνωστό ως *N-Best Paradigm* (Ostendorf et al., 1991; Schwartz and Chow, 1990). Πρώτα, ένα βασισμένο σε HMM σύστημα αναγνώρισης χρησιμοποιείται για να παράγει όχι μόνο την ανώτατης διαβάθμισης υπόθεση αλλά επίσης και τις N-διαβάθμισης υποθέσεις. Συνδεμένη με κάθε υπόθεση είναι μία ακολουθία λέξεων και φωνημάτων και η αντίστοιχη κατάτμηση. Για κάθε μία από τις διαφορετικές κατατμήσεις και ονοματολογίες, μία πιθανότητα υπολογίζεται από τα μοντέλα πιθανότητας καθενός τμήματος. Οι ιδιαίτερες; κατατμημένες διαβαθμίσεις τότε συνδυάζονται για να σχηματίσουν μία διαβάθμιση για ολόκληρη την υπόθεση. Οι υποθέσεις τότε αναδιατάσσονται από τις διαβαθμίσεις τους, και επιλέγεται η υψηλότερης διαβάθμισης υπόθεση. Τυπικά, η κατατμημένη διαβάθμιση συνδυάζεται με την HMM διαβάθμιση για να βελτιώσουν την απόδοση.

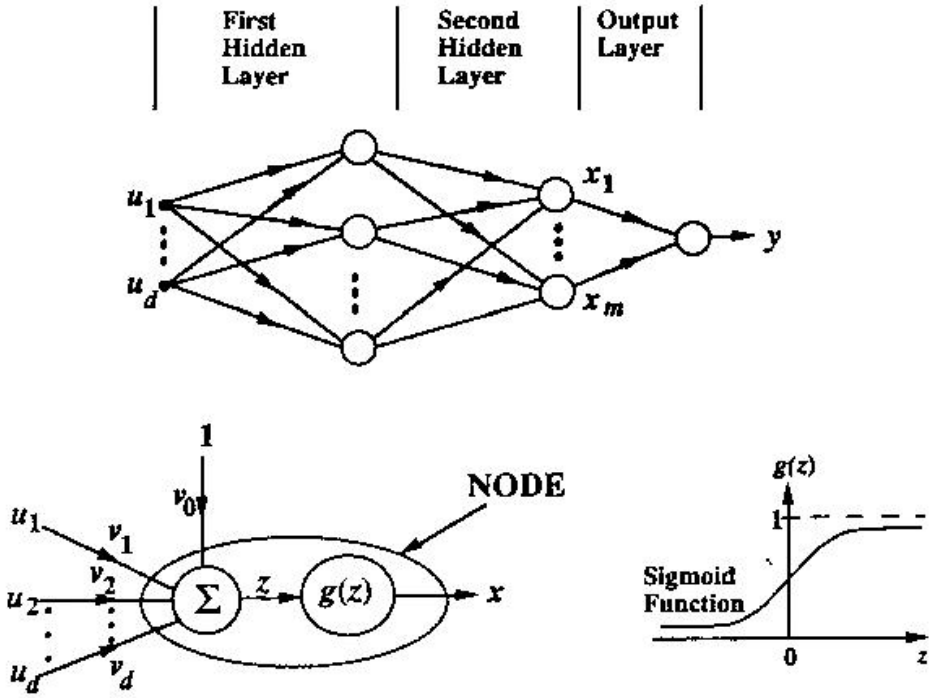
Χρησιμοποιώντας το N-καλύτερο παράδειγμα με κατατμημένα μοντέλα, με N=100, έχει μειώσει τους ρυθμούς σφάλματος λέξης κατά 20%. Το N-καλύτερο παράδειγμα είναι επίσης χρήσιμο στην μείωση των υπολογισμών όποτε μία ή περισσότερες πηγές γνώσεων χρειάζεται να συνδυαστούν, για παράδειγμα, μοντέλα διασταύρωσης-λέξεων και n-γράμματες πιθανότητες για n>2. Το N-καλύτερο είναι ένα χρήσιμο παράδειγμα εφόσον η σωστή πρόταση έχει υψηλή πιθανότητα να είναι μεταξύ των N υψηλότερων υποθέσεων. Έτσι το παράδειγμα είναι χρήσιμο για λεξιλόγια μέχρι 5000 λέξεις, ακόμα και για σχετικά μακρές προτάσεις.

Νευρωνικά Δίκτυα

Οποιανδήποτε και αν είναι τα βιολογικά κίνητρα για την ανάπτυξη των «τεχνητών νευρωνικών δικτύων» (Lippman, 1987), η χρήση των νευρωνικών δικτύων υπηρετείται καλύτερα καταλαβαίνοντας τις μαθηματικές τους ιδιότητες (Makhoul, 1991). Βλέπουμε ένα νευρωνικό δίκτυο σαν ένα δίκτυο αλληλοσυνδεόμενων απλών μη γραμμικών υπολογιστικών μονάδων και την έξοδο ενός νευρωνικού δικτύου σαν μία περίπλοκη μη γραμμική συνάρτηση των εισόδων του. Το Σχήμα 15 δείχνει ένα τυπικό τροφοδοτούμενο από μπροστά νεωρωνικό δίκτυο, δηλαδή, δεν έχει στοιχεία ανάδρασης. Αν και πολλοί διαφορετικοί τύποι νεωρωνικών δικτύων έχουν προταθεί, ο τύπος του δικτύου που φαίνεται στο Σχήμα 15 χρησιμοποιείται από την πλειονότητα των εργαζομένων σ' αυτόν τον τομέα. Στο σχήμα φαίνεται ένα δίκτυο τριών στρωμάτων, με κάθε στρώμα να αποτελείται από εισόδους, έναν αριθμό κόμβων, και αλληλοσυνδεόμενα βάρη. (Ο όρος "κρυμμένο" έχει χρησιμοποιηθεί για να περιγράψει στρώματα που δεν είναι άμεσα συνδεμένα με την έξοδο.) Όλοι οι κόμβοι είναι συνήθως ίδιοι στην κατασκευή όπως φαίνεται στο σχήμα. Οι είσοδοι u σε έναν κόμβο πολλαπλασιάζονται από ένα σύνολο βαρών v και αθροίζονται για να σχηματίσουν μία τιμή z. Μία μη γραμμική συνάρτηση των $z, g(z)$ υπολογίζεται τότε. Το σχήμα 15 δείχνει μία από τις πιο τυπικές μη γραμμικές χρησιμοποιήσιμες συναρτήσεις, το σιγμοειδές. Η έξοδος y του δικτύου είναι τότε μια μη γραμμική συνάρτηση του διανύσματος εισόδου. Γενικά, το δίκτυο μπορεί να έχει ένα διάνυσμα εξόδων επίσης.

Υπάρχουν δύο σημαντικές μαθηματικές ιδιότητες των νευρωνικών δικτύων που σχηματίζουν τους θεμέλιους λίθους πάνω στους οποίους έχουν αναπτυχθεί επιτυχείς εφαρμογές. Το πρώτο είναι

μία συνάρτηση προσέγγισης: έχει δειχθεί ότι ένα νευρωνικό δίκτυο δύο στρωμάτων είναι ικανό να προσεγγίσει οποιαδήποτε συνάρτηση αυθαίρετα κοντά σε οποιοδήποτε πεπερασμένο τμήμα του χώρου εισόδου (Cybenko, 1989). Κάποιος θα μπορούσε να χρησιμοποιήσει περισσότερα από δύο στρώματα για χάρη της οικονομίας, αλλά αυτή η ιδιότητα λέει πως δύο στρώματα είναι αρκετά (με ένα πιθανώς μεγάλο αριθμό κόμβων στο κρυμμένο στρώμα) για να προσεγγίσουμε οποιαδήποτε συνάρτηση στην είσοδο. Για εφαρμογές όπου επιθυμείται στην είσοδο μία μη γραμμική συνάρτηση, το νευρωνικό δίκτυο μπορεί να εκπαιδευτεί να ελαχιστοποιεί, για παράδειγμα, το μέσο τετραγωνικό σφάλμα μεταξύ της πραγματικής και της επιθυμούμενης εξόδου. Επαναληπτικές μη γραμμικές συναρτήσεις βελτιστοποίησης, συμπεριλαμβανομένων μεθόδων βαθμωτού υποβιβασμού, μπορούν να χρησιμοποιηθούν για να υπολογίσουν τις παραμέτρους του νευρωνικού δικτύου (Rumelhart et al., 1986).



Σχήμα Error! Unknown switch argument. **Τροφοδοτούμενο από μπροστά νευρωνικό δίκτυο**

Η δεύτερη σημαντική ιδιότητα των νευρωνικών δικτύων σχετίζεται με την χρήση τους σε εφαρμογές ταξινόμησης: ένα νευρωνικό δίκτυο μπορεί να εκπαιδευτεί να δώσει μία προσέγγιση της μεταγενέστερης πιθανότητας μίας κλάσης, δοθέντος της εισόδου. Μία δημοφιλής μέθοδος εκπαίδευσης του νευρωνικού δικτύου σε αυτήν την περίπτωση είναι να πραγματοποιήσουμε μία ελαχιστοποίηση ελαχίστων τετραγώνων όπου η επιθυμητή έξοδος τίθεται 1 όταν είναι παρούσα η επιθυμητή κλάση στην είσοδο και η επιθυμητή έξοδος τίθεται 0 σε κάθε άλλη περίπτωση. Κάποιος μπορεί να δείξει ότι πραγματοποιώντας αυτήν την ελαχιστοποίηση η έξοδος θα είναι μία εκτίμηση ελαχίστων τετραγώνων της πιθανότητας της κλάσης δοθέντος της εισόδου (White, 1989). Εάν το πρόβλημα ταξινόμησης ασχολείται με διάφορες κλάσεις, κατασκευάζεται ένα δίκτυο με τόσες εξόδους όσες κλάσεις, και, για μία δεδομένη είσοδο, επιλέγεται η κλάση που αντιστοιχεί στην υψηλότερη έξοδο.

Όπως αναφέρθηκε παραπάνω, ο υπολογισμός των παραμέτρων σε ένα νευρωνικό δίκτυο απαιτεί μία διαδικασία μη γραμμικής βελτιστοποίησης, η οποία είναι πολύ έντονη υπολογιστικά, ειδικά για μεγάλα προβλήματα όπως αναγνώριση συνεχούς ομιλίας.

Τα νευρωνικά δίκτυα έχουν χρησιμοποιηθεί για αναγνώριση συνεχούς ομιλίας μεγάλων λεξιλογίων με δύο τρόπους:

- Έχουν χρησιμοποιηθεί για μοντελοποίηση της πυκνότητας πιθανότητας εξόδου για κάθε κατάσταση σε μία HMM (Renal et al., 1992).

- Κατατμημένα νευρωνικά δίκτυα έχουν χρησιμοποιηθεί για μοντελοποίηση φωνητικών τμημάτων άμεσα από τον υπολογισμό της μεταγενέστερης πιθανότητας του φωνήματος δοθέντος της εισόδου (Austin et al., 1992).

Και στις δύο μεθόδους το σύστημα νευρωνικού δικτύου συνδυάζεται με το σύστημα HMM για την αύξηση της απόδοσης. Στην περίπτωση των κατατμημένων νευρωνικών δικτύων, το N-καλύτερο παράδειγμα χρησιμοποιείται για να παράγει πιθανές κατατμήσεις για το δίκτυο για να διαβαθμίσει. Χρησιμοποιώντας οποιανδήποτε μέθοδο, έχουν σημειωθεί μειώσεις στον ρυθμό σφάλματος λέξης περίπου 10% με 20%. Άλλες μέθοδοι νευρωνικών δικτύων έχουν επίσης χρησιμοποιηθεί σε διάφορα πειράματα αναγνώρισης συνεχούς ομιλίας με παρόμοια αποτελέσματα (Hild and Waibel, 1993; Nagai et al., 1993).

ΤΕΛΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ

Είμαστε στο χείλος μιας έκρηξης στην ενσωμάτωση της αναγνώρισης ομιλίας σε ένα μεγάλο αριθμό εφαρμογών. Η ικανότητα να πραγματοποιούμε βασισμένη σε λογισμικό, πραγματικού χρόνου αναγνώριση σε ένα υπολογιστή θα αλλάξει χωρίς αμφιβολία τον τρόπο που βλέπουν οι άνθρωποι την αναγνώριση ομιλίας. Καθένας με έναν υπολογιστή μπορεί τώρα να έχει αυτήν την δυνατότητα στο γραφείο του. Σε λίγα χρόνια, η αναγνώριση ομιλίας θα είναι ευρέως διαδεδομένη και θα μπει σε πολλές όψεις της ζωής μας. Αυτή η εργασία εξέτασε τις τεχνολογίες που έκαναν αυτές τις προόδους δυνατές.

Παρά όλες αυτές τις προόδους, πολλά είναι να γίνουν ακόμα. Η απόδοση της αναγνώρισης ομιλίας για πολύ μεγάλα λεξιλόγια και μεγάλες περιπλοκές δεν είναι ακόμα ικανοποιητική για χρήσιμες εφαρμογές, ακόμα και κάτω από ήπιες ακουστικές συνθήκες. Οποιοσδήποτε υποβιβασμός στο περιβάλλον ή αλλαγές μεταξύ των συνθηκών εκπαίδευσης και ελέγχου προκαλεί μία μείωση στην απόδοση. Έτσι, η εργασία πρέπει να συνεχίσει να αυξάνει την ρωμαλεότητα σε διάφορες συνθήκες: νέοι ομιλητές, νέες διάλεκτοι, διαφορετικά κανάλια (μικρόφωνα, τηλέφωνο), θορυβώδη περιβάλλοντα, και νέοι τομείς και λεξιλόγια. Πιο ειδικά χρειάζονται βελτιωμένα μαθηματικά μοντέλα ομιλίας και γλώσσας και μέθοδοι για γρήγορη προσαρμογή στις νέες συνθήκες.

Πολλές από τις περιοχές έρευνας θα απαιτήσουν πιο ισχυρούς υπολογιστές-περισσότερη υπολογιστική ταχύτητα και μνήμη. Μπορούμε τώρα να δούμε ωφέλειες για δύο τάξεις μεγέθους αύξηση στην ταχύτητα και στην μνήμη. Ευτυχώς, η ταχύτητα και η μνήμη των υπολογιστών θα συνεχίσει να αυξάνει με τα χρόνια που έρχονται. Το πιο ισχυρό υπολογιστικό περιβάλλον θα διευκολύνει την έρευνα πιο φιλόδοξων τεχνικών μοντελοποίησης και θα έχει, αναμφίβολα, σαν αποτέλεσμα επιπρόσθετες σημαντικές προόδους στην τελευταία λέξη της τεχνολογίας.

ΛΕΞΙΛΟΓΙΟ ΑΓΓΛΙΚΩΝ ΟΡΩΝ

allophone	αλλόφωνο	ο συστηματικός διαφορετικός τρόπος εκφοράς ενός φωνήματος σε μια γλώσσα.
allophonic	αλλοφωνικός	
articulation	άρθρωση	
cepstral smoothing	λείανση με cepstrum	εφαρμογή της μεθόδου cepstrum για φασματική λείανση, δες και ομομορφική επεξεργασία
Cepstrum		το αποτέλεσμα του μετασχηματισμού Fourier του λογαριθμικού φάσματος
corpus	σώμα υλικού ή κειμένων	
cluster	Συσσώρευση, συστάδα	
clustering	Συσταδοποίηση	
co-articulation	Συνάρθρωση	
code word	Κωδικολέξη	παράμετρος εισόδου μιας διαδικασίας ανυσματικής κβάντισης που αντιπροσωπεύει ένα σύνολο από ανύσματα (π.χ. ένα σύνολο φασματικών συνιστωσών)
codebook	Κωδικολεξικό, κωδικοβιβλίο	λεξικό κωδικολέξεων
coder	Κωδικοποιητής	
connected word recognition	Αναγνώριση συνδεδεμένων λέξεων	
Consonant	Σύμφωνο	
continuous speech	Συνεχής ομιλία	
Decoder	Αποκωδικοποιητής	
dental	Οδοντικό	
diphoneme	Διφώνημα	
Encoder	κωδικοποιητής	
Envelope	περιβάλλουσα	
frication	τυρβώδης θόρυβος	
fricative	τυρβώδες, εξακολουθητικό	/f/, /x/, /s/ κλπ
isolated word recognition	αναγνώριση μεμονωμένων λέξεων	
labial	χειλικό	
log spectrum	λογαριθμικό φάσμα	φασματική αναπαράσταση στην οποία ο κατακόρυφος άξονας εκφράζεται σε λογαριθμική κλίμακα
mel		μονάδα της αντίστοιχης μη γραμμικής κλίμακας που αναπαριστά την υποκειμενική κλίμακα αντίληψης του ύψους των ήχων από το αυτί του ανθρώπου
model	μοντέλο	
nasal	έρρινο	
neural network	νευρωνικό δίκτυο	
noise reduction	μείωση θορύβου	τεχνικές που μειώνουν ή απαλείφουν τις αρνητικές επιπτώσεις του θορύβου περιβάλλοντος σε ένα σύστημα αναγνώρισης ομιλίας
palatal	ουρανικό	
Perplexity	περιπλοκή	ποσότητα που εκφράζει τη μέση δυσκολία ή αβεβαιότητα αναγνώρισης λέξεων, όταν η αναγνώριση βασίζεται σε ένα μοντέλο γλώσσας. Ονομάζεται και μέσος παράγοντας διακλάδωσης λέξης του μοντέλου γλώσσας [*δες και* average word branching factor]
Phone	(φωνητικός) φθόγγος	μια συγκεκριμένη εκφώνηση ή μια συγκεκριμένη περίπτωση ενός φωνήματος
Phoneme	φώνημα	το φώνημα αποτελεί έναν ήχο αντίθεσης ή διάκρισης μεταξύ άλλων ήχων σε μια γλώσσα, ή φωνήματα είναι οι ήχοι που χρησιμοποιούνται σε μια γλώσσα για τη δημιουργία διαφορετικών λέξεων

phonetic transcription	φωνητική μεταγραφή	
Phonetics	φωνητική	
Phonology	φωνολογία	
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.
Place of articulation	θέση άρθρωσης	οι διάφορες θέσεις του στόματος που χρησιμοποιούνται για την παραγωγή των συμφώνων
plosive	Κλειστό μη έρρινο	
quantization	Κβάντωση, κβαντοποίηση, κβαντισμός	
quefrequency		μονάδα μέτρησης στον οριζόντιο άξονα του cepstrum
sampler	Δειγματολήπτης	
sampling	Δειγματοληψία	
segmentation	Κατάτμηση	διαδικασία που καθορίζει που αρχίζει και που τελειώνει ένα φώνημα
short-time spectrogram	Φασματογράφημα βραχείας διάρκειας	
speaker adaptation	Προσαρμογή ομιλητή	τεχνικές που προσπαθούν να τροποποιήσουν ένα σύνολο υποδειγμάτων σε ένα σύστημα αναγνώρισης ομιλίας ανεξάρτητης του ομιλητή, χρησιμοποιώντας νέα δεδομένα εκμάθησης από έναν συγκεκριμένο ομιλητή
speaker dependent recognition	Αναγνώριση εξαρτώμενη από τον ομιλητή	
speaker independent recognition	Αναγνώριση ανεξάρτητη του ομιλητή	
spectrogram	Φασματογράφημα	
spectrograph	Φασματογράφος	ειδική συσκευή παραγωγής φασματογραφημάτων
speech processing	Επεξεργασία ομιλίας	
speech synthesis	Σύνθεση ομιλίας	
tone	Τόνος	(στη γλωσσολογία) ο ιδιαίτερος τρόπος χρησιμοποίησης του μουσικού τόνου σε μια γλώσσα
triphoneme	Τριφώνημα	
utterance	Εκστόμιση, λεγόμενο	
vector quantization	Ανυσματική (ή διανυσματική) κβάντωση	
velar	Υπερωικό	
voiced	Ηχηρό, έμφωνο	ήχοι κατά την παραγωγή των οποίων πάλλονται οι φωνητικές χορδές, όπως οι [b, d, g, v, …]
vowel	Φωνήεν	
wide-band spectrogram	Φασματογράφημα ευρείας ζώνης	

ΠΙΝΑΚΑΣ ΣΥΝΤΜΗΣΕΩΝ

AI	Artificial Intelligence	
ARPA	Advanced Research Project Agency	
DFT	Discrete Fourier Transform	
FFT	Fast Fourier Transform	
HMM	Hidden Markov Model	