A Comparative Study of Mel Cepstra and EIH for Phone Classification under Adverse Conditions

by

Sumeet Sandhu

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

Bachelor of Science

and

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1995

Author ....
Department of Electrical Engineering and Computer Science
December 20, 1994

Certified by ....
James R. Glass
Research Scientist, MIT
Thesis Supervisor

Certified by ....
Oded Ghitza
Member of Technical Staff, AT&T Bell Labs
Thesis Supervisor

Certified by ....
Chin-Hui Lee
Member of Technical Staff, AT&T Bell Labs
Thesis Supervisor

Accepted by ....
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

A Comparative Study of Mel Cepstra and EIH for Phone Classification under Adverse Conditions

by

Sumeet Sandhu

Submitted to the Department of Electrical Engineering and Computer Science
on December 20, 1994, in partial fulfillment of the
requirements for the degrees of
Bachelor of Science
and
Master of Science

Abstract

The performance of current Automatic Speech Recognition (ASR) systems deteriorates severely in mismatched training and testing conditions. Signal processing techniques based on the human auditory system have been proposed to improve ASR performance, especially under adverse acoustic conditions. This thesis compares one such scheme, the Ensemble Interval Histogram (EIH), with the conventional mel cepstral analysis (MEL).

These two speech representations were implemented as front ends to a state-of-the-art continuous speech ASR and evaluated on the TIMIT database (male speakers only). To characterize the influence of signal distortion on the representation of different sounds, phonetic classification experiments were conducted for three acoustic conditions - clean speech, speech through a telephone channel and speech under room reverberations (the last two are simulations). Classification was performed for static features alone and for static and dynamic features, to observe the relative contribution of time derivatives. Automatic resegmentation was performed because it provides boundaries consistent with a well-defined objective measure. Confusion matrices were derived to provide diagnostic information.

The most notable outcomes of this study are (1) the representation of spectral envelope by EIH is more robust to noise - previous evidence of this fact from studies conducted on limited tasks (speaker dependent, small vocabulary, isolated words) is now extended to the case of speaker (male) independent, large vocabulary, continuous speech, (2) adding dynamic features (delta and delta-delta cepstrum) substantially increases the performance of MEL in all signal conditions tested, while adding delta and delta-delta cepstrum of EIH cepstrum - computed with the same temporal filters as those used for MEL - results in a smaller improvement. We suggest that in order to improve recognition performance with an EIH front end, appropriate integration of dynamic features must be devised.

Thesis Supervisor: James R. Glass
Title: Research Scientist, MIT

Thesis Supervisor: Oded Ghitza
Title: Member of Technical Staff, AT&T Bell Labs

Thesis Supervisor: Chin-Hui Lee
Title: Member of Technical Staff, AT&T Bell Labs

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech as constituted by articulated sounds is one of the most important modes of communication between human beings. Speaking is a skill usually learnt in infancy and used almost effortlessly from then onwards. The naturalness associated with speaking and hearing gives little indication of the complexity of the problems of speech processing. Several decades of research in different avenues of speech processing, such as the production, transmission and perception of speech, have yielded remarkable progress, but many fundamental questions still lack definitive answers. Part of the problem lies in the unique nature of speech as a continuous acoustic signal carrying a very large amount of information.

Speech is created by human beings by first forcing air through a constriction in the throat causing a vibration of the vocal cords, then by carefully shaping this air flow with the mouth by changing the relative position of the tongue, teeth and lips. The air pressure variations emitted from the mouth effect acoustic waves which usually undergo a number of changes while traversing the surrounding medium, before being perceived by the human ear. The decoding of these acoustic waves in the ear is an intricate process starting with the vibration of inner ear membranes and auditory nerve firings, converging via higher level neural analysis into psychological comprehension by the human being. Problems in various areas of speech processing have generally been approached in two ways, one of them involves modeling of the actual physiological processes responsible for speech production and perception, while the other treats speech as an acoustic signal exhibiting certain well-defined properties, regardless of the mechanism of speech production.

Several voice communications applications in use today are based on the second approach [37]. Their success rate has improved with the advent of low-cost, low-power Digital Signal Processing (DSP) hardware and efficient processing algorithms. Models of *speech synthesis* have been implemented in systems used in voice mail, voice banking, stock price quotations, flight information and recordings of read-aloud books. There are three main features of concern in speech synthesis systems - intelligibility and naturalness of the synthesized speech, the fluency and range of vocabulary of output speech, and the cost or complexity of the required software and hardware. For instance, announcement machines using pre-recorded speech provide high quality speech with low complexity but also low fluency, parametric systems like the Speak-n-Spell toy by Texas Instruments provide low quality, medium fluency speech at low to medium complexity, and full text-to-speech (TTS) systems such as those produced by Prose, DEC, Infovox and AT&T provide low to medium quality speech with high fluency using high complexity. The ideal synthesizer would provide high quality, high fluency speech using low complexity.

Research into the *transmission of speech* has yielded applications in wireless telecommunications and audio-video teleconferencing. *Speech coding* is used to compress the acoustic signal for transmission or storage at a lower bit-rate or over a narrow frequency bandwidth channel. Transmission applications include the wired telephone network where tight specifications are imposed in terms of quality, delay and complexity, the wireless network which has tighter requirements on bit-rate than the wired network but has more tolerance in quality and delay, and the voice security and encryption systems which generally use lower quality, longer delay and lower bit rate algorithms because of low available bandwidths. Applications of speech coding in storage are voice messaging and voice mail such as those used in telephone answering machines, and voice response systems used as telephone query processors in many large businesses. A growing area in speech transmission is the digital coding of wideband speech for applications like Digital Audio Broadcasting (DAB) of compact disk (CD) audio over Frequency Modulation (FM) channels, and surround sound for High-Definition Television (HDTV).

Automatic speech recognition (ASR) is largely aimed at facilitating and expediting human-machine interaction, e.g. replacing keyboards and control knobs with spoken commands interpreted through a voice interface. Commercial applications include voice dialing on mobile wireless phones, dictation, database access (e.g., flight reservations), eyes-free and hands-free machine control in factories and laboratories. Speech recognition techniques are also applied to *speaker verification* in banking, private branch exchanges (PBX), and along with speech synthesis techniques to *spoken language translation* and *spoken language identification.* There are three main features of practical ASR systems. One is the type of speech - isolated words (e.g., single words like "Collect" used for automatic collect-calling), connected words (e.g., credit card validation, digit dialing), or continuous speech (e.g., fluently spoken sentences). The second feature is speaker-dependence or speaker-independence, indicating that the system requires 'retraining' for new speakers if it is speaker-dependent and (generally) does not require retraining if it is speaker-independent. The third feature is the vocabulary size, which currently ranges from 10-digit phone numbers to about 20,000 words. Grammar constraints are often imposed on the recognized output to narrow down the possible choices via syntactic and semantic analyses. In the case of unrestricted input such as normal conversational speech, interpreting the meaning of recognized strings leads into another large area of research - *natural language processing.* The ultimate goal of speech recognition technology is to correctly recognize everything spoken by any person in any acoustic conditions, in the minimum possible time with the least cost.

The problem of speech recognition is made especially difficult by the variability in the speech signal arising from speaker-related and environment-related factors. These factors cause variations both across speakers and within a speaker. Speaker-related factors include pitch (e.g., male/female), articulation, intonation, accent, loudness, emotional or physical stress on the speaker, extraneous sounds (e.g., coughs, laughs, filled pauses such as "ums", "ahs") and unnatural pauses. The main environment-related factors are background noise (e.g., machine hum, side conversations, door slams, car engines), transmission channels (e.g., local or long-distance telephone lines, analog or digital telephone lines, wired or wireless telephone network) and recording microphones (e.g., carbon button or electret microphones, speakerphones, cellular phones). Since most of these are external variables and cannot be eliminated, their effects must be incorporated within the recognition process.

1.1 Basic Continuous Speech Recognition System

Work on the recognition problem was started four decades ago by using clean (relatively free of noise and distortion) isolated-word speech from a single speaker and bounding the problem with constraints such as a small vocabulary and simple grammar. After a series of breakthroughs in processing algorithms, ASR technology today has advanced to speaker-independent, large-vocabulary, continuous-speech recognition being developed on several systems around the world. Some examples of modern ASR systems are the HTK at Cambridge University, UK [51], the speech recognition system at LIMSI, France [21], the SPHINX system developed at CMU [25], the SUMMIT system at MIT [52], the TANGORA system at IBM [17], the Tied-Mixture system at Lincoln Labs [33] and the speech recognition system at AT&T Bell Labs [23]. Most of these systems adopt the pattern matching approach to speech recognition, which involves the transformation of speech into appropriate representations with signal processing techniques, creation of speech models via statistical methods, and the testing of unknown speech segments using pattern recognition techniques. Some of these systems also employ explicit acoustic feature determination (e.g., [52]) based on the theory of acoustic phonetics, which postulates that speech consists of distinct phonetic units characterized by broad acoustic features originating in articulation such as nasality, frication and formant locations. The ASR system used in this study is based on the pattern recognition approach and does not use explicit acoustic features. It has three main components - the signal representation module, the pattern matching module and the word and sentence matching module - as illustrated in Figure 1-1 [24].

The first module, *signal processing module*, serves to extract information of interest from the large quantity of input speech data, in the form of a sequence of 'feature' vectors (not to be confused with the acoustic-phonetic features mentioned earlier). This parametric representation generally compresses speech while retaining relevant acoustic information such as location and energy of formants.

Speech is divided into smaller units (along time) called subword units, such as syllables, phones and diphones, for the purposes of recognition. In the 'training' phase, feature vectors

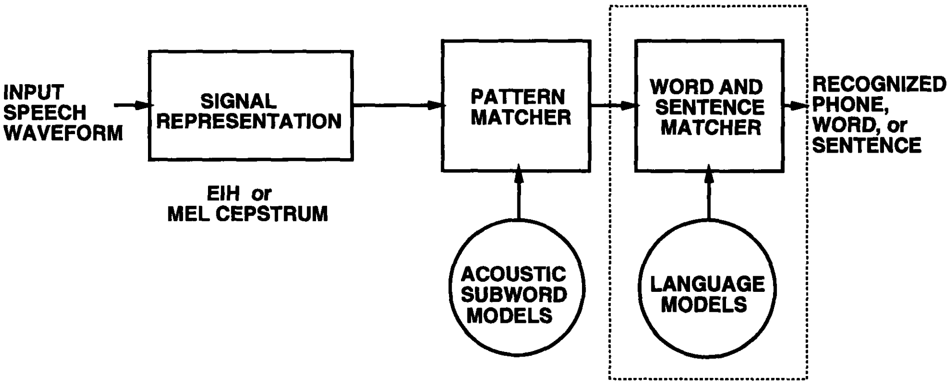


Figure 1-1: Schematic of an automatic speech recognition system.

extracted from speech segments corresponding to a subword unit are clustered together and averaged using certain optimal algorithms to obtain a 'characteristic' unit model. A set of models is obtained for all subword units in this fashion. In the 'testing' phase, the second module, the *pattern matcher*, uses some distance measure to compare the input feature vectors representing a given unknown speech segment to the set of trained unit models. The choice of subword units affects recognition accuracy (e.g., context-dependent or context-independent models [25, 24]), as does the choice of distance measure (e.g., Euclidean, log-likelihood [12]).

A list of top N candidates recognized for each given unit is passed to the third module, the *word and sentence level matcher.* This module performs lexical, syntactic and semantic analysis on candidate strings using a language model determined by the given recognition task, to yield a meaningful output. The third module is disconnected for the classification experiments conducted here, and the pattern matcher is modified to perform classification instead of recognition.

This study focuses on the choice of the signal representation module. Traditional spectral analysis schemes window the input speech at short intervals (10-20 milliseconds) and perform some kind of short-time Fourier transform analysis (STFT in chapter 6, [40]) to get a frequency distribution of the signal energy, preferably as a smooth spectral envelope. Two popular spectral analysis schemes are the filter bank model and the linear predictive coding (LPC) model (chapter 3 in [39]). The filter bank model estimates signal energy in uniformly or non-uniformly distributed frequency bands. LPC models the speech-sound production process, representing the vocal tract configuration which carries most of the speech related information. The filter bank and LPC representations are often transformed to the cepstral domain because cepstral coefficients show a high degree of statistical independence besides yielding a smoothened spectral envelope [29]. Cepstral analysis ([3]) is based on the theory of homomorphic systems (chapter 12 in [32]). Homomorphic systems obey a generalized principle of superposition; homomorphic analysis can be used to separate vocal tract and source information for speech signals [31]. A variation on cepstral analysis is the mel cepstrum, which filters speech with non-uniform filters on the mel frequency scale (based on human perception of the frequency content of sounds). These spectral analysis schemes and others derived from them are primarily based on energy measurements along the frequency axis.

1.2 Recognition in Noise

The speech signal is affected differently by various environment-related adverse acoustic conditions such as reduced signal-to-noise ratio (SNR) in the presence of additive noise, signal cancellation effects in reverberation or nonlinear distortion through a microphone [18]. It is impractical to train for diverse (and often unknown) signal conditions, therefore it is advisable to make the ASR system more robust. Several techniques for improving robustness have been proposed, including signal enhancement preprocessing [36, 7, 1], special transducer arrangements [49], robust distance measures [47] and alternative speech representations [28, 14, 45, 9]. Each of these techniques modifies different aspects of the ASR system shown in Figure 1-1.

In the case of distortion by additive noise, well-established *speech enhancement* techniques [27] can be used to suppress the noise. Such techniques generally use some estimate of the noise such as noise power or SNR to obtain better spectral models of speech from noise-corrupted signals. In particular, in [36] and [7], the enhancement techniques have been directly applied to speech recognition. In [36], the optimal least squares estimator of short-time spectral components is computed directly from the speech data rather than from an assumed parametric distribution. Use of the optimal estimator increased the accuracy for a speaker-dependent connected digit recognition task using a 10 dB SNR database from 58% to 90%. This method, however, uses explicit information about the noise level, which the algorithm in [7] avoids. In [7], the short-time noise level as well as the short-time spectral model of the clean speech are iteratively estimated to minimize the Itakura-Saito distortion [15] between the noisy spectrum and a composite model spectrum. The composite model spectrum is a sum of the estimated clean all-pole spectrum and the estimated noise spectrum. For a speaker-dependent isolated word (alphabets and digits) recognition task using 10 dB SNR test speech and clean training speech, the accuracy improved from 42% without preprocessing to 70% when both the clean training speech and the noisy testing speech were preprocessed. The main limitation is the assumption of the composite model spectrum for the noisy signal.

The work in [1] presents two methods for making the recognition system microphone-independent, based on additive corrections in the cepstral domain. In the first, SNR-dependent cepstral normalization (SDCN), a correction vector depending on the instantaneous SNR is added to the cepstral vector. The second method, codeword-dependent cepstral normalization (CDCN), computes a maximum likelihood (ML) estimate for both the noise and spectral tilt and then a minimum mean squared error (MMSE) estimate for the speech cepstrum. Cross-microphone evaluation was performed on an alphanumeric database in which utterances were recorded simultaneously using two different microphones (with average SNR's of 25 dB and 12 dB). SDCN improved recognition accuracy from 19%-37% baseline to 67%-76%, and CDCN improved accuracy to 75%-74%. The main drawbacks of SDCN are that it requires microphone-specific training, and since normalization is based on long-term statistical models it cannot be used to model a non-stationary environment. CDCN does not require retraining for adaptation to new speakers, microphones or environments.

In [49], several single-sensor and two-sensor configurations of speech transducers were evaluated for isolated-word recognition of speech corrupted by 95 dB and 115 dB sound pressure level (SPL) broad-band acoustic noise similar to that present in a fighter aircraft cockpit. The sensors used were an accelerometer, which is attached to the skin of the speaker and measures skin vibrations, and several pressure-gradient (noise-cancelling) microphones. Performance improvements were reported with various multisensor arrangements as compared to each single sensor alone, but the task was limited since the testing and training conditions were matched. Without adaptive training, there was no allowance for time-varying noise levels such as those caused by changing flying speed and altitude.

Robust distance measures aim to emphasize those regions of the spectrum that are less corrupted by noise. In [47], a weighted Itakura spectral distortion measure which weights the spectral distortion more at the peaks than at the valleys of the spectrum is proposed. The weights are adapted according to an estimate of the SNR (becoming essentially constant in the noise-free case). The measure is tested with a dynamic time warping (DTW) based speech recognizer on an isolated digit database for a speaker-independent speech recognition task, using additive white Gaussian noise to simulate different SNR conditions. This measure performed as well as the original unweighted Itakura distortion measure at high SNR's and significantly better at medium to low SNR's (at an SNR of 5 dB, this measure achieved a digit accuracy of 88% versus the original Itakura distortion which yielded 72%).

A technique for robust spectral representation of all-pole sequences, called the Short-Time Modified Coherence (SMC) representation, is proposed in [28]. The SMC is an all-pole modeling of the autocorrelation sequence of speech with a spectral shaper. The shaper, which is essentially a square root operator in the frequency domain, compensates for the inherent spectral distortion introduced by the autocorrelation operation on the signal. Implementation of the SMC in a speaker-dependent isolated word recognizer showed its robustness to additive white noise. For 10 dB SNR spoken digit database, the SMC maintained an accuracy of 98%, while the traditional LPC all-pole spectrum representation fell from 99% accuracy in clean conditions to 40%.

Rasta (RelAtive SpecTrAl) [14] methodology suppresses constant or slowly-varying components in each frequency channel of the speech spectrum by high-pass filtering each channel with a filter that has a sharp spectral zero at zero frequency [14]. This technique can be used in the log-spectral domain to reduce the effect of convolutive factors arising from linear spectral distortions, or it can be used in the spectral domain to reduce the effect of additive stationary noise. For isolated digits, with training on clean speech and testing on speech corrupted by simulated convolutional noise, the Rasta-PLP (Perceptual Linear Predictive) technique yielded 95% accuracy while LPC yielded 39% accuracy.

1.3 Motivation

While the energy based spectral analysis schemes such as LPC work well under similar acoustic training and testing conditions, the system performance deteriorates significantly under adverse signal conditions [4]. For example, for an alphanumeric recognition task, the performance of the SPHINX system falls from 77-85% accuracy with matched training and testing recording environments to 19-37% accuracy on cross conditions [1]. Robustness improving techniques explicitly based on matched training and testing conditions, noise level, SNR or the particular signal distortion, such as those described in Section 1.2, are clearly not desirable for robustness to multiple adverse acoustic conditions. The human auditory system seems to exhibit better robustness than any machine processor under different adverse acoustic conditions; it is successful in correctly perceiving speech in a wide range of noise levels and under many kinds of spectral distortions.

For robust speech recognition, speech processing models based on the human auditory system have been proposed, such as the representation in [45]. In this model, first-stage spectral analysis is performed with a bank of critical-band filters, followed by a model of nonlinear transduction in the cochlea that accounts for observed auditory features such as adaptation and forward masking [46, 13]. The output is delivered to two parallel channels, one of them yields an overall energy measure equivalent to the average rate of neural discharge, called the mean rate response, the other is a synchrony response which yields enhanced spectral contrast showing spectral prominences for formants, fricatives and stops. This auditory model has yielded good results for isolated word recognition [16, 26].

The EIH is another speech representation motivated by properties of the auditory system [9]. It employs a coherence measure as opposed to the direct energy measurement used in conventional spectral analysis. It is effectively a measure of the spatial (tonotopic) extent of coherent neural activity across a simulated auditory nerve. The EIH is computed in three stages - bandpass filtering of speech to simulate basilar membrane response, processing of the output of each filter by level-crossing detectors to simulate inner hair cell firings, and the accumulation of an ensemble histogram as a heuristic for information extracted by the central nervous system.

An evaluation of the EIH [9] was performed with a DTW based recognizer on a 39-word alpha-digit speaker-dependent task in the presence of additive white noise. In high SNR the EIH performed similarly to the DFT front end whereas at low SNR it outperformed the DFT front end markedly. Another study [16] involved the comparison of mel cepstrum and three auditory models - Seneff's mean rate response and synchrony response [44], and the EIH. It was performed on a speaker-dependent, isolated word task (TI-105 database) using continuous density Hidden Markov Models (HMMs) with Gaussian state densities. On the average, the auditory models gave a slightly better performance than mel cepstrum for training on clean speech and testing on speech distorted by additive noise or spectral variability (e.g., soft or loud speech, telephone model, recording environment model, head shadow).

This study differs from the previous evaluations in six ways : it uses a *continuous speech database* instead of isolated or connected word databases, the *size of the database* is much larger than those used earlier (4380 sentences as compared to the 39-word and 105-word vocabularies), *phone classification* is performed instead of word recognition, *mixture Gaussian HMMs* are used in contrast to the DTW based recognizer or the Gaussian HMMs used in previous experiments, *static*, and *static and dynamic features* are evaluated separately, and in addition to the average results, a *breakdown of the average results* into results for different phonetic groups is provided along with a qualitative analysis of confusion matrices of these groups.

The two speech representations, mel cepstrum and EIH, are implemented as front ends to a state-of-the-art continuous speech ASR and evaluated under different conditions of distortion on the TIMIT database (male speakers). The TIMIT is used because it is a standard, phonetically rich, hand segmented database. The recognizer is first trained on clean speech and then tested under three acoustic conditions - clean speech, speech through a telephone channel and speech under room reverberations (the last two conditions are simulated; training speech is also evaluated).

Evaluation is based on phone classification, where the left and right phone boundaries are assumed fixed and only the identity of the phone is to be established. Classification is performed, instead of recognition (which assumes no such prior information about the input speech), to focus on the front end and eliminate issues like grammar, phone insertion and phone deletion that are involved in the recognition process. The objective here is to observe the effects of signal distortion on the signal representation and statistical modeling.

The performance is displayed as average percent of phones correctly classified, and in the form of confusion matrices to provide diagnostic information. Classification experiments are conducted for (1) different sets of feature vectors, with and without time-derivatives, to observe the relative contribution of dynamic information and for (2) different iterations of automatic resegmentation of the database, which provides boundaries that are consistent with a well-defined objective measure, and is used in most current ASR systems.

The organization of the thesis is as follows :

1. Chapter 2 sketches the process of sound production, and contains a short description of different sound classes, which should serve to give some background for the discussion of results in Chapter 5. The classification system used in this work is based on an HMM recognition framework; a brief outline of Hidden Markov Models is also included.

2. Chapter 3 describes the two signal representations, mel cepstrum and Ensemble Interval Histogram, evaluated in this study. A brief description of part of the human auditory mechanism is given as background for EIH. The details of implementation for both representations are provided. The method of calculation of dynamic features

is described.

3. Chapter 4 contains a description of the experimental framework. The training and testing subsets used from the TIMIT database are described, along with the set of phones used for classification. The phone classification system and the distortion conditions - telephone channel and room reverberation simulations - are described.

4. Chapter 5 lists the results obtained for different classification experiments, with static features and static and dynamic features, and with automatically resegmented phone boundaries. The average classification results are discussed, and broad trends in the confusion matrices are observed.

5. Chapter 6 summarizes the work done in this thesis and the conclusions drawn from the results. Possible directions for future research are provided.

1.4 Summary

This chapter introduced the problem of automatic speech recognition and some of the issues in current speech research. The problem of recognizing noisy speech and different approaches taken to attain robustness in speech recognition were described. The next chapter contains a brief description of Hidden Markov Models and speech sounds, that should serve as background for the results in Chapter 5.

Chapter 2

Background

Speech is composed of a sequence of sounds carrying symbolic information in the nature, number and arrangement of these sounds. *Phonetics* is the study of the manner of sound production in the vocal tract and the physical properties of the sounds thus produced. This chapter contains an outline of the physiology of sound production in Section 2.1.1 and a brief description of the different sounds in American English classified by the manner of production in Section 2.1.2. The references are drawn mainly from Chapter 3 in [40] and Chapters 4 and 5 in [30]. A brief description of Hidden Markov Models (HMM's) is provided in Section 2.2. Further reading on HMM techniques can be found in [38].

2.1 Speech Production and Speech Sounds

2.1.1 Speech Production

The message to be conveyed via spoken sounds is formulated in the brain and then uttered aloud with the execution of a series of neuromuscular commands. Air is exhaled from the lungs with sufficient force, accompanied by a vibration of the vocal cords (or vocal folds) at appropriate times, and finally shaped by motion of the articulators - the lips, jaw, tongue, and velum. Figure 2-1 is a sketch of the mid-sagittal cross-section of the human vocal apparatus [39]. The *vocal tract* begins at the opening between the vocal cords called the glottis, and ends at the lips. It consists of two parts, the pharynx (the section between the esophagus and the mouth) and the oral cavity or the mouth. The nasal tract is the cavity between the velum and the nostrils.

Depending on the pressure gradient across the glottis, the mass and tension of vocal folds, two kinds of air flow are generated upwards through the pharynx, quasi-periodic and noise-like. Quasi-periodic (harmonically related frequencies) pulses of air are produced when the vocal folds vibrate in a relaxation oscillation at a fundamental frequency. These excite the resonant cavities in the vocal tract resulting in *voiced sounds* like vowels. Broad-spectrum (wide range of unrelated frequencies) noise-like air flow is generated by forcing air at a high enough velocity through a constriction in the vocal tract (while the vocal cords are relaxed), so as to cause turbulence. This produces *unvoiced sounds* like /s/ (as spoken in *s*ad), /f/ (*f*it) by exciting the vocal tract with a noise-like waveform, and ***plosive sounds*** like /b/ (*b*ig), /p/ (*p*ig) when air pressure is built up behind a complete closure in the oral cavity and then abruptly released.

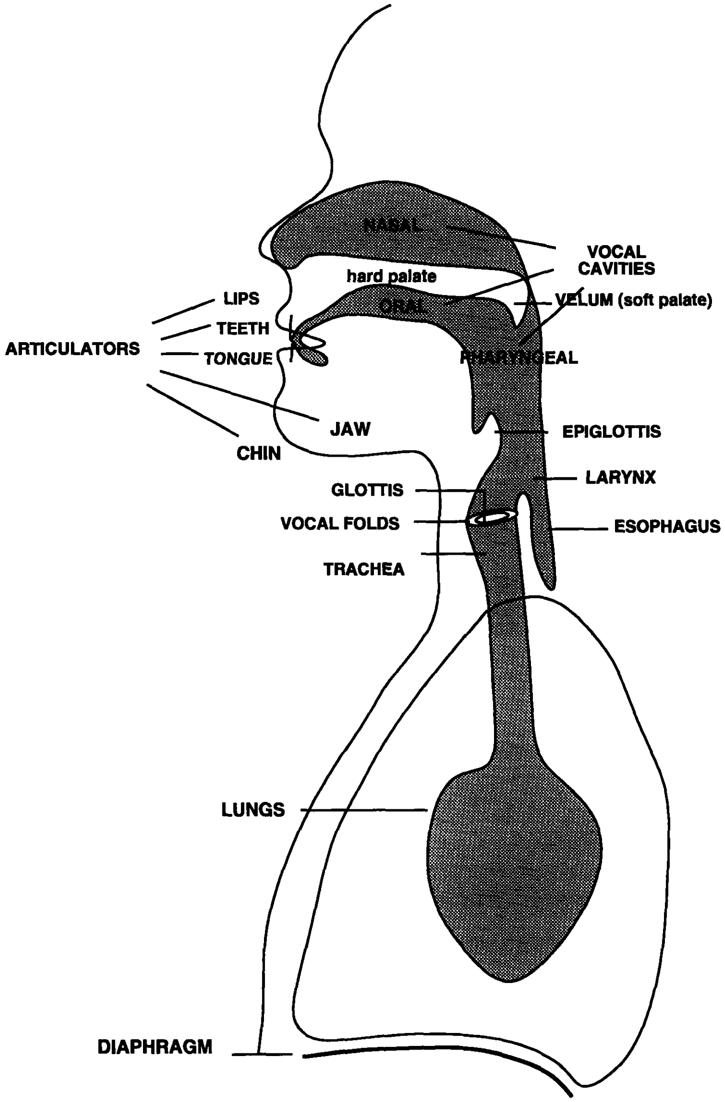


Figure 2-1: Cross section of the speech-producing mechanism

The vocal tract is a tube with a non-uniform and variable cross-sectional area. The nasal tract remains disconnected from the vocal tract for most sounds and gets acoustically coupled to it when the velum is lowered to produce nasal sounds. The resonant frequencies of the vocal tract are called *formants*; they depend on its dimensions in a manner similar to the resonances in wind instruments. Different sounds are formed by varying the shape of the tract to change its frequency selectivity. The rate of change of vocal tract configuration categorizes sounds as continuant and noncontinuant. The former are produced when a non time-varying vocal tract configuration is appropriately excited, for e.g., vowels and nasals (/m/ in *m*ild) , and the latter are produced by a time-varying vocal tract, for e.g., stop consonants (/d/ in *d*ip) and liquids (/l/ in *l*ip). Vowels, which are continuant sounds, can be differentiated into close, open, high, low and rounded based on the positions of the articulators. The consonants can also be alternatively classified by their *place-of-articulation*, for e.g., labial (lips), alveolar (gums), velar (velum), dental, palatal or glottal, along with the *manner-of-articulation* (plosive, fricative, nasals etc.). There are different ways of characterizing sounds based on the physical mechanism of production.

2.1.2 Speech Sounds

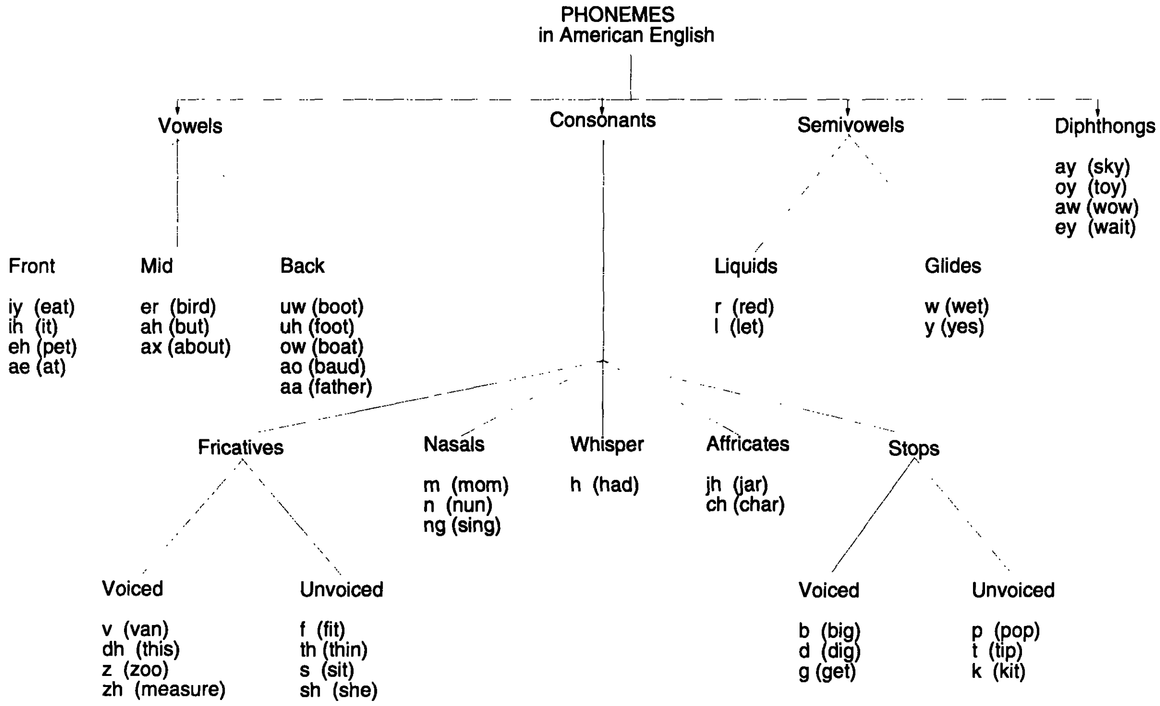


Figure 2-2: Phonemes in spoken American English

For the purposes of speech recognition, speech is segmented into subword units called *phones*, which are the acoustic realizations of abstract linguistic units called *phonemes.* Phonemes are the sound segments that carry meaning distinctions, identified by minimal pairs of words that differ only in part ([30]). For example, *fat* and *cat* are different sounding words that have different meanings. They differ from each other only in the corresponding first sounds /f/ and /c/, which makes /f/ and /c/ phonemes. Figure 2-2 shows the set of phonemes in spoken American English. The figure is based on the categorization in Chapter 4 in [30].

1. Vowels

Vowels are produced by an essentially time-invariant vocal tract configuration excited by quasi-periodic air pulses, and are the most stable set of sounds. They are generally longer in duration than consonants, and are spectrally well-defined. All vowels are voiced sounds.

Different vowels can be characterized by their first, second and/or third formants , which are determined by the area function - the dependence of the area on the distance along the vocal tract. The area function depends mainly on the tongue hump, which is the mass of the tongue at its narrowest point. Vowels can be classified by either the tongue hump position as front, mid, back vowels (shown in Figure 2-2), or by the tongue hump height as high, mid, low vowels. The vowel formant space is illustrated in Figure 2-3 based on formant values taken from Chapter 4 in [30]. The vowels are divided into different categories used later in Chapter 5, front, central and back, and high, mid and low.

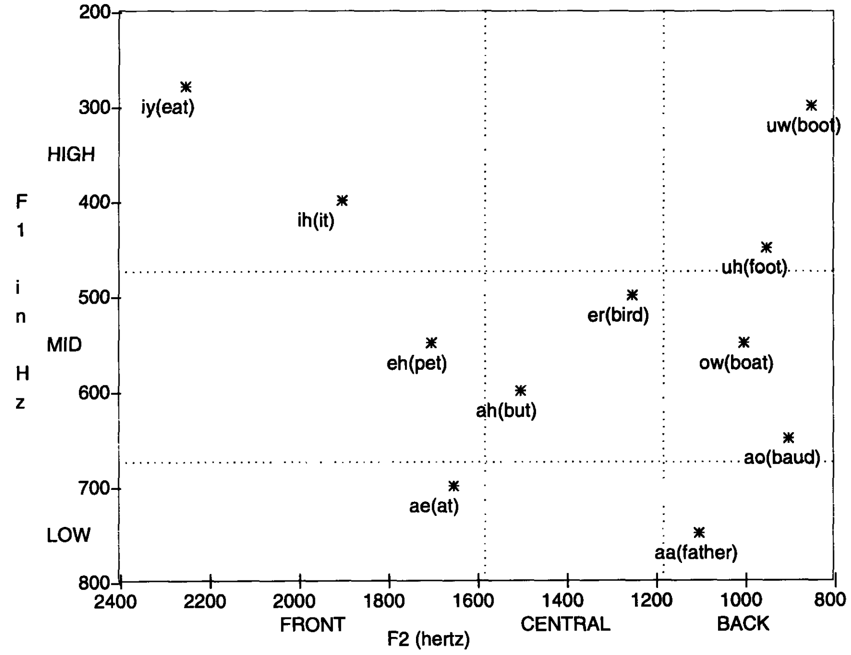


Figure 2-3: The vowel formant space

2. Diphthongs

   Diphthongs are described as combinations of certain 'pure' vowels with a second sound and are voiced. Figure 2-2 shows four diphthongs /ay/,/oy/, /aw/, /ey/.

   According to Chapter 3 of [40], a diphthong is defined as a gliding monosyllabic speech item that starts at or near the articulatory position for one vowel and moves to or toward the position for another. It is produced by smoothly varying the vocal tract configuration from that for the first vowel to the configuration for the second vowel.

   In Chapter 5 in [30], diphthongs are described as combinations of vowels and glides (glides are described after the liquids, which come next). /ay/, /ey/ and /oy/ are combinations of the corresponding vowels and the glide /y/. /aw/ is a vowel followed by the glide /w/, as /ow/ sometimes is (the diphthong /ow/ in m*ow*, as opposed to the vowel /ow/ in b*oa*t).

3. Liquids

   Liquids and glides are sometimes described as semivowels. Both are gliding transitions in vocal tract configuration between the adjacent phonemes. They consist of brief vowel-like segments and are voiced sounds, but their overall acoustic characteristics depend strongly on the context (neighboring phonemes).

   The liquids are /r/ and /l/. /r/ has low values for all first three formants, and is the only phoneme in American English that has a third formant below 2000 Hz. /l/ shows a slight discontinuity at vowel junctures, and is the only English lateral consonant (created by allowing air to pass on either side of the tongue). Both phonemes have an alveolar place-of-articulation.

4. Glides

   The glides are /w/ and /y/, and are voiced sounds. They are always followed by vowels, and have a labial and alveolar place-of-articulation respectively. They do not exhibit formant discontinuities at vowel junctures.

5. Nasals

   The nasal sounds /m/, /n/ and /ng/ are produced by lowering the velum to acoustically couple the nasal tract to the vocal tract, with glottal excitation and a complete constriction of the vocal tract at some point in the oral cavity. Sound is radiated through the nostrils in this case and the mouth acts as a resonant cavity. The resonances of the spoken nasal consonants and nasalized vowels are spectrally broader (more damped) than those of vowels.

   The total constriction made in the vocal tract is different for the three nasals: /m/ is a labial, /n/ an alveolar and /ng/ is a velar sound. The closures for these three sounds are made in the same place as the corresponding stops. Both kinds of sounds are produced with complete closure of the oral cavity, the difference is in the aperture of the velic port. For nasals, air is released through the nose and since there is no pressure buildup, no burst occurs when the oral closure is released.

   All three nasals have a prominent low frequency first formant, called the *nasal formant.* There are clear and marked discontinuities between the formants of nasals and those of adjacent sounds.

6. Fricatives

   Fricatives are characterized by the frictional passage of air flow through a constriction at some point in the vocal tract. The steady air stream used to excite the vocal tract becomes turbulent near the constriction. The back cavity below the constriction traps energy like the oral cavity in the case of nasals and introduces anti-resonances in the sound radiated from the lips.

There are two kind of fricatives, voiced and unvoiced. /v/, /dh/, /z/, /zh/ are voiced and /f/, /th/, /s/, /sh/ are unvoiced. The sounds in the two sets correspond exactly in terms of places-of-articulation, which are labio-dental, dental, alveolar and palatal respectively.

Out of the unvoiced fricatives, /th/ and /f/ have little energy, whereas /s/ and /sh/ have a considerable amount. All the unvoiced fricatives except /f/ show no energy below 1200 Hz. The voiced fricatives show energy in the very low frequency range, referred to as the *voice bar.*

7. Stops

   Stops (or stop consonants) are noncontinuant, plosive sounds. There are two kinds of stops, /b/, /d/, /g/ are voiced and /p/, /t/, /k/ are unvoiced. These also correspond in their places-of-articulation, which for both sets are labial, alveolar and velar respectively.

   No sound is radiated from the mouth during the time of total constriction of the vocal tract; this time interval is called the 'stop gap'. However, in the case of voiced stops, some low frequency energy is often radiated through the walls of the throat when the vocal cords can vibrate in spite of a total constriction in the tract. The distinctive characteristic of stops visible in a spectrogram is the presence of two distinct time segments, the closure and the burst. If the stop is followed by a voiced sound, the interval after the burst and before the sound is called the *voice onset time* (VOT). In the case of unvoiced stop consonants, after the stop gap, there is a brief period of friction followed by an interval of aspiration before voiced excitation begins. The duration and frequency content of the frication noise and aspiration vary with the unvoiced stop.

   Stops are noncontinuants, generally of short duration, and are more difficult to identify from the spectral information alone. Their properties are greatly influenced by the context.

8. Affricates

   An affricate consists of a stop and its immediately following release through the articulatory position for a continuant nonsyllabic consonant. The English affricates are /ch/ and /jh/, /ch/ is a concatenation of the stop /t/ and the unvoiced fricative /sh/. Similarly /jh/ is a combination of the voiced stop /d/ and the voiced fricative /zh/. The two affricates function as a single unit but their spectral properties are like other stop-fricative combinations.

   Affricates also display a closure-burst in the spectrogram, followed by the fricative region.

9. Whisper sounds

   The phoneme /h/ is produced by exciting the vocal tract with a steady air flow without vibration of the vocal cords. Turbulence is produced at the glottis. This is also the mode of excitation for whispered speech. The spectral characteristics of /h/ depend on the vowel following /h/ because the vocal tract assumes the position for the following vowel during the production of /h/.

2.2 Hidden Markov Models

The two speech representation, mel cepstrum and EIH, are described in Chapter 3. It is shown how a speech utterance, consisting of periodic measurement samples of the acoustic waveform, can be converted into a sequence of observation vectors, each of which corresponds to a fixed-length window or frame of N speech samples. The sequence of feature vectors, called a template, can therefore serve as a model of the speech utterance. However, different templates obtained from the same person speaking at different times are not identical. Clearly, there is even more variation across different speakers. One way to capture this variability is to model speech as a stochastic sequence.

A commonly used stochastic framework used in speech recognition is hidden Markov modeling (HMM). An overview of Hidden Markov Models can be found in [38]. HMMs are stochastic models of the speech signal designed to represent the short time spectral properties of sounds and their evolution over time, in a probabilistic manner. An HMM as used in speech recognition is a network of states connected by transitions. Each state represents the hidden probability distribution for a symbol from finite set of alphabets, such as a phone from a list of allowed phones. This output probability density function is used to determine the observation sequence by maximum likelihood estimation, using the transition probabilities associated with going from one state to the next.

Three key problems of HMMs are [38]:

1. Given the observation sequence $\mathbf{O} = \mathbf{O}_1\mathbf{O}_2 \dots \mathbf{O}_T$, and a model λ, what is the efficient way to find $P(\mathbf{O}|\lambda)$, the probability of the observation sequence given a model ? This is referred to as the scoring problem, i.e., given a model, compute the likelihood score of an observation. This is a measure of how well the utterance matches the model.

2. Given the observation sequence and the model, how should a state sequence $Q = q_1q_2 \dots q_T$ which is optimal in some meaningful sense be chosen ? This is referred to as the segmentation problem, because each vector $\mathbf{O}_i$ in the sequence is assigned to a state. For left-to-right models, since each transition can only be to the same state or the next state, this is equivalent to dividing the observation sequence into N segments, each corresponding to a state. That is, a set of transition times $\tau_1, \tau_2, \dots, \tau_N$, are obtained so that the vectors $\mathbf{O}_j$, where $\tau_i \leq j < \tau_{i+1}$ are assigned to state S_i.

3. How should the model parameters $\lambda = (A, B, \pi)$ be adjusted so as to maximize

$P(\mathbf{O}|\lambda)$? This is called the training problem, encountered when a maximum likelihood model for a word, syllable, etc. must be built from some training utterances.

The solutions of these problems are detailed in [38] and further references contained therein. A brief description of the solutions to each of the problems is given next.

The objective of *training* is to find a λ so as to maximize the likelihood of $P(\mathbf{O}|\lambda)$. The absolute maximum is difficult to find, and only iterative algorithms like the estimate-maximize (EM) algorithms can be used to find locally maximum solutions. A widely known training algorithm of this type is the Baum-Welch reestimation algorithm [34].

The actual training procedure used in experiments here is also an EM algorithm, called the segmental k-means training algorithm [41]. It trains a subword model out of several utterances of the same subword through the following steps:

1. **Initialization**: Each speech utterance, represented by a sequence of feature vectors, is uniformly segmented into states i.e. each state in a subword initially has roughly the same number of feature vectors assigned to it. The TIMIT hand segmentation is used to obtain phone boundaries.

2. **Estimation**: Using the data thus segmented, the parameters of each state are estimated. The feature vectors in each state are clustered into different mixture components using a K-means algorithm. The mean vectors, the variance vectors and the mixture weights are estimated for each cluster. The sample mean and covariance of the vectors in each cluster are used as maximum likelihood estimates of the mean vector and covariance matrix of each mixture component in the state. The weight of each mixture component is estimated based on the proportion of vectors in each cluster. This gives a first set of HMMs for all the states of all subword units.

3. **Segmentation**:

   The HMM thus estimated is used to resegment the training data into new units with Viterbi decoding.

4. **Iteration**: Steps 2 and 3 are repeated until convergence.

The acoustic space can be modeled by a finite number of distinguishable spectral shapes, which correspond to the states of an HMM. Alternatively, the speech signal can be viewed as being composed of quasi-stationary segments produced by stable configurations in the articulatory structures, connected by transitional segments produced during the time the articulatory structures evolve from one stable configuration to another. Each state of an HMM can thus either represent a quasistationary segment in the speech signal or a transitional segment. More generally, each state of an HMM can be used to model some salient features of sound so that it can be distinguished from other neighboring states.

In addition, HMMs can be hierarchical, such that each state or node can be recursively expanded into another HMM. Thus at the highest level, there can be a network of words, where each word is represented by a state in the HMM. Each word can expand into an HMM, whose states each represent a particular phone. Each of these word states can further be expanded into an HMM whose states each represent some acoustic sound or phone model.

2.3 Summary

This chapter outlined the speech production process, and described some characteristics of sounds classified by the manner of production. An outline of the Hidden Markov Modeling (HMM) procedure was also provided. The next chapter discusses the two feature extraction schemes that are used to represent speech, the mel cepstrum and the EIH.

Chapter 3

Signal Representations - Mel Cepstrum and Ensemble Interval Histogram

The two signal representations evaluated in this thesis are described in this chapter. The mel cepstrum (MEL) representation is based on Fourier analysis, and the Ensemble Interval Histogram (EIH) is based on auditory modeling. A brief introduction to the theoretical basis for both front ends is provided, followed by a description of the computational procedures. A short comparison of the two front ends is given at the end.

3.1 Mel Cepstrum

3.1.1 Mel Scale

Psychophysical studies have shown that human perception of the frequency content of sounds, for either pure tones or speech signals, does not correspond to a linear scale. The human ear is more sensitive to lower frequencies than the higher ones [48]. For each tone with a certain actual frequency in hertz, a subjective pitch is measured on the 'mel' scale. The pitch of a 1 kHz tone at 40 dB higher than the perceptual hearing threshold is defined as 1000 mels. The subjective pitch is essentially linear with the logarithmic frequency above 1000 Hz.

MEL accounts for this frequency sensitivity by first filtering speech with a filterbank which consists of filters that have increasing bandwidth and center-frequency spacing with increasing frequency [5].

3.1.2 Computation of MEL

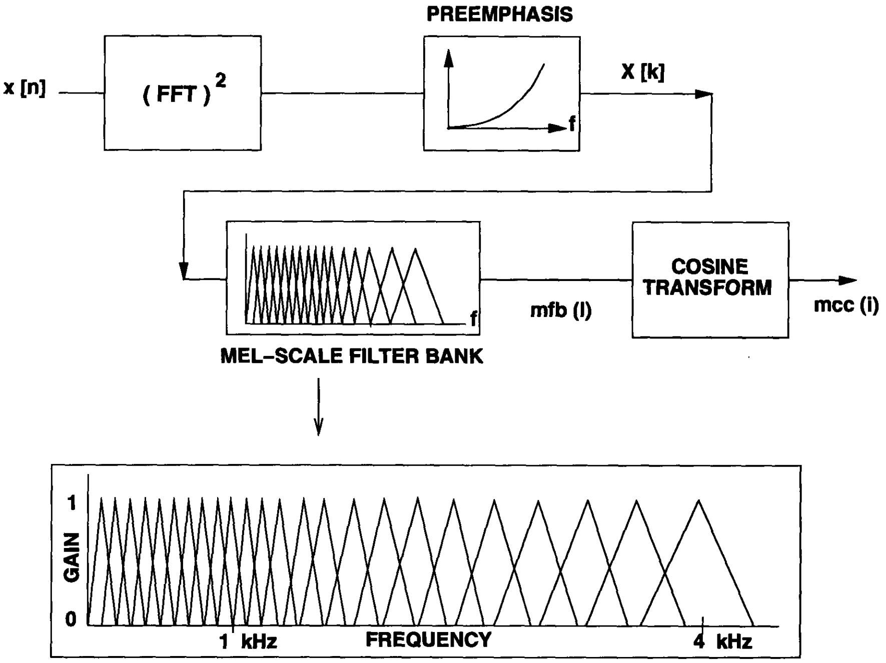


Figure 3-1: Computation of MEL Cepstrum coefficients.

MEL is computed in a standard manner [5, 42], as shown in Figure 3-1. The input speech at 8 kHz is windowed by a 20 ms long Hamming window every 10 ms, its magnitude-squared spectrum is pre-emphasized and passed through the mel-scale filter bank.

The mel filter bank consists of 24 triangular bandpass filters covering a bandwidth of 4 kHz. From 0 to 1 kHz there are 10 filters spaced linearly with a 200 Hz bandwidth per filter. Above 1 kHz there are 14 variable length filters spaced logarithmically (the center frequencies are powers of 1.1). The number of filters, NF, is selected to cover the signal bandwidth $[0, f_s/2]$ Hz, where f_s is the sampling frequency, $NF = 24$ and $f_s = 8$ kHz.

Let f_{c_l} be the center frequency of filter l, $l \in [1, NF]$. The lower and upper passband frequencies are $f_{c_{l-1}}$ and $f_{c_{l+1}}$ respectively, with $f_{c_0} = 0$ and $f_{c_l} < f_s/2 \, \forall l$. The triangular filter for an N-point DFT magnitude spectrum $X[k]$ is defined over DFT frequency index $k \in [0, N/2]$ as

$$F_l[k] = \begin{cases} ((\frac{k}{N})f_s - f_{c_{l-1}})/(f_{c_l} - f_{c_{l-1}}) & L_l \leq k \leq C_l \\ (f_{c_{l+1}} - (\frac{k}{N})f_s)/(f_{c_{l+1}} - f_{c_l}) & C_l \leq k \leq U_l \, , \end{cases}$$

where $C_l = \frac{f_{c_l}}{f_s}.N$, $U_l = \frac{f_{c_{l+1}}}{f_s}.N$, $L_l = \frac{f_{c_{l-1}}}{f_s}.N$ are the DFT indices corresponding to the l^{th} filter's center, upper and lower frequencies respectively.

The log energy output of filter l, denoted as mfb(l), is computed as

$$mfb(l) = log(\frac{1}{A_l}\sum_{k=L_l}^{U_l} F_l[k]X[k]), \qquad where\ \ A_l = \sum_{k=L_l}^{U_l} F_l[k], \tag{3.1}$$

A_l is a normalizing factor introduced to account for the varying bandwidths of the filters. The outputs from all *NF* filters constitute a mel-filter bank vector.

The mel filter bank vectors are further transformed into cepstral coefficients which show a high degree of statistical independence. This is because the cosine transform used to obtain cepstra is close to the optimal orthogonalizing *Karhunen-Loève* transform [29]. Cepstral coefficients are computed from the mel filter bank vectors using the inverse cosine transform as given by

$$mcc(i) = \frac{1}{NF}\sum_{l=1}^{NF} mfb(l)cos(i(l-\frac{1}{2})\frac{\pi}{NF}) \qquad i = 1, ..., NF-1. \tag{3.2}$$

The mcc(0) coefficient is a measure of the average log energy in the speech frame. The energy used in the feature vectors in this work is the frame-normalized energy (normalized 0 to -75 db as required for the classification system). One MEL frame is processed every 10 ms.

3.2 Ensemble Interval Histogram

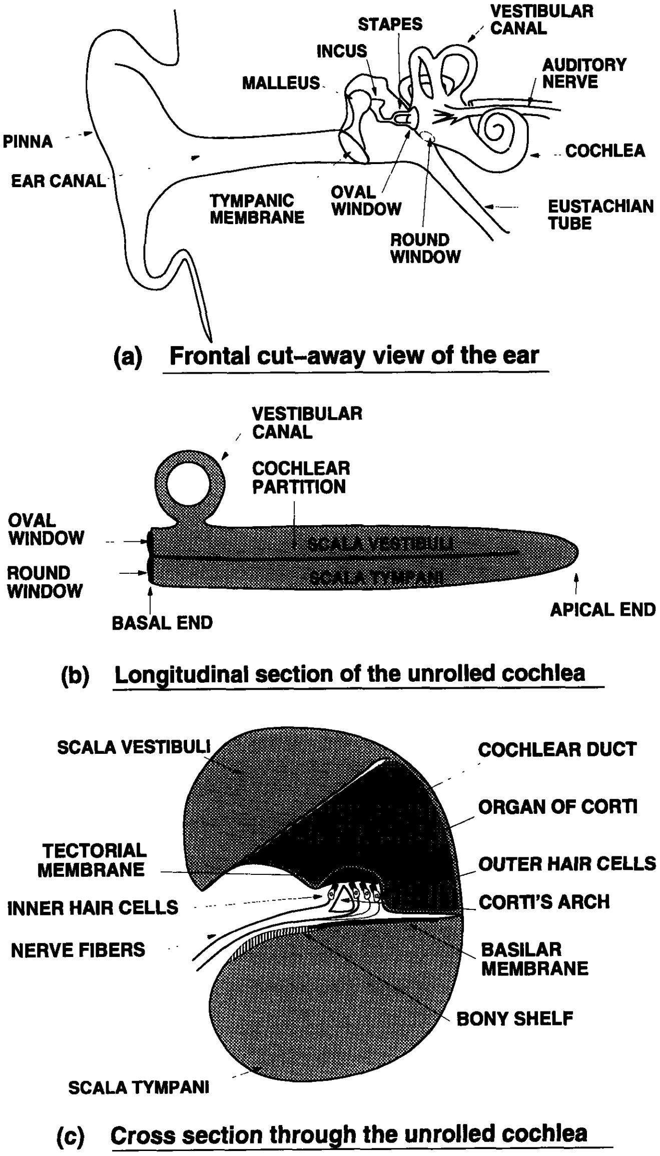
In recent years, the use of auditory models that perform comparably to the human auditory system in tasks related to speech perception has been proposed for speech processing. In speech coding, for example, bit rate can be lowered based on perceptual tolerance to acoustic deviations in speech [43]. In a similar fashion, spectral feature extraction modules based on auditory models that incorporate perceptual invariance to adverse signal conditions (for e.g., noise, channel distortions) and phonetic variability (for e.g., due to inter and intra speaker differences) may prove to be viable front ends for robust speech recognition. These models are based on physiological and psychophysical studies of the auditory system [35, 19]; the pre-auditory nerve region of the human auditory system is briefly described in Section 3.2.1.

3.2.1 The Human Auditory System

An introductory treatment of the science of speech and hearing is given in [6], from where most of the following description has been extracted.

Figure 3-2 is a sketch of a part of the auditory system. The ear has three parts, the outer ear, the middle ear and the inner ear. The outer ear consists of the pinna, the largely cartilaginous projecting portion, and the ear canal, an air-filled passageway terminated in the eardrum, that acts like an acoustic resonator (amplifying frequencies around 3-4 kHz).

The middle ear consists of the tympanic membrane (eardrum) and a mechanical transducer consisting of the three auditory ossicles, malleus (hammer), incus (anvil) and the stapes (stirrup), which converts sound impinging on the eardrum into mechanical vibrations in the inner ear. The hammer is rigidly attached to the eardrum and successively transmits its motion to the anvil and stirrup, where the stirrup is connected to the oval window on the inner ear. The middle ear amplifies sound pressure from the outer ear to the inner ear by about 35 times with the lever action of the ossicles.



(a) Frontal cut-away view of the ear

(b) Longitudinal section of the unrolled cochlea

(c) Cross section through the unrolled cochlea

Figure 3-2: Parts of the human auditory mechanism

The inner ear consists of the cochlea and the auditory (cochlear) nerve, a small intricate system of cavities in the bones of the skull. The first major transformation of sound from mechanical vibrations to nerve impulses takes place in the cochlea. The cochlea is a coiled fluid-filled chamber that can be 'unrolled' as shown in Figure 3-2(b). The cochlear partition is a membranous structure that divides the cochlea into two parts (scala vestibuli and scala tympani) along most of its length. The interior of this partition forms a third region called the cochlear duct, as shown in the cross section of the cochlea in Figure 3-2(c). The basilar membrane separates the cochlear duct from the scala tympani, and is connected to a bony shelf that extends out of the central core of the cochlea. The relative widths of the basilar membrane and the bony shelf vary gradually along the length of the cochlea, the basilar membrane being the narrowest at the basal end (about 0.04 mm) and the widest at the apical end (about 0.5 mm). The membrane is very stiff and light near the oval window but slack and massive at the apical end.

The cochlear structure is excited through the oval window by motion of the stirrup. The mechanical properties of the basilar membrane determine to a large extent the response of the cochlear partition to this excitation. A pulse-like excitation, such as a click, causes the partition to first bulge downwards at the basal end into the scala tympani. This bulge in the partition then travels along the cochlea toward the apical end, broadening as it moves. A sine-wave excitation causes vibrations in the entire partition, but the amplitude of vibration at different points along the partition depends on the applied frequency. For high frequencies, the point of maximum amplitude is near the basal end, and for low frequencies, the vibration is highest near the apical end (the partition goes through a more complex motion than a simple up and down vibration in each cycle). All this leads to a spatial separation of the maximum response to stimulation at different frequencies.

Conversion of the mechanical motion of the cochlear partition and the attached basilar membrane into neural signals takes place in the Organ of Corti, a collection of cells lying on the basilar membrane in the cochlear duct. The sensory receptors in the Organ of Corti are the hair cells, which have very fine hairs at one end that make contact with the tectorial membrane, the other end rests on the basilar membrane. Two kinds of hair cells lie on either side of a 'V-shaped' pair of rods constituting the Corti's Arch. The inner hair cells (IHC's) lie on the side of Corti's Arch closest to the central core of the cochlea. The outer hair cells lie on the other side of the arch. There is one row of inner hair cells and three rows of outer hair cells along most of the length of the basilar membrane, adding to about 3500 inner cells and 20,000 outer cells. Nerve fibers from the auditory nerve extend into the Organ of Corti, and the endings of these fibers lie close to the hair cells. When the basilar membrane vibrates in response to sound waves, the hair cells get bent and stimulate the nerve fibers producing electrochemical pulses that are sent to the brain along the auditory nerve.

The mechanical displacement of the basilar membrane at any given place can be viewed as the output signal of a band-pass filter with a spectral resonance peak at the *characteristic frequency* (CF). The CF is characteristic of the place along the membrane, and its log is approximately proportional to the distance along the membrane. The displacement of the basilar membrane is reflected in the AC component of the IHC receptor potential. Receptor cells, such as the hair cells in the Organ of Corti, receive sensory information from their environment and help to code it into the electrochemical pulses that are transmitted to the nervous system. The transformation from mechanical motion to receptor potential (voltage) involves several nonlinearities. One of them is the half-wave rectification resulting from the unidirectional depolarization of the IHC. Each IHC is innervated by about 10 auditory fibers whose spontaneous activity discharge depends on the fiber diameter and size of the synaptic region between the fiber and inner hair cell, and on the threshold of response.

Studies of cochlear mechanics and of the mechanical to neural transduction in the cochlea provide valuable information about the processing of sounds in the pre-auditory nerve stage of the auditory periphery. The auditory nerve connects the cochlea to the central nervous system and is the only afferent [1] path between them. This makes it the sole carrier of speech-signal information to the central nervous system, and therefore the first step in the modeling. Studies of the population response of single auditory nerve fibers in cats to speech-like signals have provided information about the encoding of such sounds in the auditory nerve. However, relatively little is known about the functioning of the auditory mechanism beyond the auditory nerve, lumped as the 'central processor'.

Most auditory models differ in the structural properties of the central processor, which can be described by the following two characteristics: place-nonplace component, which determines whether the processor uses explicit information about the nerve fiber's tonotopic place of origin in the cochlear partition, and the rate-temporal component, which determines whether the central processor uses instantaneous firing rate measurements alone or higher order statistics as well. The EIH belongs to the nonplace-temporal category, it omits place information and uses only the temporal properties of the global neural response.

The EIH utilizes detailed physiological modeling of the auditory periphery to simulate the firing activity of the auditory nerve, followed by a heuristic transformation to account for the higher order processing. It is a measure of the spatial (tonotopic) extent of coherent neural activity across the simulated auditory nerve [10].

3.2.2 Computation of EIH

The EIH is computed in three stages - bandpass filtering of speech to simulate basilar membrane response, processing of the output of each filter by level-crossing detectors to simulate IHC firings, and the accumulation of an ensemble histogram as a heuristic for information extracted by the central nervous system [9] as shown in Figure 3-3.

The first stage models the middle ear with a bank of mel scale filters. The mel-like bandpass filters (similar in shape and distribution to mel filters) are spaced from 0-4 kHz. In this study, 85 filters are used to cover the frequency range, which is approximately equivalent to 2 more filters being placed in between adjacent mel scale filters equally spaced in logarithmic frequency. This stage represents the auditory periphery up through the level of the auditory nerve.

The mechanical to neural transduction is modeled in the second stage. The ensemble of

[1] i.e., conveying nerve spikes towards a nerve center

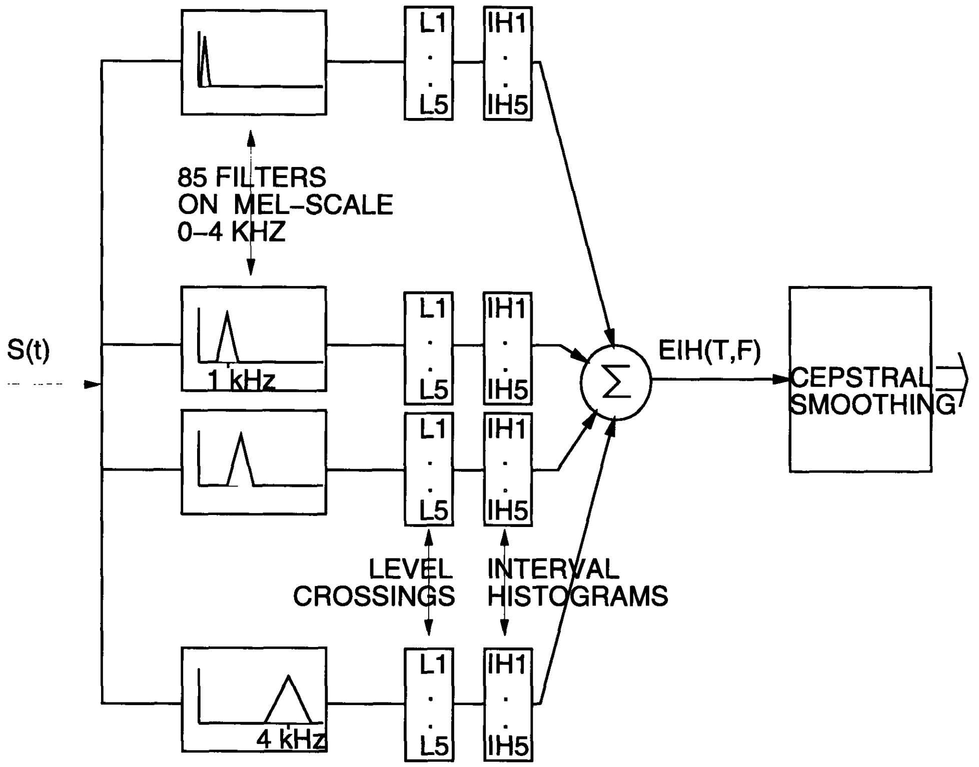


Figure 3-3: Computation of EIH Cepstrum coefficients.

nerve fibers innervating a single IHC is simulated with an array of level-crossing detectors at the output of each mel-like filter used in the first stage. Each detector models a fiber of a specific threshold, and a neural firing is simulated as the positive-going level crossing. The thresholds are distributed across a range of positive values which accounts for the half-wave rectification of the IHC receptor potential. The value assigned to detector-level j of every filter is a random Gaussian variable with mean L_j and standard deviation $0.2L_j$. The means L_j are uniformly distributed on a log scale over the amplitude range characteristic of speech sounds. The detectors are pseudorandomly assigned a range of positive levels to reflect the variability of the nerve fiber diameters and synapse-connection sizes. The nerve fiber firing activity is therefore modeled as a point process produced by a level-crossing detector. This neglects the probabilistic nature of the neural firing mechanism, so the level crossings are to be interpreted as the combined firing activity of a collection of fibers originating in different IHC's located close enough along the basilar membrane to exhibit similar cochlear tuning characteristics.

The coherent activity across the simulated fiber array is measured by determining the similarity in the short-term interval probability density functions of individual level-crossing detectors. An estimate of the interval probability density function of a level is obtained by computing a histogram of the intervals from the point process data (time intervals between successive upward-going level crossings). To obtain a frequency-domain function, the histogram of the reciprocal of the intervals is computed. The similarity across all levels and channels is measured by collecting individual histograms into one ensemble interval histogram (EIH) for 128 frequency bins ranging from 0-4 kHz. This is done every 3.2 ms with a back-in-time window whose width depends on different values of CF. The width is greater at the lower frequencies than the higher frequencies to model observed properties of auditory nerve fibers. At time t_0, intervals produced by a level-crossing detector located at characteristic frequency CF_0 are collected over a window of length $\frac{10}{CF_0}$ that ends at time t_0. This representation exhibits fine frequency resolution at low CF's and fine temporal resolution at high CF's.

A data reduction and smoothing is performed on the 128-component EIH vectors. The frame 'energy' is calculated from the histogram as the sum over 128 bins. Cepstral-like analysis is then performed on the normalized EIH (normalized so that the sum equals 1) to get 12 coefficients. The dynamic range of the frame energy is about 0 to -2.0 units of 'loudness'. An EIH frame is computed every 3.2 ms, then three successive frames are averaged to get an EIH frame every 9.6 ms.

3.3 Comparison of MEL and EIH

For both the speech representations, the first stage of analysis is performed using mel-scale filters. In the case of MEL, the 'filters' are not used to filter incoming speech as such, in the case of EIH, they are used as filters in the conventional sense. The significant difference in approach is in the second stage, as shown in Figure 3-4, namely the energy-based approach versus the coherence-based approach.

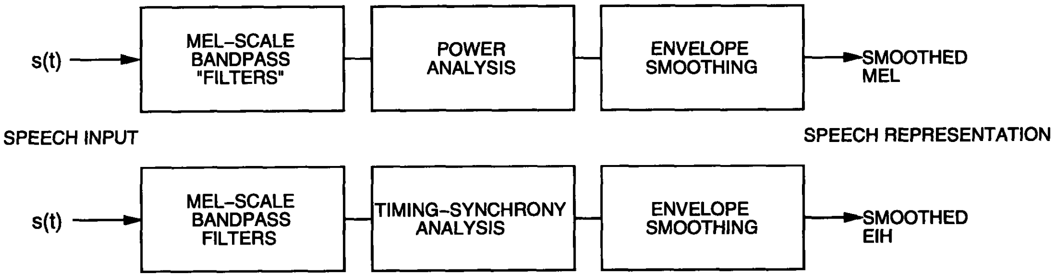


Figure 3-4: Comparison of MEL and EIH

The third stage in both cases is the envelope smoothing, based on *homomorphic analysis* (Chapter 12 in [32], [31]), yielding 12 cepstral coefficients.

A plot of the frame energy of an utterance is shown in the Appendix in Figure A-3 for MEL and for EIH.

3.4 Dynamic features

The 12-component vectors obtained from both front-ends are variously augmented by the energy term and/or the corresponding dynamic features for different experiments. Two sets of static features and two sets of static and dynamic features are used. The four sets of parameters are : **spectral envelope** alone (12 cepstral coefficients for MEL and for EIH), **spectral envelope** and **energy** (13 coefficients), **spectral envelope** and its **time derivatives** (12 cepstral coefficients, 12 delta cepstrum and 12 delta-delta cepstrum, giving 36 coefficients for MEL and for EIH), and **spectral envelope** and **energy** and their **respective time derivatives** (12 cepstral coefficients, 12 delta cepstrum, 12 delta-delta cepstrum, 1 energy, 1 delta energy and 1 delta-delta energy, giving 39 coefficients for MEL and for EIH).

For both the front ends, dynamic features are computed as follows.

The *delta cepstrum* of the sequence of cepstral vectors is approximated by an orthogonal polynomial over a finite length window of $(2K+1)$ frames centered around the current vector ($K = 2$ corresponding to a 5-frame window) ([23], Chapter 4 in [39]), as

$$\Delta\hat{c}_l(m) = [\sum_{k=-K}^{K} k\hat{c}_{l-k}(m)] \cdot G_{corr}, \qquad l \leq m \leq Q \tag{3.3}$$

where G_{corr} is a gain term chosen to equalize the variances of $\hat{c}_l(m)$ and $\Delta\hat{c}_l(m)$.

The *delta-delta cepstrum* is calculated from the delta cepstrum as

$$\Delta_2\hat{c}_l(m) = K \cdot [\Delta\hat{c}_{l+1}(m) - \Delta\hat{c}_{l-1}(m)], \tag{3.4}$$

where $\Delta\hat{c}_l(m)$ is the estimated m^{th} delta cepstrum coefficient evaluated at frame l as in Equation 3.3, and K is a scaling constant fixed at 0.375 [23].

It should be noted that the dynamic features for both speech representations were calculated with the same method. Since there was no previous study on the calculation of dynamic information for EIH, the MEL temporal filters were used for EIH as well.

3.5 Summary

This chapter discussed the characteristic features of the two speech representations - the mel cepstrum analysis and the Ensemble Interval Histogram. The motivation behind the two front ends was briefly discussed along with the procedures for computing their respective parameters. The next chapter describes the experimental framework used to compare the two front ends for phone classification.

Chapter 4

Experimental Framework

The database used is the TIMIT [22], because it is a standard, phonetically rich database available with phonetic transcriptions. The recognizer is first trained on clean TIMIT training set (male speakers only) and then the TIMIT test set (male speakers only) is tested under three acoustic conditions - clean speech, speech through a telephone channel and speech under room reverberations (the last two conditions are simulated). Evaluation is based on phone classification, where the left and right phone boundaries are assumed fixed and only the identity of the phone is to be established. To focus on the front end, classification is performed instead of recognition, eliminating issues like grammar, phone insertion and phone deletion. The goal of this study is to observe the effects of signal distortion on the signal representation and statistical modeling.

The TIMIT speech files are provided at 16 kHz sampling rate with 16 bit Pulse Code Modulation (PCM) samples. They are first lowpass filtered and downsampled to 8 kHz. The MEL and EIH cepstral coefficients are then calculated with appropriate spectral analysis; 12 coefficients are computed for both. These frames are augmented with either the energy and/or first and second order derivatives to obtain static and dynamic features, as described at the end of Chapter 3.

4.1 Database

The TIMIT database is available on a CDROM (NIST Speech Disc CD1-1.1,October 1990). The speech was recorded using a Sennheiser HMD-414-6, close-talking, noise-cancelling, headset-boom microphone in sound-treated room. The database consists of about 6225 words and 6300 sentences spoken by 630 male and female speakers from 8 major dialect regions of the U.S. Each speaker utters 10 sentences which are designed to cover a variety of phonetic contexts in spoken American English. The database is divided into training and testing sections with no overlapping speakers. Out of the 10 sentences per speaker, 2 are common to all speakers in both testing and training and have been left out to avoid undue bias towards certain phonetic contexts. The remaining 8 sentences per speaker are used for testing or training. Only the utterances by male speakers (326 train, 112 test speakers) are used in this study.

The hand-labeled phonetic transcriptions of the speech files provided in the TIMIT database are used to obtain phone boundaries for classification. To focus on broader phone classes, the 61 phones used in TIMIT segmentation are collapsed into a set of 47 phones shown in Table 4.1. The allophone listed for each phone is the original phone that was combined or removed. The closures for all the stops (/b/,/d/,/g/,/p/,/t/,/k/) and the affricates (/jh/,/ch/) are merged with the respective stops and affricates. Some other similar sounding phones have been combined, such as the /hh/ (*h*ay) and /hv/ (a*h*ead) pair, the /uw/ (b*oo*t) and /ux/ (t*oo*t) pair, and the three different schwa sounds in *a*bout, deb*i*t and s*u*spect. The glottal stop (merged with epenthetic silence) was found to possess too small a variance in the model training process, so it has been ignored. A single silence model is used for the silences in the beginning and end of utterances and the mid-utterance pauses.

Table 4.1: Set of 47 phones used and their TIMIT allophones

Phone	Word	Allophn	Phone	Word	Allophn	Phone	Word	Allophn
h#	silence	pau	aa	f***a***ther		ae	b***a***t	
ah	b***u***tt		ao	b***ou***ght		aw	b***ou***t	
ax	n***u***ll	ax-h ix	axr	butt***er***		ay	b***i***te	
b	***b***ee	bcl	ch	***ch***ild		d	***d***ay	dcl
dh	***th***en		eh	b***e***t		el	bott***le***	
em	bott***om***		en	butt***on***		er	b***i***rd	
ey	w***ai***t		f	***f***in		g	***g***ame	gcl
hh	***h***ome	hv	ih	b***i***t		iy	b***ee***t	
jh	***j***oke		k	***k***ey	kcl	l	***l***ike	
m	***mom***		n	***noon***		ng	si***ng***	eng
ow	b***oa***t		oy	b***oy***		p	***p***ay	pcl
r	***r***ed		s	***s***ea		sh	***sh***e	
t	***t***ea	tcl	th	***th***in		uh	b***oo***k	
uw	b***oo***t	ux	v	***v***ery		w	***w***ell	
y	***y***es		z	***zoo***		zh	mea***s***ure	
dx	mu***dd***y		nx	wi***nn***er				

4.2 Classification System

The HMM continuous speech recognition framework used in this study is described in detail in [23]. Each speech unit is modeled as a left-to-right 3-state HMM (except silence, which is single state). A continuous density is used to describe the observation probability density of each state as a weighted sum (mixture) of multivariate Gaussian densities (a maximum of 32 Gaussian mixture components are used per state). The covariance matrix for each Gaussian mixture component is assumed diagonal. The self and forward transitions within and between states are assumed equally likely.

Context-independent subword unit models are trained using a variant of the segmental k-means algorithm [41] with the given TIMIT segmentation. This process does not modify the TIMIT phone boundaries or the state boundaries within a phone (which are determined by uniformly segmenting each subword, except silence, into 3 states); these boundaries are maintained in the experiments with different feature vectors.

In the testing phase each speech segment is compared with all phone models using the Viterbi algorithm [50]. Likelihood scores are obtained for the top 1 and top 3 candidate phones.

New phone (and state) boundaries for the training data can be determined via Viterbi decoding with the current HMM models. The new training data segmentation can be used to build a new set of HMM models. The phone boundaries for the test data are modified in correspondence to the model-building iterations. This is done for the automatic resegmentation experiments. Initially, the TIMIT segmentation is used to train a set of HMM's. For each iteration after that, a set of HMM's is first trained using the current training data segmentation. These HMM's are used to classify the speech data (clean training speech, clean test speech, test speech passed through the two distortion simulations) using phone boundaries given by the current segmentation. This set of HMM's is also used to resegment the respective data (both the train and test sets) via Viterbi decoding. The process is then repeated with this new set of phone boundaries for the next iteration of resegmentation. In this study, the results for 3 iterations are recorded. The process of resegmentation changes phone and state boundaries, therefore the resultant sub-word units used in these classification experiments are phone-like units and do not necessarily agree with the phoneme labels associated with them.

4.3 Distortions

The TIMIT (male) testing data is run through two kinds of distortion simulations : telephone channel and room reverberation. Waveforms and spectrograms of a clean, sample TIMIT sentence and its two "noisy" versions are attached in the Appendix in Figure A-2 for illustration purposes.

4.3.1 Telephone Channel Distortion

The telephone channel simulation is illustrated in Figure 4-1. White noise is first added to the signal. The acoustic conditions in the telephone channel are simulated, but the telephone handset or receiver is not modeled.

The telephone channel simulation "wire" [20] provides several choices of telephone channels and noise, for example, AT&T data or voice channels, doppler shift, phase jitter, har-

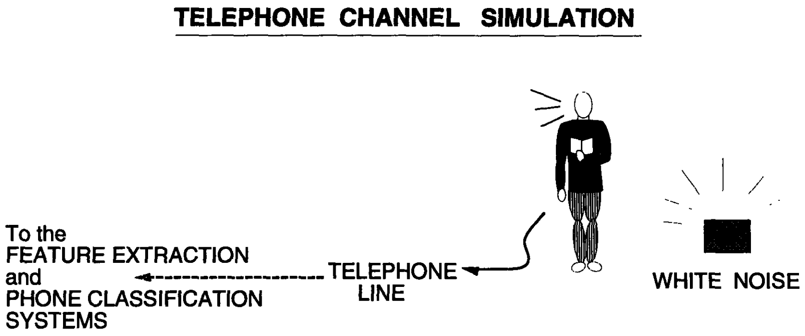


Figure 4-1: Telephone channel simulation

monic disturbance, sinusoidal tones and gaussian noise. The frequency response of the different telephone channels is calculated from actual channel measurements (attenuation observed at different delays along the channel).

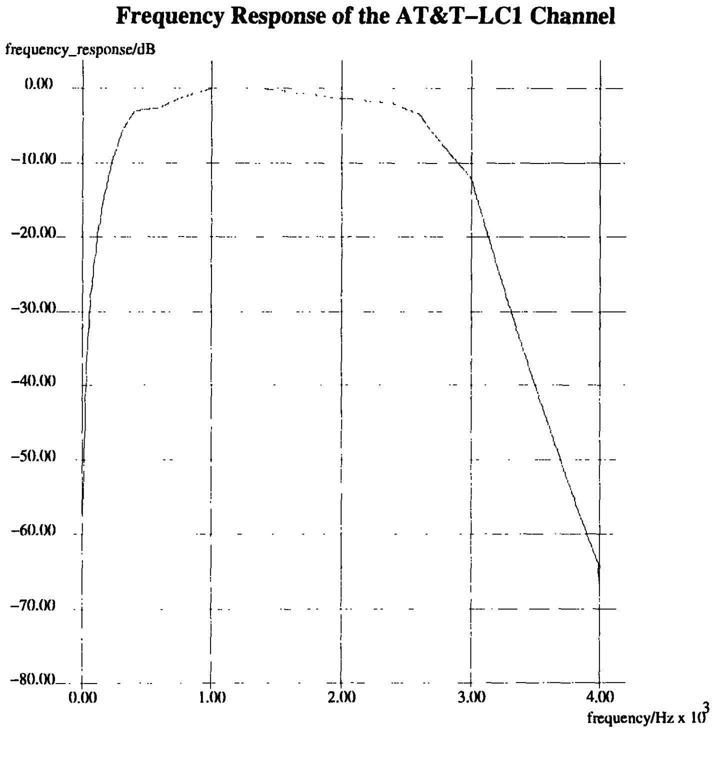


Figure 4-2: Telephone channel frequency response

The AT&T LC1 characteristic channel used has a pass-band of 300 Hz to 2600 Hz. Gaussian noise is added to the test sentence, which is then passed through the telephone channel. A plot of the channel frequency response is attached in Figure 4-2.

4.3.2 Room Reverberation Distortion

The room reverberation simulation is illustrated in Figure 4-3. The microphone shown is not modeled.

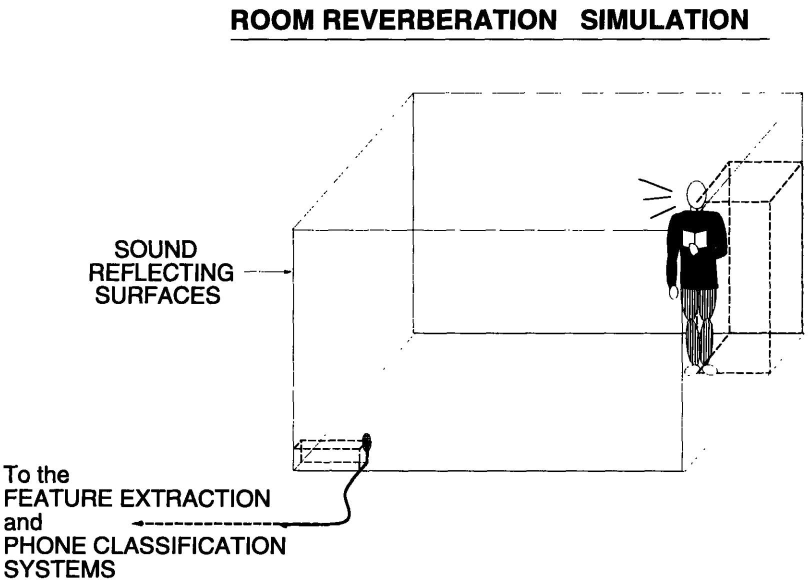
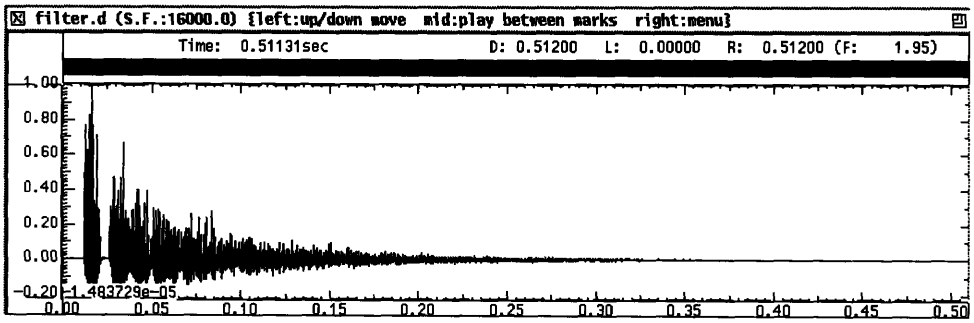

ROOM REVERBERATION SIMULATION
SOUND REFLECTING SURFACES
To the FEATURE EXTRACTION and PHONE CLASSIFICATION SYSTEMS

Figure 4-3: Room reverberation simulation

The reverberation program models the effects of echo and reverberation encountered in an enclosure with sound-reflecting walls. It calculates the source-to-receiver impulse response in a rectangular room, using a time-domain image expansion method [2]. The resulting impulse response, when convolved with a speech signal, simulates room reverberation of the speech.



```
filter.d (S.F.:16000.0) {left:up/down move  mid:play between marks  right:menu}
Time:  0.51131sec          D: 0.51200  L:  0.00000  R:  0.51200 (F:   1.95)
```

Figure 4-4: Room reverberation impulse response

The length, width and height of the room, the reflection coefficients of the six surfaces and the locations of the source and observer are adjusted so as to get a realistic reverberation time [1] between 250 and 550 ms. This is convolved with a test speech utterance (sentence) to get the reverberated speech waveform. The conditions used here are a room 10 feet by

[1] For present purposes, roughly defined as the time it takes the impulse response to fade to 10^{-3} of its maximum value.

11 feet by 12 feet, reflection coefficients equal to .90 for all six surfaces, with the speaker at coordinates (1',1',2') and the microphone at (9',8',11'). A plot of the impulse response is shown in Figure 4-4.

4.4 Summary

This chapter described the experimental framework used to compare the two speech representations, mel cepstrum and EIH - the speech database used, the phonetic classification system, and the distortions used to simulate realistic adverse conditions. The next chapter reports the results obtained from several experiments conducted within this framework.

Chapter 5

Results and Discussion

The Mel Cepstrum (referred to as MEL) and Ensemble Interval Histogram (referred to as EIH) signal representations were evaluated for phone classification on the TIMIT database (male speakers) under different conditions of distortion, in the speech recognition framework provided by a continuous-speech HMM recognition system. These experiments were conducted as a primary means to characterize the performance of the two front ends in representing different sounds under various adverse conditions. Different signal conditions were realized by using clean speech, speech passed through telephone channel simulation and speech passed through room reverberation simulation (the two distortion simulations were described in Chapter 4). The classifier was trained on clean train speech, and tested on the three signal conditions of test speech. Clean train speech was also evaluated to provide a benchmark performance. Context-independent grammar-free phone classification was performed with static features alone, and with static+dynamic features. The influence of automatic resegmentation of the phone boundaries was also tested. Section 5.1 describes the experimental conditions in greater detail.

The average percent accuracy for each signal condition is listed in the tables in Section 5.2 along with comments on the broad trends in these tables. For each percentage in the top 1 tables, a corresponding phonetic confusion table is attached in the Appendix to provide diagnostic information. Some remarks on the on-diagonal and off-diagonal behavior of different phones are presented in Section 5.3.

5.1 Experimental Conditions

Classification experiments were conducted for the following two conditions.

1. **Feature vectors :** Four sets of feature vectors were used : two sets of static features (spectral envelope, and spectral envelope and energy), and two sets of static+dynamic features (spectral envelope and its derivatives, and spectral envelope and energy and

their respective derivatives). The dynamic features were calculated as described in Chapter 3.

Static features were evaluated separately to observe the relative contribution of dynamic information.

For each feature set, appropriate feature vectors were extracted from *clean training speech* and used for training HMM's; the corresponding feature vectors were extracted from test speech in different signal conditions and used for testing.

2. **Phone boundaries:** Two kinds of phone boundaries were used for classification : the given hand segmented TIMIT phone boundaries, and phonetic boundaries obtained by automatic resegmentation of both training and testing data within the HMM framework. The resegmentation experiment was performed only for the full feature vector (39 elements consisting of 12 cepstral coefficients and energy, and their respective first and second time derivatives). The HMM's were trained on clean training speech.

   The effects of *automatic* segmentation were studied because of two reasons : (1) it provides boundaries that are consistent with a well-defined objective measure, and (2) most current ASR systems use speech models based on automatic segmentation.

Likelihood scores for phonetic classification were obtained via Viterbi decoding. No grammar or context information was used. Section 5.2 contains tables of average percent correct classified phones under different conditions.

5.2 Average Results

Section 5.2.1 summarizes the feature vector conditions. It contains tables listing the percentage of phones classified as the top 1 or 3 candidates with the TIMIT segmentation. The static and dynamic features are listed in the columns, where ***Env*** is the cepstral envelope (as represented by cepstral coefficients), ***Ener*** is the frame energy, $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta_2}$ are the first and second order time derivatives respectively. In the remainder of the chapter, the braces [] are used to represent the feature set used. For example, [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener] represents the full static+dynamic feature set (39 parameters).

Section 5.2.2 summarizes the phone boundary conditions. It contains tables listing the percentage of phones classified as the top 1 or 3 candidates with TIMIT hand segmentation and with 1, 2 or 3 iterations of automatic resegmentation. This is done for the full feature vector (39 coefficients). The successive iterations are listed in the columns.

The rows in all the tables represent different *testing conditions.* The first row, ***Tr***, represents the clean training speech. The other three rows represent the performance of the three acoustic conditions of the testing speech : ***Cl*** is clean speech, ***Te*** is the speech passed through the telephone channel simulation and ***Rv*** is the speech passed through the room reverberation simulation.

In all tables, the average percentage is calculated as shown in Equation 5.1,

$$\bar{X} = \sum_{i=1}^{46} (\frac{t_i}{\sum_{i=1}^{46} t_i}) (\frac{c_i}{t_i} 100) \quad = \quad 100 \frac{\sum_{i=1}^{46} c_i}{\sum_{i=1}^{46} t_i} \quad . \tag{5.1}$$

where $\bar{X}$ is the average percent accuracy, c_i is the number of times phone i gets classified correctly out of a total of t_i occurrences of phone i, $\frac{c_i}{t_i} 100$ is the percent correct for each phone, and $\frac{t_i}{\sum_{i=1}^{46} t_i}$ is the a-priori probability of the occurrence of phone i (out of 46 phones listed in Table 4.1).

5.2.1 Static and Dynamic Features

Table 5.1: Correct phone as top 1 candidate, TIMIT phone boundaries

	Static Features				Static+Dynamic Features			
	Env		Env , Ener		Env Δ-Δ_2		Env , Ener Δ-Δ_2	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
Tr	52.1	48.4	55.3	50.4	69.5	61.7	72.9	64.0
Cl	46.3	43.2	49.6	45.3	62.3	55.0	66.2	57.6
Te	10.1	20.8	12.8	22.7	30.0	35.0	37.2	37.0
Rv	9.7	9.7	11.2	11.5	16.7	15.2	18.7	17.3

Table 5.1 shows the percent accuracy when the correct phone is classified as the top candidate, using the original TIMIT hand-segmentation. Looking at the rows, the training speech (***Tr***) yields the highest accuracy for all feature sets for both front ends, followed by clean test speech (***Cl***). The two "noisy" versions of test speech yield lower accuracy than clean speech, and out of them, ***Te*** performs better than ***Rv***. Looking at the columns, the static features augmented by the dynamic features yield higher accuracy than the static features alone, which is expected [8].

Table 5.2: Relative increase in accuracy, in percent, with the addition of features to [Env] ($\frac{Feature-Env}{Env}$.100), correct phone as top 1 candidate, TIMIT phone boundaries

	Addition to Env of					
	Ener		Δ-Δ_2 Env		Ener,Δ-Δ_2 Env Δ-Δ_2 Ener	
	MEL	EIH	MEL	EIH	MEL	EIH
Tr	6.1	4.1	33.4	27.5	39.9	32.2
Cl	7.1	4.9	34.6	27.3	43.0	33.3
Te	26.7	9.1	197.0	68.3	268.3	77.9
Rv	15.5	18.6	72.2	56.7	92.8	78.4

Table 5.2 lists increase in accuracy, in percent, relative to the accuracy with spectral envelope (calculated as $\frac{Feature-Env}{Env}.100$), obtained across different feature vector conditions. Adding energy to the spectral envelope results in small improvements overall. Addition of dynamic features results in large improvements, especially for MEL in the case of ***Te***. EIH shows smaller percent increases than MEL in most cases. ***Tr*** generally yields the lowest improvements with the addition of parameters.

Table 5.3 lists the relative differences in performance between MEL and EIH, calculated relative to the average of MEL and EIH as $\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100$ from Table 5.1.

Table 5.3: Relative differences, in percent, between MEL and EIH, ($\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100$), correct phone as top 1 candidate, TIMIT phone boundaries

	Static Features		Static+Dynamic Features	
	Env	Env , Ener	Env Δ-Δ_2	Env , Ener Δ-Δ_2
Tr	7.4	9.3	11.9	13.0
Cl	6.9	9.1	12.4	13.9
Te	-69.3	-55.8	-15.4	0.5
Rv	0.0	-2.6	9.4	7.8

Five remarks can be made about the differences in performance between MEL and EIH.

1. MEL yields higher accuracy than EIH for clean train and test speech. This is in accordance with results seen in [16, 9].

2. EIH outperforms MEL for ***Te***, test speech passed through telephone channel simulation, for the first three feature vector types.

3. For ***Rv*** in Table 5.3, the numbers seem to be haphazard ; the percent accuracy from Table 5.1 is a very low number, suggesting mostly chance "hits".

4. Adding dynamic information improves performance for both representations, as shown in Table 5.2.

5. Adding dynamic information to MEL is more effective than it is for EIH, as Table 5.2 shows.

The increase in accuracy for clean speech exhibited by both speech representations with the addition of dynamic features can be attributed to the incorporation of information about coarticulation and changing spectral structures in the dynamic features. For ***Te***, which is essentially bandpass filtered speech (the simulation also adds white noise, which is assumed to be negligible in the following discussion), the effect of dynamic information can be explained in the frequency domain as follows.

$$Y(\omega) = H_{tc}(\omega) \cdot X(\omega) \tag{5.2}$$

where ω is the DFT frequency, $X(\omega)$ is the Fourier transform of the original speech signal, $H_{tc}(\omega)$ is the frequency response of the telephone channel (assumed to be linear) and $Y(\omega)$ is the Fourier transform of the distorted signal. In the cepstral domain, where the (real) cepstrum is defined as the logarithm of the magnitude of the Fourier transform of the speech signal, this relation becomes

$$y = h_{tc} + x \tag{5.3}$$

where x is the cepstrum of the original speech signal, h_{tc} is the cepstrum of the channel frequency response and y is the cepstrum of the distorted speech signal.

Delta cepstrum of the signal computed across times t1 and t2 (this is for illustration only, the actual delta cepstrum is calculated as described at the end of Chapter 3) is

$$\Delta y = y1 - y2 = h_{tc1} + x1 - (h_{tc2} + x2) = x1 - x2 \tag{5.4}$$

assuming a constant channel response, i.e., $h_{tc1} = h_{tc2} =$ constant.

One reason why dynamic features have a deleterious effect on the EIH representation could be the frequency-dependent time window used in EIH analysis. As explained in Chapter 3, an EIH feature vector is computed by averaging three consecutive frames of "cepstral" coefficients. Each of these "cepstral" vectors corresponds to the histogram computed at a rate of one every 3.2 ms. Each histogram is accumulated from level crossings made by speech windowed with a back-in-time window. The width of this window is not constant over frequency; it depends on CF, the characteristic frequency of the mel-like bandpass filter corresponding to the level-crossing detector contributing to the histogram. The width is greater at the lower frequencies than at the higher frequencies, as shown in Figure 5-1.

At time t_0, intervals produced by a level-crossing detector located at characteristic frequency CF_0 are collected over a window of length $\frac{10}{CF_0}$ that ends at time t_0. This is done once every 3.2 ms. Therefore, for low frequencies like 100 Hz, the past time considered is 100 ms, for middle frequencies like 1 kHz, the time considered is 10 ms, and for high frequencies like 4 kHz, the past time considered is 2.5 ms. Also for the (averaged) EIH feature vectors at the 9.6 ms rate, this corresponds to past times equal to 100 ms, 10 ms and 2.5 ms respectively. Delta cepstrum is calculated over 5 frames, centered at the current frame. For MEL, this corresponds to 20 ms of speech before the current time t_0 and 20 ms of speech after it, since MEL is calculated at the rate of a frame every 10 ms, with a frequency-independent time window of 20 ms. For EIH, however, this method of derivative calculation involves different time segments of speech for different frequencies which causes a mismatch across different cepstral coefficients. An appropriate method of calculating delta coefficients must be devised for EIH. The experiments in this work were performed with the available method as a start.

Besides the correctly classified cases reported in the previous discussion, statistics were collected for phones classified in the top 3 candidates. These serve to give an estimate of

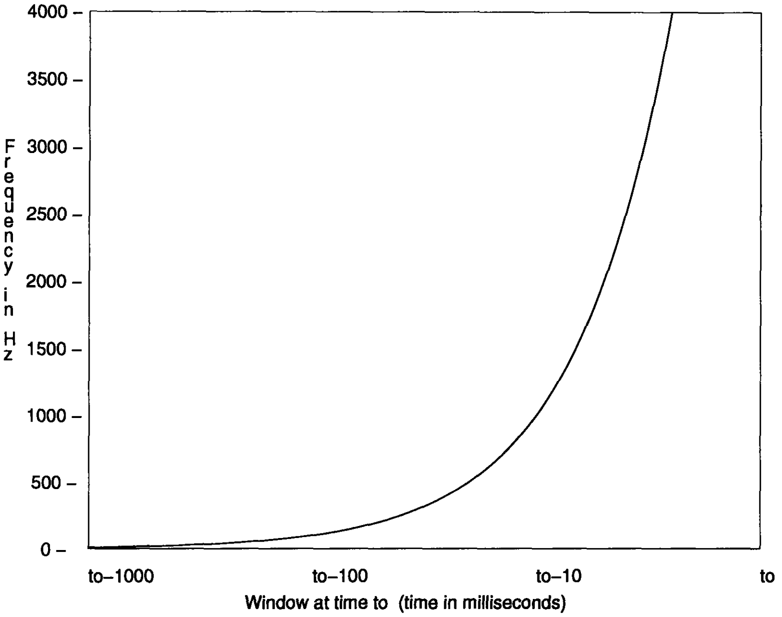


Figure 5-1: Variation of EIH time window with frequency

Table 5.4: Correct phone in top 3 candidates, TIMIT phone boundaries

	Static Features				Static+Dynamic Features			
	Env		Env , Ener		Env Δ-Δ_2		Env , Ener Δ-Δ_2	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
Tr	79.6	75.1	82.3	77.2	90.9	85.9	92.9	87.7
Cl	75.7	71.6	78.7	74.0	87.9	82.2	90.4	84.4
Te	30.2	44.6	34.6	47.5	56.7	59.7	64.4	62.3
Rv	23.1	24.6	27.8	27.8	31.9	34.2	39.5	37.1

Table 5.5: Relative differences, in percent, between MEL and EIH, $(\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100)$, correct phone in top 3 candidates, TIMIT phone boundaries

	Static Features		Static+Dynamic Features	
	Env	Env , Ener	Env Δ-Δ_2	Env , Ener Δ-Δ_2
Tr	5.8	6.4	5.7	5.8
Cl	5.6	6.2	6.7	6.9
Te	-38.5	-31.4	-5.2	3.3
Rv	-6.3	0.0	-7.0	6.3

the performance expected with grammar constraints on the classification process. Table 5.4 shows fairly high accuracies in the case of static+dynamic features, high 80's for the clean test speech and 60's for telephone channel simulation speech. These numbers are still too low to be considered for practical applications, especially since these are results for classification, not recognition.

Table 5.5 shows the relative differences in performance between MEL and EIH, calculated relative to the average of MEL and EIH as $\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100$ from Table 5.4. Comparing it to Table 5.3, it is seen that for ***Tr*** and ***Cl***, the relative differences between MEL and EIH across different feature vectors have a smaller range for top 3 candidates than for top 1 candidate. Also for ***Te***, the differences are in general smaller in magnitude for top 3 candidates. For ***Rv***, the performance trends of top 1 and top 3 candidates are different.

Section 5.2.2, average results obtained with different iterations of automatic resegmentation of the training and testing data are displayed.

5.2.2 Automatic Resegmentation

This part of the study was done to examine the effects of automatic resegmentation as compared to hand segmentation. *Automatic* resegmentation is of interest because of two reasons : (1) it provides subword boundaries that are consistent with a well-defined objective measure and can be changed automatically to achieve some sort of *optimization* within the given framework, and (2) most current ASR systems use speech models based on automatic segmentation, one of the goals of this study is to compare the two front ends on a state-of-the-art system. The second reason also explains the use of static+dynamic features (39) in the following experiments. It should be kept in mind, however, that the results reported here were obtained with dynamic features that are not necessarily suited to EIH, as discussed in Section 5.2.1.

Table 5.6 lists results obtained from iterations of resegmentation of training and testing data, with the HMM model trained anew at each iteration, as explained in Chapter 4.

Table 5.6: Correct phone as top 1 candidate: Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]

	Automatic resegmentation: iteration number							
	None		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
Tr	72.9	64.0	72.5	65.7	78.3	73.4	79.2	74.2
Cl	66.2	57.6	67.6	60.9	72.0	67.3	72.8	67.9
Te	37.2	37.0	36.1	39.6	42.1	44.5	44.8	45.4
Rv	18.7	17.3	44.4	38.9	50.3	43.4	50.7	44.3

Overall, accuracy improves with an increasing number of iterations, which is expected because the process of resegmentation and retraining yields a better fit between the HMM and the given training speech each time. Convergence is quickly reached; little improvement can be observed in Table 5.6 going from the second to the third resegmentation. The last column corresponding to the third resegmentation contains the highest results obtained with the given classification system for top 1 candidates for MEL and EIH. EIH outperforms MEL for ***Te***, and the difference in performance is statistically significant for a significance level of 0.001 with the McNemar test, as explained in greater detail in Section 5.2.3.

Table 5.7 shows the increase in accuracy, in percent, relative to the accuracy with TIMIT segmentation (calculated as $\frac{Iteration-TIMIT}{TIMIT}.100$), obtained with different resegmentation iterations. The most dramatic change in performance going from TIMIT segmentation to one resegmentation is observed in ***Rv*** for both MEL and EIH, an increase in accuracy of 137.4% and 124.9% respectively. The smallest percent increase is seen in ***Tr*** for most iterations. EIH shows a greater percent increase than MEL for ***Tr***, ***Cl*** and ***Te***.

Table 5.7: Relative increase in accuracy, in percent, with successive iterations of automatic resegmentation ($\frac{Iteration-TIMIT}{TIMIT}.100$), correct phone as top 1 candidate, Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]

	Error rate reduction : iteration number					
	One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH
Tr	-0.5	2.7	7.4	14.7	8.6	15.9
Cl	2.1	5.7	8.8	16.8	10.0	17.9
Te	-3.0	7.0	13.2	20.3	20.4	22.7
Rv	137.4	124.9	169.0	150.9	171.1	156.1

One possible reason for the marked effect of resegmentation in the case of ***Rv*** is the "echo effect" of reverberation. From Figure 4-4 which shows the reverberation impulse response used, an estimate of the "echo time" can be read off to be 20 to 40 ms. The spectrograms in Figure A-2 show a "smearing" of 40 to 60 ms in reverberated speech, beyond the boundaries for clean speech. This time range corresponds to 4 to 6 frames for MEL and EIH. TIMIT phone boundaries given for clean test speech roughly shift forward by a few frames when the speech undergoes reverberation. The first automatic resegmentation of the reverberated speech takes into account the shifted boundaries and that is why classification accuracy improves with the newly resegmented boundaries. Besides the phone boundaries, it also changes the state boundaries within phones which were uniform for experiments with the initial (TIMIT) segmentation.

The first iteration alleviates the predominant mismatch between the phone boundaries of clean test speech and the reverberated test speech, after which the improvement becomes gradual. It is worthwhile to recall that the training data is resegmented independently of the classification performance of the test data, and the "improved" phonetic boundaries of the reverberation speech (and other test sets) do not affect subsequent model building.

Sample TIMIT segmentation and one iteration of resegmentation are shown in the Appendix in Tables A.1 and A.2, for MEL and EIH representations of the test speech sentence in different signal conditions. The waveforms and spectrograms of the sentence are shown in Figures A-1 and A-2. The tables list each phone, its starting frame number, and the number of frames in the phone. Major changes with resegmentation of the utterance are the insertion of silence at different places for the two front ends, and for ***Rv***, a forward shift of an average of 3.4 frames for MEL and 4.9 frames for EIH.

Table 5.8 lists the relative differences in performance between MEL and EIH, calculated relative to the average of MEL and EIH as $\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100$ from Table 5.6. The magnitude of the difference roughly decreases with increasing iterations for ***Tr*** and ***Cl***, the least difference for both is at the second iteration. For ***Te***, the difference shows a large decrease after one iteration, after which its starts increasing. For ***Rv***, the difference shows a large increase after one iteration, and increases upto the second iteration.

Table 5.8: Relative differences, in percent, between MEL and EIH, ($\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100$), correct phone as top 1 candidate, Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]

	Automatic resegmentation: iteration number			
	None	One	Two	Three
Tr	13.0	9.8	6.5	6.5
Cl	13.9	10.4	6.7	7.0
Te	0.5	-9.2	-5.5	-1.3
Rv	7.8	13.2	14.7	13.5

Table 5.9 shows the cases where the correct phone is in the top 3 candidates. The last column projects the best performance of MEL and EIH in the HMM classification system with the given training and testing data and training parameters, assuming some form of lexical and semantic analysis on the output of the classifier. The accuracies for ***Te*** and ***Rv*** are in the 70's and 80's; these numbers give promise for further research.

Table 5.9: Correct phone in top 3 candidates, Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]

	Automatic resegmentation: iteration number							
	None		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
Tr	92.9	87.7	92.3	88.4	95.1	92.8	95.6	93.5
Cl	90.4	84.4	90.7	86.5	93.2	90.6	93.7	91.4
Te	64.4	62.3	62.3	65.8	66.8	69.4	69.5	70.6
Rv	39.5	37.1	73.6	67.6	80.1	72.6	80.2	73.7

Table 5.10 lists the relative differences in performance between MEL and EIH, calculated

relative to the average of MEL and EIH as $\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100$ from Table 5.9. Comparing it to Table 5.8, it is seen that for ***Tr***, ***Cl*** and ***Rv***, the relative differences between MEL and EIH across different resegmentations are smaller for top 3 candidates than for top 1 candidate. For ***Te***, the differences are in smaller in magnitude for top 3 candidates, for the first two resegmentations.

Table 5.10: Relative differences, in percent, between MEL and EIH, $(\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100)$, correct phone in top 3 candidates, Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]

	Automatic resegmentation: iteration number			
	None	One	Two	Three
Tr	5.8	4.3	2.4	2.2
Cl	6.9	4.7	2.8	2.5
Te	3.3	-5.5	-3.8	-1.6
Rv	6.3	8.5	9.8	8.4

5.2.3 Statistical Significance

For the top 1 average percent correct results discussed in Section 5.2, the McNemar significance test [11] was conducted. The McNemar test takes into account the speech segments (tokens) correctly classified by one front end and incorrectly classified by the other front end, to determine whether the difference in performance between two front ends tested on the same data is statistically significant.

The following method is described in detail in [11], it is briefly outlined here. Let N_{10} be the number of tokens classified correctly by MEL and classified incorrectly by EIH, and N_{01} be the number of tokens classified incorrectly by MEL and classified correctly by EIH. Let $k = N_{10} + N_{01}$. If k is large enough, ($k > 50$), and N_{10} is not too close to k or 0, let

$$W = \frac{\mid N_{10} - \frac{k}{2} \mid - \frac{1}{2}}{\sqrt{\frac{k}{4}}} \tag{5.5}$$

P, the probability that the two front ends have equal error-rates, can be calculated as

$$P = 2Pr(Z \geq w) \tag{5.6}$$

where Z is a Gaussian random variable $\aleph(0,1)$ and w is the realized value of W. For a significance level of α, values of P smaller than α indicate that the performance of the two front ends is significantly different, while values of P larger than α indicate that the performance of the two front ends is very similar.

At a significance level of 0.001, all the results in Tables 5.1 and 5.6 are statistically significant, except for ***Te*** with [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener], and ***Rv*** with the first three feature sets with TIMIT segmentation.

5.3 Confusion Matrices

In Section 5.2, the average results yielded information on the overall performance of the two representations for clean and "noisy" signal conditions. In Appendix B, more detailed analysis is attempted by using confusion matrices. A confusion matrix shows how often a given phone is confused with any of the other phones. The $a[i, j]$ element of each $N\ X\ N$ confusion matrix A is the number of times the phone P_i gets classified as phone P_j, expressed as a percentage of the total occurrences of phone P_i. A diagonal element, $a[i, i]$, represents percent accuracy of P_i, and an off-diagonal element, $a[i, j]$, $i \neq j$, represents the percentage of times P_i gets mis-classified as P_j. Elements along a row add up to 100%. An additional $(N + 1)$ column lists the number of occurrences of each phone P_i.

Table 5.11: Grouping of 47 phones into 18 groups used in the confusion matrices

Symbol	Group	Phones
FH	Front,High vowels	***iy*** be*e*t , ***ih*** b*i*t
FM	Front,Mid vowels	***eh*** b*e*t
FL	Front,Low vowels	***ae*** b*a*t
CM	Central,Mid vowels	***ah*** b*u*d
CH	Central,High vowels	***er*** b*ir*d
BH	Back,High vowels	***uw*** b*oo*t , ***uh*** b*oo*k
BM	Back,Mid vowels	***ao*** b*au*d , ***ow*** b*oa*t
BL	Back,Low vowels	***aa*** f*a*ther
Dp	Diphthongs	***ey*** w*ai*t , ***ay*** b*i*te ***aw*** b*ou*t , ***oy*** b*oy*
Lq	Liquids	***r*** *r*ed , ***axr*** butt*er* ***l*** *l*ike , ***el*** bott*le*
Gl	Glides	***y*** *y*es , ***w*** *w*ell
Ns	Nasals	***m*** *m*ob , ***em*** bott*om* ***n*** *n*od , ***en*** butt*on* ***nx*** wi*nn*er , ***ng*** si*ng*
FV	Fricatives,Voiced	***z*** *z*oo , ***zh*** mea*s*ure ***v*** *v*ery , ***dh*** *th*en
FU	Fricatives,Unvoiced	***s*** *s*ea , ***sh*** *sh*e ***f*** *f*in , ***th*** *th*in
SV	Stops,Voiced	***b*** *b*ee , ***d*** *d*ay ***g*** *g*ame , ***dx*** mu*dd*y
SU	Stops,Unvoiced	***p*** *p*ea , ***t*** *t*ea , ***k*** *k*ey
Af	Affricates	***ch*** *ch*ild , ***jh*** *j*oke
Wh	Whisper sound	***h*** *h*ot

Confusion matrices were extracted for the 46 phones (leaving out silence) listed in Table 4.1 and collapsed into confusions between 18 acoustically similar subgroups formed as shown in Table 5.11. The acoustic groups were created according to Chapter 2 in [30], especially the vowel groups.

The confusion matrices are organized according to the signal conditions of testing speech. Appendix D contains $\boldsymbol{Tr}$, clean training speech. Appendix E contains $\boldsymbol{Cl}$, clean test speech. Appendix F contains $\boldsymbol{Te}$, test speech passed through the telephone channel simulation. Appendix G contains $\boldsymbol{Rv}$, test speech passed through the room reverberation simulation. In each appendix, confusion matrices for the four feature sets are listed first, in order as [Env], [Env, Ener], [Env, Δ-Δ_2 Env], and [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener], followed by confusion matrices for one, two and three iterations of automatic resegmentation. For each condition, MEL and EIH are listed on the same page to make comparison easier.

A qualitative analysis of the confusion matrices involving observation of accuracies of distinct phone groups, of major confusions among groups, and of the confusion patterns in noisy conditions is provided in Appendix B. Some inferences drawn therein highlight the usefulness of confusion matrices for diagnostic purposes. For example, while the low values of average percent accuracies for $\boldsymbol{Rv}$ for both speech representations have no meaning for practical purposes, the confusion matrices corresponding to these averages contain information that differentiates the two speech representations. Quantitative methods can be used for rigorous analysis of the confusion matrices, but that is the subject of future research.

5.4 Summary

The results of the evaluation of the two speech representations, MEL and EIH, on a phonetic level using average percent classification accuracies were discussed in detail. Results of the McNemar significance test were reported. A qualitative analysis of confusion matrices for phone groups was provided (in the Appendix). Some previous results were confirmed, and interesting trends emerging in this study were noted, as summarized in the next chapter.

Chapter 6

Summary and Future Work

6.1 Summary of Work

The goal of this study was to compare the performance of two speech representations, the traditional Fourier-based mel cepstrum and the auditory-based Ensemble Interval Histogram (EIH), for large vocabulary, speaker independent continuous speech recognition using a state-of-the-art speech recognition system.

Phonetic classification was performed instead of recognition to focus on the speech representation, eliminating other issues like grammar, phone insertion and phone deletion involved in the recognition process.

Similar first-stage filters and frame-rates were used for the two representations, and both were transformed into the standard cepstral coefficients (12 for each). Dynamic features were calculated in identical fashion for the two representations.

Static features only, and static and dynamic features were evaluated separately to observe the relative contribution of dynamic information on the two representations under different signal conditions.

The TIMIT database was used because it is a standard, phonetically rich database, available with phonetic transcriptions.

Simulations of acoustic distortions found in realistic situations - through a telephone channel and under room reverberations - were used.

TIMIT segmentation as well as automatic resegmentations within the HMM classification framework were used.

Along with the average performance, confusion matrices were extracted for phone groups to observe the behavior of individual phone classes.

6.2 Conclusions

To summarize, the comparative study of the two speech representations, MEL and EIH, yielded the following results :

- In general, addition of dynamic parameters to the feature vector results in an increase in performance, and automatic resegmentation also results in an increase in performance.

- MEL outperforms EIH in clean continuous speech, as it does for isolated speech reported in [9, 16]. The difference is small with static features alone, and increases with the addition of dynamic features. The smaller contribution of dynamic features for EIH as compared to MEL is a trend found in all acoustic conditions. One explanation for it is that the method of computation of cepstral time derivatives is inappropriate for EIH. Delta cepstrum is calculated over five frames centered at the current frame, thus accounting for 20 ms of speech behind and 20 ms of speech ahead of the current time, at a frame rate of 10 ms. Delta-delta cepstrum is also calculated over time frames taken to be uniform over all cepstral coefficients. For EIH, however, the time-window is frequency dependent and it varies inversely with frequency. Determining the set of dynamic parameters appropriate to EIH is beyond the scope of this study. Here, dynamic features for EIH were computed using the same temporal filters as those used for MEL.

- EIH outperforms MEL for the speech passed through the telephone channel simulation. This is in agreement with [16] where the auditory models including EIH performed better than MEL under spectral distortion (for conditions with higher baseline error rates). Here the difference is the greatest for static features (about 10% for top 1 candidate and 14% for top 3 candidates). The magnitude of this difference decreases with the inclusion of dynamic features, possibly for reasons discussed earlier.

- On clean speech, for both front ends, the frequency with which voiced fricatives are confused as unvoiced fricatives is higher than the frequency with which unvoiced fricatives are confused as voiced fricatives. Also, the frequency with which voiced stops are confused as unvoiced stops is higher than the frequency with which unvoiced stops are confused as voiced stops.

- Under the telephone channel distortion, the sounds most affected are voiced and unvoiced fricatives for MEL, voiced and unvoiced stops for EIH, and affricates for both. With static features only, for MEL, most sounds are mis-classified as voiced stops and nasals. For EIH with static features, most sounds are mis-classified as nasals and liquids.

- Both front ends perform poorly for speech passed through the room reverberation simulation, with the TIMIT segmentation. The numbers are very low, suggesting mostly chance "hits". The performance of both front ends improves markedly with automatic resegmentation of the test and train data.

- Under the room reverberation distortion, the sounds most affected for MEL are most of the vowels; for EIH they are some of the vowels, the voiced stops and voiced fricatives. For all feature sets for both MEL and EIH, many sounds are mis-classified very frequently as the whisper sound (*h* as in *h*elp).
- From some examples of clean speech studied in detail, the addition of dynamic information to the feature vector improves performance for sounds with slowly varying formant structures, such as diphthongs, but not for sounds containing abrupt changes in their spectral configuration, such as stops and affricates.

Previous studies suggested that EIH performs worse than MEL in clean speech, but is more robust in adverse conditions. These studies were conducted on a limited task, i.e., speaker dependent isolated words (small vocabulary) speech recognition. Our study extends these observations to the task of speaker (male) independent, continuous speech recognition. The most notable outcomes of our study are (1) the representation of spectral envelope by EIH is more robust to noise - previous evidence of this fact is now extended to the case of speaker independent, continuous speech, (2) adding dynamic features (represented by delta and delta-delta cepstrum) substantially increases the performance of MEL in all signal conditions that were tested. Adding delta and delta-delta cepstrum of EIH cepstrum - computed by using the same temporal filters as those used for MEL - results in much smaller improvement. We suggest that in order to improve recognition performance with an EIH front end, appropriate integration of dynamic features must be devised.

6.3 Future Work

As shown by the results of this work, the next step is to determine a suitable method of incorporating dynamic information into the EIH representation.

A more objective and better defined analysis of confusion matrices is needed for quantifying the information contained therein.

Classification tests with matched acoustic conditions of training and testing speech, under distortion, can be performed to observe phonetic confusions for the two speech representations.

Two more general avenues of research are (1) to use a different distance measure in the Viterbi search in the HMM recognition framework, and (2) to try context-dependent phone modeling.

Appendix A

Waveforms, spectrograms and phonemic boundaries

This Appendix is included for illustration purposes. Waveforms, spectrograms and phonemic boundaries for the utterance "Y'all should have let me do it" under different signal conditions are attached. In the main thesis, these are referred to in Sections 4.3 and 5.2.2. A frame-energy plot for MEL and EIH is also attached; it is referred to in Section 3.3.

Figure A-1 shows the waveforms of a test utterance in three signal conditions: clean, under telephone channel simulation, and under room reverberation simulation. The abscissa in all cases represents time in seconds, the ordinate represents the signal value.

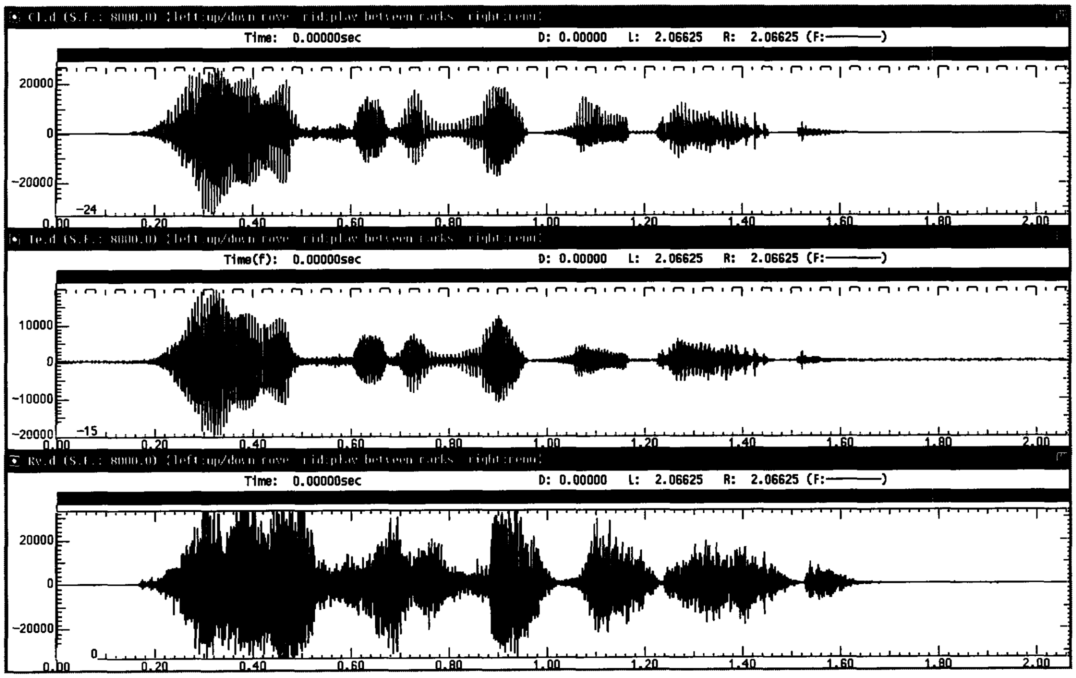
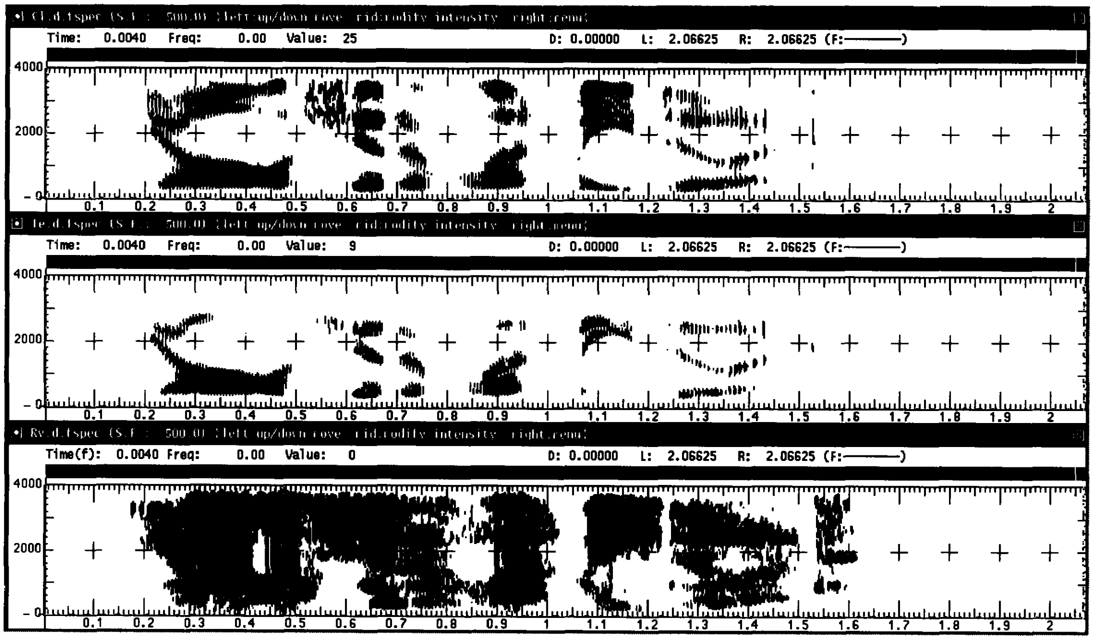

Cl.d (S.F.: 8000.0) {left:up/down move mid:play between marks right:menu}
Time: 0.00000sec
D: 0.00000 L: 2.06625 R: 2.06625 (F:———)
20000
0
-20000
-24
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
te.d (S.F.: 8000.0) {left:up/down move mid:play between marks right:menu}
Time(F): 0.00000sec
D: 0.00000 L: 2.06625 R: 2.06625 (F:———)
10000
0
-10000
-20000
-15
Rv.d (S.F.: 8000.0) {left:up/down move mid:play between marks right:menu}
Time: 0.00000sec
D: 0.00000 L: 2.06625 R: 2.06625 (F:———)
20000
0
-20000
0

Figure A-1: Waveforms of clean, telephone-channel and room-reverberation versions of the sentence " Y'all should have let me do it."

Figure A-2 shows the spectrograms of a test utterance in three signal conditions: clean, under telephone channel simulation, and under room reverberation simulation. The abscissa in all cases represents time in seconds, the ordinate represents frequency in hertz.



Time: 0.0040 Freq: 0.00 Value: 25 D: 0.00000 L: 2.06625 R: 2.06625 (F:———)

Time: 0.0040 Freq: 0.00 Value: 9 D: 0.00000 L: 2.06625 R: 2.06625 (F:———)

Time(f): 0.0040 Freq: 0.00 Value: 0 D: 0.00000 L: 2.06625 R: 2.06625 (F:———)

Figure A-2: Spectrograms of clean, telephone-channel and room-reverberation versions of the sentence " Y'all should have let me do it."

Table A.1 shows sample TIMIT segmentation and one iteration of automatic resegmentation for the utterance represented by MEL.

Table A.1: Phone boundaries for the sentence "Y'all should have let me do it" under different signal conditions, represented by MEL

	TIMIT segmentation		1 Automatic resegmentation for :					
			Cl		Te		Rv	
Unit	First Frame	No. of Frames	First Frame	No. of Frames	First Frame	No. of Frames	First Frame	No. of Frames
silence	1	14	1	13	1	14	1	15
y	15	9	14	11	15	10	16	12
ao	24	18	25	18	25	20	28	6
silence							34	5
l	42	9	43	7	45	5	39	16
sh	51	9	50	11	50	9	55	7
uh	60	7	61	6	59	8	62	8
dx	67	3	67	4	67	4	70	5
ax	70	6	71	5	71	5	75	5
v	76	7	76	7	76	3	80	5
l	83	5	83	4	79	10	85	4
eh	88	8	87	8	89	5	89	11
t	96	5	95	5	94	6	100	3
silence					100	1		
m	101	5	100	6	101	4	103	5
iy	106	11	106	11	105	13	108	16
d	117	8	117	7	118	5	124	5
uw	125	9	124	13	123	11	129	11
ah	134	12	137	8	134	9	140	8
silence							148	1
t	146	13	145	12	143	14	149	14
silence	159	47	157	50	157	50	163	44

Table A.2 shows sample TIMIT segmentation and one iteration of automatic resegmentation for the utterance represented by EIH.

Table A.2: Phone boundaries for the sentence "Y'all should have let me do it" under different signal conditions, represented by EIH

	TIMIT segmentation		1 Automatic resegmentation for :					
			Cl		Te		Rv	
Unit	First Frame	No. of Frames	First Frame	No. of Frames	First Frame	No. of Frames	First Frame	No. of Frames
silence	1	14	1	13	1	17	1	15
y	15	10	14	12	18	8	16	12
ao	25	19	26	20	26	22	28	26
l	44	9	46	6	48	4	54	4
sh	53	9	52	12	52	11	58	9
uh	62	8	64	6	63	7	67	7
dx	70	3	70	4	70	4	74	3
ax	73	6	74	4	74	4	77	5
v	79	7	78	8	78	7	82	3
l	86	6	86	8	85	9	85	9
eh	92	8	94	5	94	6	94	11
silence							105	1
t	100	6	99	7	100	4	106	3
m	106	4	106	5	104	7	109	4
iy	110	12	111	11	111	12	113	14
silence							127	2
d	122	8	122	9	123	3	129	3
uw	130	9	131	14	126	19	132	22
ah	139	13	145	7	145	8	154	3
t	152	13	152	13	153	10	157	8
silence	165	49	165	52	163	54	165	52

A plot of the frame energy of an utterance is shown in Figure A-3 for MEL and for EIH.

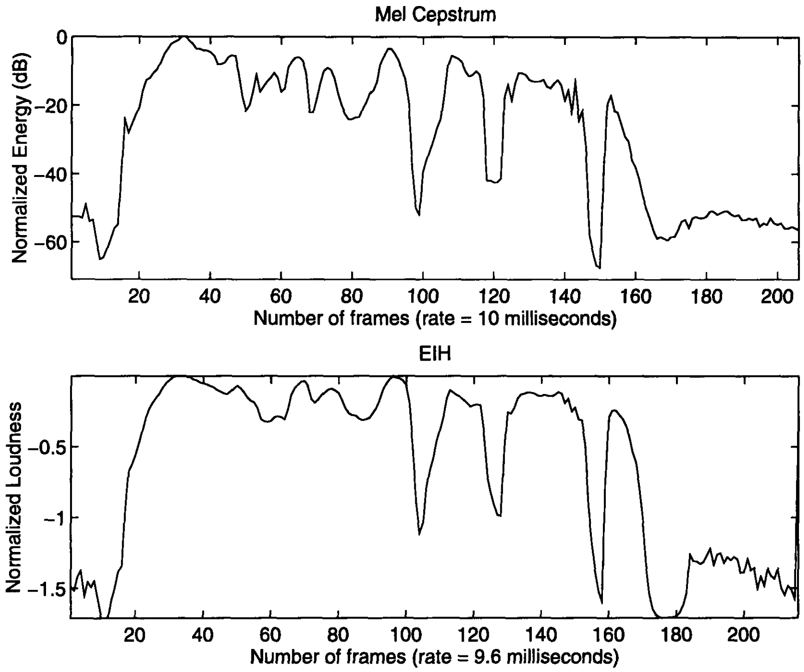


Figure A-3: Frame energy for the sentence "Y'all should have let me do it" (clean), represented by MEL and by EIH

Appendix B

Observations from confusion matrices

As described in Section 5.3, the confusion matrices are organized according to the signal conditions of testing speech, ***Tr*** (clean training speech), ***Cl*** (clean test speech), ***Te*** (test speech passed through the telephone channel simulation) and ***Rv*** (test speech passed through the room reverberation simulation), in Appendices D, E, F and G respectively. Here, some observations drawn from these confusion matrices are listed.

Section B.1 points out some observations on the on-diagonal behavior of the confusion matrices. Changes in performance of both speech representations with different feature vectors or resegmentation iterations are observed for particular sounds.

Section B.2 points out some observations on the off-diagonal behavior of MEL and EIH. Frequent confusions for particular classes and changes in the "distribution" of confusion matrices with different experimental conditions are observed.

B.1 On-diagonal trends of confusion matrices

The *on-diagonal elements* of the confusion matrices are listed in Tables C.1, C.2, C.3 and C.4, to compare the performance of the two front ends for different acoustic groups. In all the tables, the four feature sets with TIMIT segmentation are listed first, in order as [Env], [Env, Ener], [Env, Δ-Δ_2 Env], and [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener], followed by one, two and three iterations of automatic resegmentation with the full feature set [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]. The percent accuracies for phone groups are rounded to integers, and the larger number out of MEL and EIH is printed boldface. These numbers are a breakdown of the average classification accuracies in Tables 5.1 and 5.6 into accuracies for each of the 18 acoustic groups [1].

Two kinds of statistics were extracted from the on-diagonal accuracies of MEL and EIH for 18 phone groups, *relative differences* in performance between MEL and EIH for all conditions, and *relative performance increases* with the addition of features or with iterations of resegmentation for MEL and for EIH.

The *relative differences* in performance between MEL and EIH for 18 acoustic groups are displayed in Tables C.5, C.6, C.7 and C.8. These are calculated relative to the average of MEL and EIH as $\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100$.The negative differences are printed boldface; they represent cases where EIH outperforms MEL. These tables correspond to average performance Tables 5.3 and 5.8.

The *relative increases in accuracy* for MEL and for EIH for 18 phone groups are displayed in Tables C.9, C.10, C.11 and C.12. In each table, the first three columns contain accuracy increases for three feature sets relative to the accuracy with spectral envelope, calculated as $\frac{Feature-Env}{Env}.100$, and the last three columns contain accuracy increases for 3 resegmentation iterations relative to the accuracy with TIMIT segmentation, calculated as $\frac{Iteration-TIMIT}{TIMIT}.100$. The larger number out of MEL and EIH is printed boldface. These correspond to Tables 5.2 and 5.7.

Some observations from these three sets of tables are listed below for ***Cl***, ***Te*** and ***Rv*** (***Tr*** is not discussed).

1. ***Cl***, clean test speech

   - Relative differences between MEL and EIH (Table C.6)

     EIH scores higher than MEL with at least one of the *static features* for the vowels CH, CM, FH, BM and BL (in descending order). EIH scores close (within 5% relative difference) to MEL for FL, Lq and FU for both static features. It performs much worse (more than 10% relative difference) than MEL with either of the static features for Wh, Ns, Gl, FM, BH, Af, Dp, FV and SV.

     With both sets of *static+dynamic* features, EIH scores higher than MEL for the vowel CH. It scores close (within 5%) to MEL with either feature set for the vowels BL, Dp, FL, CM, and FH. It performs much worse (more than 10%) than MEL with either set for Wh, FM, SV, BH, Gl, FU, SU, Ns, Lq and BM.

     At the second and third iterations of *resegmentation*, EIH outperforms MEL for the vowels CH, FL, FM and CM. In the third resegmentation, EIH scores close (within 5%) to MEL for Dp, Lq, BM, FH and Gl, and it scores lower than 10% (relative difference) below MEL for SV, Wh and BL.

     CH is classified better by EIH for all conditions.

   - Relative increases in accuracy for MEL and for EIH (Table C.10)

[1] The percent accuracy for each phone group is computed by weighing the accuracy of each constituent phone by its relative occurrence, as explained for the average accuracies in Section 5.2.

With the addition of Ener to Env, the sounds that show large improvements are CM, Wh, Gl, FV and SU for MEL and Wh, Af, CH and CM for EIH (in descending order of relative increase). The smallest increases are shown by BL, FM, Dp, CH and FL for MEL, and BH, FM, FL, Dp and FU for EIH (in ascending order of relative increase).

With the addition of Ener and all the derivatives, the sounds with the most improvements are CM, BM, FH and FM for MEL, and Wh, CM, BL and Gl for EIH. Those with the smallest improvements are Af, Ns, FU, SU, Dp and FL for MEL, and FU, SU, SV, Lq, BH and Af for EIH.

In the third resegmentation, the highest improvements relative to TIMIT segmentation are shown by CM, BL, BH and SV for MEL, and CM, FM, BH and Wh for EIH. The lowest increases are shown by Af, Ns, Dp and CH for MEL, and Af, Dp, BL and Ns for EIH.

Stops are composed of a closure followed by a burst, visible as a dark vertical line in the spectrogram, seen on pages 84 through 90 in [30]. Affricates are considered to be a combination of a stop followed by a fricative, and they also exhibit sharp spectral changes from stop closure to stop burst as seen on pages 242 and 243. It is reasonable to expect that dynamic information which represents the time changes of the spectral envelope would significantly improve classification accuracy. This does not, however, seem to be the case in Table C.10; in the second and third columns looking within the phone groups for each front end, for MEL, Af shows low error-rate reduction, and for EIH, SV shows low error-rate reduction. Af shows a high error-rate reduction for EIH.

2. ***Te***, test speech passed through telephone channel simulation

   - Relative differences between MEL and EIH (Table C.7)

     With both sets of *static* features, EIH outperforms MEL for all sounds except SV, Gl, CH and Ns with the first set and SV, Gl, Af and Ns with the second set.

     With the second set of *static+dynamic* features, MEL outperforms EIH for SU, SV, Gl, Lq, FL, CM, Ns, FM and Wh.

     In the third *resegmentation*, MEL outperforms EIH for SU, SV, Af, Lq, FL, Gl, Ns and BL.

     Sounds that are classified better by EIH across almost all conditions are the vowels FH, CM, CH, BH, BM, Dp and the fricatives FV, FU.

     It must be kept in mind that high accuracies for certain sounds do not necessarily reflect better performance. The next section, B.2, which contains observations on the off-diagonal behavior of confusion matrices, points out examples of phone groups that a large number of other groups get classified as, including the correct group itself. As a result, the accuracy of these groups seems high. This is significant for the noise conditions, ***Te*** and ***Rv***. Here, the sounds in question are SV, Ns and Wh for MEL, and Ns, SV, Lq and Wh for EIH.

- Relative increases in accuracy for MEL and for EIH (Table C.11)

  With the addition of Ener to Env, the sounds that show large improvements are Af, FH, BH and SV for MEL and Af, CH, BH and FH for EIH. The smallest increases are shown by FV, FU, BL and Ns for MEL, and SU, Gl, SV, FV and FU for EIH.

  With the addition of Ener and all derivatives to Env, largest relative increases are shown by CM, FH, SU, FV and BH for MEL, and Gl, CH, FV, Af and BH for EIH. The lowest increases are shown by Ns, BL, Lq and SV for MEL and SU, SV, Ns and Lq for EIH.

  In the third resegmentation, the highest increases relative to TIMIT segmentation are shown by FU, Af, FV and SU for MEL and SV, SU, CM and Gl for EIH. The lowest improvements are shown by CH, Wh, FM, Dp and Ns for MEL and FH, Dp, CH, Ns and BM for EIH.

An interesting trend in Table C.3 is the consistently low accuracy of FV and FU for MEL, of SU and SV for EIH, and of Af for both MEL and EIH. The spectrograms of FU, unvoiced fricatives, on page 93 of [30] show very little energy in low frequency regions, and a large amount of energy distributed almost uniformly in frequency regions above 1200 Hz. FV, voiced fricatives shown on page 94, also exhibit a large distribution of energy in high frequency regions above 1200 Hz, and some in the low frequency regions below 500 Hz (called a "voice bar"). Affricates are similar to fricatives in the region after the stop burst, where they have energy primarily in the high frequency regions.

The telephone channel simulation adds white noise to clean speech, and then filters the noisy speech with a bandpass filter that has a passband of 300-2600 Hz, shown in Figure 4-2. The effect of channel filtering is additive in the cepstral domain and is countered slightly for these sounds by the inclusion of cepstral derivatives in the feature vector. These accuracies are such low numbers possibly because the speech spectrum outside the passband, below 300 Hz and above 2600 Hz, gets irretrievably attenuated by the bandpass filter. Also, the mel-scale filterbank used as the first stage in both front ends models the frequency content of speech beyond 1 kHz in much less detail than the part of the spectrum before 1 kHz.

3. ***Rv***, test speech passed through room reverberation simulation

   - Relative differences between MEL and EIH (Table C.8)

     EIH outperforms MEL for all vowels, with both sets of *static* features. The absolute classification accuracies are very low.

     With the second set of *static+dynamic* features, EIH outperforms MEL for BL, BM, FL, Dp, CH and FU.

At the third *resegmentation*, EIH outperforms MEL for BL, BM, Dp, CH, FL, FU, and BH.

Across all conditions, EIH outperforms MEL for FL, CH, BM, BL, Dp and FU. It consistently falls short of MEL for Af, FV, SV, SU, Wh and Ns.

The absolute percent accuracies for all feature sets obtained with TIMIT segmentation are very low numbers, so the observations made here could be misleading. Also, as pointed out in the previous section on ***Te***, the accuracies of certain sound groups must not be taken at face value because several groups are very frequently classified as these groups. For ***Rv***, the accuracies for all feature sets with TIMIT segmentation are in question, and Wh is the most confused-with sound. The others are SU, FU and FV for MEL and FU, BM, BL and Dp for EIH (from confusion matrices G.1 through G.14).

- Relative increases in accuracy for MEL and for EIH (Table C.12)

  With the addition of Ener and all derivatives to Env, the sounds with the largest improvements are CM, FM, FH, FL and Dp for MEL and FL, FH, FM, SV and CH for EIH. Lowest improvements are shown by Wh, FU, Af, FV and SV for MEL, and FU, Ns, Wh, Lq and Gl for EIH.

  With the third resegmentation, the performance improvements relative to TIMIT segmentation are the largest for CM, FL, FM, BL and BM for MEL, and for CM, Ns, SV, FM and Gl for EIH. The lowest increases are shown by Af, Wh, BH, CH, Dp and FV for MEL for MEL and BL, Dp, CH, BM, FL and FU for EIH.

It would be reasonable to expect that the sounds with very short durations, such as stops, would suffer the most under reverberation, but here stops seem to perform better than some of the vowels, especially SV for MEL and SU for EIH. In Table C.4, with the full static+dynamic feature set (the fourth column), MEL yields the lowest scores in the case of CM, BL and FM and EIH yields the lowest scores for CM, FM and SV. Sounds that score the lowest in the third resegmentation are BL, CH, FM and BM for MEL, and CM, FM, Af and SV for EIH.

B.2 Off-diagonal trends of confusion matrices

In this section, some prominent characteristics of the confusion matrices are listed for ***Cl***, ***Te*** and ***Rv***. One set of static features and one set of static+dynamic features is considered in detail in most cases - [Env, Ener], and [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener].

The pairs of phone groups that are confused most often with each other are enclosed in square brackets [P1,P2].

1. ***Cl***, clean test speech

   - [Env, Ener] (Matrices E.3,E.4)

For MEL, [CH,Lq] and [Dp,BL] are the main confusable pairs. For EIH, [Dp,FH] and [Dp,BL] are the main confusable pairs, Dp is confused equally with FH, CM and BL. For both MEL and EIH, [FV,FU] and [SV,SU] are confusable pairs. For both MEL and EIH, FV is classified as FU more often than FU is classified as FV. For MEL, SV is classified as SU more often than SU is classified as SV.

- [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener] (Matrices E.7 and E.8)

  For both MEL and EIH, [FH,Dp], [CH,Lq], [FV,FU] and [SV,SU] are the main confusable pairs. For both MEL and EIH, FV is classified as FU more often than FU is classified as FV, and SV is classified as SU more often than SU is classified as SV. For MEL, [FM,FL] is also a confusable pair, FL is confused equally with FM and Dp.

  With static+dynamic features for both MEL and EIH, CM is confused with several other groups. For EIH, Wh is very scattered as well. Here in Matrices B.1 and B.2, the detailed confusions for some phones that CM is frequently confused with in matrices E.7 and E.8 are shown. CM consists of the TIMIT phones *ah* (b*u*t), *ax* (*a*bout), *ax-h* (s*u*spect) and *ix* (deb*i*t). For the classification experiments with 46 phones, *ax-h* and *ix* were combined into *ax* as shown in Table 4.1.

Performance of MEL and EIH for the sound group BL is considered in some detail here. BL consists of the back, low vowel, *aa* (f*a*ther). From the spectrograms on pages 102 and 111 in the book [30], it is expected that BL, shown as /a/ in the book, would be confused the most with BM, specifically with the vowel shown as the inverted "c" (*ao* in b*au*d), because these two sounds have similar locations of formants.

For MEL, with static features as shown in matrix E.3, BL (accuracy 40%) is confused the most with Dp (23%), followed by BM (16%). With the addition of dynamic information, in matrix E.7, BL (accuracy 62%) is confused the most with BM (16%) followed by Dp (12%). The decrease in mis-classified instances of BL is the most for Dp, (0% decrease in BM, 11% decrease in Dp). This is reasonable because the addition of dynamic information is expected to reflect the changing formant locations of diphthongs.

For EIH, with static features in matrix E.4, BL (accuracy 43%) is confused the most with Dp and BM (equally with both, 17%). With the addition of dynamic information, in matrix E.8, BL (accuracy 61%) is confused the most with BM (15%) followed by Dp (13%). The decrease in mis-classified instances of BL is more for Dp, (2% decrease in BM, 4% decrease in Dp). The overall increase in EIH accuracy with the addition of dynamic information is less than that in MEL, possibly because of reasons discussed in Section 5.2.1.

- 3 iterations of automatic resegmentation (Matrices E.13 and E.14)

  For both MEL and EIH, the confusable pairs are [FM,FL], [CH,Lq], [FV,FU] and [SV,SU]. The voiced and unvoiced fricatives and the voiced and unvoiced stops

Table B.1: Detailed confusions for CM, central vowels; Test set: Clean; TIMIT segmentation; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; MEL

	ah	ax	er	iy	ih	uh	uw	ey	aw	ay	oy	l	r	Total
ah	**75**	4			2	3			5	5	2		3	397
ax	8	**50**		5	17	7	4	2			1	1	2	2032
er			**84**		1								12	304
iy		2		**85**	3			7						1087
ih	3	12	1	6	**59**	3	5	8			1		1	731
uh	12	11	3		10	**49**	9				2	2	2	117
uw		3		7	5	2	**78**					1	1	327
ey		1		7	4			**83**		2	2			505
aw	6								**91**			2		116
ay	5							2	2	**89**	1			404
oy	1		1							8	**86**	2	1	80
l	1	1					1		1		1	**90**		814
r	1		12								1		**82**	995

Table B.2: Same conditions as in Matrix B.1; EIH

	ah	ax	er	iy	ih	uh	uw	ey	aw	ay	oy	l	r	Total
ah	**69**	5	2		1	3	1		4	6	4	1	3	385
ax	8	**51**	2	6	14	5	5	2		1	2	2	2	1873
er			**83**		1		2						11	333
iy		3		**82**	5			9						1051
ih	6	13	2	8	**53**	2	5	9			2			723
uh	19	14			8	**40**	10	2			3	2	3	115
uw	2	6	2	7	6	2	**71**					1	1	294
ey		1		9	4			**79**		2	3			499
aw	8								**90**					120
ay	7							1	2	**86**	2			408
oy	1		1	1						8	**89**			79
l	1	3				2	1		3		1	**86**	1	752
r	1	2	14							1	1		**77**	958

show the same behavior as discussed for [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]. For MEL, other confusable pairs are [FH,CM] and [BM,BL], and for EIH, [FH,Dp] is another confusable pair.

2. ***Te***, test speech passed through telephone channel simulation

    - [Env, Ener] (Matrices F.3,F.4)

      For MEL, all phone groups are confused very frequently with SV, several groups are confused frequently with Ns and some with Lq. Very few phone groups are classified as FV, BH or FH, and none are classified as CM or FU (possibly for reasons discussed in Section B.1). For EIH, most phone groups are confused very frequently with Ns, and less frequently with Lq and SV. Very few groups are classified as SU. The consonants are confused much more frequently with Ns and Wh than in the case of MEL.

    - [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener] (Matrices F.7 and F.8)

      The addition of dynamic features to spectral envelope and energy reduces confusion dramatically. For MEL, much fewer groups are confused with SV and Ns than with static features. Very few groups are classified as FU or FV. Most of the consonants are confused frequently with Wh. Some confusable pairs are [FM,FL], [FL,Dp] and [CH,Lq]. For EIH, much fewer groups are confused with Ns than with static features, also fewer groups are confused with SV or Lq. Very few groups are classified as SU. [FH,Dp] is the only confusable pair observed.

    - 3 iterations of automatic resegmentation (Matrices F.13 and F.14)

      For MEL, confusions of consonants with Wh decrease from those with TIMIT segmentation. Confusions with SV and Ns increase slightly. Some confusable pairs are [FL,Dp], [CH,Lq] and [Ns,SV]. For EIH, confusions with most of the vowels (especially CM, BH and Dp) and Lq decrease. The confusable pairs are [FH,Dp] and [CH,Lq].

3. ***Rv***, test speech passed through reverberation simulation

    - [Env, Ener] (Matrices G.3,G.4)

      For both MEL and EIH, all phone groups are confused very frequently with Wh. With smaller frequency, several groups are confused with SU, SV and FV for MEL and Dp, Lq and SU for EIH. For MEL, the majority of phone groups are classified as groups towards the right side of the matrix, in the consonants. For EIH most confusions fall into the middle of the matrix, in the back vowels, semivowels and nasals. For MEL, very few groups are classified as vowels. For EIH, few groups are classified as the front and central vowels and SV.

    - [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener] (Matrices G.7 and G.8)

      With the addition of derivatives to envelope and energy, the general pattern of the confusion matrices does not change appreciably.

After one iteration of automatic resegmentation, a dramatic change takes place in the distribution of phone groups in the confusion matrices, seen by comparing matrices G.7 and G.9, and matrices G.8 and G.10.

- 3 iterations of automatic resegmentation (Matrices G.13 and G.14)

  For MEL, most groups are confused with Wh, then with FU, SU and Ns. For EIH, most groups are confused with Wh, most vowels are confused with the back vowels, Dp and Lq.

Appendix C

Diagonal elements of confusion matrices

Table C.1: Train set : Clean; Percent correct for 18 phone groups; correct phone as top 1 candidate

	Static Features				Static+Dynamic Features				Automatic resegmentation: iteration number					
	Env		Env , Ener		Env Δ-Δ_2		Env , Ener Δ-Δ_2		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
FH	55	**58**	58	**59**	**73**	68	**76**	69	**75**	69	**79**	77	**80**	77
FM	**41**	35	**41**	36	**55**	49	**59**	50	**57**	53	**63**	62	**64**	62
FL	62	**63**	**64**	64	**76**	71	**76**	72	**76**	72	79	**79**	80	**80**
CM	29	**32**	**35**	35	**47**	44	**54**	48	**51**	50	63	**63**	64	**65**
CH	60	**63**	62	**66**	**77**	76	**80**	78	**79**	77	80	**85**	81	**85**
BH	**52**	47	**52**	47	**71**	63	**75**	64	**73**	65	**81**	75	**81**	76
BM	58	**59**	59	**61**	**76**	71	**77**	75	**77**	76	**80**	79	**81**	80
BL	**48**	48	48	**49**	**69**	67	**72**	68	**69**	67	**75**	74	**77**	75
Dp	**74**	70	**74**	69	**89**	86	**88**	87	**89**	89	**90**	90	90	**90**
Lq	**62**	59	**64**	60	**76**	71	**78**	73	**77**	74	**83**	81	**84**	81
Gl	**64**	54	**66**	55	**83**	74	**86**	78	**87**	81	**91**	87	**91**	88
Ns	**83**	71	**86**	75	**91**	84	**93**	86	**94**	88	**95**	91	**95**	91
FV	**61**	54	**65**	58	**74**	68	**77**	70	**77**	72	**83**	78	**84**	79
FU	**82**	78	**84**	80	**91**	82	**92**	83	**93**	85	**95**	91	**95**	91
SV	**63**	56	**65**	57	**76**	66	**80**	68	**79**	68	**86**	74	**86**	74
SU	**73**	69	**77**	71	**81**	74	**85**	76	**86**	79	**89**	81	**89**	82
Af	**84**	75	**87**	79	**88**	80	**92**	82	**90**	82	**92**	86	**93**	86
Wh	**63**	39	**69**	47	**79**	62	**87**	70	**86**	72	**92**	84	**92**	84

Table C.2: Test set : Clean; Percent correct for 18 phone groups; correct phone as top 1 candidate

	Static Features				Static+Dynamic Features				Automatic resegmentation: iteration number					
	Env		Env , Ener		Env Δ-Δ_2		Env , Ener Δ-Δ_2		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
FH	49	**53**	51	**54**	**67**	64	**70**	65	**70**	66	**73**	72	**76**	72
FM	**35**	31	**34**	31	**50**	38	**50**	41	**50**	47	51	**54**	55	**55**
FL	**60**	58	**60**	58	**71**	68	**72**	70	**75**	71	73	**78**	76	**78**
CM	26	**29**	**32**	31	**43**	42	**49**	45	**48**	48	60	**60**	60	**62**
CH	45	**50**	45	**55**	62	**64**	61	**67**	64	**70**	64	**73**	62	**75**
BH	**48**	44	**50**	43	**59**	52	**63**	52	**64**	59	**72**	64	**72**	68
BM	50	**53**	52	**55**	**69**	65	**74**	67	**75**	74	**76**	75	**78**	75
BL	**45**	41	40	**43**	**60**	60	**62**	61	**63**	61	**70**	69	**71**	64
Dp	**69**	62	**68**	62	**82**	80	**83**	81	**85**	83	**85**	85	**85**	85
Lq	**58**	56	**60**	57	**70**	64	**73**	66	**72**	70	**77**	76	**78**	76
Gl	**59**	50	**64**	53	**79**	69	**83**	74	**85**	78	**88**	84	**89**	85
Ns	**80**	66	**83**	70	**88**	79	**91**	83	**92**	85	**93**	88	**93**	88
FV	**56**	51	**60**	54	**64**	61	**68**	63	**71**	66	**75**	71	**75**	70
FU	**80**	77	**81**	77	**88**	80	**91**	82	**92**	85	**93**	90	**94**	89
SV	**59**	54	**62**	56	**69**	59	**72**	61	**73**	63	**80**	67	**80**	67
SU	**70**	66	**75**	68	**79**	72	**84**	74	**86**	77	**87**	79	**88**	80
Af	**83**	69	**85**	78	**83**	78	**88**	82	**89**	79	**91**	84	**89**	85
Wh	**60**	34	**66**	41	**77**	48	**84**	60	**86**	65	**90**	76	**91**	78

Table C.3: Test set : Telephone channel simulation; Percent correct for 18 phone groups; correct phone as top 1 candidate

	Static Features				Static+Dynamic Features				Automatic resegmentation: iteration number					
	Env		Env , Ener		Env Δ-Δ_2		Env , Ener Δ-Δ_2		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
FH	2	**32**	5	**42**	42	**63**	51	**65**	45	**64**	47	**67**	54	**67**
FM	7	**22**	9	**24**	**39**	37	**42**	41	39	**44**	46	**52**	43	**52**
FL	29	**63**	30	**67**	68	**71**	**77**	71	**76**	73	**80**	78	**85**	78
CM	1	**17**	1	**22**	18	**31**	**37**	35	26	**37**	44	**46**	48	**49**
CH	**20**	18	20	**31**	43	**60**	52	**67**	50	**65**	48	**70**	52	**71**
BH	2	**19**	4	**25**	26	**44**	34	**47**	29	**52**	41	**58**	44	**60**
BM	24	**36**	31	**43**	58	**66**	61	**73**	58	**75**	67	**77**	66	**78**
BL	41	**46**	29	**53**	55	**57**	59	**60**	**66**	60	62	**67**	**66**	65
Dp	24	**42**	39	**46**	74	**81**	74	**83**	67	**85**	75	**86**	77	**87**
Lq	49	**58**	53	**56**	**73**	59	**75**	61	**76**	65	**79**	71	**80**	72
Gl	**20**	12	**26**	10	**65**	55	**70**	56	**72**	64	**75**	70	**78**	74
Ns	**89**	82	**83**	83	**86**	79	**86**	82	**87**	87	**89**	87	**90**	87
FV	0	**12**	0	**11**	5	**34**	7	**34**	3	**31**	8	**39**	10	**38**
FU	0	**15**	0	**14**	0	**30**	2	**32**	4	**40**	7	**37**	7	**37**
SV	**42**	20	**80**	18	**61**	14	**67**	15	**74**	27	**79**	22	**79**	23
SU	1	**8**	1	**4**	**13**	4	**22**	2	**22**	3	**28**	3	**30**	3
Af	4	**7**	**15**	13	6	**10**	16	**18**	**22**	18	**31**	26	**30**	22
Wh	14	**57**	20	**58**	68	**73**	**78**	76	75	**82**	79	**89**	78	**89**

Table C.4: Test set : Reverberation simulation; Percent correct for 18 phone groups; correct phone as top 1 candidate

	Static Features				Static+Dynamic Features				Automatic resegmentation: iteration number					
	Env		Env , Ener		Env Δ-Δ_2		Env , Ener Δ-Δ_2		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
FH	3	**4**	6	**7**	**19**	14	**23**	18	41	**41**	**47**	47	**47**	45
FM	1	**2**	2	**4**	4	**6**	**10**	8	**30**	23	**36**	23	**38**	26
FL	2	**8**	16	**18**	9	**32**	13	**43**	43	**54**	54	**63**	51	**65**
CM	0	**2**	1	**3**	**4**	4	**5**	5	**35**	14	**46**	19	**48**	22
CH	6	**11**	15	**16**	16	**25**	20	**32**	33	**44**	35	**47**	33	**44**
BH	11	**11**	19	**21**	**26**	26	**36**	30	54	**56**	59	**62**	59	**67**
BM	11	**36**	9	**41**	15	**50**	14	**51**	41	**69**	43	**73**	41	**71**
BL	2	**23**	3	**39**	7	**40**	7	**51**	18	**64**	28	**67**	23	**65**
Dp	6	**24**	17	**35**	17	**47**	30	**56**	44	**73**	55	**73**	53	**73**
Lq	14	**17**	**15**	14	**26**	18	**31**	20	**61**	49	**67**	58	**68**	58
Gl	14	**16**	14	**16**	**32**	19	**35**	19	**63**	52	**70**	59	**70**	60
Ns	**24**	17	**29**	16	**33**	18	**36**	18	**79**	76	**84**	74	**84**	76
FV	**32**	8	**36**	9	**31**	17	**35**	19	**58**	33	**63**	40	**62**	43
FU	35	**50**	30	**51**	34	**51**	33	**50**	66	**87**	70	**86**	69	**86**
SV	**20**	3	**21**	5	**21**	7	**22**	9	**53**	25	**60**	30	**61**	31
SU	**22**	14	**26**	16	**42**	17	**40**	18	**72**	45	**71**	46	**73**	48
Af	**53**	12	**48**	16	**59**	15	**58**	15	**63**	26	**64**	25	**63**	28
Wh	**66**	41	**60**	42	**57**	44	**61**	47	**92**	78	**92**	86	**93**	88

Table C.5: Train set : Clean; Relative differences, in percent, between MEL and EIH $(\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100)$ for 18 phone groups, correct phone as top 1 candidate

	Static Features		Static+Dynamic Features		Automatic resegmentation: iteration number		
	Env	Env , Ener	Env Δ-Δ_2	Env , Ener Δ-Δ_2	One	Two	Three
FH	**-5**	**-3**	7	9	8	3	4
FM	15	13	13	17	7	2	3
FL	**0**	0	6	5	6	**0**	**0**
CM	**-10**	0	6	11	1	**0**	**-1**
CH	**-5**	**-6**	2	2	2	**-6**	**-4**
BH	11	12	12	16	11	8	7
BM	**-2**	**-3**	6	3	1	1	2
BL	1	**-1**	2	4	3	1	3
Dp	7	7	3	2	0	0	**0**
Lq	5	7	7	7	5	2	3
Gl	18	19	11	10	8	3	4
Ns	15	14	8	7	6	5	4
FV	11	11	8	10	7	6	6
FU	5	5	10	10	9	4	5
SV	11	13	14	16	15	15	14
SU	6	8	9	12	9	9	8
Af	11	10	10	11	9	7	8
Wh	46	39	24	21	18	8	10

Table C.6: Test set : Clean; Relative differences, in percent, between MEL and EIH $(\frac{MEL-EIH}{\frac{MEL+EIH}{2}}.100)$ for 18 phone groups, correct phone as top 1 candidate

	Static Features		Static+Dynamic Features		Automatic resegmentation: iteration number		
	Env	Env , Ener	Env Δ-Δ_2	Env , Ener Δ-Δ_2	One	Two	Three
FH	**-7**	**-5**	4	7	6	2	4
FM	12	11	26	20	5	**-6**	**0**
FL	3	2	4	4	6	**-6**	**-2**
CM	**-10**	3	3	7	0	**0**	**-3**
CH	**-10**	**-21**	**-3**	**-9**	**-9**	**-14**	**-19**
BH	10	14	14	19	9	11	6
BM	**-7**	**-5**	7	10	2	1	4
BL	9	**-8**	1	1	3	2	10
Dp	11	9	2	3	2	0	1
Lq	3	5	9	10	3	1	2
Gl	16	19	13	12	9	5	4
Ns	19	17	10	9	8	6	6
FV	9	11	6	8	7	6	7
FU	4	5	10	11	8	4	5
SV	9	11	15	16	14	18	17
SU	6	9	9	12	11	10	9
Af	18	8	7	7	11	8	5
Wh	55	46	45	33	27	17	15

Table C.7: Test set : Telephone channel simulation; Relative differences, in percent, between MEL and EIH ($\frac{MEL-EIH}{\frac{MEL+EIH}{2}}$.100) for 18 phone groups, correct phone as top 1 candidate

	Static Features		Static+Dynamic Features		Automatic resegmentation: iteration number		
	Env	Env , Ener	Env Δ-Δ_2	Env , Ener Δ-Δ_2	One	Two	Three
FH	**-176**	**-154**	**-40**	**-24**	**-35**	**-35**	**-22**
FM	**-104**	**-91**	5	3	**-11**	**-11**	**-19**
FL	**-75**	**-75**	**-4**	7	5	3	9
CM	**-178**	**-188**	**-50**	6	**-35**	**-5**	-1
CH	10	**-43**	**-34**	**-25**	**-26**	**-39**	**-29**
BH	**-158**	**-148**	**-53**	**-34**	**-58**	**-36**	**-31**
BM	**-41**	**-33**	**-14**	**-17**	**-25**	**-14**	**-16**
BL	**-12**	**-58**	**-4**	**-2**	10	**-8**	2
Dp	**-54**	**-17**	**-8**	**-11**	**-23**	**-13**	**-12**
Lq	**-17**	**-6**	22	21	16	11	10
Gl	47	88	17	22	11	7	5
Ns	8	1	9	5	0	2	3
FV	**-186**	**-192**	**-148**	**-134**	**-161**	**-136**	**-117**
FU	**-200**	**-199**	**-196**	**-180**	**-162**	**-140**	**-134**
SV	72	127	127	125	93	115	110
SU	**-164**	**-109**	113	172	152	159	161
Af	**-54**	14	**-56**	**-12**	20	18	29
Wh	**-120**	**-99**	**-7**	3	**-10**	**-11**	**-13**

Table C.8: Test set : Room reverberation simulation; Relative differences, in percent, between MEL and EIH ($\frac{MEL-EIH}{\frac{MEL+EIH}{2}}$.100) for 18 phone groups, correct phone as top 1 candidate

	Static Features		Static+Dynamic Features		Automatic resegmentation: iteration number		
	Env	Env , Ener	Env Δ-Δ_2	Env , Ener Δ-Δ_2	One	Two	Three
FH	**-37**	**-10**	26	21	**-2**	0	4
FM	**-68**	**-53**	**-25**	30	26	45	38
FL	**-112**	-11	**-114**	**-105**	**-23**	-14	**-23**
CM	**-139**	**-84**	6	6	85	81	75
CH	-54	**-6**	**-41**	**-45**	**-28**	**-30**	**-28**
BH	**0**	**-8**	1	17	-2	**-5**	**-12**
BM	**-109**	**-128**	**-110**	**-113**	**-50**	**-52**	**-54**
BL	**-165**	**-168**	**-143**	**-153**	-111	**-81**	**-96**
Dp	**-123**	**-70**	**-92**	**-61**	**-49**	**-28**	**-31**
Lq	**-20**	12	36	42	21	15	14
Gl	-11	**-10**	53	59	19	17	15
Ns	38	59	62	66	5	12	11
FV	123	120	58	58	53	44	37
FU	**-35**	**-52**	**-41**	**-41**	**-27**	**-20**	-21
SV	147	129	105	79	71	67	66
SU	43	46	83	74	47	43	42
Af	129	99	120	116	83	88	75
Wh	47	37	25	25	16	7	6

Table C.9: Train set : Clean; Relative increase in accuracy, in percent, for 18 phone groups, correct phone as top 1 candidate

	Addition to Env of						Resegmentation iteration number					
	Ener		Δ-Δ_2 Env		Ener,Δ-Δ_2 Env , Ener		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
FH	**5**	2	**33**	17	**38**	19	-1	**0**	4	**12**	5	**12**
FM	0	**3**	34	**40**	**44**	43	-3	**6**	7	**24**	8	**24**
FL	**3**	2	**23**	13	**23**	14	**0**	0	4	**10**	5	**11**
CM	**21**	9	**62**	38	**86**	50	-6	**4**	17	**31**	19	**35**
CH	3	**5**	**28**	21	**33**	24	-1	-1	0	**9**	1	**9**
BH	**0**	0	**37**	34	**44**	36	-3	**2**	8	**17**	8	**19**
BM	2	**3**	**31**	20	**33**	27	0	**1**	4	**5**	5	**7**
BL	0	**2**	**44**	40	**50**	42	-4	**-1**	4	**9**	7	**10**
Dp	**0**	-1	20	**23**	19	**24**	1	**2**	2	**3**	2	**3**
Lq	**3**	2	**23**	20	**26**	24	-1	**1**	6	**11**	8	**11**
Gl	**3**	2	30	**37**	34	**44**	1	**4**	6	**12**	6	**13**
Ns	4	**6**	10	**18**	12	**21**	1	**2**	2	**6**	2	**6**
FV	7	**7**	21	**26**	26	**30**	0	**3**	8	**11**	9	**13**
FU	2	**3**	**11**	5	**12**	6	1	**2**	3	**10**	3	**10**
SV	**3**	2	**21**	18	**27**	21	-1	**0**	8	**9**	8	**9**
SU	**5**	3	**11**	7	**16**	10	1	**4**	5	**7**	5	**8**
Af	4	**5**	5	**7**	**10**	9	-2	**0**	0	**5**	1	**5**
Wh	10	**21**	25	**59**	38	**79**	-1	**3**	6	**20**	6	**20**

Table C.10: Test set : Clean; Relative increase in accuracy, in percent, for 18 phone groups, correct phone as top 1 candidate

	Addition to Env of						Resegmentation iteration number					
	Ener		Δ-Δ_2 Env		Ener,Δ-Δ_2 Env , Ener		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
FH	**4**	2	**37**	21	**43**	23	0	**2**	4	**11**	9	**11**
FM	-3	**0**	**43**	23	**43**	32	0	**15**	2	**32**	10	**34**
FL	**0**	0	**18**	17	20	**21**	**4**	1	1	**11**	6	**11**
CM	**23**	7	**65**	45	**88**	55	-2	**7**	22	**33**	22	**38**
CH	0	**10**	**38**	28	**36**	34	**5**	4	5	**9**	2	**12**
BH	**4**	-2	**23**	18	**31**	18	2	**13**	14	**23**	14	**31**
BM	**4**	4	**38**	23	**48**	26	1	**10**	3	**12**	5	**12**
BL	-11	**5**	33	**46**	38	**49**	**2**	0	13	**13**	**15**	5
Dp	-1	**0**	19	**29**	20	**31**	2	**2**	2	**5**	2	**5**
Lq	**3**	2	**21**	14	**26**	18	-1	**6**	5	**15**	7	**15**
Gl	**8**	6	34	**38**	41	**48**	2	**5**	6	**14**	7	**15**
Ns	4	**6**	10	**20**	14	**26**	1	**2**	2	**6**	2	**6**
FV	**7**	6	14	**20**	21	**24**	4	**5**	10	**13**	10	**11**
FU	**1**	0	**10**	4	**14**	6	1	**4**	2	**10**	3	**9**
SV	**5**	4	**17**	9	**22**	13	1	**3**	**11**	10	**11**	10
SU	**7**	3	**13**	9	**20**	12	2	**4**	4	**7**	5	**8**
Af	2	**13**	0	**13**	6	**19**	1	-4	**3**	2	1	**4**
Wh	10	**21**	28	**41**	40	**76**	2	**8**	7	**27**	8	**30**

Table C.11: Test set : Telephone channel simulation; Relative increase in accuracy, in percent, for 18 phone groups, correct phone as top 1 candidate

	Addition to Env of						Resegmentation iteration number					
	Ener		Δ-Δ_2 Env		Ener,Δ-Δ_2 Env , Ener		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
FH	**150**	31	**2000**	97	**2450**	103	-12	**-2**	-8	**3**	**6**	3
FM	**29**	9	**457**	68	**500**	86	-7	**7**	10	**27**	2	**27**
FL	3	**6**	**134**	13	**166**	13	-1	**3**	4	**10**	**10**	10
CM	0	**29**	**1700**	82	**3600**	106	-30	**6**	19	**31**	30	**40**
CH	0	**72**	115	**233**	160	**272**	-4	**-3**	-8	**4**	0	**6**
BH	**100**	32	**1200**	132	**1600**	147	-15	**11**	21	**23**	**29**	28
BM	**29**	19	**142**	83	**154**	103	-5	**3**	**10**	5	**8**	7
BL	-29	**15**	**34**	24	**44**	30	**12**	0	5	**12**	**12**	8
Dp	**62**	10	**208**	93	**208**	98	-9	**2**	1	**4**	4	**5**
Lq	**8**	-3	**49**	2	**53**	5	1	**7**	5	**16**	7	**18**
Gl	**30**	-17	225	**358**	250	**367**	3	**14**	7	**25**	11	**32**
Ns	-7	**1**	**-3**	-4	-3	**0**	1	**6**	3	**6**	5	**6**
FV	-100	**-8**	**1150**	183	**1650**	183	-57	**-9**	14	**15**	**43**	12
FU	-100	**-7**	-100	**100**	**400**	113	**100**	25	**250**	16	**250**	16
SV	**90**	-10	**45**	-30	**60**	-25	10	**80**	18	**47**	18	**53**
SU	**0**	-50	**1200**	-50	**2100**	-75	0	**50**	27	**50**	36	**50**
Af	**275**	86	**50**	43	**300**	157	**38**	0	**94**	44	**88**	22
Wh	**43**	2	**386**	28	**457**	33	-4	**8**	1	**17**	0	**17**

Table C.12: Test set : Room Reverberation Simulation; Relative increase in accuracy, in percent, for 18 phone groups, correct phone as top 1 candidate

	Addition to Env of						Resegmentation iteration number					
	Ener		Δ-Δ_2 Env		Ener,Δ-Δ_2 Env , Ener		One		Two		Three	
	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH	MEL	EIH
FH	**100**	75	**533**	250	**667**	350	78	**128**	104	**161**	104	**150**
FM	**100**	100	**300**	200	**900**	300	**200**	188	**260**	188	**280**	225
FL	**700**	125	**350**	300	**550**	438	**231**	26	**315**	47	**292**	51
CM	**150**	50	**900**	100	**1150**	150	**600**	180	**820**	280	**860**	340
CH	**150**	45	**167**	127	**233**	191	**65**	38	**75**	47	**65**	38
BH	73	**91**	**136**	136	**227**	173	50	**87**	64	**107**	64	**123**
BM	-18	**14**	36	**39**	27	**42**	**193**	35	**207**	43	**193**	39
BL	50	**70**	**250**	74	**250**	122	**157**	25	**300**	31	**229**	27
Dp	**183**	46	**183**	96	**400**	133	**47**	30	**83**	30	**77**	30
Lq	**7**	-18	**86**	6	**121**	18	97	**145**	116	**190**	119	**190**
Gl	**0**	0	**129**	19	**150**	19	80	**174**	100	**211**	100	**216**
Ns	**21**	-6	**38**	6	**50**	6	119	**322**	133	**311**	133	**322**
FV	**12**	12	-3	**112**	9	**138**	66	**74**	80	**111**	77	**126**
FU	-14	**2**	-3	**2**	-6	**0**	**100**	74	**112**	72	**109**	72
SV	5	**67**	5	**133**	10	**200**	141	**178**	173	**233**	177	**244**
SU	**18**	14	**91**	21	**82**	29	80	**150**	78	**156**	82	**167**
Af	-9	**33**	11	**25**	9	**25**	9	**73**	10	**67**	9	**87**
Wh	-9	**2**	-14	**7**	-8	**15**	51	**66**	51	**83**	52	**87**

Appendix D

Confusion Matrices - Train set : Clean

Table D.1: Train set : Clean; Static features [Env]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**55**	3	1	6	1	7			14	2	7	2							6102
FM	8	**41**	11	14	3	1	2		15	3		1							2273
FL	1	13	**62**	3				2	17										1632
CM	16	6	2	**29**	1	9	6	1	8	7	2	4	6		2				9099
CH	2	2			**60**	1			1	29					1				1154
BH	13			5		**52**	4		2	6	9	2	4		2				1721
BM				6		1	**58**	6	7	15	5								2539
BL			3	8			13	**48**	21	6									1602
Dp	4	4	4	4			1	4	**74**	1									3684
Lq		1		2	11	3	9	2	2	**62**	3	1	2		1				8585
Gl	7					5	5		2	9	**64**	3	3		1				2102
Ns	1			2		1				2		**83**	4		4				8204
FV				2		3				3	1	2	**61**	17	4	3	2		5195
FU													10	**82**		3	4		7384
SV	2			1		2				2	2	7	5	1	**63**	11	2		5568
SU												2	4	7	10	**73**	4		8932
Af													6	6	2	2	**84**		1291
Wh	3		2	2					2	2	2	4	3	4	2	8	1	**63**	1123

Table D.2: Same conditions as in Matrix D.1; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**58**	3	1	6	2	6	1		12	2	5	2							6084
FM	13	**35**	11	13	4	2	2		14	2		1							2273
FL	2	11	**63**	3				1	17										1633
CM	14	4	2	**32**	2	7	6	2	6	7	2	5	3	4	1	1		3	9033
CH	2	2		2	**63**	1			2	26									1154
BH	16	1		7	1	**47**	7		2	7	8	2	2						1717
BM				6		2	**59**	7	5	14	4	1							2539
BL		1	3	9			14	**48**	17	6									1602
Dp	7	5	5	4	1		1	5	**70**	1									3685
Lq		1		3	9	2	10	2	2	**59**	3	3	2			1			8506
Gl	6					3	6		2	9	**54**	7	3	2	2	3	1	2	2039
Ns	2	1	1	2		1	2		2	4	2	**71**	4		5				8148
FV	2			1		3	2			3	1	3	**54**	17	5	3	4		5028
FU													11	**78**		3	6		7360
SV	3		1		1	1	1		2	3	2	7	3	1	**56**	13	2		5364
SU												2	2	6	13	**69**	5		8378
Af													5	7	3	8	**75**		1271
Wh	6	3	3	4	1	1	2	1	5	4	3	4	2	10	1	8	3	**39**	1089

Table D.3: Train set : Clean; Static features [Env, Ener]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**58**	3	1	5	2	7			13	2	6	1							6102
FM	9	**41**	11	14	3	1	2		14	2		1							2274
FL	1	12	**64**	3				2	16										1633
CM	13	4	2	**35**	1	9	6	1	7	8	2	4	5		2			1	9104
CH	3	2			**62**	1			1	27									1154
BH	14			6		**52**	4		2	6	9	2	2		1				1721
BM				6		1	**59**	6	7	13	5								2539
BL			2	8			14	**48**	20	5									1602
Dp	4	4	5	5			1	4	**74**										3685
Lq		1		3	9	3	8	2	2	**64**	3	1	2						8586
Gl	7					4	5		2	8	**66**	3	3						2101
Ns				1		1				2		**86**	4		3				8208
FV				2		2				3	1	2	**65**	16	4	2	1		5204
FU													10	**84**		2	3		7391
SV	1			1		1				2	1	7	5	1	**65**	12	2		5570
SU												1	2	5	10	**77**	4		8933
Af													4	3	2	2	**87**		1291
Wh	2		1	3					1	2	1	4	3	4	2	6		**69**	1117

Table D.4: Same conditions as in Matrix D.3; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**59**	4	1	6	2	6			11	1	5	1							6097
FM	13	**36**	11	14	5	2	2		13	2		1							2273
FL	3	10	**64**	4				1	16										1633
CM	11	3	1	**35**	2	7	6	1	5	7	3	5	3	3	2	1		3	9104
CH	3	3		1	**66**	1			2	22									1154
BH	17	1		7	1	**47**	7		2	6	8	2	2						1722
BM				7		2	**61**	7	5	12	4								2540
BL		1	3	9	1		14	**49**	17	5									1602
Dp	7	5	6	4	1		1	5	**69**										3685
Lq	1	1		3	8	2	10	2	2	**60**	3	3	2						8574
Gl	6					4	5		2	9	**55**	7	3	2	2	2	1	2	2025
Ns	1		1	2		1	1		2	4	1	**75**	4		5				8100
FV	1			1		2	2			3		4	**58**	16	5	3	2		5004
FU													12	**80**		3	4		7369
SV	2			1	1	1			2	3	1	8	4	1	**57**	13	1		5353
SU												2	2	5	14	**71**	4		8498
Af													6	6	2	6	**79**		1281
Wh	5	2	2	4		1			3	4	4	5	2	8	1	6	3	**47**	1071

Table D.5: Train set : Clean; Static+Dynamic features [Env, Δ-Δ_2 Env]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**73**	4	1	5		3			7	1	4								6102
FM	7	**55**	14	8	2	1	1		9	2									2274
FL	2	9	**76**	2				1	10										1632
CM	16	5	1	**47**		7	4	2	6	6	2	2							9095
CH		1		1	**77**					17									1154
BH	10			4		**71**	3		1	4	5	1							1721
BM				3		1	**76**	6	3	7	2								2539
BL			2	5	1		10	**69**	11	1									1602
Dp	3	1	2	2				2	**89**										3685
Lq				2	8	2	4		2	**76**	3	1							8587
Gl	4					2	1		1	5	**83**	1							2100
Ns				1		1						**91**	2		2				8206
FV						1				1		1	**74**	14	4	2			5225
FU													6	**91**		2	1		7382
SV										1		4	4	1	**76**	10	1		5582
SU												1	2	4	8	**81**	2		9005
Af													3	5	1	2	**88**		1291
Wh	2	1		1					1	1	1	1	2	2		5		**79**	1135

Table D.6: Same conditions as in Matrix D.5; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**68**	4	1	6		3			10	1	5								6099
FM	11	**49**	12	10	2	2	1		10	2									2274
FL	2	9	**71**	3					13										1633
CM	15	4	1	**44**	1	6	3	2	7	6	3	3	1	1		2		1	9128
CH	1	2		1	**76**				1	17									1154
BH	10	1		6	1	**63**	5		2	4	7								1723
BM				4		1	**71**	6	4	9	3								2540
BL			1	5	1		11	**67**	12	2									1602
Dp	4	1	2	2				2	**86**										3685
Lq				2	7	2	5		2	**71**	4	2							8603
Gl	4			1		3	3		2	7	**74**	2				1			2108
Ns				1		1				2	1	**84**	3		4				8221
FV						2	1			1		2	**68**	14	4	3	2		5197
FU													11	**82**		2	3		7372
SV						1				2		6	5		**66**	15	1		5544
SU												2	3	3	14	**74**	4		8861
Af													5	6	2	6	**80**		1278
Wh	5			2				1	3	1	3	2	2	7		4	3	**62**	1123

Table D.7: Train set : Clean; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**76**	4	1	4		3			7	1	3								6102
FM	8	**59**	12	8	2		1		8	1									2274
FL	2	8	**76**	2				1	10										1633
CM	15	4	1	**54**		7	4	1	4	5	1	2							9101
CH	1	1			**80**					15									1154
BH	9			3		**75**	3			4	3								1721
BM				4		1	**77**	6	3	6	2								2539
BL			1	4	1		10	**72**	10	2									1602
Dp	3	1	2	2				2	**88**										3685
Lq				1	7	1	4		2	**78**	3								8586
Gl	3					2	2		1	4	**86**	1							2100
Ns												**93**	2		2				8205
FV												1	**77**	13	4	1			5223
FU													6	**92**		1	1		7392
SV												3	4		**80**	10	1		5583
SU												1	1	3	7	**85**	2		9001
Af													2	3	1	2	**92**		1291
Wh	1								1		1		1	3		3		**87**	1135

Table D.8: Same conditions as in Matrix D.7; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**69**	4		6		3			9	1	4								6102
FM	12	**50**	12	10	2	1	1		10	1									2274
FL	2	8	**72**	3					13										1633
CM	13	3	1	**48**	1	6	3	2	6	6	3	3	1	1		1		1	9151
CH	1	2			**78**				1	15									1154
BH	10			6	1	**64**	5		2	4	7								1725
BM				3		2	**75**	6	4	7	2								2540
BL			1	5	1		10	**68**	11	2									1602
Dp	4	1	2	2				2	**87**										3685
Lq				2	7	2	5		2	**73**	4	1							8611
Gl	4					3	2		2	6	**78**	2							2117
Ns				1						1		**86**	3		4				8230
FV						2				1		3	**70**	14	5	3	1		5228
FU													11	**83**		2	2		7387
SV										1		6	5		**68**	15	1		5590
SU												1	2	3	14	**76**	3		8983
Af													5	4	2	5	**82**		1291
Wh	4			2					2	2	2	2	2	5		3	1	**70**	1129

Table D.9: Train set : Clean; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 1 iteration of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**75**	4	1	4		3			7		3								6102
FM	7	**57**	15	7	2				8	2									2274
FL	2	7	**76**	2					12										1633
CM	15	5	2	**51**		7	4	2	4	5		1							9094
CH	1	2			**79**					16									1154
BH	9			4		**73**	3			3	5								1721
BM				3		1	**77**	7	3	7	2								2539
BL			2	4	1		10	**69**	12	1									1602
Dp	3	1	2	1				2	**89**										3685
Lq				2	8	2	4	1	1	**77**	3								8585
Gl	2					2	2		1	4	**87**								2101
Ns												**94**	1		2				8205
FV												1	**77**	13	4	2			5224
FU													4	**93**		2	1		7393
SV												3	3		**79**	11	1		5579
SU													1	2	7	**86**	2		9007
Af													3	3	1	3	**90**		1291
Wh	1			1					1		1	1	2	2		2		**86**	1138

Table D.10: Same conditions as in Matrix D.9; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**69**	5	1	5		2			10		5								6100
FM	10	**53**	11	10	2	1	1		9	1									2274
FL	1	7	**72**	2					16										1633
CM	13	4	1	**50**		7	3	2	5	5	3	2	1			1		1	9119
CH	1	2			**77**					17									1154
BH	10			5	1	**65**	6		1	4	6								1724
BM				3		1	**76**	6	3	7	3								2540
BL			1	3	2		12	**67**	13	2									1602
Dp	3	1	2	2				2	**89**										3685
Lq				2	9	2	5		2	**74**	3	1							8601
Gl	4					2	3		1	5	**81**	2							2106
Ns				1						1		**88**	3		4				8190
FV										1		3	**72**	14	4	2	1		5169
FU													10	**85**		2	2		7388
SV										1		6	4		**68**	16	1		5492
SU												1	2	3	12	**79**	3		8811
Af													5	6	2	5	**82**		1289
Wh	3		1	2					2	2	3	3	3	3		2	2	**72**	1102

Table D.11: Train set : Clean; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 2 iterations of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**79**	3	1	5		3			5		3								6101
FM	7	**63**	12	7	1	1			7										2274
FL	1	7	**79**	1				1	9										1633
CM	12	4	1	**63**		6	2	2	3	4		1							9090
CH		1			**80**					16									1154
BH	7			3		**81**	2			2	2								1721
BM				3			**80**	6	2	5	2								2538
BL			1	4			8	**75**	9	1									1601
Dp	3		2	2				2	**90**										3685
Lq				2	6	1	3			**83**	2								8585
Gl	2					1	1			3	**91**								2102
Ns												**95**	1		2				8203
FV													**83**	11	3				5222
FU													3	**95**		1			7394
SV												2	3		**86**	6			5572
SU													1	2	5	**89**	2		8995
Af													2	3	1	2	**92**		1291
Wh	1			1									1					**92**	1138

Table D.12: Same conditions as in Matrix D.11; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**77**	4		5		3			7		2								6103
FM	8	**62**	10	9	2				7	1									2274
FL	1	7	**79**	1					9										1633
CM	12	3		**63**		5	2	1	2	4		2						1	9151
CH					**85**					12									1154
BH	8	1		5		**75**	4			2	2	1							1724
BM				3		1	**79**	6	2	6	2								2539
BL			2	4			8	**74**	10	2									1601
Dp	3	1	1	1				2	**90**										3685
Lq				2	7	1	3		1	**81**	2	1							8623
Gl	2					1	1		1	3	**87**	1							2122
Ns										1		**91**	2		3				8233
FV												2	**78**	13	3	1			5240
FU													6	**91**		1			7386
SV												3	5	1	**74**	14			5512
SU													2	3	9	**81**	3		8963
Af													4	6	1	3	**86**		1290
Wh	2			2						1	2	2	1	1				**84**	1134

Table D.13: Train set : Clean; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 3 iterations of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**80**	3	1	4		2			5		3								6102
FM	7	**64**	12	7					6										2274
FL	1	7	**80**	2				1	9										1633
CM	12	4	1	**64**		6	2	2	2	4		1							9098
CH					**81**					16									1154
BH	7			4		**81**	2			2	2								1720
BM				3		1	**81**	6	2	5	1								2539
BL			1	4			8	**77**	8	1									1601
Dp	3	1	2	2				2	**90**										3685
Lq				1	6	1	3			**84**	2								8582
Gl	2					1	1		1	2	**91**								2102
Ns												**95**	1		2				8204
FV													**84**	10	3				5225
FU													3	**95**		1			7394
SV												2	3		**86**	7			5576
SU													1	2	5	**89**	2		9004
Af													2	3	1	1	**93**		1291
Wh				1														**92**	1140

Table D.14: Same conditions as in Matrix D.13; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**77**	4		5		2			7		2								6103
FM	8	**62**	10	8	1	1			7	1									2274
FL	1	7	**80**	2					9										1633
CM	11	3		**65**		5	2	1	2	4		2						1	9145
CH		1			**85**					12									1154
BH	7	1		5		**76**	4			2	1								1722
BM				3		1	**80**	6	2	5	2								2539
BL			2	3			8	**75**	10	1									1602
Dp	3		1	1				2	**90**										3685
Lq				2	7	1	2		1	**81**	2								8621
Gl	2						1		1	3	**88**								2120
Ns												**91**	2		3				8244
FV												2	**79**	12	3	1			5241
FU													7	**91**		1			7389
SV												3	5		**74**	14			5545
SU													2	3	9	**82**	3		8991
Af													4	5	1	3	**86**		1291
Wh	1			3						1	2	2	1	2				**84**	1136

Appendix E

Confusion Matrices - Test set : Clean

Table E.1: Test set : Clean; Static features [Env]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**49**	3	1	8	1	8			14	2	9	2			1				2059
FM	8	**35**	12	14	4		2		18	4		1							867
FL	1	13	**60**	1				1	19			2						1	519
CM	15	6	2	**26**	1	10	7	2	9	9	2	4	5		2				3049
CH	3	2		2	**45**	1	2		2	41									414
BH	14			6	1	**48**	4		3	7	10	2	2		1				525
BM				5		1	**50**	7	8	18	8								918
BL		1	2	8			15	**45**	21	7									559
Dp	4	4	4	5	1		2	7	**69**	1		1							1207
Lq	1			2	12	2	10	2	2	**58**	4	2	2						3378
Gl	9					4	6		2	13	**59**	3	1		1				810
Ns	1			1		2				2	1	**80**	5		6				2765
FV				2		3	1			3	1	3	**56**	18	5	4	4		1744
FU													10	**80**		4	5		2518
SV	2			1		3				3	2	7	5	2	**59**	13	2		1913
SU												2	4	9	10	**70**	4	1	2883
Af													5	4	3	4	**83**		359
Wh	4			2				1	4	2	4	3	3	3	2	10	1	**60**	367

Table E.2: Same conditions as in Matrix E.1; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**53**	3		7	2	7	1		11	2	7	2							2048
FM	14	**31**	11	14	4	2	3	1	13	4		1						1	867
FL	3	12	**58**	2					21										520
CM	13	4	2	**29**	2	7	8	2	5	8	2	5	3	3	2	1		3	3033
CH	2	4		2	**50**	1			1	36			1						414
BH	14			9	1	**44**	6		2	9	9	2	2	1					524
BM				5		2	**53**	6	6	19	6								918
BL		1	2	10	1		18	**41**	17	9									559
Dp	7	5	6	7	1		2	7	**62**	2									1207
Lq				2	9	2	11	2	3	**56**	3	4	2					1	3362
Gl	6					5	6		2	12	**50**	7	3	2	3	2	1	1	785
Ns	2	1	1	2		1	2		2	5	2	**66**	4	1	6			1	2739
FV				1		3	2			3	1	4	**51**	18	5	4	6		1693
FU													13	**77**		3	6		2514
SV	2		1		1	1	1		2	4	2	7	4	2	**54**	14	2		1816
SU											1	2	2	6	14	**66**	5		2715
Af												1	3	8	5	13	**69**		354
Wh	7		2	4		2	2		5	7	6	4	3	10		9	5	**34**	354

Table E.3: Test set : Clean; Static features [Env, Ener]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**51**	3	1	8	2	8			13	2	9	1							2058
FM	9	**34**	13	14	4	1	2	1	16	4									867
FL	2	13	**60**	2					20			1							520
CM	12	5	2	**32**	1	10	7	2	7	9	2	4	4		2			1	3052
CH	3	3		2	**45**		2		2	41									414
BH	13			7	1	**50**	6		2	6	10	2	2		2				525
BM				4		2	**52**	7	8	16	9								918
BL		2	2	7			16	**40**	23	9									559
Dp	5	4	5	6	1		2	7	**68**	1									1207
Lq	1	1		2	10	3	9	2	2	**60**	4	2	2						3379
Gl	8					3	6		2	12	**64**	3	1						810
Ns				1		1				2	1	**83**	5		5				2762
FV				2		2				3	2	3	**60**	18	4	3	3		1743
FU													11	**81**		3	4		2522
SV						2				3	2	7	5	2	**62**	14	2		1915
SU												1	3	5	11	**75**	4	1	2885
Af													3	3	3	6	**85**		359
Wh	4			2					2	1	2	5	3	4		9		**66**	357

Table E.4: Same conditions as in Matrix E.3; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**54**	4	1	8	2	7	1		11	2	7	1							2055
FM	14	**31**	11	16	5	1	2		13	3									867
FL	3	13	**58**	1					20			1							520
CM	11	4	1	**31**	2	8	7	2	5	9	3	6	3	3	1	1		4	3036
CH	2	4		2	**55**	1	1		2	30					1				414
BH	15			9	1	**43**	7		2	7	8	2	2	1					525
BM				5		2	**55**	6	8	15	6								918
BL		1	2	10	1		17	**43**	17	7									559
Dp	7	4	5	7	1		2	7	**62**	1									1207
Lq				3	9	2	11	2	2	**57**	3	4	2					1	3375
Gl	6					5	6		1	11	**53**	7	3	2	2	2	1	1	780
Ns	1		1	2		1	2		2	5	2	**70**	5		6	1		1	2720
FV				1		3	2			2	1	4	**54**	17	6	3	4		1683
FU													13	**77**		3	5		2518
SV	2					1	1		2	4	1	8	4	1	**56**	14	2		1811
SU												2	2	5	15	**68**	4		2752
Af											1		2	4	4	8	**78**		355
Wh	6		1	3			1		6	6	6	6	3	7	1	7	3	**41**	345

Table E.5: Test set : Clean; Static+Dynamic features [Env, Δ-Δ_2 Env]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**67**	4	1	7		4			8	1	6								2059
FM	9	**50**	13	9	3		2		9	2									867
FL	2	13	**71**						13										520
CM	16	5	2	**43**		8	4	2	6	7	2	2							3053
CH					**62**		1		1	32									414
BH	10	1		7	1	**59**	6		2	5	8	1							525
BM				3		2	**69**	7	4	10	4								918
BL			2	5	1		17	**60**	13										559
Dp	5	1	2	4			1	4	**82**										1207
Lq				2	10	1	6		2	**70**	4	1	1						3379
Gl	6					1	2		2	6	**79**	1			1				810
Ns				2						1		**88**	2		3				2763
FV						1				2		2	**64**	19	5	3	1		1752
FU													6	**88**		3	2		2523
SV										1	1	5	4	1	**69**	15	1		1919
SU												1	2	5	8	**79**	2		2913
Af													3	4	4	4	**83**		359
Wh	2								1		2		3	3	1	5		**77**	372

Table E.6: Same conditions as in Matrix E.5; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**64**	4		7	1	3			9	1	7	1							2058
FM	13	**38**	11	14	4	2	2		12	2									867
FL	4	12	**68**	1				1	14										520
CM	13	3	1	**42**	1	7	4	2	7	6	3	3	1	1		2		1	3052
CH		2		2	**64**	1	1		1	27									414
BH	9	1		10	1	**52**	7		2	8	8								524
BM				3		2	**65**	7	4	13	5								918
BL			1	6	1		15	**60**	13	2									559
Dp	5	1	3	4			1	4	**80**										1207
Lq				3	9	2	7	1	3	**64**	5	2	1						3387
Gl	6					2	3		2	7	**69**	4	1			1		1	817
Ns	1			2					1	2	1	**79**	4		5	1			2767
FV						2						4	**61**	18	5	4	2		1749
FU													12	**80**		3	4		2516
SV						1				2	1	7	6		**59**	19	2		1902
SU												1	3	4	14	**72**	4		2865
Af													3	4	6	8	**78**		357
Wh	6	1		2		1		1	3	3	5	2	3	11	1	5	4	**48**	360

Table E.7: Test set : Clean; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**70**	4	1	6		3			8	1	4								2058
FM	9	**50**	14	10	3		2	1	8	2									867
FL	1	12	**72**	1					12										520
CM	15	4	2	**49**		8	4	2	5	6	1	2							3053
CH	1			1	**61**	1	2			33									414
BH	10			7	1	**63**	5		1	5	6								525
BM				3		1	**74**	7	3	8	3								918
BL			2	6	1		16	**62**	12										559
Dp	5	1	3	3			1	3	**83**										1207
Lq				2	8	1	5		2	**73**	4	1							3378
Gl	5					1	2		1	6	**83**								810
Ns				1						1		**91**	2		2				2763
FV										1		2	**68**	19	5	2	1		1755
FU													6	**91**		2	1		2524
SV												4	5	1	**72**	15	1		1917
SU												1	1	3	8	**84**	2		2914
Af													2	2	4	5	**88**		359
Wh	1										2		2	3	1	3		**84**	371

Table E.8: Same conditions as in Matrix E.7; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**65**	5		8	1	3			9	1	6								2059
FM	13	**41**	12	13	3	2	2		10	2									867
FL	3	10	**70**	1					14										520
CM	12	3	1	**45**	1	6	4	2	6	6	3	3	1	1		1		1	3057
CH	1	1			**67**	1	2		2	24									414
BH	9			12	2	**52**	7		2	7	8								525
BM				3		1	**67**	7	5	11	4								918
BL			1	6	1		15	**61**	13	2									559
Dp	5	1	2	4				4	**81**										1207
Lq				3	9	2	7		2	**66**	4	2	1						3389
Gl	6			1		2	2		2	6	**74**	3	1			1			820
Ns				1						2		**83**	4		5	1			2771
FV						2				1		4	**63**	18	5	4	2		1761
FU													12	**82**		2	3		2522
SV										1	1	8	6		**61**	19	1		1923
SU													3	3	15	**74**	3		2907
Af													4	3	3	7	**82**	1	359
Wh	5			4		1			2	2	4	4	4	6		4	2	**60**	367

Table E.9: Test set : Clean; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 1 iteration of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**70**	5	1	6		3			7		5								2059
FM	7	**50**	16	10	3		1	1	8	3									867
FL	1	11	**75**						13										520
CM	15	5	2	**48**	1	8	5	3	4	6		1							3048
CH	1	1		1	**64**	1			1	30									414
BH	9			6	2	**64**	6		1	4	5								525
BM				2			**75**	6	3	8	3								918
BL				5	1		14	**63**	13	3									559
Dp	4	1	2	2				3	**85**	1									1207
Lq				2	10	2	5	1	1	**72**	4	1							3378
Gl	5					2	2			4	**85**								810
Ns												**92**	2		3				2763
FV												3	**71**	17	5	1			1753
FU													5	**92**		2	1		2526
SV												4	4		**73**	15	1		1914
SU													1	3	7	**86**	2		2918
Af													2	1	3	5	**89**		359
Wh	1								1		2		3	3		2		**86**	373

Table E.10: Same conditions as in Matrix E.9; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**66**	5		7		3			10		5								2058
FM	10	**47**	11	12	3	1	2		9	2									867
FL	2	8	**71**	1					16										520
CM	12	5	1	**48**	1	7	4	3	4	6	2	2				1		1	3046
CH	1	1			**70**	1			1	23									414
BH	9			10		**59**	8			5	6								524
BM				2		1	**74**	6	4	8	4								918
BL			1	5			16	**61**	13	2									559
Dp	5	1	2	3			1	4	**83**										1207
Lq				3	9	2	6		2	**70**	4	1							3385
Gl	5					2	2			6	**78**	2	1					1	821
Ns				1						2		**85**	3		4				2760
FV												4	**66**	17	5	3	2		1739
FU													11	**85**		2	2		2523
SV										1	1	6	6		**63**	18	1		1877
SU													3	3	13	**77**	2		2855
Af													4	5	3	8	**79**		358
Wh	4			2						2	4	3	4	6		3	1	**65**	357

Table E.11: Test set : Clean; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 2 iterations of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**73**	4	1	7		3			6		4								2059
FM	8	**51**	16	11	2		1		7	2									866
FL	1	12	**73**	1					12										520
CM	12	4	1	**60**		6	4	2	3	4		2							3047
CH				1	**64**	1				32									414
BH	7			7	1	**72**	4			3	3								525
BM				3		2	**76**	6	3	7	3								917
BL			1	5			11	**70**	10	1									559
Dp	4	1	2	2			1	3	**85**										1206
Lq				2	8	2	4			**77**	3		1						3377
Gl	4						2			3	**88**								810
Ns												**93**	2		2				2761
FV												2	**75**	16	4				1752
FU													4	**93**		2			2526
SV												4	4	1	**80**	8			1914
SU													1	3	6	**87**	2		2911
Af													1	2	3	3	**91**		359
Wh	1										2		2	1		1		**90**	369

Table E.12: Same conditions as in Matrix E.11; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**72**	6		7		3			8		3								2057
FM	8	**54**	12	13	2				6	2									867
FL	2	8	**78**	1					10										520
CM	11	4	1	**60**		5	3	3	2	4		3						1	3061
CH		1		1	**73**					22									414
BH	9			11		**64**	7			4	1	1							525
BM				2		1	**75**	6	3	8	4								917
BL			1	5			13	**69**	10	1									559
Dp	5	1	2	2				3	**85**										1207
Lq				2	9	2	4		1	**76**	3	1							3392
Gl	4					2	1			4	**84**								820
Ns										1		**88**	3		4				2771
FV												2	**71**	17	4	2	1		1763
FU													7	**90**		1	1		2524
SV											1	5	6	1	**67**	16	1		1889
SU													3	3	11	**79**	2		2901
Af													3	6	2	5	**84**		358
Wh	4			2						2	4	3	1	5				**76**	367

Table E.13: Test set : Clean; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 3 iterations of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**76**	4		7		3			5		3								2059
FM	7	**55**	15	10	2				6	3									867
FL	1	12	**76**	1					9										520
CM	13	4	1	**60**		6	3	3	2	4		2							3051
CH					**62**	2				34									414
BH	7			8		**72**	5			3	2	1							525
BM				2		2	**78**	6	3	6	2								918
BL				4			12	**71**	9	1									558
Dp	4	1	2	2				4	**85**										1207
Lq				2	8	1	4			**78**	3		1						3378
Gl	3					1	2			3	**89**								809
Ns												**93**	2		2				2760
FV												1	**75**	16	4				1753
FU													4	**94**		1			2526
SV												3	4		**80**	9			1914
SU													1	3	6	**88**	2		2912
Af													2	2	3	3	**89**		359
Wh											1		3	1		1		**91**	372

Table E.14: Same conditions as in Matrix E.13; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**72**	5		7		2			8		3								2056
FM	9	**55**	13	10	2	1			6	2									867
FL	2	10	**78**	1					9										520
CM	11	3	1	**62**		6	3	2	2	4		2						1	3060
CH		1			**75**					21									414
BH	6			11		**68**	6			4	2								524
BM				2			**75**	6	3	8	4								918
BL			1	5			14	**64**	12	1									559
Dp	5	1	2	2				4	**85**										1207
Lq				2	8	2	4		1	**76**	3	1							3396
Gl	4					2	1			4	**85**	1							820
Ns										1		**88**	4		4				2771
FV												2	**70**	17	5	2	1		1764
FU													8	**89**		1			2521
SV												6	6	1	**67**	16			1902
SU													3	3	10	**80**	2		2910
Af													3	6	1	5	**85**		357
Wh	3			1					1	2	4	2	1	4				**78**	369

Appendix F

Confusion Matrices - Test set : Telephone Channel Simulation

Table F.1: Test set : Telephone channel simulation; Static features [Env]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**2**			1					4	3	3	31			53				2059
FM		**7**	10		2			1	11	6		18			45				867
FL		1	**29**						10			47			12			1	520
CM		1	3	**1**	1		2	4	7	11		19			49				3054
CH					**20**			1	1	38					37				414
BH						**2**			4	11	2	21			57				525
BM							**24**	12	18	30	2				13				918
BL			2				13	**41**	16	9		2			17				559
Dp			6				1	6	**24**	3		22			35				1207
Lq					6		5	3	7	**49**	2	7			20				3379
Gl							8		5	16	**20**	35			14				810
Ns										3		**89**			7				2765
FV										3		42			44	3	2	4	1755
FU												19			54	9	3	15	2527
SV										1		56			**42**				1924
SU												30			66			3	2923
Af												10			75	4	**4**	7	359
Wh								3	3	1		49			29			**14**	374

Table F.2: Same conditions as in Matrix F.1; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**32**	4	2	12		4			3	5		23			10			5	2057
FM	6	**22**	16	10	2	2		1	9	10		9			10			2	867
FL	2	6	**63**					1	12	2		8			4			1	520
CM	7	4	2	**17**	1	4	3	3	5	14		25			7			9	3066
CH					**18**				1	71		2			6				414
BH	11			13		**19**	2		2	16	3	23			6			5	525
BM		1		3		3	**36**	8	11	25	4	4			3				918
BL			2	4			10	**46**	20	12		1			3				559
Dp	6	5	10	4			1	8	**42**	4		7			9			1	1207
Lq				1	4	1	7	2	3	**58**	3	15			2			2	3391
Gl	5			2		1	5		3	7	**12**	45	1		2			16	821
Ns			1						1	5		**82**			3			4	2781
FV				1						4		40	**12**	9	8	2		19	1772
FU										1		20	7	**15**		5		51	2526
SV										4		43	9	2	**20**	3		16	1941
SU										3		26	5	2	7	**8**		48	2926
Af												34	4	11	4	4	**7**	36	357
Wh	2		1	2				1	4	4		23			2	2		**57**	375

Table F.3: Test set : Telephone channel simulation; Static features [Env, Ener]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**5**	1			2				4	3	3	34			46				2059
FM	1	**9**	8		2			1	16	7		14			40				867
FL		2	**30**						22			33			13				520
CM		1	1		1		2	3	6	13		22			48				3054
CH					**20**		1		1	46	1	1			27				414
BH					3	**4**	1		4	9	3	22			54				525
BM							**31**	8	17	30	2				11				918
BL			2				20	**29**	28	9					11				559
Dp			3				2	4	**39**	3		18			28				1207
Lq					5		7	2	4	**53**	2	7			19				3379
Gl	1						8		3	13	**26**	32			14				810
Ns										2		**83**			14				2765
FV										2		17			70	3	2	5	1755
FU												4	1		61	9	7	19	2527
SV												17			**80**				1924
SU												3			91	**1**		4	2923
Af												3			74	6	**15**	2	359
Wh									1	2		32			44			**20**	374

Table F.4: Same conditions as in Matrix F.3; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**42**	5	2	11	2	6			3	4		13			8			3	2059
FM	7	**24**	19	9	4	3	1	1	11	8		5			6			1	867
FL	2	8	**67**					1	13	1		4			3				520
CM	7	4	2	**22**	1	6	3	4	5	13		17			6			10	3070
CH				1	**31**				2	57		1			5				414
BH	13	1		14	1	**25**	3		3	14	3	13			5			5	525
BM		1		3		4	**43**	10	13	18	1	3			2				918
BL			3	4			10	**53**	18	9					1				559
Dp	8	6	11	4	1	1	1	9	**46**	2		3			7				1207
Lq				2	6	2	9	2	3	**56**	2	13			2			2	3396
Gl	7			2		2	8		1	8	**10**	45			2			14	821
Ns									1	5		**83**			3			4	2781
FV				2						5		41	**11**	8	8	1	1	21	1773
FU										2		14	7	**14**		3	2	57	2527
SV										4		47	10	1	**18**	1		14	1941
SU										3		29	7	2	6	**4**		49	2929
Af										1		26	3	10	1	3	**13**	42	359
Wh	2		2	2					3	5		23			2	1		**58**	376

Table F.5: Test set : Telephone channel simulation; Static+Dynamic features [Env, Δ-Δ_2 Env]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**42**	6	2	7	1	1			7	4	7	20			2				2059
FM	9	**39**	21	3	3		1	1	14	7		1							867
FL	1	12	**68**						15			2							520
CM	10	9	3	**18**	1	2	4	4	13	17	1	14			3				3054
CH		1			**43**		1	1	2	50									414
BH	11	2		8	2	**26**	7		5	12	8	17			2				525
BM		1		2			**58**	9	8	17	4								918
BL			2				16	**55**	22	3									559
Dp	6	2	6					4	**74**	2		3							1207
Lq					6		4	1	3	**73**	5	3			2				3378
Gl	4						3		3	9	**65**	11			2				810
Ns										2		**86**			8				2765
FV										3		14	**5**		30	5		39	1755
FU													4		4	4	1	86	2523
SV												11			**61**	2		23	1923
SU												3			23	**13**		61	2920
Af													2		18	37	**6**	36	359
Wh	1							1	3	3	1	13			7	1		**68**	374

Table F.6: Same conditions as in Matrix F.5; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**63**	6	1	6	1	3			10	2	2	4						2	2056
FM	11	**37**	16	8	4	2	2		13	4									867
FL	3	10	**71**						15										520
CM	15	6	2	**31**	1	8	5	2	11	9	1	5						4	3063
CH	1	1		1	**60**		2		1	32									414
BH	12	2		8	2	**44**	10		3	8	6	3						1	525
BM				2		1	**66**	8	6	11	4								918
BL			2	3	1		16	**57**	18	3									559
Dp	6	2	2	3			1	4	**81**	1									1207
Lq				2	9	2	9	1	4	**59**	6	4						1	3394
Gl	9			2		3	4		2	6	**55**	10						7	818
Ns	1			1		1			2	2		**79**	2		2			7	2775
FV						2	1			1		13	**34**	14	4			28	1750
FU												3	4	**30**				61	2517
SV						1				2		21	25	8	**14**	1		24	1913
SU				1						2		10	11	14	4	**4**		55	2892
Af											1	9	6	33		1	**10**	37	356
Wh	4						1	2	5	1	1	7						**73**	369

Table F.7: Test set : Telephone channel simulation; Static+Dynamic features [Env,Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**51**	6	2	12	1	2			7	4	3	11							2058
FM	6	**42**	24	5	4		2	1	10	5									867
FL		10	**77**						10	1									520
CM	7	7	3	**37**	1	3	4	4	8	17		7							3051
CH	1				**52**		2	1		41									414
BH	13	2		13	3	**34**	7		2	12	5	9							525
BM		1		3			**61**	10	5	15	3								918
BL			2				15	**59**	19	3									559
Dp	7	2	7	1			1	4	**74**	2									1207
Lq				1	6		3	1	2	**75**	5	3			1				3379
Gl	4			1		1	3		1	8	**70**	9			2				810
Ns										2		**86**			9			1	2755
FV										2		12	**7**	1	30	2		44	1738
FU													5	**2**	4	3		86	2518
SV												8			**67**	4		20	1891
SU															24	**22**		53	2884
Af													1		25	28	**16**	29	356
Wh										3		10			7			**78**	373

Table F.8: Same conditions as in Matrix F.7; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**65**	6	1	6	2	2			10	1	2	2						1	2059
FM	12	**41**	16	8	4	1	1		13	2									867
FL	3	9	**71**						16										520
CM	14	5	2	**35**	1	7	5	2	9	9	1	4						5	3069
CH	1	1			**67**	1	3	1	2	23									414
BH	11	1		9	2	**47**	10		3	7	4	2						1	525
BM				2		1	**73**	9	6	6	2								918
BL			2	3	1		15	**60**	18										559
Dp	6	1	3	2				3	**83**										1207
Lq				2	9	2	10	1	3	**61**	5	3						1	3396
Gl	8			2		2	5		2	8	**56**	9						7	820
Ns				1		1				2		**82**	2		2			6	2778
FV						2	1			1		15	**34**	13	3			27	1748
FU												2	6	**32**				59	2520
SV										2		26	25	7	**15**			21	1925
SU				1						2		14	12	14	4	**2**		49	2904
Af												9	4	26	2		**18**	39	359
Wh	3			2		1			2	2		9	1					**76**	369

Table F.9: Test set : Telephone channel simulation; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 1 iteration of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**45**	6	3	9		2			6	3	3	15			6				2058
FM	5	**39**	27	3	3			2	8	6		2			4				867
FL		10	**76**						11			1			1				520
CM	6	8	4	**26**	1	3	4	5	9	13		9			12			1	3053
CH					**50**		1	1		43					2				414
BH	9	2		10	1	**29**	6		3	12	5	13			7			1	524
BM							**58**	14	8	15	3								918
BL			3				9	**66**	17	3									559
Dp	5	4	9					4	**67**	3		3			3				1207
Lq				1	6		4	1	3	**76**	3	3			2				3379
Gl	5			1			2		1	8	**72**	8			2				809
Ns										1		**87**			11				2763
FV										2		16	**3**		32	2		42	1727
FU													2	**4**	13	4		76	2470
SV												7			**74**	4		13	1902
SU															29	**22**		48	2906
Af													1		38	10	**22**	26	358
Wh												17			7			**75**	374

Table F.10: Same conditions as in Matrix F.9; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**64**	7	1	7		2			10	1	2	3						2	2059
FM	10	**44**	17	8	2		1	1	11	3								1	867
FL	2	7	**73**						17										520
CM	12	7	2	**37**		5	4	3	7	7		5			1			7	3070
CH		2			**65**		1		1	28									414
BH	10	1		9	2	**52**	10		2	5	3	3						3	525
BM				2			**75**	7	7	4	3								918
BL			2	3	1		14	**60**	18	3									559
Dp	6	1	2	1				3	**85**	1									1207
Lq				1	9	2	10		2	**65**	4	3						2	3394
Gl	6			1		2	3			5	**64**	5						11	821
Ns										1		**87**	2		2			6	2783
FV										1		21	**31**	16	7			22	1773
FU												5	5	**40**	1			48	2521
SV										1		30	19	3	**27**			16	1934
SU										2		16	8	7	13	**3**		49	2906
Af											2	7	5	40	4		**18**	23	359
Wh	1								1	2		8	1	1				**82**	376

Table F.11: Test set : Telephone channel simulation; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 2 iterations of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**47**	6	3	14		2			6	3	2	14			2				2058
FM	4	**46**	25	6	3				7	4		2			1				867
FL		7	**80**						11										520
CM	5	7	3	**44**		3	3	4	5	10		10			5			1	3049
CH		1			**48**			1		47									414
BH	6	1		15		**41**	4		3	8	2	13			4				520
BM				2			**67**	9	5	12	3								918
BL			3	1			13	**62**	18	2									559
Dp	3	2	8	2				4	**75**	1		2							1207
Lq				1	6		4		2	**79**	2	3			2				3379
Gl	5						2		1	7	**75**	7			2				810
Ns												**89**			9				2761
FV										2		18	**8**	2	35	3		30	1686
FU													2	**7**	11	5	1	73	2243
SV												8			**79**	3		8	1878
SU												1			26	**28**		44	2846
Af												1			38	8	**31**	20	358
Wh												12			5			**79**	374

Table F.12: Same conditions as in Matrix F.11; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**67**	7	1	7		2			8	1		4						2	2059
FM	8	**52**	15	7	3	1		1	7	3								1	867
FL	2	7	**78**						11										520
CM	10	6	1	**46**		5	3	2	4	7		5			2			7	3071
CH					**70**		1		1	25									414
BH	8			10		**58**	9			4		4						3	525
BM				1			**77**	7	5	4	3								918
BL				3			9	**67**	18	2									559
Dp	5	1	2	2				3	**86**										1207
Lq				1	8	2	6		2	**71**	3	2						3	3396
Gl	5			1		1	4			5	**70**	4						8	821
Ns										1		**87**	2		2			6	2780
FV												15	**39**	12	6			24	1769
FU												3	5	**37**				55	2519
SV										1		28	22	8	**22**			17	1930
SU										1		12	6	15	5	**3**		54	2915
Af											1	5	7	30	3		**26**	28	359
Wh	1									1		5						**89**	376

Table F.13: Test set : Telephone channel simulation; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 3 iterations of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**54**	5	3	13		2			5	2	2	11			2				2054
FM	6	**43**	29	7	2				6	4		1							867
FL	1	5	**85**						8										520
CM	5	7	4	**48**		3	2	4	6	9		7			4				3046
CH					**52**			1		43									414
BH	8	1		17	1	**44**	4		1	7	2	10			4				521
BM				3			**66**	10	6	10	3								918
BL			3				12	**66**	16										559
Dp	5	1	8	1				4	**77**	1		2							1206
Lq				1	5		3	1	1	**80**	3	2			1				3376
Gl	4						1		1	6	**78**	6			1				809
Ns												**90**			8				2756
FV				1						2	1	16	**10**	2	35	4		29	1660
FU												2	5	**7**	8	7	1	69	2064
SV												8	2		**79**	3		8	1875
SU												2			31	**30**		35	2804
Af													2	1	35	8	**30**	22	357
Wh												13			6			**78**	372

Table F.14: Same conditions as in Matrix F.13; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**67**	7	1	6		2			8		1	3						2	2059
FM	8	**52**	18	7	3	1			7	3									867
FL	2	6	**78**						13										520
CM	10	5	1	**49**		6	3	3	5	6		5			1			5	3071
CH					**71**		1		1	24									414
BH	9	1		9		**60**	8			4	1	4	2					2	525
BM				2			**78**	7	5	5	2								918
BL			1	3			10	**65**	18	1									559
Dp	4	1	2	1			1	3	**87**										1207
Lq				1	8	2	5		1	**72**	3	3						3	3395
Gl	6						3			4	**74**	3						7	820
Ns												**87**	2		2			6	2781
FV												17	**38**	14	7			23	1766
FU												3	4	**37**				55	2522
SV												29	23	6	**23**			16	1929
SU				1						2	1	15	9	15	7	**3**		45	2897
Af											2	6	6	42	3		**22**	19	359
Wh									1	2		4		1				**89**	376

Appendix G

Confusion Matrices - Test set : Room Reverberation Simulation

Table G.1: Test set : Room reverberation simulation; Static features [Env]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**3**			1	1	3			2	2	11	3	13	4	8	19	2	28	2026
FM		**1**							2	2		2	4	11	2	29		43	828
FL			**2**						1				1	18		16		60	501
CM						3	2		3	3	1	7	11	9	5	14	2	37	2927
CH		1			**6**	1	2			22	2	1	2	5	9	20		27	397
BH						**11**	1			3	12	7	18	2	12	13	2	17	518
BM						3	**11**	1	5	5	2	2	1	11	1	28		29	900
BL							1	**2**	3	1				22	1	29		39	543
Dp		1						1	**6**	1			2	17		25		42	1190
Lq					2	4	6	2	2	**14**		4	3	11	6	17		29	3272
Gl	1					6	3			9	**14**	15	7	2	7	7	4	24	777
Ns	1		2	1		1	1		2	2	2	**24**	7	5	5	8		36	2570
FV						2	1		1	4	2	7	**32**	10	9	9	7	12	1700
FU													21	**35**	2	14	14	12	2465
SV	3		1	1	1	2	1		2	7	4	15	10	5	**20**	8	4	14	1887
SU	1								1	1	2	3	11	7	7	**22**	10	32	2867
Af												2	18	8	2	4	**53**	11	352
Wh						1	2		1	2	3	2	3	6	2	9	2	**66**	364

Table G.2: Same conditions as in Matrix G.1; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**4**	1		2	2	6	1		3	5	5	4	1	7		8	2	46	2030
FM	1	**2**	3		1	3	2	3	7	6	2	5		4		7		52	864
FL		2	**8**					2	26	2		2						56	520
CM	1		1	**2**		2	7	2	4	8	5	8	3	14		8	2	32	3039
CH				1	**11**	1	6	4	2	37	2	7		1		3		22	412
BH	3			2		**11**	4	1	1	8	13	5	1	11		6		32	517
BM				1		4	**36**	8	5	11	8	3				8		15	917
BL							21	**23**	10	5	4					8		26	559
Dp	2	2	2	1	1	2	6	14	**24**	4	1	2				2		36	1207
Lq					2	2	17	7	4	**17**	6	6		7		9		22	3352
Gl	1				1	4	10		2	9	**16**	10	1	11		5	4	23	776
Ns	2	2	5	2		2	5	4	7	5	5	**17**		7		2		32	2750
FV	4	1	1	5		6	5	2	4	7	3	12	**8**	20		3	3	15	1706
FU	1			1		2			2	4	1	4	5	**50**		6	3	18	2518
SV	6	2	4	3	2	4	4	3	7	11	5	17	3	9	**3**	4	2	12	1894
SU	3	1	3	1		2	2	1	5	5	4	10	3	20	3	**14**	5	18	2865
Af	2					1			2	6	5	8	10	32		4	**12**	17	355
Wh	3	2	2			3	4	3	3	6	2	6	1	9		7	6	**41**	347

Table G.3: Test set : Room reverberation simulation; Static features [Env, Ener]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**6**	1	1	2	2	7			2	2	11	5	18	1	5	9	3	24	2057
FM	1	**2**	4	2	1	4			7	3	1	1	9		3	20		37	867
FL		1	**16**						8				7		4	24		38	520
CM				**1**	1	5	2		3	4	2	8	12	5	5	10	1	38	3039
CH		1	1	1	**15**	3	2		3	22	2	2	3		6	14		23	414
BH				2		**19**			2	3	13	8	21	1	6	4		18	525
BM						8	**9**	2	13	8	3	2	2		8	21		23	918
BL							1	**3**	11	2			2		16	34		29	559
Dp	3	2	4	1		1		2	**17**	1		2	4		7	26		28	1207
Lq					2	6	5	1	4	**15**	1	5	4	2	5	18		29	3371
Gl	1					6	2		1	10	**14**	16	9	2	6	3	5	25	763
Ns	2		3	1		2	1		4	2	2	**29**	8	1	4	8		30	2732
FV			1	2		2			2	5	2	10	**36**	7	8	8	4	9	1701
FU													29	**30**	3	7	11	17	2521
SV	4	1	2	2	1	2			3	6	4	17	10	2	**21**	10	4	8	1900
SU			1	1					1	1	2	4	11	6	10	**26**	9	25	2883
Af												4	24	7	1	4	**48**	9	358
Wh							2			2	4	5	4	6	2	10	1	**60**	348

Table G.4: Same conditions as in Matrix G.3; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**7**	2	2	2	3	9			5	4	4	3	1	7		8	3	38	2055
FM	2	**4**	4	2	2	5	3	4	14	5	1	2		4		7		41	866
FL		4	**18**					4	42	1								28	520
CM	2		1	**3**	1	5	7	4	6	6	4	6	3	13		7	1	30	3040
CH	2		1	1	**16**	2	8	10	4	31		2		2		1		17	414
BH	3			2		**21**	4	2	2	7	11	2	1	14		5	1	24	524
BM				1		9	**41**	14	8	7	3	2				2		12	918
BL							23	**39**	17	2						3		12	559
Dp	2	5	4	1	2	3	6	18	**35**	3		1				1		19	1207
Lq					3	4	18	10	6	**14**	4	4		7		6		22	3377
Gl	2				1	7	9	1	2	9	**16**	10	2	10		4	4	22	768
Ns	2	3	7	2		4	5	5	11	5	2	**16**	1	7		1		28	2745
FV	4	2	1	5		9	5	2	5	6	2	12	**9**	19		2	2	13	1681
FU	1		1	1		4			2	3		3	5	**51**		5	3	16	2522
SV	6	2	5	2	2	6	4	3	9	11	3	17	4	7	**5**	5	1	9	1879
SU	2	1	3	1		2	1	1	5	5	3	11	2	19	4	**16**	4	17	2864
Af	2			1		3			2	6	4	5	8	29		4	**16**	17	355
Wh	4	1	3			4	5	2	5	5	4	5	2	8		6	4	**42**	334

Table G.5: Test set : Room reverberation simulation; Static+Dynamic features [Env, Δ-Δ_2 Env]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**19**			5		3			4	5	12	2	7	2	7	10	2	21	1969
FM	2	**4**	1	3		2			5	5	3		3	7	3	29	1	30	750
FL		2	**9**											15		12		58	358
CM	2			**4**		4			3	6	3	6	12	6	8	19	1	21	2841
CH					**16**	3	1			38	1		2	4	4	16		10	373
BH	2			5		**26**				7	16	5	6	1	4	13	2	10	499
BM				2		4	**15**	1	1	13	5	2		10		27		18	807
BL							2	**7**	2	2				20		39		26	450
Dp	4	2		3				2	**17**	2				13		18		36	1005
Lq				1	2	3	5	1	2	**26**	2	4	2	9	7	23		12	3163
Gl	2			1		3	1		1	11	**32**	13	3	1	7	13	3	11	781
Ns	4	2	1	9		2	1		3	4		**33**	4	3	3	9		19	2339
FV	2			7		4				6	1	8	**31**	7	7	13	5	6	1592
FU				2									26	**34**	2	19	11	5	2404
SV	5			9		3			2	10	2	15	7	4	**21**	14	2	5	1832
SU	2			4					1	2		4	7	9	15	**42**	4	7	2773
Af										1		3	7	4	3	20	**59**	3	356
Wh	2			1		1	2			6	1	3	4	5	5	11	1	**57**	339

Table G.6: Same conditions as in Matrix G.5; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**14**		1	3	1	5			6	5	4	2	1	9		5	3	40	2034
FM	5	**6**	5	5		5	1	1	8	7	2	2		4		6		42	864
FL		2	**32**	1				2	15									45	520
CM	2		1	**4**		5	5	4	6	6	4	8	6	12		8	1	26	3020
CH	1	1	1	2	**25**	1	4	8	2	29	2	1		2		1		19	413
BH	2			1		**26**	4	2	3	5	8	3	2	11		4	2	25	521
BM				2		6	**50**	12	6	5	5	2						10	918
BL				1		1	23	**40**	14	3						2		13	559
Dp	4	1	2	3		2	4	11	**47**	4						1		19	1207
Lq		1		1	3	3	17	9	6	**18**	3	4		7		6		18	3342
Gl	2					5	11		3	8	**19**	8	3	10	1	4	3	22	795
Ns	5	3	7	3		5	4	4	10	5	1	**18**	2	6		2		25	2745
FV	4	1	1	4		12	5	1	4	5	1	10	**17**	16	3	2	2	12	1670
FU	1					2			2	1		4	13	**51**	1	1	1	20	2481
SV	7	3	3	3	2	7	3	2	8	10	3	18	5	7	**7**	4	2	7	1863
SU	3	1	2	2		2			3	5	2	9	8	18	9	**17**	3	13	2801
Af									1	3	3	9	22	21	6	2	**15**	15	341
Wh	4		1	1		2	4	2	5	5	2	5	3	12		5	5	**44**	353

Table G.7: Test set : Room reverberation simulation; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; TIMIT segmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**23**	1		7		5			5	6	15	2	7	1	4	4	2	17	2058
FM	4	**10**	3	6		6			7	6	4		3	3	2	17	2	25	867
FL	2	8	**13**	1					7	1			2	7		9	1	48	520
CM	3			**5**		4			4	8	5	7	12	5	7	13	1	23	3045
CH	1			1	**20**	2	1	1	1	39	2		3	2	1	9		14	414
BH	2			4		**36**			1	7	15	3	6	2	2	6	2	12	525
BM				3		8	**14**	2	5	14	7		1	4		20		18	918
BL						1	3	**7**	7	4			1	9		35	1	30	559
Dp	6	4	2	2		1		2	**30**	2			1	8		14	1	26	1207
Lq				1	2	4	5	1	3	**31**	2	4	6	6	5	13		15	3376
Gl	2			1		4			1	10	**35**	13	5		5	8	2	11	782
Ns	4	3	3	13		2			4	5		**36**	5	3	2	5		13	2725
FV	2	1		10		5	1		1	7		12	**35**	6	5	5	3	5	1696
FU				3								1	36	**33**	2	5	6	12	2516
SV	6	2	1	11		3			4	9		17	9	3	**22**	10	1	2	1896
SU	2			6					1	2		6	9	7	16	**40**	5	3	2870
Af	1			2						2		4	12	4	5	8	**58**	3	359
Wh	3			1						7	1	6	6	5	3	4		**61**	353

Table G.8: Same conditions as in Matrix G.7; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**18**		1	3	2	6			7	5	4	2	2	8		4	2	35	2058
FM	8	**8**	10	5	1	5	1	1	10	7	1			3		5	1	33	867
FL	1	3	**43**			1		2	23									24	520
CM	2		1	**5**	1	5	5	4	7	7	4	7	6	12		6	1	26	3056
CH	2	2	1	1	**32**	3	5	9	5	22	1			1				15	414
BH	2			2		**30**	5	2	3	5	8	1	1	12		3	2	21	525
BM				2		7	**51**	14	8	5	3							6	918
BL			1	1		2	19	**51**	16	2						1		6	559
Dp	5	1	4	3		1	3	11	**56**	4						1		10	1207
Lq		1		1	4	4	15	8	8	**20**	2	3		7		4		19	3389
Gl	2					5	8		4	9	**19**	8	4	10	2	4	2	23	793
Ns	6	3	8	3		6	5	3	12	5		**18**	2	6				20	2763
FV	4	1	1	4	1	11	5		5	6		11	**19**	17	2	2		10	1701
FU	1			1		2			2	1		4	14	**50**	1		1	19	2512
SV	7	2	3	3	2	7	2	2	8	9	1	19	7	7	**9**	5		5	1884
SU	3	1	3	2		2			4	4	1	11	9	18	11	**18**	1	11	2870
Af				1					2	3	2	9	23	19	5	3	**15**	15	358
Wh	4					3	4	2	5	4	2	5	4	10		5	3	**47**	362

Table G.9: Test set : Room reverberation simulation; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 1 iteration of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**41**	3	1	13		7			5	1	12	2	3	1		1		9	2052
FM	4	**30**	15	8	2	3	2		7	3				6		3		16	860
FL	2	7	**43**					1	6					8		2		29	516
CM	7	3	2	**35**		8	4	2	5	6		5	3	2		3		14	3002
CH					**33**	3	1			50				2		2		4	412
BH	3			9		**54**	4		2	5	7	6	2			1		4	522
BM				3		6	**41**	3	6	12	2	1		4		6		13	913
BL			2	4		1	8	**18**	12	2				10		9		32	557
Dp	8	3	6	3		1	1	2	**44**					6		4		20	1202
Lq				2	5	5	6	1	2	**61**	1	3	1	2	1	3		6	3358
Gl	1					3	4		2	12	**63**	4	2	1	1			3	807
Ns										1		**79**	2		5	1		9	2734
FV										1		7	**58**	11	7	4		9	1708
FU													22	**66**		3	2	7	2509
SV										1		8	5	1	**53**	18	3	9	1887
SU													4	2	9	**72**	4	7	2861
Af													16	14	1	5	**63**		359
Wh													1		1	2		**92**	368

Table G.10: Same conditions as in Matrix G.9; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**41**	4	3	4		12			10	1	3	1		2				17	2058
FM	4	**23**	22	6	2	3	1	3	21	1		2						11	867
FL		3	**54**					1	31									9	519
CM	10	3	2	**14**		14	5	6	10	3	1	6		2		1		20	3061
CH		1	2		**44**	2	1	2	4	39		1						3	414
BH	4			3	1	**56**	8	2	4	2	2	3		2				10	522
BM				1		2	**69**	16	5	2	2								918
BL			1				15	**64**	16	2									559
Dp	5	1	5			1	2	8	**73**									3	1207
Lq					6	3	17	4	3	**49**	3	3				1		8	3389
Gl	2					2	11		1	6	**52**	5	2	5		1	1	10	820
Ns							1			2	1	**76**	1	1				13	2762
FV						1	1			2		12	**33**	34	2	3		9	1733
FU													6	**87**				6	2524
SV	1					2			1	3	3	21	15	5	**25**	12	2	8	1867
SU											1	6	6	12	10	**45**	3	16	2854
Af												1	12	58			**26**	2	356
Wh							2				2	3	1	6		2	2	**78**	368

Table G.11: Test set : Room reverberation simulation; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 2 iterations of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**47**	3		14		9			4		10	2	2					6	2059
FM	8	**36**	14	10	1	5		1	6	3				3		2		9	865
FL	2	9	**54**	3					7					4		2		18	519
CM	6	3	2	**46**		9	3	2	5	5		5	2	2				9	3025
CH				1	**35**	3				53				1				4	414
BH	2			10		**59**	2		2	4	3	8	3					3	522
BM				3		9	**43**	5	4	18	4	2		2		3		6	911
BL			2	10		3	10	**28**	18	2		1		5		4		17	555
Dp	9	3	6	4		2	1	2	**55**	1		1		2		1		9	1207
Lq				3	4	5	4	1	1	**67**	1	3		1	1	2		4	3371
Gl						3	3			11	**70**	3	2		1			2	807
Ns												**84**	1		4			6	2741
FV						1				1		6	**63**	11	8	3		6	1733
FU													23	**70**		2	1	3	2511
SV										1		9	6	1	**60**	11	2	7	1897
SU												1	3	4	11	**71**	5	4	2864
Af													18	14	1		**64**		359
Wh											2	1				2		**92**	372

Table G.12: Same conditions as in Matrix G.11; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**47**	3	3	3		13			7		2	1		2				16	2056
FM	5	**23**	27	6	3	5		2	18	2		1						8	867
FL		2	**63**					1	25									7	520
CM	9	2	2	**19**		17	4	6	7	4		8		2				18	3058
CH			1		**47**	1	1	1	3	41								2	414
BH	5			4		**62**	9	1	3	2	1	2		2				9	522
BM				1		2	**73**	12	6	2	2								918
BL			1				15	**67**	13	1								1	559
Dp	5	2	4	1		2	2	6	**73**									5	1207
Lq				1	5	3	13	3	3	**58**	2	3						7	3389
Gl	2					2	9		1	5	**59**	4	2	4				9	821
Ns							1		1	2	1	**74**	2	1				14	2769
FV						1				2		10	**40**	30	2	2		10	1761
FU													6	**86**				7	2525
SV						1				1	2	19	16	6	**30**	11	1	9	1863
SU											1	4	6	14	9	**46**	3	16	2873
Af													12	58			**25**	3	358
Wh							1			1	2	2	1	4				**86**	372

Table G.13: Test set : Room reverberation simulation; Static+Dynamic features [Env, Ener, Δ-Δ_2 Env, Δ-Δ_2 Ener]; 3 iterations of automatic resegmentation; MEL

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**47**	3		15		8			4		10	1	3	1				6	2058
FM	6	**38**	14	10	1	3	1		7	2			1	5		1		9	865
FL	2	9	**51**	2				1	7					5		1		19	520
CM	6	2	1	**48**		9	2	2	5	4		5	2	3		1		9	3025
CH		1			**33**	3				54		1				1		3	413
BH	3			10		**59**	4		1	4	2	8	2					4	522
BM				4		9	**41**	3	6	15	3	1		3		3		9	913
BL		1	2	9		2	7	**23**	17	2		2		8		3	1	21	556
Dp	8	4	5	4		1		2	**53**	1		1		5		2		13	1206
Lq				3	4	5	3		1	**68**	1	3		2		1		4	3374
Gl	1					3	3			10	**70**	3	2	1	2			2	805
Ns												**84**	1	1	4	1		6	2744
FV												6	**62**	11	9	2		7	1741
FU													23	**69**		2	2	4	2514
SV												8	7	2	**61**	12	2	7	1895
SU												1	3	4	10	**73**	4	4	2877
Af													20	14		3	**63**		359
Wh															1	2		**93**	372

Table G.14: Same conditions as in Matrix G.13; EIH

	FH	FM	FL	CM	CH	BH	BM	BL	Dp	Lq	Gl	Ns	FV	FU	SV	SU	Af	Wh	Total
FH	**45**	4	3	3		13			7		4	1		2				16	2056
FM	6	**26**	25	6	3	4		2	17	2								8	867
FL	1	3	**65**					2	23									6	520
CM	9	3	2	**22**		17	4	6	7	5	1	5		2				18	3060
CH		2	1		**44**	1		2	3	43								3	414
BH	3			4		**67**	7	2	2	2	1	2		1				8	523
BM				2		3	**71**	14	6	2	2								918
BL			2				15	**65**	15	1								2	559
Dp	4	2	5			2	2	6	**73**									5	1207
Lq				1	4	3	14	3	3	**58**	2	3						7	3389
Gl	2					2	9		1	6	**60**	3	2	4				10	820
Ns									1	2		**76**	2	1				14	2772
FV						2				2		9	**43**	29	2	2		10	1767
FU													6	**86**				7	2524
SV						1				2	2	19	16	6	**31**	11	1	10	1872
SU											1	4	7	13	9	**48**	3	13	2867
Af													15	54			**28**	1	358
Wh							1			1	1	2	1	3				**88**	373

References

[1] ACERO, A., AND STERN, R. M. Environmental Robustness in Automatic Speech Recognition. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (April 1990), 849–852.

[2] ALLEN, J. B., AND BERKELEY, D. A. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America 65*, 4 (April 1979), 943–950.

[3] BOGERT, B., HEALY, M., AND TUKEY, J. The Quefrency Alanysis of Time Series for Echoes. In *Proc. Symp. on Time Series Analysis*, M. Rosenblatt, Ed. J. Wiley, 1963, ch. 15, pp. 209–243.

[4] DAUTRICH, B. A., RABINER, L. R., AND MARTIN, T. B. On the effects of varying filter bank parameters on isolated word recognition. *IEEE Trans. Acoust., Speech, Signal Proc. 31* (August 1983), 793–806.

[5] DAVIS, S. B., AND MERMELSTEIN, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Acoust., Speech, Signal Proc. 28*, 4 (August 1980), 357–366.

[6] DENES, P. B., AND PINSON, E. N. *The Speech Chain.* Bell Telephone Laboratories, 1963.

[7] EPHRAIM, Y., WILPON, J. G., AND RABINER, L. R. A linear predictive front-end processor for speech recognition in noisy environments. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (April 1987), 1324–1327.

[8] FURUI, S. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Trans. Acoust., Speech, Signal Proc. 34*, 1 (February 1986), 52–59.

[9] GHITZA, O. Auditory Nerve Representation as a Basis for Speech Processing. In *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. Marcel Dekker,Inc., 1992, ch. 15, pp. 453–485.

[10] GHITZA, O. Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recogntion. *IEEE Trans. on Speech and Audio Proc. 2*, 1 (January 1994), 115–132.

[11] Gillick, L., and Cox, S. J. Some statistical issues in the comparison of speech recognition algorithms. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (May 1989), 532–535.

[12] Gray, A. H., and Markel, J. D. Distance Measures for Speech Processing. *IEEE Trans. Acoust., Speech, Signal Proc. 24*, 5 (October 1976), 380–391.

[13] Harris, D. M., and Dallos, P. Forward masking of auditory nerve fiber responses. *Journal of Neurophysiology 42* (1979), 1083–1107.

[14] Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. Rasta-PLP Speech Analysis Technique. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (March 1992), 121–124.

[15] Itakura, F., and Saito, S. A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and Communications in Japan*, 53A (1970), 36–43.

[16] Jankowski, C. R. A comparison of auditory models for automatic speech recognition. Master's thesis, Massachusetts Institute of Technology, May 1992.

[17] Jelinek, F. The Development of an Experimental Discrete Dictation Recognizer. *Proc. IEEE 73*, 11 (March 1985), 1616–1624.

[18] Juang, B.-H. Speech recognition in adverse environments. *Computer Speech and Language 5*, 3 (July 1991), 275–294.

[19] Kiang, N. Y. S., and Peake, W. T. Physics and physiology of hearing. In *Stevens Handbook of Experimental Psychology*, 2 ed., vol. 1. Wiley, 1988, pp. 277–326.

[20] Kupin, J. *Personal Communication* (1993).

[21] Lamel, L. F., and Gauvain, J. L. Continuous Speech Recognition at LIMSI. *DARPA Continuous Speech Recognition Workshop* (September 1992).

[22] Lamel, L. F., Kassel, R. H., and Seneff, S. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. *Proc. DARPA Speech Recognition Workshop*, SAIC-86/1546 (February 1986), 100–109.

[23] Lee, C.-H., Rabiner, L. R., and Pieraccini, R. Speaker Independent Continuous Speech Recognition Using Continuous Density Hidden Markov Models. In *Speech Recognition and Understanding. Recent Advances*, P. Laface and R. D. Mori, Eds., vol. F 75 of *NATO ASI*. Springer-Verlag, 1992, pp. 135–163.

[24] Lee, C.-H., Rabiner, L. R., Pieraccini, R., and Wilpon, J. G. Acoustic modeling for large vocabulary speech recognition. *Computer Speech and Language 4*, 2 (April 1990), 127–165.

[25] LEE, K.-F. *Automatic Speech Recognition - The Development of the SPHINX-System.* Kluwer Academic Publishers, 1989.

[26] LEUNG, H. C., CHIGIER, B., AND GLASS, J. R. A Comparative Study of Signal Representations And Classification Techniques For Speech Recognition. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (1993), 680–683.

[27] LIM, J. S., Ed. *Speech Enhancement.* Signal Processing Series. Prentice-Hall, 1983, pp. 3–162.

[28] MANSOUR, D., AND JUANG, B.-H. The Short-Time Modified Coherence Representation and Noisy Speech Recognition. *IEEE Trans. Acoust., Speech, Signal Proc. 37*, 6 (June 1989), 795–804.

[29] MERHAV, N., AND LEE, C.-H. On the Asymptotic Statistical Behaviour of Empirical Cepstral Coefficients. *IEEE Trans. Signal Proc. 41*, 5 (May 1993), 1990–1993.

[30] OLIVE, J. P., GREENWOOD, A., AND COLEMAN, J. *Acoustics of American English Speech.* Springer-Verlag, 1993.

[31] OPPENHEIM, A. V., AND SCHAFER, R. W. Homomorphic Analysis of Speech. *IEEE Trans. on Audio and Electroacoustics AU-16*, 2 (June 1968), 221–226.

[32] OPPENHEIM, A. V., AND SCHAFER, R. W. *Discrete-time Signal Processing.* Prentice Hall, 1989.

[33] PAUL, D. B. The Lincoln Robust Continuous Speech Recognizer. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (May 1989), 449–452.

[34] PAUL, D. B. Speech Recognition Using Hidden Markov Models. *The Lincoln Laboratory Journal 3*, 1 (1990), 41–62.

[35] PICKLES, J. O. *An Introduction to the Physiology of Hearing.* Academic Press, 1982.

[36] PORTER, J. E., AND BOLL, S. F. Optimal estimators for spectral restoration of noisy speech. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (March 1984), 18A.2.1–4.

[37] RABINER, L. R. Applications of Voice Processing to Telecommunications. *Proc. IEEE 82*, 2 (February 1994), 199–228.

[38] RABINER, L. R., AND JUANG, B.-H. Hidden Markov Models for Speech Recognition - Strengths and Limitations. In *Speech Recognition and Understanding. Recent Advances*, P. Laface and R. D. Mori, Eds., vol. F 75 of *NATO ASI.* Springer-Verlag, 1992, pp. 3–29.

[39] RABINER, L. R., AND JUANG, B.-H. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

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

[41] Rabiner, L. R., Wilpon, J. G., and Juang, B.-H. A Segmental K-Means Training Procedure for Connected Word Recognition. *AT&T Technical Journal 65*, 3 (1986), 21–31.

[42] Rose, R. *Personal Communication* (1993).

[43] Schroeder, M. R., Atal, B. S., and Hall, J. L. Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America 66* (1979), 1647–1652.

[44] Seneff, S. A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (April 1986).

[45] Seneff, S. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics 16* (1988), 55–76.

[46] Smith, R., and Zwislocki, J. J. Short-term adaptation and incremental responses of single auditory-nerve fibers. *Biological Cybernetics 17* (1975), 169–182.

[47] Soong, F. K., and Sondhi, M. M. A Frequency-Weighted Itakura Spectral Distortion Measure and Its Application to Speech Recognition in Noise. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (April 1987), 625–628.

[48] Stevens, S. S., and Volkmann, J. The relation of pitch of frequency: A revised scale. *American Journal of Psychology 53* (1940), 329–353.

[49] Viswanathan, V., Henry, C., Schwartz, R., and Roucos, S. Evaluation Of Multisensor Speech Input For Speech Recognition In High Ambient Noise. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (April 1986), 85–88.

[50] Viterbi, A. J. Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. *IEEE Trans. on Information Theory* (April 1967), 260–269.

[51] Young, S. The general use of tying in phoneme-based HMM speech recognisers. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.* (March 1992), 569–572.

[52] Zue, V., Glass, J. R., Phillips, M., and Seneff, S. The MIT SUMMIT Speech Recognition System: A Progress Report. *Proc. DARPA Speech and Natural Language Workshop* (March 1989), 179–189.