A COMPARATIVE STUDY OF MEL CEPSTRA AND EIH FOR PHONE CLASSIFICATION UNDER ADVERSE CONDITIONS

Sumeet Sandhu †‡, *Oded Ghitza* †, *

† AT&T Bell Laboratories, Murray Hill NJ
‡ Massachusetts Institute of Technology, Cambridge MA

ABSTRACT

The performance of current large-vocabulary 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 paper compares one such scheme, the Ensemble Interval Histogram (EIH), with the conventional mel cepstral analysis (MEL). These two spectral feature extraction methods were implemented as front ends to a state-of-the-art continuous speech recognizer and evaluated on the TIMIT database (male). To characterize the influence of signal distortion on the representation of different sounds, phone 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. The performance is displayed here as percentage of phones correctly classified. Confusion matrices were also derived from phone classification to provide diagnostic information.

1. INTRODUCTION

Current automatic speech recognition (ASR) systems perform well when trained and tested in similar acoustic environments but their performance deteriorates significantly under adverse signal conditions or in mismatched training and testing conditions. For example, for an alphanumeric recognition task, the performance of the SPHINX system developed at CMU falls from 77-85% accuracy with matched training and testing recording environments to 19-37% accuracy on cross conditions [1]. It is impractical to train for diverse (and often unknown) distortions, therefore it is advisable to make the ASR system more robust. Several techniques for improving robustness have been proposed, including signal enhancement preprocessing, robust distance measures and alternative speech representations.

The EIH (Ensemble Interval Histogram) is an alternative speech representation motivated by properties of the auditory system [4]. 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. This study differs from the previous evaluations [4, 5] of the EIH in several 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 respectively), *phone classification* (no grammar) is performed instead of word recognition, *mixture Gaussian Hidden Markov Models (HMM's)* are used in contrast to the dynamic time warping (DTW) based recognizer or the Gaussian HMM's used in previous experiments, *static, and static and dynamic features* are evaluated separately, and in addition to the average results, the study includes a *breakdown of the average results* into results for different phonetic groups (formed based on the manner of articulation from the phones listed in Table 1) and a qualitative analysis of confusion matrices of these groups. For lack of space, only the conclusions drawn from the detailed analysis are reported here. The overall aim is to compare the performance of the EIH and mel cepstral analysis (MEL) for continuous speech on a state-of-the-art HMM recognizer using a well known database..

2. EXPERIMENTAL FRAMEWORK

The database used is the TIMIT [7], 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, telephone channel speech 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. Classification is performed, instead of recognition, to focus on the front end (and isolate out issues like grammar, phone insertion and deletion that are involved in the recognition process). The aim is to observe the effects of signal distortion on the signal representation and statistical modeling.

2.1. Database

The TIMIT database is divided into training and testing sections with no overlapping speakers; the same division is followed in this study. Out of the 10 sentences per speaker,

* This work was jointly done with Chin-Hui Lee of AT&T Bell Laboratories. Lee's name can not appear in the author list because of the ICASSP three paper limitation.

0-7803-2431-5/95 $4.00

2 are common to all speakers in both testing and training sections and are left out. The remaining 8 sentences per speaker are used in experiment. Only the utterances by male speakers (326 train, 112 test speakers) are used in this study. The hand marked phonetic transcriptions of the speech files provided with the 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 Figure 1. In the table, *Ph* stands for the phone used (47 total), *Wd* shows an example of occurence of the phone, and *Al* lists the TIMIT allophones collapsed with the phone.

Table 1: Set of 47 phones used and their TIMIT allophones

Ph	Wd	Al	Ph	Wd	Al
h#	silence	pau	aa	father	
ae	bat		ah	butt	
ao	bought		aw	bout	
ax	null	ax-h ix	axr	butter	
ay	bite		b	bee	bcl
ch	child		d	day	dcl
dh	then		eh	bet	
el	bottle		em	bottom	
en	button		er	bird	
ey	wait		f	fin	
g	game	gcl	hh	home	hv
ih	bit		iy	beet	
jh	joke		k	key	kcl
l	like		m	mom	
n	noon		ng	sing	eng
ow	boat		oy	boy	
p	pay	pcl	r	red	
s	sea		sh	she	
t	tea	tcl	th	thin	
uh	book		uw	boot	ux
v	very		w	well	
y	yes		z	zoo	
zh	measure		dx	muddy	
nx	winner				

2.2. Speech Recognition System

The Hidden Markov Model (HMM) continuous speech recognition (or classification) framework used in this study is described in detail in [8]. Each speech unit is modeled as a left-to-right 3-state HMM. 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). Context-independent subword unit models are trained using a variant of the segmental k-means algorithm with the given TIMIT segmentation. In the testing phase each speech segment is compared with all phone models using the Viterbi algorithm. Likelihood scores are obtained for the top 1,2 and 3 candidate phones.

2.3. Front Ends

The TIMIT speech files are provided at 16 kHz sampling rate with 16 bit PCM samples. They are first lowpass filtered and downsampled to 8 kHz. The **static features** for the two speech representations are computed as follows.

- 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 inner hair cell firings, and the accumulation of an ensemble histogram as a heuristic for information extracted by the central nervous system [4]. The first stage consists of 85 bandpass filters (similar in bandwidth and distribution to mel filters) spaced from 0-4 kHz. The second stage consists of 5 level-crossing detectors at the output of each bandpass filter. The interval counts are derived from the upward-going level crossings of the input time-waveform, allocated into 128 frequency bins from 0-4 kHz. 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. One EIH frame is obtained every 9.6 ms. The dynamic range of the frame energy is about 0 to -2.0 units of "loudness".

- Computation of MEL

  The mel scale cepstra are computed in a standard manner [3]. The input speech is windowed by a 20 ms long Hamming window every 10 ms, pre-emphasized and passed through the standard mel scale filter bank. The mel filter bank consists of 24 triangular bandpass filters covering the frequency band 0-4 kHz : 10 uniform filters placed linearly from 0 to 1 kHz and 14 variable bandwidth filters placed logarithmically from 1 to 4 kHz. The log energy output of every filter is computed as the integral of the product of the filter and input magnitude spectrum. The outputs of all the filters constitute a mel filter bank vector, from which 12 cepstral coefficients are computed using an inverse cosine transform. The frame energy is normalized 0 to -75 db for the classification system. One MEL frame is processed every 10 ms.

The 12-component vectors obtained from both front-ends are augmented by the frame energy and/or the corresponding **dynamic features** such as delta energy, delta delta energy, delta cepstrum and delta delta cepstrum for different experiments. The delta cepstrum is calculated as a first-order orthogonal polynomial over a finite-length (5) window centered around the current vector [8]. The delta-delta cepstrum is calculated as the difference between the delta cepstra for one frame ahead and one frame behind the current time. Two sets of static features and two sets of static and dynamic features are used : 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).

2.4. Distortion simulations

- Telephone Distortion

  The telephone channel simulation "wire" [6] provides a simulation of several choices of telephone channels and noise, for example, AT&T data or voice channels, phase jitter, 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). The AT&T LC1 characteristic channel is used here; it has a pass-band of 300 Hz to 2600 Hz. Gaussian noise is added to the test sentence, which is then filtered with the telephone channel response.

- Reverberation Distortion

  The room reverberation program 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. 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. The conditions used were a room 10 feet by 11 feet by 12 feet, all six surfaces had reflection coefficients equal to .90, with the speaker at coordinates (1',1',2') and the microphone at (9',8',11'). This room impulse response is convolved with a test speech utterance (sentence) to get the reverberated speech waveform.

3. RESULTS

Top 1 classification results are listed in Table 2. The first row, *Tr*, represents the clean training speech. The other three rows represent three acoustic conditions of the testing speech: *Cl* is clean speech, *Te* is speech through the telephone channel simulation and *Rv* is speech through the room reverberation simulation. The static and dynamic features are listed in columns, where **Env** is the cepstral envelope alone, **Ener** is the frame energy, Δ and Δ_2 are the first and second order time derivatives. Each entry shows the percentage of phones correctly classified as the top 1 candidate.

In Table 3, the top 3 candidates are shown for the same conditions as in Table 2, to give an idea of how the two speech representations would perform in a complete continuous speech ASR system (with lexical and syntactic constraints).

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

Table 2: Correct phone as top 1 candidate, TIMIT phone boundaries

	Static Features				Static and 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 3: Correct phone in top 3 candidates, TIMIT phone boundaries

	Static Features				Static and 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

Conclusions drawn from observation of the average results shown here, and of confusion matrices corresponding to the average results (top 1), are listed next.

4. DISCUSSION

Analysis of the performance of the two front ends, MEL and EIH, yielded the following results :

- MEL outperforms EIH in clean continuous speech, as it does for isolated speech reported in [4, 5]. 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 possible 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 past and 20 ms of speech ahead for a frame rate of 10 ms. Delta-delta cepstrum is also calculated over time frames taken to be uniform for 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 agree-

ment with [5] 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 largest 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.

- Addition of dynamic parameters to the feature vector results in an increase in performance. This is true for all signal conditions.
- 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 worst 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.
- Under the room reverberation distortion, the sounds worst affected for MEL are most of the vowels; for EIH, they are some of the vowels, voiced stops and fricatives. For all feature sets for both MEL and EIH, many sounds are confused very frequently as the whisper sound (*h* as in *h*elp).
- In 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.

5. SUMMARY

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 relative contribution of the static features alone versus that of static and dynamic features was studied, using measurements of average percent correct as well as phonetic confusions (not reported here). 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 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. REFERENCES

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

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

[3] DAVIS, S., 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.

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

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

[6] KUPIN, J. *Personal Communication* (1993).

[7] LAMEL, L., KASSEL, R., AND SENEFF, S. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. *Proc. DARPA Speech Recognition Workshop* (February 1986), 100–109.

[8] LEE, C.-H., RABINER, L., 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.