Proceedings of Workshop on Robust Methods for Speech Recognition in Adverse Conditions,
Tampere, Finland, May 1999, pp. 11-14

Perceptually motivated measures for automatic speech recognition

Oded Ghitza and M. Mohan Sondhi

Bell Labs, Lucent Technologies
Multimedia Communications Research Laboratory
Murray Hill, New Jersey 07974, USA

Abstract

This paper is concerned with developing measures of perceptual dissimilarity (or distance) between speech segments, and utilizing such measures for automatic speech recognition. The speech segments that we use are diphones (i.e., segments from the midpoint of one phoneme to the midpoint of the adjacent phoneme). To derive the perceptual measures, we implement speech recognizers based on some parameterized distance measure, and adaptively adjust the parameters so as to make the recognizers mimic human recognition performance. We show how to use such distance measures to implement a non-stationary state Hidden Markov Model. Finally, we propose a speech recognition system in which midpoints of vowel segments are first detected, and then a non-stationary HMM is used to recognize the diphone sequence between every pair of successive vowels.

1. Introduction

In this paper we will discuss our attempts at speech recognition based on the perceptually-motivated model shown in Fig. 1.

Our hypothesis is that the peripheral auditory system is capable of extracting "acoustic edges", i.e., rapid changes in spectral properties (that occur, for example, at phoneme boundaries in consonant-vowel (CV) and vowel-consonant (VC) transitions). Having detected these edges, it extracts a segment corresponding to the diphone surrounding each transition. This is the segment roughly from the middle of the phoneme preceding the transition to the middle of the phoneme following it. A diphone extracted in this manner is then identified by finding the best matching template from a library of diphone templates stored in memory (presumably acquired during early stages of language acquisition). A model for the perceptual distance used for making this selection is the main thrust of this paper.

In the next Section we discuss the psychophysical experiment which we call the "tiling" experiment ([2]), which establishes the perceptual importance of the diphone. In Sections 3 and 4 we discuss the derivation of two perceptual distance metrics for speech segments of durations roughly corresponding to those of diphones (50-150 milliseconds).

To implement an automatic speech recognizer in terms of such diphones, we postulate that the string of diphones constituting a given utterance is the output of a Hidden Markov Model (HMM) with non-stationary (diphone) states ([4]), whose parameters are obtained from training samples. A given test utterance is recognized as the state sequence that yields maximum likelihood for that utterance. In Section 5 we show how a given distance metric may be utilized for implementing such a non-stationary state HMM. We also present a recognizer which is based on inter-vocalic segments. Here, as a first step, time markers are identified at or near the midpoints of vowel portions of the utterance. The non-stationary HMM is then used to recognize the diphone strings between successive markers.

2. The tiling experiment

In this experiment (described in detail in [2]) we attempt to quantify the relative importance of various time-frequency regions of a diphone, which we call "tiles", by studying the perceptual effects of modifying these tiles

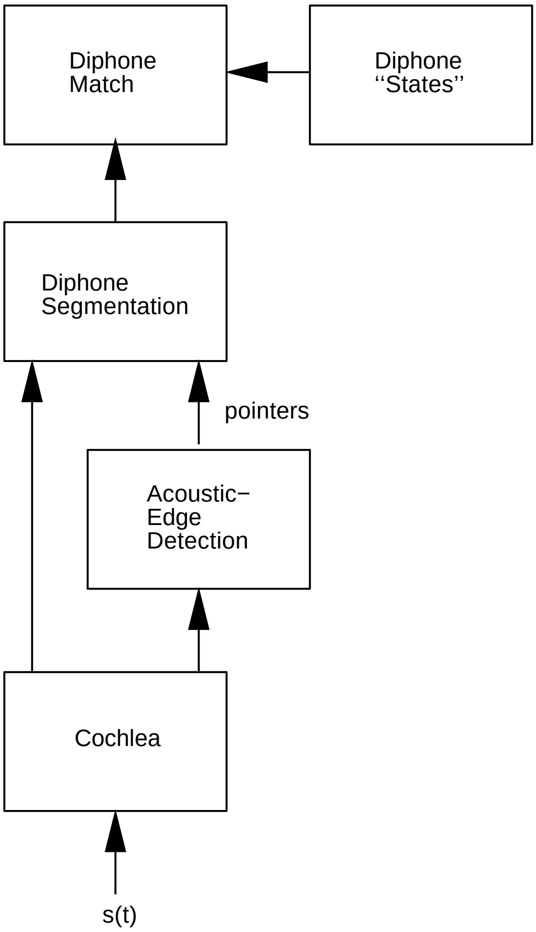

Figure 1: A schematic diagram describing the perceptually-motivated model

one (or several) at a time while keeping the rest of the time-frequency plane unaltered. For the psychophysical paradigm we choose the Diagnostic Rhyme Test (DRT), which was first suggested by Voiers ([6]), and which has been in extensive use for evaluating speech coders.

In the DRT, Voiers uses 96 pairs of confusable words spoken by several male and female speakers. All the words are of the CVC type, and the words in each pair differ only in the initial consonant. (Two modifications of this test have recently been devised. In one modification the word pairs differ in the final consonant. In the other modification the words are of the VCV type, and differ in the medial consonant. We will not discuss these modifications here.)

The target diphones are equally distributed among six phonemic distinctive features (16 word pairs per feature) and among eight vowels. The feature classification follows the binary system suggested by Jakobson, Fant and Halle ([1]). The dimensions are voicing, nasality, sustension, sibilation, graveness and compactness, and the target consonants in each pair differ in the presence or absence of one of these dimensions. An explanation of these attributes, as well as the complete list of words may be found in [2].

The database is used in a very carefully controlled psychophysical procedure. The listeners are well trained and quite familiar with the database, including the voice quality of the individual speakers. A one interval two alternative forced choice paradigm is used. A word pair is selected at random and displayed as text on a screen. One of the words in the pair (selected at random) is next presented aurally, and the subject is required to indicate which of the two words was heard. The procedure is repeated until all the words in the database have been presented. The errors made by the subjects are recorded.

For the tiling experiment, the DRT was conducted on several distorted versions of Voiers' standard database. The details of the signal processing involved in creating those distortions may be found in [2]. Briefly, we divided the time-frequency plane into six non-overlapping regions called "tiles" that cover the target diphone in each pair of words in the DRT, as shown in Fig. 2. We will call the three frequency bands – (0-1 kHz), (1-2.5 kHz), and (2.5-4 kHz) – as bands 1, 2, and 3, respectively.

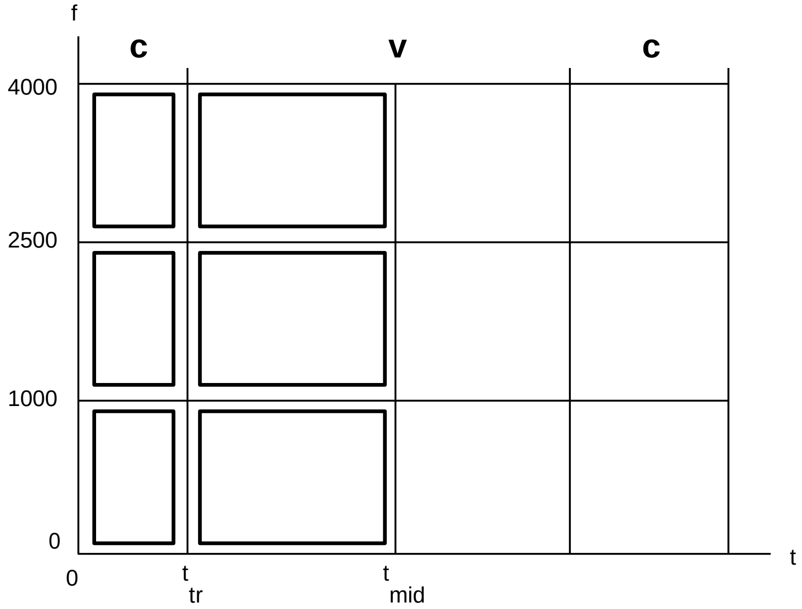

Figure 2: Configuration of the six time-frequency tiles chosen for the tiling experiment

Each distorted database was generated by interchanging a particular tile (or a combination of tiles) between the target diphones of each of the 96 pairs of words in the database. Such an interchange is illustrated in Fig. 3, in which the tile selected is the consonant part of the target diphone between 1 kHz and 2.5 kHz. Several such distorted versions of the database were created. As described in [2], special care was taken to minimize artifacts in the speech signals due to the interchange operation.

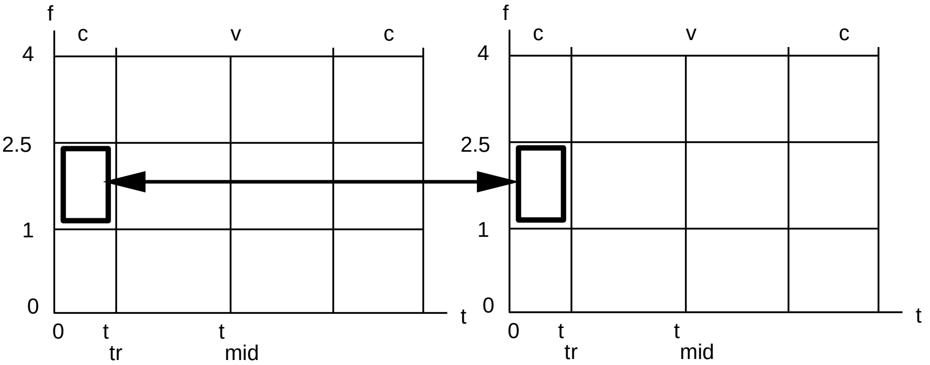

Figure 3: Illustrating the interchange of tiles for a pair of words in the DRT database

A DRT test was performed on the original database as well as on each of these distorted versions. The error for each word pair, for each of these distortion conditions, was recorded for each of 3 speakers and each of 8 listeners. As described in [2], these experiments demonstrated that perceptually, the interchange of the entire diphone in each band is far more dominant than the interchange of the consonant part or the vowel part alone. This establishes the perceptual importance of the diphone.

In view of the dominance of the diphone, the parameters of the perceptual distance were derived using only the original, undistorted, database and the distorted versions corresponding to the interchange of the entire diphone in band-1, band-2 and band-3, respectively. The error patterns of the 8 listeners (for each of the 3 speakers) for these distortions constitute the psychophysical data from which we derive a distance measure, as discussed in the next Section.

3. The perceptual distance

In order to derive the perceptual distance we first *simulate* the DRT by means of an array of recognizers. These recognizers use a parametric form of a distance between two sequences of measurement vectors (e.g., cepstra, EIH ([3]), etc.) derived from the speech waveforms of the utterances being compared. The "perceptual" distance is obtained by adjusting the parameters in such a way as to mimic the human error patterns in the psychophysical experiment described in the previous Section. The procedure is shown schematically in Fig. 4.

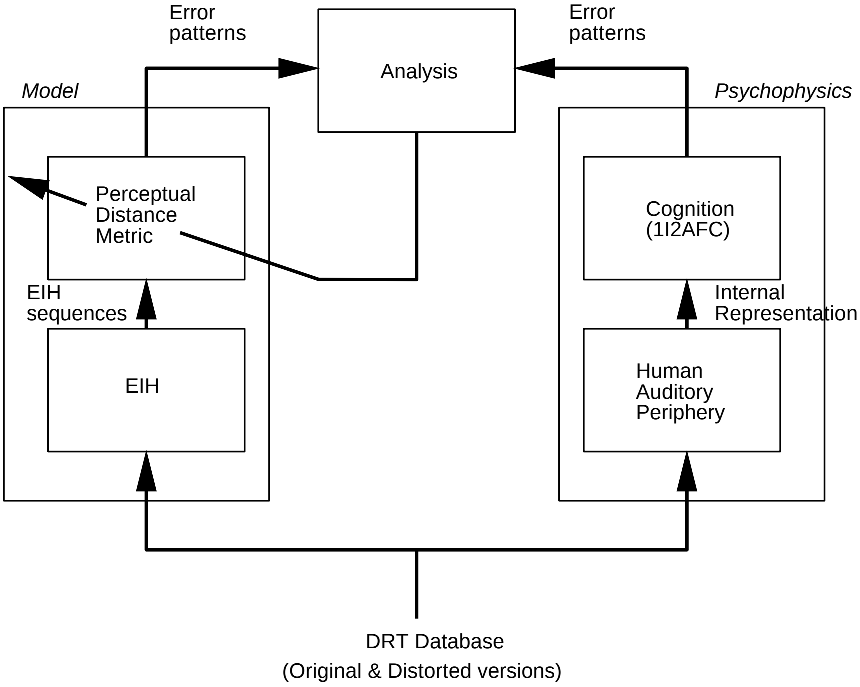

Figure 4: A schematic diagram describing the optimization procedure used to derive the parameters of the perceptual distance measure

Our justification for using an *array* of recognizers is as follows: For each comparison in the DRT, the subject knows that the word to be presented is one of two. We postulate, therefore, that the subject is able to retrieve from memory a recognizer optimized for that pair of words. We allow the simulated "subject" the same capability.

The success or failure of the optimization described by Fig. 4 depends crucially on the way in which the distance between two sequences of observation vectors is parameterized.

In the non-stationary HMM recognizer to be discussed in the next section, each state of the HMM is defined by a template sequence $\mathbf{S}$ and a covariance matrix Σ. The templates and covariance matrices are, in general, different for different states. (How the templates and matrices are derived during training is discussed in Section 5.) For a recognition task one needs to define the squared distance $D^2(\mathbf{S},\mathbf{X})$ between a template sequence $\mathbf{S}$ with observation vectors $\mathbf{s}_n, n = 1, 2, \ldots, N$, and a given test sequence $\mathbf{X}$ with observation vectors $\mathbf{x}_k, k = 1, 2, \ldots, K$. In general $K \neq N$. Also, the length of the template sequence, N, is different for different templates. One way to define D is to first define a squared distance $d^2(\mathbf{s}_n, \mathbf{x}_k)$ between any two given observation vectors. Next, the two sequences are aligned by the usual method of dynamic time warping (DTW). That is, the quantity

$$\Delta^2(\mathbf{S},\mathbf{X}) = \frac{1}{N} \min_{k(n)} \sum_{n=1}^{N} d^2(\mathbf{s}_n, \mathbf{x}_{k(n)}) \qquad (1)$$

is determined. The minimizing function $\bar{k}(n)$ then provides the optimal alignment. Finally, the distance is defined by

$$D^2(\mathbf{S},\mathbf{X}) = \frac{1}{N}\sum_{n=1}^{N} \gamma_n d^2(\mathbf{s}_n, \tilde{\mathbf{x}}_n) \qquad (2)$$

Here, $\tilde{\mathbf{x}}_n = \mathbf{x}_{\bar{k}(n)}$ is the n-th vector in the "warped" sequence, $\tilde{\mathbf{X}}$, and the weights γ_n are used to emphasize or de-emphasize various portions of the diphone. (For example, γ_n can be made large at the transition and smaller towards the beginning and end of the diphone).

The distance d can be defined in a variety of ways. The most common form is

$$d^2(\mathbf{s}_n, \mathbf{x}_k) = (\mathbf{s}_n - \mathbf{x}_k)'\Sigma^{-1}(\mathbf{s}_n - \mathbf{x}_k) \qquad (3)$$

where Σ is the covariance matrix associated with the given diphone state, and the symbol $'$ denotes matrix transpose. The template, $\mathbf{S}$, and the covariance matrix, Σ, together define the state[1] The union of the templates and covariance matrices for all the states is the set of parameters that is adjusted to obtain the perceptual distance, as outlined in Fig. 4. In [5] the definition of distance as given in eq (2) is refined by splitting the observation vectors into 4 sub-vectors, each covering roughly one octave of the signal spectrum. The distance for each sub-vector is defined in a manner similar to that in eq (2) and the *perceptually weighted* sum of these individual distances is defined as the total distance (the weights are obtained as part of the optimization procedure outlined in Fig. 4).

In terms of evaluating the validity of our approach, two questions come in mind. First, how close are the machine error patterns to the human error patterns once the optimum set of parameters have been found? And second, how does the performance of the "optimal" metric — derived by optimizing on "tiling" type of distortions — generalize to other kinds of distortions?

As for the first question, we have shown ([5]) that the simulated DRT experiment produces error patterns that match the *mean* human error patterns to within approximately one standard deviation (across three speakers and eight subjects that constitute our DRT database).

As for the second question, we ran a simulated DRT experiment on a distorted DRT database, using additive Gaussian white noise. Within the DRT simulation, we used Mel-Cepstrum (MEL-CEP) observation vectors associated with the L_2 norm, EIH ([3]) observation vectors associated with the L_2 norm, and EIH observation vectors associated with the perceptual metric (with the optimal parameters). The Table below shows the total number of errors for those DRT simulations and for the human as a function of SNR. We conclude that although the machine performance using EIH with perceptual metric does not match human performance, it is superior to the performance using EIH with L_2 metric (or to the MEL-CEP with L_2 metric).

Total number of switches, in %

	Clean	*30dB*	*20dB*	*10dB*
Human	3	2	3	7
EIH (Perceptual)	5	8	13	24
EIH (L_2)	18	17	21	27
MEL-CEP (L_2)	11	16	25	38

4. The segmental distance

The distance defined in eq (2) is additive, i.e., the distance of $\mathbf{X}$ from $\mathbf{S}$ is just the sum of the distances between corresponding vectors in the two sequences after alignment. Such an additive distance greatly simplifies the recognition algorithm because dynamic programming can be used to find optimal alignment of sequences. However, except for computational complexity there is no basic reason to limit ourselves to such a definition of distance. Recently we have considered a more general definition which we call the "segmental" distance measure. For this definition, the template sequence $\mathbf{S}$ is converted into one long vector, V_s, by concatenating the N vectors

[1] The matrix Σ can, in principle, be allowed to depend on the time index n of the template sequence. In [5] we used two different matrices – one for the first phoneme and one for the second phoneme of the diphone.

$\mathbf{s}_n, n = 1, 2, \ldots, N$. Similarly, the warped observation sequence $\tilde{\mathbf{X}}$ is converted to the stacked vector V_x. The dimension of the vectors V_s and V_x is $N \cdot p$, where p is the dimension of each observation vector. If $\mathbf{\Psi}$ is the $N \cdot p \times N \cdot p$ covariance matrix for the entire template sequence, $\mathbf{S}$, then we may define the distance between V_x and V_s as

$$D^2(\mathbf{S}, \mathbf{X}) = (V_s - V_x)' \mathbf{\Psi}^{-1} (V_s - V_x) \qquad (4)$$

The difficulty with this definition is that it is quite infeasible to estimate the matrix $\mathbf{\Psi}$ because of its size. (The dimension p for our observation vectors is 13, and typically, $N = 40$, making $\mathbf{\Psi}$ a 520×520 matrix). To deal with this problem we partition the template sequence into, say, $M = 4$ sub-sequences, average the observation vectors in each sub-sequence, and concatenate these *averaged* vectors. The concatenated vector now has $M \cdot p = 52$ elements, and the corresponding matrix (which we will call $\mathbf{\Phi}$) is 52×52. We define the distance of a test sequence $\mathbf{X}$ from a template sequence $\mathbf{S}$ as follows: Let J be some partition of $\mathbf{X}$ into M sub-sequences, and let $\mathbf{x}^J$ be obtained by averaging the vectors in each sub-sequence and concatenating the averaged vectors. (We call $\mathbf{x}^J$ the *compressed sequence* of $\mathbf{X}$.) Then the segmental distance $D_{\mathbf{S}}(\mathbf{X})$ is defined as

$$D^2_{\mathbf{S}}(\mathbf{X}) = \min_J (\mathbf{x}^J - \mathbf{s})' \mathbf{\Phi}^{-1} (\mathbf{x}^J - \mathbf{s}) \qquad (5)$$

Here, $\mathbf{s}$ is the compressed sequence of $\mathbf{S}$, with a predetermined (fixed) partition.

5. Implementing a non-stationary HMM

Let us first describe briefly how we derive, from training data, the template $\mathbf{S}$ and covariance matrix $\mathbf{\Phi}$ for a diphone. Suppose the training data for the diphone consists of L tokens, whose observation *sequences* are $\mathbf{X}_l, l = 1, 2, \ldots, L$. Note that these sequences are not all of equal length. Let the lengths of these sequences be $N_l, l = 1, 2, \ldots, L$, respectively. The template $\mathbf{S}$ (and the corresponding compressed template $\mathbf{s}$), and the covariance matrix $\mathbf{\Phi}$ are obtained by using the following iterative procedure:

1. Choose one of the token sequences, say $\mathbf{X}_m$, some initial partition J^m and some initial covariance matrix $\mathbf{\Phi}_m$, and find the segmental distance $D_{\mathbf{X}_m}(\mathbf{X}_i)$ (as defined in eq (5)) together with the *optimal* partition of $\mathbf{X}_i$, J_i^m, for all $i \neq m$.

2. Compute the cumulative distance
$$D_m^2 = \sum_{i \neq m}^{L} D^2_{\mathbf{X}_m}(\mathbf{X}_i) \qquad (6)$$

3. Repeat steps 1 and 2 for each of the tokens in the database.

4. The sequence $\mathbf{X}_m$ for which the cumulative distance is a *minimum* is chosen as the template, $\mathbf{S}$, for the diphone. Using the optimal partitions associated with $\mathbf{X}_m$ (i.e., J^m and $J_i^m, i \neq m$), obtain the compressed sequences $\mathbf{s}^{J^m}$ and $\mathbf{x}^{J_i^m}$, $i \neq m$.

5. Compute $\mathbf{\Phi}_m$, with respect to $\mathbf{s}^{J^m}$:
$$\mathbf{\Phi}_m = \frac{1}{L-1} \sum_{i \neq m}^{L} (\mathbf{x}^{J_i^m} - \mathbf{s}^{J^m})(\mathbf{x}^{J_i^m} - \mathbf{s}^{J^m})' \qquad (7)$$

6. Go to step 1, with the optimal partition (J^m), the template and the compressed template ($\mathbf{X}_m$ and $\mathbf{s}^{J^m}$, respectively) and the covariance matrix ($\mathbf{\Phi}_m$).

In this manner the template and covariance matrix are determined for each diphone. Finally, since HMMs are defined in terms of likelihoods rather than distances, we convert distance to likelihood by assuming Gaussian probability densities. Thus the log likelihood of a sequence $\mathbf{X}$ being a diphone $\mathbf{S}$ is

$$\mathbf{L} = -log|\,\mathbf{\Phi}\,| - D^2_{\mathbf{S}}(\mathbf{X}) \qquad (8)$$

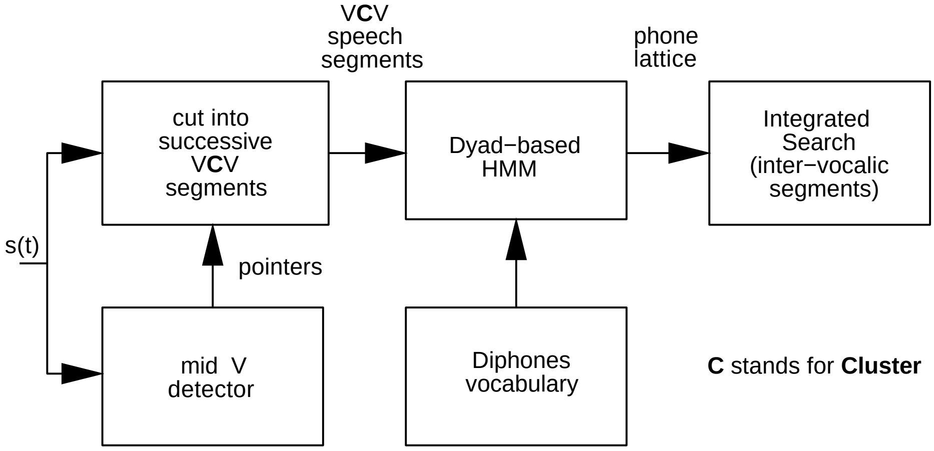

Figure 5: A schematic diagram describing the inter-vocalic ASR system

where D^2 is as defined in eq (5), and | $\mathbf{\Phi}$ | is the determinant of $\mathbf{\Phi}$. To complete the specification of the HMM, we need to specify the state transition matrix. For that we assume that all legitimate transitions are equally probable, where by "legitimate" we mean that the second phoneme of a diphone must be identical to the first phoneme of the following diphone.

We close this section by mentioning a variation of the diphone-states HMM framework ([4]), which we call the inter-vocalic ASR (Fig. 5). Although the human auditory mechanism appears to be able to detect sharp transitions, it is not easy to detect them by machine. Vowels, on the other hand, are much easier to detect, even in the presence of noise. We therefore propose that as a first step we identify time instants at or near the *midpoints* of vowel segments, and then use the diphone-states HMM framework to decode the intervals between these pointers. We have implemented such a system and are currently evaluating its performance. So far, we have compared the inter-vocalic ASR system with a traditional HMM system (with 8 mixtures, diagonal covariances, and context-independent lexicon) in a clean environment, using the 13-th order Mel Cepstrum representation. The database was a large vocabulary, fluent speech, speaker dependent (male), with some 1200 phonemically balanced sentences for training and 50 phonemically balanced sentences for testing. The comparison shows a similar performance in terms of error-rate. Future work will address the performance of inter-vocalic ASR systems which will utilize the perceptually-motivated metrics discussed in this paper.

References

[1] Jakobson, R., Fant, C. G. M., and Halle, M. (1952). "Preliminaries to speech analysis: the distinctive features and their correlates", *Technical Report No. 13, Acoustic Laboratory*, Massachusetts Institute of Technology, Cambridge, Mass.

[2] Ghitza, O. (1993). "Processing of spoken CVCs in the auditory periphery: I. Psychophysics", *Journal of the Acoustical Society of America, 94(5)*, 2507-2516.

[3] Ghitza, O. (1994). "Auditory models and human performance in tasks related to speech recognition and speech coding," recognition and speech coding", *IEEE Trans. on Speech and Audio, 2(1)*, 115-132.

[4] Ghitza, O. and Sondhi, M. M. (1993). "Hidden Markov Models with Templates as Non-Stationary States: An Application to Speech Recognition", *Computer Speech and Language, 7(2)*, 101-119.

[5] Ghitza, O. and Sondhi, M. M. (1997). "On the perceptual distance between speech segments", *Journal of the Acoustical Society of America, 101(1)*, 522-529.

[6] Voiers, W. D. (1983). "Evaluating processed speech using the Diagnostic Rhyme Test", *Speech Technology, 1(4)*, 30-39.