ROBUST PHONEME DISCRIMINATION USING ACOUSTIC WAVEFORMS

Zoran Cvetković[1], *Baltasar Beferull-Lozano*[2], *Andreas Buja*[1]

[1]AT&T Shannon Laboratory, Florham Park, New Jersey, USA
[2] University of Southern California, Los Angeles, California,USA
zoran@research.att.com, beferull@sipi.usc.edu, andreas@research.att.com

ABSTRACT

We present a study of separability of acoustic waveforms of speech at phoneme level. The analyzed data consist of 64ms segments of acoustic waveforms of individual phonemes from TIMIT data base, sampled at 16kHz. For each phoneme, by means of principal component analysis, we identify subspaces which contain a given proportion of the total energy of the available waveforms in time-domain, and also in spectral-magnitude domain. In order to assess separation between phonemes in the two domains, we perform pairwise classification of phonemes on clean data and on data immersed in white additive Gaussian noise up to 0dB signal to noise ratio. While the classification based on spectral magnitudes exhibits high sensitivity to additive noise, the time-domain classification proves to be very robust.

1. INTRODUCTION

The major problem of state of the art algorithms for automatic speech recognition is their high sensitivity to additive noise and environment changes. The first step in all speech recognition algorithms is to represent consecutive speech segments using a set of features which are supposed to facilitate recognition [1]. One of the major purposes of representing speech using the set of features is to reduce the dimension of the space in which the recognition is performed, thus make the task computationally less intensive. Typically, every 10ms feature extraction algorithms generate a new set of 14 cepstral coefficients, which, at 16kHz sampling rate, results in dimension reduction by factor around 10. Another major purpose of using feature vectors is to represent speech in a manner which obliterates its recognition irrelevant variability, *e.g.* speaker related nuances such as pitch, time alignment, etc. It turns out that magnitude spectrum of speech waveforms either abstracts these irrelevancies immediately, or facilitates their elimination by making them explicit. Hence, in the process of feature extraction, each of consecutive speech segments is first represented using its magnitude spectrum, and then, through several more stages, based on heuristic findings on how is speaker related variability reflected in the magnitude spectrum, additional information is removed until a low dimensional feature vector is reached. However, we are not certain if in this process of dimension reduction, and peeling off what seems to be speech component unnecessary for recognition, we are not discarding information which makes speech such a robust message representation, consequently ending up with automatic speech recognition systems which are very sensitive to noise and other forms of degradations.

The motivation for this work is to asses whether the information which is lost when acoustic waveforms are represented by their spectral magnitudes is important for providing better separation of distinct units of speech. For that purpose we consider pairwise classification of phonemes and compare results of classification based on raw acoustic waveforms and classification based on magnitude spectra of acoustic waveforms, for clean data and for data immersed in white additive Gaussian noise up to 0dB SNR. Classification is in both cases performed based on the distances of a particular phoneme realization from the subspaces which describe two candidate phonemes, while the subspaces are identified by means of principal component analysis performed individually on each phoneme using its realizations extracted from TIMIT data base. Note that our purpose here is not to propose a new phoneme recognition algorithm, but only to get some idea about the degree of separation of distinct phonemes in these two representation spaces, and for that reason we chose to use this particular classification method. We found that while classifications in both domains give similar results on clean data, the time-domain classification exhibits strikingly better robustness to noise than the classification based on spectral magnitudes.

2. DATA ANALYSIS

The data we use are 64ms phoneme segments (1024 samples) from TIMIT data base windowed using function w shown in Figure 1. Window function w is designed to satisfy the power-complementary condition

$$\sum_{k=0}^{1023} |W(e^{j(\omega+k\omega_0)})|^2 = 1 \text{ for all } \omega \in (-\pi, \pi)\,, \qquad (1)$$

where $\omega_0 = 2\pi/1024$, so that in the 1024-point discrete Fourier transform of these speech segments all frequencies are equally represented.

For each phoneme we collect all of its realizations in TIMIT, and then retain for the analysis only those which are in listening experiments, when presented in isolation, perceived as realizations of the corresponding phoneme. The decision to consider 64ms segments is made based on our observation that it is very difficult even for a human listener to distinguish well isolated phonemes shorter than 60ms. The window is in the case of phonemes other than stops positioned in the center of a particular realization, whereas in the case of stops the beginning of the window is placed 24ms prior to the closure-release transition. For each phoneme Φ_i we obtain in this manner a set of N_i realizations represented by raw vectors $\phi_{i,j},\ j = 1, 2, \ldots, N_i$ in R^{1024}. All realizations of a particular phoneme are contained within a lower-dimensional subspace of R^{1024}. We identify the subspace of each phoneme Φ_i by considering the eigen-structure of the corresponding covariance matrix

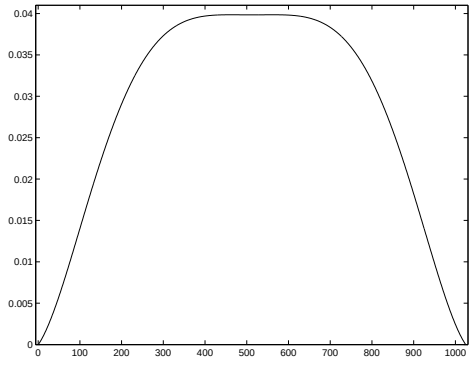

Figure 1: Window function w.

C_i,

$$C_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \phi_{i,j}^T \phi_{i,j} \ . \tag{2}$$

The mean of realizations of each particular phoneme is zero, so when doing the principal component analysis [2] we do not subtract explicitly corresponding mean vectors from realizations $\phi_{i,j}$. The eigen value decomposition of C_i gives a set of eigenvalues $\lambda_{i,1} \geq \lambda_{i,2} \geq \ldots \geq \lambda_{i,1024}$ and the corresponding set of eigen vectors $\varphi_{i,1}, \varphi_{i,2}, \ldots, \varphi_{i,1024}$. The space spanned by eigenvectors $\varphi_{i,1}, \ldots, \varphi_{i,k}$, therefore, contains $\lambda_{i,1} + \lambda_{i,2} + \ldots + \lambda_{i,k}$ of the energy of the data set $\phi_{i,j},\ j = 1, 2, \ldots, N_i$ [2]. Note that vectors $\phi_{i,j}$ are all normalized to unit norm. As an illustration, in Figure 2 and Figure 3 we show first principal components and eigen-value profiles of the front-vowel IY and the unvoiced stop KCL-K, respectively.

We perform also principal component analysis on vectors $\hat{\phi}_{i,j},\ j = 1, 2, \ldots, N_i$ which are magnitudes of the 1024-point discrete Fourier transform of respective waveforms $\phi_{i,j},\ j = 1, 2, \ldots, N_i$. For each phoneme Φ_i, the analysis in the Fourier domain gives principal components $\hat{\psi}_{i,1}, \hat{\psi}_{i,2}, \ldots, \hat{\psi}_{i,1024}$ and the corresponding eigen values $\hat{\lambda}_{i,1} \geq \hat{\lambda}_{i,2} \geq \ldots \geq \hat{\lambda}_{i,1024}$. When doing principal component analysis of spectral magnitudes, we also do not subtract mean vectors from corresponding phoneme realizations, hence, the first principal $\hat{\psi}_{i,1}$ component of each phoneme Φ_i is very close to the mean $\hat{\phi}_i^o = 1/N_i \sum_{j=1}^{N_i} \hat{\phi}_{i,j}$ of the available spectral realizations.

3. PAIRWISE CLASSIFICATION

In order to make an assessment of separability and distances between distinct phonemes in the acoustic waveform domain and in the spectral magnitude domain we conduct the following pairwise classification experiment. For each pair of phonemes, Φ_i and Φ_k, we classify 20% of the realizations $\phi_{i,j}$ and $\phi_{k,j}$ according to the following rule:

$$\begin{array}{lll} \sum_{l=1}^{L} |\langle \phi, \varphi_{i,l} \rangle|^2 \geq \sum_{l=1}^{L} |\langle \phi, \varphi_{k,l} \rangle|^2 & \Rightarrow & \phi \in \Phi_i \ , \\ \sum_{l=1}^{L} |\langle \phi, \varphi_{i,l} \rangle|^2 < \sum_{l=1}^{L} |\langle \phi, \varphi_{k,l} \rangle|^2 & \Rightarrow & \phi \in \Phi_k \ . \end{array} \tag{3}$$

In other words, given an acoustic waveform ϕ, either a $\phi_{i,j}$ or a $\phi_{k,j}$, we classify it as Φ_i if its Euclidean distance from the subspace spanned by the first L principal components of Φ_i is smaller

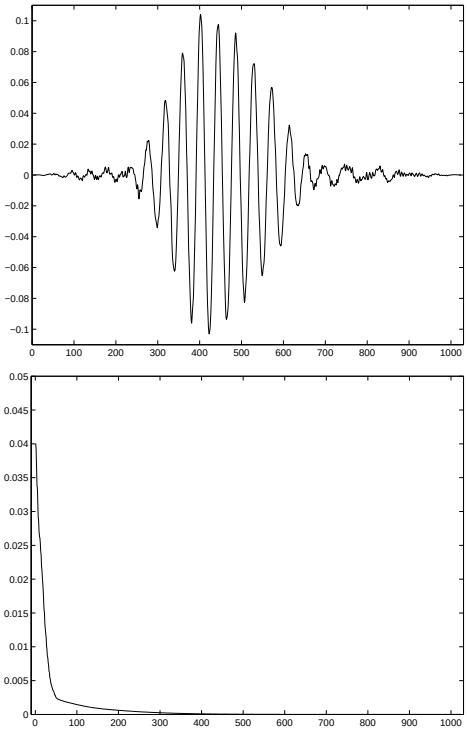

Figure 2: *The first principal component (upper graph) and the eigen-value profile (lower graph) of phoneme* IY.

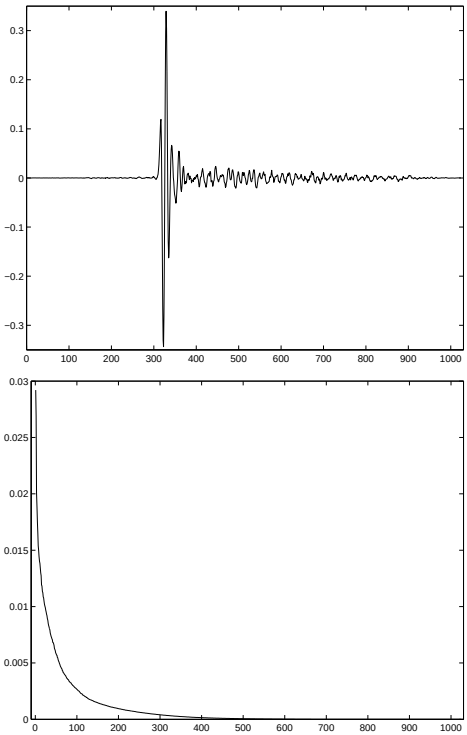

Figure 3: *The first principal component (upper graph) and the eigen-value profile (lower graph) of phoneme* KCL-K.

than its distance from the subspace spanned by the first L principal components of Φ_k; otherwise, we classify it as Φ_k. We calculated empirical classification error probability for clean data and data degraded by white additive Gaussian noise at 10dB SNR and 0dB SNR, and for values of L in the range from one to 300. We also performed analogous pairwise classification using spectral-magnitude representation. Figures 4 - 9 show classification error probabilities of several representative phonemes in function of L, parameterized by noise level. For each representative phoneme, the plotted curves represent error probability of pairwise classification averaged over all other phonemes. The error plots are shown for the vowel IY, diphthong AY, voiced stop BCL-B, unvoiced stop PCL-P, semivowel R, and the fricative Z, which are typical of the plots obtained for phonemes in their respective groups.

The results of our experiment show that there are minor differences in the classification error between acoustic-waveform and spectral-magnitude representations in the case of clean data. Attaining good results for classification using acoustic waveforms normally requires considering spaces of dimensions L greater than 100, whereas classification using spectral magnitudes requires much lower dimensions. However, classification using spectral magnitudes is very sensitive to noise, while classification using acoustic waveforms exhibits remarkable robustness to noise (we considered so far only noise up to 0dB SNR). We can also observe that in the acoustic-waveform domain best classification results are consistently obtained for L between 100 and 150, whereas optimal value of L for classification using magnitude spectra depends on the particular phoneme and on the noise level. Note that error probabilities ultimately start increasing with L and that finally at $L = 1024$ all discrimination ability is lost.

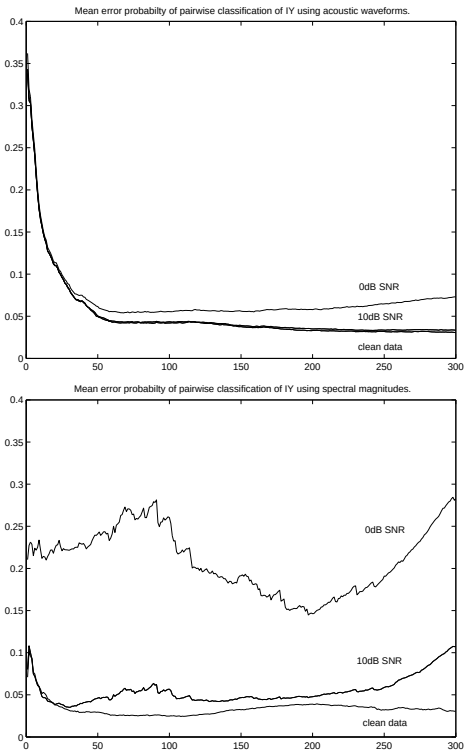

Figure 4: *Mean error probabilities of pairwise classification of phoneme* IY *using acoustic waveforms and spectral magnitudes.*

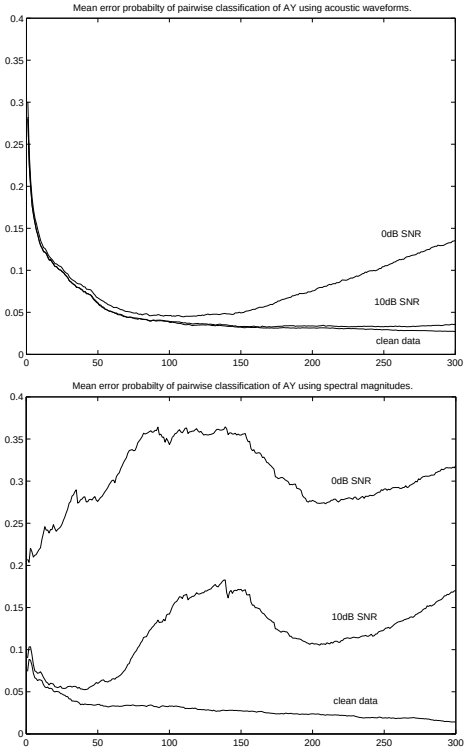

Figure 5: *Mean error probabilities of pairwise classification of phoneme* AY *using acoustic waveforms and spectral magnitudes.*

4. CONCLUSION

The classification experiment reported in this paper indicates that while spectral magnitude of speech segments captures information relevant for phoneme discrimination, it does not fully preserve distances between sections of the representation space occupied by distinct phonemes to make the discrimination task robust to noise. Note again that the sole purpose of our pairwise classification algorithm was to compare distances between phonemes in the two representation spaces and that the acoustic-waveform based classification may not be robust to linear all-pass filtering. However, all-pass filtering acts as a rotation operator, hence, the geometry and distances between the phonemes remain the same regardless of possible all-pass filtering. Our current work is concerned with optimal, not only pairwise, classification of phonemes using acoustic waveforms, in a manner which is robust to linear filtering as well.

Acknowledgment

We would like to thank very much B. Atal and S. Parthasarathy for many instructive and motivating discussions, and T. Stephenson for his tremendous help in setting up the experiments.

5. REFERENCES

[1] L. Rabiner and B. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[2] J.T. Kent and J.M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

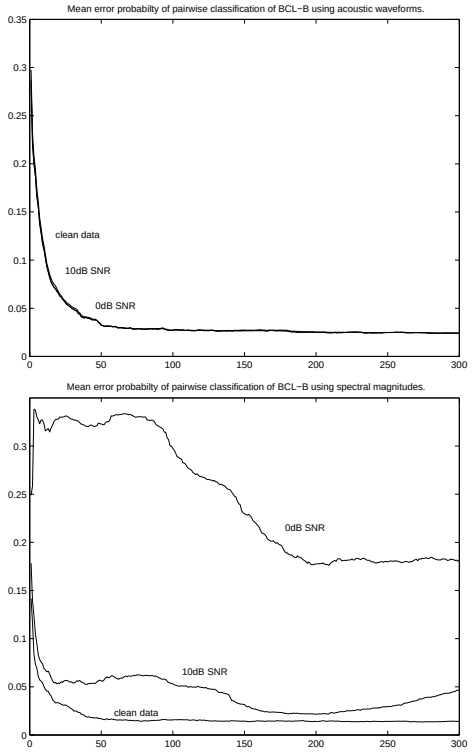

Figure 6: *Mean error probabilities of pairwise classification of* BCL-B *using acoustic waveforms and spectral magnitudes.*

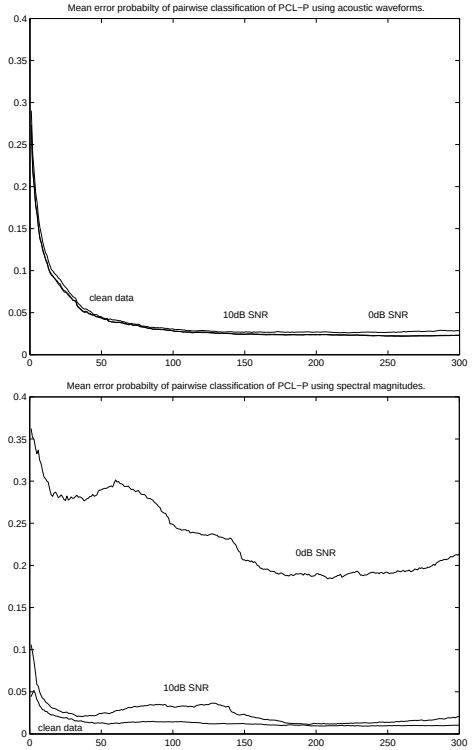

Figure 7: *Mean error probabilities of pairwise classification of* PCL-P *using acoustic waveforms and spectral magnitudes.*

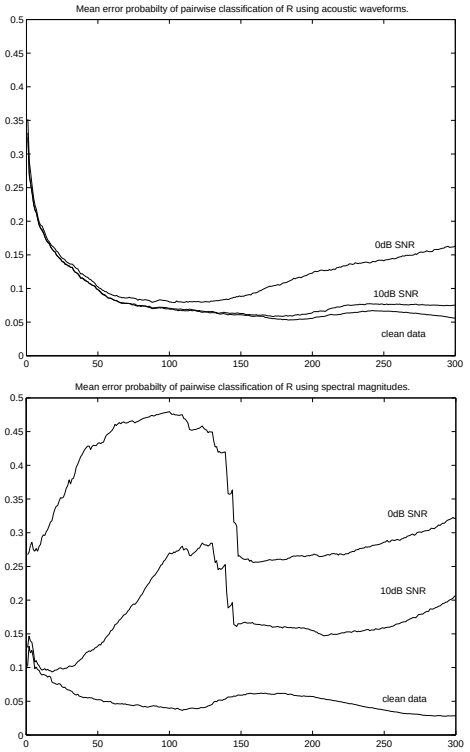

Figure 8: *Mean error probabilities of pairwise classification of phoneme* R *using acoustic waveforms and spectral magnitudes.*

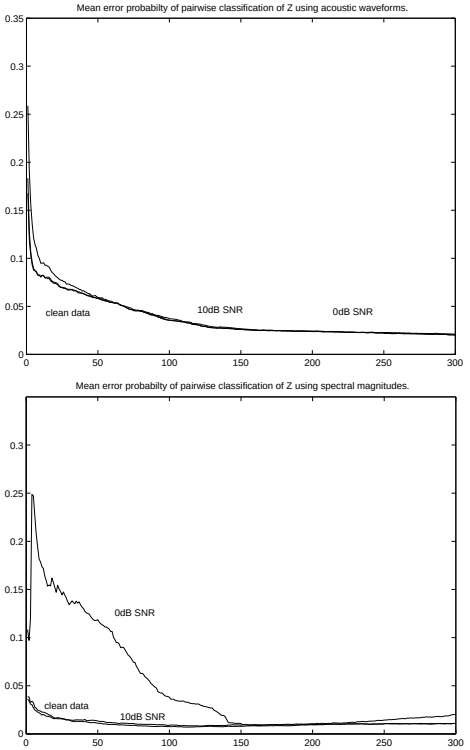

Figure 9: *Mean error probabilities of pairwise classification of phoneme* Z *using acoustic waveforms and spectral magnitudes.*