AN EVALUATION OF USING MUTUAL INFORMATION FOR SELECTION OF ACOUSTIC-FEATURES REPRESENTATION OF PHONEMES FOR SPEECH RECOGNITION

Mohamed Kamal Omar, Ken Chen, Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign,
Department of Electrical And Computer Engineering,
Urbana, IL 61801.

Yigal Brandman

Phonetact Inc.,
Los Altos CA, 94002.

ABSTRACT

This paper addresses the problem of finding a subset of the acoustic feature space that best represents the phoneme set used in a speech recognition system. A maximum mutual information approach is presented for selecting acoustic features to be combined together to represent the distinctions among the phonemes.

The overall phoneme recognition accuracy is slightly increased for the same length of feature vector for clean speech and at 10 dB compared to FFT-based Mel-frequency cepstrum coefficients (MFCC) by using acoustic features selected based on a maximum mutual information criterion.

Using 16 different feature sets, the rank of the feature sets based on mutual information can predict phoneme recognition accuracy with a correlation coefficient of 0.71 compared to a correlation coefficient of 0.28 when using a criterion based on the average pair-wise Kullback-Liebler divergence to rank the feature sets.

1. INTRODUCTION

Many recent speech recognition systems combine multiple speech recognizers to achieve more robustness and better performance. Using different feature streams within each recognizer allows the overall system to benefit from the ability of these streams to reveal complementary information of the original speech signal. Combination of multiple recognizers is consistently reported to outperform baseline systems. In [1], for example, a hybrid speech recognition system based on the combination of acoustic and articulatory information achieved better word recognition results than the baseline systems. Choices of the level of the combination and the best feature streams to be combined together remain as key issues for successful combination. These choices are currently made through intuition and empirical comparison. An approach for selecting the level of the combination based on conditional mutual information of the feature streams given the underlying phoneme identity was suggested in [2]. In [3], the mutual information was used to estimate the distribution of partial phonetic information in the time-frequency plane relative to acoustic landmarks. A framework for defining the theoretically optimal method for feature subset selection was presented by Koller and Sahami in [4]. They proved that for a feature to be unnecessary to model a certain property, it should have a Markov blanket within the complete feature set. However, this optimal feature selection approach is computationally intractable. In [5], the speech signal is modeled as a combination of independent phonological factors. These phonological factors are represented by the best fixed-length subset of the available acoustic feature space. An algorithm was presented there that calculates a good approximation of the acoustic feature subset that has the maximum mutual information with each phonological factor.

In this work, we use the maximum mutual information algorithm described in [5] to derive the best acoustic feature representation for a TIMIT phoneme recognizer. We present an evaluation of mutual information as a criterion for acoustic feature selection for speech recognition. This evaluation is based on comparing the phoneme recognition accuracies achieved using mutual information with those achieved using another distance measure based on Kullback-Liebler divergence.

The next section describes how the mutual information is calculated for different acoustic features and phoneme sets from the training data. An overview of the algorithm used for maximum mutual information selection of the acoustic features is described. Section 3 summarizes the different phoneme recognition experiments and results achieved. Finally, the implication of the results and some future directions of this work are discussed.

2. APPROACH

In this work, the feature stream used in each recognizer is selected from an acoustic feature space formed from lin-

ear prediction cepstrum coefficients (LPCC), MFCC cepstrum coefficients based on FFT, perceptual linear prediction (PLP) cepstrum coefficients, FM coefficients, energy, their deltas, and the average of the deltas. FM coefficients were provided by Phonetact, Inc. The FM coefficients are a nonlinear measure of local spectral compactness, based on the theory of band-limited phase demodulation. The selection of a fixed-length acoustic feature representation for the phonemes is based on maximizing the mutual information between the acoustic feature stream and the corresponding phoneme set.

The mutual information between two random variables X, representing the phoneme, and Y, representing the acoustic feature, is

$$I(X,Y) \quad = \quad \int \sum_{i=1}^{J} P(y|x_i)P(x_i)\log\frac{P(y|x_i)}{P(y)}dy \quad (1)$$

where J is the number of values that the phoneme set can take, x_i is the ith value of the phoneme, $P(x_i)$ is calculated from the training data, $P(y|x_i)$ is modeled by a Gaussian mixture probability density function. The Expectation-Maximization (EM) algorithm is used to calculate the parameters of $P(y|x_i)$ for all phonemes using the training data. The number of distributions within each mixture varies from 2 to 13 depending on the amount of training data assigned to phoneme x_i . Then the mutual information between each acoustic feature available and the set of phonemes under consideration is calculated. The integration in Eq. (1) is approximated by a summation over all possible values that appear in the training data.

To maximize the mutual information between a phoneme set and the subset of acoustic features that are used to model it, we used an algorithm, described in [5], which approximates this maximization within a small number of iterations. The algorithm is guaranteed to achieve a subset with high mutual information but not necessarily the optimal one .

Instead of using the feature set that achieves the maximum mutual information, we select the N-best feature sets generated by this algorithm and select from them the one with the highest recognition accuracy.

We used this algorithm also to select the acoustic-feature representation of phonemes using a criterion based on the average Kullback-Liebler divergence between all phoneme pairs.

The Kullback-Liebler divergence, $L(i,j)$, between the probability density functions $P(y|x_i)$ and $P(y|x_j)$ is

$$L(i,j) \quad = \quad \int P(y|x_i)\log\frac{P(y|x_i)}{P(y|x_j)}dy \quad (2)$$

For two uncorrelated n-dimensional Gaussian probability density functions, $P(y|x_i)$ and $P(y|x_j)$, with mean vectors $\mu_i = [\mu_{i1}\mu_{i2}\ldots\mu_{in}]$ and $\mu_j = [\mu_{j1}\mu_{j2}\ldots\mu_{jn}]$, and diagonal covariance matrices Σ_i and Σ_j respectively, $L(i,j)$ is

$$L(i,j) \quad = \quad \sum_{r=1}^{n} d_r(i,j), \quad (3)$$

where $d_r(i,j)$ is the Kullback-Liebler divergence between the marginal distributions corresponding to the rth feature, and is given by

$$d_r(i,j) \quad = \quad \frac{-1}{2} + \frac{\sigma_{ir}^2}{2\sigma_{jr}^2} + \frac{(\mu_{ir}-\mu_{jr})^2}{2\sigma_{jr}^2} + \log\frac{\sigma_{jr}}{\sigma_{ir}} \quad (4)$$

σ_{ir}, and σ_{jr} are the variances of the rth feature in Σ_i and Σ_j respectively. Then the average pair-wise Kullback-Liebler divergence for each phoneme c with phoneme k, čitekamal, is

$$D(c,k) \quad = \quad \sum_{i=1}^{m_c}\sum_{j=1}^{m_k} W_i L(i,j) + W_i \log\frac{W_i}{W_j}, \quad (5)$$

where W_i is the weight of the ith Gaussian PDF in the mixture of Gaussian's that model phoneme c,
W_j is the weight of the jth Gaussian PDF in the mixture of Gaussian's that model phoneme k,
m_c is the number of Gaussian PDF's in the mixture of Gaussian's modeling phoneme c,
m_k is the number of Gaussian PDF's in the mixture of Gaussian's modeling phoneme k,
and the average Kullback-Liebler divergence based criterion for the entire phoneme set, D, is

$$D \quad = \quad \sum_{c=1}^{J}\sum_{k=1,k\neq c}^{J} P(c)D(c,k), \quad (6)$$

where $P(c)$ is the apriori probability of phoneme c.

3. EXPERIMENTS AND RESULTS

The speech is sampled at 16 KHZ, and preemphasized, then a Hamming window with a width of 20 ms is applied every 10 ms.

The acoustic features are calculated for 4500 utterances from the TIMIT database. These acoustic features are 12 LPC based Cepstrum coefficients, 12 MFCC coefficients, 12 PLP coefficients, 14 FM coefficients energy, and the average of their deltas over periods of 150 msec, and their differences over periods of 5 msec. These acoustic features sum up to 153 features.

The 61 phonemes defined in the TIMIT database are mapped to 48 phoneme labels for each frame of speech.

The algorithm described in [5] is used to select a 39-feature vector from the 153 features available. The indexes of the acoustic features of each category in the final representation of the phonemes is shown in table 1. This acoustic feature representation has a mutual information with the phoneme set that is about 30% over the best single-type acoustic features. The average mutual information achieved its maximum in ten iterations.

Table 1. Indexes of Acoustic Features in Final MMIA Representation of the Phoneme set

Acoustic Type	Indexes of Coefficients	Total Number
Energy	1	1
△ Energy	None	1
Energy Avg.	None	1
LPCC	1,11	12
△ LPCC	1,2,3,4,5,6	12
LPCC Avg.	1,2,3,4, 5	12
MFCC	5,6,8,9,10,11,12	12
△ MFCC	9,10,11,12	12
MFCC Avg.	4,5,6,7,8,9,10,11,12	12
PLP	None	12
△ PLP	5	12
PLP Avg.	None	12
FM	1	14
△ FM	7,8	14
FM Avg.	1	14

Using intermediate feature sets generated while trying to maximize the mutual information, we trained many three-state left-to-right HMM phoneme models. These models are built based on maximum likelihood estimation (MLE) training using the EM algorithm. HMM phoneme models based on MFCC, FM, LPCC, Energy, and their delta's and average delta's were trained for comparison. Tables 2, and 3 show the phoneme recognition accuracy of the maximum mutual information acoustic (MMIA) features compared to MFCC, LPCC, and FM for clean speech and at 10 dB. MMIA has consistent slightly superior performance with and without using bigram language model. Experiments based on all set of features except FM were done to test how well can mutual information predict phoneme recognition accuracy base on certain features. As shown in figure 1, the MMIA representation algorithm can predict the acoustic representation that will give a better recognition accuracy. The correlation coefficient between the rank and the phoneme recognition accuracy is 0.71. Also low values of phoneme recognition accuracy based on an acoustic feature set results in low values of the mutual information between this feature set and phonemes.

The rank of the acoustic feature set that achieves best

Table 2. Phoneme Recognition Accuracy For Clean Speech And At 10 dB With Bigram model

Acoustic Features	Clean Speech	At 10 dB
MMIA	62.91	49.24
MFCC	62.82	47.53
FM	61.64	46.19
LPCC	58.74	45.95

Table 3. Phoneme Recognition Accuracy For Clean Speech And At 10 dB Without Language model

Acoustic Features	Clean Speech	At 10 dB
MMIA	58.59	42.65
MFCC	58.15	41.93
FM	57.98	42.37
LPCC	55.85	40.18

results is second based on mutual information. However, testing the two feature sets ranked first and second based on mutual information in noisy environment (at 10 DB with additive white Gaussian noise), their phoneme recognition accuracies are 49.24% and 48.78% respectively compared to 47.53 for MFCC. These results suggest that the correlation coefficient of the rank and phoneme recognition accuracy improves in noisy environment.

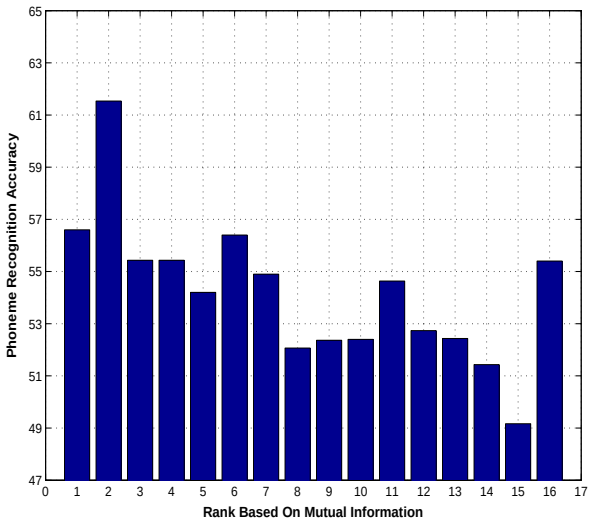

Fig. 1. Phoneme Recognition Accuracy Of Feature Sets selected Based on Mutual Information

The same algorithm is used to select an acoustic feature representation maximizing the criterion based on average Kullback-Liebler divergence, D, between all phoneme pairs

in the phoneme set.

As shown in figure 2, the average Kullback-Liebler divergence as a criterion for feature selection is not as good as mutual information in predicting the acoustic representation that will give a better recognition accuracy. The correlation coefficient between the rank, in this case, and the phoneme recognition accuracy is only 0.28. The rank of the acoustic-feature representation that gives the best phoneme recognition accuracy is 6 based on maximum average Kullback-Liebler divergence criteria. Low values of phoneme recognition accuracy based on an acoustic-feature set do not necessarily result in low values of the average Kullback-Liebler divergence among phoneme models based on this feature set.

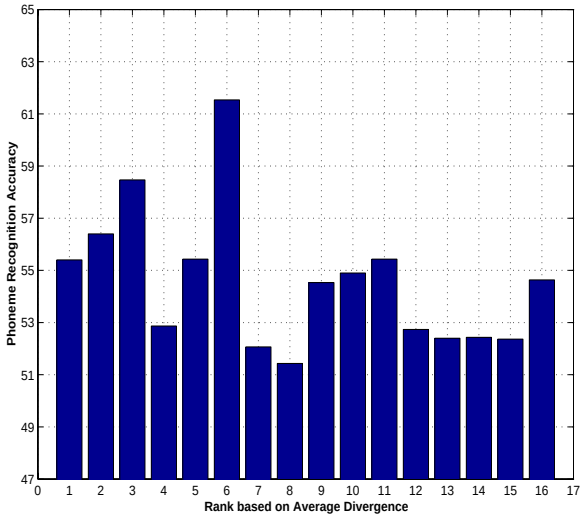

Fig. 2. Phoneme Recognition Accuracy For Feature Sets Selected Based on Average Divergence

4. DISCUSSION

The maximum mutual information feature selection approach introduced in this paper increases the average mutual information between phonemes and their acoustic-features representation by 30% more than the best single category (LPCC, MFCC, FM, or PLP) representation of the same length. These results were achieved in ten iterations of the algorithm, so the time and computational requirement are negligible compared to trying to achieve the optimal maximum mutual information feature vector even over a moderate set of acoustic features. This provides a promising approach for speech recognition systems based on combination of different acoustic features. This approach increases robustness of the speech recognition system and can be easily combined with adaptive techniques to get better recognition accuracy in noisy environments. Due to approximations in the assumed probabilistic model of phonemes, an increase in the mutual information between phonemes and their acoustic feature representation estimated using these models does not guarantee an increase in the recognition accuracy. However, the results we achieved prove that it's a good approximation of the phoneme recognition accuracy that could be achieved using certain acoustic features. This allows testing different acoustic-feature combinations using this approach before even designing the recognizer, and then selecting few possible combinations based on our approach. A small number of recognition experiments is enough to select the best acoustic representation that should be used in modeling phonemes in the recognizer. An important advantage of this approach is that it can be easily generalized to use the same probabilistic model used in any recognition system. It can be applied also to any speech unit not necessarily phonemes.

5. ACKNOWLEDGMENT

This work was supported in part by a gift from Phonetact, Inc.

6. REFERENCES

[1] Katrin Kirchhoff, Gernot A. Fink, and Gerhard Sagerer, "Conversational Speech Recognition Using Acoustic And Ariculatory Input" *IEEE Proceedings of ICASSP*, Istanbul, 2000.

[2] Daniel P. W. Ellis, Jeff A. Bilmes "Using Mutual Information To Design Feature Combinations" *Int. Conf. on Spoken Language Processing*, 2000,vol. 3, pp. 79-82.

[3] Mark Hasegawa-Johnson "Time-Frequency Distribution of Partial Phonetic Information Measured using Mutual Information" *Int. Conf. on Spoken Language Processing* , Beijing, 2000, pp. 133-136.

[4] Daphne Koller, Mehran Sahami "Toward Optimal Feature Selection" *Proceedings of the 13th International Conference on Machine Learning (ML)*, Bari, Italy, July 1996, pp. 284-292.

[5] M. Kamal Omar, Mark Hasegawa-Johnson "Maximum Mutual Information Based Acoustic-Features Representation of Phonlogical Features For Speech Recognition" *IEEE Proceedings of ICASSP*, Florida, 2002.

[6] Mohamed Kamal Omar, "Phonetic Segmentation of Arabic Speech for Verification Using HMM," M.Sc. thesis, *Cairo Univerity*, Egypt, January 1999.