MAXIMUM MUTUAL INFORMATION BASED ACOUSTIC-FEATURES REPRESENTATION OF PHONOLOGICAL FEATURES FOR SPEECH RECOGNITION

M.Kamal Omar and Mark Hasegawa-Johnson

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

ABSTRACT

This paper addresses the problem of finding a subset of the acoustic feature space that best represents a set of phonological features. A maximum mutual information approach is presented for selecting acoustic features to be combined together to represent the distinctions coded by a set of correlated phonological features. Each set of phonological features is chosen on the basis of acoustic phonetic similarity, so the sets can be considered approximately independent. This means that the output of recognizers that recognize these sets independently using the acoustic representation achieved by an algorithm presented in this paper can be combined together to increase efficiency and robustness of speech recognition systems. The mutual information between the phonological feature sets and their achieved acoustic representation is increased by up to 220% over the best single-type acoustic representation in the feature space of the same length.

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 main issues for successful combination that 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. In this paper, 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. A framework for defining the theoretically optimal method for feature subset selection was presented in [4]. It proves 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. So we present an algorithm that calculates a good approximation of the acoustic feature subset that has the maximum mutual information with each phonological factor.

The next section describes the phonological features used and the values that can be assigned to them. and how mutual information can be calculated for different acoustic features and phonological sets from the training data. In section 3, the algorithm used for maximum mutual information selection of the acoustic features to model each factor of speech is described. Section 4 summarizes the different 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 within each recognizer is selected from an acoustic feature space formed from LPC cepstrum coefficients, MFCC cepstrum coefficients based on FFT, PLP cepstrum coefficients, energy, their deltas, and their averages. The selection of a fixed-length acoustic feature representation for each phonological features set is based on maximizing the mutual information between the acoustic feature stream and the corresponding phonological feature set. An algorithm that tries to approximate this maximization within a small number of iterations is described in the next section.

Voicing, manner of articulation, place of articulation,

and duration are the main aspects of the speech signal that are selected as factors to be modeled in our system. This selection satisfies two main requirements: that these factors are enough to discriminate among all phonemes of English, and these factors can be assumed to be independent. Each factor can be assigned one of a set of values which is shown in table 1

Table 1. Phonological Factors of Speech and Their Values

Phonological Factor	Values
Voicing	voiced, unvoiced, silence
Manner of Articulation	vowel, nasal, fricative, stop, glide, silence
Place of Articulation	17 combinations of binary features: (round, anterior, distributed,lateral, low, high, back)
Duration	tense/strident, lax/ nonstrident, reduced/flap

These values are chosen based on the set of distinctive features given by K. Stevens in [5] such that all distinct configurations of different phonemes can be identified.

2.1. Maximum Mutual Information Estimation

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

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

where J is the number of values that the phonological set can take, x_i is the ith value of the phonological set, $P(x_i)$ 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 these distributions 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 this specific value of the phonological feature that the probability density function models. Then the mutual information between each acoustic feature available and the phonological feature 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 phonological set and the subset of acoustic features that are used to model it, we need an intractable amount of computation to test all possible subsets of the acoustic features we have. Hence, an algorithm is developed that guarantee to achieve a subset with high mutual information but not necessarily the optimal one.

3. FEATURE SELECTION BASED ON MAXIMUM MUTUAL INFORMATION

This section describes an algorithm for selecting a vector of M acoustic features that provide a relatively large amount of information about some phonological factor r. First, Eq. (1) is evaluated for every acoustic feature individual, to find the mutual information between each individual feature and the acuostic factor. Second, the M features with the highest individual mutual information scores are combined to form an M-dimensional initial feature vector V_r describing the phonological factor r. Then, Gaussian mixture models of the different classes of speech based on this phonological feature are built using this feature vector. Then, the average mutual information of the phonological feature set with the corresponding acoustic features vector are calculated. Also, the mutual information between each acoustic feature used and the phonological features set is calculated based on the marginal probability density function of this feature. The acoustic features are ordered based on this mutual information values. The worst F features are replaced by the same number of features from the ordered list of features based on their individual mutual information with the phonological feature. The new Gaussian mixture models based on the new feature vector are calculated and the process repeats. When the average mutual information decreases, the F worst features are replaced with the best $\frac{F}{2}$ features that were removed in previous stage and $\frac{F}{2}$ acoustic features from the ordered list based on mutual information between phonological feature and the individual acoustic feature.

The algorithm can be summarized as follows:
For $r = 1$ to H, where H is the number of phonological feature sets, repeat the following steps:

1. For $i = 1$ to N_r,
   Calculate the apriori probabilities of different values of phonological features, $P(x_{ri})$. Where N_r is the number of values that the rth phonological feature can have.

2. For $i = 1$ to N_r,
   For $a = 1$ to J,
   Using EM algorithm, build a Gaussian mixture model of the conditional PDF of the acoustic feature given certain value of the phonological feature, $P(y_a|x_{ri})$, where J is the total number of acoustic features under test.

3. For $i = 1$ to N_r,
   For $a = 1$ to J,

Calculate the mutual information between the phonological feature and the acoustic feature, $I(X_r, Y_a)$, as

$$I(X_r, Y_a) = \sum_{k=1}^{O} \sum_{i=1}^{N_r} P(y_{ak}|x_{ri}) P(x_{ri}) \log \frac{P(y_{ak}|x_{ri})}{P(y_{ak})},$$

where O is the number of frames in the training data.

4. Order the acoustic features in descending order based on mutual information calculated in previous step and save the ordered list L_r.

5. Initialize the acoustic feature vector V_r with the top M features in L_r.

6. While L_r is not empty, do the following steps:

   (a) For $i = 1$ to N_r,
   Using EM algorithm, build a Gaussian mixture model of $P(V_r|x_{ri})$.

   (b) For $i = 1$ to N_r,
   For $a = 1$ to M,
   Calculate $I(X_p, V_{ra})$ as

$$I(X_r, V_{ra}) = \sum_{k=1}^{O} \sum_{i=1}^{N_r} P(v_{rak}|x_{ri}) P(x_{ri}) \log \frac{P(v_{rak}|x_{ri})}{P(v_{rak})}$$

   (c) Sort the features in V_r in descending order.

   (d) Calculate the average mutual information

$$I(X_r, V_r) = \frac{1}{M} \sum_{i=1}^{M} I(X_p, V_{ri})$$

   (e) Remove the worst F features from the list of features in V_r

   (f) Compare the value of $I(Xr, Vr)$ with its value in previous iteration. If more, add the next best F features from L_r to V_r. If less, add the next best $\frac{F}{2}$ features from L_r to V_r and the best $\frac{F}{2}$ features removed from V_r in previous iteration

7. End

4. 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 3300 utterances from the TIMIT database. These acoustic features are 12 LPC based Cepstrum coefficients, 12 MFCC coefficients, 12 PLP coefficients, energy, and their averages over periods of 150 msec, and their differences over periods of 5 msec. These acoustic features sum up to 111 features.

The 61 phonemes defined in the TIMIT database are mapped to values of the phonological features, and hence the phoneme labels are mapped to phonological features labels for each frame of speech.

The algorithm described in the previous section is used to select a 39-feature vector from the 111 features available to represent each phonological feature. The number of acoustic features of each category in the final representation of each phonological factois shown in table 2.

High mutual information between an acoustic feature and the phonological feature does not necessarily imply that it will have high mutual information with the phonological feature when included in a certain acoustic feature vector. This is due to the correlation between the features in the vector. However, our experiments show that more than 70% of the features that end up in the final representation of the phonological feature were on the top 39 features of high mutual information with the phonological feature based on initial individual features probability density functions. That's why the initialization of the feature vector with these features decreases the time and amount of computation required to get good results.

Table 2. Number of Acoustic Features in each MMIA Representation of the Phonological Factors

Acoustic Type	Voicing	Duration	Manner	Place
Energy	1	1	1	1
△ Energy	1	1	1	1
Energy Avg.	1	0	1	0
LPCC	2	3	2	5
△ LPCC	0	6	4	5
LPCC Avg.	0	9	7	2
MFCC	3	3	2	3
△ MFCC	5	0	5	6
MFCC Avg.	12	12	5	8
PLP	2	2	3	5
△ PLP	6	2	6	2
PLP Avg.	5	0	2	1

As shown in figure 1, the maximum mutual information acoustic (MMIA) representation achieved by our algorithm came out with an increase in mutual information of about 100% in the case of voicing and 220% in the case of duration. The average mutual information achieved its maximum in five iterations in the case of voicing and in eight iterations in the case of duration. Although the achieved feature vector is not necessarily the optimal over the avail-

able features, it has an average mutual information that is two to three times better than the best feature representation with the same feature vector length.

Also, shown in figure 2, the maximum mutual information acoustic (MMIA) representation achieved by our algorithm came out with an increase in mutual information of about 9% in the case of place of articulation and 2.5% in the case of manner of articulation. The average mutual information achieved its maximum in eight iterations in the case of place of articulation and in ten iterations in the case of manner of articulation .

The small improvement in average mutual information in the latter case compared to improvements in figure 1 is due to using acoustic features that are calculated to model the vocal tract not the source of excitation or the phoneme duration.

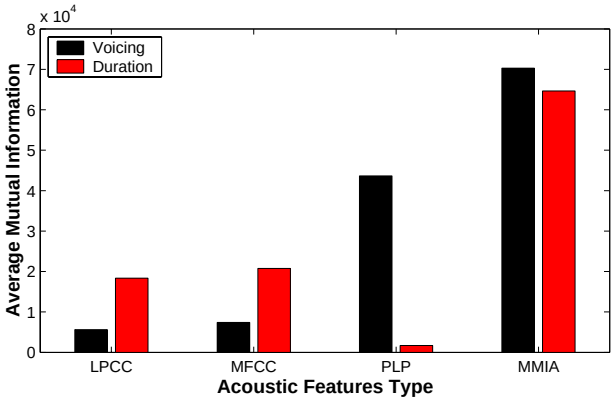

Fig. 1. Average Mutual Information Between Voicing, Duration And Their Acoustic Representation

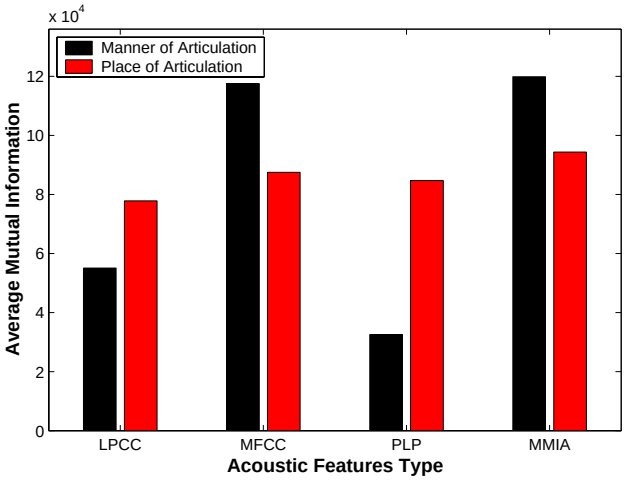

Fig. 2. Average Mutual Information Between Place And Manner of Articulation And Their Acoustic Representation

5. DISCUSSION

The maximum mutual information feature selection algorithm introduced in this paper increases the average mutual information between phonological features and their acoustic-features representation by two to three times for voicing and duration more than the best single category (LPCC, MFCC, or PLP) representation of the same length, and by 2.5% to 9% for place and manner of articulation. These results were achieved in five to twelve 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 dimension of acoustic features space. This provides a promising approach for speech recognition systems based on combination of speech recognizers based on different representations of the speech signal. The choice of categorizing the phonological features into four categories: voicing, place of articulation, manner of articulation, and duration allows the combination of these phonological features recognizers at the output level, as they can be better assumed to be independent than the outputs from phoneme recognition systems with different representations. Also the quantization of the values taken by these categories to finite set was chosen such that these values are nearly equiprobable. This has the advantage of making Gaussian mixture models built using the efficient available technique of EM algorithm more capable of minimizing probability of error in phonological features recognition than phonetic speech recognition systems.

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. Blimes "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] Kenneth N. Stevens "Acoustic Phonetics". The MIT press, Cambridge, Massachusettes, 1998.