FEATURE GENERATION BASED ON MAXIMUM CLASSIFICATION PROBABILITY FOR IMPROVED SPEECH RECOGNITION

Xiang Li and Richard M. Stern

Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA

{xiangl, rms}@cs.cmu.edu

ABSTRACT

Feature representation is a very important factor that has great effect on the performance of speech recognition systems. In this paper we focus on a feature generation process that is based on linear transformation of the original log-spectral representation. We first discuss several three popular linear transformation methods, Mel-Frequency Cepstral Coefficients (MFCC), Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA). We then propose a new method of linear transformation that maximizes the normalized acoustic likelihood of the most likely state sequences of training data, a measure that directly related to our ultimate objective of reducing Bayesian classification error rate in speech recognition. Experimental results show that the proposed method decreases the relative word error rate by more than 9.1% compared to the best implementation of LDA, and by more than 25.9% compared to MFCC features.

1. INTRODUCTION

As is the case with all pattern classification systems, the performance of speech recognition systems depends critically on the features it uses. In addition to the commonly-used cepstral representation Mel-Frequency Cepstral Coefficients (MFCC), there are other front-end representations generated with different methods and different objectives. These front-end representations can be categorized according to their generation processes, such as features based on linear transformation (*e.g.* MFCC [1], PCA [2], LDA [3][4]), and features based on non-linear transformations (*e.g.* PLP [6], tandem features [7]). Other methods of feature generation include discriminative-based feature generation, whose objective is to make the representation of different classes as different from one another as possible in the resulting feature space, or maximum likelihood-based representations, whose goal is to make the data in the new feature space fit the model assigned to them as well as possible.

Despite the success of these feature representations, most of them are based on heuristics, as neither the objectives of maximal separation [3][4] nor maximum likelihood [1][2] directly relates to our real objective of minimal word error rates (WERs). The aim of this paper is to derive a linear feature generation process based on an objective function which is intimately linked to this goal of minimal WERs. Specifically, we will use as our objective function the normalized acoustic likelihood of the most likely state sequences generated from forced-alignment, a measure can be thought of as the *a posteriori* probability of those most likely state sequences assuming that the *a priori* probabilities of the state sequences are equal. We will use the optimization procedure of gradient ascent to tune a transformation matrix used in our feature generation process in order to achieve this goal.

In the following section we will describe the most commonly used feature generation methods based on linear transformation as well as our new method. In Section 3, we will present our experimental results, and we present our conclusions in Section 4.

2. FEATURE GENERATION BASED ON LINEAR TRANSFORMATIONS

2.1. DCT-based linear transformation

The most commonly used feature representation in speech recognition is Mel-Frequency Cepstral Coefficients (MFCC [1]), where the log energies of the outputs of Mel-frequency filters are transformed via the Discrete Cosine Transform (DCT) as in Eq. (1)

$$MFCC_i = \sum_{k=1}^{N} X_k \cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{N}\right] \qquad i = 1, 2, \ldots, M \quad (1)$$

where X_k is the log-energy output of the k^{th} Mel-frequency filter, N is the total number of Mel-frequency filters, and M is the number of cepstral coefficients.

Due to the large amount of recognition classes used in modern speech recognition systems based on Hidden Markov Models (HMMs), many people use a diagonal covariance matrix to model the observation probabilities of each recognition class, which implicitly assumes mutual independence among the components of feature vector. Even though the DCT transform does not provide a theoretical guarantee of independence of the transformed components, the MFCC coefficients generated via the DCT generally become more mutually independent compared to the original log-spectral energies from which they are obtained. This is perhaps one reason for the general success of MFCC feature representations in most speech recognition systems.

2.2. Principal Component Analysis

Principal Component Analysis (PCA [2]) is a method of dimensionality reduction base on linear transformation that attempts to obtain the best representation of the original data in the least-squares sense in the projected space. Letting A represent the transformation matrix, and letting the column vector X_k represent the original data, the goal of PCA is to find A that minimizes the accumulated squared difference between the projection of the data in the new space and the original data as in Eq. (2)

$$A = Argmin\left\{\sum_{k=1}^{N}\left(\sum_{j=1}^{M}A_jX_kA_j - X_k^T\right)\left(\sum_{j=1}^{M}A_jX_kA_j - X_k^T\right)^T\right\} \quad (2)$$

where *N* is the total number of data samples, *M* is the dimension after transformation, A_j is the j^{th} row vector of the transformation matrix A, and T represents the matrix transpose operator.

The resulting matrix A is actually the eigenmatrix of the covariance matrix of the original data, and each row vector A_j is a principal axis of the original data, ranked according to the value of its corresponding eigenvalue.

In contrast to the DCT used in MFCC feature generation, the projected data obtained using PCA are guaranteed to be mutually uncorrelated due to the orthogonality of their principal axes A_j. Consequently, projected data obtained using PCA are in principle more in accord with the common assumption of diagonal covariance matrices.

2.3. Linear Discriminant Analysis

While the objective of PCA is to preserve the original data in the projected space to the extent possible, linear discrimination analysis (LDA) [3][4] is a linear transformation-based dimensionality reduction method that attempts to maximally separate the data of different recognition classes in the new space. It uses the Fisher ratio of the determinants of the *between-class* and *within-class* scatter matrices as the measure of data separation, and maximizes this measure.

If we use S_B and S_W to represent the *between-class* and *within-class* scatter matrices respectively[4], then the goal of LDA [3][4] is to find the transformation matrix *A* that maximizes the Fisher ratio:

$$A = Argmax\left(\frac{|A^T S_B A|}{|A^T S_W A|}\right) \quad (3)$$

As is shown in [4], the transformation matrix that maximizes the ratio is actually the matrix of eigenvectors of the matrix $S_w^{-1}S_B$.

As in the case of PCA, LDA transforms the data into a space where components are uncorrelated with each other. On the other hand, LDA requires that each sample of training data be labelled according to decision class, which causes the result to depend on the way that the classes are defined. Hence, defining the decision classes to be according to states versus phonemes will affect the resulting system recognition accuracy. In addition, LDA assumes that covariance matrices for all recognition classes are the same, which is a strong assumption.

2.4. Maximum Bayes Classification Probability Based Linear Feature Generation

Despite their success, methods of feature generation based on MFCC, PCA and LDA all share the same drawback: the objective of linear transformation is not directly related to the true goal of minimizing Bayesian classification error, which is usually expressed in terms of WER in speech recognition systems. The method we describe in this paper also performs a linear transformation to generate new features with reduced dimensions. Unlike other methods, the present transformation matrix is generated via an optimization procedure whose objective function is the normalized *acoustic likelihood* P_c of the true recognition classes in the training data as in Eq. (4),

$$P_c = \prod_{i \in TraingFrame} \frac{P(X_i|C_{h,i})}{\sum_{j=1}^{C} P(X_i|C_j)} \quad (4)$$

where X_i is the data sample in frame *i*, $C_{h,i}$ is the most likely state in frame *i* from the forced-alignment result, and *C* is the total number of recognition classes.

As reflected in Eq. (4), P_c can be treated as the *a posteriori* probability of true recognition classes with the flat *a priori* probability assumption, which is directly related to the Bayes classification error.

Since the training data sample X_i in each frame *i* is fixed in the feature space before transformation, the normalized acoustic likelihood P_c will only depend on the parameters of the model used for the observation probability for each recognition class as reflected from Eq. (4). Suppose the training data can be partitioned into recognition classes in a frame-by-frame manner based on forced-alignment to the correct hypothesis, and those parameters of the model for observation probability can be generated via a maximum likelihood estimation approach based on the training data assigned (using an approach such as the *K*-means training algorithm). P_c will hence only depends on the partitioning based on forced alignment of the training data in the feature space before the transformation. When we apply a linear transform to the original data space, both P_c and the new partition of the transformed training data will change according to the exact nature of the transformation matrix. If we assume that (1) the partition of the transformed data is the same as the partition in the original data space (which can be easily enforced via forced-alignment of the training data) and (2) the model for the observation probabilities of each recognition class is a single Gaussian, then we can write the equations for P_c as a function of the transformation matrix, and maximize it with respect to the transformation matrix *A*.

The method can be described as follows: we rewrite the term P_c in the transformed space as:

$$P_c = \prod_{i \in TraingFrame} \frac{P(X_i' | C_{h,i})}{\sum_{j=1}^{C} P(X_i' | C_j)} \tag{5}$$

where $X_i' = AX_i$ is the transformed feature vector. The conditional density $P(X_i' | C_j)$ can be written as:

$$P(X_i' | C_j) = \frac{1}{(2\pi)^{\frac{m}{2}} |\Sigma_j'|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(X_i' - \mu_j')^T \Sigma_j'^{-1} (X_i' - \mu_j') \right\} \tag{6}$$

where *m* is the dimensionality of the transformed data, and μ_j' and Σ_j' are the transformed means and variance of the recognition class *j*, respectively:

$$\begin{aligned} \mu_j' &= A\mu_j \\ \Sigma_j' &= A\Sigma_j A^T \end{aligned} \tag{7}$$

With μ_j and Σ_j the original mean and variance estimated from the partition of the training data in the original data space, they are fixed with respect to the transformation matrix *A*.

By substituting Eq. (6) and Eq. (7) into Eq. (5), we obtain:

$$P_c =$$

$$\prod_i \frac{\exp\left\{ -\frac{1}{2}(AX_i - A\mu_{h,i})^T (A\Sigma_{h,i} A^T)^{-1} (AX_i - A\mu_{h,i}) \right\} |A\Sigma_{h,i} A^T|^{-\frac{1}{2}}}{\sum_{j=1}^{C} \exp\left\{ -\frac{1}{2}(AX_i - A\mu_j)^T (A\Sigma_j A^T)^{-1} (AX_i - A\mu_j) \right\} |A\Sigma_j A^T|^{-\frac{1}{2}}} \tag{8}$$

For simplicity, we replace P_c by $LogP_c$ as our objective function and optimize it with respect to *A*. $\nabla LogP_c$, which is the first derivative of $LogP_c$ with respect to transformation matrix *A*, can be expressed as:

$$\nabla LogP_c = \sum_i \left\{ \nabla LogP(X_i' | C_{h,i}) - \nabla Log\left[\sum_{j=1}^{C} P(X_i' | C_j) \right] \right\} \tag{9}$$

Where:

$$\begin{aligned} \nabla Log\left[\sum_{j=1}^{C} P(X_i' | C_j) \right] &= \frac{\sum_{j=1}^{C} \nabla P(X_i' | C_j)}{\sum_{j=1}^{C} P(X_i' | C_j)} \\ &= \frac{\sum_{j=1}^{C} P(X_i' | C_j) \nabla LogP(X_i' | C_j)}{\sum_{j=1}^{C} P(X_i' | C_j)} \end{aligned} \tag{10}$$

Since $P(X_i' | C_j)$ is the acoustic likelihood of recognition class *j* based on the transformed data, it can be computed easily. Hence all that is needed is to compute $\nabla LogP(X_i' | C_j)$, the first derivative of the log acoustic likelihood of class *j* in the transformed data X_i' with respect to the transformation matrix *A*.

By assuming that the output probabilities can be modeled by a single Gaussian with a diagonal covariance matrix , we obtain:

$$\nabla LogP(X_i' | C_j) = -\sum_{k=1}^{m} \nabla \left[\frac{(x'_{i,k} - \mu'_{j,k})^2}{2\sigma'^2_{j,k}} + \frac{\log(\sigma'^2_{j,k})}{2} \right] \tag{11}$$

where $x'_{i,k}$, $\mu'_{j,k}$ and $\sigma'^2_{j,k}$ are the individual components of the transformed feature values, and the mean and variance of class *j:*

$$\begin{aligned} x'_{i,k} &= \sum_{p=1}^{N} a_{p,k} x_{i,p}; \qquad \mu'_{j,k} = \sum_{p=1}^{N} a_{p,k} \mu_{j,p} \\ \sigma'^2_{j,k} &= \sum_{i=1}^{N} \left(a_{j,i} \sum_{p=1}^{N} a_{k,p} \sigma^2_{i,p} \right) \end{aligned} \tag{12}$$

where *N* is the original dimension, *m* is the reduced dimension, $a_{p,j}$ represents component (*p, j*) of the transformation matrix *A*, and $\mu_{j,k}$ and $\sigma^2_{j,k}$ are the mean and variance of the k^{th} component of recognition class *j*.

Finally, by combining Eq. (11) and Eq. (12) we can write the closed-form solution to $\nabla LogP(X_i' | C_j)$. Substituting this term into Eq. (9) produces the closed-form solution of the first derivative of the log-normalized acoustic likelihood $\nabla LogP_c$ with

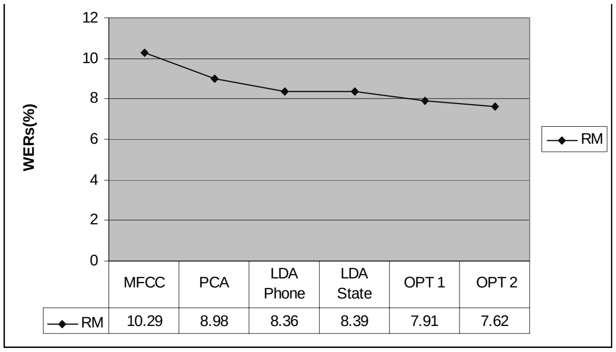

Figure 1: Word Error Rates obtained using various feature generation scheme. LDA Phone is obtained using phone classes, LDA State is obtained using state classes; OPT 1 is the proposed method derived by maximizing acoustic likelihood over all training frame, and OPT2 is the proposed method derived by maximizing the normalized acoustic likelihood only in mis-classified frames of training data.

respect to A. This quantity is then used as the increment in the gradient ascent approach to find the A matrix that optimizes the log-normalized acoustic likelihood term.

3. EXPERIMENTAL RESULTS

To compare the performance of our proposed method with that of other feature generation methods, we carried out a series of experiments using the DARPA Resource Management (RM) database. All of these the experiments were conducted using the CMU SPHINX-III speech recognition system, using 3-state continuous HMMs. Since the proposed method was derived by assuming a single Gaussian for the output distributions, we used a single-Gaussian observation probabilities for all feature sets to which the proposed method was compared. In addition to our own method, we evaluated the performance of MFCC, PCA and LDA features for comparison. Since there is no theoretically-motivated conclusion about the best class level to use for LDA, we obtained results using both phoneme based-classes and state-based classes, with class labels generated from the forced-alignment of the MFCC model.

Since the distributions of correctly-classified and mis-classified frames may be different from one another, optimizing the normalized acoustic likelihood P_c of the whole training data may be biased to correctly-classified frames given the large amount of such frames compared with mis-classified frames. We optimize P_c both for the entire ensemble of training frames as well as only for the frames in which misclassifications occurred.

We used the method of steepest gradient ascent as the optimization procedure, using the transformation matrix obtained by the phoneme-based LDA method as an initial value. The optimization process was terminated when the results converged.

Our experimental results are reported in Figure 1. These results results show that the proposed optimization over only the mis-classified frames decreases the relative word error rate by more than 9.1% compared to the best implementation of LDA, and by more than 25.9% compared to MFCC features. We also computed the statistical significance measure between our proposed methods with the best-performing previous method of phoneme-based LDA. The matched pairs method [5] provides a significance measure of 0.09 between phoneme-based LDA and our optimization over the entire training data, and 0.03 between phoneme-based LDA and our optimization over only the mis-classified training data.

4. DISCUSSION AND CONCLUSIONS

We first note that our proposed method outperforms conventional feature generation methods, and that the improvement, particularly for the optimization of normalized acoustic likelihood in mis-classified frames, is significant.

While we made several assumptions before deriving our methods, we can relax these assumptions with some modification of our current method. We can use an iterative procedure, which partitions the training data according to the model generated from the previous iteration, to relax the assumption of fixed partitioning of training data. We are also attempting to use iterative procedures to develop a modified version of our method that is based on Gaussian mixture output distributions, with the parameters and coefficients of each Gaussian mixture based on the model from the last iteration. Of course, the iterative procedure will require substantial additional computation, given that the optimization process will have to be performed for each iteration.

Another improvement that can be applied to our method is to partition the training data based on the Baum-Welch method, the actual ML training method used in most current speech recognition system, instead of the K-means algorithm. We expect even better performance to be obtained with the “soft” partitioning of training data based on the Baum-Welch algorithm.

ACKNOWLEDGEMENTS

This research was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

REFERENCES

[1] Davis, S. B., and Mermelstein, P. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences” *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4): 357-366, Aug. 1980.

[2] Jolliffe, I. T. *Principal Component Analysis*, Springer-Verlag, New York, 1986

[3] Hunt, M. J, Richardson, S.M., Bateman, D.C, Piau, A. “An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination,” *Proc. ICASSP-91*, Toronto, pp. 881-884, May 1991.

[4] Duda, R. O, Hart, P. E., and Stork, D. G. *Pattern Classification*, 2nd Edition, John Wiley & Sons, Inc., 2001

[5] Gillick, L. and Cox, S. J. “Some statistical issues in the com-

parison of speech recognition algorithms*”*, *Proc. ICASSP-89*, Glasgow, pp. 532-535, June 1989

[6] Hermansky, H., Hanson, B., and Wakita, H. “Perceptually linear predictive(PLP) analysis of speech“, Journal of the Acoustic Society of America, Vol. 87, pp 1738-1752, April 1990.

[7] Hermansky, H., Ellis, D. P. W., and Sharma, S. “Tandem connectionist feature extraction for conventional HMM system*s*” *Proc. ICASSP-2000*, Istanbul, **3:**1635 -1638, 2000.