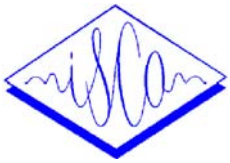

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

PARALLEL FEATURE GENERATION BASED ON MAXIMIZING NORMALIZED ACOUSTIC LIKELIHOOD

Xiang Li and Richard M. Stern
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213, USA
{xiangl, rms}@cs.cmu.edu

Abstract

Combining information from parallel feature streams generally improves speech recognition accuracy. While many studies have attempted to determine the stage of the recognition system that provides best combination performance and the specific nature of how features are combined, relatively little attention has been paid to the design or selection of parallel feature sets when used in combination. In this paper we propose a new parallel feature generation algorithm based on the criterion of maximizing the normalized acoustic likelihood of the features after they are combined, which is closely related to the recognition accuracy obtained using the combination of these features. We use a gradient ascent procedure to manipulate the values of a set of transformation matrices through which individual features are passed before they are combined in a fashion that maximizes the normalized acoustic likelihood term after the features are combined. The function that combine the parallel features together is an intrinsic part of the optimization process. The use of the optimal linear transformation provides a relative decrease of 12.7 percent Word Error Rate on the DARPA Resource Management task.

1. Introduction

Many studies have demonstrated the advantages of combining information from complementary parallel feature streams in speech recognition system (*e.g.* [1-10]). Generally speaking, there are two major issues associated with feature combination, the features to be combined and the method by which these features are combined, and the performance of systems that use combined features depends on both of these factors.

While many studies have attempted to determine the stage of the recognition system that provides best recognition accuracy and the specific nature of how features are combined, much less attention has been paid to the design or selection of the parallel features in order to provide the best performance when used in combination. There are a number of different ways in which complementary feature sets can be generated. For example, Ellis and Bilmes used the criterion of conditional mutual information to select parallel features to be combined from a set of predetermined candidates [6]. Other groups have developed parallel features through the use "splitting techniques" (*e.g.* [8], [9], [10]). For example, Halberstadt [8] split the speech recognition task into several sub-tasks, and designed parallel features that perform well within each sub-set. In [9] and [10], Bourlard and Hermansky split the whole spectrum into several frequency bands, extracting feature within each sub-band and subsequently combining them. Sets of parallel features have also been developed by adjusting specific system parameter values as in the variation of analysis frame rate between 80 and 125 frames per second by Billa *et al.* in the BBN BYBLOS system [7]. While all of these studies have demonstrated the potential of parallel feature combination methods for improved recognition accuracy, we are motivated in the present work to develop a way to optimize the choice of features to be combined, and in a takes into account the combination function itself.

The method of generation of parallel features that described in this paper will take a different approach from existing methods. It will transform the parallel feature generation process into an optimization process, whose objective function is directly related to the word error rate (WER) of the combined system. Specifically, we use the normalized acoustic likelihood of the most likely state sequence as our objective function, and search for the feature generation function that maximize this objective function. To simplify the process, we limit the parallel feature sets to be the linear transformations of traditional log-spectral feature. The objective function then becomes a function of the feature transformation matrices and the specific function that is used to combine the feature streams. Again, we note that the choice of combination function is an intrinsic part of the optimization process, which we believe to be helpful in reducing WER.

In the following section we describe how we generate parallel feature streams through the linear optimization process. We begin this discussion with the generation of a single optimal feature stream, and then extend our approach to the case of parallel feature stream generation. We present our experimental result in Section 3, and a discussion and conclusion in Section 4.

2. Linear feature generation by maximizing normalized acoustic likelihood

As described in the introduction section, new features are generated using an optimization process whose objective function is the normalized acoustic likelihood of the true (or most likely) recognition class over all the classes. We first consider the simplest case of the generation of linear feature stream, then develop the case of parallel feature streams generation.

2.1 Generation of a single feature stream

The task of linear feature generation is to find a matrix A that transforms the original feature vector (such as log-spectral features) into some new feature space. If the feature vector, mean, and variance of each phoneme or recognition state in the original feature space is labelled X, μ and Σ respectively, the corresponding parameters in the transformed feature space will become AX, $A\mu$, and $A\Sigma A^T$.

As noted above, our feature generation process is an optimization process. The objective function, which is the normalized likelihood P_c of the most likely state sequence in the new feature space can be written as

$$P_c = \prod_i \frac{P(AX_i|AC_{h,i})}{\sum_{j=1}^{C} P(AX_i|AC_j)} \quad (1)$$

where $AC_{h,i}$ represents the model parameters (*i.e.* the mean, variance, and Gaussian mixture weights) of the most-likely recognition Class $C_{h,i}$ (which could be a state or a phoneme) in frame *i,* C is the total number of recognition classes.

Under the general Gaussian mixture model assumption, the acoustic likelihood term $P(AX_i|AC_j)$ can be written as:

$$P(AX_i|AC_j) = \sum_m w_m P(AX_i|AC_{j,m}) \quad (2)$$

where w_m is the coefficient of component *m* within the mixture, $P(AX_i|AC_{j,m})$ is the individual Gaussian probability of component *m,* which can be written as

$$P(AX_i AC_{j,m}) = \frac{\exp\left\{-\frac{1}{2}(AX_i - A\mu_{j,m})(A\Sigma_{j,m}A^T)^{-1}(AX_i - A\mu_{j,m})^T\right\}}{(2\pi)^{D/2}\left|A\Sigma_{j,m}A^T\right|^{1/2}} \quad (3)$$

where D is the dimensionality of the new feature space.

Substituting Eq. (2) and Eq. (3) into Eq. (1), it is clear that the normalized acoustic likelihood term P_c becomes a function of the transformation matrix *A,* and that this function can be optimized using a gradient ascent approach. Because of space limitations, will not provide details about this procedure here, but the interested reader is referred to our related previous papers [11][12].

2.2 Generation of parallel feature streams

The generation of parallel feature streams that maximize the normalized acoustic likelihood in the combined system is very similar to the generation of a single feature stream. The only difference is that when we generate parallel feature streams, the acoustic likelihood term $P(AX_i|AC_j)$ now becomes a function of acoustic likelihood terms from each individual feature stream. In the case of two streams, the likelihood term becomes

$$P(AX_i|AC_j) = F\{P(A_1X_i|A_1C_j), P(A_2X_i|A_2C_j)\} \quad (4)$$

where $P(A_1X_i|A_1C_j)$ and $P(A_2X_i|A_2C_j)$ are the acoustic likelihoods from Feature Streams 1 and 2 respectively as in Eqs. (2) and (3). The symbol F specifies the function used to combine the probabilities (typically *summation, multiplication,* or *maximization* in our work).

Substituting Eq. (4) into Eq. (1), it is clear that the normalized acoustic likelihood term P_c now becomes a function of the transformation matrices A_1 and A_2, and the combination function *F.* Given a particular combination function *F,* P_c becomes only a function of the transformation matrices A_1 and A_2.

The gradient ascent procedure is based on the derivative of the acoustic likelihood $P(A_iX|A_iC_j)$ of each individual feature stream *i* with respect to its transformation matrix A_i as in [11][12]. If the combination function F is differentiable (as in the case of . *summation, linear regression* and *multiplication*), we can compute the derivative of P_c with respect to an individual transformation matrix A_i using the chain rule, and optimize it using the gradient ascent method.. While the *maximization* function is not directly differentiable, its derivative can be obtained by computing the limit of the derivative of the R-norm function

$$\left\|P(AX_i|AC_{h,i})\right\|_R = \left(\sum_l (P(A_lX_i|A_lC_{h,i}))^R\right)^{1/R} \quad (5)$$

as R approaches infinity..

As an example, consider the generation of two linear feature streams (via transformation matrices A_1 and A_2) using *summation* as the combination function. P_c, the normalized acoustic likelihood of the combined system, now becomes

$$P_c = \prod_i \frac{P(A_1X_i|A_1C_{h,i}) + P(A_2X_i|A_2C_{h,i})}{\sum_{j=1}^{C}[P(A_1X_i|A_1C_j) + P(A_2X_i|A_2C_j)]} \quad (6)$$

where the meaning of the parameters is the same as in Eq. (1).

By taking the log of P_c and computing the partial derivative of $LogP_c$ with respect to the transformation matrix A_1, we obtain

$$\nabla_{A_1} LogP_c = \sum_i \left\{ \frac{\nabla_{A_1} P(A_1X_i|A_1C_{h,i})}{P(A_1X_i|A_1C_{h,i}) + P(A_2X_i|A_2C_{h,i})} - \frac{\sum_{j=1}^{C} \nabla_{A_1} P(A_1X_i|A_1C_j)}{\sum_{j=1}^{C}[P(A_1X_i|A_1C_j) + P(A_2X_i|A_2C_j)]} \right\} \quad (7)$$

where $P(A_1X_i|A_1C_j)$ and $P(A_2X_i|A_2C_j)$ are the acoustic likelihoods of Class C_j obtained from Feature Streams 1 and 2, as computed from Eq. (3). Similarly, $\nabla_{A_1}P(A_1X_i|A_1C_j)$ is the derivative of acoustic likelihood of Class C_j from feature stream 1 with respect to transformation matrix A_1 as in [11][12].

Following the same approach, we can also compute the partial derivative of $LogP_c$ with respect to the transformation matrix A_2. By using the gradient ascent approach, we then find the transformation matrices A_1 and A_2 that maximize the normalized acoustic likelihood term P_c iteratively.

At this point we offer several comments about the procedure. First, we have made a fundamental assumption in our application of the linear feature generation method that the partitioning of the training data according to decision class is the same before and after the transformation via the matrix A_i. This enables us to apply Eq. (3) to compute the acoustic likelihood term in the space. This assumption can be relaxed through the use of an iterative procedure in which we use the partition of a previous iteration to initialize the parameters of the current iteration. Second, the mean and variance of the features in the transformed feature space as $A\mu$ and $A\Sigma A^T$ will only be used in the process of obtaining the transformation matrices *A*. While the actual model parameters of recognition classes (e.g. states) used in speech recognition in the transformed feature space will be Maximize Likelihood (ML) estimated directly based on the new feature using the conventional Baum-Welch training algorithm. Finally, we only use training data to generate the transformation matrices, and the transformation matrices generated are non-square. (e.g. from 40 log spectral coefficients to 13 cepstral coefficients.)

3. Experimental results

We carried out a series of experiments using the DARPA Resource Management (RM) database to compare the performance of our parallel feature generation method with conventional linear feature generation methods. All experiments were conducted using the CMU Sphinx III continuous speech recognition system with a 3-state continuous HMM structure. We used a context-dependent tied model structure with the total number of tied states equal to 2000. A bigram language model was used.

All the features tested were linearly transformed from log-spectral features. Augmentation by delta and double-delta components, log-spectral features produced a 120-dimensional feature vector before transformation that was reduced to a new 39-dimensional feature space by the transformation matrices. A baseline system was developed that used a feature set that was obtained by applying linear discriminant analysis (LDA) and principal component analysis (PCA) to the original log-spectral features. Feature combination was performed at the state probability level (sometimes "middle combination" [4][9]) by either summing or multiplying the probabilities of the PDA and LDA features together. The parallel features were generated using a gradient ascent proedure, with the transformed LDA and PCA taken as initial values. The ascent process was terminated when the normalized acoustic likelihood term P_c converges.

Experimental results are reported in terms of WER in Figure 1. Results in Fig. 1 are presented in three groups. The first group of results was obtained using the LDA and PCA features directly, either in isolation, or in multiplicative or additive from left to right. The center block of results was parallel features generated using the optimal linear transformations using multiplicative combination, and the third group represents results obtained using parallel features generated using additive combination. The best results were obtained using parallel features generated using summation, with summation also used to combine these features in the decoding process. This configuration produces a 12.7% relative decrease in WER compared to the comparable result without the optimizing linear transformation, and a relative decrease in WER of more than 30% compared to the WER using the LDA or PCA features in isolation.

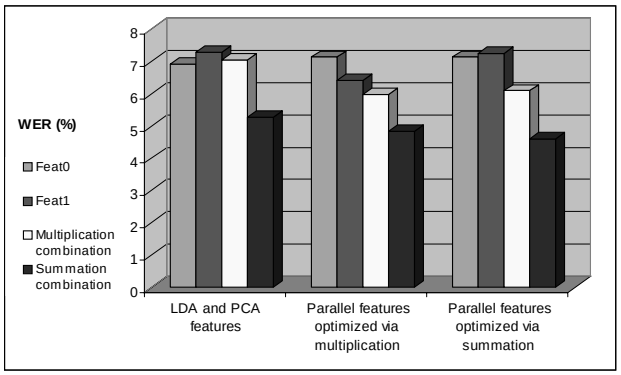

Figure 1: Word error rates (WER) for the DARPA RM corpus obtained using LDA features, PCA features, and optimized features that are derived from them, both individually and in combination. Results are clustered into three blocks according to how the features are generated. Results within each block differ according to whether features are used in isolation or in either additive or multiplicative combination.

We also compare a subset of the results with the normalized acoustic likelihood term P_c in Table 1. We observe in Table 1 that best performance is obtained using parallel features when those features are generated using the same combination function as the one with which they are combined in the decoding process. This confirms the sensitivity of the parallel features that to the nature of the combination function used to generate them, and also indicates (unsurprisingly) that .lowest WER is achieved when parallel features are combined in decoding in the manner with which they are generated.

	Combination function	**LDA& PCA**	**Parafeat (prod)**	**Parafeat (sum)**
WER (%)	Sum	5.28	4.83	4.61
	Prod	7.06	5.99	6.1
Log P_c	Sum	-60.5	-58.1	-57.6
	Prod	-78.3	-72.7	-74.9

Table 1. Normalized acoustic likelihood log P_c and WER obtained for various sets of parallel features combined with different combination functions. parafeat(sum) and parafeat(prod) represent parallel features that are generated by using *summation* and *multiplication* as combination function respectively.

	LDA & PCA	Parafeat (prod)	Parafeat (sum)
S	0.7368	0.8067	0.9932

Table 2. Complementarity measure *S* for LDA and PCA features, and for parallel feature streams that are generated using multiplicative and additive combination functions (see text) .

By comparing the "Sum" and "Prod" rows in Table 1, we also note that for each type of combination function, when WER scores decreases the corresponding values of the P_c defined in Eq. (1) increase. These limited comparisons suggest that maximizing P_c as the objective of the gradient ascent will indeed tend to reduce WER.

It is frequently considered desirable for parallel sets to represent complementary attributes of the speech waveform. Although our linear parallel feature streams were generated without any regard to the complementarity of the features that were produced, we were curious about the extent to which increases or decreases of complementarity would correspond to changes in WER. We defined an *ad hoc* measure of complementarity in the form of

$$S = (1-p)p \cdot \sum_i \frac{|e_{1,i} - e_{2,i}|}{e_{1,i} + e_{2,i}} \qquad (8)$$

The variables $e_{1,i}$ and $e_{2,i}$ in the above equation refer to the error rates obtained by a phonetic recognizer for phoneme *i* using feature 1 and feature 2, respectively. The fraction *p* represents the fraction of phonemes for which feature 1 performed better than feature 2. Hence the complementarity measure *S* reflects both the differences between results obtained using the two feature sets and the extent to which the phonemes are approximately evenly split between those that perform better with those that perform better with feature 2.

Table 2 compares the values of *S* measure that were observed by developing parallel features from LDA and PCA directly, and through the transformation matrices using *multiplication* and *summation* respectively. Comparing across these three conditions, we note that the complementarity measure *S* increases as WER decreases. While any serious confirmation of the value of this measure can only be obtained after many more results are considered, we regard the results of this limited experiment as both interesting and promising.

4. Summary and conclusions

We describe an approach to the development of parallel features for use in automatic speech recognition systems. Parallel sets were generated by passing initial log-spectral through several linear transformations that are manipulated using gradient ascent approach to maximize normalized acoustic likelihood. Features obtained using the LDA and PCA methods were used to initialize the gradient ascent process. The function that determines how feature streams are combined was an intrinsic part of the optimization process. The use of parallel features that were generated using the linear transformation matrices and gradient ascent provided a decreases in WER on the DARPA Resource Management task of 12.7% compared to the comparable result without the optimizing linear transformation, and a relative decrease in WER of more than 30% compared to the WER obtained using either the LDA or PCA features in isolation.

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 no official endorsement should be inferred.

References

[1] Fiscus, J. G., "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[2] Singh, R., Seltzer, M., Raj, B., and Stern, R.M, "Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination", *Proc. ICASSP 2001*, Salt Lake City.

[3] Li, X., Singh, R., and Stern, R.M. "Combining search spaces of heterogeneous recognizers for improved speech recognition", *Proc. ICSLP 2002,* Denver..

[4] Li, X., and Stern, R.M., "Training of stream weights for the decoding of speech using parallel feature streams", *Proc. ICASSP 2003,* Hong Kong.

[5] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., and Vergyri, D, "Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins summer 2000 workshop", *IEEE Fourth Workshop on Multimedia Signal Processing*, pp. 619-624, 2001

[6] Ellis, D.P.W., and Bilmes, J., "Using mutual information to design feature combinations", *Proc. ICSLP2000,* Oct. Beijing, 2000

[7] Billa, J., Colhurst, T., El-Jaroudi, A., Iyer, R., Ma, K., Matsoukas, S., Quillen, C., Richardson, F., Siu, M., Zavaliagkos, G., Gish, H., "Recent experiments in large vocabulary conversational speech recognition", *Proc. ICASSP 1999*, Phoenix.

[8] Halberstadt, A.K. "Heterogeneous acoustic measurements and multiple classifiers for speech recognition", *Ph.D. Thesis, MIT,* 1998.

[9] Bourlard, H., and Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands", *Proc. ICSLP 1996*, Philadelphia.

[10] Hermansky, H., Tibrewala, S., and Pavel, M., "Toward ASR on partially corrupted speech", *Proc. ICSLP1996,* Philadelphia.

[11] Li, X. and Stern, R. M. "Feature generation based on maximum classification probability for improved speech recognition", *Proc. Eurospeech2003,* Geneva.

[12] Li, X. and Stern, R.M. "Feature generation based on maximum normalized acoustic likelihood for improved speech recognition", *Proc. ICASSP 2004*, Montreal.

[13] Gillick, L., Cox, S.J., "Some statistical issues in speech recognition algorithms", *Proc. ICASSP 1989*, Glasgow.