ACOUSTIC FEATURE COMBINATION FOR ROBUST SPEECH RECOGNITION

András Zolnay, Ralf Schlüter, and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University, 52056 Aachen, Germany
{zolnay, schlueter, ney}@informatik.rwth-aachen.de

ABSTRACT

In this paper, we consider the use of multiple acoustic features of the speech signal for robust speech recognition. We investigate the combination of various auditory based (Mel Frequency Cepstrum Coefficients, Perceptual Linear Prediction, etc.) and articulatory based (voicedness) features. Features are combined by a Linear Discriminant Analysis based and by a log-linear model combination based techniques. We describe the two feature combination techniques and compare the experimental results. Experiments performed on the large-vocabulary task *VerbMobil II* (German conversational speech) show that the accuracy of automatic speech recognition systems can be improved by the combination of different acoustic features.

1. INTRODUCTION

Most automatic speech recognition systems use auditory based representation of the speech signal, e.g. Mel Frequency Cepstrum Coefficients (MFCC), Perceptual Linear Prediction (PLP), and variations of these methods. There have been also attempts at using articulatory information in the acoustic front-end, e.g. autocorrelation based voicedness feature [1]. In this paper we investigate the combination of different auditory based and articulatory based acoustic features.

Combination of acoustic features can be carried out directly on the level of feature vectors. In [1], liftered cepstral coefficients derived from all-poles magnitude spectrum has been directly concatenated with a voicedness feature. Using the concatenated features, a large relative improvement in word error rate (WER) has been achieved by applying discriminative training. Significant reduction in WER has been presented using LDA based feature combination in [2] when combining MFCCs with a phase feature and in [3] when combining MFCCs with a voicedness feature.

Combination of acoustic features can also be performed by log-linear model combination. In [4], different acoustic models have been combined by log-linear combination of acoustic and language model probabilities. The combination of 5 acoustic and language models (within-word and across-word acoustic models, bigram, trigram, and fourgram language models) has led to a significant improvement in WER, compared to the best pairwise combinations. In [5], significant reduction in WER has been achieved by using log-linear model combination to combine MFCC and main spectral peak features. Combination of PLP and modulation spectrogram features is described in [6]. Significant reduction in WER has been achieved by feature combination via the acoustic posterior probabilities, determined by an artificial neural network (ANN) based acoustic model.

In this work, we have tested a LDA based and a log-linear based feature combination methods on various feature types. Experiments have been performed on the large-vocabulary German conversational speech corpus *VerbMobil II*. On the one hand, we will compare the two feature combination methods on the same set of features. On the other hand, we will present experiments in which combinations of several auditory and articulatory based acoustic features have been tested. Experiments have yielded improvements in WER up to 7% relative to our best system optimized on the MFCC feature.

In the following we will first review the different feature extraction methods in Section 2. We will describe the LDA based feature combination in Section 3 and the log-linear model combination based feature combination in Section 4. We will present recognition results in Section 5 using LDA based and log-linear based combination of various acoustic features.

2. SIGNAL ANALYSIS

In this section, we present the feature extraction methods used in our speech recognition system. First we describe the Mel Frequency Cepstrum Coefficients (MFCC), followed by its variant derived from all-poles magnitude spectrum. In the next group of features, we describe the Perceptual Linear Predictive (PLP) feature [7] along with its alternative using a Mel scale triangular filter bank (MF-PLP). Finally, we present an autocorrelation based voicedness feature.

2.1. Mel Frequency Cepstral Coefficients (MFCC)

In the first step of the MFCC feature extraction algorithm, we perform a preemphasis of the sampled speech signal. The preemphasized samples $d[n]$ are obtained from the original samples $s[n]$ by the differencing $d[n] = s[n] - s[n-1]$. Every 10ms, a Hamming window is applied to preemphasized 25ms long speech segments. We compute the short-term spectrum by Fast Fourier Transform (FFT) along with an appropriate zero padding. Next, we compute the outputs of 20 overlapping Mel scale triangular filters. For each filter, the output is the sum of the weighted spectral magnitudes. Logarithm is next applied to the filter bank outputs followed by Discrete Cosine Transform which generates 16 cepstrum coefficients.

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under the post graduate program "Software für Kommunikationssysteme" and by the Europian Commission under the project TC-Star (FP6-506738).

Subsequently, a cepstral mean normalization is applied in order to account for different audio channels. Normalization is carried out with a symmetric sliding window of 2s. In this manner, every 10ms a feature vector consisting of 16 normalized cepstrum coefficients is computed.

2.2. MFCC Derived from All-Poles Magnitude Spectrum

In this method, MFCCs are derived from the all-poles magnitude spectrum estimate instead of the magnitude spectrum estimated by using Fast Fourier Transform. Thus the only step changed in the data flow of the MFCC algorithm is the way of calculating the magnitude spectrum. In the all-poles estimate, the magnitude spectrum $|X^{\mathsf{t}}(\omega)|$ of a time frame t is assumed to have the form of

$$|X^{\mathsf{t}}(\omega)| \approx \frac{g^{\mathsf{t}}}{|1 + \sum_{k=1}^{M} a_k^{\mathsf{t}}\ e^{-j\omega k}|} \tag{1}$$

where g^{t} is called the gain, a_k^{t} is a autoregressive coefficient, and M is number of autoregressive coefficients. Gain and the autoregressive coefficients can be directly calculated from the autocorrelation coefficients by applying the Levinson-Durbin recursion. M controls the smoothing of the magnitude spectrum. In our experiments, M is empirically set to 18. We calculate 512 points of the all-poles magnitude spectrum and carry out the rest of the MFCC algorithm. Finally, the 16 cepstrum coefficients derived from the all-poles magnitude spectrum are normalization by cepstral mean normalization as described in Section 2.1.

2.3. Perceptual Linear Predictive Analysis (PLP)

The motivation of PLP feature extraction is similar to the one of the MFCC method but there are major differences in the data flow.

In the first step, every 10ms, Hamming window is applied to the 20 ms long speech segments. Short-term spectrum is calculated by applying the FFT along with an appropriate zero padding. In the next step, a filter bank of 20 equally spaced overlapping Bark scale trapezoid filters is applied to the power spectrum. The filter bank is extended by an output at the frequency 0 and by another output at sample-rate / 2 by copying their right and respectively left neighbor. Equal loudness preemphasis is applied to the 22 filter bank outputs followed by the application of the intensity loudness law.

In the next stage of the algorithm, the cepstrum coefficients are not directly derived from the output of the intensity loudness law but from the all-poles approximation of it. First, autocorrelation coefficients are calculated by applying the Inverse Discrete Fourier Transform to the output of the intensity loudness law. Next, the 17 autocorrelation coefficient are transformed to the gain and to 16 autoregressive coefficients by the Levinson-Durbin recursion. Instead of regenerating the smoothed, all-poles approximation of the output of the intensity loudness law, we can directly compute the 16 cepstrum coefficients by applying a simple recursion. The zeroth cepstrum coefficient is explicitly set to the logarithm of the square of the gain. Finally, cepstral mean normalization is applied to the cepstrum coefficients as described in Section 2.1.

2.4. PLP Derived from Mel scale Filter Bank (MF-PLP)

In this method, the MFCC and PLP techniques are merged into one algorithm. The first steps until generating the output of the Mel scale triangular filter bank are taken from the MFCC algorithm. The only difference here is that the filter bank is applied to the power spectrum instead of the magnitude spectrum. The last steps generating the cepstrum coefficients are taken from the PLP algorithm. The 20 filter bank outputs are modified by the intensity loudness law. The 16 cepstrum coefficients are calculated from the output of the intensity loudness law via the all-poles approximation as described in Section 2.3. Finally, cepstral mean normalization is applied as described in Section 2.1.

2.5. Voicedness Feature

Voicedness feature is a measure representing the state of the vocal cords. The measure describes how periodic the speech signal is in a given time frame t. We use the autocorrelation function to measure periodicity. Autocorrelation $R^{\mathsf{t}}(\tau)$ expresses the similarity between the time frame $x^{\mathsf{t}}(\nu)$ and its copy shifted by τ. We have used the unbiased estimate of autocorrelation $\tilde{R}^{\mathsf{t}}(t)$:

$$\tilde{R}^{\mathsf{t}}(\tau) = \frac{1}{T-\tau} \sum_{\nu=0}^{T-\tau-1} x^{\mathsf{t}}(\nu)\, x^{\mathsf{t}}(\nu+\tau) \tag{2}$$

where T is the length of a time frame. Autocorrelation of periodic signals with frequency f attains its maximum $R^{\mathsf{t}}(0)$ not only at $\tau = 0$ but also at $\tau = \frac{k}{f}\ \ k = 0, \pm 1, \pm 2, ...$ integer multiples of the period. Therefore, a peak in the range of natural pitches with a value close to $R^{\mathsf{t}}(0)$ is a strong indication of periodicity.

In order to produce a bounded measure of voicedness, autocorrelation is divided by $\tilde{R}^{\mathsf{t}}(0)$. The resulting function has values mainly in the interval $[-1..1]$ although because of the unbiased estimate, theoretically any value is possible. The voicedness measure v^{t} is thus the maximum value of the normalized autocorrelation in the interval of natural pitch periods [2.5ms..12.5ms]:

$$v^{\mathsf{t}} = \frac{\max\limits_{2.5\text{ms}\cdot f_s \le \tau \le 12.5\text{ms}\cdot f_s} \tilde{R}^{\mathsf{t}}(\tau)}{\tilde{R}^{\mathsf{t}}(0)} \tag{3}$$

where f_s denotes the sample rate. Values of v^{t} close to 1 indicate voicedness, values close to 0 indicate voiceless time frames. The autocorrelation function is determined every 10ms on speech segments of 40ms length. By applying (3) to the autocorrelation, a one dimensional voicedness feature is generated every 10ms.

3. LDA BASED FEATURE COMBINATION

The Linear Discriminant Analysis (LDA) based approach combines directly the different acoustic feature vectors. In [8], LDA has been used successfully to find an optimal linear combination of successive vectors of a single feature stream. In the following steps, we describe a straightforward way to use this method for feature combination. In the first step, feature vectors extracted by different algorithms $x_t^{\mathsf{f}_i}$ are concatenated for all time frames t. In the second step, $2L+1$ successive concatenated vectors are concatenated again for all time frames t which makes up the large input vector of LDA. With $L = 5$ and with $F = 3$ different features, size of the LDA input vector grows up to ≈ 400 components. Finally, the combined feature vector y_t is created by

projecting the large input vector on a smaller (≈ 30 dimensional) subspace:

$$y_t = \left[\ V^T\ \right] \left[\begin{array}{c} \left[\begin{array}{c} x_{t-L}^{\mathrm{f}_1} \\ \cdots \\ x_{t-L}^{\mathrm{f}_F} \end{array}\right] \\ \cdots \\ \left[\begin{array}{c} x_{t}^{\mathrm{f}_1} \\ \cdots \\ x_{t}^{\mathrm{f}_F} \end{array}\right] \\ \cdots \\ \left[\begin{array}{c} x_{t+L}^{\mathrm{f}_1} \\ \cdots \\ x_{t+L}^{\mathrm{f}_F} \end{array}\right] \end{array}\right] \tag{4}$$

where the matrix V is determined by LDA such that it conveys the most relevant classification information to y_t. The resulting acoustic vectors are used as well in training and as in recognition.

In all our experiments, we have concatenated 11 successive feature vectors ($L = 5$). The baseline experiments using a single feature apply LDA in the same way. The only difference is in the size of the LDA input vector and thus in the size of the LDA matrix. The resulting feature vector has the same size to ensure comparable recognition results.

4. LOG-LINEAR MODEL COMBINATION

In this approach, different acoustic features are combined indirectly via the log-linear combination of acoustic probabilities $P_{\mathrm{f}_i}(X^{\mathrm{f}_i}|W)$ where W denotes a sequence of words and X^{f_i} denotes a sequence of feature vectors extracted by the algorithm f_i. The basic idea is to modify the modeling of the posterior probability $P(W|X)$ in Bayes' decision rule:

$$W_{opt} = \underset{W}{\operatorname{argmax}}\, P(W|X). \tag{5}$$

In the standard case, posterior probability is decomposed into language model probability $P(W)$ and acoustic model probability $P(X|W)$:

$$P(W|X) = \frac{P(W)P(X|W)}{\sum_{W'} P(W')P(X|W')}. \tag{6}$$

In the case of log-linear model combination, the posterior probability has the following form:

$$P(W|X) = \frac{e^{\sum_i \lambda_i g_i(W,X)}}{\sum_{W'} e^{\sum_i \lambda_i g_i(W',X)}} \tag{7}$$

where g_i is a so called feature function which is an arbitrary function of the word sequence W and the feature vector sequence X, and λ_i is the corresponding log-linear weight. Applying the log-linear modeling approach to speech recognition, the basic feature function types are negative logarithm of probabilities:

- language model: $g_i^{\mathrm{LM}}(W,X) = -\log P_i(W)$,
- acoustic model: $g_i^{\mathrm{AM}}(W,X) = -\log P_i(X|W)$.

Finally, in order to combine different acoustic features, we introduce a separate acoustic model $P_{\mathrm{f}_i}(X^{\mathrm{f}_i}|W)$ for each feature. Using a single language model feature function and for each feature a separate acoustic model feature function, the Bayes' decision rule for log-linear feature combination can be written as:

$$W_{opt} = \underset{W}{\operatorname{argmax}}\, P(W)^{\lambda_{\mathrm{LM}}} \prod_i P_{\mathrm{f}_i}(X^{\mathrm{f}_i}|W)^{\lambda_{\mathrm{f}_i}}. \tag{8}$$

Acoustic training of the combined system consists of two steps: independent training of each acoustic model $P_{\mathrm{f}_i}(X^{\mathrm{f}_i}|W)$ and training of the language model weight λ_{LM} and the acoustic model weights λ_{f_i}. In this work, we have run a standard maximum likelihood training to estimate the acoustic model parameters. Model weights have been optimized empirically.

5. EXPERIMENTAL RESULTS

5.1. Baseline Recognition System

Recognition tests have been conducted on the large-vocabulary corpus *VerbMobil II*. The corpus consists of German conversational speech: 36k training-sentences (61.5h) from 857 speakers and 1k test-sentences (1.6h) from 16 speakers. The baseline recognition system can be characterized as follows:

- recognition vocabulary of 10157 words;
- 3-state Hidden Markov Model topology with skip;
- 2501 decision tree based within-word triphone states including noise plus one state for silence;
- 237k gender independent Gaussian densities with global pooled diagonal covariance;
- class-trigram language model, test set perplexity: 62.0;
- 33 acoustic feature components after applying LDA.

In Table 1, we summarize results achieved by our recognition system optimized for different acoustic features: MFCC, vocal tract length normalized MFCC (MFCC-VTLN), MFCC derived from all-poles magnitude spectrum (MFCC-AllPoles), PLP, and PLP coefficients derived from Mel scale triangular filter bank (MF-PLP).

Acoustic Feature	Error Rates [%]		
	Del	Ins	WER
MFCC	6.3	2.4	23.1
MFCC-VTLN	5.0	2.7	21.3
MFCC-AllPoles	6.2	2.7	24.2
PLP	6.6	2.3	23.1
MF-PLP	6.2	2.7	23.2

Table 1. Baseline recognition results with different features.

5.2. Comparison of LDA Based and Log-Linear Combination

In this section, we describe experiments in which we combine several different acoustic features by the LDA based and by the log-linear based combination method.

LDA based method combines the different feature vectors directly, generating a single feature. Using the single combined feature stream, a standard acoustic model is trained under the settings given in Section 5.1. When using log-linear model combination, a separate acoustic models is trained for each feature. The different acoustic models are trained as well under the settings described in Section 5.1. This implies that each training includes the estimation and the application of an LDA matrix. In these cases, LDA does not combine different features but it finds an optimal linear combination of successive vectors of the same feature stream. The number of acoustic feature components after

LDA				Log-Linear			
Combined Features	Error Rates [%]			Combined Features	Error Rates [%]		
	Del	Ins	WER		Del	Ins	WER
MFCC + Voice	5.7	2.8	22.4	MFCC + Voice	6.1	2.7	23.0
				MFCC + LDA(MFCC + Voice)	5.9	2.7	22.2
MFCC-VTLN + Voice	5.1	2.6	20.8	MFCC-VTLN + LDA(MFCC + Voice)	5.3	2.3	20.3
MFCC + MFCC-VTLN + Voice	5.1	2.5	20.7	LDA(MFCC + Voice)+LDA(MFCC-VTLN + Voice)	5.3	2.2	19.9

Table 3. Recognition results of combining MFCC, vocal tract length normalized MFCC (MFCC-VTLN), and voicedness features (Voice). On the left, features are combined by LDA, on the right by log-linear model combination. LDA(MFCC + Voice) denote an acoustic model trained on the LDA based combination of MFCC and voicedness features.

Combined Features	Error Rates [%]		
	Del	Ins	WER
MFCC	6.3	2.4	23.1
MFCC + MFCC-AllPoles	5.8	2.5	22.6
MFCC + MF-PLP	6.0	2.6	22.9
MFCC + MF-PLP + PLP	5.6	2.6	22.1

Table 2. Recognition results of combining state-of-the-art features (MFCC, MFCC derived from all-poles magnitude spectrum, MF-PLP, and PLP) by using log-linear model combination.

applying the LDA matrix is set in case of MFCC, PLP, or in case of one of their variants to 33 components and in case of the voicedness feature to 1 component. After the training of acoustic models, the log-linear weights are optimized empirically using a simple grid search. Additionally, a useful option is to reuse features combined by LDA as a separate feature stream in the log-linear combination method.

Results of experiments combining state-of-the-art, auditory based features are summarized in Table 2. In spite of their common basic data flow, the log-linear based combinations of different auditory based features yield significantly better word error rates when compared to systems optimized on a single feature.

Table 3 summarizes the experimental results of combining MFCC, MFCC-VTLN, and voicedness features. Results show that the LDA based feature combination outperforms the log-linear model combination on small dimensional features, e.g. MFCCs combined with a single voicedness feature. Nevertheless we achieve significant improvements in WER if we reuse the LDA based combination of small dimensional features nested into the log-linear model combination. As shown in Table 3, we can reuse the acoustic models trained on the LDA based combination of the voicedness feature on the one hand with the MFCC feature and on the other hand with the MFCC-VTLN feature. The log-linear combination of the resulting acoustic models yields a significant improvement in WER over the pure LDA based combination of the three concerned features. One possible interpretations of the results is that with increasing number of features, the constant amount of training data become insufficient to robustly estimate the heavily enlarged within- and between-class scatter matrices. This may lead to numerical instability when solving the generalized eigenvalue problem. Another possibility for interpretation is that since we keep the number of output coefficients of the LDA constant, applying a single LDA matrix on increasing number of different feature vectors cannot convey as many classification information as the LDA matrices applied in the separate acoustic models of log-linear model combination.

6. SUMMARY

In this paper, we have analyzed two aspects of acoustic feature combination. On the one hand, we have compared an LDA based feature combination method and a log-linear model combination based method. Experiments have shown that LDA based combination nested into the log-linear model combination yields the best recognition result. On the other hand, we have performed experiments in which we have combined several different acoustic features. Despite their common basic structure, the combination of different state-of-the-art auditory based features resulted significant improvements in WER. The combinations of auditory based and articulatory based features have yielded up to 7% relative improvements in WER over the optimized single feature systems. Our future work includes extending the number of combined acoustic features and systematically analyzing the presented feature combination techniques.

7. REFERENCES

[1] D.L. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition feature," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998, vol. 1, pp. 21 – 24.

[2] R. Schlüter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May 2001, vol. 1, pp. 133 – 136.

[3] A. Zolnay, R. Schlüter, and H. Ney, "Robust speech recognition using a voiced-unvoiced feature," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, CO, Sept. 2002, vol. 2, pp. 1065 – 1068.

[4] P. Beyerlein, "Discriminative model combination," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998, vol. 1, pp. 481 – 484.

[5] H. Tolba, S. A. Selouani, and D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL, May 2002, vol. 1, pp. 837 – 840.

[6] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. Int. Conf. on Spoken Language Processing*, Sidney, Australia, Dec. 1998, pp. 891 – 894.

[7] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738 – 1752, June 1990.

[8] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, San Francisco, CA, Mar. 1992, vol. 1, pp. 13 – 16.