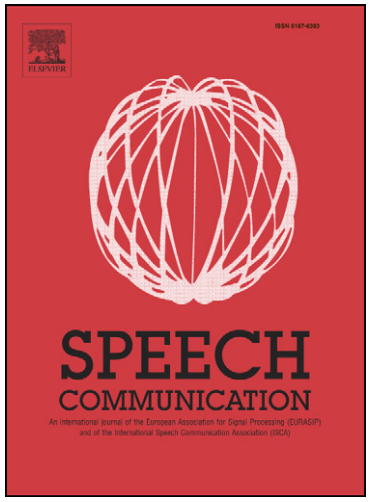
Using Multiple Acoustic Feature Sets for Speech Recognition

András Zolnay, Daniil Kocharov, Ralf Schlüter, Hermann Ney



Using Multiple Acoustic Feature Sets for Speech Recognition

András Zolnay, Daniil Kocharov, Ralf Schlüter, Hermann Ney

PII: S0167-6393(07)00070-2
DOI: 10.1016/j.specom.2007.04.005
Reference: SPECOM 1637

To appear in: *Speech Communication*

Received Date: 12 May 2006
Revised Date: 8 January 2007
Accepted Date: 11 April 2007

Please cite this article as: Zolnay, A., Kocharov, D., Schlüter, R., Ney, H., Using Multiple Acoustic Feature Sets for Speech Recognition, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.04.005

Using Multiple Acoustic Feature Sets for Speech Recognition

András Zolnay [a], Daniil Kocharov [b], Ralf Schlüter [a], and Hermann Ney [a]

[a] *Computer Science Department, Human Language Technology and Pattern Recognition, Lehrstuhl für Informatik VI, RWTH Aachen University, 52056 Aachen, Germany*

[b] *Department of Phonetics, Saint-Petersburg State University, 199034 Saint Petersburg, Russia*

Contents

Abstract

In this paper, the use of multiple acoustic feature sets for speech recognition is investigated. The combination of both auditory as well as articulatory motivated features is considered. In addition to a voicing feature, we introduce a recently developed articulatory motivated feature, the spectrum derivative feature. Features are combined both directly using linear discriminant analysis (LDA) as well as indirectly on model level using discriminative model combination (DMC). Experimental results are presented for both small- and large-vocabulary tasks. The results show that the accuracy of automatic speech recognition systems can be significantly improved by the combination of auditory and articulatory motivated features. The word error rate is reduced from 1.8% to 1.5% on the *SieTill* task for German digit string recognition. Consistent improvements in word error rate have been obtained on two large-vocabulary corpora. The word error rate is reduced from 19.1% to 18.4% on the *VerbMobil II* corpus, a German large-vocabulary conversational speech task, and from 14.1% to 13.5% on the British English part of the European parliament plenary sessions (*EPPS*) task from the 2005 TC-STAR ASR evaluation campaign.

Key words: acoustic feature extraction, auditory features, articulatory features,

voicing, spectrum derivative feature, linear discriminant analysis, discriminative model combination

1 Introduction

Most automatic speech recognition systems use at least partly acoustic features motivated by the models of the human auditory system. The most commonly used methods are the Mel frequency cepstrum coefficients (MFCC), perceptual linear prediction (PLP), and variations of these techniques. There have also been attempts at using acoustic features for speech recognition which are motivated by models of the human speech production system.

In this paper, the combination of several acoustic features is investigated. The extraction of different state-of-the-art auditory motivated acoustic features is reviewed, and detailed descriptions of the extraction of the voicing and the novel spectrum derivative features are given. In addition, investigations on the combination of these acoustic features are presented. Both the direct combination of feature sets by using linear discriminant analysis (LDA) as well as the implicit combination of feature sets via their acoustic emission distributions using discriminative model combination (DMC) and combinations thereof are described. The contributions of this paper are:

- *Voicing measure*: Former investigations showed that incorporation of the

Email addresses: zolnay@informatik.rwth-aachen.de (András Zolnay),
kocharov@phonetics.pu.ru (Daniil Kocharov),
schlueter@informatik.rwth-aachen.de (Ralf Schlüter),
ney@informatik.rwth-aachen.de (Hermann Ney).

voicing information into speech recognition can improve the word error rate (WER). In this work, an autocorrelation based voicing measure is tested in combination with different state-of-the-art acoustic features. Experiments carried out on two large-vocabulary tasks have shown that using an additional voicing measure improves even the performance of the vocal tract length normalized (VTLN) MFCC feature.

- *Spectrum derivative measure*: The novel spectrum derivative measure was first published in (Kocharov et al., 2005). In this work, the spectrum derivative feature is investigated in detail on different small- and large-vocabulary corpora. Recognition results have shown that combination of state-of-the-art acoustic features with the spectrum derivative measure improves the WER significantly.
- *Linear discriminant analysis*: In former publications, linear discriminant analysis (LDA) was used in single- and multi-feature speech recognition systems. In this work, the application of LDA to acoustic feature combination is reviewed in detail. Experiments performed on small- and large-vocabulary corpora are presented which have shown that combination of increasing numbers of auditory and articulatory motivated acoustic features can improve the recognition accuracy significantly.
- *Discriminative model combination*: In earlier publications, discriminative model combination (DMC) was used to combine different acoustic and language models. There were also attempts at applying DMC to acoustic feature combination. In this work, LDA based feature combination is nested into DMC. The nested setup leads to significant improvements in WER compared to the best underlying single LDA combined system.

The remainder of this work is organized as follows: In the subsequent sections,

we review publications closely related to this work. A review of the implementation of the MFCC, PLP, and MF-PLP features is given in Section 2 along with a short summary of the implementation of vocal tract length normalization (VTLN) used in the experiments presented here. Detailed descriptions of the voicing and the spectrum derivative measures are presented as well. The LDA and the DMC based feature combination methods are described in Section 3.1 and 3.2, respectively.

1.1 Acoustic Feature Extraction

In this section, a review of state-of-the-art auditory and articulatory motivated feature extraction techniques is given which are close related to methods investigated in this paper.

The most widespread acoustic feature, the Mel frequency cepstrum coefficients (MFCC), was first introduced in (Davis and Mermelstein, 1980). The perceptual linear predictive (PLP) feature introduced in (Hermansky, 1990) is based on ideas similar to the MFCCs. Nevertheless, there are major differences in data flow and in recognition performance as well. The third fairly widespread auditory based MF-PLP feature was derived from the two aforementioned ones, as described in (Woodland et al., 1997). The MF-PLP feature uses a Mel scale triangular filter bank embedded into the data flow of the PLP feature.

Besides methods processing the short-term magnitude spectrum, new acoustic features have been proposed recently which focus on the short-term phase spectrum. In (Paliwal and Alsteris, 2005), human perception experiments have

shown that the phase spectrum contributes to speech intelligibility even for windows of less than 1s. On a small-vocabulary task, significant reduction in word error rate (WER) has been presented in (Schlüter and Ney, 2001) by using an LDA combination of the MFCC and a set of features derived from the short-term phase spectrum.

Also, acoustic features derived from the group delay function have been investigated in different speech applications. In (Hegde et al., 2005), significant improvements in accuracy have been reported when combining a modified group delay function based feature with MFCCs.

Applications of articulatory models have already been intensively studied in speech recognition systems. In (Welling and Ney, 1996), one of the first recognition systems was presented which use formant frequencies as acoustic features. In (Holmes et al., 1997), formant frequencies were used in combination with the MFCC feature. Using a simple acoustic model, significant improvements in WER were obtained on a connected-digit recognition task when adding the formant based features. Besides formants, the voicing feature is one of the most intensively researched articulatory features. In rule-based speech recognition systems, voiced-unvoiced detection was used as one of the acoustical features. In (Atal and Rabiner, 1976), a voiced-unvoiced-silence detection algorithm is proposed using statistical approaches. A voicing measure instead of a voiced-unvoiced decision is described in (Thomson and Chengalvarayan, 1998). The authors presented results obtained by using an autocorrelation based voicing measure along with liftered cepstral coefficients. Using the concatenated features, a large relative improvement in WER was obtained by applying discriminative training. Different voicing measure extraction methods are compared in (Zolnay et al., 2003). Recognition tests were carried out

by using the different voicing measures along with the MFCC feature. In (Graciarena et al., 2004), the entropy of the high order cepstrum is used to extract voicing information. Recognition tests showed significant improvement in WER when the entropy based voicing feature was combined with an autocorrelation based one and with the MFCC feature. A sub-band based periodic and aperiodic feature set is applied to the Aurora-2J corpus in (Ishizuka and Miyazaki, 2004). Significant improvements in WER are reported when comparing the proposed feature set with the baseline MFCCs. A novel articulatory motivated feature has been proposed recently in (Kocharov et al., 2005) providing information on the distinction between obstruents and sonants. The *spectrum derivative* feature has been tested in combination with the MFCC feature leading to significant improvements in WER on small- and large-vocabulary tasks. The spectrum derivative feature captures the intensity of changes of the magnitude spectrum over the frequency axis. Similarly, the derivative of magnitude spectrum formed the basis of acoustic features for speech recognition in (Paliwal, 1999) and (Nadeu et al., 2001). A measure similar to the spectrum derivative feature is proposed in (A. H. Gray and Markel, 1974). This measure quantifies the flatness of magnitude spectrum and it has been developed to give insight into the whitening process of linear prediction.

Articulatory information can also be successfully utilized to improve the MFCC feature itself. In (Gu and Rose, 2001), the magnitude spectrum of harmonics is emphasized leading to a large improvement in WER on an isolated digit string recognition task. The idea of scaling the frequency axis of the speech signal to account for gender specific variation of the vocal tract was first proposed in (Wakita, 1977). Meanwhile, Vocal Tract Length Normalization became a standard method in the speech recognition community.

1.2 Acoustic Feature Combination

The goal of acoustic feature combination is to exploit mutually complementary classification information provided by different features. Acoustic feature combination can be carried out at different levels of a speech recognition system. In the following, we review publications closely related to linear discriminant analysis (LDA) and discriminative model combination (DMC) based feature combination.

Combination of acoustic features can be performed directly on the level of feature vectors using LDA. In this approach, different acoustic feature sets are combined by means of an optimal linear transformation. In (Häb-Umbach and Ney, 1992), LDA was used successfully to find an optimal linear combination of successive vectors of a single-feature stream. The combination of different cepstral features was tested by using LDA in (Häb-Umbach and Loog, 1999), however, without significant improvements in WER compared to using the MFCCs alone. Significant reduction in WER are presented using the LDA based feature combination in (Schlüter and Ney, 2001) when combining MFCCs and a set of phase features and in (Zolnay et al., 2002) when combining MFCCs with a voicing measure.

Combination of acoustic features can also be carried out at the level of acoustic probabilities. In this case, acoustic models trained on different feature sets produce probabilities which are combined in a log-linear manner. Log-linear model combination has already been applied to different problems in speech recognition. In (Tolba et al., 2002), acoustic features were combined by the means of log-linear modeling. The combination of the MFCCs with main

spectral peak features led to a significant reduction in WER on a noisy small-vocabulary task. Word error minimizing training of log-linear model weights for speech recognition models was proposed in (Beyerlein, 1997). Discriminative model combination (DMC) applied to 5 acoustic and language models (within-word and across-word acoustic models, bigram, trigram, and fourgram language models) led to a significant improvement in WER, compared to the best pairwise combinations, as described in (Beyerlein, 1998). An application of DMC to acoustic feature combination is published in (Häb-Umbach and Loog, 1999). Significant improvements in WER were obtained by the combination of state-of-the-art cepstral features.

2 Signal Analysis

In this section, the feature extraction methods used in the experiments presented are described. First the Mel frequency cepstrum coefficients (MFCC) are described, followed by the perceptual linear predictive (PLP) features, and the MF-PLP feature. Subsequently, vocal tract length normalization (VTLN) is shortly reviewed. Finally, two articulatory motivated features are presented, the autocorrelation based voicing feature and the spectrum derivative feature.

2.1 Mel Frequency Cepstral Coefficients

The data flow of the Mel frequency cepstral coefficients (MFCC) feature extraction is depicted on Fig. 1. Every 10ms, a Hamming window is applied to preemphasized 25ms segments and fast Fourier transform is applied along with an appropriate zero padding. The obtained spectral magnitudes are in-

tegrated within 20 triangular filters arranged on the Mel-frequency scale. The filter output is the logarithm of the sum of the weighted spectral magnitudes. The number of filters depends on the sample rate, 15 for 8kHz and 20 for 16kHz. Subsequently, a discrete cosine transform is applied to decorrelate the filter bank outputs. The optimal number of cepstrum coefficients depends on the corpus, see Table 1. Finally, normalization steps are applied to account for variations in the recording channel. Here, cepstral mean subtraction and energy normalization are carried out either utterance-wise or using a sliding window.

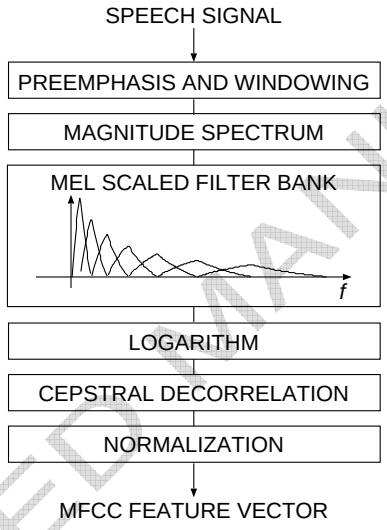


Fig. 1. Block diagram of Mel Frequency Cepstral Coefficients.

2.2 Perceptual Linear Predictive Analysis

The motivation of the Perceptual Linear Predictive (PLP) feature, proposed in (Hermansky, 1990), is similar to the one of the MFCCs. As depicted on Fig. 2, every 10ms, a Hamming window is applied to the speech signal. Unlike in the MFCC method, a window length of 20ms is used. Fast Fourier transform is applied and the resulting spectral magnitudes are integrated within 20

trapezoidal filters arranged on the Bark-frequency scale. The filter output is the weighted sum of the spectral magnitudes. The number of filters depends on the sample rate, 15 for 8kHz and 20 for 16kHz. The filter bank is virtually extended by two more filters, centered at frequency 0 and at sample-rate/2. Since these filters reach far beyond the valid frequency range, their output is discarded and replaced by the value of the right and left neighbor, respectively. Equal loudness preemphasis is applied to the extended filter bank outputs followed by the application of the intensity loudness law. Next, the cepstrum coefficients are derived from an all-poles approximation of the output of the intensity loudness law. For this, autocorrelation coefficients are calculated by applying the inverse discrete Fourier transform to the output of the intensity loudness law. To obtain the cepstrum coefficients, the autocorrelation coefficients are transformed to the gain and to autoregressive coefficients by using the Levinson-Durbin recursion. Instead of regenerating the smoothed all-poles approximation of the output of the intensity loudness law, the cepstrum coefficients are computed directly by applying a simple recursion. The zeroth cepstrum coefficient is explicitly set to the logarithm of the square of the gain. Finally, the resulting cepstrum coefficients are normalized as described in Section 2.1.

2.3 PLP Derived from Mel Scale Filter Bank

In this method, the MFCC and PLP techniques are merged into one algorithm generating the MF-PLP feature. As shown in Fig. 3, the Mel scale triangular filter bank taken from the MFCC algorithm is applied here to the power spectrum instead of the magnitude spectrum. Subsequently, cepstrum coefficients

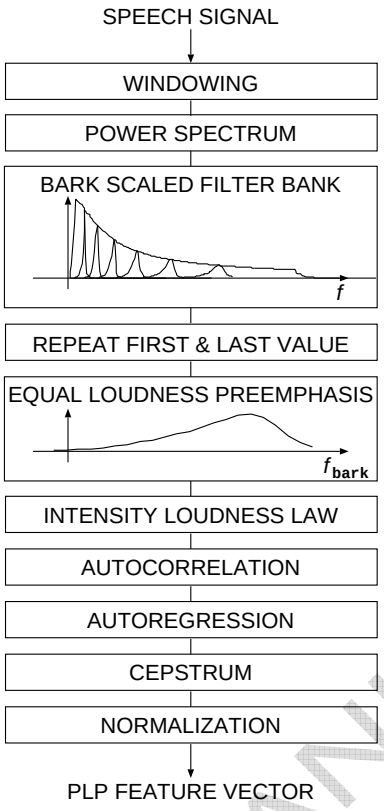


Fig. 2. Block diagram of Perceptual Linear Predictive Analysis.

are computed as described for the extraction of PLP features, where the copying of the outermost filters and the equal loudness preemphasis is skipped. The dynamic range of the filter bank outputs is compressed by the intensity loudness law. The cepstrum coefficients are calculated from the output of the intensity loudness law via the all-poles approximation as described in Section 2.2. Finally, a normalization is applied as described in Section 2.1.

2.4 Vocal Tract Length Normalization

A considerable part of the variability in the speech signal is caused by speaker dependent differences in vocal tract length. Vocal tract length normalization tries to account for this effect by warping the frequency axis of the power

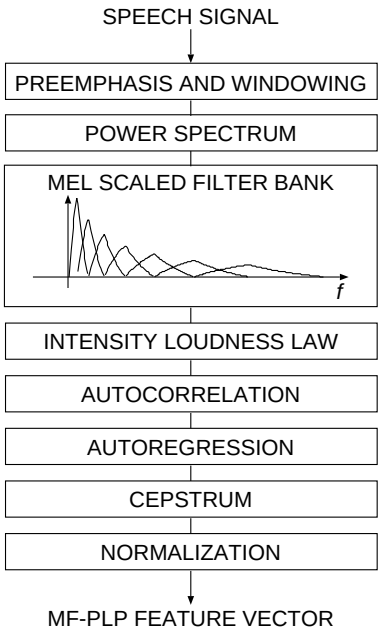


Fig. 3. Block diagram of MF-PLP feature.

spectrum. In a simplified model, the human vocal tract is treated as a straight uniform tube of length L. According to this model, a change in L by a certain factor α^{-1} results in a scaling of the frequency axis by α. Thus, for this model, the frequency axis should be scaled linearly to compensate for the variability caused by different vocal tracts of individual speakers. The warping of frequency axis can be implemented similar to the Mel warping in the MFCC data flow. Instead of a separate warping step requiring interpolation of the magnitude spectrum, the linear warping function is nested into the Mel warping function. The nested warping function can simply be integrated into the filter bank. The algorithm of the filter warping remains unchanged. The only difference is that the Mel warping function is replaced by the nested warping function. The estimation of the warping factors in test either is carried out using the maximum likelihood estimation on a preliminary recognition pass (Lee and Rose, 1996) or is based on text-independent Gaussian Mixture Models without the need of a first recognition pass (Welling et al., 2002). In

this work, maximum likelihood estimation of the warping factors has been used.

2.5 Voicing Feature

Voicing represents an important characterizing feature of phonemes. Therefore, a method explicitly extracting the degree of voicing from a speech signal can be expected to improve discrimination of phonemes and consequently to improve recognition results. Here, the goal is to produce a continuous measure representing the degree of periodic vibration of the vocal cords instead of the implementation of a voiced-unvoiced decision algorithm. The oscillation of the vocal chords produces quasi periodic segments in the speech signal. Common motivation of the voicing extraction methods is to quantify this periodicity. In (Zolnay et al., 2003), three methods were compared which produce voicing measures describing the degree of periodicity of a speech signal in a given time frame. The *harmonic product spectrum* based method measures the periodicity of a time frame in the frequency domain while the *autocorrelation* based and the *average magnitude difference* based methods operate in the time domain. Since the comparison did not show any significant differences, in this work, the autocorrelation based method has been used.

2.5.1 Extraction Algorithm

Assume the unbiased estimate of the autocorrelation $\tilde{R}_t(\tau)$ for some time frame t and a shift τ:

$$\tilde{R}_t(\tau) = \frac{1}{T-\tau} \sum_{\nu=0}^{T-\tau-1} x_t(\nu)\ x_t(\nu+\tau), \tag{1}$$

where T is the length of a time frame. The autocorrelation of periodic signals of frequency f attains its maximum not only at $\tau = 0$ but also at integer multiples of the period, i.e. for $\tau = \frac{k}{f} \quad k = 0, \pm 1, \pm 2, ...$ Therefore, a peak in the range of practically relevant pitches with a value close to $R_t(0)$ is a strong indication of periodicity. In order to produce a bounded measure of voicing, the autocorrelation is divided by $\tilde{R}_t(0)$. The resulting function provides values mainly in the interval $[-1..1]$ nevertheless because of the unbiased estimate, theoretically any value is possible. The voicing measure v_t is thus the maximum value of the normalized autocorrelation in the interval of practically relevant pitch periods [2.5ms..12.5ms]:

$$v_t = \frac{\max\limits_{2.5\text{ms}\cdot f_s \leq \tau \leq 12.5\text{ms}\cdot f_s} \tilde{R}_t(\tau)}{\tilde{R}_t(0)} \tag{2}$$

where f_s denotes the sample rate. Values of v_t close to 1 indicate voicing, values close to 0 indicate voiceless time frames. Fig. 4 summarizes the necessary steps to extract the voicing measure. The autocorrelation function is determined

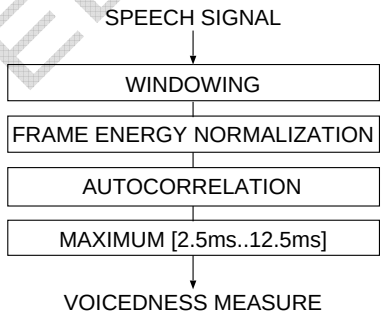


Fig. 4. Block diagram of voicing measure.

every 10ms on speech segments of 40ms length. The window length has been optimized empirically on small- and large-vocabulary corpora. The optimal value of 40ms corresponds to former results, cf. (Rabiner and Schafer, 1979, p. 318). The frame energy normalization ensures that $\tilde{R}_t(0) \equiv 1$ such that the

division in (2) can be omitted. After calculating the unbiased autocorrelation for discrete lags in the interval $[2.5\text{ms} \cdot f_s, 12.5\text{ms} \cdot f_s]$, a simple linear search is used to find the maximal value. In this way, a one-dimensional voicing feature is generated every 10ms.

2.5.2 Analysis of the Voicing Feature

To analyze the voicing measure v_t, histograms of the measure on a voiced-unvoiced sound pair have been estimated. For example, in Fig. 5, we have compared the pair of fricatives /v/-/f/ which phonetically differ only by the type of excitation (i.e. state of the vocal cords). The voicing histogram of both phonemes has been estimated on values aligned to any of the states of the triphones with the given phoneme as central phoneme. As shown in Fig. 5, the voicing measure can effectively contribute to the discrimination of voiced-unvoiced sound pairs.

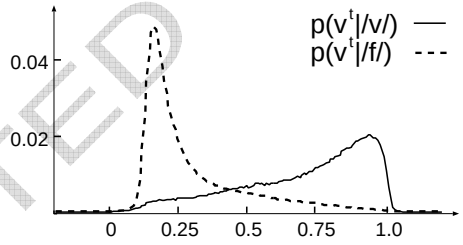


Fig. 5. Histograms of the voicing measure v_t for the voiced fricative /v/ and its unvoiced counterpart /f/ estimated on the *VM II* corpus (cf. Sec. 4.1).

2.6 Spectrum Derivative Feature

The spectrum derivative feature was first introduced in (Kocharov et al., 2005) to distinguish consonants from two articulatory classes: obstruents and so-

nants. From a phonetic point of view, these two classes differ by the presence of formants. In the magnitude spectrum of sonants, we can observe peaky formant-like structures. However, obstruents manifest in a flat and noisy magnitude spectrum. Hence, a feature summarizing the intensity of changes of the magnitude spectrum over the frequency axis can help to distinguish both phonetic classes.

2.6.1 Extraction Algorithm

The spectrum derivative feature is a measure calculated as the absolute sum of the first order derivatives of the magnitude spectrum. The extraction procedure is shown in Fig. 6. A Hamming window is applied to preemphasized

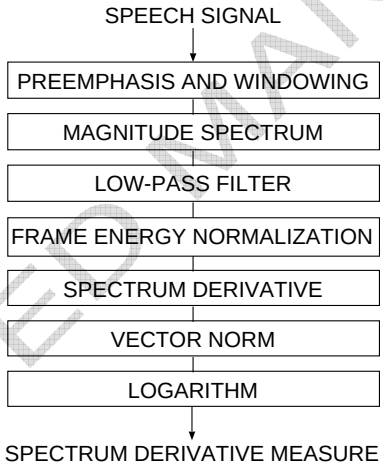


Fig. 6. Block diagram of spectrum derivative measure.

speech segments. The frame shift is chosen to 10ms. The window length has been optimized empirically in a range between 15ms and 90ms. The best results have been obtained by using 25ms window, as used for MFCC generation. The magnitude spectrum $X_t[n]$ of time frame t is calculated by using FFT along with an appropriate zero padding. The preprocessing of the magnitude spectrum begins with discarding the high frequency magnitudes. The cut-

off frequency is chosen at 1kHz. Comparative tests on different corpora have shown that processing the lower part of the magnitude spectrum only gives the best recognition results. Nevertheless, further experiments are necessary to understand the effects of filtering. In the next step, the filtered magnitude spectrum $\hat{X}_t[n]$ is energy normalized to account for frame energy variation. Experiments have been carried out by using frame-wise and utterance-wise energy normalization. Best recognition results have been obtained by applying energy normalization to every time frame:

$$\tilde{X}_t[n] = \frac{\hat{X}_t[n]}{\sqrt{\hat{X}_t[0]^2 + \hat{X}_t[\frac{N}{2}]^2 + 2\sum_{n=1}^{N/2-1} \hat{X}_t[n]^2}}, \tag{3}$$

where t denotes the time frame, n denotes the discrete frequency, and N is the number of FFT points. The first order derivative $a_t[n]$ is calculated over the normalized magnitude spectrum $\tilde{X}[n]$:

$$a_t[n] = \tilde{X}_t[n] - \tilde{X}_t[n-1], \tag{4}$$
$$a_t[0] \equiv 0. \tag{5}$$

Finally, the spectrum derivative feature is a continuous measure s_t calculated as the logarithm of the absolute sum of the discrete first order derivatives:

$$s_t = \log\left(\sum_{n=0}^{N/2} |a_t[n]|\right). \tag{6}$$

Note that this method can be straightforward extended to using higher order derivatives. The measure can be calculated for every higher order derivative of the magnitude spectrum as well. Nevertheless, experiments has not shown any consistent additional improvement in WER when using the higher order derivatives on top of the first order derivative.

2.6.2 Analysis of the Spectral Derivative Feature

In order to analyze the spectrum derivative feature, histograms of the spectrum derivative measure have been generated for a specific phoneme pair. Fig. 7 depicts distributions of s_t on the exemplary phoneme pair /v/ and /s/, which, phonetically, differ by their sonority. The histogram estimation has been carried out similar to the voicing feature described in Section 2.5.2.

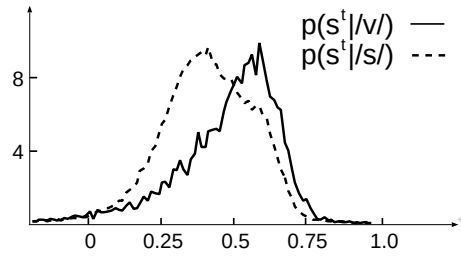


Fig. 7. Histograms of spectrum derivative measure s_t for the sonant consonant /v/ and the obstruent consonant /s/ estimated on the *VM II* corpus (cf. Sec. 4.1).

3 Feature Combination

3.1 LDA Based Feature Combination

The linear discriminant analysis (LDA) based feature combination approach can be used to combine different acoustic feature vectors directly. In (Häb-Umbach and Ney, 1992), LDA was first 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 $x_t^{f_i}$ extracted by different algorithms 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, the 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 subspace:

$$y_t = V^T \cdot \left[\left[x_{t-L}^{f_1}, \ldots, x_{t-L}^{f_F}\right], \ldots, \left[x_t^{f_1}, \ldots, x_t^{f_F}\right], \ldots, \left[x_{t+L}^{f_1}, \ldots, x_{t+L}^{f_F}\right]\right]^T, \quad (7)$$

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 both in training and in recognition.

3.2 DMC Based Feature Combination

Discriminative model combination (DMC) was first proposed in (Beyerlein, 1997). This method provides a flexible framework for log-linear combination of acoustic and language models for speech recognition. In the following, DMC is shortly reviewed and the application of DMC to acoustic feature combination is described. The DMC based approach combines acoustic features indirectly via log-linear combination of acoustic models for each acoustic feature. The log-linear model weights are trained by using a discriminative criterion minimizing word error rate.

The basic idea of DMC is to modify the modeling of the posterior probability $P(W|X)$ in Bayes' decision rule:

$$W_{opt} = \underset{W}{\operatorname{argmax}}\, P(W|X). \quad (8)$$

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

$$P(W|X) = \frac{P(W)P(X|W)}{\sum_{W'} P(W')P(X|W')}. \tag{9}$$

In case of discriminative model combination, the posterior probability is generalized to a log-linear distribution:

$$P(W|X) = \frac{e^{-\sum_i \lambda_i g_i(W,X)}}{\sum_{W'} e^{-\sum_i \lambda_i g_i(W',X)}} \tag{10}$$

When applying log-linear modeling to speech recognition, the basic feature function types are negative logarithms of probability distributions:

- language model: $g_{lm}(W,X) = -\log P(W)$,
- acoustic model: $g_{am,f_i}(W,X) = -\log P_{f_i}(X_{f_i}|W)$.

In order to combine different acoustic features, we redefine X to be a sequence of tuples containing the time-synchronous acoustic feature vectors $x_t^{f_i}$ instead of a sequence of single feature acoustic observation vectors. Furthermore, we introduce separate acoustic feature functions $g_{am,f_i}(W,X)$ for each acoustic feature f_i. Theoretically, every feature function receives all the different acoustic feature vectors. Nevertheless, in our system, the acoustic feature functions make only use of the underlying acoustic feature f_i. Consequently, the Bayes' decision rule for log-linear feature combination using a single language model and for each acoustic feature a separate acoustic model can be written as:

$$W_{opt} = \underset{W}{\operatorname{argmax}}\, P(W)^{\lambda_{lm}} \prod_i P_{f_i}(X_{f_i}|W)^{\lambda_{f_i}}. \tag{11}$$

3.2.1 DMC Training Process

The training of a DMC system consists of two major steps: independent training of the parameters of each feature function $g_i(W,X)$ and discriminative

training of the log-linear model weights λ_i. In this work, the negative logarithms of the probability distributions have been used as feature functions. In order to train the parameters of the language and acoustic model distributions, standard maximum likelihood training has been performed. The training of model weights has been carried out in a discriminative manner minimizing word error rate. Detailed descriptions of the word error minimizing training of log-linear model weights can be found in (Beyerlein, 1998, 2000).

4 Experiments

4.1 Corpora and Recognition Systems

Experiments for acoustic feature combination have been carried out on a number of small- and large-vocabulary speech recognition tasks. Small-vocabulary tests have been performed on the *SieTill* corpus. The corpus consists of German continuous digit strings recorded over telephone line. Large-vocabulary experiments have been conducted on the *VerbMobil II* (*VM II*) corpus and on the English partition of the European parliament plenary sessions (*EPPS*) corpus from the 2005 TC-STAR ASR evaluation campaign. The *VM II* corpus consists of German conversational speech whereas the *EPPS* corpus contains plenary session speeches of the European Parliament in British English. The settings of the RWTH speech recognition systems for these corpora are summarized in Table 1.

4.2 Baseline Recognition Results

Baseline recognition tests have been carried out by using the state-the-of-art MFCC, MF-PLP, PLP, and VTLN features. Table 2 summarizes the results.

Although vocal tract length normalization can be applied to any of the MFCC, PLP, and MF-PLP features, in this paper, VTLN denotes the normalization of the MFCC feature. On the *EPPS* corpus, we used text-independent Gaussian Mixture Models for warping factor estimation. For the sake of faster recognition passes, we have used supervised warping factor estimation on the *VM II* corpus. Note that only slight or insignificant differences in WER can be found when comparing the supervised warping factor estimation with other unsupervised ones.

4.3 LDA Based Feature Combination

In this section, experimental results are presented which have been obtained by the LDA based combination of state-of-the-art features with the voicing and the spectrum derivative measures.

For the different corpora, the number of concatenated successive feature vectors (L) taken as input to LDA, and the dimension of the projected feature space, cf. (7), have been optimized using the MFCC feature and are given in Table 1. For a given corpus, the size of the projected feature vectors has been kept constant throughout different experiments to ensure comparable numbers of parameters and therefore comparability of recognition results. Nevertheless, the LDA input dimension increases with the number of feature sets combined,

therefore implying a slight increase of parameters for the LDA transformation matrix. Finally, note that LDA has been applied in baseline experiments using a single feature in the same way as in feature combination experiments.

4.3.1 Recognition Results for Voicing Feature Inclusion

The one-dimensional voicing measure has to be viewed as an auxiliary feature in contrast to the baseline MFCC, PLP, MF-PLP, or VTLN features. Therefore, the use of the voicing measure necessarily implies feature combination. Here, LDA based feature combination as described in Section 3.1 has been used to incorporate the voicing feature. Table 3 summarizes the results obtained by using a single additional voicing measure (V) on different corpora. The application of the voicing measure has led to consistent improvements in WER of 11% on the small- and 3% on the large-vocabulary corpora relative to the baseline features.

In order to analyze the effects of the additional voicing feature on recognition accuracy, the difference of two confusion matrices is shown in Table 4. The confusion matrices were obtained on the small-vocabulary German digit string recognition *SieTill* task considering correctly recognized and substituted utterances. In order to show the changes caused by using the additional voicing feature, the difference of the confusion matrices is presented rather than showing the single matrices separately. The confusion matrix obtained by using the LDA combined MFCC and voicing features has been *subtracted* from the one generated by using only the MFCC feature. Consequently, in the diagonal of the difference confusion matrix, negative elements show improvements and positive ones degradations. Naturally, negative off-diagonal elements indicate

degradations and positive ones improvements.

By introducing the voicing feature, the overall number of errors decreased by more than $\frac{1}{10}$. Nevertheless, locally also degradations can be found in the difference confusion matrix. Furthermore, the amount of confusions between the words '2' and '3' has increased. In this case, a voicing feature could only contribute to the first stop consonant, and it could be observed that in this case, the word '2' is favored at the cost of word '3', with a negligible effect on the overall change in error rate. Apart from small variations, in all other cases improvements could be observed.

4.3.2 Recognition Results for Spectrum Derivative Inclusion

Similar to the voicing feature, also the spectrum derivative feature has to be viewed as auxiliary. Therefore, results for the spectrum derivative feature have also been produced using LDA based feature combination, here using the MFCC and VTLN features. Table 5 shows the results obtained by using the single additional spectrum derivative measure (SD) on different corpora. Applying the spectrum derivative measure has resulted in improvements in WER of 11% on the *SieTill* and 3% on the *VM II* corpora relative to the baseline features. The spectrum derivative feature could not improve the WER significantly on the *EPPS* corpus.

4.3.3 Combining MFCC, VTLN, Voicing, and Spectrum Derivative Features

Finally, experiments to combine the MFCC, vocal tract length normalized MFCC (VTLN), voicing (V), and spectrum derivative (SD) features have been conducted using LDA. Table 6 summarizes the corresponding recognition re-

sults.

On the small-vocabulary *SieTill* task, an improvement of 11% in WER relative to the MFCCs has been obtained when adding the voicing feature. Extending the set of features by the spectrum derivative feature, the WER has further improved by relative 6%. Nevertheless, this improvement has turned out to be significantly less than the improvements obtained by the combination of solely the MFCC and the spectrum derivative features. This observation has been confirmed on the large-vocabulary corpora as well.

On the large-vocabulary tasks, the MFCC and the best performing VTLN features have been chosen as baselines. On the *VM II* corpus, the combination of the baseline features with the voicing measure has given a relative improvement of 3% in WER. Similar to the small-vocabulary task, adding the spectrum derivative feature results in further improvement. The improvement is 1% relative to the system using the additional voicing measure which is less than the improvement obtained when combining the spectrum derivative feature solely with the baseline features, cf. Table 5. On the *EPPS* corpus, the additional voicing measure has given improvements similar to the *VM II* corpus. Nevertheless, the additional spectrum derivative measure has not resulted in further improvements in WER.

4.4 Feature Combination using DMC on Top of LDA

Although DMC is applicable to any acoustic feature set, the focus of the recognition tests presented in this work has been on the combination of LDA combined features. The output of LDA can be interpreted as a new separate acoustic feature. Consequently, acoustic models trained on LDA output vectors can

be combined in a straightforward way using the DMC framework.

Recognition tests using DMC for feature combination have been carried out using the standard word-conditioned tree search algorithm with integrated log-linear model combination. The integration of the log-linear model into a standard search method facilitates a single-pass recognition. Weighting of the language model and the acoustic models based on the different acoustic features happens on demand and can be implemented in a straightforward way.

Table 7 shows recognition results obtained by nesting LDA into DMC. For each corpus, the first two lines represent the baseline results using the MFCC feature with and without speaker normalization, cf. Table 2. The second group shows the recognition results for the LDA based combination of MFCC and VTLN features with the voicing measure (V). In the final experiment, the acoustic models trained in the second group, i.e. LDA(MFCC+V) and LDA(VTLN+V), have been combined using DMC. On both corpora, the DMC based combination resulted in improvements in WER compared to the LDA(VTLN+V) system. The most remarkable improvement has been obtained on the evaluation set of the *EPPS* corpus. On this corpus, the DMC based feature combination has lead to an improvement in WER of 4% relative to the best LDA based system.

5 Summary and Outlook

In this paper, we have analyzed the combination of several acoustic features both on feature and on model level. Besides considering four state-of-the-art baseline acoustic features, we have investigated the extraction of two articula-

tory motivated features. An autocorrelation based voicing measure has been studied first followed by the novel spectrum derivative feature. The articulatory motivated features have been first tested separately in combination with one of the baseline MFCC, MF-PLP, PLP, or VTLN features. The combination has been carried out by using LDA. The conclusions are summarized as follows:

- Additional articulatory motivated features can improve the performance of state-of-the-art acoustic features.
- On the small-vocabulary *SieTill* task, combination of MFCC feature with one of the voicing or spectrum derivative features resulted in a relative improvement of 11% in WER compared to the MFCC feature alone.
- Improvements were obtained also on the large-vocabulary *VM II* and *EPPS* tasks. The additional voicing feature led to consistent improvements of up to 3% relative to using the baseline feature alone. On the *VM II* corpus, the use of the spectrum derivative feature resulted in improvements comparable to those obtained by using the voicing feature. However, no significant improvements in WER could be observed on the *EPPS* corpus.

The combination of baseline acoustic features with both of the voicing and spectrum derivative features gave further improvements in WER. Experimental results can be summarized as follows:

- Similar improvements were obtained by using the MFCC or the speaker adapted VTLN features as baseline.
- Improvements were consistent over the three different corpora.
- When adding the spectrum derivative feature, the relative improvements over systems already using the additional voicing feature were significantly

less than those obtained by combining the spectrum derivative feature with only the baseline features.

- Experiments on the small-vocabulary *SieTill* task gave a relative improvement of 16% in WER when adding the voicing and the spectrum derivative features to the MFCCs. On large-vocabulary tasks, improvements of up to 4% were obtained relative to the best performing single feature systems.

Finally, DMC was shown to improve recognition results starting from models based on LDA combined acoustic features. The conclusions are:

- LDA based feature combination can be nested into DMC in a straightforward way.
- DMC based combination of highly optimized feature combination setups can improve the WER significantly. On the *EPPS* corpus, the application of DMC gave a relative improvement of 4% in WER compared to the best LDA based system.
- The best recognition results were obtained by using the voicing measure in combination with the MFCC and the VTLN features. On both large-vocabulary corpora, the LDA based method nested into DMC gave relative improvements of 4% in WER compared to the best single feature systems (VTLN).

Future work includes the optimization of the articulatory features presented and further development of additional acoustic features. In particular, the spectrum derivative feature will be compared to a closely related spectral flatness feature proposed in (A. H. Gray and Markel, 1974). Furthermore, alternative combination methods and especially system combination on the level of the recognition output will be investigated for its potential in feature combination.

Acknowledgments

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

References

A. H. Gray, J., Markel, J. D., Jun. 1974. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. IEEE Transactions on Acoustics, Speech, and Signal Processing 22 (3), 207 – 217.

Atal, B. S., Rabiner, L. R., 1976. A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 24 (3), 201 – 212.

Beyerlein, P., Dec. 1997. Discriminative model combination. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop. Santa Barbara, CA, pp. 238 – 245.

Beyerlein, P., May 1998. Discriminative model combination. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 1. Seattle, WA, pp. 481 – 484.

Beyerlein, P., Oct. 2000. Diskriminative modellkombination in spracherkennungssystemen mit grossem wortschatz. Ph.D. thesis, RWTH Aachen University.

Davis, S., Mermelstein, P., Aug. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-

28 (4), 357 – 366.

Graciarena, M., Franco, H., Zheng, J., Vergyri, D., Stolcke, A., May 2004. Voicing feature integration in sri's decipher lvcsr system. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 1. Montreal, Canada, pp. 921 – 924.

Gu, L., Rose, K., May 2001. Perceptual harmonic cepstral coefficients for speech recognition in noisy environment. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Salt Lake City, UT, pp. 125 – 128.

Häb-Umbach, R., Loog, M., Sep. 1999. An investigation of cepstral parameterisations for large vocabulary speech recognition. In: Proc. European Conf. on Speech Communication and Technology. Vol. 3. Budapest, Hungary, pp. 1323 – 1326.

Häb-Umbach, R., Ney, H., Mar. 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 1. San Francisco, CA, pp. 13 – 16.

Hegde, R. M., Murthy, H. A., Gadde, V., Mar. 2005. Speech processing using joint features derived from the modified group delay function. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 1. Philadelphia, PA, pp. 541 – 544.

Hermansky, H., Jun. 1990. Perceptual linear predictive (plp) analysis of speech. Journal of the Acoustical Society of America 87 (4), 1738 – 1752.

Holmes, J. N., Holmes, W. J., Garner, P. N., Sep. 1997. Using formant frequencies in speech recognition. In: Proc. European Conf. on Speech Communication and Technology. Vol. 4. Rhodes, Greece, pp. 2083 – 2086.

Ishizuka, K., Miyazaki, N., May 2004. Speech feature extraction method repre-

senting periodicity and aperiodicity in sub bands for robust speech recognition. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 1. Montreal, Canada, pp. 141 – 144.

Kocharov, D., Zolnay, A., Schlüter, R., Ney, H., Sep. 2005. Articulatory motivated acoustic features for speech recognition. In: Proc. European Conf. on Speech Communication and Technology. Vol. 2. Lisboa, Portugal, pp. 1101 – 1104.

Lee, L., Rose, R., May 1996. Speaker normalization using efficient frequency warping procedures. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 1. Atlanta, GA, pp. 353 – 356.

Nadeu, C., Macho, D., Hernando, J., 2001. Frequency and time filtering of filter-bank energies for robust hmm speech recognition. Speech Communication 34, 93–114.

Paliwal, K., Sep. 1999. Decorrelated and liftered filter-bank energies for robust speech recognition. In: Proc. European Conf. on Speech Communication and Technology. Budapest, Hungary, pp. 85 – 88.

Paliwal, K., Alsteris, L., 2005. On the usefulness of stft phase spectrum in human listening tests. Speech Communication 45, 153 – 170.

Rabiner, L. R., Schafer, R. W., 1979. Digital Processing of Speech Signals. Prentice-Hall Signal Processing Series, Englewood Cliffs, NJ.

Schlüter, R., Ney, H., May 2001. Using phase spectrum information for improved speech recognition performance. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 1. Salt Lake City, UT, pp. 133 – 136.

Thomson, D., Chengalvarayan, R., May 1998. Use of periodicity and jitter as speech recognition feature. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 1. Seattle, WA, pp. 21 – 24.

Tolba, H., Selouani, S. A., O'Shaughnessy, D., May 2002. 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. Vol. 1. Orlando, FL, pp. 837 – 840.

Wakita, H., Apr. 1977. Normalization of vowels by vocal tract length and its application to vowel identification. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. ASSP-25. pp. 183 – 192.

Welling, L., Kanthak, S., Ney, H., Sep. 2002. Speaker adaptive modeling by vocal tract normalization. IEEE Transactions on Speech and Audio Processing 10 (6), 415 – 426.

Welling, L., Ney, H., May 1996. A model for efficient formant estimation. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 2. Atlanta, GA, pp. 797 – 800.

Woodland, P., Gales, M., Pye, D., Young, S., Apr. 1997. Broadcast news transcription using htk. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Vol. 2. Munich, Germany, pp. 719 – 722.

Zolnay, A., Schlüter, R., Ney, H., Sep. 2002. Robust speech recognition using a voiced-unvoiced feature. In: Proc. Int. Conf. on Spoken Language Processing. Vol. 2. Denver, CO, pp. 1065 – 1068.

Zolnay, A., Schlüter, R., Ney, H., Sep. 2003. Extraction methods of voicing feature for robust speech recognition. In: Proc. European Conf. on Speech Communication and Technology. Vol. 1. Geneva, Switzerland, pp. 497 – 500.

Table 1

Settings of the RWTH recognition system for the *SieTill*, *VM II*, and *EPPS* corpora.

	corpus name	*SieTill*		*VM II*		*EPPS*		
		train	eval	train	eval	train	dev	eval
	speech seg. [h]	11.6	11.7	61.5	1.6	40.8	3.7	3.5
	# speakers	362	356	857	16	154	16	36
lexicon	vocabulary size	11		10 157		54 265		
language model	type	zerogram		class-trigram		trigram		
			test		test		dev	eval
	perplexity		11		62.0		87	99
feature extraction	sample rate [kHz]	8		16		16		
	# MFCCs	12		16		16		
	LDA window	11		11		9		
	LDA output	30		45		45		
model units	type	whole-word		triphone		triphone		
	gender dep.	yes		no		no		
	across-word	no		yes		yes		
HMM topology	# states per unit	39 (ave.)		3		6		
	# silence states	1		1		1		
state tying	type	none		decision tree		decision tree		
	# GMMs	215		3 501		4 501		
emission modeling	# densities	7k		396k		446k		
	pooled covar.	yes		yes		yes		

Table 2

Word error rates for baseline acoustic features. *SieTill* and *VM II* corpora: results only on *evaluation* set. *EPPS* corpus: results on both *development/evaluation* sets.

corpus	acoustic feature	error rates [%] del	ins	WER
SieTill	MFCC	0.3	0.5	1.8
VM II	MFCC	4.5	2.9	21.0
	VTLN	3.8	2.9	19.1
	MF-PLP	5.2	2.3	21.0
	PLP	5.9	2.3	21.4
EPPS	MFCC	4.3/3.8	1.4/1.7	14.7/15.3
	VTLN	4.3/3.7	1.3/1.5	14.2/14.1
	MF-PLP	4.2/3.7	1.5/1.7	14.8/15.3
	PLP	4.3/3.5	1.6/1.8	15.4/15.8

Table 3

Word error rates obtained by the LDA based combination of baseline features (MFCC, VTLN, MF-PLP, PLP) and the voicing feature (V). *SieTill* and *VM II* corpora: results only on *evaluation* set. *EPPS* corpus: results on both *development/evaluation* sets.

corpus	baseline feature	V	error rates [%] del	ins	WER
SieTill	MFCC	no	0.3	0.5	1.8
		yes	0.3	0.4	1.6
VM II	MFCC	no	4.5	2.9	21.0
		yes	4.6	2.7	20.3
	VTLN	no	3.8	2.9	19.1
		yes	4.1	2.7	18.7
	MF-PLP	no	5.2	2.3	21.0
		yes	4.7	2.6	20.5
	PLP	no	5.9	2.3	21.4
		yes	4.6	3.0	20.6
EPPS	MFCC	no	4.3/3.8	1.4/1.7	14.7/15.3
		yes	3.9/3.4	1.5/1.9	14.3/14.8
	VTLN	no	4.3/3.7	1.3/1.5	14.2/14.1
		yes	4.0/3.3	1.5/1.6	13.8/14.0
	MF-PLP	no	4.2/3.7	1.5/1.7	14.8/15.3
		yes	3.7/3.2	1.7/2.0	14.3/15.2
	PLP	no	4.3/3.5	1.6/1.8	15.4/15.8

Table 4

Difference confusion matrix generated by subtracting the confusion matrix obtained by using the LDA based combination of the MFCC and voicing features from the one obtained by using solely the MFCC feature. Larger **improvements** and *degradations* are emphasized.

	spoken										
recognized	’1’	’2’	’3’	’4’	’5’	’6’	’7’	’8’	’9’	’0’	’zwo’
’1’ /aɪns/	**-12**	3	3	0	-1	2	1	0	1	0	-1
’2’ /tsvaɪ/	1	**-16**	*-16*	0	-1	0	0	0	1	0	0
’3’ /draɪ/	2	**9**	*6*	0	0	0	0	0	3	1	0
’4’ /fiːɐ/	-1	-1	0	-4	1	0	0	1	0	1	0
’5’ /fʏnf/	2	0	1	1	**-6**	-1	-1	0	-1	-2	0
’6’ /zɛks/	0	1	1	0	-1	-4	0	-1	1	0	0
’7’ /ziːbən/	1	-1	-1	-2	1	1	0	0	0	0	0
’8’ /axt/	0	0	3	0	0	-1	1	-2	0	0	0
’9’ /nɔʏn/	3	-1	4	0	3	0	0	0	**-6**	**10**	0
’0’ /nʊl/	0	2	1	0	2	0	0	0	-1	**-12**	1
’zwo’ /tsvoːɐ/	0	3	-1	2	1	2	0	0	1	1	-2

Table 5

Word error rates obtained by the LDA based combination of baseline features (MFCC, VTLN) and the spectrum derivative feature (SD). *SieTill* and *VM II* corpora: results only on *evaluation* set. *EPPS* corpus: results on both *development/evaluation* sets.

corpus	baseline feature	SD	error rates [%] del	ins	WER
SieTill	MFCC	no	0.3	0.5	1.8
		yes	0.3	0.4	1.6
VM II	MFCC	no	4.5	2.9	21.0
		yes	4.5	2.9	20.3
	VTLN	no	3.8	2.9	19.1
		yes	3.7	3.0	18.6
EPPS	MFCC	no	4.3/3.8	1.4/1.7	14.7/15.3
		yes	4.5/4.0	1.2/1.5	14.7/15.1
	VTLN	no	4.3/3.7	1.3/1.5	14.2/14.1
		yes	4.0/3.3	1.5/1.7	14.2/14.1

Table 6

Word error rates obtained by the LDA based combination of increasing number of acoustic features. *SieTill* and *VM II* corpora: results only on *evaluation* set. *EPPS* corpus: results on both *development/evaluation* sets.

corpus	acoustic feature	error rates [%] del	ins	WER
SieTill	MFCC	0.3	0.5	1.8
	+V	0.3	0.4	1.6
	+SD	0.2	0.3	1.5
VM II	MFCC	4.5	2.9	21.0
	+V	4.6	2.7	20.3
	+SD	4.4	2.9	20.1
	VTLN	3.8	2.9	19.1
	+V	4.1	2.7	18.7
	+SD	3.9	2.9	18.4
EPPS	MFCC	4.3/3.8	1.4/1.7	14.7/15.3
	+V	3.9/3.4	1.5/1.9	14.3/14.8
	+SD	4.2/3.7	1.4/1.6	14.4/14.9
	VTLN	4.3/3.7	1.3/1.5	14.2/14.1
	+V	4.0/3.3	1.5/1.6	13.8/14.0
	+SD	3.6/3.1	1.6/1.8	13.7/14.0

Table 7

Word error rate obtained by nesting of discriminative model combination (DMC) based and linear discriminant analysis (LDA) based feature combination methods. *VM II* corpus: results only on *evaluation* set. *EPPS* corpus: results on both *development/evaluation* sets.

corpus	acoustic features	error rates [%] del	ins	WER
VM II	MFCC	4.5	2.9	21.0
	VTLN	3.8	2.9	19.1
	LDA(MFCC+V)	4.6	2.7	20.3
	LDA(VTLN+V)	4.1	2.7	18.7
	DMC[(LDA(MFCC+V)+ LDA(VTLN+V)]	4.1	2.5	18.4
EPPS	MFCC	4.3/3.8	1.4/1.7	14.7/15.3
	VTLN	4.3/3.7	1.3/1.5	14.2/14.1
	LDA(MFCC+V)	3.9/3.4	1.5/1.9	14.3/14.8
	LDA(VTLN+V)	4.0/3.3	1.5/1.6	13.8/14.0
	DMC[(LDA(MFCC+V)+ LDA(VTLN+V)]	4.1/3.5	1.2/1.4	13.6/13.5