Articulatory Motivated Acoustic Features for Speech Recognition

Daniil Kocharov

Department of Phonetics
Faculty of Philology
Saint-Petersburg State University
199034 Saint Petersburg, Russia
kocharov@phonetics.pu.ru

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 continuous speech recognition. A novel articulatory motivated acoustic feature is introduced, namely the spectrum derivative feature. The new feature is tested in combination with the standard Mel Frequency Cepstral Coefficients (MFCC) and the voicedness features. Linear Discriminant Analysis is applied to find the optimal combination of different acoustic features. Experiments have been performed on small and large vocabulary tasks. Significant improvements in word error rate have been obtained by combining the MFCC feature with the articulatory motivated voicedness and spectrum derivative features: improvements of up to 25% on the small-vocabulary task and improvements of up to 4% on the large-vocabulary task relative to using MFCC alone with the same overall number of parameters in the system.

1. Introduction

Most automatic speech recognition systems use auditory motivated representation of the speech signal, e.g. Mel Frequency Cepstrum Coefficients (MFCC), Perceptual Linear Prediction (PLP), and variations of these methods. There have also been attempts at using articulatory information in the acoustic front-end, e.g. an autocorrelation based voicedness feature [1]. These experiments yielded significant improvements in word error rate when combining standard auditory motivated features with the articulatory ones. In this paper, we describe and investigate a novel articulatory motivated feature, namely the spectrum derivative feature.

Extraction and application of articulatory motivated features have already been intensively studied in speech recognition systems. The first related studies go back to rule based speech recognition. On a digit string recognition task, formant frequencies were first applied as acoustic features by [2]. In [3], formant frequencies have been used successfully in combination with the MFCC feature. Significant improvements in word error rate (WER) have been obtained on a connected-digit recognition task when using the additional articulatory motivated features. Acoustic feature derived from the group delay function have been researched in different speech applications. In [4], significant improvements have been reported when combining modified group delay function based feature with MFCCs. In [1], liftered cepstral coefficients have been concatenated with an autocorrelation based voicedness measure. Using the concatenated features, a large relative improvement in WER has been achieved by applying discriminative training. A significant reduction in WER has been presented using Linear Discriminant Analysis (LDA) based feature combination in [5] when combining MFCCs with a voicedness feature.

In this work, a novel acoustic feature is introduced. The feature was first developed to distinguish obstruent from sonant consonants. Sonants differ from obstruents by the presence of formant structure. Thus a measure summarizing the changes in the magnitude spectrum over the frequency axis can contribute to differentiating the phoneme classes above. The implementation is based on the derivatives of the magnitude spectrum over the frequency axis. In every time frame, the spectrum derivative feature is a vector of single measures derived from the first, second, third, and higher order derivatives of the magnitude spectrum.

We tested the spectrum derivative feature in combination with different acoustic features by a Hidden Markov Model (HMM) based recognition system. Experiments showed significant improvements in word error rate when using additional articulatory motivated features: relative improvements of up to 25% on the small-vocabulary task and relative improvements of up to 4% on the large-vocabulary task over the best optimized MFCC based systems.

The rest of the paper is organized as follows. In Section 2, details of the different feature extraction methods are described including the spectrum derivative feature. In Section 3, we review an LDA based feature combination algorithm. Experiments are presented in Section 4, followed by a summary in Section 5.

2. Signal Analysis

In this section, we present the feature extraction methods used in our speech recognition system. First we describe the standard Mel Frequency Cepstrum Coefficients (MFCC), followed by the autocorrelation based voicedness feature. Finally, we present the new spectrum derivative feature.

2.1. Baseline Feature Extraction

In this section, the standard MFCC signal analysis component of our speech recognition system is described. First we perform a preemphasis of the sampled speech signal. Every 10ms, a Hamming window is applied to pre-emphasized 25ms speech segments. We compute the short-term spectrum by Fast Fourier Transform (FFT) along with an appropriate zero padding (e.g. 256 points in the case of 8kHz sampling rate). Next, we compute the outputs of overlapping Mel scale triangular filters,

the number of which depends on the sampling rate and varies 15 to 20 in our system. 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 the cepstrum coefficients. The optimal number of cepstrum coefficients varies from 12 to 16 depending on the number of filters in the filter bank.

Subsequently, a cepstral mean and variance normalization is carried out in order to account for different audio channels. We distinguish two types of normalization: sentence-wise and session-wise. For sentence-wise recorded corpora, normalization is performed on whole sentences. In addition, the zeroth coefficient is shifted so that the maximum value within every sentence is zero (energy normalization). Session-wise recorded corpora consist of recordings containing several sequentially spoken sentences. For these corpora, normalization is carried out with a symmetric sliding window of 2s without energy normalization. In this way, a vector consisting of normalized cepstrum coefficients is computed every 10ms.

2.2. Voicedness Feature (V)

Voiced and unvoiced sounds form two complementary classes. Thus, a feature explicitly expressing the voicedness of a time frame can lead to better discrimination of phonemes and consequently to better recognition results. 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 $\tilde{R}^{\mathsf{t}}(t)$:

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

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, any value is theoretically 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]:

$$v^{\mathsf{t}} = \frac{\max\limits_{2.5\text{ms}\cdot f_s \le \tau \le 12.5\text{ms}\cdot f_s} \tilde{R}^{\mathsf{t}}, (\tau)}{\tilde{R}^{\mathsf{t}}(0)} \qquad (2)$$

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 in length. The segment length is larger than for MFCCs to increase the possible number of periods covered by a time frame. By applying (2) to the autocorrelation, a one dimensional voicedness feature is generated every 10ms.

2.3. Spectrum Derivative Feature (SD)

The spectrum derivative feature was first introduced to distinguish two articulatory classes: obstruent and sonant consonants. 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. Thus, a feature summarizing the intensity of changes of the magnitude spectrum over the frequency axis can help to differentiate these two phoneme classes.

In Section 2.3.1, we describe the extraction algorithm of the spectrum derivative feature. In Section 2.3.2, we analyze histograms of the feature estimated over different phoneme classes.

2.3.1. Extraction Algorithm

The spectrum derivative feature is a vector of measures. The measures are calculated as the absolute sum of different order derivatives of the magnitude spectrum. The extraction procedure is shown on Figure 1.

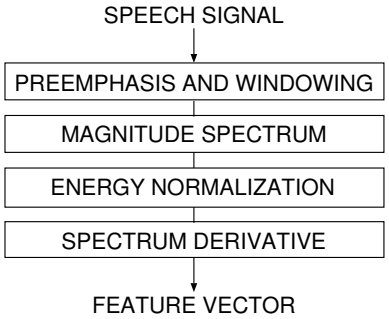

Figure 1: Extraction of spectrum derivative feature

A Hamming window is applied to preemphasized speech segments. The frame shift is chosen at 10ms. The frame length has been optimized empirically in a range from 15ms to 90ms. The best results have been obtained by using 25ms which is the same frame length used when generating MFCC features.

Normalization of the magnitude spectrum $X_t[n]$ is performed to account for different frame energies. Experiments have been carried out by using frame-wise and utterance-wise energy normalization. Best recognition results have been obtained by normalizing the energy of every time frame:

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

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

$$a_t^{(i)}[n] = a_t^{(i-1)}[n] - a_t^{(i-1)}[n-1], \qquad (4)$$

$$a_t^{(1)}[n] = \tilde{X}_t[n] - \tilde{X}_t[n-1], \qquad (5)$$

$$a_t^{(i)}[0] = 0. \qquad (6)$$

Finally, the spectrum derivative feature is a vector containing measures. The measures $S_t^{(i)}$ are calculated as the logarithm of the absolute sum of the i-th order derivative:

$$S_t^{(i)} = \log\left(\sum\nolimits_{n=0}^{N/2} |a_t^{(i)}[n]|\right). \qquad (7)$$

We carried out experiments including different order spectrum derivatives. The optimal number of spectrum derivatives depends highly on the corpora (see Section 4).

2.3.2. Histograms of Spectrum Derivative Feature

To analyze the spectrum derivative feature, we have generated histograms of the measure derived from the first order derivative for different phoneme pairs. Figure 2. depicts distributions of $S_t^{(1)}$ on the phoneme pair /v/ and /s/, which, from the point of view of phonetics, differ in their sonority. The histogram of a given phoneme has been estimated on values aligned to the central states of one of the triphones with the given phoneme as a central phoneme. Although the overlap of the histograms is rather large, the spectrum derivative feature can contribute to the differentiation of these two phonemes.

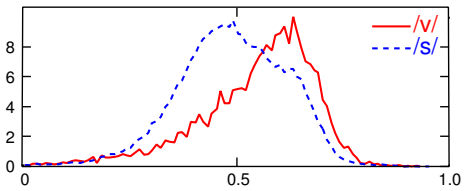

Figure 2: Histograms of the first order spectrum derivative measure for the phonemes /v/ and /s/ estimated over *VerbMobil II* corpus.

We have estimated histograms for another pair of phonemes to verify if the spectrum derivative feature contains information about voicedness of sounds. As Figure 3. shows, the spectrum derivative feature can not distinguish voiced-unvoiced phoneme pairs such as the alveolar fricatives /s/ and /z/.

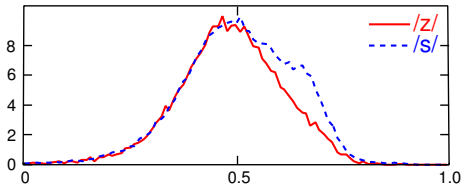

Figure 3: Histograms of the first order spectrum derivative measure for the phonemes /z/ and /s/ estimated over *VerbMobil II* corpus.

3. Feature Combination

We have used Linear Discriminant Analysis (LDA) to combine the different acoustic features. In [6], 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 combining the different acoustic features. For all time frames t, the MFCC feature vectors are concatenated with the voicedness and the spectrum derivative measures. In the second step, 11 successive concatenated vectors of the sliding window $t-5, t-4, ..., t, ..., t+4, t+5$ are concatenated again for all time frames t which makes up the large input vector of LDA. Finally, the combined feature vector is created by projecting the large input vector onto a smaller subspace. The projection matrix is determined by LDA such that it conveys the most relevant classification information. The resulting acoustic vectors are used as well in training and as in recognition.

The baseline experiments apply LDA in the same way. The only difference is in the size of the LDA input vector and thus in the number of columns of the projection matrix. The resulting feature vector has the same size to ensure comparable recognition results.

4. Experimental Results

Experiments have been performed on the small-vocabulary task *SieTill* and on the large-vocabulary task *VerbMobil II*.

4.1. Small-vocabulary Task

The small-vocabulary tests were performed on the *SieTill* corpus. The corpus consists of German continuous digit strings recorded over telephone line: approximately 43k spoken digits in 13k sentences in both the training and the test set. The number of female and male speakers is balanced.

The baseline recognition system for the *SieTill* corpus is built with whole word HMMs using continuous emission distributions. It can be characterized as follows:

- vocabulary of 11 German digits including 'zwo';
- gender-dependent whole-word HMMs;
- 214 distinct states for each gender plus one for silence;
- Gaussian densities, global pooled diagonal covariance;
- sample rate is 8kHz $\rightarrow$ number of filter banks is 15;
- 30 acoustic features after applying LDA;
- max. likelihood training using Viterbi approximation.

The baseline system has a word error rate of 1.89% which is the best reported so far using MFCC features and maximum likelihood training . In Table 1., the experimental results are summarized for using additional articulatory motivated features and for their combinations. Experiments were performed with single and with 32 Gaussian densities per mixture. A relative improvements of up to 20% has been obtained in both cases.

Table 1: Word error rates on the *SieTill* test corpus which were obtained by combining MFCC with different articulatory motivated features: voicedness measure (V) and the first order spectrum derivative measure (SD1). #dns gives the average number of densities per mixture.

#dns	acoustic features	error rates [%]		
		del	ins	WER
1	MFCC	0.50	0.71	3.83
	MFCC + V	0.40	0.46	3.23
	MFCC + SD1	0.35	0.73	3.50
	MFCC + V + SD1	0.44	0.45	2.93
32	MFCC	0.54	0.30	1.89
	MFCC + V	0.27	0.38	1.52
	MFCC + SD1	0.26	0.52	1.73
	MFCC + V + SD1	0.26	0.34	1.51

In Table 2., we present experimental results by using different number of spectrum derivative measures. Note that, when adding the i-th order spectrum derivative measure, we keep on using all the lower order measures.

4.2. Large-vocabulary Task

The large-vocabulary tests were conducted on the *VerbMobil II* corpus. The corpus consists of German large-vocabulary conversational speech: 36k training-sentences (61.5h) from 857

Table 2: Word error rates on the *SieTill* test corpus which were obtained by combining MFCC with voicedness measure (V) and spectrum derivative measures (SD) of different order. #dns gives the average number of densities per mixture. #SD denotes the number of spectrum derivatives measures (e.g. 3 means that the experiment has included the first, second, and third order spectrum derivative measures).

#dns	acoustic feature	#SD	error rates [%]		
			del	ins	WER
1	MFCC + V		0.40	0.46	3.23
	MFCC + V + SD	1	0.44	0.45	2.93
		2	0.43	0.39	2.93
		3	0.44	0.37	**2.92**
		4	0.45	0.34	2.98
		5	0.47	0.35	3.08
32	MFCC + V		0.27	0.38	1.52
	MFCC + V + SD	1	0.26	0.34	1.51
		2	0.26	0.37	1.60
		3	0.24	0.33	**1.45**
		4	0.25	0.34	1.51
		5	0.24	0.35	1.53

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-HMM topology with skip;
- 3501 decision tree based across-word triphone states including noise plus one state for silence;
- 385k gender independent Gaussian densities with global pooled diagonal covariance;
- sample rate is 16kHz $\rightarrow$ number of filter banks is 20;
- 33 acoustic features after applying LDA;
- max. likelihood training using Viterbi approximation;
- class-trigram language model, test set perplexity: 62.0.

The baseline system has a word error rate of 21.6% which is the best reported so far using MFCC features and across-word acoustic modeling . In Table 3., the experimental results are summarized for using different articulatory motivated features. Relative improvements in word error rate of up to 4% have been achieved by using both additional features. The bootstrap estimate of the probability of improvement is 98.9%.

Table 3: Word error rates on *VerbMobil II* test corpus which were obtained by combining MFCC with different articulatory motivated measures: voicedness measure (V) and the first order spectrum derivative measure (SD1).

acoustic features	error rates [%]		
	del	ins	WER
MFCC	5.2	2.8	21.6
MFCC + V	4.8	2.8	21.2
MFCC + SD1	4.8	2.9	21.5
MFCC + V + SD1	4.6	2.8	20.8

The optimization of the number of spectrum derivative measures has also been carried out on the *VerbMobil II* corpus. As shown in Table 4., the optimal number differs from the one obtained on the *SieTill* corpus. The inconsistent results show that there are further investigations necessary on the role of the higher order spectrum derivative measures.

Table 4: Word error rates on the *VerbMobil II* corpus which were obtained by combining MFCC with voicedness measure (V) and spectrum derivative measures (SD) of different order. #SD denotes the number of spectrum derivatives measures (e.g. 3 means that the experiment has included the first, second, and third order spectrum derivative measures).

acoustic features	#SD	error rates [%]		
		del	ins	WER
MFCC + V		4.8	2.8	21.2
MFCC + V + SD	**1**	4.6	2.8	**20.8**
	2	4.4	3.0	21.4
	3	5.1	2.8	21.3

5. Summary

In this paper, we have introduced a novel articulatory motivated acoustic feature. The new spectrum derivative feature aims to summarize the changes in the formant structure over the frequency axis. Recognition results showed that the spectrum derivative feature supplies mutually complementary information compared to the MFCC and the voicedness features. The best recognition results have been obtained by combining the MFCC, the voicedness, and the spectrum derivative features. Significant improvement in word error rate has been obtained on both of the recognition tasks: improvements of up to 25% on the small-vocabulary task *SieTill* and improvements of up to 4% on the large-vocabulary task *VerbMobil II* relative to the optimized systems using the MFCC feature alone.

In our future work, we will focus on better understanding the spectrum derivative feature. Important issues to investigate include the effect of spectrum smoothing and the role of the higher order derivatives.

6. Acknowledgements

This work was partly funded by the European Union under the integrated project TC-Star (Technology and Corpora for Speech to Speech Translation, IST-2002-FP6-506738, http://www.tc-star.org.

7. References

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

[2] L. Welling and H. Ney, "A model for efficient formant estimation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, Atlanta, GA, May 1996, pp. 797 – 800.

[3] J. N. Holmes, W. J. Holmes, and P. N. Garner, "Using formant frequencies in speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, vol. 4, Rhodes, Greece, Sept. 1997, pp. 2083 – 2086.

[4] R. M. Hegde, H. A. Murthy, and V. Gadde, "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, Mar. 2005, pp. 541 – 544.

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

[6] 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*, vol. 1, San Francisco, CA, Mar. 1992, pp. 13 – 16.