USING PHASE SPECTRUM INFORMATION FOR IMPROVED SPEECH RECOGNITION PERFORMANCE

Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology
D-52056 Aachen, Germany
Email: {schlueter,ney}@cs.rwth-aachen.de

ABSTRACT

In this work, new acoustic features for continuous speech recognition based on the short-term *Fourier* phase spectrum are introduced for mono (telephone) recordings. The new phase based features were combined with standard *Mel Frequency Cepstral Coefficients* (MFCC), and results were produced with and without using additional linear discriminant analysis (LDA) to choose the most relevant features. Experiments were performed on the *SieTill* corpus for telephone line recorded German digit strings. Using LDA to combine purely phase based features with MFCCs, we obtained improvements in word error rate of up to 25% relative to using MFCCs alone with the same overall number of parameters in the system.

1. INTRODUCTION

Acoustic features of today's state-of-the-art automatic speech recognition systems do not take into account phase spectrum information. Currently, signal analysis methods for speech recognition are mostly based on power spectra, where phase information has been removed. Representative and widely used basic signal analysis methods for speech recognition are both the *Mel Frequency Cepstral Coefficients* (MFCC) [1], and, as an extension the *Perceptual Linear Prediction* (PLP) coefficients [4]. For an extensive discussion on signal analysis for speech recognition see e.g. [9].It is a well known fact that the standard signal analysis methods are better suited to extract the attributes of quasi-stationary speech signals like vowels, than those of non-stationary signals like plosives.

In [8], investigations on the role of phase information for the human perception of intervocalic plosives have been presented. In one of the experiments presented therein, stimuli are constructed from vowel-consonant-vowel sequences, where the original short-term Fourier spectra are combined with different phase spectra for analysis window sizes between 10 and 30 ms. The results indicate that the short-term amplitude spectra cannot exclusively be specified by plosives. Moreover, the authors conclude that the perception of voicing for plosives relies strongly on phase information, whereas the perception of the place of articulation is mainly determined by amplitude information. It should be pointed out that these investigations were performed for analysis window sizes that are comparable to those used in speech recognition.

Until today, considerable effort has been devoted to including knowledge on the human auditory system into signal analysis methods for speech recognition, see e.g. [6]. As reported in [2], such approaches usually are concentrated on modeling the ascending pathway of the auditory periphery. In order to also incorporate efferent effects and the descending pathway in the auditory system, a feedback control model is introduced in [2]. The model tries to represent an efferent-induced depression mechanism which leads to considerable improvements in speech recognition performance under noisy conditions. After performing further experiments with phoneme recognition [3], the authors conclude that the improvements introduced by the feedback mechanism are mainly due to an improved discrimination of consonants, such as fricatives, affricates and plosives.

In neurobiological experiments it could be shown that an efferent mechanism based on the outer hair cells exists in the auditory system, which both leads to negative as well as positive mechanical feedback in the cochlea [7, 10].

This paper does not try to investigate or use detailed models of the human auditory system. Recently, it has even been questioned whether speech recognition really has benefitted from detailed simulation of properties of the human auditory system [5]. Rather than simulating physiological processing, we will try to point out functional *principles*, which introduce new information into speech recognition. These functional principles will be used to build models to compute acoustic features, which both reflect the structure represented by these principles and which comply with the requirements of an statistical pattern recognition system.

The aim of this work is to present methods to incorporate information from short-term *Fourier* phase spectra in order to improve speech recognition performance for mono (telephone) recording conditions. Two analysis methods will be discussed: acoustic features based on interference, and acoustic features, which are purely based on phase information. These phase features were combined with standard MFCC features, and results were produced with and without using additional linear discriminant analysis (LDA) to choose the most relevant features. Experiments were performed on the *SieTill* corpus for telephone line recorded German digit strings. Using LDA to combine purely phase based features with MFCCs, we obtained significant improvements in word error rate of up to 25% relative to using MFCCs alone with the same overall number of parameters in the system.

The rest of the paper is organized as follows. In Section 3, acoustic features based on frequency selective interference are introduced. In Section 4, purely phase based acoustic features are derived. Experiments will be presented in Section 5, followed by the conclusions in Section 6.

2. STANDARD SIGNAL ANALYSIS

In this section, the standard short-term power spectrum based signal analysis component of our speech recognition system is described. First we perform a preemphasis of the sampled speech signal. The preemphasized samples $d[n]$ are obtained from the original samples $s[n]$ by

$$d[n] \ = \ s[n] - s[n-1].$$

Every 10 ms, a Hamming window is applied to preemphasized 25 ms speech segments. We compute the short-term spectrum by a 256–point fast Fourier transform together with zero padding. For further processing we use the frequency range from 0 to 4 kHz, since the signal is sampled with 8 kHz. Next, we compute 15 mel scale triangular filters [15], where the mel scale is defined by

$$\mathrm{Mel}(f) \ = \ 2595 \log_{10} \left(1 + \frac{f}{700Hz}\right).$$

A filter bank is applied to the mel spectrum, in which each filter has a triangular bandpass frequency response with bandwidth and spacing determined by a constant mel frequency interval. For each filter the output is the logarithm of the sum of the weighted spectral magnitudes. Due to overlapping filters, filter bank outputs of adjacent filters are correlated. The covariance matrix of a vector consisting of the filter bank outputs has Toeplitz form. Thus the filter bank outputs are decorrelated by a discrete cosine transform [1]. $M = 12$ cepstrum coefficients c_m are computed from $N = 15$ filter bank outputs f_n by

$$c_m \ = \ \sum_{n=1}^{N} f_n \cos\left(\frac{\pi m(n-0.5)}{N}\right) \ , \ 0 \leq m < M.$$

Subsequently, a cepstral mean normalization is carried out for every utterance in order to account for different transfer functions. In addition, the zeroth coefficient is shifted so that the maximum value within every utterance is zero (energy normalization). Every 10 ms, a vector $y[t]$ consisting of $M = 12$ cepstrum coefficients at time t is computed. Each vector $y[t]$ is augmented by first–order regression coefficients [13], and the second–order regression coefficient of the energy. In this paper, a time difference of $\pm 3 \cdot 10$ ms was used for the calculation of the regression coefficients. The resulting acoustic vector $x[t]$ is used for recognition.

3. ACOUSTIC INTERFERENCE

According to the relation between the short-term phase spectrum information and the human perception of intervocalic plosives [8], the expectation is that the consideration of phase information for speech recognition bears at least the potential to improve the discrimination of plosives.

3.1. Interference

We recall the fact that there exists a feedback mechanism via the outer hair cells in the cochlea [7, 10]. Such a feedback loop will imply a time delay, which, as we suppose, will produce frequency selective phase differences. More explicitly, our hypothesis is that a feedback loop corresponding to a particular frequency band leads to interference with the original signal within this frequency band.

Let $x_t[k]$ and $\varphi_t[k]$ be the amplitude and phase in frequency channel k, obtained by a short-term *Fourier* transformation centered at time t after applying preemphasis and Hamming window to the original speech samples. Suppose that the feedback loop in frequency channel k introduces a phase shift of θ_k. Moreover, let η be the number of speech samples per time unit. The superposition of a signal at time t with amplitude $x_t[k]$ and phase $\varphi_t[k]$, and a signal at time $t - \Delta\tau$ with amplitude $x_{t-\Delta\tau}[k]$ and phase $\varphi_{t-\Delta\tau}[k]$ then leads to the amplitude $\xi_{t,t-\Delta\tau}[k]$ of the superimposed signal, as given in Eq. 1.

At this point, the questions remain how the time delay $\Delta\tau$ as well as the phase shift θ_k are to be chosen. Since the details of the above mentioned feedback loop are not completely known, these parameters will be optimized empirically so as to maximize speech recognition performance. For the time being, we chose a phase shift of $\theta_k = 0$. The value of the time delay $\Delta\tau$ will be combined with a smoothing in the time domain. The characteristics of the smoothing including the time delay will be varied in the corresponding experiments.

3.2. Smoothing

Let us define time intervals $\Delta\tau_i = i \cdot \Delta t$ with $i = -I, \ldots, I$ and a time step size Δt. Based on the superposition amplitudes $\xi_{t,t-\Delta\tau}[k]$ we define the following smoothed superposition feature $\xi_t[k]$:

$$\xi_t[k] = \sqrt{\frac{1}{2I+1} \sum_{i=-I}^{I} \xi^2_{t,t+\Delta\tau_i}[k]}.$$

In our experiments, smoothed superposition features were calculated every 10ms. In combination with the value of the time step size Δt, the value of I defines the range of the temporal smoothing. The step size has been chosen to be $\Delta t = 2ms$. The range has been set to $I = 2$, which corresponds to a time window of $2 \cdot I \cdot \Delta t = 8$ms.

3.3. Data Reduction

Subsequently, a mel filterbank identical to the case of the MFCCs is applied, where each output is the logarithm of the sum of the weighted smoothed superposition features, cf. Section 2. Then, 15 smoothed superposition cepstral coefficients are computed by applying a discrete cosine transform for the 15 outputs of the smoothed superposition mel filterbank. Finally, the MFCCs based on the amplitude, as presented in Section 2, are subtracted from the corresponding smoothed superposition cepstral coefficients, in order to enforce the phase contribution of these features. As for the case of standard MFCCs, a cepstral mean normalization is carried out on the superposition cepstrum for every utterance.

$$\xi_{t,t-\Delta\tau}[k] = \sqrt{x_t^2[k] + x_{t-\Delta\tau}^2[k] - 2x_t[k] \cdot x_{t-\Delta\tau}[k] \cdot \cos\left(\varphi_t[k] - \varphi_{t-\Delta\tau}[k] - \frac{2\pi k}{N} \cdot \eta \cdot \Delta\tau + \theta_k\right)}. \qquad (1)$$

4. ACOUSTIC PHASE FEATURES

In this section, we will present acoustic features, which are based purely on phase information.

4.1. Phase Dependence of Interference

In the case of interference, cf. Eq. (1), the only phase dependence is introduced by the cosine of the phase difference between two signals at times t and $t - \Delta\tau$, including the phase shift θ_k introduced by the feedback loop. This defines our base phase feature $\zeta_{t,t-\Delta\tau}[k]$ for frequency channel k:

$$\zeta_{t,t-\Delta\tau}[k] = \cos\left(\varphi_t[k] - \varphi_{t-\Delta\tau}[k] - \frac{2\pi k}{N}\cdot\eta\cdot\Delta\tau + \theta_k\right).$$

4.2. Phase Evolution

The evolution of phase over a restricted period of time could now be obtained by fixing a center time t and varying the time step $\Delta\tau$. In order to obtain an estimate of the local phase change, we calculate the average of the absolute difference between the base phase features for adjacent time steps $\Delta\tau_i, \Delta\tau_{i-1}$ over a restricted period of time , i.e. for $i = -I+1, \ldots, I$. The resulting feature $\zeta_t[k]$ will be called smoothed phase feature in the rest of this paper:

$$\zeta_t[k] = \frac{1}{2I}\sum_{i=-I+1}^{I}\left|\zeta_{t,t+\Delta\tau_i}[k] - \zeta_{t,t+\Delta\tau_{i-1}}[k]\right|.$$

In our experiments, smoothed phase features were calculated each 10ms. In all cases, the range $2 \cdot I \cdot \Delta t$ of the averaging has been chosen to be 20ms. The time step sizes considered were $\Delta t = 0.125$ms (equivalent to 1 sample) with $I = 80$, $\Delta t = 2$ms with $I = 5$, and $\Delta t = 10$ms with $I = 1$.

4.3. Data Reduction

Subsequently, a mel filterbank identical to the case of the MFCCs is applied, where each output is the logarithm of the sum of the weighted smoothed phase features, cf. Section 2. Then, 15 smoothed phase cepstral coefficients are computed by applying a discrete cosine transform for the 15 outputs of the smoothed phase mel filterbank. As for the case of standard MFCCs, a cepstral mean normalization is carried out on the smoothed phase cepstral coefficients for every utterance.

5. EXPERIMENTAL RESULTS

Experiments were performed on the *SieTill* corpus [11, 12] for telephone line recorded German continuous digit strings. The *SieTill* corpus consists of approximately 43k spoken digits in 13k sentences for both training and test. 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 is characterized as follows:

- vocabulary of 11 German digits including '*zwo*'
- gender-dependent whole-word HMMs, with every two subsequent states being identical
- for each gender 214 distinct states plus one for silence,
- Gaussian mixture emission distributions,
- global pooled diagonal covariance matrix,
- 12 cepstral features plus first order regression coefficients and the second order regression coefficient of the energy,
- optional LDA with three adjacent frames as input features.

The baseline recognizer applies ML training using the Viterbi approximation. A detailed description of the baseline system could be found in [14]. In particular cases, we also applied discriminative training. The corresponding training system is described in [11, 12] and the papers cited therein.

5.1. Results for Superposition Features

In Table 5.1, the experimental results are summarized for the case of single Gaussian densities using additional interference features in comparison to MFCCs only. Without using LDA, a relative improvement in word error rate of about 7% is obtained using additional interference features. On the other hand, with LDA, a minor relative improvement of only 2.5% is obtained. Therefore, we assume that the improvement without LDA is mainly due to the increased acoustic feature dimension. All in all, although some improvement could be observed, the interference features do not perform as well as expected.

Table 1: Word error rates on the *SieTill* test corpus obtained for standard MFCCs and for additional interference features (IF), as defined in Section 3. In these experiments, single Gaussian emission distributions were used. In case of LDA, the number of features gives the output dimension of the LDA transformation.

LDA	acoustic features		error rates [%]		
	type	#	del - ins	WER	SER
no	MFCC	25	1.07-0.78	4.33	10.87
	+IF	37	0.70-0.82	4.02	10.43
yes	MFCC	40	0.67-0.52	3.57	9.19
	+IF	40	0.66-0.49	3.48	9.10

5.2. Results for Phase Features

In Table 5.2, the experimental results using additional phase features in comparison to MFCCs only are summarized for the case of Gaussian mixture densities. For the case without LDA, experiments have only been performed using single Gaussian emission distributions. For a time step size of $\Delta t = 10ms$, a relative improvement in word error rate of more than 10% is obtained. On the other hand, for $\Delta t = 2ms$ and $0.125ms$, considerable deteriorations in word error rate could be observed. Since the Gaussian emission distributions are modeled with diagonal variances, this effect could be due to correlations introduced by the new features. For the improvement in case of $\Delta t = 10ms$ without LDA, it should again be noted that the feature dimension differs considerably and could partly be responsible for the improvement.

With LDA, experiments have been performed for single Gaussian densities, as well as mixture densities with empirically optimized number of densities per mixture. Using LDA, the results for the phase features are clear cut: for single densities and all choices of the time step size, relative improvements in word error rate of at least 12% are obtained for comparable numbers of acoustic features after LDA. Using mixture densities, the results remains the same: for time step sizes of both 2ms and 10ms relative improvements of at least 13% are obtained in comparison to the best result using standard MFCC features before LDA. Using ML training, the best result is 1.51% word error rate. Using subsequent *Discriminative Training* (DT) [12], this could even be improved to 1.28% word error rate, which compares to 1.67% word error rate using MFCC features only.

Table 2: Word error rates on the *SieTill* test corpus obtained for standard MFCCs and for additional phase features (PF), as defined in Section 4. In these experiments, Gaussian mixture emission distributions were used, where 'dns' gives the average number of densities per mixture. In case of LDA, the number of features is the output dimension of the LDA transformation.

LDA	dns	train crit.	acoust. feat. type	acoust. feat. #	Δt [ms]	error rates [%] del - ins	WER	SER
no	1	ML	MFCC	25	–	1.07-0.78	4.33	10.87
			+PF	40	10	0.67-0.61	3.85	9.94
					2	1.09-1.37	5.68	14.55
					1/8	1.04-1.17	5.08	12.85
yes	1	ML	MFCC	40	–	0.67-0.52	3.57	9.19
			+PF	40	10	0.65-0.32	3.14	8.31
					2	0.50-0.40	3.01	8.04
					1/8	0.66-0.43	3.14	8.30
	32	ML	MFCC	40	–	0.49-0.40	1.78	4.72
			+PF	40	10	0.39-0.31	1.54	4.16
					2	0.33-0.35	1.51	4.20
		DT	MFCC	40	2	0.41-0.37	1.67	4.50
			+PF	40	2	**0.36-0.18**	**1.28**	**3.55**

6. CONCLUSIONS

In this work, new acoustic features for continuous speech recognition based on the short-term *Fourier* phase spectrum were introduced for mono (telephone) recordings. The new phase based features were added to standard *Mel Frequency Cepstral Coefficients* (MFCC). Experiments were performed on the *SieTill* corpus for telephone line recorded German digit strings. Using standard MFCCs and LDA gave a baseline word error rates of 1.78% using ML training and 1.67% using discriminative training. Using additional acoustic features, which are purely based on short-term *Fourier* phase, the baseline results could be improved to word error rates of 1.51% absolute with ML training and finally 1.28% absolute using discriminative training. In the case of discriminative training the relative improvement in word error rate using short-term phase is 25% compared to MFCC features alone.

7. REFERENCES

[1] S. B. Davis, P. Mermelstein. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28, Nr. 4, pp. 357-366, August 1980.

[2] Y. Gao, T. Huang, S. Chen, J.-P. Haton. "Auditory Model Based Speech Processing," Proc. 1992 Intern. Conf. on Spoken Language Processing, Banff, Alberta, Canada, October 1992.

[3] Y. Gao, T. Huang, J.-P. Haton. "Central Auditory Model for Spectral Processing," Proc. 1993 Intern. Conf. on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 704-707, Minneapolis, MN, April 1993.

[4] H. Hermansky. "Perceptual Linear Predictive (PLP) Analysis of Speech," Journ. Acoustic Soc. America, Vol. 87, Nr. 4, pp. 1738-1752, June 1990.

[5] M. J. Hunt. "Spectral Signal Processing for ASR," Proc. *1999 Automatic Speech Recognition and Understanding (ASRU) Workshop*, Vol. 1, pp. 17-25, Keystone, CO, December 1999.

[6] J.-C. Junqua, J.-P. Haton. *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Norwell, MA, 1996.

[7] M. Kössl. "Sound Emission from Cochlear Filters and Foveae - Does the Auditory Sense Organ Make Sense?" Naturwissenschaften, Vol. 84, pp. 9-16, 1997.

[8] L. Liu, J. He, G. Palm. "Effects of Phase on the Perception of Intervocalic Stop Consonants," Speech Communication, Vol. 22, No. 4, pp. 403-417, 1997.

[9] J. W. Picone. "Signal Modeling Techniques in Speech Recognition," Proc. of the IEEE, Vol. 81, Nr. 9, September 1993.

[10] I. J. Russell, M. Kössl. "Modulation of Hair Cell Voltage Responses to Tones by Low-Frequency Biasing of the Basilar Membrane in the Guinea Pig Cochlea," Journ. of Neuroscience, Vol. 12, Nr. 5, pp. 1587-1601, May 1992.

[11] R. Schlüter, W. Macherey. "Comparison of Discriminative Training Criteria," Proc. 1998 Intern. Conf. on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 493-496, Seattle, WA, April 1998.

[12] R. Schlüter, W. Macherey, B. Müller, H. Ney. "Maximum Mutual Information and Maximum Likelihood Approach for Mixture Density Splitting," Proc. 1999 Europ. Conf. on Speech Communication and Technology, Vol. 4, pp. 1715-1718, Budapest, Hungary, September 1999.

[13] L. Welling, N. Haberland, H. Ney. "Acoustic Front-End Optimization for Large Vocabulary Speech Recognition," Proc. 1997 Europ. Conf. on Speech Communication and Technology, Vol. 4, pp. 2099-2102, Rhodos, Greece, September 1997.

[14] L. Welling, H. Ney, A. Eiden, C. Forbrig. "Connected Digit Recognition using Statistical Template Matching," Proc. *1995 Europ. Conf. on Speech Communication and Technology*, Madrid, Vol. 2, pp. 1483-1486, September 1995.

[15] S. J. Young. "HTK: Hidden Markov Model Toolkit V1.4," User Manual, Cambridge University Engineering Department, Cambridge, England, February 1993.