ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Filterbank-based Feature Extraction for Speech Recognition and Its Application to Voice Mail Transcription

Jun Huang[1] *and Mukund Padmanabhan*[2]

[1] University of Illinois, Urbana, IL61801, USA

[2] IBM T.J.Watson Research Center, Yorktown Heights, NY 10598, USA

ABSTRACT

In this paper, we propose a filterbank-based technique to extract more robust and discriminative features for the application of telephony speech recognition. First, we propose an extended Lerner grouping method to approximate the shape of the Mel filters in MFCC while reducing the cross-correlation between filterbank outputs. Then we used welch processing to reduce the variance of the spectral features while retaining the spectral resolution. Finally, we describe experiments where we augment the cepstral features with formant related features, computed using an adaptive filterbank. The new features represent the trajectory of the frequency components within different formant bands. Experimental results showed that the welch processing consistently improved the word error rate on a task of large vocabulary voice mail transcription and the formant related features provide higher discriminability than the MFCC features.

1. INTRODUCTION

Conventional speech features are extracted based on the power spectrum which contains information related to the source signal and the vocal tract. Typically, the power spectrum of a speech frame is computed via an FFT or a filter-bank analysis, and a number of studies have shown that perceptually based filter-bank analysis is more robust than an FFT-based representation [1],[2]. Conventional speech features include LPC, mel-frequency cepstral coefficients (MFCC), perceptually based linear prediction analysis (PLP), *etc.* In this paper, we describe some alternative feature extraction schemes for telephone speech recognition. First, we try to design filters that mimic the shape of the Mel filters (the MFCC features are generally extracted by summing up the FFT power spectrum outputs weighted by Mel filter coefficients, and the Mel filters could be thought of as a filterbank with triangular passband characteristics), but have much lower sidelobes. (Note that it is also possible to reduce the sidelobes by applying a Hamming window to the speech signal, and in fact most feature extraction techniques do use this method – the Lerner grouping of channels was investigated to see if it provided a reasonable alternative to standard feature extraction techniques). Then we investigate a welch processing method [6] to reduce the variance of the feature vectors while still retaining the temporal resolution. Then we propose a new set of features which represent the dynamics of the local energy concentration within different formant bands. This paper is organized as following. In section 2, we describe a lerner grouping technique to design bandpass filters whose passband reflects the shape of the Mel filters, and that also provide low sidelobes. In section 3, we introduce a welch processing technique to get smoother speech features. A new set of features based on locating spectral peaks is proposed in section 4. We present some experimental results in voice mail transcription and discriminant analysis (LDA) of the new feature in section 5.

2. LERNER GROUPED FEATURES

Lerner grouping is a method of realizing bandpass filters with relatively low sidelobes, by grouping together the outputs of uniformly spaced element bandpass filters. The lerner grouping design was originally proposed for realizing continuous-time filter-banks having almost linear-phase bandpass outputs with good stopband performance [8],[9]. In the original lerner grouping method, each bandpass filter was realized by a weighted sum of adjacent parallel second-order biquadratic filters with the weighting coefficients alternating in sign for adjacent resonators. All the weights were +- 1 except for the bandpass-edge resonators, for which the weights were +- 0.5. In our method, we use an "extended" lerner design technique in which each lerner weighting coefficient can have an arbitrary value constrained by the fact that the alternating sign condition is still maintained. In our experiment, we want to design a set of lerner grouping coefficients which has a passband similar to the Mel filters in MFCC feature extraction. The criterion to design the lerner coefficients is now given by:

$$\hat{\vec{l}}^{(k)}_{1\times N} = \arg\min_{\vec{L}^{(k)}_{1\times N}} \left\| \vec{l}^{(k)}_{1\times N} - \vec{h}^{(k)}_{1\times N} F^{-1}_{N\times N} \right\|^2, \quad k = 1,2,\cdots,K \qquad (1)$$

$$\text{subject to } l^{(k)}_{i\text{-}j} = l^{(k)}_{i+j},\ sign(l^{(k)}_{i}) \times sign(l^{(k)}_{i+1}) = -1 \qquad (2)$$

Where $\hat{\vec{h}}_{1\times N}^{(k)}$ is the desired impulse response of the *k*-th Mel filter in the time-domain; *K* is the number of channels; $F_{N\times N}$ is the DFT matrix; $\vec{l}_{1\times N}^{(k)}$ is the lerner grouping of coefficients for the *k*-th channel. This is a least-squares problem and we can get a closed form solution:

$$\vec{g}^{(k)} = \vec{h}^{(k)} F^{-1},\ \vec{g}^{(k)} = [g_1^{(k)}\ g_2^{(k)}\ \cdots\ g_N^{(k)}] \qquad (3)$$

$$\vec{l}^{(k)} = [l_1^{(k)}\ l_2^{(k)}\ \cdots\ l_N^{(k)}],\ l_i^{(k)} = g_i^{(k)} \qquad (4)$$

$$l_{i+j}^{(k)} = l_{i-j}^{(k)} = (-1)^j \times \tfrac{1}{2}(g_{i+j}^{(k)} + g_{i-j}^{(k)}),\ j = 1,2,\cdots,D \quad (5)$$

In our experiment, we use lerner coefficients computed in Eqs. (3) - (5) to group the real part of the spectrum of the speech signal, then pass it through a $|\cdot|$ operator and a DCT transformation matrix to get features similar to MFCC speech feature. The experimental results on voice mail transcription are given in section 5.

3. WELCH PROCESSING

In our experiment, we found that the log operation that follows the binning of the power spectrum outputs causes the variance of the features to increase greatly. In order to get a smoother feature while keeping the spectral resolution, we applied a lowpass filter to smooth the feature before the log operation. As in [6], we used a simple average for the lowpass filtering operation. The binned outputs were computed every 2 ms, and then averaged over 10ms to yield an estimate of the binned outputs every 10ms. Experimental results using this welch processing on voice mail transcription are presented in section 5.

4. PSEUDO-GRADIENT-BASED NEW SPEECH FEATURES

4.1 Pseudogradients of the resonator filter-bank

It is well known that spectral transitions and dynamics play an important role in the perception of speech [9]. In this paper, we propose a set of new features based on an adaptive filter to capture the dynamics of the local concentration of energy within different formant bands of the speech signal. The adaptive filter comprises of a number of digital resonators in a feedback loop [5] – the result being a multiple notch transfer function with notches at the resonator frequencies. The adaptive algorithm adapts these resonator frequencies to minimize the power at the notch output – hence the resonator frequencies track the frequencies of the sinusoidal components in the input. The i'th resonator frequency is controlled by a single coefficient, k_i, and a gradient or "pseudo-gradient" method can be used to adapt the parameter k_i:

$$k_i(n+1) = k_i(n) - \mu \times \nabla_i(\omega) \qquad (6)$$

Where $\nabla_i(\omega)$ denotes the gradient of the error power $E\{x_e^2\}$ with respect to the coefficient k_i, μ is the step size. As the error surface is extremely multimodal, gradient descent techniques could lead to the filter converging to local minima, consequently we use a pseudogradient (rather than the gradient) to adapt the parameters k_i. It was proved in [6] that this pseudogradient can guarantee global convergence under certain conditions, and has lower computational complexity ($O(N)$) as compared to the true gradient. The pseudogradient is computed by correlating the error signal x_e with the output of a pseudo-sensitivity filter:

$$\nabla_{ps,i} = \frac{1}{2\pi j}\oint H_e(z^{-1}) H_{ps,i}(z) X(z) X(z^{-1}) z^{-1} dz \quad (7)$$

where $X(z)$ denotes the *z*-transform of the input signal. For a single sinusoidal input signal with frequency ω, the pseudogradient $\nabla_{ps,i}(\omega)$ is given by:

$$\nabla_{ps,i}(\omega) = -2G \frac{da_i}{dk_i} |A(\omega)|^2 [2\cos(\omega) - 2a_i] \quad (8)$$

In order to reduce the variance of the instantaneous pseudo-gradient estimates, and to normalize out the amplitude of the input signal, we used a normalized version of the pseudogradient, given by:

$$\hat{\nabla}_i^{ps}(\omega) = \frac{E\{H_e^*(\omega)\cdot H_{ps,i}(\omega)\}}{E\{H_{ps,i}^*(\omega)\cdot H_{ps,i}(\omega)\}} = \frac{\sum_{t=1}^{T} x_e(t)\cdot x_{ps,i}(t)}{\sum_{t=1}^{T} |x_{ps,i}(t)|^2} \quad (9)$$

where $x_e(t)$ and $x_{ps,i}(t)$ denotes the error signal and the *i*-th pseudo-sensitivity filter output of the *t*-th input data sample, respectively. *T* is the number of samples in a speech frame.

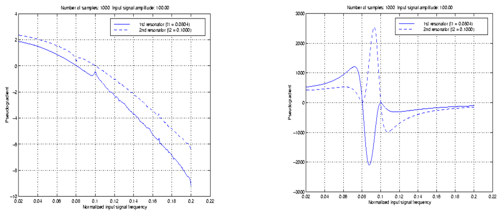

Fig. 6. Pseudogradient (a) and normalized pseudogradients (b)

Fig. 6(a) and 6(b) show the pseudogradients and normalized pseudogradients for a two-resonator system with two closely spaced normalized resonator frequencies $f_1 = 0.0804$ and $f_2 = 0.10$, respectively. We can see from these two figures that for both the pseudogradients and the normalized pseudogradients , the sign of the signal correlates with whether the input frequency is greater or less than the resonator frequency, in the neighbourhood of the resonator frequency, there is in fact a linear relationship between the pseudogradient and the input sinusoidal frequency, consequently, the pseudogradient can be used to get an estimate of the input sinusoidal frequency. Further, the pseudogradient has a much stronger dependence on the amplitude of the input signal than the normalized pseudogradient. We constructed a set of new features based on the pseudogradients of a 4-resonator system. The frequencies of the resonators are set at the center frequencies of four formant bands, i.e. F1 band: 250 Hz – 850 Hz; F2 band: 700 Hz – 2300 Hz; F3 band: 2200 Hz – 2900 Hz; F4 band: 3100 Hz – 3900 Hz. In order to further investigate the discriminability of this set of new features, we use linear discriminant analysis (LDA) to quantify the ability

$$T = \sum_t (\vec{y}_t - \vec{\mu}_t)(\vec{y}_t - \vec{\mu}_t)^T \qquad (10)$$

$$B_i = \sum_{s \in C_i} (\vec{y}_s - \vec{\mu}_s(\vec{y}_s - \vec{\mu}_s)^T, B = \sum_i \tfrac{n_i}{N} B_i \quad (11)$$

$$W = T\text{-}B \qquad (12)$$

of these new features to distinguish between different phonemes.

4.2 Linear discriminant analysis (LDA)

LDA is a measure of the capability of a linear projection in separating out the different classes. For the whole speech feature space, define the covariance matrices by: where $\vec{y}_t$ and $\vec{\mu}_t$ denotes the *t*-th feature vector and it's mean, C_i denotes the *i*-th class, n_i and N denotes the number of features in the *i*-th class and the whole feature set, respectively. Then *B* and *W* in Eqs. (11) and (12) denote the between-class covariance matrix and the within-class covariance matrix, respectively. If the dimension of the feature space is *K*, we can define the parameters d_i by:

$$d_i = \max_{\vec{v}_i} \left\{ \frac{\vec{v}_i^T B \vec{v}_i}{\vec{v}_i^T W \vec{v}_i} \right\}, \qquad i = 1,2,\cdots K \quad (13)$$

Then d_i for i=1denotes the maximum of ratio of the inter-class distance and the average within class variance. Hence, d1 is a measure of the discriminative ability of the linear projection $\vec{v}_i$ which is called the leading linear discriminant of the data set. We'll compare the leading eigenvalue d1 of the new feature set with the MFCC feature set in the next section.

5. EXPERIMENTAL RESULTS

5.1 Experiment results on voice mail transcription

In this part, we present results on a large vocabulary, speaker-independent, conversational telephone speech recognition (voicemail transcription) task. We primarily report results of speech recognition experiments that used the Welch smoothing technique on baseline MFCC features, and some preliminary analysis of the discriminant ability of the new pseudo-gradient features. The vocabulary size of the voice mail database is 20K words. The recognition system has 2313 context-dependent states and about 70K Gaussians. The language model perplexity is about 100. The baseline extracted features consist of 13 MFCC and their first and second temporal derivatives. We experimented with two different training and test sets - the first training set "vmtrgab" has 20 hours of telephone speech and the second set "vmtrgabcd" contains 70 hours of telephone speech. The first test data set "vmt.test2b" has 43 phone-mail messages with 1986 words and the second test data set "vmt.test2c.sort" has 86 phone-mail messages with 6925 words. Experimental results of the phone recognition accuracy and word error rate (WER) are listed in Table I and II, respectively.

Table I. Phone recognition accuracy of the baseline and welch processing

	Vmtrgab	Vmtrgabcd
Vmt.test2b (baseline)	28.43%	28.88%
Vmt.test2b (welch)	28.64%	29.10%
Vmt.test2c (baseline)	30.35%	31.08%
Vmt.test2c (welch)	30.55%	31.42%

Table II. WER of the baseline and the welch processing

	Vmtrgab	Vmtrgabcd
Vmt.test2b+Vmt.test 2c (baseline)	43.72%	43.74%
Vmt.test2b+Vmt.test 2c (welch processing)	43.68%	43.52%

We can see from Table I and II that welch processing consistently improves the phone accuracy over the baseline for different combinations of training and testing conditions. The WER of the lerner grouped features is 47.89% with the training set "vmtrgab" and test set "vmt.test2b". The higher error rate appears to be because the lerner features have a much larger variance than the Mel features (possibly because the binning is done in the amplitude domain, rather than in the power domain, leading to much smaller-valued features as the input to the log function).

5.2 LDA of the pseudo-gradient-based feature set

In this part, we carried out some experiments on the LDA analysis of the pseudo-gradient-based feature set and it's combination with the conventional MFCC feature set. First, we try to augment the conventional MFCC features (MFCC1-12+C0) with the two pseudo-gradient-based features and to compare the leading eigenvalue of LDA analysis for the new feature (MFCC1-12 + pseudo-gradients 1-2+ C0) and the MFCC feature (MFCC1-12 + C0). The leading eigenvalue of the first 3 MFCC features and the first 3 augmented features are listed in Table III.

Table III. LDA measurement d_i of the MFCC feature and new augmented feature.

MFCC+C0(1-5)	2.9414	2.3814	1.9402
Augmented feature (1-5)	2.9466	2.4313	1.9922

Furthermore, we replace the last two components (MFCC 11 - 12) of the MFCC features by the first two pseudo-gradient-based features. We call the new features mixed features. The leading eigenvalue of the mixed features and the MFCC features are listed in Table IV.

Table IV. LDA measurement d_i of the MFCC feature and new mixed feature.

MFCC+C0(1-5)	2.9414	2.3814	1.9402
Mixed feature (1-5)	2.9453	2.4310	1.9904

We can see from Table III and Table IV that both the augmented features and the mixed new features help to increase the discriminative ability of the features over conventional MFCC features.

6. DISCUSSION

The welch processing helps to decrease the variance of the speech features and consistently improve the phone recognition accuracy and the WER by some extend. The pseudo-gradient-based features represent the trajectory of the movement of the frequency components within different formant bands. LDA analysis shows that this new feature improves the discriminative ability to distinguish between different phone classes. Future work includes incorporating this new feature into a speech recognition system to improve the recognition performance.

REFERENCES

1. Van Alphen P. and Poles L., "Comparing various feature vectors in automatic speech recognition", in proceedings of *EUROSPEECH'91*, pp. 533 – 536.
2. Tierney J., "A study of LPC analysis of speech in additive noise", *IEEE Trans. ASSP*, Vol. 23, No. 7. , 1990.
3. Hermansky H., "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, Vol. 87, No. 4, pp. 1738-1752, 1990.
4. Wegman S., McAllaster, D., Orloff J. and Peskin B., "Speaker normalization on conventional telephone speech", in proceedings of *IEEE ICASSP'96*, pp. 339-341.
5. Padmanabhan M. and Martin K., "Resonator-based filter-banks for frequency-domain applications", *IEEE Trans. Circuits and Systems*, Vol. 38, No. 10, pp. 1145-1159, 1991.
6. S. Dharanipragda and R. Gopinath, "Techniques for capturing temporal variations in speech signals with fixed rate processing", Proceedings of the ICSLP, 1998.
7. Lerner R.M., "Band-pass filters with linear phase", *Proceedings of the IEEE*, Vol. 52, pp. 249-268, 1975.
8. Martin K. and Padmanabhan M., "Resonator-in-a-loop filter-banks based on a lerner grouping of outputs"" in proceedings of *IEEE ICASSP''2*, pp. 329-332.
9. Junqua J. and Haton J. *Robustness in automatic speech recognition*, Boston, MA: Kluwer academic publishers, 1996.