Statistical Integration of Temporal Filter Banks for Robust Speech Recognition Using Linear Discriminant Analysis (LDA)

Jia-lin Shen, Wen-liang Hwang
Institute of Information Science, Academia Sinica
Taipei, Taiwan, Republic of China
{jlshen, whwang}@iis.sinica.edu.tw

ABSTRACT

This paper presents a study on statistical integration of temporal filter banks for robust speech recognition using linear discriminant analysis (LDA). The temporal properties of stationary features were first captured and represented using a bank of well-defined temporal filters. Then these derived temporal features can be integrated and compressed using the LDA technique. Experimental results show that the recognition performance can be significantly improved both in clean and in noisy environments.

1. INTRODUCTION

Various temporal features were proposed to capture time dependency of stationary features for the improvements of accuracy and robustness for speech recognition [1-6]. These temporal features were derived form the temporal filters designed by partitioning and filtering the frequency response of temporal trajectories of frequency bands (called modulation frequency). Among these temporal filters, several were developed by emphasizing the relatively important modulation frequencies of speech signals which are obtained from human hearing perception [1] or speech data analysis [2-4], while the others were designed by improving the discriminative ability using the statistical approaches [4-6]. Because the purpose of the extraction of temporal features is for speech recognition, it can be found that the recognition performance can be further improved by integrating the statistical methods such as linear discriminative methods (LDA) [5] and minimum classification error (MCE) [6] techniques. In comparison with the widely used delta features obtained by using a FIR filter (called delta filter) to approximate the first order temporal derivative of the stationary features over a short duration of window, we found that although those sophisticated temporal features outperform delta features in both clean and noisy environments, the improvements will be greatly reduced when the window length used in delta feature extraction is enlarged for capturing longer temporal information. This is because the passband of the delta filter is shifted to lower modulation frequency bands when longer window length is applied, which meets the superiority of the RASTA filter by adding an extra pole to the delta filter [1]. On the other hand, the incorporation capabilities with the stationary features for those sophisticated temporal features are much poorer than do the delta features.

In this paper, we intend to develop an approach to integrate the temporal information in the speech features with improved discriminant ability and incorporation capability. In other words, the relatively important modulation frequencies must be emphasized and the separability for speech recognition must be improved. First, a set of temporal filters will be well-defined to span the modulation frequency domain and concentrate on the relatively important frequencies. Then, to integrate these temporal features into stationary features, the linear discriminant analysis (LDA) technique is used for reducing the dimensionality of the resultant feature vector as well as improving the discriminative ability in a maximum class separability manner. It is very different from the previous work [4] that applies the LDA technique to a rather long window of stationary features and the

derived eigenvectors with high corresponding eigenvalues were used as the temporal filters. Instead, we try to combine different temporal features obtained by different temporal filters and maximize the separability information for speech recognition using the LDA technique.

2. STATISTICAL INTEGRATION OF TEMPORAL FILTER BANKS

In order to capture the temporal properties of speech signals, various temporal filters were proposed to filter and partition the modulation frequency bands [1-6]. However, what frequencies are important for the purpose of speech recognition ? Two approaches were investigated in the previous studies. First, some analyses were performed to understand the relatively important modulation frequencies for speech signals [1-4]. Although the recognition performance is indeed improved by concentrating those relatively important modulation frequencies, it is probable to achieve further improvements by properly including the disregarded ones. Secondly, the temporal filters were developed using the statistical approach in a maximum class separability manner [4-6]. Here the temporal filters can be directly designed using the data-driven approach based on the linear discriminant analysis (LDA) technique [5,7] or refined using the minimum classification error (MCE) algorithm [6,8]. Because the purpose of the temporal filters is to improve the robustness and accuracy for speech recognition, it should be a right way to integrate the statistical methods in developing the temporal filters.

In this paper, a two-stage approach by combining the knowledge and the statistical methods was developed to extract the temporal features. In the first stage, we select a bank of temporal filters to span and partition all the modulation frequency bands. Because there exists redundant and insignificant information for speech recognition in these temporal filters, the LDA technique is further used to reduce the dimensionality as well as to improve the discriminating ability in the second stage. In other words, all the temporal information is first included by a bank of temporal filters and then compressed using LDA technique to improve the separability for speech recognition. An example set is the delta filters with different duration of windows. One can note that longer window duration means the passband locating on lower modulation frequencies, for instance, when 10 ms of frame shift is used, the peak modulation frequencies appear on 25.0Hz, 13.8Hz, 9.7Hz, 7.5Hz, 6.1Hz, 5.1Hz, 4.4Hz and 3.9Hz with respect to 20ms, 40ms, 60ms, 80ms, 100ms, 120ms, 140ms and 160ms of window length, respectively, as shown in Fig. 1.

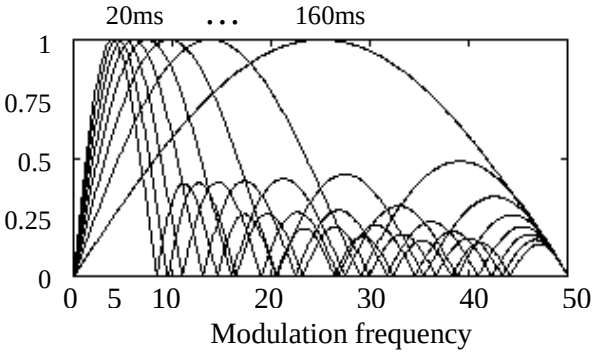

Figure 1 : The frequency response for delta filters with window length ranging from 20ms to 160ms.

The block diagram of the derived features is shown in Fig. 2. Before applying the LDA approach, the feature vector at time *t* can be expressed as :

$$[x_t, x^1{}_t, x^2{}_t, ..., x^N{}_t] \qquad (1)$$

where x^i_t means the stationary feature vector x_t filtered by the *i*-th temporal filter. In fact, in many speech recognition systems, the first and second order derivatives of stationary features are used, which means a coarse partition of the modulation frequency plane. As an alternative, a fine representation of the modulation frequency plane is first derived from a bank of well-defined temporal filters. Then the LDA technique is used to compress the resultant feature vectors in equation 1. In some studies [5,9], the LDA was directly applied to the stationary feature and the neighbor frames such that equation 1 can be changed in the following :

$$[x_t, x_{t+1}, ..., x_{t+d}], \qquad (2)$$

where *d* denotes the window length. Note that no temporal filters are used. In fact, the eigenvectors of the LDA matrix can be regarded as the temporal filters instead. In [5], relatively long window length (1 sec) was used.

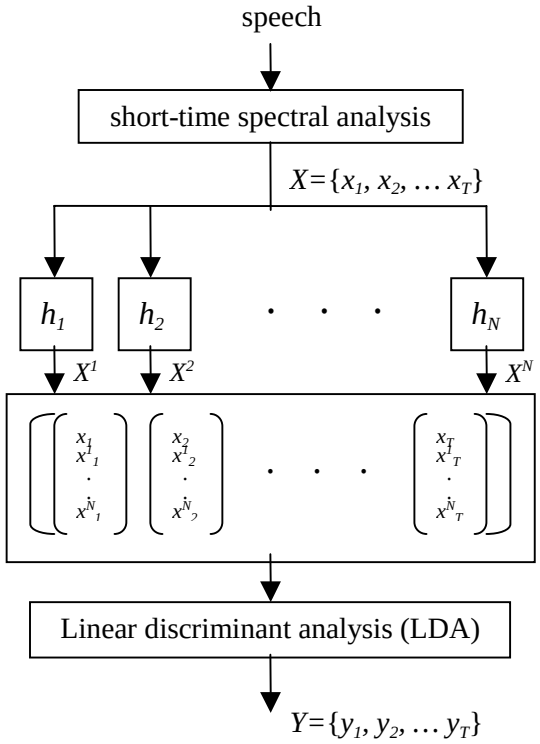

Figure 2 : The block diagram of the derived features based on the statistical integration of the temporal filter banks h_1, h_2,…, h_N using LDA technique.

3. EXPERIMENTAL RESULTS

3.1 Speech Database

In the following experiments, the speech database was produced by two male speakers. For each speaker, 4 utterances for each of all the 1345 Mandarin syllables using two types of microphones were produced at a sampling frequency of 16kHz, where 3 sets are used for training and 1 set is used for testing. A 14-order mel-frequency cepstral vector derived from the power spectrum filtered by a set of 30 triangular band-pass filters was used for each frame. Each syllable was modeled using the left-to-right HMM.

3.2 Microphone Variations

In the first experiment, we combined 14-order cepstral features and the 8 sets of temporal features derived from the delta filters with different window sizes mentioned above (a total of 14*9=126 dimensions). As shown in Table 1, the recognition rates were increased from 87.51% and 69.89% to 90.56% and 84.24% when the 8 sets of temporal features were incorporated into cepstral features without and with microphone variations, respectively. Then when the LDA technique was applied, the recognition rates can be further improved to 93.53% and 91.08%, respectively, without and with microphone variations using 42 dimensions only. In comparison with the data-driven RASTA-like filters in [5] using the LDA in a window of stationary features for obtaining the temporal filters, the recognition rates were 93.31% and 89.14% when the first three eigenvectors are used (a total of 42 dimensions). It is obvious that the error rates were reduced by 3.29% and 17.86%, respectively, without and with microphone variations. Moreover, the experiments using the widely used features by incorporating the first and second order delta features into stationary cepstral features were performed as also shown in Table 1.

As listed in Table 2, a bank of RASTA filters was used as a comparison. Here 3 different parameters of the one-pole filter in RASTA filter were chosen, including 0.94, 0.8 and 0.6, in which the peak modulation frequencies were located on 3.9Hz, 7.0Hz and 9.7Hz, respectively. The results shown in Table 2 were worse than those in Table 1 by using delta filters. This is because the 8 delta filters lead to better partition of the modulation frequency plane.

3.3 Additive Noise

As shown in Table 3, the recognition rates in the presence of white noise were improved from 48.33% and 14.42% to 51.30% and 27.92% when the 8 sets of delta features mentioned above were incorporated at SNR of 30dB and 20 dB, respectively. Next, when the robust features proposed in this paper were used, the recognition

rates were further improved to 67.49% and 36.42%, respectively, at SNR of 30dB and 20 dB when 42 dimensions were chosen with LDA. In comparison with the previous method in [5], the recognition rates were 65.65% and 28.70% (42 dimensions). Thus the error-rate reductions of 5.36% and 10.83%at SNR of 30dB and 20 dB can be obtained.

It is believed that the recognition performance can be further improved when more effective and promising temporal filters are included in the first stage. In addition, the MCE algorithm can be applied to the discriminant matrix obtained in LDA for further improvements.

4. CONCLUSION

This paper presents an approach to derive robust features for speech recognition by statistically integrating the temporal filters. A bank of temporal filters was first used to span and partition all the modulation frequency bands. Then the derived temporal features and the stationary feature are combined and compressed by the LDA algorithm. Improved accuracy and robustness can be therefore obtained.

Feature(dimension)	clean	microphone variations
cepstral features(14)	87.51	69.89
cepstral + 8 delta features(126)	90.56	84.24
plus LDA(42)	93.33	91.08
cepstral+delta+2nd delta features + LDA(42)	93.09	87.41
data-driven RASTA-like(42)	93.31	89.14

Table 1: Recognition rates for various features in the presence of microphone variations.

Feature(dimension)	clean	microphone variations
cepstral + 3 RASTA features(56)	88.22	78.29
plus LDA(42)	93.09	89.52

Table 2: Recognition rates by incorporating a bank of RASTA features in the presence of microphone variations.

REFERENCES

1. H. Hermansky, N. Morgan, "RASTA Processing of Speech", *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 4, Oct. 1994, pp. 578-589.
2. J.L. Shen, W.L. Hwang, L.S. Lee, "Robust Speech Recognition Features Based on Temporal Trajectory Filtering of Frequency Band Spectrum", *ICSLP*, pp. 881-884, 1996.
3. C. Nadeu, P.P. Leal, B.H. Juang, "Filtering the Time Sequences of Spectral Parameters for Speech Recognition", *Speech Communication*, 22, 1997, pp. 315-332.
4. J.L. Shen, W.L. Hwang, "New Temporal Features for Robust Speech Recognition with Emphasis on Microphone Variations", *Computer Speech and Language (accepted).*
5. Van Vuuren S., H. Hermansky, "Data-driven Design of RASTA-like Filters", *Eurospeech*, pp. 409-412, 1997.
6. J.L. Shen, "Discriminative Temporal Feature Extraction for Robust Speech Recognition", *Electronics letters*, Vol. 33, No. 19, pp. 1598-1600, 1997.
7. B.H. Juang, W. Chou, C.H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition", *IEEE Trans. On Speech and Audio Processing*, Vol. 5, No. 3, pp. 257-265, May 1997.
8. L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, M.A. Picheny, "Robust Methods for Using Context-dependent Features and Models in a Continuous Speech Recognizer", *ICASSP*, pp. 533-536, 1994.

Feature(dimension)	clean	30dB	20dB
cepstral features(14)	87.51	48.33	14.42
plus 8 delta features (126)	90.56	51.30	27.92
plus LDA (42)	93.33	67.49	36.42

Table 3: Recognition rates for various features in the presence of additive noise.