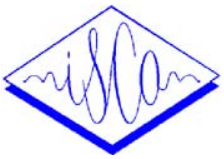
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

ROBUST SPEECH RECOGNITION USING DATA-DRIVEN TEMPORAL FILTERS BASED ON INDEPENDENT COMPONENT ANALYSIS

Junhui Zhao, Jingming Kuang, Xiang Xie

Research Center of Digits Communication Technologies,
Beijing Institute of Technology, Beijing, 100081, China
{zkes1000, jmkuang, xiexiang}@bit.edu.cn

10.21437/Interspeech.2004-625

Abstract

In this paper, a data-driven temporal processing method based on Independent Component Analysis (ICA) is proposed for obtaining a more robust speech representation. Two different schemes of dominant temporal filters based on ICA are investigated. The one is the perceptually-based filter which always focuses on the modulation frequency range between 1 and 16 Hz and the other is the most independent component discovered by ICA algorithm. Detailed comparative analysis between the proposed ICA-derived temporal filters and the previous statistical methods including Linear Discriminant Analysis (LDA) and Principle Component Analysis (PCA) is presented. The preliminary experiments show that the performance of the ICA based temporal filtering is much better in comparison with the LDA and PCA based methods in noisy environment.

1. Introduction

Temporal filtering is becoming a common technique for improving the robustness of Automatic Speech Recognition (ASR) system in mismatched training and test environment. Cepstral mean subtraction (CMS), generalized dynamic features [1] and RASTA filtering [2] are typical examples of temporal processing that have been successfully applied. In particular, data-driven temporal filtering is considered as more optimal for a special recognition task or application environment, which has brought out many research and engineering efforts. Van Vuuren and Hermansky [3] introduced a data-driven technique for direct derivation of temporal basis functions through Linear Discriminant Analysis (LDA). In a recent work, Jeih-weih Hung [4] used Principal Component Analysis (PCA) to obtain an optimal set of data-driven time trajectory filters. Such data-driven temporal filtering techniques were reported to yield better recognition performance than the conventional CMS method [5].

Independent component analysis (ICA) is signal processing technique that can transform linearly mixed signals into components that statistically as independent from each other as possible [6]. ICA has many promising applications such as Blind Signal Separation (BSS), feature extraction and exploratory data analysis.

In this paper, we proposed a new data-driven temporal filter based on ICA for improving the robustness of speech recognition system in noisy environment. Perceptual experiments [7] indicate that some components of modulation spectrum are more important for the intelligibility of speech than others. We applied ICA technique to extract these relatively important frequency components from the modulation spectrum of speech signals, which is based on the assumption that the sub-signals occupying different frequency bands are linearly mixed. Two sets of dominant temporal filters are extracted from ICA algorithm according to different criterion in our study. Some preliminary experiments are performed and the recognition results reveal that ICA based temporal filtering can always achieve better performance in noisy environments of various SNR conditions than the previous methods such as PCA and LDA.

2. Speech Modulation Spectrum

Speech reflects the geometry of the vocal organs, and the linguistically dominant component is in the shape of the vocal tract [8]. At any given point in time, the shape of the vocal tract is reflected in the short-time spectral envelope of the speech signal. The change rate of the vocal tract shape appears to be important for the identification of linguistic components. This change rate can be described by the modulation spectrum, i.e., the spectrum of the time trajectories described by the short-time spectral envelope.

Generally, the time trajectory sequence S can be viewed as the sum of the sub-signals occupying different modulation frequency band, such as

$$S = \sum_i s_{[f_{L_i}, f_{U_i}]} \tag{1}$$

where f_{L_i} and f_{U_i} are the lower cutoff frequency and upper cutoff frequency of the ith sub-signal s, respectively. The research of Hermansky [2] indicates that the modulation spectrum of speech mainly covers the frequency

range of 1-16 Hz and exhibits a maximum at about 4 Hz, the average syllabic rate. Disturbances often have modulation frequency components outside the speech range, and could in principle be attenuated without significantly affecting the range with the relevant linguistic information. The motivation of temporal filtering is to reduce the disturbed frequency components in modulation spectrum and reserve enough linguistic information for subsequent recognition processing by filtering along the time trajectory of the change of vocal tract.

3. ICA based Temporal Filters

Given an ordered sequence of K-dimensional feature vectors $x(n)$ as shown in Fig.1, with time index $n = 1, 2, .., N$, and feature index $k = 1, 2.., K$,

$$x(n) = [x(n,1)x(n,2), .., x(n,k), .., x(n,K)]^T \quad (2)$$

then the kth time trajectory of $x(n)$ is the sequence

$$y_k(n) = x(n,k) \qquad \text{n=1,2,..,N} \quad (3)$$

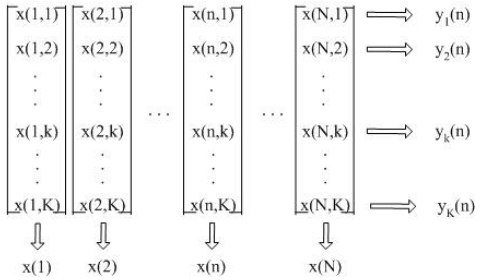


Figure 1: The representation of the time trajectories of feature sequences

In this study, the implementation of data-driven temporal filtering is to design an L-point FIR filter applied on the kth time trajectory $y_k(n)$, as depicted in Fig.2.

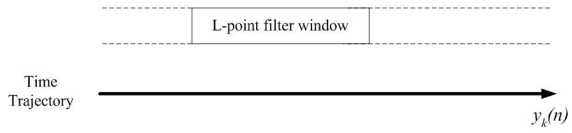


Figure 2: Data-driven temporal filtering operation

With the filter sliding along the kth time trajectory, a set of L-dimensional vectors Z can be obtained, as shown in Fig. 3. Below, the desired L-point FIR filter is derived by using ICA.

3.1. Independent Component Analysis

In ICA algorithm, Z is assumed to consist of the linear mixture of the independent source signals X, so Z can be

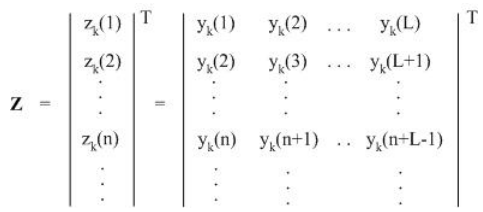


Figure 3: The representation of the L-dimensional vectors from the time trajectory $y_k(n)$

denoted as the multiplication of X and the corresponding mixing matrix A.

$$Z = AX \quad (4)$$

The aim of ICA is to find the estimation of source signals and the corresponding linear combination matrix A. If W is the inverse matrix of A, P is the number of independent components and $S = [s_1, s_2, .., s_P]$ is an estimation of X, Equ.4 can be expressed as

$$S = WZ \quad (5)$$

where $W = [w_1, w_2, .., w_P]^T$.
Thus, the ICA problem is reduced to finding a linear combination W which transforms the observed signals into independent signals. The learning rule is based on maximization of the approximation of negentropy $J(S)$ and is represented as [9]

$$max\ J(s_i) = J(w_i^T Z) = [E\{G(w_i^T Z)\} - E\{G(\upsilon)\}]^2 \quad (i = 1, 2, \cdots, P) \quad (6)$$

where G is contrast function and υ is a standardized Gaussian variable. In this work, we take the exponential function as contrast function:

$$G(u) = -exp(-\frac{1}{2}u^2) \quad (7)$$

Using a fast fix-point algorithm [10], $w_i (i = 1, 2, \cdots, P)$ can be iteratively updated until convergence.

$$w_i^+ = E\{Zg(w_i^T Z)\} - E\{g^{'}(w_i^T Z)\}w_i \quad w_i = w_i^+ / ||w_i^+|| \quad (8)$$

where g is the derivative of the contrast function G. In this algorithm, a deflation scheme based on a Gram-Schmidt-like decorrelation is applied to prevent the different weighted vectors from converging to the same maxima.

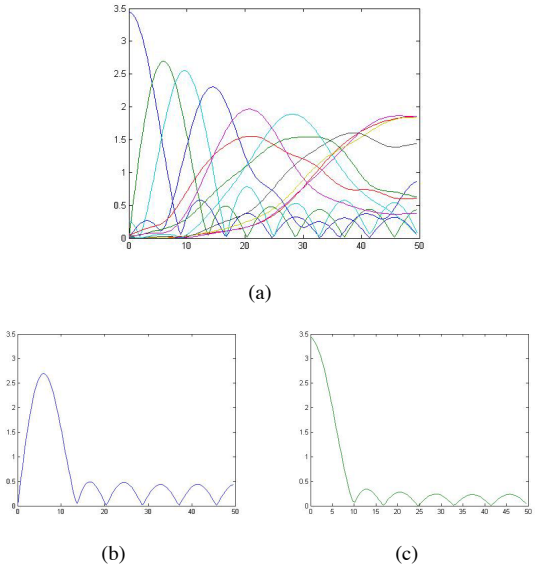
3.2. Dominant ICA-derived Temporal Filters

From the above ICA processing, we can acquire a set of orthogonal FIR filters $w_i (i = 1, 2, \cdots, P)$ for each time trajectory. For instance, Fig.4 (a) shows the frequency responses of the filters obtained by ICA for the 10th time trajectory of MFCC feature in the following experiment.

These filters covering different frequency band can be used to extract the latent sub-signals from the original signal. For subsequent temporal processing, we may select a dominant FIR filter from these orthogonal filters. Here, we have basically the following two choices for dominant temporal filter.

- According to section 2, the modulation frequency components between 1 and 16 Hz are more important for speech intelligibility than others. Thus the filter whose main-lobe covers the intelligibility frequency range is considered as the perceptually based dominant filter. Fig.4 (b) gives the perception-based dominant temporal filter.

- w_1 is the first independent component extracted from ICA algorithm using the deflation scheme based on Gram-Schmidt-like decorrelation. By means of this method, the independent components can be estimated one by one. So w_1 is definitely the most non-gaussian component which may indeed be the most important in many applications [10]. We also used w_1 as the filter coefficients of the dominant temporal filter. Fig.4 (c) gives the independence-based dominant temporal filter.



(a)

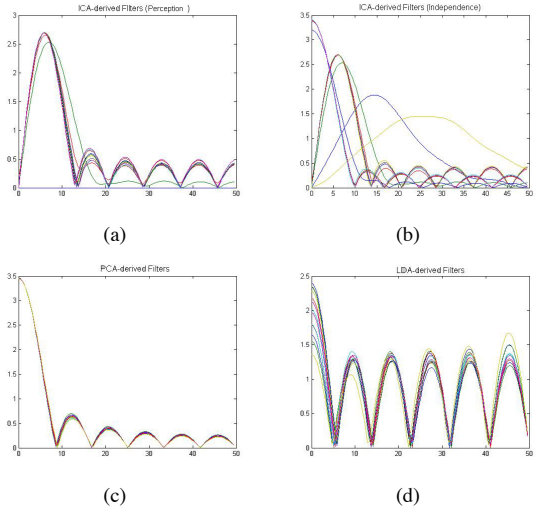
(b) (c)

Figure 4: (a) The frequency responses of the orthogonal FIR filters from ICA for the 10th time trajectory of MFCC; (b) the perception-based dominant filter; (c) the independence-based dominant filter

4. Recognition Experiments

The speech database chosen for this experiment is Mandarin digits database recorded by 125 speakers (63 male and 62 female) at 8k Hz, 16 bits. The vocabulary includes 11 Mandarin digits ranging from zero to nine adding "yao1" (equivalent to "one"). The training data consists of 3388 utterances and recognition performs on other 1540 utterances. In this experiment, mel-frequency cepstral coefficients (MFCC) are calculated from Hamming-windowed, pre-emphasized 20 ms segments of the input signal with a frame period of 10 ms. Here, each MFCC feature vector contains 13 features (12 coefficients, and log energy). For obtaining these data-driven temporal filters, we perform the ICA, PCA, and LDA process on the original feature vectors of the training speech set of 3388 utterances. The length of filters is preliminarily set as 12.

The frequency responses of 13 ICA-derived temporal



(a) (b)

(c) (d)

Figure 5: The frequency responses of 13 temporal filters according to (a) ICA (Perception), (b) ICA (Independence), (c) PCA and (d) LDA

filters based on perception and independence are shown in Fig.5 (a) and (b). For comparison, PCA-derived and LDA-derived filters are also given in Fig.5 (c) and (d) respectively. According to the above figures, we can see some important phenomena.

- The perception-based temporal filters always focus on the frequency range between 1 and 16 Hz that gives most contributions to speech perception. The center frequency of these filters is near 4 Hz.

- Among the independence-based dominant temporal filters, most filters concentrate on perceptual frequency range, others are lowpass or bandpass. Especially, the temporal filter used for the time trajectory of log energy dimension is bandpass and center frequency lies in near 25Hz.

- Most of ICA filters, both perception-based and independence-based, suppress the 0Hz frequency

component. In contrast, it is the center of both PCA- and LDA-derived filters.

The different kinds of temporal filters including ICA, PCA and LDA are respectively applied to the time trajectories of the feature vectors of both training and testing database to obtain the new feature vectors. These new feature vectors are finally used for model training and testing following ASR procedures. The recognition engine throughout our experiments is based on the HTK software package. The digits are modeled as whole word HMMs with 8 states and 3 mixtures per state (according to 10 states in HTK notation with 2 dummy states at beginning and end). The ICA-, PCA- and LDA-derived features from the original feature are used to train for their respective HMM's for comparison. In training phase, the references are obtained on clean data only. White noise is manually added into the testing utterances of 2170 at the different SNRs of 20dB, 15dB, 10dB and 5dB. Furthermore, the clean test utterances without adding noise is also taken into account as the fifth case.

Tab. 1. Recognition rates (%) for different temporal processing techniques

SNR (dB)	Clean	20	15	10	5	Ave
MFCC	98.64	41.62	31.3	20.65	15.26	41.49
PCA	97.32	43.57	37.40	27.21	18.68	44.84
LDA	97.39	43.70	38.18	26.75	13.05	43.81
ICA(P)	97.41	78.57	66.43	47.40	27.43	63.45
ICA(I)	98.92	79.57	68.43	44.40	26.17	63.50

Table 1 gives the recognition results of different temporal filtering techniques on MFCC in mismatched condition. PCA, LDA and ICA represent the three different filtering approaches on the corresponding feature parameter, respectively. P and I denote the different scheme for dominant filter selection, perception-based and independence-based. The first point to note is that the independence-based ICA temporal filtering can improve the recognition performance in clean environment, which can achieve the rate of 98.92%. However, when other data-driven temporal processing methods are applied, the recognition rates drop a little compared with the plain MFCC parameters. Secondly, for the testing speech corrupted by White noise, we can see that PCA and LDA can improve the recognition performance slightly with SNR dropped. On the other hand, by using ICA approach, significant improvement can be obtained. In ICA(P) case, the average rate is 63.45% yielding relative improvement of 52.92% compared with the baseline. ICA(I) performs best which achieves 63.50% in average rate. Finally, the recognition performance between ICA(P) and ICA (I) is very close in most SNR conditions. In high SNR such as 20 and 15 dB, ICA(I) shows better recognition rates than ICA(P), but at low SNR ICA(P) performs better.

5. Conclusions

In this paper, we present ICA based data-driven temporal filtering technique for improving the robustness of ASR system. Two different scheme for dominant temporal filter selection are investigated, one is perception-based and the other is independence-based. Detailed comparative study between ICA, PCA and LDA is also given. Significant improvements in recognition accuracy under noisy condition show the effectiveness of the proposed temporal filtering approach.

6. References

[1] R. Chengalvarayan and L. Deng, "Use of Generalized Dynamic Feature Parameters for Speech Recognition",IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 3, pp. 232-242, May 1997.

[2] H. Hermansky and N. Morgan, "RASTA Processing of Speech", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 587-589, October 1994.

[3] S. van Vuuren and H. Hermansky. "Data-driven Design of RASTA-like Filters," Eurospeech1997, vol. 1, pp. 1607-10, Greece, September 1997.

[4] J. Hung, H. Wang, and L. Lee, "Comparative Analysis for Data-Driven Temporal Filters Obtained via Principal Component Analysis," Eurospeech2001, Aalborg Demark, Sept. 2001.

[5] J. Hung, and L. Lee, "Data-Driven Temporal Filters Obtained via Different Optimization Criteria Evaluated on AURORA2 Database", ICSLP02, Denver, USA, Sep, 2002.

[6] T. W. Lee, "Independent Component Analysis-Theory and Applications," Kluwer Academic Publishers, 1998.

[7] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Intelligibility of Speech with Filtered Time Trajectories of Spectral Envelopes," ICSLP96, Philadelphia, pp. 2490 - 2493, 1996.

[8] C. Avendano and H. Hermansky, "On the Properties of Temporal Processing for Speech in Adverse Environments," Proc. IEEE Workshop on ASSP, New Paltz, N.Y., October, 1997

[9] P.Comon, "Independent Component Analysis," Signal Processing, vol. 36, pp. 287-314, 1994.

[10] A. Hyvärinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis", IEEE Transactions on Neural Networks, Vol. 10, No. 3, pp. 626–634, 1999.