ICA-Based Feature Extraction for Phoneme Recognition

Oh-Wook Kwon, Te-Won Lee

School of Electrical and Computer Engineering, Chungbuk National University, 48 Gaesin-dong, Heungdeok-gu, Cheongju, Chungbuk, 361-763, Korea
owkwon@chungbuk.ac.kr
† Institute for Neural Computation, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0523, USA
tewon@ucsd.edu

Abstract

We propose a new scheme to reduce phase sensitivity in independent component analysis (ICA)-based feature extraction using an analytical description of the ICA-adapted basis functions. Furthermore, since the basis functions are not shift invariant, we extend the method to include a spectral-domain ICA stage that removes redundant time shift information. The performance of the new scheme is evaluated for TIMIT phoneme recognition and compared with the standard mel frequency cepstral coefficient (MFCC) feature.

1. Introduction

It has been a key focus in pattern recognition tasks to find an efficient data representation. Popular methods for capturing the structure of data have been principal component analysis (PCA), which yields a compact representation, and more recently independent component analysis (ICA). In ICA, the data are linearly transformed such that the resulting coefficients are statistically as independent as possible. In other words, it is assumed that independent source signals activate basis functions that describe the observation. For audio signals, Bell and Sejnowski proposed ICA to learn features [1].

Feature extraction for speech recognition aims at an efficient representation of spectral and temporal information of nonstationary speech signals. Conventionally the mel frequency cepstral coefficients (MFCCs) are one of the most common base features to represent spectral characteristics of speech signals. The MFCCs are nearly uncorrelated each other, which is desirable when back-end speech recognizers are based on continuous hidden Markov models (HMMs) using Gaussian mixture observation densities with diagonal covariance matrices. ICA-based feature extraction is data driven and attempts to find a linear transformation such that the resulting coefficients are as independent as possible.

To reduce the temporal correlation, conventionally delta and acceleration components were appended to the base features in the standard HMM-based speech recognizers [2]. Recently research efforts have been made to replace the conventional method by using spectro-temporal orthogonal transforms [3] [4]. Recently a few researchers applied ICA for the same purpose [5] [6] [7].

Prior research on using ICA features for speech recognition resulted in significant improvements [7] but experiments were conducted under constraint settings (small training data). Our goal was to investigate this approach without any constraint setting and provide new analysis and options to cope with the main problems of the standard ICA features in providing features that are phase insensitive and time-shift invariant.

In Section 2, we describe the speech model assumed in the paper, explain the phase problem and propose a new feature extraction method. In Section 3 we analyze the ICA basis functions in the time and spectro-temporal domain and show the potential advantage of ICA by analyzing the conditional probability distribution. In Section 4, TIMIT phoneme recognition results are presented and several issues related with speech recognizers are discussed. Conclusions are presented in Section 5.

2. ICA-Based Feature Extraction

We assume that speech signals are generated by a generative model where speech signals are represented as a linear combination of basis functions weighted by independent source coefficients. A frame of $N$ observed speech samples is represented by a linear combination of $N$ source signals as $x = As$ where $x$ is an $N \times 1$ column vector of the speech samples, $A$ is an $N \times N$ mixing matrix whose column vectors constitute a set of basis functions and $s$ is an $N \times 1$ column vector of the source signals.

When processed in a short segment, speech signals are insensitive to phase variation, as opposed to the case of natural images. According to the preliminary results,
The mel filter was implemented to work with frames with the MFCC-based method as the ICA filters have different filters in the proposed method are different from those in cosine transform (DCT) and temporal filtering. The mel temporal domain (ICA2) plays the role of the discrete filter bank. PCA and ICA in the spectro-temporal domain (ICA1) replaces the FFT of the MFCC-related weights.

However, the resulting coefficients have similar characteristics to non-uniform center frequencies and non-uniform filter weights.

We alleviated the phase sensitivity problem by using the analytic ICA filters (ICA1) and taking the magnitude of complex ICA coefficients. The shift sensitivity problem was mitigated by using a mel filter and summing squared magnitude assigned to the same mel band. However, the resulting coefficients have similar characteristics to the standard mel filter bank coefficients except non-uniform center frequencies and non-uniform filter weights.

Fig. 1 shows feature extraction method using ICA. Comparing the MFCC feature extraction, the ICA in the time domain (ICA1) replaces the FFT of the MFCC-based method, and the PCA and ICA in the spectro-temporal domain (ICA2) plays the role of the discrete cosine transform (DCT) and temporal filtering. The mel filters in the proposed method are different from those in the MFCC-based method as the ICA filters have different center frequencies.

A stream of speech signals is segmented into a series of frames with \( N \) samples and each frame is windowed by a Hamming window. In the following sections, we omit the frame index \( t \) unless confused, assuming that all processing is done in the frame base.

### 2.1. Analytic ICA in the Time Domain (ICA1)

We used the Infomax algorithm [8] to obtain the basis functions and the corresponding coefficients. To reduce the phase sensitivity, we used the analytic version of the unmixing matrix, which was obtained via the Hilbert transform: \( B = B + j\bar{B} \) where \( \bar{B} \) is the Hilbert transform of \( B \) in the row direction and \( j = \sqrt{-1} \). By using the analytic version of the unmixing matrix, we can obtain a smoother estimate of the \( i \)-th coefficient magnitude \( r_n(i) = ||B_i, x||^2 \), \( i = 1, \ldots, M \) where \( B_i \) is the \( i \)-th row vector of the analytic unmixing matrix. While a fixed transformation matrix is used to compute FFT coefficients, the ICA1 learns the filters with non-uniform center frequencies and weights from speech signals.

### 2.2. Mel Filter

Mel band energies were obtained by weighting the magnitude coefficients considering the center frequency of the mel bands [2] and the center frequency of the ICA filters. Investigating the spectro-temporal characteristics of the basis functions of ICA in the time domain, which will be shown later, this is useful especially in the high frequency band where the corresponding basis functions have narrow temporal span. The logarithm of the resulting coefficients were taken from the fact that human auditory system is sensitive to speech loudness in the logarithmic scale. The output vector \( g \) is used as a base for the following feature transformation stage to remove temporal dependencies between frames and obtain components that are as independent as possible by using another ICA step.

### 2.3. ICA in the Spectro-Temporal Domain (ICA2)

We concatenate \( 2\Delta + 1 \) consecutive frames of \( g \) to form a new vector at time \( t \), \( h(t) \). Because the DC component of \( h \) does not have a sparse distribution, we subtracted the local mean of \( h \). Before applying the second ICA, we performed PCA to reduce dimension first to the number of the target coefficients \( L \), following the procedure described in Section 2.1. We attempted to reduce the dimension to a larger dimension than \( L \) and select \( L \) basis functions, which turned out to yield worse accuracy. Therefore we directly reduced the dimension to \( L \) before ICA in the spectral domain.

The basis functions of ICA in the spectro-temporal domain is learned in the same way as those of ICA in the time domain except that the input vectors are subtracted by local mean. The resulting coefficients of ICA2 can be used for recognition purpose. In this case, phase sensitivity indicates spectral change or presence of phoneme boundaries. PCA is optimal to decorrelate signals generated by a single Gaussian density. But speech recognizers use Gaussian mixture models and hence an independence constraint is more desirable as in this case.

### 3. Analysis of ICA Basis Functions

#### 3.1. Speech Database

Using the TIMIT speech database, we analyzed the basis functions of ICA and evaluated the performance of the ICA-based feature in a speaker-independent phoneme recognition task. The sampling rate of the database was down-converted to 8 kHz to reduce the training time for ICA. To train the ICA filter, we used 1 hour of speech data from the training set. For evaluation, we used the core test set which includes 192 sentences by 16 male and 8 female speakers.
3.2. Basis Functions of ICA in the Time Domain

Fig. 2 shows the Wigner-Ville distribution (WVD) of the basis functions [7] when the window size is 10 ms. Each contour line represents the locus of half of the maximum amplitude and each cross denotes the location where the maximum amplitude occurs. The use of the WVD caused narrower time span in the low frequency region than the exact time span obtained from visual inspection of the basis functions. The figures indicate that the basis functions in the middle-to-high frequency band are localized both in the frequency and temporal directions. The first 20 basis functions covered the 0-4 kHz frequency range where most of speech signal energy is distributed. The narrow width of the basis functions in the temporal direction implies the shift variant property of the basis functions.

In case of 20 ms window size, the temporal range of the basis functions in the high frequency region was similar to that of 10 ms window size. In both cases, the basis functions in the low frequency region spanned the whole frame length. The basis functions resembled sinusoidal waveforms and more basis functions covered the whole time span, which is similar to the animal vocalization case in [9]. Note that the WVD figures in this work are a little different in the low frequency region from [7] because input signals were preemphasized.

3.3. Basis Functions of ICA in the Spectro-Temporal Domain

To illustrate the basis functions of the ICA in the spectro-temporal domain, we trained the ICA filters by using the concatenated log energy coefficients, which are more Gaussian-like than power signals. In terms of signal processing, it may be reasonable to use power signals. However in view of phoneme recognition, we argue that log energy is a plausible quantity considering logarithmic loudness sensitivity curve of the human auditory system. Recent study results also shows the validity of ICA in the spectral domain [5] [6].

The parameters used for learning were the same as those for the ICA1. The window size of the ICA1 was 25 ms and the number of the source signals was 128. We decided to produce the same number of the final coefficients to compare with the standard MFCC feature, which made us use 38 basis functions from 9 frames of the 23-dimensional coefficient vectors. Fig. 3 shows the learned 38 basis functions displayed as an image patch. The horizontal axis denotes temporal index and the vertical axis denotes the index to the mel-filtered log coefficients. The gray level of the image patches denotes the value of the basis functions normalized to 0 (black) to 255 (white) between the minimum and maximum values of the basis functions. The first 5 basis functions represent the temporal changes at phoneme boundaries and the horizontal stripes represent the spectral distribution of speech signals. We also observe that some basis functions (e.g., the 21, 22, 24-th ones) are localized in the spectral direction.

We computed the conditional probability distribution to check the independence of output signals in each step of the proposed method. Fig. 4 shows the conditional probability distributions of output coefficients of ICA1, mel filter bank coefficients, coefficients transformed by DCT, coefficients transformed by PCA, and coefficients by ICA.
Table 1: Phoneme accuracy (%) with different window sizes and numbers of source signals when $\Delta = 4$

<table>
<thead>
<tr>
<th>$N$</th>
<th>$M$</th>
<th>ICA-FBANK</th>
<th>ICA-PCA</th>
<th>ICA-ICA</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>80</td>
<td>25.90</td>
<td>45.97</td>
<td>48.18</td>
</tr>
<tr>
<td>128</td>
<td>128</td>
<td>27.57</td>
<td>47.14</td>
<td>48.23</td>
</tr>
<tr>
<td>160</td>
<td>128</td>
<td>30.52</td>
<td>50.44</td>
<td>50.89</td>
</tr>
<tr>
<td>200</td>
<td>128</td>
<td>29.88</td>
<td>49.70</td>
<td>50.78</td>
</tr>
</tbody>
</table>

4. Phoneme Recognition Results

For comparison purposes, the hidden Markov toolkit (HTK) was used to extract the standard MFCC feature with 23 frequency bands [2]. The window size was 25 ms and each frame was shifted by 10 ms. The speech signals were preemphasized by an FIR filter with a factor of 0.97. We did not use any language model to evaluate only the performance of feature extraction. We used 48 context-independent phoneme models including silence. Every phoneme model except short pause was modeled by a 3-state left-to-right HMM and short pause was modeled by a single-state HMM. We used continuous HMM with observation probability distribution of 4 Gaussian mixtures for each state. Phoneme accuracy was converted to a value corresponding to the standard 39 phoneme set.

We first evaluated accuracy of the baseline system with the standard features, which was used as the reference for the subsequent experiments. With the standard 23-dimensional filter bank feature, we obtained the phoneme accuracy of 32.18%. The DCT applied to the log energy coefficients of the filter bank feature yielded 36.45% of phoneme accuracy. The DCT improved accuracy by decorrelating the filter bank coefficients. We obtained 50.87% accuracy by appending the delta and acceleration components (MFCC\_E\_D\_A) with $\Delta = 2$.

We tested the proposed ICA-based feature with different window sizes ($N$) and different numbers of source signals ($M$). We set $\Delta$ to 4 so that the number of coefficients used for temporal filtering is equivalent to the standard feature case (MFCC\_E\_D\_A). We found that the window size 160 (20 ms at the 8 kHz sampling rate) achieved the best accuracy as shown in Table 1. The ICA-based feature with multiple frames yielded recognition results comparable with the MFCC\_E\_D\_A feature.

Our results are different from [5] and [6] in that our method uses the mel-filtered band energies from the time domain ICA as the base feature while they use the MFCC instead. When the number of mixtures increases, the performance difference is narrowed as pointed out in [6].

Although one would expect better speech recognition accuracy due to improved encoding of the speech signal, we were not able to meet the expectation mostly due to the nature of current speech recognition systems that are optimized for features that produce a Gaussian density fit. ICA for speech signal representation assumes a super-Gaussian density model for source signals, which is different from a Gaussian density used as the observation density in common HMM-based speech recognizers. There is still some mismatch between some known properties of speech perception and recognition.

5. Conclusions

We investigated the effectiveness of ICA-based feature extraction. To alleviate the phase shift sensitivity and time shift problems in using the conventional ICA coefficients, the analytic version of ICA filters was used and the outputs from adjacent ICA filters were summed. It was found that applying nonlinear operations such as the log operation and converting ICA coefficients by using an additional transform was an effective method to reduce the dependencies among the source coefficients. We also analyzed the time-frequency characteristics of the learned basis functions. Phoneme recognition results were similar to the MFCC-based feature speech recognition system. Further research is needed in finding appropriate nonlinear transformations to accommodate the human perceptual mechanisms in the spectral domain.

6. References