ICA-Based Speech Features in the Frequency Domain

Włodzimierz Kasprzak, Adam F. Okazaki, and Adam B. Kowalski

Institute of Control and Computation Engineering,
Warsaw University of Technology,
ul. Nowowiejska 15-19, PL - 00-665 Warsaw, Poland
W.Kasprzak@ia.pw.edu.pl
http://www.ia.pw.edu.pl/

Abstract. We apply the technique of independent component analysis to Fourier power coefficients of speech signal frames for a blind detection of basic vectors (sources). A subset of sources corresponding to the noisy influence of basic frequency is identified and its corresponding features could be eliminated. The mixing coefficients for such sources are then determined for every speech sample. We compare our features with the Mel Frequency Cepstrum Coefficient (MFCC) features, widely used today for phoneme-based speech recognition.

1 Introduction

It is common in automatic speech recognition systems to apply a frame-based segmentation of the signal, i.e. to use short-time frames [1], [2] in which a windowed Fourier transform is performed. Although specific features of a single frame can be detected already in the time-domain (like LPC features), there are widely used Mel Frequency Cepstrum Coefficients (MFCC) [2], [3] which are computed in the "cepstral" space (this needs a homomorphic filtering via the Fourier space back to the time domain and a post-processing step called "liftering").

It was observed, that statistical cues could offer increased power to speaker recognition systems [4], [5]. In this context the two techniques - PCA and ICA - can be considered [6], [7]. Different authors derive the principal component analysis (PCA) or ICA [4] of the power spectra vectors, which are also smoothed using Mel-scale triangular filters. The authors of [5] assume that the spectra of sounds generated by a given speaker can be synthesized using a set of speaker specific basis functions - the unknown source in the ICA model.

In this paper we follow this idea and we expect the Fourier power coefficients of a single frame to be mixtures of a set of basic, statistically independent vectors. In section 2 the problem of speech feature detection is introduced. The proposed approach is described in section 3 and simulation results follow in section 4.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 609–616, 2006.

2 The MFCC Features for Speech

The task of ICA is to find the waveforms $s_i(t)$ of the sources, knowing only the mixtures $x_j(t)$ and the number m of sources [6]. A well-known iterative optimization method the stochastic gradient (or gradient descent) search [8] can be applied in this context. Especially for the ICA problem different gradient approaches were developed (e.g. the *natural gradient descent* [9]). An efficient "batch" approach is the method "FastICA" [7]. The batch processing allows a preliminary "whitening" step for the zero-mean mixture signals, which improves the convergence speed of the ICA procedure.

2.1 Energy of Speech Samples

As illustrated in Fig. 1 and 2 the energy distribution in time of the same spoken word significantly differs from sample to sample and from speaker to speaker. Hence, we need a feature scheme which is rather interested in the "waveform" or relative (normalized) energy pattern than in the global energy distribution. In some applications the possibility to achieve ICA demixing results with respect to a scaling factor only is a disadvantage of the ICA approach. In case of speech features no such drawback should appear.

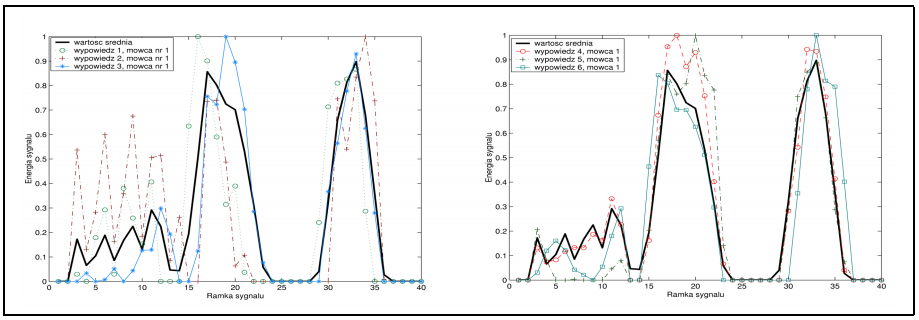

Fig. 1. Energy distribution in time of some polish word "pusc" (release): 3 (left) and 3 (right drawing) samples with their averages (bold lines) for one speaker

In order to limit the variability of energy distribution among speakers and due to different emotional attitude of the speaker we make an energy normalization step before feature detection. As our goal is to extract ICA-based features and to compare them with the MFCC features only, without performing general word recognition, we can deal with the necessary time stretching by performing an interpolation-based resampling in the time domain in advance of the feature detection step. In this way we assure that the current utterance and the pattern utterance have both the same number of samples (for every word a different number of samples is usually required).

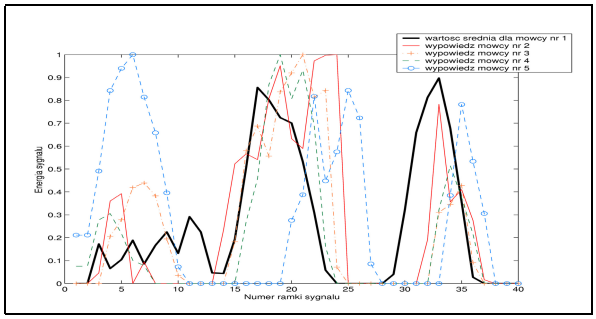

Fig. 2. Energy distribution in time of polish word "pusc" (release) - the averages for 5 different speakers

2.2 The Standard MFCC Features

The Mel-cepstrum features are the result of the characteristic (homomorphic) transformation $MFCC(h) = FT^{-1}\{MFC\{FT\{h\}\}\}$ for $\mathbf{h} = \mathbf{x} \otimes \mathbf{w}$ (a convolution of $\mathbf{x}$ with $\mathbf{w}$).

The short-term power spectrum is computed by applying the discrete Fourier Transform (DFT) (in fact the FFT) to each windowed signal and taking directly the magnitudes of Fourier coefficients raised to the power of two. The power spectrum is usually represented on a log scale.

A MEL scale (empirical result) adopts the frequency bandwidths to the bandwidths recognized by the human auditory system. The set of Fourier features is reduced by considering bandwidths, centered around some MEL scale frequencies. Usually one uses a set of l triangle filters $D(l,t)$ to compute l so called Mel-spectral coefficients $MFC(k,t)$ for every signal frame t.

A disadvantage of Fourier coefficients, even after consolidation by triangle filters, is the joint correlation of neighbor frequency coefficients. Since the vocal tract is smooth, energy levels in adjacent bands tend to be correlated. To compensate this smoothing of features the inverse DFT (in fact only the cosine transform as the transformed MFC's are real-valued) is applied, which converts the set of logarithm-scaled energies to a set of cepstrum coefficients (for example, m = 12), which are largely un-correlated:

$$MFCC(k,t) = \sum_{l=0}^{M-1} log[MFC(l,t)] \cos\left[\frac{k(2l+1)\pi}{2M}\right],\ k = 1, ..., 12. \quad (1)$$

Another disadvantage of this scheme is that noisy oscillations of the human larynx are overlayed onto the energy of basic frequency and some of its first harmonic frequencies. To reduce it a so called *liftering* of the MFCC features is finally performed [2], [3]. Let c_n be the n-th MFCC. Then its liftering is as follows:

$$c_n^{lift} = \left[1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right)\right] c_n,\ n = 1, 2, ..., K < L, \quad (2)$$

where L is related to the feature index for the basic frequency. Usually the final number of features L is set by default to the number of triangle filters, as the on-line computation of this parameter for every consecutive frame is not feasible: (a) a variable number of features could appear for different frames, (b) although the basic frequency is related to the individual speaker, it is variable even for the same speaker, as it depends on the accentuation and emotional standing.

3 The Approach

3.1 Applying ICA for Source Separation

The Fourier coefficients obtained for given frame $FC(a,t)$ constitute a vector $\mathbf{x}(t)$ - a single (mixture) input to the ICA learning procedure (the size of this vector is N). This vector is expected to be a particular mixture of $m < N$ independent sources. For every spoken word, that we can detect in the speech sample, we get a learning set of frames: $x_i(t)(i = 1, ..., n; t = 1, ..., N)$.

The basic mixing model in ICA (without noise) is assumed. $\mathbf{x}(t)$ is a matrix of n time-varying vector signals, each of size N. $\mathbf{a}_i$ is a set of n mixing vectors (each of size m) combined to a mixing matrix $\mathbf{A}$ (every $\mathbf{a}_i$ is a single row of matrix $\mathbf{A}$). $\{\mathbf{s}_i(t)\}$ is a set of m sources - each one consists of N time samples.

After running the ICA method both unknown sources and unknown mixing coefficients are determined - on base of given sequence of observations (frames) $x_i(t)$ the vector $\mathbf{s}$ and weight matrix $\mathbf{W}$ are estimated. The sources need to be normalized and reordered, while the weights are of no importance during learning.

3.2 Matching of Source Sets

During learning we need to establish a correspondence between existing source set (the reference components) and the newly created source set for current signal frame. During this comparison a proper permutation index, the scaling and even the sign of amplitude must be adjusted [10]:

(1) The amplitudes of all components are re-scaled to the interval of $< -1, 1 >$.
(2) FOR all tested components $y_i, (i = 1, ..., n)$ DO:
FOR all reference components $s_j, (j = 1., ..., n)$ DO:
compute the mean square error of approximating s_j by y_i or by $-y_i$:
$MSE[y_i, s_j]$ and $MSE[-y_i, s_j]$
and select the better one, i.e. with lower value;
(3) All selected MSE-s are transformed into elements of a new created matrix $\mathbf{P} = [a_{i,j}]_{n \times n}$, where $a_i = \frac{1}{\sqrt{MSE[y_i, s_j]}}$.
(4) The error index $EI(\mathbf{P})$ is computed as:

$$\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{a_{ij}}{max_i(a_{ij})} - n\right] + \frac{1}{n}\left[\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{a_{ij}}{max_j(a_{ik})} - n\right].$$

The first part of above sum expresses the average error for matching a tested ICA component with one reference component, whereas the second part is equivalent

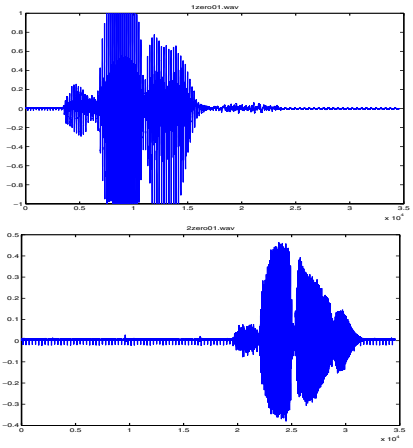

Fig. 3. Waveforms of the word "zero" pronounced by two speakers (male and female)

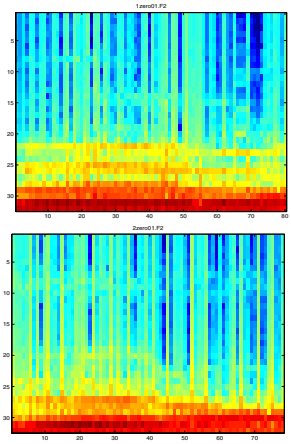

Fig. 4. The spectrograms (selected frames with sufficient energy only) for above words "zero"

to a penalty score, if a single reference component is matched with more than one tested component.

3.3 Larynx Noise Detection

Some of the sources correspond to the noisy influence of basic oscillations of the larynx. In MFCC scheme they are tried to be eliminated by the "liftering" processing. In case of our ICA scheme these "noisy" sources are detected by their continuously decreasing waveform, with its highest value at the index of 0. The remaining sources are equipped with one or several local maxima at particular frequency indices (see Fig. 3 and 5).

3.4 Feature Extraction

For every signal frame we need to determine a feature vector in the previously established ICA space determined by the selected source set. These features are equivalent to the unknown mixing coefficients of ICA sources that lead to the power spectrum vector for current frame. Hence, let us assume the matrix $\mathbf{S}$, with rows representing the reference ICA sources in frequency space $s_i(\omega)$, was established during the learning phase. One part of these sources forms the feature-relevant base $\mathbf{S_F}$ and the other part - the larynx-related part $\mathbf{S_L}$ of matrix $\mathbf{S}$. Then we estimate the unknown mixing coefficients for current window k of the speech signal as: $a_k^T = x_k(\omega)\mathbf{S}^{-1}$, where $x_k(\omega)$ is the vector of power spectra for the k-th window of speech. The final feature vector is a sub-vector of a_k corresponding to the subspace determined by $\mathbf{S_F}$.

An illustration of ICA features detected for the source set in Fig. 5 is specified in Fig.7 and 8. We observe that the coefficients $\mathbf{W}$ for different words "jeden" and "dwa" with the same ICA components are quite different, but for the same word and even different speaker - these coefficients are similar.

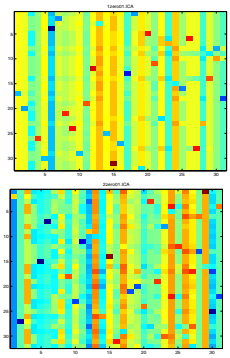

Fig. 5. The detected 31 basic vectors (one column represents one vector with 32 elements) after ICA was applied to above two spectral images

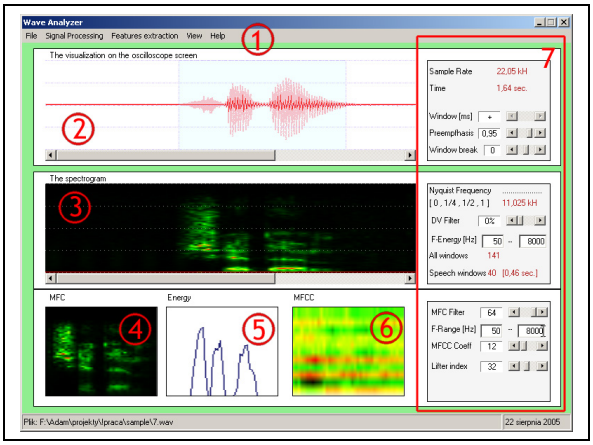

Fig. 6. The main window of our test application: (1) menu, (2) oscillogram, (3) spectrogram, (4) MFC, (5) energy, (6) MFCC or ICA, (7) analysis parameters

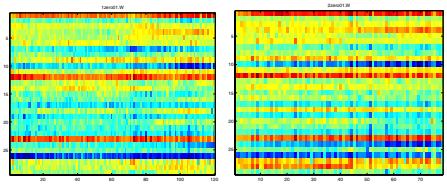

Fig. 7. The coefficients **W** (one column represents one vector of coefficients for one signal frame) for two speech samples of word "zero" from different speakers. Great similarities appear.

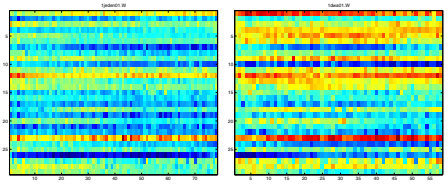

Fig. 8. The coefficients **W** for different words "jeden" (one) and "dwa" (two) from the same speaker. Large differences appear.

4 Experimental Results

Both the MFCC- and ICA-based approaches for speech frame feature detection were implemented and tested on speech signal examples, acquired with the sampling frequency of 22 kHz. Speech samples from 18 persons (both male and female) were available for testing (Fig. 6).

The MFCC and ICA features are quite stable for different samples of the same word and speaker (see Fig. 9). For different speakers a larger variability appears (Fig. 10).

Some experiments of both approaches are summarized in tables 1-2, where the EI index values were computed while matching the tested sample components

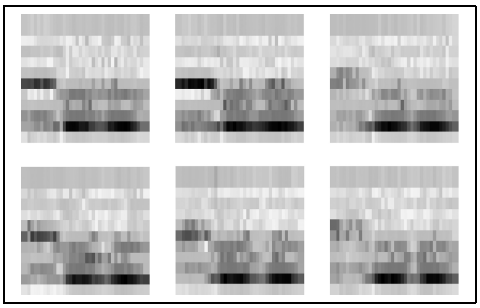

Fig. 9. MFCC features for speaker 1 and word "pusc" (release) - 6 different samples

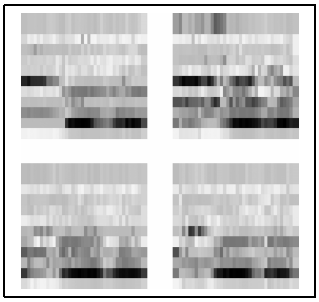

Fig. 10. MFCC features for speakers 2-4 and word "pusc" (release)

Table 1. Comparison of the error index $EI(\mathbf{P})$ for components - sources - of the same word ("zero") but for 4 different speakers (31 sources with 32 elements)

Reference / *Tested*	M1	F1	M2	F2
Male 1	6.04	4.46	5.15	3.90
Female 1	6.15	4.62	5.85	5.56
Male 2	6.21	4.47	5.13	4.70
Female 2	9.03	8.47	7.45	7.92

Table 2. Comparison of the error index for components - sources - of different words ("zero", "jeden", "dwa") but the same speaker (31 sources with 32 elements)

Reference / *Tested*	"zero"	"jeden"	"dwa"
"zero"	3.46	2.50	1.98
"jeden"	2.33	2.82	1.20
"dwa"	2.66	2.94	1.85

Table 3. The classification success rate for the MFCC- and ICA-based feature sets (20 classes with 26 learning and 12 verification samples for each class - different speakers)

Feature set / *Word*	MFCC	ICA
"zero"	66%	100%
"jeden" (one)	58%	100%
"dwa" (two)	63%	66%
"trzy" (three)	58%	100%
"cztery" (four)	58%	100%
...		
"dziewiec" (nine)	83%	100%
"start"	66%	66%
"stop"	92%	83%
"lewo" (left)	66%	91%
"prawo" (right)	66%	66%
"gora" (up)	89%	100%
"dol" (down)	75%	83%
"pusc" (release)	91%	91%
"zlap" (catch)	91%	91%
"os" (axis)	83%	91%
"chwytak" (grab)	66%	100%
average	75.6 %	90.55%

with the proper reference components. From Table 1 it is evident, that the components are quite independent from the speaker. From Table 2 we conclude that ICA sources, obtained for different words of one speaker, are also similar. Hence, ICA produces a general base for speech features.

The last table 3 summarizes a comparison between MFCC features and ICA-based features. A word reference (class) was represented by an average map of all the learned feature maps for given word. We applied a simple minimum-distance classifier for the classification of feature sets, computed in both schemas - MFCC and ICA. A success was noted if the minimum distance was achieved for the proper reference word and the distance between current feature map and reference map was below half of the standard deviation of samples for given class.

5 Conclusion

We have proposed an ICA-based method for speech feature detection in a frame-based speech recognition system. A subset of sources detected by ICA provides base vectors of the feature space in the frequency domain, whereas the mixing coefficients in ICA mixing model constitute the feature vectors. The experiments show a better quality (in terms of the recognition success rate) of such features if compared to standard MFCC features.

Acknowledgment. The work reported in this paper was supported by the grant MNiI - 4T11A 003 25.

References

1. Junqua, J.-C., Haton, J.-P.: Robustness in automatic speech recognition. Kluwer Academic Publications, Boston etc. (1996)
2. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice Hall, New York (1993)
3. Grocholewski, S.: Statystyczne podstawy systemu ARM dla jezyka polskiego. Vol. 362 of "Rozprawy", Poznan University of Technology Press (2001)
4. Ding, P., Kang, X., Zhang, L.: Personal recognition using ICA. Proceedings ICONIP (2001)
5. Rosca, J., Kopfmehl, A.: Cepstrum-like ICA representations for text independent speaker recognition. Procceedings of ICA'2003, (Nara, Japan, April 2003), Publ. by NTT Kyoto (2003) 999–1004.
6. Cichocki, A., Amari, S.: Adaptive Blind Signal and Image Processing. John Wiley, Chichester, UK (2002)
7. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, New York etc. (2001)
8. Duda, R.O., Hart, P.E., Stork, D.: Pattern classification. 2nd edition. John Wiley, New York (2001)
9. Amari, S., Douglas, S.C., Cichocki, A., Yang, H.Y.: Novel on-line adaptive learning algorithms for blind deconvolution using the natural gradient approach. IEEE Signal Proc. Workshop on Signal Processing Advances in Wireless Communications, (Paris, April 1977), 107–112
10. Kasprzak, W.: Adaptive computation methods in digital image sequence analysis. Vol. 127 of "Prace Naukowe - Elektronika", Warsaw University of Technology Press, Warszawa, Poland (2000)