A COMMON FEATURE REPRESENTATION SCHEME FOR SPEECH FRAMES AND IMAGE CONTOURS

Włodzimierz Kasprzak
Adam F. Okazaki
and Robert Seta
Institute of Control and Computation Eng.
Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warszawa, Poland
W.Kasprzak@ia.pw.edu.pl, A.Okazaki@elka.pw.edu.pl, R.Seta@elka.pw.edu.pl

Keywords: frequency-based decomposition, shape features, speech features, statistical independence.

Introduction

In contour-based shape representation different schemas exist [1], like order numbering, geometric moments [2], Fourier-based coefficients [3] or autoregressive (AR) models [4]. In [3] an interesting comparison is described between AR-methods and Fourier-based methods for contour feature representation. The authors find a best specific Fourier-based scheme, that is well suited for the recognition of contours of letters or aircrafts. Unfortunately the Fourier coefficients tend to be not independent but mutually correlated, i.e. some post-processing of them is needed.

This problem was long ago discovered by the speech recognition community. A basic problem in automatic speech recognition is to detect such features, which are of general nature and does not depend much on the speaking person [5]. Nowadays it is common to apply a frame-based segmentation of the speech signal; i.e. to use short-time frames [6]. Specific features of a single frame are detected either in the time-domain (like LPC features, Fourier space (power coefficients) or in "cepstral" space. The last scheme obtains the widely used so called *Mel Frequency Cepstrum Coefficients (MFCC)* [6]. The MFCC's require a post-processing of the power spectra of signal windows and an inverse transformation into the time domain.

Recently it was observed, that statistical cues could offer increased power to speaker recognition systems while remaining in the Fourier space [7], [8]. The principal component analysis (PCA) [9] or independent component analysis (ICA) [10] of the power spectra vectors could be performed [7]. Resulting

features are further narrowed down using a linear discriminate based criterion. The authors of [8] suggested 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.

Following this idea, in this paper we also expect that the Fourier power coefficients of a single frame are mixed from a set of independent basic vectors. Having fixed the set of basic vectors - after the learning phase - in the active analysis step we identify the mixture coefficients for each frame - they constitute the feature vector for given signal frame. The same idea motivates our application of the ICA technique for frequency representations of 2-D contours. The application of ICA for image decomposition (but dominantly in the space domain) started with the paper [11] and meantime different mixed techniques have been proposed, for example the ICA of Gabor features [12].

1. The standard MFCC features

The so called *Mel features of complex cepstrum* are the result of a characteristic (homomorphic) transformation $MFCC(h) = DFT^{-1}[MFC\{DFT(h)\}]$ for $\mathbf{h} = \mathbf{s} \star \mathbf{w}$ (a convolution of signal **s** with window **w**).

The short-term power spectrum is computed by applying the discrete Fast Fourier Transform (FFT) to each windowed signal and taking directly the magnitudes of Fourier coefficients raised to the power of two.

$$FC(k,\tau) = |\frac{1}{M}\sum_{t=0}^{M-1}[x(\tau+t)e^{-i2\pi kt/M}w_\tau(t)]|^2,\ k = 0,1,...,M-1. \tag{1}$$

A MEL scale (empirical result) adopts the frequency bandwidths to the bandwidths recognized by the human auditory system: $f_{mel} = 2595log(1 + f/700Hz)$. 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 $D(l,k)$ to compute l so called Mel-spectral coefficients MFC(l, t) for each signal window τ:

$$MFC(l,\tau) = \sum_{k=0}^{M-1}[D(l,k)FC(k,\tau)],\ \text{ for } l = 1,...,32 = N. \tag{2}$$

Since the vocal tract is smooth, energy levels in adjacent bands tend to be correlated. The inverse DFT (in fact only the cosine transform as the transformed MFC's are real-valued) converts the set of logarithm-scaled energies to a set of cepstrum coefficients (usually m = 12), which are largely un-correlated:

$$MFCC(k,\tau) = \sum_{l=0}^{M-1} log[MFC(l,\tau)]cos\left(\frac{k(2l+1)\pi}{2M}\right),\ k = 1,...,12. \tag{3}$$

2. Fourier coefficient-based features of 2-D contours

In order to apply the 1-D FFT to a contour its 1-D representation should first be acquired. We could generate at least three such representations: (1) a *curvature* function - angular changes of the boundary tangent along the contour chain; (2) a *centroidal distance* function - distances of the boundary points from the centroid of the contour (x_C, y_C); (3) a *complex coordinate* function - coordinates of the boundary pixels in an contour centered coordinate system:

$$z_i = (x_i - x_C) + j(y_i - y_C). \tag{4}$$

Without loss of generality of the approach we shall use the (3)-th representation, that was tested to behave best with respect to Fourier-based features [3].

After pixel number normalization of the contour representation to a fixed number of $N = 2^n$ samples, the Fast Fourier Transform (FFT) can be applied. The DC component in Fourier space depends only on the position of the contour, hence it is not needed (a translation invariance of contour features is achieved). Orientation invariance of Fourier-based features is usually achieved by phase normalization - here the use of absolute values is a simplified way to achieve this. In order to make it scale invariant, the Fourier coefficients are divided by the absolute value of the first non-zero frequency coefficient. Thus the feature vector obtained in the *Complex Contour Fourier* method is:

$$x_i = \left[\frac{|F_{-(N/2-1)}|}{|F_1|}, ..., \frac{|F_{-1}|}{|F_1|}, \frac{|F_2|}{|F_1|}, ..., \frac{|F_{N/2}|}{|F_1|} \right]^T. \tag{5}$$

3. The common coding scheme in frequency space

An efficient "batch" approach is the method "FastICA" [10]. The batch processing allows a preliminary "whitening" step for the zero-mean mixture signals, which improves the convergence speed of the ICA procedure.

In our approach we consider frames of speech or chain representations of image contours. The Fourier coefficients for given frame or contour $FC(., t)$ constitute a single vector $\mathbf{x}(t)$ - a single (mixture) input to the ICA learning procedure (the size of vector is N). This is expected to be a particular mixture of $m < N$ independent sources. From the spoken word we get a learning set of frames for given speaker: $x_i(t)(i = 1, ..., n; t = 1, ..., N)$.

The basic mixing model in ICA (without noise) is assumed:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}. \tag{6}$$

$\mathbf{x}(t)$ 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 $\mathbf{A}$). $\mathbf{s}_i$ is a set of m sources - each of size N.

In the ICA-demixing process both unknown sources and unknown mixing coefficients are determined - on base of given sequence of observations

(frames) $\mathbf{x}(t)$, the vector of sources $\hat{\mathbf{s}}$ and the weight matrix $\mathbf{W}$ are estimated.

$$\hat{\mathbf{s}}(t) = \mathbf{W}\mathbf{x}(t). \tag{7}$$

We assume, that the learned set of sources in ICA (basis vectors) is common to some (all available) speech samples or contour samples. Every single phoneme (in speech) or contour (in an image) is now represented by a vector of mixing coefficients. This vector can be computed from the set of equations:

$$\mathbf{x}_i(t) = \mathbf{a}_i^T \hat{\mathbf{s}}(t). \tag{8}$$

In a directly following classification step the features $\mathbf{a}$ are classified in terms of a phoneme class or a contour class, respectively.

4. Experimental results

The approach described in section 3 was implemented and tested on speech signal examples and chain representations of contours.

For speech tests Polish spoken digits from 18 persons (both male and female) were available. The figures 1 - 4 document the ICA approach for two speakers. We compared the difference of two sets of basic functions, one obtained for the first speaker while the other - for the second speaker. The differences are quite independent from the speaker (Fig. 2).

Hence, let us fix one set of reference ICA components $\hat{\mathbf{s}}$ and estimate the mixing coefficients: $\mathbf{W} = \mathbf{x}\ \mathtt{pinv}(\hat{\mathbf{s}})$, where $\mathbf{x}$ is the selectively chosen spectrogram for given speech sample. In Fig. 3 two sets of coefficients $\mathbf{W}$ are presented for the same word "zero" pronounced by two speakers. It is evident, that both sets are more similar than their spectrograms (shown in Fig. 1).

We also discovered that ICA components obtained for the same speaker but for different words were also similar. Hence our approach seems to produce a general base for speech recognition. In Fig. 4 we see that the coefficient matrices $\mathbf{W}$ for words "jeden" and "dwa" acquired from the mixture with the ICA components coming from word "zero" are quite different, than the coefficients for word "zero". An easy word detection can follow.

For image contour tests we created several hand-written letters and we selected out blood vessels from angiography images, like shown in Fig. 5. In

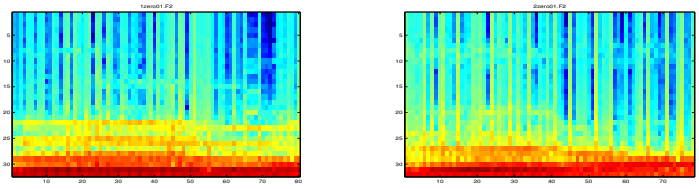

Figure 1. The spectrograms (selected frames with sufficient energy only) for some male and female pronunciation of the word "zero".

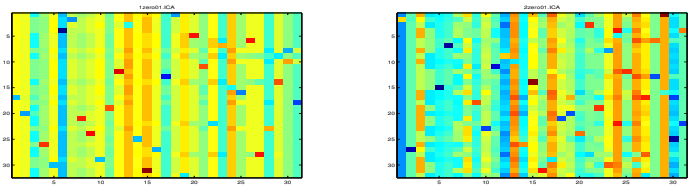

Figure 2. The detected 31 basic vectors (one column represents one vector with 32 elements) after ICA was applied to above two spectral images.

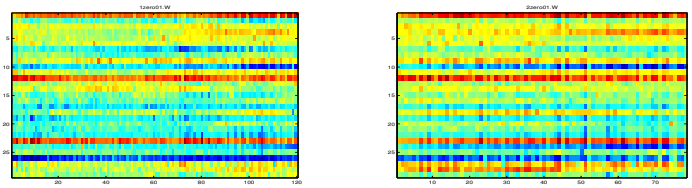

Figure 3. The coefficients $\mathbf{W}$ (one column represents one vector of coefficients for one signal frame) for two speech samples from different speakers, with the same ICA components computed for the word "zero".

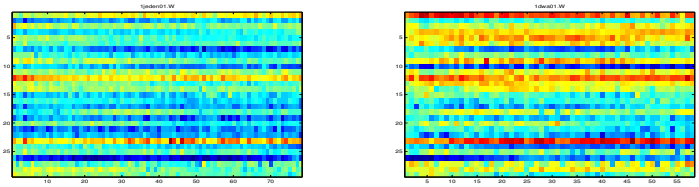

Figure 4. The coefficients $\mathbf{W}$ for different words "jeden" and "dwa" from the same speaker. The ICA components come from a different word "zero".

Table 1 results are shown, that demonstrate ”inner-class compactness” and ”outer-class separability” of different features. This verifies the usefulness of ICA in the post-processing stage of *Complex Contour Fourier* features.

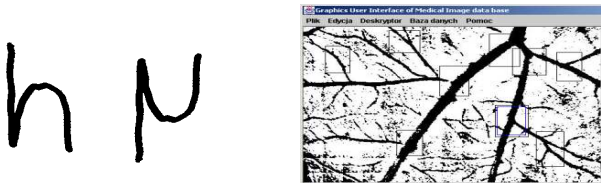

Figure 5. Examples of contours - hand-written letters and blood vessels in images.

Table 1. Comparison of contour features for two sets of images - letters and blood vessels.

Contour features	inner-class	between-class	compact. / separation
	L - B	L - B	L - B
ICA-based	0.414 - 0.330	2.102 - 1.060	0.197 - 0.311
Complex Contour Fourier	0.801 - 0.630	2.804 - 1.920	0.286 - 0.330
2nd order moments	1.28x - 1.11x	4.19x - 3.260x	0.305 - 0.340

5. Summary

We have applied the ICA technique for frequency features of speech or 2-D image contours. Many speech frame- or contour-feature schemas exist, which are heuristically motivated rather than by optimum criteria. The ICA-based features fulfill a well-defined criteria - this leads to optimum compactness and separability of features with respect to learning samples.

Acknowledgments

The authors would like to thank the Rector of Warsaw University of Technology for supporting this work by a grant in year 2004.

References

[1] W. Skarbek: *Multimedia - sprzet i oprogramowanie*. PLJ, Warszawa, 1999.

[2] H. Niemann H.: *Pattern Analysis and Understanding*. Springer, Berlin etc., 1990.

[3] H. Kauppinen, T. Seppanen, M. Pietikainen M.: An Experimental Comparison of Autoregressive and Fourier-Based Descriptors in 2D Shape Classification. *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 17 (1995), No. 2, 201-206.

[4] I. Sekita, T. Kurita, N. Otsu: Complex autoregressive model for shape recognition. *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 14 (1992), 489-496.

[5] J.-C. Junqua, J.-P. Haton: *Robustness in automatic speech recognition*, Kluwer Academic Publications, Boston etc., 1996.

[6] L. Rabiner, B. Juang: *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[7] P. Ding, X. Kang, L. Zhang: Personal recognition using ICA, *Proceedings ICONIP*, 2001.

[8] J. Rosca, A. Kofmehl: Cepstrum-like ICA representations for text independent speaker recognition, *Procceedings of ICA'2003*, (Nara, Japan, April 2003), NTT Kyoto, Japan .

[9] A. Cichocki, S. Amari: *Adaptive Blind Signal and Image Processing*, John Wiley, Chichester, UK, 2002.

[10] A. Hyvarinen, J. Karhunen, E. Oja: *Independent Component Analysis*, John Wiley & Sons, New York etc., 2001.

[11] A.J. Bell, T.J. Sejnowski: The 'independent components' of natural scenes are edge filters. *Vision Research*, vol.37 (1997), 3327-3338.

[12] C. Liu, H. Wechsler: Independent Component Analysis of Gabor Features for Face Recognition. *IEEE Transactions on Neural Networks*, vol. 14 (2003), No. 4, 919-928.