www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

ARCHIVES OF ACOUSTICS
**36**, 2, 311–331 (2011)

*DOI: 10.2478/v10168-011-0024-x*

# ICA-based Single Channel Audio Separation: New Bases and Measures of Distance

Dariusz MIKA[(1)], Piotr KLECZKOWSKI[(2)]

[(1)] *Studio sQuat Professional Sound Studio Recording*
Pl. Tysiąclecia PP 1, 22-100 Chełm, Poland
e-mail: 75mika@wp.pl

[(2)] *AGH University of Science and Technology*
*Department of Mechanics and Vibroacoustics*
Al. Mickiewicza 30, 30-059 Kraków, Poland

(*received February 4, 2011; accepted March 30, 2011*)

Independent Component Analysis (ICA) can be used for single channel audio separation, if a mixed signal is transformed into time-frequency domain and the resulting matrix of magnitude coefficients is processed by ICA. Previous works used only frequency (spectral) vectors and Kullback-Leibler distance measure for this task. New decomposition bases are proposed: time vectors and time-frequency components. The applicability of several different measures of distance of components are analysed. An algorithm for clustering of components is presented. It was tested on mixes of two and three sounds. The perceptual quality of separation obtained with the measures of distance proposed was evaluated by listening tests, indicating "beta" and "correlation" measures as the most appropriate. The "Euclidean" distance is shown to be appropriate for sounds with varying amplitudes. The perceptual effect of the amount of variance used was also evaluated.

**Keywords:** audio unmixing, blind signal separation, independent component analysis, measures of distance.

## 1. Introduction

Blind separation of signals (BSS) from mixtures has extensive bibliography. An excellent survey was written by Cardoso (1998). The most successful approaches were based on Independent Component Analysis (ICA) (Hyvarinen *et al.*, 2001). This method was originally developed to address the cocktail party effect but has gained popularity in a wide range of applications, including e.g. financial analysis.

In audio engineering, besides speech segregation and recognition, signal separation (also called unmixing or demixing) can also be used in automatic music transcription, musical information retrieval systems, speech/instrument identification, forensic audio, karaoke and special sound effects. The number of applications will grow considerably as soon as the technology permits to achieve high perceptual quality of separated sounds.

The ICA technique is most useful when signals are recorded by a set of microphones or sensors where each sensor receives a different combination of source signals. The highest efficiency of separation can be obtained when the number of microphones is greater or equal to the number of sound sources. This arrangement can be represented compactly by a so called mixing equation:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{1}$$

where $\mathbf{x}(t)$ is a column vector containing $m$ values (from $m$ sensors) of observed signals (mixes), $\mathbf{s}(t)$ is a column vector containing $n$ values (from $n$ sources) of source signals, $\mathbf{A}$ is a $m \times n$ mixing matrix, and $m \geq n$. The matrix $\mathbf{A}$ is nonsingular and the solution to the separation problem becomes:

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t), \tag{2}$$

where $\mathbf{y}(t)$ is an estimation of $\mathbf{s}(t)$ and $\mathbf{W}$ is an estimation of the inverse of $\mathbf{A}$.

Various techniques based on ICA were used for the separation of mixed signals (Lee, Lewicki, 2000; Casey, 2001; Jang, Lee, 2003; Barry *et al.*, 2005). The essence of ICA is in appropriate processing of a mixed multichannel signal so that its constituent components, seen as random variables, are statistically independent (Hyvarinen *et al.*, 2001).

In audio engineering, the need for separation usually occurs in stereophonic or single channel mixes. Therefore, the model given by (1) and (2) can not be directly used. Both stereo and single channel cases, in terms of BSS, are under-determined, as $m < n$. However, in stereo recordings intensity and phase differences between channels can be used as valuable cues to discriminate sources (Barry *et al.*, 2004; Cooney *et al.*, 2006; Master, 2006; Vinyes *et al.*, 2006).

For single channel separation, the collection of perceptually motivated techniques jointly called Computational Auditory Scene Analysis (CASA) (Wang, Brown, 2006) is typically used. The efficiency is limited and some *a priori* knowledge of the signals is often needed. Within these techniques, differences between time-frequency (t-f) distributions of sources are frequently used. Most works in this case were concentrated on the isolation of speech signals (Bach, Jordan, 2005; Yilmaz, Rickard, 2004), and an often exploited cue is so called W-disjoint orthogonality of signals, i.e. their non-overlapping in the t-f plane (Rickard, Yilmaz, 2002; Brungart *et al.*, 2006; Vinyes *et al.*, 2006).

A family of computing methods based on intelligent algorithms, including artificial neural networks, fuzzy sets or rough sets are widely used in recogni-

tion of acoustic signals (Kostek, 2005). Single channel separation is one of the application areas (Dziubiński, Kostek, 2010).

ICA in its basic form can not be applied to single channel separation, as only one observation is available. Jang and Lee (2003) proposed single channel ICA unmixing, where the basis functions were obtained by learning the properties of constituent signals, so that they capture the statistical structures of the sources (Jang, Lee, 2002).

Taghia and Doostari (2009) used band-wide decomposition of components of the mix and applied ICA to a mixed signal in the time domain. The SCICA method (Single Channel ICA) proposed by Davies and James (2007) is also based on the time signal.

Casey (2001) proposed an original approach for single channel ICA demixing. He first obtained a t-f representation of a mixed signal by the Short Time Fourier Transform (STFT) according to:

$$x_{\text{mix}}(t) \xrightarrow{\text{STFT}} \mathbf{TFD}^{\text{mix}}_{[m\times n]} \tag{3}$$

and then treated the rows of the obtained Time Frequency Distribution (**TFD**) matrix as individual channels in a multichannel signal, so that they could be seen as "multichannel" by ICA. He then used ICA estimation to obtain statistically independent components of the t-f representation of a mixed one-channel signal. The results were satisfying. An important advantage of this technique is that it does not require any prior knowledge about the signals to be separated.

A similar, but simplified approach was taken by Barry *et al.* (2005). They performed separation of two sounds by taking just two rows of the **TFD** matrix corresponding to two spectrograms separated by 330 ms and assuming that spectra were stationary over this time. The perceptual results of separation were good, but their sounds (a flute and a bass) were separated spectrally anyway. Wang and Plumbley (2006) used Nonnegative Matrix Factorisation (NMF) on STFT matrix of a single channel mix, but that algorithm needed a proper training set. Mijovic *et al.* (2010), besides the wavelet transform, used a combination of Empirical Mode Decomposition and ICA for the separation of ECG signals and compared their results to the SCICA method. These methods are usually called Spectral-Decomposition-Based methods.

Litvin and Cohen (2009) proposed the Bark Scale aligned Wavelet Packet Decomposition (BS-WPD) instead of STFT, and at the stage of separation used the Gaussian Mixture Model (GMM). Duan *et al.* (2008) presented a combination of various single-channel separation techniques including elements of CASA, Spectral-Decomposition-Based methods and Model-Based methods.

Considering the $\mathbf{TFD}^{\text{mix}}_{[m\times n]}$ matrix as a $m$-channel signal, each consisting of $n$-samples, and performing the ICA estimation upon this multichannel signal we obtain statistically independent components $\mathbf{z}_i$ of the t-f representation of a one channel signal.

The following relation holds between a $\mathbf{TFD}^{\mathrm{mag}}_{\mathrm{mix}}$ matrix, where mag denotes magnitude, and $\mathbf{Z}$ is a matrix of statistically independent components (Casey, 2001):

$$\mathbf{TFD}^{\mathrm{mag}}_{\mathrm{mix}} = [\mathbf{A}]_{[m\times n]} \cdot [\mathbf{Z}]_{[n\times n]} = \sum_i (\mathbf{A}_i \cdot \mathbf{z}_i) = \sum_i \mathbf{TFD}_i = \sum_c \mathbf{TFD}^i, \tag{4}$$

where $\mathbf{TFD}^{\mathrm{mag}}_{\mathrm{mix}}$ is the magnitude form of t-f signal representation $\mathbf{TFD}^{\mathrm{mix}}_{[m\times n]}$, matrix $[\mathbf{A}]$ of dimensions $m \times n$ is a mixing matrix, $\mathbf{A}_i$ is an $i$-th column of $[\mathbf{A}]$, $\mathbf{z}_i$ is an $i$-th row of $[\mathbf{Z}]$, $\mathbf{TFD}_i = \mathbf{A}_i \cdot \mathbf{z}_i$ is an $i$-th t-f component of a mixed one channel signal, $\mathbf{TFD}^i$ is a t-f distribution of an $i$-th constituent signal.

The statistically independent components $\mathbf{z}_i$ (rows of $\mathbf{Z}$) are called spectral bases in this work, and the columns of $\mathbf{A}$, characteristic of time functions of spectral components $\mathbf{z}_i$ – are called time bases (denoted further $\mathbf{T}_i$). The matrix $\mathbf{TFD}_i$ equal to the product of the time basis $\mathbf{T}_i$ and the spectral basis $\mathbf{z}_i$ is called $i$-th t-f component.

Casey (2001) reported that by an appropriate grouping of spectral bases $\mathbf{z}_i$ into $c$ (where $c$ is a number of the constituent signals in the mix) disjoint subgroups we can obtain $c$ bases of subspaces generating the constituent components of the mix (or their t-f representations to be more precise).

Mika (2009) noticed that it was also possible to group $\mathbf{T}_i$ or $\mathbf{TFD}_i$ into $c$ bases and obtain similar results. This option offers additional flexibility in single channel unmixing.

In this work, Casey's idea is generalised, by investigating the performance of $\mathbf{T}_i$ and $\mathbf{TFD}_i$ bases and by extensive examination of the properties of different measures of distance. The performances of those measures were mainly tested with the time ($\mathbf{T}_i$) bases. For spectral $\mathbf{z}_i$ and t-f ($\mathbf{TFD}_i$) bases only one measure of distance was tested, respectively. Two ways of verification of the results were used: perceptual, by way of listening tests, and graphical, by visual inspection of t-f spectra of separated signals. The results are discussed in Secs. 4 and 5 and in Conclusions.

## 2. Procedure

In practical applications only those components of $\mathbf{A}_i$ and $\mathbf{z}_i$, are used for which the variance exceeds a pre-determined threshold. Assuming that we will use only $\varphi$ of signals' variance, Eq. (4) takes the form:

$$\mathbf{TFD}^{\mathrm{mag}}_{\mathrm{mix}} = [\mathbf{A}]_{[m\times n]} \cdot [\mathbf{Z}]_{[n\times n]} \approx \sum_{i,\sigma=\varphi} (\mathbf{A}_i \cdot \mathbf{z}_i) = \sum_{i,\sigma=\varphi} \mathbf{TFD}_i = \sum_c \mathbf{TFD}^i, \tag{5}$$

where $\mathbf{TFD}_i$ components are summed from the largest value of variance to the lowest until the pre-determined value of variance $\varphi$ is reached. The choice of $\varphi$ determines the number of bases involved in ICA estimation, i.e. the number of ICA terms used. Those bases define an input subspace, which is a maximally informative subspace.

The process of grouping of the bases: $\mathbf{z}_i$, $\mathbf{T}_i$, or $\mathbf{TFD}_i$ is in fact a process of clustering, i.e. grouping of elements into so called *clusters* (Jain *et al.*, 1999; Jain, Dubes, 1988; McQueen, 1967). A result of clustering depends on many factors, of which the measure of distance used in this process and the clustering algorithm are the most important. A distance may be defined in a number of ways. The choice of a particular type of distance depends on a number of factors such as: frequency composition of the constituent signals in the mix, the amount of overlapping of constituent signals both in time and in frequency, precision of separation required and spectral similarity of components. Casey (2001) used one measure of distance ("Kullback–Leibler", see Subsec. 3.3) and in the clustering stage he used stochastic annealing.

The entire procedure of ICA-based single channel separation can be divided into four stages. Each one can be modified and adapted to a particular application independently of the others.

1. The generation of input data – a t-f representation of the mix: $\mathbf{TFD}^{\text{mix}}_{[m\times n]}$.
2. The estimation of independent components. Implemented by ICA or Non-negative Matrix Factorization (NMF).
3. The grouping (clustering) of bases.
4. The reverse transform of separated t-f signals $\mathbf{TFD}^{i}_{[m\times n]}$ back to the time signal.

In Stage 1 we used the STFT, and implemented it with Matlab "specgram" function. In Stage 2 we applied the "FastICA" Matlab function based on (Hyvarinen *et al.*, 2001), to the t-f representation from Stage 1. Our original contribution is concentrated in Stage 3 where we performed grouping by the application of specifically derived distance measures between particular elements ($\mathbf{z}_i$, $\mathbf{T}_i$, $\mathbf{TFD}_i$, $\mathbf{TFD}^i$). All algorithms were written in Matlab.

## 3. Definitions of distances used

The distances between spectral, time, and time-frequency vectors were computed according to a number of measures of distance often used in related literature. Cost functions estimating statistical independence of the components of the mixed signal were also computed. The computations were performed for example signals (see Sec. 4). This Section presents the measures of distance that can possibly be used for the clustering of basis components. Section 4 presents the results for those measures of distance and cost functions which provided best separation of components.

### 3.1. Standard Euclidean distance between $\mathbf{z}_i$ vectors:

$$\mathbf{D}_{i,j}^2 = \left\| \widehat{\mathbf{Z}}_i - \widehat{\mathbf{Z}}_j \right\|^2 = \sum_{k=1}^{n} (\mathbf{z}_{ik} - \mathbf{z}_{jk})^2, \tag{6}$$

where || . || symbol denotes the Euclidean metrics. The Euclidean distance can be used also for $\mathbf{T}_i$ vectors (columns of matrix $\mathbf{A}$).

### 3.2. Euclidean distance for t-f components (Paatero, Tapper, 1997), referred to as $\mathbf{TFD}_i$ in Sec. 1

This distance has been defined as:

$$\mathbf{D}_{TFD}^2(i,j) = \|\mathbf{TFD}_i - \mathbf{TFD}_j\|^2 . \tag{7}$$

As in (6) the || . || symbol denotes the Euclidean metrics.

### 3.3. The Kullback-Leibler distance

Defined as a symmetrised kind of the Kullback-Leibler's (K-L) divergence between two probability distributions (Hyvarinen *et al.*, 2001) and is given by (8):

$$KL_s(p,q) = \frac{1}{2} \int p(\widehat{u}) \log \left( \frac{p(\widehat{u})}{q(\widehat{u})} \right) \mathrm{d}\widehat{u} + \frac{1}{2} \int q(\widehat{u}) \log \left( \frac{q(\widehat{u})}{p(\widehat{u})} \right) \mathrm{d}\widehat{u}. \tag{8}$$

By denoting the probability distributions of variables $\mathbf{z}_i$ and $\mathbf{z}_j$ as $P_{z_i}(\widehat{u})$ and $P_{z_j}(\widehat{u})$ in our case we obtain (Casey, 2001):

$$\mathbf{D}_{KL}(i,j) = \mathbf{KL}_s(\mathbf{P}_{z_i}(\widehat{u}), \mathbf{P}_{z_j}(\widehat{u})). \tag{9}$$

### 3.4. Maximisation of negentropy of $\mathbf{TFD}^i$ components

Any two recordings of different acoustic sources are statistically independent, therefore their negentropy is maximum (Hyvarinen *et al.*, 2001; Cover, Thomas, 1991; Papoulis, 1991).

We need to find such $\mathbf{TFD}^i$ components for which their sum of negetropies is at the maximum (Mika, 2009):

$$(\mathbf{TFD}^i) \Rightarrow \max \sum_i J(\mathbf{TFD}_{\mathrm{rec}}^i). \tag{10}$$

The following approximation of negentropy was used (Hyvarinen *et al.*, 2001; Papoulis, 1991):

$$J(y) \sim [E\{G(y)\} - E\{G(v)\}]^2. \tag{11}$$

### 3.5. Clustering based on a $\beta$ distance of Gaussian distribution

The definition of generalised Gauss distribution is given by (Box, Tiao, 1973):

$$p(y|\mu,\sigma,\beta) = \frac{\omega(\beta)}{\sigma}\exp\left[-c(\beta)\left|\frac{y-\mu}{\sigma}\right|^{2/(1+\beta)}\right]. \tag{12}$$

The $\beta$ parameter describes a type of a random variable $y$. In the case when spectrograms of original constituent components of the mix are known we can find the $\beta_i^{\text{org}}$, i.e. the parameter of the distribution of a random variable with realisations given by those spectrograms. Then it is possible to minimise the distance between $\beta_i^{\text{org}}$ and the parameter $\beta$ characterising the $\mathbf{TFD}_{\text{rec}}^i$ distribution and hence we can determine the correct contents of a $\mathbf{TFD}^i$ spectrogram, so that it is as close as possible in the statistic sense to the original spectrogram of the $i$-th constituent component of the mix (Mika, 2009):

$$D_\beta = \left|\beta_i^{\text{org}} - \beta_i\left(\sum_a \mathbf{TFD}_a\right)\right|, \tag{13}$$

$$\mathbf{TFD}^i \rightarrow \min D_\beta. \tag{14}$$

The estimation of $\beta_i$ is performed by finding the maximum of $\beta$ parameter *a posteriori*. The *a posteriori* distribution of $\beta$ parameter when observations of $\mathbf{TFD}_{\text{rec}}^i = \{x_1, x_2, \ldots, x_N\}$ variable are available was given by (Lee, Lewicki, 2000):

$$p(\beta|x) \propto p(x|\beta)p(\beta), \tag{15}$$

where $p(x|\beta)$ is a data likelihood (Lee, Lewicki, 2000).

### 3.6. Distances based on time bases

1. Euclidean distance for time bases $\mathbf{T}_i$:
$$D_{ij}^{\text{Euclidean}^2} = \|\mathbf{T}_i - \mathbf{T}_j\|^2. \tag{16}$$
2. "Cityblock" distance (Saber, 1984):
$$D_{ij}^{\text{city}} = \|\mathbf{T}_i - \mathbf{T}_j\| = \sum_k |T_{ki} - T_{kj}|, \tag{17}$$
where $\|\ .\ \|$ denotes a Minkowski metrics ($L_1$) and $T_{ki}$ is $k$-th element of $i$-th time basis $\mathbf{T}_i$.
3. "Cosine" distance:
Defined as one minus the cosine of the angle between the $\mathbf{T}_i$ and $\mathbf{T}_j$ vectors (Saber, 1984):
$$D_{ij}^{\text{cosine}} = 1 - \cos\left(\angle(\mathbf{T}_i, \mathbf{T}_j)\right). \tag{18}$$

4. "Correlation" distance
   Defined as one minus the correlation coefficient between the $\mathbf{T}_i$ and $\mathbf{T}_j$ vectors (Saber, 1984):

$$D_{ij}^{\text{correlation}} = 1 - \mathbf{T}_i^T \mathbf{T}_j, \tag{19}$$

   where $\mathbf{T}_i$ and $\mathbf{T}_j$ – are used as column vectors.

## 4. Results of processing and discussion

The experiments consisted in demixing one-channel mixes of two and three signals. The signals were chosen so that both their respective types of sources and their spectral composition were different. $S_1(t)$ signal was a recording of an electric ringer and $S_2(t)$ signal was a baby cry. $S_3(t)$ component is a sound generated by a percussive instrument – "tom", i.e. it is a typical impulsive signal. A three-component mixed signal was generated by adding the percussive "tom" signal $S_3(t)$ to the two-component mix.

Both signals were 1.2 seconds long and recorded at the sampling frequency $F_s = 8$ kHz. The spectrograms shown below present the frequency range from 0 through 4000 Hz. The signals were analysed with the STFT, using blocks 256 samples long, 50% overlapped, and the Kaiser time window. The first 3968 and next 5888 samples (two separate blocks) were used in t-f analysis of each signal because of better stationarity of the respective spectra. Full signals of 9856 individual samples were used in the computation of $\beta$ distance of the Gaussian distribution.

Figure 1 presents the spectrograms of $S_1(t)$ and $S_2(t)$ and the spectrogram of their sum: $S_{\text{mix}}(t) = S_1(t) + S_2(t)$. In all subsequent figures the graph on the left corresponds to the "ringer" sound and on the right to the "baby" sound.

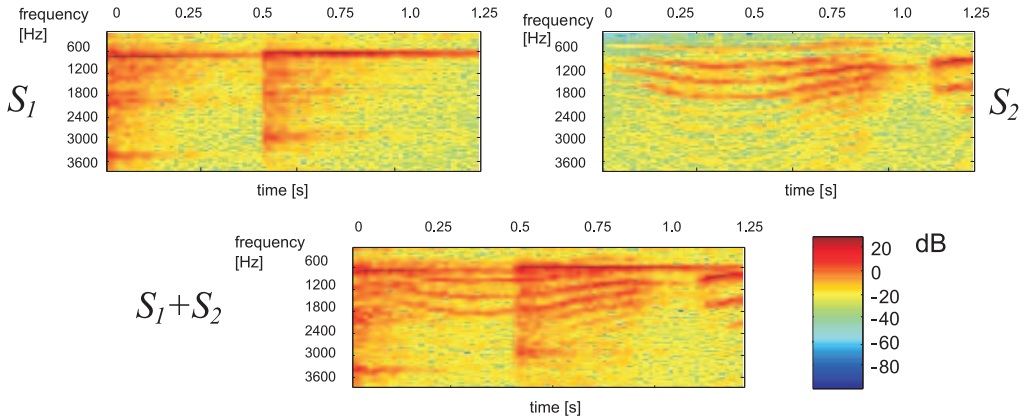


Fig. 1. Spectrograms of constituent signals: $S_1$ – "ringer", $S_2$ – "baby" (upper diagrams), and the spectrogram of the mixed signal $S_1 + S_2$ (lower diagram).

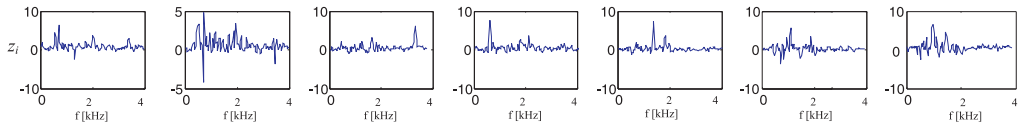


Fig. 2. The first 7 statistically independent spectral bases $\mathbf{z}_i$ obtained as a result of an ICA estimation of the spectrogram of the mixed signal for signal variance $\varphi = 85\%$.

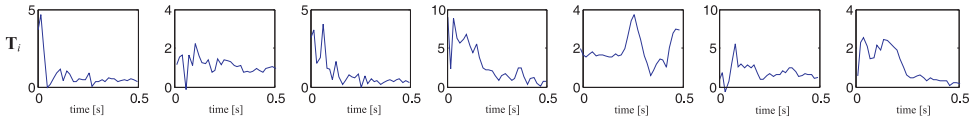


Fig. 3. The first 7 time bases $\mathbf{T}_i$ obtained for the spectrogram of the mixed signal at signal variance $\varphi = 85\%$.

The t-f spectrum of the two-component mixed signal $S_{\text{mix}}(t)$ was analysed by ICA procedure. The statistically independent components – $\mathbf{z}_i$ spectral bases were obtained. Figure 2 presents the bases with signal variance $\varphi$ of 85%, and Fig. 3 presents the corresponding time bases $\mathbf{T}_i$. In this figure, and in some of the next figures, the results of ICA performed on the first block of samples (from 0 through 0.51 s) are shown.

In stage 3 of the analysis we used hierarchical clustering (Jain, Dubes, 1988) and the $k$-mean partitional clustering (McQueen, 1967). We used the Euclidean distance as a measure of t-f components $\mathbf{TFD}_i$. The results of this clustering are depicted in Figs. 4a and 4b.

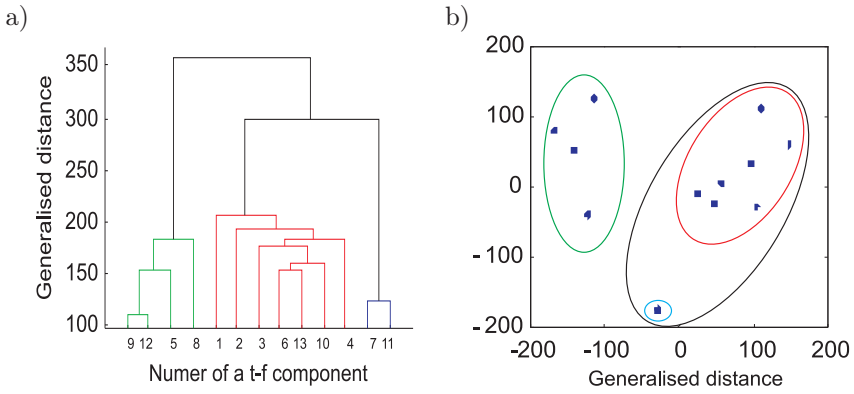


Fig. 4. a) A dendrogram obtained for the Euclidean distance for t-f components $\mathbf{TFD}_i$, b) Visualisation of groups of distances between $\mathbf{TFD}_i$ components obtained by multidimensional scaling (Saber, 1984). Coloured ellipses correspond to components grouped in the dendrogram shown in a).

By summing the $\mathbf{TFD}_i$ t-f components grouped in Fig. 4b in green and black ellipses (the latter corresponds to red and blue colours in a dendrogram of Fig. 4a) we obtain the spectrograms of the first and second components of the mixed signal:

$$\mathbf{TFD}^1 = \sum_{\substack{1,2,3,4,6,\\7,10,11,13}} \mathbf{TFD}_i, \tag{20}$$

$$\mathbf{TFD}^2 = \sum_{5,8,9,12} \mathbf{TFD}_i. \tag{21}$$

T-f signal components $\mathbf{TFD}_i = \mathbf{T}_i\mathbf{z}_i$ are presented in Fig. 5.

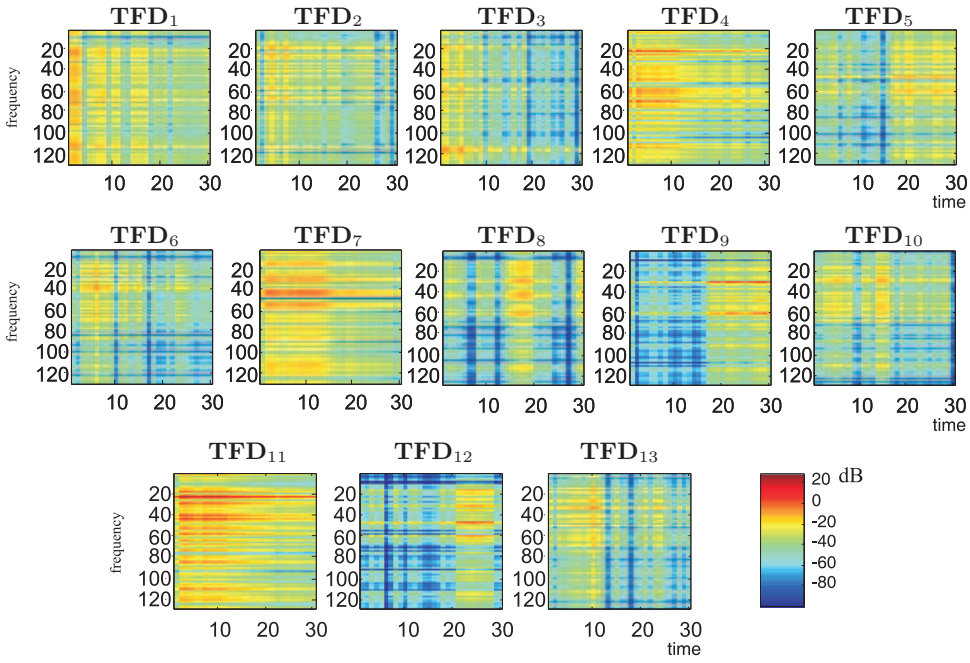


Fig. 5. The first 13 t-f components $\mathbf{TFD}_i$ of the spectrogram $\mathbf{TFD}^{\mathrm{mag}}_{\mathrm{mix}}$ for signal $\varphi = 85\%$. The sum of the presented signals is the t-f amplitude spectrum of the analysed mixed one-channel fragment $\mathbf{TFD}^{\mathrm{mag}}_{\mathrm{mix}}$. The scales for all $\mathbf{TFD}_i$ range from 0 through 129, which corresponds to the frequency range from 0 through 4 kHz. The range 0–30 in the time scale corresponds to the range 0–0.51 s. When Fig. 5 is compared to Fig. 1 it can be clearly seen that components no. 4, 7 and 11 belong to signal $S_1$ i.e. the ringer.

The reconstructed components of the mix $\mathbf{TFD}^1$ and $\mathbf{TFD}^2$ are shown in Fig. 6.

Figure 7 presents the result of separation with the use of an algorithm maximising the sum of negentropies of $\mathbf{TFD}^i$ components.

Below, the results of clustering with the use of $\beta$ distance of the Gaussian distribution are shown. As can be seen in the picture this technique seems to be efficient but the results depend on the length of a signal analysed and on its variance ($\varphi$ parameter), and consequently, on the number of $\mathbf{TFD}_i$ components. The lower this number the better are the results of clustering. However, reducing

the number of $\mathbf{TFD}_i$ components involves deterioration of reconstruction quality of the spectrograms of components.

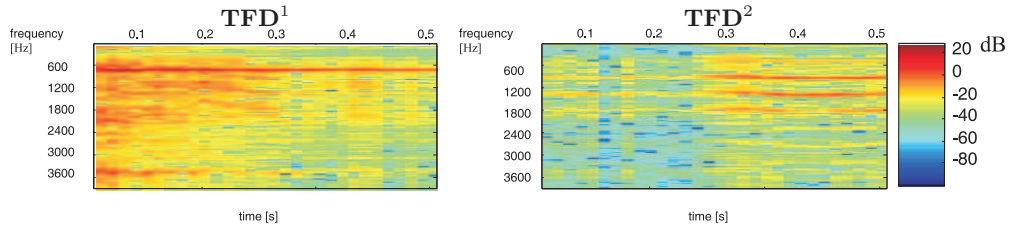


Fig. 6. Reconstructed components of the mix obtained by hierarchical clustering with the use of Euclidean distance for t-f components $\mathbf{TFD}_i$. $\mathbf{TFD}^1$ – the part of the spectrogram of $S_1$ ("ringer") corresponding to the first 0.51 s of this signal, $\mathbf{TFD}^2$ – the spectrogram of "baby". The harmonic structure corresponding to the original spectra of Fig. 1 can be clearly seen.

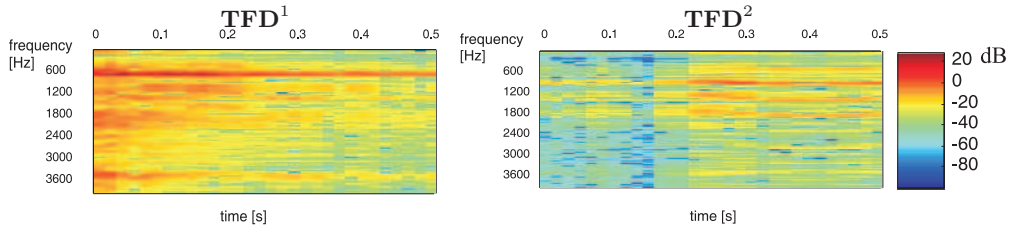


Fig. 7. $\mathbf{TFD}^i$ components obtained with an algorithm maximising the sum of their negentropies. The variance $\varphi = 90\%$ of the mixed signal was used. $\mathbf{TFD}^1$ – the spectrogram of "ringer", $\mathbf{TFD}^2$ – the spectrogram of "baby" (0.51 s of the mixed signal of Fig. 1 was analysed).

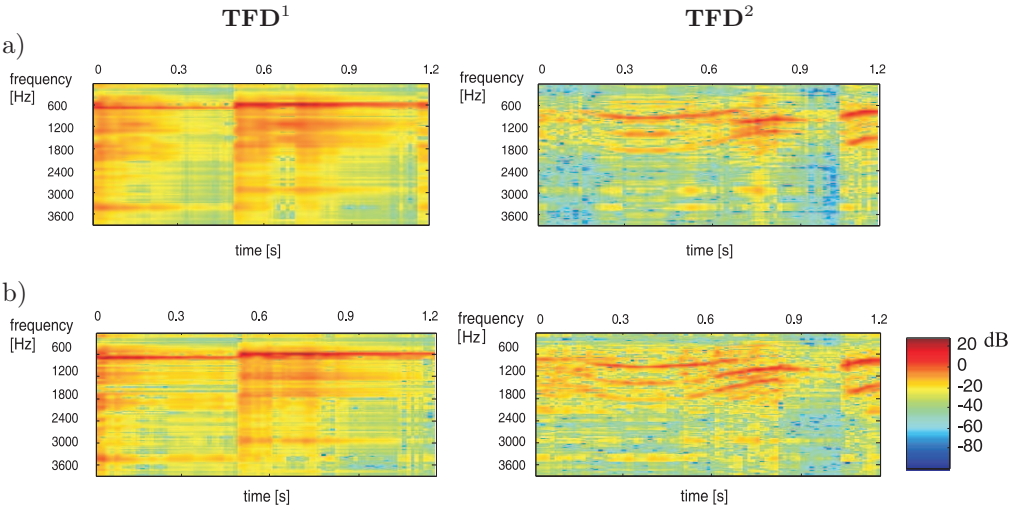


Fig. 8. Spectrograms of separated components of the mix with the use of $\beta$ distance of Gaussian distribution. The results are obtained respectively for a) $\varphi = 70\%$ and b) $\varphi = 80\%$ of variance and the signal duration of 1.2 s. $\mathbf{TFD}^1$ – the spectrogram of "ringer", $\mathbf{TFD}^2$ – the spectrogram of "baby". The similarity to original spectrograms of Fig. 1 can be clearly seen. The quality of separation is noticeably worse for $\varphi = 70\%$ which manifests in mutual penetration of spectra in a), especially for "baby" (diagrams on the right).

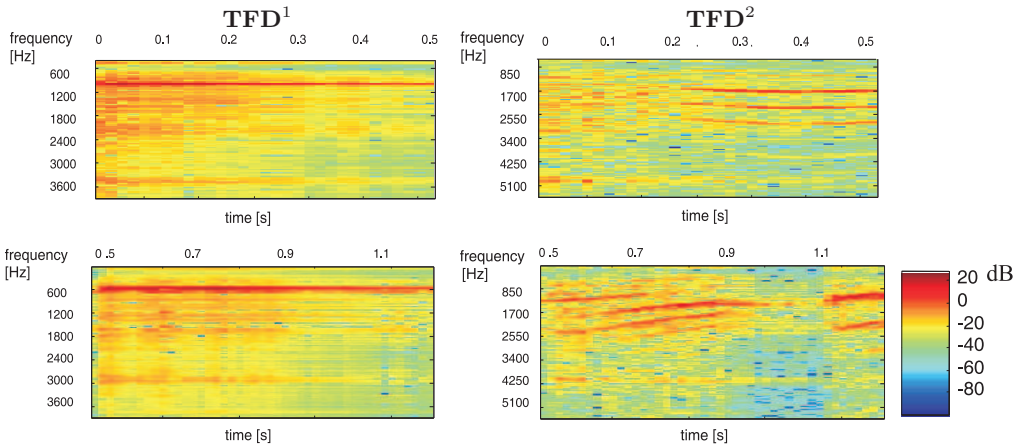


Fig. 9. The results of separation obtained by hierarchical clustering with Euclidean distance for time bases $\mathbf{T}_i$ and variance $\varphi = 90\%$. $\mathbf{TFD}^1$ – the spectrogram of "ringer", $\mathbf{TFD}^2$ – the spectrogram of "baby". The analysis was performed separately for two time frames corresponding to 0–0.51 s and 0.51–1.2 s of the analysed mixed signal of Fig. 1.

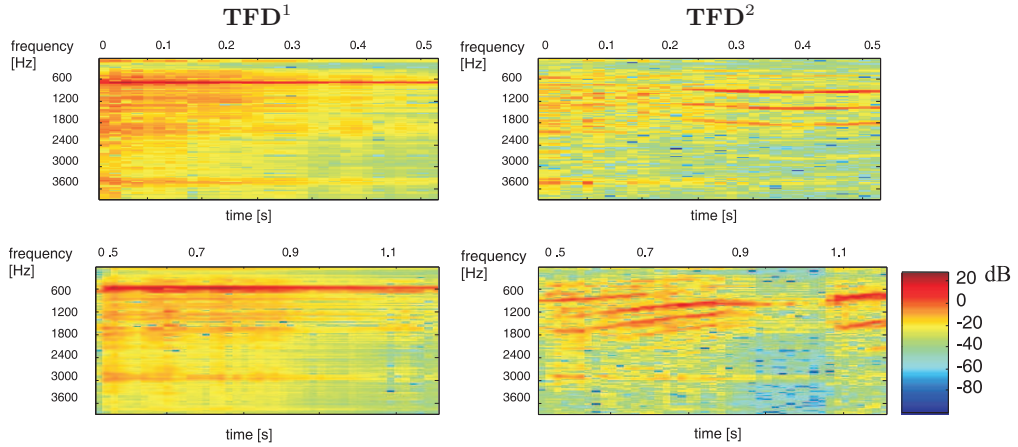


Fig. 10. The results of separation obtained with $k$-means clustering and "cityblock" distance (as defined in Subsec. 3.6) for time bases $\mathbf{T}_i$ and variance $\varphi = 90\%$. $\mathbf{TFD}^1$ – the spectrogram of "ringer", $\mathbf{TFD}^2$ – the spectrogram of "baby". The result of separation is identical to that presented in Fig. 9. This is often the case when the number of basis components is low and the measures of distance are based on the same components ($\mathbf{T}_i$ in both Fig. 9 and 10).

The results obtained with other types of distances based on time bases are shown below. The results of $k$-means clustering obtained with distance matrixes defined in par. 3.7.2, 3.7.3 and 3.7.4 will be presented.

Figure 12 presents an attempt to separate a three-component mix, obtained by summing signals $S_1$, $S_2$ and $S_3$ (see the beginning of this paragraph).

In this case stages 1 and 2 of the analysis were held exactly as for the two-component mix. After STFT and ICA estimation, in the 3rd stage the compo-

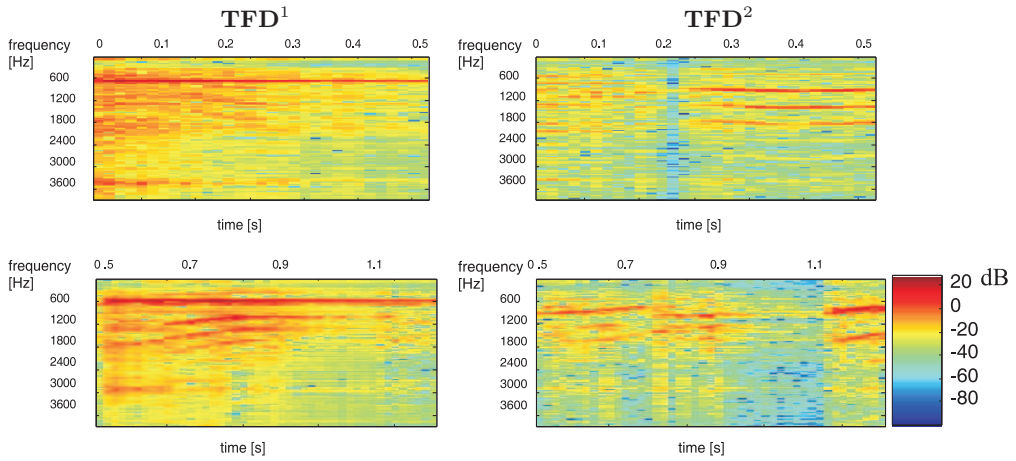


Fig. 11. The results of separation obtained with $k$-means clustering and "cosine" distance for time bases $\mathbf{T}_i$ and variance $\varphi = 90\%$. $\mathbf{TFD}^1$ – the spectrogram of "ringer", $\mathbf{TFD}^2$ – the spectrogram of "baby". In this case the separation is worse which can be seen in the mutual penetration of spectra.

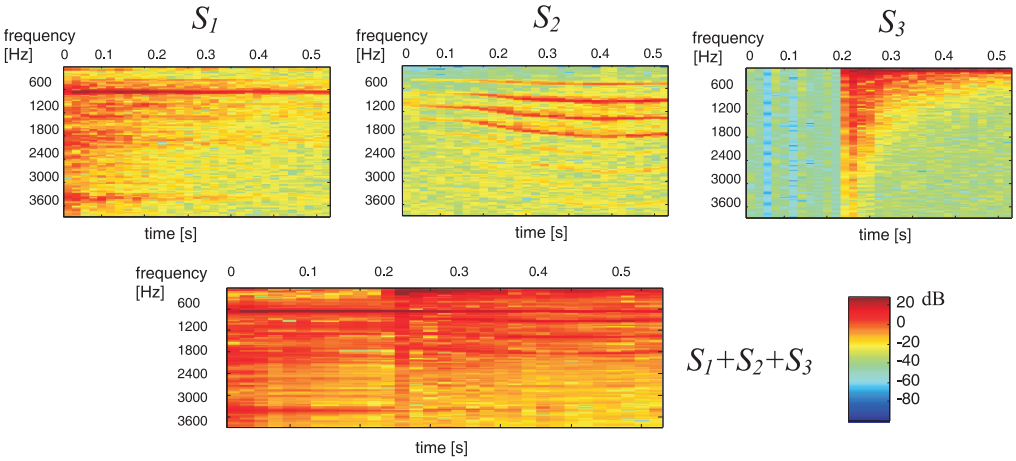


Fig. 12. Spectrograms of particular constituent signals and of the mixed signal. $S_1$ – "ringer", $S_2$ – "baby", $S_3$ – "tom". The lower diagram shows the spectrogram of the mixed signal.

nents were grouped into three classes. The results in Fig. 14 were obtained with the use of Euclidean distance for $\mathbf{TFD}_i$ bases presented in Fig. 13.

For each of the versions of decompositions presented above, the STFTs of the separated constituent signals have been converted back to the time signal. A specific algorithm written in Matlab was used. The STFT is not a perfectly invertible transform and in our implementation the inverse STFT was based on magnitude coefficients only, which resulted in the loss of phase information contained in the input signals.

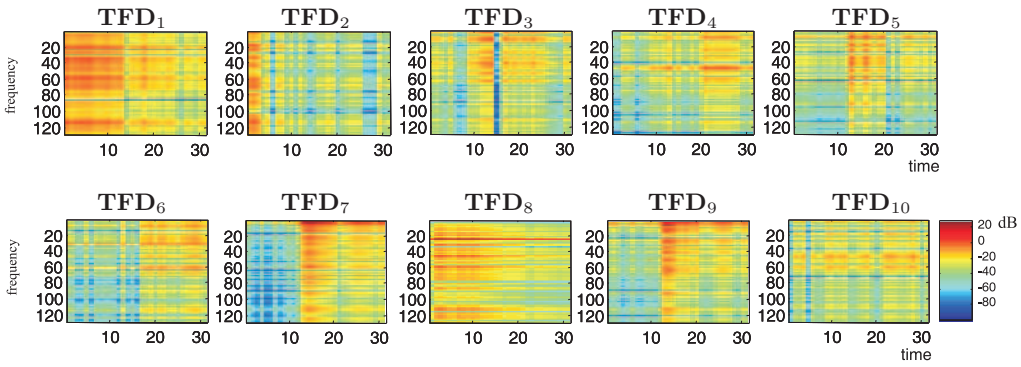


Fig. 13. $\mathbf{TFD}_i$ components, for 80% of signal variance. The scales for all $\mathbf{TFD}_i$ range from 0 through 129, which corresponds to the frequency range from 0 through 4 kHz. The range 0–30 in the time scale corresponds to the range 0–0.51 s. Similarities of $\mathbf{TFD}_i$ elements with the constituent sounds of the mixed signals can be clearly seen. For example, $\mathbf{TFD}_1$, $\mathbf{TFD}_2$ and $\mathbf{TFD}_8$ belong to "ringer", $\mathbf{TFD}_5$, $\mathbf{TFD}_7$ and $\mathbf{TFD}_9$ belong to "tom" and other components belong to "baby".

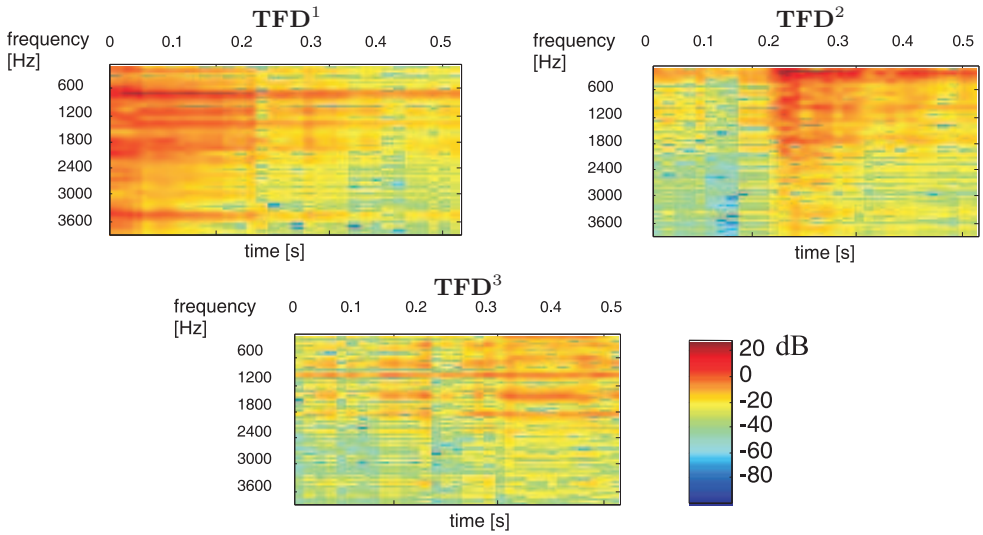


Fig. 14. The results of separation of a three-component signal obtained with hierarchical grouping and Euclidean distance for t-f bases $\mathbf{TFD}_i$. Duration 0.51 s – $\mathbf{TFD}^1$ – the spectrogram of "ringer", $\mathbf{TFD}^2$ – the spectrogram of "tom", $\mathbf{TFD}^3$ – the spectrogram of "baby". There is an obvious similarity to original spectrograms in Fig. 12.

## 5. Perceptual evaluation

### *5.1. Experiment*

The objective of the tests was to compare the original sound (a constituent of the mix) with its equivalent obtained from separation. All sounds were system sounds of Microsoft Windows and were resampled to 8 kHz. As separation was

performed with only magnitude spectrograms, all separated signals were synthesized with zero phase. The objective of the test was a mutual comparison of techniques, therefore the references for this test were also re-synthesized with zero phase, to eliminate the effect of phase distortion. The RMS values of all original and separated sounds were normalised.

Five mixes were evaluated:

1) "ringer" + „baby";
2) "notify" + "tada";
3) "ringer" + "tom";
4) "baby" + "tom";
5) "baby" + "ringer" + "tom".

Each of them was unmixed in a number of processing setups, with different bases, different measures of distance and in some cases different amounts of variance used. Many possible combinations of options were not included in the test, as they were found less efficient, on the basis of spectrograms and informal listening tests. Among these were: decorrelation and maximisation of negentropy of $\mathbf{TFD}^{j}_{\mathrm{rec}}$ components and most of the distances for spectral and $\mathbf{TFD}_{i}$ bases.

The total of 107 different pairs: original (reference) sound plus separated sound were prepared. These pairs will be referred to as "samples".

Five sets of samples were generated, each containing all 107 samples. In each set the sequence was random and different. The samples were separated by 3 to 4 s of silence and within each sample both intervals were separated by 500 ms. The reproduction of each set of samples lasted about 9 minutes. A listener took a short break and then evaluated the next set. This way, each participant listened to each sample for five times. Five listeners participated, including one audio engineer, two musicians and two subjects not active in music. Four of them were about 30 and one was 40.

Each listener heard samples at the same loudness (over 80 dB, this level was preferred by most listeners) over the closed headphones AKG K271 Studio, in a silenced studio room.

The listeners rated the quality of separation using the degradation category rating scale (Bech, Zacharov, 2006). The original 5-point scale was extended to 6 points, following suggestions from listeners. The lowest score of 1 denoted "very annoying difference" and the highest of 6 denoted "imperceptible difference". Before the actual test, each participant went through a short training session. Listeners wrote down their answers in a scoring sheet.

### *5.2. Results and discussion*

Tables 1 through 8 show mean values and standard deviations of evaluation scores for eight different processing setups, separately for each sound.

Table 9 presents the effect of the amount of variation of mixed signal used (70% or 90%) on perceptual quality of separations. Two sounds: "ringer" and "baby" were evaluated.

**Table 1**. Mean scores and standard deviations for each sound obtained with "Beta" distance. Common mean = 3.42, common $\sigma = 1.34$.

| "Baby" | "Ringer" | "Notify" | "Tada" | "Tom" |
|---|---|---|---|---|
| mean = 3.42 | mean = 4.45 | mean = 3.68 | mean = 3.84 | mean = 2.05 |
| $\sigma = 1.12$ | $\sigma = 0.99$ | $\sigma = 1.11$ | $\sigma = 1.07$ | $\sigma = 0.8$ |

**Table 2**. Mean scores and standard deviations for each sound obtained with "correlation" distance. Common mean = 3.14, common $\sigma = 1.34$.

| "Baby" | "Ringer" | "Notify" | "Tada" | "Tom" |
|---|---|---|---|---|
| mean = 2.86 | mean = 4.31 | mean = 2.2 | mean = 4.04 | mean = 2 |
| $\sigma = 1.11$ | $\sigma = 0.86$ | $\sigma = 1.15$ | $\sigma = 0.93$ | $\sigma = 0.79$ |

**Table 3**. Mean scores and standard deviations for each sound obtained with "cosine" distance. Common mean = 3.05, common $\sigma = 1.32$.

| "Baby" | "Ringer" | "Notify" | "Tada" | "Tom" |
|---|---|---|---|---|
| mean = 2.41 | mean = 4.33 | mean = 2.28 | mean = 4.2 | mean = 2.05 |
| $\sigma = 0.9$ | $\sigma = 0.79$ | $\sigma = 0.98$ | $\sigma = 1.04$ | $\sigma = 0.79$ |

**Table 4**. Mean scores and standard deviations for each sound obtained with "city" distance. Common mean = 3.03, common $\sigma = 1.25$.

| "Baby" | "Ringer" | "Notify" | "Tada" | "Tom" |
|---|---|---|---|---|
| mean = 2.67 | mean = 3.98 | mean = 2.68 | mean = 3.8 | mean = 2.12 |
| $\sigma = 0.99$ | $\sigma = 1.05$ | $\sigma = 1.14$ | $\sigma = 0.96$ | $\sigma = 0.91$ |

**Table 5**. Mean scores and standard deviations for each sound obtained with "Euclidean" distance with variance normalised to the value of 1. Common mean = 2.97, common $\sigma = 1.32$.

| "Baby" | "Ringer" | "Notify" | "Tada" | "Tom" |
|---|---|---|---|---|
| mean = 2.97 | mean = 3.88 | mean = 2.44 | mean = 3.84 | mean = 1.64 |
| $\sigma = 1.04$ | $\sigma = 1.15$ | $\sigma = 1$ | $\sigma = 0.94$ | $\sigma = 0.71$ |

**Table 6**. Mean scores and standard deviations for each sound obtained with "Euclidean" distance. Common mean = 2.86, common $\sigma = 1.26$.

| "Baby" | "Ringer" | "Notify" | "Tada" | "Tom" |
|---|---|---|---|---|
| mean = 2.73 | mean = 3.67 | mean = 2.88 | mean = 4.16 | mean = 1.52 |
| $\sigma = 0.89$ | $\sigma = 0.87$ | $\sigma = 1.01$ | $\sigma = 0.94$ | $\sigma = 0.7$ |

**Table 7**. Mean scores and standard deviations for each sound obtained with "K-L" distance. Common mean = 2.53, common $\sigma = 1.14$.

| "Baby" | "Ringer" | "Notify" | "Tada" | "Tom" |
|---|---|---|---|---|
| mean = 2.39 | mean = 3.41 | mean = 3.32 | mean = 1.52 | mean = 1.6 |
| $\sigma = 0.95$ | $\sigma = 0.82$ | $\sigma = 1.14$ | $\sigma = 0.59$ | $\sigma = 0.68$ |

**Table 8**. Mean scores and standard deviations for each sound obtained with "Euclidean" distance for TFD bases. Common mean = 2.40, common $\sigma = 1.13$.

| "Baby" | "Ringer" | "Notify" | "Tada" | "Tom" |
|---|---|---|---|---|
| mean = 2.44 | mean = 3.12 | mean = 3.44 | mean = 1.04 | mean = 1.53 |
| $\sigma = 0.9$ | $\sigma = 1.04$ | $\sigma = 0.92$ | $\sigma = 0.2$ | $\sigma = 0.66$ |

**Table 9**. The effect of the amount of variance of mixed signal used on perceptual quality of separation. Mean scores for the "ringer" + "baby" mix.

| Measure of distance | "ringer" | | "baby" | |
|---|---|---|---|---|
| | 90% | 70% | 90% | 70% |
| "Beta" | 5.00 | 4.36 | 4.32 | 3.04 |
| "city" | 4.28 | 4.20 | 2.84 | 1.96 |
| "correlation" | 4.40 | 4.32 | 3.92 | 2.60 |
| "cosine" | 4.40 | 4.08 | 2.36 | 2.36 |
| "Euclidean" | 3.64 | 3.68 | 2.44 | 2.72 |
| "Euclidean_var = 1" | 4.52 | 4.16 | 3.12 | 2.88 |
| "Kullback-Leibler" | 3.52 | 3.20 | 2.60 | 2.20 |
| "TFD Euclidan" | 3.44 | 4.00 | 2.44 | 2.64 |
| Mean score | 4.15 | 4 | 3.00 | 2.55 |

In Table 10 the mean scores of perceptual quality for "ringer" + "baby" sounds obtained from their mix are compared with the scores for the same sounds obtained from the three-component mix "ringer" + "baby" + "tom".

**Table 10**. The comparison of mean scores for "ringer" + "baby" sounds obtained from their mix and from the three-component mix "ringer" + "baby" + "tom" for different measures of distance.

| Measure of distance | "ringer" | | "baby" | |
|---|---|---|---|---|
| | Three-component mix | Two-component mix | Three-component mix | Two-component mix |
| "Beta" | 4.00 | 5.00 | 2.84 | 4.32 |
| "city" | 4.60 | 4.28 | 2.80 | 2.84 |
| "correlation" | 3.92 | 4.40 | 1.84 | 3.92 |
| "cosine" | 4.24 | 4.40 | 2.04 | 2.36 |
| "Euclidean" | 4.12 | 3.64 | 2.56 | 2.44 |
| "Euclidean_var =1" | 4.44 | 4.52 | 2.72 | 3.12 |
| "Kullback-Leibler" | 3.32 | 3.52 | 1.68 | 2.60 |
| "TFD Euclidean" | 2.28 | 3.44 | 1.64 | 2.44 |
| Mean score | 3.86 | 4.15 | 2.26 | 3.00 |

The best perceptual quality of separation was obtained for "Beta" measure of distance, and it was closely followed by "correlation", "city" and "cosine" dis-

tances. For each of these cases the mean score exceeded 3. The "ringer" sound was most efficiently unmixed in any of the distances used. For "Beta", "correlation" and "cosine" the mean score exceeded 4. Listeners also favourably evaluated the „tada" sound (the score over 4 for three of the distances). "Baby" was evaluated considerably lower, and "tom" turned out to be the most difficult to separate – the highest scores were around 2. This demonstrates that the algorithms evaluated proved to be most efficient for a sound with a quasi-stationary harmonic spectrum ("ringer") and least efficient for a highly non-stationary sound with noise-like spectrum ("tom").

The quality of separation was higher for higher amount of variation of the mixed signal used (Table 9). As could be expected, quality was higher in separation from two-component mixes (Table 10) but the scores on average were better by only 0.5 point, which is promising for demixing of larger numbers of sounds.

In previous works on single channel separation with ICA only spectral vectors were grouped and only with the K-L distance. We have proved, that this approach is often sub-optimal (see fairly poor results for this measure in Tables 9 and 10).

## 6. Conclusions

In most cases, the time bases $\mathbf{T}_i$ provided better separation than spectral bases used in earlier works.

The results demonstrate that grouping can be performed both by using distances that require some information on the constituent signals (*"Beta"* distance of Gaussian distribution) and by exploiting the bare similarity between the bases. The aim should be the grouping of bases without using any information about constituent signals. However, when this is the case, then the choice of distance depends on the analysed constituent signals $S_j(t)$. If the amplitude of signals varies in time considerably, then it is appropriate to use the Euclidean distance for time bases $\mathbf{T}_i$. For successful decomposition the constituent signals should have stationary spectra within the analysed period. It is possible to overcome this limitation by shortening the analysed period, but this manifests itself in the deterioration of audible quality of reconstructed demixed signals. The use of the K-L distance provides satisfactory results when the basis vectors $\mathbf{z}_i$ are similar in terms of their probability density distributions within any of the constituent signals (inter-signal similarity), while distinctly differing when different signals are compared (intra-signal dissimilarity). Again, for real signals, time bases belonging to one sound source often have completely different probability density distributions. The application of the K-L distance in such cases produces false results.

Satisfying results have been obtained when the Euclidean distance was applied to grouping of the t-f ($\mathbf{TFD}_i$) components. This type of distance measure is naturally suited for grouping both time and spectral features of signals.

It was demonstrated, that clustering analysis both in its hierarchical and $k$-means forms can be successfully applied to grouping of components for signal separation.

The main limitation of one-channel STFT based ICA signal decomposition is the lack of an all-purpose procedure. The choice of the type of basis components (spectral, time, or time-frequency), the measure of distance and a clustering algorithm should depend on time-frequency characteristics of constituent signals of the mix. Moreover, the result of clustering and quality of separation also depend on the choice of variance parameter $\varphi$. Too high a value of $\varphi$ makes clustering less efficient, as there are too many components to handle, while too low a value deteriorates the quality of separated components. The quality of separation also depends on inherent limitations of ICA. As the number of mixed signals increase, the degree of separation of components obtained from ICA decreases, leading to mutual penetration of spectra in separated constituent signals.

The procedure developed for this research is an open one. Each of the stages can be modified and optimized independently. For example, in t-f analysis stage other methods that STFT may be used, including bilinear transforms, and in the ICA stage its more advanced versions can be applied.

A comparison of computational load between the procedures evaluated was not performed. The processing tools developed in Matlab were strongly experiment-oriented and contained some extra procedures in order to keep better control of a process. With about 30 basis components the computation time on a 3.2 GHz PC computer was below 20 s, and computation of the distance matrix took most of the time. It may be expected that with the use of a compiler the processing time should approach real time.

Future works will be focused on the application of other t-f signal distributions, particularly those improving local t-f resolution like Wigner-Ville, and their effect on the quality of separation. In the ICA stage some more advanced methods like topographic ICA will be investigated. In order to make the clustering process automatic and more data independent the application of artificial intelligence algorithms should be evaluated. These methods could acquire information of the components of the mix at the learning stage and then efficiently cluster basis components, like in pattern recognition procedures. The current perceptual estimation can be extended by using pair comparison tests besides comparisons to the reference.

## Acknowledgment

The authors would like to thank prof. Bożena Kostek for her comments on D. Mika's doctoral dissertation, which were used as a valuable advice in work on this manuscript. The authors also appreciate helpful comments from two anonymous reviewers. P. Kleczkowski's work was supported by AGH University of Science and Technology grant no. 11.11.130.885.

## References

1. Bach F.R., Jordan M.I. (2005), *Blind one-microphone speech separation: A spectral learning approach*, Advances in neural information processing systems, **17**, 65–72.
2. Barry D., Fitzgerald D., Eugene Coyle E., Lawlor G. (2005), *Single Channel Source Separation using Short-time Independent Component Analysis*, 119th Audio Engineering Society Convention, Convention Paper 6603, New York.
3. Barry D., Lawlor B., Coyle E. (2004), *Sound source separation: azimuth discrimination and resynthesise*, Proc. of the 7th Int. Conference on Digital Audio Effects (DAFX-04), Naples, Italy.
4. Bech S., Zacharov N. (2006), *Perceptual Audio Evaluation,* John Wiley & Sons, Inc., Chichester, England.
5. Box G., Tiao G. (1973), *Bayesian Inference In Statistical Analysis*, John Wiley & Sons, Inc., England.
6. Brungart D.S., Chang P.S., Simpson B.D., Wang D.L. (2006), *Isolating the energetic component of speech-on-speech masking with ideal t-f segregation*, Journal Acoustical Society of America, **120**, 4007–4018.
7. Cardoso J.-F. (1998), *Blind Signal Separation: statistical principles*, Proceedings of the IEEE, **9**, 10, 2009–2025.
8. Casey M.A. (2001), *Separation of Mixed Audio Sources by Independent Subspace Analysis*, Merl – A Mitsubishi Electric Research Laboratory, TR-2001-31.
9. Cooney R., Cahill N., Lawlor R. (2006), *An Enhanced implementation of the ADRess (Azimuth Discrimination and Resynthesis) Music Source Separation Algorithm*, Audio Engineering Society, Convention Paper 6984, 121st Convention, San Francisco, USA.
10. Cover T.M., Thomas J.A. (1991), *Elements of Information Theory*, John Wiley & Sons, Inc., New York.
11. Davies M.E., James C.J. (2007), *Source separation using single channel ICA*, Signal Process., **87**, 8, 1819–1832.
12. Duan Z., Zhang Y., Zhang C., Shi Z. (2008), *Unsupervised Single-Channel Music Source Separation by Average Harmonic Structure Modeling*, IEEE Transactions on Audio, Speech and Language Processing, **16**, 4, 766–778.
13. Dziubinski M., Kostek B. (2010), *Evaluation of the separation algorithm performance employing ANNs*, Chapter in Advances in Soft Computing, **80**, pp. 27–37, Springer Verlag, Berlin, Heidelberg.
14. Hyvarinen A., Karhunen J., Oja E. (2001), *Independent Component Analysis*, A Wiley-Interscience Publication, John Wiley & Sons, Inc. New York.
15. Jain A.K., Dubes R.C. (1988), *Algorithms for Clustering Data,* Prentice-Hall advanced reference series, Prentice-Hall, Inc., Upper Saddle Riever, NJ, USA.
16. Jain A.K., Murty M.N., Flyn P.J. (1999), *Data Clustering: A Review*, ACM Computing Survey, **31**, 3.
17. Jang G.-J., Lee T.-W. (2002), *Learning statistically efficient features for speaker recognition*, Neurocomputing, **49**, 1–4, 329–348.

18. Jang G.-J., Lee T.-W. (2003), *A Maximum Likelihood Approach to Single-Channel Source Separation*, Journal of Machine Learning Research, **4**, 1365–1392.

19. Kostek B. (2005), *Perception-Based Data Processing in Acoustics*, Springer Verlag, Berlin.

20. Lee T.-W., Lewicki M.S. (2000), *The generalized Gaussian Mixture Model using ICA*, International workshop on ICA, 239–244.

21. Litvin, Y., Cohen I. (2009), *Single-channel source separation of audio signals using Bark Scale Wavelet Packet Decomposition*, IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2009, 1–4.

22. Masters A.S. (2006), *Stereo music source separation via Bayesian modeling*, Ph.D. dissertation, Stanford University, USA.

23. McQueen J. (1967), *Some methods for classification and analysis of multivariate observations*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297.

24. Mijovic' B., De Vos M., Gligorijevic' I., Taelman J., Van Huffel S. (2010), *Source Separation From Single-Channel Recordings by Combining Empirical-Mode Decomposition and Independent Component Analysis*, IEEE Transactions on Biomedical Engineering, **57**, 9, 2188–2196.

25. Mika D. (2009), *Separation of sounds from various sources in a mixed acoustic signal* [in Polish], Ph.D. Thesis, AGH University, Kraków, Poland.

26. Paatero P., Tapper U. (1997), *Least squares formulation of robust non-negative factor analysis*, Chemometr. Intell. Lab., **37**, 1, 23–35.

27. Papoulis A. (1991), *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 3rd edition, New York.

28. Rickard S., Yilmaz O. (2002), *On the approximate W-disjoint orthogonality of speech*, [in:] ICASSP, Orlando, Florida, 529–531.

29. Seber G.A.F. (1984), *Multivariate Observations*, John Wiley & Sons, Inc., New York.

30. Taghia J., Ali Doostari M. (2009), *Subband-based Single-channel Source Separation of Instantaneous Audio Mixtures World*, World Applied Sciences Journal, **6**, 784–792.

31. Vinyes M., Bonada J., Loscos A. (2006), *Demixing Commercial Music Productions via Human-Assisted T-f Masking*, Audio Engineering Society, Convention Paper 6719, 120th Convention, Paris, France.

32. Wang D.L., Brown G.J. (2006), *Computational auditory scene analysis, Principles, Algorithms, and Applications*. IEEE Press/Wiley-Interscience, Hoboken NJ.

33. Wang B., Plumbley M. (2006), *Investigating single-channel audio source separation methods based on non-negative matrix factorization*, ICA Research Network International Workshop.

34. Yilmaz O., Rickard S. (2004), *Blind separation of speech mixtures via t-f masking*, IEEE Transactions on Signal Processing, **52**, 7, 1830–1847.