Audio Engineering Society
Convention Paper
Presented at the 112th Convention
2002 May 10–13 Munich, Germany

Intelligent Audio Source Separation using Independent Component Analysis

Nikolaos Mitianoudis and Mike Davies

DSP Lab, Queen Mary College, University of London, Mile End Road, London E1 4NS, UK

Correspondence should be addressed to Nikolaos Mitianoudis (`nikolaos@elec.qmul.ac.uk`)

ABSTRACT

The authors introduce the idea of performing *Intelligent ICA* to focus on and separate a specific instrument, voice or sound source of interest. This is achieved by incorporating high-level probabilistic priors in the ICA model that characterise each instrument or voice. For instrument modelling, we experimented with various feature sets previously used for instrument or speaker recognition. Prior training of a Gaussian Mixture Model for each instrument was performed. The order of the feature vector, the number of gaussian mixtures and the training audio data length were kept to reasonably minimum levels.

INTRODUCTION

Audio source separation deals with the problem of isolating the audio sources that are present in an auditory scene. In order to capture the auditory scene, we place a number of microphones in different spots and record their observations. Using these recordings, we try to separate the audio objects of this auditory scene.

One way to perform audio source separation is using *Independent Component Analysis* (ICA). ICA is a newly developed technique that exploits the statistical properties of audio signals to perform separation. ICA methods perform fast and efficient separation in the case that we have equal number of sources and microphones in the auditory scene and the observation signals are modelled as instantaneous mixtures of the audio sources [1, 2, 4]. Modern ICA approaches can perform fast separation in the case of mixtures recorded in real environments [5]. ICA methods have been introduced in the *"less sensors than sources"* case without much success. One basic drawback of these approaches is that they try to separate *all* the sources that are present in the auditory scene. On the other hand, we may want to separate just a particular

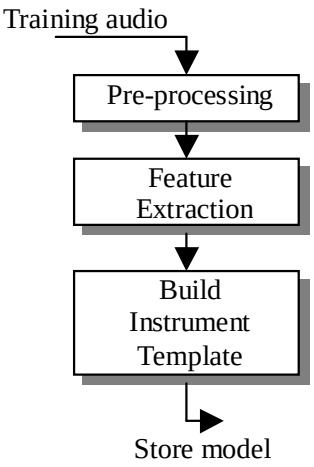

Fig. 1: A general front-end for instrument recognition model training.

source of interest that is present in the scene but not all of them, for example the guitar or a specific person speaking. In order to achieve this, we have to build statistical models that will be able to characterise a specific instrument or person and use these models to "steer" the ICA method towards the desired source. In this paper, the authors try to use the existing modelling solutions mainly employed in instrument recognition and speaker verification to model the sources efficiently and incorporate these probabilistic priors in the ICA framework, so as to perform Intelligent ICA.

INSTRUMENT RECOGNITION

Automatic musical instrument recognition is a very interesting problem, which can have many applications. It can be a useful tool in all *Musical Information Retrieval* (MIR) procedures (music indexing, music summarisation etc) and of course in automatic music transcription. Instrument recognition is also similar to *speaker recognition* or *verification*, where a person can be identified from his voice.

An instrument/speaker recognition procedure is basically split into two phases: the *training* and the *recognition* phase. During the training phase, some audio samples from the instrument or person are used to retrieve and store some information about it in a model. During the recognition process, a smaller audio sample of the instrument is shown to the system. The system retrieves the same type of information from the sample, compares it with the information available in the database and makes an inference about the instrument or the person. A general front-end for the training procedure is depicted in figure 1. The training audio is passed through a pre-processing stage, which can have various parts. Usually, possible DC bias is removed from the signal and is amplitude normalised. Some silent parts may be removed as they may affect the performance of the recogniser. Finally, the signal is pre-emphasized, using a first-order high-pass filter, like the one in (1), increasing the relative energy of the high-frequency spectrum.

$$H(z) = 1 - 0.97z^{-1} \tag{1}$$

Then, we extract several features from the signal that will be used to identify the instrument. Usually, the signal is segmented into overlapping, windowed frames and for each of these we calculate a set of parameters that constitute the *feature vector*. The performance of the recogniser is mainly determined by the feature set used in this step. Many feature sets, capable of capturing the aural characteristics of an instrument, were proposed for instrument recognition [7, 8]. A whole family of feature sets capture frequency envelopes: the *Linear Predictive coefficients*, the *Warped Linear Predictive coefficients*, the *Mel-Frequency Cepstral coefficients* (MFCC) (delta and delta-delta) and the *Perceptual Linear Predictive coefficients* can capture signal envelopes in the frequency, warped log-frequency, mel-frequency and bark-frequency domain respectively. Other features can be the *spectral centroid*, the *crest factor*, the *fundamental frequency* etc. The importance of all these features is analytically discussed in [7, 8, 9]. In our further analysis, we will use the MFCCs as they featured good performance in a study presented in [7] and also generally in speaker verification [6].

Finally, the feature vectors are used to build a model or reference template for the instrument/person. There are many techniques that can be used in this part: a *vector quantiser*, a *neural network classifier*, a *Hidden Markov Model* or a *Gaussian Mixture Model* (GMM). In the following analysis, we are going to use a GMM as the recogniser [6, 10]. A GMM describes the probability density function of the instrument's feature vectors as a weighted sum of Gaussian distributions. If $\underline{v}$ is a feature vector, then the probability model built by a GMM is given by the following equations:

$$P(\underline{v}|\lambda) = \sum_{i=1}^{M} p_i b_i(\underline{v}) \tag{2}$$

$$b_i(\underline{v}) = \frac{exp(-0.5(\underline{v} - \underline{m}_i)^T C_i^{-1}(\underline{v} - \underline{m}_i))}{\sqrt{(2\pi)^D |C_i|}} \tag{3}$$

where $p_i, \underline{m}_i, C_i$ are the weight, the mean vector and the covariance matrix of each gaussian and M is the number of gausssians used. A GMM model is usually described using the notation $\lambda = \{p_i, \underline{m}_i, C_i\}$, for $i = 1, .., M$.

In our analysis, we will assume that each Gaussian has its own covariance (*nodal covariance*), but each covariance matrix is diagonal. This model is usually trained using the *Expectation Maximisation* (EM) algorithm, as described analytically in [10].

A general front-end for performing instrument recognition is shown in figure 2. Basically, the preprocessing and the feature extraction stages are identical to the ones used during training. Suppose we have S instrument models λ_k stored and a set of feature vectors for the instrument to be identified $V = \{\underline{v}_1, \underline{v}_2, ..., \underline{v}_T\}$. The correct model should maximise the following probability:

$$\max_{1 \le k \le S} P(\lambda_k|V) = \max_{1 \le k \le S} P(V|\lambda_k) \tag{4}$$

In other words, the model maximising equation (2), given the data V, gives us the identity of the instrument.

In our experiments, we tested several feature vectors - number of gaussians configurations. The aim was to build a fairly

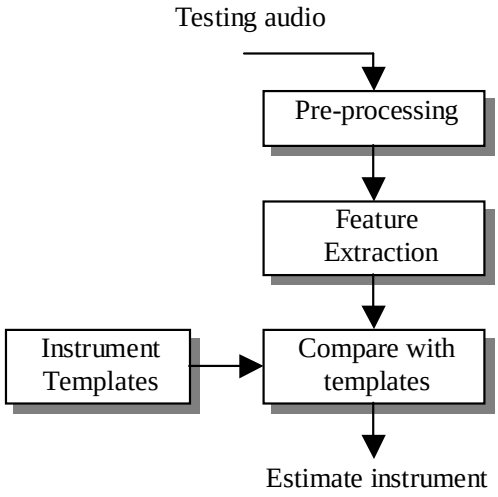

Fig. 2: A general front-end for performing instrument recognition.

simple, fast and robust system. As a result, we ended up using a combination of 18 MFCCs and 16 Gaussian Mixtures. The MFCCs performed reasonably well in our study, as well as in [7].

INDEPENDENT COMPONENT ANALYSIS

Independent Component Analysis (ICA) is a technique used to perform audio source separation. ICA usually exploits the *nongaussianity* of source signals and assumes *statistical independence* of separated signals to perform separation. Many ways were proposed to perform separation in this general ICA framework. However, we will only examine the Bayesian and the non-gaussianity approach in this study. Suppose there are N audio sources $\underline{s} = [s_1[n]s_2[n]...s_N[n]]^T$ in a room and N microphones capturing the auditory scene, by recording the observation signals (mixtures) $\underline{x} = [x_1[n]x_2[n]...x_N[n]]^T$. Ignoring the room's acoustics, the mixed signals can be modelled as summed portions of the original sources, i.e. instantaneous mixtures of the sources. In this study, we will use this rather simplified model, which can be valid in the case of signals mixed using a mixing desk. The case of possible additive noise won't be considered. If A is the mixing matrix, then:

$$\underline{x} = A\underline{s} \tag{5}$$

The solution to this problem is given by estimating $W \approx A^{-1}$, in order to unmix the sources.

$$\underline{u} = W\underline{x} \tag{6}$$

The *Bayesian approach* solves the problem by forming a Maximum Likelihood or a maximum a posteriori (MAP) estimate of W, assuming a supergaussian probabilistic model for the sources. Amari et al [4] proposed the *natural gradient algorithm* that provides a stable solution to the problem with good separation quality. However, the convergence is relatively slow.

The *nonguassianity* approach estimates the direction of the most nongaussian component in the mixtures, optimising several measures of nongaussianity like kurtosis or negentropy. Hyvarinen [2, 3] proposed several Newton-type solutions to the problem (i.e. FastICA) that are faster and more stable than the natural gradient approach. He also proposed several one-unit versions, estimating the first component that maximises the nongaussianity criterion. For example, the one-unit learning rule that maximises the absolute value of kurtosis is the following [3]:

$$\underline{w}^+ \leftarrow E\{\underline{x}(\underline{w}^T\underline{x})^3\} - 3\underline{w} \tag{7}$$

$$\underline{w}^+ \leftarrow \underline{w}^+/||\underline{w}^+|| \tag{8}$$

where $\underline{x}$ are prewhitened observations of the auditory scene. The most nongaussian component u is calculated, as follows:

$$u = \underline{w}^T\underline{x} \tag{9}$$

INTELIGENT ICA

In this study, we explore the possibility of combining the efficient probabilistic modelling performed by instrument/speaker verification in the ICA of instantaneous mixtures framework. There are two ways to perform intelligent separation: one combining non-gaussianity measurements and probabilistic inference from the model and another estimating the direction that maximises the posterior probability of the instrument's model.

One should point out that the instrument recognition problem that we are called to solve is slightly different to the one usually tackled in the literature. Usually in instrument recognition, we have an audio sample from an instrument/person and we compare the information acquired from that with the templates in the database. In this case, the problem is quite the opposite. We know the identity of the instrument/person and we want to identify the audio source that is better represented by the model. Mathematically speaking, this is formulated as follows. Suppose we have S series of feature vectors $\underline{V}_k$, belonging to different audio sources and the desired instrument model λ. The correct audio source should maximise the following likelihood:

$$\max_{1\le k\le S} P(V_k|\lambda) \tag{10}$$

Combining nongaussianity and probabilistic inference

In this section, we will propose a method to separate the desired source using the kurtosis-based one unit learning law as presented in (7),(8) and the GMM model λ that was trained for the specific instrument.

First of all, we prewhiten the observation signals, i.e. perform *Principal Component Analysis*. After this step, we know that the sources are uncorrelated, i.e. orthogonal to each other. Then, we randomly initiate a one-unit learning rule based on nongausiannity, for example the kurtosis-based one, as described in (7) and (8). Consequently, we get a first estimate $\underline{w}^+$ towards the direction of the most nongaussian component. As the sources are prewhitened, we can also get an estimate of the other sources' directions, as they will be orthogonal to the first estimate in the N-dimensional space. An example of the 2x2 case is illustrated in figure 3. The direction of the orthogonal vector will then be $\underline{w}_\perp^+ \leftarrow \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \underline{w}^+$.

The next step is to calculate the estimated sources at each direction and perform instrument verification. In other words,

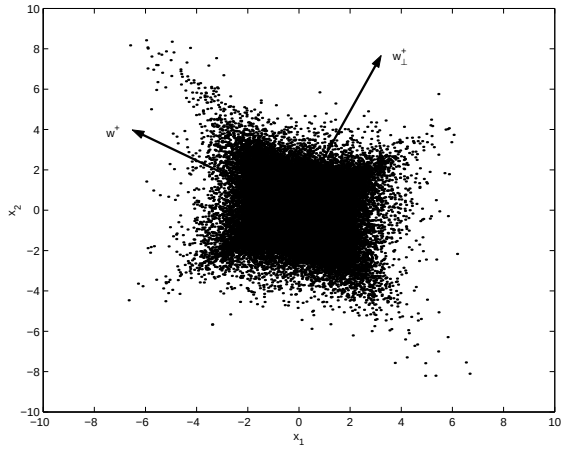

Fig. 3: A scatter plot of the two sources, two sensors case. Getting an estimate of the most nongaussian component can give an estimate of all the other components in the prewhitened N-D space.

extract the feature vectors $\underline{v}_i$ from each of the estimated signals and then calculate the probability of $P(\underline{v}_i|\lambda)$ given the model of desired source. The direction that maximises this probability is the best estimate of the direction of the desired source. This direction will be used as the next starting point for (7).

The same procedure is repeated until convergence. The likelihood comparison step prevents the nongaussian contrast function from converging to the global maximum, but instead to the desired local maximum of kurtosis. Of course, this method works only in *batch mode*, i.e. processing all or at least big blocks of the available data set.

Bayesian Approach

In the Bayesian approach, we try to maximise the posterior probability of the model and form a *Maximum Likelihood* (ML) estimate of the unmixing vector $\underline{w}$. The optimisation problem is set as follows:

$$\max_{\underline{w}} G(\underline{w}) \tag{11}$$

where $G(\underline{w}) = \log P(\underline{v}|\underline{x}, \lambda)$ is defined by equation (2).

We can form a *gradient ascent* solution to this problem, which is given by the following law:

$$\underline{w}^+ \leftarrow \underline{w} + \eta E\{\frac{\partial G}{\partial \underline{w}}\} \tag{12}$$

$$\underline{w}^+ \leftarrow \underline{w}^+ / ||\underline{w}^+|| \tag{13}$$

where η is the learning rate of the gradient ascent. Therefore, we have to calculate the $\partial G/\partial \underline{w}$. Forming a expression connecting $G(\underline{w})$ with $\underline{w}^T\underline{x}$ is not so straightforward, as it is not easy to represent feature extraction with a function f, i.e. $\underline{v} = f(\underline{w}^T\underline{x})$. However, we can split the derivative in the following parts:

$$\frac{\partial G}{\partial \underline{w}} = \frac{1}{P(\underline{v})} \frac{\partial P}{\partial \underline{v}} \frac{\partial \underline{v}}{\partial \underline{w}} \tag{14}$$

where

$$\frac{\partial P}{\partial \underline{v}} = -\sum_{i=1}^{M} p_i b_i(\underline{v}) C_i^{-1}(\underline{v} - \underline{m}_i) \tag{15}$$

The term $\partial \underline{v}/\partial \underline{w}$ is hard to define In our analysis, we performed numerical calculation of this derivative. Another approach can be to perform numerical calculation of the whole derivative.

However, a Maximum Likehood estimate may not always be accurate. This was observed in many speaker verification approaches [6, 11]. In order to improve the performance, the log-probability of the user claiming identification is normalised to the mean of the log-probabilities of all the *cohort* speakers, i.e. the speakers scoring equally well. This is called *cohort normalisation*. Therefore, it seems that in order to improve the performance, we should optimise the following cost function:

$$G(\underline{w}) = \log P(\underline{v}|\underline{x}, \lambda_t) - \frac{1}{N-1} \sum_{i=1}^{N-1} \log P(\underline{v}|\underline{x}, \lambda_i) \tag{16}$$

where λ_t is the model of the desired instrument and λ_i are the models of the other instruments present in the mixture. In other words, we try to maximise the difference between the desired instrument and all other instruments present in the mixture.

This function is equally difficult to optimise, therefore, we calculated numerically the derivative of G. The solution proposed in this paragraph is still developing. However, we demonstrate that it is possible to perform intelligent blind separation by using the posterior likelihood of the instrument/person model.

RESULTS

First of all, we configured the instrument recognition system. Our recogniser used 18 MFCCs, 16 Gaussian mixtures and non-overlapping frames of 16ms. The model for each instrument was trained using around 5-6 minutes of solo instrument recordings. The instruments used in our analysis were piano, violin, accordeon and acoustic guitar. For the recognition process, 15 secs of different recordings were used.

We tested the proposed solutions with instantaneous mixtures of the four instruments in groups of two. The *first solution* was actually capable of producing fast, good quality separation. The algorithm converged in an average of 4 iterations. Performing experiments with combinations of the four instruments that we trained the GMM, the method was always able to spot the correct source. The fast speed of the algorithm is mainly due to the Newton-type method that maximises kurtosis. The likelihood comparison points to the desired direction of convergence. The simple instrument verification setup was able to perform effectively well, as the variety of the instruments was limited. However, this kind of approach is versatile and can be adapted to any advanced instrument verification system, with any feature set configuration, and therefore deal successfully with more difficult instrument recognition cases.

The *second solution* was generally very slow in convergence. This was due to the numerical calculation of the derivative. Firstly, we experimented with maximising only the posterior likelihood with the same testing set as in the first solution. The algorithm was capable of separating one of the source perfectly. However, the separation of the other component

was not good, proving that the ML estimate may not always be accurate enough. On the other hand, maximising the normalised likelihood seemed to be capable of separating both audio sources with good quality, proving that intelligent separation is possible using stronger probabalistic priors.

CONCLUSION

In this study, we explored the possibility of imposing stronger probabilistic priors in the general ICA model, in an effort to separate a particular source of interest. Instrument/speaker verification efforts were employed to provide efficient modelling solutions to this problem. As a result, two methods were proposed for performing *Intelligent ICA*: one exploiting nongaussianity and probabilistic inference for the identity of the separated output and one maximising the posterior likelihood of the instrument/speaker model. The results acquired were encouraging. Moreover, the proposed methods are very versatile to any instrument recognition configuration.

In the future, we would like to formulate a more explicit solution for the bayesian approach, as well as speeding up its convergence with a second-order method. Secondly, we would like to expand these methods in the *"more sources than sensors"* case, as we reckon that we might be able to get better separation results than previous efforts, by imposing stronger priors. This will bring ICA efforts closer to the *Computational Auditory Scene Analysis* (CASA) approach [8] on audio source separation.

ACKNOWLEDGEMENTS

In order to train the GMM, we used the function *gaussmix.m* from Mike Brookes's Voicebox, available from [12].

REFERENCES

[1] Comon P., *Independent Component Analysis, a new concept?*, Signal Processing, Elsevier, 36(3):287–314, April 1994.

[2] Hyvarinen A., *Survey on Independent Component Analysis*, Neural Computing Surveys 2:94–128, 1999.

[3] Hyvarinen A., Oja E., *A Fast Fixed-Point Algorithm for Independent Component Analysis*, Neural Computation, 9(7):1483-1492, 1997.

[4] Amari S., Cichocki A., Yang H. H., *A new learning algorithm for blind source separation*, Advances in Neural Information Processing Systems, pp. 757-763, MIT Press, Cambridge MA, 1996.

[5] Mitianoudis N., Davies M., *New fixed-point solutions for convolved mixtures*, 3rd International Conference on Independent Component Analysis and Source Separation, San Diego, California, December 2001.

[6] Mitianoudis N., *A graphical framework for the Evaluation of Speaker Verification systems*, MSc thesis, Imperial College, University of London, September 2000.

[7] Eronen A., *Comparison of features for musical instrument recognition*, IEEE workshop on Applications of Signal Processing on Audio and Acoustics, New Paltz, New York, October 2001.

[8] Martin K.D., *Sound-Source Recognition: A Theory and Computational Model*, PhD Thesis, Media Lab, MIT, June 1999.

[9] Kostek B., Czyewski A., *Representing instrument sounds for their automatic classification*, Journal of AES, vol. 49, no 9, September 2001.

[10] Reynolds D.A., Rose R.C., *Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models*, IEEE transactions on Speech and Audio Processing, Vol.3, No. 1, January 1995.

[11] Che C.W., Lin Q., Yuk D.S., *An HMM approach to text-prompted speaker verification*, Proceedings of the ICASSP, IEEE 1996, pp. 673-676.

[12] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html