2nd COST-G6 Workshop on Digital Audio Effects (DAFx99), December 1999. Trondheim, Norway.

SEPARATION OF MUSICAL INSTRUMENTS BASED ON PERCEPTUAL AND STATISTICAL PRICIPLES

Lino García

Universidad Politécnica de Madrid
ETSI Telecomunicación
Ciudad Universitaria S/N. 28040 Madrid, Spain
lino@gaps.ssr.upm.es

Javier Casajús-Quirós

Universidad Politécnica de Madrid
ETSI Telecomunicación
Ciudad Universitaria S/N. 28040 Madrid, Spain
javier@gaps.ssr.upm.es

ABSTRACT

The separation of musical instruments acoustically mixed in one source is a very active field which has been approached from many different viewpoints. This article compares the blind source separation perspective and oscillatory correlation theory taking the auditory scene analysis as the point of departure (ASA).

The former technique deals with the separation of a particular signal from a mixture with many others from a statistical point of view. Through the standard Independent Component Analysis (ICA), a blind source separation can be done using the particular and the mixed signals' statistical properties. Thus, the technique is general and does not use previous knowledge about musical instruments.

In the second approach, an ASA extension is studied with a dynamic neural model which is able to separate the different musical instruments taking a priori unknown perceptual elements as a point of departure. Applying an inverse transformation to the output of the model, the different contributions to the mixture can be recovered again in the time domain.

1. INTRODUCTION

Source separation, in general, consists of recovering a group of independent signals from the mixture. When the sources are environmental sounds or the ones that come from musical instruments the separation can profit from the human auditory system segregation perceptual mechanisms. In this case we try to recover a perceptual description of each constituent sound source. The process of simulating the peripheral auditory system has been called auditory scene analysis [1] (ASA) or computational auditory scene analysis [2] (CASA) when we deal with a computational modeling of ASA. The output of these models is the input to the primary auditory nervous system. In this context, the algorithms that simulate the CASA relate the acoustic signals with its auditory representation [9] and are the input to higher level models of neurobiological hearing mechanisms. Trying to understand the neurobiological basis of ASA and the fact that humans can perceptually segregate sound sources with relative ease suggests the possibility of developing neural network models of ASA [10].

The separation of sounds and particularly of speech has also received special attention in the blind source separation research. As distinguished from CASA, the blind source separation is a statistical technique whose underlying model is the observation of *m* linear combinations (probably noisy) of *n* statistically independent signals. The problem lies basically in recovering the original signals from the mixture [5] with no *a priori* information about the mixture matrix coefficients. Precisely the 'blind' separation term is due to them.

2. COMPUTATIONAL AUDITORY SCENE ANALYSIS

The main task of auditory perception is to recover a mental description of each sound source from the acoustic source which is received by our ears composed of sound energy from several environmental sources [10].

Bregman [1] describes this auditory system function as *auditory scene analysis* (ASA). The computational model of ASA provides the basis for the development of psychological and physiological theories of perception.

According to Bregman, ASA can be understood as a two stage process. In the first stage (*segmentation*) the acoustic mixture which is received by the ears is decomposed into a collection of sensory elements *segments*. The second stage (*grouping*) combines segments that probably belong to the same acoustic event in a perceptual entity named *stream*. These streams can lead to higher level processes for the comprehension and recognition of the scene.

Several auditory neuroscience reports agree on the fact that the different properties of acoustic events (such as periodicity, spatial location and spectral shape) are registered in different locations of the auditory system. However, we perceive the auditory events as a whole and not in parts. That is, the auditory system is able to group characteristics represented in remote neural structures to make a perceptual entity [3].

The traditional solution to the grouping problem stablishes a hierarchy of detecting cells with increasingly specialized

characteristics. Von der Malsburg, however, suggested that the response of the characteristic detecting cells can be grouped by the temporary synchronicity of their oscillatory firing activity beginning what has been called *oscillatory correlation theory*. According to this view, the detecting cells which represent a characteristic of the same perceptual event should be synchronised while the cells which represent characteristics of different events should be desynchronised.

Wang and Brown [10] propose an oscillatory network model with two layers which use simple computational methods for the extraction of auditory characteristics.

2.1. MODEL

In the first stage of Wang's model peripheral auditory processing is simulated by passing the input signal through a bank of cochlear filters. The gains of the filters are chosen to reflect the transfer function of the outer and middle ears. In turn, the output of each filter channel is processed by a model of hair cell transduction, giving a probabilistic representation of auditory nerve firing activity which provide the input to subsequent stages of the model [10].

The second stage of the model produces auditory representation of 'mid-level'. The *correlogram* is formed by calculating a running autocorrelation of the auditory nerve activity in each filter channel. The correlogram is computed in intervals of approximately 10ms, forming a three-dimentional volume in which time, channel central frequency and autocorrelation lag are represented on orthogonal axes. Besides, a 'pooled' correlogram is formed at each time frame by summing the periodicity information in the correlogram over frequency. The largest peak in the pooled function occurs at the period of the dominant fundamental frequency (FO) in each time frame. This information is very useful for the third stage of the model to group the acoustic components according to the fundamental frequencies (FO). Besides, an analysis of cross-correlation is made motivated by the observation that filter channels with center frequency close to the same harmonic or formant exhibit similar patterns of periodicity. Due to this, the Wang model computes a running cross-correlation between adjacent correlogram channels providing the basis for the segment formation in the third stage of model.

The third stage of the Wang model consists of a network of relaxation oscillators with two layers. The first layer is locally-excitatory globally-inhibitory oscillator network (LEGION), where the auditory organization takes place. The oscillators in the second layer are linked by two kinds of lateral connections. The first kind consists of mutual excitatory connections between oscillators within the same segment. The second kind consists of lateral connections between oscillators of different segments, but within the same time frame. This layer groups the segments in a stream. The network is bi-dimensional: frequency and time. The strength of the coupling between oscillators is a 'distance' function between time and frequence. The time dimension is implemented as a delay line series.

The input to the Wang model are bi-dimentional binary matrices of time and frequency (*N* channels x *M* time frames). Each input matrix represents a auditory sequence whose binary elements corresponding to time-frequency events are on or off. The network task is simply to respond to each input matrix with active time-frequency events triggering the corresponding time-frequency oscillators [7].

The stream segregation is obtained from the emergent synchronization between simultaneously active oscillators. The synchronicity between oscillators, at the same time, is produced by the excitatory connections between oscillators and is a time-frequency distance function.

3. BLIND SOURCE SEPARATION

The blind source separation consists in recovering unobserved signals or 'sources' from the several observed mixtures. The lack of *a priori* knowledge about the mixture is compensated by a statistically strong but often physically plausible assumption of *independence* between the source signals [4]. The blind separation algorithms try to invert the mixing process in such a way that recovering the components in some way independent. Another frequent strong assumption is to consider the mixing process linear instead of convolutive.

The blind source separation profits mainly the *spatial diversity* (different sensors receive different signal mixtures). The spectral diversity, if it exists, could be profitable but the separation focus is essentially 'spatial': looking for structures through the sensors, not through time.

The two main components of a statistical model are: the mixing matrix and the probability distribution of the source vectors. The mixing matrix is considered linearly independent so that it is invertible. The probability distribution of each source is an annoying parameter because though it is not interesting, it is necessary to know or estimate it in order to estimate 'efficiently' the parameters of interest. The separation techniques in fact depend on the assumptions related to the individual distribution of the sources.

The simplest way of mixing process is

$$x(t) = As(t)$$

where $s(t) = [s_1(t), \ldots, s_n(t)]$ is a *n* x 1 column vector that contains the source signals, $x(t)$ is a vector that contains the *n* observed signals and *A* is a square matrix of *n* x *n* that contains the mixture coefficients .

The problem of blind source separation consists of recovering the vector $s(t)$ only from the observed data $x(t)$. This can be formulated as the computation of an *n* x *n* 'separating matrix' whose output $y(t)$

$$y(t) = Bx(t)$$

is an estimate of the vector $s(t)$ of the source signals.

Normally the use of second order information (decorrelation) allows to reduce the blind separation problem to a simpler form. If a vector has variance unit and covariance unit it is said that the vector is *spatially white*. Therefore, 'whitening' or 'sphering' the data reduce the mixture to a rotation matrix. It means that a separating matrix *B* can be found as a product *B = UW* where *W* is a whitening matrix and *U* is a rotation matrix. Therefore

$$y(t) = UWx(t)$$

The 'contrast' functions are real functions of the distributions of the output ***y*** = *B****x*** and they serve as objectives: they must be designed in such a way that source separation is achieved when they reach their minimum value and are generically denoted by $\phi[\mathbf{y}]$. High order statistics can be used to define contrast functions and can be simply expressed using the *cumulants*.

Statistical independence implies that the joint moments of the source signals of all the orders is zero. For zero mean random variables the second order cumulants are identical to the second order moments. The algorithms which guarantee that only the second order joint moments are zero (e.g., the covariance matrix is the unit) are classified as *principal component analysis* (PCA). However, the algorithms which explicitly operate with higher order statistics are classified as *independent component analysis* (ICA) [4].

The Comon [6] procedure minimizes the fourth order cumulants given by

$$\phi^{\circ}_{ICA}[\mathbf{y}] = \sum_{ijkl \neq iiii} C^2_{ijkl}[\mathbf{y}]$$

Independence can also be tested over a small subset of cross-cumulants with:

$$\phi^{\circ}_{JADE}[\mathbf{y}] = \sum_{ijkl \neq ijkk} C^2_{ijkl}[\mathbf{y}]$$

This is a criterion of 'joint diagonalization' criterion of eigen-matrices (JADE). It has been proved that both algorithms are equivalent but a faster optimization process exists for JADE. JADE assumes a linear mixing model and is not iterative but it reacts directly over the statistics of the complete data set.

4. SOURCES

When the sources are acoustic signals, the linear mixture model that perfectly aligns in time the source signals observed by the microphones is not true due to the differing pathlengths to the microphone. Another complication in the real acoustic environment is the distortion that signals suffer due to echoes and hall acoustic response. A more reasonable mixing model should include these effects, it should be convolutive of the kind

$$x_i(t) = \sum_j (h_{ij} * s)(t)$$

where *h* is the filter impulse response. Another choice is the election of a non-linear mixing model.

The statistic independence suggestion is strong enough when the source signals come from musical instrument. The inner harmonic nature of most of musical instruments can bring about strong non-stationary harmonic correlations. Besides, the sound that come from musical instruments are not stationary and difficult to model statistically.

Our report has been written taking into consideration a simple case of linear mixing under a strong suggestion of acoustic signal independence. In general, the mixtures with white noise are rich in higher order joint statistics and the mixtures with narrow band signals are poor in higher order joint statistics.

5. CONCLUSIONS

Both approaches can be evaluated in terms of signal noise relation (SNR) though they are not identical.

The power of Wang's oscillator network to model streaming lies in three primary components of the model [7]. First, auditory sequences are translated into a special array via a series of delay lines, so that sequence of time-frequency events are processed all-at-once. Second, lateral excitatory connections between oscillators based on time and frequency proximity enable oscillators corresponding to time-frequency events to achieve synchrony if and only if they are sufficiently near in time and frequency. Third, the global inhibitor counteracts lateral excitation by desynchronizing the oscillators. It is the competition between the synchronizing lateral connections and the desynchronizing global inhibitor that permits the model to form streams based on time and frequency proximity, providing a successful model of the basic phenomena.

The Wang model fails to capture some significant aspects like high-low tone sequences [7] (which are perceived as an integrated sequence or as two separate sequences: one only with high tones and other with low tones). In general, architecture has serious problems regarding the granularity of time dimension which does not allow to obtain details from many experiments y and losing many important characteristics [8].

According to von der Malsburg correlation theory, the brain functions imply synchronicity of neural firing but does not require neural oscillation *per se* **Error! Unknown switch argument.**. In fact, there are dynamic systems where isolated cells can show oscillatory activity without being synchronized to other cells. Equally, the firing activity of different cells can be synchronized without showing oscillations. In other words, though the neural oscillations and synchronicity occur together, they are not necessarily dependent.

The oscillatory segregation network functioning is controlled by a great amount of non directly related parameters with

neurophysiological findings and far from a standard which guarantees a universal functioning before several sources.

The segregation is led by the dominant frequency (FO) extracted from the running correlogram which excludes the identification of sounds coming from musical instruments noisy or poorly tonal.

The Wang's model keeps a record of the auditory sequence events through a series of delay lines in such a way that sequences are processed by the oscillatory network all at once.
The segregation of a stream through the oscillator synchronicity keeps itself only while the input sequence persists. However, most of the auditory streams do not remain in memory when there
is a lack of input, and when the stream formation does not take into consideration memory storing and extracting.

Memory, thus, is not a key consequence of the model. The model does not need a learning sequence. As a consequence of this, the sequence processing does not influence the previous sequences exposed to the model. The formation of auditory streams is determined by two kinds of changes which specify the dynamics of oscillator synchronization: changes in the oscillator activations and changes in the weights that modulate the oscillator interaction.

The segregation network can be a reasonable point of departure to model the auditory perception but must be improved from several viewpoints.

The statistical techniques of the blind source separation impose strong statistical restrictions to the mixed signals and to the mixture. Besides, the knowledgement of the source number is required and the number of mixed signals should be equal to the source number. Even when the mixing process is linear, the mixing matrix should be far from singular (if the microphones are relatively close, the effective mixing matrix can nearly be singular). If the sources move in space, the coefficients of the mixing matrix change in time, however, the blind source separation algorithms like JADE require stationary source signals.

Human beings also have enourmous limitations when simultaneously distinguishing many auditory events. In such scenes, the power to separate a source from the mixture consists of isolating selected events which introduces complex attentional mechanisms.

Both approaches work successfully when separating sound mixtures assuming several restrictions more or less significant but very far from a natural hearing environment.

6. REFERENCES

[1] A. S. Bregman. "*Auditory Scene Analysis*", MIT Press, Cambridge, MA, 1990.

[2] G. J. Brown and M. Cooke. "*Computational Auditory Scene Analysis*". Computer Speech and Language, vol. 8, no. 4, pp. 297-336, 1994.

[3] G. J. Brown, M. Cooke, E. Mousset. "*Are Neural Oscillations The Substrate of Auditory Grouping?*". Proceeding of the ESCA Tutorial and Workshop on the Auditory Basis of Speech Perception. Keele University. July 15-19, 1996.

[4] J-F. Cardoso. "*Blind Signal Separation: Statistical Principles*". Proceedings of the IEEE, vol. 9, no. 10, pp. 2009-2025, Oct. 1998.

[5] J-F. Cardoso. "*Equivariant Adaptative Source Separation*". to appear in IEEE Transactions on Signal Processing.

[6] P. Comon. "*Independent Component Analysis, A New Concept?*". Signal Processing, vol. 36, no. 3, pp. 287-314, Apr. 1994. Special issue on Higher-Order Statistics.

[7] J. D. McAuley. "*Auditory Scene Analysis via Emergent Synchrony*". Cognitive Modelling Workshop of the Seventh Australian Conference on Neural Networks, Australian National University Canberra, 1996.

[8] M. Norris. "*Design Decisions in an Oscillatory Model of Primitive Auditory Segregation*". Draft. University of Queensland.

[9] X. Yang, K. Wang, S. A. Shamma. "*Auditory Representations of Acoustic Signals*". IEEE Transactions on Information Theory. vol. 38, no. 2, March 1992.

[10] D. Wang, G. J. Brown. "*Separation of speech from interfering sounds based on oscillatory correlation*". Cognitive Science, Thecnical Reports, 24, June, 1998. The Ohio State University.