Experimental Evaluation of Source Separation with Only One Sensor

J. Walker*, F. Bimbot and L. Benaroya†**

***Department of Electronic and Computer Engineering, University of Limerick, Ireland, **IRISA (CNRS&|INRIA), Rennes, France, †Mist technologies, Paris, France.**

Abstract

We report on the evaluation of a new method for audio source separation using only one sensor. The method can be viewed as a generalization of Wiener filtering to locally stationary signals, where the sources are modelled using power spectral density dictionaries which are estimated during a training step. The experiments were designed to measure how separation performance varied with amount of training data, model complexity and the representativity of the training data. The results show that model complexity and training data representativity are more important than the amount of training data.

1 Introduction

Source separation techniques have many potential applications in speech and music signal processing including polyphonic music transcription (Plumbley et al., 2002). The classical blind source separation problem describes a linear instantaneous mixture where the number of available mixtures is at least equal to the number of sources. If the sources are assumed to be both independent and Gaussian, the mixing matrix may be found and used to discover the sources (Cardoso, 1998). The under-determined source separation problem (Lee et al., 1999) (where the number of sources is greater than the number of mixtures) often arises with musical recordings as these may be available only in stereo form, and, sometimes, only in mono. Here, we consider the case of a single mixture: $x(t) = s_1(t) + s_2(t)$ and as traditional blind estimation is not possible, we build a source model, based on *a priori* knowledge of the sources in a Bayesian framework. If $s_1(t)$ and $s_2(t)$ were stationary Gaussian processes, the Bayesian optimal estimates would be obtained by Wiener filtering (Wiener, 1949). For sources, such as audio sources, which are only locally stationary (non-Gaussian), we can represent each source with a Gaussian mixture model (GMM) (Bijaoui, 2002; Benaroya and Bimbot, 2003) and generalise the Wiener filtering

$$\hat{S}_i(t, f) = \sum_{k_1 = 1}^{K_1} \sum_{k_2 = 1}^{K_2} \gamma_{k_1 k_2}(t) \frac{\sigma^2_{k_1, i}(f)}{\sigma^2{}_{k_1, 1}(f) + \sigma^2_{k_2, 2}(f)} S_x(t, f) \qquad 1.1$$

where $\hat{S}_i(t, f)$ is the short-time Fourier transform (STFT) of the ith source estimate $i = 1, 2$, $S_x(t, f)$ is the STFT of the mixture, $\sigma^2_{k_i, i}(f)$ is the variance of the Gaussian density of the k_ith ($k_i = 1, 2, \ldots, K_i$) component of the ith source at the frequency component f and $\gamma_{k_1 k_2}(t)$ are the weights associated with the pair of covariances $\sigma^2_{k_i, i}(f)$ and $\sigma^2_{k_j, j}(f)$ (Benaroya, 2003) with which we model the shapes of the power spectral densities found within each source.

GMMs cannot take account of the possible variation in the amplitude of the power spectral densities. Gaussian scaled mixture models (GSMM) deal with this problem by introducing a slowly varying non-negative amplitude factor $a(t) \geq 0$ so that the source is represented as $s_i(t) = a_i(t) \times b_i(t)$, where $b_i(t)$ is the underlying Gaussian process with covariance $\sigma_i^2(f)$. The Wiener filtering then becomes (Portilla et al., 2001; Benaroya et al., 2003):

$$\hat{S}_i(t, f) = \frac{a_i(t)\sigma_i^2(f)}{a_1(t)\sigma^2{}_1(f) + a_2(t)\sigma_2^2(f)} S_x(t, f) \qquad 1.2$$

This basic approach can be used to develop a more realistic source model in which different spectral shapes may occur at different times with varying amplitudes: $s_i(t) = \sum_{k \in K_i} a_k(t) b_k(t)$, i.e., we represent the source as a combination of slowly varying amplitude factors, $a_k(t)$ and a set of stationary Gaussian processes with covariances $\sigma^2{}_k(f)$ (Benaroya et al., 2003). There are two different ways to develop this approach (Benaroya, 2003). In the GSMM approach, we assume that only one component in the model is active at any one time, and we condition our source estimates on a pair of active power spectral densities and an associated pair of amplitude factors. This is equivalent to assuming that only one instrument (per source) plays at any one time. Alternatively, it is possible to assume that more than one component in the model (instrument) can be active at a time, and that these components can add, weighted by their individual amplitudes, to generate the source. This second approach can be seen as the development of a dictionary of signal pattern prototypes specific to each source: a power spectral density (PSD) dictionary model. The separation process is then formulated as:

$$\hat{S}_1(t,f) = \frac{\sum_{k \in K_1} a_k(t)\sigma_k^2(f)}{\sum_{k \in K_1 \cup K_2} a_k(t)\sigma^2{}_k(f)} S_x(t,f)$$
$$\hat{S}_2(t,f) = \frac{\sum_{k \in K2} a_k(t)\sigma_k^2(f)}{\sum_{k \in K_1 \cup K_2} a_k(t)\sigma^2{}_k(f)} S_x(t,f) \quad 1.3$$

for the two sub-dictionaries $\{\sigma^2{}_k(f)\}_{k \in K_1}$ and $\{\sigma^2{}_k(f)\}_{k \in K_2}$ which are characteristic of the two sources $s_1(t)$ and $s_2(t)$, where $K_1 \cap K_2 = \varnothing$. The aim is to express the observed mixture as a decomposition over the two sub-dictionaries (as in the denominator of equation (1.3)), estimating the corresponding amplitude coefficients for each frame (Benaroya, 2003).

1.1 Training Step

The training step consists of the development of the PSD dictionaries from the training data. In the log-spectral domain, vector quantization is first used to split the training data into the required number of model components (PSDs). Initialising the amplitude coefficients, a_k on the basis of this initial model, we then iterate between estimating the source dictionaries, and re-estimating the amplitude coefficients (Benaroya, 2003).

2 Experiments

For the experiments we used the first movement of Sonata No. 1 in D major op. 12 No. 1 by L. van Beethoven in a midi version, including separate violin and piano tracks (Paterson, 1999-2004). The two sources are violin and piano, which are known to have quite distinct timbres. Specifically, the piano is known for its firm attack and the inharmonic nature of its partials and the violin is known for its gradual attack and the relative strength of its higher partials. The musical excerpts used as training and test sources were taken from the different parts of the separate violin and piano tracks. In most cases, the mixture, from which the sources are to be separated, was created by adding the test sources. In some cases the test mixture was taken from the part of the ensemble midi corresponding exactly to the location of the individual test signals.

Midi files were the main source of signals for the tests because of the difficulty in obtaining adequate amounts of training and test material from real recordings. Often, with real recordings, only the real mixture is available and individual instruments or instrument groups play solo for only short periods of time. Using midi also ensures that the individual training and test sources will be exactly identical to the sources as they play within the ensemble, even if an ensemble mixture is used. The chief disadvantage of using midi files is that the testing may not be as demanding as a test on real audio signals; the regularity of midi files tending to make the separation task easier than it would otherwise be. In order to use midi files, they have to be converted into audio (typically Microsoft .wav format). The quality of the rendered instruments depends on the midi patches used and can be very variable. We would think that this problem is likely to make the separation task easier because poorly rendered instruments have an “electronic” sound suggesting a simple timbre which may be more easily distinguished in the spectral domain (e.g. it may be more concentrated or more regular in its distribution in the vector space).

2.1 Evaluation Criteria

For evaluation, we use the Source-to-Interference Ratio (SIR) and the Source-to-Artefact Ratio (SAR) (Gribonval et al., 2003). Given the original sources, s_1 and s_2 and their estimates $\hat{s}_1$ and $\hat{s}_2$, the projections of the estimated sources over the vector space spanned by the real sources are: $\hat{s}_1 = \alpha s_1 + \alpha s_2 + n_1$ and $\hat{s}_2 = \beta s_1 + \beta s_2 + n_2$.

$$SIR_1 = 20\log\frac{|\alpha_1|\|s_1\|}{|\alpha_2|\|s_2\|} \qquad SIR_2 = 20\log\frac{|\beta_1|\|s_1\|}{|\beta_2|\|s_2\|}$$
$$SAR_1 = \frac{\|\hat{s}_1 - n_1\|}{\|n_1\|} \qquad SAR_2 = \frac{\|\hat{s}_2 - n_2\|}{\|n_2\|} \quad 2.1$$

Thus, the SIR measures the interference due to the residual of one source in the estimation of the other; whereas the SAR is a measure of the amount of distortion in each estimate.

2.2 Experimental Questions

Three different factors which might affect the source separation were evaluated: the length of the training data, the number of components in the model and the representativity of the training data. The effect of training data length on the source separation performance was evaluated with the model complexity held constant. Unfortunately, varying the length of the training data also affects the representativity of the training data, but does not necessarily increase it. To deal with

this problem, the training data is lengthened systematically, i.e. it is always taken to start at the beginning of the track, and continues for the indicated number of seconds. The training data lengths chosen were: 15s, 45s, 90s and 105s. 105s represents a practical computational limit under the present implementation running on a Dell workstation 360 Pentium 4. During the training data length experiments, the number of model components for each source was held constant at 16. In the model complexity experiments, the number of model components is increased to 32 and 64, while the training data length is held constant at 45s.

The key factor in how "representative" the training data is of the test data is probably whether all of the notes found in the test data are present in the training data. However, the similarity of tempo, rhythm and occurrence of distinctive patterns of note runs or particular chords are also likely to be important. In these experiments the representativity of the training data was varied while holding the length of the training data and the model complexity constant. The criterion for representativity which was used was subjective: eight different training selections were made (i.e. four for each instrument), all of length 45s. The track which was considered most representative contained sections which were very nearly identical to the test data; the track which was least representative was taken from a part of the piece where a stylistic interlude occurs and the key changes. The selections for each instrument were made separately, but resulted in corresponding selections coming from the same (to within approximately ± 5 s) part of the piece. Training was only conducted using training data pairs of corresponding representativity. The case where the training data is identical to the test material was also considered. In this case, the same 15 second segment was taken, for each instrument, for test and training. To produce a training sample of 45 seconds in length, the test sample was concatenated three times. As an objective assessment of representativity for comparison with the subjective assessment, a signal correlation was taken between the training data for a given instrument and the test data for that instrument. The experiments on representativity were run with two different degrees of model complexity, but only the results for a model size of 32 are shown here.

3 Results

In Table 1 and Table 2 we present the results obtained by varying the training data length. It can be seen that having only the same amount of training data as test data (i.e. 15 seconds in this case) is very detrimental to both separation performance (as measured by SIR) and distortion (as measured by SAR). However, increasing the length of the training data does not improve performance without limit. A plateau effect in performance improvement with increasing training data length is also noticeable for the SAR measure.

Training Data Length	**Source 1**		**Source 2**	
	SIR	**SAR**	**SIR**	**SAR**
15 secs	2.9125	5.7404	6.3710	-0.3013
45 secs	13.7520	8.0123	16.5663	7.2946
90 secs	14.6497	8.1117	15.8787	7.7896
105 secs	13.7035	8.0142	15.2619	7.5703

TABLE 1.SIR/SAR Results for training data length with additive mixture.

Separation of the sources from the real mixture appears from these results to be more difficult than from the artificial additive mixture. This is surprising because a real mixture in a midi is not that different from an additive mixture. However, in this case the additive mixture is made using the mean-subtracted time-frequency representations of the sources. Also, the SIR and SAR for the real mixture results are necessarily calculated comparing the source estimates with the individual source test data, and not directly with the individual piano and violin parts within the mixture (which should nevertheless be identical.) The subjective sound quality is comparable between the two sets of results (and between the reconstructed and the real mixture.)

Training Data Length	**Source 1**		**Source 2**	
	SIR	**SAR**	**SIR**	**SAR**
15 secs	4.0685	-3.4780	5.3919	-6.6188
45 secs	10.1969	-4.3277	16.7155	1.5826
90 secs	11.0827	-4.2927	16.9602	2.0133
105 secs	10.3203	-4.2153	16.8463	1.8384

TABLE 2.SIR/SAR results for training data length with real mixture.

As shown in Table 3, increasing model complexity has a stronger effect on separation performance than increasing training data length. There may be a limit to the possible improvement, for example in source 1 (violin), improvement is greater as number of model components is increased from 16 to 32, than from 32 to 64. However, for source 2 (the piano), the increase is greater for the change from 32 to 64 components.

Number of model components	Source 1		Source 2	
	SIR	SAR	SIR	SAR
16	13.7520	8.0123	16.5663	7.2946
32	17.2859	8.8636	17.1988	8.8155
64	18.5109	10.0023	19.8171	9.7668

TABLE 3.SIR/SAR results for varying number of model components.

The results for training data representativity are shown in Table 4. Performance improves consistently from poor separation performance when the training data is subjectively chosen as least representative to very high quality separation when the training data is the test data itself. As the length of the training data was 45 seconds in all cases and an additive mixture was always used a comparison may be made with the second row of Table 1. Performance with all training data samples except “High” and “Identical” was worse than the 45 seconds case from Table 1. The “Identical” training data did produce another further improvement in performance and may indicate the limit of what can be achieved for a given training data length and model size.

Data Representativity	Source 1			Source 2		
	Training-test Correlation	SIR	SAR	Training-test Correlation	SIR	SAR
Low	-18.5420	6.2249	4.9790	-20.6033	7.3355	3.8136
Low-medium	-19.9118	11.8581	4.8998	-16.5463	8.9985	6.3051
Medium	-19.1645	12.2724	6.3038	-22.7756	11.8009	6.4293
High	-13.8901	17.2355	8.5127	-16.7644	16.4543	8.6029
Identical	4.6885	19.4205	10.1266	4.7390	19.4271	10.0696

TABLE 4.SIR/SAR results for training data representativity and model complexity.

3.1 Discussion

The results suggest that simply increasing the length of the training data will not continue to improve separation performance, probably due to the fact that some parts of the longer pieces of training data will be unrepresentative of the test sample. Model complexity was shown to be an important factor in separation performance and the experiments suggest that improving the model is more important than simply increasing the amount of training.

Experiments in varying the representativity of the training data show that this factor is important. There was a clear link between the subjective evaluation of how representative the training data was and the separation performance. There was less correspondence between the statistical correlation of the training and test data and subsequent separation performance. An interesting area of further research would be to develop an objective measure of representativity, which would correspond more closely with the subjectively perceived representativity.

4 References

BENAROYA, L.. 2003. Séparation de plusieurs sources sonores avec un seul microphone. Thesis, Université de Rennes 1, Rennes, France.

BENAROYA, L. & BIMBOT, F. 2003. Wiener based source separation with HMM/GMM using a single sensor. *Proc. 4th International Symposium on Independent Component Analysis (ICA 2003),* 957-961.

BENAROYA, L., MCDONAGH, L., BIMBOT, F. & GRIBONVAL, R. 2003. Non negative sparse representation for Wiener based source separation with a single sensor. *Proc. IEEE Int. Conf. Acoustics Speech and Signal Proc.*, **6**, 613-616.

BIJAOUI, A. 2002. Wavelets, Gaussian mixtures and Wiener filtering. *Signal Processing,* **82**, 709-712.

CARDOSO, J. F. 1998. Blind signal separation: statistical principles. *Proc. IEEE,* **86**, 2009-2025.

GRIBONVAL, R., BENAROYA, L., VINCENT, E. & FÉVOTTEE, C. 2003. Proposals for performance measurement in source separation. In *Proc. 4th International Symposium on Independent Component Analysis (ICA 2003),* 763-768.

LEE, T., LEWICKI, M., GIROLAMI, M. & SEJNOWSKI, T. 1999. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, **4**, 87-90.

PATERSON, J. 1999-2004. Violin sonata No. 1 in D Major by L. V. Beethoven. Individual piano and violin parts and full sonata in Midi files. Available at: http://www.mfiles.co.uk/.

PLUMBLEY, M. D., ABDALLAH, S. A., BELLO, J. P. , DAVIES, M. E., MONTI, G. & SANDLER, M. B. 2002. Automatic music transcription and audio source separation. *Cybernetics and Systems.* **99**, 603-627.

PORTILLA, J., STRELA, V., WAINWRIGHT, M. J., & SIMONCELLI, E. P. 2001. Adaptive Wiener denoising using a Gaussian scale mixture model in the wavelet domain. *Proc. Of the 8th International Conference on Image Processing*, Thessaloniki, Greece, 37-40.

WIENER, N. 1949. Extrapolation, interpolation and smoothing of stationary time series. MIT Press: Boston, MA.