EVALUATION OF SEVERAL STRATEGIES FOR SINGLE SENSOR SPEECH/MUSIC SEPARATION

Raphaël BLOUET[1]*, Guy RAPAPORT*[2]*, Israel COHEN*[2] *and Cédric FÉVOTTE*[3] †

[1] MIST Technologies Research
204, rue de Crimée
75019 Paris, France

[2] Dept of Electrical Engineering
Israel Institute of Technology
Technion City, Haifa 32000, Israel

[3] CNRS-TELECOM ParisTech
37-39, rue Dareau
75014 Paris, France

[1] raphael.blouet@mist-technologies.com
[2] {guyrom@tx,icohen@ee}.technion.ac.il
[3] fevotte@tsi.enst.fr

ABSTRACT

In this paper we address the application of single sensor source separation techniques to mixtures of speech and music. Three strategies for source modeling are presented, namely Gaussian Scaled Mixture Models (GSMM), Autoregressive (AR) models and Amplitude Factor (AF). The common ingredient to the methods is the use of a codebook containing elementary spectral shapes to represent non-stationary signals, and to handle separately spectral shape and amplitude information. We propose a new system that employs separate models for the speech and music signals. The speech signal proves to be best modeled with the AR-based codebook, while the music signal is best modeled with the AF-based codebook. Experimental results demonstrate the improved performance of the proposed approach for speech/music separation in some evaluation criteria.

Index Terms— Single sensor source separation, Gaussian mixture models, spectral estimation, autoregressive model.

1. INTRODUCTION

Single sensor source separation is a challenging research topic that attracts much interest in many fields including audio processing, medical imaging, and communication. In audio, attempts to solve this task were proposed in the context of Computational Auditory Scene Analysis (CASA) [1] or binary masking techniques [2]. Other approaches involve various techniques and models such as dual Kalman filters [3], Independent Component Analysis (ICA) [4], sparse decompositions [5] or Nonnegative Matrix Factorization (NMF) [6]. We here consider separation techniques embedded in a probabilistic Bayesian framework. In this context, codebook approaches have recently been successfully employed [7]. In this paper we consider three different codebook-based strategies, namely Gaussian Scaled Mixture Models (GSMM), Autoregressive (AR) models and Amplitude Factor (AF) models. The methods are described and compared in Section 2, and then evaluated in Section 3 on speech and piano mixtures. This section also describes an hybrid approach, a combination of two of the latter techniques, in which speech is modeled by an AR-based codebook, while AF-based codebook is employed for the background music. Finally Section 4 gives conclusions and directions for future work.

*This work has been supported by the European Commission under project Memories FP6-IST-035300.

†This work was done while C. Févotte was a research engineer with Mist-Technologies.

2. SINGLE SENSOR SOURCE SEPARATION

2.1. Problem Formulation

Given an observed signal x, which is the mixture of two sources s_1 and s_2, the source separation problem consists of finding estimates for s_1 and s_2 from x. Algorithms presented in this work are applied in the Short Time Fourier Transform (STFT) domain. Denote by $X(f,t)$ the STFT of x, where t represents the frame index and f the frequency-bin index. Due to linearity of the STFT, we have :

$$X(f,t) = S_1(f,t) + S_2(f,t)\,. \tag{1}$$

We here aim at deriving estimators $\hat{S}_1(t,f)$ and $\hat{S}_2(t,f)$.[1] Codebook approaches rely on the assumption that each source can be represented by a given "dictionary" representative of the nature of the signal. They usually work in two stages :

i) An offline learning step builds the codebooks from training data;
ii) An estimation step finds the source parameters that best explain the mixture from the given codebook(s).

In the following the codebooks representative of the first and second source are noted $\phi_1 = \{\phi_{1,k}\}_{k=1,\dots K_1}$ and $\phi_2 = \{\phi_{2,k}\}_{k=1,\dots,K_2}$ respectively, where K_1 and K_2 are

[1] If only two sources are considered here, extension to n sources is straightforward.

the codebooks lengths. The nature of the codebooks is source dependent. As will be detailed in the next section, the GSMM and AF-based codebooks contain variance parameters (under Gaussian modeling), while the AR-based codebook contains Linear Predictive Coefficients (LPC). As for the reconstruction of the source signals given the estimated representation parameters, we will describe approaches either based on Maximum A Posteriori (MAP) or Minimum Mean Square Error (MMSE) principles.

2.2. GSMM-based source separation

The source separation technique presented in [7] suggests the use of GSMMs to model the sources statistical behavior. In this case, the codebook is simply formed by Gaussian Mixture Models (GMMs) parameters trained from sample data representative of the sources. The GSMM incorporates a supplementary scale parameter which aims at better taking into account non-stationarity of the sources. Each component k of source i is identified by a diagonal covariance matrix $\Sigma_{i,k}$ and a state prior probability $\omega_{i,k}$, so that in this case we have $\phi_{i,k} = \{\Sigma_{i,k}, \omega_{i,k}\}$. The GSMM model is then simply defined by :

$$p(S_i(:,t)|\{\phi_{i,k}\}_k) = \sum_{k=1}^{K_i} \omega_{i,k}\, \mathcal{N}(S_i(:,t)|0, a_{i,k}(t)\Sigma_{i,k}) \tag{2}$$

where $\sum_{k=1}^{K_i} \omega_{i,k} = 1$, $a_{i,k}(t)$ is a time-varying amplitude factor and $S_i(:,t)$ denotes the vector of frequency coefficients of source i at frame t. Though GSMMs are a straightforward extension of GMMs, they are unfortunately untractable due to the added amplitude factors. [7] suggests to estimate these amplitude factors pairwise, in a Maximum Likelihood (ML) sense, as follows :

$$\begin{aligned} \gamma_{a_{1,k},a_{2,q}}(t) &= P(\phi_{1,k}, \phi_{2,q}|X(:,t), a_{1,k}(t), a_{2,q}(t)) \\ \hat{a}_{1,k}(t), \hat{a}_{2,q}(t) &= \max_{a_{1,k},a_{2,q}} \{\gamma_{a_{1,k},a_{2,q}}(t)\} \end{aligned} \tag{3}$$

The source STFTs can then be estimated either in a 1-Best hard decision MMSE (H-MMSE) or MMSE sense, as follows.

H-MMSE estimator :

$$\hat{S}_i(f,t) = \frac{\hat{a}_{i,k^*}\sigma^2_{i,k^*}(f)}{\hat{a}_{1,k^*}\sigma^2_{1,k^*}(f) + \hat{a}_{2,q^*}\sigma^2_{2,q^*}(f)} X(f,t) \tag{4}$$

where $(k^*, q^*) = \underset{(k,q)}{\text{argmax}}\{\gamma_{\hat{a}_{1,k},\hat{a}_{2,q}}(t)\}$.

MMSE estimator :

$$\hat{S}_i(f,t) = \sum_{k,q} \gamma_{\hat{a}_{1,k},\hat{a}_{2,q}}(t) \frac{\hat{a}_{i,k}\sigma^2_{i,k}(f)}{\hat{a}_{1,k}\sigma^2_{1,k}(f) + \hat{a}_{2,q}\sigma^2_{2,q}(f)} X(f,t) \tag{5}$$

Note that since the covariance matrices are assumed diagonal, conditionally on the selected state, separation is performed independently in each frequency bin.

2.3. AR-based source separation

Spectral envelopes of speech signals in the STFT domain are efficiently characterized by AR models, which have been used for enhancement in [8, 9]. Many earlier methods for speech enhancement assume that the interfering signal is quasi-stationary, which restricts their usage for non-stationary environments, such as music interferences. Srinivasan and al. [8, 9] suggest to represent the speech and interference signals by using codebooks of AR processes. The predefined codebooks now contain the linear prediction coefficients of the AR processes, noted $\phi_1 = \{\phi_{1,k}\}_{k=1,\ldots,K_1}$ and $\phi_2 = \{\phi_{2,k}\}_{k=1,\ldots,K_2}$ ($\phi_{i,k}$ is now a vector of length equal to the AR order).

2.3.1. ML approach

[8] proposes a source separation approach based on the ML. The goal is to find the most probable pair $\{\phi_{1,k^*}, \phi_{2,q^*}\}$ for a given observation, with

$$(k^*, q^*) = \underset{k,q}{\text{argmax}} \left\{ \max_{\lambda_{1,k}(t),\lambda_{2,q}(t)} \{p(x(:,t)|\phi_{1,k}, \phi_{2,q}; \lambda_{1,k}(t), \lambda_{2,q}(t))\} \right\} \tag{6}$$

where $x(:,t)$ denotes frame t of mixture x (this time in the time domain) and $\lambda_{1,k}(t)$, $\lambda_{2,k}(t)$ are the frame-varying variances of the AR processes describing each source. In [8] a method is proposed to estimate the excitation variances pairwise. Like previously, once the optimal pair is found, source separation can be achieved through Wiener filtering on the given observation $x(:,t)$.

2.3.2. MMSE approach

[9] proposes an MMSE estimation approach for separation. In a Bayesian setting, the LPC and excitation variances are now considered as random variables, which can be given prior distributions to reflect *a priori* knowledge. Denoting by $\theta = \{\phi_1, \phi_2, \{\lambda_1(t)\}_t, \{\lambda_2(t)\}_t\}$, the MMSE estimator of θ is

$$\hat{\theta} = E[\theta|x] = \frac{1}{p(x)} \int_\theta \theta\, p(x|\theta) p(\theta) d\theta \tag{7}$$

We take $p(\theta) = p(\phi_1)\, p(\phi_2)\, p(\{\lambda_1(t)\}_t)\, p(\{\lambda_2(t)\}_t)$. [9] then shows that the likelihood function $p(x|\theta)$ decays rapidly when deviating from the true excitation variances. This gives ground to approximating the true excitation variances by their

ML estimates, (7) can then be rewritten as

$$\hat{\theta} = \frac{1}{p(x)} \int_{\phi_1,\phi_2} [\phi_1, \phi_2]\, p(x|\phi_1, \phi_2; \hat{\lambda}_1^{ML}, \hat{\lambda}_2^{ML}) \times p(\phi_1) p(\hat{\lambda}_1^{ML}) p(\phi_2) p(\hat{\lambda}_2^{ML}) d\phi_1 d\phi_2 \quad (8)$$

where $\hat{\lambda}_1^{ML}$ and $\hat{\lambda}_2^{ML}$ are the ML estimates of the excitation variances. We use codebook representatives as entries in integration (8).

Assuming that they are uniformly distributed, $\hat{\theta}$ is given by [9] :

$$\hat{\theta} = \frac{1}{K_1 K_2} \sum_{k=1}^{K_1} \sum_{q=1}^{K_2} \theta_{kq} \frac{p(x|\phi_{1,k}, \phi_{2,q}; \lambda_{1,k}^{ML}, \lambda_{2,q}^{ML})}{p(x)} \times p(\lambda_{1,k}^{ML}) p(\lambda_{2,q}^{ML}) \quad (9)$$

where $\theta_{kq} = [\phi_{1,k}, \phi_{2,q}, \hat{\lambda}_{1,k}^{ML}, \hat{\lambda}_{2,q}^{ML}]$. Given two fixed AR codebooks, (9) allows an MMSE estimation of AR processes jointly associated to source 1 and source 2. Once $\hat{\theta}$ is known, we can use Wiener filtering for the separation stage.

2.4. Amplitude Factor source separation

This source separation technique described in [10] proposes to model each STFT frame of each source as a sum of *elementary components* modeled as zero-mean complex Gaussian distribution with known Power Spectral Density (PSD), also referred to as *spectral shape*, and scaled by amplitude factors. More precisely, each source STFT is modeled as

$$S_i(f,t) = \sum_{k=1}^{K_i} \sqrt{a_{i,k}(t)} \cdot E_{i,k}(f,t) \quad (10)$$

where $E_{i,k}(f,t) \sim \mathcal{N}_c(0, \sigma_k^2(f))$. The representatives of the codebooks are now $\phi_{i,k} = [\sigma_{i,k}^2(f_1), \ldots, \sigma_{i,k}^2(f_N)]^T$.

This model is well adapted to the complexity of musical sources, as it explicitly represents the signal as linear combination of more simple components, with various spectral shapes.

Given the codebooks, the separation algorithm based on this model consists of two steps, as follows :

i) Compute of the amplitude parameters $\{a_{i,k}(t)\}$ in an ML sense; this is tantamount to performing a nonnegative expansion of $|X(f,t)|^2$ onto the basis formed by the union of the codebooks,

ii) Given the estimated $\{a_{i,k}(t)\}$, estimate each source in an MMSE sense through Wiener filtering :

$$\hat{S}_i(f,t) = \frac{\sum_{k=1}^{K_i} \hat{a}_{i,k} \sigma_{i,k}^2(f)}{\sum_{k=1}^{K_1} \hat{a}_{1,k} \sigma_{1,k}^2(f) + \sum_{k=1}^{K_2} \hat{a}_{2,k} \sigma_{2,k}^2(f)} X(f,t) \quad (11)$$

2.5. Learning the Codebooks

We assume that we have some clean training samples of each source. These training excerpts do not need to be identical to the source signals in the observed mixture, but we assume that they are *representatives* of the sources. We estimate the codebooks on the training samples according to the models of previously presented separation strategies :

Model of Section 2.2: the Expectation-Maximization algorithm [11] is used to estimate $\{\Sigma_{i,k}, \omega_{i,k}\}_{k=1}^{K_i}$,

Model of Section 2.3: the generalized Lloyd algorithm is used to learn the LPC coefficients [12],

Model of Section 2.4: the generalized Lloyd algorithm is applied to the short-term power spectra of the training samples.

3. RESULTS

3.1. Evaluation criteria

We used the standard Source to Distortion Ratio (SDR), the Signal to Interference Ratio (SIR) and the Signal to Artifacts Ratio (SAR) described in [13]. In short, the SDR provides an overall separation performance criterion, while the SIR only measures the level of residual interference and the SAR measures the level of artifacts in each estimated source. The higher are the ratios, the better is the quality of the estimation. Note that in underdetermined source separation, the SDR is usually driven by the amount of artifacts in the source estimates.

3.2. Experimental setup and results

Audio samples are available at [14]. The evaluation task consists of unmixing a mixture of speech and piano. The signals are sampled at 16 kHz and the STFT is calculated using a Hamming window of 512 samples length (32 ms) with 50% overlap between consecutive frames.

For the learning step we used piano and speech segments that were 10 minutes long. The observed signals are obtained from mixtures of 25 s long test signals. The data set consists of speech segments taken from the TIMIT database and piano segments acquired through the Web. Results are shown in Table 1. All methods use codebook with 128 components.

When observing the simulation results, one can see that no single algorithm is superior for all criteria. However, the AR/MMSE performs well when separating the speech. Another observation is that the AR model yields low SIR results for the piano source; this can be explained by the fact that AR processes are not very adequate for representing piano signals. We thus propose to combine the AF and AR based meth-

ods : an AF-based codebook is used for the piano while an AR-based codebook is used for speech. The results with this approach, shown in Table 1, show an enhancement in performance in one evaluation criteria (speech SIR) while the other criteria stay akin to those obtained with the best-performing systems.

Table 1. SIR/SDR/SAR Measures (in dB) for GMM/AR/Amplitude Factor and Amplitude Factor + AR Based Methods.

		GSMM		AR		Ampl.	AM +
		H-MMSE	MMSE	ML	MMSE	Factor	AR
Speech	SDR	5.4	4.8	4.8	3.9	4.2	4.1
	SIR	3.9	4.1	4.5	2.2	4.4	4.6
	SAR	2.4	2.7	2.0	4.7	3.5	3.8
Music	SDR	3.0	2.8	2.5	2.1	2.6	2.4
	SIR	10.8	11.1	7.1	5.0	10.4	10.8
	SAR	3.2	3.5	7.6	12.9	5.0	5.3

4. CONCLUSION

We have presented in this paper three codebook approaches for single channel source separation. Each codebook underlies different models for the sources, i.e addresses different features of the sources. The above separation results show that AR-based model efficiently captures speech features, while the AF-based model is good at representing music because of its additive nature (a complex music signal is represented as a sum of simpler elementary components). Oppositely, the GSMM assumes in its conception that the audio signal is *exclusively* in one state *or* another, which intuitively does not best explain music. The separation results presented in this paper also tend to corroborate this fact.

It is worthwhile noting that the above methods rely on the assumptions that sources are continuously active in all time frames. This is generally incorrect for audio signals, and we will try in our future work to use source presence probability estimation in the separation process. The separation algorithms define the posterior probabilities and gain factors of each pair based on the entire frequency range. This causes numerical instabilities and does not take into consideration local features of the sources, e.g., for speech signals the lower frequencies may contain most of the energy. Another aspect of our future work will consist in adding perceptual frequency weighting in the expansion coefficient estimation.

5. REFERENCES

[1] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds using oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, pp. 684–697, Sep. 1999.

[2] S. T. Roweis, "One microphone source separation," *Advances in Neural Information Processing Systems*, vol. 13, pp. 793–799, 2000.

[3] E. A. Wan and A. T. Nelson, "Neural dual extended Kalman filtering: Applications in speech enhancement and monaural blind signal separation," in *IEEE Workshop on Neural Networks and Signal Processing*, Amelia Island, FL, USA, 1997, pp. 466–475.

[4] G.-J. Jang and T.-W. Lee, "A probabilistic approach to single channel source separation," *Advances in Neural Information Processing Systems*, vol. 15, 2003.

[5] B. A. Pearlmutter and A. M. Zador, "Monaural source separation using spectral cues," in *Proc. 5th International Conference on Independent Component Analysis and Blind Source Separation (ICA'04)*, Granada, Spain, Sep. 2004.

[6] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *International Computer Music Conference (ICMC'03)*, Singapore, Sep. 2003.

[7] L. Benaroya, F. Bimbot, and R. Gribonval., "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[8] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[9] ——, "Codebook-based Bayesian speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, Philadelphia, USA, March 2005, pp. 1077–1080.

[10] L. Benaroya, R. Gribonval, and F. Bimbot, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, Hong Kong, 2003, pp. 613–616.

[11] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B.*, vol. 39, pp. 1–38, 1977.

[12] K. K. Paliwal and W. B. Kleijn, "Quantization of lpc parameters," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, pp. 433–466.

[13] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA'03)*, Nara, Japan, Apr. 2003.

[14] http://persos.mist-technologies.com/~rblouet/.