SCALABLE NON-SQUARE BLIND SOURCE SEPARATION IN THE PRESENCE OF NOISE

Radu Balan, Justinian Rosca, Scott Rickard

Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540
{radu.balan,justinian.rosca,scott.rickard}@scr.siemens.com

ABSTRACT

Few source separation and independent component analysis approaches attempt to deal with noisy data. We consider an additive noise mixing model with an arbitrary number of sensors and possibly more sources than sensors (the "degenerate separation problem") when sources are disjointly orthogonal. We show how disjoint orthogonality can be viewed as a limit of a stochastic voice modeling assumption. This is the basis for our approach to noisy model estimation by maximum likelihood, under the direct-path far-field assumptions. The implementation of the derived criterion involves iterating two steps: a partitioning of the time-frequency plane for separation followed by an optimization of the mixing parameter estimates. The solution is applicable to an arbitrary number of microphones and sources. Experimentally, we show the capability of the technique to separate four voices from two, four, six and eight channel recordings in the presence of strong noise.

1. INTRODUCTION

Source separation promises to further a variety of applications of speech enhancement and separation beyond what is possible today with classical microphone array techniques [1]. In particular for audio signals (the domain of interest in this work), a variety of BSS techniques have been introduced in recent years. Few work on real audio data (e.g. [2, 3, 4]), even fewer with noisy data [5], and most deal with the "square" case of source separation (equal number of sources and sensors). Claims of generalization to the non-square case exist, however most often it is not clear how techniques would scale, neither from an algorithmic perspective nor in terms of computational properties.

[6] introduced a BSS technique for the separation of an arbitrary number of sources from just *two* mixtures provided the time-frequency representations of sources do not overlap. The key observation in the technique is that each time-frequency (TF) point depends on at most one source and its associated mixing parameters. This deterministic hypothesis was called *W-disjoint orthogonality* and is reviewed in section 2.2. In anechoic non-noisy environments, it is possible to extract the mixing parameters from the ratio of the TF representations of the mixtures. Using the mixing parameters, one can partition the TF representation of the mixtures to produce the original sources.

The deterministic signal model was extended to a stochastic signal model in [7], where each time-frequency coefficient was modeled as a product between a continuous random variable and a 0/1 discrete Bernoulli random variable (indicating the "presence" of the source). This way signals can be modeled as independent random variables, and one can derive the maximum likelihood (ML) estimator of the mixing parameters.

The ICA literature scarcely discusses the noise case [8]. BSS and deconvolution results of a theoretical nature in dealing with noise were presented in [5]. For the two-channel system in [4], the ML estimator of the mixing parameters was derived in the presence of Gaussian sensor noise. However the noise element represented a technicality in that noise was considered in the limit zero in order to be able to derive parameter update equations. Nonetheless the approach proved effective on real non-noisy data.

In this paper we deal with the multi-channel case from an algorithmic perspective. We present a novel approach to BSS exploiting TF properties of the input data, which is readily applied to speech separation on two, four, six and eight channels. For this, we extend the ML estimators derived before (under the W-disjoint orthogonality assumption). The ML approach considers both mixing parameters and sources, unlike in [4] where the optimization was over mixing parameters only. The estimation algorithm iterates two optimization steps. First, likelihood is optimized over the set of mixing parameters for each source separately. Second the partition of TF points is optimized. For the purposes of this paper we consider the anechoic mixing model only. However the method presented can be extended to arbitrary complex mixing models.

The organization of the paper is as follows. Section 2 presents the signal mixing model and a statistical motivation of the W-disjoint orthogonality signal model. Section 3 shows the derivation of the ML estimator of mixing parameters and source signals, and its implementation by an iterative procedure. Section 4 experimentally highlights the capability of the system to deal with noisy echoic data, and its scaling properties. Experiments with two, four, six and eight inputs show increased separation capability and decreased artifacts with an increase in the number of inputs on data ranging from anechoic to echoic.

2. MIXING MODEL AND SIGNAL ASSUMPTION

2.1. The Mixing Model

Consider the measurements of L source signals by a equispaced linear array of D sensors under far-field assumption where only the direct path is present. In this case, without loss of generality, we can absorb the attenuation and delay parameters of the first

In Proceedings of ICASSP 2003, Hong-Kong China, and Presented at WASPAA, Mohonk NY, October 2003

mixture $x_1(t)$, into the definition of the sources:

$$\begin{aligned} x_1(t) &= \sum_{l=1}^{L} s_l(t) + n_1(t) \\ x_k(t) &= \sum_{l=1}^{L} (1-a_{k,l}) s_l(t-\tau_{k,l}) + n_k(t), \;\; 2 \le k \le D \end{aligned} \quad (1)$$

where $n_1, \ldots, n_D$ are the sensor noises, and $(a_{d,l}; \tau_{d,l})$ are the attenuation and delay parameters of source l to sensor d. For the far-field model and equispaced sensor array, the attenuations $a_{d,l}$ and delays $\tau_{d,l}$ are linearly distributed across the sensors (i.e. with respect to index d). Thus we can define the average attenuation a_l, and delay τ_l, so that

$$a_{d,l} = (d-1)a_l, \;\; \tau_{d,l} = (d-1)\tau_l, \;\; 1 \le d \le D, 1 \le l \le L \quad (2)$$

Clearly other mixing models can be considered at the expense of increasing the model complexity. We use Δ to denote the maximal possible delay between adjacent sensors, and thus $|\tau_l| \le \Delta, \forall l$.

We denote by $X_d(k,\omega)$, $S_l(k,\omega)$, $N_d(k,\omega)$ the short-time Fourier transform of signals $x_d(t)$, $s_l(t)$, and $n_d(t)$, respectively, with respect to a window $W(t)$, where k is the frame index, and ω the frequency index. Then the mixing model (1) turns into

$$X_d(k,\omega) = \sum_{l=1}^{L} (1-(d-1)a_l) e^{-i\omega(d-1)\tau_l} S_l(k,\omega) + N_d(k,\omega) \quad (3)$$

When no danger of confusion, we shall drop the arguments k, ω in X_d, S_l and N_d.

Our problem is: given measurements $(x_1(t), \ldots, x_D(t))_{1 \le t \le T}$ of the system (1) we want to determine the ML estimates of the mixing parameters $(a_l, \tau_l)_{1 \le l \le L}$ and the source signals $(s_1(t), \ldots, s_L(t))_{1 \le t \le T}$. When the number of sources is greater than the number of mixtures the problem is degenerate. In order to solve this we rely on the W-disjoint orthogonality assumption.

2.2. The Stochastic W-Disjoint Orthogonal Signal Model

In [4] we called two signals s_1 and s_2 *W-disjoint orthogonal*, for a given windowing function $W(t)$, if the supports of the windowed Fourier transforms of s_1 and s_2 are disjoint, that is:

$$S_1(k,\omega) S_2(k,\omega) = 0 \;\; , \;\; \forall k, \omega \quad (4)$$

For L sources $S_1, \ldots, S_L$ the assumption generalizes to:

$$S_i(k,\omega) S_j(k,\omega) = 0 \;\; , \;\; \forall\, 1 \le i \ne j \le L, \forall k, \omega \quad (5)$$

Such a deterministic constraint is not only rarely satisfied, but it also implies that the signals are in general statistically dependent[1], which is not the case for voice signals. Yet, in [9] it has been noticed that relation (4) is satisfied in an approximate sense by real speech signals. To reconcile the inconsistent basis for the theoretical development of the algorithm with the fact that the algorithm works in practice, we take a closer look at our model, and show that (4) can be seen as the limit of a stochastic model we introduced in [7].

We briefly review the model and signal class from [7]. It states that the time-frequency coefficient $S(k,\omega)$ of a speech signal $s(t)$ factors as a product of a continuous random variable, say $G(k,\omega)$, and a 0/1 Bernoulli $V(k,\omega)$:

$$S(k,\omega) = V(k,\omega) G(k,\omega) \quad (6)$$

Denoting by q the probability of V to be 1, and by $p(\cdot)$ the p.d.f. of G, the p.d.f. of S turns into

$$p_S(S) = qp(S) + (1-q)\delta(S) \quad (7)$$

with δ, the Dirac distribution. For L independent signals $S_1, \ldots, S_L$, the joint p.d.f. is obtained by conditioning with respect to the Bernoulli random variables. To simplify the notation, we assume all $G(k,\omega)$ have the same distribution $p(\cdot)$, and all $V(k,\omega)$ have the same q. We obtain:

$$\begin{aligned} p(S_1, \ldots, S_L) &= (1-q)^L \prod_{l=1}^{L} \delta(S_l) + q(1-q)^{L-1} \\ &\times \sum_{l=1}^{L} p(S_l) \prod_{j=1, j\ne l}^{L} \delta(S_j) + q^2 \mathrm{Rest}(S_1, \ldots, S_L) \end{aligned} \quad (8)$$

where $\mathrm{Rest}()$ contains terms with the condition that *at least two* sources are active simultaneously. The first two elements in the sum correspond to the condition that *at most one* source is active, which is what is used in the disjoint orthogonality condition.

On the contrary, if we do not assume that at most one source is active but rather approximate (8) when q is very small, by ignoring the q^2 factor and after renormalization we get:

$$\begin{aligned} p_{\mathrm{WDO}}(S_1, \ldots, S_L) &= \frac{1-q}{1+(L-1)q} \prod_{l=1}^{L} \delta(S_l) \\ &+ \frac{q}{1+(L-1)q} \sum_{l=1}^{L} p(S_l) \prod_{j \ne l} \delta(S_j) \end{aligned} \quad (9)$$

This is the stochastic counterpart of the deterministic constraint (4) for L sources. Equation (9) shows that the deterministic constraint on the signals (5) is a reasonable assumption in the stochastic limit, hence the name p_{WDO}. In this paper we do assume the joint p.d.f. of the source signals in the short-time Fourier domain is given by (9), with the interpretation that this is not an inconsistent assumption but rather the limit of a stochastic model.

The second ingredient of our stochastic model is given by the assumption the sensor noises are independently distributed and have Gaussian distributions with zero mean and σ^2 variance.

3. THE MAXIMUM LIKELIHOOD ESTIMATOR OF SIGNAL AND MIXING PARAMETERS

In this section we derive the joint maximum likelihood estimator of parameters and source signals under assumption 5. The source signals naturally partition the time-frequency plane into L disjoint subsets $\Omega_1, \ldots, \Omega_L$, where each source signal is non-zero (i.e. active). Thus the signals are given by the collection $\Omega_1, \ldots, \Omega_L$ and one complex variable S that defines the active signal:

$$S_l(k,\omega) = S(k,\omega) 1_{\Omega_l}(k,\omega) \quad (10)$$

Let the model parameters θ consist of the mixing parameters (a_l, τ_l), $1 \le l \le L$, the partition $(\Omega_l)_{1 \le l \le L}$ and S. Its likelihood and maximum log-likelihood estimator are given by:

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{d=0}^{D-1} \prod_{l=1}^{L} \prod_{(k,\omega)\in\Omega_l} \tfrac{1}{\pi\sigma^2} \exp\{-\tfrac{1}{\sigma^2} |Y_{d,l}(k,\omega)|^2\} \\ \hat{\theta}_{ML} &= \mathrm{argmin}_\theta \sum_{d=0}^{D-1} \sum_{l=1}^{L} \sum_{(k,\omega)\in\Omega_l} |Y_{d,l}(k,\omega)|^2 \end{aligned} \quad (11)$$

[1] Indeed, this is easily proved by the fact that the conditional distribution $p(S_1 = s_1 | S_2 \ne 0) = \delta(s_1)$ is different from the conditional $p(S_1 = s_1 | S_2 = 0)$.

where $Y_{d,l}(k,\omega) = X_{d+1}(k,\omega) - \alpha_{d,l}(\omega) S_l(k,\omega)$ and $\alpha_{d,l}(\omega) = (1 - da_l)e^{-id\tau_l\omega}$. For any partition $(\Omega_1, \ldots, \Omega_L)$ we define the selection map Σ : TF-plane $\rightarrow \{1, \ldots, L\}$, $\Sigma(k,\omega) = l$ iff $(k,\omega) \in \Omega_l$. Clearly Σ defines a unique partition. Optimizing over S in (11) we obtain

$$\hat{S} = \frac{1}{\sum_{d=1}^{D}(1-(d-1)a_l)^2} \sum_{d=1}^{D} \overline{\alpha_{d,l}} X_d \tag{12}$$

where $l = \Sigma(k,\omega)$. Let us denote by $R_l(\omega)$ the D-vector:

$$R_l(\omega) = \left[\begin{array}{cccc} 1 & \alpha_{1,l}(\omega) & \cdots & \alpha_{D-1,l}(\omega) \end{array}\right]^T$$

and by X the D-vector of measurements, $[X_1, \ldots, X_D]^T$. We use $A = (a_l, \tau_l)_{1 \le l \le L}$ to denote the mixing parameters. Inserting (12) into (11), the optimization problem reduces to:

$$(\hat{A}, \hat{\Sigma}) = \text{argmax}_{A,\Sigma} J(A, \Sigma) \tag{13}$$

where:

$$J(A,\Sigma) = \sum_{(k,\omega)} \frac{1}{\|R_{\Sigma(k,\omega)}(\omega)\|^2} |\langle R_{\Sigma(k,\omega)}(\omega), X(k,\omega)\rangle|^2$$

Note the criterion to maximize depends on a set of continuous parameters A, and a selection map Σ. A typical optimization algorithm for such a criterion works as follows. The optimization is done in two steps: first the optimization over the continuous parameters, and then the optimization over the selection map (or, equivalently, the partition). Such a procedure is iterated until the criterion reaches a saturation floor. Because the criterion is bounded above, we are guaranteed it will converge. Next we describe solutions for the two optimization problems.

3.1 Optimal Partition

Given a set of mixing parameters, $A = (a_l, \tau_l)_{1 \le l \le L}$, the optimal selection map is simply given by

$$\hat{\Sigma}(k,\omega) = \text{argmax}_l \frac{1}{\|R_{\Sigma(k,\omega)}(\omega)\|^2} |\langle R_{\Sigma(k,\omega)}(\omega), X(k,\omega)\rangle|^2 \tag{14}$$

The partition is then immediate: $\Omega_l = \{(k,\omega) | \Sigma(k,\omega) = l\}$.

3.2 Optimal Mixing Parameters

Now given a partition $(\Omega_l)_{1 \le l \le L}$, the optimal mixing parameters are obtained independently for each l by:

$$(\hat{a_l}, \hat{\tau_l}) = \text{argmax}_{a_l,\tau_l} \sum_{(k,\omega)\in\Omega_l} \frac{1}{\|R_l(\omega)\|^2} |\langle R_l(\omega), X(k,\omega)\rangle|^2 \tag{15}$$

Expanding the denominator and numerator, we obtain quadratic expressions in a_l. The criterion becomes

$$I(a_l, \tau_l) = \frac{\alpha a_l^2 - 2\beta a_l + \gamma}{\mu a_l^2 - 2\nu a_l + \rho}$$

which can be easily optimized over a_l. We obtain

$$\hat{a_l} = \frac{\mu\gamma - \alpha\rho - \sqrt{(\alpha\rho-\mu\gamma)^2 - 4(\beta\mu-\alpha\nu)(\nu\gamma-\beta\rho)}}{2(\beta\mu-\alpha\nu)} \tag{16}$$

This value should then be inserted in I above and optimization over τ_l should be carried over by a gradient descent, or an exhaustive search (because τ_l is between $-\Delta$ and $+\Delta$):

$$\hat{\tau_l} = \text{argmax}_{\tau_l} \frac{\alpha(\tau_l)\hat{a_l}^2 - 2\beta(\tau_l)\hat{a_l} + \gamma(\tau_l)}{\mu(\tau_l)\hat{a_l}^2 - 2\nu(\tau_l)\hat{a_l} + \rho(\tau_l)} \tag{17}$$

Summing these findings, the optimization algorithm becomes:

3.3 ML Algorithm

- Step 0. Initialize $(a_l^0, \tau_l^0)_{1 \le l \le L}$ with, for instance random values so that $|a_l^0| < 1$ and $|\tau_l^0| < \Delta$; Set $s = 0$, $J^s = 0$, and choose a stopping threshold ϵ;
- Step 1. Find the optimal partition $(\Omega_l^{s+1})_{1 \le l \le L}$, and selection map, Σ^{s+1} by solving (14) with $a_l = a_l^s$, $\tau_l = \tau_l^s$;
- Step 2. Find the optimal parameters $(a_l^{s+1}, \tau_l^{s+1})$ from (16,17) for each $1 \le l \le L$, and subset of time-frequency points Ω_l^{s+1};
- Step 3. Set $s = s + 1$, and compute $J^s = J(A^s, \Sigma^s)$. If $(J^s - J^{s-1})/J^s > \epsilon$ then go to Step 1; otherwise:
- Step 4. The exit values are $a_l = a_l^s$, $\tau_l = \tau_l^s$, and $\Omega_l = \Omega_l^s$, obtained after s iterations. The source signal are then computed by converting the estimated time-frequency representations back into the time domain.

The algorithm can be modified to deal with an echoic mixing model, or different array configurations at the expense of increased computational complexity. It requires knowledge of the number of sources, however this number is not limited to the number of sensors. It works also in non-square case. The algorithm is guaranteed to converge to a local minimum only.

Since we used (9) as the stochastic limit of (8), the signal estimator we derive is *the maximum á posteriori* with respect to the prior joint p.d.f. (9). However, if one adopts the deterministic point of view regarding (5), our estimator is truly the maximum likelihood estimator.

4. EXPERIMENTAL RESULTS

We implemented the algorithm and applied it on realistic synthetic mixtures generated with a ray tracing model. Mixtures consisted of four source signals in different room environments and Gaussian noise. The room size was $4 \times 5 \times 3.2$ m. We used three setups corresponding to anechoic mixing, low echoic (reverberation time 18 ms), and echoic (reverberation time 130 ms). The microphones formed a linear array with 2 cm spacing. Source signals were distributed in the room. Input signals were sampled at 16KHz. For time-frequency representation we used a Hamming window of 256 samples and 50% overlap. Noise was added on each channel. The average (individual) signal-to-noise-ratio (SNR) was 0 dB. The average input signal-to-interference-ratio (SIR) was about -5 dB. Each test was performed three times with independent noise realizations.

The optimization problem (17) was solved through an exhaustive search over a grid of 200 points (thus the precision in estimating τ was roughly 0.005 sample). Experimentally, the optimization algorithm converged very fast. In at most six iterations it reached 10^{-3}% of the local maximum.

To compare results, we used two criteria: output average signal to interference ratio gain (includes other voices and noise) and signal distortion, defined as follows:

$$\text{SIRgain} = \frac{1}{N_f} \sum_{k=1}^{N_f} 10\log_{10}\left(\frac{\| S_o \|^2}{\| \hat{S} - S_o \|^2} \frac{\| X - S_i \|^2}{\| S_i \|^2}\right) \tag{18}$$

$$\text{distortion} = \frac{1}{N_f} \sum_{k=1}^{N_f} 10\log_{10} \frac{\| S_o - S_i \|^2}{\| S_i \|^2} \tag{19}$$

D	$Anechoic$	$LowEchoic$	$Echoic$
2	-4.86 (0.94)	-4.93 (0.80)	-4.84 (1.05)
4	-3.31 (0.95)	-2.88 (0.90	-2.86 (0.89)
6	-3.85 (1.24)	-3.36 (1.02)	-2.99 (1.04)
8	-4.14 (1.28)	-4.14 (1.16)	-2.80 (0.80)

Table 1. Distortions for 0dB input SNR: mean (standard deviation) for $D = 2, 4, 6, 8$.

where: N_f is the number of frames where the summand is above -10 dB for SIR gain, and -30 dB for distortion; $\hat{S}$ is the estimated signal that contains S_o contribution of the original signal; X is the mixing at sensor 1, and S_i is the input signal of interest at sensor 1. The summands were saturated at $+30$ dB for SIR gain and $+10$ dB for distortion. Ideally, $\mathrm{SIRgain}$ should be a large positives, whereas $\mathrm{distortion}$ should be a large negative.

We present results on noisy data for which SNR level (computed for average voice on a channel) is 0 dB. SIR gains are presented in Figure 1 and the distortion values are given in Table 1. Results show separation of all voices particularly for $D \geq 4$ (a sample of input and outputs for $D = 4$ is given in Figure 2. Also SIR gains tend to improve with an increase in the number of sensors. This indicates that separation power of the system increases. Also, one can notice a decrease in performance as we move from anechoic to echoic data. Artifacts as measured by distortion do decrease, with the exception of the two channel case. In that case, separation of all voices does not take place. Three outputs out of four contain merely a mixture of the signals, therefore distortion measures are better at the cost of decreased separation.

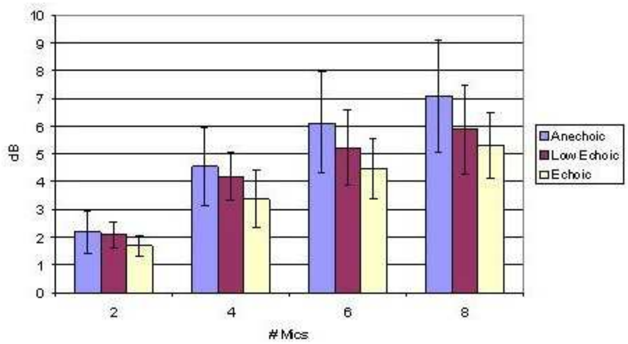

Fig. 1. SIR gains for 2-8 microphones on three data types (anechoic, low echoic and echoic). Each bar includes one standard deviation bounds.

5. CONCLUSIONS

Real source separation scenarios are rarely square. On the contrary, situations constantly vary between the so called degenerate case and the over specified case. By being able to deal evenly with such cases and in the presence of noise, the present approach opens the door to audio source separation in realistic scenarios.

This was possible by exploiting the time frequency sparseness of signals. We showed that disjointness in time-frequency, although inconsistent theoretically for a deterministic model, is justifiable from a stochastic perspective. We modeled each time-frequency coefficient as a product between a continuous random variable and a discrete 0/1 Bernoulli random variable. In the limit this corresponded to the deterministic W-disjoint orthogonality model as studied in [9].

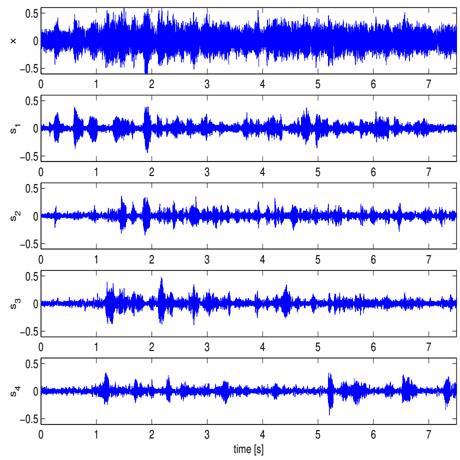

Fig. 2. Example of 4-channel algorithm behavior on mixture of noise and four voices (-8.5 and -3.5 dB input SIR). The separated outputs show an SIR gain of 7, 4, 5.3 and 9.5 dB respectively.

Our source separation algorithm implements the maximum likelihood estimator for both mixing parameters and source signals under a direct-path mixing model and for a linear array of sensors. We presented an iterative procedure to optimize the likelihood, similar in spirit to hybrid optimization algorithms. It is worthy to outline the nice scaling properties of the approach both algorithmically and experimentally. The former refers to scalability in the number of inputs (here we used two, four, six and eight microphone linear arrays). The latter views the increased separation power and decreased artifacts with an increase in the number of inputs on echoic data.

Future work could address the question whether anything is to be gained by considering an echoic model. This extension is naturally feasible in this approach.

6. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[2] L.Parra, "Convolutive blind source separation based on multiple decorrelation," 1997, vol. IEEE-ICNN.

[3] J.Annemuller and B.Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *ICA*, 2000, pp. 215–220.

[4] S. Rickard, R. Balan, and J. Rosca, "Real-time blind source separation using DUET," in *Proc. 3rd ICA-BSS*, 2001.

[5] E. Moulines, J.F. Cardoso, and E. Gassiat, "Ml for bss and deconvolution of noisy signals using mixture models," in *ICASSP*, 1997, pp. 3617–3720.

[6] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *ICASSP*, 2000.

[7] R. Balan and J. Rosca, "Statistical properties of STFT ratios for two channel systems and applications to blind source separation," in *Proc. ICA-BSS*, 2000.

[8] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley and Sons, 2001.

[9] S. Rickard and O. Yilmaz, "On the W-disjoint orthogonality of speech," in *ICASSP*, 2002.