A Stochastic Speech Model Supporting W-Disjoint Orthogonality

Radu Balan, Justinian Rosca, Scott Rickard
Siemens Corporate Research, Princeton, NJ, USA
e-mail: rvbalan,rosca,rickard@scr.siemens.com

Abstract — **In previous work, we have successfully used an ideal joint sparseness assumption: W-Disjoint Orthogonality (WDO). This assumption, that the time-frequency representations of the sources have disjoint support, is satisfied in an approximate sense by many signals of practical interest, including speech. Here we discuss results derived from a stochastic model of speech signals that justify the WDO hypothesis. If the magnitude of the time-frequency components of the source signals have Laplacian priors, a subset of their maximum á posteriori (MAP) estimates are guaranteed to satisfy the WDO assumption.**

I. Introduction

Speech is sparse. This sparseness has been exploited in the ICA-BSS community for parameter estimation and source separation (e.g., [1],[2],[3]). A time-frequency (TF) sparseness assumption was introduced in [4] and subsequently used in [5] and [6] which allows for the separation of more than two sources given just two mixtures. This sparseness property, called *W-disjoint orthogonality* (WDO), assumes that the signals have non-overlapping TF representation supports. Given source TF representations $S_1(\omega,t),\ldots,S_N(\omega,t)$, the WDO assumption can be stated,

$$S_i(\omega,t)S_j(\omega,t)=0,\ \forall i\neq j, \forall(\omega,t). \tag{1}$$

This assumption has been shown to be approximately true for speech signals [7]. In [6] we argued that WDO is approximately satisfied when one assumes a signal model of the form:

$$S(\omega,t)=B(\omega,t)G(\omega,t) \tag{2}$$

where $B(\omega,t)$ is a Bernoulli random variable (i.e. it takes a value of only 0 or 1), and $G(\omega,t)$ is a continuously distributed random variable. It follows that the joint distribution is:

$$\begin{aligned} p_{S_1,S_2}(S_1,S_2) &= (1-q)^2\delta(S_1)\delta(S_2)+q(1-q)(\delta(S_1)p(S_2) \\ &\quad +\delta(S_2)p(S_1))+q^2p(S_1)p(S_2) \end{aligned} \tag{3}$$

The purpose of this note is to point out that (1) and (3) follow as consequences of a more general stochastic model.

II. The Stochastic Model and Main Results

Our model is based on the following assumption: TF coefficients of speech signals are independent and have Laplace distributed priors. Furthermore, given a mixture, one cannot distinguish between the true input signals and their maximum à posteriori estimates. Hence, given empirical distributions of the measured mixtures, the only inference about the true distribution of source signals is given by the distribution of the MAP estimates.

Assume two signals $s_1(\cdot),s_2(\cdot)$ are mixed by a known convolutive model and measured together with some measurement noise. In the TF domain, the mixing model becomes:

$$\begin{aligned} X_1(\omega,t) &= S_1(\omega,t)+S_2(\omega,t)+N_1(\omega,t) \\ X_2(\omega,t) &= H_1(\omega)S_1(\omega,t)+H_2(\omega)S_2(\omega,t)+N_2(\omega,t) \end{aligned} \tag{4}$$

Assuming N_1,N_2 are Gaussian distributed, the MAP estimates of S_1,S_2 becomes:

$$(\hat{S}_1,\hat{S}_2)=\text{argmin}_{S_1,S_2}|X_1-S_1-S_2|^2+|X_2-H_1S_1-H_2S_2|^2+\lambda_1|S_1|+\lambda_2|S_2| \tag{5}$$

where λ_1 and λ_2 depend on the prior variance of the two signals and the variance of the noise. The problem (5) seems not to have a closed form solution. In fact, the higher dimensional equivalent problem has been the focus of many other papers, and recently an algorithm to solve it in a wavelet basis has been proposed (see [8]). Instead we will show the behavior of the solution. We state here, without proof, the two main results:

Theorem 1 *Assume $H_1\neq H_2$. Then $\exists r>0$ such that for all X_1,X_2, $|X_1|<r$, $|X_2|<r$, the solution of (5) satisfies:*

$$\hat{S}_1\hat{S}_2=0 \tag{6}$$

Theorem 1 states that the MAP estimates of the sources for all time-frequency components with magnitude smaller than some threshold r satisfy the WDO assumption.

Theorem 2 *Assume $H_1\neq H_2$. Given the joint empirical distribution of X_1,X_2, the empirical dist. of $(\hat{S}_1,\hat{S}_2)$ factors:*

$$\begin{aligned} p_{\hat{S}_1,\hat{S}_2}(S_1,S_2) &= (1-q_1)(1-q_2)\delta(S_1)\delta(S_2)+q_1(1-q_2)p_1(S_1)\delta(S_2) \\ &\quad +(1-q_1)q_2\delta(S_1)p_2(S_2)+q_1q_2p(S_1,S_2) \end{aligned} \tag{7}$$

Theorem 2 states that the empirical MAP joint distribution takes the same form as the joint distribution based on the Bernoulli TF model (Equation (3)). Thus, both (1) and (3) follow from the MAP source estimators of mixtures of Laplacian distributed sources. Future work will focus on (a) the relationship between r and the random variable parameters (we hope that r is indeed large enough to contain considerable source energy) (b) the extension to arbitrary mixing dimensions, and (c) continuing efforts to leverage (1) and (3) to provide closed form BSS algorithms rather than iterative procedures.

References

[1] J. Huang, N. Ohnishi, N. Sugie. "A biomimetic system for localization and separation of multiple sound sources", IEEE Trans. on Inst. and Measurement, 44(3):733–738, June 1995.

[2] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, Y. Kaneda, "Sound source segregation based on estimating incident angle of each freq. comp. of input signals acquired by multiple microphones", Acoust. Sci. Tech., 22(2), 149–157, 2001.

[3] M. Zibulevsky , B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary", Neural Computation, 13(4):863–882, 2001.

[4] A. Jourjine, S. Rickard, O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures", 2985–2988, ICASSP 2000.

[5] S. Rickard, R. Balan, J. Rosca, "Real-time time-frequency based blind source separation", 651–656, ICA 2001.

[6] R. Balan, J. Rosca, S. Rickard, "Scalable non-square blind source separation in the presence of noise", ICASSP 2003.

[7] S. Rickard, O. Yilmaz. "On the approximate W-disjoint orthogonality of speech", 529–532, ICASSP 2002.

[8] I. Daubechies, M. Defrise, C. De Mol, "An iterative algorithm for ill-posed inverse problems where the object has a sparse wavelet expansion", AMS Meeting, Baltimore January 2003.