Source Separation of Speech Signals using Kurtosis Maximization

James P. LeBlanc *Phillip L. De Leòn*
leblanc@nmsu.edu pdeleon@nmsu.edu
Klipsch School of ECE
New Mexico State University
Las Cruces, NM USA

Abstract

We present a computationally efficient method of separating mixed speech signals using a recursive adaptive gradient descent technique. The cost function is designed to maximize the kurtosis of the output (separated) signals. The choice of kurtosis maximization as an objective function (which acts as a measure of separation) is supported by investigation and analysis of *spherically invariant random processes* (SIRP's) [6].

Introduction

The problem of separation of speech signals is considered. Making some mild assumptions on the statistics of the voice signals we use higher order statistics to separate the voices. The use of higher-order statistics is not new to the source separation problem (see [1], [5], for example). But, many of these methods are applied to digital communications signals which belong to a different statistical class (e. g. sub-Gaussian) [1] than speech signals (super-Gaussian).

A fundamental idea of many blind separation and equalization schemes in digital communications makes note that the sum of sub-Gaussian processes (as occurs with mixing and intersymbol interference) results in a process that "looks more" Gaussian than the originals. [3] includes an excellent discussion of measures of *Gaussianity*. With such a measure, one constructs a cost function, and associated adaptive gradient descent algorithm which minimizes this Gaussianity measure resulting in source separation or intersymbol interference reduction. A common measure which appears quite often is *kurtosis*, which is defined for a zero mean random process X as $\kappa_{\mathrm{X}} = \mathrm{E}\{x^4\}/\{\mathrm{E}\{x^2\}\}^2$ Kurtosis relates to the Constant Modulus Algorithm (CMA) [4] used for blind equalization. Here, we modify a CMA-based source separation algorithm [2] by adjusting for the differing statistics (i.e. super-Gaussian) of voice signals.

Problem Setting

The generic two signal separation problem is shown in Figure 1. Two sources s_0 and s_1 are mixed through mixing matrix $\mathbb{A}$, resulting in received signals x_0 and x_1. The mixing relation is denoted, $X = \mathbb{A}S$ where $X = [\,x_0 \quad x_1\,]^t$ and $S = [\,s_0 \quad s_1\,]^t$.

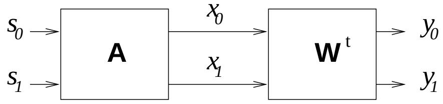

Figure 1: Separation Block Diagram

The goal is to separate the s_0 and s_1 components present in the mixed signals yielding $Y = [\,y_0 \quad y_1\,]^t$ through the use of matrix $\mathbb{W}^t$. Clearly, $\mathbb{W}^t = {\mathbb{A}^{t}}^{-1}$ achieves the desired result (assuming $\mathbb{A}$ is invertible) but, $\mathbb{A}$ is typically unknown and $\mathbb{W}$ (or $\mathbb{A}$) must be estimated only using the mixture X.

Separation by Kurtosis

Communications Signals. An interesting feature of kurtosis is now noted. Let u_0 and u_1 be two independent, identically distribute (iid), zero mean random variables with kurtosis κ_{U}. Let $w = u_0 + u_1$ and consider κ_{W}. It can be shown that

$$\begin{aligned} \kappa_{\mathrm{U}} < 3 &\Rightarrow \kappa_{\mathrm{W}} > \kappa_{\mathrm{U}} \text{ (sub-Gaussian)} \\ \kappa_{\mathrm{U}} > 3 &\Rightarrow \kappa_{\mathrm{W}} < \kappa_{\mathrm{U}} \text{ (super-Gaussian)} \end{aligned} \qquad (1)$$

Since digital communications signal are typically considered to be *sub-Gaussian*, the resulting mixture will have a higher kurtosis. Thus, a logical separation strategy is to minimize the output kurtosis, which in effect, is exactly what CMA does. In [2] an iterative separation algorithm from digital communications signals utilizing the CMA error function is presented as

$$\mathbb{W}_{n+1} = \mathbb{W}_n - \mu \bigtriangledown_{\mathbb{W}} (\phi(\mathbb{W})) \qquad (2)$$

where μ is the small adaptive stepsize, and $\bigtriangledown_{\mathbb{W}}\phi(\mathbb{W})$ denotes the gradient of ϕ,

$$\phi(\mathbb{W}) = \sum_{i=1}^{N} \mathrm{E}\left\{(y_i^2 - 1)^2\right\} - \ln(\det|\mathbb{W}|) \quad (3)$$

The first term is the CMA cost function, while the second term associates a cost to duplicating a source at the output Y.

[1] The term *sub-Gaussian* (*super-Gaussian*) is used to denote processes having a kurtosis less (more) than the kurtosis of a Gaussian.

In light of (1), such kurtosis minimization agrees with source separation.

Speech Signals. We adopt a kurtosis-based strategy for separating speech signal by recognizing that speech signal are *super-Gaussian.* In light of (1), we choose the adaptation objective to be kurtosis maximization. The adaptive algorithm becomes (ignoring for the moment the desire to prevent duplicate sources at output),

$$\mathbb{W}_{n+1} \;=\; \mathbb{W}_n + \mu \bigtriangledown_{\mathbb{W}} (\kappa_{\mathrm{Y}}(\mathbb{W})) \qquad (4)$$

where μ is the small adaptive stepsize, and $\bigtriangledown_{\mathbb{W}} \kappa_{\mathrm{Y}}(\mathbb{W})$ denotes the gradient of the kurtosis of the outputs Y. For the two channel case, performing the differentiation leads to the update law

$$\mathbb{W}_{n+1} = \mathbb{W}_n + \mu \begin{bmatrix} -\alpha_1\beta_1\gamma_1 w_{21} & -\alpha_2\beta_2\gamma_2 w_{22} \\ \alpha_1\beta_1\gamma_1 w_{11} & \alpha_2\beta_2\gamma_2 w_{12} \end{bmatrix}$$

where $\alpha_i = 4(w_{i1}x_1 + w_{2i}x_2)^3$ and

$$\begin{aligned} \beta_i &= (-x_1 w_{i1} r_{12} - x_1 w_{2i}\sigma_2^2 + x_2 w_{1i}\sigma_1^2 + \\ &\qquad w_{2i} x_2 r_{12}) \\ \gamma_i &= 1/(w_{i1}^2\sigma_1^2 + 2w_{i1}w_{2i}r_{12} + w_{2i}^2\sigma_2^2)^3 \end{aligned}$$

and $\sigma_1^2 = \mathrm{E}\{y_1^2\}$, $\sigma_2^2 = \mathrm{E}\{y_2^2\}$, $r_{12} = \mathrm{E}\{y_1 y_2\}$. Knowing the actual values of σ_1^2,σ_2^2, and r_{12} a priori is not possible but may be replaced by simple autoregressive estimators. Also, an output scaling factor must be incorporated into the algorithm, since kurtosis is a scale invariant quantity.

The critical assumption here is that the kurtosis of two mixed voice signals has a lower kurtosis than the individual kurtosis values (as *hinted* at by (1)). However, this may not be true. For speech signals there is no assurance different speakers have identical distributions, nor are the samples from any speaker temporally independent. While lacking a proof, this critical assumption holds for all sampled speech we've tested. This issue has been initially investigate using SIRP's are a good statistical speech model. Under our analysis, the critical assumption holds for a wide range of parameters representing speech. While lending credence to this approach, further study is on-going.

Separation Example

The algorithm is demonstrated for

$$\mathbb{A} \;=\; \begin{bmatrix} 0.72 & 0.34 \\ 0.41 & 0.63 \end{bmatrix},$$

s_0 is a speaker with $\kappa_{\mathrm{s}_0} = 13.9$ and s_1 is a different speaker $\kappa_{\mathrm{s}_1} = 13.1$. The received mixtures have kurtosis $\kappa_{\mathrm{x}_0} = 11.4$ and $\kappa_{\mathrm{x}_1} = 8.8$ verifying the critical assumption. The severity of the mixing renders the signals x_0 and x_1 unintelligible.

Figure 2 plots the power ratios of the s_0 and s_1 components in both y_0 and y_1 (a measure of the separation at the output). The achieved separations exceed 35dB. Qualitatively, listening to the resulting separated signals, the second speech signal was virtually imperceptible.

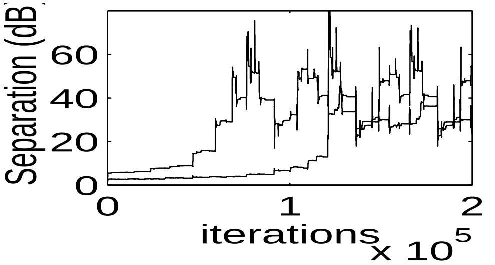

Figure 2: Example of Source Separation

Conclusion

We adopted ideas from the blind source separation/equalization problem and modified them for use in speech separation, presenting an algorithm and its motivation using the concept of kurtosis maximization with leptokurtic signals. The algorithm, originally based on heuristics, appears more plausible through use of analysis of SIRP's as statistical speech models. Continuing work on the analysis of the SIRP model, convergence analysis, and lowered computational complexity of the presented algorithm are all avenues for continuing work.

References

[1] E. Moreau, O. Macchi, "New self-adaptive algorithm for source separation based on contrast functions," *Proc. IEEE Sig. Proc. Wrkshp on Higher Order Stat.*, Lake Tahoe, CA, 1993, pp.215-219.

[2] L. Castedo, O. Macchi, "Maximizing the Information Transfer for Adaptive Unsupervised Source Separation," *Proc. IEEE Sig. Proc. Adv. in Wireless Comm. Wrkshp*, Paris, 1997, pp.65-69.

[3] D.L. Donoho, "On Minimum Entropy Deconvolution," *Applied Time Series Analysis*, D.F. Findley, Ed., New York: Academic Press, 1981.

[4] J.R. Treichler, M. G. Agee, "A New Approach to Multipath Correction of Constant Modulus Signals," *IEEE Trans. Acous., Speech, and Sig. Proc.*, Apr. 1983.

[5] J. -F. Cardoso, "Source separation using higher order moments," *Proc. ICASSP 89*, Glasgow, Scotland, May 1989, vol. 4, pp.2109-2112.

[6] Brehm, H. and Stammler, W., "Description and generation of spherically invariant speech-model signals," *Sig. Proc., vol. 12, no. 2, pp. 119-141*, Mar. 1987.