Source Separation of Speech Signals using Kurtosis Maximization

James P. LeBlanc    Phillip L. De León
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 (SIRPs) [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) 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 $\kappa_X = \frac{\text{E} \{ x^4 \} / \text{E} \{ x^2 \}^2}{\text{E} \{ x^2 \}^2}$ 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 $A$, resulting in received signals $x_0$ and $x_1$. The mixing relation is denoted, $X = A S$ where $X = [x_0 \ x_1]^t$ and $S = [s_0 \ s_1]^t$.

![Figure 1: Separation Block Diagram](image)

The goal is to separate the $s_0$ and $s_1$ components present in the mixed signals yielding $Y = [y_0 \ y_1]^t$ through the use of matrix $W^t$. Clearly, $W^t = A^{-1}$ achieves the desired result (assuming $A$ is invertible) but, $A$ is typically unknown and $W$ (or $A$) 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 distributed (iid), zero mean random variables with kurtosis $\kappa_u$. Let $w = u_0 + u_1$ and consider $\kappa_w$. It can be shown that

$$\kappa_u < 3 \Rightarrow \kappa_w > \kappa_u \text{ (sub-Gaussian)}$$
$$\kappa_u > 3 \Rightarrow \kappa_w < \kappa_u \text{ (super-Gaussian)}$$

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

$$W_{n+1} = W_n - \mu \nabla_{W_n} (\phi(W))$$

where $\mu$ is the small adaptive stepsize, and $\nabla_{W_n} \phi(W)$ denotes the gradient of $\phi$,

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

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

---

1The 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),

$$W_{n+1} = W_n + \mu \nabla_{W} (\kappa_Y(W))$$

where \(\mu\) is the small adaptive stepsize, and \(\nabla_{W} \kappa_Y(W)\) denotes the gradient of the kurtosis of the outputs \(Y\). For the two channel case, performing the differentiation leads to the update law

$$W_{n+1} = W_n + \mu \left[ -\alpha_1 \beta_1 \gamma w_{11} - \alpha_2 \beta_2 \gamma w_{22} \right]$$

$$+ \alpha_1 \beta_1 \gamma w_{11} + \alpha_2 \beta_2 \gamma w_{12}$$

where \(\alpha_i = 4(w_{1i} x_{1i} + w_{2i} x_{2i})^3\) and

$$\beta_i = (-x_1 w_{11} r_{12} - x_1 w_{12} \sigma_2^2 + x_2 w_{11} \sigma_1^2 +$$

$$w_{12} x_{12} r_{12})$$

$$\gamma_i = 1/(w_{11}^2 \sigma_1^2 + 2 w_{12} w_{22} r_{12} + w_{22}^2 \sigma_2^2)$$

and \(\sigma_1^2 = E\{y_1^2\}\), \(\sigma_2^2 = E\{y_2^2\}\), \(r_{12} = E\{y_1 y_2\}\). Knowing the actual values of \(\sigma_1^2, \sigma_2^2\), 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

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

\(s_0\) is a speaker with \(\kappa_{s_0} = 13.9\) and \(s_1\) is a different speaker \(\kappa_{s_1} = 13.1\). The received mixtures have kurtosis \(\kappa_{s_0} = 11.4\) and \(\kappa_{s_1} = 8.8\) 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.

**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**


