SPEECH SEPARATION BY KURTOSIS MAXIMIZATION

James P. LeBlanc *Phillip L. De Leòn*

Klipsch School of ECE
New Mexico State University
Las Cruces, NM USA

ABSTRACT

We present a computationally efficient method of separating mixed speech signals. The method uses a recursive adaptive gradient descent technique with the cost function 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 experiments with a number of speech signals as well as *spherically invariant random processes* (SIRP's) which are regarded as excellent statistical models for speech. Development and analysis of the adaptive algorithm is presented. Simulation examples using actual voice signals are presented.

1. INTRODUCTION

We address the speech separation problem. Making some assumptions on the statistics of the voice signals we use higher order statistics to separate mixed voices. The use of higher-order statistics is not new to the source separation problem, see ([2] [1], [9], [1], [8] for example). Many of these methods are applied to digital communications signals which inherently belong to a different statistical class than speech signals. Specifically, many such adaptive algorithms used on digital communications signals use the fact that the sequences are *sub-Gaussian* [1]. We note that there are separtation algorithms designed for use with speech signals (notably [11] and [10]). Some methods are computationally intensive, necessitating correlation matrix estimation and eigen-decompositions or polyspectra estimations. As an alternative, the speech separation developed herein is designed to be computationally efficient. We exploit the fact that speech signals are *super-Gaussian* (i.e. have high kurtosis). Noting this, we may adopt similar strategies using higher order statistics after appropriate modification.

A fundamental idea of many blind separation and blind equalization schemes in the digital communications setting is to note that the sum of sub-Gaussian processes (as occurs with mixing and intersymbol interference) results in a process with statistics that differ from the original process(es). More specifically, the mixture "looks more" Gaussian than the originals. In [7] one finds an excellent discussion of developing measures of *Gaussianity*. With such a measure, one may construct a cost function, and associated adaptive gradient descent algorithm which minimizes this Gaussianity and results in source separation or intersymbol interference reduction (for the sub-Gaussian digital communications signals). Many possible measures are possible. One which appears quite often is *kurtosis*.

[1]The term *sub-Gaussian* may have different meanings among different communities. Here it is used to denote processes having a kurtosis less than the kurtosis of a Gaussian.

The relation of kurtosis to one of the the more popular blind equalization algorithm known as the Constant Modulus Algorithm (CMA) [8] or Godard Algorithm [6] is discussed in [5]. For the separation of speech signals, we modify a source separation algorithm recently proposed for digital communications (sub-Gaussian) signals in [3] which uses the CMA (or Godard) error function. This modification adjusts for the differing statistics between digital communication signals and voice signals.

2. 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,

$$X = \mathbb{A}S \tag{1}$$

where $X = [\, x_0 \quad x_1 \,]^t$ and $S = [\, s_0 \quad s_1 \,]^t$.

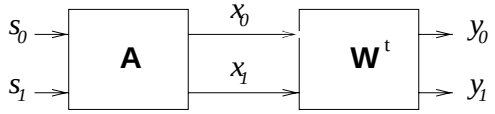

Figure 1. Separation Block Diagram

The goal is to separate out the s_0 and s_1 components present in the mixed signals x_0 and x_1 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. In the blind problem, $\mathbb{W}$ (or similarly $\mathbb{A}$) must be estimated from knowledge only of the mixture X. The second order statistics of X (i. e. the autocorrelation matrix $\mathbb{R}_{XX}$) do not provide enough information. For this reason higher order statistics are often considered. However, the use of higher order statistics often requires further assumptions on the distributions of S. In the blind separation problem (as well as the blind equalization problem), the source distributions are typically considered to be sub-Gaussian in the sense that their kurtosis is below that of a Gaussian. The kurtosis [2] of a zero mean random variable x is defined as the dimensionless, scale invariant quantity,

$$\kappa_{\rm x} = \frac{\mathrm{E}\left\{x^4\right\}}{\{\mathrm{E}\left\{x^2\right\}\}^2} \tag{2}$$

where $\mathrm{E}\{\bullet\}$ is the expectation operator. For any random variable we have $\kappa \geq 1$, and for a Gaussian distribution $\kappa =$

[2]The reader is cautioned that some texts define kurtosis a bit differently as $\kappa_{\rm x} = \frac{\mathrm{E}\{x^4\}}{(\mathrm{E}\{x^2\})^2} - 3$. We shall however follow the definition above as found in [13].

3. Distributions with $\kappa < 3$ are considered sub-Gaussian (or platykurtic), and those with $\kappa > 3$ are labelled as super-Gaussian (or leptokurtic).

3. SEPARATION BY KURTOSIS

3.1. 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 κ_{W} is always closer to 3, than κ_{U}. More specifically,

$$\text{if } \kappa_{\text{U}} < 3 \text{ (platyurtic), then } \kappa_{\text{W}} > \kappa_{\text{U}} \tag{3}$$

$$\text{if } \kappa_{\text{U}} > 3 \text{ (leptokurtic), then } \kappa_{\text{W}} < \kappa_{\text{U}} \tag{4}$$

Since, digital communications signal are typically leptokurtic. Given two iid sources, the resulting mixture will have a higher kurtosis. Thus, a logical separation strategy is to minimize the kurtosis, which in effect, is exactly what CMA does. In [3] an iterative separation algorithm from digital communications signals utilizing the CMA error function is presented as

$$\mathbb{W}_{n+1} \quad = \quad \mathbb{W}_n - \mu \bigtriangledown_{\mathbb{W}} (\phi(\mathbb{W})) \tag{5}$$

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

$$\phi(\mathbb{W}) = \sum_{i=1}^{N} \text{E}\left\{(y_i^2 - 1)^2\right\} - \ln(\det|\mathbb{W}|) \tag{6}$$

in which the first term is as the CMA cost function, while the second term associates a cost to duplicating a source at the output Y. The existence of the CMA cost term can be associated with a gradient descent algorithm performing a kurtosis minimization on the output Y. In light of (3), such kurtosis minimization agrees with source separation.

3.2. Speech Signals

We adopt a kurtosis-based strategy for separating speech signal by observing that speech signal are *leptokurtic* (in contrast to the typcially platykurtic communications signals). In light of (4), 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),

$$\mathbb{W}_{n+1} \quad = \quad \mathbb{W}_n + \mu \bigtriangledown_{\mathbb{W}} (\kappa_{\text{Y}}(\mathbb{W})) \tag{7}$$

where μ is the small adaptive stepsize, and $\bigtriangledown_{\mathbb{W}}\kappa_{\text{Y}}(\mathbb{W})$ 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

$$\begin{aligned} \alpha_i &= 4(w_{i1}x_1 + w_{2i}x_2)^3 \\ \beta_i &= (-x_1 w_{i1} r_{12} - x_1 w_{2i}\sigma_2^2 + x_2 w_{1i}\sigma_1^2 + \\ &\quad w_{2i}x_2 r_{12}) \\ \gamma_i &= 1/(w_{i1}^2\sigma_1^2 + 2w_{i1}w_{2i}r_{12} + w_{2i}^2\sigma_2^2)^3 \\ W &= \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \end{aligned}$$

and $\sigma_1^2 = \text{E}\left\{y_1^2\right\}$, $\sigma_2^2 = \text{E}\left\{y_2^2\right\}$, $r_{12} = \text{E}\left\{y_1 y_2\right\}$. In true implementation, knowing the actual values of σ_1^2,σ_2^2, and r_{12} a priori is not possible but these may be replaced by simple autoregressive estimators of the form

$$\begin{aligned} \hat{\sigma}_i^2 &= \lambda\hat{\sigma}_i^2 + (1-\lambda)y_i(k)^2 \\ \hat{r}_{12} &= \lambda\hat{r}_{12} + (1-\lambda)y_1(k)y_2(k) \end{aligned}$$

where λ is the estimator's forgetting factor. It should also be noted that a scaling factor is 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 (4)). However, in the strict sense, (4) is not necessarily always true for speech signals due to the temporal correlation of speech. There is much evidence derived from studies with sampled speech that this relation in (4) often holds true. This issue is addressed in the next section.

4. SPEECH, SIRPS, AND MIXTURE KURTOSES

The underlying assumption for the proposed speech signal separation technique is that mixtures of speech signals have a kurtosis lower than the kurtosis values of the individual speech signals. In this section we offer evidence which supports this assumption based on actual speech signals and statistical models for speech.

Eight individual speech signals were sampled and the kurtoses computed. In addition, the kurtosis of a mixture ($x_0 = as_0 + (1-a)s_1$) of the speech signals was also computed. The results are shown in Figure 2 (solid line). We note that for the 50%-50% mixture (a = 0.5), the individual speech signals are higher in kurtosis than the mixture for 93% of the studied cases. This indicates that the speech signals will begin to separate from the mixture based on the kurtosis maximization algorithm. As separation proceeds, some of these mixtures may cease further separation (as defined by power ratios), as indicated by the lower probability of both indivudual kurtosis being higher than the mixture kurtosis (Figure 2). However, we have observed in our experiments that there is always one speech signal that is higher in kurtosis than the mixture, indicating that at least this one speech signal could be separated from the mixture; the remaining speech signal might be subsequently separated from residual signal analysis and processing.

Historically, speech probability density functions (PDFs) have been modeled with either the Laplace or Gamma PDF [14] [15]. For these two PDFs, it can be shown that the kurtosis of the sum of these PDFs is lower than the kurtosis values of the individual PDFs. More recently, refinement in the speech model has been achieved through the use of Spherically Invariant Random Processes (SIRP's) also known as circularly or spherically symmetric random processes [16].

The use of SIRP's to model speech signals is based on the facts that many random processes are SIRP's including those with Laplace and Gamma PDFs and that actual speech bivariate PDFs (three dimensional amplitude histogram taken from samples of speech signals spaced $t < 5$ms apart) have been shown to exhibit SIRP-like qualities [18], [19], [17]. Based on work by Brehm, the SIRP which includes the Gaussian, Laplace, Gamma, and K_0 PDF's, was found to be especially suited to modeling the measured densities of speech signals. By continuous variation of b_1 and b_2, a whole family of modeling PDF's may be generated, one of which may be a better approximation to the speech PDF than Laplace or Gamma PDFs. Brehm and Stammler found that speech PDFs can be closely matched by choosing

$$-0.4 \leq b_1 \leq -\frac{1}{3} \tag{8}$$

and

$$b_2 \geq 0.25 \tag{9}$$

Furthermore, when the time shift between the two speech signals used to compute the bivariate speech PDF is less than 3.8 ms, the bivariate PDF corresponding to SIRP fits the observed contour lines in the bivariate PDF very well [17].

Five SIRPs were modeled according [17] and the Kurtoses computed. In addition the kurtosis of a mixture ($a\text{SIRP}_X + (1-a)\text{SIRP}_Y$) of the SIRP's was also computed. The results are shown in Figure 2 (dashed line). In all 50%-50% mixture cases, $\kappa_{s_0}, \kappa_{s_1} > \kappa_{x_i}$ the individual SIRP's are higher in kurtosis than the mixture meeting the assumed necessary conditions for the proposed algorithm.

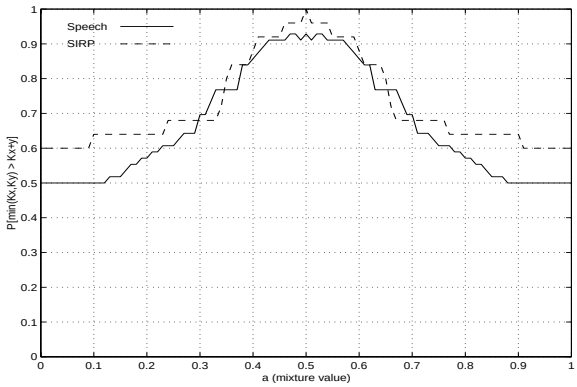

Figure 2.

5. SEPARATION EXAMPLES

In this section the performance of the algorithm in (8) is demonstrated using two separtation examples (case I and case II). Here, the mixing matrices are arbitrarily chosen.

5.1. Case I

$$\mathbb{A}_I = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}, \tag{10}$$

s_0 and s_1 are sampled speech of a male speakers reading different text segments of a text book with $\kappa_{s_0} = 14.9$ and $\kappa_{s_1} = 14.2$. The received mixtures x_0 and x_1 have kurtosis values of $\kappa_{x_0} = 11.6$ and $\kappa_{x_1} = 11.1$ verifying the assumption on mixed speech signals having lower kurtosis. The severity of the mixing renders to signals x_0 and x_1 unintelligible.

stepsize	$\mu = 10^{-6}$
moment est. forgetting factor	$\lambda = 0.999$
separation matrix initialization	$\mathbb{W}(0) = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$
initial moment estimates	$\hat{r}_{12}(0) = 0.1$
	$\hat{\sigma}_1(0) = 0.1$
	$\hat{\sigma}_2(0) = 0.1$

Table 1. Example Simulation Parameters

Plotted below are the power ratios of the s_0 and s_1 components in both y_0 and y_1 which provides a measure of the separation of the source components are the output. The adaptation parameters are found in Table 1.

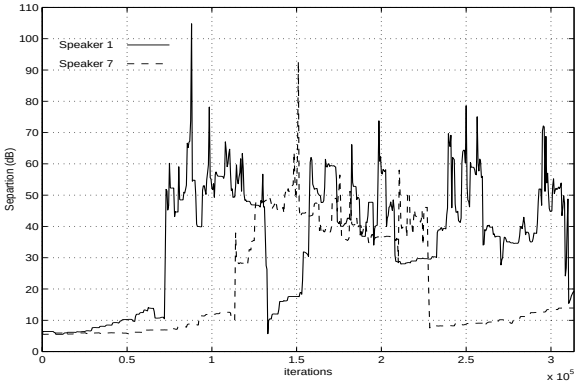

Figure 3. Case I of Source Separation using Algorithm

The achieved separation is very good, exceeding 40dB on both channels at times. Qualitatively, listening to the resulting separated signals, the second speech signal was virtually imperceptible.

5.2. Case II

Here s_0 and s_1 are different speakers reading text segments of a text book with $\kappa_{s_0} = 27.9$ and $\kappa_{s_1} = 11.3$. The mixing matrix

$$\mathbb{A}_{II} = \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix}, \tag{11}$$

is used resulting in received mixtures x_0 and x_1 have kurtosis values of $\kappa_{x_0} = 24.9$ and $\kappa_{x_1} = 20.8$. Note here that one mixture violates the assumption that the necessary assumption and only one source is separted. This is thought to be due to the widely disparate source kurtoses.

Again, the power ratios of source components are plotted using the algorithm with the same parameters. The achieved separation on one output is very good, again exceeding 40dB on one channel. However, at the other output channel little separtation is acheived (about constant at 10dB). But, since excellent separation has been acheived at

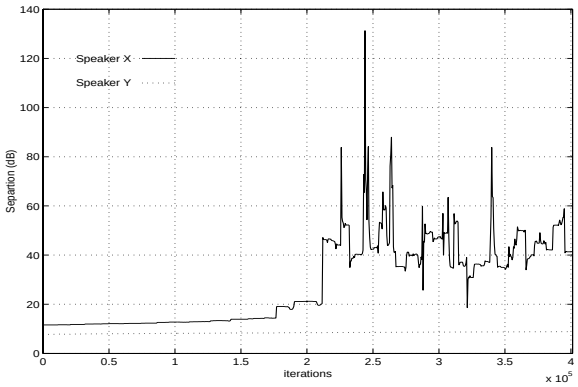

Figure 4. Case II of Source Separation using Algorithm

the other output, it is possible to perform further processing using residual signal analysis.

6. CONCLUSION

We have adopted some ideas and results from the digital communications community used in the blind source separation and blind equalization problem and modified them for use in the speech separation problem. The presentation of an algorithm along with its motivation is described through the concept of kurtosis maximization with leptokurtic signals. While originally based on heuristics, the analysis of SIRP's in relation to statistical speech models provides a strong framework for analysis. Initial analysis supports the conjecture regarding speech mixing and kurtosis effects for reasonable known SIRP modeling bounds on human voice.

REFERENCES

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

[2] J. -L. Lacoume, P. Ruiz, "Source Identification: A solution based on the cumulants," *Proc.* 4^{th} *ASSP Workshop Spectral Estimation Modeling*, Minneapolis, MN, Aug. 1988, pp.199-203.

[3] L. Castedo, O. Macchi, "Maximizing the Information Transfer for Adaptive Unsupervised Source Separation," *Proc. IEEE Signal Processing Advances in Wireless Communications Workshop*, Paris, 1997, pp.65-69.

[4] P. Comon, "Independent Component Analysis, A New Concept?," *Signal Processing*, vol. 36, pp.287-314, April, 1994.

[5] J. P. LeBlanc, I. Fijalkow, C.R. Johnson, Jr., "CMA Fractionally Spaced Equalizers: Stationary Points and Stability under IID and Temporally Correlated Sources," *International Journal of Adaptive Control and Signal Processing*, (accepted for publication).

[6] D. N. Godard, "Self-Recovering Equalization and Carrier Tracking in Two-Dimensional Data Communication Systems," *IEEE Trans. on Commun.*, vol. COM-28, No. 11, Nov. 1980.

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

[8] J.R. Treichler, M. G. Agee A New Approach to Multipath Correction of Constant Modulus Signals IEEE Trans. on Acoustics, Speech, and Signal Processing April 1983

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

[10] Y. Cao,S. Sridharn, M. Moody, "Multichannel Speech Separtation by Eigendecomposition and Its Application to Co-Talker Interference Removal," *IEEE Trans. on Speech, and Audio Proc. vol. 5, No. 3*, May 1997.

[11] S. Shamsunder, G. B. Giannakis, "Multichannel Blind Signal Separtation and Reconstruction," *IEEE Trans. on Speech, and Audio Proc. vol. 5, No. 6*, Nov. 1997.

[12] J. -F. Cardoso, "Iterative technique for blind source separtation using only fourth order cumulunts," *Proc. EUSIPCO 92*, Brussels, Belgium, Aug. 1992, vol. 2, pp.739-742.

[13] M. G. Bulmer, *Principles of Statistics*, New York: Dover Publications, 1967.

[14] Davenport, W. B An experimental study of speech-wave probability distributions Journal of the Acoustical Society of America, vol. 24, pp. 390-399, Jul. 1952

[15] Paez, M. D. and Glisson, T. H. Minimum mean squared-error quantization in speech IEEE Transactions on Communications, vol. Com-20, pp. 225-230, Apr. 1972

[16] Papoulis, A. Probability, Random Variables, and Stochastic Processes New York, NY: McGraw-Hill, 1989

[17] Brehm, H. and Stammler, W. Description and generation of spherically invariant speech-model signals Signal Processing, vol. 12, no. 2, pp. 119-141, Mar. 1987

[18] Wolf, D. and Brehm, H. Experimental studies on one- and two-dimensional amplitude probability densities of speech signals in Proceedings of the 1973 International Symposium on Information Theory, pp. B 4-6

[19] Brehm, H. Description of Spherically Invariant Random Processes by Means of G-Functions in Lecture Notes in Mathematics (A. Dold and B. Eckmann, Eds.). Springer-Verlag, 1982, vol. 969, pp. 39-73