SHORT-TIME KURTOSIS OF SPEECH SIGNALS WITH APPLICATION TO CO-CHANNEL SPEECH SEPARATION

Phillip L. De Leon

New Mexico State University
Klipsch School of Electrical and Computer Engineering
Las Cruces, New Mexico 88003-8001
pdeleon@nmsu.edu

ABSTRACT

Recent work into the separation of mixtures of speech signals has shown some success. One particular method is based on the assumption that scalar mixtures of speech signals have a kurtosis less than that for either source. Under this assumption, a simple gradient search algorithm is employed to maximize kurtosis thereby separating the source speech signals from the mixture. While this assumption has been observed to be generally true for long speech segments, it is quite reasonable to expect the assumption not to hold over short segments (windows) of speech. In this case, kurtosis maximization is not the appropriate strategy and the algorithm will fail to separate the signals. In this paper, we examine the kurtosis of speech signals over short segments of speech, i.e. short-time kurtosis. The analysis will indicate in general, how successful a kurtosis maximization strategy can be in separating speech signals from a mixture.

1. INTRODUCTION

In many audio-interface, forensic, multimedia, and speech recognition applications, mixtures of speech signals from various speakers must be separated out before processing. Given the complicated nature of speech signals this is a difficult problem compounded by environmental effects such as noise, echo, and reverberation and a strong desire for a simple algorithm suitable for real-time operation [2]. Several methods have been proposed some of which have shown moderate success but often at the expense of high computational complexity [6],[8].

The basic problem is illustrated in Figure 1. As a first step, we assume two unknown speech source signals, s_1 and s_2 are mixed in a scalar fashion (as opposed to the more realistic convolutional mixture which is also more difficult to separate) to produce two mixture signals x_1 and x_2. Thus given x_1 and x_2 and no further information, we wish to produce y_1 and y_2 which approximate s_1 and s_2. Such a problem formulation is referred as the "blind source separation" problem. The problem is illustrated in a more convenient form in Figure 2 where $\mathbf{A}$ is the mixing matrix whose elements are real numbers (scalars) and $\mathbf{W}^T$ is the separation matrix we must determine. In this case we have

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where

$$\begin{aligned} \mathbf{s} &= \left[\begin{array}{cc} s_1 & s_2 \end{array} \right]^T \\ \mathbf{x} &= \left[\begin{array}{cc} x_1 & x_2 \end{array} \right]^T \end{aligned} \tag{2}$$

Clearly, choosing $\mathbf{W}^T = \mathbf{A}^{-1}$ would separate the signals (assuming $\mathbf{A}$ is invertible) but $\mathbf{A}$ is not known.

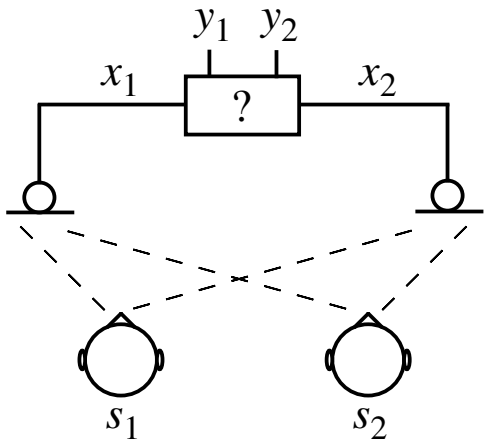

Figure 1: Speech signal separation problem

This work was supported by Air Force Research Laboratory under Grant F41624-99-0001

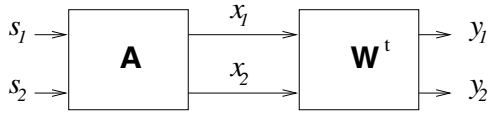

Figure 2: Basic signal separation setting

2. SPEECH SIGNAL SEPARATION

We begin by defining the kurtosis of a zero mean random variable, x as

$$\kappa_x = \frac{E\left[x^4\right]}{\{E\left[x^2\right]\}^2}. \quad (3)$$

Previous work observed that long-term mixtures of speech signals generally have a kurtosis lower than the kurtosis of the individual speech signals, i.e.

$$\begin{aligned} \kappa_{x_1} &< \min\{\kappa_{s_1}, \kappa_{s_2}\} \\ \kappa_{x_2} &< \min\{\kappa_{s_1}, \kappa_{s_2}\}, \end{aligned} \quad (4)$$

[3]. In addition, spherically-invariant random processes which have been used as statistical models for speech exhibit similar characteristics [1], [3]. In other types of source separation problems (not necessarily speech), conditions on the signal kurtosis have often been employed in algorithm design [5]. In the speech separation problem, we formulate a kurtosis maximization algorithm to adaptively compute the separation matrix $\mathbf{W}^T$ under the (critical) assumption that (4) holds. The steepest ascent (used in maximization) algorithm is given by

$$\mathbf{W}(n+1) = \mathbf{W}(n) + \mu \bigtriangledown \mathbf{W}(\kappa_{\mathbf{y}}) \quad (5)$$

where μ is the step size and $\bigtriangledown \mathbf{W}(\kappa_{\mathbf{y}})$ is the gradient of the kurtosis of the output signals,

$$\mathbf{y} = \left[\begin{array}{cc} y_1 & y_2 \end{array}\right]^T. \quad (6)$$

Computing the gradient yields the update algorithm for speech separation through kurtosis maximization

$$\begin{aligned} &\mathbf{W}(n+1) = \mathbf{W}(n) + \mu \cdot \\ &\left[\begin{array}{cc} -\alpha_1\beta_1\gamma_1 w_{21} & -\alpha_2\beta_2\gamma_2 w_{22} \\ \alpha_1\beta_1\gamma_1 w_{11} & \alpha_2\beta_2\gamma_2 w_{12} \end{array}\right] \end{aligned} \quad (7)$$

where

$$\alpha_i = 4\left[w_{1i}(n)x_1(n) + w_{2i}(n)x_2(n)\right]^3, \quad (8)$$

$$\begin{aligned} \beta_i = &-x_1(n)w_{1i}(n)r_{12} - x_1(n)w_{2i}(n)\sigma_2^2 + \\ &x_2(n)w_{1i}(n)\sigma_1^2 + x_2(n)w_{2i}(n)r_{12}, \end{aligned} \quad (9)$$

$$\begin{aligned} \gamma_i = &\left[w_{i1}^2(n)\sigma_1^2 + \right. \\ &\left. 2w_{i1}(n)w_{2i}(n)r_{12} + w_{2i}^2(n)\sigma_2^2\right]^{-3}, \end{aligned} \quad (10)$$

$$\mathbf{W}(n) = \left[\begin{array}{cc} w_{11}(n) & w_{12}(n) \\ w_{21}(n) & w_{22}(n) \end{array}\right], \quad (11)$$

and $\sigma_i^2 = E\left[x_i^2\right]$ and $r_{12} = E\left[x_1 x_2\right]$. Simple autoregressive estimators are used for σ_i^2 and r_{12},

$$\begin{aligned} \hat{\sigma}_i^2(n) &= \lambda\hat{\sigma}_i^2(n-1) + (1-\lambda)\, x_i(n)^2 \\ \hat{r}_{12}(n) &= \lambda\hat{r}_{12}(n-1) + (1-\lambda)\, x_1(n)x_2(n) \end{aligned} \quad (12)$$

since these statistics are not known a priori.

Results for the algorithm given in (7) indicated that at least one if not both speech signals could be separated from the mixture with separation ratios on the order of 40-50dB [4]. (It is assumed that if at least one speech signal can be separated from the mixture, residual processing can separate out the remaining signal.) It was also noted that during adaptation, separation ratios at times also decreased due to short-time failure of the critical assumption in (4) (kurtosis of the mixed speech signals is less than that for the source signals). The decrease in separation ratios during these failures was at times as much as pre-adaptation levels thus indicating no real separation of the mixed speech signals.

3. SHORT-TIME KURTOSIS OF SPEECH SIGNALS

In order to evaluate the susceptibility of the speech separation algorithm to failures in the critical assumption, we examine the kurtosis of speech source signals and scalar mixtures of speech signals over short-time windows. We first measure the kurtosis of each source speech signal over windows of 0.1, 0.25, and 0.5s in duration (no window overlap). Next we measure kurtosis of scalar mixtures of the speech signals $[x_1 = \alpha s_1 + (1-\alpha)s_2]$ over the same windows for the various mixing parameters $\alpha = 0.1, 0.2, \ldots, 0.9$. For each window, we determine whether the critical assumption in (4) is satisfied for one or both source signals. The proportion of signals which satisfy the critical assumption then lead to statistics on algorithm failure.

In order to control the evaluation of short-time kurtosis of speech signals, we use speech signals from 100 speakers in the standard TIMIT speech corpus [7]. The duration of the speech signals is on the order of 10s. Figures 3-5 illustrate the results. We see from each of the figures that the critical assumption is generally satisfied (>90%) by at least one of the mixture signals over segments as small as 0.1s for a wide range of mixing ratios. We note that at extreme mixing ratios of $\alpha = 0.1$ and $\alpha = 0.9$ the signals are approximately separated to begin with and thus short-time failures in these cases are not as critical.

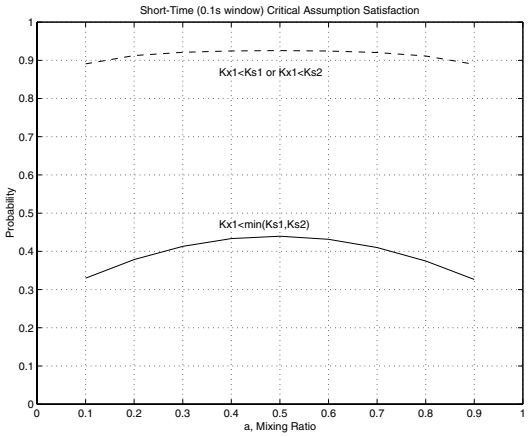

Figure 3: Probability of satisfying kurtosis condition over 0.1s window

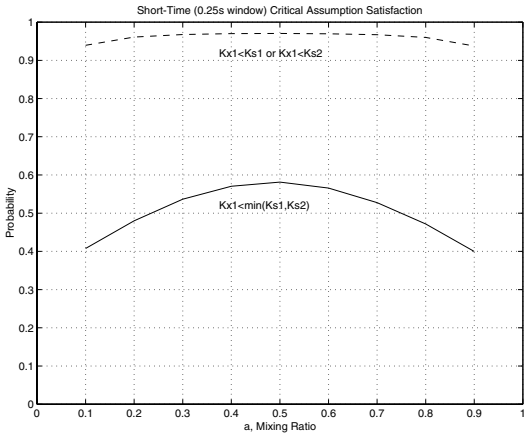

Figure 4: Probability of satisfying kurtosis condition over 0.25s window

4. CONCLUSION

In this paper we have examined the kurtosis of short segments of speech. The results indicate that a speech separation strategy based on maximizing kurtosis of the output signals will generally be effective. Our results indicate that over durations of 0.1, 0.25, and 0.5s, the mixture signal will have a kurtosis less than that of both source signals about 50% of the time and less than that either source signals about 90% of the time. The latter result indicates that nearly all the time at least one speech signal can be separated; residual processing may then be used to separate the remaining speech signal.

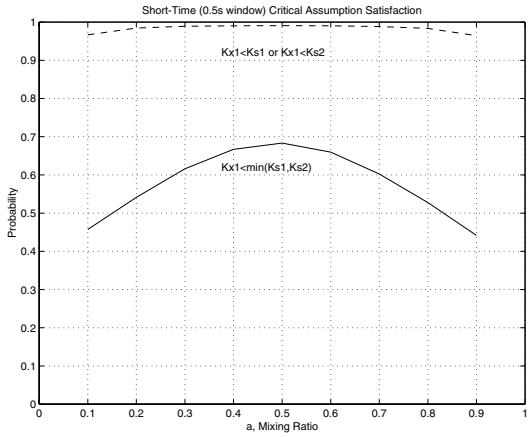

Figure 5: Probability of satisfying kurtosis condition over 0.5s window

5. REFERENCES

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

[2] F. Ehlers and H. Schuster, "Blind separation of convolutive mixtures and an applications in automatic speech recognition in a noisy environment," *IEEE Trans. Signal Processing*, vol. 45, pp. 2608–2612, Oct. 1997.

[3] J. LeBlanc and P. De Leon, "Source separation of speech signals using kurtosis maximization," *Proc. 35th Allerton Conference on Communications, Control, and Computing*, Moticello, IL., Sep. 1997.

[4] J. LeBlanc and P. De Leon, "Speech separation by kurtosis maximization," *Proc. ICASSP 98*, Seattle, WA., May 1998.

[5] A. Mansour and C. Jutten, "What should we say about the kurtosis?," *IEEE Signal Processing Letters*, vol. 6, pp. 321–322, Dec. 1999.

[6] D. Morgan, E. George, L. Lee, and S. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 407–424, Sep. 1997.

[7] W. Fisher, G. Doddington, and K. Goudie-Marshall, " The DARPA speech recognition research database: specifications and status," *Proceedings of the DARPA Workshop on Speech Recognition*, pp. 93–99, 1986.

[8] K. Yen and Y. Zhao, "Adaptive co-channel speech separation and recognition," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 138–151, Mar. 1999.