USE OF LOCAL KURTOSIS MEASURE FOR SPOTTING USABLE SPEECH SEGMENTS IN CO-CHANNEL SPEECH

Kasturi Rangan Krishnamachari, Robert E. Yantorno and Jereme M. Lovekin
Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, Pa 19122-6077, USA
ryantorn@nimbus.temple.edu, kkrish01@astro.temple.edu, jlovekin@temple.edu,
http://nimbus.temple.edu/~ryantorn/speech

Daniel S. Benincasa and Stanley J. Wenndt
Air Force Research Laboratory/IFEC, 32 Brooks Rd. Rome NY 13441-4514, USA
danb@rl.af.mil, wenndts@rl.af.mil

ABSTRACT

Recently, a novel method to process co-channel speech was proposed [1]. Previous methods include enhancing the target speech, or suppressing the interfering speech or both enhancing the target and suppressing the interferer. The proposed new method searches for usable speech frames which are usually found in clusters under co-channel conditions. The term "usability" is context dependent, i.e., usable in the context of such things as speaker identification, gisting, etc. In this paper we investigate the use of kurtosis for spotting usable speech segments under co-channel conditions. Preliminary results reveal that a kurtosis of 1.5 or greater occurs close to the beginning and ends of segments of usable speech, i.e., they usually bracket the usable speech segment. For Male/Male case, we observe that 92% of usable clusters are spotted, for Male/Female case 83% of usable clusters are spotted and for Female/Female case, 86% of usable clusters are spotted

1. INTRODUCTION

Recently, a novel method to process co-channel speech was proposed [1]. Previous methods include enhancing the target speech, or suppressing the interfering speech or both enhancing the target and suppressing the interferer. The proposed new method searches for usable speech frames which are usually found in clusters under co-channel conditions. The term "usability" is context dependent, in terms of speaker identification. However, usability could be defined as well for speaker verification or gisting.

Traditionally, Higher Order Statistics (HOS) algorithms have been used for blind source separation of digital communication signals [2][3][4][5][6], or to reduce the inter-symbol interference between signals. Even though digital communication signals fall under a different statistical class than speech signals, there are some HOS based techniques which are used for co-channel interference removal from speech signals [7]. They are usually computationally intensive, necessitating correlation matrix estimation and eigen-decomposition or polyspectra estimations, requiring signals from two separate sources, which have different channel characteristics.

Kurtosis is the measure of gaussianity of a signal and is computed using fourth and second moments. One can find a discussion of developing measures of gaussianity in Donoho, 1981 [8]. LeBlanc *et. al* [9] proposed a recursive adaptive gradient descent technique with the cost function designed to maximize the kurtosis of the separated interference and target speech. As we are not interested in separating target speech from interference per se, but in spotting usable clusters in co-channel speech, we employ kurtosis in a different way.

2. STATISTICAL MODEL AND KURTOSIS OF SPEECH:

Applying statistical notions to speech signals necessitates the estimation of a probability density function (PDF). The probability density is estimated by determining a histogram of amplitudes for a large number of samples. Davenport [10] made extensive measurements of this kind. A simple approximation to the speech pdf is the Laplacian density

$$f_x(x) = \frac{1}{\sqrt{2}s_x} * e^{\frac{-\sqrt{2}|x|}{s_x}}$$

Paez and Glisson [11], using similar measurements, have shown that a more accurate approximation to measured speech amplitude densities is a gamma distribution of the form

$$f_x(x) = \sqrt{\frac{\sqrt{3}}{8ps_x|x|}} * e^{\frac{-\sqrt{3}|x|}{2s_{\max}}}$$

More recently, a 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 [12]. Modeling speech signals using SIRP's was justified by the fact that the actual speech bivariate PDFs have been shown to exhibit SIRP-like quality [13][14][15]. Another interesting fact is that the random processes with Laplace or Gamma PDFs are SIRPs.

The kurtosis of a random variable x is defined as

$$\frac{E(x^4)}{E(x^2)^2} - 3$$

However, some textbooks define kurtosis as

$$\frac{E(x^4)}{E(x^2)^2}$$

Kurtosis is a measure of gaussianity of a Random Variable (RV), and it is a scale independent dimensionless parameter. A Gaussian RV has a kurtosis of zero. If an RV has kurtosis less than zero, it is termed platykurtic. If it has kurtosis greater than zero, it is termed leptokurtic. For PDFs like Gamma or Laplacian, the kurtosis of the sum of these PDFs is lower than the kurtosis values of the individual PDFs.

Speech signals are generally leptokurtic. This fact was used to develop an adaptive algorithm for blind separation of mixed speech signals, with the cost function designed to maximize the output kurtosis [9]. The algorithm is actually a modified version of the source separation algorithm of digital communication signals [6]. The modification adjusts for the differing statistics between digital communication signals and voice signals. The digital communication signals, being platykurtic require minimization of output kurtosis.

It has been reported that the kurtosis of co-channel speech is generally less than the kurtosis of the individual speech utterances, though this may not always be the case [9]. Our experiments deal with examining the kurtosis of individual frames rather than whole speech, and our observations confirm that a majority of the time, the kurtosis of the sum of speech signals are less than the individual kurtosis of the target and interfering speech.

3. EXPERIMENTS AND RESULTS

Five speech signals (4 male, 3 female) were taken from the TIMIT database. The original speech was sampled at the rate of 16 kHz. We re-sampled them to 8 kHz prior to analysis. Two speech signals were added together to obtain co-channel data. The range of Target-to Interfere Ratio (TIR) in individual frames varied from -20 dB to +20 dB approximately. Analysis was done on a frame-by-frame basis with 50% overlap. The frame length was 320 samples corresponding to 40 ms. Prior to mixing, the TIR for each frame was computed. All frames having |TIR| > 10 dB were *apriori* labeled as usable, and usable frames were then grouped into clusters.

We observed that setting a threshold of 1.5 for the kurtosis resulted in detection of frames adjacent to the beginning and ends of segments of usable speech, i.e., they usually bracket the usable speech segment (usable as defined by the by the TIR method). Let the frames selected by the kurtosis threshold be called Fk. We say an Fk frame is in the vicinity of a usable cluster if it is within ±2 frames of the end points of the usable cluster. If an Fk is not in the vicinity of any usable cluster, it is a false alarm. A usable cluster having no Fk in its vicinity is said to be missed.

Figure 1 shows a typical TIR magnitude and Kurtosis frame-by-frame variations of co-channel speech. The figure was for one Male/Male case. The usable regions as selected by TIR criterion are indicated as regions below the square wave. The usable clusters, as observed from the figure, occur mostly between peaks in the Kurtosis waveform. From the graph, setting a high threshold will remove many false alarms at the expense of missed clusters. However, setting a low threshold will increase the number of false alarms. After observing similar graphs for various cases like Male/Female and Female/Male, we found that 1.5 is a reasonable threshold for the kurtosis. We summarize our observations of spotted usable clusters using the kurtosis threshold method, false alarms and misses in Table 1 below.

For Male/Male case, we observe that around 92% of usable clusters are spotted, for Male/Female case around 83% of usable clusters are spotted and for Female/Female case, around 86% of usable clusters are spotted. An important thing to note is that kurtosis by itself does not point to usable clusters, rather a coarse location where usable clusters may be searched in a co-channel utterance. We repeatedly observed that most of the Fks form the end points of usable clusters.

Certain Fks not in the vicinity of any usable cluster were observed. Closer inspection revealed that these were mostly silence over silence portions. These Fks produced false alarms. The average false alarm rate for Male/Male case was 37%, for Male/Female case was 31% and for Female/Female case was 18%. One has to keep in mind that kurtosis is scale invariant and hence it is independent of the energy of the frame. The statistics of the silence portions were not exactly noise-like and certain portions showed even periodicity though with very less amplitude. A simple voicing state detection system coupled with kurtosis criterion would avoid such false alarms.

Certain clusters that had no Fks in their vicinity were also observed. Those are said to be missed. The maximum and average length of a missed cluster indicates the reliability of this kurtosis method as a tool for usable speech detection. Failure to detect one large usable cluster may overshadow the advantages in detecting number of small usable clusters. The maximum length of missed clusters for Male/Female and Female/Female cases were 2 frames and for Male/Male, 6 frames. The average lengths are 1.03 frames, 2.94 frames and 1.17 frames for Male/Male, Male/Female and Female/Female cases respectively. We see that, on an average, approximately two usable frames are missed by this method. Hence, this method is indeed a good tool for usable speech detection.

An inspection revealed that the missed clusters were bursty in nature - that is, they had unusable portions before and after them in the same frame. The TIR criterion would naturally pick these frames, but the kurtosis method missed them. Occasionally we also observed cases where the TIR of a frame with unvoiced over silence condition would go above 10 dB and hence incorrectly be flagged as usable. However, a failure to detect these types of frames should not actually be termed as a miss.

The experiments outlined in this paper are part of an effort to build a next generation co-channel speech processing system, with usable speech detection unit at the front-end. The kurtosis-based criterion offers a set of locations in the co-channel speech where the occurrence of a usable cluster is highly likely. It offers the advantage of selectively searching regions of co-channel speech, rather than searching blindly. Additionally, it may be noted that kurtosis computation is not very costly.

4. SUMMARY

The result of our preliminary investigations with kurtosis as a tool to detect usable speech segments in co-channel speech is promising. Even though kurtosis by itself may not be sufficient to identify usable clusters, it provides coarse estimates of locations that are likely to have usable clusters in their vicinity. Suppressing processing on obviously unusable portions, such as silence over silence, can reduce false alarms, as the results of kurtosis on those frames are misleading.

ACKNOWLEDGEMENT

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, USAF, under agreement number F30602-00-1-0517. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

DISCLAIMER

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory, or the U.S. Government.

5. REFERENCES

1. Krishnamachari, K. R., Yantorno, R. E., Benincasa D. S., and. Wenndt, S. J., “Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions.”, ICSPACS, 2000.

2. Cardoso, J. F, “Iterative Technique for Blind Source Separation Using Only Fourth Order Cumulants”, Proc. EUSIPCO, vol 2, pp: 739-742, 1992

3. Cardoso, J.-F., “Source Separation Using Higher Order Moments,” Proc. IEEE ICASSP, pp: 2109-2112, 1989.

4. Laocume, J. L., and Ruiz, P., “Source Identification: A Solution Based on Cumulants”, Proc. 4th ASSP Workshop Spectral Estimation Modeling, pp: 199-203, 1988.

5. Moreau, E., Macchi, O., “New Self-adaptive Algorithm for Source Separation Based on Contrast Functions”, Proc. IEEE Signal Processing Workshop on Higher Order Statistics, pp: 215-219, 1993.

6. Treichler, J. R., and Agee, M. G., “A New Approach to Multipath Correction of Constant Modulus Signals”, IEEE Trans. On Acoustics, Speech and Signal Processing, 1983.

7. Cao, Y., Sridharan, S., Moody, M., “Multichannel Speech Separation by Eigen decomposition and Its Application to Co-Talker Interference Removal”, IEEE Trans. On Speech and Audio Proc., Vol 5, no. 3, pp:1997

8. Donoho, D. L., “On Minimum Entropy Deconvolution”, *Applied Time Series Analysis*, D. F. Findley, Ed., New York: Academic Press, 1981.

9. LeBlanc, P. J., Leon, P. L. de., “Speech Separation by Kurtosis Maximization”, IEEE ICASSP, pp: 1029-1032, 1998.

10. Davenport, W. B., “An Experimental Study of Speech Wave Probability Distributions”, Journal of Acoust. Soc. America, Vol 24, pp: 390-399, 1952.

11. Paez, M. D., and Glisson, T. H., “Minimum Mean Squared Quantization in Speech”, IEEE Transactions on Communications, Vol. Com-20, pp: 225-230, 1972.

12. Papoulis, A., "Probability, Random Variables and Stochastic Processes", New York: McGraw-Hill, 1989.

13. Brehm, K., "Description of Spherically Invariant Random Processes by Means of G-Functions", in Lecture Notes in Mathematics (A. Dold and B. Eckmann, Eds)", Springer-Verlag, Vol 969, pp: 39-73, 1982.

14. Brehm, H., and Stammler, W., "Description and Generation of Spherically Invariant Speech-Model Signals", Signal Processing, Vol. 12, no. 2, pp: 119-141, 1987.

15. Wolf, D., and Brehm, H., "Experimental Studies on One and Two-dimensional Amplitude Probability Densities of Speech Signals ", Proceedings of the 1973 International Symposium on Information Theory, pp: B 4-6, 1973.

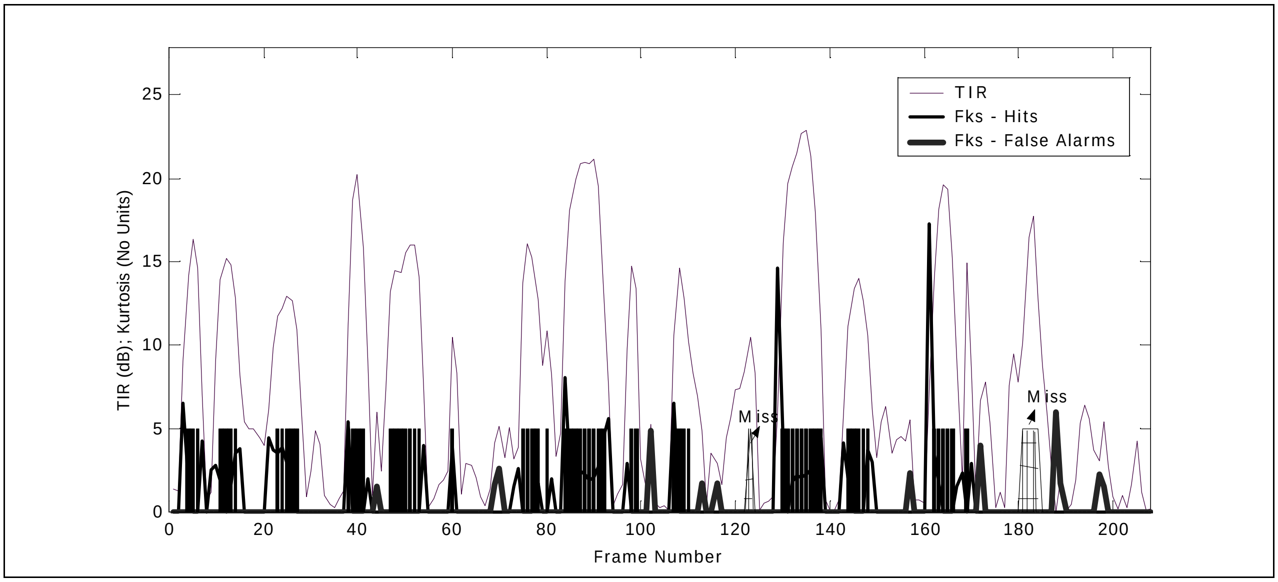

Figure 1 Typical TIR magnitude and kurtosis versus frame number for one Male/Male case. Regions under square wave are usable clusters, as spotted by TIR threshold. The solid regions are those clusters that have at least an Fk in their vicinity. The rest of the pulses (pulses with grids) correspond to missed clusters.

TABLE – I: Cluster Detection Statistics:

	Male/Male	**Male/Female**	**Female/Female**
Average percentage of spotted clusters	92.4%	82.7%	87.5%
Average percentage of missed clusters	7.6%	17.3%	12.5%
Average percentage false alarm rate	37.3%	30.9%	18.3%
Maximum length of missed cluster	6 frames	2 frames	2 frames
Average length of missed clusters	1.028 frames	2.94 frames	1.167 frames