

The Spectral Autocorrelation Peak Valley Ratio (SAPVR) – A Usable Speech Measure Employed as a Co-channel Detection System

Robert E. Yantorno, Kasturi Rangan Krishnamachari 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

The traditional approach to co-channel speech processing has been to either enhance the target speech, or suppress the interferer's speech, or perform both simultaneously. We have previously presented a novel and unique approach [Spectral Autocorrelation Peak Valley Ratio (SAPVR)] to processing co-channel speech, and that is to identify and extract segments of speech that are usable, i.e., usable in the context of speaker identification, speech recognition or speaker tracking [1]. This measure has also been found to be an effective method to identify the existence of co-channel speech; in this case only voiced portions are used. The voiced portions were identified using spectral flatness. The SAPVR, as a part of a co-channel detection system, appears not to be gender specific in terms of its operational capabilities. False alarms and missed frames occur mostly in transition regions (onset or offset of voicing). SAPVR approach to co-channel detection is part of an ongoing effort to develop a co-channel detection and separation system that would utilize multiple, independent measures.

1. INTRODUCTION

Co-channel speech has presented a challenge to the speech processing community for over 30 years with only minimal success. The traditional approach to co-channel speech is to attempt to enhance the target speech [2],

suppress the interferer's speech [3], [4], [5] and [6] or perform both enhancement and suppression [7] and [8]. More recently, a novel approach to co-channel speech processing has been to identify those segments of co-channel speech that are usable, and extract and use those segments for such processes as speaker identification [1]. Usability, however is context dependent, i.e., if one is going to use the speech for speaker identification then one must determine what defines usability in the context of speaker identification. Development of usability criteria for speaker identification is presently underway [9]. Another application for co-channel speech processing would be the detection of co-channel speech, for if co-channel speech is detected then one might wish to suspend any speech processing operation until there is no longer any co-channel speech.

The method presented here is part of an ongoing project to develop a series of usable speech measures to be used as co-channel detection measures, of which the SAPVR is the first. A block diagram of the proposed co-channel speech detection system is shown in Figure 1 below. The basic concept of the system is to use a number of different measures, each of which contains different types of information, such as time domain and frequency domain measures, so that each measure will be independent of the other measures. The measures will then be weighted and fused in such a way as to produce a system much more effective, accurate and robust than using only one measure.

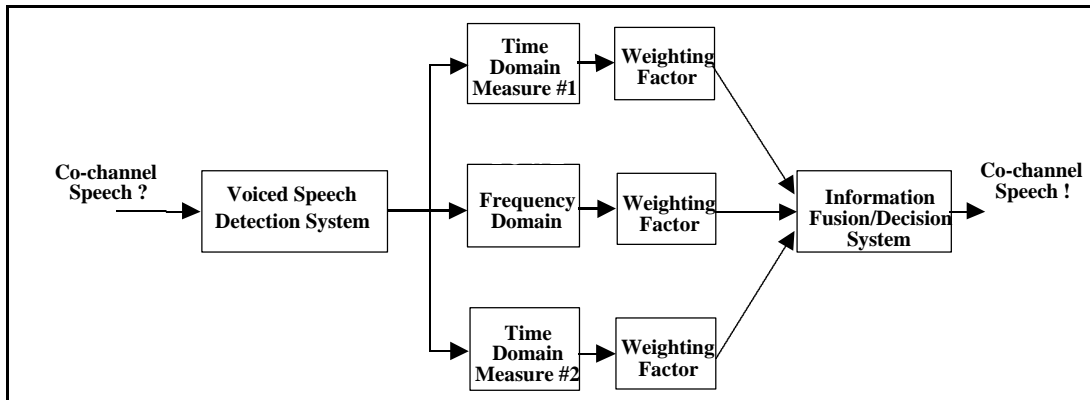


Figure 1. Block diagram of proposed co-channel detection system.

2. METHODS

Speech from the TIMIT database was used. The original TIMIT data, in 16 bit 16kHz form, was filtered and then decimated to 8kHz. The SAPVR measure has been described previously [1]. The SAPVR was conducted on 32 msec frames of speech data. The basic concept of the SAPVR approach is to perform the autocorrelation on the magnitude spectrum and then determine the ratio of the sum of the peaks in the spectral autocorrelation domain over the value of the first valley.

The SAPVR measure is used to detect structure in the spectral autocorrelation domain. This structure is well illustrated in Figure 2 (above) for the time, frequency and spectral autocorrelation domains. Also, because the spectral autocorrelation can be used to represent structure in the frequency domain, it can also be used to detect a loss of structure, but only for voiced speech. This loss of structure is shown in Figure 3 (below).

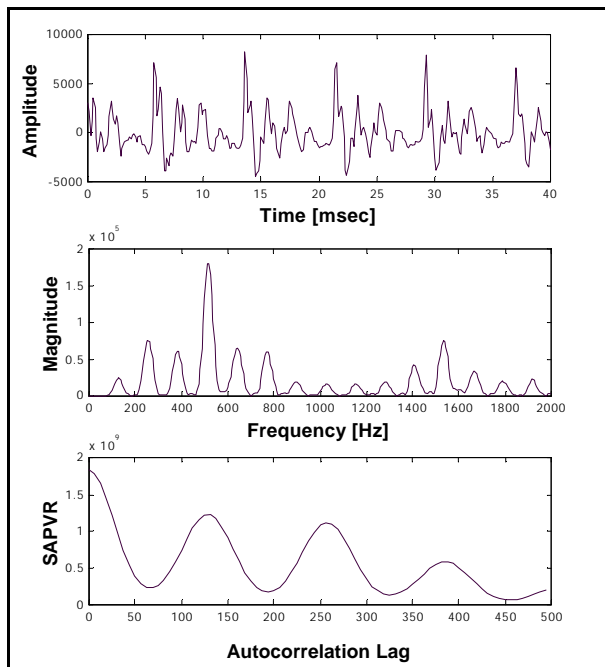


Figure 2. Single speaker data. a.) Male speech data. b.) Magnitude spectrum. c.) Autocorrelation spectral lag.

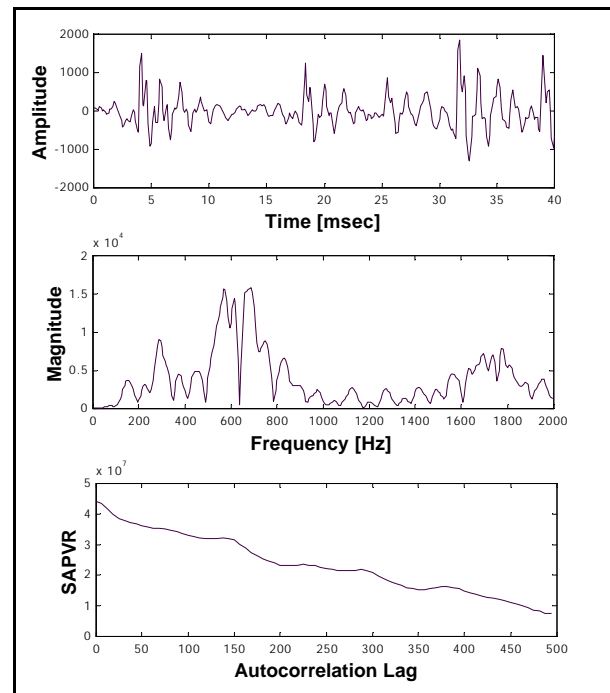


Figure 3. Multi-speaker data. a.) Two male speakers. b.) Magnitude spectrum c.) Autocorrelation spectral lag.

For co-channel speech there are a number of different voicing conditions that can exist for two speakers' speech.

For example, there can be voiced-silence, silence-voiced, silence-unvoiced, voiced-unvoiced, etc. However, the condition which would most readily indicate the existence of co-channel speech would be the voiced-voiced situation. The reason for this is that for voiced speech, the time and frequency domain representation is very structured. However, for the co-channel situation of voiced-voiced, the structure in both the time and frequency domain is lost. Therefore, our co-channel detection system will process voiced speech frames (which for this application may actually be voiced-silence, voiced-unvoiced, or voiced-voiced).

Because the SAPVR co-channel speech detection approach relies on the analysis of only voiced portions of speech, the spectral flatness measure was used for identifying voiced speech. There are advantages to using this method. First, one can more easily vary the threshold for voiced speech using spectral flatness than using a traditional approach, such as measuring the energy and zero crossing rate. Also, there is no need to consider changing the zero-crossing threshold when going from a male speaker with low pitch to a female speaker with high pitch. Finally, spectral flatness provides better discrimination between voiced and unvoiced speech, and is critical for our approach. Note, the spectral flatness can vary from 0dB for unvoiced to 60dB for voiced. This means that frames with a spectral flatness close to 60dB are more voice-like, and therefore, can be classified as voiced. The spectral flatness measure (SFM_{dB}) in dB is defined as:

$$SFM_{dB} = 10 \log_{10} \frac{Gm}{Am} \quad (1)$$

$$Gm = \frac{1}{N} \sqrt{\prod_{i=0}^N mag(i)} \quad \text{and} \quad Am = \sum_{i=0}^N mag(i) \quad (2)$$

Where, Gm is the geometric mean, Am is the arithmetic mean, $mag(i)$ are the magnitudes of each of the spectral lines i , and N is the number of FFT points or spectral lines.

The results of using the spectral flatness are shown in Figure 4, where the threshold used to extract the voiced speech is represented by the flat horizontal line. Note, as indicated in Figure 4 above, the threshold is shown as being positive when in fact it is actually negative; the positive spectral flatness threshold was used only for the convenience of illustration and simplicity of concept. Note, TIR is the Target-to-Interfere Ratio. It has been determined previously [10] that speaker identification is minimally degraded, i.e., a decrease of about 15% in accuracy is observed with a TIR of 20 dB.

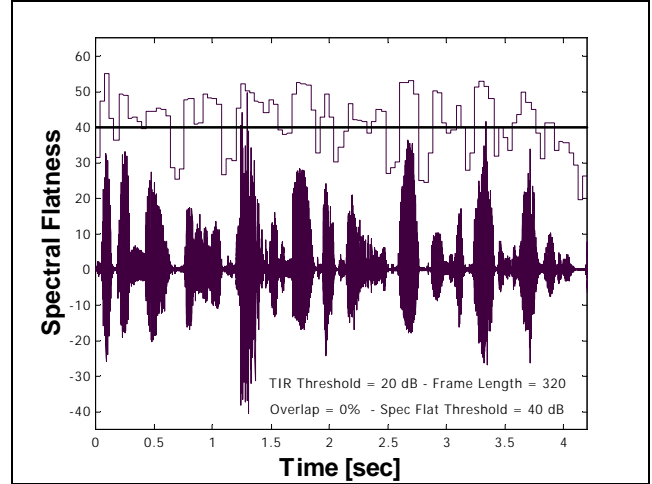


Figure 4. Speech data, Spectral flatness and Spectral Flatness threshold. Straight horizontal line represents a spectral flatness threshold of 40dB.

Figure 5 (shown below) illustrates the effectiveness of using spectral flatness; the threshold for accepting the frame as being voiced was 40dB, therefore, all frames with a spectral flatness measure greater than 40 dB are considered as being voiced. The result of using the 40 dB threshold is shown in Figure 4, where the voiced portions to be used are identified by the black rectangles. It is evident from Figure 5 that using the spectral flatness provides an accurate identification of voiced segments.

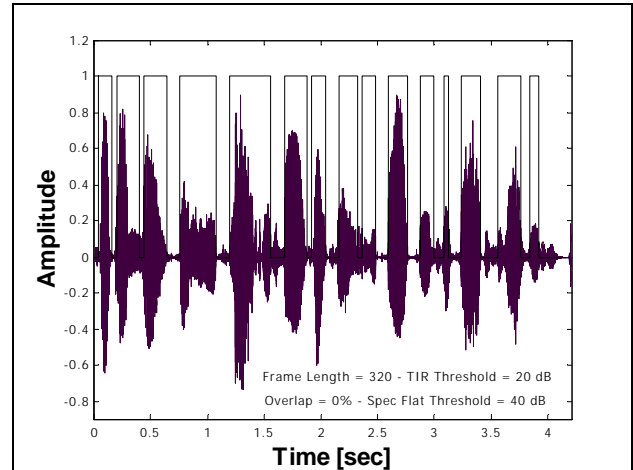


Figure 5. Speech data and spectral flatness results using information from Figure 4 above.

3. RESULTS

The result of using the SAPVR measure for detection of female-male co-channel speech is shown below in Figure 6.

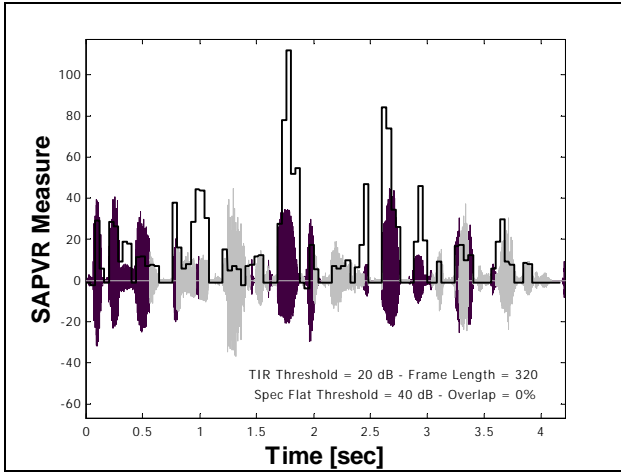


Figure 6. SAPVR data for female-male co-channel speech. Usable speech is labeled as black and unusable "co-channel" speech is identified as gray. The SAPVR data is shown as black rectangles.

Note, in Figure 6 that the usable speech (black) has much higher SAPVR values, i.e., values above the SAPVR threshold of 10, with some as large as 110, than the unusable speech (gray). The SAPVR seems to perform reasonably well identifying "usable" speech. The converse is also true, in that the SAPVR is also a good indicator of unusable speech. In this case, unusable speech would be defined as speech that is corrupted with another speaker's speech, or speech where there is no structure in the spectral autocorrelation domain, which would be the case for unvoiced speech.

Note, TIR, as listed in Figure 6, is the Target-to-Interfere Ratio. It has been determined previously [10] that speaker identification is minimally degraded, i.e., a decrease of about 15% in accuracy is observed with a TIR of 20 dB.

First, voiced-only portions of the co-channel speech were identified and tagged using the spectral flatness measure. Next, the SAPVR for each frame of voiced speech was determined. Finally, a threshold of 10 for the SAPVR was used as a measure to differentiate between usable speech, and unusable speech, i.e., co-channel speech, where any value below 10 indicates the existence of co-channel speech.

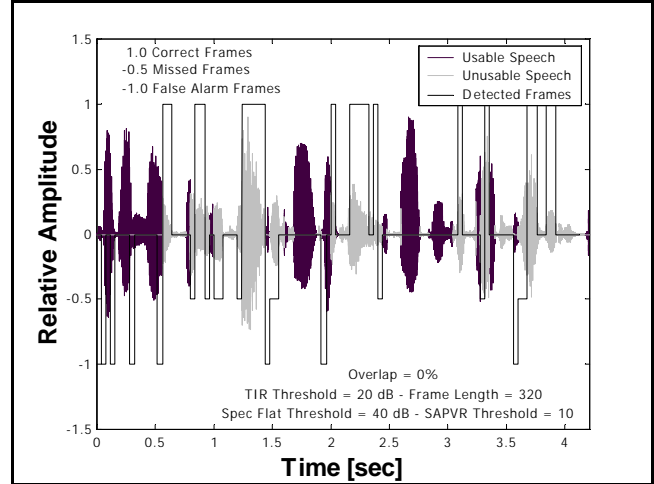


Figure 7. Detection of female-male co-channel speech using the SAPVR measure. Positive rectangles indicate detection success (1.0); negative rectangles indicate either false alarms (-0.5) or misses (-1.0).

The results of the test are shown in Figure 7 above for female-male co-channel speech. Note that the false alarms, i.e., -1.0 rectangles occur in the regions of transition. Also, the majority of the missed frames occur in transition regions.

A tabulation of the results of Figure 7 shown above along with results from a test with male-male co-channel speech (data not shown) are shown in Table 1 below.

Table 1. Comparison of Co-channel detection for male-male and female-male speech.

| | | Male-Male | Female-Male |
|--------------------|--------------|------------|-------------|
| Correct | Speaker #1 | 33% | 35% |
| | Speaker #2 | 88% | 100% |
| | Total | 52% | 66% |
| False Alarm | Speaker #1 | 0% | 6% |
| | Speaker #2 | 33% | 40% |
| | Total | 14% | 22% |
| Missed | Speaker #1 | 67% | 65% |
| | Speaker #2 | 13% | 0% |
| | Total | 48% | 34% |

SAPVR threshold = 10, 320 points, and 0% overlap. Note – Male-Male Speaker #2 is the same as Female-Male Speaker #2.

Some conclusions can be drawn from Table 1. First, with an increase in the number of correct decisions there is

an accompanying increase in false alarms. Also, because speaker #1 was male for one set of tests and female for another set of tests, this indicates that the system is not "gender" specific in terms of its operation, and also that the speech of speaker #2 has some characteristics which make it better suited for this measure.

4. SUMMARY

We have presented a novel approach for dealing with co-channel speech by detecting segments of co-channel speech rather than by extracting them, using enhancement or suppression or a combination of enhancement and suppression of co-channel speech. Initial tests indicate that the SAPVR method shows promise as a part of a co-channel detection system and further investigation is planned.

ACKNOWLEDGEMENTS

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.

We would like to thank Dr. Andy Noga for his assistance in the development of the mathematical approach for determining the false alarm rate.

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. Wennndt, S. J., "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions." IEEE International. Symposium. Intelligent. Sig. Process. and Comm. Sys., (accepted), Nov. 2000.
- [2] Shields, V. C. Jr., "Separation of Added Speech Signals by Digital Comb Filtering," Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA, September 1970.
- [3] Parsons, T. W., "Separation of Speech from Interfering Speech by Means of Harmonic Selection," *J. Acoust. Soc. Am.*, 60(4):911-918, October 1976.
- [4] Hanson, B. A., and Wong, D. Y., "The Harmonic Magnitude Suppression (HMS) Technique for Intelligibility Enhancement in the Presence of Interfering Speech," Proc. IEEE ICASSP, pp: 18A.5.1-18A5.4, 1984.
- [5] Naylor, J. A., and Boll, S. F., "Techniques for Suppression of an Interfering Talker in Co-channel Speech", Proc. IEEE ICASSP, pp: 205-208, 1987.
- [6] Zissman, M. A., and Weinstein, C. J., "Automatic Talker Activity Labeling for Co-channel Talker Interference Suppression", Proc. IEEE ICASSP, pp: 813-816, 1990.
- [7] Lee, C. K., and Childers, D. G. "Cochannel Speech Separation:", *J. Acoust. Soc. Am.*, 83(1):274-280, January 1988.
- [8] Morgan, D. P., George, E. B., Lee, L. T, and Kay, S. M., "Co-channel Speaker Separation by Harmonic Enhancement and Suppression", IEEE Trans. Speech & Audio Process., Vol. 5, No. 5, pp: 407-424, 1997.
- [9] Lovekin, J. M., Yantorno, R. E., Benincasa, D., and Wennndt, S. J., "Developing Usable Speech Criteria for Speaker Identification Technology." (submitted) ICASSP 2001.
- [10] Yantorno, R. E., "Co-Channel Speech and Speaker Identification Study", Final report for Summer Research Faculty Program, Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, 1998.