

SPECTRAL AUTOCORRELATION RATIO AS A USABILITY MEASURE OF SPEECH SEGMENTS UNDER CO-CHANNEL CONDITIONS

Kasturi Rangan Krishnamachari and Robert E. Yantorno

Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, Pa 19122-6077, USA
kkrish01@astro.temple.edu, ryantorn@nimbus.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

A Spectral Autocorrelation Ratio is used to quantify the usability of speech, which has been corrupted by another speech signal in the context of speaker identification. Recent studies [1] have revealed that sizeable portions of speech exist in a co-channel utterance that contain enough information for speaker identification. In scenarios with limited training and testing data, a system which separates out those "usable" portions of speech, is desirable. The purpose of this investigation is to identify these usable segments in the co-channel speech. Our results suggest that spectral autocorrelation approach identifies those portions of the co-channel speech that are usable. What is being proposed here, i.e., the identification of "usable" speech, represents the front end of a next generation speech processing system that will involve the use of an information fusion/decision system which will process both time as well as frequency domain information about the usability of frame of speech. The material present here is our initial effort at defining usable speech.

1. INTRODUCTION

Until recent, the classical approaches to co-channel speaker separation were to enhance the target speech, suppress the interfering speech, or enhance the target speech while suppressing the interfering speech. The question concerning co-channel speech in the past was how do we extract the speech of one of the speakers. However, if the final goal with respect to co-channel speech is to use it for such things as speaker identification, then it becomes more advantageous to determine which segments of co-channel speech will improve the performance of the speaker identification system.

We are proposing a new approach to co-channel speech. From previous studies we have determined that there exists segments of speech that we have identified as "usable" in the sense that the interferer's speech does not degrade the informational content of the target speech to be used for such things as speaker recognition.

Recent studies [1] have revealed that about 39% of the co-channel speech has enough information about the target speaker to perform reliable speaker identification even when the overall Target-to-Interferer (TIR) ratio is 0 dB. Hence those segments become "usable" for a speaker identification system. A speech segment is "usable" if it

contains enough information to identify the target speaker. The usability definition can be extended to include other applications such as speech recognition, gisting and word reconstruction. It was also found that about 32% of the co-channel speech contained enough information about the interferer such that the interferer's identity could be identified [2]. Hence, if one wishes to extract the identity of both the target and the interferer, as much as 70% of the entire speech is available. The normal situation with those usable frames is that they occur in segments rather than isolated frames.

It was determined that a 20 dB Target-to-Interferer ratio is a reasonable lower limit for speaker identification to work reliably [2]. So, a straightforward method to estimate the usability of a speech frame would be to estimate target-to-interferer ratio for each frame. This is similar to the estimation of Harmonic-to-Noise ratio, used by laryngologists to rate the degree of hoarseness of a voice.

Under voiced portion-over-voiced portion co-channel conditions, there will be a significant amount of energy within a frame, related to the stronger speaker. Hence the ratio of Harmonic energy of the stronger talker to the energy content of all other components (both noise as well as harmonic energy content of weaker talker) is a good measure of usability of that speech frame. The problem with this method is to accurately measure the harmonic energy related to the stronger talker. This requires an accurate estimate of the pitch of at least one talker. A simple pitch detection algorithm is Schroeder's frequency histogram method [3]. This method makes use of the fact that voice harmonics are harmonic to a high degree of precision. This means the fundamental frequency could be determined by dividing the harmonic peak by its harmonic number. The problem is that we do not know which harmonic it is, since we have not yet calculated the fundamental frequency. Schroeder circumvents this problem by finding integer sub-multiples of all the peaks and entering them into a histogram. The largest entry in the histogram is taken to be the pitch. In most cases this would be the correct decision, although it should be observed that if f_0 is a histogram entry, $f_0/2$ would be an entry of equal height, since every harmonic of the former is the harmonic of the latter. Hence the histogram method would be susceptible to octave errors. There are other

pitch detection algorithms, such as the Maximum Likelihood, Cepstral, harmonic matching and auditory synchrony based pitch detectors [4]. Those were developed originally without assuming co-channel situations, but later modified for such conditions [5].

2. SPECTRAL AUTOCORRELATION

Our original aim was to quantify the usability of a co-channel speech frame and without having to estimate the pitch of the stronger or weaker talker. Performing autocorrelation in the spectral domain provides an elegant solution to our original problem of finding the usability of a speech frame. Instead of finding the ratio of strong talker's harmonic energy to the weaker talker's harmonic energy (and the remaining noise), we perform the autocorrelation of the magnitude spectrum.

This technique of performing autocorrelation in frequency domain was previously used to represent heart rate variability [6] and to compare color flow imaging algorithms [7]. Ashira and Kado [8] also used a similar technique of frequency domain autocorrelation. Their method used frequency domain linear prediction to separate voiced portions of speech from additive noise. Again, a co-channel situation was not assumed in their research. However, an important conclusion from their research, one that could be exploited, is that the power spectrum of voiced speech can be predicted because of its harmonic structure, while that of noise cannot be predicted.

Consider a frame of speech that is voiced. The frequency spectrum $X(k)$ of such a frame will contain harmonically related pulses. If we use Schroeder's method (which was later adapted by Parsons [9] for pitch estimation), we have to search either side of the highest peak at its sub-multiples, for local maxima. Instead, performing a spectral autocorrelation of such a frame will always result in pulses of decreasing height with increasing lag (Figures 1a and 1b). This is clearly an advantage, as will be discussed below.

If the original magnitude spectrum $X(k)$ contained harmonics at integral multiples of the digital frequency 'p', then the major contribution to the first peak in the spectral autocorrelation, after lag zero, is due to the product of adjacent harmonics, which occurs at lag 'p'. That is, the magnitude of the first spectral peak after lag zero for a voiced frame can be approximated as

$$R(p) = X(p)X(2p) + X(2p)X(3p) + \dots (1)$$

Other terms will contain less energy, and will not contribute significantly to this peak. Note that this parameter contains all the information about significant harmonics. The next peak occurs at lag '2p' and its amplitude can be approximated as

$$R(2p) = R(p)R(3p) + R(2p)R(4p) + \dots (2)$$

By the inherent property of the autocorrelation function, this peak has lesser amplitude than $R(p)$. If the segment of speech is unvoiced, the spectral autocorrelation will not contain any prominent peaks other than the one at lag 0.

We investigated the behavior of spectral autocorrelation under co-channel condition, and the spectral autocorrelation varied, depending on whether 1.) both the target and interfering speech were voiced, 2.) either one of them were unvoiced or 3.) both of them were unvoiced. When both the speech frames were unvoiced, the spectral autocorrelation did not contain any pulses that were harmonically related to each other. If at least one of the speech frames was voiced, the spectral autocorrelation contained harmonically related pulses as expected. If both the speech frames were voiced, the spectral autocorrelation contained either two distinct trains of pulses that were harmonically related if the speakers' pitches were different by approximately 25%, otherwise there was one train of broad pulses. Figures 1a, 1b, 1c and 1d show two frames of voiced speech and their corresponding spectral autocorrelations. Figure 1e is the composite speech derived from the above frames, and figure 1f is the spectral autocorrelation of the composite speech. One important observation concerning figure 1f is that the ratio of the first local maximum after the one at lag 0, to the local minima between this maximum and the next local minimum, was significantly lower than that of the single speaker case. This is due to the fact that there are significant autocorrelation values for lags that are not harmonically related, due to co-channel conditions. This motivated us to define a spectral autocorrelation ratio, which reflects the extent of corruption of a target speech by the interfering speech.

The Spectral Autocorrelation Ratio (SAR) parameter is defined as follows:

$$SAR = 20 \text{LOG}_{10} \{R(p_1) / R(q_1)\}$$

where, $R(p_1)$ is the local maximum of spectral autocorrelation other than the one at lag 0 (occurring at lag p_1) and $R(q_1)$ is the next local maximum that is not harmonically related to the first peak, or the local minimum between p_1 and $2p_1$.

The SAR has to be properly interpreted. If speech of one of the speakers is silent or is unvoiced, a peak that is not harmonically related to the peak due to voicing state of one talker will be substantially lower in amplitude. This means the SAR will be very high, from which we would conclude that the frame of speech is usable. If, however, the speech of target and interferer were of comparable magnitude, the SAR ratio would approach zero, which would identify that particular frame as unusable. What if

there is a spurious peak of comparable magnitude along with the harmonically related pulses? The SAR will again be low, but the physical interpretation is that, a pure tone is mixed with the speech signal, and if it is of comparable magnitude, that speech frame is definitely unusable.

3. EXPERIMENTS AND RESULTS

The speech data was obtained from the TIMIT database. The original speech was sampled at 16 kHz, and re-sampled to 8 kHz after low-pass filtering to 3 kHz. The target speech and the corrupting speech were scaled and added so that the overall TIR was 0 dB. The segmental TIR varied from approximately -40 dB to +47 dB.

The TIR of the composite speech was computed on a frame-by-frame basis. The frame size was 20 ms and was hamming windowed prior to computing the magnitude spectra and the corresponding spectral autocorrelation ratios. The SAR was computed for each frame as discussed in section 2. 256 frames were processed and the results are shown in figure 2 and table I.

83 frames were flagged usable by setting the TIR threshold as 20 dB and 89 frames were flagged usable by setting the SAR threshold as 6.23 dB. 47 frames were flagged usable by both the TIR and SAR thresholds.

At first sight, it appears that only 57% of the usable frames (47 out of 83) as flagged by the TIR threshold are detected by the SAR algorithm. However, in real-life situations usable frames occur in clusters rather than single isolated frames [2]. If we consider the clusters instead of the number of frames, the TIR method spotted 16 usable clusters. The SAR algorithm also spotted 16 usable clusters. Out of these 16 usable clusters spotted by the SAR algorithm, 15 matched those spotted by the TIR threshold. If at least two consecutive frames are present in a cluster spotted by both TIR and SAR method, we considered that to be a match.

4. SUMMARY

The purpose of this paper was to identify the usable portions of co-channel speech in the context of speaker identification. It was found that the Spectral Autocorrelation Ratio is a useful measure in spotting those usable frames. Further improvements in this algorithm are possible, to make the performance more robust. One possible improvement is to process only those frames that are voiced (i.e., at least one speaker's speech is voiced). It is also worthwhile to investigate why the SAR method picks some frames adjacent to those declared usable by the TIR threshold condition.

ACKNOWLEDGEMENT

Effort 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.

REFERENCES

- [1] 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.
- [2] Yantorno, R.E., "Co-channel speech study", Final report for Summer Research Faculty Program, Research Laboratory AFRL/IF, Speech Processing Lab, Rome Labs, New York, 1999.
- [4] Schroder, M.R., "Period histogram and product spectrum: new methods for fundamental frequency measurements", *J. Acoust. Soc. Am.*, 43, 829-834 (1968).
- [5] Naylor, J.A. and Boll, S.F., "Techniques for suppression of an interfering talker in co-channel speech", *Proc. ICASSP*, 205-208, 1987.
- [6] Chazan, Y., Stettiner, Y., and Malah, D., "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation", *Proc. ICASSP*, II-728 - II-731, 1993.
- [7] Link A., L. Trahms, L., Oeff., M. and Steinhoff, U., "Heart rate variability determined as heart frequency deviation", *Proc. of 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1648-1649, vol. 4, 1997.
- [8] Shariati, M.A., Dripps, J.H. and McDicken, W.N., "A comparison of color flow imaging algorithms", *Physics in Medicine and Biology*, vol. 38, no. 11, 1589 - 1600, 1993.
- [9] Ashihara, K. and Kado, H., "Separation of speech from noise by using linear prediction in frequency domain", *Journal of Acoustical Society of Japan*, vol. 5, no. 11, 821-828, 1995.
- [10] Parsons, T.W., "Separation of speech from interfering speech by means of harmonic selection", *J. Acoust. Soc. Am.*, Vol. 60, No. 4, 911-918, 1976.

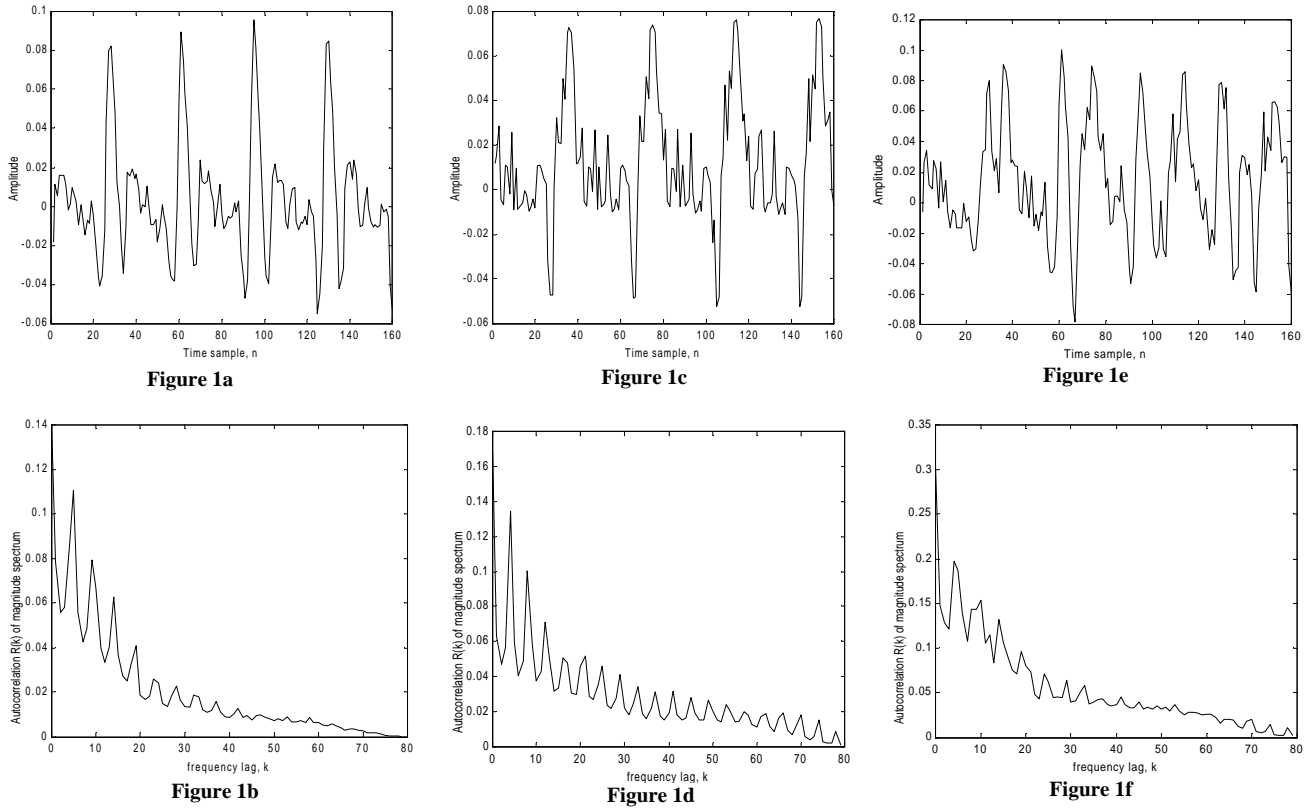


FIGURE 1. Frames of voiced speech from two speakers figures 1a and 1c and corresponding Spectral Autocorrelations figures 1b and 1d. Frame of co-channel (combination of figures 1a and 1b) figure 1e, and corresponding Spectral Autocorrelation of the co-channel speech figure 1f.

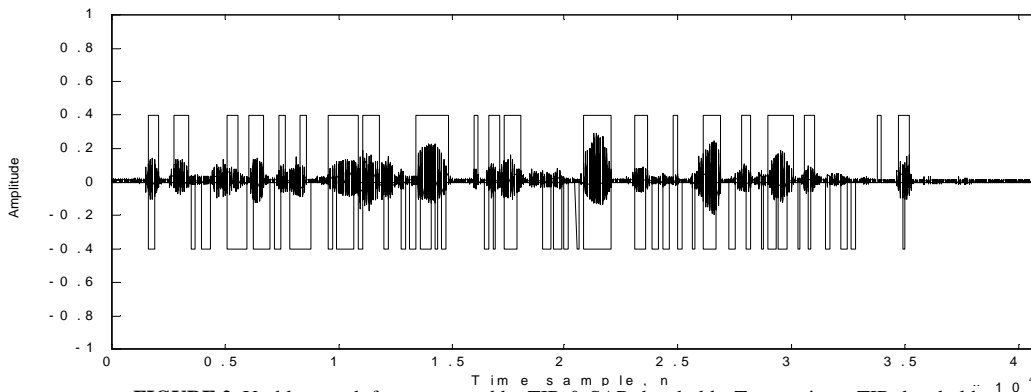


FIGURE 2. Usable speech frames spotted by TIR & SAR thresholds. Top portion – TIR threshold 20 dB. Bottom portion – SAR threshold 6.23 dB.

TABLE 1. Frames detected using either TIR threshold of 20 dB or SAR threshold of 6.23 dB

| | |
|--|-----|
| Total frames | 256 |
| Frames declared usable by TIR threshold 20 dB | 83 |
| Frames declared usable by SAR threshold 6.23 dB | 89 |
| Frames declared usable by both TIR & SAR thresholds | 47 |
| Total useful clusters spotted by TIR threshold | 16 |
| Total useful clusters spotted by SAR threshold | 16 |
| Total useful clusters spotted by both TIR & SAR thresholds | 15 |