USABLE SPEECH DETECTION USING THE MODIFIED SPECTRAL AUTOCORRELATION PEAK TO VALLEY RATIO USING…

Article

2 authors, including:

USABLE SPEECH DETECTION USING THE MODIFIED
SPECTRAL AUTOCORRELATION PEAK TO VALLEY RATIO
USING THE LPC RESIDUAL

Nishant Chandra, Robert E. Yantorno
Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, PA 19122-6077, USA
cnishant@temple.edu, robert.yantorno@temple.edu
http://www.temple.edu /speech_lab

ABSTRACT

Many applications of speech communication and speaker identification suffer from the problem of interfering or co-channel speech. The proposed method is to find the segments of co-channel speech, which can be identified as "usable" and could then be processed by a speech processing system. The Spectral Autocorrelation Peak to Valley Ratio (SAPVR) of the Linear Predictive Coding Residual (SAPVR-Residual) is a modified version of the original SAPVR used for detection usable speech segments in co-channel speech. Results obtained using the SAPVR-Residual, was useful in spotting approximately 71% of usable segments, with a corresponding false alarm rate of 37%. This method of identification of usable speech represents the front-end process of a next generation co-channel speech processing system involving an information fusion/decision system whose final goal is speech separation.

**KEY WORDS:** Co-channel speech, usable speech, SAPVR, LPC residual.

1. INTRODUCTION

Previous work [1] has shown that there exist segments of speech which can be identified as usable in the presence of interfering or co-channel speech. These segments can be labeled as usable and could be processed by a speech processing system, see Figure 1. A number of methods to identify usable speech segments have been developed [2, 3, 4]. Usable speech measures have also been used to detect co-channel speech, which could provide information to speech processing system, such as a speaker identification

Figure 1: Application of Co-channel Speech Detection System

system or speech recognition system, to suspend the operation of those speech-processing systems, whose operation would be degraded if it were processing co-channel speech [5, 6].

Usable speech is a novel approach to the co-channel speech processing problem, where the concept is to identify and extract those portions of co-channel speech which could be used for such speech processing operations as speaker identification or speech recognition, which suffer from sever operational degradation when processing co-channel speech. For example, a speech segment is "usable" if it contains enough information to identify the target speaker. A recent study [7] revealed that as much as 38% of a co-channel speech utterance has enough information about the target speaker to perform reliable speaker identification even when the Target-to-Interferer (TIR) ratio is 0 dB [averaged over the entire utterance]. It was also found that as much as 32% of a co-channel speech utterance contained enough information about the interfering speech such that the interferer's identity could be identified [7, 8, 9]. Hence, those segments become "usable" for a speaker identification system. Consequently, if one wishes to extract the identity of both the target and the interferer, as much as 70% of the entire speech is available [7]. However, the amount of usable speech gleaned from a co-channel utterance depends heavily upon the nature of the speech, i.e. whether it contains many pauses or is relatively continuous speech. The typical 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 (TIR) ratio is a reasonable lower limit for speaker identification to work reliably [5, 7]. 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 [10, 11]. 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
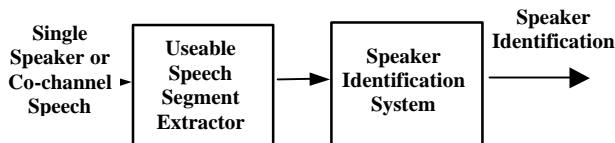
to the energy content of all other components (both noise as well as harmonic energy content of the weaker talker) is a good measure of usability of that speech frame.

An important characteristic of LPC parameters is that they preserve essentially the intelligibility information of the speech signal. LPC analysis allows one to extract the vocal tract parameters, which then can be used to perform inverse filtering, in which the output, which is the residual, does not have any of the vocal tract information.

In LPC, the vocal track is modeled as an all pole digital filter that can be expressed mathematically as:

$$H(z) = \frac{G}{1 + a_1 z^{-1} + a_2 z^{-2} \dots\dots + a_p z^{-p}} = \frac{S(z)}{E(z)} \qquad (1)$$

Where, p is the order of the model. G is the gain, s(n) is the speech output of the model, and e(n) is the excitation input. The equation above can be written in the time domain as:

$$S(n) = Ge(n) - a_1 s(n-1) - a_2 s(n-2) \dots\dots - a_p s(n-p) \qquad (2)$$

In other words, every speech sample is computed as a linear weighted sum of previous speech samples plus the excitation.

The LPC method is most accurate when it is applied to stationary signals. To be able to apply LPC to speech segments, we segment speech into quazi-stationary frames using a hamming window.

2. SAPVR USING LPC RESIDUAL APPROACH

We are proposing a modification to an approach previously developed for detection of co-channel speech [2]. In that approach the SAPVR was calculated using the speech segments. The present approach is to use the LPC residual, which helps to remove the vocal track effect resulting in the residual being highly periodic and much flatter in the frequency domain than the corresponding speech signal's spectrum.

It is our goal to select usable frames of speech using SAPVR-Residual measure without having any a priori information about the energy of either speaker. The TIR measure was used as a benchmark for the proposed measure. A successful identification of usable speech occurs when both the TIR and SAPVR-Residual

methods select a frame of co-channel speech as usable. A missed identification is said to occur when the TIR measure has selected a frame that has not been selected as useable by the SAPVR-Residual measure. A false alarm is said to occur when the SAPVR-Residual measure has selected a frame that has not been selected by the TIR measure.

Usable speech, which is composed almost entirely of voiced speech, has a periodic nature [2]. Shown at the top of Figure 2 is a 40ms segment of usable speech. Due to the periodicity of usable speech, the excitation of speech also has a periodic structure, which is evident in Figure 2 (b). Figure 2 (c) shows the FFT of the residual. Figure 2(d) is obtained by performing the autocorrelation on the FFT of residual of Figure 2 (c). Definite peaks and valleys can be seen in this subplot – as identified by the dots, which are used to compute the SAPVR. The SAPVR is calculated as the ratio of the sum of twice the first peak and next four peaks to the first valley.
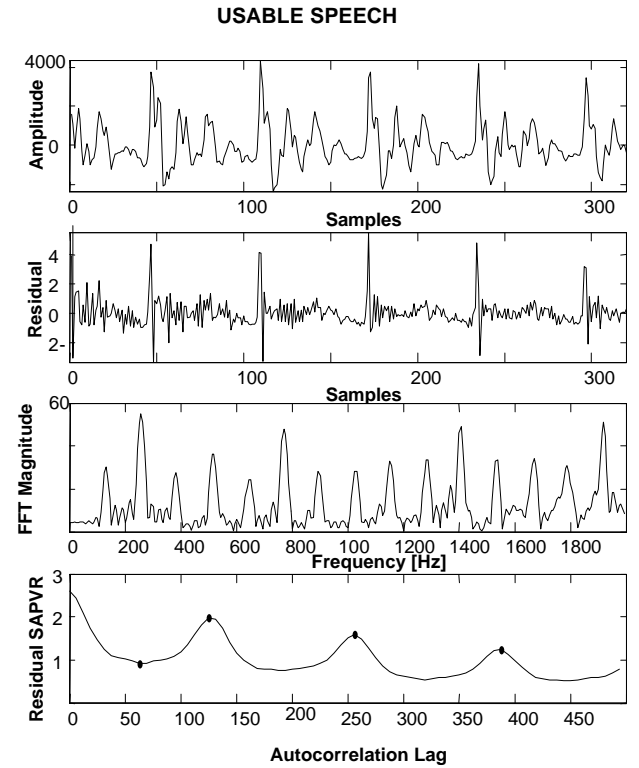
USABLE SPEECH

Figure 2: Modified SAPVR Approach, (a) Speech segment (top panel), (b) Residual of Usable Speech (second panel down), (c) FFT of residual (third panel down), (d) Spectral autocorrelation (bottom panel).

The SAPVR-Residual measure is used to detect the structure of the spectral autocorrelation. This structure is illustrated in Figure 2 for the residual, 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 for voiced speech. This loss of structure is shown in Figure 3. It can be seen that there are no good peaks and valleys in Figure 3 (d), and therefore its peak to valley ratio falls below the threshold indicating that this frame is unusable.
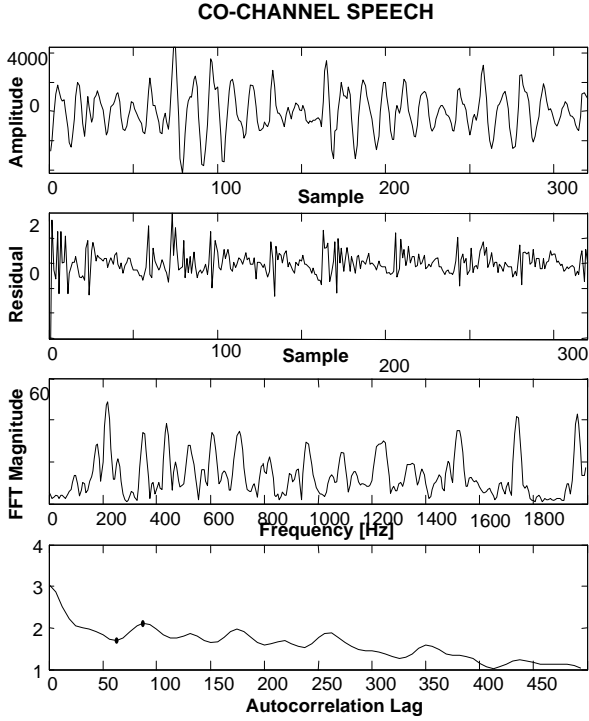
CO-CHANNEL SPEECH

Figure 3: Modified SAPVR Approach, (a) Co-channel Speech segment (top panel), (b) Residual of Unusable Speech (second panel down), (c) FFT of residual (third panel down), (d) Spectral autocorrelation (bottom panel).

3. EXPERIMENTS AND RESULTS

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.

A very important step in the process of developing a good measure is to select the proper threshold. SAPVR-Residual measure is plotted against its probability of occurrence in order to facilitate in determining an optimal threshold. As seen in Figure 4, a threshold of 6.6 results in a large number of hits and minimal number of false alarms. When reviewing the experiments, it was observed that false alarms and
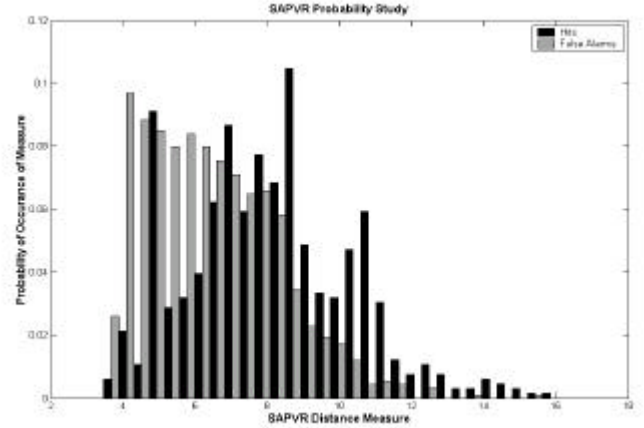
Figure 4: Probability study of SAPVR distance measure.

missed frames occur mostly in transition regions (onset and offset of voicing).

The result of a single experiment is shown in Figure 5. The TIR of composite speech was computed for each speech frame. The dashed rectangles in Figure 5 indicate detection of usable speech. Frames of speech where the TIR is at least 20 dB are shown in black. The gray sections of the co-channel utterance in Figure 5 are those frames whose TIR are less than 20 dB and therefore are considered unusable.

For the experiment shown in Figure 5, using both the TIR threshold at 20 dB and the SAPVR-Residual threshold at 6.6 resulted in an average of 71% of the frames as usable, being detected indicating correct identifications. Also, an average of 37% of the frames were flagged usable by the proposed measure but had a TIR below 20 dB, thereby indicating false alarms.

Table1: Comparison of Results of SAPVR-Speech and SAPVR-Residual Based Co-channel Speech Detection systems

Co-channel speech	% Correct		% False	
	SAPVR Speech	SAPVR Residual	SAPVR Speech	SAPVR Residual
Male-Male	51	62	27	29
Female-Female	83	81	61	50
Female-Male	72	70	40	33
Average	69	71	43	37

Table 1 presents the results of performing SAPVR on Speech segments (SAPVR-Speech) versus SAPVR-Residual. A total of 42 speech files were used for performing the experiment. It can be seen that for the Male-Male case the SAPVR-Residual system a gives much higher percent correct detection of co-channel speech. It should be noted that in the case of Female-Female and Female-Male, there was minimal change in the percent correct for the SAPVR-Residual versus SAPVR-Speech measure. However there was a sizable decrease in false alarms.

4. SUMMARY

The purpose of the research presented here was to identify the usable portions of co-channel speech. It was found that the SAPVR-Residual is a useful measure, spotting approximately 71% of those usable speech segments compared to 69% for SAPVR-Speech method. Also a false alarm rate of 37% compared with 43% for SAPVR-Speech method was obtained.

Further improvements in this algorithm are possible, to make the performance more robust. One possible improvement is to study false alarm with respect to their TIR. It might be interesting to look at how many frames, labeled as false alarms, are close to 0 dB TIR, and how many are close to 20 dB. Those frames that are close to 0 dB TIR pose a bigger problems for speaker identification.

5. ACKNOWLEDGEMENT

6. 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.

7. REFERENCES

1. 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.

2. 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". IEEE International Symposium on Intelligent Signal Processing and Communication Systems 2000, pp: 710-713, November, 2000.

3. Lovekin, J., Krishnamachari, K. R., Yantorno, R. E., Benincasa, D. S., and Wenndt, S. J., "Adjacent Pitch Period Comparison (APPC) as a Usability Measure of Speech Segments Under Co-channel Conditions". IEEE International Symposium on Intelligent Signal Processing and Communication Systems, pp:139-142, November, 2001.

4. Krishnamachari, K. R., Yantorno, R. E., Lovekin J. M., Benincasa, D. S., and Wenndt, S. J., "Use of Local Kurtosis Measure for Spotting Usable Speech Segments in Co-channel Speech." ICASSP 2001, pp:649-652, May 2001.

5. Yantorno, R. E., Krishnamachari, K. R., Lovekin, J. M., Benincasa D. S., and Wenndt, S. J., "The Spectral Autocorrelation Peak Valley Ratio (SAPVR) – A Usable Speech Measure Employed as a Co-channel Detection System". IEEE Workshop on Intelligent Signal Processing, pp: 193-197, Hungary, May, 2001.

6. Kizhanatham, A., Yantorno, R.E., "Co-Channel Speech Detection Approaches Using Cyclostationarity or Wavelet Transform", IASTED SIP 2002 (submitted).

7. 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.

8. Smolenski, B. Y., Yantorno, R., E., Benincasa D. S., and Wenndt, S. J., "Co-Channel Speaker Segment Separation", ICASSP 2002 (accepted).

9. Lovekin, J., Yantorno, R. E., Benincasa, S., Wenndt, S., and Huggins, M., "Developing Usable Speech Criteria for Speaker Identification", ICASSP 2001, pp:421-424, May 2001.

10. Yumoto, E. and Gould, W. J., "Harmonics-to-noise ratio as an index of the degree of hoarseness", J. Acoust. Soc. Am., vol. 71, No. 6, pp: 1544-1550, 1982.

11. Krom, G. de, "A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals", American Speech-Language-Hearing Association, pp: 254-265,1993.
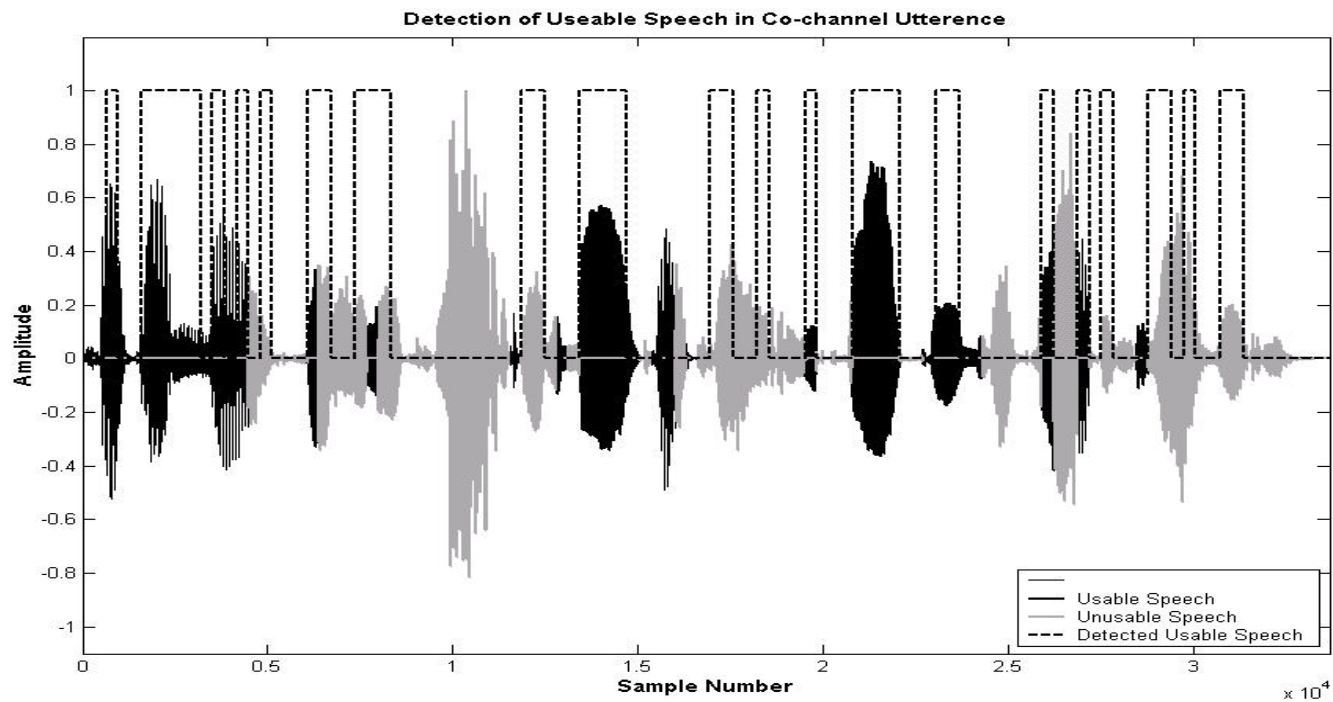
Figure 5: Usable Speech Detected by TIR & SAPVR of Residual Thresholds. TIR Usable Speech (black), TIR Unusable Speech (gray), Detected Usable Speech (dashed box)