

ADJACENT PITCH PERIOD COMPARISON (APPC) AS A USABILITY MEASURE OF SPEECH SEGMENTS UNDER CO-CHANNEL CONDITIONS

Jereme Lovekin, Kasturi Rangan Krishnamachari and Robert E. Yantorno
Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, Pa 19122-6077, USA
jlovekin@temple.edu, kkrish01@astro.temple.edu, robert.yantorno@temple.edu
http://www.temple.edu/speech_lab

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

Previous work [1] has shown that there exist segments of speech which can be identified as usable in the presence of noise. These segments can be labeled as usable and processed by a speech processing system. Some methods to identify "usable" speech segments have been developed [2], [3]. Here, we present a method which performs a comparison of adjacent pitch periods within voiced speech to quantify the usability of speech which has been corrupted by another speech signal. It was found that the method is useful in spotting approximately 75% of usable segments for speaker identification. False alarms and missed frames occur mostly in transition regions (onset or offset of voicing). In scenarios with limited training and testing data, a system which separates out those "usable" portions of speech is desirable. The technique proposed here, i.e., the identification of "usable" speech, represents the front-end process of a next generation speech processing system. This process involves an information fusion/decision system which will utilize both time as well as frequency domain information to determine the usability of a frame of speech.

1. INTRODUCTION

Until recently, 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 to 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 can be 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.

A speech segment is "usable" if it contains enough information to identify the target speaker. A recent study [4]

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 [4]. 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. 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 normal situation with 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 [4], [5]. 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 [5]. 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.

2. ADJACENT PITCH PERIOD COMPARISON

It is our goal to select usable pitch periods of speech using the Adjacent Pitch Period Comparison (APPC) measure without having any *a priori* information about the energy of either speaker. The TIR measure will be used as a benchmark for the proposed APPC measure. A successful identification of usable speech occurs when both the Adjacent Pitch Period Comparison and Transmitter to Interferer Ratio methods select a pitch period of co-channel speech as usable for speaker identification. A missed identification is said to occur when the TIR measure has selected a pitch period that has not been selected by the APPC measure as usable. A false alarm is said to occur when the

APPC measure has selected a pitch period that has not been selected by the TIR measure as usable.

Usable speech, which is composed entirely of voiced speech, has a periodic nature [2]. Due to the periodicity of usable speech, adjacent pitch periods of voiced speech are similar in ‘shape,’ which is evident in *Figure 1*. Shown at the top of *Figure 1* is approximately 20 ms. of usable speech. The length of each reference pitch period is determined by finding the distance between the zero-lag point and the next highest peak of the Autocorrelation matrix of the next 10 ms. (about the length of 3 pitch periods) of the co-channel speech. In this case, the length of the reference pitch period is selected to be 57 samples, corresponding to the first 57 samples of speech shown at the top of *Figure 1*. The adjacent pitch period begins immediately following the reference pitch period at sample 58 and ends at sample 115 (58+57). This method assumes that the any change in length from one pitch period to its neighboring pitch period is negligible. The bottom of *Figure 1* shows the amplitude of the reference and adjacent pitch periods superimposed. The APPC measure attempts to exploit the shape similarity of adjacent pitch periods within usable speech. By comparing the sample-by-sample amplitude variations of adjacent pitch periods, as shown in the bottom of *Figure 1*, a measure has been developed to identify usable speech segments within co-channel speech. First, this method requires accurate detection of the pitch periods. Using an autocorrelation method, the length of a pitch period is determined. A comparison of adjacent pitch periods with single speaker voiced speech results in minimal sample-by-sample amplitude variations. The measure is determined by summing these variations over the reference and adjacent pitch periods. The length of each pitch period is easily determined for a single speaker.

In the case of two simultaneous voiced speakers, a comparison of the adjacent pitch periods will yield large differences, because there will be peaks from interfering voiced speech, as shown in *Figure 2*. Also, with co-channel speech, pitch period estimation could be erroneous. Fortunately, this can work to our advantage. If pitch periods are ill-selected as in the co-channel speech case, dissimilar pitch periods will be compared and a very large distance measure will result. Therefore, selection of proper pitch periods in the single speaker voiced speech situation is essential for this method to operate accurately, while it is not of particular importance when there is strong interfering voiced speech.

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 TIR of the composite speech was computed for each pitch period, as discussed in section 2. Each pitch period ranged from 32 to 96 samples, equating to 4-6 ms.

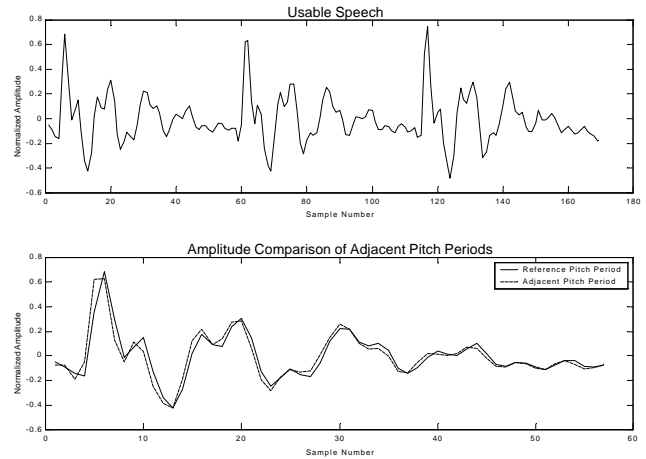


Figure 1: Usable Speech (top) and Adjacent Pitch Period Amplitude Comparison (bottom)

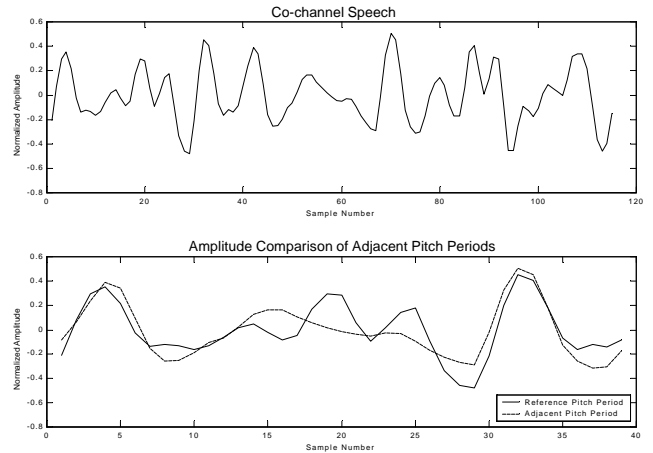


Figure 2: Co-channel Speech (top) and Pitch Period Amplitude Comparison (bottom)

The binary-valued line above the speech in *Figure 3* indicates detection of usable speech, where a peaks indicates where the APPC measure has detected usable speech. Pitch periods (frames) of speech where the TIR is at least 20 dB or the ITR (Interfere-to-Target Ratio) is 20dB are shown in black. The gray sections of the co-channel utterance in *Figure 3* do not meet the 20 dB criteria and are therefore considered unusable for speaker identification.

In this two speaker set, both the TIR threshold at 20 dB and the APPC threshold at 6 flagged 74.5% of the pitch periods as usable, indicating correct identifications. This also indicates that 25.5% of the pitch periods deemed usable by the TIR measure were missed by the APPC measure. In addition, 24.4% of the unusable pitch periods were flagged usable by the APPC measure, indicating false alarms.

Missed usable speech detections occur most often at the onset and offset of voiced speech, due to the transition that occurs at the beginning or end of a voiced speech utterance. When amplitudes of adjacent pitch periods are compared at the onset of voiced speech, there may be a large amplitude difference between the samples of the adjacent pitch period and the reference pitch period, even though they may be the same 'shape.' The same phenomenon may be observed at the end of the voiced speech as well. It is possible to partially overcome this problem by normalizing each pitch period before comparison. This way, the overall amplitude difference will have less of an effect on the point-to-point amplitude comparison. A small reduction (about 2%) was realized in the false alarm rate when performing the normalization procedure.

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 Adjacent Pitch period Comparison method is a useful measure in spotting approximately 75% of those usable portions. It was found that normalizing the overall amplitude of each pitch period before performing a point-to-point comparison slightly reduces the missed detection rate. 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) in order to improve computational efficiency. It is also worthwhile to investigate why the APPC method picks some frames declared unusable by the TIR method. Also, comparing features other than amplitude variations is currently underway, and thus far Dynamic Time Warping appears to be a promising candidate.

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 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. Intelligent. Sig. Process. and Comm. Sys., pp: , Nov. 2000.
- [3.] 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.
- [4.] 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.
- [5.] 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.
- [6.] 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.

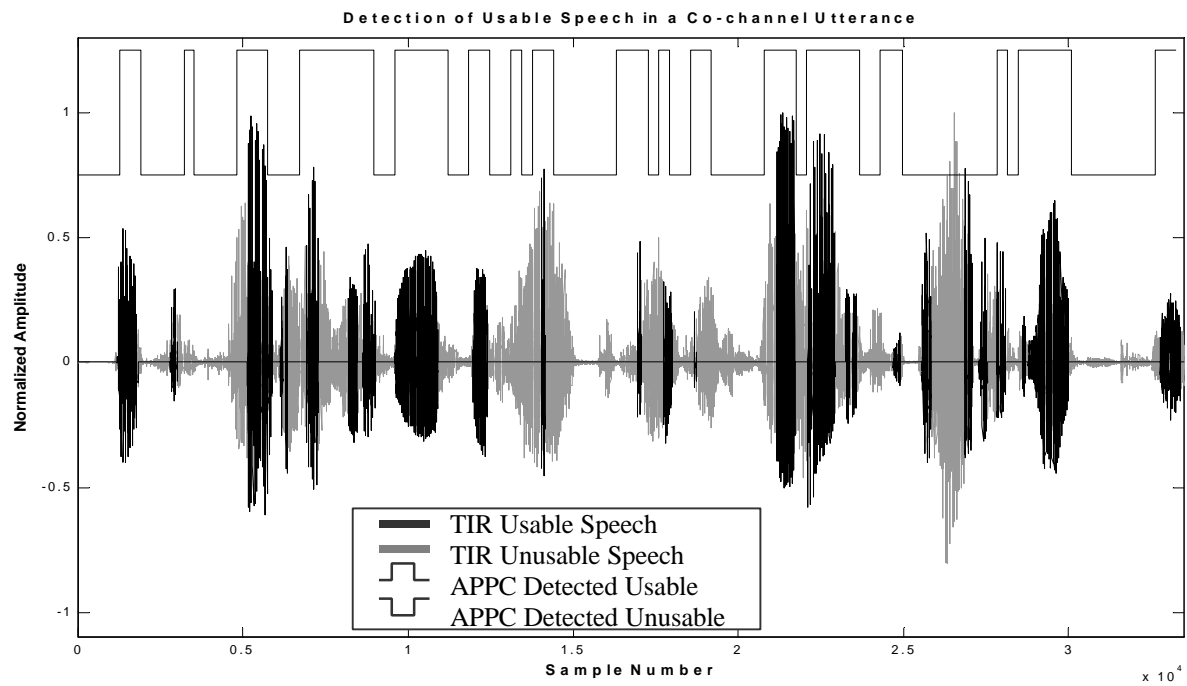


Figure 3: Usable Speech Detected by TIR & APC Thresholds. TIR Usable Speech (black), TIR Unusable Speech (gray), APC Detected Usable Speech (black box).