

## Structural Usable Speech Measure Using LPC Residual

Ananth N. Iyer, Melinda Gleiter, Brett Y. Smolenski and Robert E. Yantorno

Speech Processing Laboratory, Temple University

12th & Norris Streets, Philadelphia PA 19122-6077 USA Tel: 215-204-6984

E-mail: ananth.iyer@temple.edu, mgleiter@temple.edu,

bsmolens@temple.edu, robert.yantorno@temple.edu

**Abstract:** In an operational environment speech is degraded by many kinds of interferences. The operation of many speech processing techniques are plagued by such interferences. Usable speech extraction is a novel concept of processing degraded speech data. The idea of usable speech is to identify and extract portions of degraded speech that are considered useful for various speech processing systems. The performance reduction of speaker identification systems under degraded conditions and use of usable speech concept to improve the performance has been demonstrated in previous work. A new usable speech measure, based on the structure of Linear Predictive Coding (LPC) residual is developed to identify usable speech frames. It is shown that the method has 72% success in identifying the usable frames with 28% false rate.

### 1. Introduction

Speaker identification system performance is plagued by many kinds of situations in an operational environment. Of particular interest is when two or more people are talking at the same time over the same channel, known as co-channel speech. Traditional methods of co-channel speech processing have been to enhance the prominent speaker (target), suppress the interfering speaker speech or both [1]. However, previous studies on co-channel speech have shown that it is desirable to process only portions of the co-channel speech which are minimally degraded [2]. Such portions of speech considered usable for speaker identification are referred to as "usable speech". The concept of usable speech can also be extended to other applications such as automatic speech recognition [3].

Recently, a usable speech extraction system was proposed to classify co-channel speech frames into usable speech and unusable speech for speaker identification [4]. The speech segments can be declared usable based upon a Target-to-Interferer energy ratio (TIR) threshold. Studies indicate that by considering only usable speech segments for speaker identification, there is a considerable increase in the accuracy of speaker identification [5]. The block diagram of the usable speech extraction system and speaker identification is shown in Figure 1. The segment separator block [6] is employed to distinguish frames of speech data belonging to each

speaker.

Several usable speech measures are being developed to identify usable speech frames [7] [8] [9] [10] [11]. These measures are used in a fusion/decision system to reliably identify usable speech frames [12] [13] [14] by harnessing the complementary information provided by the usable speech measures.

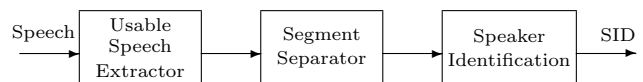


Figure 1. Block diagram showing the application of usable speech detection

Hence to achieve higher identification accuracy, one needs to fuse many measures providing complimentary information. A new usability measure for spotting the usable speech segments from a co-channel utterance is proposed here. This measure is based on the fact that there is a significant change in the excitations of speech belonging to the two classes in co-channel speech, i.e., usable speech and unusable speech.

### 2. Background

#### 2.1 Linear Prediction

The composite spectrum effects of radiation and vocal tract is assumed to be represented by a time-varying digital filter whose steady-state system function is of the form

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^p \alpha_k z^{-k}} \quad (1)$$

where  $\alpha_k$  are the digital filter coefficients. This all-pole system is excited by the voice source to produce speech data. A linear predictor with predictor coefficients,  $\alpha_k$  is defined as a system whose output is

$$\hat{s}(n) = \sum_{k=1}^p \alpha_k s(n-k) \quad (2)$$

The all-pole linear predictor models a signal  $s_n$  by a linear combination of its past values. The LPC analysis is performed to determine the predictor coefficients  $\alpha_k$ , which represents the vocal tract model of the speaker, directly from the speech signal.

## 2.2 Minimum Description Length

Under co-channel conditions one would expect the model order of the LPC analysis needed to be twice that used for a single speaker speech. Hence the model order for LPC analysis is dynamically selected for each frame of speech. Principle of Minimum Description Length (MDL) is employed to choose the model order that gives the shortest description of the speech frame data [15]. The model order  $i$  is determined by using the equation given below. The value of  $i$  is incremented and the order which leads to the lowest MDL is chosen as the order for LPC analysis.

$$M_i = N \log_{10}(P_s)^2 + i \log_{10}(N) \quad (3)$$

where  $N$  is the speech frame size and  $P_s$  is the residual power.

## 2.3 LPC Residual of Co-channel Speech

The LPC residual is spectrally flattened speech, where the vocal tract shape is removed by creating an inverse filter using the LPC coefficients. Figure 2, below shows a typical usable frame, which is voiced and having a high TIR value, exhibiting good periodic structure (upper panel). The bottom panel of the figure is the LPC residual of the speech frame. It can be seen that the periodic structure of usable speech is represented as pulses with constant frequency in the LPC residual of speech [16].

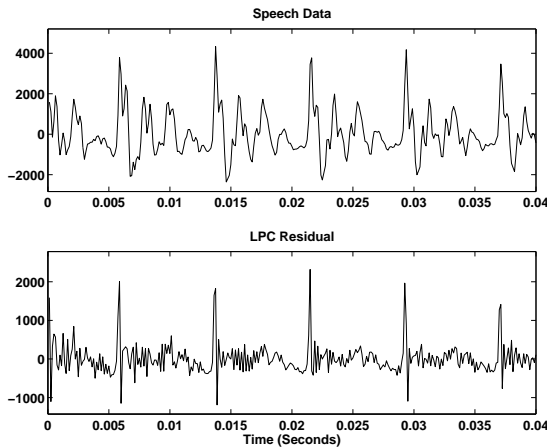


Figure 2. Usable speech frame (top panel) and the corresponding LPC residual of speech frame (lower panel) showing good structure

Figure 3 shows, an example of an unusable (mixed speech data from two speakers) speech frame (upper panel). The frame of speech shown is voiced and has low TIR value. It is seen that the speech data has no defined periodic pattern. The lower panel shows the LPC

residual of the unusable frame. The LPC residual shows no presence of definite pulses and is unstructured. The property of usable speech residual having a good periodic structure and unusable speech having no definite structure was used to develop a usable speech measure.

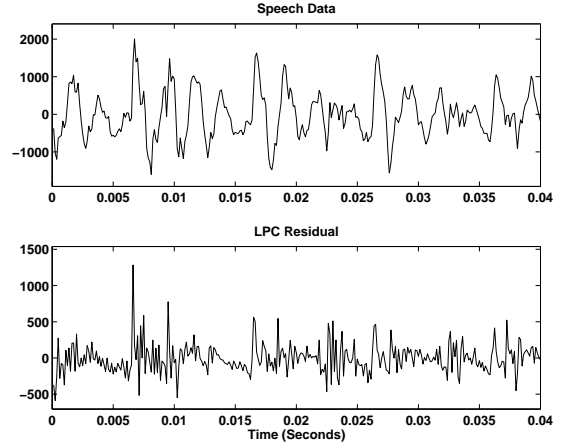


Figure 3. Unusable speech frame (top panel) and the corresponding LPC residual of speech frame (lower panel) showing no presence of structure

## 3. Algorithm Description

The presence of equidistant pulses in the residual suggested the possibility of measuring the distances between the pulses and use the distance to measure the periodic structure of speech frame. The various steps to obtain the distances is explained below -

1. Processing of data was performed on a frame-by-frame basis. A frame of length 32ms equal to 240 samples for speech data sampled at 8kHz, from the co-channel utterance was used. The frame was determined to be - voiced, unvoiced or silence. Unvoiced and silent frames were rejected from further processing. Voiced/unvoiced decision was done using the Spectral Flatness Method [17] and silence decision was done using an energy threshold.
2. LPC model order was determined based on minimum description length.
3. The LPC coefficients were used to create an inverse filter and the speech frame was filtered using the inverse filter. The output of the inverse filter is the LPC residual.
4. An intelligent peak-picking algorithm was designed to determine the prominent peaks from the residual frame. The peak-picker was implemented based on a sorting method and selecting the peaks which satisfies three conditions - a) the magnitude of the peak should be at least 40% of the highest peak; b) The distance between the peaks picked should be greater than 30 time points

(3.7ms); and c) the distance between the peaks picked should be smaller than 100 time points (12.5ms). If any of the peaks picked did not satisfy these conditions, it was dropped. The peak-picking algorithm was run iteratively until only peaks satisfying the above conditions were picked.

5. The distances between the peaks picked in the previous step were determined. These distances represent the pitch period of the speech frame. Hence the distances in male speech would nearly be twice that of a female speech. The peak distances are normalized such that the distances for both male speech and female speech lie in the same range. The normalized distances is the product of the individual distances between the peaks and the number of peaks in the frame. For example, female speech with pitch 200 Hz or 40 points and 7 peaks would result in a value of 280, and male speech of pitch 100 Hz or 80 points and 3 peaks would result in a value of 240.

6. The average of the normalized peak distances in each frame was computed and used as a usable speech measure. A threshold was determined based on the constructed probability density function.

#### 4. Experimental Results

The LPC residual peak based usable speech measure developed above was tested on co-channel data obtained from the TIMIT database. To determine the optimum threshold for the proposed usable speech measure, the probability density function of the measure values was generated and is shown in Figure 4.

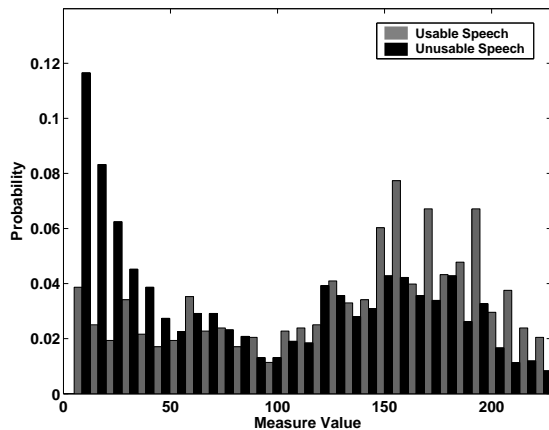


Figure 4. Probability density function - The mean value of normalized peak distances in each frame. Black bars represent unusable data and gray bars represents usable data.

For the Figure 4 above, the gray bars in the probability density function represent measure value for the usable speech frames, i.e., frames with TIR > 20dB or

TIR < -20dB [5] and the black bars represent the measure values for unusable speech frames. From the probability density function, a threshold value of 100 was considered optimum.

To determine the effectiveness of the proposed measure in spotting usable speech frames, the percent correct identifications (hits) and percent incorrect identifications (false alarms) was determined. The measure is said to have a hit if, the measure as well as TIR identify a frame as usable. The measure is said to have a false alarm if a frame is identified as usable by the measure, but unusable based on TIR. The proposed measure has resulted in 72% correct decisions or hits and 30% incorrect decisions or false alarms. These values were obtained on a set of 861 co-channel utterances. These utterances were created using 21 male speech and 21 female speech files.

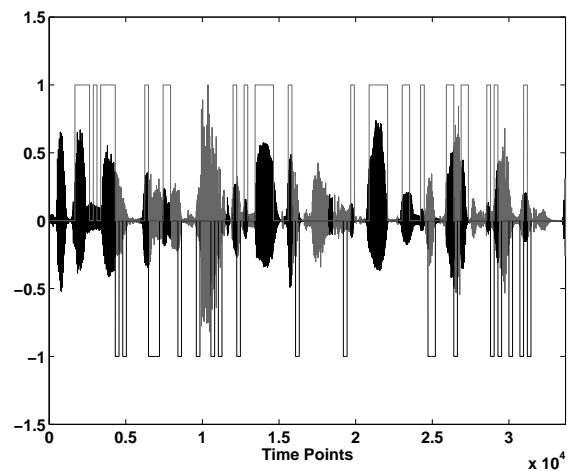


Figure 5. Usable speech detected by TIR and usable speech measure. TIR usable speech (black), TIR unusable speech (gray), Usable speech measure detected frames (black box).

The rectangles above the speech in Figure 5 indicates detection of usable speech, a value of +1 indicates a hit in identifying usable speech and a value of -1 indicates a false alarm. Portions of speech having a TIR value of 20dB or above are shown in gray. The portions of the co-channel utterance in black unusable speech frames. A closer look at the false alarms in the usable speech measure, reveals that many frames lie in a transition region - from voiced to unvoiced/silence. This suggests that the false alarm rate can be reduced if such transition frames are identified and eliminated from processing.

The identified usable speech frames were extracted and tested in a speaker identification system, which uses LPC-Cepstrum features and a Vector Quantization scheme to perform the classification. The system was tested using the extracted usable speech frames and the

resulted in 82% correct identification compared to 45% correct detection, with co-channel speech.

## 5. 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 structural changes in the LPC residual of speech can be used as an effective method to identify usable speech segments. The proposed usable speech measure detected 72% of the usable portions from co-channel data. Further improvements in this algorithm are possible, to make the performance more robust. One possible scenario would be to identify and remove outlier frames such as those lying in transition regions. It is also expected that the performance of usable speech extraction system would improve with the inclusion of this new measure. Further, the LPC coefficients can be potentially used to develop a model based usable speech measure.

## Acknowledgements

The Air Force Research Laboratory, Air Force Material Command, and USAF sponsored this effort, under agreement number F30602-02-2-0501. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Further we wish to extend thanks to Nithya Sundaram for providing information about the principle of minimum description length for model order selection.

## 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] J. A. Naylor and S. F. Boll, "Techniques for suppression of an interfering talker in co-channel speech," *Proc. IEEE ICASSP*, pp. 205–208, 1987.
- [2] R. E. Yantorno, "Co-channel speech study, final report for summer research faculty program," tech. rep., Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, New York, 1999.
- [3] W. T. Hicks, B. Y. Smolenski, and R. E. Yantorno, "Testing for intelligibility of corrupted speech with an automated speech recognition system," *IIS Systemics, Cybernetics and Informatics*, 2003.
- [4] R. E. Yantorno, "Co-channel speech and speaker identification study," tech. rep., Air Force Office of Scientific Research, Speech Processing Lab, Rome labs, New York, 1998.
- [5] J. M. Lovekin, R. E. Yantorno, K. R. Krishnamachari, D. Benincasa, and S. J. Wenndt, "Developing usable speech criteria for speaker identification," *IEEE, International Conference on Acoustics and Signal Processing*, pp. 424–427, May 2001.
- [6] B. Y. Smolenski, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Co-channel speaker segment separation," *IEEE, International Conference on Acoustics and Signal Processing*, May 2002.
- [7] K. R. Krishnamachari and R. E. Yantorno, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions.," *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Nov 2000.
- [8] J. M. Lovekin, K. R. Krishnamachari, and R. E. Yantorno, "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*, Nov 2001.
- [9] N. Chandra and R. E. Yantorno, "Usable speech detection using modified spectral autocorrelation peak to valley ratio using the lpc residual," *4th IASTED International Conference Signal and Image Processing*, pp. 146–150, 2002.
- [10] A. R. Kizhanatham, R. E. Yantorno, and B. Y. Smolenski, "Peak difference autocorrelation of wavelet transform (pdawt) algorithm based usable speech measure.," *IIS Systemics, Cybernetics and Informatics*, 2003.
- [11] N. Sundaram, A. N. Iyer, B. Y. Smolenski, and R. E. Yantorno, "Usable speech detection using linear predictive analysis - a model-based approach," *IEEE International Symposium on Intelligent Signal Processing and Communication Systems, IS-PACS 2003*, 2003.
- [12] B. Y. Smolenski and R. E. Yantorno, "Fusion of co-channel speech measures using independent components and nonlinear estimation.," *IEEE International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS*, 2002.
- [13] B. Y. Smolenski and R. E. Yantorno, "Fusion of usable speech measures using quadratic discriminant analysis.," *IEEE, International Symposium On Intelligent Signal Processing and Communication Systems (submitted)*, 2003.
- [14] J. K. Shah, B. Y. Smolenski, and R. E. Yantorno, "Decision level fusion of usable speech measures using consensus theory," *IEEE International Sympo-*

*sium on Intelligent Signal Processing and Communication Systems, ISPACS 2003*, 2003.

- [15] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.
- [16] M. Sharma, Y. Kogan, R. Ramachandran, and Y. Li, "Speaker count determination, final technical report," Tech. Rep. AFRL-IF-RS-TR-1999-57, T-Netix, Inc., Air Force Research Laboratory, Rome, NY, April 1999.
- [17] D. S. Benincasa, "Voicing state determination of co-channel speech," *IEEE, International Conference on Acoustics and Signal Processing*, vol. II, pp. 1021–1024, 1998.