

Usable Speech Detection Using Linear Predictive Analysis – A Model-Based Approach

Nithya Sundaram, Robert E. Yantorno, Brett Y. Smolenski and Ananth N. Iyer
Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, PA 19122-6077, USA
nithyas@temple.edu, robert.yantorno@temple.edu, bsmolens@temple.edu, aniyer@temple.edu
http://www.temple.edu/speech_lab

Abstract: A speech segment is defined as “usable,” if speech, which is corrupted by interfering speech, can still be used for applications like speaker identification. In tactical communications, where there are multiple signals transmitted over the same channel such as telephone or radio transmission, separation of usable speech from speech corrupted by voices of other speakers is desired. This separation is important in making automatic speaker and speech recognition systems more robust. A novel approach towards developing a usable speech measure could be model-based. Using this concept of model-based usable speech detection, the use of Linear Prediction is investigated. The method reveals that an average of 75% of the usable speech is correctly detected with false alarms of 34%.

1. Introduction

Speech that is corrupted by interfering speech or non-stationary noise, but still usable for applications such as speaker identification, is referred to as “usable” speech. The goal of usable speech research has been to identify and extract the usable portions from co-channel speech. The system, which extracts usable speech segments under co-channel conditions, could be used as a front-end unit of a next-generation speech processing system [1]. Such extracted usable segments are sent on for further processing, while discarding segments containing co-channel speech. A schematic of the usable speech extraction system is shown in Figure 1.

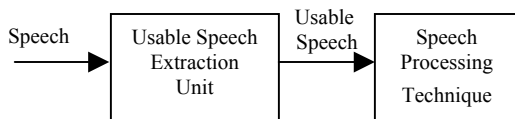


Figure 1. Application of Usable Speech Extraction System.

Speech segments can be declared “usable” for speaker identification based upon a power ratio of the speech of the target speaker to the speech of the interfering speaker.

The ratio is expressed as TIR (*Target to Interferer Ratio*, in dB). It has been shown that when the target speaker is at least 20 dB greater than the interfering speaker, 80% reliable identification of the target speaker can be obtained [1], [2]. Hence, segments with a high Target-to-Interferer Ratio (TIR) may be considered usable with respect to speaker identification.

Previous work was structure-based and shown that 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 to quantify the usability of speech [3]. Recent work has also shown that shape similarity exists in adjacent pitch periods of usable speech. This method of comparing the shape of adjacent pitch periods is known as the Adjacent Pitch Period Comparison (APPC) [4]. The peak distance between the LPC residual peaks [5], and peak difference of autocorrelation matrix of the wavelet transformed signal [6] were the two other structure-based methods to detect usable speech. A new approach to developing a usable speech measure is to use Linear Prediction as a model-based approach. This measure is devised based on the fact that the linear predictive coding (LPC) model of a single speaker (usable) speech is expected to be different from the LPC model of two speakers’ (unusable) speech.

2. LPC Analysis of Speech Signals

The human speech production system can be easily divided into the glottis or vocal cords and the vocal tract (mouth, tongue and lips). The glottal excitation acts as the source signal. The vocal tract, acting as a filter, then shapes the source signal to generate the output speech. In LPC, the vocal tract is modeled as an all pole digital filter that can be expressed mathematically as:

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

where, p is the order of the model, G is the gain, S(z) is the speech output of the model, and E(z) is the excitation input. The vocal tract information contains formants, which are the resonance of the vocal tract.

3. Model-Based Usable Speech Detection

The initial approaches to usable speech measure development were structure-based, e.g., Spectral Autocorrelation Peak Valley Ratio (SAPVR) [3] and Adjacent Pitch Period Comparison (APPC) [4]. To obtain higher identification accuracy, these measures are fused together and to achieve this one needs more than one measure providing complementary information. The introduction of a model-based approach now allows us to have a very different type of measure, which should provide complementary information as compared with the structure-based approaches. The LPC model approach is based on the premise that for a single speaker there will be approximately 5 resonances or peaks in the frequency characteristics of the LPC model whose coefficients are derived from using LPC analysis on a speech frame, and that there should be twice as many peaks for co-channel or two speaker speech. A schematic of the model-based usability approach is shown in Figure 2.

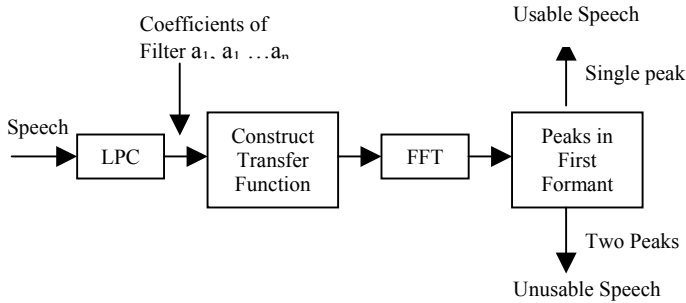


Figure 2 LPC Model-based Usable Speech Measure

LPC analysis allows extraction of vocal tract parameters, whose frequency characteristics contain resonances (formants). The extracted vocal tract information is analyzed by observing the frequency characteristics of the transfer function of LPC model. The decision of usability is made based on the number of peaks formed in the frequency range of 0 to 1 kHz. The decision is based on the fact that the first formant is the strongest formant and therefore the easiest one to detect [7].

The major obstacle of model-based approach using LPC analysis has been the occurrence of spurious peaks [8]. We have found an approach that reduces the number of spurious peaks significantly and have quantified the LPC measure.

4. Problems of Spurious Peaks

When determining the parameters of a system using linear prediction, a problem arises when the order of the predictor is higher than the order of the system. This results in spurious peaks, i.e., peaks that have no physical relevance to the system under study.

Peaks are extracted using analysis-by-synthesis or peak picking methods. The problem with peak picking algorithms is that they incorrectly identify spurious peaks as true formant peaks. Problems also arise with peak-picking algorithms when two formants occur close to each other (merged peaks) [9], which may be the case with co-channel speech. These spurious peaks and merged peaks are the challenges in detecting the difference in number of peaks formed by usable and unusable speech.

5. Spurious Peak Formation

The pattern of peak formation in the first formant in the range 0 to 1 kHz will be the focus of research presented here, because the first formant is the predominant peak, and therefore, should provide the most information about the number of speakers [10]. The initial set of experiments was performed using 12th and 24th order LPC analysis to investigate the problem of spurious peaks. One such data showing the spurious peak problem is shown in Figure 3.

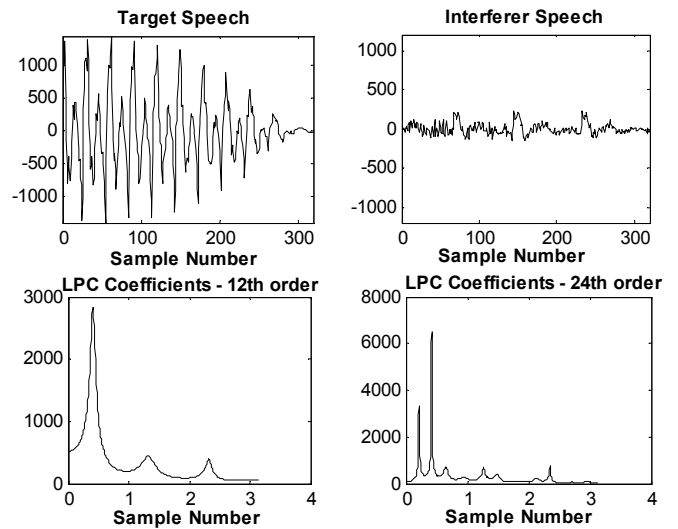


Figure 3. LPC Order Study. Target speech (upper left), Interferer speech (upper right), 12th order LPC (lower left), 24th order LPC frequency characteristics (lower right).

Note the added peak on each side of the first formant in the bottom right panel of Figure 3. Because the interferer speech is very low energy compared to target speech ($|TIR| \geq 20$ dB), the co-channel speech can be considered as single speaker speech, and therefore one would expect only the usual number of formants or peaks, i.e., 4 or 5, which is not the case for the 24th order LPC- right panel of Figure 3.

The method of using the LPC as a usable measure/detector involves being able to detect the difference between legitimate peaks related to speakers and spurious peaks. One approach is to reduce the LPC analysis model order, thereby reducing the number of possible spurious peaks.

6. Peak Pattern in First Formant

Experiments were performed to determine if it is possible to differentiate between single speaker speech and co-channel speech using linear prediction with a lower order model than 12th and 24th order so that the problem of spurious peaks would be reduced. The set of figures shown below illustrates two different conditions, i.e., one speaker (Figure 4) speech and two-speaker (co-channel) speech (Figure 5). 8th order and 16th order LPC analysis results are shown in Figures 4 and 5.

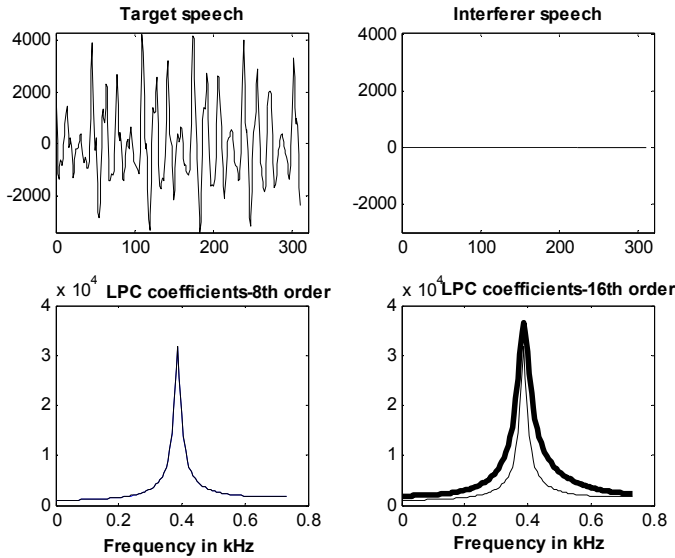


Figure 4. Usable Speech Detection Using LPC Analysis (usable speech shown). Target speech (upper left), interferer speech (upper right). 8th order LPC (lower left) and 16th order LPC model frequency characteristics shown in thick lines (lower right), target LPC frequency characteristics shown in thin lines (lower right).

The time form of the target is top left plot and the time form of the interferer is the top right plot for both Figures 4 and 5. Our interest is only in the first formant range, i.e., 0 to 1 kHz, and therefore we expect one peak for single speaker speech and two peaks for co-channel speech. In the bottom panels, the frequency characteristics of the target speech are shown in thin lines, composite speech in thick lines and interferer speech in dotted lines (where explicitly visible), with the left panel showing the 8th order characteristics and the right panel showing the 16th order characteristics.

It should be noted that for the co-channel speech in Figure 4, the interferer speech is so small as to be insignificant, and therefore, the co-channel speech can be considered single speaker (usable) speech. This is due to the insignificant contribution from the interferer, as can be observed in the upper right-hand panel of Figure 4.

When the same experiment was performed on co-channel speech, i.e., where the interferer was no longer insignificant, a notable change was observed in the number of peaks in the first formant (0-1kHz) as shown for the 16th order LPC depicted in the lower right plot in Figure 5. As we observe in Figure 5, it is clear from the energies of the target and the interferer that the combined speech is co-channel. Note that there is a significant change in the number of peaks. As explained above, we expect there to be twice as many peaks for two speakers' speech, and we observe there are two distinct peaks shown in thin line within the range 0 to 1 kHz in Figure 5.

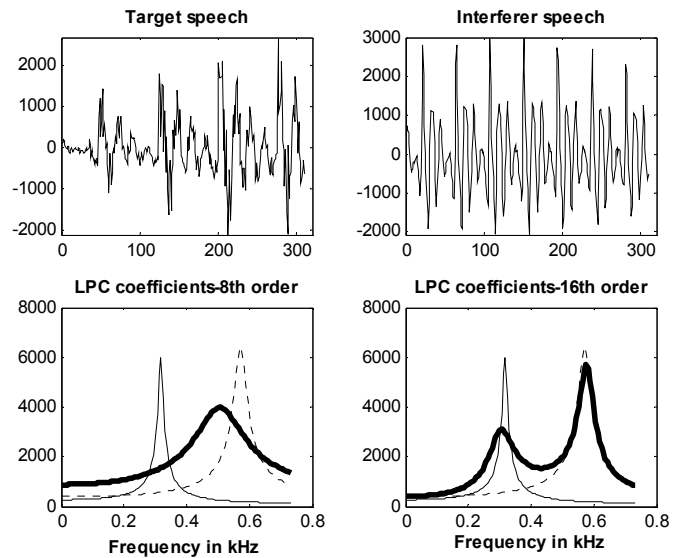


Figure 5. Usable Speech Detection Using LPC Analysis (unusable speech shown). Target speech (upper left), Interferer speech (upper right), 8th order LPC (lower left), 16th order LPC model frequency characteristics shown in thick lines (lower right), target LPC frequency characteristics shown in thin lines (lower right), interferer LPC frequency characteristics shown in dotted lines (lower right).

From Figure 5, it can be observed that the 8th order LPC of target and interferer were not able to model the composite speech, whereas the 16th order LPC (thick) was able to model the co-channel speech by producing one peak for target and one peak for interferer. These experiments suggest that using lower order LPC analysis provides a method of overcoming the problem of spurious peaks.

7. Experiments and Results

Speech data for experiments discussed here was obtained from the TIMIT database. Forty five different combinations of co-channel data from ten speech files, with equal number of male and female speakers, were used in the experiments. 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. Each frame was hamming windowed prior to computing the LPC. LPC analysis was done for each frame of length 10 ms. By processing the voiced frames, the accuracy of the algorithm was expected to improve. Hence, the low energy (unvoiced) frames and silences were removed using the spectral flatness measure with the help of a preset threshold [1].

The problem of spurious peaks is reduced appreciably by reducing the order of the LPC analysis to 8 and 16, which therefore reduces the number of false alarms. A peak-picking algorithm was designed to determine the number of peaks present in the first formant region. The peak-picking algorithm looks for the local maxima in that search interval and counts the number of such maximas. The speech segment was then declared usable if there were one peak detected and unusable if there were two peaks detected. The results of using the LPC measure for detection of male-female co-channel speech is shown in Figure 6, where the number of peaks formed by the LPC analysis is plotted versus the TIR values.

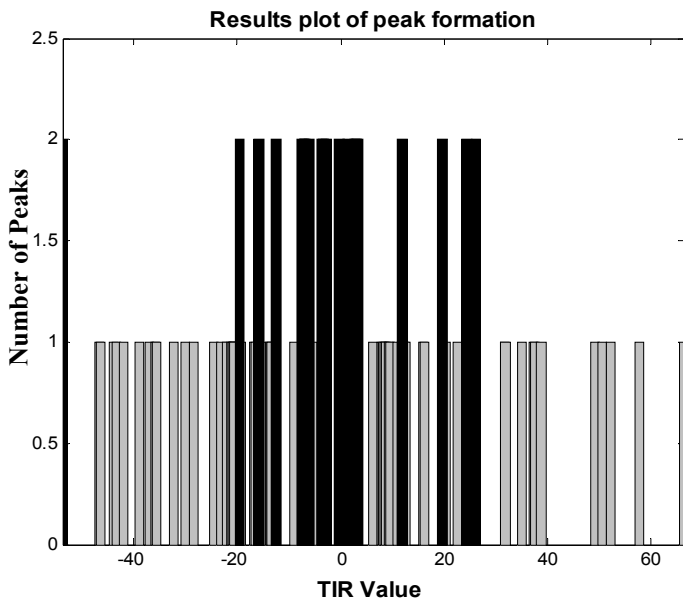


Figure 6: Results plot of peak formation. Usable data – Single peak (gray); Unusable data – Two peaks (black).

From Figure 6, it is observed that a majority of the single peaks (gray) found are in the range of $|TIR| \geq 20$ dB (usable), whereas a majority of two peaks (black) were found for $|TIR| < 20$ dB (unusable). The measure shows around 75 % correct detection and 34 % incorrect decision or false alarms. Figure 7 shows LPC model usable speech correct detection, which is represented by black rectangles.

The utterance shown in Figure 7 is TIR identified speech (gray – usable segments; black – unusable segments) from co-channel data and LPC model detection information (black rectangles). The results indicate that there is good correlation between the usable speech detection by the TIR and LPC model-based approach.

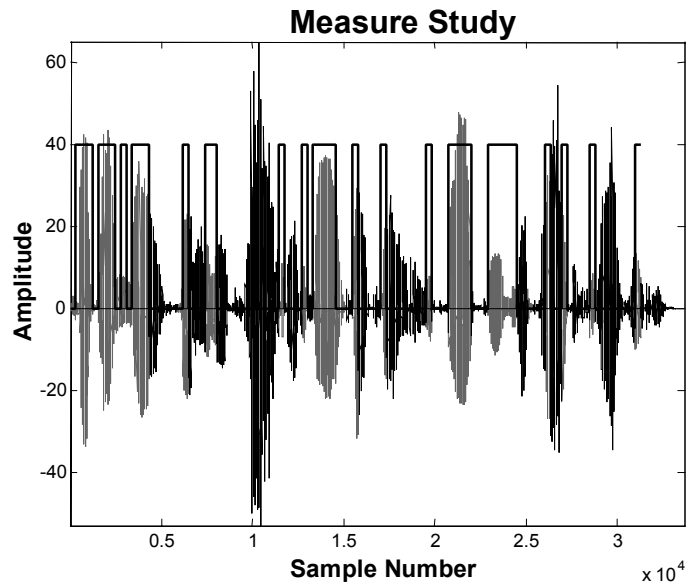


Figure 7: Comparison of detection using TIR and LPC measure. TIR detection – usable segments (gray), unusable segments (black), LPC detection, rectangles- correct usable speech detection.

8. Summary

In this paper we have presented a new method of detecting usable speech based on LPC analysis. The performance of speaker identification systems can be improved by using the detected usable speech segments as most of the corrupted data (unusable speech) has been removed. The goal of the LPC measure is to extract the maximum amount of usable speech with a minimum false alarm rate. On average, the LPC based usable speech measure detects at least 75% of the usable speech.

9. Future Areas of Research

The challenge of improving the LPC model approach for usable speech detection is to devise a method for detecting spurious peaks. One such possible approach is to lowpass (0 to 1 kHz) filter the speech, where we are working to devise a way to detect whether one of the two peaks is the first formant of the second speaker or is a spurious peak. Another approach is to lowpass filter the speech to 1 kHz, and to reduce the system to a simple 2nd order (single speaker system) and 4th order (two speakers system) system. The decision of usability can then be made by performing LPC on the residual of the first stage of LPC analysis.

Acknowledgement

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, 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.

Disclaimer

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory, or the U.S. Government.

10. References

- [1] Yantorno, R.E., "Co-channel Speech Study" Final Report to Air Force Office of Scientific Research and Speech Processing Lab, Rome Labs, 1999.
- [2] Lovekin, J. M., Yantorno, R. E., Krishnamachari, K. R., Benincasa, D. S., and Wenndt, S. J., "Developing Usable Speech Criteria for Speaker Identification Technology", IEEE, ICASSP, pp: 424-427, May 2001.
- [3] 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-ISPACS, pp: 710-713, Nov. 2000.
- [4] Lovekin, J. M., Yantorno, R. E., Krishnamachari, K. R., Benincasa, D. B., and Wenndt, S. J., "Adjacent Pitch Period Comparison (APPC) as a Usability Measure of Speech Segments Under Co-channel Conditions", IEEE, ISPACS, pp: 139-142, Nov. 2001.
- [5] Iyer, A.N., Gleiter, M., Smolenski, B.Y., and Yantorno, R.E., "Structural Usable Speech Measures Using LPC Residual", IEEE International Symposium on Intelligent Signal Processing and Communication Systems, Dec 2003 (Accepted).
- [6] Kizhanatham, A.R., and Yantorno, R.E., "Peak Difference Autocorrelation of Wavelet Transform Algorithm Based Usable Speech Measure", 7th World Multi-conference on Systemic, Cybernetics, and Informatics, Aug 2003 (Submitted).
- [7] Schafer, R.W., and Rabiner, L.R., "System for Automatic Formant Analysis of Voiced Speech", Journal on Acoustical Society of America, Volume: 47, pp-637-648, 1970.
- [8] Mccandless, S.S., "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra" IEEE Transaction on Acoustical Society of America, Volume: 22, pp-135-141, 1974.
- [9] Yegnanarayana, B., "Formant Extraction from Linear-Prediction Phase Spectra" Acoustical Society of America, Volume: 63, no.5, pp- 1638-1640, 1978.
- [10] Parsons. T., "Separation of Speech from Interfering Speech by Means of Harmonic Selection", Journal of Acoustical Society of America, Volume: 60, pp-911-918, Oct 1976.