

DETECTION OF USABLE SPEECH UNDER CO-CHANNEL CONDITIONS

**Department of Electrical And Computer Engineering
Temple University**

Thesis Proposal

In Partial Fulfillment Of The Requirements For The Degree

Master of Science In Electrical Engineering

Author: Nishant Chandra

Advisor: Dr. Robert Yantorno

ABSTRACT

Speech that is corrupted by interfering speech, but is still usable for applications such as speaker identification, is referred to as “usable” speech. Since, the concept of usability is context dependent, it is necessary to define usability based upon its intended application. Recently, a usable speech extraction system was proposed to separate a co-channel utterance into segments that are usable for speaker identification and those that are unusable. Speech segments can be declared usable for speaker identification based upon a Target-to-Interferer energy ratio (TIR). Portions of usable speech occur when high energy voiced speech from the target speaker overlaps with low-energy speech from an interfering speaker, or visa versa. The proposed research is to develop various methods for identification of usable speech in a co-channel environment. One such new usability measure developed is the Spectral Autocorrelation Peak to Valley Ratio of the linear predictive coding Residual (SAPVR-Residual). Preliminary results obtained using the SAPVR-Residual, indicate it is useful in spotting approximately 71% of usable segments, with a corresponding false alarm rate of 37%. Corresponding results obtained by SAPVR-Speech was 69% correct and 43% false alarms. Cyclostationarity and higher order statistics are other potential usability measures and research is being carried out to develop these other candidates as measures to be used in usable speech extraction processes. To make the co-channel speech processing system robust several different methods for usable speech extraction can be intelligently fused together. Fusion is expected to give good results when each measure gives complimentary information. This method of identification and extraction of usable speech, which also involves an information fusion system, represents the front-end process of a next generation co-channel speech processing system whose final goal is separation and reconstruction of target and interferer speech.

TABLE OF CONTENTS

ABSTRACT.....	ii
LIST OF FIGURES.....	iv
CHAPTER 1 INTRODUCTION	1
1.1. Usability of speech and speech measure.....	2
1.2. Co-Channel speech separation.....	6
1.3. Next generation co-channel speech processing system.	7
1.4. Overview.	10
CHAPTER 2 BACKGROUND	11
2.1. Previous work on co-channel speech separation.	11
2.2. Usable speech measures.	14
2.2.1 Spectral autocorrelation peak to valley ratio.	14
2.2.2 Adjacent pitch period comparison.	16
CHAPTER 3 PROPOSED NEW MEASURES.....	19
3.1. Spectral autocorrelation peak to valley ratio of the LPC residual.	19
3.1.1 Spectral autocorrelation of speech.	20
3.1.2 Linear predictive coding (LPC).	24
3.2. Cyclostationary as a potential usable speech detection method.	25
CHAPTER 4 PRELIMINARY RESULTS	28
4.1. Experimental setup.....	28
4.2. Results and discussion.....	28
4.3. Comparison of SAPVR-Residual versus SAPVR-Speech.....	34
4.4. Discussion.	35
REFERENCES.....	37

LIST OF FIGURES

Figure	Page
1. Application of co-channel speech detection system.....	2
2. Co-channel speech utterance. Usable speech (black) and unusable speech (gray).....	3
3. Usable speech extracted from co-channel speech.....	4
4. Percent correct verses TIR in dB	6
5. Block diagram of the next-generation co-channel speech processing system for speaker identification and speech extraction.	8
6. Usable speech extractor sub-unit of proposed next generation speech processing system for speaker identification and speech extraction.	9
7. Voicing state decision tree for co-channel speech.	13
8. Block diagram of the spectral autocorrelation peak-to-valley ratio (SAPVR) method for detecting usable speech frames.	14
9. SAPVR approach, Co-channel speech.	15
10. SAPVR approach. Voiced speech.	16
11. Usable speech and adjacent pitch period amplitude comparison (bottom).	17
12. Target and interferer speakers and their spectral autocorrelation.	21
13. Co-channel speech and its spectral autocorrelation.....	21
14. Calculation of SAPVR.....	22
15. Block diagram of generation of residual of speech using LPC.....	23
16. Cyclostationarity based usable speech detection system.	25
17. Conjugated cyclic correlation.	26

18. Usable speech, modified SAPVR-Residual approach,.....	28
19. Co-channel speech, SAPVR-Residual approach.....	29
20. Probability study of SAPVR distance measure.....	30
21. Usable speech detected by TIR & SAPVR-Residual thresholds.	31
22. Comparison of co-channel speech processed by SAPVR-Speech and SAPVR-Residual.....	33

1CHAPTER 1

INTRODUCTION

Co-channel speech occurs when one speaker's speech is corrupted by another speaker's speech. The performance of a speaker identification system degrades under co-channel conditions. The level of degradation depends on the amount of corruption by the interfering speech. Previous work [Yantorno, 1999] has shown that there exist segments of speech in the presence of interfering or co-channel speech, which can be identified as "usable" which could be used by a speech processing system. The focus of this thesis is to develop approaches that identify and extract those usable segments. Usable segments occur in co-channel speech when the low-energy unvoiced or silence portions of the interfering speech overlaps with the voiced portions of the target speech or vice versa. If those segments are extracted, and then processed by a speaker identification system, one would expect the results to be much more reliable than if the entire co-channel utterance was processed. Apart from speaker identification, the extracted usable speech could also be used for speech processing system, such as a speech recognition system.

A number of methods to identify usable speech segments have been developed. [Krishnamachari, *et al.*, 2000; Lovekin, *et al.*, 2001; Krishnamachari, *et al.*, 2001; Chandra and Yantorno, 2002] Usable speech measures have also been used to detect co-channel speech. This could provide information to speech processing system, to suspend its operation whose performance would be degraded if it were processing co-channel speech. [Yantorno, *et al.*, 2001; Kizhanatham and Yantorno, 2002] A usable speech

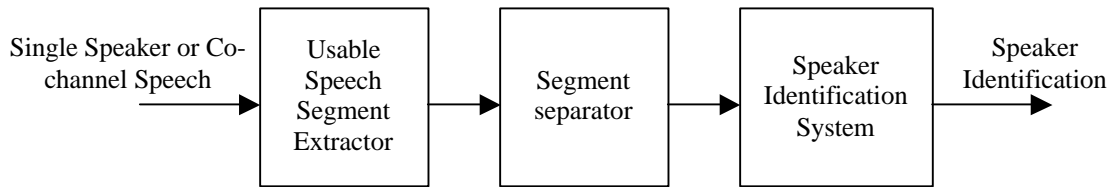


Figure 1: Application of co-channel speech detection system.

segment extractor as shown in Figure 1, can be used to extract usable speech and therefore maintain high performance of speech processing system.

1.1. Usability of speech and speech measure.

Usable speech is speech that is degraded in some way, but is still usable for certain applications [Yantorno, 1999]. Typically these applications are also of interest in general digital signal processing, such as speaker identification, speaker verification, speech recognition, etc. Because these applications of speech processing are all inherently different, it is necessary to define usability based upon its intended application. For example, speech defined as usable for speaker identification will not necessarily be usable for speech recognition. The concept of usability in this research refers to usability or usable speech in the context of speaker identification only. Usability measures are mathematical parameters that provide a quantitative measure of usability of a speech segment. Typically they would be correlated to Target to Interferer Ratio (TIR). In this research, the mathematical parameters are validated against TIR to ascertain their validity as usability measures.

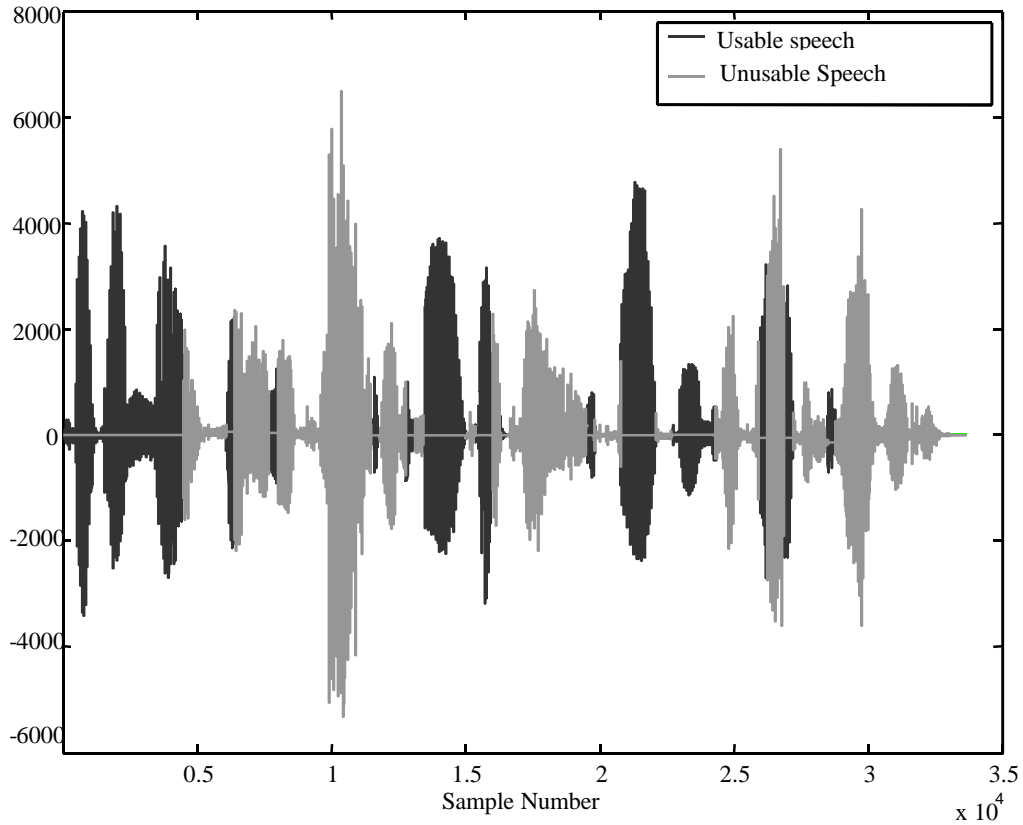


Figure 2: Co-channel speech utterance. usable speech (black) and unusable speech (gray).

In order to perform speaker identification, a logical approach to the problem is to detect the portions of an entire co-channel speech utterance that contains minimal or no speech from a different speaker. Such segments can then be defined as usable and sent on for further processing. A system for extracting usable segments was proposed for usable segment extraction [Yantorno 1999]. The usable segment extraction system is based on TIR. TIR can be expressed for entire utterances or individual frames of speech. For usability, previous experimentation (Figure 2, shown in black) has shown that frames above 20 dB TIR, are considered usable, and those shown in gray are considered

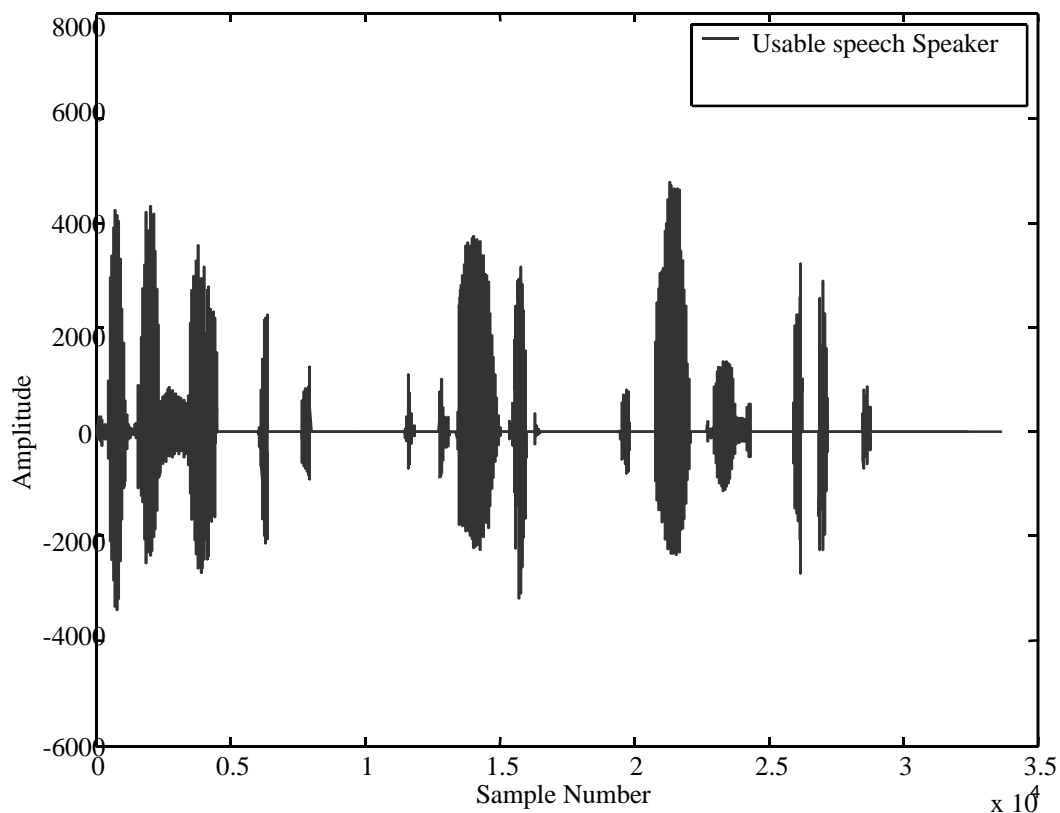


Figure 3: Usable speech extracted from co-channel speech.

unusable segments. The concept of usable segments relies on the fact that for any given time frame, the energy of each of the speakers may differ. The usable speech concept takes advantage of the situation when the energy of the primary speaker is much greater than the energy of the interfering speaker for a given time frame. It is important to note that although the overall co-channel utterance has a 0 dB TIR (both target and interfering speaker have the same average energy), there are segments where the TIR is at least 20 dB. These segments occur when one speaker has a much higher energy than the other speaker, i.e. one speaker is voiced and the other speaker is unvoiced or silent.

Shown in Figure 3 is the result of the voiced segment extraction process. The data was extracted by removing all unvoiced and silent speech from single speaker utterances. This represents the information that will be sent to the speaker identification system for testing. It is important to note that the speech signal represented in Figure 3 has the silence removed and therefore is a time-compressed signal. The co-channel utterance showed in Figure 2 contains approximately 3.3 seconds of speech, while the usable speech shown in Figure 3 contains about 1.3 seconds of speech or approximately 25% of the original speech. Because usable speech does not include unvoiced speech or silences, the signal presented to the speaker identification system for testing is very 'information rich' for purposes of speaker identification. 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. Shown in Figure 4 is the result of experiments conducted to establish this, [Yantorno, *et al.*, 2001; Yantorno, 1998]. 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. [Yumoto and Gould, 1982; Krom, 1993] For 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.

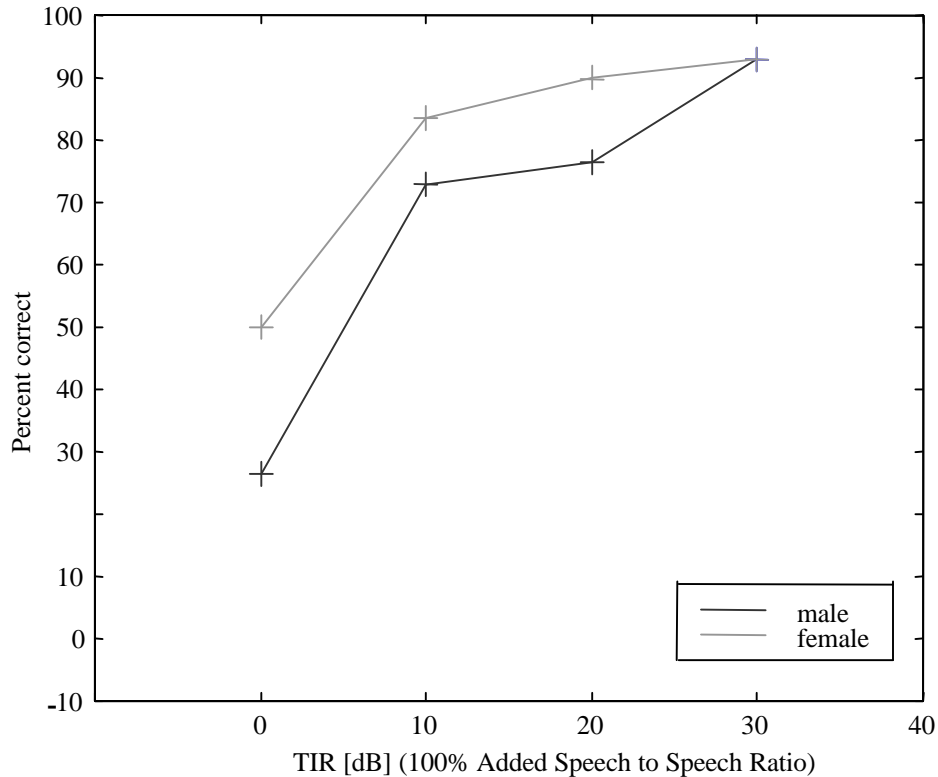


Figure 4: Percent correct versus TIR in dB.

1.2. Co-Channel speech separation.

Multi-speaker speech is a common occurrence resulting from the combination of speech from simultaneous and independent sources into composite speech at the receiver. The multi-speaker phenomenon may also be observed in speech originating from different sources occurs over a common communications channel such as a telephone or cross talk in a radio transmission. Although the human auditory system is quite proficient at focusing on a particular speaker or speakers in a mixture (known commonly as the

cocktail party effect) [Sayers and Cherry, 1957; Mitchell *et al.*, 1971], computer algorithms designed to do the same task have demonstrated only a limited degree of success. Co-channel speech more accurately describes the scenario of composite speech over a common communications channel. Thus, co-channel speech separation algorithms cannot rely on spatial location of the speakers, as do some algorithms that use a multiple microphones, to obtain a binaural input [Nakatani, *et al.*, 1996]. The research presented in this proposal will be to investigate only co-channel (two speakers) speech separation, with the condition that the manner in which speech is entered into the automatic speaker identification system should be monophonic.

As discussed earlier, usable speech detection is a novel approach to the co-channel situation. This differs from co-channel speech separation because the initial goal is not separation of the speech. Instead, it is of interest to find usable segments within co-channel speech that contain very little or no interfering speech, to this end. Mathematical measures are needed to detect the presence of usable speech frames within a composite co-channel utterance.

1.3. Next generation co-channel speech processing system.

A next generation co-channel speech processing system relies upon a combination of measures and technologies. The ultimate long-term goal of separation of co-channel speech is to automatically take co-channel speech and reconstruct the speech of both speakers into separate speech streams. The system block diagram of separation of co-

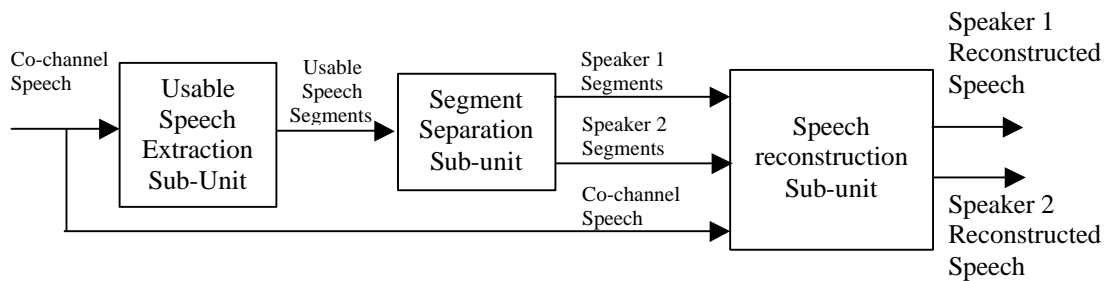


Figure 5: Block diagram of the next-generation co-channel speech processing system for speech extraction.

channel speech is shown in Figure 5. The first block in the Figure is the proposed usable speech extraction sub-unit. The objective of the speech segmentor is to identify those co-channel speech frames, which are usable. The extracted usable speech is fed into a segment separation unit. The segment separation unit might use speaker identification as a mechanism for separating the speech of two speakers.

Speech reconstruction would be performed in the following way. The frame following and preceding a segment would be analyzed using information from the abutting frames of the usable segment of speech. Information that might be used for reconstruction would be such things as pitch as well as formants; this assumes that neither of parameters changes dramatically from frame to frame, which is usually the case, then, those parameters are extended into the abutting frames. Identification of those features of the abutting the frames would then allow for extraction of the “predominant” speaker for that frame.

Since one is working with speech segments instead of frames, it is possible that other more sophisticated speech processing tools can be used for speech extraction. For

example, speech recognition tools could be used to construct words from segments. Also, information about isolated words in segments could be used to construct words in corrupted segments, as shown in the “constructing complete utterance” block.

To make the co-channel speech processing system robust several different methods for usable speech extraction can be developed and intelligently fused together. Fusion is expected to give good results when each measure gives unique information. To achieve this it is desired that the usable speech extraction methods come from different domains, shown in Figure 6. Several different ways for usable speech extraction has already been developed and are discussed in a later section. The current research is to develop new usable speech measure, which would become a part of usable speech extraction sub-unit of Figure 6.

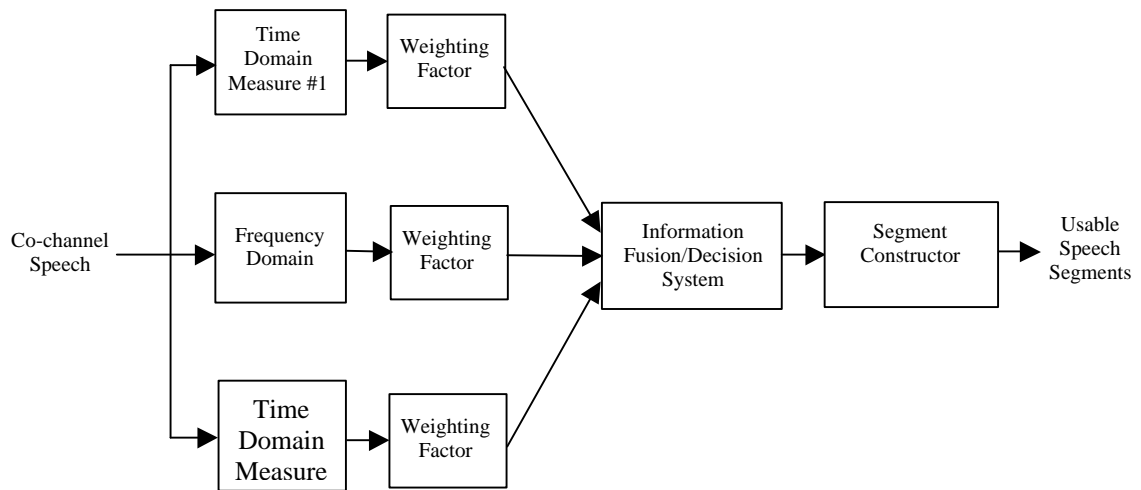


Figure 6: Usable speech extractor sub-unit of proposed next generation speech processing system for speaker identification and speech extraction.

1.4. Overview.

This thesis covers different approaches to the usable speech detection problem. Chapter 2 provides background information reviewing the classic and current approaches to processing co-channel speech. Chapter 3 develops a method for detecting usable frames of speech in a co-channel utterance. Finally, chapter 4 provides preliminary results and discussion.

2CHAPTER 2 BACKGROUND

Separation of co-channel speech has been an area of research for more than three decades. Many different approaches had been used. In the first section of this chapter an effort has been made to summarize the most significant developments in the area. In section two of the chapter usability speech measures, which have been developed in the past, have been discussed along with their approaches and results.

2.1. Previous work on co-channel speech separation.

Many techniques for addressing the co-channel speech separation problem have been developed. The earliest attempts at separating target speech from composite speech relied upon pitch estimation techniques that were originally developed to enhance a speech signal corrupted by noise [Shields, 1970; Frazier, 1975]. From these enhancement techniques, others were adapted for separating one or more speakers by selective filtering of composite speech based upon the pitch estimation of the stronger talker [Parsons and Weiss, 1975]. These techniques relied upon estimation of the pitch of at least one speaker, and enhancing the harmonics of the target speaker, suppressing the interfering speech [Hanson and Wong] and a combination of enhancement and suppression [Morgan, *et al.*, 1995]. However, these techniques suffered many drawbacks, including:

- They required a high TIR
- At least one speaker must be voiced within the analysis frame
- They are not robust to pitch estimation errors

Parthasarathy and Tufts [1987] used glottal opening onset and synchronous processing for separation of co-channel speech. However, this method requires a high TIR, is restricted to voiced speech and is computationally intensive. Lee and Childers [1988] used the harmonic magnitude suppression (HMS) system [Hanson and Wong] as a front-end for pitch estimation for a system that minimizes the cross entropy of two talkers in order to obtain enhanced spectral separation. A cepstral pitch suppression technique was proposed by Stubbs and Sommerfield [1991] for removing the harmonics of the interfering speaker. However, poor “quefreny” resolution in the cepstral domain resulted in muffled speech and poor suppression.

Quatieri and Danisewicz [1990] used a least squares estimation technique to determine the sinusoidal components of each talker, but it suffered from stability problems when the frequencies of the speakers were very close, and required *a priori* pitch information. Naylor and Porter [1991] proposed an algorithm that resolves speaker’s components into the complex spectrum, arguing that signals added together in the time domain do not necessarily add in the magnitude spectrum domain. However, because phase estimates are used, this method is very sensitive to additive noise [Morgan, *et al.*, 1997].

Benincasa and Savic [1998] developed a voicing state determination algorithm (VDSA) to estimate the voicing state of two speakers in a segment of co-channel speech. The VDSA system uses a Bayesian classifier that uses supervised training of the voicing states and automatic detection for testing the voicing states. However, the VDSA relies upon the assumption that two speakers are continuously talking. Therefore, as seen in Figure 7, there is no differentiation from voiced speech originating from a single speaker or simultaneous voiced speech from two speakers.

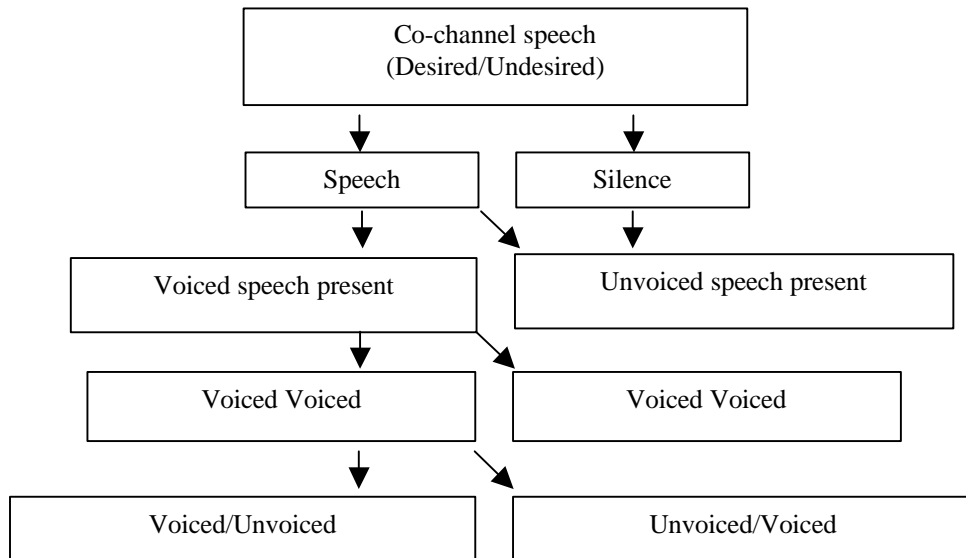


Figure 7: Voicing state decision tree for co-channel speech.
[Benincasa and Savic, 1998]

Many other algorithms are available in the speech processing community for separation of co-channel speech. An effort here has been made to summarize the most significant developments in the area.

2.2. Usable speech measures.

Two techniques for detection of usable speech have already been developed, these are spectral autocorrelation peak-to-valley ratio (SAPVR) and adjacent pitch period comparison (APPC). In the next subsections these methods are briefly discussed. SAPVR method takes advantage of the structure of voiced speech in the frequency domain whereas, APPC is a time domain measure.

2.2.1 Spectral autocorrelation peak to valley ratio.

Spectral autocorrelation peak-to-valley ratio (SAPVR) [Krishnamachari, *et al.*, 2000] measure was the first to be developed in the series of usable speech measures. The block diagram of the SAPVR method is shown in Figure 8. The speech is first windowed using a Hamming window. The magnitude of the FFT is found and autocorrelation is performed on windowed speech. A peak/valley-picking algorithm is utilized on the resulting autocorrelation and a ratio of the peaks to valley is calculated. Finally, a peak to valley ratio is compared to a threshold value. Frames exceeding the threshold are deemed usable.

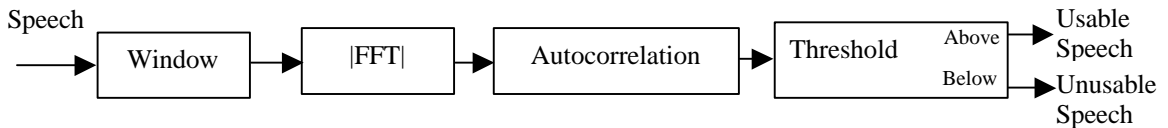


Figure 8: Block diagram of the spectral autocorrelation peak-to-valley ratio (SAPVR) method for detecting usable speech frames.

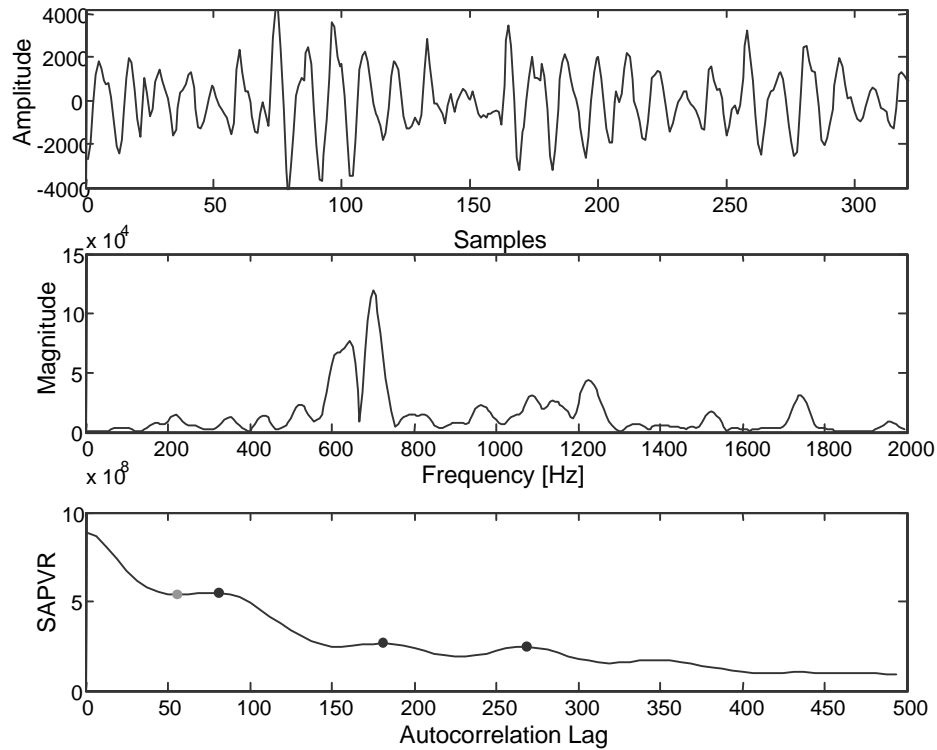


Figure 9: SAPVR approach. Co-channel speech (a) Speech segment (top panel), (b) FFT of speech (second panel down), (c) Spectral autocorrelation (bottom panel).

Figure 9 shows the graphical SAPVR results of a single frame of speech. The original speech frame is shown in the upper window. The middle window is the FFT of the speech frame. The bottom window is the spectral autocorrelation of the middle window. For a well-structured single speaker frame of speech (Figure 9, top panel), there is a well-defined harmonic structure in the frequency domain, (Figure 9, middle panel) and the spectral autocorrelation produces well-defined peaks and valleys (Figure 9, bottom panel). On the other hand, for a co-channel frame of speech, shown in Figure 10, the frequency domain is not well structured, (Figure 10, middle panel) and therefore the spectral autocorrelation does not produce well-defined peaks and valleys (Figure 10,

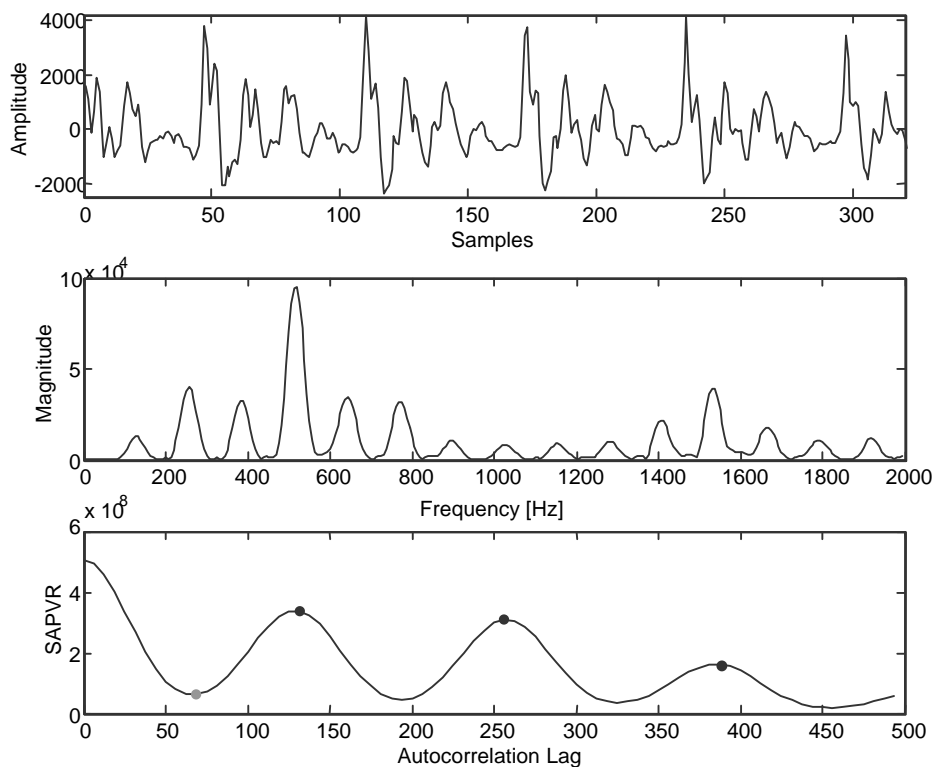


Figure 10: SAPVR Approach. Voiced speech (a) Speech segment (top panel), (b) FFT of speech (second panel down), (c) Spectral autocorrelation (bottom panel).

bottom panel). The SAPVR method takes advantage of the well-defined peaks and valleys to produce a high peak to valley ratio. This promising measure has been shown to detect 73% of usable frames with a corresponding false alarm rate of 28%.

2.2.2 Adjacent pitch period comparison.

Usable speech, which is composed entirely of voiced speech, is periodic in nature [Krishnamachari, *et al.*, 2001]. Due to the periodicity of usable speech, adjacent pitch periods of voiced speech are similar in ‘shape,’ which is evident in Figure11 (top panel).

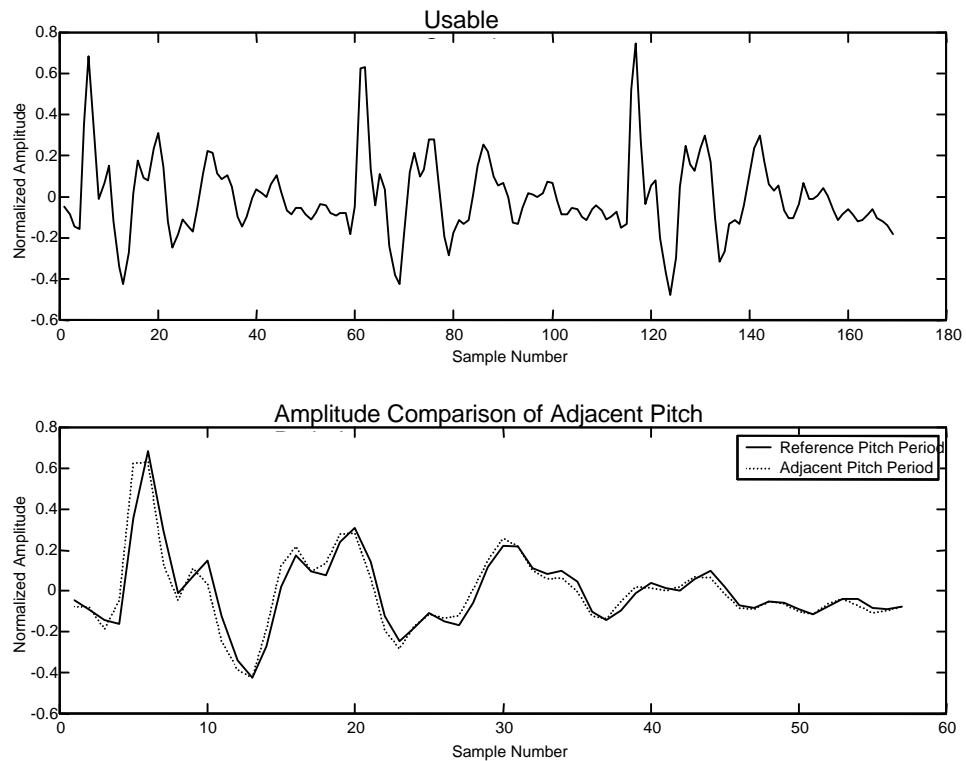


Figure 11: Usable speech (top) and adjacent pitch period amplitude comparison (bottom).

Shown at the top panel of Figure 11 is a 20-ms. usable speech segment. Observing the shape resemblance in adjacent pitch periods of usable speech, a mathematical measure was devised to quantify this shape similarity. The hypothesis was that if two pitch periods were of similar shape, then an absolute distance measure between the two waveforms would be minimal. This method of comparing the shape of adjacent pitch periods is known as the adjacent pitch period comparison, or APPC. [Lovekin, *et al.*, 2001]. The APPC measure is shown to be accurate in detecting usable frames of speech

in a co-channel utterance, with an average of 86% correct detection rate, a 14% missed detection rate, and a false alarm rate of 14% for a set of 38 speakers.

3CHAPTER 3 PROPOSED NEW MEASURES

As a step towards the goal of the research, spectral autocorrelation peak to valley ratio of the linear predictive coding (LPC) residual (SAPVR-Residual) is developed as a usability measure. The first section of this chapter deals with SAPVR-Residual. Then the concepts of SAPVR are explained, in the later subsection, LPC fundamentals are discussed. Initial study have suggested that cyclostationarity and higher order statistics are potential usability measures, which is discussed in the last section of this chapter.

3.1. Spectral autocorrelation peak to valley ratio of the LPC residual.

Spectral Autocorrelation Peak to Valley Ratio of the LPC residual (SAPVR-Residual) is a modification of an approach previously developed for detection of usable speech. In the previous approach [Krishnamachari, *et al.*, 2000] the SAPVR was calculated on the speech segments. The proposed 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 provided for 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.

3.1.1 Spectral autocorrelation of speech.

The technique of performing autocorrelation in frequency domain was previously used to represent heart rate variability [Link et. al, 1997] and to compare color flow imaging algorithms [Shariati *et al.*, 1993]. Ashira and Kado [1995] 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, that could be exploited, is that the magnitude spectrum of voiced speech can be predicted because of its harmonic structure, while that of the 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 peaks. If we use Schroeder's method (which was later adapted by Parsons [1976] 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. 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

$$R(p) = X(p)X(2p) + X(2p)X(3p) + \dots \quad (4.1)$$

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

$$R(2p) = R(p)R(3p) + R(2p)R(4p) + \dots \quad (4.2)$$

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.

The behavior of spectral autocorrelation under co-channel condition, varies depending on whether:

- Both the target and interfering speech are voiced,
- Either one of them are unvoiced or
- Both of them are unvoiced.

When both speech frames are unvoiced, the spectral autocorrelation does not contain any peaks that are harmonically related to each other. If at least one of the speech frames was voiced, the spectral autocorrelation contains harmonically related peaks as expected. If

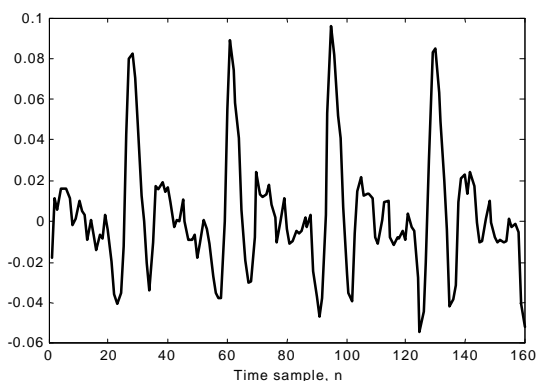


Figure 12a

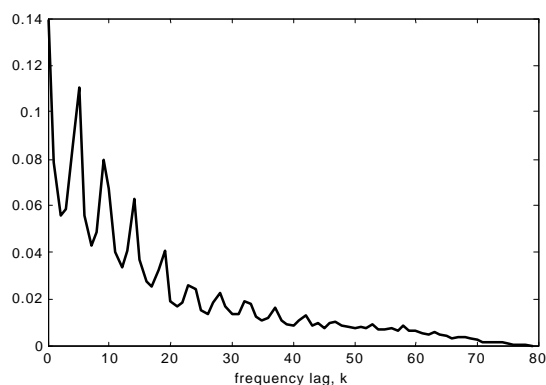


Figure 12b

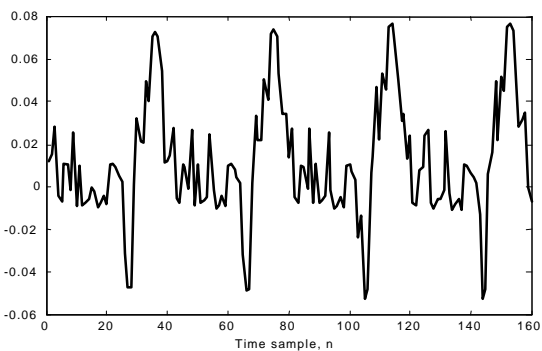


Figure 12c

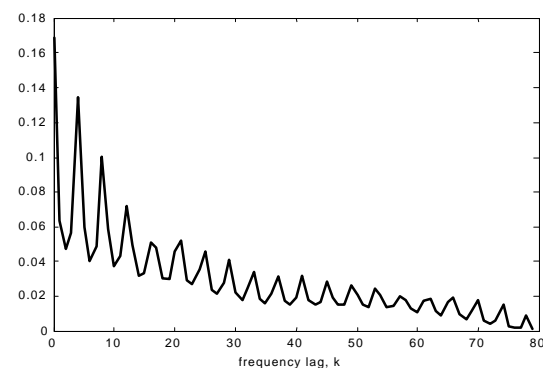


Figure 12d

Figure 12: Target and interferer speaker and their spectral autocorrelation, (a) and (c) Time waveforms of two-voiced speech frames. (b) and (d) are their corresponding spectral autocorrelations.

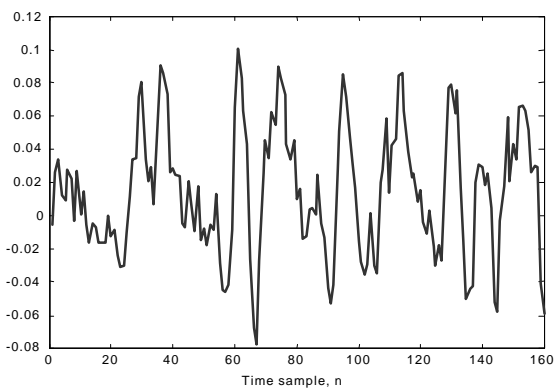


Figure 13 a

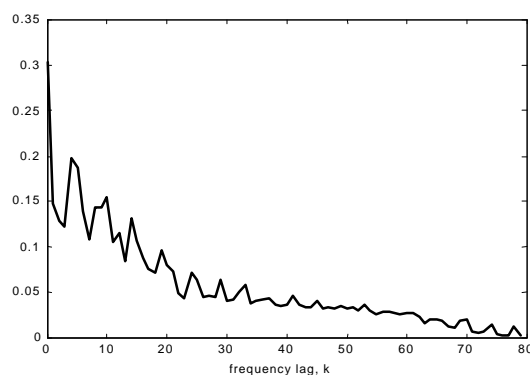


Figure 13b

Figure 13: Co-channel speech and its spectral autocorrelation, (a) Composite waveform generated by adding voiced speech of figure 12(a) and (c). (b) Spectral autocorrelation of (a).

both the speech frames are voiced, the spectral autocorrelation contains either two

distinct trains of peaks that are harmonically related or no harmonically related peaks.

Figures 12(a) and 12(c) show two frames of voiced speech and Figure 12(b) and 12(d) their corresponding spectral autocorrelations. Notice the well defined peaks at harmonically related frequency lags in Figures 12b and 12d. The voiced speech frames shown in Figures 12a and 12c were taken from TIMIT database, which was used for the experiments. Figure 13a is the composite speech derived from the above frames, and Figure 13b is the spectral autocorrelation of the composite speech. One important observation concerning Figure 13b 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 autocorrelation values for lags are not harmonically related, due to co-channel conditions. (Krishnamachari *et al.*, 2000).

The Spectral autocorrelation peak to valley ratio parameter is defined as follows:

$$SAPVR = (2 * P0 + P1 + P2 + P3 + P4)/(V0) \quad (4.3)$$

The SAPVR is calculated as the sum of twice the first peak and next four peaks to the first valley as illustrated in Figure 14.

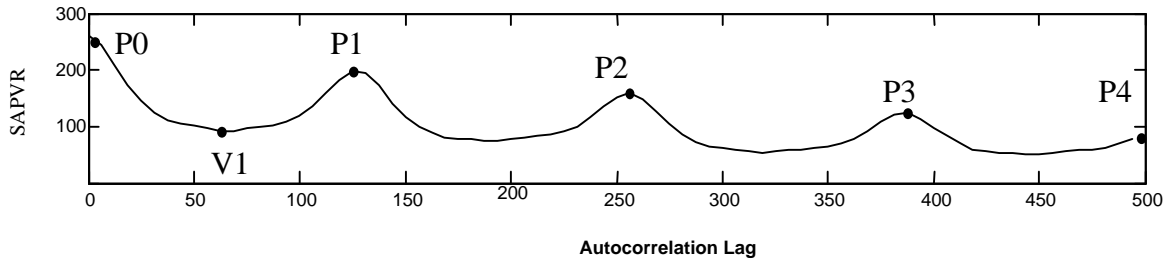


Figure 14: Calculation of SAPVR

3.1.2 Linear predictive coding (LPC).

Linear Predictive Coding (LPC) is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate. An important characteristic of LPC parameters is that they essentially preserve the intelligibility information of the speech signal.

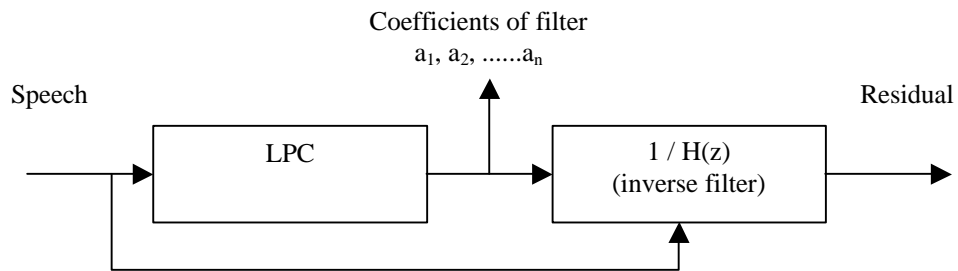


Figure 15: Block diagram of generation of residual of speech using LPC.

LPC analysis allows extracting 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. This is shown in Figure 15.

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

$$H(z) = \frac{G}{1 + a_1z^{-1} + a_2z^{-2} \dots + a_pz^{-p}} = \frac{S(z)}{E(z)} \quad (4.4)$$

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 equation above can be written in the time domain as:

$$S(n) = Ge(n) - a_1s(n-1) - a_2s(n-2) \dots - a_p s(n-p) \quad (4.5)$$

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.

3.2. Cyclostationary as a potential usable speech detection method.

A common assumption made by conventional statistical signal processing methods is that the random signals operated upon are stationary. That is, the parameters of the physical system that generates the random signal are invariant with time. For most man-made signals, some parameters do vary periodically with time and in some cases harmonically unrelated periodicities are involved [Gardner, 1991]. These random signals can be modeled as cyclostationary, in which the statistical parameters vary in time with single or multiple periodicities. Investigations by Gardner [1991] revealed that an inherent property of cyclostationarity signals is spectral redundancy, which corresponds to the correlation that exists between the random fluctuations of components of the signal residing in distinct spectral bands. This property could be exploited to perform various signal processing tasks like:

- Detecting the presence of signals buried in noise and/or severely masked by interference.
- Recognizing such corrupted signals according to modulation type.
- Reduction of signal corruption due to co-channel interference and/or channel fading for single receiver systems.
- Linear periodic time-variant prediction.

Initial research suggests that these properties can be exploited to detect the presence of

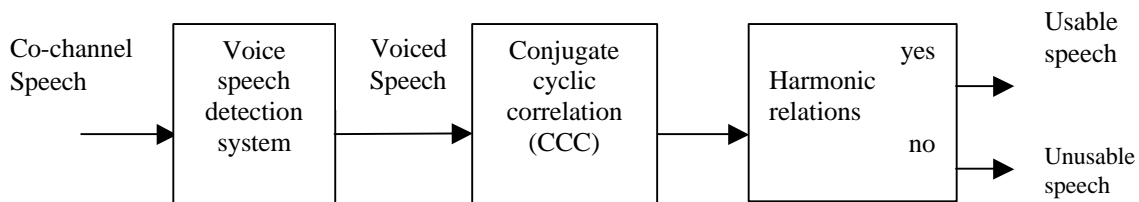


Figure 16: Cyclostationarity based usable speech detection system.

usable speech segments in co-channel speech signal.

The input to the system in Figure 16 is a co-channel speech signal. We consider only voiced speech. There are two types of voiced speech detection system, energy and zero crossings voiced detection system and spectral flatness voiced detection system.

1. Energy and zero crossings voiced detection system.

The composite signal is analyzed on frame-by-frame basis. For each frame, energy and zero crossings are computed and compared with the preset threshold to determine whether it is voiced speech or unvoiced speech.

2. Spectral flatness voiced detection system.

The spectral flatness for each frame is computed and it is compared with the preset threshold (35 dB) to determine if the frame under investigation is voiced or unvoiced speech.

Once the voiced frames are identified the Conjugate cyclic correlation (CCC) is found out as shown in Figure 17.

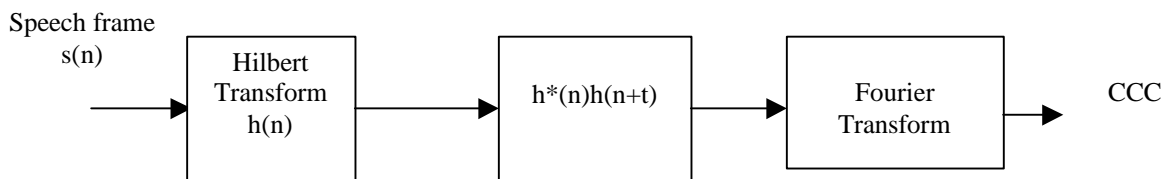


Figure 17: Conjugated cyclic correlation.

Experiments are being carried out using this property to establish a reliable usable measure.

4CHAPTER 4 PRELIMINARY RESULTS

Results obtained from experiments for SAPVR-Residual are given in this chapter. Later in the chapter a comparison is made between SAPVR-Speech and SAPVR-Residual, usable speech detection methods.

4.1. Experimental setup.

The experiments were performed on speech data obtained from the TIMIT database. In all 80 speaker files were used for the experiment. The original speech was sampled at 16 kHz and resample 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 on a frame-by-frame basis. The frame size was 40 ms. each frame was hamming windowed prior to computing the magnitude spectra and the corresponding spectral autocorrelation ratio. On an average each file had 125 frames and 40 files were used to generate data. There were equal number of male and female files from different speakers and speaking different phrases.

4.2. Results and discussion.

A successful identification of usable speech occurs when SAPVR-Residual measure and TIR measure select a frame of co-channel speech as usable. A missed identification is

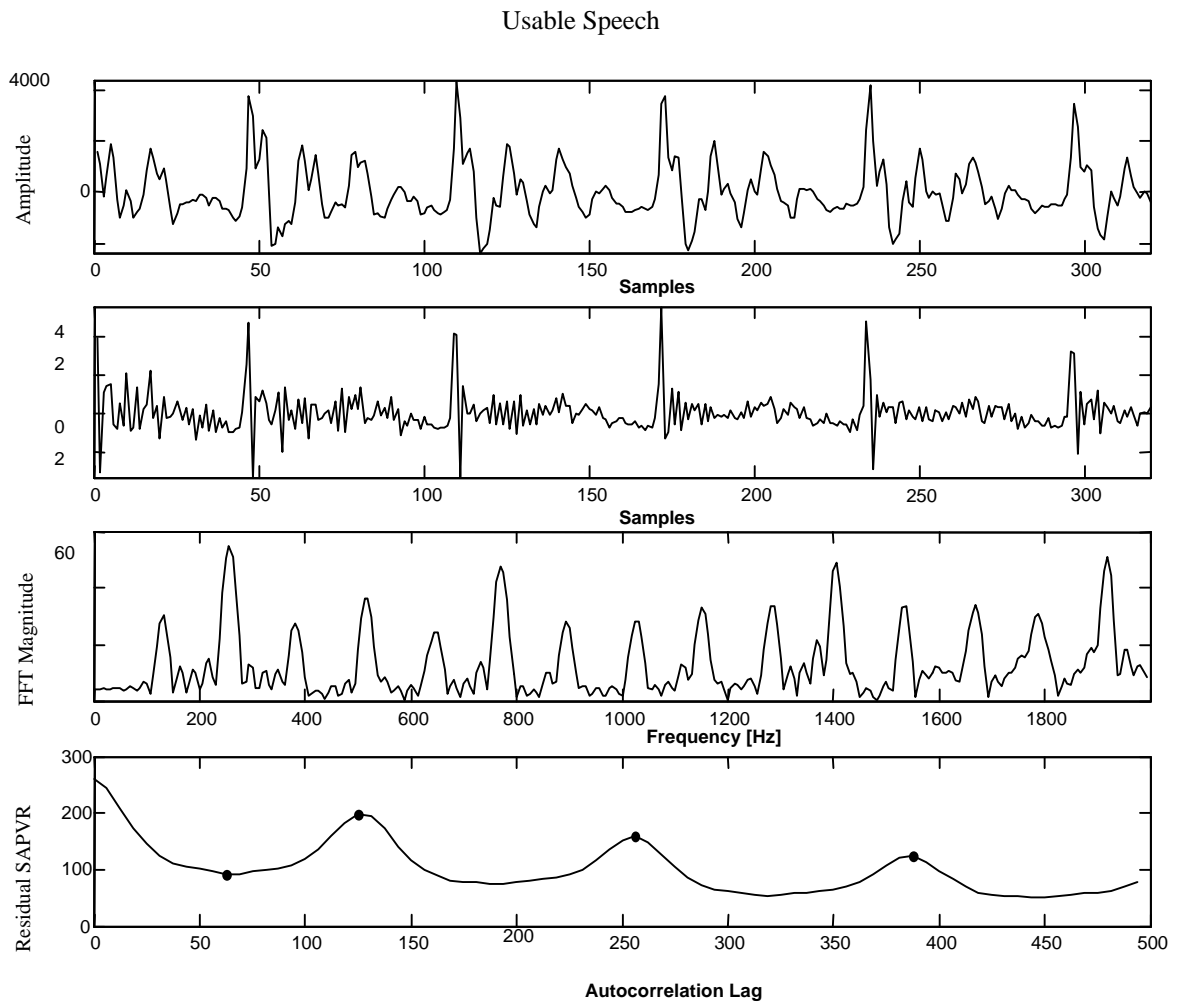


Figure 18: Usable speech, modified SAPVR approach, (a) Speech segment (top panel), (b) Residual of usable speech (second panel down), (c) FFT of residual (third panel down), (d) Spectral autocorrelation (bottom panel).

said to occur when the TIR measure has selected a frame that has not been selected by the SAPVR-Residual measure as usable. 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 as usable. Usable speech, which is composed almost entirely of voiced speech, has a periodic nature and so its spectral autocorrelation also has a periodic structure. Shown at the top of Figure 18 is a 40ms segment of usable speech. Due to the periodicity of usable

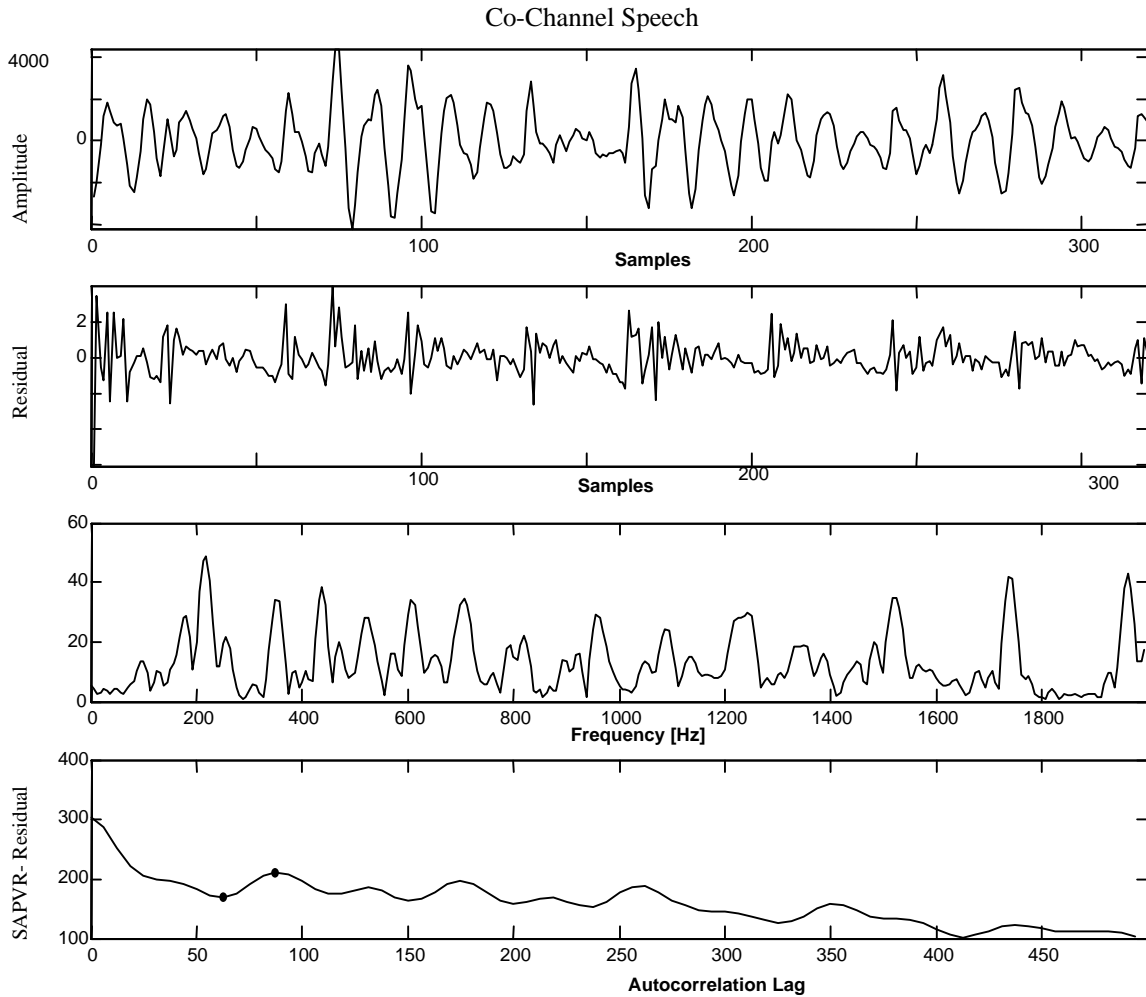


Figure 19: Co-channel speech, modified SAPVR approach, (a) Co-channel speech segment (top panel), (b) Residual of unusable speech (second panel down), (c) FFT of residual (third panel down), (d) Spectral autocorrelation (bottom panel).

speech, the excitation (residual) of speech also has a periodic structure, which is evident in Figure 18(b). Figure 18(c) shows the FFT of the residual. Figure 18(d) is obtained by performing the autocorrelation on the FFT of residual. Definite peaks and valleys can be seen in this subplot – as identified by the dots, which are used to compute the SAPVR.

The SAPVR-Residual measure is used to detect the structure of the spectral autocorrelation. This structure is illustrated in Figure 18 for the residual, frequency and spectral autocorrelation domains. Also, because the spectral autocorrelation can be used

to illustrate 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 19. It can be seen that there are no good peaks and valleys in Figure 19 (d), and therefore its peak to valley ratio falls below the threshold indicating that this frame is unusable.

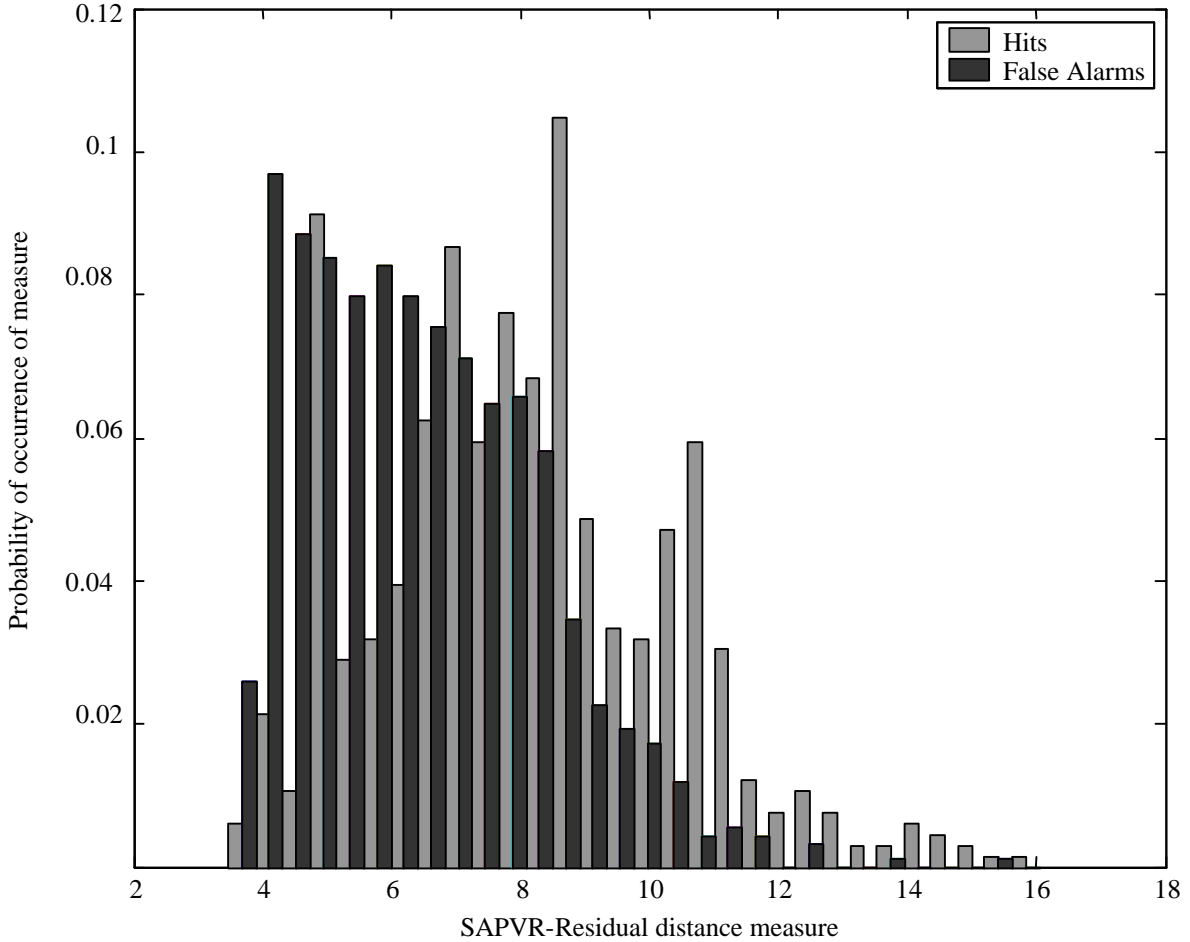


Figure 20: Probability study of SAPVR-Residual distance measure.

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 value. This is shown in Figure 20. 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 missed frames occur mostly in transition regions (onset and offset of voicing).

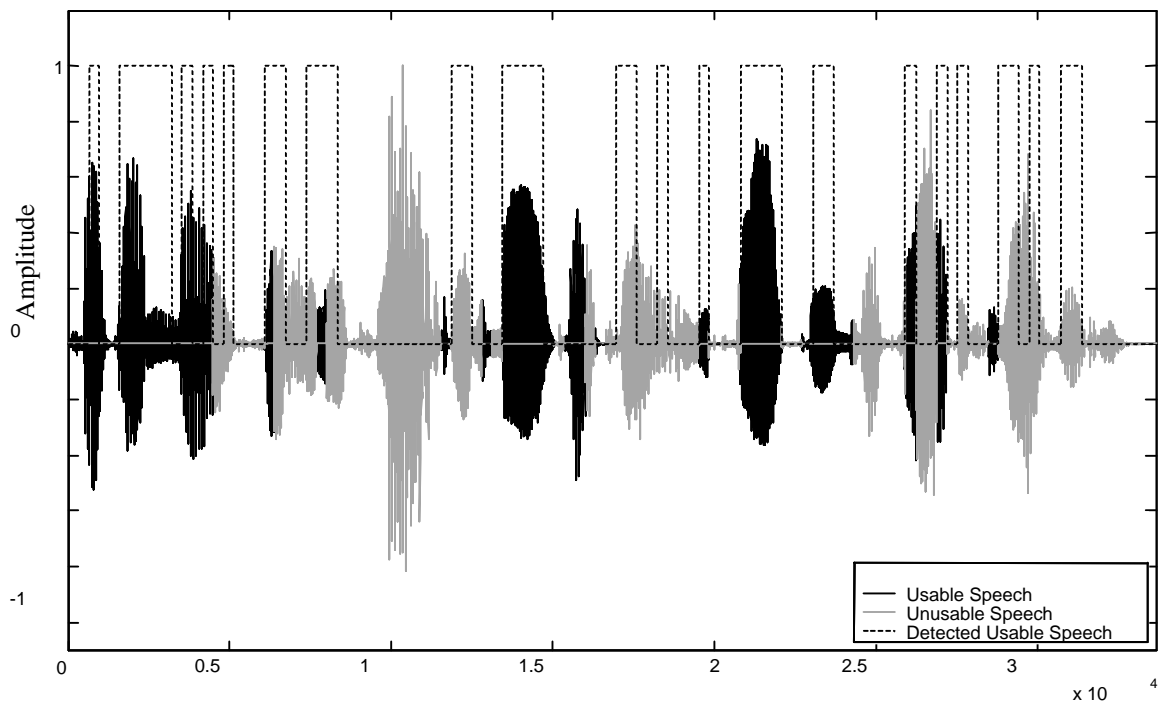


Figure 21: Usable speech detected by TIR & SAPVR-Residual thresholds. TIR usable speech (black), TIR unusable speech (gray), detected usable speech (dashed box).

The result of a single experiment is shown in Figure 21. The dashed rectangles in Figure 21 indicate detection of usable speech. For comparison, the TIR is plotted at a threshold ± 20 dB. Frames of speech where the TIR is at least 20 dB are shown in black, since these frames are picked both by TIR and SAPVR-Residual measure they are hits. The

gray sections of the co-channel utterance in Figure 21 whose TIR are less than 20 dB are considered unusable for speaker identification, these frames are picked by SAPVR-Residual as usable but were not usable by TIR measure, so they are called as false alarms.

For the experiment shown in Figure 20, 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 detected as usable. 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.

4.3. Comparison of SAPVR-Residual versus SAPVR-Speech.

Figure 22 shows the comparison of SAPVR-Speech and SAPVR-Residual on the same

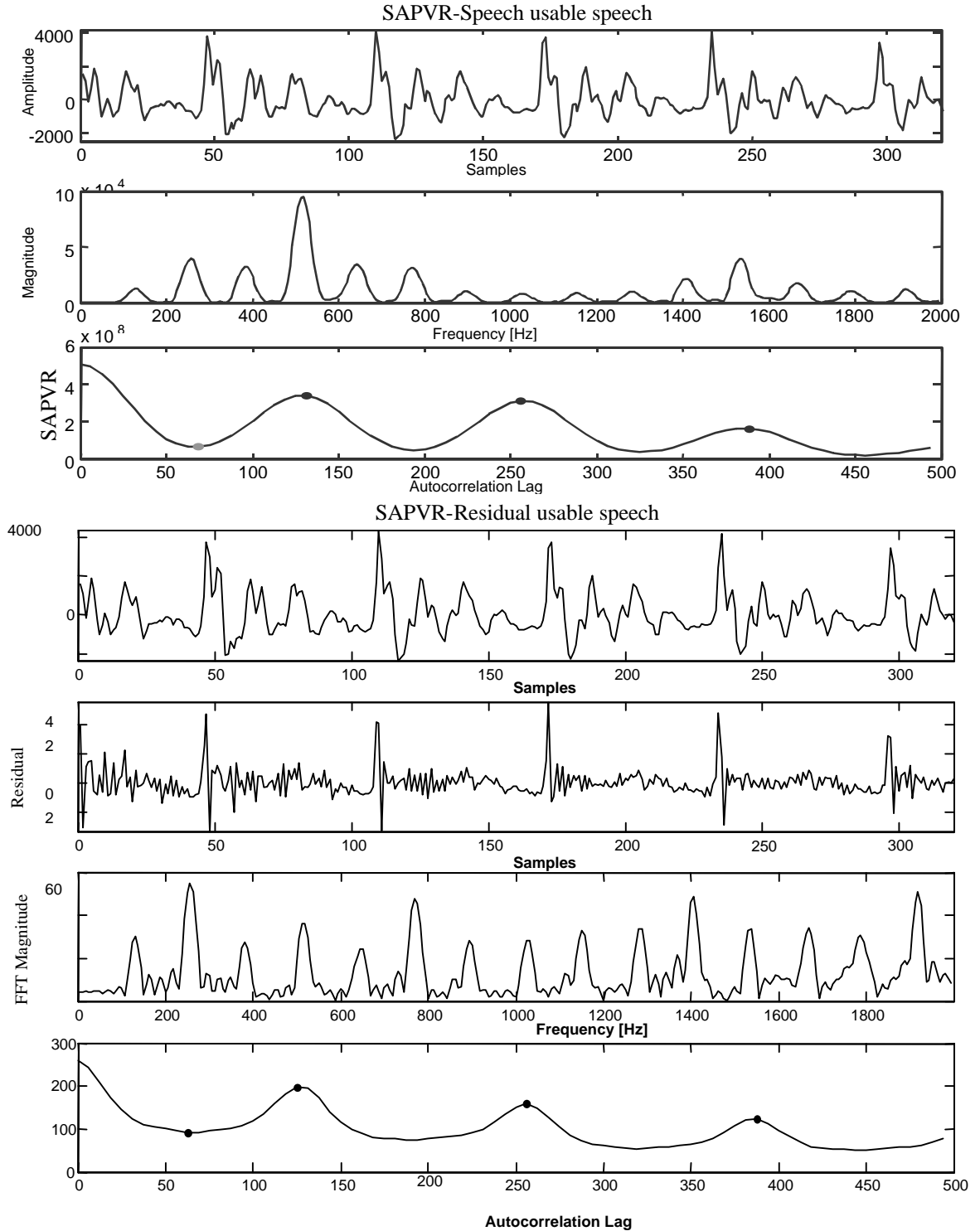


Figure 15:

speech segment. It can be seen here that the magnitude of Spectral autocorrelation in the two cases differs a lot, Figure 22 (c) and (g). Since Residual is the excitation of the speech it has a very low energy.

Table 1 compares results from SAPVR-Speech versus SAPVR-Residual. It can be seen that for the male-male case the SAPVR-Residual system gives 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 decrease in 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

4.4. Discussion.

The purpose of the thesis research is 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 observed.

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.

Research is being carried out to establish cyclostationary as another usable speech measure.

5 REFERENCES

1. Benincasa, D. S., and Savic, M.I., "Voicing state determination of co-channel speech," proc. IEEE ICASSP, pp: 1021-1024, 1998.
2. Chandra, N, Yantorno, R.E., "Usable speech detection using modified spectral autocorrelation peak to valley ration using the LPC residual", IASTED SIP 2002 (submitted).
3. Frazier, R.H. "An adaptive filtering approach toward speech enhancement" M.S. thesis, Dept. of Electrical Engineering, MIT, 1975
4. Gardner, W. A, "Exploration of spectral redundancy in cyclostationary signals." IEEE signal processing magazine, April 1991.
5. Hanson, B.A. and Wong, D.Y. "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech.
6. Kizhanatham, A., Yantorno, R.E., "Co-channel speech detection approaches using cyclostationarity or wavelet transform", IASTED SIP 2002 (submitted).
7. 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.
8. 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.
9. Lee, C.K., and Childers, D.G., "Co-channel speech separation", J. Acoust. Soc. Am., Vol. 83, No.1, pp:274-280, 1988.
10. 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.
11. 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.

12. Morgan, D. P., George, E. B., Lee, L. T., and Kay, S. M. "Co-channel speaker separation" ICASSP-95. pp: 828 -831 vol.1 1995.
13. Morgan, D.P.; George, E.B.; Lee, L.T.; Kay, S.M., "Co-channel speaker separation by harmonic enhancement and suppression", IEEE transactions on speech and audio processing, Volume: 5 Issue: 5, Sept. 1997 pp: 407 – 424.
14. Nakatani, T., Masataka, G., and Hiroshi, G.O. "Localization by harmonic structure and its application to harmonic sound stream segregation", Acoustics, speech, and signal processing, 1996. ICASSP-96. Conference proceedings., 1996, pp: 653 -656 vol. 2.
15. Naylor, J.A., and Porter, J., "An effective speech separation system which requires no a priori information," Proc. IEEE ICASSP, pp: 937-940, 1991.
16. Parsons, P.W., and Weiss, M.R., "Enhancing intelligibility in noise or multi-talker environments," Rome Air Development Center report RADC-TR-75-155, 1975.
17. Parthasarathy, S. and Tufts, D.W. "Excitation synchronous modeling of voiced speech," IEEE Trans. on acoustics, speech, signal processing, vol. ASSP-35, 1987, pp. 1241-1249.
18. Quatieri T.F. and Danisewicz, R.G. "An approach to co-channel talker interference suppression using a sinusoidal model for speech" IEEE Trans. on acoustics, speech, signal processing, vol. 38, pp: 56-69. Jan. 1990.
19. Sayers, B. McA., and Cherry, E.C, "Mechanism of binaural fusion in the hearing of speech," J. Acoust. Soc. Am., 29, pp: 973-987, 1957.
20. Shields, V. C. "Separation of added speech signals by digital comb filtering" M.S thesis, Dept. of Electrical Engineering, MIT, 1970.
21. Smolenski, B. Y., Yantorno, R., E., Benincasa D. S., and Wenndt, S. J., "Co-channel speaker segment separation", ICASSP 2002 (accepted).
22. Stubbs, R.J. and Summerfield, Q., "Effects of signal to noise ratio, signal periodicity, and degree of hearing impairment on the performance of voice separation algorithms." J. Acoust. Soc. Amer., vol. 89, pp. 1383-1396, Mar. 1991.
23. 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.
24. 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.

25. 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.