

DETECTION OF COCHANNEL SPEECH

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: ARVIND RAMAN KIZHANATHAM

ADVISOR: Dr. ROBERT YANTORNO

ABSTRACT

Co-channel speech occurs when one speaker's speech is corrupted by another speaker's speech. Speech recognition systems, speaker identification systems, speech coding systems, gisting and natural language processing systems work on the basis that there is only one speaker's speech. If there is more than one speaker (co-channel speech) in the speech signal the performance of the speech recognition system, speaker identification system etc. will be degraded, for example, speaker identification system will not be able to make a correct decision about the speaker. A co-channel detection system could be used to detect multiple speakers in a speech signal, thereby stopping the processing of speech, hence preventing the degradation of the performance of speech recognition system, speaker identification system etc. The focus of this research is to develop mathematical parameters and criteria, which will be able to identify if there is a single speaker or multiple speakers in the speech signal. Some possible candidates can be cyclostationarity, spectral correlation density function, wavelets, modulation maps, statistical shape analysis and higher order statistics. A general scheme will be presented on how the mathematical approaches can be incorporated into a proposed co-channel detection system.

TABLE OF CONTENTS

ABSTRACT.....	II
1. INTRODUCTION.....	1
2. BACKGROUND.....	5
2.1 CO-CHANNEL SPEECH DETECTION PROBLEM.....	5
2.2 PREVIOUS WORK ON SPEECH DETECTION.....	5
2.2.1 DETECTION OF SPEECH CORRUPTED BY BACKGROUND NOISE.....	5
2.2.1.1 VOICE ACTIVITY DETECTION USING PERIODICITY MEASURE.....	7
2.2.1.2 SPEECH DETECTION USING MICROPHONE ARRAY.....	8
2.2.2 DETECTION OF SPEECH CORRUPTED BY ANOTHER SPEECH.....	10
3. RATIONALE FOR RESEARCH.....	11
3.1 CO-CHANNEL SPEECH DETECTION BASED ON SAPVR METHOD.....	12
3.2 RESULTS AND DISCUSSION OF SAPVR BASED DETECTION SYSTEM.....	14

4. CO-CHANNEL SPEECH DETECTION SYSTEM.....	17
4.1 PROPOSED CO-CHANNEL DETECTION SYSTEM.....	17
4.2 CO-CHANNEL SPEECH DETECTION.....	17
4.3 POTENTIAL CO-CHANNEL DETECTION CANDIDATES.....	18
4.3.1 CYCLOSTATIONARITY AND SPECTRAL REDUNDANCY.....	18
4.3.1.1 CYCLIC AUTOCORRELATION FUNCTION.....	19
4.3.1.2 SPECTRAL CORRELATION DENSITY FUNCTION.....	20
4.3.2 WAVELETS	21
4.3.3 MODULATION MAPS.....	24
4.3.4 STATISTICAL SHAPE ANALYSIS	24
4.3.5 HIGHER ORDER STATISTICS(POLYSPECTRA).....	25
5. RESULTS AND DISCUSSION.....	27
5.1 INITIAL RESULTS AND DISCUSSION OF CYCLOSTATIONARY BASED CO-CHANNEL DETECTION SYSTEM.....	27
5.2 WAVELETS	32
5.3 MODULATION MAPS.....	32
5.4 STATISTICAL SHAPE ANALYSIS.....	33
5.5 HIGHER ORDER STATISTICS.....	33
REFERENCES.....	34

CHAPTER 1

INTRODUCTION

Co-channel speech has been an area of interest and research for over three decades. Co-channel speech occurs when a speaker's speech (target) is degraded by another speaker's speech (interferer). The fact that co-channel speech is still being investigated indicates the problems that speech processing researchers are still confronted with co-channel speech.

Co-channel speech can occur in many common situations, such as when two AM signals are transmitted on the same frequency, or when two people are speaking simultaneously (e.g. when talking on the telephone), or due to cross-talk from a neighboring communications channel. The goal of co-channel speech research has been to extract the target speech or suppress the interfering speech from the composite signal or both. The human auditory system is adept at resolving the speech of one talker amongst many (the cocktail-party effect, Sayers and Cherry, 1957). Computer algorithms, which try to emulate this feature of human auditory system, are successful only to a limited degree (Morgan, *et al.*, 1995; George, *et al.*, 1995; Lee, *et al.*, 1995; and Kay, *et al.*, 1995). co-channel situation poses problems to speech processing tasks like speaker identification, speech recognition, etc., for example, speaker identification system will not be able to make a correct decision about the speaker.

There are systems that are capable of resolving the speech signals mixed with noise, one of them is a Voice Activity Detector (VAD). A Voice Activity Detector (VAD) was developed to extract voiced segments from the mixture of voiced segments and noise (Sohn *et al.*, 1999; Kim *et al.*, 1999; Sung *et al.*, 1999). An advanced voice activity detection algorithm to detect the voice segments from the noise was developed later (Woo *et al.*, 2000; Yang *et al.*, 2000; Park *et al.*, 2000; Lee *et al.*, 2000). When voices interfere over a monophonic channel such as telephone, detection is much more difficult and voice may mask the noise (Chen and Ser, 1999).

Until recently, researchers have approached the detection of a speech mixed with noise, thereby using Voice Activity Detector (VAD) to extract the speech. If detection of speech mixed with noise were the case, intelligibility and quality of the extracted speech would be the important characteristic considered in co-channel speech processing. However, in co-channel speech where a speech is corrupted by another speech, the existing VAD method cannot be used for detecting the target speech and a new approach to detection of speech signal needs to be formulated. Cepstral and pitch prediction features are used for determining number of speakers in co-channel speech (Lewis and Ramachandran, 1998). A Spectral Autocorrelation Peak Valley Ratio (SAPVR) based co-channel speech detection system was developed (Yantorno, 2000) where a spectral autocorrelation domain is used to make a decision about whether the speech frame under analysis is co-channel .

1.2 PROBLEM STATEMENT

Co-channel speech occurs when one speaker's speech is corrupted by another speaker's speech. Speech recognition systems, speaker identification systems, speech coding systems, gisting and natural language processing systems work on the basis that there is only one speaker's speech. If there is more than one speaker (co-channel speech) in the speech signal the performance of the speech recognition system, speaker identification system etc. will be degraded, for example, speaker identification system will not be able to make a correct decision about the speaker.

1.3 SCOPE OF RESEARCH

The goal of the research presented in this proposal is to develop a system, which will be able to detect co-channel speech segments in a co-channel environment. A co-channel detection system could be used to detect multiple speakers in a speech signal, thereby stopping the processing of speech, hence preventing the degradation of the performance of speech recognition system, speaker identification system etc.

1.4 ORGANIZATION OF THE PROPOSAL

Chapter 2 talks about the background of speech detection, basically it talks about detection of speech corrupted by background noise and detection of speech corrupted by another speech.

Chapter 3 talks about the research on co-channel speech detection and how it initiated this research.

Chapter 4 talks about the co-channel speech detection unit and the possible measures of co-channel speech detection.

Chapter 5 gives the initial results and discussion of cyclostationarity based co-channel speech detection system and how some other measures can be used in co-channel speech detection.

CHAPTER 2

BACKGROUND

2.1 CO-CHANNEL DETECTION PROBLEM

Co-channel detection problem can be explained as follows, given a composite speech signal which is the mixture of one or more speaker we need to detect the frames of speech where there is more than one speaker is present.

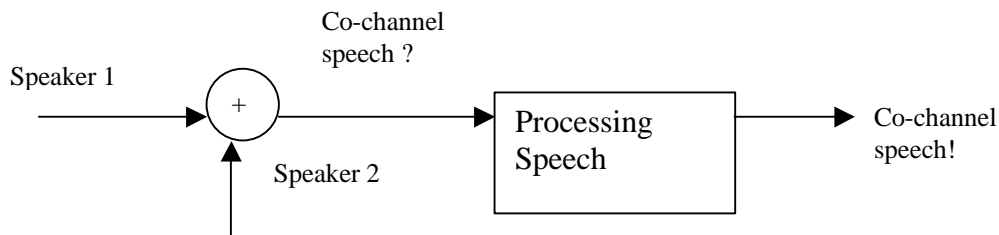


Figure 2.1 co-channel speech detection system

In the above fig 2.1 T is the target speech (speaker 1) and I interfering speech (speaker2), the composite speech signal will have frames of speaker 1 alone, frames of speaker 2 alone and frames of speaker 1 and speaker 2 together. When we say co-channel speech we are assuming that there are only two speakers. The goal of the research is to develop a system, which detects frames that have a combination of speaker 1 and speaker 2.

2.2 PREVIOUS WORK ON SPEECH DETECTION

2.2.1 DETECTION OF A SPEECH CORRUPTED BY BACKGROUND NOISE

The initial work on speech detection were based on the fact that the voiced speech were mixed with noise. A Voice Activity Detection (VAD) algorithm was developed in which noise is filtered out in the frequency domain (Woo *et al.*, 1999, Yang *et al.*, 1999, Park *et*

al., 1999). Their algorithm uses log energy parameters to detect the noise period with which time-varying noise characteristics can be reliably estimated. The advantage of the technique is that it can prevent incorrect detections caused by unvoiced or nasal sounds with high frequency components being covered by noise with low frequency components. The algorithm is suitable for real time implementation with one microphone. Also, a speaker independent speech recognition system has been implemented for navigation using fixed Point Oak DSP systems, which incorporates the proposed VAD algorithm. The system enhanced the recognition rates for 12 isolated command words to 94.52% compared with the 80.7% of the baseline recognizer.

A modified Voiced Activity Detector (VAD) was developed for the application to variable-rate speech coding (Sohn *et al.*, 1999, Kim *et al.*, 1999, Sung *et al.*, 1999). The developed VAD employs the decision-directed parameter estimation method for the likelihood ratio test. In addition, they developed an effective hangover scheme, which considers the previous observations by a first-order Markov process modeling speech occurrences. According to their simulation results, the proposed VAD shows significantly better performance than the earlier VAD in low signal-to-noise ratio (SNR) and vehicular noise environment.

Visible speech cues, which are the visible movements of the speech were used to detect speech (Grant, 1999 and Seitz, 1999). Estimation of periodicity measure of the target speaker was used in speech detection (Tucker, 1992).

A microphone array was used in speech detection (Chen *et al.*, 1999 and Ser *et al.*, 1999). An explicit expression is first deduced for representing the signal-to-noise ratio (SNR) of each signal segment. A constant SNR threshold is then used to discriminate between speech and non-speech signals.

Some of the co-channel speech detection systems, which were outlined above are briefly explained below:

2.2.1.1 VOICE ACTIVITY DETECTION USING A PERIODICITY MEASURE

Figure 2.2.1.1 shows the structure of a Least-Squares Periodicity Estimator (LSPE) based VAD. In addition to the main LPSE section, there are the pre-processing and post processing sections and an energy detector. The LPSE calculation uses nonoverlapping 25ms frames on a 200-1000 Hz bandwidth signal. A narrow bandwidth is used to minimize the probability of an inband interference signal, while still allowing effective periodicity detection. The signal is two-times over-sampled at 4kHz to give extra resolution for the periodicity detection.

The energy detector is included to prevent the detection of very low-level signals in the presence of larger signals. If Automatic Gain Control (AGC) has not been applied to the input signal, the detection threshold needs to adapt to the input signal level. The purpose of energy detector is to pass signals that might be speech, while rejecting any signal that is definitely not speech.

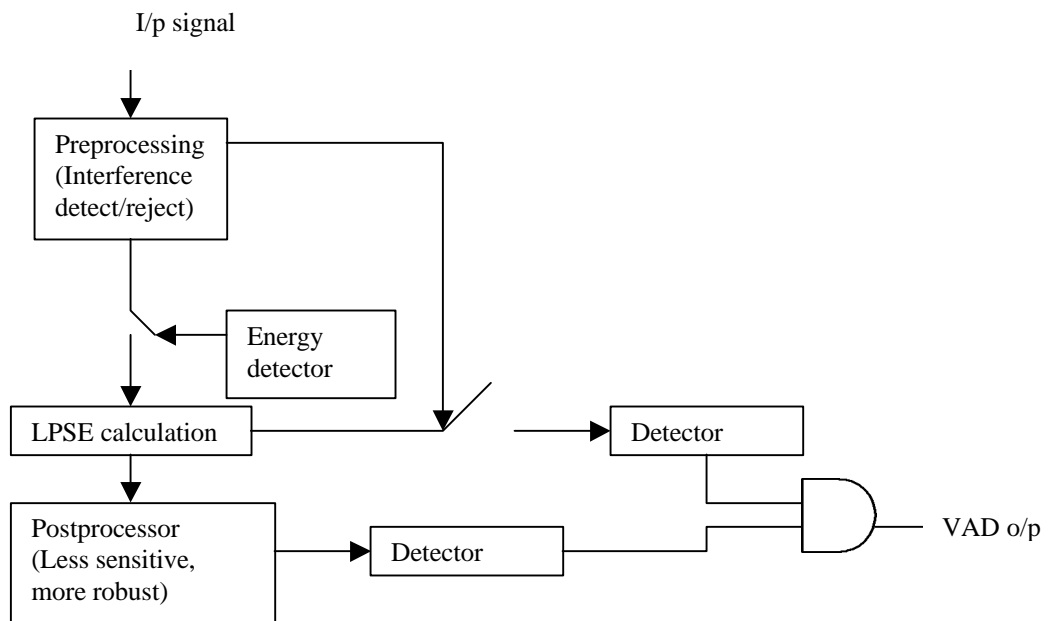


Figure 2.2.1.1 Least-squares periodicity estimator (LSPE) based VAD.

The preprocessor is needed to detect, and if possible remove periodic interference. When it detects interference, the preprocessor suppresses the LPSE detector, because there may be residual periodicity even after tone removal. The postprocessor is required to provide less sensitive but more robust detector in the presence of interference. Different environments will have different interferences, so that exact nature of the preprocessor will depend on the expected type of interference.

2.2.1.2 SPEECH DETECTION USING MICROPHONE ARRAY

Microphone array was used to detect speech in a noisy environment (Chen and Ser, 2000). An explicit expression was first deduced for representing the signal-to-noise ratio (SNR) of each signal segment. A constant SNR threshold was then used to discriminate

between speech and nonspeech signals. They considered a linear microphone array of M wideband sensors, which receive the plane wave generated by a wideband source (speaker) in the presence of Gaussian white noise wavefield. The noise was assumed to be independent of the speaker, and the noises in different frequency bins are assumed to be uncorrelated.

From the algorithm described above, we can obtain the SNRs for all the frequency bins. These SNR's are normally not of the same value. For example, for pitch or formant frequencies, the SNR would be much higher. Through simulations and experiments, Chen and Ser, (2000) found that for the purpose of detecting the presence of speech signal, the first two or three largest SNR's could be averaged to represent the SNR of the analysed segment. A predetermined SNR threshold can then be used to classify all the segments into speech and nonspeech signals.

Estimation errors would result from the above method due to the nonstationary nature of speech signals, the limited length of observed data and the size of the microphone array. In addition, some modeling inaccuracies, such as coloured and correlated noise, array sensor position errors, etc. could also affect the performance of the method. This is the drawback of the system developed by Chen and Ser, but however since they considered only one speaker to be present whose power is dominant most of the time, the proposed system works well.

2.2.2 DETECTION OF SPEECH CORRUPTED BY ANOTHER SPEECH

Spectral Auto Correlation Peak Valley Ratio (SAPVR) based co-channel detection system was developed by Yantorno (2000). For the SAPVR method, it has been determined that one can use the spectral autocorrelation domain information to make a decision about whether the speech frame under analysis is usable. Therefore, if a frame of speech is determined to be voiced, and there is only one speaker talking or the other speaker's energy is low, either because it is unvoiced speech or simply just low energy, then one would expect it to have well defined spectral autocorrelation. This means that SAPVR number should be large for the frame under investigation should have a large. Therefore, if one were to use a method for detecting voiced speech frames and then perform the SAPVR, one would expect the SAPVR value to be high for speech with a well-defined spectrum with nice harmonic structure. If there is little or no harmonic structure (as is usually the case for co-channel speech) then one would expect the SAPVR value to be low thereby detecting the frame as co-channel.

It should be noted that, none of the methods discussed above except the ones by Lewis and Ramachandran (1998) and Yantorno (2000) dealt with co-channel speech.

CHAPTER 3

RATIONALE FOR RESEARCH

As discussed earlier, the usual approach of co-channel speech processing is to enhance the target speaker's speech or suppress the interferer's speech or both. This research is focused on developing a novel method in which extraction of speech is not the initial concern per se but rather determining which portions of speech is co-channel. This co-channel speech detector, will avoid the co-channel speech to degrade the performances of speaker identification system etc., will be an important unit of a more complex system. Hence, the scope of this research is the development of co-channel speech detection techniques.

The motivation for the current research is based on co-channel speech detection system performance experiments under additive noise and corrupting speech conditions by Yantorno (2000). His investigations, where he used SAPVR, have revealed that co-channel speech detection system works better for male speech when compared to female speech. The investigations and results are outlined below

3.1 CO-CHANNEL SPEECH DETECTION BASED ON SAPVR METHOD

Figure 3.1 shown below is the system, which detects the co-channel speech based on SAPVR (Yantorno, 2000). This system is very similar to the one, which has been used for usable speech detection (Krishnamachari *et al.*, 2000). The input speech signal is first windowed and speech is then classified as voiced speech or not using voiced speech detection system (spectral flatness). FFT is performed on the voiced speech and then the autocorrelation of the speech signal is computed and the resulting function is compared with a preset threshold and if it is above the threshold it is co-channel speech, else it is not. If a frame of speech is determined to be voiced, and there is only one speaker talking or the other speaker's speech is low energy, either because it is unvoiced speech or simply just low energy, then one would expect it to have a well-defined spectral autocorrelation, which means that it should have a large SAPVR number. Therefore, if one were to use a method for detecting voiced speech frames and then perform the SAPVR, one would expect the SAPVR value to be high for speech with a well-defined spectrum with nice harmonic structure, whereas if there is little or no harmonic structure (as is usually for co-channel voiced speech) then one would expect the SAPVR value to be low.

It should be noted that we cannot detect co-channel speech all of the time but only part of the time and that is during voicing, i.e. for voiced speech. Therefore one must have a reliable method for ensuring that the portions to be tested are voiced.

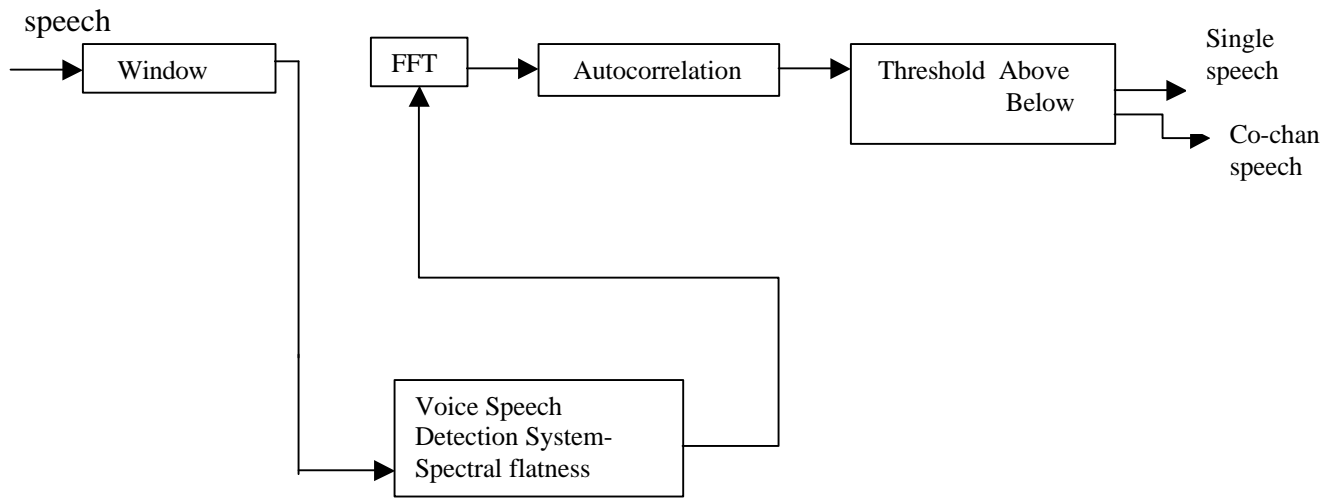


Figure 3.1 Co-channel speech detection system based on SAPVR.

Yantorno (2000) performed a series of tests using a standard approach for determining voiced-unvoiced-silence of an utterance, i.e. energy and zero-crossings. However, the traditional voicing-unvoiced-silence approach has limitation, in that one must determine the zero crossings thresholds. However, there is another method which allows one to extract voiced speech in a straightforward way, and that method is called Spectral Flatness. For speech, the spectral flatness can vary from 0dB for unvoiced to -60dB for voiced. A plot of spectral flatness study is shown below in figure 3.2a

Spectral Flatness Measure is given by,

$$SFM_{dB} = 10\log_{10}(G_m/A_m) \quad (3.1)$$

where,

$G_m = \sqrt[N]{\prod \text{mag}(i)}$ and $A_m = \frac{\sum \text{mag}(i)}{N}$. G_m -geometric mean and A_m -arithmetic mean,

$\text{mag}(i)$ is each of the magnitude spectral lines and N is the number of FFT points or spectral lines.

3.2 RESULTS AND DISCUSSION OF SAPVR BASED DETECTION SYSTEM

Figure 3.1 above shows the SAPVR based speech detection system. Initial results using the spectral flatness and SAPVR measure for co-channel detection are shown in the figure 3.2. Figure 3.2 shows the speech file and the associated spectral flatness per frame. Figure 3.3 illustrates the effectiveness of using spectral flatness, the threshold for accepting the frame was -40dB , therefore all frames with a measure less than -40dB are considered as being voiced. The frames that are indicated as above the threshold in figure 3.3 are actually frames below the threshold, i.e., more negative. We know that if one were to use a method for detecting voiced speech frames and then perform the SAPVR, one would expect the SAPVR value to be high for speech with a well-defined spectrum with nice harmonic structure. If there is little or no harmonic structure (as is usually for co-channel voiced speech) then one would expect the SAPVR value to be low, results using this concept are shown in figure 3.4. Note, positive rectangles indicate co-channel speech detected and negative rectangles indicate false alarms.

Spectral Flatness Study - Co-channel Data fvmh₁₀₈.k.bin & macd_zero_s.il.bin 23-Apr-2001

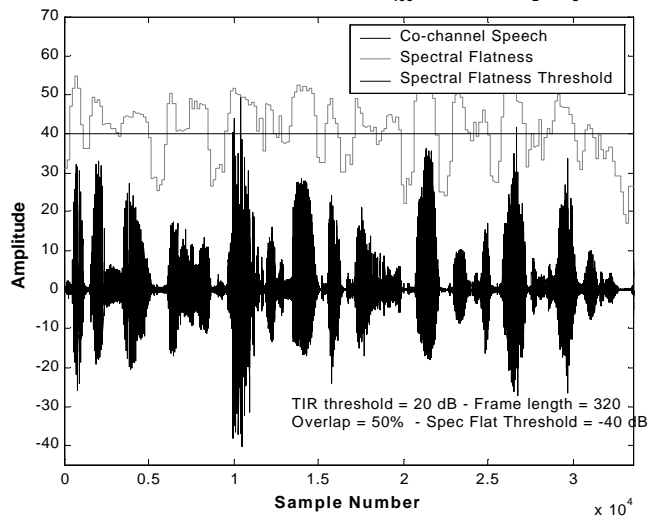


Figure 3.2 Spectral Flatness Study- a measure of detecting voiced speech

Spectral Flatness Study - fvmh₁₀₉.k.bin & macd_zero_s.il.bin 23-Apr-2001

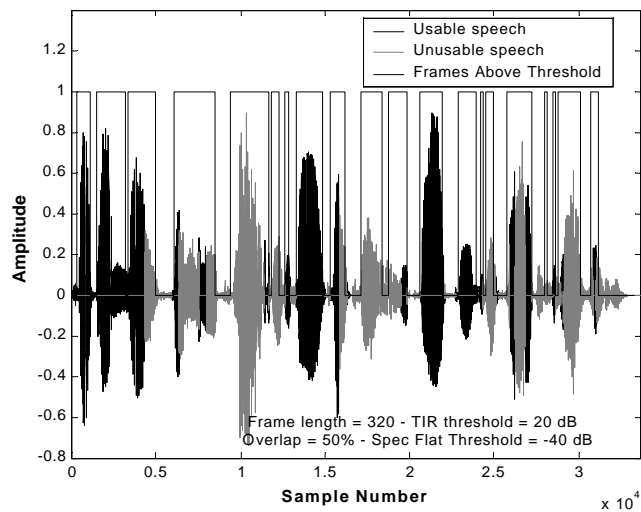


Figure 3.3 Spectral Flatness Study- black - usable speech, grey – unusable speech
black lines – voiced speech

Co-channel Detection Study - fvmh₁₀₈.k.bin & madc₂ero_sil.bin 23-Apr-2001

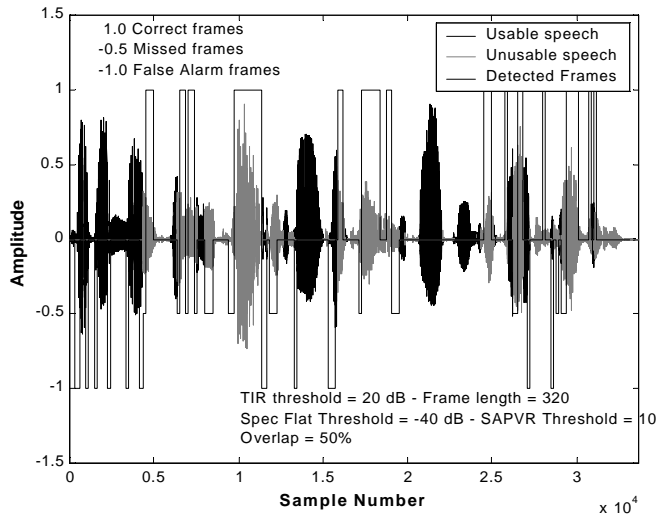


Figure 3.4 SAPVR method for co-channel detection- 1.0 correct frames, -0.5 missed frames
-1.0 false alarm frames

The operational characteristics in terms of percent correct, percent missed and percent false alarms for both speakers were performed. Results of the system showed that for a female-male co-channel speech, 66% were correct, 42% were false alarms, 34% were missed with a confidence level of 49% and for a male-male co-channel speech the results were 66% were correct, 21% were false alarms and 34% were missed and with a confidence level of 74%. It is interesting to note that when using the SAPVR for co-channel detection that the system performs better with male speech than with female.

CHAPTER 4

CO-CHANNEL SPEECH DETECTION SYSTEM

4.1 DETECTION OF CO-CHANNEL SPEECH

The problem we are attempting to solve is, given a speech signal, which has co-channel speech (target speech corrupted by interfering speech) segments in it, how to spot and detect the co-channel speech segments of the speech signal.

4.2 PROPOSED CO-CHANNEL SPEECH DETECTION SYSTEM

The co-channel speech detection unit, which can be used as the front-end of the potential next generation speech processing system, is shown in figure 4.2

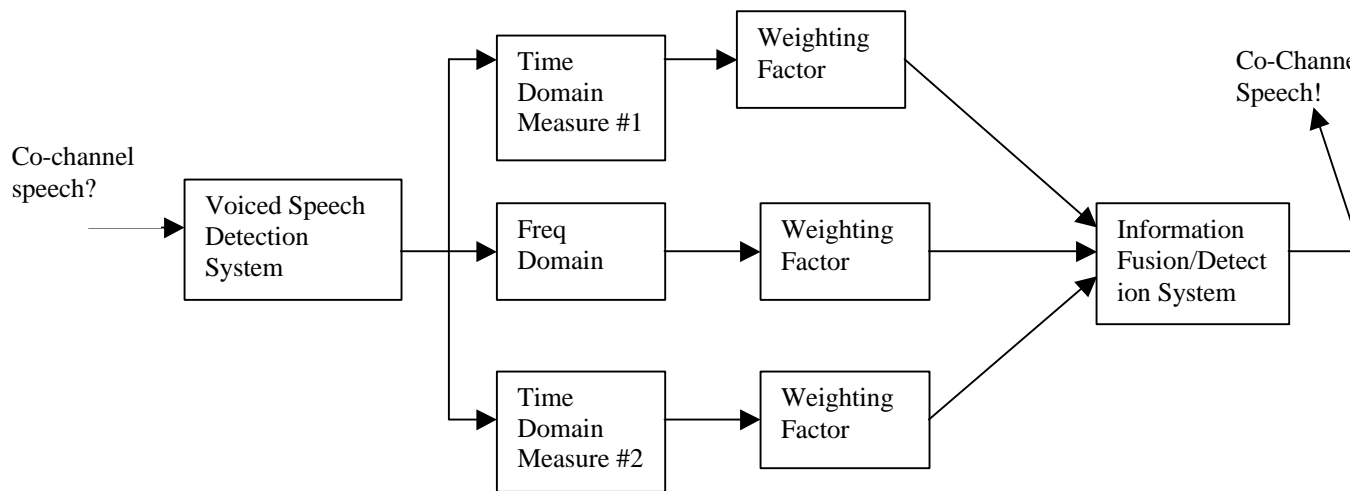


Figure 4.2 Proposed co-channel speech detection system

4.3 POTENTIAL CO-CHANNEL DETECTION CANDIDATES

Some potential candidates for detecting co-channel speech are:

1. Cyclostationarity
2. Wavelets
3. Statistical Shape Analysis.
4. Higher Order Statistics (PolySpectra).
5. Modulation Maps

4.3.1. CYCLOSTATIONARITY AND SPECTRAL REDUNDANCY

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

4.3.1.1 CYCLIC AUTOCORRELATION FUNCTION

A signal can be modeled as cyclostationary if its statistical parameters vary in time with single or multiple periodicities. Cyclic autocorrelation function is used to identify whether a random signal exhibits cyclostationarity. The conventional autocorrelation function of a signal $x(t)$ for lag τ is defined as

$$R_x(\tau) = \langle x(t+\tau/2) x^*(t-\tau/2) \rangle \quad (4.1)$$

where $*$ denotes complex conjugation and $\langle \cdot \rangle$ denotes ensemble averaging. The cyclic autocorrelation function is defined as

$$R_x^{\alpha}(\mathbf{t}) \stackrel{\Delta}{=} \langle x(t+\mathbf{t}/2) x^*(t-\mathbf{t}/2) e^{-j2p\alpha t} \rangle \quad (4.2)$$

Note that $R_x^{\alpha}(\mathbf{t})$ reduces to $R_x(\tau)$ for $\alpha = 0$. A random signal $x(t)$ exhibits second order periodicity if and only if the Power Spectral Density (PSD) of the delay-product signal for some delays τ contains spectral lines at some nonzero frequencies $\alpha \neq 0$, such a signal

is said to exhibit cyclostationarity of second order. The cyclic autocorrelation function can actually be expressed as a conventional cross correlation function

$$R_{uv}(\mathbf{t}) \stackrel{\Delta}{=} \langle u(t + \mathbf{t}/2)v^*(t - \mathbf{t}/2) \rangle = R_x^a(\mathbf{t}) \quad (4.3)$$

where,

$$u(t) = x(t) e^{-j\pi\alpha t} \quad (4.4)$$

$$v(t) = x(t) e^{+j\pi\alpha t} \quad (4.5)$$

4.3.1.2 SPECTRAL-CORRELATION DENSITY FUNCTION:

The relation between conventional autocorrelation function and Power Spectral Density (PSD) function is given by the well known Wiener-Khintchin theorem, that is, autocorrelation function and PSD form Fourier transform pairs. If $R_x(t)$ is the autocorrelation function of a signal $x(t)$ and $S_x(f)$ is its PSD, then we have the relation

$$S_x(f) = \int_{-\infty}^{+\infty} R_x(\mathbf{t}) e^{-j2\pi f \mathbf{t}} d\mathbf{t} \quad (4.6)$$

Likewise, the Fourier transform of the cyclic autocorrelation function $R_x^a(\mathbf{t})$ is called Cyclic Spectral Density or Spectral Correlation Density (SCD). Thus, SCD is defined as

$$S_x^a(f) = \int_{-\infty}^{+\infty} R_x^a(\mathbf{t}) e^{-j2\pi f \mathbf{t}} d\mathbf{t} \quad (4.7)$$

Since the cyclic autocorrelation could be expressed as conventional cross-correlation, the SCD can also be expressed as conventional cross-spectral density $S_{uv}(f)$. That is

$$S_x^a(f) \equiv S_{uv}(f),$$

Where, $S_{uv}(f)$ is the cross-spectral density of the signals $u(t)$ and $v(t)$ defined in the previous section. The above identity suggests an appropriate normalization for $S_x^a(f)$, as

long as the PSD of $x(t)$ contains no spectral lines at either of the frequencies $f \pm \alpha/2$, the correlation of the spectral components is actually a covariance, since the means of the spectral components are zero (Gardner, 1987). When normalized by the geometric mean of the corresponding variances, which are given by

$$S_u(f) = S_x(f + \alpha/2) \quad (4.8)$$

and

$$S_v(f) = S_x(f - \alpha/2) \quad (4.9)$$

the covariance becomes a correlation coefficient

$$\frac{S_{uv}(f)}{[S_u(f)S_v(f)]^{1/2}} = \frac{S_x^a(f)}{[S_x(f + \mathbf{a}/2)S_x(f - \mathbf{a}/2)]^{1/2}} \stackrel{\Delta}{=} \mathbf{r}_x^a(f) \quad (4.10)$$

Since $|\rho_x^\alpha(f)|$ is bounded to the interval $[0,1]$, it is a convenient measure of the degree of local spectral redundancy that results from spectral correlation. For example, if

$|\rho_x^\alpha(f)| = 1$, we have complete spectral redundancy at $f + \alpha/2$ and $f - \alpha/2$.

4.3.2 WAVELETS

The Wavelet Transformation (WT) is defined as the convolution of a signal $x(t)$ with a wavelet function $\Psi(t)$ shifted in time by a translation parameter and dilated by a scale parameter. Wavelets are mathematical functions that divide data into different frequency components, and then analyze each component with a resolution matched to its scale. A wavelet is a function with some special properties. Wavelet is one that has small

concentrated burst of finite energy in the time domain and it exhibits some oscillations in time. The continuous Wavelet Transform (CWT) is defined as the convolution of a signal $x(t)$ with a wavelet function $\Psi(t)$ shifted in time by a translation parameter 'b' and a dilation parameter 'a' and is given by

$$\text{CWT}_x(b, a) = \int_{-\infty}^{\infty} x(t) \Psi^*(t-b)/a dt \quad (4.11)$$

Where the $\Psi(t)$ is the wavelet transform and $\Psi^*(t)$ is the complex conjugate of $\Psi(t)$. The wavelet is either compressed or expanded depending on the choice of 'a'. Hence the CWT can extract both the local and global variations in the signal $x(t)$. The CWT is defined as the Dyadic Wavelet Transform ($D_y\text{WT}$) if the scale parameter 'a' in equation 4.11 is discretized along the dyadic sequence 2^j , where $j=1,2,\dots$. The $D_y\text{WT}$ is defined as

$$D_y\text{WT}(b, 2^j) = \int_{-\infty}^{\infty} x(t) \Psi^*(t-b)/2^j dt \quad (4.12)$$

The $D_y\text{WT}$ exhibits various interesting properties such as linearity, time shift invariance and the detection of sharp and slow variations in the signal, which makes it an useful tool for the analysis of speech signals and hence could be a good candidate for detection of co-channel speech.

4.3.3 MODULATION MAPS

Modulation mapping is a technique that provides an effective method for segregating voiced speech from competing background activity. The maps are constructed by computing modulation spectra in a bank of auditory filters. The reassigned spectrum, which is a new time frequency representation, allows a reduction in the window size from 200 ms to 50 ms without a loss of performance (Kodera *et al.*, 1978; Auger *et al.*, 1995).

The reassignment method was first proposed by Kodera *et al.*, (1978) in order to improve the resolution of the spectrogram. The method assigns the value of the spectrogram to the center of gravity in the analysis window as in the normal Fourier Transform. When applied to the spectrogram, the point of assignment is moved in time and frequency. However, when time information is not required, only frequency displacement is computed, leading to the definition of the reassigned spectrum. The frequency displacement uses the phase of the Fourier Spectrum, and can be computed using the ratio of the Fourier Transforms as

$$V_r = V - \text{Im} \left\{ \frac{STFT_{dh}(V) \cdot STFT_h^*(V)}{|STFT_h(V)|^2} \right\} \quad (4.12)$$

Where,

V_r is the reassigned frequency point,

V is the frequency point using the normal Fourier Transform,

$STFT_h$ is the Short Time Fourier Transform using the window h , and

dh is the time derivative of the window h .

The reassigned spectrum is resampled at regular points. The value of the spectrum at each point is computed by summation of all the points falling into the same bin. Bins containing no frequency points are set to 0.

Meyer *et al.*, (1997) tested the modulation maps on a ‘double vowel’ identification task that has been used extensively in psychophysical experiments. The approach used was to split the signal into 32 bandpass filtered channels, using an auditory filter bank. Fourier transforming the half-wave rectified and low pass filtered output further expanded each

channel. The map codes amplitude modulation frequency against the channel frequency. This representation allows two simultaneous voiced sounds to co-exist as separate but interleaved patterns, provided they have different pitch. Investigations by Meyer *et al.*, (1997) revealed that, in a 'double vowel' identification experiment, if 204.8 ms windows are used, both conventional and reassigned spectrogram require around 6 Hz separation to identify the target. If 51.2 ms windows are used, the conventional spectrogram requires 25 Hz pitch separation to identify the second vowel, while the reassigned spectrogram is still able to separate vowels with only 6% pitch difference.

From the discussion above, the reassigned spectrum seems to be a candidate in identifying co-channel speech under co-channel conditions. Further investigations are needed to identify parameters that could be used to spot co-channel speech segments from reassigned spectrum.

4.3.4 STATISTICAL SHAPE ANALYSIS

Shape is defined as the geometrical information that remains when location, scale and rotational effects are filtered out from an object. For an object, a landmark is a point of correspondence on each object that matches between and within populations (Dryden and Mardia, 1998). Shape analysis is of great interest in a wide variety of disciplines like image analysis, archaeology, geography, geology, agriculture and genetics.

We note that for a voiced frame, its time domain waveform, its LPC spectrum and its autocorrelation of magnitude spectrum have well defined structure, and this structure is degraded by interfering speech. There are distance measures available, which indicate how dissimilar the shape of two objects is. An example is full Procrustes distance. Once we fix the landmarks and consequently the shape of a voiced speech (or its LPC spectrum or its spectral autocorrelation), the corresponding shape of frames from a co-channel speech could be compared with this reference by computing the distance measure. A threshold could then be fixed on the magnitude of the distance measure, depending on how serious the interference is, and hence it could be determined if co-channel speech is present in a frame.

4.3.5 HIGHER ORDER STATISTICS (POLYSPECTRA)

We are familiar with power spectrum, which is the Fourier Transform of the autocorrelation function of a random process. The k-th order polyspectrum is defined by (Priestly, 1981; Nikias and Raghuveer, 1987)

$$C_k(\omega_1, \omega_2, \dots, \omega_{k-1}) =$$

$$\sum_{t_1=-\infty}^{\infty} \dots \sum_{t_{k-1}=-\infty}^{\infty} |c_k(t_1, t_{21}, \dots, t_{k-1})| < \infty \cdot \exp(-j(\mathbf{w}_1 \mathbf{t}_1 + \mathbf{w}_2 \mathbf{t}_2 + \dots + \mathbf{w}_{k-1} \mathbf{t}_{k-1})) \quad (4.13)$$

A sufficient condition for the existence of the polyspectrum $C_k(\omega_1, \omega_2, \dots, \omega_{k-1})$ is that the associated kth-order cumulant $c_k(\tau_1, \tau_2, \dots, \tau_{k-1})$ be absolutely summable as shown by

$$\sum_{t_1=-\infty}^{\infty} \dots \sum_{t_{k-1}=-\infty}^{\infty} |c_k(t_1, t_{21}, \dots, t_{k-1})| < \infty \quad (4.14)$$

The power spectrum, bispectrum and trispectrum are special cases of the k th-order polyspectrum. Specifically, we may state the following:

1. For $k=2$, we have the ordinary power spectrum.

$$C_2(\mathbf{w}_1) = \sum_{t_1=-\infty}^{\infty} c_2(t_1) \exp(-j\mathbf{w}_1 t_1) \quad (4.15)$$

which is a restatement of Einstein-Wiener-Khintchine relation.

2. For $k = 3$, we have the bispectrum, defined by

$$C_3(\mathbf{w}_1, \mathbf{w}_2) = \sum_{t_1=-\infty}^{\infty} \sum_{t_2=-\infty}^{\infty} c_3(t_1, t_2) \exp[-j(\mathbf{w}_1 t_1 + \mathbf{w}_2 t_2)] \quad (4.16)$$

3. For $k=4$, we have the trispectrum, defined by,

$$C_4(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3) = \sum_{t_1=-\infty}^{\infty} \sum_{t_2=-\infty}^{\infty} \sum_{t_3=-\infty}^{\infty} c_4(t_1, t_2, t_3) \exp[-j(\mathbf{w}_1 t_1 + \mathbf{w}_2 t_2 + \mathbf{w}_3 t_3)] \quad (4.17)$$

The k th-order cumulant $c_k(\tau_1, \tau_2, \dots, \tau_{k-1})$ and k th-order polyspectrum $C_k(\omega_1, \omega_2, \dots, \omega_{k-1})$ form a multidimensional Fourier transform. In general, $C_k(\omega_1, \omega_2, \dots, \omega_{k-1})$ is complex for order k higher than 2. Moreover, the polyspectrum is a periodic function with period 2π .

An outstanding property of polyspectrum is that all polyspectra of higher order than two vanish when the process is Gaussian. This property is a direct consequence of the fact that all the joint cumulants of order greater than 2 are identically zero. Thus higher-order spectra measure the departure of a stochastic process from Gaussianity (Priestly, 1981; Nikias and Raghuveer, 1987).

CHAPTER 5

RESULTS AND DISCUSSION

As mentioned before the problem we are trying to solve is to detect co-channel speech (a speaker's speech corrupted by another speaker's speech). The following section discusses how cyclostationarity can be used and initial results of the cyclostationarity based co-channel detection system.

5.1 INITIAL RESULTS AND DISCUSSION OF CYCLOSTATIONARITY BASED DETECTION SYSTEM

The cyclostationarity method for detecting the number of speakers in a speech signal was proposed Wenndt, S. (2000). It works as follows, first each frame of speech is classified as voiced, unvoiced or silence and the sampling rate are chosen to be 16000 using voicing state determination algorithm. To determine whether the frame is voiced, unvoiced or silence the zero threshold is set depending upon the frame size which is 512. Now the total number of frames in the data (length of the speech data/frame size) is calculated. The beginning and end point of each speech frame are located. The energy, zero crossing and zero thresholds are calculated. From the above data it is determined if the frame is voiced, unvoiced or silence.

A frame is termed to be a "voiced" if the zero crossings is less than or equal to zero threshold and zero crossings is greater than 0, and energy is greater than energy threshold.

A frame is termed to be a “unvoiced” if the energy is greater than energy threshold and zero crossing is greater than zero threshold, and it is termed as a “silence” if energy is less than or equal to energy threshold.

After determining if the frame is voiced, unvoiced or silence the following procedure is adopted.

If it is a voiced frame then Conjugate Cyclic Correlation is used where in the Hilbert transform of the signal is performed. Say $s(n)$ is the Hilbert transform of the given voiced frame then $s(n)*s(n+\tau)$ is computed and the magnitude is computed. If there are harmonic relations then there is a single speaker and if not there are 2 speakers.

If it is an unvoiced frame then DFT of the frame is performed and if there are harmonic relations then 2 speakers are present and if not there is only one speaker is present. This procedure is shown in the block diagram (figure 5.1).

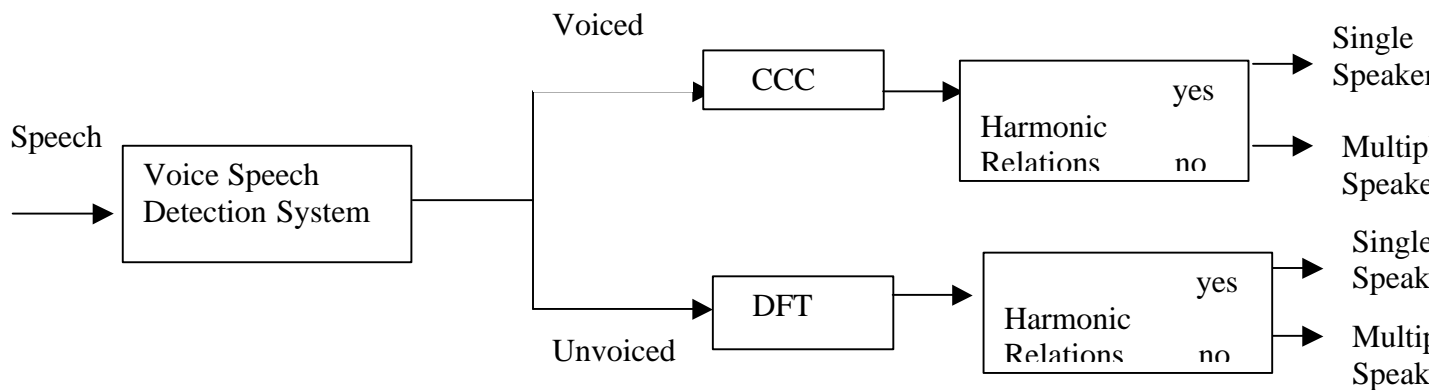


Figure 5.1 Cyclostationarity Based Co-channel Speech Detection System

The results of co-channel speech detection based on cyclostationarity are shown in figures below. Figure 5.2 shows the result using the cyclostationarity as a measure for co-channel detection for a voiced speech, as we can see it has well defined structure with harmonically related peaks. Since it is a voiced speech, as described before we look at the Hilbert transform of the signal to see if there are harmonic relations then there is only one speaker or else multiple speakers. When we come down to figure 5.3 which is for unvoiced speech we see some distortion in the structure and since it is a unvoiced speech we look into the DFT of the signal and look for harmonic relations and based on that we decipher the number of speakers in the speech signal. Figure 5.4 which is for silence it is nothing but distortion (no proper structure). Further investigations are needed to estimate the efficiency of the co-channel system based on Cyclostationarity.

The point to be noted is though the speech signal which is fed as input to the cyclostationarity system is a combination of two speech signals the co-channel speech has randomly mixed frames of single speaker and two speakers and hence we can't detect co-channel speech all the time. If we have a co-channel speech, which has sequential

combination of single speaker, and two speakers it will be easy to estimate the efficiency of the system since we know which frame is single speaker and which frame is two speaker. This is another area of research which will be generating artificial data which has a known combination of single speaker and two speakers.

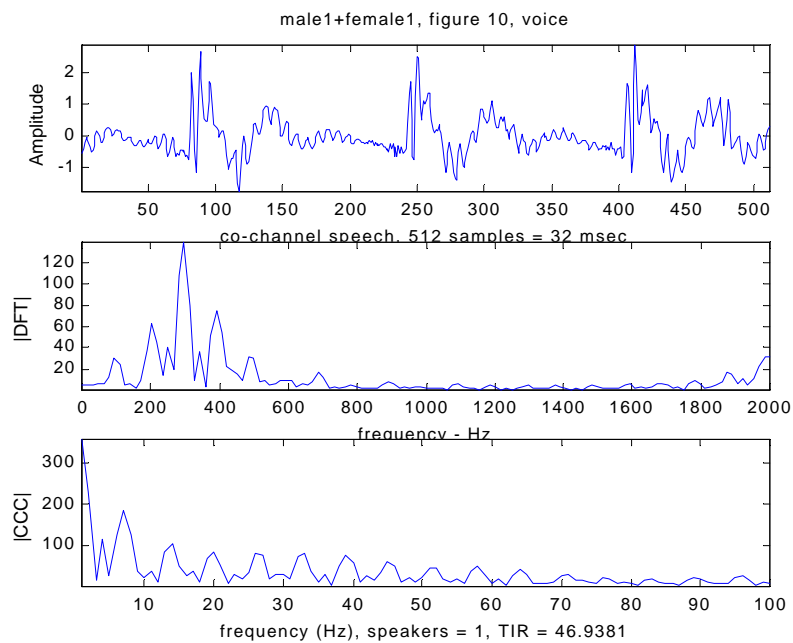


Figure 5.2 Cyclostationarity as a method of co-channel detection system for voiced speech

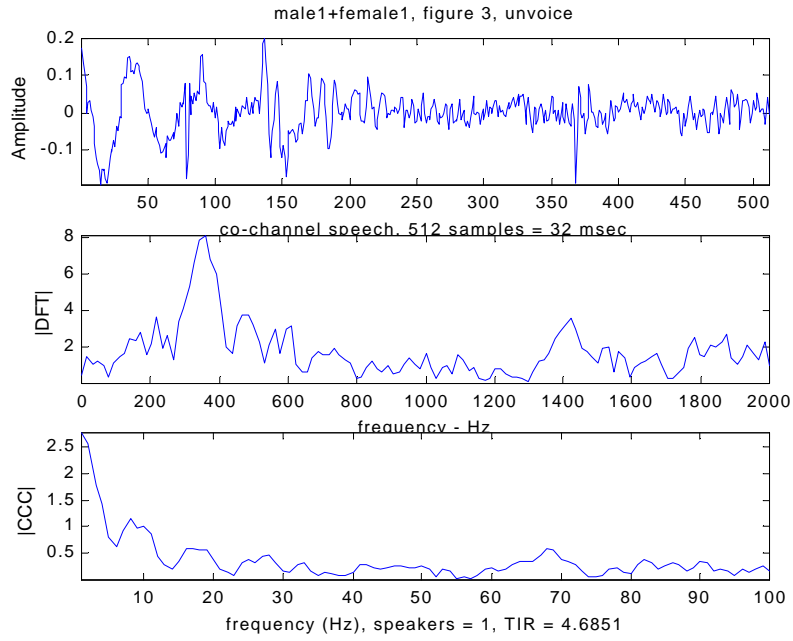


Figure 5.3 Cyclostationarity as a method of co-channel detection system for unvoiced speech

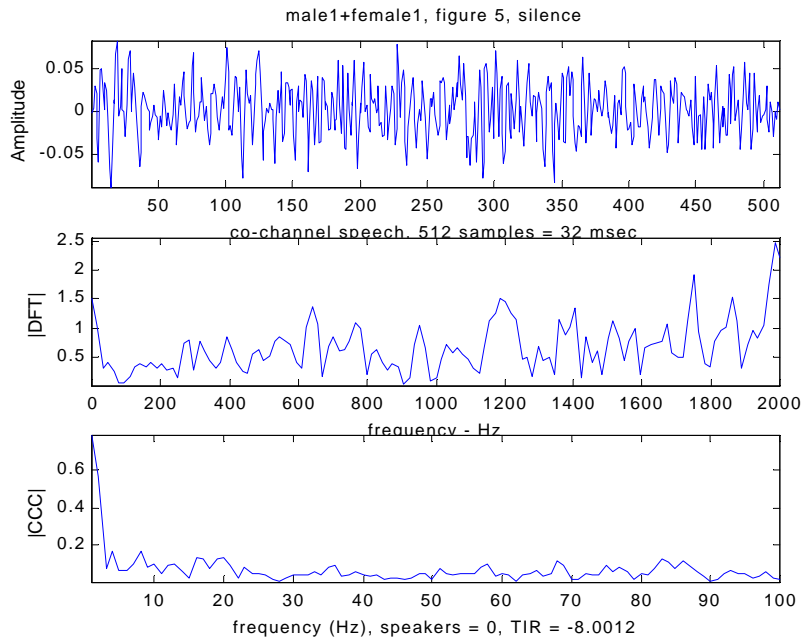


Figure 5.4 Cyclostationarity as a method of co-channel detection system for silence

Some other measures which iam planning to investigate as said before are wavelets, modulation maps, statistical shape analysis and higher order statistics.

5.2 WAVLETS

Wavelets as said before are finite bursts of energy in time domain and these are mathematical parameters, which divide data into different frequency components. A Wavelet transform is a convolution of signal $x(t)$ and a wavelet function $\Psi(t)$. The wavelet functions can either be linear phase wavelets (symmetrical or anti-symmetrical about any point), e.g. haar wavelets, spline wavelets and gaussian wavelets or wavelet functions can be minimum phase wavelets (neither symmetrical nor anti-symmetrical about any point). Abrupt changes in the signal level can be determined using the wavelet transform which makes it a possible measure for detecting co-channel speech since if there is a co-channel speech in a signal there will be a distinct change in the signal level.

5.3

MODULATION MAPS

Modulation maps are constructed by computing the modulation spectra (modulated power spectra) that give rise to a new spectrum, which is called a reassigned spectrum (shifted in frequency). The maps divide the spectrum into different frequency components and it codes the amplitude modulation frequency versus the channel frequency, thereby allowing two sounds to co-exist if they have different pitch periods. It has also been determined that if two vowels are separated by 6Hz modulation maps are

able to identify them distinctly. Since modulation maps can identify two different sounds separated by 6Hz it can be used for detecting co-channel speech.

5.4 STATISTICAL SHAPE ANALYSIS

Shape by definition is geometrical information, which is invariant. In statistical shape analysis a landmark (point of correspondence) is fixed and the distance of signal samples from the landmark is calculated. This distance measure is known to indicate the dissimilarity in the shape. First a landmark and shape of the voiced speech is fixed and the frames of the input speech signal is compared with the reference by computing the distance measure and if there is a co-channel speech coming in there will a distinct alteration in the distance measure by which we can detect co-channel speech.

5.5 HIGHER ORDER STATISTICS (HOS)

Higher order statistics such as skewness, kurtosis etc. have been used in detection of speech from background noise. An expression for 3rd order and 4th order cumulants is derived first and the properties of these cumulants have shown that HOS for speech noise is distinct and since co-channel speech is noise like though not noise this property can be exploited in detecting co-channel speech.

In the next few months iam going to investigate more about the possibility of wavelets, modulation maps, statistical shape and higher order statistics being a co-channel detection measure and probably focus on two of most promising ones to develop a detection measure based on them.

REFERENCES

- Ahmadi, S. and Spanias, A., "Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm", IEEE Transactions on Speech and Audio Processing, vol.7, no.3, pages: 333-338, May 1999.
- Bennincasa, D. S. and Savic, M. I., "Voicing State Determination Of Co-channel Speech", ICASSP, part2, vol.2, pages:1021-1024, May 12-15,1998.
- Cavallaro, A., Beritelli, F. and Casale, S., " A Fuzzy Logic-Based Speech Detection Algorithm For Communications In Noisy Environments", Istituti di Informatica e Telecomunicazioni- University of Catania, pages: 565-568, 1998.
- Chen, F. and Ser, W., "Speech Detection Using Microphone Array", IEE 2000, Electronic Letters, 15 November 1999.
- Dryden, I. L. and Mardia, K. V., "Statistical Shape Analysis". John Wiley & Sons. 1998.
- Einicke, G., "Adaptive Speech Estimation For Low SNR and Co-channel FM", Australian and New Zealand Conference on Intelligent Information Systems, vol.4, pages: 312-315, Nov 1996.
- Gary, E. and Marcia A., "LPC- Based Spectral Similarity Measure For Speech Recognition In The Presence Of Co-Channel Speech Interference", ICASSP, vol.1, pages: 270-273, May 23-26, 1989.
- Grant, W. and Seitz, F., "The Use Of Visible Speech Cues For Improving Auditory Detection Of Spoken Sentences", J. Acoustic Society of America.108 (3), pt.1, pages: 1197-1206, Sep 2000.

- Gresham, C. and Collins, M., "A Comparison Using Signal Detection Theory Of The Ability Of Two Computational Auditory Models To Predict Experimental Data", ICASSP, vol.2, pages: 933-936, 1999.
- Fite, D., Bruzzone, P. and Brian, G., "Blind Separation Of Voice Modulation Single-Band Using The Multi-Target Variable Modulus Algorithm", ICASSP, part 5, pages: 2726-2729, May7-10, 1996.
- Hamkins, J., "A Joint Viterbi Algorithm To Separate Co-Channel FM signals", ICASSP, vol.6, pages: 3297-3300, May1998.
- Kadambe, S. and Boudreaux-Bartels, G.F., "A Comparison Of A Wavelet Functions For Pitch Detection Of Speech Signals", ICASSP, vol.1, pages: 449-452, May 1991.
- Lewis, A. and Ramachandran, P., "On the Use Of Cepstral and Pitch Prediction Features For Speaker Count Labelling Of Co-channel Speech", ICSPAT Conference Proceedings, Toronto, Canada. Pages: 1020-1024, September 13-16,1998.
- Loizou, C., Dorman, M., Poroy, O. and Spahr, T., "Speech Recognition By Normal-Hearing And Cochlear Implant Listeners As A Function Of Intensity Resolution", J. Acoustic Soc. America.108 (5), pt.1, pages: 2377-2386, Nov2000.
- Morgan, P., George, E., Lee, T. and Kay, M., "Co-Channel Speaker Separation", ICASSP, part1, vol.1, pages: 828-831, May9-12, 1995.
- Naylor, J. and Porter, J., "An Effective Speech Separation System Which Requires No a Priori Information", ICASSP, pages: 937-940, May14-17, 1991.

- Quatieri, F. and Danisewicz, G., "Approach To Co-Channel Talker Interference Suppression Using A Sinusoidal Model For Speech", IEEE Transactions on Acoustics, Speech and Signal Processing, vol.38, pages: 56-59, Jan 1990.
- Sohn, J., Kim, S. and Sung, W., "A Statistical Model-Based Voice Activity Detection", IEEE Signal Processing Letters, vol.6, no.1, pages: 1-3, Jan 1999.
- Tucker R., "Voice Activity Detection Using A Periodicity Measure", IEE proceedings-I, vol.139, no.4, pages: 377-380, Aug 1992.
- Wenddt, S. J., "Personal Communications", 2000.
- Woo, H., Yang, Y., Park, J. and Lee, C., "Robust Voice Activity Detection Algorithm For Estimating Noise Spectrum", IEE 2000, Electronic Letters, 9 December 1999.
- Yantorno, R. E, "A Study Of Spectral Autocorrelation Peak Value Ratio (SAPVR) As Method For Identification Of Usable Speech And Detection Of Co-channel Speech", AFR Rome Labs Report, May 2000.
- Yen, K-C. and Zhao Y., "Adaptive Co-Channnel Speech Separation", IEEE Transactions On Speech and Audio Processing, vol.7, no.2, pages: 138-151, Mar 1999.
- Yen, K-C. and Zhao, Y., "Improvements On Co-Channel Speech Separation Using ADF: Low Complexity, Fast Convergence, and Generalization", ICASSP, part2, pages:1025-1028, May12-15, 1998.
- Yen, K-C. and Zhao, Y., "Co-Channel Speech Separation For Robust Automatic Speech Recognition, Stability and Efficiency", ICASSP, part2 (of 5), vol.2, pages: 859-862, Apr21-24,1997.
- Yoma, McInnes and Jack, "Robust Speech Pulse Detection Using Adaptive Noise Modeling", IEE 1996, Electronic Letters, 7 May 1996.

Zhang, B., "Harmonic Peaks Method For Voice Separation", International Conference On Signal Processing, ICSP, pages: 678-681, Oct12-Oct16, 1998.

Zissman, M. A., Weinstein, C. J. and Braid, L. D., " Automatic Talker Activity Labeling For Co-channel Talker Interference Suppression", ICASSP, part2, vol.2, pages: 813-816, Apr3-6, 1990.