

CO-CHANNEL SPEECH DETECTION APPROACHES USING CYCLOSTATIONARITY OR WAVELET TRANSFORM

Arvind Raman Kizhanatham, Nishant Chandra, Robert E. Yantorno

Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, PA 19122-6077, USA

akizhana@astro.temple.edu, cnishant@astro.temple.edu, robert.yantorno@temple.edu

http://www.temple.edu/speech_lab

Stanley J. Wenndt

Air Force Research Laboratory/IFEC, 32 Brooks Rd. Rome NY 13441-4514, USA

wenndts@rl.af.mil

ABSTRACT

Co-channel speech occurs when one speaker's speech is corrupted by another speaker's speech. A co-channel detection system could provide information to suspend the operation of any speech processing system whose operation would be degraded if it were processing co-channel speech. In this paper we present two new methods of co-channel speech detection, one based on cyclostationarity and the other based on wavelet transform. Detection of co-channel speech in this paper refers to the detection of co-channel voiced speech, as it is not yet possible to detect unvoiced co-channel speech. Cyclostationary-based co-channel speech detection reveals that at least 65% of co-channel speech is correctly detected for different combinations of speech, e.g., male-male, female-female etc., with false alarms of approximately 24%. Investigation of the wavelet transform based co-channel speech detection reveals that at least 94% of co-channel speech is correctly detected with false alarms of approximately 28% making both methods tools for detecting co-channel speech.

1. INTRODUCTION

Co-channel speech processing 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). Co-channel speech can occur in many 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 neighboring communication channels.

An application of a co-channel detection system is shown in Figure 1. The co-channel detection system would provide information to suspend the operation of

any speech processing system such as a speech recognition system, speaker identification system, etc., whose operation would be degraded if it were processing co-channel speech. Previously two approaches used for detection of co-channel speech have been presented; one was Spectral Autocorrelation Peak Valley Ratio (SAPVR) [1], and the other used cepstral and pitch prediction features to determine the number of speakers in co-channel speech [2]. In this paper we present two new methods for detecting co-channel speech.

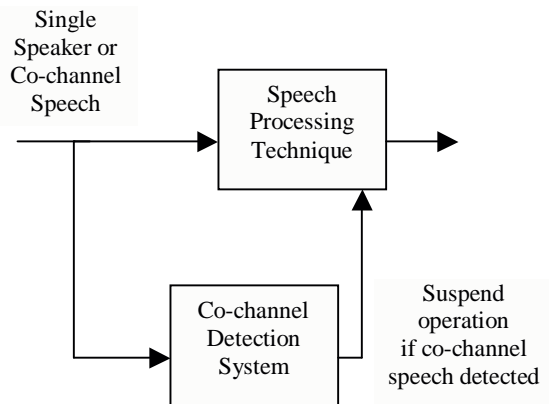


Figure 1. Block Diagram of Application of Co-channel Speech Detection System.

A random signal can be modeled as cyclostationary, if its statistical parameters vary in time with single or multiple periodicities. The property of cyclostationarity of a signal can be exploited to detect the presence of signals buried in noise and/or severely masked by interference. This property makes this cyclostationarity a good candidate for detection of co-channel speech.

A wavelet transform uses localized basis functions, and hence is capable of yielding good signal approximations with very few terms of the wavelet transform. Because wavelets are localized within an interval, resolution in

time can be traded for resolution in frequency, making it feasible to investigate a particular signal interval efficiently. A wavelet transform can be used to detect the abrupt changes in the amplitude of the speech signal; this property makes this approach a good candidate for detection of co-channel speech.

2. 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 time invariant. For most man-made signals, some parameters do vary periodically with time and in some cases harmonically unrelated periodicities are involved [3]. Investigations 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 [3]. This property could be exploited to perform various signal processing tasks, such as:

- 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.

3. PROCEDURE FOR DETECTING CO-CHANNEL SPEECH USING CYCLOSTATIONARITY

The procedure for using the cyclostationary approach to detect co-channel speech is described in Figure 2.

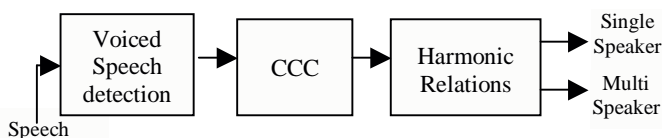


Figure 2. Block Diagram of a Co-channel Speech Detection System Using Cyclostationarity. CCC-Conjugated Cyclic Correlation

1. Spectral flatness (SF) of the speech was used to determine if the speech frame under consideration

is voiced [4]. A preset spectral flatness threshold (35dB) was found by performing a set of experiments. The spectral flatness of each frame of speech was compared against a preset threshold, and any frame above the threshold is determined to be voiced.

2. If the frame is determined to be voiced, a Hilbert transform is performed on the speech frame.
3. Then the convolution $h^*(n)*h(n+t)$ is performed on the Hilbert transformed signal.
4. The Fourier transform is performed on the output of convolution. From the output of the Fourier Transform three maxima are found. The time difference between the first and second maxima and the time difference between the second and third maxima is calculated. If both the differences are greater than a preset threshold (which was found to be 6 by performing a series of experiments and calculating the pitch difference), harmonic relations exist and therefore the frame is considered to be from a single speaker.

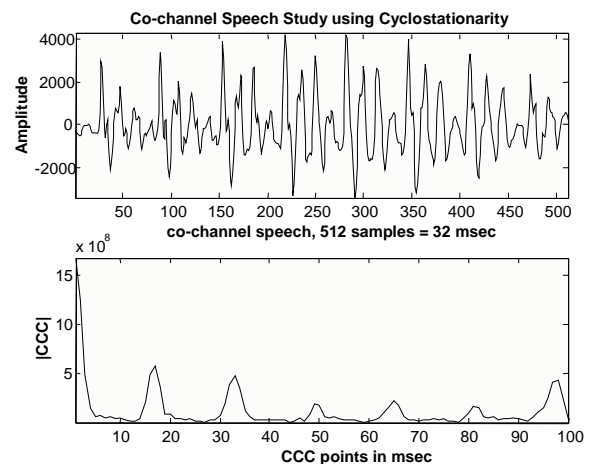


Figure 3. Harmonic Structure for Single Speaker Speech, Original Speech Frame (top panel), Conjugated Cyclic Correlation (bottom panel).

Figure 3 shows the original speaker speech frame in the top panel, conjugate cyclic correlation (CCC) of speech signal in the bottom panel. The time difference between the first and second maxima and the time difference between the second and third maxima in the bottom panel of Figure 3 is greater than the threshold (harmonic relations exist), hence it is a frame of single speaker.

Figure 4 shows the original speech frame in the top panel, conjugate cyclic correlation (CCC) of speech signal in the bottom panel. The time difference between the first and second maxima and the time difference between the second and third maxima in the bottom panel of Figure 3 is less than the threshold (no harmonic relations exist), hence it is a frame of co-channel speech.

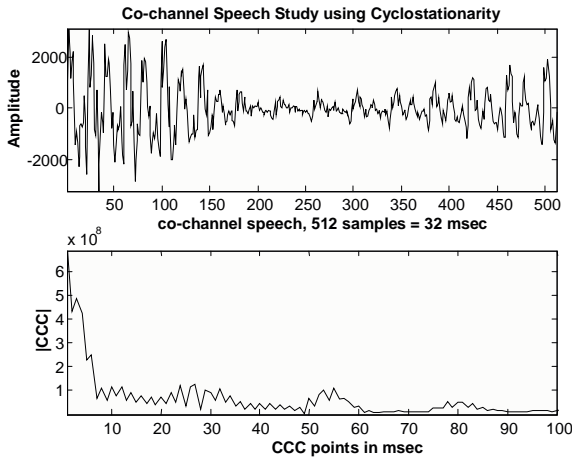


Figure 4. Lack of Harmonic Structure of Co-channel Speech, Original Speech Frame (top panel), Conjugated Cyclic Correlation (bottom panel).

4. WAVELET ANALYSIS OF SPEECH SIGNALS

The Fourier, sine and cosine transforms are non-local and hence there are limitations in time-frequency resolution of these transforms. Wavelets are small concentrated bursts of finite energy in time domain.

A wavelet transform can be used to detect the abrupt changes in the amplitude of the speech signal, e.g., for pitch detection, [5] [6], or for voiced/unvoiced detection [6] [7], and is also a good candidate for detection of co-channel speech.

If one observes a signal in a large window, gross features will be observed. However, small features are best observed by using small windows, which wavelet transforms can do. This allows the wavelets to reveal all the hidden features in the signal. This multi-resolution capability relies on being able to dilate (squeeze or/and expand) and translate the wavelet. Dyadic dilation (dilation by powers of 2) of the wavelet is the most popular of the wavelet functions and is also easy to implement [8] [9]. The wavelet prototype function

used for analysis is called the mother wavelet [10] [11] [12] [13]. This function is dilated and translated to achieve the basis function at different scales.

If $x(t)$ is the signal and $\Psi(t)$ is the wavelet function then a continuous wavelet transform (CWT) [CWT(b,a)] is a convolution of signal $x(t)$ and wavelet function $\Psi(t)$ expressed as:

$$CWTx(b, a) = \frac{1}{\sqrt{a}} \int x(t) \Psi^*[(t - b) / a] dt \quad (1)$$

where “a” is the dilation parameter and “b” is the translation parameter.

5. PROCEDURE FOR DETECTING CO-CHANNEL SPEECH USING WAVELETS

The procedure for using the wavelet approach to detect co-channel speech is described in Figure 5.

1. To determine the voiced speech, a discrete wavelet transform (DWT) is performed on the speech signal, on a frame-by-frame basis to obtain the approximate

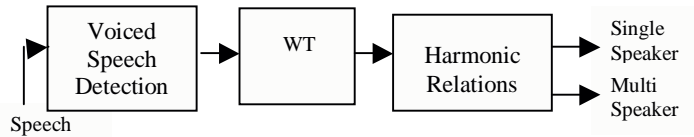


Figure 5. Block Diagram of a Co-channel Speech Detection System Using Wavelet Transform. WT-Wavelet Transform

and detail coefficients. The complex continuous coefficients are obtained by computing CWT. Both DWT and CWT are used in order to ensure the reliability of voiced speech detection.

2. If the length of the frame is N, then for voiced speech the output of DWT and CWT will have about 90% of the total energy in the first N/2 samples and only about 10% of the total energy in the N/2+1 to N samples.
3. For unvoiced speech, the output of DWT and CWT will not have 90% or more of the total energy for the first N/2 samples.
4. If a frame of speech is determined to be voiced, three maxima are found. The time difference between the first and second maxima and the time difference between the second and third maxima is calculated. If both the time differences are greater than a preset threshold (which was found to be 9 by performing series of experiments) harmonic relations exist and

therefore the frame is considered to be from a single speaker.

Figure 6 shows the original single speaker speech frame in the top panel, discrete wavelet transform (DWT) of speech signal in the middle panel and the harmonics of the discrete wavelet transformed speech signal in the bottom panel. Looking at the middle panel of Figure 5 we find that at least 90% of the energy of the output of the DWT is concentrated in first $N/2$ samples. Also, the time difference between the first and second maxima and the time difference between the second and third maxima in the bottom panel of Figure 6 is greater than the threshold (harmonic relations exist), hence it is a frame of single speaker speech.

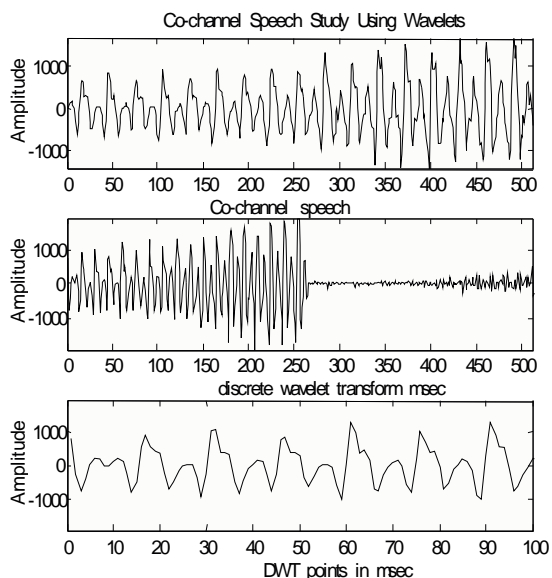


Figure 6. Harmonic Structure for a Single Speaker Speech, Original Speech Frame (top panel), Discrete Wavelet Transform (DWT) (middle panel), Lower Portion of the DWT of the Middle Panel Showing “Harmonic Relations” (bottom panel).

Looking at the middle panel of Figure 7 we find that at least 90% of energy of the output of the DWT is concentrated in first $N/2$ samples but the time difference between the first and second maxima and the time difference between the second and third maxima in the bottom panel of Figure 3 is less than the threshold (no harmonic relations exist), hence it is a frame of co-channel speech.

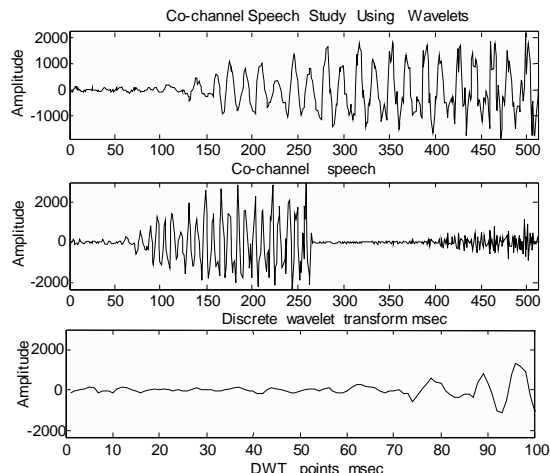


Figure 7. Harmonic Structure for a co-channel speech, original speech frame (top panel), discrete wavelet transform (DWT) (middle panel), lower portion of the DWT of the middle panel showing lack of “Harmonic Relations” (bottom panel).

6. EXPERIMENTS AND RESULTS

Ten speech signals (5 male, 5 female) were taken from the TIMIT database. For each experiment, speech signals from two different talkers were combined to form a composite speech signal having overall TIR (Target-to-Interferer Ratio) of 0 dB (equal energy). The speech signals were sampled at 16kHz and then down sampled to 8 kHz. The frame size was 32ms. Three different sets of experiments (male-male, female-female, male-female) were performed. Care was taken to ensure that the false alarm rate of the cyclostationary method was as close as possible to the false alarm rate of the wavelet transform approach so that comparison of both methods was possible.

The cyclostationary-based co-channel detection system uses the procedure described in Section 3 to detect the existence of co-channel speech. As described in section 3, spectral flatness is performed on the input speech signal to detect voiced speech. For female-female speech after ten experiments 59.9% of co-channel speech was detected correctly and 42.1% was missed, with false alarms of 23.1%. For male-male speech 67.8% of co-channel speech was detected correctly and 32.2% missed, with false alarms of 22.3%. Similar experiments were performed for female-male speech and the results are presented in Table 1.

Figure 8 shows the detection of co-channel speech based on cyclostationarity for female-female speech. The grey segments in Figure 8 are single speaker speech, the black

segments are the co-channel speech, and the rectangles are the detected co-channel speech segments using cyclostationarity.

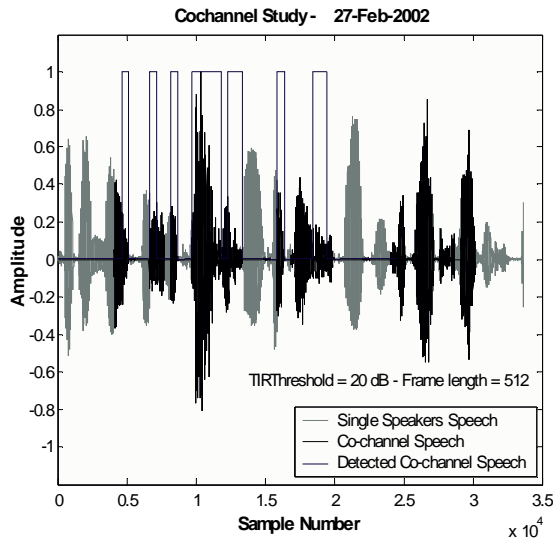


Figure 8. Detection of Co-channel Speech Using Cyclostationarity, Single Speaker’s Speech (gray), Co-channel Speech (black), Detected Co-channel Speech (rectangles).

The wavelet-based co-channel detection system uses the procedure described in Section 5 to detect the existence of co-channel speech. As described in section 5, DWT and CWT are performed on the input speech signal to detect voiced speech. Three maximas (peaks) are then found and the time differences between the maximas are compared to a preset threshold to detect the existence of co-channel speech. Figure 9 shows the detection of co-channel speech of female-female speech using wavelets. The grey segments in Figure 9 are single speaker speech, the black segments are the co-channel speech, and the rectangles are the detected co-channel speech segments using wavelet transform.

For female-female speech we determined, after ten experiments, that 95.0 % of co-channel speech was detected correctly and 5.0% were missed, with false alarms of 26.0%. For male-male speech, 93.6% was determined as being correct and 6.4% was missed, with 27.2% false alarms. Similar sets of experiments were performed for female-male speech and results are tabulated in Table 1 below.

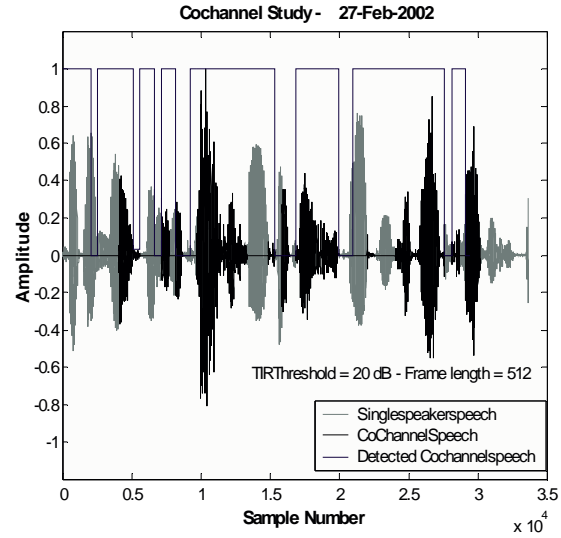


Figure 9. Detection of Co-channel Speech Using Wavelets, Single Speaker’s Speech (black), Co-channel Speech (gray), Detected Co-channel Speech (rectangles).

Table1: Results of Cyclostationary and Wavelets Based Co-channel Detection Systems.

Co-channel speech	% Correct		% False	
	Cyc	Wav	Cyc	Wav
Female-Female	59.9	95.0	23.1	26.0
Male-Male	67.8	93.6	22.3	27.2
Female-Male	66.7	94.1	26.1	29.6
Average	64.8	94.2	23.8	27.6

Cyc-cyclostationary, Wav- wavelets

As can be observed from Table 1, for female-female speech, wavelets-based co-channel detection system gives a higher percentage correct as compared with the cyclostationary-based co-channel detection system. Similar observations can be made for male-male speech and female-male speech from Table 1.

7. SUMMARY

In this paper we have presented two new methods of detecting co-channel speech, one based on cyclostationarity and the other based on wavelets. The results of our investigation of the cyclostationary and the wavelet methods of co-channel detection reveal that both systems can be used for detecting co-channel speech. The wavelet approach has an advantage over cyclostationary method in that it gives higher percent correct detection

when compared to cyclostationary, without a corresponding increase in false alarms.

8. FUTURE AREAS OF RESEARCH

The possibilities of fusing the cyclostationary method and the wavelets method to determine a better co-channel speech detection system could also be explored. Further, the properties of statistical shape analysis and modulation maps could be studied in detail to explore the possibility of using them as methods of detecting co-channel speech.

ACKNOWLEDGEMENT

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, under agreement number F30602-02-2-0501. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

DISCLAIMER

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

9. REFERENCES

- [1] Yantorno, R. E. "A Study of Spectral Auto-Correlation Peak Value Ratio (SAPVR) as Method for Identification of Usable Speech and Detection of Co-channel Speech", Intelligent Signal Processing Workshop, Hungary, pp: 193-197, May 2001.
- [2] 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, pp: 1020-1024, Sep13-16, 1998.
- [3] Gardner, "Exploitation of spectral redundancy in cyclostationary signals", Signal Processing, pp: 14-36, April 1991.
- [4] Johnston, J. "Transform Coding of Audio Signals Using Perceptual Noise Criteria" IEEE J. on Select Areas in Comm., vol. SAC-6, pp: 314-323, 1988.
- [5] Gonzalez, N. and Docampo, D. "Application of Singularity Detection with Wavelets for Pitch Estimation of Speech Signals", Proc. EUSIPCO, pp: 1657-1660, 1994.
- [6] Janer, L. "New Pitch Detection Algorithm Based on Wavelet Transform", IEEE-SP, pp: 165-168, 1998.
- [7] Nam, H., Kim, H. and Yang, S. "Speaker Verification Using Hybrid Model with Pitch Detection By Wavelets", Proc. IEEE ICASSP, pp: 153:156, 1998.
- [8] Kadambe, S. and Boudreaux-Bartels, G.F. "A Comparison of A Wavelet Functions for Pitch Detection of Speech Signals", ICASSP, vol.1, pp: 449-452, May 1991.
- [9] Johnson, I. A. "Discrete Wavelet Transform Techniques in Speech Processing", IEEE TENCON, pp: 514-519, 1996.
- [10] Davenport, M. R. and Garudadri, H. "A Neural Net Acoustic Phonetic Feature Extraction Based on Wavelets", IEEE- Computers and Signal Processing, pp: 449-452, 1991.
- [11] Daubechies, I. "Orthonormal Basis of Compactly Supported Wavelets", Comm. on Pure and Appl. Math. vol.41, pp: 909-996, Nov1988.
- [12] Kaisheng, Y. and Zhigang, C. "A Robust Feature-Perceptive Scalogram Based on Wavelet Analysis", ICSP, pp: 662-665, 1998.
- [13] Pinter, I. "Perceptual Wavelet-Representation of Speech Signals", Computer, Speech and Language, pp: 1-22, 1996.