

PEAK DIFFERENCE OF AUTOCORRELATION OF WAVELET TRANSFORM (PDAWT) ALGORITHM BASED USABLE SPEECH MEASURE

Arvind Raman Kizhanatham, Robert E. Yantorno, Brett Y. Smolenski
Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, PA 19122-6077, USA
akizhana@astro.temple.edu, robert.yantorno@temple.edu, bsmolens@temple.edu
http://www.temple.edu/speech_lab

ABSTRACT

Speech that is corrupted by interfering speech or nonstationary noise, but is still usable for applications such as speaker identification is referred to as “usable speech”. Recently, some usable speech extraction measures have been developed to separate a co-channel utterance into two different groups, those segments that are usable and those that are unusable. Portions of usable speech occur when high energy voiced speech from the target speaker overlaps with low-energy speech from an interfering speaker, or vice versa. A new method of usable speech detection based on the auto correlation of the wavelet transform is developed. Investigation of the method reveals that at least 80% of the usable speech is correctly detected with false alarms of 30%.

1. INTRODUCTION

Usable speech can be defined as degraded speech that is still usable for certain applications, such as speaker identification, speech recognition, and gisting. Because the concept of usability is context dependent, it is necessary to define usability based upon its intended application. For example, a speaker identification system can determine the identity of a particular speaker, relying on information that is sequence independent. In contrast, for speech recognition, it is necessary to have sequential speech information so that words and sentence structure can be properly identified [1].

In order to perform speaker identification, it is necessary to detect the portions of an entire co-channel speech utterance that contains maximum information about the target speaker. Such segments can then be defined as “usable” and sent on for further processing, while discarding segments containing co-channel speech as shown in Figure 1. It has been shown that when the target speaker is at least 20 dB greater than the interfering speaker, 80% reliable identification of the target speaker can be obtained [2]. Hence, these segments with a high Target-to-Interferer Ratio (TIR) may be considered usable with respect to speaker identification. Previously, Spectral Autocorrelation Peak Valley Ratio (SAPVR) [2], Kurtosis [3] and Adjacent Pitch Period Comparison (APPC) [4], modified SAPVR using LPC residual [5] were used for detection of usable speech. In this paper we present a new method of detection of usable speech based on autocorrelation of the wavelet transform.

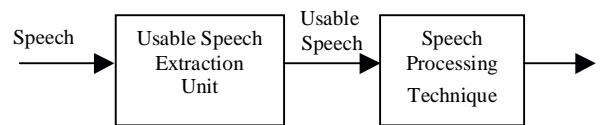


Figure 1. Block Diagram of Application of Usable Speech Extraction System.

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. 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 level, energy level of the speech signal, e.g., for pitch detection, [6] [7], or for voiced/unvoiced detection [7] [8].

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 and/or 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 [9] [10]. These properties makes the wavelet approach a good candidate for detection of usable speech as the amplitude levels, energy levels of usable speech and unusable or co-channel speech are different. The wavelet prototype function used for analysis is called the mother wavelet [11][12]. 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:

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

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

2. PROCEDURE FOR DETECTING USABLE SPEECH USING WAVELETS

The procedure for using the wavelet approach to detect usable speech is described in Figure 2. Voiced speech is first detected from the speech signal as it is not yet possible to determine the usable speech from unvoiced speech.

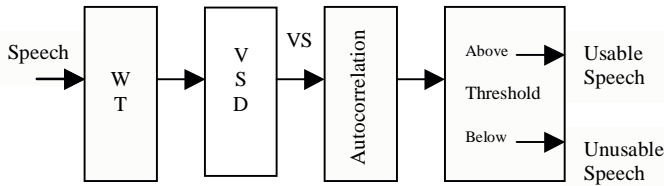


Figure 2. Block Diagram of a Usable Speech Detection System Using Wavelet Transform -VSD-Voiced Speech Detection, WT- Wavelet Transform, VS-Voiced Speech.

1. To determine voiced speech, a discrete wavelet transform (DWT) is performed on the speech signal, on a frame-by-frame basis, to obtain the approximate 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. For voiced speech the output of DWT and CWT will have about 90% of the total energy for the first $N/2$ samples (detail and approximate coefficients, N being 512) and only about 10% of the total energy in the remaining $N/2$ samples. Conversely, 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.
3. If a frame of speech is determined to be voiced, an autocorrelation is performed on the first half of the output of DWT on a frame-by-frame basis. Three maxima are then determined from this autocorrelation signal. For a usable speech frame the autocorrelation lag between the first and second maxima and the autocorrelation lag between the second and third maxima will be larger than the preset threshold. For an unusable speech frame the autocorrelation lags between the maxima will be less than the preset threshold (selection of threshold is discussed in Section 3).

Figure 3 shows a frame of single speaker (usable) speech in the top panel, discrete wavelet transform (DWT) of speech signal in the middle panel and the autocorrelation of the first half of the discrete wavelet transformed speech signal in the bottom panel. For a usable speech frame (Figure 3) the autocorrelation lag between the first and second maxima and the autocorrelation lag between the second and third maxima will be larger than the preset threshold of 15, the selection of which is discussed in Section 3. Also, autocorrelation of usable speech frame will show a very well defined structure (bottom panel of Figure 3).

Similarly, Figure 4 shows a frame of single speaker (unusable) speech in the top panel, discrete wavelet transform (DWT) of speech signal in the middle panel and the autocorrelation of the first half of the discrete wavelet transformed speech signal in the bottom panel. For an unusable speech frame (Figure 4), the autocorrelation lags between the maxima (bottom panels of Figure 4) will be less than the preset threshold of 15 (no harmonic relations exist). Autocorrelation of unusable speech frame will not show a well defined structure (bottom panel of Figure 4).

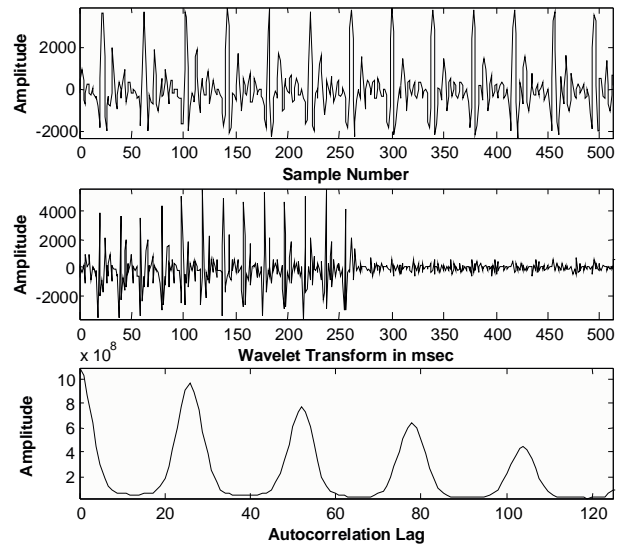


Figure 3. Harmonic Structure for a Frame of Single Speaker (Usable) Speech (top panel), Discrete Wavelet Transform (DWT) (middle panel), Autocorrelation of the DWT of the Middle Panel Showing “Harmonic Relations” (bottom panel).

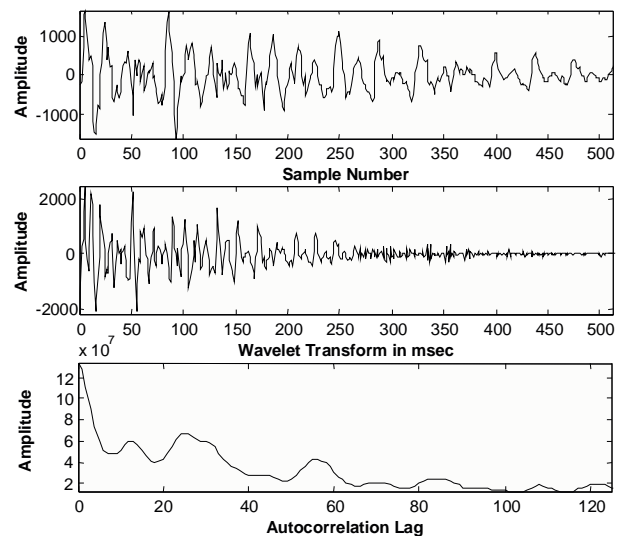


Figure 4. Harmonic Structure for Frame of Co-channel (Unusable) Speech (top panel), Discrete Wavelet Transform (DWT) (middle panel), Autocorrelation of the DWT of the Middle Panel Showing Lack of “Harmonic Relations” (bottom panel).

3. EXPERIMENTS AND RESULTS

Twenty speech signals (10 male, 10 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 an overall TIR (Target-to-Interferer Ratio) of 0 dB (equal energy). The speech signals were sampled at 16 kHz and then down sampled to 8 kHz. Three different sets of experiments (male-male, female-female, and male-female) were conducted and for each set of experiment 20 different speech files were used to identify the usable speech in the speech signal on a frame-by-frame basis.

SELECTING WAVELETS THRESHOLD

One of the most important aspects of a usable speech detection measure is selecting a proper threshold value. Only segments meeting the selected threshold criteria will be flagged as usable. Shown in Figure 5 is the probability study of the wavelets measure. Selecting a low threshold will ensure fewer false alarms; however, the trade-off is that fewer of the usable speech frames will be flagged as usable by the wavelets detection approach. Hence considering the tradeoffs associated, the threshold is selected as 15 (Figure 5) as there is a good compromise between percent correct detection and percent false alarms.

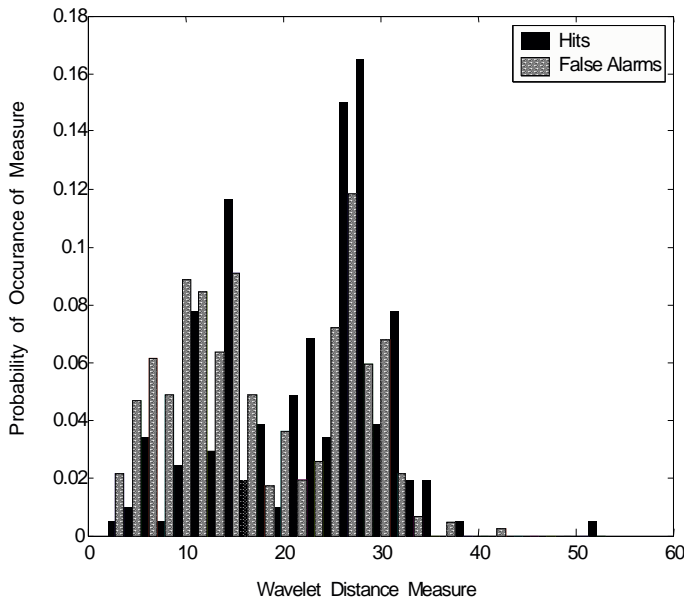


Figure 5. Probability of Hits and False Alarms for Wavelet Distance Measure

The wavelet-based usable detection system uses the procedure described in Section 2 to detect the existence of usable speech. As described in Section 2, DWT and CWT are performed on the input speech signal to detect voiced speech. Three maxima (peaks) are then found from the autocorrelated speech signal and the time differences between the maxima are compared to a preset threshold to detect the existence of usable speech. Figure 6 shows the wavelet transform usable speech detection

approach. The black segments in Figure 6 are (single speaker) usable speech, the grey segments are the unusable speech, and the rectangles are the detected usable speech segments using auto correlated wavelet transform.

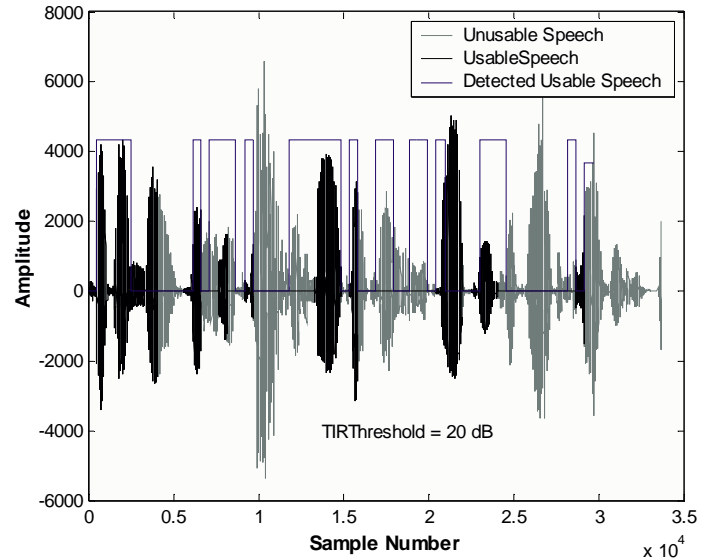


Figure 6. Detection of Usable Speech Using Wavelets, Usable Speech (black), Unusable Speech (gray), Detected Usable Speech (rectangles)

As shown in Table 1, for female-female speech it was determined, that 82.0 % of usable speech was detected correctly with false alarms of 32.0%. For male-male speech, 80.5% was determined as being correct with false alarms of 30%. For female-male speech, 81.3% was determined as being correct with false alarms of 29%. An interesting point to note is that wavelets based usable speech measure does not show much difference in percent correct detection for male speech and female speech.

Table 1: Results of Wavelets Based Usable Speech Measure.

Co-channel Speech	% Correct	% False
Female-Female	82.0	32.3
Male-Male	80.5	30.6
Female-Male	81.3	29.6
Average	81.2	30.8

4. SUMMARY

In this paper we have presented a new method of detecting usable speech based on wavelet transform. The performance of speaker identification systems can be improved by using the detected usable speech segments as most of the corrupted data

(unusable speech) has been removed. The goal of the wavelet transform measure is to extract the maximum amount of usable speech segments as possible with a minimum false alarm rate. On average the wavelet based usable speech measure detects at least 80% of the usable speech with a false alarm rate of 30%.

5. FUTURE AREAS OF RESEARCH

Further research on different wavelet functions such as haar wavelets, biorthogonal wavelets, gaussian wavelets etc., and even better understanding of the properties of wavelets could help in achieving increased correct detection rate and reduced false alarm rate. Modifying the PDAWT algorithm so that it finds the peak to valley ratio of the autocorrelated wavelet transform could result in a better measure. The possibilities of fusing the PDAWT based usable speech measure, SAPVR based usable speech measure, and APPC based usable speech measure to produce better results are to be explored [13][14].

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] Ma, K.W., Zavalagkos, G. and Meter, M., "Sub-sentence Discourse Models for Conversational Speech Recognition", IEEE Trans. On Acoustics, Speech and Signal Processing, Volume: 2, pp: 693-696, 1998.
- [2] Krishnamachari, K. R., Yantorno, R. E, Benincasa, D. S. and Wenndt, S. J., "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions", IEEE-ISPACS, pp: 710-713, Nov. 2000.
- [3] Krishnamachari, K. R, Yantorno, R. E., Lovekin, J. M., Benincasa, D. B., and Wenndt, S. J., "Use of Local Kurtosis Measure for Spotting Usable Speech Segments", Proc. IEEE-ICASSP, pp: 649-652, May 2001.
- [4] Lovekin, J. M., Yantorno, R. E., Krishnamachari, K. R., Benincasa, D. B and Wenndt, S. J. "Adjacent Pitch Period Comparison (APPC) as a Usability Measure of Speech Segments Under Co-channel Conditions", IEEE, ISPACS, pp: 139-142, Nov. 2001.
- [5] Lovekin, J. M., Yantorno, R. E., Krishnamachari, K. R., Benincasa, D. B and Wenndt, S. J., "Adjacent Pitch Period Comparison (APPC) as a Usability Measure of Speech Segments Under Co-channel Conditions", IEEE, ISPACS, pp: 139-142, Nov. 2001.
- [6] Gonzalez, N. and Docampo, D., "Application of Singularity Detection with Wavelets for Pitch Estimation of Speech Signals", Proc. EUSIPCO, pp: 1657-1660, 1994.
- [7] Janer, L. "New Pitch Detection Algorithm Based on Wavelet Transform", IEEE-Signal Processing, pp: 165-168, 1998.
- [8] Nam, H., Kim, H. and Yang, S., "Speaker Verification Using Hybrid Model with Pitch Detection By Wavelets", Proc. IEEE ICASSP, pp: 153:156, 1998.
- [9] Kadambe, S. and Boudreaux-Bartels, G. F., "A Comparison of A Wavelet Functions for Pitch Detection of Speech Signals", Proc. IEEE ICASSP, vol.1, pp: 449-452, May 1991.
- [10] Johnson, I. A., "Discrete Wavelet Transform Techniques in Speech Processing", IEEE TENCON, pp: 514-519, 1996.
- [11] 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.
- [12] Daubechies, I. "Orthonormal Basis of Compactly Supported Wavelets", Comm. on Pure and Appl. Math. vol.41, pp: 909-996, Nov. 1988.
- [13] Smolenski, B. Y., Yantorno, R. E. and Wenndt, S. J., "Fusion of Co-channel Speech Measures Using Independent Components and Nonlinear Estimation", IEEE, ISPACS, Nov 2002.
- [14] Smolenski, B. Y., and Yantorno, R. E., "Fusion of Usable Speech Measures Using Quadratic Discriminant Analysis", Proc. IEEE, ICASSP, April 2003 (Submitted)