Unsupervised Detection of Whispered Speech in the Presence of Normal Phonation

Michael A. Carlin [1], Brett Y. Smolenski [2], Stanley J. Wenndt [3]

[1]Speech Processing Laboratory, Temple University, Philadelphia, PA 19122 USA
[2]Research Associates for Defense Conversion, Marcy, NY 13403 USA
[3]Air Force Research Laboratory/IFEC, Rome, NY 13441 USA

mcarlin@temple.edu, {Brett.Smolenski, Stanley.Wenndt}@rl.af.mil

10.21437/Interspeech.2006-240

Abstract

The results of an investigation into unsupervised detection of whispered speech segments in the presence of normally phonated speech are presented. The Whispered Speech Detection system presented here extracts features which exploit both waveform energy and periodicity. Unsupervised classification of these features was performed to identify and label long segments (approx. 2 - 2.5 seconds) of whispered speech which is typically an indication of criminal activity over telephone networks, for instance, in a correctional facility environment. Experiments indicate that it is possible to automatically detect long segments of whispering in the presence of normally phonated speech; testing of the algorithm presented in this paper yields promising results in correct identification of whispered speech segments.

Index Terms: whispering, automatic whisper detection

1. Introduction

It is of interest to detect long segments (approx. 2 - 2.5 seconds) of whispered speech in the presence of normal phonation. For example, in a correctional facility telephone network, whispering in between sections of normally phonated speech is often an indication of criminal activity. It is labor-intensive and often difficult to monitor large numbers of individual phone calls for criminal activity. As such, this necessitates a system to automatically monitor telephone network communications to detect various types of criminal activity. This research has focused on automatically identifying whispered speech segments.

Whispered speech is generated when the glottis is only partially opened and a turbulent, aspirated flow of air is articulated by the vocal tract. The observed waveform is highly noise-like with spectral characteristics depending on the size of the glottis and the shape of the vocal tract. In studying these spectra, it has been observed that normally phonated and whispered speech exhibit differences in formant characteristics; specifically, shifts in the first formant frequency F_1, have been observed in whispered vowels. Formants F_2-F_4 are vowel-dependent and exhibit no consistent trends which can be exploited for reliably separating normally phonated and whispered speech sounds. Additionally, while significant F_1 shifts were found in Serbian vowels [1], when English vowels were observed [2] there were no consistent shifts in F_1 frequency and amplitude. It was also found that these shifts can be masked by the shifts caused by different speakers, conversation content, and widely varying amplitude levels between speakers and/or audio sources. Further, it has also been observed that the spectral tilt of whispered speech is less sloped than for normal speech [3].

A modified Speaker Identification (SID) approach has also been used in an attempt to classify unknown segments of speech as either whispered or normally phonated [4]. Essentially, multilingual, gender-independent models were built which represented whispered and normally phonated speech. An unknown speech segment was then presented to the SID system and the closest match declared as the winning "speaker" model. However, this procedure required training data which may not be readily available or easily gathered in a prison environment.

The approach in this research was to detect and identify whispered speech segments in speech data known *a priori* to contain both whispering and normal phonation using unsupervised classification techniques. It was assumed that whispering is generated from a noisy, highly aperiodic source with a wideband energy distribution. Based on this assumption, energy and periodicity features were extracted on a frame-by-frame basis, post-processed using nonlinear smoothing techniques, and examined by a data clustering algorithm. Contiguous segments of speech labeled as whispering of sufficient length (user-defined, typically 2 - 2.5 seconds) would flag the Whispered Speech Detection (WSD) system and segments containing data classed as whispering returned to the user.

The paper is organized as follows: Section 2 describes the development and details of the the WSD algorithm; Section 3 describes the test file layout and experimental setup for evaluating the WSD algorithm; Section 4 describes the performance of the WSD algorithm; Section 5 provides a summary of the research.

2. Algorithm Development

The WSD algorithm shown in Figure 1 was composed of four primary sections: (1) feature extraction, (2) post-processing of features, (3) cluster analysis of features, and (4) voicing state classification as either normally phonated or whispered speech.

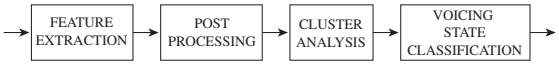

Figure 1: *Unsupervised WSD algorithm.*

Features which tracked energy content and periodicity were investigated per extracted data frame: (1) the Energy Ratio, and (2) Modified LPC Residual Autocorrelation Coefficients. After appro-

priate smoothing, these feature data were then made available to a clustering algorithm and the clusters appropriately labeled. The WSD system indicates if a user-defined length of data labeled as whispered speech was observed.

2.1. Feature Extraction

2.1.1. Energy Ratio

Shown in Figure 2 is a test waveform composed of three normally phonated segments interspersed with three whispered (W) sections. A window of approximately 2 seconds in length and overlapped 50% was used. The Energy Ratio (ER) was defined as the ratio of energy above 2.5 kHz (E_{HB}) to energy below 1 kHz (E_{LB}):

$$ER = \frac{E_{HB}}{E_{LB}} \quad (1)$$

The motivation that this feature should detect whispering was based on (1) most of the energy of normally phonated speech is below 1 kHz and (2) the source for whispered speech is a wideband glottal noise excitation. Thus, if the majority of the information in the data window was noisy, i.e., whispered, then the ratio would be larger.

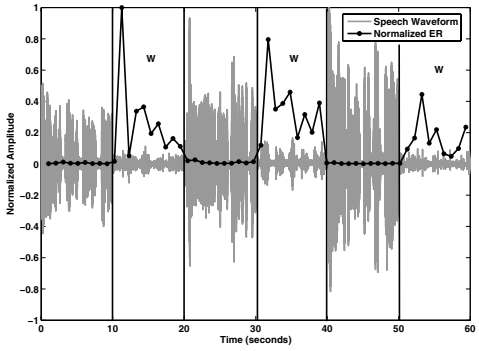

Figure 2: *Original waveform with normally phonated and whispered (W) sections and normalized energy ratio.*

Notice in Figure 2 above that the feature value increases as expected during sections of whispering. However, in an effort to bolster the performance of the WSD system, it was desired to develop another feature which exploits the inherent lack of periodicity present in whispering; this is discussed next.

2.1.2. Modified LPC Residual Autocorrelation Coefficients

Figure 3 below illustrates the process to extract the Modified LPC Residual Autocorrelation Coefficients (MLPCs).

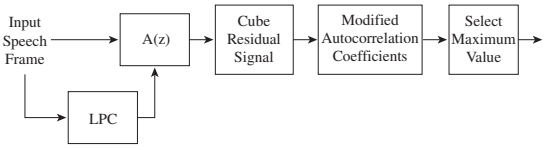

Figure 3: *Algorithm for extracting the MLPCs.*

For a given 32 ms input frame of speech data extracted using a rectangular window, the 14th-order linear prediction (LP) coefficients were calculated. The input segments were then inverse-filtered using an all-pole model from the LP coefficients. The residual signal was then cubed to accentuate peaks and attenuate small, noisy data. The residual signals are shown in Figure 4 for voiced and whispered segments. The main idea was to isolate the approximated glottal source, accentuate the peaks and develop a measure of the *periodicity* of the input frame, but not necessarily measure the specific pitch of the data [5].

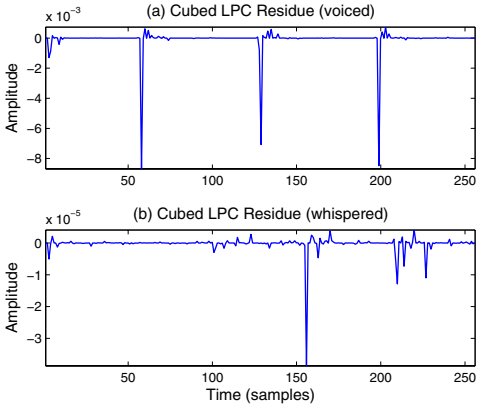

Figure 4: *Cubed LPC Residues for (a) voiced and (b) whispered segments. After cubing the residue, the voiced residual exhibited a distinct periodic structure useful for estimating the amount of periodicity in the waveform.*

Modified autocorrelation, as described in [6], was then performed on the residual signal by computing Pearson's linear correlation coefficients between the first half of the input data frame and the rest of the data in the frame. Next, the maximum value from these correlation coefficients per frame was selected between lag 30 and lag 90. On average, it was observed that for a voiced residual signal a significant peak in the modified autocorrelation would arise closer to the center of the given frame, i.e., the first half of a periodic frame should be highly correlated with itself about halfway through. This is shown for a whispered and normally phonated test segment in Figure 5.

Shown in Figure 6 is the output of the MLPCs for an input test waveform. Note the general decrease in the value of the MLPCs corresponding to sections of whispering. Further, it was assumed that distinct clusters would form to distinguish between whispering and normal phonation following some post-processing of the feature data. The next section describes such techniques.

2.2. Post-Processing

2.2.1. Nonlinear Ranked-Order Statistics Filtering

Following raw computation, the features were post-processed first using a nonlinear ranked-order statistics filter. Using a length $Nw = 199$ sliding window, the window elements were increasingly ordered and the $r = (Nw-1)$ statistic selected as the output of the filter. The motivation for selecting $r = (Nw - 1)$ was to accentuate the transitions in the average values between normally phonated and whispered sections, as observed in Figure 7(b) and (c). Likewise, the technique performed well as a smoothing filter

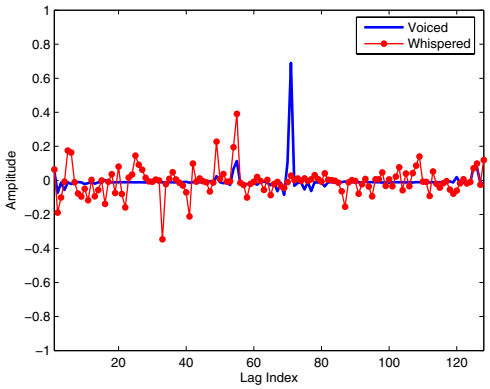

Figure 5: *Modified autocorrelation obtained from preprocessed LPC residues of voiced and whispered segments of speech. The output corresponding to the voiced segment exhibits a flat structure with a distinct peak after about 70 lags while the whispered segment results in a noisy curve and lower maximum correlation coefficient.*

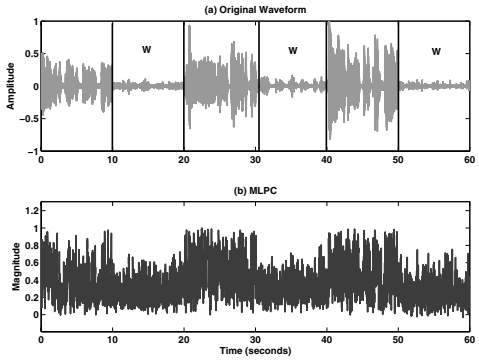

Figure 6: *(a) Original waveform, and (b) MLPC feature output.*

for the raw features as it tracked the envelope changes corresponding to whispering and normal phonation.

2.2.2. *Normalization and Scaling*

The features were normalized to maintain the dynamics of the feature curve, as this was the indication of whispering. Additionally, it was desired to squash any outlier data points such that more accurate cluster analysis could be performed. The so-called *softmax* scaling function [7] seemed to be optimal. Essentially, all outlier points sufficiently far from the mean of the distribution of the feature were squashed exponentially and the resulting values of the feature vector limited to the range (0,1].

2.3. Cluster Analysis

In a correctional facility environment, for example, telephone data is sometimes severely degraded on one or both ends of the conversation. Therefore, rather than simply setting a threshold to make a voicing state decision (as was done in [4]), implementing a classification technique which can adapt to varying channel conditions was desired.

Shown in Figure 7 are the smoothed feature data used for clustering. Note how the feature contours correspond to distinctions between normal phonation and whispering. Specifically, the MLPC feature contour falls during sections of whispering and rises during sections of normal phonation. Conversely, the ER feature contour rises during sections of whispering and falls during sections of normal phonation.

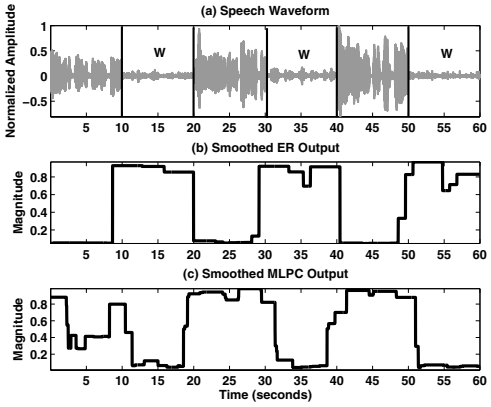

Figure 7: *(a) Original speech waveform, (b) smoothed ER output, and (c) smoothed MLPC output.*

Given varying channel conditions, it is possible that valid clusters which have formed could translate about the feature space while still maintaining a cluster structure. This further substantiates using a clustering approach rather than setting an arbitrary threshold. Clustering of the features was performed using the k-means clustering algorithm. Once determined, the clusters were labeled appropriately as either "normally phonated" and "whispered." Knowing that whispered sections were indicated by larger ER and smaller MLPC values and vice versa for normally phonated sections, the corresponding centroid values were examined and used to label the whispered sections appropriately. The output of the clustering algorithm for this particular example is shown in Figure 8. Note that the clusters labeled as whispering (cluster output value = 0.5) correspond to the sections of whispering labeled in the test file.

2.4. Voicing State Classification

The final stage of the WSD algorithm involves examining the length of contiguous segments of speech labeled as "whispering" ('W'). Note that an unvoiced frame of speech is highly aperiodic and would therefore be classified as 'W' by the algorithm. However, since it was desired to detect long segments of aperiodic data, when a contiguous set of 'W' frames exceeding 2.5 seconds were observed, the WSD system indicated having detected a segment of whispering.

3. Experimental Setup

The data used for testing the WSD system was generated from an Air Force database of male and female speakers with normally phonated and whispered utterances as separate files. Artificial con-

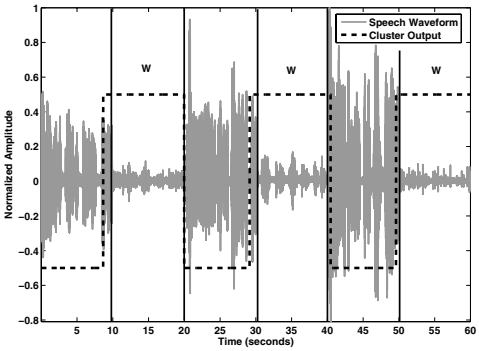

Figure 8: *Original speech waveform and clustering algorithm output. Note that frames labeled as whispered have an output value = 0.5 for clarity.*

versations were then generated from these utterances. The original waveforms were sampled at 8 kHz with 16 bits of resolution. The waveforms were bandlimited to 2800 Hz on average. Multiple phonation scenarios were then considered. For example, male normally phonated segments (M-NP) were concatenated with female whispered (F-W) segments and vice versa. No single speaker was used more than once in a test file. Table 1 shows each of the test files F1-F15 and their respective layouts. Test file F11 was created by concatenating F1 with F2, test file F12 created by concatenating F3 with F4, and so on through F15. Test files F1-F10 were one minute in length, F11-F15 were therefore 2 minutes in length. In all, there were a total of 41 possible whispered segments to be detected. Further, as the system was unsupervised, no training data was needed.

Table 1: Test file layouts.

File	Layout Detail					
F1	M-NP	F-W	M-NP	F-W	M-NP	F-W
F2	M-W	F-NP	M-W	F-NP	M-W	F-NP
F3	F-NP	M-W	F-NP	M-W	F-NP	M-W
F4	F-W	M-NP	F-W	M-NP	F-W	M-NP
F5	M-NP	M-W	M-NP	F-W	F-NP	F-W
F6	F-W	F-NP	F-W	M-NP	M-W	M-NP
F7	M-NP	M-NP	M-NP	F-W	F-W	F-W
F8	F-NP	F-NP	F-NP	M-W	M-W	M-W
F9	M-W	M-W	M-W	F-NP	F-NP	F-NP
F10	F-W	F-W	F-W	M-NP	M-NP	M-NP
F11	F1-F2					
F12	F3-F4					
F13	F5-F6					
F14	F7-F8					
F15	F9-F10					

4. Testing & Results

The performance of the WSD algorithm is shown in Table 2. As each test file was presented to the algorithm, detection of a whispered segment of speech was counted if a contiguous set of whisper-labeled frames longer than 2.5 seconds was observed. As seen in Table 2, the WSD system detected 40 out of a possible 41 whispered segments - a correct system performance of 97.5%.

Table 2: Results of unsupervised whisper detection using the WSD algorithm.

File	Known Whispered Sections	Correctly Detected	File	Known Whispered Sections	Correctly Detected
F1	3	3	F9	1	1
F2	3	3	F10	1	1
F3	3	3	F11	5	4
F4	3	3	F12	5	5
F5	3	3	F13	5	5
F6	3	3	F14	2	2
F7	1	1	F15	2	2
F8	1	1	—	—	—
			% Correct Detected		*97.5%*

5. Summary

A system to automatically detect segments of whispered speech in the presence of normal phonation has been investigated. The system exploits the aperiodicity and wideband energy distribution of whispered speech to perform unsupervised classification of whispered speech segments. The WSD system described has been able to accurately detect whispering with 97.5% correct detection. This is useful as a possible indication of criminal activity in a correctional facility telephone network since whispering would most likely occur in between sections of normally phonated speech.

6. References

[1] Jovicic, S. T. "Formant feature difference between whispered and voice sustained vowels." Acoustica, 84:739-743, 1998.

[2] Wilson, J. B. "A comparative analysis of whispered and normally phonated speech using an LPC-10 vocoder." RADC Final Report TR-85-264, Air Force Research Laboratory, Rome, NY, December 1985.

[3] Itoh, T., Takeda, K., and Itakura, F. "Acoustic analysis and recognition of whispered speech." In International Conference on Acoustic, Speech, and Signal Processing, volume 1, pages I-389-92, 2002.

[4] Wenndt, S. J., Cupples, E. J., and Floyd, R. M. "A study on the classification of whispered and normally phonated speech." In International Conference on Spoken Language Processing, Denver, CO, 2002.

[5] Smolenski, B. Y. *A Filterless Approach to Speaker Identification in the Presence of Non-Stationary Interference*. PhD Thesis, Temple University, May 2005.

[6] Rabiner, L. R., and Schaefer, R. W. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[7] Theodoridis, S. and Koutroumbas, K. *Pattern Recognition*. Academic Press, 2003.