

Analysis and Classification of Speech Mode: Whispered through Shouted¹

Chi Zhang and John H.L. Hansen

Center of Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas 75083, USA
{chi.zhang, john.hansen}@utdallas.edu <http://crss.utdallas.edu>

Abstract

Variation in vocal effort represents one of the most challenging problems in maintaining speech system performance for coding, speech and speaker recognition. Changes in vocal effort (or mode) result in a fundamental change in speech production which is not simply a change in volume. This is the first study to collectively consider the five speech modes: whispered, soft, neutral, loud and shouted. After corpus development, analysis is performed for i) sound intensity level, ii) duration and silence percentage, iii) frame energy distribution and iv) spectral tilt. The analysis shows vocal effort dependent traits which are used to investigate speaker recognition. Matched vocal mode conditions result in a closed-set speaker ID rate of 97.62%, with mismatch vocal conditions producing 54.02%. Finally, a speech mode classification system is developed, which has a range of classification rate from 44.5% to 98.5% confusing with adjacent vocal modes. These advancements can provide improved speech/speaker modeling information, as well as classified vocal mode knowledge to improve speech and language technology in real scenarios.

Index Terms: Whispered, vocal effort, Speaker recognition, classification

1. Introduction

The development of speech and language technology has resulted in the deployment of automatic speech systems for coding, speech and speaker recognition in real environments. These systems are designed to work well in normally/neutrally phonated speech conditions. Vast amounts of research has concentrated on the characterization of normally phonated speech for speech recognition, speaker identification and language identification performance, but limited research has been devoted to characterizing and recognizing speech which is not normally phonated speech. This fact drives us to explore the characteristics of speech in different phonated modes related to vocal efforts and how speech mode can affect performance of automatic speech systems.

Speech signals can be generally classified into five categories based on the modes of speech production: whispered speech, soft speech, neutral speech, loud speech, and shouted speech. The whispered speech is defined as the lowest vocal mode of speech with limited vocal cord vibration. On the other hand, shouted speech is referred to as the highest vocal mode of speech, which requires the most dramatic change in vocal excitation. Normally phonated speech is called neutral speech. Soft speech is viewed as a

moderate mode between whispered and neutral speech, with loud speech mode viewed as the mode between neutral and shouted speech mode.

The difference in speech production mechanism causes differences in characteristics of the five speech modes. For example, the differences in formant characteristics between whispered speech and neutral speech has been observed by Jovicic in [4] and Wilson in [5]. The differences in production characteristics typically impact performance for automatic speech systems in different ways. For example, for a speech recording system with Automatic Gain Control (AGC), the shouted speech, possibly the rollercoaster riders' scream [2], will overwhelm the AGC and cause clipping in the speech waveform. Recording of soft or whispered speech may result in an acoustic waveform that does not allow the listener to distinguish the text content of what was spoken. If the characteristics of the speech in different modes are known, a pre-processing phase can be applied to classify the speech mode and compensate the negative impact from the speech in non-neutral modes. Furthermore, with the knowledge of characteristics of speech in different modes, an on-line process to identify the speech modes would be a valuable pre-processing step to any automatic speech systems [10].

Some prior studies have analyzed properties of different speech modes. In [2], the characteristics of neutral speech have been analyzed and compared with those of soft, loud and cognitive/physical stressed speech. In [6] [7], the acoustic features and phonetic structure of shouted speech has been presented. However, no research has focused on the characteristic analysis of whispered through shouted speech.

In this paper, we report on the analysis of characteristics of the speech in five different vocal modes: whispered, soft, neutral, loud and shouted; and to identify discriminating features of speech modes. The influence of speech mode on closed-set speaker recognition is also considered. Finally, a classification system for vocal efforts is developed to identify the speech mode. For this study we collected a corpus of speech data in five modes, carefully measuring Sound Pressure Level (SPL) throughout the corpus development.

The paper is organized as follows, Sec. 2 describes corpus developed of five speech modes. The next section details the various analyses performed on the corpus. In Sec. 4 experiments and results on speaker recognition are presented. Sec. 5 describes a proposed vocal effort classification technique and its performance. We close in Sec. 6 with some concluding remarks.

¹ This work was funded by RADC under contract A40104, and by University of Texas at Dallas under Project EMMITT.

2. Corpus Development

All speech data used for this study are recorded in an ASHA certified, single walled sound booth using a multi-track FOSTEX 8-channel synchronized digital recorder with gain adjustments for individual channels. A total of 12 male, native English-spoken subjects participated in the data collection. For each subject, the speech was recorded in several tokens using three microphones- a throat microphone, a close-talking and a far field microphone. A 1 kHz sinusoid signal generated by an NTI analog audio generator was played by an ALTEC speaker as the calibration test tone in all recordings. At the beginning of each token, the volume of the test tone was carefully adjusted to make the SPL of the test tone measure 75 dB using a QUEST sound level meter. The test tone was recorded with all three microphones. The position of the subject, the location of the calibration test tone speaker, and the location of the sound level meter were all positioned in an equidistant triangle separated by 75cm.

The data collection procedure was divided into 3 phases for each subject. Phase I consists of 2 sessions which has 5 tokens corresponding to the five speech modes. In each token, 5 sentences from the TIMIT database were read in one of five speech modes and recorded. Phase II consists of 20 sentences which were read in the neutral mode. Phase III includes the spontaneous speech of one-minute duration in each vocal mode.

3. Analysis of speech in five modes

Speech data in Phase I from the 12 subjects, produced with the close-talking microphone was used in the following analyses and experiments. The speech data was analyzed in terms of: i) sound intensity level, ii) sentence duration and silence duration, iii) frame energy distribution, and iv) spectral tilt. The results of the study are discussed below.

3.1. Sound intensity level

Sound intensity is the power of the sound transmitted along the wave [8]. Sound Intensity Level (SIL) is an important measurement describing the sound quality in decibel level. In this study, carefully measured and recorded 75 dB-SPL test tones were used to calculate the statistic properties of the sound intensity level of the speech in five modes. The following figure shows the mean and standard deviation of the sound intensity level in the five speech vocal modes.

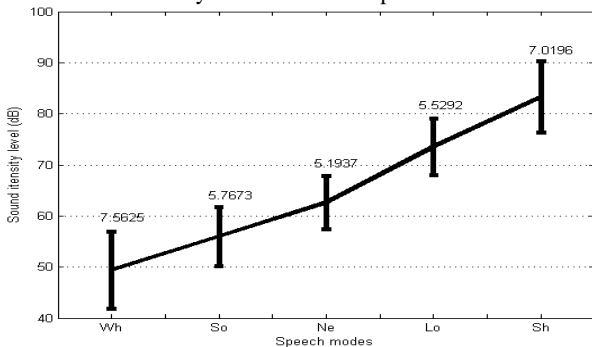


Figure 1: Mean and standard deviation of the sound intensity level of sentences under five speech modes. [Wh: whispered, So: soft, Ne: neutral, Lo: loud, Sh: shouted].

In Fig. 1, the increasing SIL shows the speech intensity increasing as the speech mode changes from whispered to shouted. The standard deviations of the SIL in five speech

modes indicate that the variation of the SIL in neutral mode is lower than that in the other four speech modes. As long as the speech mode shifts from neutral mode to shouted mode or to whispered mode, the variation of the SIL increases. Since the SIL of five speech modes are discriminative, the speech intensity can be viewed as a representative trait of the speech mode. We note that in many speech applications, it may not be possible to know the exact dB-SPL level since recording levels and an AGC may be used to improve A/D conversion. In spite of this, discriminating ratios such as consonant-to-vowel amplitude ratio (CVAR), consonant-to-semivowel amplitude ratio (CSVAR), and vowel to semivowel amplitude ratio (VSVAR) have been shown to be efficient for soft, neutral and loud speech [2].

3.2. Sentence duration and silence percentage

The duration of the 10 sentences in five speech modes was analyzed in this study. The duration of each sentence in whispered, soft, loud and shouted modes was normalized by the corresponding duration under neutral speech mode to remove the effect of sentence length, which varied within the chosen set of TIMIT sentences. It was found that on average, sentence duration increases in all four speech modes relative to the neutral mode. However the increases of duration in soft and whispered modes were mainly caused by silence duration increases, while the increases of sentence duration in loud and shouted speech modes are mainly caused by increases in word duration.

To clarify this, a frame-energy based approach was used to extract the speech frames from the sentences and to identify the silence frames between speech frames. Frames of 20ms in duration with a 10ms overlap were used. The frames above a certain threshold were selected as speech frames, and those between speech frames and lower than the threshold were identified as silence frames. From the speech and silence frame counts, the sentence duration and percentage of silence duration were computed. For each speech mode, the mean and variance of normalized sentence duration, and the percentage of silence frames are illustrated in Figure 2.

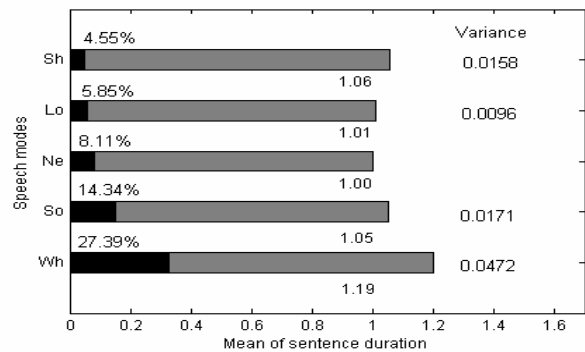


Figure 2: Mean and variance of the normalized sentence duration and silence percentage.

Fig. 2 shows overall sentence duration, normalized to neutral (Ne) as 1.00, along with the sentence duration variance (listed on the right). The mean percentage change in duration from neutral varied from 1% to 19%, whispered having the greatest duration changes. Whispered also had the largest variance in sentence duration. Also shown in Fig. 2 is the percentage of frames within the sentences which are devoted to silence frames. We see for loud and shouted, the silence frame percentage decreases, while for soft and whispered there are significant increases, especially for

whispered. This illustrates that under vocal effort changes, sentence duration and silence content are discriminating features.

3.3. Frame energy distribution

Fig. 3(a, b, c, d, e) show the histograms of frame energy distributions from the sentences for 12 subjects in five speech modes. It is evident that, a fundamental change in the speech energy distribution result. The amount of low energy frames, which can be viewed as silence and fricative frames, in whispered speech is much more than the other four modes. The energy spread change and the shift of the peak position in the histogram in each speech mode indicates that as long as the intensity level of speech increases, the frame energy increases, and the amount of high energy frames, which may represent voiced frames increases as well.

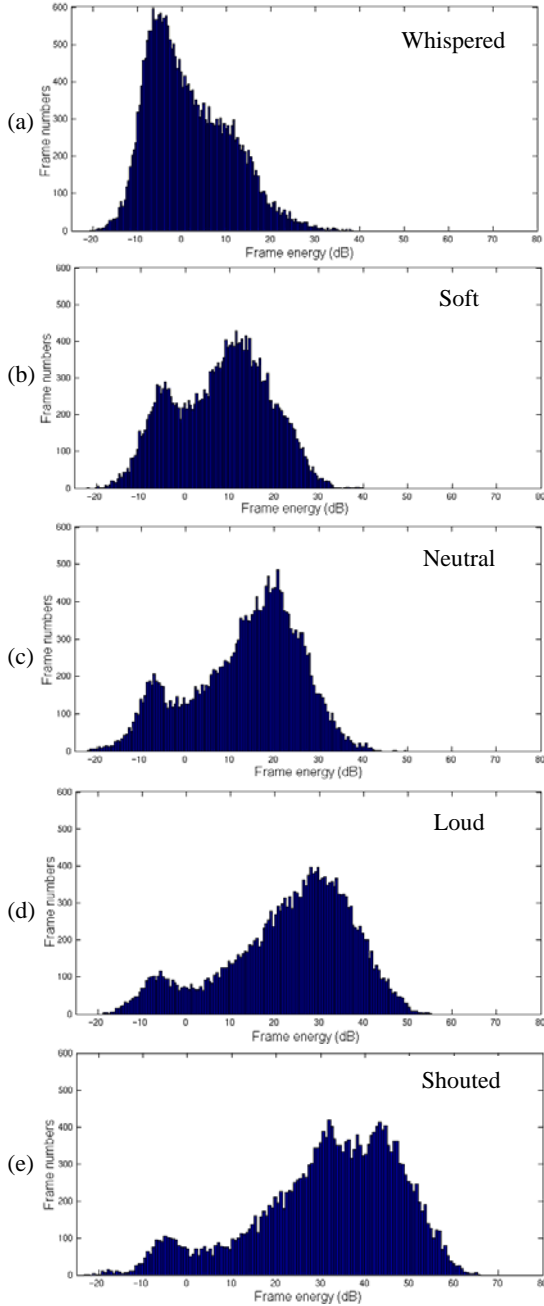


Figure 3: *Frame energy distribution for speech under (a) whispered, (b) soft, (c) neutral, (d) loud, and (e) shouted vocal modes.*

3.4. Spectral tilt

Next, variations in spectral tilt were investigated for the 12 subjects. The approach for estimating the spectral tilt can be found detailed in [2]. According to the frame energy distribution, a certain threshold is chosen for the frame energy, which is close to the upper peak in the frame energy distribution figures. The frames with the frame energy above the threshold were extracted as potential voiced frame, such that the low energy frames which possibly represent the silence and fricative were removed. The spectral periodograms slope computation was performed on the selected frames. The resulting periodograms were averaged and a linear regression was deployed to compute the slope of the spectrum. Table 1 summarizes the average spectral slope for the five speech modes.

Table 1: *Mean spectral tilt (dB/Octave) for five speech modes*

Speech mode	Wh	So	Ne	Lo	Sh
Mean slope	-2.86	-6.71	-8.29	-8.27	-7.51

This table indicates that the spectral slope decreases for whispered, soft, loud and shouted speech with respect to neutral speech. The decreases for whispered and soft speech imply the energy in high frequency increases when speech intensity decreases. This could be due to increased consonant frame energy. The decreases of spectral slope for loud and shouted speech may be caused by the glottal pulse with more regular shapes.

4. Performance of speaker-ID system

4.1. System details

To explore the impact of vocal mode on the automatic speech systems, speech from the five vocal modes were employed in a closed in-set speaker recognition evaluation. The system selects 1 of N closed speakers for recognition, where training and test materials are open. Due to limited sized training material, a Universal Background Model (UBM) was constructed from a separate set of speakers from the UT-SCOPE corpus [1]. A speaker specific MAP adapted Gaussian Mixture Model (GMM) is obtained from the UBM for each of the trained close-set speakers. The scores obtained by comparing the test utterances with the trained speaker models were sorted and the model with the highest score is identified as the model which fits the test utterance best. Further details of the GMM-UBM system can be found in [9].

4.2. Experimental setup

Speaker and development set: The same speech data used in analysis part were chosen to test the automatic speaker recognition system. The speech data consists of 110 sentences collected from the other 11 subjects in the neutral mode, which was also used for training the UBM.

Front-end processing: Speech from all speakers was windowed with a Hamming window of 20ms duration with 10ms overlap rate. A 19-dimensional MFCC feature vector was extracted from all speech data.

4.3. Experiments and results

The speech data which was used in Sec. 3 has been used to train the GMM for each subject and tested for automatic speaker recognition. The 10-12 seconds (6 utterances) speech data from each subject was used for training the GMM for that subject. The other 4 utterances were used for system test. The round-robin technique was used to obtain average performance of the Speaker-ID system for each speech mode. The experiment results are shown in Table 2.

Table 2: Accuracy rate (%) of the ASI system for training and testing data under five speech modes

Train \ Test	Wh	So	Ne	Lo	Sh
Whispered	94.6	33.3	30.4	23.3	17.9
Soft	57.9	97.5	86.3	61.7	41.7
Neutral	46.7	86.7	98.8	86.3	56.3
Loud	39.2	66.7	92.1	98.3	64.2
Shouted	27.1	40.4	53.8	68.3	97.1

The comparison of the highlighted diagonal elements of the table suggests that matched speech mode performs well, and that a mismatch in vocal mode can seriously impact speaker-ID performance (e.g. an average reduction from 97.26% in matched to 54.02% in mismatched modes). Particularly, the whispered speech affects the performance of the Speaker-ID system most. The table also shows that the test data in specific speech mode fits the GMM trained by the same speech mode best. This result motivates us to develop a speech mode classification system based on the Speaker-ID system.

5. Speech mode classification

In [3], a whisper/phonated classification techniques based on a VQ speaker ID system has been proposed. The prior work and the experiment results described in Sec. 4 motivate us to explore a classification technique which can automatically identify the speech mode based on the Speaker-ID system. Instead of training the GMM for each subject under five modes, the GMM for each speech mode was trained to compute the score for the test utterances. Experimental evaluation was performed on the modified GMM system. For each speech mode, 100 utterances from 10 subjects under that speech mode were used to train the GMM for that speech mode. A total of 20 sentences in each speech mode from the other 2 subjects were tested on the GMM based speech mode classifier. The round-robin technique was used to obtain the average performance of the classification technique. The results are shown in Table 3.

Table 3: Classification rate (%) of the GMM based speech mode classification technique

Test Mode \ Assessed as	Wh	So	Ne	Lo	Sh
Wh	98.5	1.5	0	0	0
So	0	71.5	25.5	3	0
Ne	0	36	44.5	19.5	0
Lo	0	3	22	64.5	10.5
Sh	0	0	0	32.5	67.5

According the highlighted parts in Table 3, we can sense that the sentence can mainly be assessed as its true vocal effort mode or its adjacent modes. Furthermore the largest difference of classification rates (97%) between adjacent

whispered mode and soft mode indicates the fundamental difference of whispered speech in production mechanism from other four vocal effort modes. The results here for speech mode classification suggest a consistent detector for whispered speech, and useful performance for soft and shouted speech. This knowledge would therefore be helpful in developing more effective speech systems across changes in vocal effort.

6. Conclusion and future work

This study has considered an analysis of the characteristics of speech in different speech modes. It has been shown that the sound intensity levels of five speech modes are significantly different and the average speech intensity can represent the speech mode. The duration of sentence increases as long as the speech intensity goes higher or lower than that of neutral speech. The percentage of silence frames decreases when the speech intensity increases. The frame energy distribution indicates an energy shift from lower to higher levels when the speech intensity increases. The average spectral slope of the speech decreases, indicating an increase in high frequency components when speech intensity decreases. The results of an experiment performed on Speaker-ID show the negative impact from changes in vocal effort. It also shows that whispered speech which has the lowest speech intensity degrades the performance of the Speaker-ID system most. A preliminary system for speech mode classification was also performed in this study.

Since whispered speech has the most discriminative characteristics and it has a more significant impact on the Speaker-ID system than other speech modes, it motivates a more efficient and robust technique to identify whispered speech embedded in neutral speech and improvement in speech system for future.

7. References

- [1] Varadarajan, V. S. and Hansen, J.H.L., "Analysis of Lombard Effect under Different Types and Levels of Noise with Application to In-set Speaker ID Systems", INTERSPEECH 2006-ICSLP, pp.937-940, 2006.
- [2] Hansen, J.H.L., "Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatics Recognition", Georgia Inst. Tech. July, 1988.
- [3] Wendt, J, Cupples, J., Floyd, M., "A study on the Classification of whispered and Normally Phonated Speech", INTERSPEECH2006, pp.649-652, 2002.
- [4] Jovicic, S. T., "Formant Feature Difference between Whispered and Voice Sustained Vowels", Acoustica, Vol. 84, 1998, pp.739-743.
- [5] Wilson, J.B., "A Comparative Analysis of Whispered and Normally Phonated Speech Using An LPC-10 Vocorder", RADC, Final Report TR-75-264.
- [6] Rostolland D., "Acoustic Features of Shouted Voice Part I", Acustica, Vol. 50 pp.118-125, 1982.
- [7] Rostolland D., "Phonetic Structure of Shouted Voice Part II", Acustica, Vol. 51 pp.80-89, 1982
- [8] Denes, B., Pinson, N., "The Speech Chain", W.H. Freeman and Company, New York, 2001
- [9] Angkititrakul, et. al, "Cluster-dependent Modeling and Confidence Measure Processing", ICSLP 2004
- [10] Womack, B.D., Hansen, J.H.L., "N-channel hidden Markov models for combined stressed speech classification and recognition", vol. 7, Issue 6, IEEE Trans. 1999