AUDIOVISUAL CORPUS TO ANALYZE WHISPER SPEECH

Tam Tran, Soroosh Mariooryad and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Dep. of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

tmt090020@utdallas.edu, soroosh.ooryad@utdallas.edu, busso@utdallas.edu

ABSTRACT

Current *automatic speech recognition* (ASR) systems cannot recognize whisper speech with high accuracy. ASR systems are trained with neutral speech, which have significant acoustic differences with whisper speech (i.e., energy, duration, harmonics structure, and spectral slope). Given the limitations of speech-based systems to process whisper speech, we propose to explore the benefits of visual features describing the orofacial area. We hypothesize that the lips' articulation between whisper and neutral speech is similar, providing a valuable whisper-invariant modality. This paper introduces the first audiovisual corpus of whisper speech. While we are targeting over 40 speakers, the current corpus has recordings from eleven subjects who were asked to read TIMIT sentences, and isolated digits alternating between neutral and whisper speech. The corpus also includes spontaneous recordings, in which the subject answered a series of general questions. The paper also analyzes an exhaustive set of audiovisual features, including *action units* (AUs), lip spreading, fundamental frequency, intensity, MFCCs, and formants. We study the differences in the features' distributions between whisper and neutral speech using *Kullback-Leibler divergence* (KLD). Then, we conducted statistical test to determine whether the differences in the features are statistically significant. The results support our hypothesis that visual features are less affected by whisper speech.

Index Terms— Audiovisual corpus, whisper speech

1. INTRODUCTION

Whisper speech is characterized by the absence of periodic excitation, changes in energy and duration characteristics, shift of lower formant locations, and changes in the spectral slope [1–3]. Given these differences, the performance of speech systems such as *automatic speech recognition* (ASR) [1], *speaker identification* (SID) [4, 5] or keyword spotting [6] significantly decrease in presence of whisper speech. Recognizing whisper speech is important in situations when individuals give important private information such as social security numbers, credit card numbers, or pin numbers in public areas where such information is easily overheard. This problem is also important for vocally impaired individuals such as heavy smokers or quiet speakers. Likewise, recognizing the identity of people who whisper is relevant in the area of national security and defense.

To address the intrinsic limitations of current speech systems to process whisper speech, studies have proposed robust features such as *modified temporal patterns* (m-TRAPs) [7], feature warping over *Mel-frequency cepstral coefficients* (MFCCs) [5] and model adaptation schemes such as *maximum likelihood linear regression* (MLLR) [1]. Others have considered alternative sensing technologies such as throat microphones [8]. We propose to explore additional modalities that are invariant to the changes observed in whisper speech. In particular, we hypothesize that features describing the appearance and displacement of the orofacial area will not be significantly affected by whisper speech. Studies have validated the benefits of using audiovisual speech recognition, particularly when the acoustic signal is noisy [9–11]. Also, the advent of laptops, tablets and smart phones with frontal camera can facilitate the popularity of audiovisual interfaces in future communication systems. Therefore, using visual features is a viable method to improve the performance of whisper speech recognition. Toward this goal, this paper introduces the first audiovisual corpus for whisper speech, and our initial analysis on the changes observed in acoustic and orofacial features caused by whisper speech.

The *audiovisual whisper* (AVW) corpus currently contains recordings from 11 subjects, who speak English as native language (our goal is to collect over 40 individuals). The corpus contains audiovisual data of read and spontaneous speech in whisper and neutral modes. The subjects read sentences and isolated digits (1-9, "zero" and "oh"), and answered predefined questions in both modes. The corpus is collected in a sound booth with 2 professional LED light panels, providing ideal conditions for audio and video recordings.

We analyze the corpus to determine audiovisual features that are more affected by whisper speech. First, we estimate the deviation in the distributions of an exhaustive set of acoustic and visual features caused by whisper speech. The analysis relies on the *Kullback-Leibler divergence* (KLD). Then, we evaluate whether the differences are statistically significant using matched pair two-tailed t-test. Both analyses reveal that visual features are less affected by whisper speech, supporting our hypothesis.

2. RELATION TO PRIOR WORK

Previous studies have identified significant differences between whisper and neutral speech [1–3, 12]. These differences include changes in the vocal excitation and vocal tract function [4]. For example, studies have reported larger frequency shifts at lower formant frequencies, and little or no shifts at high formant frequencies [13, 14]. The differences between modes have stronger effects on certain phonetic units [12, 15]. For example, consonants have prolonged durations and their intensity depends on whether they are voiced or unvoiced [12]. Fan et al. [15] studied the properties of whisper speech and the dependencies across speakers and phonemes. The study concluded that compensation schemes should consider speaker dependency.

Given the degradation on the speech signal, studies have proposed schemes to compensate the mismatches introduced by the differences in acoustic features [5, 7]. Jou et al. [8] proposed to use complementary sensing technologies, such as throat microphone. Our research explores visual features for whisper speech recognition. Our hypothesis is that facial features will not be significantly affected by whisper speech. This hypothesis is not completely clear since studies have reported adaptation strategies in visual modal-

This work was funded by NSF (IIS-1217183) and Samsung Telecommunications America.

(a) Subject (b) Equipment

Fig. 1. Data recording in the sound booth.

ity in the presence of other vocal effort (e.g., hyper-articulation in Lombard speech). For example, Garnier et al. [16] found that visual features are subconsciously altered when there is a change in vocal effort. This change occurs whether or not they are interacting with another subject. However, we expect that the differences in facial features between speech modes are less pronounced than the ones in acoustic features.

In our previous work, we presented a preliminary evaluation of audiovisual whisper speech recognition for isolated digits [17]. The approach was implemented with separate *hidden Markov models* (HMMs) for acoustic and visual modalities. By including the visual modality, the system improved the word accuracy from 42.7% to 79.7%. While the results are impressive, the main limitation was the limited corpus collected from a single subject with only 30 minutes per speech mode. This work motivates our group to design and record the first audiovisual corpus of whisper speech, which we expect to make available to the community. This paper presents the corpus and our initial analysis of acoustic and visual features.

3. DATABASE DESCRIPTION

This section introduces the *audiovisual whisper* (AVW) corpus, which is designed and recorded to study the benefits of using visual information to recognize whisper speech. The corpus consists of read (i.e., sentences and isolated digits) and spontaneous (i.e., answering to general questions) speech. While our goal is to record over 40 subjects, the corpus includes audiovisual recordings from 11 speakers at the present time (8 males and 3 females). The subjects are students at the *The University of Texas at Dallas* (UTD). All the subjects speak English as native language.

Figure 1 describes the recording setting. The corpus is collected in a 13ft $\times$ 13ft ASHA certified single-walled sound booth. The audio is recorded with a close-talk microphone at 48 KHz. Two high definition cameras are placed to capture frontal and side views of the subjects (1440$\times$1080 pixels, 29.97 fps). The cameras record the participants' upper body including shoulders and head (see Fig. 1(a)). The sound booth is illuminated with two professional LED light panels (Fig. 1(b)). While we realize that portable devices in real environment may not provide the quality of our data, our goal is to collect audiovisual data under ideal conditions to study the benefits of facial information in whisper speech recognition (e.g., no noise, good lighting, frontal view).

The corpus is recorded in three parts with suitable breaks in between. In the first part, the subjects are asked to read sentences in whisper and neutral mode. We selected 129 TIMIT sentences. A fixed subset of 30 sentences are used to record read speech in both whisper and neutral modes. This subset is used across speakers. In addition, we randomly selected 60 sentences per subject which are read in either whisper (30 sentences) or neutral (30 sentences) modes. Altogether, each subject read 120 sentences, which were presented in blocks of ten sentences alternating between modes – ten sentences in neutral mode followed by ten sentences in whisper mode. We implement this protocol to reduce the fatigue caused by whispering over long periods, and the cognitive load associated with switching too often between modes. In the second part, the subjects are asked to read isolated digits (i.e., 1-9, "zero", and "oh"). Each digit is read ten times in each mode producing 220 samples per speaker. Similar to the sentences, the order of the digits is randomized per subject and presented in blocks of ten, alternating between modes. In the third part, we collect spontaneous speech. The subjects are asked to respond to general questions (see Tab. 1). Each subject selected 10 out of 15 questions. After the selection, the questions are randomized and presented alternating between whisper and neutral modes. The average duration of their answers is 45 sec. The duration of each session is approximately 1 hour, including breaks. Some aspects of the protocol were adjusted as we collected the corpus (e.g., fixing the common sentences that are read in neutral and whisper mode across subjects, and the number of sentences and digits). Therefore, some of the early recordings slightly deviate from the described protocol.

Table 1. Questions used to elicit spontaneous speech

1.	Describe your typical schedule during the summer.
2.	If you could travel anywhere, where would it be and why?
3.	Describe your favorite book or movie.
4.	How do you like the weather in Texas?
5.	When (What time during the year) do you like to take vacations and why?
6.	How do you like to travel? (By air, boat, car, etc.)
7.	What do you like about UTD?
8.	What do you usually do in your spare time?
9.	What do you think of Arnold Schwarzeneggar?
10.	What kinds of food do you like and why?
11.	Describe a sport or activity you like to watch or do.
12.	Who is your favorite politician and why?
13.	Where do you like to go off-campus and why?
14.	If you could meet anyone in history that is no longer alive, who would it be and why?
15.	Describe your field of study.

The sentences, digits and questions are displayed in a screen using slides (see Fig. 1(b)). The background color of the slides is used to indicate whether the material should be read with whisper or neutral speech, which helps the subjects to use the right speech mode. A ringing sound is produced whenever we presented a different slide. This sound is recorded in a second channel and it is used to segment the speech. Transcriptions for the sentences and digits are extracted from the slides. The open-source software SAILAlign is used to forced-align the transcription to the speech signal [18]. As expected, the toolkit provides reliable alignment for neutral speech. We observed some errors for whisper speech given the differences in acoustic domain between speech modes. For this reason, some of the sentences are not included in the analysis. As part of our future work, we will transcribe the spontaneous portion of the recordings.

4. STUDY OF WHISPER AND NEUTRAL SPEECH

4.1. Feature extraction

The analysis compares visual and acoustic features under whisper and neutral conditions. The visual features are extracted from the frontal videos with the *computer expression recognition toolbox* (CERT) [19]. The toolkit estimates some of the *action units* (AUs) that define the *facial action coding system* (FACS) [20]. The AUs describe the movement of individual or groups of facial muscles. The study explores the AUs related to the lips area (see Tab. 2). In addition to AUs, we estimate the horizontal mouth opening. CERT provides the locations of the lips corners. The information is used to estimate their distance, given in pixels. Given the differences in facial anatomy between subjects, and their relative position with respect to the camera, the horizontal lip distance is normalized. For each subject, we estimate the average distance during neutral

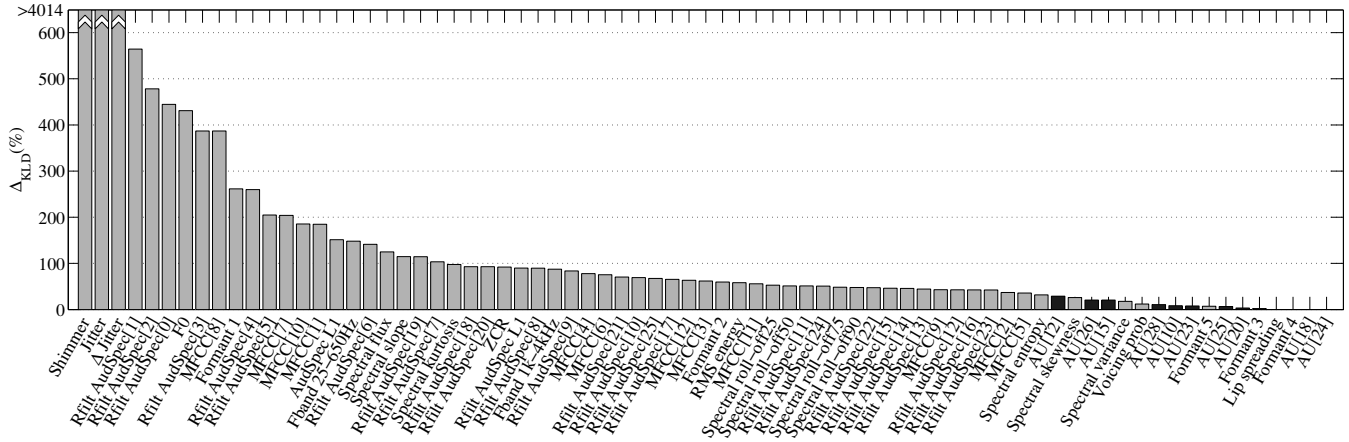

Fig. 2. Relative increase in KLD measurements, caused by whisper speech, across all acoustic (gray) and facial (black) features. Higher values indicate a larger deviation between the features' distributions in neutral and whisper speech (see Sec. 4.2).

Table 2. Facial features, which include *action units* (AUs) extracted by CERT [19] and distance between lip corners.

action unit	description	action unit	description
Action Units			
AU 10	Lip Raise	AU 23	Lip Tightener
AU 12	Lip Corner Pull	AU 24	Lip Presser
AU 15	Lip Corner Depressor	AU 25	Lips Part
AU 18	Lip Pucker	AU 26	Jaw Drop
AU 20	Lip stretch	AU 28	Lips Suck
Lip Features			
Lip spreading	Horizontal Lip Spreading		

condition, which is used to divide the horizontal lip value for each frame. Notice that the vertical opening of the mouth is given by AU25 (see Tab. 2). Notice that lip spreading and lip aperture have being considered to analyze visual features in Lombard speech [16].

The acoustic features include the exhaustive set of *low level descriptors* (LLD) given for the Interspeech 2011 speaker state challenge [21]. The set includes spectral, prosodic and voice quality features which are estimated using openSMILE [22] (see Tab. 3). In addition, we estimate the first five spectral formants using Praat [23].

4.2. Analysis of Audiovisual Feature Distribution

The first part of the analysis consists in comparing the distributions of the audiovisual features in whisper and neutral speech. By comparing the distributions, instead of second order statistics such as means and variances, we expect to unveil the effect of whisper speech on acoustic and visual features. The study relies on the *Kullback-Leibler divergence* (KLD) (Eq. 1), which is used in information theory to assess the similarity between two *probability mass functions* (PMFs). KLD is a suitable metric to quantify the deviation in various features caused by whisper speech. The features' PMFs are estimated using K-means algorithm. First, a single distribution is estimated across the entire corpus including neutral and whisper speech. Then, global, nonuniform bins are estimated using K-means algorithms. While we achieved similar results for different values of K, we report results using $K = 40$. These bins are then used to estimate the features' PMFs for each speech mode condition. This analysis considers the entire corpus.

$$KLD(P||Q) = \sum_i \ln(\frac{P(i)}{Q(i)})P(i) \quad (1)$$

The proposed approach estimates the relative increase in KLD caused by whisper speech (i.e., Δ^f_{KLD} – see Eq. 2). We split the

Table 3. Frame-level acoustic features, which include the *low level descriptors* (LLDs) introduced in the Interspeech 2011 speaker state challenge [21] and formants extracted with Praat [23]

Spectral LLDs	
Rfilt AudSpec [X]	RASTA-style filtered auditory spectrum bands 1-26 (0-8kHz)
MFCC [X]	Mel-frequency cepstral coefficients 1-12
Fband [F1-F2]	Spectral energy 25-650Hz, 1000-4000Hz
Spectral roll-off [X]	Spectral roll-off point 0.25, 0.50, 0.75, 0.90
Spectral [statistic]	Spectral flux, entropy, variance, skewness, kurtosis, slope
Formants [X]	*Spectral Formants 1-5*
Prosody LLDs	
AudSpec L1	Auditory spectrum L1-norm (loudness)
Rfilt AudSpec L1	RASTA-style filtered auditory spectrum L1-norm
RMS Energy	RMS Energy
ZCR	Zero-crossing rate (ZCR)
F0	Fundamental frequency
Voicing prob	Voicing probability
Voice Quality LLDs	
Jitter	Frame-to-frame F0 deviations
Δ Jitter	Frame-to-frame Jitter deviations
Shimmer	Frame-to-frame amplitude deviations

corpus into two speaker independent partitions, one from five subjects and the other from six subjects. One of the partition is used as reference. For each feature f, we estimate the reference PMF, P^f_{Ref}, by considering only its neutral samples. The second partition is used to estimate the PMFs of neutral (P^f_N) and whisper (P^f_W) speech. By having a reference distribution, estimated from neutral speech, we can compensate for the differences in the intrinsic distributions across features. Therefore, we can directly compare the values of Δ_{KLD} estimated from different features. Notice that the higher the values of Δ_{KLD}, the stronger the differences in the features' PMFs. To maximize the usage of the corpus, we estimate this metric using two-fold cross-validation. We report the average Δ^f_{KLD} values.

$$\Delta^f_{KLD} = \frac{KLD(P^f_W||P^f_{Ref}) - KLD(P^f_N||P^f_{Ref})}{KLD(P^f_N||P^f_{Ref})} \times 100 \quad (2)$$

Figure 2 shows that acoustic features (gray bars) in whisper speech present the largest deviations from neutral speech. Visual features (black bars) are not as affected by the changes in speech modes as the acoustic features. The AU with the highest deviation from neutral speech is lip corner pull (AU12), which suggest some differences in articulation. However, the results agree with our hypothesis

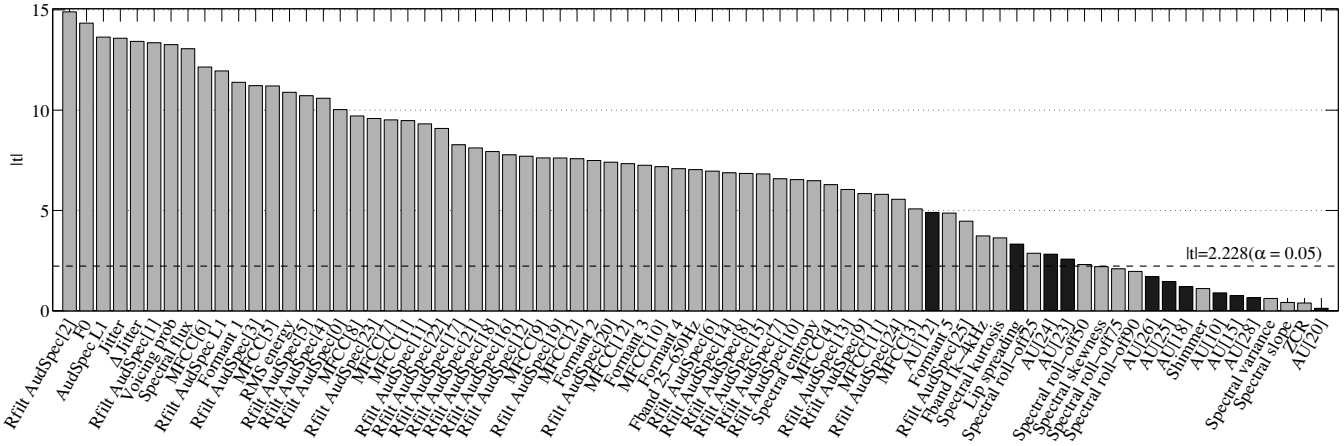

Fig. 3. Absolute t-values of the statistical test for all acoustic (gray) and facial (black) features. The rejection region is given when $|t| > 2.228$ (dashed line). Values in this region imply statistically significant differences in the features for neutral and whisper speech.

that visual features are more invariant to whisper speech. Features describing the fundamental frequency are the ones with the highest deviation from neutral speech. This result is expected given the lack of voicing during whisper speech. We also observe some predicted behaviors identified in previous work. For example, we observe that low frequency formants (F1 and F2) are more affected by whisper speech than higher formants (F3, F4, and F5) [13, 14]. Also, the energy in high frequencies (*Fband 1k-4kHz*) is less affected by whisper speech than the energy in low frequencies (*Fband 25-650Hz*). This result is consistent with the statistical test presented in Sec. 4.3 (Fig. 3). This result is expected since africatives, fricatives and stops are less affected by whisper speech than vowels.

4.3. Statistical Analysis of Features' Changes in Whisper Speech

The second part of the analysis addresses whether the differences in the features between speech modes are statistically significant. To compensate for the underlying lexical information, we consider only the portion of the corpus with isolated digits. We conduct a matched pair two-tailed t-test [24], in which the digit is the matched variable. Given that there are 11 digits (1-9, "zero" and "oh"), the t-test has 10 degrees of freedom. We estimate the t statistic for each feature using Eq. 3,

$$t = \frac{\bar{d}}{\sigma_d/\sqrt{n}} \approx \frac{\bar{d}}{s_d/\sqrt{n}} \tag{3}$$

where $\bar{d}$ and s_d are the sample mean and standard deviation of the differences in the features in neutral and whisper speech, for each of the digits. By setting the $p - value = 0.05$, the rejection region $|t| > t_{\alpha/2}$ for the null hypothesis (i.e., $|\mu_{N^{digit}} - \mu_{W^{digit}}| = 0$) is any t with absolute value greater than 2.228 (dashed line in Fig. 3). The phonetic alignment is used to remove the segments with silences at the beginning and ending of the samples.

Figure 3 shows the t scores for each of the audiovisual features. Most of the acoustic features (gray bars) present significant differences in the presence of whisper speech. In contrast, most of our visual features (black bars) are below the rejection region, which indicates that there are no enough evidences to reject the hypothesis that the patterns are similar in both speech modes. Only four visual features are found in the rejection region, with two of them very close to the critical t value. *AU[12]* and *Lip spreading* are the visual features with the highest t values. However, the t value for lips part (AU25) is out of the rejection region. While the lip spreading (i.e., horizontal distance) is affected by whisper speech, the lip aperture (i.e., vertical distance) presents similar patterns in both speech modes. Unfortunately, the features do not provide details about lip protrusion.

The results of the statistical test provide similar insights to the ones observed in the KLD analysis (Sec. 4.2 – Fig. 2). For example, we observe that the t values for the first two formants are higher than the ones for higher formants. Furthermore, we observe that many spectral values are greatly affected by whisper speech. This is clearly observed for MFCCs, for which all of the coefficients present statistically significant differences (also consistent with Fig. 2). This result is relevant because current ASR systems are built with MFCCs. Therefore, it is important to identify alternative features for automatic whisper speech recognition. Visual features may be the answer, as suggested by these results.

5. CONCLUSION

This paper presented the first audiovisual whisper database consisting of read (sentences and isolated digits) and spontaneous speech in both neutral and whisper modes. We analyzed the differences observed in audiovisual features during neutral and whisper speech. The study relies on KLD and matched pair two-tailed t-test. The analyses show that visual features are less affected by whisper speech than acoustic features. For speech features, the results agree with findings observed in previous studies (e.g., patterns in formants and energies in frequencies). For visual features, we observe that most of the differences between modes are observed in lip spreading. Other aspects describing the orofacial area are preserved during whisper speech.

Building upon this work, our next step is to gather more data and train an ASR system for whisper speech using audiovisual features. We expect to achieve similar results as the ones presented in our pilot experiment [17], which suggested an important improvement in whisper recognition when the visual features were included. As mentioned, our goal is to collect data from over 40 subjects, which will provide the necessary resources to systematically evaluate out multimodal solution. We are particularly interested in recognizing spontaneous speech, which will be the most challenging task.

Acknowledgment

The authors would like to thank the Machine Perception Lab (MPLab) at The University of California, San Diego for providing the CERT package.

6. REFERENCES

[1] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, February 2005.

[2] H.R. Sharifzadeh, I.V. McLoughlin, and M.J. Russell, "A comprehensive vowel space for whispered speech," *Journal of Voice*, vol. 26, no. 2, pp. 49–56, March 2012.

[3] S. Zhang, Z. Wu, H.M. Meng, and L. Cai, "Head movement synthesis based on semantic and prosodic features for a Chinese expressive avatar," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, HI, USA, April 2007, vol. 4, pp. 837–840.

[4] X. Fan and J.H.L. Hansen, "Acoustic analysis for speaker identification of whispered speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 5046–5049.

[5] Q. Jin, S. Jou, and T. Schultz, "Whispering speaker identification," in *IEEE International Conference on Multimedia and Expo (ICME 2007)*, Beijing, China, July 2007, pp. 1027 –1030.

[6] C. Zhang, *Whisper Speech Processing: analysis, modeling, and detection with application to keyword spotting*, Ph.D. thesis, University of Texas at Dallas, Richardson, TX, USA, May 2012.

[7] X. Fan and J.H.L. Hansen, "Speaker identification for whispered speech using modified temporal patterns and MFCCs," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 896–899.

[8] S. Jou, T Schultz, and A Waibel, "Adaptation for soft whisper recognition using a throat microphone," in *Interspeech 2004 - ICSLP*, 2004.

[9] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, September 2003.

[10] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proc. Human Language Technology Conference*, San Diego, CA, 2002.

[11] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Technical Report 764, Workshop 2000 Final Report, October 2000.

[12] S.T. Jovičić and Z. Šarić, "Acoustic analysis of consonants in whispered speech," *Journal of Voice*, vol. 22, no. 3, pp. 263–274, 2008.

[13] K.J. Kallasil and F.W. Emanuel, "Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects," *Journal of Speech and Hearing Research*, vol. 27, no. 2, pp. 245–251, June 1984.

[14] S.T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels," *Acustics united with Acta acustrica*, vol. 84, no. 4, pp. 739–743, July/Augustl 1998.

[15] X. Fan, K.W. Godin, and J.H.L. Hansen, "Acoustic analysis of whispered speech for phoneme and speaker dependency," in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011, pp. 181–184.

[16] M. Garnier, L. Ménard, and G. Richard, "Effect of being seen on the production of visible speech cues. a pilot study on lombard speech," in *Interspeech 2012*, Portland, OR, USA, September 2012.

[17] X. Fan, C. Busso, and J.H.L. Hansen, "Audio-visual isolated digit recognition for whispered speech," in *European Signal Processing Conference (EUSIPCO-2011)*, Barcelona, Spain, August-September 2011, pp. 1500–1503.

[18] A. Katsamanis, M. P. Black, P.G. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, PA, USA, January 2011.

[19] M.S. Bartlett, G.C. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J.R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, pp. 22–35, September 2006.

[20] P. Ekman and W.V. Friesen, *Facial Action Coding System: A Technique for Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, CA, USA, 1978.

[21] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Firenze, Italy, October 2010, pp. 1459–1462.

[23] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, 1996, http://www.praat.org.

[24] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*, Prentice-Hall, Upper Saddle River, NJ, USA, 2006.