AUDIO-VISUAL ISOLATED DIGIT RECOGNITION FOR WHISPERED SPEECH

Xing Fan, Carlos Busso, and John H.L. Hansen

Center for Robust Speech Systems(CRSS)
University of Texas at Dallas, Richardson, Texas 75083
xxf064000, busso, john.hansen@utdallas.edu

ABSTRACT

Whisper is used by talkers intentionally in certain circumstances to protect personal privacy. Due to the absence of periodic excitation in the production of whisper, there are considerable differences between neutral and whispered speech in the spectral structure. Therefore, performance of speech recognition systems trained with high energy voiced phonemes, degrades significantly when tested with whisper. In this study, we investigate the use of multi-stream models in isolated digit recognition of whispered speech. A small digit corpus with one subject speaking both whisper and neutral speech is collected. The eigenlips approach is used to extract visual features describing the lips appearance. MFCCs are employed as feature set for speech. Two HMM systems are trained for each stream independently and their scores are linearly combined. The resulted word accuracy shows significant improvement (37%, absolute). The study represents one of the first advancements in whisper recognition using audiovisual features. It also supports the use of multistream HMM to improve the performance on whisper/neutral speech conditions.

1. INTRODUCTION

Whisper is a natural alterative speech production mode. It is commonly used in public circumstances to avoid being overheard or to protect personal information. For example, a speaker may prefer using whisper when providing their date of birth, credit card number and billing address when making a hotel/flight/car reservation over the phone. Aphonic individuals, as well as those with low vocal capability, including heavy smokers, also employ whisper as a primary form of oral communication. Due to absence of periodic excitation, there is a significant difference between whisper and neutral speech in both time and spectral domain [1][2][3]. Those differences cause significant degradation when a neutral [1] trained ASR system is tested on whispered utterances. For example, Ito et al. showed that when a neutral trained continuous speech recognition system was tested with whispered speech, an absolute degradation of 63% was observed for word accuracy[2].

Considering the bimodal nature of human speech production, this study proposes to combine audio with visual information for an isolated digit recognition task for whispered speech. We hypothesize that normal lip configuration is preserved, up to some extent, in presence of whispered speech. Therefore, a multimodal system will increase the robustness and accuracy of the speech recognition system under this condition. In [4][5][6], audio-visual speech recognition systems based on multi-stream or coupled HMMs were proposed. These statistical frameworks offers varying degrees of asynchrony between acoustic and visual state sequence. However, no study has considered using visual information to help the recognition of low vocal effort speech, such as whisper.

Our study represents a first step in improving the performance of ASR systems for whisper/neutral mismatch conditions by using audio-visual speech recognition systems. As a case study, we investigate the use of multi-stream models in isolated digit recognition of whispered speech. We record a small digit corpus from one subject speaking in both conditions (whisper/neutral). Two HMM systems are independently trained (one for each stream). In the decoding step, their scores are linearly combined. The results confirm the potential of the proposed system when a mismatch between speech modes is observed. The audiovisual system achieves an accuracy of 79.7%, which is 37% (absolute) higher than the performance achieved by a system trained with only speech features. With the rapid development of new interfaces, the proposed approach can prove to be useful in many domains ranging from mobile devices to in-car applications.

The remainder of this paper is organized as follows. Section 2 provides a review of whispered speech studies. Section 3, introduces the corpus we collected for this study, and presents a description of feature extraction for both audio and visual information. Section 4 introduces the multimodal framework and section 5 provides the experimental results. Summary and discussion are included in Section 6.

2. REVIEW OF OF WHISPERED SPEECH

In neutral speech, voiced phonemes are produced through a periodic vibration of the vocal folds to produce glottal air flow into the pharynx and oral cavities. However, for whispered speech, the vocal folds remain open without vibration, resulting in no periodic excitation. The air flow from the lungs is used as the excitation sound source, and the shape of the pharynx is adjusted such that the vocal folds will not vibrate. Fig. 1 shows the waveforms of the speech signal " Guess the question from the answer" from the same speaker in both whisper and neutral mode. Clearly, the waveform for whisper speech is significantly lower in amplitude and longer in duration.

Due to significant differences in speech production, the acoustic characteristics of whisper are different from neutral mainly in formant locations, spectral slope, and energy.

This project was funded by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

[1] In this study, the term "neutral speech" refers to modal speech produced at rest in quiet soundbooth.

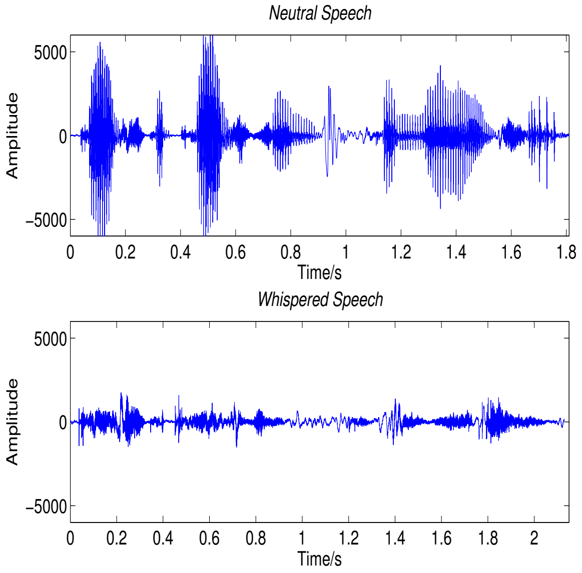


Figure 1: *Waveforms of whispered and neutral speech.*

Those differences present challenges when a neutral trained ASR system is tested with whispered speech. To the authors' knowledge, no study has analyzed the lip contour shape differences. We hypothesize that normal lip configuration is preserved, up to some extent, in presence of whispered speech. Therefore, a multimodal system is expected to increase the robustness and accuracy of the speech recognition system under this condition. Also, our recognition results in Section 5 will reflect some useful observations and will be presented in Section 4.

3. METHOLOGY

3.1 Corpus setup

To validate the hypothesis that visual information is helpful for improving the ASR performance for whisper, a database was recorded from 1 male native speaker of American English using a SONY dcr-sx41 video camera. The transcription contains the digits 0-9, including “zero” and “oh”, in English in random order. To obtain natural whispered speech, the recoding is conducted in a total of 4 sections. Both neutral and whispered speech were recorded, and each mode contains 0.5 hour of speech. Both audio and video recording devices were set up to record acoustic and visual information. Only the front of the face is considered in our corpus though the side of the face may contain useful information. Figure 1 shows an example frame from the visual data. The sampling frequency for visual data is 25 Hz, with a resolution of 720×576. The acoustic data is recorded using the microphone in the video camera. The camera distance to the subject was maintained at approximately 70 cm. The sample frequency for audio data is 44.1 kHz.

3.2 Feature extraction

For the acoustic signal, 13 dimensional static MFCCs are appended to first and second order derivatives to constitute a 39 dimensional feature vector. For the visual signal, an Eigenlips method [7] similar to Eigenfaces is employed. Instead of applying an active contour based method (i.e., snakes) for extraction of Region of Interest (ROI), which is usually time consuming, we employ a method based on Gaussian Mixture Models (GMMs) to detect the lip region. A total of 30 samples from lip areas are manually extracted from our corpus and employed as training data to represent the general color of the lips. The 3-dimensional RGB color vector for each pixel from those training data is used as feature vectors. By using the obtained training vectors, 4-mixture GMMs are trained by the EM algorithm to detect the lip area in each frame.

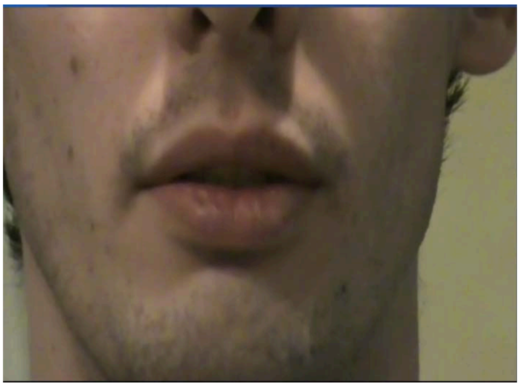
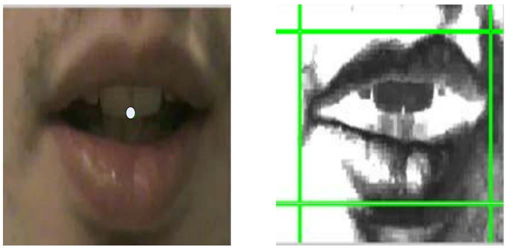


Figure 2: *An sample frame from the visual data.*



(a) Detected Lip center using Gravity Center

(b) Detected Lip boundary using GMM scoring

Figure 3: *Details about extraction of ROI.*

For each frame of visual information, a rough ROI is first extracted to remove the nose and the background. Next, the center of gravity of each frame is detected to indicate the center of the lips. The 3 dimensional RGB color vector for each pixel in the current ROI is tested with the trained GMMs to obtain a score matrix. Combining the knowledge of the center information and score matrix, the boundary of the lips can be estimated. In Fig. 3, we show the center and the boundary of the lips, estimated using the proposed method. Fig. 2(b) shows the score of each pixel using the trained GMM. After we obtain the ROI, it is resized to 100×200.

A Principle Components Analysis (PCA) is conducted for the final feature extraction. The Eigenface Matlab toolkit [8] is used to calculate the eigenlips. First, 400 frames are randomly selected from the neutral section of our corpus for calculating the eigenlips. Each image is transformed into a vector of size 20000 (100×200) and placed into the set *S*.

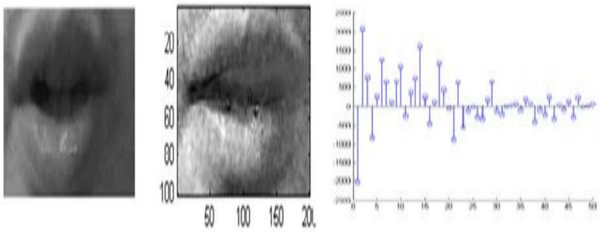


Figure 4: *An example of the original image of lips, reconstructed lips and corresponding weights.*

$$S = \{\Gamma_1, \Gamma_2, \Gamma_3, \ldots \Gamma_{400}\}. \quad (1)$$

After that, the mean image Ψ can be obtained as follows:

$$\Psi = \frac{1}{400}\sum_{n=1}^{400} \Gamma_n. \quad (2)$$

After obtaining the mean image, we find the difference between the input image and the mean image Φ:

$$\Phi_i = \Gamma_i - \Psi. \quad (3)$$

Next PCA is‘employed to determine a set of M orthogonal vectors, u_n, that best describe the distribution of the data. The k^{th} vector u_k is chosen such that,

$$\lambda_k = \frac{1}{400}\sum_{n=1}^{400} (u_k^T \Phi_n)^2 \quad (4)$$

Due to the fact that the vector length for each image is 100×200, the covariance matrix of Φ will be very large. In order to reduce the computational requirement, we calculate the covariance matrix of Φ^T. In this way, the size of the covariance matrix is just 400 × 400. The eigenvector for the covariance matrix of Φ can be easily obtained through a transformation of the eigenvector for the covariance matrix of Φ^T. In our study, we take the first 15 eigenvectors with the largest number of eigenvalues as the eigenlips. For testing images, the weight of each eigenlips can be found using Eq. (5). In Fig. 4 shows the weight and reconstructed lips using the extracted eigenlips. The dimension of the final feature vector is 30, containing 15 dimensional static weight appended with first order derivatives.

$$\omega_k = u_k^T(\Gamma - \Psi) \quad (5)$$

4. MULTIMODAL FRAMEWORK

The HTK toolkit is used in this study to build the HMMs for audio and video streams separately. For both streams, word level HMMs are trained instead of monophone or triphone models due to the limited amount of training data. Specifically, for the audio data, a HMM of 16 states (with 2 non-emitting state) is trained. Because there is only one subject in the corpus, each state has one Gaussian mixture. For the visual data, HMM of 7 states (with 2 non-emitting states) is trained. For the same reason, each state only has one Gaussian mixture. The structure for the audio and visual HMMs has the topology shown in Fig. 5. It should be noted that all training data comes only from neutral data unless specially mentioned. Compared to other feature fusion method, this multimodal framework has the advantage of considering the asynchrony between the audio and visual streams inside a word and also forces both streams to be synchronized at the word boundary. Also, the training process requires less computation compared to product HMMs or coupled HMMs.

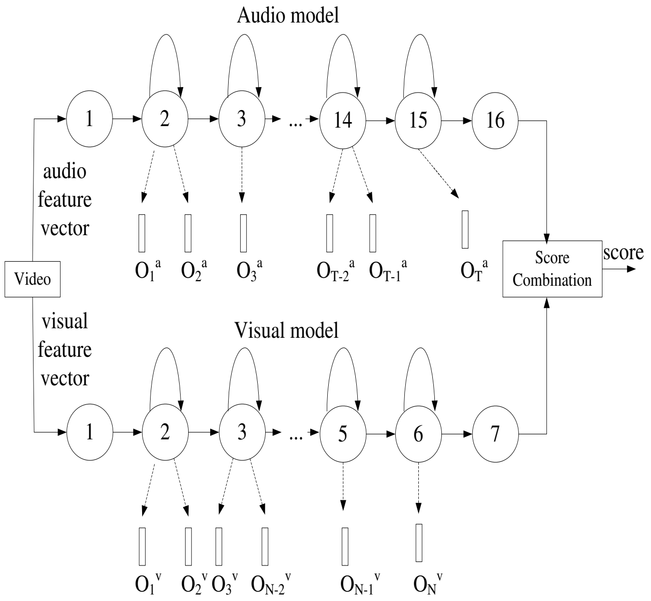


Figure 5: *An example of HMM typology used in this study with states number of 6.*

For both audio and visual data, their corresponding HMMs are trained in a continuous way without detecting the word boundary. In order to simplify the problem, we conduct isolated digit word recognition in this study. First, HMMs trained with neutral speech are adapted using all the available whisper data using the Maximum-Likelihood Linear Regression (MLLR) method. These adapted HMMs are used only for segmenting the word. During the recognition (results presented in section 5), we assume no whisper adaptation data is available. Next, the isolated whisper digit audio data is extracted using forced alignment. By using the searched boundary in audio data, we can extract the corresponding visual data for each isolated word. The audio and visual scores are combined as follows:

$$\begin{aligned} score^{final} = \lambda^a log\{\sum_X a^a_{x(0)x(1)} \prod_{t=1}^{T} b^a_{x(t)}(o_t) a^a_{x(t)x(t+1)}\} + \\ \lambda^v log\{\sum_X a^v_{x(0)x(1)} \prod_{t=1}^{T} b^v_{x(t)}(o_t) a^v_{x(t)x(t+1)}\} \end{aligned} \quad (6)$$

where X is the state sequence and a is the state transmission probability matrix. λ^a and λ^v are the score weights of the audio HMMs and video HMMs, respectively. The weight is under the constraint of Eq. (7):

$$\lambda^a + \lambda^b = 1. \quad (7)$$

5. EXPERIMENTAL RESULTS

First, we will only use single stream for recognition. The results are shown in Table 1. For the audio data, the word

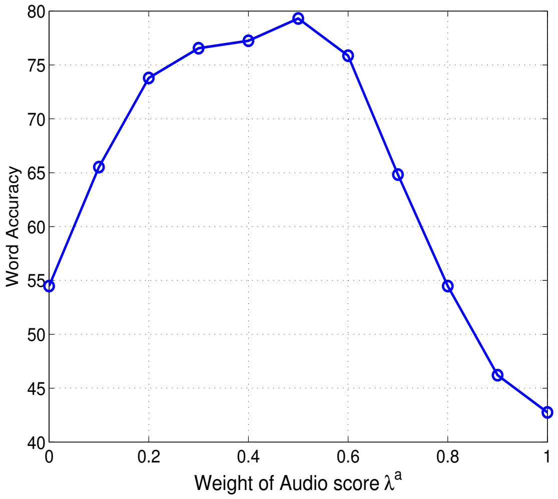


Figure 6: *Relation between accuracy and λ^a.*

accuracy for whisper/whisper train/test condition is 15.4% lower compared to neutral/neutral condition. This result suggests that whisper contains less speech dependent information than neutral in acoustic domain due to the absence of periodic excitation and lower energy. Also, a significant degradation of more than 40% in performance is observed for neutral/whisper train/test mismatch condition, which confirms the difference between whisper and neutral in spectral structure. For the visual data, the performance between neutral/neutral and whisper/whisper is close. Meanwhile, for the neutral/whisper train/test condition, the performance is 16% lower than the neutral/neutral matched condition. Even though, this result suggests that there are differences between whisper and neutral speech in the visual domain, the performance does not dramatically decrease as the case in the acoustic domain.

Table 1: *Word accuracy using single one stream data.*

stream	training	test	Word Accuracy
audio data	neutral	neutral	98.7%
audio data	whisper	whisper	83.3%
audio data	neutral	whisper	42.7%
video data	neutral	neutral	70.7%
video data	whisper	whisper	68.0%
video data	neutral	whisper	54.7%
combined (best)	neutral	whisper	79.7%

When combining the two streams together, we can obtain a set of word accuracy for the whisper test data by adjusting the value of λ^a. The results can be seen in Figure 6. The best performance is achieved when λ is equal to 0.5, which yield a word accuracy of 79.7%. This represents an absolute improvement of 37% compared with audio HMMs system and 25% compared with visual HMMs system.

6. CONCLUSIONS

In this study, we investigate the use of multi-stream models for developing seamless speech recognition system for whisper/neutral mismatch condition. A small digit corpus is collected using video camera in the soundbooth room. Both audio and video data have been collected in a synchronized way. A feature extraction method based on eigenlips is applied for the video data, and conventional MFCCs is extracted from the audio data. HMMs for audio and visual streams are trained independently. The test is conducted for isolated word recognition by linearly combining the score from audio and visual streams. The synchrony between audio and visual data is constrained at the word boundary. The final system presents a word accuracy of 79.7%. Compared with using audio data or video data alone, this result presents a significant improvement and confirms the potential of using multistream framework for improving the performance of mismatched speech mode in the training and testing set.

While the results are encouraging, it is clear that a much larger and more comprehensive corpus collection is needed to demonstrate the effectiveness of the proposed algorithms in actual voiced communication/voice dialog systems. The results, though, do represent one of the first advancements in address speech recognition for whispered speech, and confirms the viability of the overall multistream framework to improve the performance on whisper/neutral speech conditions.

REFERENCES

[1] C. Zhang and J.H.L. Hansen, "Analysis and Classification of Speech Mode: Whisper through Shouted", *ISCA INTERSPEECH*, 2007.

[2] T. Ito and K. Takeda and F. Itakura, " Analysis and Recognition of Whispered speech", *Speec Communication*, pp. 139-152, vl. 45, issue. 2, 2005.

[3] S. Jovicic and Z. Saric, "Acosutic Analysis of Consonants in Whispered Speech",*J. of Voice*, pp. 263-274, vl. 22, 2005.

[4] G. Gravier and G. Potamianos and C. Neti, "Asynchrony modeling for audio-visual speech recognition", *Human Language Technology Conference*, pp. 1-6, November, 2002.

[5] C. Neti and G. Potamianos and I. Matthews and H. Glotin and J Luettin, "Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins", *Workshop on Multimedia signal processing, special session on join audio-visual processing*, October, 2001.

[6] A.V. Nefian and L. Liang and X. Pi and X. Liu and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition", *EURASIP Journal in Applied Signal Processing*, pp. 1274-1288, issue 11, 2002.

[7] C. Bregler and Y. Konig, "Eigenlips for Robust Speech Recognition", *ICSI Technical Report TR-94-002*, January, 1994.

[8] http //www.pages.drexel.edu/ sis26/Eigenface Tutorial.htm