Lipreading Approach for Isolated Digits Recognition under Whisper and Neutral Speech

Fei Tao and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
Email: fxt120230@utdallas.edu, busso@utdallas.edu

Abstract

Whisper is a speech production mode normally used to protect confidential information. Given the differences in the acoustic domain, the performance of *automatic speech recognition* (ASR) systems decreases with whisper speech. An appealing approach to improve the performance is the use of lipreading. This study explores the use of visual features characterizing the lips' geometry and appearance to recognize digits under normal and whisper speech conditions using *hidden Markov models* (HMMs). We evaluate the proposed features on the digit part of the *audiovisual whisper* (AVW) corpus. While the proposed system achieves high accuracy in speaker dependent conditions (80.8%), the performance decreases when we evaluate speaker independent models (52.9%). We propose supervised adaptation schemes to reduce the mismatch between speakers. Across all conditions, the performance of the classifiers remain competitive even in the presence of whisper speech, highlighting the benefits of using visual features.

Index Terms: Lipreading, whisper speech, multimodal corpus

1. Introduction

Whisper is a speech mode characterized by low energy and lack of vocal cord vibrations. It is an important speech mode that people use to protect confidential information, speak in quiet places, and cope with temporary or permanent speech disorders (e.g., amygdalitis or cold and heavily smoker condition). The differences between neutral and whisper speech in acoustic features degrade the performance of *automatic speech recognition* (ASR) systems when they are trained with neutral speech and tested with whisper speech [1, 2]. Different approaches have been proposed to reduce the mismatches in speech modes. The solutions include robust features [3], feature normalization [4], model adaptation [5], and alternative sensing technologies such as throat microphones [6].

An appealing approach to improve whisper speech recognition is the use of visual features describing the geometry or appearance of the lips. The visual information plays a key role in speech intelligibility, especially under noise or soft/whisper speech. We have demonstrated that, compared to acoustic features, visual features are less affected by whisper speech [7]. Furthermore, the advances and ubiquity of portable devices with frontal cameras make lipreading an attractive modality that is relatively invariant to whisper speech. This paper explores lipreading for isolated digit recognition under normal and whisper speech conditions. With the exception of our early work which proofed the concepts [2], this is the first study that use lipreading to recognize whisper speech.

We represent visual information with geometric and appearance based features, which provide a reasonable trade-off between accuracy and generalization of the models. The isolated digit recognition is implemented with *hidden Markov model* (HMM). The results demonstrate that the accuracies only decrease 8.9% (absolute) under mismatched conditions (i.e., trained with neutral speech, tested with whisper speech). The evaluation considers speaker dependent and speaker independent models. We explore model adaptation schemes to reduce the drop in performance observed under speaker independent condition, achieving over 24% accuracy improvements. To the best of our knowledge, this is the first study on whisper speech recorded from multiple speakers (40 subjects from the AVW corpus) that uses lipreading for isolated digit recognition.

2. Relation to Prior Work

We have studied the use of lipreading to improve whisper speech recognition [2, 7]. Fan et al. [2] demonstrated that visual features describing the lips can compensate the drop in performance caused by whisper speech in isolate digit recognition. When the system was trained with neutral speech and tested with whisper speech, the audiovisual features increased the accuracy by 37% (absolute) compared to the case when only acoustic features were used. However, the study considered a limited corpus recorded from one subject. The promising results led us to collect the AVW corpus, a multimodal database for whisper speech research. Our feature analysis demonstrated that facial features were more robust against whisper speech than acoustic features [7]. This paper extents our effort by building a lipreading system for whisper speech evaluated across multiple speakers.

This study leverages the advances on lipreading. A research group at IBM investigated an audiovisual speech recognition system [8, 9, 10]. They created a large database from 290 subjects, containing 24325 utterances with a vocabulary size of about 10500 words. They explored two types of features: *discrete cosine transform* (DCT) and *active appearance model* (AAM). Given differences in appearance across speakers, they showed that the accuracy decreases for both features under speaker independent evaluations. The system did not compensate for head rotations.

Other research groups have worked on lipreadings using smaller corpora. Zhang et al. [11, 12] used geometric lip features describing the shape and opening of the mouth. The results on isolated word recognition were significantly better in speaker dependent conditions than in speaker independent con-

This work was funded by NSF (IIS-1217104) and Samsung Telecommunications America.

(a) (b) (c) (d)

Figure 1: The MSP-AVW corpus: (a-b) show the setting used to record the corpus. (c-d) show sample images taken from one subject.

ditions. Yau et al. [13] used *motion history image* (MHI), achieving 84.7% accuracy in recognizing nine visemes (consonants). Shaikh et al. [14] used vertical optical flow as feature, and *support vector machine* (SVM) as classifier. They focused on viseme classification. The number of feature frames per viseme was fix to reduce speaker variability. They observed good performance on speaker independence tests. Bregler et al. [15] implemented a *neural network* (NN) classifier to recognize words that were spelled by two subjects. Benhaim et al. [16] applied local features and multiple kernel learning to recognize isolated digits (CUAVE corpus). They reported 85% accuracy in speaker independent tests (36 speakers).

The advances on the area of lipreading offer an attractive approach to improve whisper speech recognition systems, since visual features are less sensitive to this speech mode than acoustic features [7]. Our paper evaluates this approach using HMMs trained with geometric and appearance based features, over a corpus recorded from 40 subjects.

3. Data Preparation

3.1. The MSP-AVW Database

This study uses the *audiovisual whisper* (AVW) corpus [7]. The corpus is being recorded at the *multimodal signal processing* (MSP) laboratory, and contains data from 20 females and 20 males speakers. For each subject, we record three sessions consisting of read sentences, isolated digits and spontaneous speech. The data is recorded under neutral and whisper conditions. The corpus is being collected in a 13ft $\times$ 13ft ASHA certified single-walled sound booth, illuminated by two professional LED light panels (see Fig. 1). The audio is recorded with a close-talk microphone at 48 kHz; the video is collected with two high definition cameras which provide 1440×1080 resolution at 30 fps. One camera captures frontal view of the subjects including shoulder and head. The second camera captures profile view of the subjects (see Fig. 1). For the latest recordings, we included green screens to facilitate video processing steps. The corpus is described in Tran et al. [7].

This study only uses the recordings corresponding to isolated digits from the 40 speakers. The isolated digit recording includes sequences of numbers that are spoken under normal and whisper conditions (i.e. 1-9, "zero" and "oh"). Each digit is recorded ten times per speech condition, which are randomly presented in groups of ten, alternating between modes. At the beginning of the recordings, the subjects are asked to pose a neutral face for 2-3 secs. These clips are used to extract a frontal face template image for each subject (see Sec. 3.2 – Fig 2). After that, the participants can move their body and head, without any constraint. Some subjects wore eye glasses, hat, and ear rings. These recording conditions are less restrictive than the ones used to collect similar audiovisual corpora such as the CUAVE database [17]. Our settings provide more realistic conditions for practical multimodal interfaces. Figure 1 shows samples from two subjects.

As part of the data processing procedure, the recordings are segmented into turns. We use the open-source software SAILAlign [18] to force-align the transcription to the speech signal. The phonetic boundaries are used to estimate the video sequences for each digits. Given the acoustic differences between whisper and normal speech, the toolkit fails to find the alignment for some samples. In these cases, we manually annotated the boundaries.

3.2. Face Analysis Processing and Facial Features

The estimation of facial features from videos requires preprocessing steps to detect and normalize the face against head rotation. Figure 2 describes the block diagram with the proposed procedure.

The first step consists in detecting facial landmarks with the CSIRO face analysis SDK [19]. A template image is manually created for each subject using his/her neutral facial pose collected at the beginning of the session. A total of 66 facial landmarks are carefully located from the frontal video. The landmarks include rigid feature points such as the nose tip, eye corners and outline of the face. The CSIRO toolkit implements a deformable model fitting by regularized landmark mean-shift [19] to align the facial expressions displayed in each frame with the frontal face template image. As a result, the facial landmarks are automatically detected in the frames.

As mentioned, participants were allowed to move their head during the recordings. Therefore, we normalize the face against head rotation using an affine transformation. The parameters of the affine transformation are estimated by comparing a subset of the landmarks between the frames and the template image [20]. The selected landmarks included rigid facial points, excluding the points describing the lips and eyelids. We apply this affine transformation to normalize the head pose (see Fig. 2).

We implemented a quality control step to minimize errors on the extraction of facial landmarks after normalization. First, we detect the face using the Viola-Jones face detector [21]. Then, we use a generative model [22, 23] to identify a set of nine facial landmarks including mouth corners and nose tip. The location of the nine landmarks are compared with the corresponding ones identified with the deformable model. We consider mouth width, eyes corners, nose points, face size, and mouth corner coordinates (see Fig. 2). We use the frames only when the face is transformed correctly and the landmarks are accurately detected. Otherwise, the frame is discarded and the facial features are considered as missing values. If less than 10% of the frames are missing, they are interpolated from other frames.

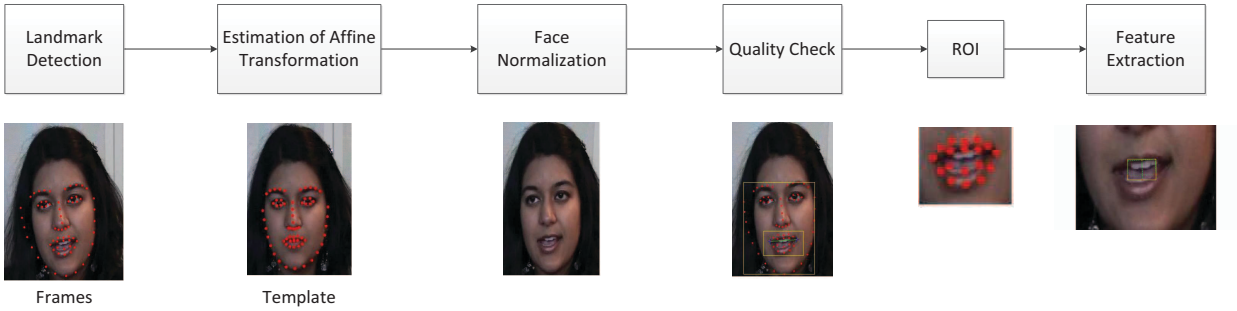

Figure 2: The block diagram to estimate facial features. Facial landmarks are detected using a deformable model that compares each frame with a template image. After normalizing head rotation, the location of the landmarks are evaluated during the quality control step. Finally, we determine the region of interest and we estimate the geometric and appearance based features.

Otherwise, we discard the video (838-neutral and 809-whisper). Across all the speakers, we use 3562 videos with neutral speech and 3591 videos with whisper speech. We estimate the *region of interest* (ROI) from the facial landmarks describing the lips (Fig. 2). From this region, we estimate facial features consisting of both geometric and appearance based features.

Studies show that geometric features are more robust against speaker variability [8]. Therefore, we consider five distances between six of the facial landmarks. We estimate the distances of the square formed by the left upper lip, left lower lip, right upper lip and right lower lip. In addition, we add the distance between the tips of upper and lower lips. All the distances are normalized by the mouth width from each frame to reduce the variability in shape and size across frames.

Under optimal recording conditions, appearance based features yield better performance for lipreading [11]. During speech production, the appearance of the tongue and teeth are important to distinguish between sounds. Geometric features cannot capture this information. This study considers a 25D *discrete cosine transform* (DCT) vector from the area enclosed by six lip landmarks highlighted in Figure 2 (see yellow box below the "feature extraction" block). We selected this region to reduce the inter speaker variability while still preserving important articulatory information. The trajectory of the features are normalized such that they all have the same length across digits. The approach consists of resampling and interpolation. The first and second order derivatives of all the features are calculated and appended to the features, resulting in a 90D feature vector ([25D-DCT +5D-distance] $\times$ 3) .

During our preliminary evaluation, we considered other facial features including local appearance based features, DCT coefficients estimated over larger lip areas, and optical flow. These features did not improve the performance presented here, so these results are not included.

4. Experimental Evaluation

This study uses HMM to recognize isolated digits (one model per digit). The facial features are normalized by subtracting the corresponding mean from each feature frame across the videos. The evaluation considers *speaker dependent* (SD) and *speaker independent* (SI) conditions. In both cases, we implement a leave-one-out cross-validation approach to maximize the usage of the corpus. The reported accuracies are the average across folds. We build the HMMs using different training and testing

Table 1: Accuracy for isolated digit recognition under various training/testing conditions (speaker independent (SI), speaker dependent (SD) and model adaptation (ADPT) cases).

Train	Test	SI(%)	SD(%)	ADPT(%)
NEU	NEU	52.93	80.78	77.31
WHI	WHI	52.34	82.64	76.24
NEU	WHI	50.87	71.85	68.14

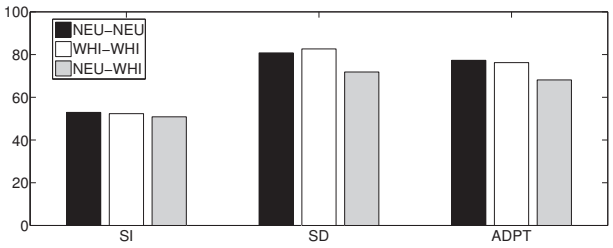

Figure 3: Accuracy for isolated digit recognition in speaker independent (SI), speaker dependent (SD) and model adaptation (ADPT) cases. We use three samples per digits for ADPT.

conditions: training and testing with neutral speech; training and testing with whisper speech; and, mismatched condition in which we train with neutral speech and test with whisper speech (most interesting case). The accuracies of the system under these conditions are listed in Table 1 and illustrated in Figure 3.

In the speaker dependent condition, we train the models for each of the subjects using cross-validation. In each fold, we use one sample per digit as testing set and the rest of the samples as training set. Table 1 shows that the accuracies under the matched conditions is above 80%. The performance drops to 72% in mismatched conditions (training in neutral speech and testing in whisper speech). The drop by 9% in accuracy suggests slight visual differences in speech production between neutral and whisper mode. Notice that the degradation in accuracy in speech based ASR can reach 71% in mismatched conditions [5]. The results for matched and mismatched conditions observed across the 40 subjects are significantly higher than the results reported in our previous work [2]. In that study, we achieved lipreading accuracies below 71% in matched conditions, and 55% in mismatched conditions, using a single

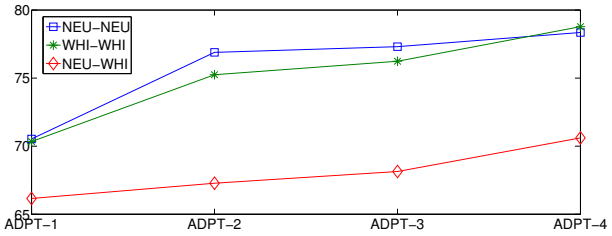

Figure 4: Accuracy of lipreading using adaptation. ADPT-1, ADPT-2, ADPT-3 and ADPT-4 give the results when the models are adapted with 1, 2, 3 and 4 samples per digit, respectively. The vertical axis gives the accuracy in percentage.

speaker. The proposed features are more reliable than the PCA based features that we previously used.

For the speaker independent condition, the cross-validation scheme is implemented such that in each fold one speaker is used for testing and the others for training (speaker independent partitions). Table 1 shows that the performance drops approximately 27% compared with the speaker dependent cases (chance level is 9%). This result is commonly observed in lipreading systems under speaker independent conditions [8, 9, 10, 11, 12]. For example, Zhang et al. [11] reported a drop in performance of almost 22% (absolute) from speaker dependent (48.9%) to speaker independent (26.94%) conditions. As a comparison, studies on the CUAVE database [17] have achieved similar accuracies for speaker independent conditions. Perez et al. [24] reported 47%, Gowdy et al. [25] reported 53.3% and Gurban and Thiran [26] reported 64% accuracies. Since the recording of the AVW database are less controlled than the recordings of the CUAVE database, a performance of 52.93% is very competitive. Notice that the drop in performance is worse when only appearance based features are considered (approximately 40% in matched conditions). Geometric features help to maintain the performance. The accuracy also decreases when the HMM system is trained and tested with whisper speech (from 82.64% to 52.34%), suggesting that people use different whisper strategies (e.g., hyper/hypo-articulation).

To address the drop in performance in speaker independent conditions, we combine the *maximum likelihood linear regression* (MLLR) and *maximum a posteriori* (MAP) adaptation algorithms. To understand the amount of data required for adaptation, we evaluate various configurations by increasing the data used for adaptation from 1 sample to 4 samples, per digits. The speech mode of the data used for adaptation matches the speech mode used for training (either neutral or whisper conditions). Figure 4 shows that the performance increases as the amount of data used to adapt the models increases. With four samples per digit, the accuracies reach over 77% in matched conditions. Table 1 reports the accuracies when three samples per digit are used to adapt the models. The gap in accuracy between matched and mismatched conditions is only 9.17%. This result suggests that when small set of data is available for the target subject, generic models trained across speakers can be easily adapted to improve the performance of the classifiers.

5. Conclusions and Future Work

This study investigated the use of lipreading in whisper speech recognition. We proposed a HMM approach that combines geometric and appearance based features. The evaluation considered training-testing mismatches with normal and whisper speech modes. The results show that the lipreading approach works well in speaker dependent tests (80.78%). The accuracy only decreases 8.9% when neutral speech models are tested with whisper speech, while the accuracy in speech based ASR can degrade as much as 63% [5]. To address the drop in performance in speaker independent conditions, the study considered MLLR and MAP speaker adaptation schemes. We presented a systematic analysis to determine the size of the corpus that is needed for adaptation. The results demonstrate that adaptation with only few samples per digit can improve the accuracy of the system over 18%.

The study suggests that lipreading approach is a feasible alternative to improve the performance of whisper speech recognition. Our next step is to fuse the proposed system with acoustic features. The results reported in this study are encouraging, since we have demonstrated that, even with lower accuracy, an audiovisual isolated digit recognition system tested with whisper speech can lead to 37% improvement in accuracy over a system based only on speech features [2].

We are working on statistical methods that capture the rich interaction between audiovisual modalities. We are also working on alternative strategies to reduce speaker and speech mode mismatches. Finally, we are exploring the use of phoneme/viseme models to recognize speech to extend our current solution to large vocabulary continuous speech recognition.

6. References

[1] C. Zhang and J. Hansen, "Analysis and classification of speech mode: Whisper through shouted," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2289–2292.

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

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

[4] 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.

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

[6] S. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, October 2004, pp. 1493–1496.

[7] T. Tran, S. Mariooryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 8101–8105.

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

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

[10] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Salt Lake City, UT, USA, May 2001, pp. 169–172.

[11] X. Zhang, R. Mersereau, and M. Clements, "Audio-visual speech recognition by speechreading," in *International Conference Digital Signal Processing (DSP 2002)*, vol. 2, Santorini, Greece, July 2002, pp. 1069–1072.

[12] X. Zhang, C. Broun, R. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP Journal on Advances in Signal Processing*, vol. 1, pp. 1228–1247, January 2002.

[13] W. Yau, D. Kumar, and S. Arjunan, "Voiceless speech recognition using dynamic visual speech features," in *HCSNet Workshop on the Use of Vision in Human-Computer Interaction (VisHCI '06)*, Canberra, Australia, November 2006, pp. 93–101.

[14] A. Shaikh, D. Kumar, W. Yau, M. Che Azemin, and J. Gubbi, "Lip reading using optical flow and support vector machines," in *International Congress on Image and Signal Processing (CISP 2010)*, Yantai, China, October 2010, pp. 327–330.

[15] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1993)*, vol. 1, Minneapolis, MN, USA, April 1993, pp. 557–560.

[16] E. Benhaim, H. Sahbi, and G. Vitte, "Designing relevant features for visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 2420–2424.

[17] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2002)*, Orlando, FL, USA, May 2002, pp. 2017–2020.

[18] A. Katsamanis, M. P. Black, P. 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] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, January 2011.

[20] R. Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, January 2006.

[21] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1, Kauai, HI, USA, December 2001, pp. 511–518.

[22] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! my name is ... buff" – automatic naming of characters in TV video," in *Proceedings of the British Machine Vision Conference (BMVC 2006)*, Edinburgh, Scotland, September 2006, pp. 899–908.

[23] J. Sivic, M. Everingham, and A. Zisserman, ""Who are you?" – learning person specific classifiers from video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami Beach, FL, USA, June 2009, pp. 1145–1152.

[24] J. Pérez, A. Frangi, E. Solano, and K. Lukas, "Lip reading for robust speech recognition on embedded devices," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 473–476.

[25] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "DBN based multi-stream models for audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)*, vol. 1, Montreal, Quebec, Canada, May 2004, pp. 993–996.

[26] M. Gurban and J. Thiran, "Information theoretic feature extraction for audio-visual speech recognition," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4765–4776, December 2009.