AUDIO VISUAL WORD SPOTTING

Ming Liu, Ziyou Xiong, Stephen M. Chu, Zhenqiu Zhang, and Thomas S. Huang*

Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign, Urbana, IL 61801
*Human Language Technologies Department,
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598
E-mail: {mingliu1, zxiong, zzhang6, huang}@ifp.uiuc.edu, *schu@us.ibm.com

ABSTRACT

The task of word spotting is to detect and verify some specific words embedded in unconstrained speech. Most Hidden Markov Model(HMM)-based word spotters have the same noise robustness problem as a speech recognizer. The performance of a word spotter will drop significantly under noisy environment. Visual speech information has been shown to improve noise robustness of speech recognizer[1][2][3]. In this paper, we combine the visual speech information to improve the noise robustness of the word spotter. In visual frontend processing, the Information-Based Maximum Discrimination(IBMD)[4] algorithm is used to detect the face/mouth corners. In audio-visual fusion, the feature-level fusion is adopted. We compare the audio-visual word-spotter with the audio-only spotter and show the advantage of the former approach over the latter.

Keywords: word spotting, face detection, mouth corner detection, audio-visual fusion

1. INTRODUCTION AND RELATED WORK

Although there has been significant progress on automatic continuous speech recognition(ACSR) in recent decades, its success is restricted to clean environments such as quiet office, and well-defined tasks such as dictation or medium vocabulary transactions. The performance of current ACSR systems has yet to reach the requirement for spontaneous/unconstrained speech. In scenarios such as command controls, car navigation, spoken document retrieval, word spotting can play an important role where the full-scale speech recognition is not necessary. For one, since word spotting does not need to transcribe all the words in the utterance, it can be made much faster than an entire decoding. For another, word spotting systems can deal with utterances that are not covered by the pre-defined grammar.

In recent years, visual speech information has been successfully used to improve noise robustness of speech recognizers. There have been a lot of audio-visual ACSR systems reported in literature. For audio-visual ACSR, the main issue is the visual frontend processing and audio-visual fusion strategy. There are mainly two categories of visual features: geometry based and appearance based. The appearance based features have been shown to outperform the other one[1]. There are three kinds of fusion strategies: feature-level, model-level and decision-level. The model-level fusion has been shown to outperform the other two[3]. Due to the limitation of the bimodal data, most of the systems are tested on very small number of speakers. The vocabulary of words of these systems are usually quite small. Only in recent years, the improvement has been demonstrated on large vocabulary tasks[3]. In contrast, few research efforts have been made on audio-visual word spotting systems. This motivates us to combine the visual speech information to improve the noise robustness of the word spotter.

The rest of the paper is organized as follows. An overview of our system is presented in Section 2, which includes face and mouth corner detection, audio-visual feature extraction and fusion for word spotting. An audio-visual database is presented in Section 3. Experimental results are described in Section 4. Conclusions and future work are in Section 5.

2. OVERVIEW OF THE SYSTEM

2.1. The System Diagram

Figure 1 show the diagram of the system. It is composed of following modules: an audio feature extraction component, a visual feature extraction component, an audio-visual fusion component and a word spotting component. The audio feature is more mature compared to visual feature. In this paper, the Mel-Frequency Cepstral Coefficients(MFCCs) are extracted as audio features.

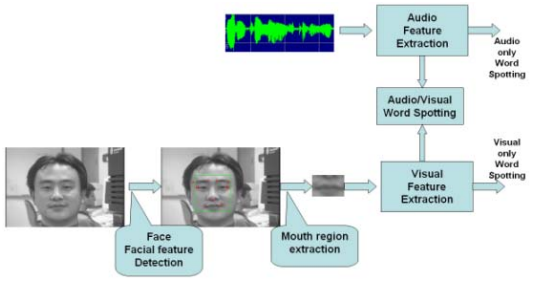

Fig. 1. The diagram of the proposed audio-visual word spotting system.

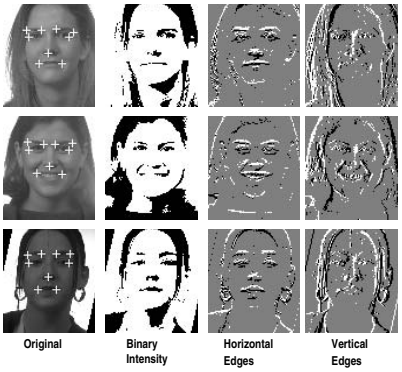

Fig. 2. Example of features detection

2.2. The Visual Feature Extraction Component

The module is illustrated in the lower part of the Fig. 1. It is composed of face detection, mouth corners detection and feature extraction from the mouth region.

2.2.1. Face Detection

For more detailed description of the Information-Based Maximum Discrimination(IBMD) algorithm, please refer to [5]. Here we summarize it from an intuitive point of view, instead of solid information theoretic analysis.

Our assumption is that the **permuted** version of the face image $O = (o_1, \cdots, o_n)$, $O' = (o_{s_1}, \cdots, o_{s_n})$ comes from a 2_{nd} order Markov process $X' = (X'_1, \cdots, X'_n) = (X_{s_1}, \cdots, X_{s_n})$ where $s_1, \cdots, s_n$ is a permuted sequence from $1, \cdots, n$. This optimal permutation is the one that maximizes a cost function which we choose to be the Kullback divergence(cross entropy) between $P_F(O')$ and $P_N(O')$. This optimization problem is solved by a modified version of the Kruskal's Algorithm for minimum-weight spanning tree.

2.2.2. Mouth Corner Detection

During training, We combine the results of three low-level image processing techniques: (i) binary quantization of the intensity, (ii) three-level of horizontal edges, and (iii) three-level of vertical edges. The threshold values used to re-quantize these low-level feature images are based on a fixed percentage of the pixels in a region in the center of the face. Combining these sets of discrete images, we construct the discrete image $I = I_i + 2I_v + 6I_h$ that is used to locate the facial features.

We trained the IBMD classifiers using 150 images in which the mouth corner positions were located by hand. We used three rotation angles and three scale factors to produce the image examples of facial features. Negative examples were obtained from image sub-windows at neighbor locations around the corresponding feature positions. The relative locations of the facial features in these training images were also used to determine the size and location of the facial feature search areas of the detection procedure.

In test, mouth corner detection is carried out on those face candidates found by the face detector. An example of the training faces, mouth corners and the low-level features is shown in Fig. 2.

2.2.3. Mouth Feature Extraction

After the mouth corners are detected, a mouth region is extracted from the image frame. Scale normalization is carried out by putting the detected two mouth corners to the pre-defined, fixed positions of an 18×12 image. Intensity histogram of the 18×12 image is then normalized. Principal Component Analysis(PCA) and Linear Discriminative Analysis(LDA) are preformed on the normalized mouth region to reduce the dimensionality of the feature space from 216 to 20.

2.3. Audio-visual Fusion Component

After visual feature extraction, the audio-visual fusion is the main issue for audio-visual speech processing. There are three types of fusion strategies: feature level fusion, decision level fusion and model level fusion. The first method, feature level fusion, is to concatenate audio feature and visual feature to a single larger feature.

$$O_{av,t} = [O_{a,t}, O_{v,t}] \tag{1}$$

where $O_{av,t}$ is the joint audio-visual feature, $O_{a,t}, O_{v,t}$ are the audio-only and visual-only feature respectively. Based on the combined feature, single stream HMMs are trained to represent the phonemes. In contrast, for the second method, decision fusion, it combines the output of two single modality classifiers.

$$P(O_{av,t}|C) = P(O_{a,t}|C)P(O_{v,t}|C) \tag{2}$$

where $P(O_{av,t}|C)$ is the likelihood of joint audio-visual feature $O_{av,t}$ from class C, $P(O_{a,t}|C)$ and $P(O_{v,t}|C)$ are

the likelihoods of audio-only feature $O_{a,t}$ and visual-only feature $O_{v,t}$ respectively from class C. The two classifiers $P(O_{a,t}|C)$ and $P(O_{v,t}|C)$ are independent. For the third method, the model level fusion, it tries to use more complex models, such as coupled HMMs, to model the audio-stream and visual-stream jointly.

Although, feature level fusion is limited to fuse the audio-visual information [3], in this paper, we adopt this method to combine audio and visual cues to develop a baseline system for further investigation.

2.4. Word Spotting Component

In our audio-visual word spotting system, context independent subword model was trained on the training set. In this paper, we choose the phoneme as the subword unit. A 3-state HMM is used to model each phoneme, and each state is a 16-component Gaussain mixtures model. A keyword model is represented by the concatenation of its phoneme models. The non-keyword model, which is also called the filler model, is represented by the combination of all the phoneme models and the silence model.

The decoding network of the word spotting system is shown in Figure 3. To allow sequences of keywords to be recognized, a null transition is added from the right-most node to the left-most node, as shown by the curve in the upper part of Figure 3. The network in Figure 3 can be used to identify the most likely sequence of keywords in the speech input. After the Viterbi decoding, absolute values of the likelihoods of the keywords can be thresholded to give a decision. However since these likelihoods are influenced by the noise and the speaker characteristics, we use the likelihood ratio of the keywords instead of the absolute values, as follows:

$$L(O_i^j|W_k, Filler) = \frac{P(O_i^j|W_k)}{P(O_i^j|Filler)} \qquad (3)$$

where O_i^j is the observation sequence from the i_{th} frame to the j_{th} frame, $P(O_i^j|W_k)$ is the likelihood score of the observation sequence from the keyword model W_k, while $P(O_i^j|Filler)$ is the likelihood score from the filler model. This likelihood ratio score is more robust. By thresholding the likelihood ratio score, we make the accept/reject decision.

3. AUDIO-VISUAL WORD SPOTTING DATABASE

In order to test the algorithm we proposed, we have collected an audio-visual word spotting database. It is collected by a video camera in a quiet studio environment. The resolution of video frame is 352×240, and the frame rate is 30Hz. The video is MPEG-1 encoded at compression rate 30:1. Exemplar video frames are shown in Figure 4. The

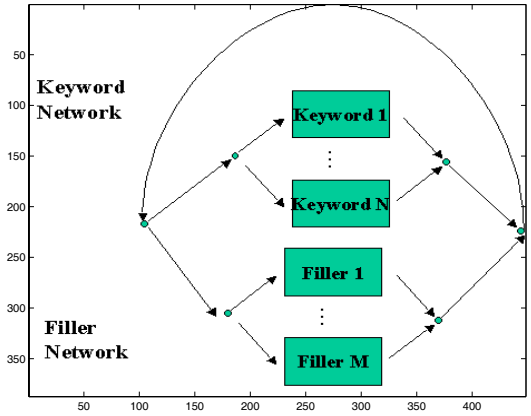

Fig. 3. The grammar diagram of the word spotting system. This diagram contain N keywords models which are glued by M filler models.

move fifth amphibious brigade five miles to the north.
deploy first motorized infantry seven miles to the southwest.
put first tank division forty ten miles to the bank of the river.
... ...

Table 1. typical utterances of the database

lighting conditions, image background are quite uniform in the dataset, thus simplifying the visual frontend processing. In addition, high quality wideband audio is synchronously collected at 16kHz sampling rate and with a signal-to-noise ratio (SNR) of 25dB. The database consists of 100 subjects, uttering scripts generated by a simple grammar on a 100-word vocabulary. There are 200 utterances for each subject, the duration of the entire database is approximately 50 hours. Table 1 lists several typical utterances in the database.

4. EXPERIMENTAL RESULTS

The word spotting algorithm has been evaluated on our audio-visual database. Half of the data are used for training, and the other half are used for evaluation. The keywords are listed in Table 2. The noisy speech is generated by artificially adding white noise at various SNRs. The experimental results of audio-only, visual-only and audio-visual systems are presented in Figure 5. In evaluation of different algorithms, we use Figure of Merit (FOM) as the performance measurement. FOM is the average detection rate of keywords under the false alarm rate of $10fa/kw/hr$.

Figure 5 shows that the audio-visual fusion significantly

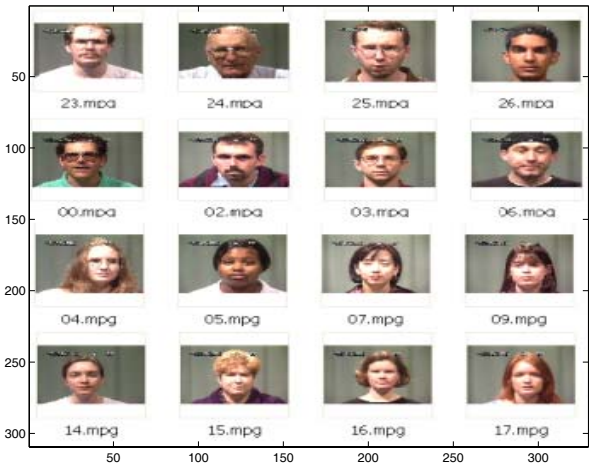

Fig. 4. Exemplar images in the database.

amphibious, brigade, tank, division, platoon, motorized, infantry, gunfire, liaison, company, communications, group, artillery, regiment, airborne, squadron, squad, armored, cavalry

Table 2. Examples of keywords

improves the noise robustness of the word spotter system. For example, the improvement is about 23% at a SNR of $15dB$. However, under extreme low SNRs, such as $0dB$, the audio cues actually degrade the performance of the visual-only system. The reason of this phenomenon is under extreme high noise, the audio stream does not provide useful but harmful cues. For detailed explanations, please refer to section V of [3].

5. CONCLUSIONS AND FUTURE WORK

In conclusion, we have integrated a face/mouth corner detection algorithm into an audio-based word spotting system. We have tested the system on a database of 100 subjects and the entire 50 hours video audio data. By comparing the joint audio-visual approach with audio-only word-spotter,

SNR(dB)	25	20	15	10	5	0
A HMM	77.7	69.7	53.0	36.4	17.2	6.0
V HMM	25.9	25.9	25.9	25.9	25.9	25.9
AV HMM	78.0	72.6	64.7	46.5	30.3	21.2

Table 3. A comparison of Audio-only(A HMM), visual-only(V HMM) and Audio-Visual(AV HMM) system at different acoustic SNRs in term of FOM, average detection rate below the false alarm rate 10 fa/kw/hr.

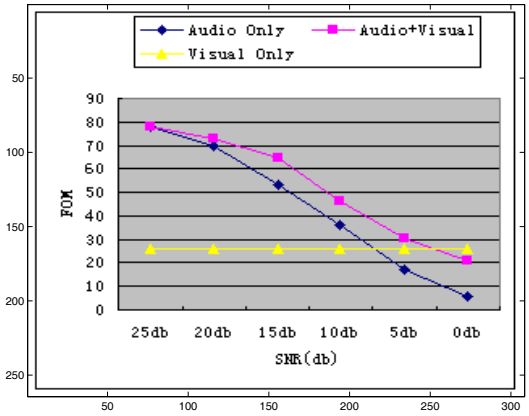

Fig. 5. FOM curve of the audio-only, visual only and audio-visual systems under different SNRs. X axis: Signal Noise Ratio. Y axis: FOM, average detection rate below the false alarm rate 10 fa/kw/hr.

we have shown that even a simple fusion technique such as feature-level fusion has been promising.

In the future, we will investigate more sophisticated fusion strategy such as the model-level fusion and other visual frontend feature extraction methods to further improve the performance of the audio-visual word spotting system.

6. REFERENCES

[1] C. Neti, G. Potamianos, J. Luettin, I. Matthews, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio visual speech recognition," in *Final Workshop 2000 Report*, 2000.

[2] A. V. Nefian, L. Liang, X. Pi, X. Liu, C. Mao, and K. Murphy, "A coupled hmm for audio-visual speech recognition," in *Proceedings of the IEEE Int. Conf. Acoust.,Speech, Signal Processing*, 2002.

[3] G. Potamianos, C. Neti, G. Gravier, and A. Garg, "Recent advances in the automatic recognition of audio-visual speech," *Proceedings of the IEEE*, vol. 91, no. 9, Sep. 2003.

[4] A. Colmenarez and T. S. Huang, "Face detection with information-based maximum discrimination," *Conference on Computer Vision and Pattern Recognition*, 1997.

[5] A. Colmenarez and T. S. Huang, "Maximum likelihood face detection," in *Proceedings International Conference on Automatic Face and Gesture Recognition*, Killington, VT, Oct. 1996, pp. 139–152.