Improving Speech Related Facial Action Unit Recognition by Audiovisual Information Fusion

Zibo Meng[1] and Yan Tong[1]
[1] Department of Computer Science and Engineering, University of South Carolina, Columbia, USA

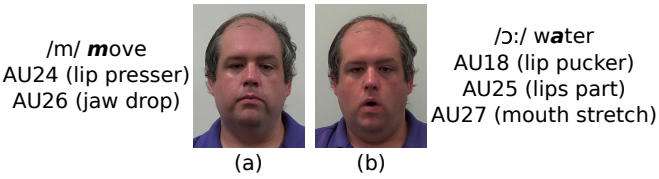

Fig. 1. Examples of speech-related facial activities, where different AUs are activated non-additively to pronounce speech. (a) The gap between teeth is occluded by the pressed lips in a combination of AU24 and AU26 when sounding /m/ and (b) the space between teeth is partially visible due to the protruded lips in a combination of AU18, AU25, and AU27 when producing /ɔ:/.

I. Introduction

Facial behavior is the most powerful and natural means of expressing the affective and emotional states of human being [12]. The Facial Action Coding System (FACS) developed by Ekman and Friesen [4] is a comprehensive and widely used system for facial behavior analysis, where a set of facial *action units* (AUs) are defined. According to the FACS [5], each facial AU is anatomically related to the contraction of a specific set of facial muscles, and combinations of AUs can describe rich and complex facial behaviors. Besides the applications in human behavior analysis, an automatic system for facial AU recognition has emerging applications in advancing human-computer interaction (HCI) such as interactive games, computer-based learning, and entertainment. Extensive research efforts have been focused on recognizing facial AUs from static images or image sequences as discussed in the survey papers [13], [20], [15].

In spite of progress achieved on posed facial display and controlled image acquisition, recognition performance degrades significantly for spontaneous facial displays [17], [16].

Furthermore, recognizing AUs that are responsible for producing speech is extremely challenging, since they are generally activated at a low intensity with subtle facial appearance/geometrical changes during speech and, more importantly, often introduce ambiguity in detecting other co-occurring AUs [5], e.g., producing non-additive appearance changes. For instance, as illustrated in Fig. 1(a), recognizing AU26 (jaw drop) from a combination of AU24 (lip presser) + AU26, when voicing a */m/*, is almost impossible from visual observations. The reason is that the gap between teeth, which is the major facial appearance clue to recognize AU26 [5], is small and invisible due to the occlusion by the pressed lips. In another example, when producing */ɔ:/*, as shown in Fig. 1(b), AU27 (mouth stretch) would probably be recognized as AU26 because the lips are protruded due to the activation of AU18 (lip pucker), which makes the opening of mouth smaller than that when only AU27 is activated. The failure in recognition of speech-related AUs is because we extract information from a single source, i.e., the visual channel, in the current practice. As a result, all speech-related AUs are represented by a uniform code [5], [17], i.e., AD 50, or totally ignored [16], during speech. However, identifying and differentiating the speech-related AUs from the others that express emotion and intention is critical to emotion recognition, especially during emotional speech.

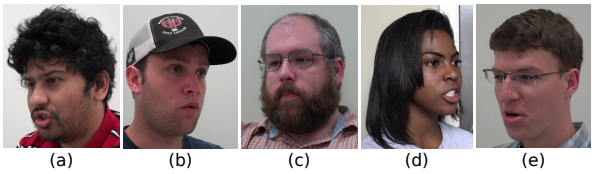

Fig. 3. Example images in the challenging subset collected from different illuminations, varying view angles, and with occlusions by glasses, caps, or facial hairs.

Instead of solely improving visual observations of AUs, we proposed to explore and exploit the information from both audio and visual channels for AU recognition. Specifically, facial AUs and voice are highly correlated in two ways. First, voice/speech has strong physiological relationships with some lower-face AUs such as AU24, AU26, and AU27, because jaw and lower-face muscular movements are the major mechanisms to produce differing sounds. These relationships are well recognized and have been exploited in natural human communications. For example, without looking at the face, people will know that the other person is opening his/her mouth by hearing ah. Following the example of recognizing AU26 from a combination of AU24 and AU26 as illustrated in Fig. 1(a), people can easily guess both AU24 and AU26 are activated because of a sound /m/, although AU26 is invisible from the visual channel. Second, both facial AUs and voice/speech convey human emotions in human communications. Since the second type of relationships is emotion and context dependent, we will focus on studying the physiological relationships between AUs and speech, which are more objective and will generalize better to various contexts.

II. Summary of the Work Till Date

A. Audiovisual AU-coded Dataset

As far as we know, the current publicly available AU-coded databases only provide information in visual channel. Furthermore, all the speech-related AUs have been either

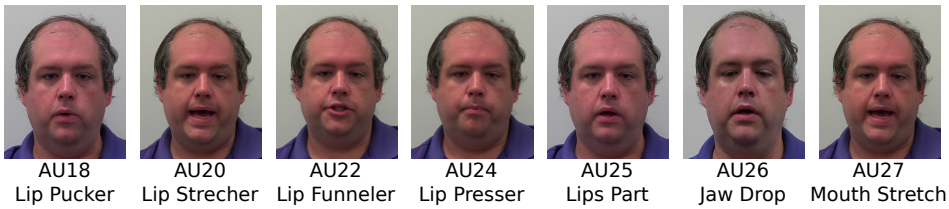

Fig. 2. A list of speech related AUs and their interpretations included in the audiovisual database.

annotated by a uniform label, i.e., AD 50 [17] or not labeled [16], during speech. In order to learn the semantic and dynamic physiological relationships between AUs and phonemes, as well as to demonstrate the proposed audiovisual AU recognition framework, we have constructed a pilot AU-coded audiovisual database consisting of two subsets, i.e. a clean subset, and a challenging subset. Fig. 2 illustrates example images of the speech-related AUs in the audiovisual database.

There are a total of 13 subjects in the audiovisual database, where 2 subjects appear in both the clean and challenging subsets. All the videos in this database were recorded at 59.94 frames per second at a spatial resolution of 1920×1080 with a bit-depth of 8 bits; and the audio signals were recorded at 48kHz with 16 bits.

In the clean subset, videos were collected from 9 subjects covering different races, ages, and genders. It consists of 12 words [1], which contain 28 phonemes and the most representative relationships between AUs and phonemes. Each subject was asked to speak the selected 12 words individually, each of which will be repeated 5 times. In addition, all subjects were required to keep a neutral face during data collection to ensure all the facial activities are only caused by speech.

Videos in the challenging subset were collected from 6 subjects covering different races and genders speaking the same words for 5 times as those in the clean set. As illustrated in Fig. 3, the subjects were free to display any expressions on their face during speech and were not necessary to show neutral face before and after speaking the word. In addition, instead of being recorded from the frontal view, videos were collected mostly from the sideviews with free head movements and occlusions by glasses, caps, and facial hairs, introducing challenges to AU recognition from the visual channel.

Groundtruth phoneme segments and AU labels were recorded in the database. Specifically, the utterances were transcribed using the Penn Phonetics Lab Forced Aligner (p2fa) [19], which takes an audio file along with its corresponding transcript file as input and produces a Praat [2] TextGrid file containing the phoneme segments. 7 speech-related AUs, i.e. AU18, AU20, AU22, AU24, AU25, AU26, and AU27, as shown in Fig. 2, were frame-by-frame labeled manually by two certified FACS coders.

[1]The 12 words including "beige","chaps","cowboy","Eurasian", "gooey","hue","joined","more","patch","queen", "she", and "waters" were selected from English phonetic pangrams (`http://www.liquisearch.com/list_of_pangrams/english_phonetic_pangrams`) that consists of all the phonemes at least once in 53 words.

B. Feature-level fusion for facial AU recognition

First, we propose to directly employ information from the visual and the audio channels by integrating the features extracted from the two channels. Figure 4 illustrates the proposed audiovisual feature-level fusion framework for facial AU recognition. Given a video, visual features and acoustic features are extracted from the images and the audio signal, respectively. To deal with the difference in time scales as well as the time shift between the two signals, the audio features need to be aligned with the visual features such that the two types of features are extracted at the same point in time. Then, the aligned audio and visual features are integrated and used to train a classifier for each target AU.

In order to demonstrate the effectiveness of using audio information in facial AU recognition, two different types of visual features are employed, based on which two feature-level fusion methods are proposed. The first method is based on a kind of human-crafted visual feature. Then, the audio and visual features are directly concatenated to form a single feature vector, which is used to train a classifier for each target AU. The other method employs visual features learned by a deep convolutional neural network (CNN). Then the audio and visual features are integrated into a CNN framework.

Experimental results on the audiovisual AU-coded dataset have demonstrated that both fusion methods outperform their visual counterparts in recognizing speech-related AUs. The improvement is more impressive with occlusions on the facial images, which would not affect the audio channel.

C. Audio-based facial AU recognition

Second, we proposed a novel approach to recognize speech-related AUs from speech by modeling and exploiting the dynamic and physiological relationships between AUs and phonemes through a Continuous Time Bayesian Network (CTBN) [11]. CTBNs are probabilistic graphical models proposed by Nodelman [11] to explicitly model the temporal evolutions over continuous time. CTBNs have been spotted in different applications, including users' presence and activities modeling [10], robot monitoring [9], sensor networks modeling [8], object tracking [14], host level network intrusion detection [18], dynamic system reliability modeling [3], social network dynamics learning [6], cardiogenic heart failure diagnosis and prediction [7], and gene network reconstruction [1].

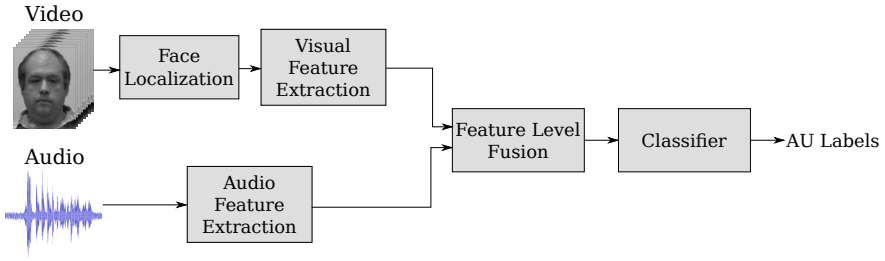

Fig. 4. The flowchart of the proposed feature-level fusion framework for bimodal facial AU recognition.

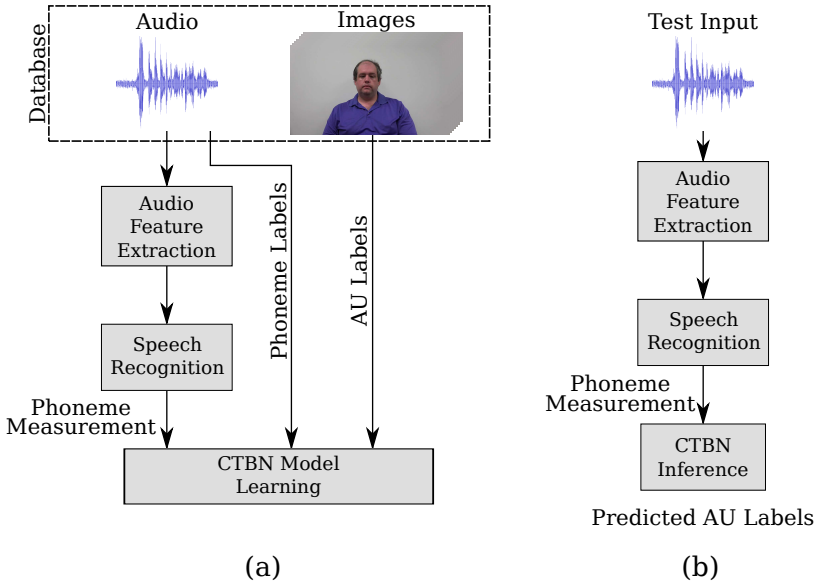

Fig. 5. The flowchart of the proposed audio-based AU recognition system: (a) an offline training process for CTBN model learning and (b) an online AU recognition process via probabilistic inference.

By considering AUs and phonemes as dynamic events, the CTBN model aims to explicitly characterize the relationships between AUs and phonemes, and more importantly, to model the temporal evolution of the relationships as a stochastic process over continuous time. Fig. 5 illustrates the proposed audio-based AU recognition system. During the training process (Fig. 5(a)), ground truth labels of AUs and phonemes are employed to learn the relationships between AUs and phonemes in a CTBN model. Furthermore, this model should also account for the uncertainty in speech recognition. For online AU recognition, as shown in Fig. 5(b), measurements of phonemes are obtained by automatic speech recognition and employed as evidence by the CTBN model; then AU recognition is performed by probabilistic inference over the CTBN model.

Experimental results on the AU-coded audiovisual dataset show that the proposed CTBN model achieves promising recognition performance for 7 speech-related AUs and outperforms the state-of-the-art visual-based methods especially for those AUs that are activated at low intensities or "invisible" in the visual channel. Furthermore, the CTBN model yields more impressive recognition performance on the challenging subset, where the visual-based approaches suffer significantly.

D. Audiovisual facial AU recognition via DBN

Third, Fig. 6 depicts the flowchart of the proposed audiovisual AU recognition system. During the training stage (6(a)), given a set of videos, ground truth labels of AUs and phonemes are employed to learn the relationships between AUs and phonemes in a DBN model. In addition, this model should represent the uncertainties in both AU and speech recognition. For online AU recognition, as illustrated in 6(b), given a video, visual-based AU recognition is performed to obtain the measurements of AUs; and speech recognition is conducted to obtain the measurements of phonemes. Then, all measurements are fed into a DBN model, which captures the physiological relationships between AUs and phonemes as well as measurement uncertainty. AU recognition is performed by audiovisual information fusion via DBN inference.

Experiments on the AU-coded audiovisual dataset have demonstrated that the proposed audiovisual fusion framework yields significant improvement in recognizing speech-related AUs compared to the state-of-the-art visual-based methods and the feature-level fusion method. In addition, the proposed framework achieves consistent improvements for speech-related facial AUs based on two state-of-the-art

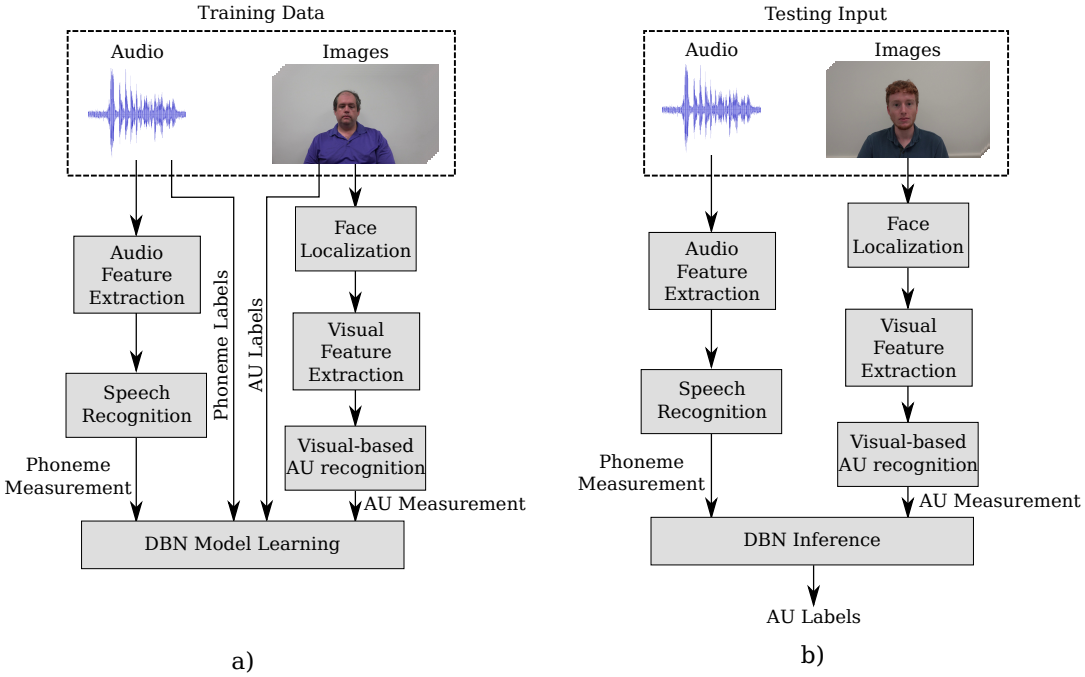

Fig. 6. The flowchart of the proposed audiovisual AU recognition system: (a) an offline training process for DBN model learning and (b) an online AU recognition process via probabilistic inference over the DBN model.

visual based methods. Furthermore, drastic improvement has been achieved for those AUs, whose visual observations are impaired during speech.

III. Future Work and Challenges

We plan to use CTBN to model the semantic and dynamic relationships between AUs and phonemes, and utilize information from both visual and audio channels to recognize speech related facial AUs. Since CTBN does not work well with discrete measurements, we need to figure out how to integrate the measurements into CTBN to produce facial AU recognition results over continuous time.

IV. Acknowledgement

This work is supported by National Science Foundation under CAREER Award IIS-1149787.

References

[1] E. Acerbi and F. Stella. Continuous time bayesian networks for gene network reconstruction: a comparative study on time course data. In *Bioinformatics Research and Applications*, pages 176–187. Springer, 2014.

[2] P. Boersma and D. Weenink. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[3] H. Boudali and J. B. Dugan. A continuous-time bayesian network reliability modeling, and analysis framework. *Reliability, IEEE Transactions on*, 55(1):86–97, 2006.

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

[5] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System: the Manual*. Research Nexus, Div., Network Information Research Corp., Salt Lake City, UT, 2002.

[6] Y. Fan and C. R. Shelton. Learning continuous-time social network dynamics. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 161–168. AUAI Press, 2009.

[7] E. Gatti, D. Luciani, and F. Stella. A continuous time bayesian network model for cardiogenic heart failure. *Flexible Services and Manufacturing Journal*, 24(4):496–515, 2012.

[8] R. Herbrich, T. Graepel, and B. Murphy. Structure from failure. In *Proceedings of the 2nd USENIX workshop on Tackling computer systems problems with machine learning techniques*, pages 1–6, 2007.

[9] B. Ng, A. Pfeffer, and R. Dearden. Continuous time particle filtering. In *IJCAI*, volume 19, page 1360, 2005.

[10] U. Nodelman, C. Shelton, and D. Koller. Learning continuous time Bayesian networks. In *UAI*, pages 451–458, 2003.

[11] U. Nodelman, C. R. Shelton, and D. Koller. Continuous time bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in Artificial Intelligence*, pages 378–387. Morgan Kaufmann Publishers Inc., 2002.

[12] M. Pantic and M. S. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, Vienna, Austria, 2007.

[13] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang. Human computing and machine understanding of human behavior: A survey. In T. S. Huang, A. Nijholt, M. Pantic, and A. Pentland, editors, *Artificial Intelligence for Human Computing*, LNAI. Springer Verlag, London, 2007.

[14] S. Qiao, C. Tang, H. Jin, T. Long, S. Dai, Y. Ku, and M. Chau. Putmode: prediction of uncertain trajectories in moving objects databases. *Applied Intelligence*, 33(3):370–386, 2010.

[15] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation and recognition. *IEEE T-PAMI*, 37(6):1113–1133, 2015.

[16] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn. FERA 2015 - second facial expression recognition and analysis challenge. *FG*, 2015.

[17] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE T-SMC-B*, 42(4):966–979, 2012.

[18] J. Xu and C. R. Shelton. Continuous time bayesian networks for host level network intrusion detection. In *Machine learning and knowledge discovery in databases*, pages 613–627. Springer, 2008.

[19] J. Yuan and M. Liberman. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878, 2008.

[20] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE T-PAMI*, 31(1):39–58, Jan. 2009.