2018

Improving Speech-related Facial Action Unit Recognition by Audiovisual Information Fusion

Zibo Meng
University of South Carolina

Improving Speech-related Facial Action Unit Recognition by Audiovisual Information Fusion

by

Zibo Meng

Bachelor of Engineering
Changchun University of Science and Technology 2009

Master of Engineering
Zhejiang University 2013

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Computer Science and Engineering

College of Engineering and Computing

University of South Carolina

2018

Accepted by:

Yan Tong, Major Professor

Srihari Nelakuditi, Chair, Examining Committee

Michael Huhns, Committee Member

Song Wang, Committee Member

XiaoFeng Wang, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

I am eternally grateful to my adviser, Prof. Yan Tong, for her guidance, patience, and encouragement during my graduate studies. She was always there to give advice on my research with great patience. She gave me enough time and space to implement my ideas and helped to build up my confidence when I was frustrated with my experiments. She is the best adviser I can ever ask for.

I would like to express my sincere gratitude to my dissertation committee members, Prof. Michael Huhns, Prof. Srihari Nelakuditi, Prof. Song Wang, and Prof. Xiaofeng Wang, for their constructive and valuable advice on my work. It is a great honor to have them served on my committee.

I want to thank my colleagues including: Ping Liu, Shizhong Han, James O'Reilly, Shehab Khan, Jie Cai, Zhiyuan Li, Xiaochuan Fan, Yuewei Lin, Youjie Zhou, Dazhou Guo, Kang Zheng, Hongkai Yu, Yang Mi, Haozhou Yu, Hao Guo, Jun Chen, Jing Wang, Yuhang Lu, Xiangyu Hu, and other members in our research group for their help and fruitful discussions.

I also want to thank Hanhan Xu for being such a great room-mate and Cookie, a four-year-old golden retriever, for bringing me a lot of joy.

Last but not the least, my special thanks go to my mother for her unconditional love and support when I am pursuing my Ph.D degree at University of South Carolina.

This research was supported by National Science Foundation under CAREER Award IIS-1149787.

Abstract

In spite of great progress achieved on posed facial display and controlled image acquisition, performance of facial action unit (AU) recognition degrades significantly for spontaneous facial displays. Furthermore, recognizing AUs accompanied with speech is even more 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, e.g., producing non-additive appearance changes. All the current AU recognition systems utilized information extracted only from visual channel. However, sound is highly correlated with visual channel in human communications. Thus, we propose to exploit both audio and visual information for AU recognition.

Specifically, a feature-level fusion method combining both audio and visual features is first introduced. Specifically, features are independently extracted from visual and audio channels. The extracted features are aligned to handle the difference in time scales and the time shift between the two signals. These temporally aligned features are integrated via feature-level fusion for AU recognition. Second, a novel approach that recognizes speech-related AUs exclusively from audio signals based on the fact that facial activities are highly correlated with voice during speech is developed. Specifically, dynamic and physiological relationships between AUs and phonemes are modeled through a continuous time Bayesian network (CTBN); then AU recognition is performed by probabilistic inference via the CTBN model. Third, a novel audiovisual fusion framework, which aims to make the best use of visual and acoustic cues in recognizing speech-related facial AUs is developed. In particular, a

dynamic Bayesian network (DBN) is employed to explicitly model the semantic and dynamic physiological relationships between AUs and phonemes as well as measurement uncertainty. AU recognition is then conducted by probabilistic inference via the DBN model.

To evaluate the proposed approaches, a pilot AU-coded audiovisual database was collected. Experiments on this dataset have demonstrated that the proposed frameworks yield significant improvement in recognizing speech-related AUs compared to the state-of-the-art visual-based methods. Furthermore, more impressive improvement has been achieved for those AUs, whose visual observations are impaired during speech.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Figure 1.1 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:/.

Facial behavior is the most powerful and natural means of expressing the affective and emotional states of human being [75]. The Facial Action Coding System (FACS) developed by Ekman and Friesen [22] 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 [23], 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 facial AU recognition system 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 [76, 119, 86, 58].

In spite of progress achieved on posed facial display and controlled image acquisition, recognition performance degrades significantly for spontaneous facial displays [104, 102].

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 [23], e.g., producing non-additive appearance changes. For instance, as illustrated in Fig. 1.1(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 [23], is small and invisible due to the occlusion by the pressed lips. In another example, when producing */ɔː/*, as shown in Fig. 1.1(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 [23, 104], i.e., AD 50, or totally ignored [102], 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.

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. 1.1(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. Instead of solely improving visual observations of AUs, *this work aims to explore and exploit the relationships between facial activity and voice to recognize speech-related*

AUs. 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.

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. In particular, a feature-level audiovisual fusion framework, an audio-based AU recognition system employing a continuous time Bayesian network (CTBN) and an audiovisual AU recognition approach based on a dynamic Bayesian network are developed.

1.1 Related Work

As discussed in the survey papers [76, 119, 86, 58], existing approaches for facial AU recognition directly employed either spatial or temporal features extracted from only the visual channel, i.e. static images or videos, to capture the visual appearance or geometry changes caused by a specific AU or AU combinations.

Human-designed Facial Features

General purpose human-crafted features are widely employed for facial activity analysis. These features include magnitudes of multi-scale and multi-orientation Gabor wavelets extracted either from the whole face region or at a few fiducial points [99, 7, 125, 124, 107, 101, 100], Haar wavelet features [107] considering the intensity difference of adjacent regions, and Scale Invariant Feature Transform (SIFT) features [116] extracted at a set of keypoints that are invariant to uniform scaling and orientation. Histograms of features extracted from a predefined facial grid have been also employed such as histograms of Local Binary Patterns (LBPs) [91, 89, 103], Histograms of Oriented Gradients (HOG) [6], histograms of Local Phase Quantization (LPQ) features [45], and histograms of Local Gabor Binary Patterns (LGBP) [90, 102].

In addition, spatiotemporal extensions of the aforementioned 2D features, such as LBP-TOP [126], LGBP-TOP [4, 3], LPQ-TOP [45], HOG-TOP [18], and dynamic Haar-like features [111, 112], have been employed to capture the spatiotemporal facial appearance changes caused by AUs.

Facial Features Learned from Data

In addition to the human-crafted features, features can also be learned in a data-driven manner by sparse coding or deep learning. As an over-complete representation learned from given input, sparse coding [74] can capture a wide range of variations that are not targeted to a specific application and has achieved promising results in facial expression recognition [118, 114, 52, 57, 128]. By taking advantages of both sparse coding [74] and Nonnegative Matrix Factorization (NMF) [50], Nonnegative Sparse Coding (NNSC) [40] has been demonstrated to be effective in facial expression recognition [9, 127, 117, 56]. To become more adaptable to the real world that consists of combination of edges [24], deep learning has been employed for facial expression recognition including deep belief network based approaches [80, 82, 53, 55] and convolutional neural network (CNN) based approaches [29, 59, 83, 97, 54, 46, 98, 34, 43, 27, 110, 113, 21, 21, 36]. Most of these deep-learning based methods took the whole face region as input and learned the high-level representations through a set of processing layers.

All the aforementioned visual-based approaches extracted information solely from the visual channel, and thus are inevitably challenged by imperfect image/video acquisition due to pose variations, occlusions, and more importantly, by the non-additive effects as illustrated in Fig. 1.1 in recognizing speech-related AUs.

1.1.1 Audio-based Facial AU recognition

Most recently, facial activity recognition from the audio channel has been briefly studied in [51, 84, 62]. Lejan et al. [51] detected three facial activities, i.e. eyebrow movement, smiling, and head shaking, using acoustic information. Assuming that these facial activities are not correlated, different groups of low-level acoustic features are extracted for each facial activity, respectively. Ringeval et al. [84] utilized low-level acoustic feature sets, i.e. ComParE and GeMAPS, for predicting facial AUs for emotion recognition. Our early work [62] employed Mel-Frequency Cepstral Coefficients (MFCC) features extracted from the audio channel for speech-related facial AU recognition. These methods only utilized low-level acoustic features without considering the semantic and dynamic relationships between facial activity and voice. As shown in our previous work [62], AU recognition using low-level acoustic features performed worse than the visual-based approaches for most of the speech-related AUs.

1.1.2 Audiovisual Information Fusion

The proposed framework takes advantage of information fusion of both visual and audio channels, and thus is also related to audiovisual information fusion, which has been successfully demonstrated in automatic speech recognition (ASR) [30, 47] and audiovisual affect/emotion recognition [119]. In the following, we will present a brief review on audiovisual affect/emotion recognition. There are three typical ways to perform audiovisual information fusion.

Feature-level fusion directly employs audio and visual features as a joint feature vector for affect/emotion recognition [105, 119]. Recently, deep learning has been employed for learning features from both visual and audio input [48, 21]. In our previous work [62], two feature-level fusion methods were developed for speech-related facial AU recognition. Specifically, one method combined LBP and MFCC features selected from AdaBoost independently; and the other one integrated visual features

learned by a CNN with MFCC features. However, these feature-level fusion methods often suffer from differences in time scales, metric levels, and noise levels in the two modalities [119].

Model-level fusion [93, 123, 31, 14, 88, 122, 68, 18] exploits correlation between audio and visual channels [119] and is usually performed in a probabilistic manner. For example, coupled [68], tripled [93] or multistream fused HMMs [123, 122] were developed by integrating multiple component HMMs, each of which corresponds to one modality, e.g., audio or visual, respectively. Fragpanagos et al. [31] and Caridakis et al. [14] used an ANN to perform fusion of different modalities. Sebe et al. [88] employed Bayesian network to recognize expressions from audio and facial activities. Chen et al. [18] employed Multiple Kernel Learning (MKL) to find an optimal combination of the features from two modalities. Most of the existing *feature-level fusion* or *model-level fusion* approaches utilize only the low-level features from each modality, e.g. prosody [123, 105, 68], MFCC [105, 68, 62] and formants [105]) for audio channel.

Decision-level fusion combines recognition results from two modalities assuming that audio and visual signals are conditionally independent of each other [121, 120, 119, 27, 110, 113, 21], while there are strong semantic and dynamic relationships between audio and visual channels.

In contrast to the major stream of visual-based facial AU recognition, we propose to utilize information extracted from audio channel for improving the performance for speech-related facial action unit recognition.

1.2 Scope of the Proposed Research

This research aims to improve the performance on facial AU recognition especially when they are accompanied by speech by developing: 1) a feature-level fusion approach; 2) an audio-based method, and 3) a decision-level fusion system for audiovisual facial AU recognition.

First, a feature-level fusion strategy is employed to illustrate the effectiveness of integrating information extracted from both audio and visual channels for speech-related facial AU recognition. Specifically, two feature-level fusion methods are proposed, which used local binary patterns (LBPs) and features learned by a deep convolutional neural network (CNN), respectively. For both methods, features are independently extracted from visual and audio channels. These features are aligned to handle the difference in time scales and the time shift between the two signals and integrated into a joint feature vector for AU recognition. Experimental results on a new 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.

Second, a novel approach that recognizes speech-related AUs exclusively from audio signals during speech is presented. Specifically, dynamic and physiological relationships between AUs and phonemes are analyzed, and modeled through a continuous time Bayesian network (CTBN); then AU recognition is performed by probabilistic inference via the CTBN model given only measurements produced from audio channel. Experimental results on this database show that the proposed CTBN model achieves promising recognition performance for 7 speech-related AUs and outperforms both the state-of-the-art visual-based and audio-based methods especially for those AUs that are activated at low intensities or "hardly visible" in the visual channel. The improvement is more impressive on the challenging subset, where the visual-based approaches suffer significantly.

Third, a novel audiovisual fusion framework, which aims to make the best use of visual and acoustic cues in recognizing speech-related facial AUs is developed. In particular, a dynamic Bayesian network (DBN) is employed to explicitly model the semantic and dynamic physiological relationships between AUs and phonemes as

Figure 1.2 A list of speech related AUs and their interpretations included in the audiovisual database.

well as measurement uncertainty. AU recognition is then conducted by probabilistic inference via the DBN model. Experiments on this database have demonstrated that the proposed framework yields significant improvement in recognizing speech-related AUs compared to the state-of-the-art visual-based methods especially for those AUs whose visual observations are impaired during speech, and more importantly also outperforms audio-based methods as well as feature-level fusion methods by explicitly modeling and exploiting physiological relationships between AUs and phonemes.

1.3 Audiovisual AU-coded Dataset

To the best of our knowledge, the current publicly available AU-coded databases only provide information in visual channel. Furthermore, all the speech-related AUs have been either annotated by a uniform label, i.e., AD 50 [104] or not labeled [102], 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. 1.2 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. The statistics, i.e., the numbers of occurrences, of the speech-related AUs in the clean and challenging subsets are reported in Table. 1.1.

Table 1.1 Statistics of the speech-related AUs in the audiovisual database.

Subsets	**AU18**	**AU20**	**AU22**	**AU24**	**AU25**	**AU26**	**AU27**	**Total Frames**
Clean	7,014	1,375	4,275	2,105	25,092	18,280	4,444	34,622
Challenging	4,118	1,230	3,396	1,373	17,554	11,830	3,242	23,274

In the clean subset, videos were collected from 9 subjects covering different races, ages, and genders. It consists of 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. The selected 12 words 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. 1.3, 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

Figure 1.3 Example images in the challenging subset collected from different illuminations, varying view angles, and with occlusions by glasses, caps, or facial hairs.

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) [115], which takes an audio file along with its corresponding transcript file as input and produces a Praat [10] TextGrid file containing the phoneme segments. 7 speech-related AUs, i.e. AU18, AU20, AU22, AU24, AU25, AU26, and AU27, as shown in Fig. 1.2, were frame-by-frame labeled manually by two certified FACS coders. Roughly 10% of the data was labeled by both coders independently to estimate inter-coder reliability measured by Matthews Correlation Coefficient (MCC) [78]. As illustrated in Fig. 1.2, the MCC for each AU ranges from 0.69 for AU27 to 0.98 for AU25 and has an average of 0.88 on the clean subset, and ranges from 0.80 for AU26 to 0.96 for AU25 on the challenging subset, which indicates strong to very strong inter-coder reliability of AU annotation.

Table 1.2 Inter-coder reliability measured by MCC for the 7 speech-related facial AUs on the audiovisual database.

Subsets	AU18	AU20	AU22	AU24	AU25	AU26	AU27	Total Frames
Clean	0.945	0.930	0.864	0.944	0.985	0.791	0.695	0.879
Challenging	0.941	0.917	0.930	0.824	0.963	0.799	0.842	0.888

1.4 Structure of the Dissertation

This dissertation is organized as follows. Chapter 2.4 presents a novel feature level fusion approach utilizing features extracted from audio and visual channels for speech-related facial AU recognition. Chapter 3.4 introduces a novel audio-based system, where the physiological and dynamic relationships between facial AUs and audio are explicitly modeled by a continuous time Bayesian network (CTBN). Facial AU recognition is performed via probabilistic inference over the CTBN given the speech recognition results as measurements. Chapter 4.5 describes a comprehensive audio-visual speech-related facial AU recognition framework, where a dynamic Bayesian network (DBN) is employed to model the semantic and dynamic relationships between speech and facial AUs. The predictions can be obtained over the DBN model given both audio and visual measurements. The conclusion and future research are discussed in Chapter 5.

Chapter 2

Audiovisual Facial Action Unit Recognition using Feature Level Fusion

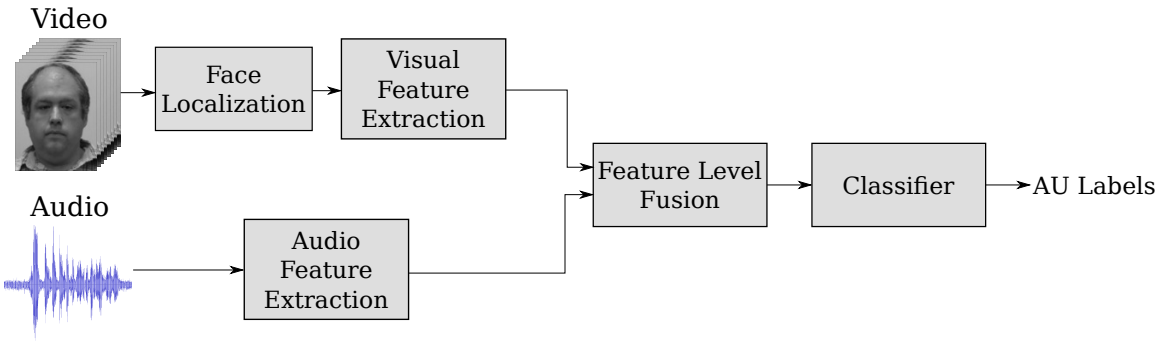

Figure 2.1 The flowchart of the proposed feature-level fusion framework for bimodal facial AU recognition.

2.1 Introduction

All existing approaches on facial AU recognition extract information solely from the visual channel. In contrast, this paper proposes a novel approach, which exploits the information from both visual and audio channels, to recognize speech-related AUs. This work is motivated by the fact that facial AUs and voice are highly correlated in natural human communications. Specifically, voice/speech has strong physiological relationships with some lower face AUs such as AU25 (lips part), AU26 (jaw drop), and AU24 (lip presser) because jaw and lower-face muscle movements together with the soft palate, tongue and vocal cords produce the voice.

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 when hearing laughter. Following the example of recognizing AU26 (jaw drop) in the Stop phase of pronouncing the phoneme /b/, we can infer that AU26 (jaw drop) has been activated when hearing the sound /b/, even when it is "invisible" in the visual channel.

Specifically, we propose to directly employ information from the visual and the audio channels by integrating the features extracted from the two channels. Figure 2.1 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.

This work falls into the category of feature-level audiovisual fusion by employing features extracted from the two channels. Different from the prior feature-level fusion approaches, which often suffer from differences in time scale [61], we propose a method to align the audio and visual features frame-to-frame such that the two types of features are extracted at the point in time.

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.

There are four major contributions in this work.

- To the best of our knowledge, it is the first utilization of both audio and visual features to recognize the speech-related facial AUs.
- Two feature-level fusion methods are proposed based on human-crafted visual features and a CNN, respectively.
- To facilitate feature-level fusion, we propose a method to align the audio and visual features.

- An AU-coded audiovisual database is constructed to evaluate the proposed feature-level fusion framework and can be employed as a benchmark database for AU recognition.

Experimental results on the new audiovisual AU-coded dataset have demonstrated that the proposed bimodal AU recognition framework achieved promising recognition performance. Specifically, both fusion methods outperform those only employing visual information in recognizing speech-related AUs. The improvement is more impressive when the face regions are occluded, which, however, would not affect the audio channel.

2.2 Methodology

Since speech is anatomically produced by a specific set of jaw and lower facial muscle movements, there are strong physiological relationships between the lower-face AUs and speech. Taking the word beige for instance, a combination of AU24 (lip presser) and AU26 (jaw drop) is first activated to produce the Stop phase of /b/ (Figure 2.2a). Then, AU25 (lips part) and AU26 are activated together to sound /b/ in its Aspiration phase and /ei/ (Figure 2.2b). Finally, AU22 (lip funneler) and AU25 are activated for sounding /ʒ/ (Figure 2.2c). Inspired by this, we propose to utilize the information from both visual and audio channels for recognizing speech-related facial AUs. In addition, signals in different channels are usually sampled at different time scales and are not synchronized perfectly. In this work, we show how to extract visual and audio features and how to align the features from each channel to perform the feature level fusion.

2.2.1 Audio Feature Extraction

In this work, 13-dimensional Mel Frequency Cepstral Coefficients (MFCCs) [20], which are widely used in speech recognition, are employed as the audio features.

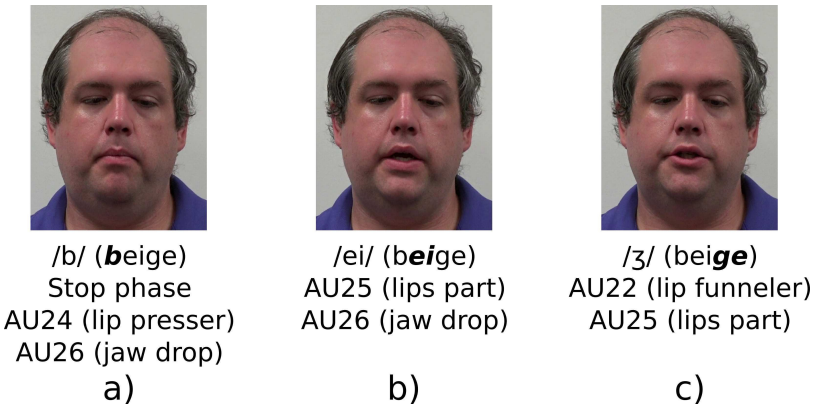

Figure 2.2 Examples of physiological relationships between speech and AUs. To pronounce a word "beige", different combinations of AUs are activated sequentially.

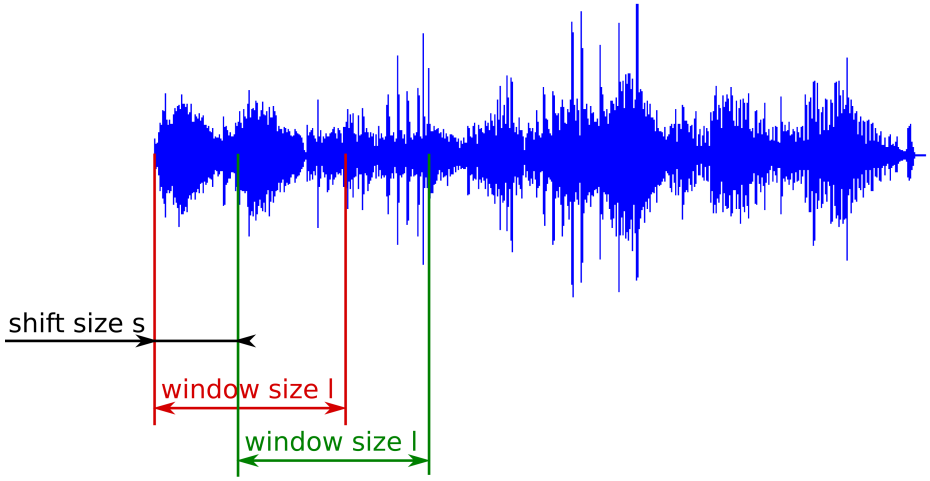

Figure 2.3 Illustration of audio feature extraction, where 13-dimensional MFCC features are obtained from a temporal window of size l and shift to the next window by a stride of size.

Specifically, given an input wave file, the size of the temporal window denoted by l, and a stride denoted by s, a state-of-the-art speech recognition method, i.e., Kaldi toolkit [77], is employed to obtain MFCC features. As illustrated in Figure 2.3, the Kaldi toolkit extracts the MFCC features within a temporal window with a size of l and shifts to the next window by a stride of s.

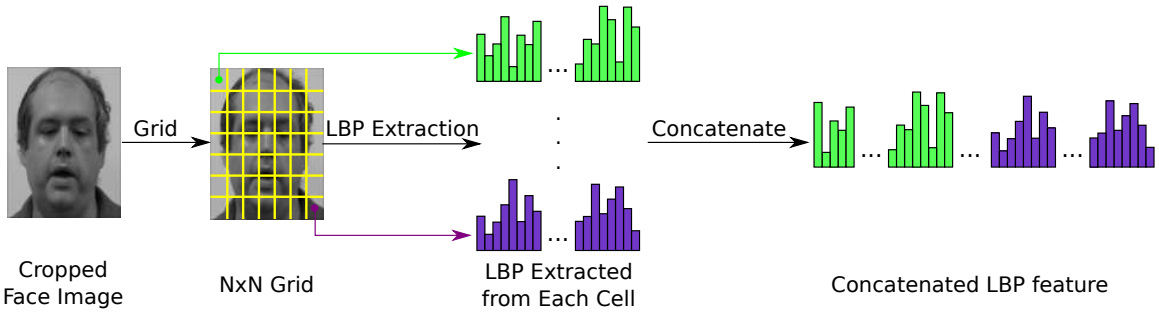

Figure 2.4 An illustration of extracting LBP features from a face image. The face image is divided into a grid, from each of which, LBP histograms are extracted. Then, LBP features are obtained by concatenating all the LBP histograms from each cell. Best viewed in color.

2.2.2 Visual Feature Extraction

In this work, two types of visual features are employed, including human-crafted features and features learned by deep learning.

LBP feature extraction

Among the human-crafted features, LBP features [73] are employed as the visual feature descriptor because of its good performance in facial expression/AU recognition [90, 37, 102]. As shown in Figure 2.4, the face region is divided into an $N \times N$ grid. From each cell, LBP features are extracted as follows:

$$\mathrm{LBP}(p) = \sum_{k=0}^{7} \phi(v_k - v_p) \cdot 2^k,$$

where

$$\phi(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

p is a pixel with an intensity of v_p, and v_k, $k = 0, \cdots, 7$, are the intensities of its eight neighboring pixels. Since only a subset of the LBPs, i.e. the uniform patterns containing at most two bitwise transitions from 0 to 1, are crucial for encoding the

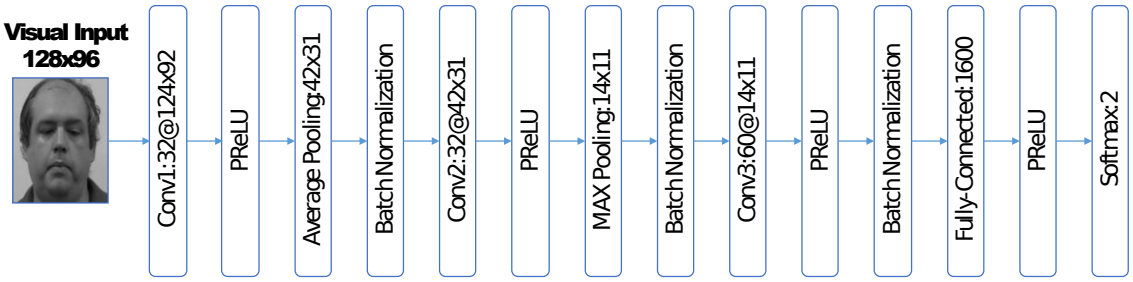

Figure 2.5 The architecture of the V-CNN used to learn the visual features. For each layer, the neuron number and the map dimension are given by the numbers before and after "@", respectively. 1600 neurons are employed in the fully connected layer.

texture of images, histograms of the 59 uniform patterns are calculated for each cell of the grid. Then, LBP histograms extracted from all cells are concatenated as LBP features. Figure 2.4 illustrates the process of extracting LBP features from an image.

Visual features learned by Deep Networks

Recently, CNNs have been demonstrated to be effective on various computer vision tasks [49, 26], as well as on audiovisual fusion [67, 41, 96]. In this work, a CNN, denoted as V-CNN, is developed to learn feature representations from the visual channel. As shown in Figure 6, the V-CNN consists of three convolutional layers followed by a fully-connected layer. After the first convolutional layer, there is an average-pooling layer; and a max-pooling layer is following the second convolutional layer. Following each convolutional layer, there is also a batch normalization layer [42], which normalizes each scalar feature in the training batch to zero mean and unit variance. The batch normalization has been shown to improve classification performance and accelerate the training process [42]. A Softmax layer is employed to generate the predictions and calculate the loss according to the groundtruth labels.

As shown in Figure 2.5, parametric rectified linear units (PReLU) [38] are employed after the convolutional layers and the fully-connected layer to produce nonlin-

earity for hidden neurons. As an extension of a rectified linear unit (ReLU) activation function, PReLU has better fitting capability than the sigmoid function or hyperbolic tangent function [49] and further boosts the classification performance compared to the traditional ReLU. The PReLU activation function is defined as [38]:

$$\text{PReLU}(\mathbf{y}^j) = \begin{cases} \mathbf{y}^j, & \text{if } \mathbf{y}^j > 0 \\ \alpha \mathbf{y}^j, & \text{if } \mathbf{y}^j \leq 0 \end{cases}$$

where $\mathbf{y}^j$ is the input of PReLU of the j^{th} layer, and α is a parameter used to control the slope when the input is negative, which is adaptively learned during the training process.

The output of the fully-connected layer is employed as the visual features learned by the CNN. In addition, the output of the V-CNN, i.e., a 2-way softmax, is used as a binary classifier predicting the probability distribution over 2 different classes, i.e. the "presence" or "absence" status of a target AU, which is used as a baseline visual-based method in our experiment.

2.2.3 Audiovisual Feature Alignment

The visual and audio features are usually extracted at different time scales. Furthermore, since the video clips in the audiovisual database are cut from long streaming videos, there is a random shift between the visual and audio signals, even if they have the same sampling rates. To perform feature level fusion, the time scale of audio features should be adjusted to that of visual features and more importantly, these two types of features should be extracted at the same time. Hence, an "alignment" process is needed and described as follows.

As depicted by Figure 2.6, given a sequence of MFCC features, $\mathbf{v} = (v_0, \cdots, v_n)$, and its corresponding time points, $\mathbf{t} = (t_0, \cdots, t_n)$ with n+1 points and n time

intervals, a cubic spline for each interval $[t_i, t_{i+1}]$ is estimated as follows:

$$S_i(t) = a_i(t - t_i) + b_i(t - t_i)^2 + c_i(t - t_i)^3 + d_i$$

where a_i, b_i, c_i, and d_i are coefficients to be estimated for the spline for the i^{th} interval. After estimating the splines for all the intervals, the MFCC values at each time point t'_j, where the j^{th} image frame is sampled, can be estimated by interpolation according to the corresponding cubic spline.

The audio features resulting from interpolation may contain errors due to imperfect alignment. Furthermore, information from the neighboring time frames may contain important information for AU recognition. For example, facial activities are usually activated slightly earlier than the sound is made. It is especially true for AU24 (lip presser), which is activated and relaxed before the sound /b/ is emitted. To address this issue, MFCC features from multiple frames are concatenated as the feature vector for the current frame.

2.2.4 Audiovisual Feature-Level Fusion

Audiovisual fusion based on LBP features

The extracted LBP features are concatenated with the aligned MFCC features into a unified feature vector, which is employed as input to train a classifier for each target AU.

Audiovisual fusion based on CNN

As depicted in Figure 2.7, a CNN, denoted as AV-CNN, is designed to perform the audiovisual fusion for facial AU recognition. In particular, the visual stream of the proposed AV-CNN has the same structure as V-CNN. The visual features, i.e. the output of the fully-connected layer in V-CNN, are combined with the aligned MFCC features as the input for a softmax layer, the output layer of the AV-CNN. The output

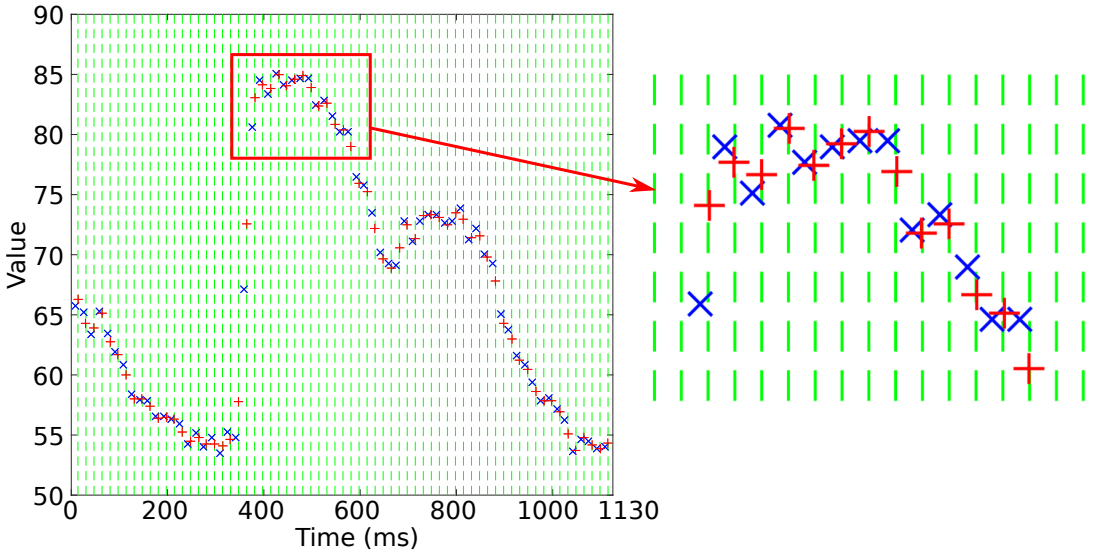

Figure 2.6 An illustration of aligning MFCCs to image frames. The left image gives a sequence of data and the right one shows the close up of a portion of the sequence, where the blue crosses represent the original values of MFCCs at their respective times; the green vertical dash lines give the time points of image frames; and the red crosses denote the aligned MFCC features.

of the AV-CNN is the probability of the "presence" or "absence" status of a target AU.

2.2.5 Implementation Details and Experimental Setup

MFCC feature extraction

In this work, the MFCC features are extracted with window size $l = 16.67ms$ using a time shift of $s = 16.67ms$. To include more temporal information, 7 frames, i.e. 3 frames before and after the current frame along with the current one are concatenated as the final MFCC feature for each frame.

LBP-based audiovisual fusion

For preprocessing purpose, the face regions across different facial images are aligned to remove the scale and positional variance based on eye positions using a state-of-

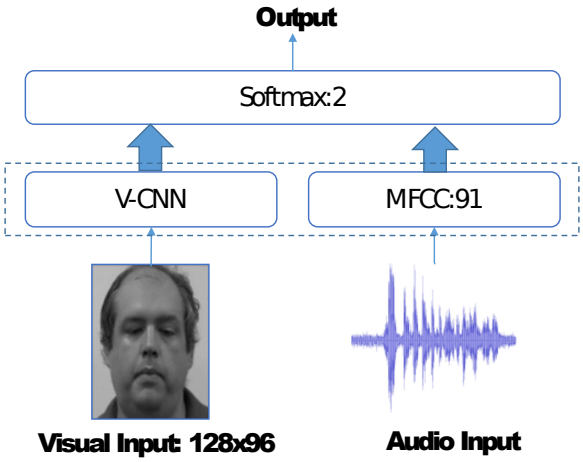

Figure 2.7 Architecture of a CNN used for audiovisual fusion, where the fully-connected layer in V-CNN is combined with the 91 dimension MFCC feature as the input to a softmax layer, which is employed to predict the probability of the "presence" and "absence" status of a target AU.

the-art facial landmark detection method [5] and then cropped to a size of 96×64. Following Han et al. [37], each of the face images is divided into a 7×7 grid, from each of which, LBP histograms with 59 bins are extracted. All extracted LBP histograms are then concatenated as LBP features.

To handle the difference in metrics, a normalization process is required to ensure that the LBP features are within the same range as the MFCC features. Hence, both features are normalized to the same interval [0, 1]. The normalized MFCC and LBP features are concatenated into a uniform feature vector, from which an AdaBoost classifier is employed in a supervised manner to select the most discriminative features, i.e., a set of weak classifiers, based on the classification errors and to construct a strong classifier to perform AU recognition.

CNN-based audiovisual fusion

For preprocessing purposes, the face regions are aligned and cropped to a size of 128×96. The V-CNN model is trained from a CNN model pretrained on the FERA 2015

database [102] using stochastic gradient decent with a batch size of 128, momentum of 0.9, and a weight decay parameter of 0.005. A base learning rate of 5e-4 is employed at the beginning of the training process and decreased by a factor of 0.5 every 500 iterations. The kernel size for average pooling layer and the max pooling layer is with a stride of 3. Dropout is applied to each fully-connected layer with a probability of 0.5, i.e. zeroing out the output of a neuron with probability of 0.5. The CNN models are implemented using the Caffe library [44].

2.2.6 Results on Audiovisual Dataset

To demonstrate the effectiveness of utilizing audio information in AU recognition, we compared the two proposed feature-level fusion methods, i.e., LBP-based fusion method, denoted as LBP-Fusion, and CNN-based fusion method, denoted as AV-CNN, with their visual-based counterparts, i.e., the LBP-based method [37] and the V-CNN, respectively. In addition, we reported the results using the information extracted from only audio channel denoted as Ada-MFCC and MFCC-CNN based on AdaBoost and CNN, respectively. For the methods compared, a leave-one-subject-out training/testing strategy is employed, where the data from 8 subjects is used for training and the remaining data is used for testing. The experimental results are computed as the average of 9 runs.

2.2.7 Experimental results

Comparison of LBP, Ada-MFCC, and LBP-Fusion: Quantitative experimental results based on the LBP features and the MFCC features are reported in Table 2.1 in terms of false alarm rate (FAR), true positive rate (TPR), and F1 score. The F1 score is defined as $F1 = \frac{2TP}{2TP + FP + FN}$, where TP is the number of positive samples that are recognized correctly, FP is the number of negative samples that are recognized as positive, and FN is the number of positive samples that are recognized as negative.

Table 2.1 Performance comparison of LBP, Ada-MFCC, and LBP-Fusion in terms of F1 score, TPR, and FAR.

AUs	***LBP***			***Ada-MFCC***			***LBP-Fusion***		
	F1	FAR	TPR	F1	FAR	TPR	F1	FAR	TPR
AU18	0.641	0.149	0.746	0.558	0.166	0.635	**0.679**	**0.125**	**0.768**
AU20	0.181	0.178	0.661	0.15	0.203	0.641	**0.221**	**0.15**	**0.7**
AU22	0.442	0.186	0.657	0.445	0.168	0.651	**0.493**	**0.166**	**0.703**
AU24	0.348	0.187	0.746	0.201	0.271	0.592	**0.375**	**0.163**	**0.755**
AU25	0.855	0.119	0.784	0.783	0.144	0.681	**0.886**	**0.089**	**0.825**
AU26	0.582	0.273	0.516	0.568	0.201	0.476	**0.624**	**0.239**	**0.556**
AU27	0.329	0.249	0.526	0.419	0.211	0.654	**0.455**	**0.186**	**0.669**
AVG	0.482	0.191	0.662	0.448	0.195	0.619	**0.533**	**0.16**	**0.711**

As shown in Table 1, the proposed LBP-Fusion method achieves promising recognition performance for the 7 speech-related AUs and outperforms both the visual-based method, ***LBP***, and the audio-based method, Ada-MFCC, in terms of the F1 score, FAR, and the TPR for all target AUs.

Compared to LBP and Ada-MFCC, which employs information only from the visual or the audio channel, the overall AU recognition performance is improved from 0.482 (LBP) and 0.448 (Ada-MFCC) to 0.533 (LBP-Fusion) in terms of the average F1 score, which demonstrates the effectiveness of using information from both the audio and visual channels. Compared to the LBP method, the performance improvement is more obvious for AU27 (mouth stretch) when using audio information: the F1 score is improved from 0.329 (LBP) to 0.419 (Ada-MFCC) and is further improved to 0.455 (LBP-Fusion) by integrating both audio and visual information. This is because the visual observation of AU27 is not reliable during speech due to the occlusion caused by lip movements, whereas the information from the audio channel plays an important role in detecting AU27.

Comparison of AV-CNN, MFCC-CNN, and V-CNN: Table 2 gives the experimental results using features learned by CNNs. The proposed AV-CNN outperforms both the V-CNN and the MFCC-CNN in terms of the average F1 score, average FAR,

Table 2.2 Performance comparison of V-CNN, MFCC-CNN, and AV-CNN in terms of F1 score, TPR, and FAR.

AUs	***V-CNN***			***MFCC-CNN***			***AV-CNN***		
	F1	FAR	TPR	F1	FAR	TPR	F1	FAR	TPR
AU18	0.517	0.09	0.581	0.582	0.117	0.506	**0.74**	**0.061**	**0.735**
AU20	0.162	**0.033**	0.19	0.125	0.036	0.201	**0.267**	0.034	**0.201**
AU22	0.364	0.081	0.391	0.514	0.072	0.481	**0.534**	**0.066**	**0.513**
AU24	**0.397**	**0.037**	**0.379**	0.111	0.058	0.201	0.319	0.038	0.341
AU25	**0.944**	**0.113**	0.961	0.829	0.435	0.828	0.943	0.115	**0.962**
AU26	0.692	0.309	0.787	0.625	0.385	0.63	**0.712**	**0.274**	**0.792**
AU27	0.221	0.106	0.264	0.467	0.078	0.472	**0.477**	**0.077**	**0.466**
AVG	0.471	0.11	0.507	0.465	0.169	0.474	**0.57**	**0.095**	**0.573**

and the average TPR. In addition, compared to V-CNN, the performance on AU27 gains a dramatic improvement using AV-CNN, i.e. from 0.221 by V-CNN to 0.447 by AV-CNN in terms of the average F1 score.

Comparison between LBP-Fusion and AV-CNN: As shown in Table 2.1 and Table 2.2, AV-CNN (0.570) outperforms LBP-Fusion (0.533) in terms of F1 score, because the feature representations learned by CNN can better capture the discriminative information in data than the hand-crafted features.

Experimental results on the data with occlusions The visual-based facial AU recognition is made more challenging with head movements and occlusions of the face region, e.g. moustache and beards, since the extracted features include the noise due to the misalignment of face regions and occlusions. However, the audio channel will not be affected by the aforementioned challenges in the visual channel. Hence, the information extracted from the audio signal is more robust to head movements and occlusions for facial AU recognition. To better demonstrate the effectiveness of the proposed audiovisual fusion methods, we randomly add occlusions to face images, as illustrated in Figure 10. Specifically, black blocks are randomly added to the mouth region of each image to synthesize occlusions.

From the images with occlusions, we retrained the visual-based LBP method de-

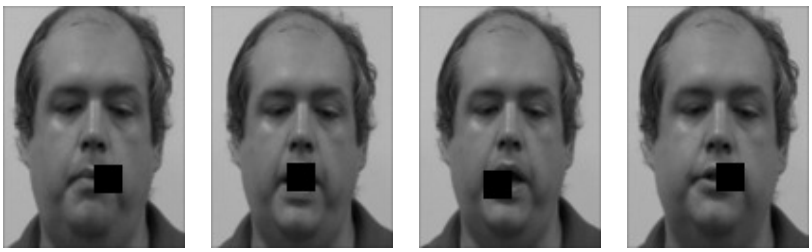

Figure 2.8 Example images of adding 15 by 15 pixel black blocks randomly to the mouth region in face images to synthesize occlusions.

noted as LBP-Occlusion and V-CNN-based method denoted as V-CNN-Occlusion as well as the fusion methods denoted as LBP-Fusion-Occlusion and AV-CNN-Occlusion, respectively. The quantitative experimental results on the images with occlusions are reported in Table 2.3 and Table 2.4 for LBP and CNN-based methods, respectively.

2.3 Experiments

Not surprisingly, the performance of the visual-based methods, especially the V-CNN-Occlusion, drops significantly when the images contain occlusions. In contrast, the performance of the proposed fusion methods, i.e., LBP-Fusion-Occlusion (0.516) and AV-CNN-Occlusion (0.506), are less affected by the occlusions, since the information from audio channel is more reliable when the images contain occlusions. Particularly, by employing audio information, the recognition performance on the occluded data is improved dramatically from 0.365 (V-CNN-Occlusion) to 0.506 (AV-CNN-Occlusion) in terms of the average F1 score.

In addition, the performance of V-CNN-Occlusion decreases more significantly than that of LBP-Occlusion. In the CNN, the decision is made by a fully-connected layer, where each output node is connected to every single node in the previous layer, and thus, will be affected by occlusions in any position. In contrast, the LBP features are less correlated, and thus, the recognition performance will not be affected significantly by the failure of one or several LBP features extracted in the occluded

Table 2.3 Performance comparison between LBP and LBP-Fusion on the data with occlusions in terms of F1 score, TPR, and FAR.

AUs	LBP-Occlusion			LBP-Fusion-Occlusion		
	F1	FAR	TPR	F1	FAR	TPR
AU18	0.63	0.157	0.744	**0.672**	**0.133**	**0.771**
AU20	0.207	0.163	0.701	**0.254**	**0.133**	**0.714**
AU22	0.423	0.201	0.643	**0.477**	**0.179**	**0.694**
AU24	0.305	0.183	0.656	**0.329**	**0.175**	**0.709**
AU25	0.805	0.151	0.72	**0.853**	**0.109**	**0.778**
AU26	0.514	0.288	0.442	**0.577**	**0.255**	**0.509**
AU27	0.34	0.244	0.544	**0.452**	**0.189**	**0.67**
AVG	0.461	0.198	0.636	**0.516**	**0.167**	**0.692**

Table 2.4 Performance comparison between V-CNN and AV-CNN on the data with occlusions in terms of F1 score, TPR, and FAR.

AUs	V-CNN-Occlusion			AV-CNN-Occlusion		
	F1	FAR	TPR	F1	FAR	TPR
AU18	0.465	0.093	0.579	**0.674**	**0.082**	**0.638**
AU20	0.091	0.036	0.085	**0.197**	**0.035**	**0.177**
AU22	0.196	0.097	0.338	**0.423**	**0.081**	**0.402**
AU24	0.107	0.054	0.101	**0.153**	**0.051**	**0.161**
AU25	0.876	0.174	0.911	**0.913**	**0.153**	**0.942**
AU26	0.666	0.303	0.753	**0.722**	**0.274**	**0.808**
AU27	0.161	0.11	0.221	**0.458**	**0.085**	**0.39**
AVG	0.365	0.124	0.427	**0.506**	**0.109**	**0.503**

region.

2.4 Conclusion

Recognizing speech-related AUs is challenging due to the subtle facial appearance/ geometrical changes and occlusions introduced by frequent lip movements. Motivated by the fact that facial activities are highly correlated with voice, in this section, a novel feature-level fusion framework employing information from both the audio channel and the visual channel is proposed. Specifically, two feature-level fusion methods were developed based on LBP features and features learned by a CNN. To

handle the differences in time scale and metrics, the audio and visual features are aligned frame-to-frame and normalized into the same range. Experimental results on a new audiovisual AU-coded dataset have demonstrated that both LBP-based and CNN-based feature-level fusion methods outperform the methods only using visual features, especially for those AUs whose visual observations are "invisible" during speech. The improvement is more impressive when evaluated on the image data containing occlusions.

Chapter 3

Listen to Your Face: Inferring Facial Action Units from Audio Channel

3.1 Motivation

Facial AU recognition from static images or videos has received an increasing interest during the past decades as elaborated in the survey papers [76, 119, 86]. In spite of progress on posed facial displays and controlled image acquisition, recognition performance degrades significantly for spontaneous facial displays with free head movements, occlusions, and various illumination conditions [103]. More importantly, it is extremely challenging when recognizing AUs involved in speech production, since these AUs are usually activated at low intensities with subtle facial appearance/geometrical changes and often introduce ambiguity in detecting other co-occurring AUs [23], i.e., non-additive effects of AUs in a combination. For example, pronouncing a phoneme /p/ has two consecutive phases, i.e., *Stop* and *Aspiration* phases. As shown in Fig. 3.1(b), the lips are apart and the oral cavity between the teeth is visible in the *Aspiration* phase, based on which AU25 (lips part) and AU26 (jaw drop) can be detected from the image. Whereas, during the *Stop* phase as shown in Fig. 3.1(a), the lips are pressed together due to the activation of AU24 (lip presser). Since the oral cavity is occluded by the lips, AU26 is difficult to be detected from the visual channel. In another example, the oral cavity is partially occluded by the lips when producing /ɔ:/ in Fig. 3.1(d) due to the activation of AU18 (lip pucker). Hence, even the mandible is pulled down significantly, it is difficult to detect AU27 (mouth stretch) from Fig. 3.1(d).

These facial activities actually can be "heard", i.e., inferred from the information extracted from the audio channel. 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. In addition, eyebrow movements and fundamental frequency of voice have been found to be correlated during speech [15]. As demonstrated by the McGurk effect [60], there is a strong correlation

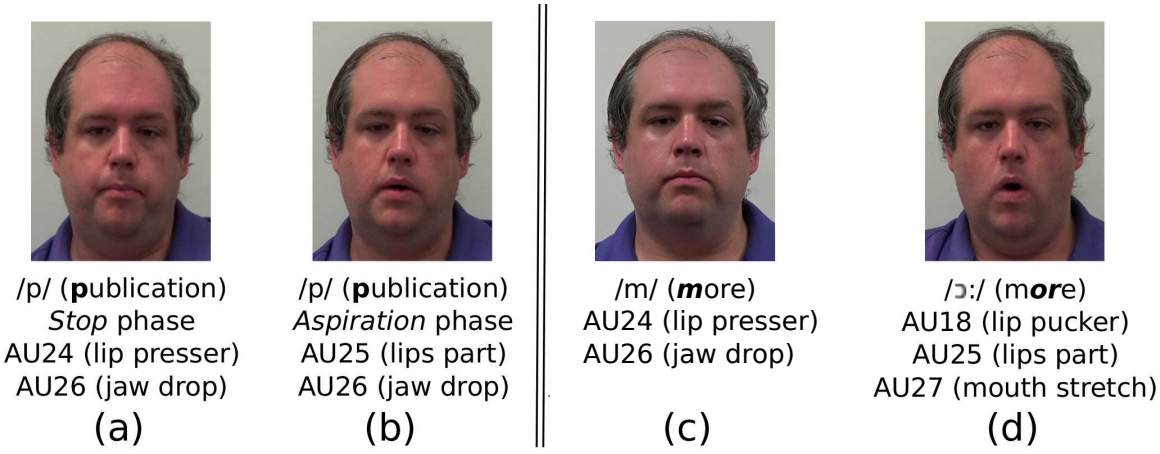

Figure 3.1 Example images of speech-related facial behaviors, where different AUs are activated to pronounce sounds. Note non-additive effects of AUs co-occurring in a combinations in (a) and (d).

between visual and audio information for speech perception. 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 lower-face AUs and speech, which are more objective and will generalize better to various contexts.**

In audiovisual automatic speech recognition (ASR), a *viseme* has been defined to represent facial muscle movements that can visually distinguish the sound [94, 12, 8]. Since some phonemes have similar facial appearance when produced, the mapping from phoneme to viseme is usually derived by statistical clustering [33, 85, 65, 109], but without a universal agreement. Furthermore, the mapping is not always one-to-one because the number of visemes is usually less than the number of phonemes. For example, Neti et al. [65] clustered 44 phonemes into 13 visemes. However, one viseme may be produced by different AU or AU combinations or by a sequence of AU or AU combinations. For example, */p/* and */m/* are in the same cluster of bilabial consonants [65]. */p/* is produced by AU24 (lip presser) + AU26 (jaw drop) (Fig. 3.1a)

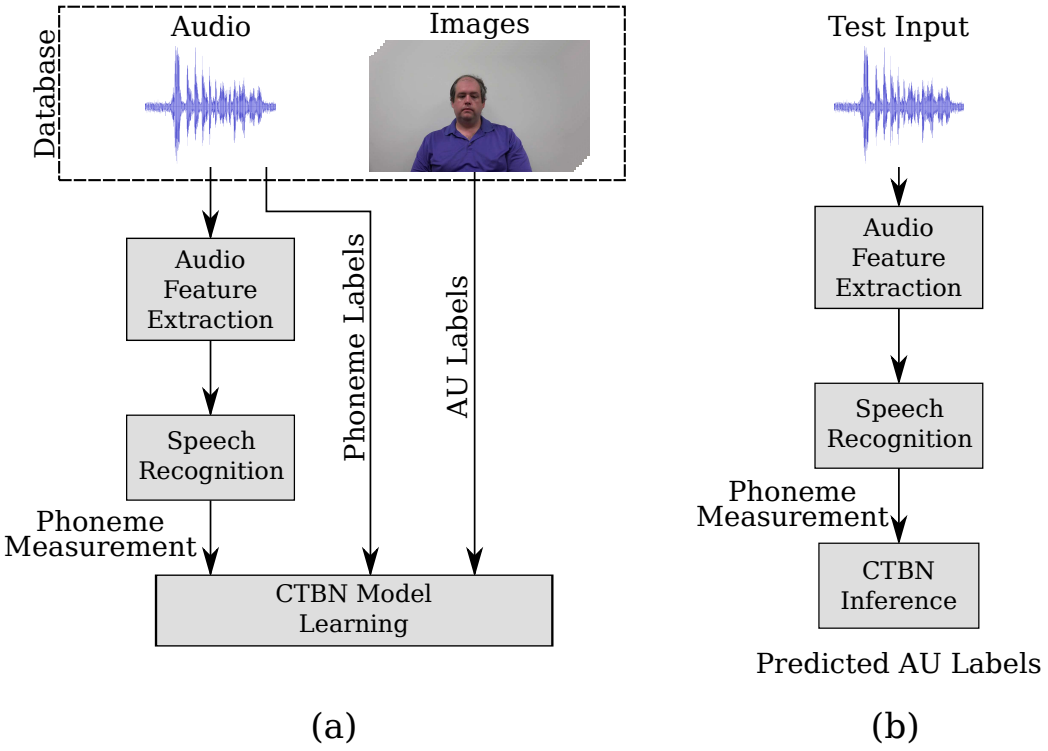

Figure 3.2 The flowchart of the proposed audio-based AU recognition system: (a) an offline training process for CTBN model learning and (b) an online AU recognition process via probabilistic inference.

followed by AU25 (lips part) + AU26 (jaw drop) (Fig. 3.1b); while */m/* is produced by AU24 (lip presser) + AU26 (jaw drop) as shown in Fig. 3.1c. Based on these observations, we proposed to directly study the relationships between facial AUs and phonemes rather than utilizing visemes as intermediate descriptors.

Specifically, a phoneme, which is the smallest phonetic unit of speech, is pronounced by activating a combination of AUs as illustrated in Fig. 3.1. Due to the variation in individual subjects, such relationships are stochastic. Furthermore, different combinations of AUs are responsible for sounding a phoneme at different phases as depicted in Fig. 3.1(a) and (b). Therefore, the dynamic dependencies between AUs and phonemes also undergo a temporal evolution rather than stationary.

Inspired by these, we proposed a novel approach to recognize speech-related AUs from speech by modeling and exploiting the dynamic and physiological relation-

ships between AUs and phonemes through a Continuous Time Bayesian Network (CTBN) [70]. CTBNs are probabilistic graphical models proposed by Nodelman [70] to explicitly model the temporal evolutions over continuous time. CTBNs have been found in different applications, including users' presence and activities modeling [69], robot monitoring [66], sensor networks modeling [39], object tracking [79], host level network intrusion detection [108], dynamic system reliability modeling [11], social network dynamics learning [28], cardiogenic heart failure diagnosis and prediction [32], and gene network reconstruction [2].

Dynamic Bayesian networks (DBNs) are widely used dynamic models for modeling the dynamic relationships among random variables, and have been employed for modeling relationships among facial AUs in the visual channel [101, 100]. However, the dynamic events need to be discretized into discrete time points and thus, the relationships between them are modeled discontinuously. In addition, an alignment strategy should be employed to handle the difference in time scales and the time shift between the two signals. In contrast, considering AUs and phonemes as continuous dynamic events, the CTBN model can explicitly characterize the relationships between AUs and phonemes, and more importantly, model the temporal evolution of the relationships as a stochastic process over continuous time. Fig. 3.2 illustrates the proposed audio-based AU recognition system. During the training process (Fig. 3.2(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. 3.2(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.

This work has three major contributions.

- The dynamic and physiological relationships between AUs and phonemes are

theoretically and probabilistically modeled using a CTBN model.

- Instead of using low-level acoustic features, accurate phoneme measurements are employed benefiting from advanced speech recognition techniques.
- A pilot AU-coded audiovisual database is constructed to evaluate the proposed audio-based AU recognition framework and can be employed as a benchmark database for AU recognition.

The audiovisual AU-coded database consists of a "clean" subset with frontal and neutral faces and a challenging subset collected under unconstrained conditions with large head movements, occlusions from facial hair and accessories, and illumination changes. Experimental results on this database show that the proposed audio-based AU recognition framework achieves significant improvement in recognizing 7 speech-related AUs as compared to the state-of-the-art visual-based methods. The improvement is more impressive for those AUs that are activated at low intensities or "hardly visible" in the visual channel. More importantly, dramatic improvement has been achieved on the challenging subset: the average F1 score of the 7 speech-related AUs is almost doubled compared to those of the visual-based approaches. Furthermore, the proposed CTBN model also outperforms the baseline audio-based methods by a large margin owing to explicitly modeling the dynamic interactions between phonemes and AUs.

3.2 Methodology

3.2.1 Phoneme-AU Relationship Analysis

A phoneme is defined as the smallest phonetic unit in a language. In this work, a set of phonemes defined by Carnegie Mellon University Pronouncing Dictionary (CMUdict) [106] is employed, which is a machine-friendly pronunciation dictionary designed for speech recognition, where 39 phonemes are used for describing North

(a) G(***g***ooey) AU25+26 (b) UW(g***oo***ey) AU18+25+26 (c) IY(goo***ey***) AU25+26

Figure 3.3 Examples of physiological relationships between phonemes and AUs. To pronounce a word *gooey* , different combinations of AUs are activated sequentially. (a) AU25 (lip part) and AU26 (jaw drop) are responsible for producing *G* (***g****ooey*); (b) AU18 (lip pucker), AU25, and AU26 are activated to pronounce *UW* (*g****oo****ey*); and (c) AU25 and AU26 are activated to sound *IY* (*goo****ey***).

American English words. The 39 phonemes defined by CMUdict, along with sample words in parenthesis, are as follows: *AA* (***o***dd), *AE* (***a***t), *AH* (h***u***t), *AO* (***aw***ful), *AW* (c***ow***), *AY* (h***i***de), *B* (***b***e), *CH* (***ch***eese), *D* (***d***ee), *DH* (***th***ee), *EH* (***E***d), *ER* (h***ur***t), *EY* (***a***te), *F* (***f***ee), *G* (***g***reen), *HH* (***h***e), *IH* (***i***t), *IY* (***ea***t), *JH* (***g***ee), *K* (***k***ey), *L* (***l***ee), *M* (***m***e), *N* (***kn***ee), *NG* (pi***ng***), *OW* (***oa***t), *OY* (t***oy***), *P* (***p***ee), *R* (***r***ead), *S* (***s***ea), *SH* (***sh***e), *T* (***t***ea), *TH* (***th***eta), *UH* (h***oo***d), *UW* (t***wo***), *V* (*v***ee**), *W* (***w***e), *Y* (***y***ield), *Z* (***z***ee), *ZH* (sei***z***ure) [106].

Since each phoneme is anatomically related to a specific set of jaw and lower facial muscular movements, there are strong physiological relationships between the speech-related AUs and phonemes. Taking the word *gooey* for instance, a combination of AU25 (lip part) and AU26 (jaw drop) is first activated to produce *G* (***g****ooey*) (Fig. 3.3a). Then, AU18 (lip pucker), AU25, and AU26 are activated together to sound *UW* (*g****oo****ey*) (Fig. 3.3b). Finally, AU25 and AU26 are responsible for producing *IY* (*goo****ey***) (Fig. 3.3c).

Furthermore, these relationships also undergo a temporal evolution rather than stationary. *There are two types of temporal dependencies between AUs and phonemes.*

First, a phoneme is produced by a combination of AUs as shown in Fig. 3.3. The

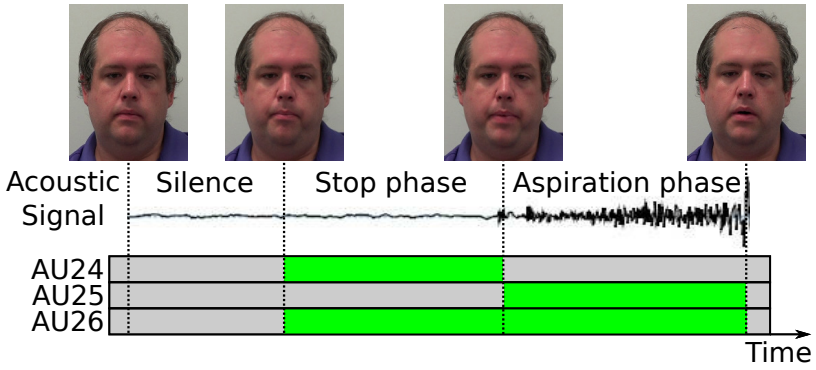

Figure 3.4 Illustration of the dynamic relationships between AUs and phonemes while producing *P*. Specifically, AU24 (lip presser) and AU26 (jaw drop) are activated in the first phase, i.e., the Stop phase, while AU25 (lips part) and AU26 are activated in the second phase, i.e., the Aspiration phase. The activated AUs are denoted by green bars with a diagonal line pattern; while the inactivated AUs are denoted by grey bars. Best viewed in color.

probabilities of the AUs being activated increase prior to voicing the phoneme and reach an apex when the sound is fully emitted, and then decrease while preparing to voice the next phoneme.

Second, different combinations of AUs are responsible for producing a single phoneme at different phases. For example, as illustrated in Fig. 3.4, the phoneme *P* in the word *cha**p**s* has two consecutive phases, i.e., *Stop* and *Aspiration* phases. During the *Stop* phase, AU24 (lip presser) is activated as lips are pressed together to hold the breath without making sound [13], when the upper and lower teeth are apart indicating the presence of AU26 (jaw drop). In the *Aspiration* phase, the lips are apart by activating AU25 and releasing AU24 to release the breath with an audible explosive sound [13]. Thus, AU24 and AU26 are activated before the sound is heard, and AU24 is released as soon as the sound is made when AU25 is activated.

Note that these dynamic and physiological relationships are stochastic and vary among individual subjects and different words. For example, according to phonetics [13], AU20 is responsible for producing the phoneme AE; and AU22 is responsible for producing the phonemes CH (***ch***aps), ZH (Eura***s***ian), and SH (***sh***e). However,

some subjects are not activating those facial AUs while producing the corresponding phonemes. For example, in our audiovisual database, there are 8 out of 13 subjects did not activate AU20 when sounding the phoneme AE as in ch***a***ps; and two subjects did not activate AU22 when sounding the phoneme CH as in pat***ch***. Moreover, one subject activated AU22 for producing G as in ***g***ooey, H as in ***h***ue, and K as in ***q***ueen, where AU22 is not responsible for producing those phonemes according to phonetics [13].

In addition, speech recognition is not perfect. For speech recognition, uncertainties are not only introduced by mis-classification of phonemes, but also by the temporal shift of the recognized phoneme segments compared to the groundtruth phoneme segments. Moreover, the AUs are usually activated slightly before the phoneme is produced [35]. Therefore, we employ a probabilistic framework, a CTBN [70] in particular, to explicitly model the dynamic relationships between phonemes and AUs over continuous time.

3.2.2 Modeling Phoneme-AU Relationships by a CTBN

Figure 3.5 A CTBN model for audio-based AU recognition.

A CTBN is a directed, possibly cyclic, graphical model [70], which consists of an initial distribution specified as a Bayesian network and a set of random variables. As shown in Fig. 3.5, a CTBN model is employed to capture the dynamic and physiolog-

ical relationships between AUs and phonemes as well as the measurement uncertainty in speech recognition. There are two types of nodes in the model: the *unshaded nodes* represent hidden nodes, whose states should be inferred through the model; whereas the *shaded node* denotes the measurement node, whose states can be observed and used as evidence for inference.

Specifically, the phoneme node denoted by "**Phone**" has 29 states representing 28 phonemes in the audiovisual dataset and one silence state, and is employed to model the dynamics of phonemes: durations of phonemes and transitions between phonemes. A measurement node denoted as "$\mathbf{O}_p$" with 29 states is used to represent the phoneme measurement obtained by speech recognition. The directed link between "**Phone**" and "$\mathbf{O}_p$" represents the measurement uncertainty in speech recognition, e.g., misdetection and temporal misalignment.

Based on the study in [101], there are semantic and dynamic relationships among AUs. In this work, AUs often occur in combinations to produce sounds. However, the CTBN model has an assumption that no two variables change at exactly the same time, which, unfortunately, is not held in this application, where two or even more different AUs can change simultaneously. For example, AU25 is activated at the same time when AU24 is released as illustrated in Fig 3.

Instead of using 7 separate nodes for 7 speech-related AUs, respectively, a single "**AU**" node is employed to model the joint distributions of all speech related AUs. Since each AU can be at one of "absence" or "presence" status, "**AU**" has $2^7 = 128$ states, each of which is corresponding to one combination of 7 AUs. For example, the state 0 of the "**AU**" node represented by a binary number "0000000", means no AU is activated, while the state 1 with a binary number "0000001", means only AU27 is present. This way, the relationships between all AUs are naturally modeled without learning the CTBN structure. The directed links between "**AU**" and "**Phone**" capture the dynamic and physiological relationships between them.

3.2.3 Model Parameterization

In a CTBN, each node, e.g., "**Phone**" and "**AU**" in this work, evolves as a Markov process, whose dynamics is described by a set of transition intensity matrices, called conditional intensity matrices (CIMs) denoted by $\mathbf{Q}$, in which the transition intensity values are determined by the instantiations of parent node(s).

Model Parameterization for "**Phone**"

The directed link from "**AU**" to "**Phone**" represents the relationships that AUs are activated prior to pronounce a phoneme and thus, the dynamic of "**Phone**" is based on the instantiations of "**AU**". Given the k^{th} state of "**AU**" denoted as a_k, $k = 0, \cdots, 127$, the CIM for "**Phone**", a 29×29 matrix denoted as $\mathbf{Q}_{\mathbf{Phone}|\mathbf{AU}=a_k}$, is defined as follows:

$$\mathbf{Q}_{\mathbf{Phone}|\mathbf{AU}=a_k} = \begin{bmatrix} -q_0^{ph|a_k} & q_{0,1}^{ph|a_k} & \cdots & q_{0,28}^{ph|a_k} \\ q_{1,0}^{ph|a_k} & -q_1^{ph|a_k} & \cdots & q_{1,28}^{ph|a_k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{28,0}^{ph|a_k} & q_{28,1}^{ph|a_k} & \cdots & -q_{28}^{ph|a_k} \end{bmatrix} \tag{3.1}$$

where $q_i^{ph|a_k}$ denotes the conditional intensity value when "**Phone**" remains at its i^{th} state denoted by ph_i, $i = 0, \cdots, 28$, given $\mathbf{AU} = a_k$; $q_{i,j}^{ph|a_k}$ ($j = 0, \cdots, 28$ and $j \neq i$) denotes the conditional intensity value when "**Phone**" transitions from its i^{th} state to its j^{th} state, given $\mathbf{AU} = a_k$; and $q_i^{ph|a_k} = \sum_{j \neq i} q_{i,j}^{ph|a_k}$.

Based on Eq. 3.1, the dynamics of "**Phone**" may change following the state of "**AU**". For example, if "**AU**" is at its a_0 state, the dynamics of "**Phone**" will be controlled by its CIM $\mathbf{Q}_{\mathbf{Phone}|\mathbf{AU}=a_0}$; while the intensity matrix $\mathbf{Q}_{\mathbf{Phone}|\mathbf{AU}=a_1}$ will be employed after "**AU**" transitions to its a_1 state.

Given the initial states of "**Phone**" and "**AU**" at time $t = 0$ (**Phone** $= ph_i$ and $\mathbf{AU} = a_k$), the probability of **Phone** remaining at its initial state ph_i is specified by

the probability density function as [70]:

$$f(t) = q_i^{ph|a_k} e^{-q_i^{ph|a_k} t}, \quad t \geq 0 \tag{3.2}$$

Then, the expected time of transition of "**Phone**", i.e., leaving from the i^{th} state to any of the other states, can be computed as $\frac{1}{q_i^{ph|a_k}}$. When transition occurs, "**Phone**" transitions from its i^{th} state to its j^{th} state with probability denoted by $\theta_{i,j|a_k} = \frac{q_{i,j}^{ph|a_k}}{q_i^{ph|a_k}}$ [70].

Model Parameterization for "**AU**" and "$\mathbf{O}_p$"

The state of "**AU**" may also change according to the state of "**Phone**". Following the previous example of producing a phoneme P, the probability of AU24 (lip presser) should decrease rapidly if the sound is emitted in the *Aspiration* phase. Such relationships can be captured by a directed link from "**Phone**" to "**AU**". Then, the CIM of "**AU**" given the i^{th} state of "**Phone**" is denoted by $\mathbf{Q}_{\mathbf{AU}|\mathbf{Phone}=ph_i}$ and can be defined similarly as Eq. 3.1. Likewise, the CIM of "$\mathbf{O}_p$" given the i^{th} state of "**Phone**" ($\mathbf{Q}_{\mathbf{O}_p|\mathbf{Phone}=ph_i}$) captures the measurement uncertainty of speech recognition and is defined similarly as Eq. 3.1.

3.2.4 Parameter Estimation

The model parameters of a CTBN include the initial distribution Pr_0 specified by a Bayesian network, the structure of CTBN, and the CIMs. The initial distribution Pr_0 can be estimated given the groundtruth AU and phoneme labels of the first frames of all sequences. It becomes less important in the context of CTBN inference and learning when we assume the model is irreducible, especially when the time range becomes significantly large [108]. Thus, as the CTBN model structure is given as shown in Fig. 3.5, the model parameters we should learn are the expected time of transitions, i.e., $\frac{1}{q_i^{ph|a_k}}$, and the transition probabilities, i.e., $\theta_{i,j|a_k}$. In this work, the groundtruth AU labels and the phoneme labels are manually annotated, and thus

the training data $\mathcal{D}$ is complete, i.e., for each time point along each trajectory, the instantiation of all variables is known. Then, we can estimate the parameters of a CTBN efficiently using Maximum Likelihood estimation (MLE) [71]. In particular, the likelihood function can be factorized as the product of a set of local likelihood functions as below:

$$\begin{aligned} L(\mathbf{q}, \boldsymbol{\theta} : \mathcal{D}) &= \prod_{X \in \mathbf{X}} L_X(\mathbf{q}_{X|\mathbf{v}_X}, \boldsymbol{\theta}_{X|\mathbf{v}_X} : \mathcal{D}) \\ &= \prod_{X \in \mathbf{X}} L_X(\mathbf{q}_{X|\mathbf{v}_X} : \mathcal{D}) L_X(\boldsymbol{\theta}_{X|\mathbf{v}_X} : \mathcal{D}) \end{aligned} \tag{3.3}$$

where $\mathbf{X}$ consists of all random variables in the CTBN, i.e., "**AU**", "**Phone**", and "$\mathbf{O}_p$" in this work; $X \in \mathbf{X}$ is a random variable with M states and has a set of parent nodes denoted by $\mathbf{V}_X$. $\mathbf{q}_{X|\mathbf{v}_X}$ is a set of parameters characterizing the expected time of transition from the current state of X to any of the other states given its parent nodes $\mathbf{V}_X$, i.e., the diagonal elements of $\mathbf{Q}_{X|\mathbf{v}_X}$; and $\boldsymbol{\theta}_{X|\mathbf{v}_X}$ represents the transition probabilities of X given its parent nodes $\mathbf{V}_X$, i.e., the off-diagonal elements of $\mathbf{Q}_{X|\mathbf{v}_X}$.

Given an instantiation of the parent nodes, i.e., $\mathbf{V}_X = \mathbf{v}_X$, the sufficient statistics are $T[x_i|\mathbf{v}_X]$ representing the total length of time that X stays at the state x_i and $N[x_i, x_j|\mathbf{v}_X]$ representing the number of transitions of X from the state x_i to the state x_j. With the sufficient statistics, $L_X(\mathbf{q}_{X|\mathbf{v}_X} : \mathcal{D})$ and $L_X(\boldsymbol{\theta}_{X|\mathbf{v}_X} : \mathcal{D})$ in Eq. 3.3 can be calculated as follows [71],

$$\begin{aligned} &L_X(\mathbf{q}_{X|\mathbf{v}_X} : \mathcal{D}) \\ &\qquad = \prod_{\mathbf{v}_X} \prod_{i \in M} (q_i^{X|\mathbf{v}_X})^{N[x_i|\mathbf{v}_X]} \exp\left(-q_i^{X|\mathbf{v}_X} T[x_i|\mathbf{v}_X]\right) \end{aligned} \tag{3.4}$$

where $q_i^{X|\mathbf{v}_X}$ is the i^{th} diagonal element in the CIM of X given an instantiation of its parent nodes ($\mathbf{Q}_{X|\mathbf{v}_X}$, referring to Eq. 3.1); and $N[x_i|\mathbf{v}_X] = \sum_{j \in M, j \neq i} N[x_i, x_j|\mathbf{v}_X]$ represents the total number of transitions leaving from the state x_i.

$$L_X(\boldsymbol{\theta}_{X|\mathbf{v}_X} : \mathcal{D}) = \prod_{\mathbf{v}_X} \prod_{i \in M} \prod_{j \in M, j \neq i} (\theta_{i,j|\mathbf{v}_X})^{N[x_i, x_j|\mathbf{v}_X]} \tag{3.5}$$

where $\theta_{i,j|\mathbf{v}_X} = \frac{q_{i,j}^{X|\mathbf{v}_X}}{q_i^{X|\mathbf{v}_X}}$ represents the transition probability from the i^{th} state of X to the j^{th} state, given an instantiation of its parent nodes $\mathbf{v}_X$.

By substituting Eq. 3.4 and Eq. 3.5 into Eq. 3.3, the log-likelihood for X can be obtained as below

$$\begin{aligned}
&\ell_X(\mathbf{q}_{X|\mathbf{v}_X}, \boldsymbol{\theta}_{X|\mathbf{v}_X} : \mathcal{D}) = \ell_X(\mathbf{q}_{X|\mathbf{v}_X} : \mathcal{D}) + \ell_X(\boldsymbol{\theta}_{X|\mathbf{v}_X} : \mathcal{D}) \\
&= \sum_{\mathbf{v}_X} \sum_{i \in M} N[x_i|\mathbf{v}_X] \ln(q_i^{X|\mathbf{v}_X}) - q_i^{X|\mathbf{v}_X} T[x_i|\mathbf{v}_X] \\
&+ \sum_{\mathbf{v}_X} \sum_{i \in M} \sum_{j \in M, j \neq i} N[x_i, x_j|\mathbf{v}_X] \ln \theta_{i,j|\mathbf{v}_X}
\end{aligned} \tag{3.6}$$

By maximizing Eq. 3.6, the model parameters can be estimated as follows [72]:

$$\hat{q}_i^{X|\mathbf{v}_X} = \frac{N[x_i|\mathbf{v}_X]}{T[x_i|\mathbf{v}_X]} \tag{3.7}$$

$$\hat{\theta}_{i,j|\mathbf{v}_X} = \frac{N[x_i, x_j|\mathbf{v}_X]}{N[x_i|\mathbf{v}_X]} \tag{3.8}$$

3.2.5 Phoneme Measurements Acquisition

In this work, a state-of-the-art speech recognition approach, i.e., Kaldi toolkit [77], is employed to obtain the phoneme measurements. In particular, 13-dimensional MFCC features [20] are first extracted, based on which, Kaldi is used to produce word-level speech recognition results, which are further aligned into phoneme-level segments. These phoneme-level segments are then fed into the CTBN model as the evidence. Note that, the evidence is given as a continuous event and the gaps between two successive phonemes are considered as silence.

3.2.6 AU Recognition via CTBN Inference

Given the fully observed evidence, i.e., phoneme measurements denoted by $\mathbf{E}_p$, and a prior distribution, Pr_0, over the variables at time t_0, AU recognition is performed by estimating the posterior probability $Pr(\mathbf{AU}|\mathbf{E}_p)$ via the CTBN model. Exact inference can be performed by flattening all CIMs into a single intensity matrix $\mathbf{Q}$

using amalgamation, which will be treated as a homogeneous Markov process [70], where the intensity values in **Q** stay the same over time. However, exact inference is infeasible for this work as the state space grows exponentially large as the number of variables increases. In this work, we employ auxiliary Gibbs sampling [81], which takes a Markov Chain Monte Carlo (MCMC) approach to estimate the distribution given evidence, implemented in the CTBN reasoning and learning engine (CTBN-rle) [92] to perform CTBN inference.

Since the state of the "**AU**" node corresponds to the joint states of 7 speech-related AUs, the inference results would be the joint probability of those AUs. Then, the posterior probability of a target AU given the evidence can be obtained by marginalizing out all the other AUs. Optimal states of the target AUs can be estimated by maximizing the posterior probability.

3.3 Experiments

To demonstrate the effectiveness of the proposed approach, the proposed method was compared with four state-of-the-art visual-based methods, five state-of-the-art audio-based methods, and two baseline methods based on probabilistic modeling [1].

3.3.1 Baseline Methods for Comparison

To demonstrate the effectiveness of the proposed audio-based AU recognition framework, we compared the proposed method, denoted as *CTBN*, with four state-of-the-art visual-based methods, five state-of-the-art audio-based methods, and two baseline audio-based methods using probabilistic modeling on the AU-coded audiovisual database.

[1]The two baseline methods using probabilistic modeling are developed in this work

Ada-LBP: The first visual-based baseline method, denoted as *Ada-LBP* [37, 62], employs histogram of LBP features, which have been shown to be effective in facial AU recognition. Specifically, face regions across different facial images are aligned to remove the scale and positional variance based on a face and eye detector and then cropped to 96×64 for preprocessing purposes [2]. Then, the cropped face region is divided with a 7×7 grid, from each subregion of which, LBP histograms with 59 bins are extracted. AdaBoost is employed to select the most discriminative features, which are used to construct a strong classifier for each AU.

Ada-LPQ: The second visual-based baseline method, denoted as *Ada-LPQ* [45], employs histogram of LPQ features. Specifically, the face region is divided into 7×7 grid, from each of which, LPQ histograms with 256 bins are extracted. Similar to the *Ada-LBP*, AdaBoost is employed for feature selection and classifier construction for each AU.

SVM-LGBP: The third visual-based baseline method, denoted as *SVM-LGBP*, employed histogram of LGBP features [90, 102, 61]. Particularly, the face region is convolved with 18 Gabor filters, i.e. three wavelengths $\lambda = \{3, 6.3, 13.23\}$ and six orientations $\theta = \{0, \frac{\pi}{6}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{2\pi}{3}, \frac{5\pi}{6}\}$, with a phase offset $\psi = 0$, a standard deviation of the Gaussian envelope $\sigma = \frac{5\pi}{36}$, and a spatial aspect ratio $\gamma = 1$, which results in 18 Gabor magnitude response maps. Each of the response maps is divided into a 7×7 grid, from each of which, LBP histograms with 59 bins are extracted and concatenated as LGBP features. For each AU, AdaBoost is employed to select 400 LGBP features, which are employed to train an SVM classifier.

IB-CNN-LIP: The fourth visual-based baseline method, denoted as *IB-CNN-LIP*, employed a deep learning based model, i.e. Incremental Boosting Convolutional

[2] All the visual-based baseline methods employed the same preprocessing strategy. Except the IB-CNN-LIP, which employed a 96×96 face region, all methods used a 96×64 face region.

Neural Network (IB-CNN) [36] for facial AU recognition. Since only the lower-part of the face is responsible for producing the speech-related AUs, the aligned and cropped lip region along with the landmarks on lips are employed in a two-stream IB-CNN to learn both appearance and geometry information for each target AU.

Baseline Audio-based Methods

SVM-GeMAPS and **SVM-ComParE**: The first and second audio-based baselines employ low-level audio features, i.e. 18-dimensional GeMAPS features and 130-dimensional ComParE features, extracted from the audio channel using openSMILE [25], denoted as *SVM-GeMAPS* and *SVM-ComParE*, respectively. In addition, a $z-score$ normalization is performed for each subject to compensate the inter-subject variations. An SVM is trained for each target AU using LIBSVM toolbox [17].

LSTM-GeMAPS and **LSTM-ComParE**: The third and fourth audio-based baselines employ the same GeMAPS and ComParE features described above. For each target AU, a Long-Short Term Memory (LSTM) network is employed to learn the temporal dependencies. The LSTM network, implemented using TensorFlow library [1], consists of 3 hidden layers with 156, 256, and 156 hidden units, respectively.

Ada-MFCC: The last audio-based baseline method, denoted as *Ada-MFCC*, employs low-level audio features, i.e., 13-dimensional MFCC features, extracted from the audio channel. In addition, 7 frames of the MFCC features, i.e. 3 frames before and after the current frame along with the current one, are concatenated as the final MFCC features employed as the input to train an AdaBoost classifier for each AU.

Because of the different sampling rate used in visual and audio channels, a cubic spline interpolation method is employed to synchronize the acoustic features with the image frames [62] for all audio-based methods.

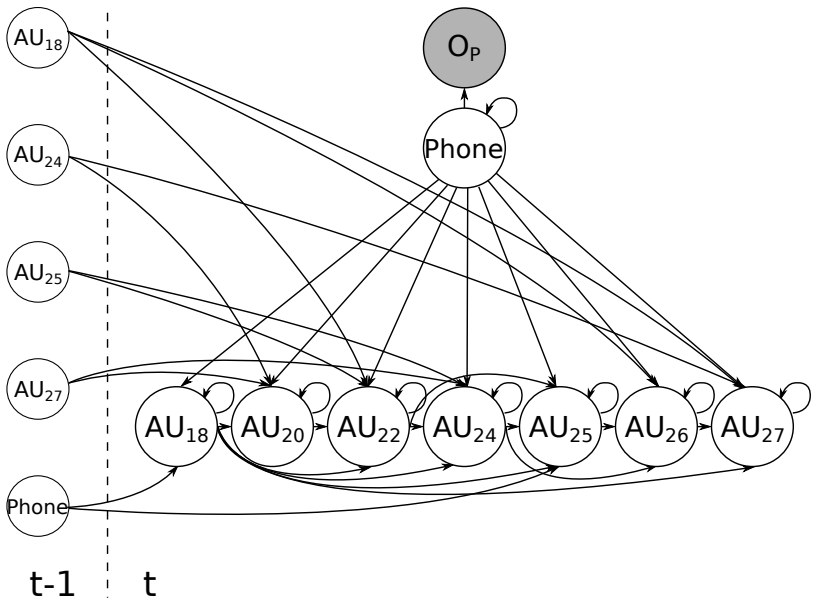

Figure 3.6 A DBN model learned from the clean subset for modeling the semantic and dynamic relationships between AUs and phonemes. The directed links in the same time slice represent the semantic relationships among the nodes; the self-loop at each node represents its temporal evolution; and the directed links across two time slices represent the dynamic dependency between the two nodes. The shaded node is the measurement node and employed as evidence for inference; and the unshaded nodes are hidden nodes, whose states can be estimated by inferring over the trained DBN model.

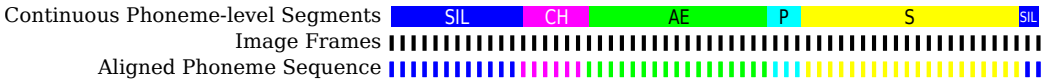

Figure 3.7 An illustration of discretizing continuous phoneme segments into frame-by-frame phoneme measurements for the word "chaps". The first row gives the phoneme-level segments obtained by Kaldi [77]. The second row shows a sequence of image frames, to which the phonemes will be aligned. The last row depicts the aligned sequence of phoneme measurements. Best viewed in color.

DBN: The first baseline probabilistic-model-based method, denoted as *DBN*, employs a Dynamic Bayesian Network (DBN) to model the semantic and dynamic relationships between phonemes and AUs. Specifically, the DBN structure as shown in Fig. 3.6 as well as the DBN parameters are learned using the Bayes Net Toolbox [64] from the clean subset. In order to synchronize phoneme measurements with the image frames, the continuous phoneme segments obtained by speech recognition are discretized according to the sampling rate of the image frames, as illustrated in Fig. 3.7. Then, AU recognition is performed by DBN inference given the discretized phoneme measurements.

CTBN-F: The last baseline method, denoted as *CTBN-F*, which is short for CTBN-Factorized, employs a factorized CTBN to explicitly model the dynamic and physiological relationships between phonemes and each AU as well as the dynamic relationships among AUs. As shown in Fig. 3.8, each AU is represented by an individual node with 2 states, i.e. "absence" and "presence", in contrast to a combined node in Fig. 3.5. The directed link between "**Phone**" and each AU node represents the dynamic and physiological relationships between phonemes and the AU. Those between AU nodes capture the dynamic interactions among AUs and are learned from the data using CTBN-rle [92]. The model parameters, i.e., the CIMs, are estimated as described in Section 3.2.4 from the training data.

Both *DBN* (Fig. 3.6) and *CTBN-F* (Fig. 3.8) capture dynamic relationships between AUs and phonemes. However, the dynamic dependencies from AUs to phonemes are not learned and modeled in a *DBN*. This is because the penalty for adding a link from an AU node to the phoneme node is much higher than that from the phoneme node to AU nodes for the 29-state phoneme node. In addition, since loops are allowed in a CTBN model, there is a loop between AU24 and AU25 in *CTBN-F* indicating the strong dynamic relationships between those two AUs.

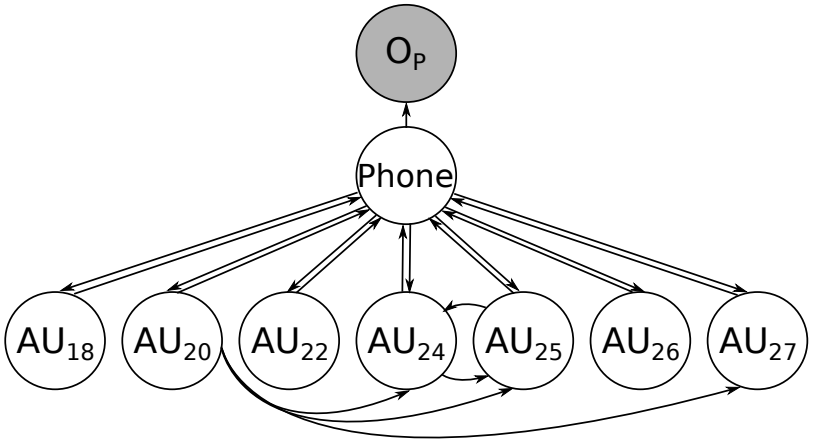

Figure 3.8 The structure of a *CTBN-F* model trained on the clean subset for modeling the dynamic physiological relationships between AUs and phonemes.

Note that the *Ada-MFCC*, *DBN*, and *CTBN-F* methods are proposed in this work for recognizing speech-related AUs using only audio information.

3.3.2 Experimental Results and Data Analysis on the Clean Subset

We first evaluate the proposed CTBN model on the clean subset. For all methods compared, a leave-one-subject-out training/testing strategy is employed, where the data from 8 subjects is used for training and the remaining data is used for testing. Quantitative experimental results on the clean subset are reported in Fig. 3.9 in terms of F1 score, true positive rate, and false positive rate. As shown in Fig. 3.9, the proposed CTBN model outperformed all the baseline methods significantly in terms of the average F1 score (**0.653**).

Compared with *Ada-LBP*, *Ada-LPQ*, and *SVM-LGBP*, which employ appearance information from the visual channel, the overall AU recognition performance is improved from **0.416** (*Ada-LBP*), **0.448** (*Ada-LPQ*), and **0.386** (*SVM-LGBP*) to **0.653** by the proposed *CTBN* model in terms of the average F1 score. As shown in Fig. 3.9, *CTBN* outperforms *Ada-LBP*, *Ada-LPQ*, and *SVM-LGBP* for all target AUs, which demonstrates the effectiveness of using information extracted from the

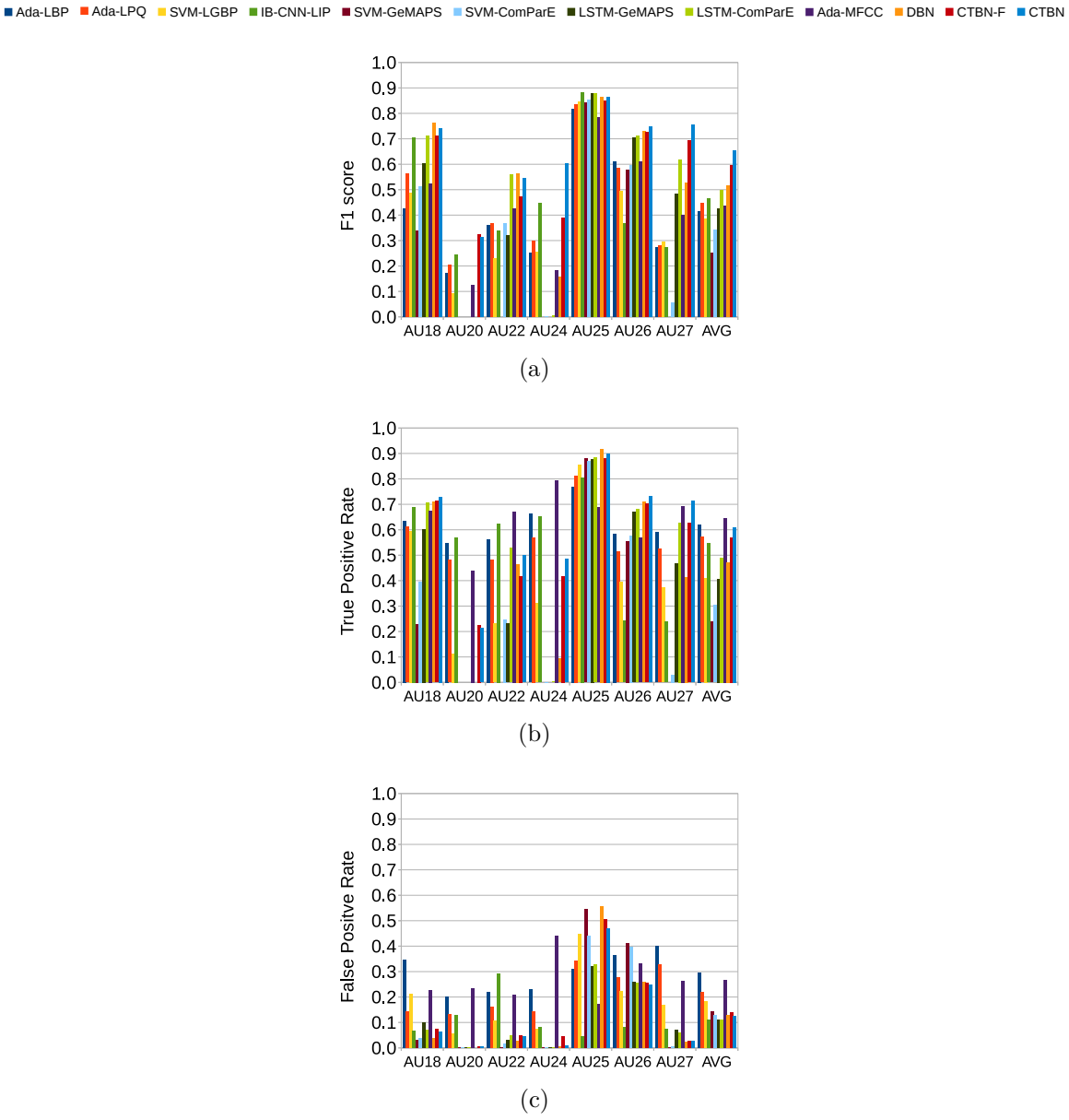

Figure 3.9 Performance comparison on the clean subset in terms of (a) F1 score, (b) true positive rate, and (c) false positive rate for the 7 speech-related AUs.

audio channel. The improvement is more impressive for AU27 (mouth stretch), i.e., **0.755** by *CTBN* versus **0.273** by *Ada-LBP*, **0.279** by *Ada-LPQ*, and **0.296** by *SVM-LGBP*, since the visual observation of AU27 is not reliable during speech due to the occlusion caused by lip movements as illustrated in Fig. 3.1.

Compared with *IB-CNN-LIP*, which employs both appearance and geometry

information from the visual channel, the overall AU recognition performance is improved from **0.465** (*IB-CNN-LIP*) to **0.653** by the proposed *CTBN* in terms of the average F1 score. Not surprisingly, the *IB-CNN-LIP* outperforms the other visual-based approaches that employ only appearance features. In addition, it also performs better than the proposed *CTBN* on AU25 (lips part) because both the appearance and geometry clues from the visual channel are strong for AU25. In contrast, drastic improvement is achieved for AU26 (jaw drop), from **0.367** by *IB-CNN-LIP* to **0.748** by *CTBN*, because the appearance information for AU26 is invisible due to the occlusion as depicted in Fig. 3.1 and the geometrical change is subtle during speech.

Compared with *SVM-GeMAPS* and *SVM-ComParE*, which employ low-level acoustic features for facial AU recognition, the overall AU recognition performance in terms of F1 score is improved from **0.251** (*SVM-GeMAPS*) and **0.342** (*SVM-ComParE*) to **0.653** using the proposed *CTBN*. GeMAPS and ComParE have been designed to model short-term paralinguistic [25], i.e. non-lexical, states for emotion recognition, and thus are not favorable to capture the relationships between speech and facial appearance changes. In addition, both *SVM-GeMAPS* and *SVM-ComParE* do not perform well on AU20 and AU24 because they have the lowest numbers of occurrence in our dataset as shown in Table 1.1 in Chapter 1.

Compared with *LSTM-GeMAPS* and *LSTM-ComParE*, which employ low-level acoustic features and LSTM that models dynamics in the audio channel, the overall AU recognition performance in terms of F1 score is improved from **0.427** (*LSTM-GeMAPS*) and **0.498** (*LSTM-ComParE*) to **0.653** using the proposed *CTBN*. The performance improvement is owing to explicit modeling relationships between phonemes and AUs as well as relationships among AUs.

Compared with *Ada-MFCC*, which employs low-level acoustic features extracted in a sequence (7 frames), the proposed *CTBN* improves the overall AU recog-

nition performance by **0.217** in terms of the average F1 score. Furthermore, the CTBN outperforms the *Ada-MFCC* for all AUs notably by employing more accurate and reliable higher level audio information, i.e., the phonemes, thanks to the advanced speech recognition techniques, and more importantly, by exploiting the dynamic physiological relationships between AUs and phonemes.

Compared with *DBN*, the proposed *CTBN* improves the overall AU recognition performance by **0.138**, in terms of the average F1 score. Particularly, the F1 scores of *CTBN* are better than or at least comparable to those of *DBN* for all AUs, as shown in Fig. 3.9. The primary reason for the performance improvement is that the dynamic dependencies modeled in *DBN* are stationary; whereas the relationships between AUs and phonemes actually undergo a temporal evolution as modeled in the *CTBN*. For example, the F1 score of AU24 (lip presser) is dramatically improved from **0.158** by *DBN* to **0.603** by *CTBN*, because AU24 is activated before the sound is produced and released once the sound is heard, which can be better modeled in *CTBN*. Note that *DBN* fails to recognize AU20 (lip stretcher). Although AU20 is required to produce AE in ch***a***ps according to Phonetics [13], some subjects did not activate AU20 as observed in our audiovisual dataset and thus, the semantic relationship between **Phone** and **AU20** is rather weak. However, no dynamic link is learned between **Phone** and **AU20** in *DBN*. In contrast, dynamic relationships between AUs and phonemes modeled by *CTBN* are more crucial for inferring AU20. As a result, the F1 score of AU20 is improved from **0** by *DBN* to **0.314** by *CTBN*. We found that *DBN* performs slightly better on AU18 (lip pucker) and AU22 (lip funneler) than *CTBN*. This is because AU18 and AU22 have the strongest static relationships with phonemes: when pronouncing UW in t***wo*** and CH in ***ch***eese, they are activated for most of subjects.

Compared with *CTBN-F*, the proposed *CTBN* further improves the overall AU recognition performance by **0.057**, in terms of the average F1 score. By employing

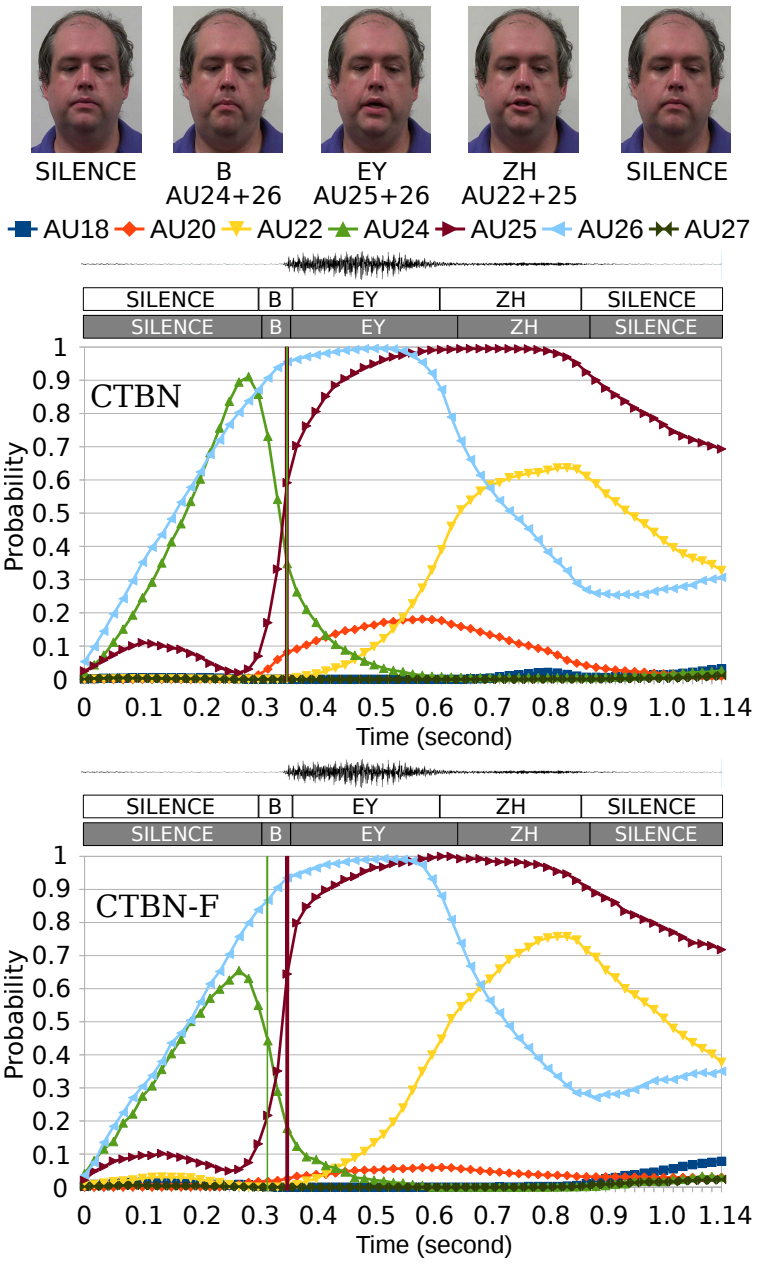

Figure 3.10 An example of the system outputs by CTBN inference using *CTBN* and *CTBN-F*, respectively. The top row shows key frames from an image sequence where a word "beige" is produced, where AU22, AU24, AU25, and AU26 are involved. The bottom two figures depict the probabilities of AUs changing over time by *CTBN* and *CTBN-F*, respectively. The shaded phoneme sequence is used as evidence of the CTBN models and the unshaded one is the ground truth phoneme labels. The vertical green line denotes the time point when AU24 is released, while the vertical garnet line denotes the time point when AU25 is activated. The two lines are overlapped with each other in the *CTBN* output. Best viewed in color.

one single node to model the joint distribution over the 7 target AUs, comprehensive relationships between AUs and phonemes, i.e., AUs occur in combinations to produce sounds, can be well characterized as discussed in Section 3.2.2.

In addition, Fig. 3.10 gives an example of the system outputs of the *CTBN* and *CTBN-F*, i.e., the probabilities of AUs given the phoneme measurements (the shaded phoneme sequence), by the CTBN inference over continuous time. For both *CTBN* and *CTBN-F*, the probabilities of AUs change corresponding to the transitions of phonemes, when sounding a word "beige". For example, the probability of AU24 increases and reaches its apex before the phoneme *B* is produced. AU26 can be recognized even though the gap between upper and lower teeth is invisible in visual channel because the presence of AU24. When the sound *B* is emitted, the probability of AU24 drops rapidly, while the probability of AU25 increases. The vertical green line denotes the time point when AU24 is released, while the vertical garnet line denotes the time point when AU25 is activated. Ideally, they should overlap with each other due to the transition from the "Closure" phase to the "Release" phase of sounding *B*, as the result of the *CTBN* (the top plot). Whereas, a noticeable gap between the two lines can be observed in the result of the *CTBN-F* (the bottom plot) because that no two AUs are allowed to change states at the same time in the factorized CTBN.

Moreover, we analyzed the relationships learned by *CTBN*. Table 3.1 depicts a part of the CIM associated with the "**AU**" node given the state of "**Phone**" as *B*, where the first row and column give the states of "**AU**" with the corresponding AU/AU combinations. For example, if "**AU**" is at its 10^{th} state, i.e., AU24 and AU26 are activated, the corresponding conditional intensity value in the CIM is -16.72. As described in Section 3.2.3, the "**AU**" node is expected to transit in $\frac{1}{16.72}s$. Upon transition, it has a higher chance to transit to its 6^{th} state (AU25+AU26) with a probability of $\frac{9.12}{16.72}$. However, if the lip movement is not fast, i.e., AU25 is not acti-

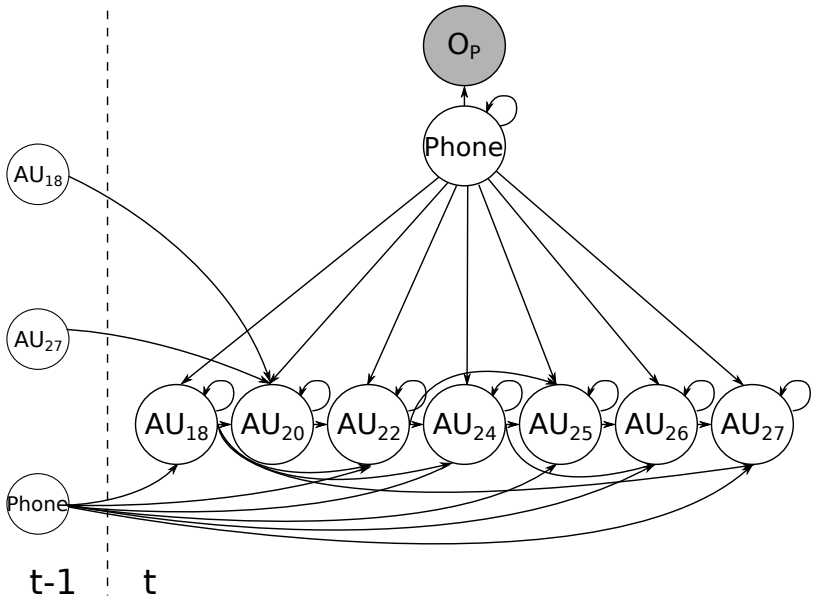

Figure 3.11 A DBN model learned from the challenging subset for modeling the semantic and dynamic relationships between AUs and phonemes.

vated when AU24 is released, it may transit to its 2^{nd} state (AU26) with a probability of $\frac{6.59}{16.72}$. Then, the "**AU**" node will leave this state quickly in $\frac{1}{40.47}s$ and transit to its 6^{th} state with a high probability of $\frac{32.37}{40.47}$.

Compared with other AUs, the performance of *CTBN* on AU20 and AU22 is relatively low. This is mainly because of the high inter-subject variations for these two AUs as discussed in Section 3.2.1.

3.3.3 Experimental Results on the Challenging Subset of the Audiovisual Database

Experiments were conducted on the challenging subset to demonstrate the effectiveness of the proposed audio-based facial AU recognition under real world conditions, where facial activities are accompanied by free head movements, illumination changes, and often with occlusions of the face regions caused by facial hairs, caps, or glasses.

The proposed *CTBN* model, the state-of-the-art audio-based and visual-based methods, and the baseline methods were trained and tested on the challenging subset

Table 3.1 A part of the CIM associated with the "**AU**" node given the state of "**Phone**" as B, where the first row and column give the states of "**AU**" node with the corresponding AU/AU combinations in the parenthesis.

	0	$\cdots$	2(AU26)	$\cdots$	6(AU25+AU26)	$\cdots$	10(AU24+AU26)	$\cdots$
0	-10.07	$\cdots$	0	$\cdots$	0	$\cdots$	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	0	$\cdots$	-40.47	$\cdots$	32.37	$\cdots$	8.09	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
6	1.23	$\cdots$	0	$\cdots$	-1.23	$\cdots$	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	0	$\cdots$	6.59	$\cdots$	9.12	$\cdots$	-16.72	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

using a leave-one-subject-out strategy. Since there are only 6 subjects in the challenging subset, we employed the data in the clean subset except those of the two subjects, who also appear in the challenging subset, as additional training data to ensure a subject-independent context. Specifically, the data of 5 subjects from the challenging subset along with the data of 7 subjects from the clean subset is used as the training data, and the remaining one subject from the challenging subset is employed as the testing data.

The structures of the *DBN* and *CTBN-F* trained on the challenging data are shown in Fig. 3.11 and Fig. 3.12, respectively. Comparing Fig. 3.6 and Fig. 3.11, we can see that the dynamic relationships from phonemes to AUs become more important on the challenging subset in the *DBN* model, i.e. more temporal links are learned from the phoneme node of the $t-1^{th}$ slice to AU nodes of the t^{th} slice in Fig. 3.11. This is because the labeling uncertainty of AUs is alleviated in the challenging subset, especially for non-frontal faces, since lip movement is often asymmetrical during speech [15].

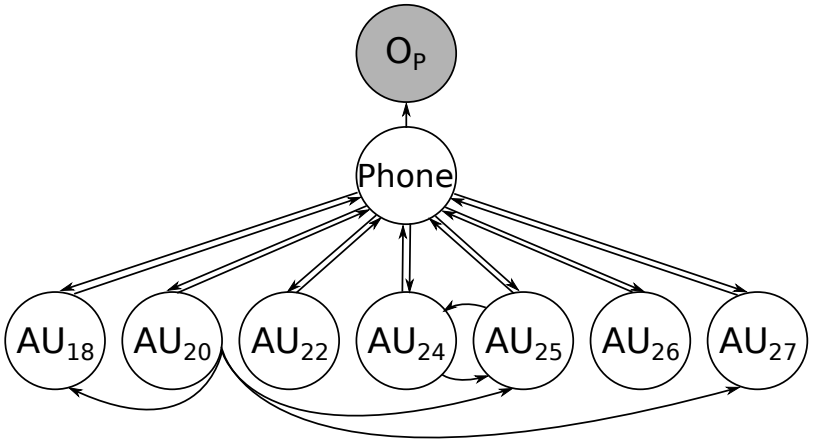

Figure 3.12 A *CTBN-F* model trained on the challenging subset for modeling the dynamic physiological relationships between AUs and phonemes.

Experimental Results and Discussion

Quantitative experimental results are reported in Fig. 3.13 in terms of F1 score, true positive rate, and false positive rate. As shown in Fig. 3.13, the proposed *CTBN* achieved the best recognition performance among all the methods compared, in terms of the average F1 score (**0.682**).

Note that the performance of the visual-based methods degrades significantly on the challenging subset even with more training data. As shown in Table 3.2, the average F1 score of *Ada-LBP* decreases from **0.416** (clean) to **0.372** (challenging); that of *Ada-LPQ* decreases from **0.448** (clean) to **0.362** (challenging); that of *SVM-LGBP* decreases from **0.386** (clean) to **0.339** (challenging); and that of *IB-CNN-LIP* drops from **0.465** to **0.382** due to large face pose variations and occlusions on the face regions. In contrast, the information extracted from the audio channel is robust to head movements and occlusions for facial AU recognition. As a result, the performance of the audio-based methods on the challenging subset is comparable or even slightly better than that on the clean subset because of employing additional training data, as reported in Table 3.2.

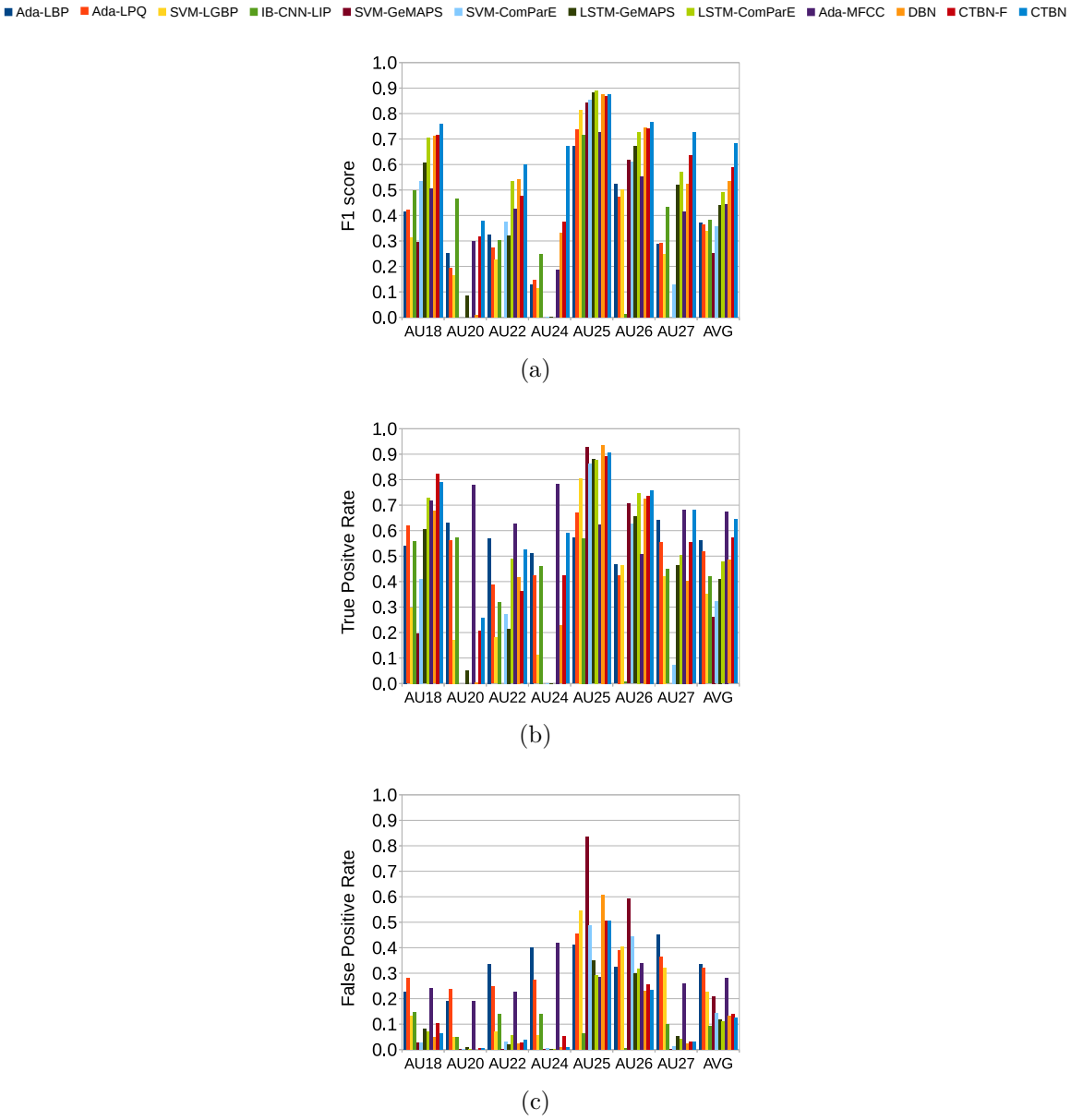

Figure 3.13 Performance comparison on the challenging subset in terms of (a) F1 score, (b) true positive rate, and (c) false positive rate for the 7 speech-related AUs.

3.3.4 Analysis on Phoneme Measurement

The proposed audiovisual fusion framework benefits from the remarkable achievements in speech recognition. In our experiments, the speech recognition performance of the Kaldi Toolkit [77] is 2.8% (15/540) on the clean subset and 2.8% (10/360) on the challenging subset in terms of the word-level error rate (WER,

Table 3.2 Performance comparison on the two subsets in terms of the average F1 score.

Approaches	**Clean**	**Challenging**
Ada-LBP [62]	0.416	0.372
Ada-LPQ [45]	0.448	0.362
SVM-LGBP [61]	0.386	0.339
IB-CNN-LIP [36]	0.465	0.382
SVM-GeMAPS [84]	0.251	0.251
SVM-ComParE [84]	0.342	0.358
LSTM-GeMAPS [84]	0.427	0.442
LSTM-ComParE [84]	0.498	0.490
Ada-MFCC [62]	0.436	0.445
DBN	0.515	0.534
CTBN-F	0.596	0.589
CTBN	**0.653**	**0.682**

[insert+delete+substitute]/[number of words]). Moreover, the phoneme recognition rates measured in phoneme error rates (phonemes that have been mis-classified / number of phonemes) are 100/1350 (7%) and 142/2025 (7%) for the clean and challenging subsets, respectively. To evaluate the effect of phoneme measurement on fusion, we have conducted an experiment using the ground-truth phoneme segments as the evidence for the CTBN model, denoted as *CTBN-perfect*. As shown in Fig. 3.15, *CTBN* using phoneme measurements from speech recognition yields comparable performance with *CTBN-perfect* using ground-truth phoneme segments.

3.4 Conclusion

It is challenging to recognize speech-related AUs due to the subtle facial appearance and geometrical changes as well as occlusions introduced by frequent lip movements. In this work, we proposed a novel audio-based AU recognition framework by exploiting information from the audio channel, i.e., phonemes in particular, because facial activities are highly correlated with voice. Specifically, a CTBN model is employed to model the dynamic and physiological relationships between phonemes and AUs,

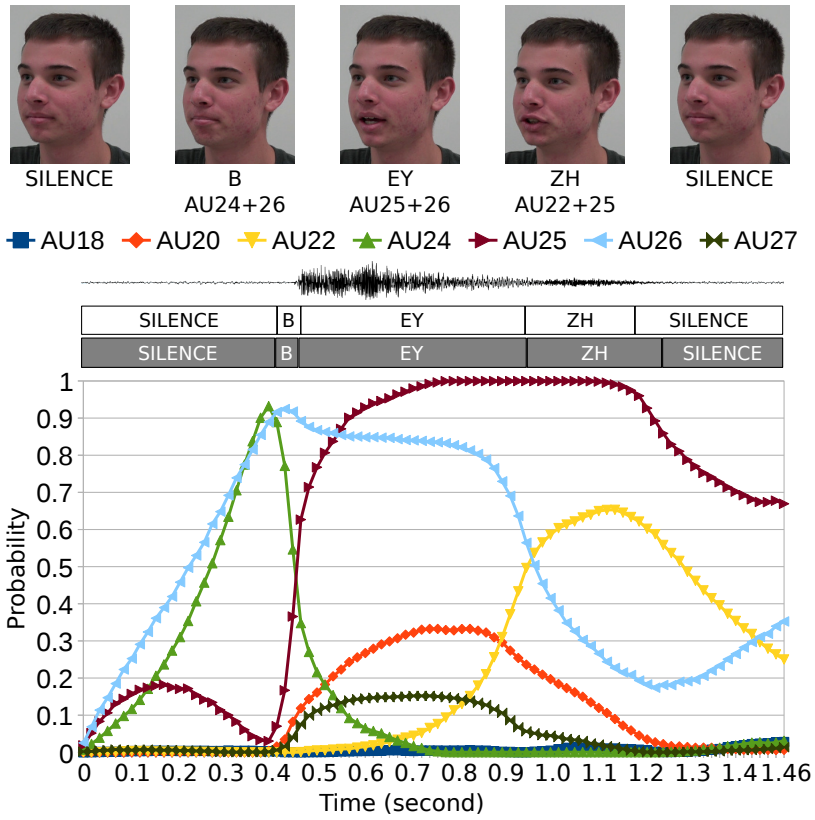

Figure 3.14 An example of the system outputs by *CTBN* inference on the challenging subset. The top row shows key frames from an image sequence where a word "beige" is produced and AU22, AU24, AU25, and AU26 are involved. The bottom figure depicts the probabilities of AUs changing over time. The shaded phoneme sequence is used as evidence of the *CTBN* and the unshaded one is the ground truth phoneme labels.. Best viewed in color.

as well as the temporal evolution of these relationships. Given the phoneme measurements, AU recognition is then performed by probabilistic inference through the CTBN model.

Experimental results on a new audiovisual AU-coded dataset have demonstrated that the CTBN model achieved significant improvement over the state-of-the-art visual-based AU recognition methods. The improvement is more impressive for those AUs, whose visual observations are impaired during speech. More importantly, the experimental results on the challenging subset have demonstrated the effectiveness of utilizing audio information for recognizing speech-related AUs under real world

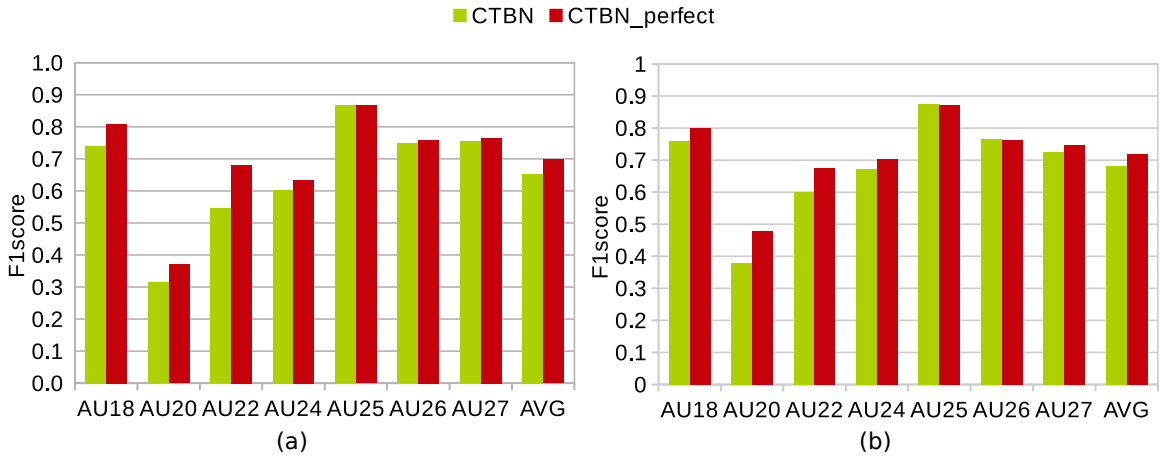

Figure 3.15 Performance comparison between *CTBN* and *CTBN-perfect* in terms of F1 score on (a) the clean subset and (b) the challenging subset.

conditions, where the visual observations are not reliable. Furthermore, the proposed CTBN model also outperformed the other baseline methods employing audio signals, thanks to explicitly modeling the dynamic interactions between phonemes and AUs in the context of human communication.

Chapter 4

Improving Speech Related Facial Action Unit Recognition by Audiovisual Information Fusion

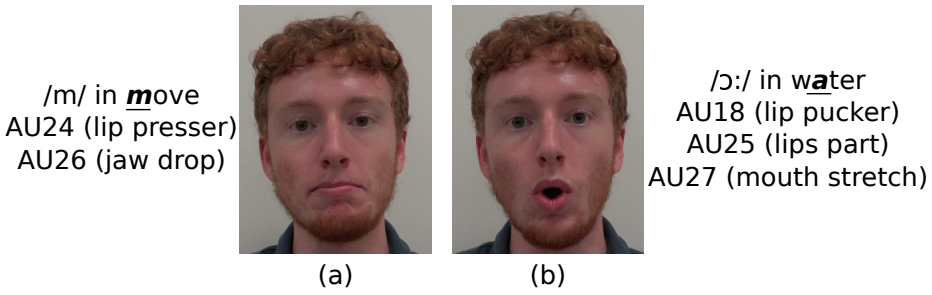

Figure 4.1 Examples of speech-related facial activities, where different AUs are activated non-additively to produce sound. (a) The gap between teeth is occluded by the pressed lips in a combination of AU24 and AU26 when sounding /m/ and (b) the space between teeth is partially occluded due to the protruded lips in a combination of AU18, AU25, and AU27 when producing /ɔː/.

4.1 Motivation

Extensive research efforts have been focused on recognizing facial AUs from static images or image sequences as discussed in the survey papers [76, 119, 86, 58]. Although great progress has been achieved on posed or deliberate facial displays, facial AU recognition suffers significantly for spontaneous facial behavior [104, 102]. Furthermore, it is extremely challenging to recognize AUs that are responsible for producing speech. During speech, these AUs are generally activated at a low intensity with subtle changes in facial appearance and facial geometry and more importantly, often produce non-additive appearance changes, which introduces challenges for detecting co-occurring AUs [23]. For instance, as illustrated in Fig. 4.1(a), recognizing AU26 (jaw drop) from a combination of AU24 (lip presser) and AU26 is almost impossible from visual observations when voicing */m/*. The reason is that the gap between teeth, which is the major facial appearance clue to recognize AU26 [23], is invisible due to the occlusion by the pressed lips. In another example, when producing */ɔː/*, as shown in Fig. 4.1(b), AU27 (mouth stretch) would probably be recognized as AU26 because the gap between teeth is partially occluded by protruding lips, i.e., AU18 (lip pucker), and thus, looks much smaller than that when only AU27 is activated. The

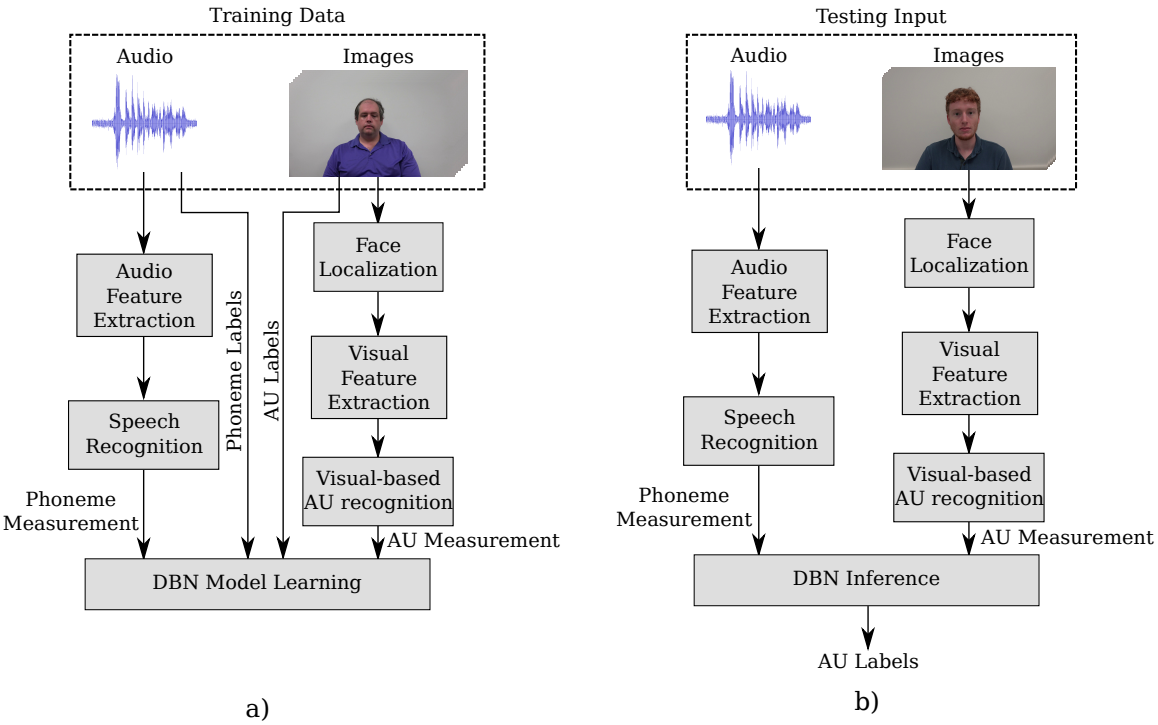

Figure 4.2 The flowchart of the proposed audiovisual AU recognition system.

failure in recognition of speech-related AUs is because information is extracted from a single source, i.e., the visual channel. As a result, all speech-related AUs are either represented by a uniform code [23, 104], i.e., AD 50, or totally ignored [102], during speech. However, it is critical to identify and differentiate the AUs that are responsible for producing voice from those for expressing emotion and intention, especially during emotional speech.

Instead of solely improving visual observations of AUs, *this work aims to explore and exploit the relationships between facial activity and voice to recognize speech-related AUs*. Specifically, there are two types of correlations between facial AUs and phonemes. First, some lower-face facial AUs and voice can be *physiologically correlated* since jaw and lower-face facial muscles are highly involved in speech production. 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 mouth as hearing "ah". Following the example of recognizing AU26 from a combination of AU24 and AU26 as illustrated in Fig. 4.1(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, facial AUs and speech can be emotionally correlated, since both facial AUs and voice/speech convey human emotions in human communications. In this work, we will focus on studying and exploiting the physiological relationships between facial AUs and speech, since these relationships are emotion and context independent, and can generalize better to various contexts.

Since speech can be represented by a sequence of phonemes, each of which is defined as the smallest sound unit in a language, *the relationships between AUs and phonemes will be investigated and explicitly modeled* in this work. Specifically, a phoneme is usually produced by combinations of AUs; and, more importantly, different combinations of AUs are activated sequentially to produce different phases for the same phoneme. For example, */b/* is produced in two consecutive stages, i.e., *closure* and *release*, where AU24 + AU26 and AU25 (lips part) + AU26 are activated, sequentially. Because the physiological relationships between AUs and phonemes are dynamic and stochastic, varying in subjects and languages, we propose to systematically and probabilistically model these relationships by a dynamic Bayesian network (DBN).

Fig. 4.2 depicts the flowchart of the proposed audiovisual AU recognition system. During training (Fig. 4.2(a)), a DBN model is learned from the ground truth labels of AUs and phonemes to capture the physiological relationships between AUs and phonemes as well as modeling measurement uncertainty of AUs and phonemes. For online AU recognition (Fig. 4.2(b)), AU measurements obtained by visual-based AU recognition and phoneme measurements obtained by speech recognition are employed as evidence for the DBN model. Then, AU recognition is performed by audiovisual

information fusion via DBN inference.

In summary, our work has two major contributions.

- A novel audiovisual AU recognition framework is proposed to make the best use of visual and acoustic cues, as humans do naturally.
- Semantic and dynamic physiological relationships between AUs and phonemes as well as measurement uncertainty of AUs and phonemes are systematically modeled and explicitly exploited by a DBN model to improve AU recognition.

Experimental results on a pilot audiovisual AU-coded database [63] demonstrate that the proposed framework yields significant improvement for speech-related AU recognition compared with the state-of-the-art visual-based methods, especially for those AUs, whose visual observations are severely impaired during speech. Moreover, the proposed method also outperforms the audio-based methods and the feature-level fusion methods, owing to explicitly utilizing the semantic and dynamic physiological relationships between facial AUs and phonemes.

4.2 Methodology

4.2.1 Modeling Semantic Physiological Relationships between Phonemes and AUs

A phoneme is defined as the smallest sound unit in a language. In this work, a phoneme set defined by the CMU pronouncing dictionary (CMUdict) [106] is employed, which is developed for speech recognition and describes North American English words using 39 phonemes [1]. Since each phoneme is anatomically related to a

[1] According to the CMUdict, the phoneme set and example words, given in the parenthesis, are listed as follows: *AA* (***o***dd), *AE* (***a***t), *AH* (h***u***t), *AO* (***aw***ful), *AW* (c***ow***), *AY* (h***i***de), *B* (***b***e), *CH* (***ch***eese), *D* (***d***ee), *DH* (***th***ee), *EH* (***E***d), *ER* (h***ur***t), *EY* (***a***te), *F* (***f***ee), *G* (***g***reen), *HH* (***h***e), *IH* (***i***t), *IY* (***ea***t), *JH* (***g***ee), *K* (***k***ey), *L* (***l***ee), *M* (***m***e), *N* (***kn***ee), *NG* (pi***ng***), *OW* (***oa***t), *OY* (t***oy***), *P* (***p***ee), *R* (***r***ead), *S* (***s***ea), *SH* (***sh***e), *T* (***t***ea), *TH* (***th***eta), *UH* (h***oo***d), *UW* (t***wo***), *V* (*v***ee**), *W* (***w***e), *Y* (***y***ield), *Z* (***z***ee), *ZH* (sei***z***ure) [106].

Figure 4.3 Examples of the semantic physiological relationships between phonemes and AUs. To produce a word *beige*, different combinations of AUs are activated successively.

specific set of jaw and lower-face muscular movements, different combinations of AUs are activated to produce different phonemes. Taking the word *beige* for example, a combination of AU24 and AU26 is first activated to produce the *Closure* phase of phoneme B (Fig. 4.3a), and then a combination of AU25 and AU26 is activated in the *Release* phase of B and move on to sound EY (Fig. 4.3b). Finally, AU26 is released and AU22 and AU25 are responsible for sounding ZH (Fig. 4.3c). These physiological relationships are "semantic" because they are context-dependent. For example, the physiological relationships learned and modeled in this work depend on North America English phonetics, where legal words are constructed by producing phonemes sequentially based on certain phonetics rules.

These semantic physiological relationships are stochastic and vary among subjects. For example, AU20 (lip stretcher) is responsible for producing AE in ***a***t based on Phonetics [13]; while some subjects do not activate AU20 in practice as found in our audiovisual AU-coded database. In addition, due to the noise in both visual and audio channels, the measurements of AUs and phonemes are not perfect. Instead of employing direct mappings from phonemes to AUs, we propose to use a Bayesian network (BN) to model these semantic relationships probabilistically.

Specifically, each target AU is associated with a node having two discrete states

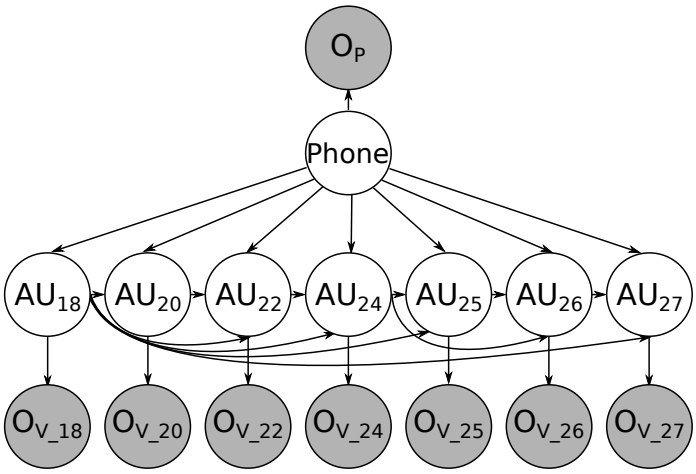

Figure 4.4 A BN models semantic physiological relationships between AUs and phonemes as well as the relationships among AUs.

$\{0, 1\}$ representing its *absence* or *presence* status. Phonemes are defined as unique acoustic events and are mutually exclusive at the same time. During speech, a set of phonemes are produced sequentially to speak a meaningful word; and the same phoneme would not repeat itself in two consecutive sound events. Thus, we employ a single node *Phone* with 29 discrete states representing 28 phonemes, which are involved in producing the words in our audiovisual database, plus one *silence* state denoted as *SIL*. Then, the node *Phone* can be in one of its 29 states at a certain time to ensure the mutually exclusive relationships among the phonemes.

Learning Semantic Physiological Relationships

Given a complete training set including the groundtruth labels of AUs and phonemes, a K2 algorithm [19], implemented in the Bayes Net Toolbox (BNT) [64], is employed to learn the semantic physiological relationships between *Phone* and the AU nodes. For a node X_i, the K2 algorithm finds the set of parents, denoted as $Par(X_i)$ by maximizing the following function [19]:

$$f(X_i, Par(X_i)) = \prod_{j=1}^{M_i} \frac{(K_i - 1)!}{(N_{ij} + K_i - 1)!} \prod_{k=1}^{K_i} N_{ijk}! \tag{4.1}$$

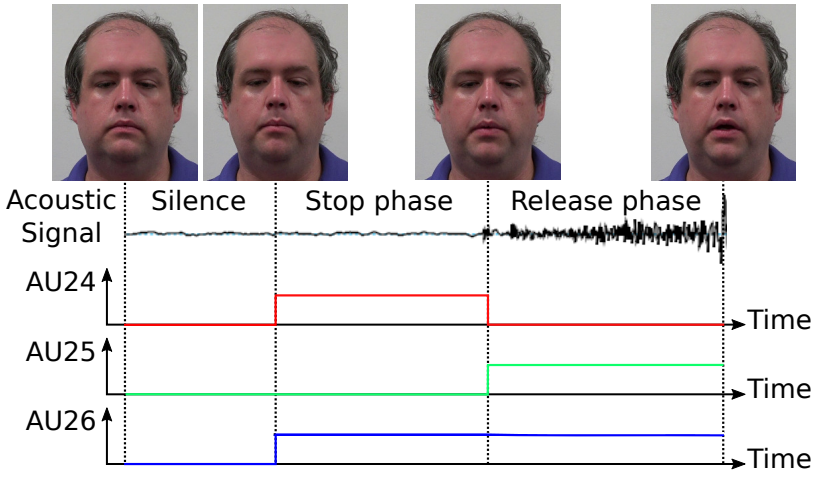

Figure 4.5 Illustration of the dynamic relationships between AUs and the phoneme while producing B, where on-axis and off-axis colored lines represent *absence* and *presence* of the corresponding AUs, respectively. Best viewed in color.

where K_i is the number of all possible states that X_i may take; M_i is the number of all possible configurations of the parents of X_i; N_{ijk} denotes the number of instances when X_i is in its k^{th} state and its parents take the j^{th} configuration; and $N_{ij} = \sum_{k=1}^{K_i} N_{ijk}$.

A BN model learned by the K2 algorithm is shown in Fig. 4.4, where the nodes represent random variables (*Phone* and AUs) and the directed links between them represent the conditional dependency. Particularly, the links between *Phone* and the AU nodes capture their semantic physiological relationships. In addition, since AUs are activated in combinations to produce a meaningful sound, the relationships among AUs are also captured in the BN model by the directed links among them.

4.2.2 Modeling Semantic and Dynamic Relationships using A Dynamic Bayesian Network

By studying Phonetics [13], we know that there are strong physiological relationships between AUs and phonemes. More importantly, these relationships also undergo a temporal evolution. *In particular, there are two kinds of dynamic relationships*

between AUs and phonemes.

On the one hand, as the facial muscular movements are activated before a sound is generated [35, 95, 16], the probabilities of the AUs being activated increase and reach an apex as the phoneme is fully made, and then decrease while preparing to produce the next phoneme. On the other hand, different combinations of AUs are activated in different phases for sounding a single phoneme. For example, as illustrated in Fig. 4.5, the phoneme B in ***b***e has two sequential phases. In the first phase, i.e. the *Closure* phase, the lips are pressed together as activating AU24; the upper and the lower front teeth are usually parted as activating AU26; and "the breath is held and compressed" [13] without emitting sound, i.e., the *Phone* node is in its silence state *SIL*. As a result, the lip movements (AU24 and AU26) occur earlier than the sound can be heard. In the second phase, i.e. the *Release* phase, the lips part by activating AU25 and the compressed breath is released suddenly as releasing AU24 [13]. Therefore, the physiological relationships between *Phone* and AUs change over time. In addition, since the duration of the *Closure* phase varies across different subjects and different words, these dynamic relationships are stochastic. Such semantic and dynamic relationships can be well captured by extending the BN (Fig. 4.4) to a DBN, which not only models the temporal evolution of AUs and phonemes but also models the temporal dependencies among them.

Learning Dynamic Physiological Relationships

Given a complete set of training data sequences $\mathcal{D} = \{D_1, \ldots, D_S\}$, the dynamic dependencies among AUs and phonemes, i.e., the transition model of a DBN, can be learned by maximizing the following score function

$$Score(B_{tr}) = \log\, p(\mathcal{D}|B_{tr}) + \log\, p(B_{tr}) \tag{4.2}$$

where B_{tr} is a candidate structure of the transition model, and the two terms are the log likelihood and the log prior of B_{tr}, respectively. For a large data set, the first

term can be approximated using Bayesian information criterion (BIC) [87] as:

$$\log p(\mathcal{D}|B_{tr}) \approx \log p(\mathcal{D}|\hat{\theta}_{B_{tr}}, B_{tr}) - \frac{q}{2}\log S \tag{4.3}$$

where $\theta_{B_{tr}}$ is a set of model parameters; $\hat{\theta}_{B_{tr}}$ is the maximum likelihood estimation of $\theta_{B_{tr}}$; q is the number of parameters in B_{tr}; and S is the number of data samples in $\mathcal{D}$. In Eq. 4.3, the first term gives the log maximum likelihood of B_{tr} and the second term penalizes the model complexity. In this way, we can learn a DBN model as shown in Fig. 4.6a.

Specifically, there are two types of dynamic links: self-loops and directed links across two time slices, i.e., from the $(t-1)^{th}$ time slice to the t^{th} time slice. The self-loop at each AU node represents the temporal evolution of each AU; while the self-loop at the *Phone* node denotes the dynamic dependency between two phonemes. For instance, a consonant is followed by a vowel at the most of time, and hence the probability of a consonant followed by a vowel is much higher than that of a consonant followed by another consonant. The directed links across two time slices characterize the dynamic dependency between two variables, e.g., the dynamic physiological relationships between phonemes and AUs as well as the dynamic relationships among AUs.

Incorporating domain knowledge in the DBN Structure: As shown in Fig. 4.6, the dynamic dependency between AUs in the $(t-1)^{th}$ time slice and *Phone* in the t^{th} time slice, however, are not learned from data. This is because the penalty, i.e. the second term in Eq. 4.3, for adding a link from an AU node to *Phone* is much higher than that from *Phone* to AU for the 29-state *Phone* node. Therefore, we refine the learned DBN model by combining the expert knowledge, i.e., the facial muscular movements are activated before sounding a phoneme [35, 95, 16]. Specifically, the dynamic links from AUs to *Phone* across two successive time slices are manually added as depicted by the red dashed links in Fig 4.6b.

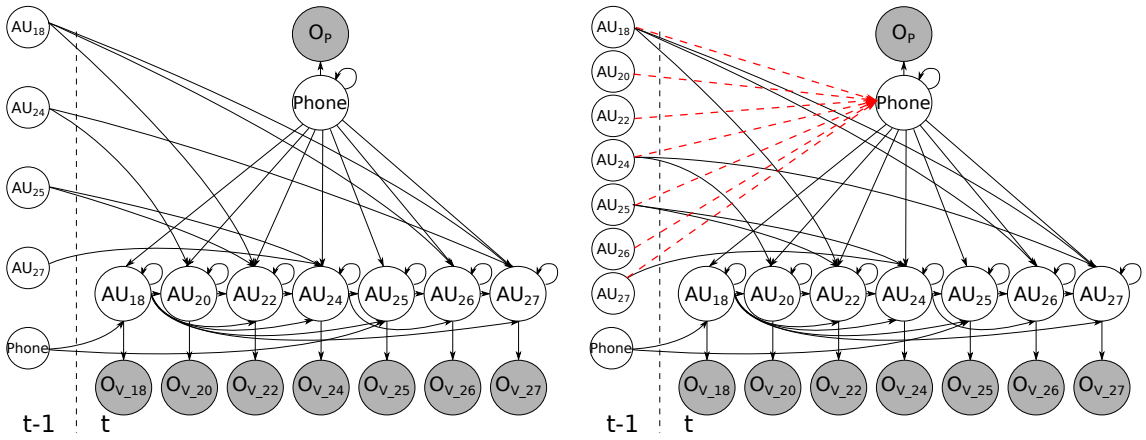

Figure 4.6 A DBN model for audiovisual AU recognition: (a) the DBN structure learned from data, and (b) the DBN structure by integrating expert knowledge into the learned structure. Shaded nodes are the measurement nodes for the corresponding AU nodes and the phoneme node *Phone*, respectively. The links between the unshaded nodes and the shaded nodes characterize the measurement uncertainty.

As shown in Fig. 4.6b, a comprehensive DBN model is constructed. There are two types of nodes in the DBN model: measurement nodes and hidden nodes. The measurement nodes, denoted by the shaded nodes, represent the measurements of AUs denoted by $\mathbf{O}_v$ and the measurement of the phoneme denoted by $\mathbf{O}_p$, whose states can be obtained by visual-based AU recognition and speech recognition, respectively. The hidden nodes are denoted by the unshaded nodes, whose states need to be estimated via probabilistic inference. This DBN model is capable of modeling various interactions in the scenario of audio-visual AU recognition including the semantic and dynamic physiological relationships between AUs and phonemes, semantic and dynamic relationships among AUs, the dynamic relationships between different phonemes, the temporal evolution of AUs, as well as measurement uncertainty.

4.2.3 Learning Model Parameters

Given the model structure as shown in Fig. 4.6b, the DBN parameters, specified as a set of conditional probabilistic tables (CPTs) associated with each node, can be learned from a set of training data $\mathcal{D} = \{D_1, D_2, ..., D_S\}$. The DBN can be considered as an expanded BN consisting of two time slices of static BNs connected by dynamic links. Hence, in addition to learning the CPTs within the same time slice as that does for a static BN, the transition probabilities associated with the dynamic links are also learned. Since the training data is complete in this work, the parameters of the DBN can be estimated using Maximum Likelihood estimation (MLE).

4.2.4 Audiovisual AU Recognition via DBN Inference

Given all available observations from both visual and audio channels until the t^{th} time slice, i.e., $\mathbf{O}_v^{1:t}$ and $\mathbf{O}_p^{1:t}$, AU recognition can be performed through probabilistic inference via the DBN model. Specifically, the posterior probability of the target AUs given all the observations, i.e., $p(\mathbf{AU}^t|\mathbf{O}_v^{1:t}, \mathbf{O}_p^{1:t})$, where $\mathbf{AU}^t$ represents all target AUs at the t^{th} time slice, can be factorized and computed by DBN inference. In this work, a forward-backward inference algorithm implemented in the BNT is employed [64]. Then, predictions of the target AUs can be made by maximizing the posterior probability:

$$\mathbf{AU}^{t*} = \underset{\mathbf{AU}^t}{\operatorname{argmax}}\, p(\mathbf{AU}^t|\mathbf{O}_v^{1:t}, \mathbf{O}_p^{1:t}) \tag{4.4}$$

4.3 Measurements Acquisition

To perform the probabilistic inference using the DBN model, the measurements of AUs and the phoneme at each time slice are required. However, signals in different channels are usually sampled at different time scales. For example, the images are sampled at 60 frame per second (fps) and audio tracks are continuous in our audio-

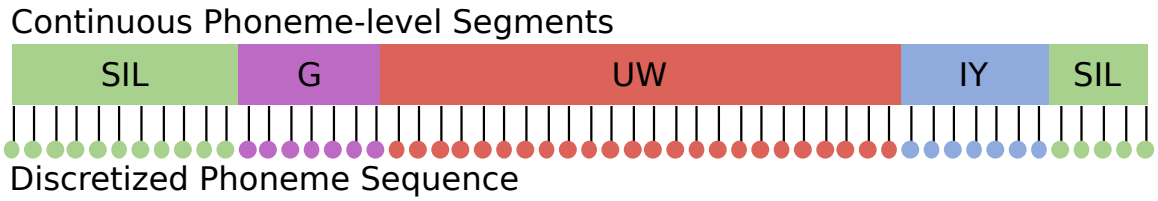

Figure 4.7 Aligning continuous phoneme segments with image frames for the word *gooey*. The top row gives the phoneme-level segments obtained by Kaldi toolkit [77]. The bottom row depicts the discretized sequence of phoneme measurements, where the same color indicates the same phoneme in the continuous phoneme-level segments. The vertical lines in-between represent a sequence of image frames, to which the phonemes will be aligned. Best viewed in color.

visual database. Here, we show how to get AU measurements and phoneme measurements and how to align the measurements from two channels frame by frame.

4.3.1 Extracting AU Measurements

To illustrate *the proposed method can be build upon any advanced visual-based facial AU recognition method and achieve improvement on speech-related AU recognition*, two different state-of-the-art visual based AU recognition methods are employed to extract AU measurements.

Extracting AU Measurements using LBP

In this work, a state-of-the-art visual-based AU recognition method is adopted [102] to obtain AU measurements. For preprocessing purpose, the face regions across different facial images are aligned to remove scale and positional variance [5] and then cropped to 96×64, which are further divided into 7×7 grid. From each grid, LBP histograms with 59 bins are extracted and then, concatenated into a single vector, which is denoted as LBP feature. For each target AU, an AdaBoost classifier is employed to select the most discriminative features from the LBP feature pool and construct a strong classifier to perform facial AU recognition for each target AU. The binary

classification results obtained from AdaBoost classifiers will be fed into the DBN model as the AU measurements.

Extracting AU Measurements using LPQ

The second state-of-the-art visual-based AU recognition method adopted to obtain AU measurements in this work is LPQ [45]. After the same preprocessing operations as described in Section. 4.3.1, the facial images are divided into 7×7 grid. From each grid, LPQ histograms with 256 bins are extracted and then, concatenated into a single vector, which is denoted as LPQ feature. AdaBoost classifier is utilized to select the most discriminative features and construct a strong classifier, which is employed to produce binary classification results as the measurements for the DBN model for each target AU.

4.3.2 Extracting Phoneme Measurements

In this work, a state-of-the-art speech recognition method, i.e., Kaldi toolkit [77], is employed to obtain the phoneme measurements. Specifically, 13-dimensional Mel Frequency Cepstral Coefficients (MFCCs) [20] features are extracted and employed in Kaldi to get word-level speech recognition results, which are further divided into phoneme-level segments as shown in Fig. 4.7. In order to obtain a phoneme measurement for each time slice, which should be also synchronized with the AU measurements, the continuous phoneme segments are discretized according to the sampling rate of the image frames, i.e., 60 fps in our experiment. As illustrated in Fig. 4.7, the first row shows the continuous phoneme-level segments for the word *chaps*; the second row shows a sequence of image frames to be aligned to; and the last row shows the frame-by-frame phoneme measurements, which are synchronized with the image frames and will be fed into the DBN model as the measurements for *Phone* for audiovisual AU recognition.

4.4 Experiments

4.4.1 Methods in Comparison

Experiments have been conducted on the new audiovisual AU-coded dataset to illustrate the effectiveness of the proposed framework, as shown in Fig. 4.6, in improving the recognition performance for speech-related AUs. In this work, we built the proposed approach upon two state-of-the-art visual-based methods, i.e. LBP-based and LPQ-based methods, denoted as *DBN-LBP-AV+E* and *DBN-LPQ-AV+E*[2], respectively. Each method is first compared with the state-of-the-art visual-based methods utilizing the same features. Furthermore, they are compared with three baseline audiovisual fusion methods to demonstrate the effectiveness of explicitly modeling and employing the semantic and dynamic physiological relationships between AUs and phonemes in audiovisual fusion. The *LBP-based baseline methods* are described as follows.

Ada-LBP-V employs a state-of-the-art visual-based AU recognition approach [102] using LBP features, as described in Section 4.3.

DBN-LBP-V employs a state-of-the-art DBN-based model [101] to model the relationships among AUs and used as a dynamic visual-based baseline. The structure of *DBN-LBP-V* is obtained by eliminating the "**Phone**" node and its measurement node from the *DBN-LBP-AV+E* structure depicted by Fig. 4.6.

Ada-LBP-AV developed in our early work [62], employs a *feature-level fusion scheme* that extracts features from both visual and audio channels, i.e. histograms of LBP features and MFCC features. Specifically, given an input wave file, MFCCs are extracted using window size $l = 16.67ms$ with a frame-shift $s = 16.67ms$ by Kaldi toolkit [77]. To include more temporal information, 7 frames, i.e. 3 frames before and after the current frame along with the current one, are concatenated as the

[2]The suffix *-V* indicates the visual-based approach; the suffix *-AV* means the audiovisual fusion method; and the suffix *+E* indicates the integration of expert knowledge in the DBN model.

final MFCC feature for each frame. The extracted LBP and MFCC features are integrated into a single feature vector and employed as the input for AdaBoost to make predictions for the target AU.

BN-LBP-AV employs a static BN model with a structure illustrated in Fig. 4.4 plus measurement nodes for all AUs and the *Phone* node. The *BN-LBP-AV* only considers the semantic relationships between AUs and phonemes as well as the semantic relationships among AUs, while the dynamics of AUs and phonemes are ignored.

DBN-LBP-AV employs the learned DBN structure denoted by the solid links in Fig. 4.6, which does not model the dynamic dependencies between AUs in the $(t-1)^{th}$ time slice and phonemes in the t^{th} time slice.

The LPQ-based baseline methods are defined in the same way, whose model structures are the same as those of the LBP-based equivalents. For all methods evaluated, a leave-one-subject-out training/testing strategy is employed, where the data from one subject is used for testing and the remaining data is used for training.

4.4.2 Experimental Results and Data Analysis on the Clean Subset

We first evaluate the proposed *DBN-LBP-AV+E* and *DBN-LPQ-AV+E* on the *clean* subset.

Experimental Results of LBP-based methods

Quantitative experimental results of LBP-based methods are reported in Fig. 4.8 in terms of F1 score, false positive rate (FPR), and true positive rate (TPR) for the 7 speech-related AUs[3]. As shown in Fig. 4.8, all the audiovisual fusion methods outperform the static visual-based method, i.e., *Ada-LBP-V*. Specifically, the overall recognition performance is improved from **0.416** using the *Ada-LBP-V* to **0.463** (*Ada-LBP-AV*), **0.658** (*BN-LBP-AV*), **0.666** (*DBN-Learned*), and **0.696** by

[3]The results are obtained on all the predictions using a default threshold, i.e. 0 for AdaBoost-based methods, and 0.5 for graphical model-based methods.

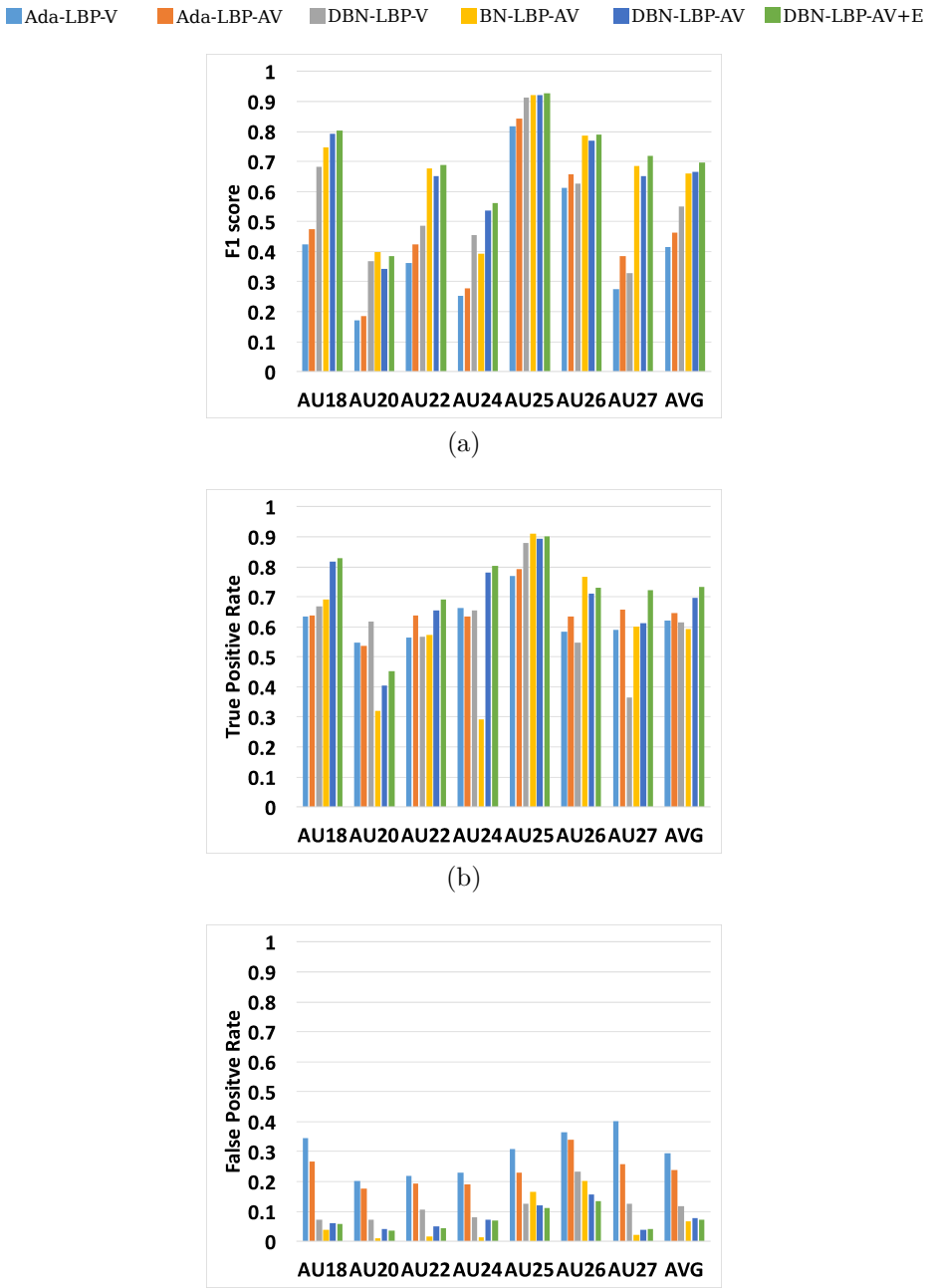

Figure 4.8 Performance comparison of AU recognition on the clean subset in terms of (a) F1 score, (b) true positive rate, and (c) false positive rate.

the proposed *DBN-LBP-AV+E*, in terms of the F1 score, which demonstrates that information from the audio channel indeed helps the recognition of speech-related AUs.

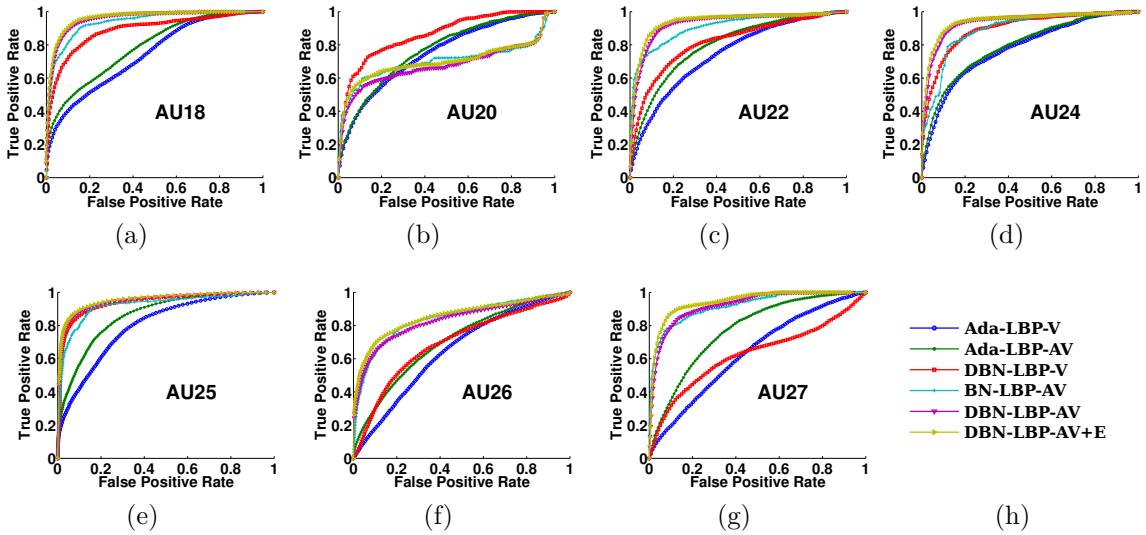

Figure 4.9 ROC curves for 7 speech-related AUs on the clean subset using LBP features. Best viewed in color.

All the fusion methods, except the feature level fusion method, i.e. *Ada-LBP-AV*, perform better than the dynamic visual-based method, i.e. *DBN-LBP-V*. The performance of *Ada-LBP-AV* is inferior to that of *DBN-LBP-V* because subjects in the *clean* subset were asked to produce the words and display the lip movements clearly, and thus the relationships between AUs, explicitly modeled by *DBN-LBP-V*, are strong. Furthermore, as shown in Fig. 4.8, the proposed *DBN-LBP-AV+E* framework outperforms all methods compared in terms of the F1 score (**0.696**), the FPR (**0.071**), and the TPR (**0.732**). In the following, we will compare the proposed *DBN-LBP-AV+E* with each baseline method side by side.

Comparison of *Ada-LBP-V*, *DBN-LBP-V*, and *DBN-LBP-AV+E* The proposed *DBN-LBP-AV+E* model significantly outperforms the state-of-the-art visual based method *Ada-LBP-V* and *DBN-LBP-V* on all target AUs. Notably, *DBN-LBP-AV+E* drastically improves the recognition performance of AU26 (jaw drop) and AU27 (mouth stretch). For example, the F1 score of AU27 increases from **0.273** (*Ada-LBP-V*) and **0.328** (*DBN-LBP-V*) to **0.720** (*DBN-LBP-AV+E*). The perfor-

mance improvement is primarily because of the integration of audio information in recognition of these AUs. As shown in Fig. 4.1, the visual cues for recognizing AU26 and AU27 are severely impaired by occlusions introduced by the presence of other AUs during speech. However, the information from the audio channel is not affected and thus, more reliable.

Comparison between *Ada-LBP-AV* and *DBN-LBP-AV+E* As shown in Fig. 4.8, *DBN-LBP-AV+E* outperforms *Ada-LBP-AV*, a feature-level fusion method, for all target AUs. Specifically, the F1 score is improved from **0.463** (*Ada-Fusion*) to **0.696** (*DBN-LBP-AV+E*). The performance improvement mainly comes from two aspects. First, the proposed *DBN-LBP-AV+E* benefits from the remarkable achievements in speech recognition: the speech recognition performance of the Kaldi Toolkit [77] in terms of the word-level and phoneme-level error rates are 2.8% and 7%, respectively, on the *clean* subset in our experiments. Hence, it makes more sense to employ accurate phoneme measurements than to use low-level acoustic features directly. Second, it is more effective to explicitly model and exploit the semantic and dynamic physiological relationships between AUs and phonemes than to employ the audiovisual features directly. For example, the TPR of AU26 increases from **0.635** (*Ada-LBP-AV*) to **0.730** (*DBN-LBP-AV+E*) with a drastic decrease in the FPR from **0.338** (*Ada-LBP-AV*) to **0.136** (*DBN-LBP-AV+E*) by employing the physiological relationships between the phonemes and AUs as shown in Fig. 4.1.

Comparison between *BN-LBP-AV* and *DBN-LBP-AV+E* Both AUs and phonemes are dynamic events and their dynamics are crucial in natural communications. By modeling both the semantic and dynamic relationships between AUs and phonemes, the recognition performance of using *DBN-LBP-AV+E* is better than that of using *BN-LBP-AV* for all target AUs in terms of all metrics. For example, there are strong dynamic relationships between phonemes and AU24 (lip presser), e.g., AU24 is activated in the *Stop* phase of *B* in ***b***e before the sound is emitted,

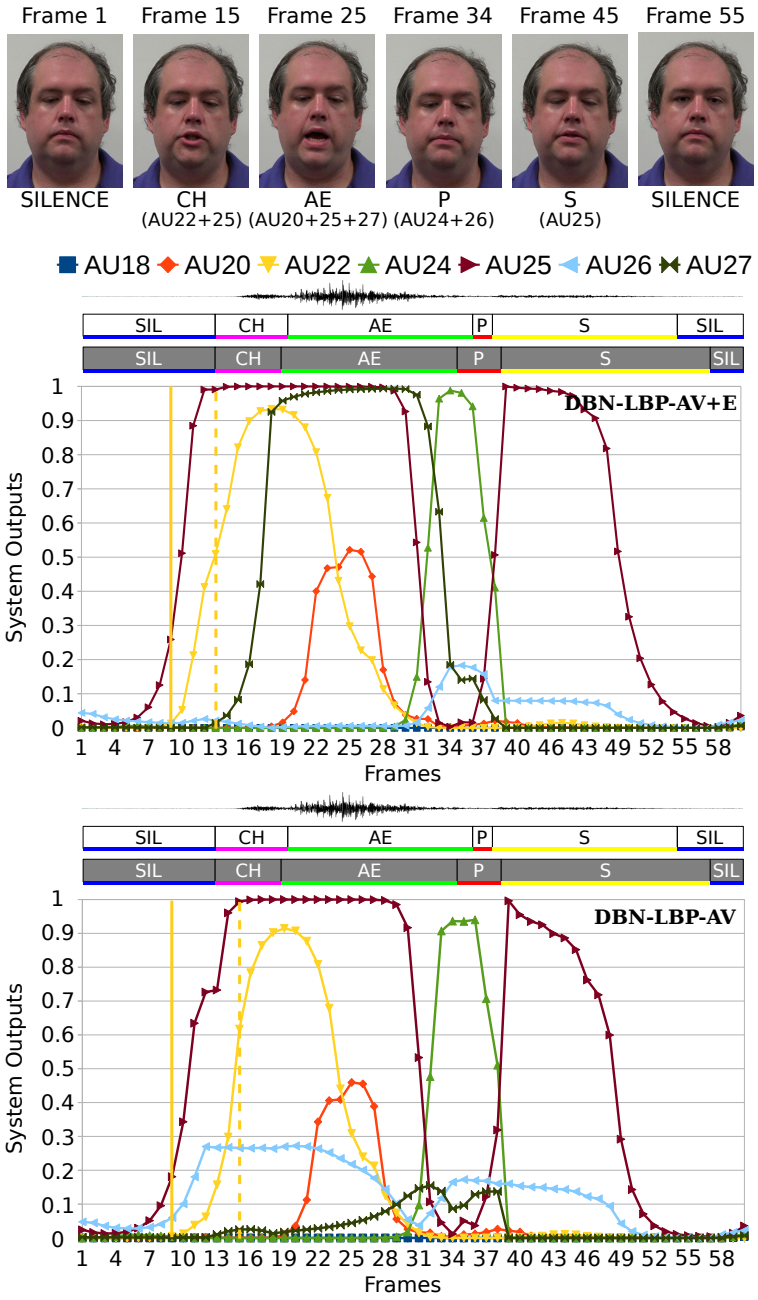

Figure 4.10 An example of the system outputs by DBN inference using *DBN-LBP-AV+E* and *DBN-LBP-AV*, respectively. A word *chaps* is produced and AU20, AU22, AU24, AU25, and AU27 have been activated. The top row shows key frames from the image sequence, as well as the AU combinations for producing the corresponding phonemes. The two bottom figures depict the probabilities of AUs estimated by *DBN-LBP-AV+E* and *DBN-LBP-AV*, respectively. The unshaded phoneme sequence is the ground truth and the shaded one represents the evidence utilized by DBN models. The dashed and solid vertical lines denote the ground truth and the predicted time point, where AU22 is activated, respectively. The dashed vertical line is closer to the solid vertical line in *DBN-LBP-AV+E*. Best viewed in color.

as depicted in Fig. 4.5. As shown in Fig. 4.8, the recognition performance of AU24 gains a significant improvement using *DBN-LBP-AV+E*: the F1 score increases from **0.394** (*BN-LBP-AV*) to **0.560** (*DBN-LBP-AV+E*).

Comparison between *DBN-LBP-AV* and *DBN-LBP-AV+E* Since AUs are the major mechanism to produce the voice, they are activated before the sound is produced [35]. As shown in Fig. 4.8, the *DBN-LBP-AV+E* outperforms the *DBN-LBP-AV* for all target AUs in terms of F1 score, which demonstrates the effectiveness of integrating this expert knowledge into the learned DBN model.

In addition to the three metrics, an ROC analysis is conducted for each AU to further demonstrate the performance of the proposed framework. As show in Fig. 4.9, each ROC curve is obtained by plotting the TPR against FPR at different thresholds over the predicted scores. The performance of the proposed *DBN-LBP-AV+E* model is better or at least comparable with that of the baseline methods on all the target AUs except for AU20. As shown in Fig. 4.9, the performance of the fusion-based methods is inferior to that of the dynamic visual-based method, i.e., *DBN-LBP-V*, because the relationships between AU20 and the phonemes are weak due to large variations among subjects. For example, although AU20 is responsible to producing the phoneme AE in *ch**a**ps* according to Phonetics [13], some subjects did not activate AU20 during speech.

Furthermore, Fig. 4.10 gives an example of the system outputs, i.e., the estimated probabilities of AUs, by *DBN-LBP-AV+E* and *DBN-LBP-AV*, respectively. As shown in Fig. 4.10, when sounding a word *chaps*, the probabilities of AUs increase when they are preparing to sound a phoneme and decrease rapidly after the sound is emitted. As the facial movements are activated before the sound is generated, the probability of AU22 increases and reaches above 0.5, i.e. the activation threshold, before the phoneme *CH* is made. The solid vertical yellow line represents the onset time point of AU22 labeled manually, while the dashed line shows the estimated activation time

Table 4.1 Performance comparison on the two subsets in terms of the F1 score.

Subsets	*Ada-LBP-V*	*Ada-LBP-AV*	*DBN-LBP-V*	*BN-LBP-AV*	*DBN-LBP-AV*	*DBN-LBP-AV+E*
Clean	0.416	0.463	0.551	0.658	0.666	**0.696**
Challenging	0.372	0.448	0.368	0.608	0.548	**0.622**
Subsets	*Ada-LPQ-V*	*Ada-LPQ-AV*	*DBN-LPQ-V*	*BN-LPQ-AV*	*DBN-LPQ-AV*	*DBN-LPQ-AV+E*
Clean	0.448	0.482	0.536	0.651	0.651	**0.679**
Challenging	0.362	0.430	0.370	0.619	0.579	**0.651**

of AU22. The closer those two lines are, the better the prediction is. By integrating the expert knowledge, *DBN-LBP-AV+E* can better predict the activation of an AU given the measurements. In another example, AU27 was not detected by *DBN-LBP-AV* shown in the bottom plot, because the visual-based classifier fails to detect AU27. However, the dynamic relationships between AUs and phonemes utilized by *DBN-LBP-AV+E* can help to predict AU27 despite the poor visual measurements.

Experimental Results of LPQ-based methods

The quantitative experimental results using LPQ features are reported in Table 4.1 in terms of the F1 score. Not surprisingly, the proposed *DBN-LPQ-AV+E* outperforms all the compared methods in terms of F1 score (**0.679**).

The performance improvement on both the LBP-based method and the LPQ-based method demonstrates that the proposed method can be built upon any advanced visual-based AU recognition method, and consistently improve the performance for speech-related AUs recognition.

4.4.3 Experimental Results and Data Analysis on the Challenging Subset

Experiments were conducted on the *challenging* subset to further demonstrate the advantage of integrating audio information with visual cues for speech-related AU recognition in the wild, where the AU recognition system is challenged by large head movements and occlusions introduced by facial hair and accessaries.

The proposed *DBN-LBP-AV+E* and *DBN-LPQ-AV+E* models and the baseline methods were trained and tested on the *challenging* subset using a leave-one-subject-out strategy. Since the *challenging* subset contains only 6 subjects, the data in the *clean* subset was employed as additional training data, except those of the two subjects who also appear in the *challenging* subset to ensure a subject-independent context. In particular, the data of 5 subjects from the *challenging* subset along with the data of 7 subjects from the *clean* subset is used for training, while the remaining one subject from the *challenging* subset is employed for testing. The structures of the *DBN-LBP-AV* and *DBN-LBP-AV+E* trained on the *challenging* subset are shown in Fig. 4.11.

Experimental Results and Discussion

Quantitative experimental results on the *challenging* subset are reported in Fig. 4.12 for LBP-based methods, in terms of F1 score, TPR, and FPR for the 7 speech-related AUs. From Fig. 4.12, we can find that all the audiovisual fusion methods outperform the methods employing only visual information (*Ada-LBP-V* and *DBN-LBP-V*)[4]. Specifically, as the head movements and occlusions are introduced in the *challenging* subset, the visual observations of AUs become unreliable, which is reflected by the drastic drop in performance of the visual-based methods, i.e. **0.372** (*Ada-LBP-V*) and **0.368** (*DBN-LBP-V*) in terms of the F1 score. In contrast, the information

[4]The speech recognition performance of the Kaldi Toolkit [77] in terms of the word-level and phoneme-level error rates are 2.8% and 7.4%, respectively, on the *challenging* subset in our experiments.

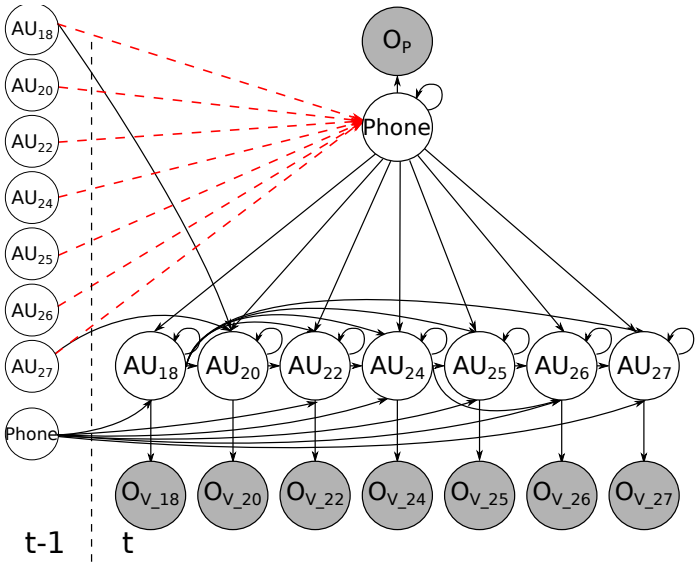

Figure 4.11 A DBN model learned from the challenging data for audiovisual AU recognition: the solid links representing the learned DBN structure and the red dashed links denoting the expert knowledge integrated into the learned structure.

extracted from the audio channel is less affected. Thus, the performance is improved from **0.372** (*Ada-LBP-V*) to **0.448** (*Ada-LBP-AV*), **0.608** (*BN-LBP-AV*), **0.548** (*DBN-LBP-AV*), and **0.622** by the proposed *DBN-LBP-AV+E* in terms of F1 score.

Experimental results for LBP-based and LPQ-based methods on both subsets are reported in Table 4.1, in terms of the F1 score. On the *challenging* subset, both *DBN-LBP-V* and *DBN-LPQ-V* have performance comparable to *Ada-LBP-V* and *Ada-LPQ-V*, respectively, since the visual observations become unreliable under challenging settings. Moreover, the dynamic dependencies between AUs and phonemes become more important. More temporal links between phonemes and AUs are learned from the challenging subset (Fig. 4.11) than those learned from the clean subset (Fig. 4.6). In addition, by incorporating the expert knowledge, the proposed *DBN-LBP-AV+E* and *DBN-LPQ-AV+E* improve the performance by **0.074** and **0.072** compared with the *DBN-LBP-AV* and *DBN-LPQ-AV*, respectively, in terms of F1 score.

A paired t-test has been conducted to demonstrate that the performance improve-

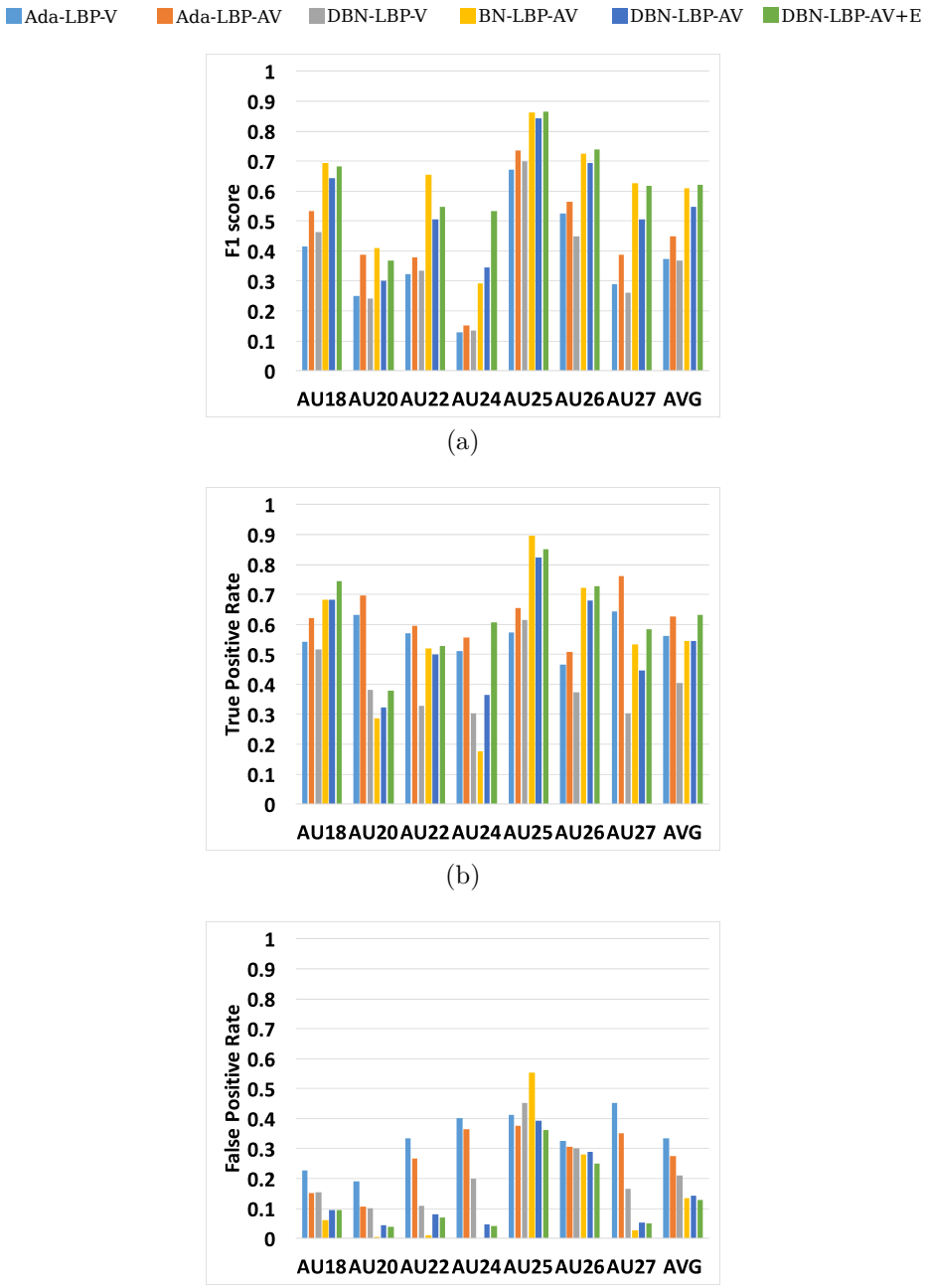

Figure 4.12 Performance comparison of AU recognition on the challenging subset in terms of (a) F1 score, (b) true positive rate, and (c) false positive rate.

ment is statistically significant. The null hypothesis is that the performances of the baseline visual-based method or the feature-level fusion method and the proposed method are "no difference". The p-values of the proposed method against the base-

Table 4.2 Performance comparison on the two subsets in terms of the F1 score.

Approaches		Clean	Challenging
Visual-based	*SVM-LGBP* [102]	0.386	0.339
	IB-CNN-LIP [36]	0.465	0.382
Audio-based	*SVM-GeMAPS* [84]	0.251	0.251
	LSTM-GeMAPS [84]	0.427	0.442
	SVM-ComParE [84]	0.342	0.358
	LSTM-ComParE [84]	0.498	0.490
	Ada-MFCC [62]	0.436	0.445
	DBN-A [63]	0.515	0.534
	CTBN-A [63]	0.653	**0.682**
Audiovisual Fusion	*DBN-LBP-AV+E*	**0.696**	0.622
	DBN-LPQ-AV+E	**0.679**	0.651

line visual-based methods or the feature-level fusion methods are all less than 0.001 on both clean and challenging subsets for LBP/LPQ-based methods. Therefore, the performance improvement of the proposed method is statistically significant over the visual-based methods, more importantly, the feature-level fusion method.

4.4.4 Comparison with More State-of-the-art Visual-based and Audio-based Methods

To further demonstrate the effectiveness of the proposed framework, two more state-of-the-art visual-based methods and five state-of-the-art audio-based methods are implemented and evaluated on the AU-coded audiovisual database using a leave-one-subject-out cross-validation strategy.

Visual-based methods

LGBP-SVM: One of the visual-based methods employs a kind of human-crafted feature, i.e. LGBP features [90, 102]. Specifically, 400 LGBP features are selected by AdaBoost and employed to train an SVM classifier for each target AU.

IB-CNN-LIP: The other visual-based method is based on a deep learning model, i.e. incremental boosting convolutional neural network (IB-CNN) [36]. Since only the lower-part of the face is responsible for producing the speech-related AUs, a two-stream IB-CNN is developed to learn both appearance and shape information, particularly, from the lip region along with the landmarks on the lips.

Audio-based methods

SVM-GeMAPS and LSTM-GeMAPS: The first two audio-based methods employ a kind of low-level acoustic feature set, i.e. GeMAPS [84], on top of which, SVMs and Long-Short Term Memory (LSTM) networks are employed to produce the predictions for target AUs, denoted as *SVM-GeMAPS* and *LSTM-GeMAPS*, respectively. Specifically, 18-dimensional GeMAPS features are extracted given the audio signals using openSMILE [25]. An SVM is trained for each target AU using LIBSVM toolkit for *SVM-GeMAPS* and an LSTM network implemented in TensorFlow library [1] is trained to learn the temporal dependencies over time for *LSTM-GeMAPS*. The employed LSTM network consists of 3 hidden layers with 156, 256, and 156 hidden units, respectively.

SVM-GeMAPS and LSTM-ComParE: The third and the fourth audio-based methods are based on another kind of low-level acoustic feature set, i.e. ComParE [84]. Similar to *SVM-GeMAPS* and *LSTM-GeMAPS*, an SVM and an LSTM network are trained for each target AU on top of the extracted 130-dimensional ComParE features, respectively.

Ada-MFCC: The fifth audio-based method is based on MFCC features. In particular, 13-dimensional MFCC features are extracted from audio signals, and a cubic spline interpolation approach [62] is employed to align the extracted features with image frames. Moreover, 3 frames before and after the current frame along with the current one, which makes it 7 frames in total are concatenated as the MFCC feature

vector for the current image frame. The final MFCC features are employed as the input to train an AdaBoost classifier for each AU.

DBN-A and CTBN-A: The last two audio-based methods are from our previous work [63], which recognize speech-related AUs using phoneme segments by modeling the relationships between facial AUs and phonemes with a DBN model and a CTBN model, respectively.

Experimental results and discussion

Experimental results of all methods in comparison can be found in Table 4.2, in terms of F1 score. *LGBP* and *IB-CNN-LIP* recognize AUs based on only visual clues, where the appearance changes for speech-related AUs are subtle, and sometimes "invisible" in the visual channel. *SVM-GeMAPS*, *LSTM-GeMAPS*, *SVM-ComParE*, *LSTM-ComParE* and *Ada-MFCC* employ only low-level acoustic features without considering the physiological relationships between AUs and phonemes. As illustrated in Table 4.2, the proposed *DBN-LBP-AV+E* and *DBN-LPQ-AV+E*, as well as *DBN-A* and *CTBN-A* developed in our previous work [63], outperform all the other methods consistently by a large margin, in terms of the F1 score. The improvement mainly comes from explicitly modeling the semantic and dynamic relationships between phonemes and AUs.

By utilizing both acoustic and visual information, the proposed methods, i.e. *DBN-LBP-AV+E* (**0.696**) and *DBN-LPQ-AV+E* (**0.679**) outperform the *DBN-A* (**0.515**) and *CTBN-A* (**0.653**) employing only audio information on the clean subset, thanks to the usage of the decent visual measurements obtained from the "clean" images. The proposed methods also outperform the audio-based *DBN-A* on the challenging subset. However, *CTBN-A* achieved better performance than the proposed audiovisual fusion approaches on the challenging subset, which is primarily due to two reasons. First, since most of challenges are intentionally introduced into the

visual channel in the challenging subset, the visual measurements are much less reliable than those on the clean subset as shown in Table 3.2. Second, since CTBNs directly model the dynamic relationships along continuous time without predefining any granularities as in DBNs, the relationships between facial AUs and phonemes are better captured by CTBNs. However, CTBNs cannot handle noisy measurements well and thus, are not suitable for audiovisual fusion, especially with noisy visual measurements. Thus, in our work, DBNs are employed for the audiovisual fusion for speech-related facial AU recognition.

4.5 Conclusion

Facial activity is not the only channel for human communication, where voice also plays an important role. This paper presents a novel audiovisual fusion framework for recognizing speech-related AUs by exploiting information from both visual and audio channels. Specifically, a DBN model is employed to capture the comprehensive relationships for audiovisual AU recognition including the semantic and dynamic relationships among AUs, the temporal development of AUs, the dynamic dependencies among phonemes, and more importantly, the semantic and dynamic physiological relationships between phonemes and AUs. Experimental results on a pilot audiovisual AU-coded dataset have demonstrated that the proposed DBN framework significantly outperforms the state-of-the-art visual-based methods as well as audio-based methods by integrating audio and visual information in facial activity analysis. More importantly, the DBN model also beats the feature-level fusion by comprehensively modeling and exploiting the relationships in a context of audiovisual fusion.

In the future, we plan to enrich the pilot audiovisual database with more challenging data including emotional speech, and then extend the current framework to recognizing a larger set of AUs including upper-face AUs by modeling and exploiting more complicated relationships in natural human communications.

Chapter 5

Conclusion

5.0.1 Summary of Contribution

In summary, three approaches exploiting the fact that facial activities are highly correlated with voice are proposed to utilize the audio information along with the visual features for speech-related facial AU analysis: 1) a feature-level fusion method for speech-related facial AU recognition based on both audio features, i.e. MFCC, and visual features, i.e. LBP, is presented; 2) an audio-based approach that recognizes speech-related facial AUs during speech using information exclusively from the audio channel is introduced; and 3) an audiovisual fusion framework is developed, which aims to make the best use of visual and acoustic cues in recognizing speech-related facial AUs.

5.0.2 Future Research

In the future, we plan to extend the audiovisual database to include continuous and emotional speech, which is more challenging for speech recognition. In addition, more challenges will be introduced to the audio channel, such as higher environmental noise and different microphone locations. The framework learned from the enriched database is expected to capture more comprehensive relationships in natural human communication. Under the contexts of emotion, more AUs especially the upper-face AUs can be modeled. In addition, it is expected to be more robust to imperfect phoneme measurements by modeling the relationships in the words.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S Corrado, A. Davis, J. Dean, M. Devin, et al., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, arXiv preprint arXiv:1603.04467 (2016).

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

[3] T. Almaev, A. Yüce, A. Ghitulescu, and M. Valstar, *Distribution-based iterative pairwise classification of emotions in the wild using LGBP-TOP*, Proc. Int. Conf. on Multimodal Interfaces (ICMI), 2013, pp. 535–542.

[4] T. R. Almaev and M. F. Valstar, *Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition*, Proc. Int. Conf. on Affective Computing and Intelligent Interaction (ACII), Sept 2013, pp. 356–361.

[5] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, *Robust discriminative response map fitting with constrained local models*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3444–3451.

[6] T. Baltrusaitis, M. Mahmoud, and P. Robinson, *Cross-dataset learning and person-specific normalisation for automatic action unit detection*, Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG), vol. 6, 2015, pp. 1–6.

[7] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel, and J. R. Movellan, *Recognizing facial expression: Machine learning and application to spontaneous behavior*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 568–573.

[8] F. Berthommier, *Audio-visual recognition of spectrally reduced speech*, Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP), 2001, pp. 183–189.

[9] I. Bociu and I. Pitas, *A new sparse image representation algorithm applied to facial expression recognition*, Proc. Int. Workshop on Machine Learning for Signal Processing (MLSC), 2004, pp. 539–548.

[10] P. Boersma and D. Weenink, *Praat, a system for doing phonetics by computer*, Glot International **5** (2001), no. 9/10, 341–345.

[11] H. Boudali and J. Dugan, *A continuous-time bayesian network reliability modeling, and analysis framework*, Reliability, IEEE Transactions on **55** (2006), no. 1, 86–97.

[12] D. Burnham, R. Campbell, and B. Dodd, *Hearing eye II: The psychology of speechreading and auditory-visual speech*, Psychology Press, 1998.

[13] D. R. Calvert, *Descriptive phonetics*, Thieme, 1986.

[14] G. Caridakis, L. Malatesta1, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, *Modeling naturalistic affective states via facial and vocal expression recognition*, Proc. Int. Conf. on Multimodal Interfaces (ICMI), 2006.

[15] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, *About the relationship between eyebrow movements and fo variations*, Proc. of the Int. Conf. on Spoken Language Processing (ICSLP), vol. 4, 1996, pp. 2175–2178.

[16] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, *The natural statistics of audiovisual speech*, PLoS computational biology **5** (2009), no. 7.

[17] C. Chang and C. Lin, *Libsvm: a library for support vector machines*, ACM transactions on intelligent systems and technology (TIST) **2** (2011), no. 3, 27.

[18] J. Chen, Z. Chen, Z. Chi, and H. Fu, *Emotion recognition in the wild with feature fusion and multiple kernel learning*, Proc. Int. Conf. on Multimodal Interfaces (ICMI), 2014, pp. 508–513.

[19] G. F. Cooper and E. Herskovits, *A Bayesian method for the induction of probabilistic networks from data*, Machine Learning **9** (1992), no. 4, 309–347.

[20] S. Davis and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Trans. on Acoustics, Speech, and Signal Processing **28** (1980), no. 4, 357–366.

[21] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, *Recurrent neural networks for emotion recognition in video*, Proc. Int. Conf. on Multimodal Interfaces (ICMI), ACM, 2015, pp. 467–474.

[22] 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.

[23] 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.

[24] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, *Visualizing higher-layer features of a deep network*, Dept. IRO, Université de Montréal, Tech. Rep (2009).

[25] F. Eyben, F. Weninger, F. Gross, and B. Schuller, *Recent developments in opensmile, the munich open-source multimedia feature extractor*, Proc. ACM Int. Conf. Multimedia, 2013, pp. 835–838.

[26] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang, *Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 1347–1355.

[27] Y. Fan, X. Lu, D. Li, and Y. Liu, *Video-based emotion recognition using CNN-RNN and C3D hybrid networks*, Proc. Int. Conf. on Multimodal Interfaces (ICMI), 2016, pp. 445–450.

[28] Yu Fan and Christian R Shelton, *Learning continuous-time social network dynamics*, Proc. Uncertainty in Artificial Intelligence (UAI), 2009, pp. 161–168.

[29] B. Fasel, *Robust face analysis using convolutional neural networks*, Pattern Recognition **2** (2002), 40–43.

[30] C. Y. Fook, M. Hariharan, S. Yaacob, and A. Adom, *A review: Malay speech recognition and audio visual speech recognition*, ICoBE, 2012, pp. 479–484.

[31] N. Fragopanagos and J. G. Taylor, *Emotion recognition in human-computer interaction*, IEEE Trans. on Neural Networks **18** (2005), no. 4, 389–405.

[32] E Gatti, D Luciani, and F Stella, *A continuous time bayesian network model for cardiogenic heart failure*, Flexible Services and Manufacturing Journal **24** (2012), no. 4, 496–515.

[33] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, Speechreading by Humans and Machines (D.G. Stork and M.E. Hennecke, eds.), Rationale for phoneme-

viseme mapping and feature selection in visual speech recognition, Springer, Berlin, Germany, 1996, pp. 505–515.

[34] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis, *Deep learning based FACS action unit occurrence and intensity estimation*, Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG), 2015.

[35] H. Gunes and M. Pantic, *Automatic, dimensional and continuous emotion recognition*, Int. J. of Synthetic Emotions (2010).

[36] S. Han, Z. Meng, A. S. Khan, and Y. Tong, *Incremental boosting convolutional neural network for facial action unit recognition*, Proc. Advances in Neural Information Processing Systems (NIPS), 2016, pp. 109–117.

[37] S. Han, Z. Meng, P. Liu, and Y. Tong, *Facial grid transformation: A novel face registration approach for improving facial action unit recognition*, Proc. Int. Conf. on Image Processing (ICIP), 2014.

[38] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, Proc. Int. Conf. on Computer Vision (ICCV), 2015, pp. 1026–1034.

[39] Ralf Herbrich, Thore Graepel, and Brendan Murphy, *Structure from failure*, Proceedings of the 2nd USENIX workshop on Tackling computer systems problems with machine learning techniques, 2007, pp. 1–6.

[40] P. O. Hoyer, *Non-negative matrix factorization with sparseness constraints*, J. Machine Learning Research **5** (2004), 1457–1469.

[41] J. Huang and B. Kingsbury, *Audio-visual deep learning for noise robust speech recognition*, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 7596–7599.

[42] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, Proc. Int. Conf. on Machine learning (ICML), 2015, pp. 448–456.

[43] S. Jaiswal and M. F. Valstar, *Deep learning the dynamic appearance and shape of facial action units*, IEEE Workshop on Applications of Computer Vision (WACV), 2016.

[44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, *Caffe: Convolutional architecture for fast feature embedding*, Proc. ACM Int. Conf. Multimedia, ACM, 2014, pp. 675–678.

[45] B. Jiang, M.F. Valstar, and M. Pantic, *Action unit detection using sparse appearance descriptors in space-time video volumes*, Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG), 2011.

[46] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim, *Deep temporal appearance-geometry network for facial expression recognition*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2015).

[47] A. P. Kandagal and V. Udayashankara, *Automatic bimodal audiovisual speech recognition: A review*, IC3I, 2014, pp. 940–945.

[48] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, *Contrasting and combining least squares based learners for emotion recognition in the wild*, Proc. Int. Conf. on Multimodal Interfaces (ICMI), 2015, pp. 459–466.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, Proc. Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.

[50] D. D. Lee and H. S. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature **401** (1999), no. 6755, 788–791.

[51] G. Lejan, N. Souviraà-Labastie, and F. Bimbot, *Facial expression recognition from speech*, Tech. report, INRIA, 2013.

[52] Y. Lin, M. Song, D.T.P. Quynh, Y. He, and C. Chen, *Sparse coding for flexible, robust 3d facial-expression synthesis*, Computer Graphics and Applications **32** (2012), no. 2, 76–88.

[53] M. Liu, S. Li, S. Shan, and X. Chen, *AU-aware deep networks for facial expression recognition*, Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–6.

[54] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, *Deeply learning deformable facial action parts model for dynamic expression analysis*, Proc. Asian Conf. on Computer Vision (ACCV), 2014.

[55] P. Liu, S. Han, Z. Meng, and Y. Tong, *Facial expression recognition via a boosted deep belief network*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1805–1812.

[56] P. Liu, S. Han, and Y. Tong, *Improving facial expression analysis using histograms of log-transformed nonnegative sparse representation with a spatial pyramid structure*, Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–7.

[57] W. Liu, C. Song, and Y. Wang, *Facial expression recognition based on discriminative dictionary learning*, Proc. Int. Conf. on Pattern Recognition (ICPR), 2012.

[58] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, *Automatic analysis of facial actions: A survey*, IEEE Trans. on Affective Computing **13** (2017), no. 9, 1–22.

[59] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, *Subject independent facial expression recognition with robust face detection using a convolutional neural network*, IEEE Trans. on Neural Networks **16** (2003), no. 5, 555–559.

[60] H. McGurk and J. MacDonald, *Hearing lips and seeing voices*, (1976).

[61] Z. Meng, S. Han, M. Chen, and Y. Tong, *Feature level fusion for bimodal facial action unit recognition*, Proc. IEEE Int. Symposium on Multimedia (ISM), IEEE, 2015.

[62] Z. Meng, S. Han, M. Chen, and Y. Tong, *Audiovisual facial action unit recognition using feature level fusion*, International Journal of Multimedia Data Engineering and Management (IJMDEM) **7** (2016), no. 1, 60–76.

[63] Z. Meng, S. Han, and Y. Tong, *Listen to your face: Inferring facial action units from audio channel*, (2017).

[64] K. Murphy, *The Bayes net toolbox for Matlab*, Computing Science and Statistics **33** (2001), no. 2, 1024–1034.

[65] Chalapathy Neti, Gerasimos Potamianos, and Juergen Luettin, *Large-vocabulary audio-visual speech recognition: A summary of the johns hopkins summer 2000 workshop*, Proc. Workshop on Multimedia Signal Processing, 2001, pp. 619–624.

[66] B. Ng, A. Pfeffer, and R. Dearden, *Continuous time particle filtering*, Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI), vol. 19, 2005, p. 1360.

[67] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, *Multimodal deep learning*, Proc. Int. Conf. on Machine learning (ICML), 2011, pp. 689–696.

[68] M. A. Nicolaou, H. Gunes, and M. Pantic, *Audio-visual classification and fusion of spontaneous affect data in likelihood space*, Proc. Int. Conf. on Pattern Recognition (ICPR), 2010, pp. 3695 – 3699.

[69] U. Nodelman and E. Horvitz, *Continuous time Bayesian networks for inferring users presence and activities with extensions for modeling and evaluation*, Tech. Report MSR-TR-2003-97, Microsoft Research, December 2003.

[70] U. Nodelman, C. R. Shelton, and D. Koller, *Continuous time bayesian networks*, Proc. Uncertainty in Artificial Intelligence (UAI), 2002, pp. 378–387.

[71] U. Nodelman, C. R. Shelton, and D. Koller, *Learning continuous time bayesian networks*, Proc. Uncertainty in Artificial Intelligence (UAI), 2002, pp. 451–458.

[72] Uri D. Nodelman, *Continuous time Bayesian network*, Ph.D. thesis, Standford University, 2007.

[73] T. Ojala, M. Pietikäinen, and D. Harwood, *A comparative study of texture measures with classification based on featured distributions*, Pattern Recognition **29** (1996), no. 1, 51–59.

[74] B. A. Olshausen and D. J. Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, Nature **381** (1996), no. 6583, 607–609.

[75] M. Pantic and M. S. Bartlett, *Machine analysis of facial expressions*, Face Recognition (K. Delac and M. Grgic, eds.), I-Tech Education and Publishing, Vienna, Austria, 2007, pp. 377–416.

[76] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang, *Human computing and machine understanding of human behavior: A survey*, Artificial Intelligence for Human Computing (T. S. Huang, A. Nijholt, M. Pantic, and A. Pentland, eds.), Lecture Notes in Artificial Intelligence, Springer Verlag, London, 2007.

[77] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, *The Kaldi Speech Recognition Toolkit*, ASRU, December 2011.

[78] D. M. Powers, *Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation*, Journal of Machine Learning Technologies **2** (2011), no. 1, 37–63.

[79] Shaojie Qiao, Changjie Tang, Huidong Jin, Teng Long, Shucheng Dai, Yungchang Ku, and Michael Chau, *Putmode: prediction of uncertain trajectories in moving objects databases*, Applied Intelligence **33** (2010), no. 3, 370–386.

[80] M Ranzato, J. Susskind, V. Mnih, and G. Hinton, *On deep generative models with applications to recognition*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 2857–2864.

[81] Vinayak Rao and Yee Whye Teh, *Fast mcmc sampling for markov jump processes and extensions*, J. Machine Learning Research **14** (2013), no. 1, 3295–3320.

[82] S. Reed, K. Sohn, Y. Zhang, and H. Lee, *Learning to disentangle factors of variation with manifold interaction*, Proc. Int. Conf. on Machine learning (ICML), 2014.

[83] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, *Disentangling factors of variation for facial expression recognition*, Proc. European Conf. on Computer Vision (ECCV), 2012, pp. 808–822.

[84] F. Ringeval, E. Marchi, M. Mehu, K. R. Scherer, and B. W. Schuller, *Face reading from speech-predicting facial action units from audio cues*, Proc. Conf. of the Int. Speech Communication Association (INTERSPEECH), 2015.

[85] A. Rogozan, *Discriminative learning of visual data for audiovisual speech recognition*, Int. J. Artificial Intelligence Tools **8** (1999), no. 1, 43–52.

[86] E. Sariyanidi, H. Gunes, and A. Cavallaro, *Automatic analysis of facial affect: A survey of registration, representation and recognition*, IEEE Trans. on Pattern Analysis and Machine Intelligence **37** (2015), no. 6, 1113–1133.

[87] G. Schwarz, *Estimating the dimension of a model*, The Annals of Statistics **6** (1978), 461–464.

[88] N. Sebe, I. Cohen, T. Gevers, and T.S. Huang, *Emotion recognition based on joint visual and audio cues*, Proc. Int. Conf. on Pattern Recognition (ICPR), 2006, pp. 1136–1139.

[89] T. Sénéchal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, *Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units*, Proc. Int. Conf. on Automatic Face and Gesture Recognition Workshop (FGW), 2011, pp. 860 – 865.

[90] T. Sénéchal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, *Facial action recognition combining heterogeneous features via multikernel learning*, IEEE Trans. on Systems, Man, and Cybernetics-Part B: Cybernetics **42** (2012), no. 4, 993–1005.

[91] C. Shan, S. Gong, and P.W. McOwan, *Facial expression recognition based on Local Binary Patterns: A comprehensive study*, J. Image and Vision Computing **27** (2009), no. 6, 803–816.

[92] Christian R Shelton, Yu Fan, William Lam, Joon Lee, and Jing Xu, *Continuous time bayesian network reasoning and learning engine*, J. Machine Learning Research **11** (2010), 1137–1140.

[93] M. Song, J. Bu, C. Chen, and N. Li, *Audio-visual based emotion recognition – a new approach*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2004, pp. 1020–1025.

[94] D. G. Stork and M. E. Hennecke, *Speechreading by humans and machines: Models, systems, and applications*, Springer, Berlin, Germany, 1996.

[95] K. Strelnikov, J. Foxton, M. Marx, and P. Barone, *Brain prediction of auditory emphasis by facial expressions during audiovisual continuous speech*, Brain topography **28** (2015), no. 3, 494–505.

[96] C. Sui, M. Bennamoun, and R. Togneri, *Listening with your eyes: Towards a practical visual speech recognition system using deep boltzmann machines*, Proc. Int. Conf. on Computer Vision (ICCV), 2015, pp. 154–162.

[97] Y. Tang, *Deep learning using linear support vector machines*, Proc. Int. Conf. on Machine learning (ICML), 2013.

[98] A Teixeira Lopes, E de Aguiar, and T Oliveira-Santos, *A facial expression recognition system using convolutional networks*, Graphics, Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on, IEEE, 2015, pp. 273–280.

[99] Y. Tian, T. Kanade, and J. F. Cohn, *Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity*, Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG), May 2002, pp. 229–234.

[100] Y. Tong, J. Chen, and Q. Ji, *A unified probabilistic framework for spontaneous facial action modeling and understanding*, IEEE Trans. on Pattern Analysis and Machine Intelligence **32** (2010), no. 2, 258–274.

[101] Y. Tong, W. Liao, and Q. Ji, *Facial action unit recognition by exploiting their dynamic and semantic relationships*, IEEE Trans. on Pattern Analysis and Machine Intelligence **29** (2007), no. 10, 1683–1699.

[102] 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*, Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG) (2015).

[103] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, *Meta-analyis of the first facial expression recognition challenge*, IEEE Trans. on Systems, Man, and Cybernetics-Part B: Cybernetics **42** (2012), no. 4, 966–979.

[104] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, *Meta-analysis of the first facial expression recognition challenge*, IEEE Trans. on Systems, Man, and Cybernetics-Part B: Cybernetics **42** (2012), no. 4, 966–979.

[105] Y. Wang and L. Guan, *Recognizing human emotional state from audiovisual signals*, IEEE Trans. on Multimedia **10** (2008), no. 5, 936–946.

[106] R. L. Weide, *The CMU Pronouncing Dictionary*, 1994, `http://www.speech.cs.cmu.edu/cgi-bin/cmudict/`.

[107] J. Whitehill, M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, *Towards practical smile detection*, IEEE Trans. on Pattern Analysis and Machine Intelligence **31** (2009), no. 11, 2106–2111.

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

[109] J. Xue, J. Jiang, A. Alwan, and L. E. Bernstein, *Consonant confusion structure based on machine classification of visual features in continuous speech*, Proc. Int. Conf. Auditory-Visual Speech Processing, 2005, pp. 103–108.

[110] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun, *Multi-clue fusion for emotion recognition in the wild*, Proc. Int. Conf. on Multimodal Interfaces (ICMI), 2016, pp. 458–463.

[111] P. Yang, Q. Liu, and D. N. Metaxas, *Boosting coded dynamic features for facial action units and facial expression recognition*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), June 2007, pp. 1–6.

[112] P. Yang, Q. Liu, and Metaxas D. N., *Boosting encoded dynamic features for facial expression recognition*, Pattern Recognition Letters **30** (2009), no. 2, 132–139.

[113] Anbang Yao, Junchao Shao, Ningning Ma, and Yurong Chen, *Capturing AU-aware facial features and their latent relations for emotion recognition in the wild*, Proc. Int. Conf. on Multimodal Interfaces (ICMI), 2015, pp. 451–458.

[114] Z. Ying, Z. Wang, and M. Huang, *Facial expression recognition based on fusion of sparse representation*, Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence (D.-S. Huang, X. Zhang, C.A. Reyes García, and L. Zhang, eds.), Lecture Notes in Computer Science, 2010, pp. 457–464.

[115] J. Yuan and M. Liberman, *Speaker identification on the scotus corpus*, Journal of the Acoustical Society of America **123** (2008), no. 5, 3878.

[116] A. Yuce, H. Gao, and J. Thiran, *Discriminant multi-label manifold embedding for facial action unit detection*, Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG), 2015.

[117] S. Zafeiriou and M. Petrou, *Nonlinear non-negative component analysis algorithms*, IEEE Trans. on Image Processing **19** (2010), no. 4, 1050–1066.

[118] S. Zafeiriou and M. Petrou, *Sparse representations for facial expressions recognition via L1 optimization*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 32–39.

[119] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, *A survey of affect recognition methods: Audio, visual, and spontaneous expressions*, IEEE Trans. on Pattern Analysis and Machine Intelligence **31** (2009), no. 1, 39–58.

[120] Z. Zeng, J. Tu, T. S. Huang, B. M. Pianfetti, D. Roth, and S. Levinson, *Audio-visual affect recognition*, IEEE Trans. on Multimedia **9** (2007), no. 2, 424–428.

[121] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T.S. Huang, D. Roth, and S. Levinson, *Bimodal HCI-related affect recognition*, Proc. Int. Conf. on Multimodal Interfaces (ICMI), 2004, pp. 137–143.

[122] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, *Audio-visual affective expression recognition through multistream fused HMM*, IEEE Trans. on Multimedia **10** (2008), no. 4, 570–577.

[123] Z. Zeng, J. Tu, P. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T.S. Huang, and S. Levinson, *Audio-visual affect recognition through multi-stream fused HMM for HCI*, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 967–972.

[124] Y. Zhang and Q. Ji, *Active and dynamic information fusion for facial expression understanding from image sequences*, IEEE Trans. on Pattern Analysis and Machine Intelligence **27** (2005), no. 5, 699–714.

[125] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, *Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron*, Proc. Int. Conf. on Automatic Face and Gesture Recognition (FG), 1998, pp. 454–459.

[126] G. Zhao and M. Pietiäinen, *Dynamic texture recognition using local binary patterns with an application to facial expressions*, IEEE Trans. on Pattern Analysis and Machine Intelligence **29** (2007), no. 6, 915–928.

[127] R. Zhi, M. Flierl, Q. Ruan, and W.B. Kleijn, *Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition*, IEEE Trans. on Systems, Man, and Cybernetics-Part B: Cybernetics (2010), no. 99, 1–15.

[128] L. Zhong, Q. Liu, P. Yang, J. Huang, and D.N. Metaxas, *Learning multiscale active facial patches for expression analysis*, IEEE Trans. on Cybernetics **45** (2015), no. 8, 1499–1510.