Domain Adaptation for Speech-Driven Affective Facial Features Synthesis

by

Rizwan Sadiq

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Electrical and Electronics Engineering

KOÇ ÜNİVERSİTESİ

September 04, 2020

Domain Adaptation for Speech-Driven Affective Facial Features Synthesis

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Rizwan Sadiq

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Engin Erzin (Advisor)

_____

Prof. Yücel Yemez

_____

Prof. Çiğdem Eroğlu Erdem

_____

Prof. A. Murat Tekalp

_____

Prof. Murat Saraçlar

Date: _____

Dedicated to my Sister Saima Sadiq

ABSTRACT

Domain Adaptation for Speech-Driven Affective Facial Features Synthesis

Rizwan Sadiq

Doctor of Philosophy in Electrical and Electronics Engineering

September 04, 2020

Although speech-driven facial animation has been studied extensively in the literature, works focusing on the affective content of the speech are limited. This is mostly due to the scarcity of affective audio-visual data. In this thesis, we present three major studies that lead us to speech-driven affective facial synthesis. First, we investigate the use of lip articulations for affect recognition to better understand dependencies across lip articulations, phonetic classes and affect. Then, we propose a multimodal system, consisting of text and speech, for affective facial feature synthesis. A phoneme-based model driven from text qualifies generation of speaker independent animation, whereas a speech based model enables capturing affective variation during the facial feature synthesis. Finally, we improve the affective facial synthesis using domain adaptation by partially reducing the data scarcity. In this last study, we first define a domain adaptation to map affective and neutral speech representations to a common latent space in which cross-domain bias is smaller. Then, the domain adaption is used to augment affective representations for each emotion category, including angry, disgust, fear, happy, sad, surprise and neutral, so that we can better train emotion-dependent deep audio-to-visual (A2V) mapping models. Based on the emotion-dependent deep A2V models, the proposed affective facial synthesis system is realized in two stages: first, speech emotion recognition extracts soft emotion category likelihoods for the utterances; then a soft fusion of the emotion-dependent A2V mapping outputs form the affective facial synthesis. Experimental evaluations are performed on the SAVEE audio-visual dataset with objective and subjective evaluations. The proposed affective A2V system achieves significant mean square error loss improvements in comparison to the recent literature. Furthermore, the resulting facial animations of the proposed

system are preferred over the baseline animations in the subjective evaluations.

ÖZETÇE

Doktora Tez Başlığı
Rizwan Sadiq
Elektrik ve Elektronik Mühendisliği, Doktora
04 Eylül 2020

Literatürde konuşma ile sürülen yüz animasyonu yoğun olarak çalışılmış olsa da, konuşmanın duygusal içeriğine odaklanan çalışmalar sınırlıdır. Bu çoğunlukla duygusal görsel-işitsel verilerin az erişilebilir olmasından kaynaklanmaktadır. Bu tezde, bizi konuşma ile sürülen duygusal yüz sentezine götüren üç ana çalışma sunuyoruz. İlk olarak, dudak devinimleri, fonetik sınıflar ve duygulanım arasındaki bağımlılıkları daha iyi anlamak için dudak devinimlerini kullanan duygu tanıma modellerini araştırdık. Ardından, duygusal yüz sentezi için metin ve konuşmadan oluşan çok-kipli bir sistem önerdik. Metinden türetilen fonem tabanlı bir model, konuşmacıdan bağımsız yüz sentezini nitelendirirken, konuşma tabanlı bir model, yüz sentezi için duygusal değişimlerin yakalanmasını sağlar. Son olarak, ulaşılabilir veri azlığını kısmen azaltmak için alan uyarlaması kullanarak duygusal yüz sentezini iyileştirdik. Bu son çalışmada, öncelikle alanlar arası istatistiksel farkın daha küçük olduğu ortak bir gizli alanla duygusal ve nötr konuşma özniteliklerini eşlemek için bir alan uyarlaması tanımlıyoruz. Ardından, alan uyarlaması öfke, tiksinti, korku, mutluluk, üzgünlük, sürpriz ve nötr dahil olmak üzere her farklı duygu kategorisi için duygu içerikli öznitelikleri artırmak için kullanılır, böylece duyguya bağlı derin görsel-işitsel (A2V) dönüşümü daha iyi eğitebiliriz. Duyguya bağlı derin A2V modellerine dayanarak, önerilen duygusal yüz sentez sistemi iki aşamada gerçekleştirilmiştir: birinci aşamada, konuşmadan duygu tanıma, konuşma parçaları için yumuşak duygu kategorisi olasılıklarını çıkarır; daha sonra duyguya bağlı A2V dönüşümü çıktılarının yumuşak bir füzyonu, duygusal yüz sentezini oluşturur. SAVEE görsel-işitsel veri seti üzerinde nesnel ve öznel deneysel değerlendirmeler yaptık. Önerilen duygusal A2V dönüşüm sistemi, yakın literatüre kıyasla ortalama kareler hatası için önemli iyileşmeler sağladı. Ayrıca öznel değerlendirmelerde, önerilen duygusal yüz sentezleri baz referans yüz sentezlerine göre tercih edildi.

ACKNOWLEDGMENTS

First and foremost I would like to express my sincere gratitude to my advisor, Prof. Engin Erzin, who constantly offered his guidance and support. I am grateful for his patience, motivation, immense knowledge and spent time in helping me during my research. His doors were always open for guidance. I would also like to thank Prof. Yücel Yemez and Prof. Çiğdem Eroğlu Erdem for their valuable suggestions throughout the course of this study. I would like to thank my friends and family for their moral support. Last but not the least, I would like to thank my wife, who made me realize that PhD is not the toughest thing to deal with.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ASM	Active Shape Model
PCA	Principal Component Analysis
MSE	Mean Squared Error
DNN	Deep Neural Network
CNN	Convolutional Neural Network
HMM	Hidden Markov Model
HCI	Human Computer Interaction
A2V	Audio to Visual
MFSC	Mel Frequency Spectral Coefficients
EDDA	Emotion Dependent Domain Adaptation
SER	Speech Emotion Recognition
NTL	No Transfer Learning
DA	Domain Adaptation
MA	Model Adaptation
SED	Soft Decision Emotion Dependent
HED	Hard Decision Emotion Dependent

Chapter 1

INTRODUCTION

Making machines interact with humans as efficiently as possible is an active area of research for decades. The main goal is to make human computer interaction (HCI) as natural as possible. Among all the interaction methods, speech, the most natural and reliable form of communication, is increasingly gaining popularity and recently deployed in various HCI systems, such as mobile personal assistants powered by speech recognition and text-to-speech synthesis technologies. Exploiting the advantages embedded in speech signals, speech driven facial animations also span decades of research. Facial animations are essential for their practical use in various applications, such as virtual agents and avatars [Capin et al., 1997, Konstantinidis et al., 2009]. Researchers in the fields of computer graphics, machine learning, psychology as well as medicine are relentlessly working towards generating more natural and realistic looking animated avatars [Gaggioli et al., 2003, Konstantinidis et al., 2009, Vougioukas et al., 2018, Taylor and et all, 2017]. Audio-to-visual (A2V) mapping involves prediction of speech-related facial motion from the acoustic signal, or automatically generating animations of articulatory features of speech, such as lips, [Taylor et al., 2016] as exemplified in Figure 1.1. A2V mapping also targets to carry the emotional content in the speech signal, so that the speech driven animations can reflect and convey the emotional content in audio-visual channels in a more natural manner.

It's natural for humans to identify the emotions and react accordingly. However, perception of emotion and affective response generation are still challenging problems for machines. We target to estimate facial animation representations from affective speech. For this purpose, we construct a two stage process, where the first stage

uncovers the emotion in the affective speech signal, and the second stage defines a mapping from the affective speech signal to facial movements conditioned on the emotion.

Emotion recognition from speech has been an active research area in HCI for various applications, specially for humanoid robots, avatars, chat bots, virtual agents etc. Recently, various deep learning approaches have been utilized to improve emotion recognition performance, which includes significant challenges in realistic HCI settings. Before the deep learning era, classical machine learning models, such as hidden Markov models (HMMs), support vector machines (SVMs), and decision tree-based methods, have been used in speech emotion recognition [Pan et al., 2012, Schuller et al., 2003, Lee et al., 2011, Sadiq and Erzin, 2017].

As deep neural networks become widely available and computationally feasible, wide range of studies adapted complex deep neural network models for the speech emotion recognition problem. Among these models, Convolutional neural network (CNN) based models are successfully utilized for improved speech emotion recognition [Bertero and Fung, 2017, Badshah et al., 2017]. In another setup, a recurrent neural network model based on text and speech inputs has been successfully used for emotion recognition [Yoon et al., 2018]. Emotion recognition from the speech has been an active research area over the decades [Vinola and Vimaladevi, 2015, El Ayadi et al., 2011, Schuller et al., 2011]. Speech emotion recognition (SER) is generally performed on a short time frame level or utterance level information. SER can be performed either as a classification problem for categorical emotions (happy, sad, angry, neutral etc.) or a regression problem (activation, valance and dominance). In this study we employed SER as classification problem for seven class emotion categories. Ideally, A2V mapping systems can deliver automatic and affective animation generation from speech in real time without using resources from skilled human animators. These systems can be utilized in various applications, such as lip syncing of dubbed movies or facial animations for video conferencing. In video conferencing, one can effectively use the bandwidth of video communications by transmitting only voice data over a smaller bandwidth

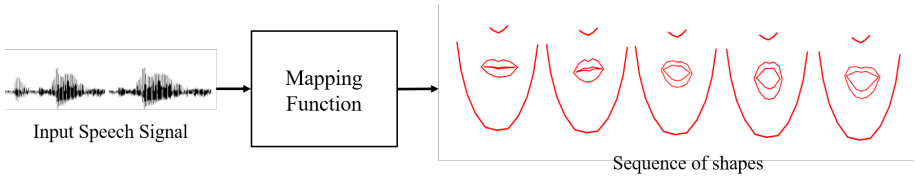


Figure 1.1: Block diagram of a general audio-to-visual (A2V) mapping system

and later generating the visual content based on the speech or sending only key frames and partially filling the remaining frames.

In this study, we propose a transfer learning based A2V mapping framework for affective data. Our framework includes a domain adaptation based on the feature representation transfer, which is conditioned on emotion prediction. Note that the feature representation transfer for affective speech driven facial animation is distinctly new and different from the two related recent works [Taylor and et all, 2017, Asadiabadi et al., 2018], where both do not address affective facial animation and domain adaptation paradigms. Hence, the use of domain adaptation will be a first in the literature for the affective A2V mapping problem. In this study we choose to consider mapping of speech signals to the features of lower face area, i.e. lips and jaws only. The significance of the lower facial area for mapping speech to facial gestures, in the presence of emotional content, was discussed by [Busso and Narayanan, 2007]. They also explored this relation for different phoneme classes (vowel, nasals, glides, stops, consonants etc.), and showed that lower facial area is more active as compared to upper face, when exposed to emotional content. As a motivation to use the lower face area for audio to visual mapping, we investigate the effect of phonetic classes for affect recognition from lip articulations. The affect recognition problem is formalized in discrete activation, valence and dominance attributes. A set of discriminative phonemes, which improve the affect recognition performance, is identified.

Generation of facial gestures using speech-driven systems mostly is lacking the reflection of emotions in the speech onto the facial gestures produced. Overcoming this limitation and capturing emotions in the speech signal and reflecting the

emotional content on facial gestures represents one of our main motivations.

1.1 Contributions

This thesis presents an affective speech-driven facial feature synthesis frame-work. We are interested in fully systems with no human involvement of fixing the shapes of facial features over the frames of video. For this purpose, we proposed a system that inputs a speech signal (with or without emotional content) and outputs a sequence of video frames representing the facial features in terms of landmarks. The main contribution of this study has multiple folds:

1. We analysed the significance of lip articulations under emotional scenario and present a set of discriminative phonemes for better emotion recognition from lip articulations.[Sadiq and Erzin, 2017]

2. We proposed a multimodal approach to enhance the previously developed methods for high fidelity production of data driven facial animations [Asadiabadi et al., 2018].

3. We propose a state-of-the-art domain adaptation scheme, that is integrated for the affective A2V mapping problem [Sadiq and Erzin, 2020].

4. We propose a new feature selection scheme to define a parallel data set across source and target domains for the neutral and affective speech signals that makes feature representation transfer possible [Sadiq and Erzin, 2020] .

5. We propose A2V mapping as a weighted combination of emotion dependent facial regression outputs [Sadiq et al., 2019, Sadiq and Erzin, 2020].

6. We perform extensive experimental comparison of the different approaches using subjective and objective tests.

1.2 List of Publications

1. Sadiq, R. and Erzin, E. (2020). Emotion Dependent Domain Adaptation for Speech Driven Affective Facial Feature Synthesis, in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2020.3008456.

2. Sadiq, R., AsadiAbadi, S., and Erzin, E.(2019). Emotion Dependent Facial Animation from Affective Speech. in 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP 2020).

3. Asadiabadi,S., Sadiq, R., and Erzin, E. (2018). Multimodal Speech Driven Facial Shape Animation using Deep Neural Networks. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pages 1508–1512

4. Sadiq, R. and Erzin, E. (2017). Affect Recognition from Lip Articulations. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pages 2432–2436.

1.3 Organization

The remainder of the thesis is organized as follows:

- Chapter 2 presents the literature survey of audio to visual mapping and emotion recognition.

- Chapter 3 presents the methodologies and data set used to determine the effect of phonetic classes for affect recognition from lip articulations along with the results.

- Chapter 4 provides the details of A2V mapping process and significance of multimodality for generation of speech driven facial shape features. Results of using multimodality for speech driven facial feature synthesis is also presented.

- Chapter 5 explains the use of domain adaptation in emotional speech scenarios for generating visual shape features. A2V models selection based on emotion dependence is also discussed in this chapter. Results of all the experimental studies related to domain adaptation and emotion dependent A2V mapping are covered in this chapter.

- Chapter 6 presents the concluding remarks along with some future directions to this research.

Chapter 2

LITERATURE REVIEW

For the A2V mapping, various approaches have been studied to generate the naturalistic animations using different cues. In a pioneering study, the strong correlation between visual and speech features has been studied for A2V mapping [Yehia et al., 1998]. A2V mapping is widely studied based on statistical methods including Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) and Dynamic Bayesian Networks (DBNs). In [Yamamoto et al., 1997], visemes are represented by HMM states, which are mapped to the visual lip parameters in a generative synthesis model. Note that a viseme is defined as the cluster of speech sounds that generates the same visual articulation pattern. Utilization of visemes for speech driven facial animation has been studied in the literature [Bozkurt et al., 2007, Verma et al., 2003]. In general association of phonemes to visemes is not one-to-one. A coupled HMM has been proposed to model this many-to-one audio-visual dependency characteristics in comparison with other single and multi-stream HMM architectures [Xie and Liu, 2007]. In another early work [Brand, 1999], 3D facial movements are predicted from LPC and RASTA-PLP acoustic features using an entropy-minimization algorithm that learns both the structure and the parameters of an HMM to perform frame to frame mapping. Later a KNN model has been used to map speech features (RASTA-PLP) to visual cues by forming a lookup table of audio-visual dynamic pairs, where pairs are constructed by associating each video frame with the acoustic features from past and future audio frames [Gutierrez-Osuna and et al, 2005].

A GMM based statistical mapping between articulatory movements and speech signal has been studied in [Toda et al., 2008]. A relatively recent work extends this GMM based approach to map speech spectrum features to PCA representation of

the facial landmarks by refining GMM parameters for a minimum conversion error [Luo et al., 2014]. A DBN based model has been studied to capture dependencies not only between speech and facial behaviors, but also between head and eyebrow motions [Mariooryad and Busso, 2012]. Note that most of this early research do not address affective A2V mapping problems and they are mostly based on the viseme representations, which limit the capability of A2V mapping.

Although HMMs are extensively utilized in temporal clustering of speech signals [Vougioukas et al., 2018], with the recent advances in deep learning, deep neural networks (DNNs) have demonstrated great impact on various tasks including the speech signal representation. Particularly convolutional neural networks (CNNs) hold tremendous potential. Although CNNs have been dominantly used in the field of computer vision, especially for object detection where performances exceeding the accuracy of human raters have achieved [Ioffe and Szegedy, 2015], they have utilized for applications of the speech signal processing, such as speech recognition [Qian et al., 2016, Abdel-Hamid et al., 2014, Song and Cai, 2015] and speech emotion recognition [Huang et al., 2014, Trigeorgis et al., 2016, Mao et al., 2014].

In the recent literature, DNN based A2V mapping studies are getting popular. A recent study on lip synchronous synthesis of high-quality video, demonstrating a sample video of Ex-USA president Obama speaking with accurate lip sync, has popularized the potential of DNN models for the task [Suwajanakorn et al., 2017]. In this study, a recurrent neural network (RNN) has been trained on large collection of the president's weekly address footages that learns the mapping from raw speech features to mouth shapes. Although the study claims to model emotion dependent animation generation, since it is speaker dependent, it requires extensive speaker dependent training and hence it can not be generalized to other speakers with ease. A similar approach using bidirectional LSTMs has also been studied for photo-realistic talking head synthesis that also lacks the generalization [Fan et al., 2015]. In a recent study, generalization problem has been addressed by synthesizing facial animations solely from phoneme sequence as input [Taylor and et all, 2017]. They use a DNN based sliding window predictor that learns arbitrary nonlinear mappings from

phoneme label input sequences to facial lip movements. Contrary to conventional methods, which map phoneme sequence to fixed number of visemes [Verma et al., 2003, Bozkurt et al., 2007], their model successfully maps the input phoneme sequence to output video representation of the continuous speech. Although their system provides some promising results with high-quality speech-driven animations in a generalizable setup, affective speech scenario has not been addressed. As an extension to the phoneme sequence to lip movements mapping, we recently investigate a deep multimodal CNN based framework to fuse the textual phoneme sequence information with the spectral acoustic features [Asadiabadi et al., 2018]. We observed objective improvements for both natural and affective speech inputs. In this study, we use the CNN based framework as a baseline for A2V mapping problem.

On the other hand, A2V mapping problem has been investigated in several recent studies under affective scenarios. An LSTM based mapping from acoustic to affective facial features has been investigated in [Pham et al., 2017]. The affective LSTM model has been trained on the RAVDESS audio-visual dataset [Livingstone et al., 2012], which includes affect variations articulated on only two sentences that lacks phonetic richness.

A recent study from the NVIDIA labs investigates a deep network for end-to-end generation of facial animations from raw speech signal [Karras et al., 2017]. Along with facial movements, their model learns a latent code, which can be used to identify the emotional state of the speaker. The DNN based model has been trained on 3-5 minutes of audio-visual data. Although the DNN model performs well in most cases, it fails when the tempo and emotional mood are different in test compared to the training data.

A more recent study defines an A2V mapping model using the generative adversarial network (GAN) [Vougioukas et al., 2018]. Rather than crafted speech features, the GAN model takes the raw speech and a single facial image of the subject to generate animations with expressions, such as eyebrow movements and eye blinks. Randomly generated expressions are the major shortcoming of the approach.

Training a deep A2V mapping model requires rich and large amount of multimodal data. Furthermore, building affective models needs the multimodal data to include rich emotional content. Although, many audio-visual data sets, such as GRID [Cooke et al., 2006], SAVEE [Haq and Jackson, 2011], RAVDESS [Livingstone et al., 2012] and BBC-LRW [Chung et al., 2017], are available for the task, they either lack emotional content or have limited emotional content to train realistic A2V mapping models. Hence, it is required to have large multimodal datasets with rich emotional content to better train affective A2V mapping models. However, it is often time consuming and costly to collect and annotate such datasets. Transfer learning has the potential to avail alternative solution to the training problem, where one can use the knowledge of abundant neutral multimodal data to improve modeling of the scarce affective data.

Transfer learning aims to generate a model for a task that suffers the lack of training data in a target domain, by utilizing and exploiting the knowledge from a different but related source domain with abundance of available data [Pan et al., 2010, Weiss et al., 2016]. In the A2V mapping problem, we can respectively define the neutral and affective audio-visual data as the source and target domains, where source domain has abundant neutral data and target domain has scarce affective data. Note that for both domains the input feature space is defined by the acoustic speech signal representation. However, marginal probability distributions of the feature space for the source and target domains are expected to differ due to emotion dependent articulations. In this case, model adaptation is a widely used transfer learning approach where parameters of the model can be adapted from source domain to target domain [Bonilla et al., 2008]. Figure 2.1 presents a general block diagram of the model adaptation approach for the A2V mapping problem.

Alternatively, domain adaptation can be used to transfer feature representation into a latent space where the marginal probability distributions converge for the source and target domains [Blitzer et al., 2006, Glorot et al., 2011]. Autoencoders have been widely used for domain adaptation. A study on emotional speech analysis successfully utilized adaptive denoising autoencoders for improved unsupervised

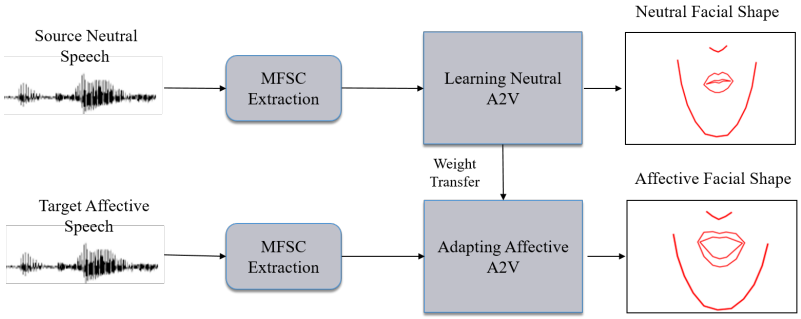


Figure 2.1: Model adaptation (MA) for the affective A2V mapping

domain adaptation where prior knowledge learned from a limited target set is used to regularize the training on a abundant source set [Deng et al., 2014]. In a more general framework, stack denoising autoencoders (SDAs) have been successfully used to adapt new representations for feature augmentations [Glorot et al., 2011]. It is shown that SDAs are able to disentangle hidden factors, which explain the variations in the input data and automatically group features in accordance with their relationship to these factors. A modified version of the SDA, marginalized SDA (mSDA), adopts the layer by layer training of the SDA [Chen et al., 2012]. In contrast to the SDA, mSDA marginalizes noise and defines a closed-form solution. Therefore, it does not need any optimization algorithm to learn parameters. Furthermore, augmentation capabilities of the mSDA are par with the traditional SDAs, which are attaining almost identical accuracy in benchmark tasks [Zhou et al., 2014].

Chapter 3

AFFECT RECOGNITION FROM LIP ARTICULATIONS

Emotional content in a conversation helps deliver the actual meaning. Affect recognition received increased attention from researchers due to its practical applications in the fields of human-machine interaction, psychology and sociology. Multiple cues have been used in the literature for the purpose of affect recognition. Some studies used single modality [Busso et al., 2007, Cowie et al., 2001], while some others utilized multimodal systems, which include speech, facial expressions, head and full body movements [Metallinou et al., 2012, Nicolaou et al., 2011, Bozkurt et al., 2013]. In this chapter,we investigate the effect of phonetic classes for affect recognition from lip articulations. The affect recognition problem is formalized in discrete activation, valence and dominance attributes. A set of discriminative phonemes, which improve the affect recognition performance, is identified. we first define the database, which includes a rich set of lip articulation data from affective interactions. Then, the lip feature representation is given. Later, we describe a scoring function to select phonemes, which are most discriminating for the affect recognition problem. Finally, we define the affect recognition framework.

3.1 Data Set

In our analysis and experimental evaluations of affect recognition from lip articulations, we used the interactive emotional dyadic motion capture database (IEMOCAP), which is designed to study expressive human interactions [Busso et al., 2008]. The IEMOCAP is an audiovisual database, which also provides motion capture data of face, head and partially hands. It comprises spontaneous conversations between pairs of professional actors in dyadic interaction sessions. The corpus has five sessions with ten actors taking part in dyadic interactions. In

each session, actors play three scripts and improvise eight hypothetical scenarios to elicit rich emotional reactions. Recordings of each session are split into clips. The total number of clips is 150 with a total duration of approximately 8 hours. Motion capture of the face has 53 markers, where 8 markers are placed on the lips. For every sentence phonetic transcriptions are also available in the database. The emotional content was annotated by human annotators on the sentence level either in categorical labels (angry, happy, excited, sad, frustrated, fearful, surprised, disgusted, neutral and other) or in dimensional descriptions of valence, activation and dominance. Value 1 denotes very low activation, dominance and very negative valence and 5 denotes very high activation, dominance and very positive valence. Those labels were assigned values between 1 and 5 and were averaged across 2-3 annotators [Busso et al., 2008].

In our experiments, we quantize AVD attribute values, A, into three levels as used in [Metallinou et al., 2012],

$$Q(A) = \begin{cases} A_1 & \text{if } 1 \leq A \leq 2, \\ A_2 & \text{if } 2 < A < 4, \\ A_3 & \text{if } 4 \leq A, \end{cases} \tag{3.1}$$

where A_1, A_2 and A_3 are representing the low, medium and high levels of activation, dominance and valence attributes.

3.2 Feature Extraction

We computed four distance features using the x, y, z coordinates of 8 lip markers to define the frame level lip feature vector. The lip feature is comprised of one horizontal and three vertical lip distances as shown in Figure 3.1. We extracted statistical functionals of the lip feature vectors over temporal windows to define the segment level lip feature vectors. A total of 11 statistical functionals used were: mean, standard deviation, skewness, kurtosis, range, min, max, first quantile, third quantile, median quantile, and inter-quantile range [Metallinou et al., 2013]. The

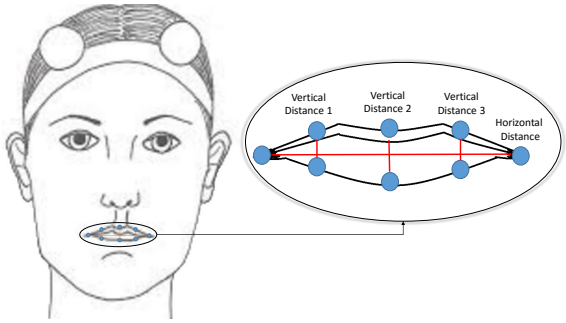


Figure 3.1: Horizontal and vertical lip distances

segment level 44-dimensional lip feature vectors, f, was computed over phoneme duration. Phoneme boundaries were used as provided by the IEMOCAP database.

3.3 Discriminative Phoneme Selection

We investigated how statistical characterization of the segment level lip features changes across different levels of affect state for each phoneme. For this purpose, we defined a symmetric Kullback–Leibler divergence (KLD) of lip features f given affect state and phonetic class as

$$D_{mn}(f|p) = KLD(P(f|A_m, p), P(f|A_n, p)), \tag{3.2}$$

where $P(f|A_m, p)$ is the conditional probability mass function of f given affect level A_m for phoneme p. The symmetric KLD is defined over probability mass functions $X()$ and $Y()$ as

$$KLD(X, Y) = \sum_j X(j) log \frac{X(j)}{Y(j)} + \sum_j Y(j) log \frac{Y(j)}{X(j)}. \tag{3.3}$$

As the discrimination of affect requires significant changes of the distributions across different affect levels, we defined a cumulative distance function $S(f|p)$ for

each phoneme as,

$$S(f|p) = w_p(D_{12}(f|p) + D_{23}(f|p) + D_{13}(f|p)), \tag{3.4}$$

where w_p is the frequency of occurrence of phoneme. Note that, the cumulative distance function $S(f|p)$ is expected to output larger distances if lip feature distribution changes largely over affective states.

3.4 Affect Classification

We used the segment level lip features to classify 3-level discrete affect values, and used the Support vector machine of LIBSVM [Chang and Lin, 2011] with radial basis kernel function as the classifier. Affect labels at sentence level, which are assigned by annotators, were used as ground truth to train and validate the classifiers. Classification task was performed at phoneme and sentence levels. These two approaches are briefly described in the following sub-sections.

3.4.1 Phoneme Level Classification

In the phoneme level classification, a classifier was constructed for each phoneme. The 44-dimensional segment level lip feature vectors, which were extracted over each occurrence of a phoneme, were used as input features. Over all the dataset, using the provided phoneme boundaries, all segments of each phoneme were extracted. Affect state level of each phoneme segment was extracted from the sentence level annotations. Classifiers were constructed for each phoneme and affect classification was evaluated over phoneme segments.

3.4.2 Sentence Level Classification

In the sentence level classification, we investigated two strategies: decision fusion and data fusion. In decision fusion, we applied the majority voting over the phoneme level classifier outputs. In the data fusion, we constructed segment level lip feature vectors over all duration of the sentence, and trained sentence level classifiers. On

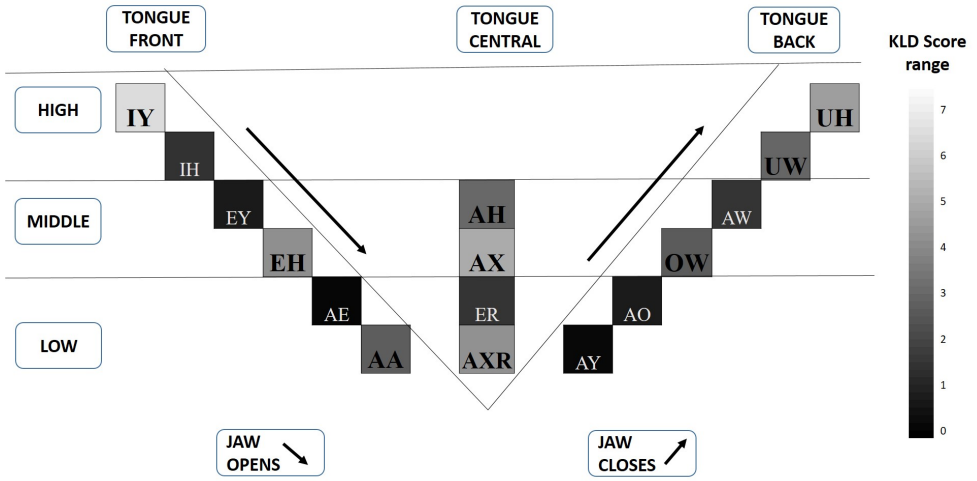


Figure 3.2: Cumulative discriminative distances of vowels according to manner and placement of articulation

sentence level, we also investigated the use of discriminative set of phonemes to the classification performance.

3.5 Results on Discriminative Phonemes

Over the whole database, we first computed the conditional probability mass functions for the lip features, then the cumulative distance functions were evaluated for each phoneme and for each affect attribute. We observed that more than 80% of the phonemes exhibiting higher KLD score were common in all three dimensions of affect, so we sorted the phonemes with respect to the average of distances over three affect attributes. Figure 3.2 and 3.3 respectively plots the cumulative discriminative distances of vowels and consonants according to manner and place of articulation. Note that discriminative distance gets lower as color gets darker. Hence, phonemes that are discriminating affective states better with the lip features have lighter colors. Among the vowels, the least discriminative region is observed as jaw opens and tongue is at back, and we select */IY/, /EH/, /AA/, /AX/, /AXR/, /AH/, /UH/, /UW/, /OW/* as the discriminative vowels for lip driven affect recognition.

Similarly among the consonants, the voiced palatal */Y/* and the voiced alveeolar */Z/* are the top two most discriminative phonemes for the affect recognition. On the

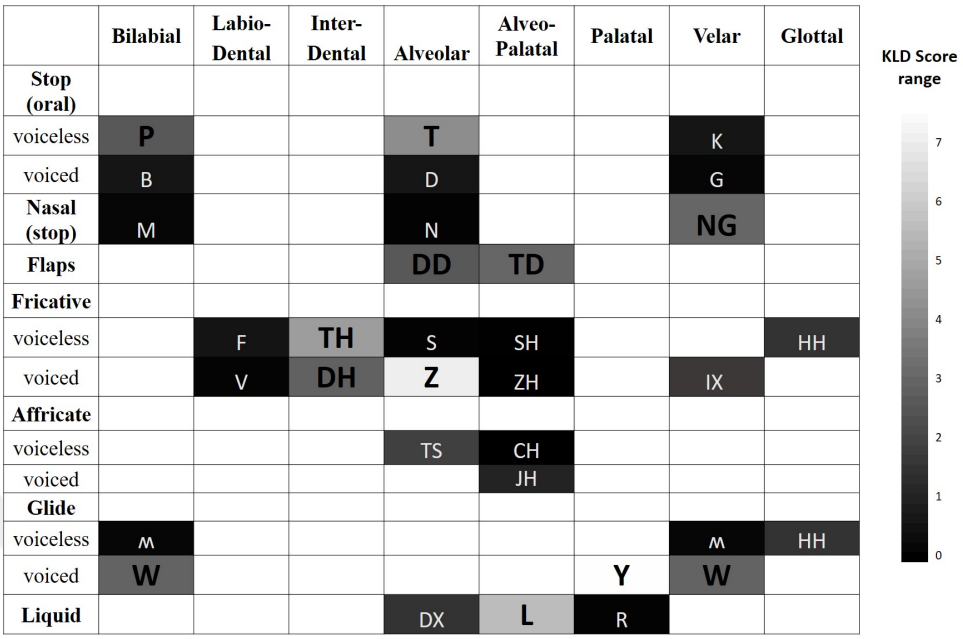


Figure 3.3: Cumulative discriminative distances of consonants according to manner and placement of articulation

other hand, the labio-dentals */F/* and */V/* and alveo-palatals */SH/, /ZH/, /CH/, /JH/* are among the least discriminative phonemes. We select */P/, /W/, /TH/, /T/, /TD/, /DH/, /Z/, /DD/, /L/, /Y/, /NG/* as the discriminative consonants for lip driven affect recognition.

In Figure 3.4, we present the top 10 most discriminative and bottom 5 least discriminative phonemes for the affect recognition task.

3.5.1 Affect classification Results

Classification experiments are organized in a leave-one-speaker-out cross validation scheme. Within our experiments, we use two performance evaluation methods: unweighted average accuracy (UA), which is the sum of all class accuracies, divided by the number of classes, and weighted accuracy (WA), which is the number of correctly recognized labels divided by total number of occurrences.

The results presented here are generated as speaker independent results. We compare classification performance of affect recognition task using all phonemes

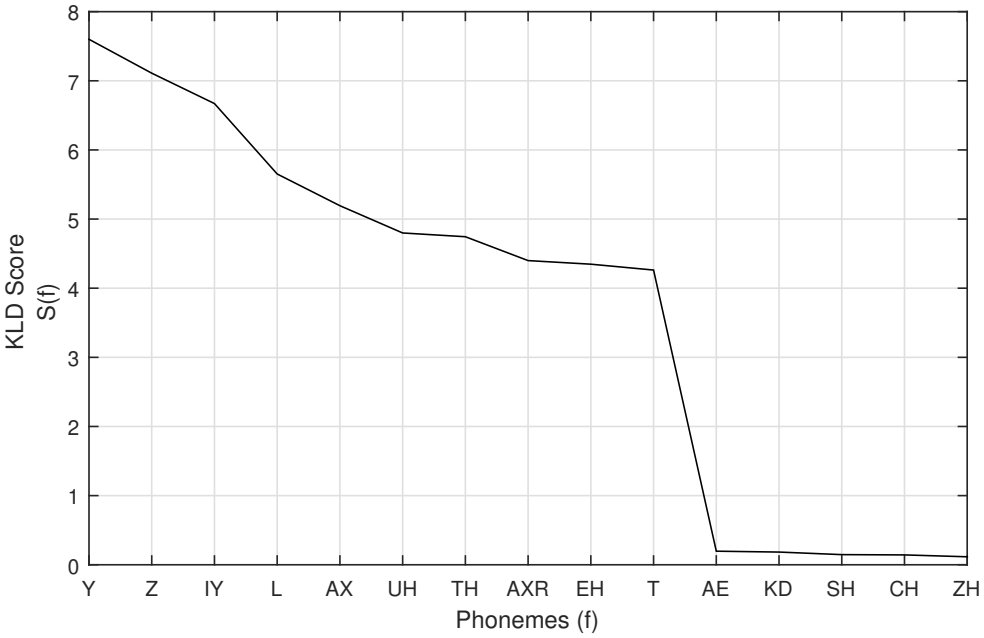


Figure 3.4: KLD Score of top 10 and bottom 5 phonemes

and the selected list of discriminative phonemes. Table 3.1 reports the sentence level classification performances using the data fusion. Similarly, Table 3.2 presents the sentence level classification performances using the decision fusion. Note that unweighted accuracy scores are in general lower, since number of samples from the affect state levels are unbalanced. We observe classification performance improvements with the selected list of discriminative phonemes for all affect attributes.

3.6 Summary

In this chapter, we explain the role of phonemes for affect recognition under lip articulations. We observe that using only lip features can attain affect recognition performance higher than the random. Our results also suggest that a selected list of discriminative phoneme articulations can better classify all affect attributes using only the lip features. These results are encouraging for the use of lip articulations in multimodal affect recognition systems. We also performed an analysis of the

Table 3.1: Comparison of sentence level weighted (WA) and unweighted (UA) classification accuracies of AVD attributes with selected phonemes and all phonemes using data fusion.

	Classification Accuracy (%)					
	All Phonemes			**Selected Phonemes**		
	A	V	D	A	V	D
WA	71.40	45.43	61.79	72.13	**46.44**	64.47
UA	40.37	38.82	38.53	41.42	**42.33**	**39.61**

Table 3.2: Comparison of sentence level weighted (WA) and unweighted (UA) classification accuracies of AVD attributes with selected phonemes and all phonemes using decision fusion.

	Classification Accuracy (%)					
	All Phonemes			**Selected Phonemes**		
	A	V	D	A	V	D
WA	71.86	45.59	64.74	**72.16**	46.16	**64.92**
UA	39.77	35.12	38.11	**42.65**	36.33	38.98

discriminative phonetic classes for the lip driven affect recognition as a function of manner and place of articulation. Extensive studies in this direction have potential to improve contribution of lip modality to the affect recognition systems. The results are also encouraging for the use of lower face area (lip and jaws) in speech driven facial features synthesis systems explained in the next chapters.

Chapter 4

MULTIMODAL SPEECH DRIVEN FACIAL SHAPE ANIMATION USING DEEP NEURAL NETWORKS

Speech driven facial animation translates dynamics of speech to articulation of facial gestures. Facial animations are essential, for their practical use in various applications such as on line virtual agents and other interactive human-computer interfaces. High quality speech driven animations are usually generated either by a skilled animator, or by re-targeting motion capture of an actor. The benefit of hand made animation is that the animator can accurately synthesize, style and time synchronize the animation, but it is costly and time consuming. The main alternative to this method is data driven (text or speech) animation by capturing and tracking facial motion of an actor's face [Beeler et al., 2010] [Cao et al., 2015] [Taylor and et all, 2017]. But in later method, there is trade off between quality and cost/time. In this targeted area of research, speech and text are among the most popular and effective modalities to generate facial animations.

In this chapter, we will see in detail of the proposed multimodal approach to enhance the previously developed methods for high fidelity production of data driven facial animations. Our proposed system can, not only benefit from the sequence to sequence mapping of text based generation of facial shape animations, but also utilizes the acoustic variability in the speech and merge them together to enhance the quality of generated animations. Most of conventional studies utilized the data either gathered in a neutral emotionless way or with highly emotional states, in a controlled studio environment. We show that using speech features along with the text features not only improves the accuracy on the affective data remarkably, but turns out to be more efficient for neutral data as well.

This chapter covers the details of the datasets we utilized for multimodal

speech driven facial shape animation; GRID, an audio visual dataset recorded in neutral without expressing emotions, and Surrey Audio-Visual Expressed Emotion (SAVEE), recorded with different categorical emotions. Then, the feature representation is given, followed by description of the network architectures employed for the multimodal speech-driven facial animation system.

4.1 Datasets

GRID [Cooke et al., 2006], an audio visual-corpus, consists of audio and video recordings of 1000 sentences spoken by each of 34 speakers. The sentences are drawn from the following simple grammar: command (4) + color (4) + preposition (4) + letter (25) + digit (10) + adverb (4). The digits in parenthesis represent the number of choices for each of the 6 word categories. None of the spoken sentences contain any emotional content and uttered in a neutral way. Videos are recorded at a rate of 25 fps and audio is sampled at 25 kHz. Word transcriptions are provided along with dataset. A total of approximately 2.5 million frames are available to train and validate models.

The Surrey Audio-Visual Expressed Emotion (SAVEE) database [Haq and Jackson, 2011] consists of footage of 4 British male actors with six basic emotions(disgust, anger, happy, sad, fear surprise) and neutral state. A total of 480 phonetically balanced sentences are selected from the standard TIMIT corpus [Garofolo et al., 1993] for every emotional state. Audio is sampled at 44.1 kHz with video being recorded at a rate of 60 fps. Phonetic transcriptions are provided with the dataset. A total of approximately 102 K frames are available to train and validate models. Actors' face is painted with blue markers for tracking of facial movements during the recordings.

4.2 Feature Extraction

A key factor in training the neural networks is representation of the inputs and outputs of the model. In this study we utilized speech features along with

Figure 4.1: An example of facial feature extraction on SAVEE dataset. Blue points are original markers painted on actor's face. Dlib extracted points (red) are used for a more descriptive shape of the lower face.

the underlying phoneme labels to train a model to estimate the visual facial features.

4.2.1 Text Features

We utilized the underlying phoneme labels of the input speech as text features. In the preparation of text features, we used Montreal forced aligner [McAuliffe et al., 2017], to generate the phoneme transcription files. Each phoneme can span variable length video frames, depending upon the length of its occurrence in a specific sentence. A standard set of 41 phonemes was used for transcribing the data, including silence and short pause. Following the work presented in [Taylor and et all, 2017], One hot encoding was used to represent the phoneme indicator feature corresponding to each video frame. The set of extracted phoneme features is represented as $\left\{f_j^p\right\}_{j=1}^{N}$, where $f_j^p \in \mathbb{R}^{41\times 1}$ and N is the number of instances in the training set.

4.2.2 Speech Features

Raw speech signal contains confined spectral information that can be valued for speech to facial mapping. Inspired by the work of [Abdel-Hamid et al., 2014] we used Mel-Frequency Spectral Coefficients, also denoted as MFSC features. For each speech frame, 40 MFSC features were extracted using HTK toolkit [Young and et al, 2006] to define the acoustic energy distribution over 40 mel-frequency bands. These features were computed on short term overlapping Hamming-windows over the speech, with sampling interval set according to the frame rate of the corresponding videos. The window size was chosen as 8 ms and 10 ms for the SAVEE and GRID datasets respectively, resulting in a 2-to-1 and 4-to-1 audio to video frame correspondence. All speech features were z-score normalized to have zero mean and unit variance in each feature dimension. We represent the set of acoustic feature vectors as $\left\{f_j^a\right\}_{j=1}^{N}$, where $f_j^a \in \mathbb{R}^{40\times r}$ and r is the audio to video frame ratio.

4.2.3 Visual Features

The raw visual features are described with a set of coordinate points on the lower face region, along the jaw line, nose, inner and outer lips, and represented as:

$$S = [x_1, y_1, x_2, y_2, ..., x_M, y_M]^T. \tag{4.1}$$

The Dlib facial landmark detector [King, 2009] was used to detect a total of $M = 36$ landmark points on the lower face region for each video frame in both datasets. Figure 4.1 presents extraction of sample landmark points using the Dlib detector.

We utilized Active Shape Model (ASM) to remove the correlation in the training set and for dimensionality reduction as well [Cootes and Taylor, 1993]. In the ASM, a statistical model is trained for a set of shapes $\left\{S_1,S_2,...,S_N\right\}$, by applying Principal Component Analysis (PCA) to model the variation of data around the mean shape. Prior to building the shape model, it is necessary to filter out the possible rotation,

scale and position difference of shapes from one frame to another. The Generalized Procrustes Analysis (GPA) was used for the shape alignment purpose [Goodall, 1991]. In GPA, all the shapes from the training set are mean removed and scaled to have unit norm. The training set is then aligned iteratively by minimizing the distance of each shapes to the reference mean until convergence. After alignment of the shapes, the mean shape and the covariance matrix are then computed as:

$$\bar{S} = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{4.2}$$

$$C = \frac{1}{N-1}\sum_{i=1}^{N}(S_i - \bar{S})(S_i - \bar{S})^T \tag{4.3}$$

In the shape model, each shape is described by a set of parameters in a lower dimensional space. Given the model parameters, the corresponding shape S in the original space is generated using:

$$S = \bar{S} + Pf_v, \tag{4.4}$$

Upon training the shape model, the visual features were chosen as the projection of shapes in the training set to the PCA space, i.e. the model parameters f_v, which are computed as:

$$f_v = P^T(S - \bar{S}). \tag{4.5}$$

We represent the set of visual feature vectors as $\left\{f_j^v\right\}_{j=1}^{N}$, where $f_j^v \in \mathbb{R}^{18\times 1}$.

4.3 Feature Representation

The raw training data used in this study comprises a sequence of 40-dimensional MFSC features and a sequence of 41-dimensional phoneme indicator features, as input sequences, and a sequence of 18-dimensional PCA features as the output representation. To capture the temporal nature of the speech and visual features, often temporal sliding windows are utilized, as in [Taylor and et all, 2017]. We

also utilized sliding windows of size K_a, K_p, K_v over the MFSC, phoneme and PCA sequences, respectively. For each instance in the raw training data, the sliding window covers a neighborhood around the instance, with the instance at the window's center, converting the original feature sequences into a set of overlapping features. Therefore at each training instance the temporal feature representation is calculated as:

$$F_j^m = \left\{(f_{j-k_m}^m, ..., f_j^m, ..., f_{j+k_m}^m)\right\}_{j=1}^N \tag{4.6}$$

Where m indicates the modality i.e. $m \in \{a, p, v\}$ and $k_m = (K_m - 1)/2$ is the half window size.

For the output PCA and input phoneme indicator sequences, the feature vectors covered inside the window are concatenated column-wised to create samples $F_j^v \in \mathbb{R}^{18K_v \times 1}$ and $F_j^p \in \mathbb{R}^{41K_p \times 1}$, respectively. For the MFSC input sequence, feature vectors inside the window are stacked through the 3-d dimension yielding an instance $F_j^a \in \mathbb{R}^{40 \times 1 \times rK_a}$, which could be interpreted as a $(40, 1)$ image with depth rK_a. Note that the scaler r is the speech to video frame ratio, hence $r = 2$ and $r = 4$ for the SAVEE and GRID datasets respectively.

4.4 Network Architecture

From a machine learning point of view, speech animation is described as a multi-variate regression model i.e. the real-valued temporal output features are estimated given the input features.

We investigate three different DNN for the speech animation task. The text-based method proposed in [Taylor and et all, 2017] is used as a baseline in this study. Despite having some benefits such as making the model speaker and language independent, a shortcoming of the text-based approach will arise when processing affective data. We show that using speech features along with the text features, improves the performance of the speech animation system for both affective

and neutral datasets.

4.4.1 Text-based Architecture

For text-based experiments, a deep Multilayer Perceptron (MLP) similar to what is described in [Taylor and et all, 2017] was used. The input layer, accepting indicator features, is connected to three fully connected (FC) hidden layers with 1024 neurons each and a final output layer. To induce the non-linearity, each fully connected layer is followed by a hyperbolic tangent activation function. We employed standard mini batch stochastic gradient descent algorithm for training. To counteract over-fitting, dropout [Srivastava et al., 2014] with 50% probability was used. Mini batch size was selected as 128 along with Adam optimizer [Kingma and Ba, 2014] for learning rate adaptation. The final output layer is standard multi-variate regression layer predicting the PCA sequence and trained to minimize the MSE loss.

4.4.2 Speech-based Architecture

Convolutional Neural Networks (CNN) are shown to be efficient in extracting discriminative features from speech [Abdel-Hamid et al., 2014]. We employed a CNN architecture for speech-based model training, as another baseline to our proposed multimodal approach.

The image-like temporal speech features are fed to the input layer of the speech-based network. The network contains two convolution layers with first layer having 64 filters of size 7, followed by a pooling layer with window size of 4 and stride of 2. In the second convolution layer, we decreased the filter size to 5 and increased the number of filters to 128 and a pooling layer with window size and stride of 2 was chosen. The network is then connected to two fully connected layers with 1024 hidden neurons each. We used dropout regularization method with probability 50% in the fully connected layers only, to overcome over-fitting. Hyperbolic tangent activation function was used at each layer with Adam optimizer for hyper learning rate optimizations.

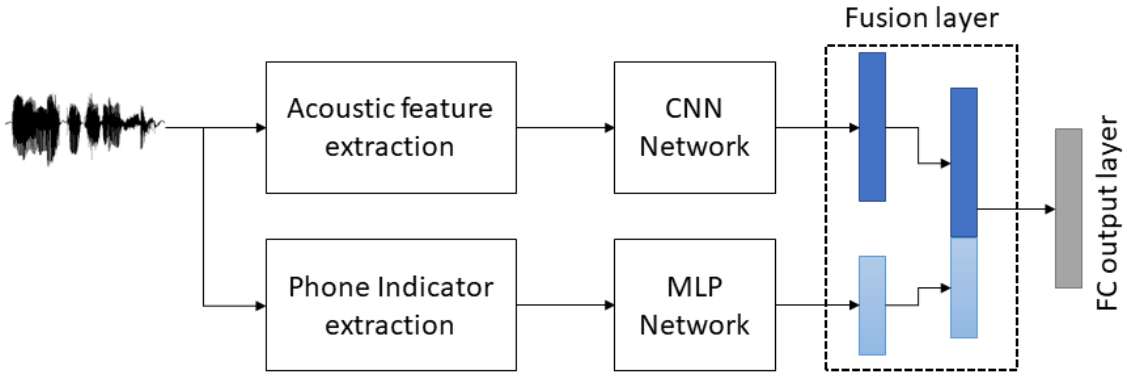


Figure 4.2: Proposed multimodal-based network. Acoustic and Phoneme features are extracted from speech and fed to the network separately, merged at an intermediate level and connected to the fully connected output layer.

4.5 Proposed Multimodal Architecture

The proposed multimodal approach is a combination of the text-based and speech based networks, hence gaining the advantage of text features for a speaker independent model and benefiting from the speech features for discriminating different affective content. We utilized a *fusion* strategy to update the output layer's weights according to the merged hidden neurons of the two modalities, during optimization. Text and speech features are fed separately to the network and later concatenated in last layer of the network (before the output layer). The fused neurons are then connected to the fully connected output regression layer as shown in Figure 4.2. Similar deep MLP and CNN structures were employed, as described in sections 4.4.1 and 4.4.2, for text and speech inputs, respectively.

4.6 Results

To evaluate the performance of the proposed method, MSE loss between the ground truth and the ones estimated by the model was calculated on both parameter and original shape space. The predicted output gives the shape model parameters in a

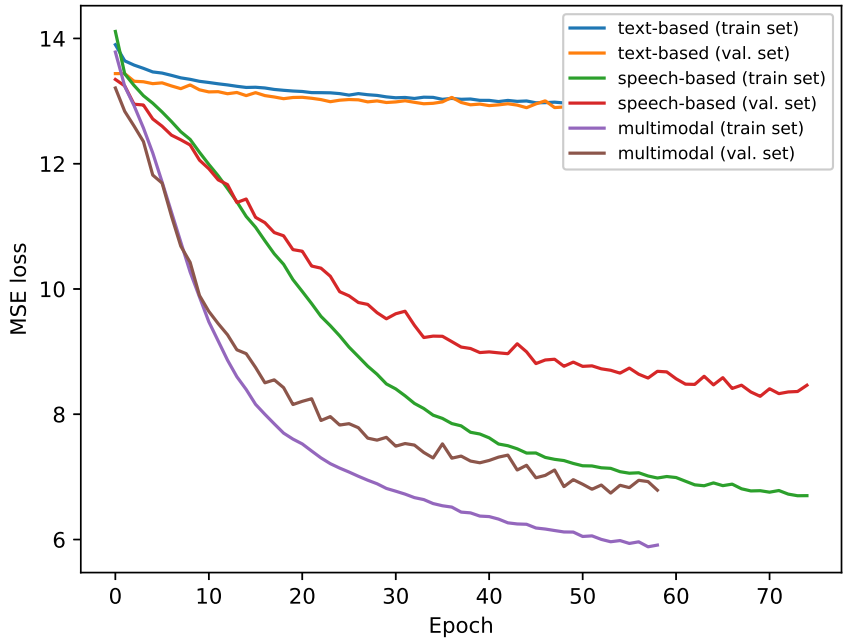


Figure 4.3: Train and validation loss curve for three different network architectures on SAVEE dataset.

temporal window of size K_v. The model parameters at a given frame are calculated as the temporal mean of the network's output and later used to estimate the shape contours in the original shape space, using (4.4). The shape space loss is reported per landmark point in a 150×150 frame scale.

For SAVEE dataset with a total number of $102K$ samples, 90% of data was used for training the model and the rest 10% for validation set. Though the number of data samples in GRID dataset is a lot more than in SAVEE (approximately $2.5M$ samples), for a fair comparison we used same number of samples to train the models.

Aside from the network characteristics, the sliding window sizes are important hyper parameters which need to be selected carefully. The optimal sliding window sizes over the raw feature sequences were chosen as $(rK_a, K_p, K_v) = (30, 7, 3)$ and $(28, 11, 5)$ after fine tuning the proposed multimodal network, for SAVEE and GRID datasets respectively.

Figure 4.3 shows the MSE loss curve through the learning process, for three

Table 4.1: Performance evaluation of existing and proposed method over emotional and neutral datasets. Values indicate the MSE in the Parameter and shape space.

MSE	**Dataset**	**Text-based [Taylor and et all, 2017]**	**Speech-based**	**Multimodal**
PCA space	SAVEE	12.9	8.28	**6.75**
	GRID	10.7	7.86	**7.26**
Shape space	SAVEE	0.76	0.53	**0.42**
	GRID	0.53	0.44	**0.39**

different discussed architectures, trained and validated on the emotional SAVEE dataset. As it is obvious from the figure, the proposed multimodal method performs remarkably better than the text-based network on emotional data. The non-decreasing learning curve for the text-based network indicates the failure of the model to learn the variation in the emotional output data using only phoneme features. Whereas in the multimodal case, adding speech features enables the model to capture the output sequence more accurately, hence yielding a smaller MSE loss.

The proposed method performs better on the neutral GRID dataset as well in terms of the MSE in the shape and parameter space. However we observe that the benefit which the multimodal network brings to the text-based model is slightly less than the emotional case. Table 4.1 presents the comparison of mean squared error between synthesized and original features in the parameter and original shape space. We observe that merging the two modalities give a clear edge over using either of the modalities, with both neutral and emotional datasets.

4.7 Summary

In this chapter, we explained our novel multimodal approach for generating face animations synchronized with the input speech. We explained the dataset used in the study of multimodal facial feature synthesis process. We discussed about the audio visual features. A statistical shape model was trained to project the shape data into an uncorrelated lower dimensional parameter space, capturing 98% of variation in the training data and used as visual features. We proposed a multimodal deep neural network with acoustic and phoneme features as inputs, to estimate the facial

shape parameters. The proposed approach was shown to perform better on neutral and emotional datasets, in terms of MSE loss in the shape and parameter space. We observed that the benefit of the proposed multimodal approach over the baseline text and speech-based methods was more remarkable for the affective dataset. As a next step we further investigate speech animation under different emotional states. We also deploy the concept of domain adaptation to overcome the major problem of scarce affective data. Since, the emotions are quite subjective and vary from person to person, We also use the generated shape animations for a user study performance evaluation.

Chapter 5

EMOTION DEPENDENT DOMAIN ADAPTATION FOR SPEECH DRIVEN AFFECTIVE FACIAL FEATURE SYNTHESIS

In this chapter, we will discuss the affective speech driven facial animation generation process. We model the affective speech animation problem as a cascade of emotion classification followed by facial parameter regression as depicted in Figure 5.1. The affective content of the input speech is first classified into one of the 7 emotion categories, then the corresponding emotion dependent domain adaptation (EDDA) model maps the spectral speech representation to the visual shape parameters. Block diagram of the proposed EDDA framework is given in Figure 5.2. The EDDA framework consists of a learning phase and a synthesis phase. In the learning phase, emotion dependent A2V mapping models are extracted by mSDA based domain adaptation, where the knowledge of abundant neutral data is transferred through a parallel dataset across neutral and affective speech signals. In the model, six emotion and the neutral categories have been considered, and hence the learning phase repeats model generation for each emotion separately. In the synthesis phase, an emotion recognition system produces emotion class likelihoods that are used to fuse visual synthesis outputs of all emotion dependent A2V mapping models.

Detailed descriptions of the feature representations, parallel data generation, domain adaptation, speech emotion recognition and visual synthesis are given in the following sections.

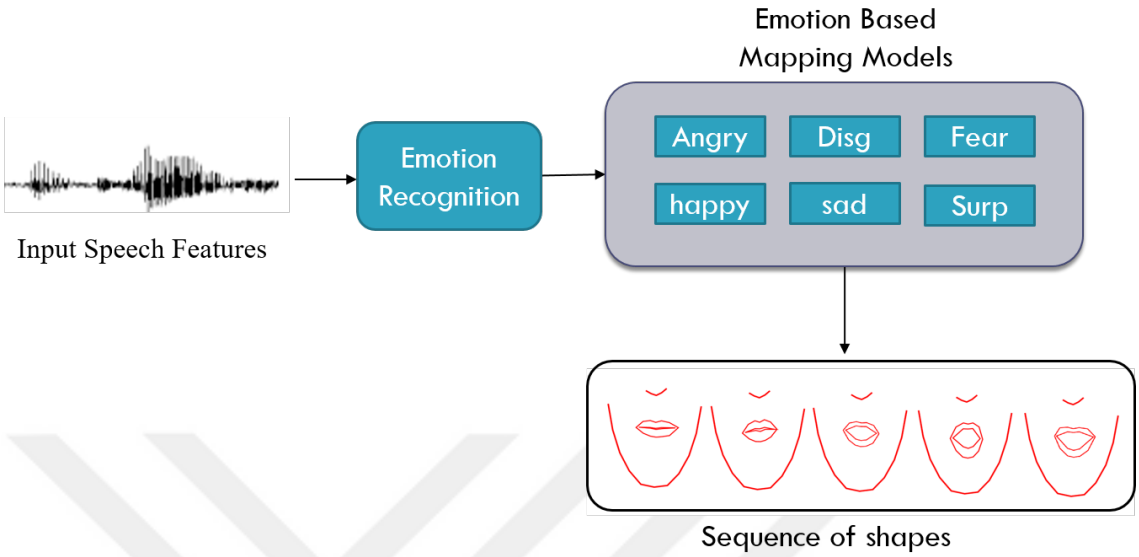


Figure 5.1: Emotion dependent framework for facial animation system

5.1 Feature Representations

5.1.1 Speech Feature Extraction

Speech signal confines useful spectral information for A2V mapping. Any utterance with different emotion articulations is expected to have variations in the spectral representation. Figure 5.3 presents sample spectrograms for different emotion settings.

We use the Mel-Frequency Spectral Coefficient (MFSC) representation as the speech signal features [Abdel-Hamid et al., 2014]. We use the Mel-Frequency Spectral Coefficient (MFSC) representation as the speech signal features [Abdel-Hamid et al., 2014]. Note that MFSC features are the mel-scaled spectral band energies and the widely used MFCC features are defined after the discrete cosine transform (DCT) of the MFSC features. Recently, increasing number of studies tend to use the MFSC features as the inputs to the deep CNN architectures [Sehgal and Kehtarnavaz, 2018, Abdel-Hamid et al., 2014, Tóth, 2015]. There are two main motivations for this choice. First, CNN architectures are good feature extractors and able to replace the DCT feature compaction through learning.

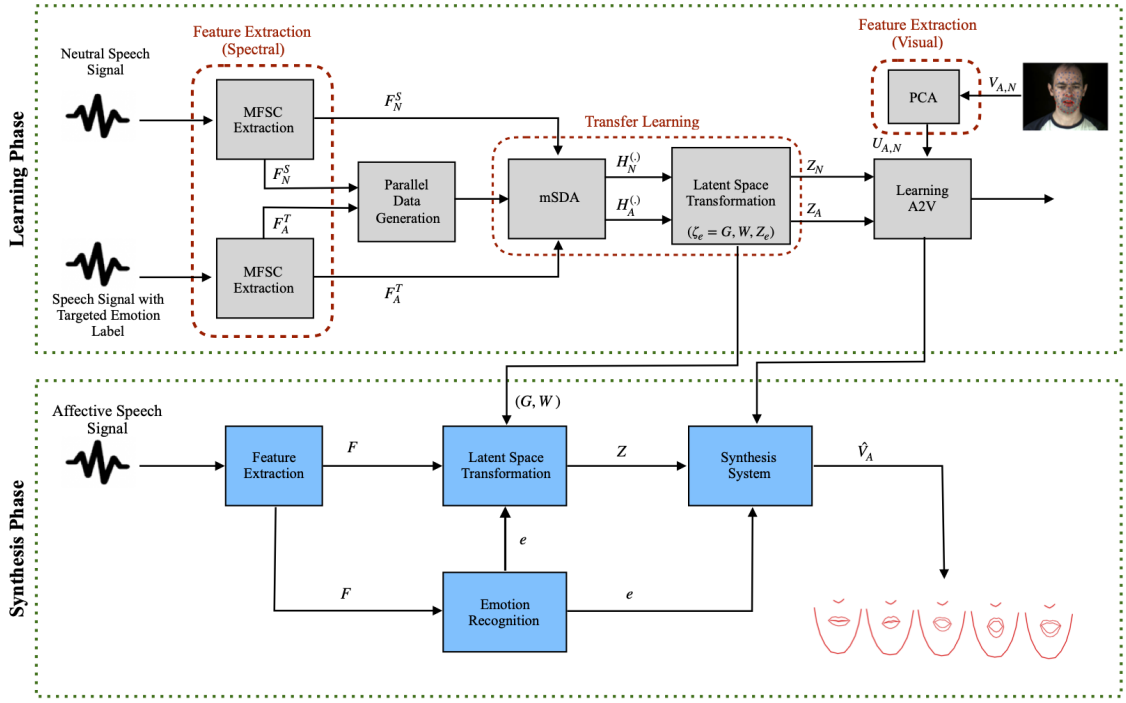


Figure 5.2: The proposed emotion dependent domain adaptation (EDDA) framework for the A2V mapping

Second, preserving the locality in frequency domain is expected to be beneficial for the convolution operator within the CNN architecture.

For each speech frame, MFSC features are extracted to define the acoustic energy distribution over $M = 40$ mel-frequency bands. These features are computed on short-term overlapping Hamming-windows of 20 ms over the speech signal with a frame shift of 10 ms. Hence, the frame rate of MFSC features are set to 100 fps. All speech features are z-score normalized to have zero mean and unit variance in each feature dimension. We represent the MFSC feature vector at the j-th frame as $f_j \in \mathbb{R}^M$.

A spectral image over a temporal sliding window is defined to capture the spatio-temporal characterization of the speech signal as $F_j = [f_{j-\Delta}, ..., f_j, ..., f_{j+\Delta}] \in \mathbb{R}^{M \times T}$, where $T = 2\Delta + 1$ is the temporal length of the spectral image at time frame j.

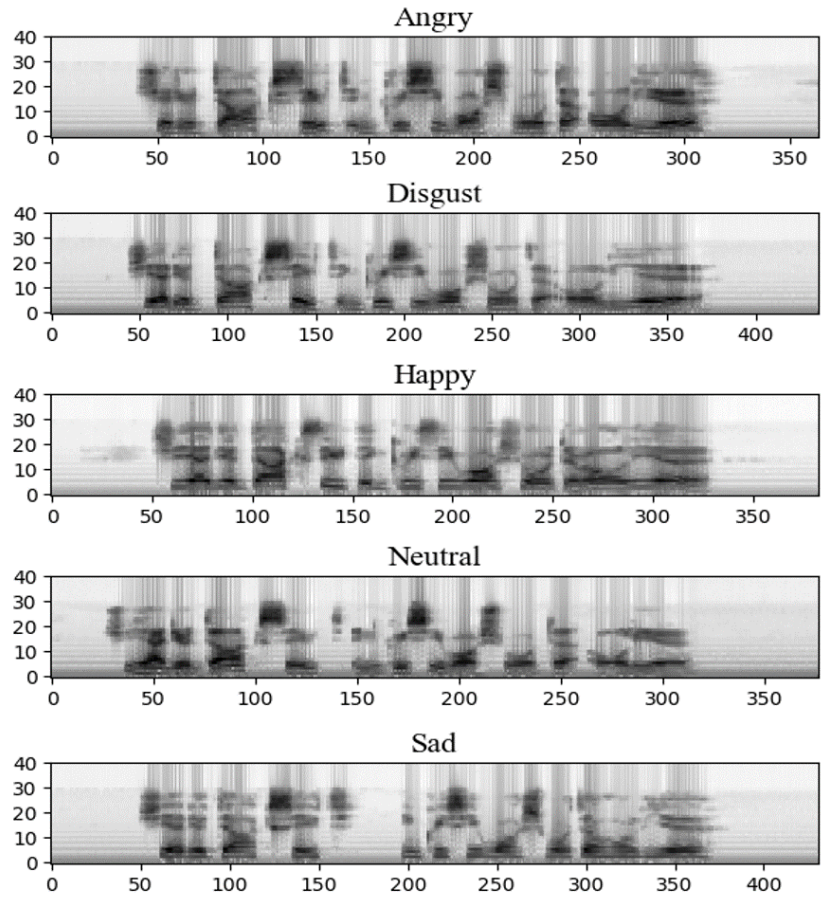


Figure 5.3: Spectrograms of the sentence "she had your dark suit in greasy wash water all year" from the SAVEE data set with five different emotions.

5.1.2 Visual Feature Extraction

A similar approach as described in 4.2.3 of multimodal study has been employed in the emotion dependent experiments also. The raw visual features are extracted as a set of 2D coordinate points on the lower face region, along the jaw line, nose, inner and outer lips. For this purpose, the Dlib facial landmark detector is used to extract D landmark points on the lower face region for each video frame [King, 2009]. The Dlib detector is based on the classic Histogram of Oriented Gradients (HOG) feature along with a linear classifier, an image pyramid and a sliding window detection scheme. Since the video frame rate is 25 fps, the resulting raw visual

feature vectors are up-sampled by 4 to match the frame rate of the speech features, and the raw visual feature at frame j is represented as

$$v_j = [x_1, y_1, x_2, y_2, ..., x_D, y_D]^\intercal, \tag{5.1}$$

where x and y are the indexed coordinates of the $D = 36$ facial landmark points. Figure 5.4 presents a sample set of extracted landmark points using the Dlib detector.

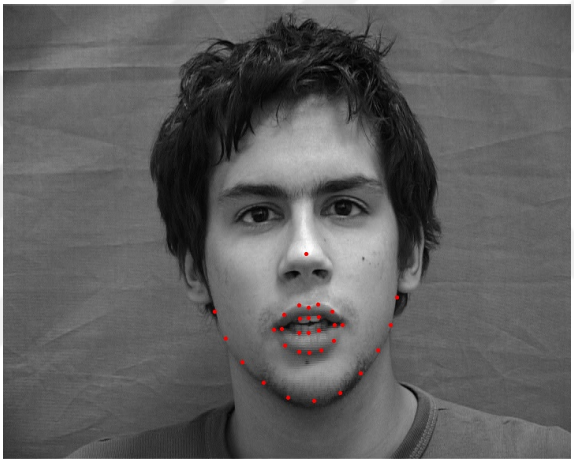


Figure 5.4: A sample set of facial landmark points using the Dlib detector

We utilized Active Shape Model (ASM) to remove the correlation in the facial landmark representation and for dimensionality reduction [Cootes and Taylor, 1993]. In the ASM, a statistical model is extracted over the set of visual feature vectors in the training dataset using the Principal Component Analysis (PCA), which models the variation of data around the mean shape. Prior to building the shape model, it is necessary to filter out the possible rotation, scale and position differences of shapes. In the literature, Generalized Procrustes Analysis (GPA) has been successfully used for the shape alignment problem [Goodall, 1991]. In GPA, all the shapes from the training dataset are mean removed and scaled to have unit norm. The training set is then aligned iteratively by minimizing the distance of each shape to the reference mean until convergence. Once the shapes are aligned, they are transformed to the

PCA domain. Hence, the final visual feature vector at frame j is extracted after the GPA and PCA transformations as,

$$u_j = PCA(GPA(v_j)) \in \mathbb{R}^N, \tag{5.2}$$

where the feature dimension is reduced to $N = 18$ by capturing 98% of the variation in the dataset.

Figure 5.5 presents the three main modes of variation around the mean facial shape, varying the model parameters between ± 3 of their standard deviation.

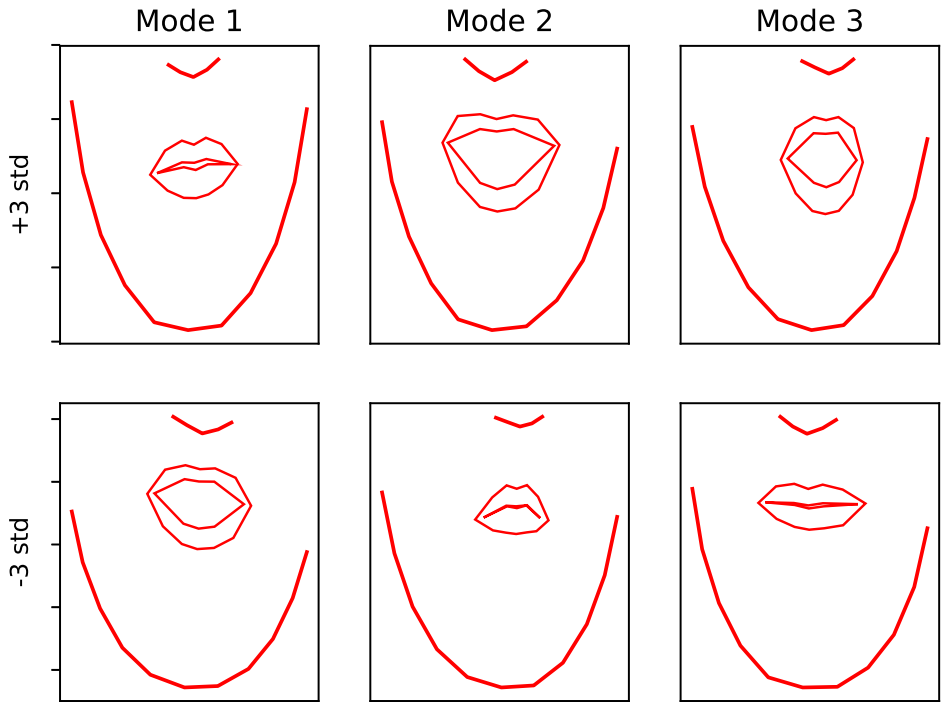


Figure 5.5: Three modes of variation around mean facial shape

5.2 Parallel Data Generation

In the proposed EDDA framework, learning phase requires to have a parallel dataset of speech feature representations across neutral and affective speech signals to perform the mSDA based domain adaptation. Note that it is not straight forward to identify a parallel sequence of feature representations between neutral and affective speech signals. For this purpose, we propose a two-stage parallel data generation scheme. First, we perform a phoneme-level alignment between the neutral and affective domains, so that we can sample the parallel feature sequence over the matching phonetic context. Second, we perform a frame-by-frame alignment across the neutral and affective pair of feature sequences using the dynamic time-warping algorithm.

The smallest perceptually distinct speech sound is defined as phoneme. The pronunciations of a phoneme differ depending on the neighboring phonemes that defines the context of the phoneme. In this study we define the context of a phoneme ϕ_i considering the two previous and two following phonemes in the utterance as a phoneme-context,

$$C(\phi_i) = \{\phi_{i-2}, \phi_{i-1}, \phi_i, \phi_{i+1}, \phi_{i+2}\}, \tag{5.3}$$

where i is the phoneme sequence index and each phoneme ϕ_i includes a sequence of MFSC features starting from frame $n_i + 1$ for a frame duration of d_i as

$$\ddot{f}_{\phi_i} = \{f_{n_i+1}, f_{n_i+2}, \ldots, f_{n_i+d_i}\}. \tag{5.4}$$

Then the parallel feature dataset is defined as the union of phone feature sequences from the source and target domains with matching phone-context as

$$D_p = \bigcup_{\substack{\forall \phi_i, \phi_j \\ \text{s.t. } C(\phi_i)=C(\phi_j)}} (\ddot{f}^S_{\phi_i}, \ddot{f}^T_{\phi_j}), \tag{5.5}$$

where $\ddot{f}^S_{\phi_i}$ and $\ddot{f}^T_{\phi_j}$ are representing phone feature sequences from the source and

target domains respectively. Furthermore, the matching phone-context requires equality of all five phones in the $C(\phi_i)$ and $C(\phi_j)$. Note that duration of the feature sequences from the source and target domains could differ. In order to construct a parallel association between the source and target domain feature sequences, a dynamic time warping algorithm [Salvador and Chan, 2007] is executed across the phone feature sequences $\ddot{f}^S_{\phi_i}$ and $\ddot{f}^T_{\phi_j}$. A graphical representation of the phone-context matching and feature sequence association are given in Figure 5.6.

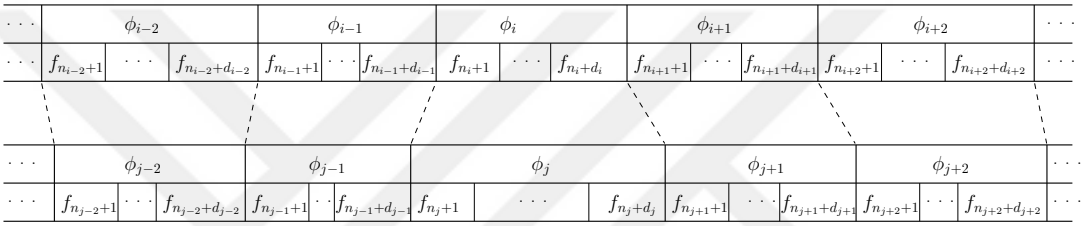


Figure 5.6: Phone-context matching and feature sequence association for the parallel data generation

5.3 Domain Adaptation

In this section, we define a domain adaptation scheme across affective and neutral speech representations. Given that spectral representations of affective and neutral speech statistically differ, our domain adaptation approach tries to separately map affective and neutral speech representations to a common latent space in which their statistical distributions are more similar and the cross-domain bias is smaller. For this purpose we adopt the mSDA approach for mapping the affective and neutral domain representations to the latent space [Zhou et al., 2014]. The details of the mSDA based domain adaptation are presented in the remaining of this section.

We can set the affective speech representation as the target domain, $\mathbf{D}_T = \{f_1^T, f_2^T, \ldots\}$, and the neutral speech representation as the source domain, $\mathbf{D}_S = \{f_1^S, f_2^S, \ldots\}$, where f_i^T and f_j^S are representing the MFSC feature vectors at frames i and j in the target and source domains, respectively. The parallel dataset of

speech feature representations across neutral and affective speech signals is defined in Section 5.2. Note that the parallel data, $\mathbf{D}_p = \{\mathbf{D}_S^p, \mathbf{D}_T^p\}$, is a collection of source, $\mathbf{D}_S^p$, and target, $\mathbf{D}_T^p$, domain data that are associated to each other in a one-to-one manner. Then the mSDA learns weight matrices $\mathbf{W}_T$ and $\mathbf{W}_S$, and a mapping matrix $\mathbf{G}$ to transform target and source domains to the latent space representation.

We first learn the weight matrix $\mathbf{W}_T \in \mathbb{R}^{M\mathtt{x}M}$ by minimizing the squared loss function,

$$\mathcal{L} = \sum_{i=1}^{R} \left\| \mathbf{D}_T - \mathbf{W}_T \, \mathbf{D}_T^{(i)} \right\|^2, \tag{5.6}$$

where $\mathbf{D}_T^{(i)}$ denotes the i-th corrupted version of $\mathbf{D}_T$. For the corruptions we use additive white Gaussian noise (AWGN) with variable signal to noise ratio. Then, the closed form solution of the weight matrix is extracted as

$$\mathbf{W}_T = \mathbf{P}_T \mathbf{Q}_T^{-1} \tag{5.7}$$

with $\mathbf{Q}_T = \widetilde{\mathbf{D}}_T \widetilde{\mathbf{D}}_T^{\intercal}$ and $\mathbf{P}_T = \overline{\mathbf{D}}_T \widetilde{\mathbf{D}}_T^{\intercal}$, where $\overline{\mathbf{D}}_T$ is R-times repeated version of target data as $\overline{\mathbf{D}}_T = [\mathbf{D}_T, \ldots, \mathbf{D}_T]$ and $\widetilde{\mathbf{D}}_T$ is the corrupted version of $\overline{\mathbf{D}}_T$. Later, the hyperbolic tangent squashing function is applied on the output of each mSDA layer, $\mathbf{H}_T = \tanh(\mathbf{W}_T \mathbf{D}_T)$, to extract non-linear feature representation. Similarly, the weight matrix for the source, $\mathbf{W}_S \in \mathbb{R}^{D\mathtt{x}D}$, and $\mathbf{H}_S$ can be calculated.

The first phase of the mSDA computes the non-linear feature representations $\mathbf{H}_T$ and $\mathbf{H}_S$. In the second phase, a mapping matrix $\mathbf{G} \in \mathbb{R}^{M\mathtt{x}M}$ is extracted to transform target domain representation to source domain by using the cross-domain parallel data $\mathbf{D}_p$. Let us denote $\mathbf{H}_T^p$ and $\mathbf{H}_S^p$ as encoded cross-domain non-linear features for the parallel target and source data, respectively. A feature transformation $\mathbf{G}_k$ for the k-th mSDA layer is extracted by minimizing the objective,

$$\left\| \mathbf{H}_{S,k}^p - \mathbf{G}_k \mathbf{H}_{T,k}^p \right\|^2 + \lambda \left\| \mathbf{G}_k \right\|^2, \tag{5.8}$$

where $\lambda > 0$ is a regularization term on $\mathbf{G}_k$ and the subscript k refers to the layer index. The optimization problem has a closed-form solution,

$$\mathbf{G}_k = \left[\mathbf{H}^p_{S,k}\left(\mathbf{H}^p_{T,k}\right)^\intercal\right]\left[\mathbf{H}^p_{T,k}\left(\mathbf{H}^p_{T,k}\right)^\intercal + \lambda\mathbf{I}\right]^{-1}, \tag{5.9}$$

where $\mathbf{I} \in \mathbb{R}^{M\mathrm{x}M}$ is the identity matrix.

The mSDA approach consists of stacked layers of transformations, which are represented by the weight and mapping matrices. Algorithm 1 presents the flow of the domain adaptation scheme for a K-layer architecture using the above mSDA formulation. The final latent space representations are extracted as $\mathbf{Z}_T$ and $\mathbf{Z}_S$ for the target and source domains, respectively.

Algorithm 1: Domain adaptation through the mSDA

Input : Target and source domain data: $\mathbf{D}_T$ and $\mathbf{D}_S$
Parallel data: $\mathbf{D}^p_T$, $\mathbf{D}^p_S$
Regularization parameter: λ
Total number of layers: K

Output : Transformed target data: $\mathbf{Z}_T$
Transformed source data: $\mathbf{Z}_S$

Initialize: $\mathbf{H}_{T/S,1} = \mathbf{D}_{T/S}$ and $\mathbf{H}^p_{T/S,1} = \mathbf{D}^p_{T/S}$
$\mathbf{G}_1 = \left[\mathbf{H}^p_{S,1}\left(\mathbf{H}^p_{T,1}\right)^\intercal\right]\left[\mathbf{H}^p_{T,1}\left(\mathbf{H}^p_{T,1}\right)^\intercal + \lambda\mathbf{I}\right]^{-1}$
for $k = 2, \ldots, K$ **do**
    Get $\mathbf{W}_{T,k}$ and $\mathbf{W}_{S,k}$ using $\mathbf{H}_{T,(k-1)}$ and $\mathbf{H}_{S,(k-1)}$
    $\mathbf{H}_{T/S,k} = \tanh\left(\mathbf{W}_{T/S,k}\mathbf{H}_{T/S,(k-1)}\right)$
    $\mathbf{H}^p_{T/S,k} = \tanh\left(\mathbf{W}_{T/S,k}\mathbf{H}^p_{T/S,(k-1)}\right)$
    $\mathbf{G}_k = \left[\mathbf{H}^p_{S,k}\left(\mathbf{H}^p_{T,k}\right)^\intercal\right]\left[\mathbf{H}^p_{T,k}\left(\mathbf{H}^p_{T,k}\right)^\intercal + \lambda\mathbf{I}\right]^{-1}$
end
Compute the outputs $\mathbf{Z}_T = [(\mathbf{G}_1\,\mathbf{H}_{T,1})^\intercal, (\mathbf{G}_2\,\mathbf{H}_{T,2})^\intercal, \ldots, (\mathbf{G}_K\,\mathbf{H}_{T,K})^\intercal]^\intercal$
$\mathbf{Z}_S = [(\mathbf{H}_{S,1})^\intercal, (\mathbf{H}_{S,2})^\intercal, \ldots, (\mathbf{H}_{S,K})^\intercal]^\intercal$

5.4 Speech Emotion Recognition

Emotion recognition from speech has been widely studied in the literature and the speech spectrum is known to discriminate emotions well. As demonstrated

in Figure 5.3, articulation of the same utterance within different emotional states exhibits observable variations in the MFSC spectrograms.

We construct the speech emotion recognition (SER) as a classification problem to one of the seven emotion classes, which are angry, disgust, fear, happy, sad, surprise and neutral. In the proposed EDDA framework, SER has been performed at the utterance level, that is SER output for each utterance has been used as a side information in the affective A2V mapping. Since reliability of the SER output is significantly higher for the top two predicted emotion classes with respect to the most likely class, we consider to extract the top two predicted emotion classes with their normalized likelihood scores.

We set the speech spectral image F_j as the input of a CNN based classifier to recognize one of the seven emotion classes for each frame j. The SER-CNN classifier includes three convolutional layers with max pooling and ReLu as the non-linear activation function. The three convolutional layers are followed by a fully connected layer and a softmax layer at the output. Table 5.1 presents detailed description of the SER-CNN.

Table 5.1: Network architecture of the SER-CNN classifier

Layer	**Type**	**Depth**	**Kernel**	**Output**
1	CONV+ReLu	32	5x5	40x15x32
2	MaxPooling	32	3x3	19x7x32
3	CONV+ReLu	64	5x5	19x7x64
4	MaxPooling	64	3x3	9x3x64
5	CONV+ReLu	128	5x5	9x3x128
6	MaxPooling	128	3x3	4x1x128
7	FC+ReLu	-	-	256
8	Dropout	-	-	256
9	FC+Softmax	-	-	7

The SER-CNN outputs an emotion label $\tilde{e}_j$ for each frame j. At the utterance level, having L frames for the utterance, the recognized emotion sequence is represented with $\{\tilde{e}_j\}_{j=1}^{L}$. Then, the probability of the i-th emotion class, e^i, for

the utterance is computed as

$$p_i = \frac{1}{L}\sum_{j=1}^{L} 1(\tilde{e}_j = e^i) \;\text{ for } i = 1, \ldots, 7, \tag{5.10}$$

where ones function, $1(\cdot)$, returns 1 when condition is true else a zero. Then, the most likely top two emotion classes for the utterance can be identified as

$$\alpha = \arg\max_{1 \le i \le 7} p_i \;\text{ and }\; \beta = \arg\max_{\{1 \le i \le 7\} - \{\alpha\}} p_i. \tag{5.11}$$

We choose to utilize the top second emotion class when the confidence to the top emotion class is weak. For this purpose, probability of the top second emotion is modified to be zero when the probability of the top emotion class is higher than 0.65, as

$$\tilde{p}_\beta = \begin{cases} 0 & \text{if } p_\alpha > 0.65 \\ p_\beta & \text{otherwise.} \end{cases} \tag{5.12}$$

Then the normalized likelihoods for the utterance level top two emotions are set as

$$\bar{p}_\alpha = \frac{p_\alpha}{p_\alpha + \tilde{p}_\beta} \;\text{ and }\; \bar{p}_\beta = \frac{\tilde{p}_\beta}{p_\alpha + \tilde{p}_\beta}, \tag{5.13}$$

corresponding to the top two emotion classes e^α and e^β, respectively.

5.5 Emotion Dependent A2V Mapping

We construct a deep A2V mapping network (DA2V) to generate visual lower face representations from the speech spectral features. Table 5.2 presents the DA2V network structure, which consists of 17 layers with two alternative output dimensions based on the speech spectral representation it receives at the input. Note that convolution, ReLu non-linearity and batch normalization are packed as a repeated layer with different depths and kernel sizes. Visual representation constitutes the target output sequence of the DA2V network. The visual target at frame j is

defined as the concatenation of three visual feature vectors, including the left and right neighboring frames, as $U_j = [u_{j-1}^\intercal, u_j^\intercal, u_{j+1}^\intercal]^\intercal \in \mathbb{R}^{3N\times 1}$. On the other hand, the input to the DA2V network can be set as the speech spectral image $F_j \in \mathbb{R}^{M\times T}$ or as the latent space image $Z_j = [z_{j-\Delta}, \ldots, z_j, \ldots, z_{j+\Delta}] \in \mathbb{R}^{KM\times T}$, where z_j is representing latent space vector as defined in section 5.3, $z_j \in \mathbf{Z}$. In Table 5.2, output dimensions are listed respectively for the speech spectral image F_j and the latent space image Z_j for the first 13 layers.

Table 5.2: Network architecture of the DA2V with convolutional, max pooling, dropout and dense layers and batch normalization (BN)

Layer	**Type**	**Depth**	**Kernel**	**Output**
1	CONV+ReLu+BN	32	5x1	$[M, KM]$xTx32
2	MaxPooling	32	3x1	[19,59]xTx32
3	CONV+ReLu+BN	64	5x1	[19,59]xTx64
4	MaxPooling	64	3x1	[9,29]xTx64
5-9	CONV+ReLu+BN	128	5x1	[9,29]xTx128
10	MaxPooling	128	2x1	[4,14]xTx128
11	CONV+ReLu+BN	128	5x1	[4,14]xTx128
12	CONV+ReLu+BN	128	3x1	[4,14]xTsx128
13	MaxPooling	128	2x1	[2,7]xTx128
14	Dense+ReLu	-	-	1024
15	Dropout	-	-	1024
16	Dense+ReLu	-	-	512
17	Dense	-	-	54

In the proposed emotion dependent A2V mapping framework, the DA2V networks are populated for each emotion class separately. We refer the emotion dependent networks as ED-DA2V to discriminate them from the rest. Hence, in the synthesis phase the ED-DA2V networks output seven candidate visual representations, one for each emotion. Based on the SER-CNN emotion recognition outputs, we use a hard and a soft decision approach to generate the final visual representation. In the hard decision, the visual output $\tilde{U}_j$ is constructed as

$$\tilde{U}_j = \hat{U}_j^\alpha, \tag{5.14}$$

where $\hat{U}_j^\alpha$ is the output of the network for the top matching emotion class e^α. The

hard decision network is denoted as HED-DA2V.

Since the performance of the emotion recognition system is not always robust, the hard decision network is prune to modeling errors. In order to improve robustness, we construct a soft decision output by fusing the network outputs for the top two matching emotion classes as

$$\tilde{U}_j = \bar{p}_\alpha \hat{U}_j^\alpha + \bar{p}_\beta \hat{U}_j^\beta, \tag{5.15}$$

where $\bar{p}_\alpha$ and $\bar{p}_\beta$ are the utterance level normalized likelihoods for the top two matching emotions. The soft decision network is denoted as SED-DA2V.

5.6 Visual Shape Synthesis

The goal of the synthesis is to map estimated visual output sequence $\tilde{U}_j$ from the PCA space to the visual shape space as a sequence of facial shape features $\hat{v}_j$. Let us represent the estimated output at frame j as $\tilde{U}_j = [\tilde{u}_{j-1}^{j\intercal}, \tilde{u}_j^{j\intercal}, \tilde{u}_{j+1}^{j\intercal}]^\intercal$, where $\tilde{u}_{j-1}^{j}$ is representing an estimation of the u_{j-1} at frame j. The estimated visual output $\tilde{U}_j$ in the PCA space can be transformed to the original shape space as $\tilde{V}_j = [\tilde{v}_{j-1}^{j\intercal}, \tilde{v}_j^{j\intercal}, \tilde{v}_{j+1}^{j\intercal}]^\intercal$.

Hence, there are three estimates of the v_j appearing in $\tilde{V}_{j-1}$, $\tilde{V}_j$ and $\tilde{V}_{j+1}$. The final estimate of the visual shape space facial feature at frame j is then defined as the average of the estimates as

$$\hat{v}_j = \frac{1}{3}(\tilde{v}_{j+1}^{j-1} + \tilde{v}_j^j + \tilde{v}_{j-1}^{j+1}). \tag{5.16}$$

Averaging over the estimates maintain smooth transitions of the landmark points at the frame transitions. Finally, the sequence of facial shape features $\{\dots, \hat{v}_{j-1}, \hat{v}_j, \hat{v}_{j+1}, \dots\}$ can be synthesized as facial landmark sequence for visualization.

5.6.1 Affective Audio-Visual Dataset

In this study, we use the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset to evaluate the affective A2V models [Cooke et al., 2006]. The data set consists of video clips from 4 British male actors with six basic emotions (disgust, anger, happy, sad, fear surprise) and the neutral state. Each actor utters 120 sentences, in which 30 are in neutral state and 15 are in each emotion class. Hence there are 480 utterances in the dataset that are using phonetically balanced sentences from the TIMIT corpus [Garofolo et al., 1993]. Speech is sampled at 44.1 kHz with video at a rate of 25 fps. Total duration of the dataset is approximately 30.7 minutes. Face recordings are all frontal view. Utterance level phonetic transcriptions with their respective start time is also provided.

We take the neutral data as the source domain, $\mathbf{D}_S$, and the affective data from each of the six emotions as the target data, $\mathbf{D}_T$. The parallel data generation scheme defined in section 5.2 has been applied to the six emotion classes separately to define the parallel data between each emotion class and neutral state. Total duration of the parallel data is 22.37 minutes for all six emotion classes, where the parallel data duration is around 3.7 minutes per emotion class.

5.6.2 A2V Models in Evaluation

In the experimental evaluations, we consider two baseline models. The first model, which is referred as the no transfer learning (NTL) model, is extracted by training the DA2V network under emotion dependent or independent settings using only the available affective data. The second model is defined based on model adaptation (MA), where the DA2V network is first trained over the neutral data and then adapted with the affective data to the target domain.

The baseline and the proposed models are considered under emotion dependent and independent settings with hard or soft outputs and with different inputs. A short description of these models are given in Table 5.3. Note that both baseline models NTL and MA are investigated under emotion dependent settings as well. Furthermore, in order to isolate classification accuracy of the emotion recognition

system and to define a performance upper bound for the emotion dependent models, we construct the emotion dependent models with hard decision outputs using the true emotion class labels (T-HED-*).

Table 5.3: Short descriptions of the baseline and proposed A2V models

Model	Description
NTL	No Transfer Learning: Single DA2V network is trained on all affective and neutral data. Network input is the speech spectral image F.
MA	Model Adaptation: Single DA2V network is trained on the neutral data, then the model is adapted with the all affective data. Network input is the speech spectral image F.
DA	Domain Adaptation: Single DA2V network is trained on all affective and neutral data. Network input is the latent space image Z.
SED-NTL SED-MA SED-DA	Soft decision emotion dependent models: Respectively same as NTL, MA and DA but seven DA2V models are trained for all seven affective states (including neutral) with soft decision visual outputs.
HED-NTL HED-MA HED-DA	Hard decision emotion dependent: Respectively same as SED-NTL, SED-MA and SED-DA except the visual outputs are hard decisions.
T-HED-NTL T-HED-MA T-HED-DA	Emotion dependent with true emotion labels: Respectively same as HED-NTL, HED-MA and HED-DA except the true emotion labels are used in hard decisions.

5.6.3 Training Deep Models

The mSDA based domain adaptation model has several parameters to tune. In order to set the number of mSDA layers, we extract several performance comparisons as a function of number of mSDA layers as presented in Figure 5.7. We observe linear increases in the time-to-train per epoch and the number of model parameters with

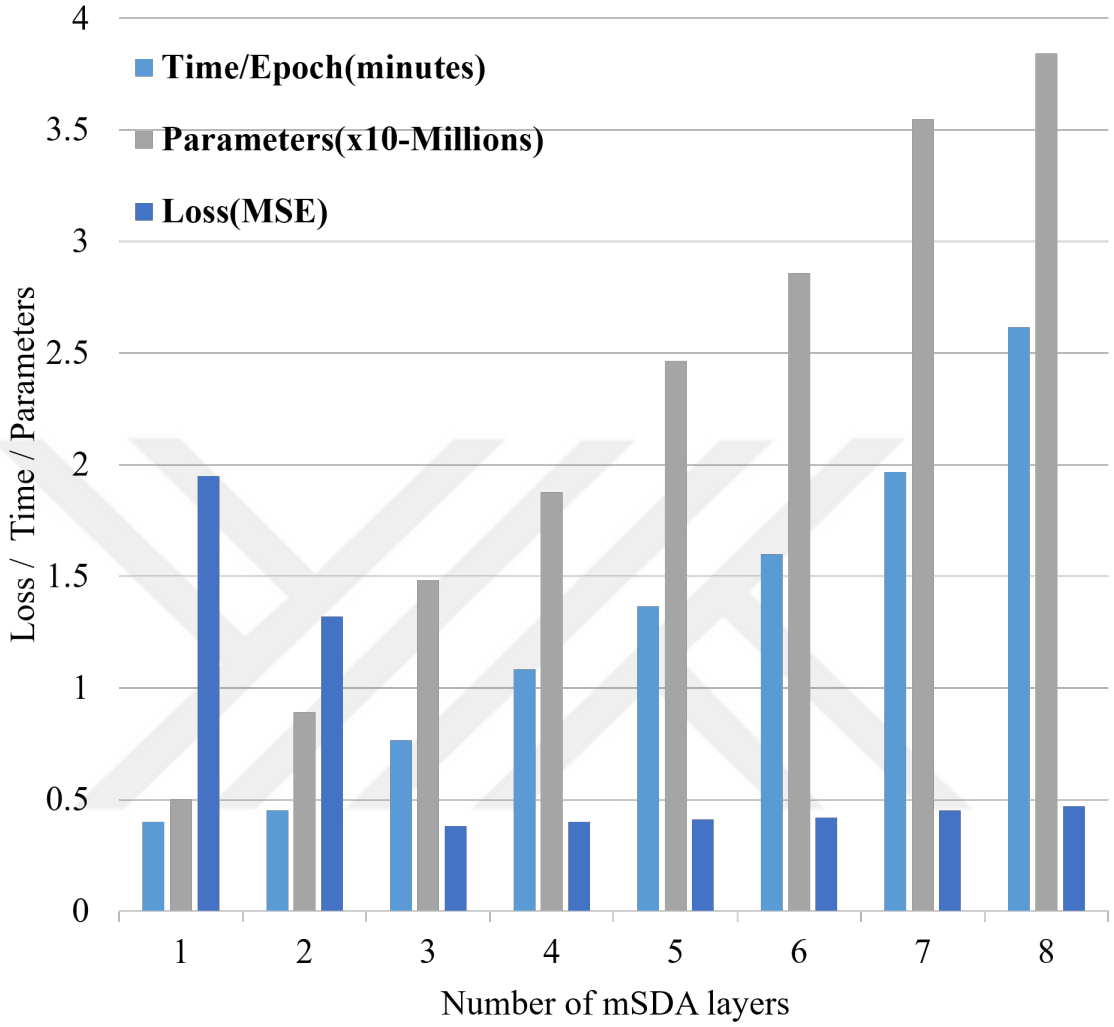


Figure 5.7: Number of mSDA layers vs performance in terms of MSE loss, time-to-train per epoch, and model complexity in terms of total number of parameters

the number of mSDA layers. However, the resulting MSE loss saturates after 3 layers of mSDA. Hence, we set the number of mSDA layers as $K = 3$. Furthermore, we use $R = 4$ repeated versions of input and the regularization term is selected as $\lambda = 0.00001$.

We train the proposed all DA2V models from scratch, where the weights are initialized randomly using a normal distribution with mean zero and standard deviation 0.05. We use Adam optimizer [Kingma and Ba, 2014] with initial learning

rate of 0.001, and further reduce it with a factor of 10 if there is no improvement in the loss value for five consecutive epochs. Early stopping is employed to avoid any over fitting if there is no improvement in loss for 20 consecutive epochs. The SER model is trained over 200 epochs using the categorical cross-entropy loss function, while DA2V models are trained for 150 epochs with mean square error (MSE) loss between the original and network generated PCA components. In both networks, drop-out layer is introduced only after the dense layer with the probability of 0.5. Batch size of 64 is selected for both models. The temporal window size for acoustic features is set as $T = 15$. Both of the deep models, deep emotion recognition and facial shape regression models, are trained using the Keras[1] with Tensorflow [Abadi and et al, 2016] backend on a NVIDIA TITAN XP GPU.

5.7 Emotion Dependent A2V Experimental Evaluations

5.7.1 Objective Evaluations

We deploy 5-fold cross-validation (CV) train/test scheme. Data samples are partitioned into five equal parts for 5-fold CV, where, in each fold, four parts are used for training and one part for testing. This way each sample in data set is used for training and also been tested at least once.

The models resulting from training are used for synthesizing and animating facial shapes over the test utterances.

5.7.2 SER results

In section 5.4, we define the top and top second recognized emotion classes at the utterance level as e^{α} and e^{β}, respectively. Speech emotion recognition performance is defined using two different recognition rate metrics, which are based on only the top emotion class and based on the top two emotion classes. Emotion recognition

[1] `https://keras.io/`

rate for the top emotion class is defined as

$$ERR_t = \frac{1}{N_U} \sum_j 1(e_j^\alpha = e_j) \tag{5.17}$$

where index j runs over all test utterances, N_U is the total number of test utterances and the ones function is used as defined in (5.10). Similarly, emotion recognition rate for the top two emotion classes can be defined as

$$ERR_{t2} = \frac{1}{N_U} \sum_j \left[1(e_j^\alpha = e_j) + 1(e_j^\beta = e_j)\right]. \tag{5.18}$$

Utterance level emotion recognition rates are reported for each emotion category in Table 5.4. Emotion recognition performances are significantly higher for the ERR_{t2} metric that indicates the strong confidence of the top two recognized emotion classes for the emotion recognition task. Hence, a more robust emotion dependent A2V mapping using the soft decision output, which depends on the top two recognized emotion classes, as defined in section 5.5 can be expected.

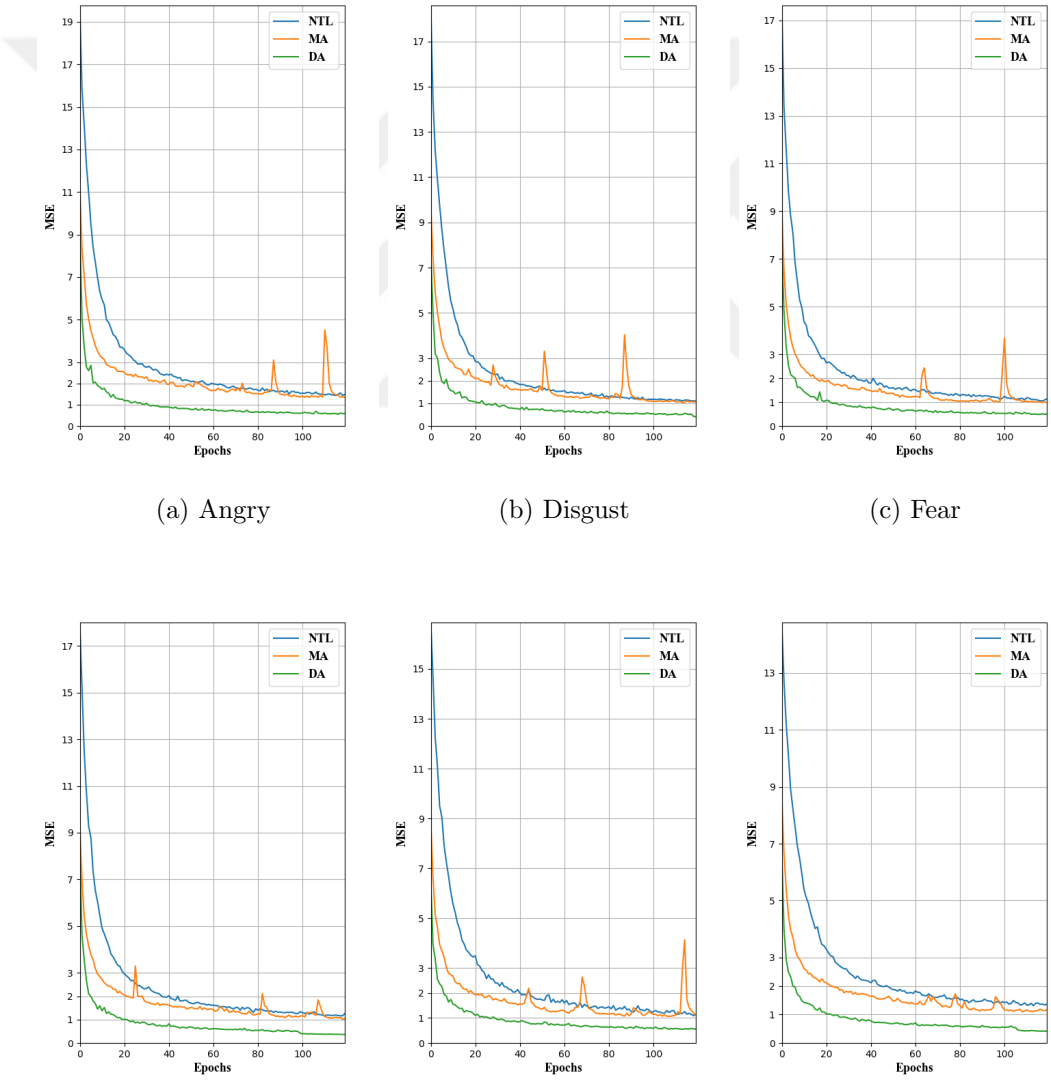
Table 5.4: Utterance level emotion recognition rates (%) for each emotion category in terms of the top and top two recognized emotion classes.

	Angr	Disg	Fear	Happ	Neut	Sad	Surp
ERR_t	72.15	75.38	73.21	71.71	91.93	75.19	67.98
ERR_{t2}	95.86	89.29	89.74	88.26	97.45	90.43	91.96

5.7.3 DA2V regression results

We use the MSE loss between the ground truth and the estimated PCA features to evaluate the training performance of the NTL, MA and DA models. Figure 5.9 presents the MSE loss curve for six emotion categories trained separately using emotion dependent approach and a single model using all emotional data combined. Figure 5.8 presents the MSE loss curves for the six emotion categories during the training. The proposed DA model achieves significantly lower MSE loss compared

to the NTL and MA models. The decreasing trend in the MSE loss curves indicates that all models have a successful learning phase. However, we observe no significant improvement with the model adaptation (MA) over the NTL model. On the other hand, the proposed domain adaptation enables a further MSE loss reduction and better modeling.



(a) Angry (b) Disgust (c) Fear

(d) Happy (e) Sad (f) Surprise

Figure 5.8: MSE loss curves of the DA2V networks during the learning process for the six emotion categories

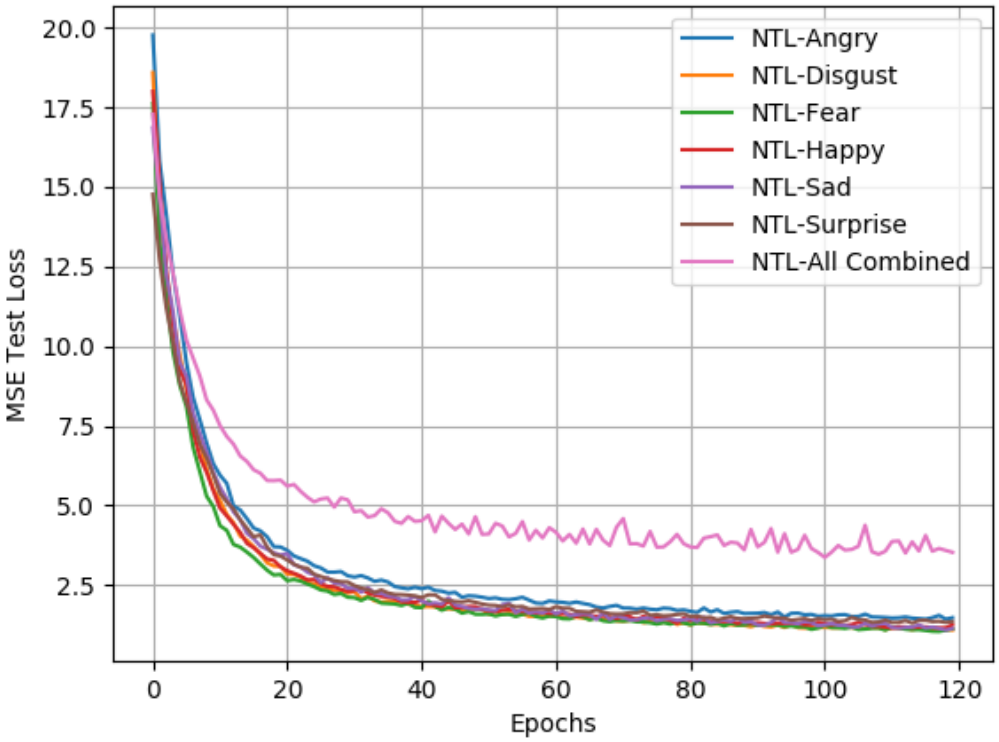


Figure 5.9: The MSE loss over the validation data along the epochs for the emotion dependent (separate for each emotion-No Transfer Learning) and independent (NTL) models.

Table 5.5 provides the MSE loss values over the test set of all the models described in Table 5.3. We clearly observe that the proposed soft decision emotion dependent model, SED-DA, outperforms all the other models by consistently achieving the lowest MSE loss. Although the deep learning approach for generalized speech animation in [Taylor and et all, 2017] performs good under neutral speech, it fails with emotional speech. We also provide the MSE loss values of the emotion dependent models with the true emotion labels. These MSE loss values are presented in the last three rows of the Table 5.5 and they exhibit upper bound performances for the emotion dependent models. Performance of the domain adaptation based model, T-HED-DA, clearly surpasses the other two that indicates strength of the domain adaptation.

Table 5.5: MSE loss values at the test for the proposed and baseline models together with two other models from the recent literature

Method	**MSE**
Taylor et al. [Taylor and et all, 2017]	12.90
Asadiabadi et al. [Asadiabadi et al., 2018]	6.75
NTL	6.73
MA	6.80
DA	**3.72**
HED-NTL	9.54
HED-MA	9.22
HED-DA	**5.32**
SED-NTL	5.57
SED-MA	5.61
SED-DA	**2.83**
T-HED-NTL	3.19
T-HED-MA	3.22
T-HED-DA	**1.60**

5.7.4 Subjective Evaluations

Unlike the objective evaluations, subjective tests can evaluate realism and naturalness of the lower-face animations by capturing human perception. We use a mean opinion score (MOS) test to subjectively evaluate animations of the three soft decision emotion dependent synthesis methods, SED-NTL, SED-MA and SED-DA, and the ground truth (GT), which is extracted with the DLib to capture the original lower-face contour. The test is run with 20 participants who are between 22 and 35 years of age. we conducted two set of experiments. In the first part of experiment, participants were shown video clips from emotion dependent and independent models. The proposed emotion dependent facial shape animation scheme is preferred over the baseline emotion independent scheme. Furthermore, in each emotional category a similar preference tendency, except for happy and sad categories, is observed as presented in Figure 5.10.

In the second part of experiment, each test session for a participant contains

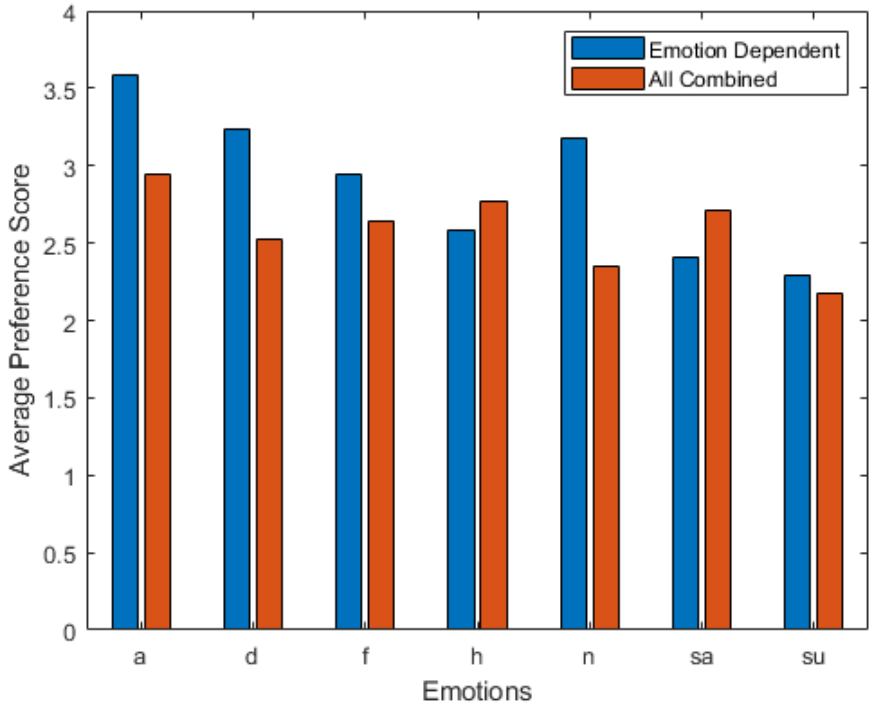


Figure 5.10: Emotion category based average preference scores for the emotion dependent and independent (all combined) model animation evaluations.

28 clips for all 4 methods (GT,SED-NTL, SED-MA and SED-DA) , where each method is tested with 7 clips. During the test, all clips are shown in random order. In the test, each participant is asked to evaluate naturalness, emotional content and synchronization of animations with the speech using a five-point assessment scale (1: Poor, 2: Fair, 3: Good, 4: Very Good, 5: Excellent). Participant had the option to replay a video multiple times before submitting the score. Samples of animation clips from the subjective tests are available for online demonstration[2].

Table 5.6 presents the mean opinion scores for each lower-face animation method SED-NTL, SED-MA, SED-DA and ground truth (GT). As expected, the ground truth (GT) animation is preferred with the highest MOS score. The proposed soft decision emotion dependent domain adaptation method, SED-DA, is preferred with

[2]Sample clips for the speech driven affective facial feature synthesis are available at http://mvgl.ku.edu.tr/demos/

Table 5.6: Mean opinion score (MOS) subjective test results of the lower-face animations of the SED-NTL, SED-MA, SED-DA and ground truth (GT) with the score scale 1: Poor, 2: Fair, 3: Good, 4: Very Good, 5: Excellent

Method	**Mean**	**Std**
SED-NTL	2.64	1.14
SED-MA	2.40	0.96
SED-DA	2.91	1.13
GT	3.20	1.13

the second high MOS score of 2.91 after the GT.

A pairwise significance test is also performed for all methods in the MOS evaluations and presented in Table 5.7. We can observe that the p-value between the GT and the proposed SED-DA method is greater than 0.05, which indicates that quality of landmarks generated with the proposed method are not significantly different from the ground truth animations. Furthermore, the proposed SED-DA and as well the GT animations are preferred significantly different from the SED-NTL and SED-MA animations.

Table 5.7: Significance test of the MOS subjective evaluations with p-values for the ANOVA (* The mean difference is significant at p=0.01 level)

Method	SED-MA	SED-DA	GT
SED-NTL	0.08	0.01*	0.01*
SED-MA		0.01*	0.01*
SED-DA			0.09

The MOS subjective test results for each emotion category are presented in Figure 5.11. We can observe that the animations generated with the proposed SED-DA method are preferred over the SED-NTL and SED-MA methods. Only exception is the neutral category, where the SED-NTL is preferred higher than the proposed SED-DA. This is expected since neutral data is more abundant and delivers a fair enough learning for the no transfer learning approach.

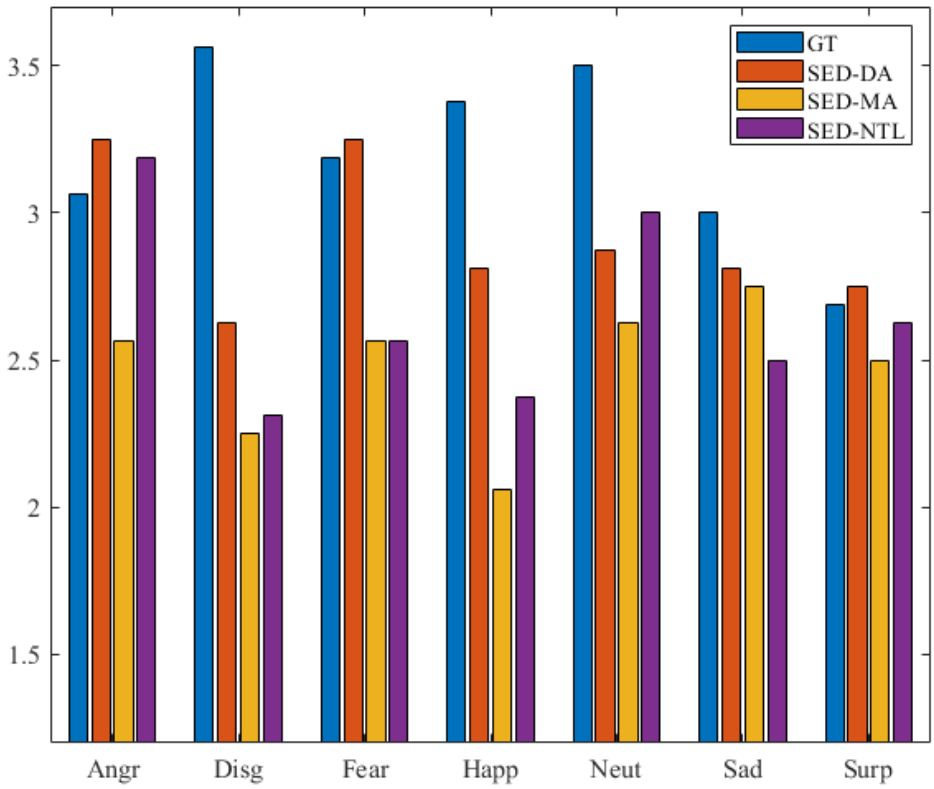


Figure 5.11: Mean opinion score (MOS) subjective test results of the lower-face animations of the SED-NTL, SED-MA, SED-DA and ground truth (GT) for each emotion category

5.8 Summary

In this chapter we explained the significance of emotion dependent speech driven facial animation system. Details of audio visual dataset and feature representation for emotion dependent A2V models is also provided in this chapter. In depth details of using the domain adaptation and parallel data generation process are provided. This chapter also covers the SER process. Network architectures for SER and DA2V models are also explained. The results for objective evaluations to compare the performance of all the models mentioned earlier in the chapter in table 5.3 are explained in this chapter. Followed by objective evaluations, subjective evaluations

to assess the quality of synthesized facial features are also reported. The results reported proves the superiority of proposed methodology over the base line methods in both objective and subjective evaluations.

Chapter 6

CONCLUSION AND FUTURE WORK

In this study, at first we investigate the role of phonemes for affect recognition under lip articulations. We observe that using only lip features can attain affect recognition performance higher than the random. Our results also suggest that a selected list of discriminative phoneme articulations can better classify all affect attributes using only the lip features. These results are encouraging for the use of lip articulations in multimodal affect recognition systems.

We also deliver an analysis of discriminative phonetic classes for the lip driven affect recognition as a function of manner and place of articulation. Extensive studies in this direction have potential to improve contribution of lip modality to the affect recognition systems.

In order to develop speech driven facial animations, we introduced a novel approach for generating face animations synchronized with the input speech. A statistical shape model was trained to project the shape data into an uncorrelated lower dimensional parameter space, capturing 98% of variation in the training data. We trained a multimodal deep neural network with acoustic and phoneme features as inputs, to estimate the facial shape parameters. The proposed approach was shown to perform better on neutral and emotional datasets, in terms of MSE loss in the shape and parameter space. We observed that the benefit of the proposed multimodal approach over the baseline text and speech-based methods was more remarkable for the affective dataset. As a future work we will investigate speech animation under different emotional states and will use the generated shape animations for a user study performance evaluation.

Later we expand our work for emotional speech and introduce a framework for emotional speech driven facial shape generation. Our system has several advantages

compared to previous studies. Using the domain adaptation concept, we achieve improved performance even by using a relatively small affective data set. Our model is constructed specifically for emotional speech and can handle basic emotions embedded in speech especially for the angry or surprised categories. Soft decision emotion dependent mapping provides us an extra boost in the performance even with rather decent SER system. We should note that the speech animation performance can be further improved by improving the SER performance towards the T-HED-DA animations, which are using the true emotion labels. Furthermore, combining the proposed approach with generative models might enable us to better generate the speech animations with full face features. Although we achieve improved results in a rather small data set, a larger and richer affective data set would also be used to attain speaker and gender independence with improved animation performance.

Several future work directions can be pointed to extend the scope of this work. Firstly, combining the proposed approach with generative models might enable us to better generate the speech animations with full face features. Secondly, a larger and richer affective data set would also be used to attain speaker and gender independence with improved animation performance. Finally, investigating cross-dataset adaptation would be interesting to deliver more robust affective facial synthesis.

BIBLIOGRAPHY

[Abadi and et al, 2016] Abadi, M. and et al (2016). Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA. USENIX Association.

[Abdel-Hamid et al., 2014] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545.

[Asadiabadi et al., 2018] Asadiabadi, S., Sadiq, R., and Erzin, E. (2018). Multimodal speech driven facial shape animation using deep neural networks. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1508–1512. IEEE.

[Badshah et al., 2017] Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE.

[Beeler et al., 2010] Beeler, T., Bickel, B., Beardsley, P., Sumner, B., and Gross, M. (2010). High-quality single-shot capture of facial geometry. In *ACM Transactions on Graphics (ToG)*, volume 29, page 40. ACM.

[Bertero and Fung, 2017] Bertero, D. and Fung, P. (2017). A first look into a convolutional neural network for speech emotion detection. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5115–5119. IEEE.

[Blitzer et al., 2006] Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Bonilla et al., 2008] Bonilla, E. V., Chai, K. M., and Williams, C. (2008). Multi-task gaussian process prediction. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 153–160. Curran Associates, Inc.

[Bozkurt et al., 2013] Bozkurt, E., Asta, S., Özkul, S., Yemez, Y., and Erzin, E. (2013). Multimodal analysis of speech prosody and upper body gestures using hidden semi-markov models. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3652–3656. IEEE.

[Bozkurt et al., 2007] Bozkurt, E., Erdem, C. E., Erzin, E., Erdem, T., and Ozkan, M. (2007). Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. In *2007 3DTV Conference*, pages 1–4. IEEE.

[Brand, 1999] Brand, M. (1999). Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28. ACM Press/Addison-Wesley Publishing Co.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

[Busso et al., 2007] Busso, C., Lee, S., and Narayanan, S. S. (2007). Using neutral speech models for emotional speech analysis. In *Interspeech*, pages 2225–2228.

[Busso and Narayanan, 2007] Busso, C. and Narayanan, S. S. (2007). Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2331–2347.

[Cao et al., 2015] Cao, C., Bradley, D., Zhou, K., and Beeler, T. (2015). Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)*, 34(4):46.

[Capin et al., 1997] Capin, T. K., Noser, H., Thalmann, D., Sunday Pandzic, I., and Thalmann, N. M. (1997). Virtual human representation and communication in vlnet. *IEEE Computer Graphics and Applications*, 17(2):42–53.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

[Chen et al., 2012] Chen, M., Xu, Z., Weinberger, K., and Sha, F. (2012). Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*.

[Chung et al., 2017] Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE.

[Cooke et al., 2006] Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

[Cootes and Taylor, 1993] Cootes, T. and Taylor, C. (1993). Active shape model search using local grey-level methods: a quantitative approach. *Proc. British Machine Vision Conference*, pages 639–648.

[Cowie et al., 2001] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

[Deng et al., 2014] Deng, J., Zhang, Z., Eyben, F., and Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072.

[El Ayadi et al., 2011] El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

[Fan et al., 2015] Fan, B., Wang, L., Soong, F. K., and Xie, L. (2015). Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE.

[Gaggioli et al., 2003] Gaggioli, A., Mantovani, F., Castelnuovo, G., Wiederhold, B., and Riva, G. (2003). Avatars in clinical psychology: a framework for the clinical use of virtual humans. *Cyberpsychology & behavior*, 6(2):117–125.

[Garofolo et al., 1993] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., and Pallett, D. (1993). Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93.

[Glorot et al., 2011] Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520.

[Goodall, 1991] Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society B*, 53(2).

[Gutierrez-Osuna and et al, 2005] Gutierrez-Osuna, R. and et al (2005). Speech-driven facial animation with realistic dynamics. *IEEE transactions on multimedia*, 7(1):33–42.

[Haq and Jackson, 2011] Haq, S. and Jackson, P. (2011). Multimodal emotion recognition. In *Machine audition: principles, algorithms and systems*, pages 398–423. IGI Global.

[Huang et al., 2014] Huang, Z., Dong, M., Mao, Q., and Zhan, Y. (2014). Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804. ACM.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

[Karras et al., 2017] Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94.

[King, 2009] King, D. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758.

[Kingma and Ba, 2014] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Konstantinidis et al., 2009] Konstantinidis, E. I., Hitoglou-Antoniadou, M., Luneski, A., Bamidis, P. D., and Nikolaidou, M. M. (2009). Using affective avatars and rich multimedia content for education of children with autism. In *Proceedings of the 2nd international conference on pervasive technologies related to assistive environments*, pages 1–6.

[Lee et al., 2011] Lee, C.-C., Mower, E., Busso, C., Lee, S., and Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171.

[Livingstone et al., 2012] Livingstone, S. R., Peck, K., and Russo, F. A. (2012). Ravdess: The ryerson audio-visual database of emotional speech and song. In *Annual meeting of the canadian society for brain, behaviour and cognitive science*, pages 205–211.

[Luo et al., 2014] Luo, C., Yu, J., Li, X., and Wang, Z. (2014). Realtime speech-driven facial animation using gaussian mixture models. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE.

[Mao et al., 2014] Mao, Q., Dong, M., Huang, Z., and Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia*, 16(8):2203–2213.

[Mariooryad and Busso, 2012] Mariooryad, S. and Busso, C. (2012). Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2329–2340.

[McAuliffe et al., 2017] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: trainable text-speech alignment using kaldi. In *Proceedings of interspeech*.

[Metallinou et al., 2013] Metallinou, A., Katsamanis, A., and Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152.

[Metallinou et al., 2012] Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., and Narayanan, S. (2012). Context-sensitive learning for enhanced

audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198.

[Nicolaou et al., 2011] Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105.

[Pan et al., 2010] Pan, S. J., Yang, Q., et al. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

[Pan et al., 2012] Pan, Y., Shen, P., and Shen, L. (2012). Speech emotion recognition using support vector machine. *International Journal of Smart Home*, 6(2):101–108.

[Pham et al., 2017] Pham, H. X., Cheung, S., and Pavlovic, V. (2017). Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *CVPR Workshops*, pages 2328–2336.

[Qian et al., 2016] Qian, Y., Bi, M., Tan, T., and Yu, K. (2016). Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2263–2276.

[Sadiq et al., 2019] Sadiq, R., AsadiAbadi, S., and Erzin, E. (2019). Emotion dependent facial animation from affective speech. *arXiv preprint arXiv:1908.03904*.

[Sadiq and Erzin, 2017] Sadiq, R. and Erzin, E. (2017). Affect recognition from lip articulations. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2432–2436. IEEE.

[Sadiq and Erzin, 2020] Sadiq, R. and Erzin, E. (2020). Emotion dependent domain adaptation for speech driven affective facial feature synthesis). *IEEE transactions on affective computing*.

[Salvador and Chan, 2007] Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580.

[Schuller et al., 2011] Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087.

[Schuller et al., 2003] Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–1. IEEE.

[Sehgal and Kehtarnavaz, 2018] Sehgal, A. and Kehtarnavaz, N. (2018). A convolutional neural network smartphone app for real-time voice activity detection. *IEEE Access*, 6:9017–9026.

[Song and Cai, 2015] Song, W. and Cai, J. (2015). End-to-end deep neural network for automatic speech recognition. *arXiv preprint, arxiv:1701.02720.*

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A.and Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

[Suwajanakorn et al., 2017] Suwajanakorn, S., Seitz, S., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95.

[Taylor and et all, 2017] Taylor, S. and et all (2017). A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):93.

[Taylor et al., 2016] Taylor, S., Kato, A., Matthews, I. A., and Milner, B. P. (2016). Audio-to-visual speech conversion using deep neural networks. In *INTERSPEECH*. International Speech Communication Association.

[Toda et al., 2008] Toda, T., Black, A. W., and Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech Communication*, 50(3):215–227.

[Tóth, 2015] Tóth, L. (2015). Phone recognition with hierarchical convolutional deep maxout networks. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–13.

[Trigeorgis et al., 2016] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE.

[Verma et al., 2003] Verma, A., Rajput, N., and Subramaniam, L. V. (2003). Using viseme based acoustic models for speech driven lip synthesis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5. IEEE.

[Vinola and Vimaladevi, 2015] Vinola, C. and Vimaladevi, K. (2015). A survey on human emotion recognition approaches, databases and applications. *ELCVIA: electronic letters on computer vision and image analysis*, pages 00024–44.

[Vougioukas et al., 2018] Vougioukas, K., Petridis, S., and Pantic, M. (2018). End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*.

[Weiss et al., 2016] Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1):9.

[Xie and Liu, 2007] Xie, L. and Liu, Z.-Q. (2007). A coupled hmm approach to video-realistic speech animation. *Pattern Recognition*, 40(8):2325–2340.

[Yamamoto et al., 1997] Yamamoto, E., Nakamura, S., and Shikano, K. (1997). Speech to lip movement synthesis by hmm. In *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*.

[Yehia et al., 1998] Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43.

[Yoon et al., 2018] Yoon, S., Byun, S., and Jung, K. (2018). Multimodal speech emotion recognition using audio and text. *arXiv preprint arXiv:1810.04635*.

[Young and et al, 2006] Young, S. and et al (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.

[Zhou et al., 2014] Zhou, J. T., Pan, S. J., Tsang, I. W., and Yan, Y. (2014). Hybrid heterogeneous transfer learning through deep learning. In *Twenty-eighth AAAI conference on artificial intelligence*.