Multimodal Speech Driven Facial Shape Animation Using Deep Neural Networks

Sasan Asadiabadi, Rizwan Sadiq and Engin Erzin
Multimedia, Vision and Graphics Laboratory,
College of Engineering, Koç University, Istanbul, Turkey
E-mail: [sabadi15, rsadiq13, eerzin]@ku.edu.tr

***Abstract*—In this paper we present a deep learning multimodal approach for speech driven generation of face animations. Training a speaker independent model, capable of generating different emotions of the speaker, is crucial for realistic animations. Unlike the previous approaches which either use acoustic features or phoneme label features to estimate the facial movements, we utilize both modalities to generate natural looking speaker independent lip animations synchronized with affective speech. A phoneme-based model qualifies generation of speaker independent animation, whereas an acoustic feature-based model enables capturing affective variation during the animation generation. We show that our multimodal approach not only performs significantly better on affective data, but improves performance over neutral data as well. We evaluate the proposed multimodal speech-driven animation model using two large scale datasets, GRID and SAVEE, by reporting the mean squared error (MSE) over various network structures.**

***Index Terms*—Deep Learning, Speech Driven Animations, Deep Neural Network (DNN), Active Shape Models (ASM)**

I. Introduction

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. To understand natural speech in noisy environments or non grammatical expressions and to give healthy response, facial animation can be more beneficial in human computer interaction, [1] [2]. Computer simulation of human faces, which can accurately reflect facial movements, has been a thriving field of research for decades, resulting in a large number of facial models and animation systems [3]. Researchers in the field of computer graphics, machine learning, psychology as well as medicine are relentlessly working towards generating more realistic animated avatars.

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 actors face [4] [5] [6]. 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. The problem of mapping a speech signal to the facial animations can be investigated on several different levels as: acoustic signal level, phoneme level and word level [7]. At signal level, raw speech or its low/high level features are used to generate automated facial animations. Whereas phoneme and word level are part of text driven animations. At the phoneme level, speech is first segmented into a sequence of phonemes. Mapping is then found for each phoneme in the speech signal using a viseme table, which contains one visual feature set for each phoneme. The standard set of visemes is specified in MPEG-4 and contains 15 static visemes that can be easily distinguished [8].

Early research on acoustic to visual mappings was based on different methods including Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) and Dynamic Bayesian Networks (DBN's). In an early work of [9], authors predicted the 3-D facial movements from LPC and RASTA-PLP acoustic features. Their research was based on entropy-minimization algorithm that learned both the structure and the parameters of an HMM to perform frame to frame mapping. Later, [1] proposed to consider context by tagging video frames to audio frames from past and future. Most of this mapping from speech to visual features is performed off line by training a model and then utilizing that model to predict facial movements. A recent work presented a method to capture not only facial movements in real-time at high fidelity, but also some fine details such as wrinkles [5]. Another recent study on lip-sync produced a high-quality video of USA former president Obama, speaking with accurate lip-sync [10]. A recurrent neural network was trained on many hours of his weekly address footage, that learns the mapping from raw audio features to mouth shapes. Although the method claimed to generate lip shapes correctly, the limitation is generalization as it was trained and tested on a single speaker. This problem was addressed in [6], where they focused on generating facial animations solely from phoneme sequence as input. The idea was based on using a DNN with sliding window predictor that learns arbitrary nonlinear mappings from phoneme label input sequences to mouth movements. Contrary to conventional method of mapping phoneme sequence to fixed number of visemes, their model successfully mapped the text based input of phonetic sequence to output video representation of continuous speech. Although their system provided some promising results with high quality speech driven animations and it was claimed to be generalizable for any speaker, despite being trained on a single reference speaker, they did not address possible variabilities

due to affective speech animation.

In this study, we proposed a multimodal approach to enhance the previously developed methods for high fidelity production of data driven facial animations. We propose a system that 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. In this study, we applied the proposed method on different neutral and affective datasets. 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. Remaining of the paper is organized as follows. In section II we describe the proposed multimodal speech-driven facial shape animation system. We give experimental evaluations in Section III. Finally, the article is concluded in Section IV.

II. Methodology

This section covers the details of the datasets we utilized for this study; 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.

A. Datasets

GRID [11], 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 [12] 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 [13] 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.

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

Fig. 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.

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 [14], 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 [6], 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 $\{f_j^p\}_{j=1}^N$, where $f_j^p \in \mathbb{R}^{41\times 1}$ and N is the number of instances in the training set.

2) Speech Features: Raw speech signal contains confined spectral information that can be valued for speech to facial mapping. Inspired by the work of [15] we used Mel-Frequency Spectral Coefficients, also denoted as MFSC features. For each speech frame, 40 MFSC features were extracted using HTK toolkit [16] 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 $\{f_j^a\}_{j=1}^N$, where $f_j^a \in \mathbb{R}^{40\times r}$ and r is the audio to video frame ratio.

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:

$$S = [x_1, y_1, x_2, y_2, ..., x_M, y_M]^T. \quad (1)$$

The Dlib facial landmark detector [17] was used to detect a total of $M = 36$ landmark points on the lower face region for each video frame in both datasets. Figure 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 [18]. In the ASM, a statistical model is trained for a set of shapes $\{S_1, S_2, ..., S_N\}$, 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 [19]. 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:

$$\bar{S} = \frac{1}{N}\sum_{i=1}^{N} S_i \quad (2)$$

$$C = \frac{1}{N-1}\sum_{i=1}^{N}(S_i - \bar{S})(S_i - \bar{S})^T \quad (3)$$

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:

$$S = \bar{S} + Pf_v, \quad (4)$$

where $P = [P_1 P_2 ... P_k]$ is the truncated eigenvectors of the covariance matrix(C), corresponding to the k largest eigenvalues, and f_v is the k-dimensional model parameters. We chose $k = 18$ which allows the trained shape model to capture 98% of the variation in the training set. Figure 2 shows three main modes of variation around the mean, varying the model parameters between ± 3 of their standard deviation.

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:

$$f_v = P^T(S - \bar{S}). \quad (5)$$

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

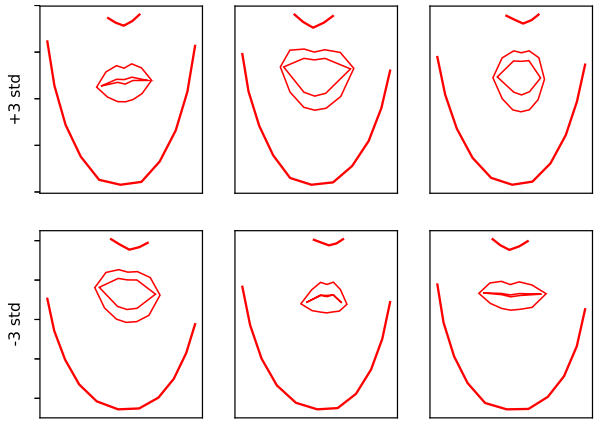

Fig. 2. Three main modes of variation in the shape model for the SAVEE dataset. Parameters range is restricted between ± 3 of their std to allow generation of plausible shapes.

C. 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 [6]. 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:

$$F_j^m = \{(f_{j-k_m}^m, ..., f_j^m, ..., f_{j+k_m}^m)\}_{j=1}^N \quad (6)$$

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.

D. 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 [6] is used as a baseline in this study. Despite having some benefits such as making

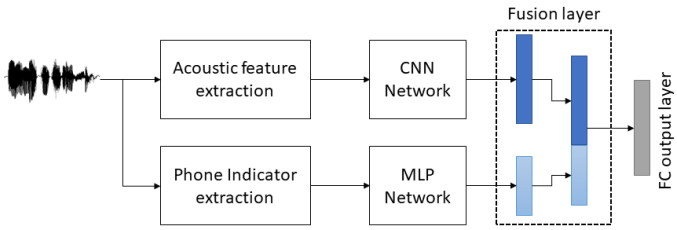

Fig. 3. Proposed multimodal-based network. Acoustic and Phoneme features are extracted from speech and fed to the network separately, merged at an intermediate level and connected to the fully connected output layer.

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.

1) Text-based Architecture: For text-based experiments, a deep Multilayer Perceptron (MLP) similar to what is described in [6] 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 [20] with 50% probability was used. Mini batch size was selected as 128 along with Adam optimizer [21] 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.

2) Speech-based Architecture: Convolutional Neural Networks (CNN) are shown to be efficient in extracting discriminative features from speech [15]. 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.

3) 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

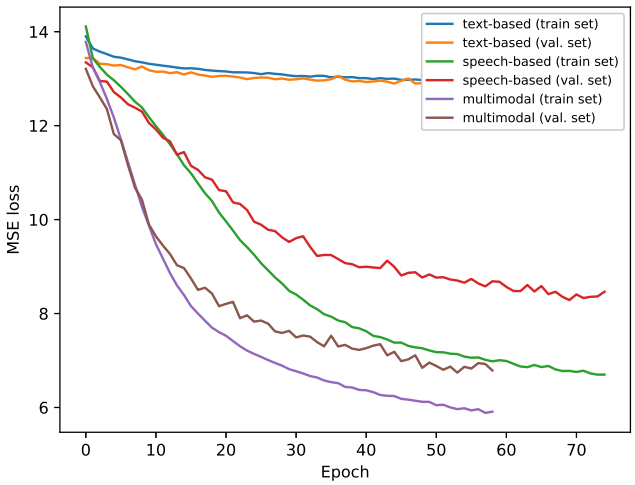

Fig. 4. Train and validation loss curve for three different network architectures on SAVEE dataset.

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 3. Similar deep MLP and CNN structures were employed, as described in sections II-D1 and II-D2, for text and speech inputs, respectively.

All models were trained using Keras[1] with Tensorflow [22] backend on a NVIDIA TITAN XP GPU.

III. Experimental Evaluations

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 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). 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.

[1]https://keras.io/

TABLE I
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 [6]	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**

Figure 4 shows the MSE loss curve through the learning process, for three 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 I 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.

IV. CONCLUSIONS

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.

ACKNOWLEDGMENT

We thank to NVIDIA for donating Titan XP within the GPU Grant Program.

REFERENCES

[1] P. Kakumanu and et al. Speech driven facial animation. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5. ACM, 2001.

[2] D.W. Massaro. *Perceiving talking faces: From speech perception to a behavioral principle*, volume 1. Mit Press, 1998.

[3] J. Ostermann. Animation of synthetic faces in mpeg-4. In *Computer Animation 98. Proceedings*, pages 49–55. IEEE, 1998.

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

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

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

[7] G. Zorić. *Automatic lip synchronization by speech signal analysis*. PhD thesis, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, 2005.

[8] I.S. Pandzic and R. Forchheimer. *MPEG-4 facial animation: the standard, implementation and applications*. John Wiley & Sons, 2003.

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

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

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

[12] S. Haq and Ph.J.B. Jackson. Multimodal emotion recognition. In *Machine audition: principles, algorithms and systems*, pages 398–423. IGI Global, 2011.

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

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

[15] O. Abdel-Hamid and et al. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.

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

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

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

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

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

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

[22] M. Abadi and et al. 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, 2016. USENIX Association.