Noise-Resilient Training Method for Face Landmark Generation From Speech

Sefik Emre Eskimez, *Member, IEEE*, Ross K. Maddox, Chenliang Xu, *Member, IEEE*, and Zhiyao Duan, *Member, IEEE*

***Abstract*—Visual cues such as lip movements, when available, play an important role in speech communication. They are especially helpful for the hearing impaired population or in noisy environments. When not available, having a system to automatically generate talking faces in sync with input speech would enhance speech communication and enable many novel applications. In this article, we present a new system that can generate 3D talking face landmarks from speech in an online fashion. We employ a neural network that accepts the raw waveform as an input. The network contains convolutional layers with 1D kernels and outputs the active shape model (ASM) coefficients of face landmarks. To promote smoother transitions between video frames, we present a variant of the model that has the same architecture but also accepts the previous frame's ASM coefficients as an additional input. To cope with background noise, we propose a new training method to incorporate speech enhancement ideas at the feature level. Objective evaluations on landmark prediction show that the proposed system yields statistically significantly smaller errors than two state-of-the-art baseline methods on both a single-speaker dataset and a multi-speaker dataset. Experiments on noisy speech input with five types of non-stationary unseen noise show statistically significant improvements of the system performance thanks to the noise-resilient training method. Finally, subjective evaluations show that the generated talking faces have a significantly more convincing match with the input audio, achieving a similarly convincing level of realism as the ground-truth landmarks.**

***Index Terms*—Speech animation, noise-resilient, convolutional neural networks, speech processing, face landmarks.**

I. Introduction

SPEECH communication between humans is often not merely via the acoustic channel; visual cues can also play an important and even critical role. Extensive studies have shown that seeing lip movements besides hearing speech can significantly improve speech comprehension for both the general and hearing impaired population [1]–[4], especially when background noise or compression effects corrupt the acoustic signal.

Therefore, having ways to generate talking faces from acoustic speech signals would significantly improve speech communication and comprehension in many scenarios and enable many applications. It improves access to abundantly available speech content on the web for the hearing impaired population. It is also useful in AR/VR professional training applications for pilots, drivers, machine operators, doctors, police officers, and soldiers, where the training scenarios are often noisy, and audio-only speech comprehension can be challenging. It is also useful for developing visual dubbing applications for movies.

To this end, researchers proposed end-to-end and module-based systems. End-to-end data-driven methods can learn the mapping between speech and visual cues; as a result, they can generate natural looking talking faces [5]–[7]. However, utilizing separate modules to generate the key parameters (articulation, mouth shapes) and fine details (texture, identity) has benefits. The key parameters, such as face landmarks, are driven by the speech content directly, and they play the skeleton role in such systems. Another module can further process the generated face landmarks to impose photo-realistic textures and details of the face. This modular design provides more flexibility than end-to-end generation systems. For example, the face landmarks can be manipulated before being processed by the texture module to change the facial expression, emotion and the fine articulation of words.

Speech signals encountered in the wild often contain background noise that degrades the performance of automatic speech processing systems. It is vital that the talking face generation system is resilient to such background noise in practice. To our best knowledge, however, most of the existing systems do not consider background noise in their system design and evaluation.

In this paper, we improve on our previous work [8] and present a new method that generates 3D face landmarks directly from the raw waveform. We propose a novel pre-processing method to normalize the identities of the face landmarks. In addition, we propose a neural network that processes the waveform with convolutional layers with 1D filters and predicts the active shape model (ASM) parameters of 3D face landmarks with a following fully connected (FC) layer. We train the network with pairs of speech audio and 3D face landmarks extracted from the GRID dataset [9]. To cope with background noise in speech input, we further propose a noise-resilient training method that uses speech enhancement ideas in feature learning. Objective

Manuscript received January 25, 2019; revised June 12, 2019 and September 24, 2019; accepted October 7, 2019. Date of publication October 16, 2019; date of current version December 24, 2019. This work was supported in part by the National Science Foundation Grant 1741472 and in part by the University of Rochester Pilot Award Program in AR/VR. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lei Xie. *(Corresponding author: Sefik Emre Eskimez.)*

S. E. Eskimez and Z. Duan are with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14623 USA (e-mail: eeskimez@ur.rochester.edu; zhiyao.duan@rochester.edu).

R. K. Maddox is with the Department of Biomedical Engineering and Department of Neuroscience, University of Rochester, Rochester, NY 14623 USA (e-mail: rmaddox@ur.rochester.edu).

C. Xu is with the Department of Computer Science, University of Rochester, Rochester, NY 14623 USA (e-mail: chenliang.xu@rochester.edu).

Digital Object Identifier 10.1109/TASLP.2019.2947741

evaluations show that our proposed method yields better results than two state-of-the-art baseline methods. Results also show significant improvement thanks to the noise-resilient training method in non-stationary noise conditions. Through subjective evaluations, we show that the generated 3D face landmarks demonstrate a convincing match with the speech audio signals. To promote scientific reproducibility, we release several generation examples, code of the proposed system, and pre-trained models.[1]

Compared to our preliminary work [8], we make the following contributions in this work: 1) We generate 3D face landmarks as opposed to 2D as our previous work. Including the 3 rd dimension allows novel applications such as AR/VR, video games and movie dubbing. 2) Instead of Mel-Frequency Cepstral Coefficients (MFCC) and their temporal derivatives, we directly input the raw waveform to the network. 3) We propose a new network architecture that replaces Long Short-Term Memory (LSTM) layers with convolutional layers for improving the results on raw waveform inputs. 4) We propose a noise-resilient training method to incorporate speech enhancement ideas at the feature level to increase the system's robustness to non-stationary background noise. This noise-resilient training method can be applied to other speech processing tasks such as automatic speech recognition, emotion recognition, and speaker identification/verification.

The rest of the paper is organized as follows: We first present related work on speech animation in Section II. We then describe the proposed neural network system and its data preparation in Section III. Then, we present implementation details, the objective and subjective evaluations, and the analysis of the architecture in Section IV. Finally, we conclude the paper in Section V.

II. Related Work

Audio-visual and multi-modal approaches have received much attention in signal processing in recent years. When multi-modal inputs are available, performance of many tasks such as speech enhancement, source separation, speech recognition, emotion recognition and voice activity detection can be significantly improved [10]–[15]. When one modality (e.g., text, audio, visual) is missing, cross-modal generation approaches have been proposed to reconstruct that modality from other modalities [16]–[18].

Generating talking faces from speech belongs to cross-modal generation. It has drawn much attention from researchers in recent years. There are *shape model-oriented* methods and *image-oriented* methods. Shape model-oriented methods usually employ a deformable face shape model, where the face shape is represented by sparse points in a 2D or 3D space. These models can be controlled by low dimensional parameters that are often obtained by principal component analysis (PCA) or other dimensionality reduction methods. Image-oriented models predict the RGB face or mouth image sequences directly from speech. Some of these methods use intermediate representations as constraints, such as the face or mouth landmarks.

Some works generate talking faces from text [19]–[24]. There is a key difference between text-driven and speech-driven talking faces. Speech signals show large variations across speakers, emotions, and accents for the same text, and the generated talking face must be in sync with the input speech. However, for text-driven faces, any plausible talking face is sufficient. These two tasks require different approaches. Therefore, in the following, we only review speech-driven talking face generation methods.

A. Image-Oriented Methods

Suwajanakorn *et al.* [5] demonstrated an LSTM-based system on synthesizing videos of President Barack Obama from his speech. This system uses a two-stage approach. It first uses an LSTM to predict PCA coefficients of 18 mouth landmark points from 13 MFCCs plus the energy term. Then, according to the predicted PCA coefficients, few nearest candidate frames are selected from the dataset that contains the images of the target identity, and the weighted median is applied to synthesize the texture. Therefore, this method is a hybrid shape-image model since it predicts mouth landmarks first. Although the results are photo-realistic and impressive, the method requires a large amount of training data for the target identity. It can accept speech from a different person; however, it can only generate the face of the person in the training data. It is also computationally heavy, making it difficult to run on edge devices.

Chung *et al.* [6] proposed a method that accepts 12 MFCCs and a single frame target image to generate a talking face video. The system uses an audio encoder and an identity encoder to convert audio features and the target image to their respective embeddings. It then uses an image decoder to generate face images from these embeddings. Since the generated images are blurry, the system utilizes a separate deblurring module to sharpen the images. All modules are based on 2D convolutional neural networks. This method can run in real-time on a GPU. Similar to this work, Chen *et al.* [25] proposed a method that leverages an adversarial loss function in addition to a pixel-level reconstruction loss and a perceptual loss, to generate sequences of images from speech. The network accepts the speech and a target lip image as inputs and outputs 16 frames of lip images that are synchronized with the speech. The network contains an audio encoder, an identity encoder, and 3D convolutional residual blocks. Compared to Chung et al.'s method, the generated images are sharper, and a deblurring module is not needed.

A disadvantage of these systems is that facial expressions, animations, and for some systems, the identity information, are difficult to manipulate during generation. The shape model-oriented methods usually predict an intermediate representation that can be manipulated before rendering the details.

B. Shape Model-Oriented Methods

Early works focused on Hidden Markov Models (HMMs) to map from speech to talking faces [26]–[32]. Voice puppetry [26]

[1][Online]. Available: http://www.ece.rochester.edu/projects/air/projects/3Dtalkingface.html

was one of the early works. It models 26 points of a face using HMM and drove them using linear predictive coding and relative spectral transform - perceptual linear prediction audio features. Choi *et al.* [27] used HMM inversion (HMMI) to estimate visual parameters from 12 MFCCs of speech, where the visual parameters include the left and right corners of the mouth and the heights of the upper and lower lips.

Cosker *et al.* [28], [29] employed a hierarchical model that models subareas of the face independently by an active appearance model (AAM) [33] and then merges them into a full face containing a total of 82 landmark points. Each sub-area is driven by 12 MFCCs of speech. Xie *et al.* [30] proposed a system that generates only the mouth region using coupled HMMs (cHMMs) to compensate audio-visual asynchrony. They used MFCCs and their first- and second-order derivatives as speech features and PCA coefficients of the mouth region as the visual parameters. Zhang *et al.* [32] also used PCA coefficients of the mouth region as the visual parameters, but estimated HMM states from speech features with a deep neural network (DNN).

Recent works use deep neural networks to map speech features to face landmarks. Pham *et al.* [34] proposed an LSTM network that predicts the 3D face model parameters from speech input features, namely MFCCs and the chromagram. The authors later improved their work by using spectrograms as input and employing convolutional and recurrent layers [7] in the network architecture. Karras *et al.* [35] employed a network that maps speech into 3D positions of 5022 landmark points. The network can generate realistic faces with emotions using only 5 minutes of training data. However, their system is designed for the generation of a single speaker.

Compared to these listed works, our approach includes a novel pre-processing method that normalizes the identities of the target data (face landmarks). This normalization improves the quality of results, leads to faster convergence for neural network training and can work using fairly simple network architectures. In addition, our approach uses a noise-resilient training mechanism to ensure its robustness in noisy conditions. To our best knowledge, this is the first consideration of background noise in the system design and evaluation of talking face generation from speech. Furthermore, compared to shape model-oriented methods described above, our system predicts landmarks by the multi-pie 68 point markup convention [36], which is used by most of the existing systems for facial landmark detection, face morphing, and face swapping applications. This allows our system to be seamlessly integrated into a pipeline for facial manipulation with such systems. A quantitative side-by-side comparison with the most closely related methods, however, is difficult. Karras *et al.* [35] and Pham *et al.* [34] are the two most similar methods to ours, but their systems are optimized for different face models making a side-by-side comparison difficult with ours. In particular, Karras *et al.* used facial motion capture to obtain a 3D mesh model of the face. Pham *et al.* used a 3D mesh model of the face built from a Kinect point cloud, and they developed a technique to map videos into this 3D mesh model. We do not have access to their code of these face models, and we believe that it is not fair to those methods to re-implement them but using our 68 point face model to compare with ours.

Considering the above-mentioned difficulties, we eventually chose to compare with the landmark generation part of the system proposed by Suwajanakorn *et al.* [5] and our prior preliminary system [8]. The system in [5] is a state-of-the-art image-oriented system that generates realistic face images of a single speaker. As an intermediate step, it also predicts PCA coefficients of mouth landmarks similar to our method. Therefore, we believe that it is a reasonable baseline for our method.

III. Method

In this section, we describe face landmark extraction, landmark pre-processing before training the neural network, the proposed neural network architecture, the proposed method to increase the system's resilience against background noise, and how it works during the inference process.

A. Pre-Processing

1) Face Landmark Extraction: We first use the open source library DLIB [37] to extract 2D face landmarks (x and y coordinates), and then use the method described in [38] to estimate 3D face landmarks from these 2D landmarks and their corresponding video frames. We extract a total of 68 landmarks, following a standard in the mark-up convention described in [36]. Face shapes formed by connecting these landmarks are shown in the first row of Fig. 1.

2) Face Landmark Alignment: The extracted raw landmarks are in pixel coordinates and can be at different positions, scales and orientations. These variations make it difficult to train our neural network, as they are largely irrelevant to the input speech. To minimize these variations, we use Procrustes analysis [39] to align the 3D landmarks. This is a common practice for creating active shape models (ASMs) [40] and active appearance models (AAMs) [33], [41]. Face shapes after alignment are shown in the second row of Fig. 1.

3) Face Landmark Identity Removal: Different speakers have different face shapes, where mouth, nose, and eyes may not be well aligned across speakers even after the Procrustes analysis. These variations are also less correlated to the input speech. Therefore, we want to remove this *identity variation* from our 3D face landmarks. To achieve that, for each landmark sequence, we detect one reference frame that contains a closed mouth by thresholding the distance between the upper lip and lower lip coordinates. We then calculate the landmark coordinate deviations from this reference frame for each frame in the sequence, and impose these deviations onto a template face across all sequences of all identities. This template face is calculated as the average of aligned faces with a closed mouth across all identities. The 3D face landmarks can be represented as:

$$\mathbf{s} = (x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_N, y_N, z_N)^T, \tag{1}$$

where N is the number of vertices and T denotes vector transpose. The identity removal operation can be described as:

$$\mathbf{s}_{IR} = \mathbf{s} - \mathbf{s}_{CM} + \mathbf{s}_T, \tag{2}$$

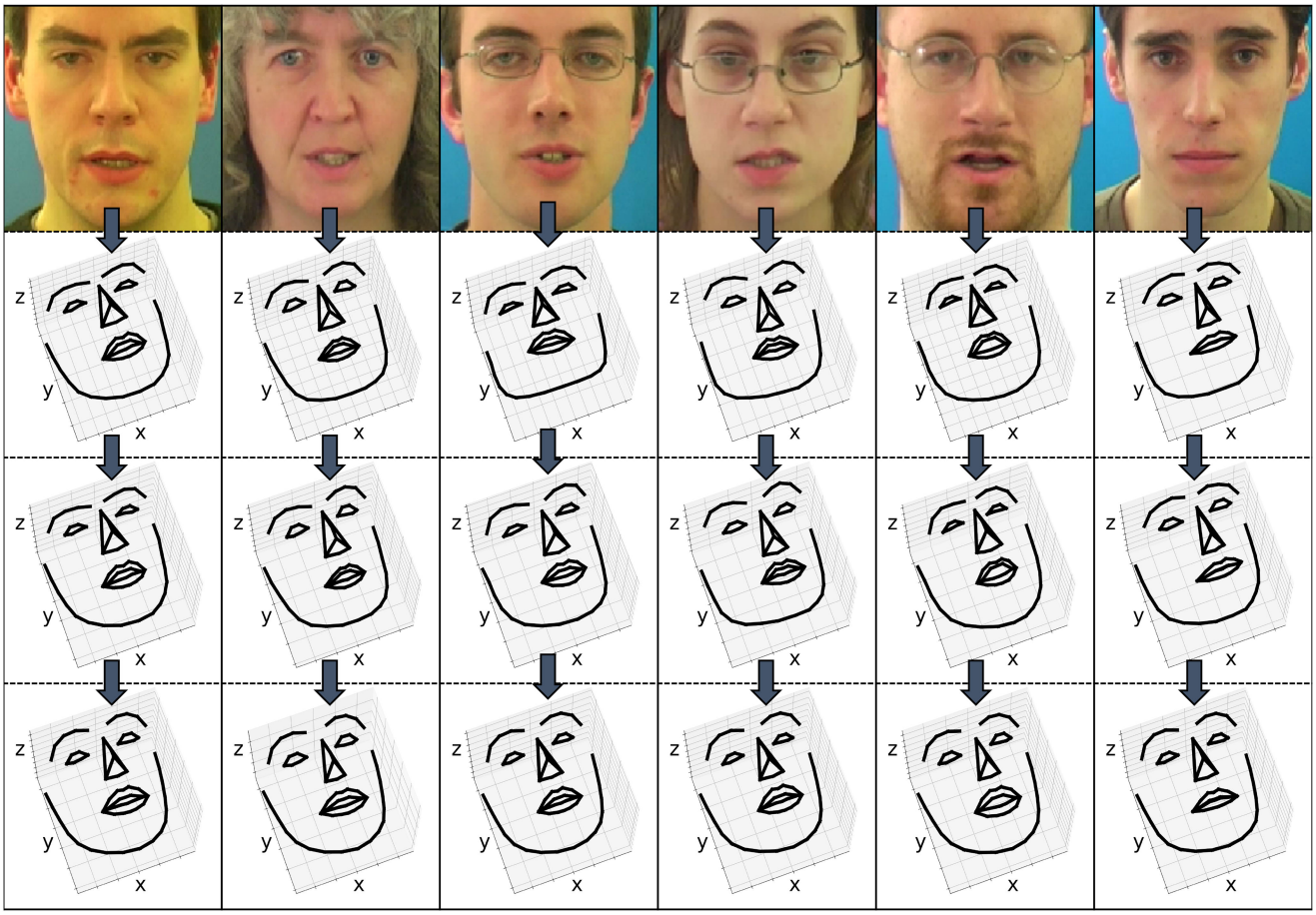

Fig. 1. Data preparation steps for face landmarks illustrated on six different speakers, where each column corresponds to a speaker. We draw lines between certain landmarks to form face shapes. The first, second, and third rows show raw face landmarks extracted from video images, landmarks after Procrustes alignment, and landmarks after identity removal, respectively.

where $\mathbf{s}_{IR}$ represents the identity removed face shape, $\mathbf{s}_{CM}$ is a face frame with mouth closed that is automatically selected from the video, and $\mathbf{s}_T$ is the template (reference) shape. Face shapes after identity removal are shown in the third row of Fig. 1.

4) Active Shape Model (ASM): ASMs [40] are deformable shape models that can represent the variations in the training set by a set of coefficients. These coefficients are the weights for eigenvectors that are obtained by PCA. By using the parameters obtained from PCA, $\mathbf{s}$ in Equation (2) can be described as follows:

$$\mathbf{s} = \mathbf{s}_\mu + \mathbf{wS}, \tag{3}$$

where s_μ is the mean shape vector, $\mathbf{w} = [c_1, \ldots, c_P]$ is a vector that contains the weights and $\mathbf{S} = [\mathbf{s_1}, \ldots, \mathbf{s_P}]$ is a matrix that contains the eigenvectors. P is the number of PCA components, which is smaller than their dimensionality ($P < N$).

We create pairs of raw speech waveforms and corresponding ASM weights $\mathbf{w}$ as the input-output pairs for neural network training.

5) Data Augmentation: By removing the target identity from the 3D face landmarks, we already standardized the face landmarks, which is described in Section III-A3. We do not further augment the landmark data.

For the speech input, we perform data augmentation to improve our system's robustness to pitch and loudness variations. Augmentation is not performed before but rather during training iterations. For each sample in each training batch, we randomly choose whether we use the original training sample or an augmented sample. If it is the latter, two augmentation steps are applied in a sequence. We first pitch shift the sample by one or two semitones up or down. We then apply a gain factor to the amplitude of the sample between -12 dB and 6 dB with a 3 dB granularity. It is noted that this dynamic augmentation is random, but it saves memory compared to a preset augmentation beforehand.

B. Network Architecture

The deep neural network (DNN) accepts a frame (280 ms) of the raw waveform as an input and outputs the ASM weights of that frame. There are four convolutional layers with 1D filter kernels operating on the raw waveform. The number of filters grows as the time dimension shrinks. We use strides for each convolutional layer, which halves the time-steps. Each convolutional layer is followed by LeakyReLU activation with a slope of 0.3 and a dropout layer that discards 20% of the units.

TABLE I
DETAILED PARAMETERS OF THE PROPOSED NETWORK ARCHITECTURE. THE NUMBER OF FILTERS AND HIDDEN UNITS, FILTER SIZES, STRIDES, ACTIVATIONS, AND OUTPUT SHAPES ARE SHOWN FOR EACH LAYER. *1D_CNN_TC* IS IDENTICAL TO *1D_CNN*; FURTHER, IT ACCEPTS CONDITION INPUT AND CONCATENATES IT WITH THE OUTPUT OF THE FULLY CONNECTED (FC) LAYER THAT IS SHOWN IN THE LAST TWO ROWS OF THE TABLE. THIS CONCATENATED TENSOR IS FED TO ANOTHER FC LAYER THAT OUTPUTS THE FINAL ASM WEIGHTS

Net	Layers	Number of Filters or Hidden Units	Filter Size	Strides	Activation	Output Shape
1D_CNN	Input	-	-	-	-	(2240, 1)
	Conv	64	(21, 1)	(2, 1)	LeakyReLU	(1110, 64)
	Conv	128	(21, 1)	(2, 1)	LeakyReLU	(545, 128)
	Conv	256	(21, 1)	(2, 1)	LeakyReLU	(263, 256)
	Conv	512	(21, 1)	(2, 1)	LeakyReLU	(122, 512)
	FC	6	-	-	LeakyReLU	(6)
1D_CNN_TC	Condition	-	-	-	-	(6)
	FC	6	-	-	LeakyReLU	(6)

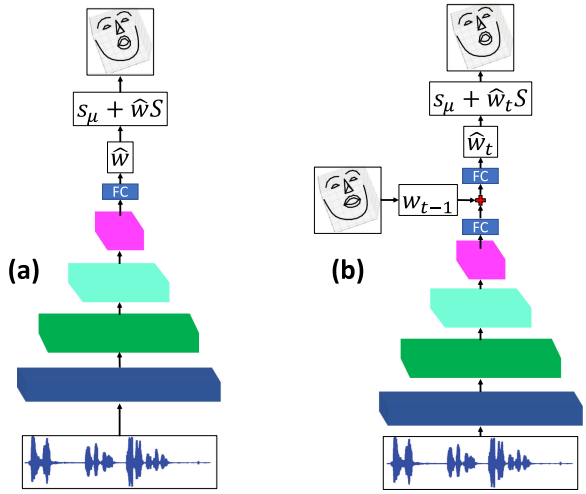

Fig. 2. The network architecture for (a) *1D_CNN* network and (b) *1D_CNN_TC* network. *1D_CNN_TC* is identical to *1D_CNN*, except that it accepts the previous frame's ASM weights as a condition to enforce temporal constraint. Raw waveform is fed to four convolutional layers, followed by a fully connected (FC) layer.

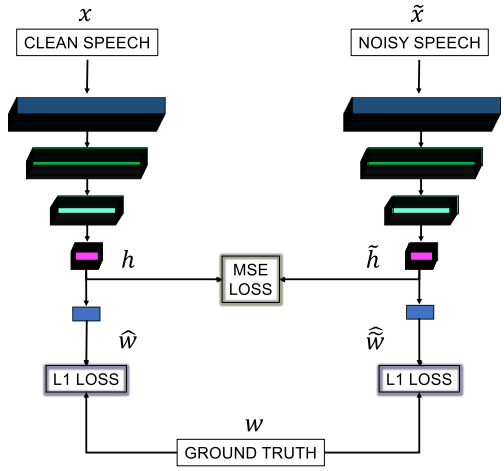

Fig. 3. The noise-resilient training scheme. The networks on the left and right sides are the same, and their weights are shared. The clean and noisy speech goes through the left and the right networks, respectively, to reconstruct their face landmarks. A mean-squared error (MSE) constraint is applied to the latent representations to incorporate the supervised speech enhancement idea at the feature level.

The final layer is a fully connected layer that outputs the ASM weights. The network architecture is shown in Table I and Fig. 2.

In order to have smooth transitions between generated talking faces across frames, we further added a temporal constraint to the network architecture. It accepts the previous frame's ASM weights as a condition in order to obtain smoother results over time. The condition is concatenated to the intermediate tensor immediately after the fully connected layer, and we add another fully connected layer as shown in Table I. We discuss the trade-off between these two models in Section IV. We denote our proposed method as *1D_CNN* and the temporally constrained version as *1D_CNN_TC* throughout the rest of this paper.

The network minimizes the L1 loss between the predicted and ground-truth ASM weights, as follows:

$$J_{\ell 1}(\mathbf{w}, \widehat{\mathbf{w}}) = \|\mathbf{w} - \widehat{\mathbf{w}}\|_1 \,, \tag{4}$$

where $\widehat{\mathbf{w}}$ is the ASM weight vector predicted by the network. Equation (4) shows the loss for a single sample. During training, the average of all training samples is minimized.

C. Noise-Resilient Training

To make the system robust to noise, we propose a novel, yet simple method for noise-resilient training. The idea is to match the intermediate features obtained from the clean and noisy speech, as in theory, they contain the same speech information hence the extracted features are ideally be the same. This is shown in Fig. 3. The clean features h is obtained by feeding the clean speech x to the network. The corrupted features $\tilde{h}$ is obtained by feeding the corrupted speech $\tilde{x}$ to the same network. In addition to the ASM coefficient loss on both networks, we also add the weighted MSE between h and $\tilde{h}$:

$$J = J_{\ell 1}(\mathbf{w}, \widehat{\mathbf{w}}) + J_{\ell 1}(\mathbf{w}, \widehat{\tilde{\mathbf{w}}}) + \lambda \|\mathbf{h} - \tilde{\mathbf{h}}\|_2 \,, \tag{5}$$

where λ is the weighting coefficient, and $\widehat{\tilde{\mathbf{w}}}$ is the ASM parameters generated from corrupted speech $\tilde{x}$.

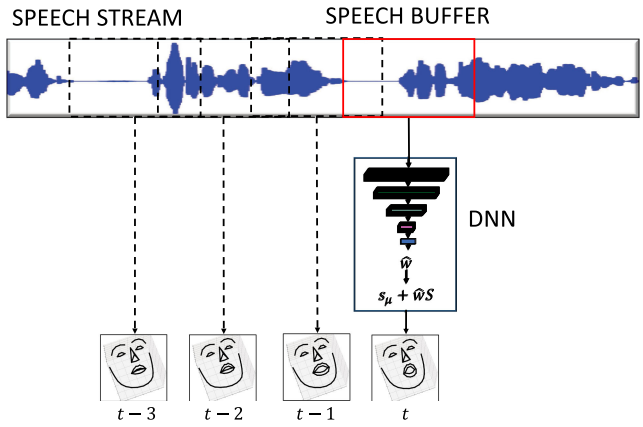

Fig. 4. System overview. A talking face is generated every 40 ms (frame hop size) from 320 ms (frame length) of audio. t represents the time.

D. System Overview

During inference, our system utilizes a speech buffer that acts as first-in-first-out (FIFO) queue. First, the speech buffer is initialized with zeros. When the system receives new speech data, it is pushed to the speech buffer, and the network predicts the next frame's weights. There is no pre-processing applied to the speech; the raw speech is directly fed to the neural network. The predicted weights are converted to 3D landmark points using Equation (3). The system overview is shown in Fig. 4.

For the *1D_CNN_TC* network, the system utilizes another buffer, called the conditioning buffer, that stores the last frame. The conditioning buffer is initialized with the template face shape weights.

IV. Experiments

A. Datasets

We start our experiments in a single-speaker setting. To this end, we follow Suwajanakorn *et al.* [5] and utilize President Obama's weekly address videos, which are available online.[2] We downloaded 315 videos that have 3 minutes average duration, totaling to approximately 18 hours of content. The videos are provided in 30 frames per second (FPS) and we down-sampled the videos to 25 FPS. We split the dataset into training (70%), validation (15%), and testing (15%) sets.

For multi-speaker experiments, we use a publicly available audio-visual dataset called GRID [9] to train our system. There are 34 native English speakers in this dataset, with 16 female and 18 male speakers, who are ranging from 18 to 49 years old. All of the speakers are from England except one from Scotland and one from Jamaica. Each speaker has 1000 recordings that are 3 seconds in duration. The recordings contain sentences that are identical for each speaker. The structure of the sentences is in the following form: *command (4) - color (4) - preposition (4) - letter (25) - digit (10) - adverb (4)*, where the numbers of choices are shown in parenthesis for each component. An example sentence can be given as "*set blue at C 5 please*".

[2][Online]. Available: https://obamawhitehouse.archives.gov/briefing-room/weekly-address

Recordings are provided both in audio and video format. In this study, we use the high-resolution videos included in the GRID dataset. These videos have a frame rate of 25 FPS and a resolution of 720×576 pixels. Since each recording is 3 seconds in duration, each video has a total of 75 frames. The video files contain the corresponding audio that has a sampling rate of 44.1 kHz. We down-sample the audio to 8 kHz which is a typical sampling rate for speech signals in telecommunication. We split the GRID dataset into training (92%) and validation (8%) sets.

We employ another multi-speaker dataset that is disjoint from the GRID dataset for testing on unseen speakers, namely Speech Test Video Corpus (STEVI) [42]. Specifically, we employ the *High-Probability speech perception in noise (SPIN) Sentences* and *Nonsense Sentences* listed in [42]. The videos are provided in 29.97 FPS and 1920×1080 resolution. The audio stream has a sampling rate of 48 kHz. We down-sample the audio to 8 kHz and generate 3D talking faces. Since our system is trained to generate 25 FPS videos, we use cubic spline interpolation to up-sample the generated videos to 29.97 FPS to match with the ground truth face landmarks. There are a total of 4 speakers, each of which has 400 sentences, 200 *High-Probability SPIN Sentences* and 200 *Nonsense Sentences*. The duration of each sentence is around 2 to 3 seconds.

We use DLIB [37] and [38] to extract face landmarks from these videos according to Section III-A1 for training, validation and testing. To verify the validity of the extracted face landmarks, we employ a two-step approach. First, we run a script that automatically identifies wrong landmarks by comparing the upper and lower lip landmark positions and eliminates invalid landmark sequences. This script was applied to all extracted landmarks. In the second step, we manually check the landmarks and eliminate problematic sequences. Since manual verification is costly, the second step is only applied to the STEVI dataset and the test set of the Obama dataset. In this way, we further improve the quality and validity of the evaluation data.

To create noisy speech input, we employ a noise dataset named Sound Ideas [43] that contains 138 different noise types including non-stationary noises from various environments such as nature, city, domestic, office, traffic, and industry. A noisy speech is created by mixing a clean speech file with a randomly selected noise file in 6 to 30 dB SNRs with 3 dB increments.

B. Implementation Details

Our system was trained to generate 25 FPS videos, i.e., the system generates a talking face every 40 ms. We include the context information to our input speech. Specifically, we concatenate 3 frames from past and future, totaling 7 frames. For 8 kHz speech signals, a 40 ms window contains 320 data points. The input speech size becomes $7 \times 320 = 2240$ as shown in Table I. The networks were trained for 100 epochs, and the weights were saved only if the validation loss was improved for each epoch. We implemented our method in PyTorch [44]. The mini-batch size and learning rate were set to 128 and 10^{-4}, respectively. We used Adam [45] optimizer during training.

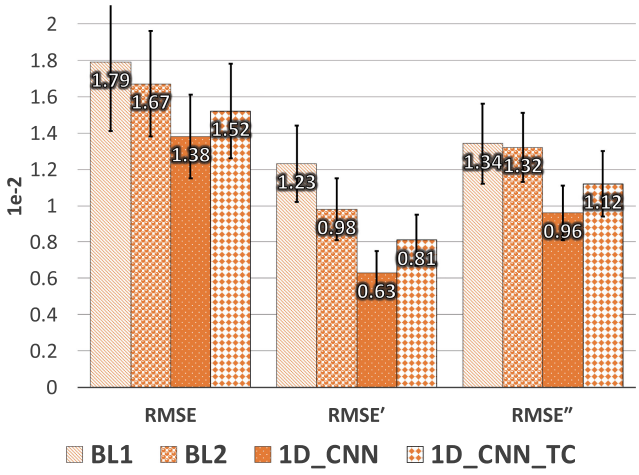

Fig. 5. Single-speaker objective evaluation results for the *BL1* [5], *BL2* [8], *1D_CNN* and *1D_CNN_TC* methods. We calculate the root-mean-squared error (RMSE) between generated and ground-truth 2D mouth landmarks, and its first- and second-order temporal derivatives. Error bars show the standard deviation.

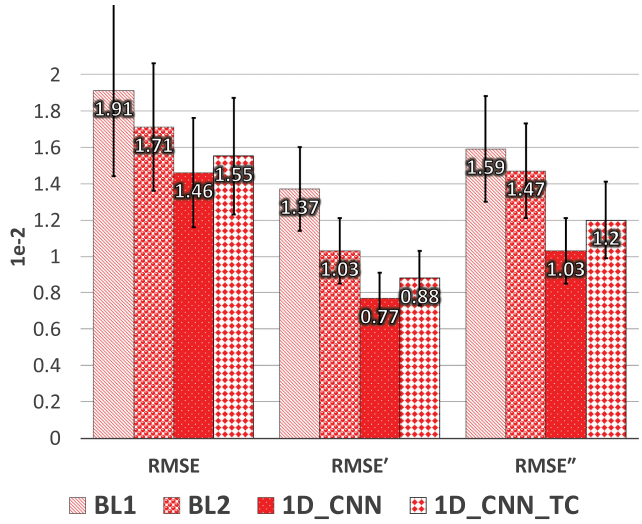

Fig. 6. Multi-speaker objective evaluation results for the *BL1* [5], *BL2* [8], *1D_CNN* and *1D_CNN_TC* methods. We calculate the root-mean-squared error (RMSE) between generated and ground-truth 2D full face landmarks, and its first- and second-order temporal derivatives. Error bars show the standard deviation.

We compared our method with Suwajanakorn et al.'s [5] landmark generation method denoted as *BL1* and our preliminary work [8] denoted as *BL2*. *BL1* utilizes a single LSTM layer with a time delay to generate 20 PCA coefficients for the mouth landmarks. The input of their network is the 13 MFCCs plus the log mean energy and their first temporal derivatives. *BL2* accepts first and second derivatives of 13 MFCCs of speech as input and outputs PCA coefficients for the whole face landmarks. There are 4 LSTM layers in the network architecture.

For single-speaker experiments, we trained all of the above-mentioned methods on the Obama dataset, while for multi-speaker experiments, we trained them on GRID and evaluated them on STEVI.

C. Objective Evaluation

We used the root-mean-squared error (RMSE) between the ground-truth and predicted face landmark sequences and their first and second derivatives for evaluation. Although our system generates 3D landmarks, we used only x and y coordinates (2D landmarks) of the results of our system for these calculations since the baseline can only generate 2D face landmarks. Therefore, all numbers reported in this section were obtained from 2D landmarks. Before we evaluated the landmarks, we normalized the values between 0 and 1. Therefore, each 0.01 RMSE value corresponds to approximately 1 percent of the face length.

For the single-speaker setting, we evaluated our systems and the baseline systems with the test set of Obama dataset by using only the mouth landmarks. For the multi-speaker setting, we used unseen speakers from STEVI corpus. Figs. 5 and 6 show the single- and multi-speaker results, respectively, for the baseline methods (*BL1* and *BL2*), and two versions of our proposed methods (*1D_CNN* and *1D_CNN_TC*).

For the single-speaker setting, the results show that the *1D_CNN* method yields the best objective results with an RMSE value of 1.38×10^{-2} followed by *1D_CNN_TC* with an RMSE value of 1.52×10^{-2}. For the multi-speaker setting, the trends are similar: the *1D_CNN* method yields the best objective results with an RMSE value of 1.46×10^{-2} followed by *1D_CNN_TC* with an RMSE value of 1.55×10^{-2}. There is a significant improvement over the *BL1* method that has an RMSE value of 1.91×10^{-2}.

1D_CNN_TC results are smoother due to the temporal constraint. However, the resulting mouth movement of the talking faces has weaker high-frequency movements. This can also be observed from the multi-speaker objective results. The RMSE′ and RMSE″ are higher for *1D_CNN_TC* (0.88×10^{-2}, 1.2×10^{-2}) compared to *1D_CNN* ($0.77 \times 10^{-2}, 1.03 \times 10^{-2}$); and both of them are better than the baseline methods. A paired t-test shows that results of both proposed systems are statistically significantly better than the baseline at a significance level of 0.01 for all the three measures.

There is a trade-off between these two versions of our method. From our observations of the generated outputs, *1D_CNN* yields better mouth movement and mouth shape match, where *1D_CNN_TC* yields more stable and smoother shape changes over time. One may prefer *1D_CNN* for applications that focus on improving speech comprehension since high-frequency mouth movement is essential in such cases, and one may prefer *1D_CNN_TC* for general speech animation applications. An example result for the word "ashes" has been shown in Fig. 7.

D. Analysis of the Network

We further analyze the *1D_CNN* network architecture by changing the number of convolutional layers, the number of filters in each layer, and the input speech size in the multi-speaker setting on the STEVI corpus.

1) Number of Layers: The original configuration contains four convolutional layers. We conducted experiments with 2,

Fig. 7. The example output showing the pronunciation of the word "ash". The speech sample was taken from STEVI corpus. The first row shows the result generated by *1D_CNN*. The second row shows the comparison of the result generated by *1D_CNN* and the ground-truth (dotted red line). The third and fourth rows show the result generated by *1D_CNN_TC* and comparison with the ground-truth (dotted red line). Columns show every three frames.

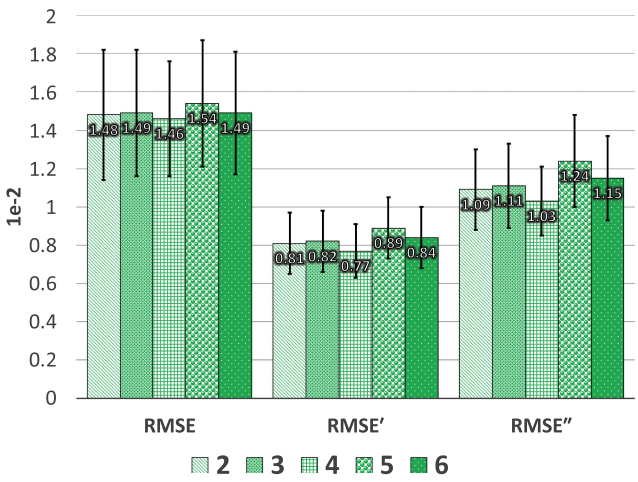

Fig. 8. Comparison of *1D_CNN* configurations with different number of convolution layers. The number of filters for Layers 1 to 4 is shown in Table I. The number of Layers 5 and 6 is both 512. We compare the root-mean-squared error (RMSE) between generated and ground-truth landmarks, and its first- and second-order temporal derivatives. Error bars show the standard deviation.

3, 4, 5 and 6 convolutional layers, and compared the objective results. For fifth and sixth layers, we used 512 filters.

The results are shown in Fig. 8. The 4-layer configuration achieves the best results, where 5-layer configuration has the worst results. An interesting outcome is that the 2-layer configuration has the second best results. For RMSE″, there is a big gap between the 4-layer configuration and others. A paired t-test shows that RMSE′ and RMSE″ results of 4-layer configuration is statistically better at a significance level of 0.01 compared to other configuration results. In conclusion, we selected 4-layer configuration in our final models.

2) Number of Filters: Table I shows the number of filters for the convolutional layers, which are $x = 64, 2x = 128, 4x = 256$ and $8x = 512$ for the four layers, respectively, in the original configuration. We varied x to have values of 16, 32, 64, and 128 and compared the objective results.

Fig. 9 shows the results. Networks with $x = 16$ and $x = 32$ perform similarly for all metrics. The network with $x = 128$ has the worst performance compared to other configurations; We suspect that this is due to over-fitting given its largest capacity. The network with $x = 64$ performs better than other configurations. A paired t-test shows that RMSE′ and RMSE″ results of $x = 64$ configuration is statistically better at a significance level of 0.01 compared to other configuration results. Therefore, we selected $x = 64$ as the final parameter for our networks.

3) Input Speech Size: The input speech includes context information of past and future frames as described in Section IV-A. In the original configuration, we use 7 frames of speech, including 3 frames before and 3 frames after the current frame. Each frame corresponds to 40 ms of speech. In this section, we vary the input size from 5, 7, and 9 frames and compare the performance.

The results are shown in Fig. 10. The RMSE results are similar; However, for RMSE′ and RMSE″, 7 frames

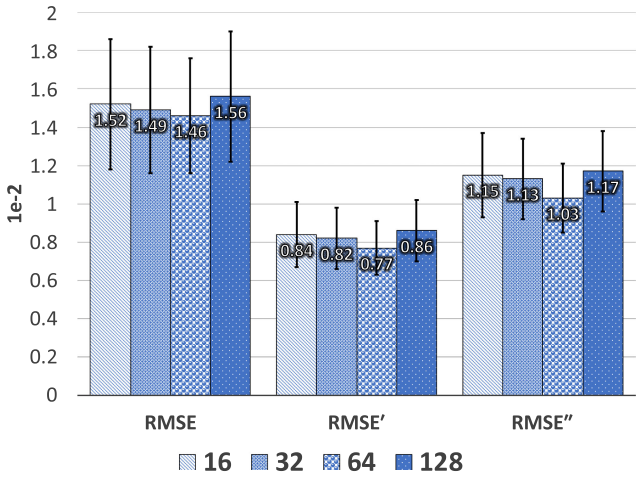

Fig. 9. The comparison of *1D_CNN* configurations that has a different number of filters in convolutional layers is shown. The number of filters in the first layer is displayed, which are 16, 32, 64, and 128. After the first layer, the filters are doubled with each following convolutional layer. We compare the root-mean-squared error (RMSE) between generated and ground-truth landmarks, and its first- and second-order temporal derivatives. Error bars show the standard deviation.

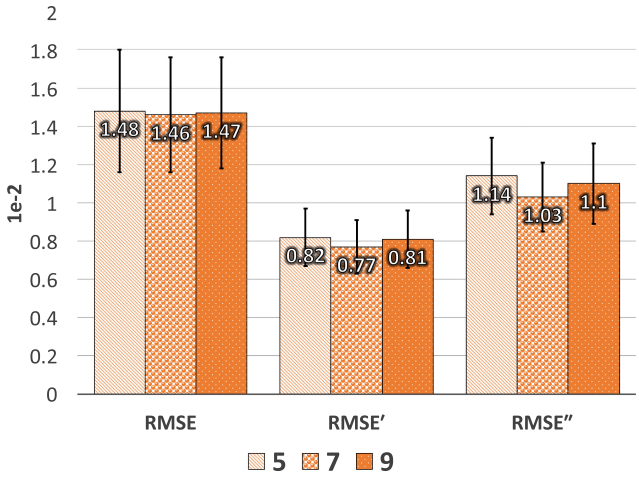

Fig. 10. The comparison of results for different sizes of the input speech is shown for *1D_CNN* network. The number of frames is displayed, which are 5, 7, and 9. Each frame spans 40 ms speech. We predict the middle frame and use previous and past frames as context information. We compare the root-mean-squared error (RMSE) between generated and ground-truth landmarks, and its first- and second-order temporal derivatives. Error bars show the standard deviation.

configuration has better results. A paired t-test shows that RMSE′ and RMSE″ results of 7 frames configuration is statistically better at a significance level of 0.01 compared to 5 and 9 frames configuration results. In our final network, we selected 7 frames of speech as our input.

E. Resilience Against Noise

In this section, we evaluate our system on noisy conditions. We consider five types of noise for the evaluation, namely babble, factory, speech-shaped noise (SSN), motorcycle and cafeteria. We mix the speech files of STEVI corpus with the

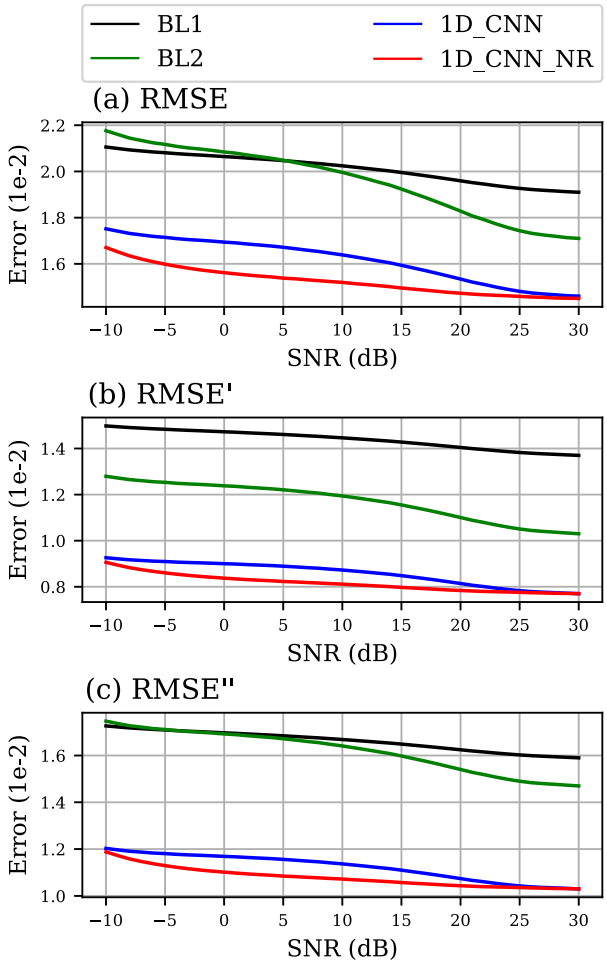

Fig. 11. Average RMSE, RMSE′, and RMSE″ of face landmark generation with five unseen noise types (babble, factory, SSN, motorcycle and cafeteria) at an SRN from −10 to 30 dB with 1 dB increments.

noises described above in 5 and 10 dB signal-to-noise ratio (SNR) values and report the RMSE values of the generated faces. Note that the noise types used in evaluation were not included in the training set, and were obtained from a different source (i.e., different recording conditions).

For the noise-resilient training method, we initialized the weights using the pre-trained weights from the clean version of our network and reduced the learning rate to 10^{-5}. We conducted experiments and varied the λ parameter in Equation (5) between 1 and 0, and found that 10^{-2} performs the best. Therefore, we set λ to 10^{-2}. The network was trained for 100 epochs. The noise resilient (NR) version of our network is denoted as *1D_CNN_NR* throughout the rest of the paper.

Fig. 11 shows the average results of the baseline systems and the proposed system with and without the noise-resilient training method on noisy speech with five kinds of noises at SNRs ranging from −10 dB to 30 dB with 1 dB increments. The test noises include babble, factory, SSN, motorcycle and cafeteria noise types, none of which were included in the training. The results show that the proposed method outperforms both baselines on all three measures across all SNRs. In particular, the proposed

method with noise-resilient training at the worst SNR (−10 dB) achieves better performance than both baselines at the highest SNR (30 dB). Comparing *1D_CNN_NR* with *1D_CNN*, we see a significant performance boost thanks to noise-resilient training. When the SNR is between −5 dB and 20 dB, the reduction of errors is equivalent to an increase of SNR for 15–20 dB (e.g., errors for *1D_CNN_NR* at −5 dB SNR are about the same as those for *1D_CNN* at 15 dB). It is noted that the training noises were mixed in SNRs ranging from 6 dB to 30 dB with 3 dB increments (as described in Section IV-A); therefore, these results show that the noise-resilient training can generalize to the unseen SNRs, even in extremely noisy conditions (negative SNRs). We conducted a paired t-test between the results of *1D_CNN* and *1D_CNN_NR* for each noise category in −5 dB to 20 dB SNRs. The results show that the three versions of RMSE values are all statistically significant at a significance level of 0.01.

F. Subjective Evaluation

To further evaluate the match between generated face landmarks and the input speech, we conducted a subjective Turing test in the multi-speaker setting. We recruited 20 volunteers as our evaluators. We presented each evaluator a random selection of 16 samples generated by *BL1*, 16 samples generated by *BL2*, 16 samples generated by the proposed system, and 16 samples of ground-truth landmarks. All of the speech samples were taken from the STEVI dataset, which was not used for training. For the *BL1*, we retrained the system with full 68 face landmarks' PCA coefficients instead of just the mouth landmarks' PCA coefficients in order to conduct the subjective tests. We found out that using only the mouth region compared to using all 68 face landmarks does not change mouth movements. This is due to the alignment of face landmarks in pre-processing; the regions besides the mouth region do not change much.

The generated landmarks were painted and added teeth and eyes in order for evaluators to easily recognize the faces and mouth movements. During evaluation, a few ground-truth talking face samples were shown to each evaluator. Then, the 64 samples were presented to the evaluator in a random order, and the evaluator was asked to assign a score between 0 (worst) and 100 (best) based on the match between the speech and mouth movement. Each sample was presented twice before the evaluator was asked to assign a score.

The results are shown in Fig. 12. The proposed method significantly scores higher than the baseline methods. These results show comparable scores for our method and the ground-truth face landmarks, indicating that our system can generalize well to unseen speakers and can convince evaluators that speech and articulation match strongly. A paired t-test shows that the *1D_CNN* results are statistically significantly better compared to the both baseline results at the significance level of 0.01.

G. Limitations

As a data-driven approach, the performance of our method highly depends on the the training data. The dataset should contain a wide variety of phonemes, ideally uniformly distributed. However, the GRID dataset is limited in terms of the words and phonemes it includes. Our future work includes expanding the training set to include more data that has rich phonetic content and balancing the data in order to have uniformly distributed phoneme content.

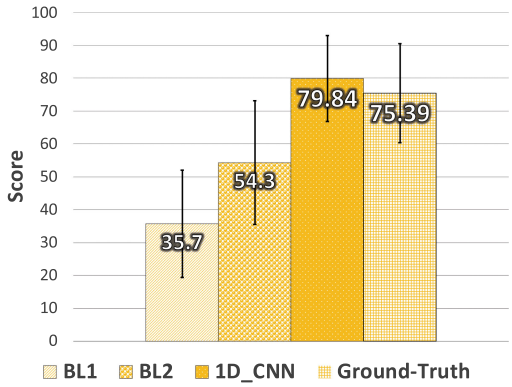

Fig. 12. The results for the subjective test of speech-mouth match. The bars show the average score for the baseline method, proposed method (1D_CNN) and ground-truth face landmarks. Error bars show the standard deviation.

The performance of our system is proportional to the performance of the face landmarks extractor on the training data. The extractor we used in this study works on each single frame and does not consider temporal relations across frames. This might be the main reason for noisy mouth movements in the extracted landmarks. We believe that by utilizing a video-based face landmark extractor that models temporal dependencies of landmarks, the quality of landmark extraction and our trained model will be improved.

V. Conclusion

In this work, we proposed a new noise-resilient neural network architecture to generate 3D face landmarks from speech in an online fashion that is robust against unseen non-stationary background noise. The network predicts active shape model (ASM) coefficients of face landmarks from input speech. In one version of the system, we further added the predicted ASM coefficients in the previous frame to the network input to improve the smoothness of frame transitions. We conducted objective evaluations on landmark prediction errors and subjective evaluations on audio-visual coherence. Both objective and subjective evaluations showed that the proposed method statistically significantly improves over state-of-the-art baseline methods. Detailed analyses of network hyper-parameters were also provided to gain insights into the architecture design. To promote scientific reproducibility, we provided the research community with our pre-trained models, code and generation examples.

Acknowledgment

The authors would like to thank Voice Biometrics Group[3] for donating the noise files used in this study.

[3][Online]. Available: https://www.voicebiogroup.com

REFERENCES

[1] C. A. Binnie, "Bi-sensory articulation functions for normal hearing and sensorineural hearing loss patients," *J. Acad. Rehabilit. Audiol.*, vol. 6, no. 2, pp. 43–53, 1973.

[2] K. S. Helfer and R. L. Freyman, "The role of visual speech cues in reducing energetic and informational masking," *J. Acoust. Soc. Am.*, vol. 117, no. 2, pp. 842–849, 2005.

[3] J. G. Bernstein and K. W. Grant, "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3358–3372, 2009.

[4] R. K. Maddox, H. Atilgan, J. K. Bizley, and A. K. Lee, "Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners," *Elife*, vol. 4, 2015, Art. no. e04995.

[5] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 95.

[6] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" presented at the British Machine Vision Conf., 2017.

[7] H. X. Pham, Y. Wang, and V. Pavlovic, "End-to-end learning for 3D facial animation from raw waveforms of speech," in *Proc. 20th ACM Int. Conf. Multimodal Interaction*, New York, NY, USA, 2018, pp. 361–365.

[8] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2018, pp. 372–381.

[9] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[10] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 112-1–112-11, Jul. 2008.

[11] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *INTERSPEECH*, 2018.

[12] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2130–2134.

[13] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 2592–2596.

[14] M. S. Hossain, G. Muhammad, M. F. Alhamid, B. Song, and K. Al-Mutib, "Audio-visual emotion recognition using big data towards 5G," *Mobile Netw. Appl.*, vol. 21, no. 5, pp. 753–763, 2016.

[15] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 732–745, Apr. 2015.

[16] Z.-H. Ling *et al.*, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, May 2015.

[17] M. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, 2019, Art. no. 118.

[18] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2015, pp. 1–6.

[19] L. Wang, W. Han, F. K. Soong, and Q. Huo, "Text driven 3D photo-realistic talking head," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 3307–3308.

[20] V. Wan *et al.*, "Photo-realistic expressive text to talking head synthesis." in *Proc. Interspeech*, 2013, pp. 2667–2669.

[21] S. Cassidy *et al.*, "Expressive visual text-to-speech as an assistive technology for individuals with autism spectrum conditions," *Comput. Vision Image Understanding*, vol. 148, pp. 193–200, 2016.

[22] H. Tang, Y. Hu, Y. Fu, M. Hasegawa-Johnson, and T. S. Huang, "Real-time conversion from a single 2D face image to a 3D text-driven emotive audio-visual avatar," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2008, pp. 1205–1208.

[23] L. Xie, N. Sun, and B. Fan, "A statistical parametric approach to video-realistic text-driven talking avatar," *Multimedia Tools Appl.*, vol. 73, no. 1, pp. 377–396, 2014.

[24] B. Fan, L. Wang, F. K. Soong, and L. Xie, "Photo-real talking head with deep bidirectional LSTM," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4884–4888.

[25] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *Proc. Eur. Conf. Comput. Vision*, Sep. 2018.

[26] M. Brand, "Voice puppetry," in *Proc. 26th Annu. Conf. Comput. Graph. Interactive Techn.*, 1999, pp. 21–28.

[27] K. Choi, Y. Luo, and J.-N. Hwang, "Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system," *J. VLSI Signal Process. Syst. Signal, Image, Video Technol.*, vol. 29, pp. 51–61, 2001.

[28] D. Cosker, D. Marshall, P. Rosin, and Y. Hicks, "Video realistic talking heads using hierarchical non-linear speech-appearance models," *Proc. Mirage*, vol. 147, pp. 20–27, 2003.

[29] D. Cosker, D. Marshall, P. L. Rosin, and Y. Hicks, "Speech driven facial animation using a hidden Markov coarticulation model," in *Proc. 17th Int. Conf. Pattern Recognit.*, 2004, vol. 1, pp. 128–131.

[30] L. Xie and Z.-Q. Liu, "A coupled HMM approach to video-realistic speech animation," *Pattern Recognit.*, vol. 40, pp. 2325–2340, 2007.

[31] L. D. Terissi and J. C. Gómez, "Audio-to-visual conversion via HMM inversion for speech-driven facial animation," in *Proc. Brazilian Symp. Artif. Intell.*, 2008, pp. 33–42.

[32] X. Zhang, L. Wang, G. Li, F. Seide, and F. K. Soong, "A new language independent, photo-realistic talking head driven by voice only," in *Proc. Interspeech*, 2013, pp. 2743–2747.

[33] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[34] H. X. Pham, S. Cheung, and V. Pavlovic, "Speech-driven 3D facial animation with implicit emotional awareness: A deep learning approach," in *Proc. Comput. Vision Pattern Recognit Workshop*, 2017, pp. 2328–2336.

[35] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 94.

[36] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vision Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[37] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[38] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1021–1030.

[39] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[40] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-Their training and application," *Comput. Vision Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[41] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[42] 2018. [Online]. Available: http://www.sens.com/products/stevi-speech-test-video-corpus/

[43] 2018. [Online]. Available: https://www.sound-ideas.com/

[44] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *Proc. Neural Inf. Process. Syst. Workshop*, 2017.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, 2014.

Sefik Emre Eskimez (S'14–M'19) received the B.Sc. and M.Sc. degrees in mechatronics engineering from Sabanci University, Istanbul, Turkey, in 2011 and 2013, respectively, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Rochester, Rochester, NY, USA, in 2015 and 2019, respectively. He joined Microsoft Speech and Dialog Research Group in 2019. His research interests include generative models, speech processing, natural language processing, multi-modal learning, and deep learning.

Ross K. Maddox received the B.S. degree in sound engineering from the University of Michigan, Ann Arbor, MI, USA, in 2006, and the M.S. and Ph.D. degrees in biomedical engineering from Boston University, Boston, MA, USA, in 2009 and 2011, respectively. He completed his Postdoctoral training with the University of Washington Institute for Learning and Brain Sciences. He is currently an Assistant Professor with the Departments of Biomedical Engineering and Neuroscience, University of Rochester, Rochester, NY, USA. He is also a member of the Center for Visual Science and the Del Monte Institute for Neuroscience, University of Rochester. His research is focused on the neural processes that support listening in noisy environments, ranging from basic science to development of diagnostic tools. His lab combines behavioral studies, electroencephalography recordings of neural activity, and novel applications of signal processing techniques. Among the awards and honors, he has received the Pathway to Independence Award from the National Institutes of Health (K99/R00).

Chenliang Xu (M'11) received the B.S. degree in information and computing science from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2010, the M.S. degree in computer science from SUNY Buffalo, Buffalo, NY, USA, in 2012, and the Ph.D. degree in computer science from the University of Michigan, Ann Arbor, MI, USA, in 2016. He is currently an Assistant Professor with the Department of Computer Science, University of Rochester, Rochester, NY, USA. His research interests include computer vision and its relations to natural language, robotics, and data science. His work primarily focuses on problems in video understanding, such as video segmentation, activity recognition, and multimodal vision-and-x modeling. He was the recipient of NSF BIGDATA Award 2017, NSF CDS&E Award 2018, University of Rochester AR/VR Pilot Award, and the Best Paper Award as SMC 2017.

Zhiyao Duan (S'09–M'13) received the B.S. degree in automation and the M.S. degree in control science and engineering from Tsinghua University, Beijing, China, in 2004 and 2008, respectively, and the Ph.D. degree in computer science from Northwestern University, Evanston, IL, USA, in 2013. He is currently an Assistant Professor with the Electrical and Computer Engineering Department, University of Rochester, Rochester, NY, USA. His research interest is in the broad area of computer audition, i.e., designing computational systems that are capable of understanding sounds, including music, speech, and environmental sounds. He copresented a tutorial on Automatic Music Transcription at ISMIR 2015. He received a Best Paper Award at the 2017 Sound and Music Computing (SMC) conference, a Best Paper nomination at the 2017 International Society for Music Information Retrieval (ISMIR) conference, and a CAREER Award from the National Science Foundation.