Emotion Dependent Domain Adaptation for Speech Driven Affective Facial Feature Synthesis

Rizwan Sadiq and Engin Erzin

Abstract—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 paper, we improve the affective facial animation using domain adaptation by partially reducing the data scarcity. 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. The proposed models are assessed with objective and subjective evaluations. The proposed affective A2V system achieves significant MSE 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.

Index Terms—Speech driven facial animation, affective facial animation, domain adaptation, transfer learning.

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 [1], [2]. 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 [2], [3], [4], [5]. 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, [6] as exemplified in Figure 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.

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

For the A2V mapping, various approaches have been studied to generate naturalistic animations using different cues. In a pioneering study, the strong correlation between visual and speech features has been studied for A2V mapping [7]. A2V mapping is widely studied based on statistical methods including Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) and Dynamic Bayesian Networks (DBNs). In [8], 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 [9], [10]. 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 [11].

A GMM based statistical mapping between articulatory movements and speech signal has been studied in [12]. 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 [13]. A DBN based model has

- *All the authors are with the Electrical & Electronics Engineering and Computer Engineering Departments of the Koç University, Istanbul, Turkey.*

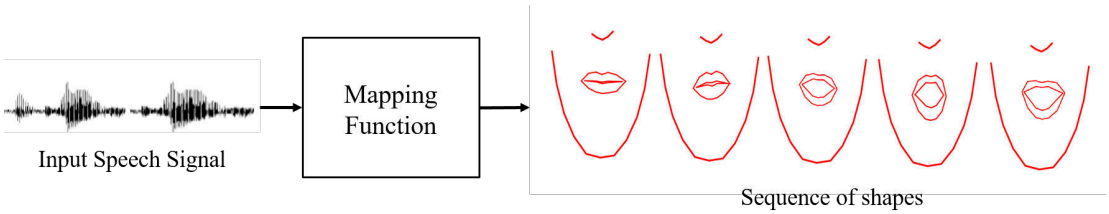


Fig. 1: Block diagram of a general audio-to-visual (A2V) mapping system

been studied to capture dependencies not only between speech and facial behaviors, but also between head and eyebrow motions [14]. 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 used in temporal clustering of speech signals [4], 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 [15], they have used for applications of the speech signal processing, such as speech recognition [16], [17], [18] and speech emotion recognition [19], [20], [21].

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 [22]. 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 [23]. In a recent study, generalization problem has been addressed by synthesizing facial animations solely from phoneme sequence as input [5]. 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 [9], [10], 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 [24]. 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 [25]. The affective LSTM model has been trained on the RAVDESS audio-visual dataset [26], 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 [27]. 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) [4]. 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 [28], SAVEE [29], RAVDESS [26] and BBC-LRW [30], 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 [31], [32]. 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 [33]. Figure 2 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 [34], [35]. Autoencoders have been widely used for domain adaptation. A study on emotional speech analysis successfully used adaptive denoising autoencoders for improved unsupervised domain adaptation where prior knowledge learned from a limited target set is used to regularize the training on a abundant source set [36]. In a more general framework, stacked denoising autoencoders (SDAs) have been successfully used to adapt new representations for feature augmentations [35]. 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 [37]. 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 on par with the traditional SDAs, which are attaining almost identical accuracy in benchmark tasks [38].

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 [5], [24], where neither 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 these regards, the main contribution of this paper is threefold: 1) A state-of-the-art domain adaptation scheme is integrated for the affective A2V mapping problem; 2) a new feature selection scheme is proposed to define a paired set of data across source and target domains namely parallel data, for the neutral and affective speech signals that makes feature representation transfer possible; and 3) the proposed A2V mapping is defined as a weighted combination of emotion dependent facial regression outputs.

2 METHODOLOGY

In this paper, we present an emotion dependent domain adaptation (EDDA) framework for the A2V mapping problem. Block diagram of the proposed EDDA framework is given in Figure 3. The EDDA framework consists of a learning phase and a synthesis phase. In the learning phase, emotion dependent A2V mapping models are extracted by the marginalized stacked denoising autoencoders (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 sub-sections.

2.1 Feature Representations

2.1.1 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 4 presents sample spectrograms for different emotion settings.

We use the Mel-Frequency Spectral Coefficient (MFSC) representation as the speech signal features [17]. 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 [17], [39], [40]. 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. 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.

2.1.2 Visual Feature Extraction

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 [41]. 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

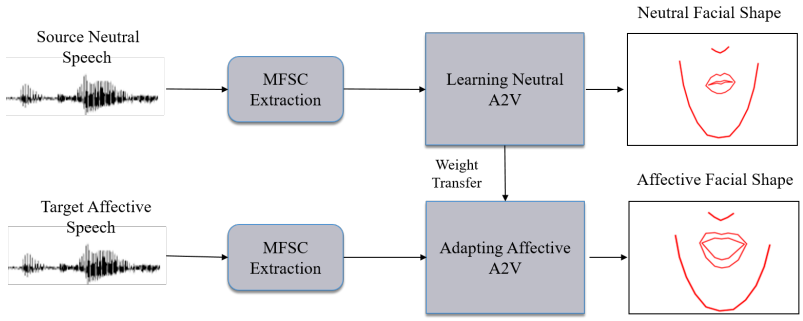


Fig. 2: Model adaptation (MA) for the affective A2V mapping

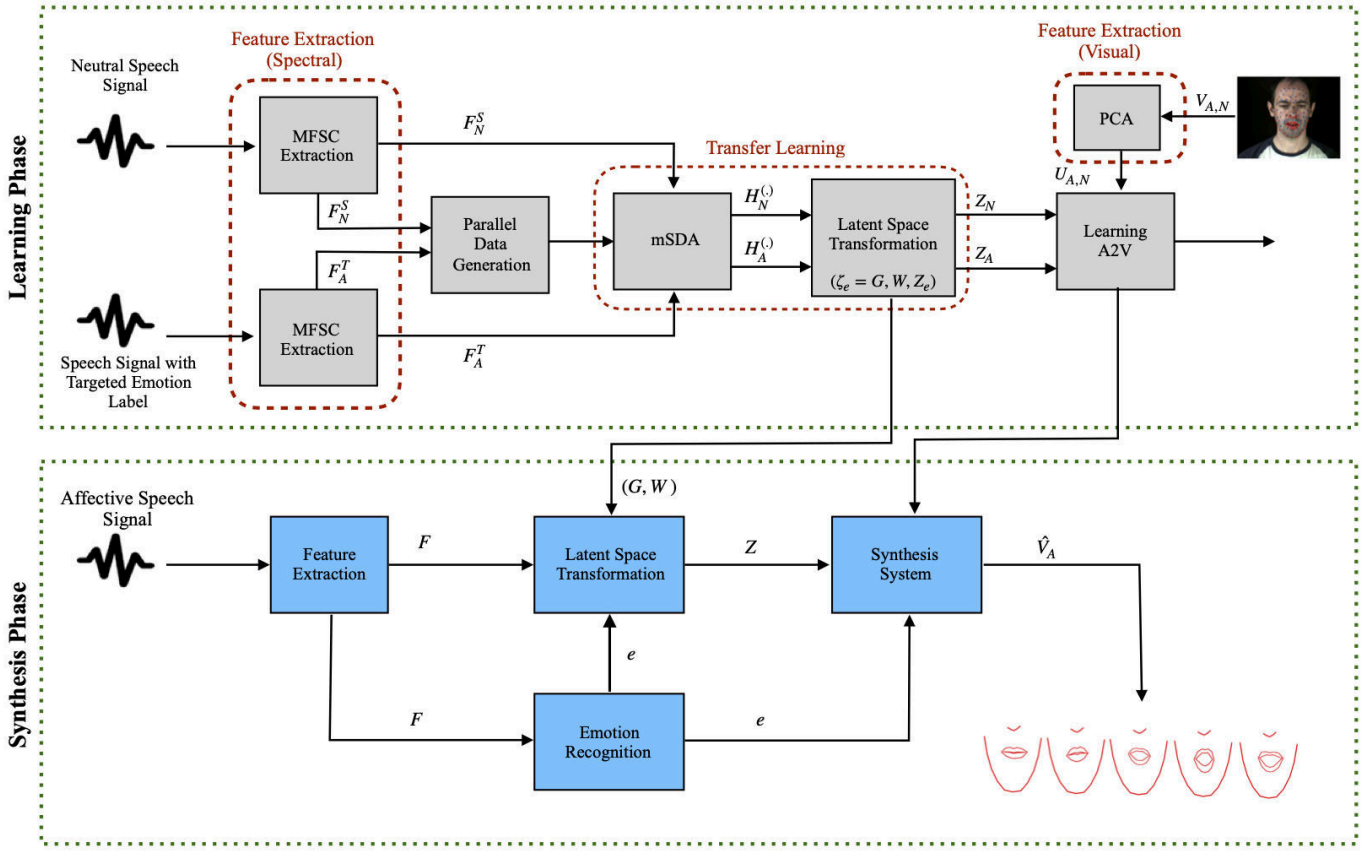


Fig. 3: The proposed emotion dependent domain adaptation (EDDA) framework for the A2V mapping

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

$$v_j = [x_1, y_1, x_2, y_2, ..., x_D, y_D]^\intercal, \quad (1)$$

where x and y are the indexed coordinates of the $D = 36$ facial landmark points. Figure 5 presents a sample set of extracted landmark points using the Dlib detector.

We used Active Shape Model (ASM) to remove the correlation in the facial landmark representation and for dimensionality reduction [42]. 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 [43]. 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,

$$u_j = PCA(GPA(v_j)) \in \mathbb{R}^N, \quad (2)$$

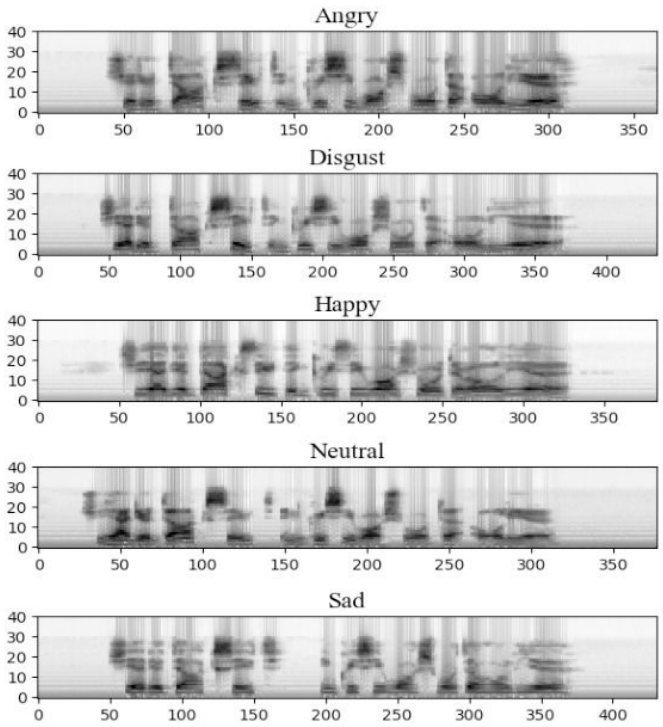


Fig. 4: Spectrograms of the sentence "she had your dark suit in greasy wash water all year" from the SAVEE data set with five different emotions.

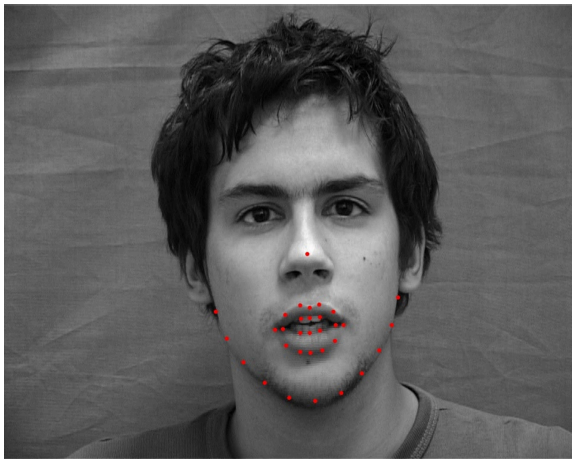


Fig. 5: A sample set of facial landmark points using the Dlib detector

where the feature dimension is reduced to $N = 18$ by capturing 98% of the variation in the dataset. Figure 6 presents the three main modes of variation around the mean facial shape, varying the model parameters between ± 3 of their standard deviation.

2.2 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. 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,

$$C(\phi_i) = \{\phi_{i-2}, \phi_{i-1}, \phi_i, \phi_{i+1}, \phi_{i+2}\}, \tag{3}$$

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

$$\ddot{f}_{\phi_i} = \{f_{n_i+1}, f_{n_i+2}, \ldots, f_{n_i+d_i}\}. \tag{4}$$

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

$$D_p = \bigcup_{\substack{\forall \phi_i, \phi_j \\ \text{s.t. } C(\phi_i)=C(\phi_j)}} (\ddot{f}^S_{\phi_i}, \ddot{f}^T_{\phi_j}), \tag{5}$$

where $\ddot{f}^S_{\phi_i}$ and $\ddot{f}^T_{\phi_j}$ are representing phoneme feature sequences from the source and target domains respectively. Furthermore, the matching phoneme-context requires equality of all five phonemes in the $C(\phi_i)$ and $C(\phi_j)$. Note that duration of the feature sequences from the source and target domains may differ. In order to construct a parallel association between the source and target domain

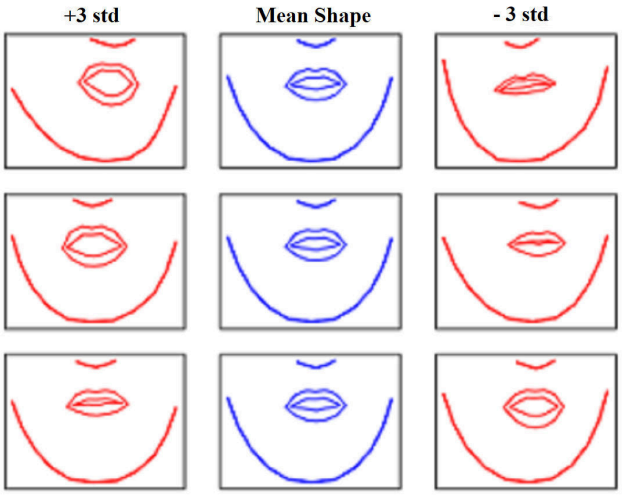


Fig. 6: Three modes of variation around mean facial shape

feature sequences, a dynamic time warping algorithm [44] is executed across the phoneme feature sequences $\ddot{f}^S_{\phi_i}$ and $\ddot{f}^T_{\phi_j}$. A graphical representation of the phoneme-context matching and feature sequence association are given in Figure 7.

2.3 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 [38]. 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, $\mathbf{D}_T = \{f^T_1, f^T_2, \ldots\}$, and the neutral speech representation as the source domain, $\mathbf{D}_S = \{f^S_1, f^S_2, \ldots\}$, where f^T_i and f^S_j 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 2.2. Note that the parallel data, $\mathbf{D}_p = \{\mathbf{D}^p_S, \mathbf{D}^p_T\}$, is a collection of source, $\mathbf{D}^p_S$, and target, $\mathbf{D}^p_T$, 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,

$$\mathcal{L} = \sum_{i=1}^{R} \left\| \mathbf{D}_T - \mathbf{W}_T \, \mathbf{D}^{(i)}_T \right\|^2, \tag{6}$$

where $\mathbf{D}^{(i)}_T$ 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

$$\mathbf{W}_T = \mathbf{P}_T \mathbf{Q}^{-1}_T \tag{7}$$

with $\mathbf{Q}_T = \widetilde{\mathbf{D}}_T \widetilde{\mathbf{D}}^\intercal_T$ and $\mathbf{P}_T = \overline{\mathbf{D}}_T \widetilde{\mathbf{D}}^\intercal_T$, 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 $\mathbf{H}^p_T$ and $\mathbf{H}^p_S$ 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,

$$\left\| \mathbf{H}^p_{S,k} - \mathbf{G}_k \mathbf{H}^p_{T,k} \right\|^2 + \lambda \left\| \mathbf{G}_k \right\|^2, \tag{8}$$

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,

$$\mathbf{G}_k = \left[\mathbf{H}^p_{S,k} \left(\mathbf{H}^p_{T,k} \right)^\intercal \right] \left[\mathbf{H}^p_{T,k} \left(\mathbf{H}^p_{T,k} \right)^\intercal + \lambda \mathbf{I} \right]^{-1}, \tag{9}$$

where $\mathbf{I} \in \mathbb{R}^{M \mathtt{x} M}$ 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
$\mathbf{Z}_T = [(\mathbf{G}_1 \, \mathbf{H}_{T,1})^\intercal, (\mathbf{G}_2 \, \mathbf{H}_{T,2})^\intercal, \ldots, (\mathbf{G}_K \, \mathbf{H}_{T,K})^\intercal]^\intercal$
$\mathbf{Z}_S = [(\mathbf{H}_{S,1})^\intercal, (\mathbf{H}_{S,2})^\intercal, \ldots, (\mathbf{H}_{S,K})^\intercal]^\intercal$

2.4 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 4, 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

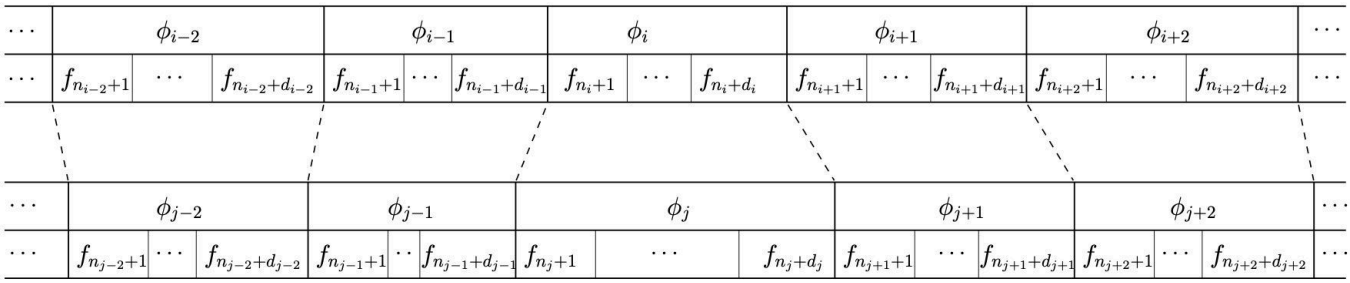


Fig. 7: The proposed emotion dependent domain adaptation (EDDA) framework for the A2V mapping

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 1 presents detailed description of the SER-CNN.

TABLE 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

$$p_i = \frac{1}{L}\sum_{j=1}^{L} 1(\tilde{e}_j = e^i) \;\text{ for } i = 1, \ldots, 7, \tag{10}$$

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

$$\alpha = \underset{1\le i\le 7}{\arg\max}\, p_i \;\text{ and }\; \beta = \underset{\{1\le i\le 7\}-\{\alpha\}}{\arg\max}\, p_i. \tag{11}$$

We choose to use 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

$$\tilde{p}_\beta = \begin{cases} 0 & \text{if } p_\alpha > 0.65 \\ p_\beta & \text{otherwise.} \end{cases} \tag{12}$$

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

$$\bar{p}_\alpha = \frac{p_\alpha}{p_\alpha + \tilde{p}_\beta} \;\text{ and }\; \bar{p}_\beta = \frac{\tilde{p}_\beta}{p_\alpha + \tilde{p}_\beta}, \tag{13}$$

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

2.5 Emotion Dependent A2V Mapping

We construct a deep A2V mapping network (DA2V) to generate visual lower face representations from the speech spectral features. Table 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 2.3, $z_j \in \mathbf{Z}$. In Table 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.

In the proposed emotion dependent A2V mapping framework, the DA2V networks are trained 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

$$\tilde{U}_j = \hat{U}_j^\alpha, \tag{14}$$

where $\hat{U}_j^\alpha$ is the output of the network for the top matching

TABLE 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

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

$$\tilde{U}_j = \bar{p}_{\alpha}\hat{U}_j^{\alpha} + \bar{p}_{\beta}\hat{U}_j^{\beta}, \tag{15}$$

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.

2.6 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

$$\hat{v}_j = \frac{1}{3}(\tilde{v}_{j+1}^{j-1} + \tilde{v}_j^{j} + \tilde{v}_{j-1}^{j+1}). \tag{16}$$

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

3 Experimental Evaluations

In this section, we first introduce the affective audio-visual dataset in use, baseline models to be compared with the proposed approach and the experimental details. Later, objective and subjective evaluations of the proposed system are given with their discussion.

3.1 Affective Audio-Visual Dataset

In this study, we use the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset to evaluate the affective A2V models [28]. 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 [45]. 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 2.2 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.

3.2 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 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 lower bound for the emotion dependent models, we construct the emotion dependent models with hard decision outputs using the true emotion class labels (T-HED-*).

3.3 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 8. We observe linear increases in the time-to-train per epoch and the number of model parameters with 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

TABLE 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.

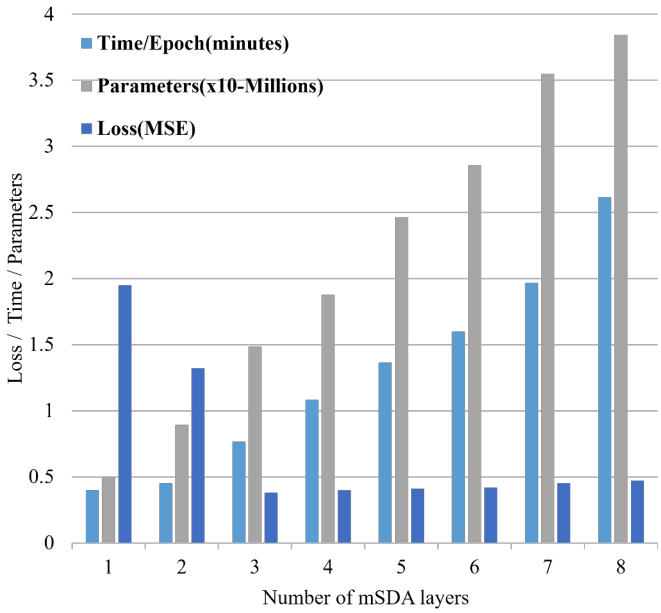


Fig. 8: Number of mSDA layers vs performance in terms of MSE loss, time-to-train per epoch, and model complexity in terms of total number of parameters

distribution with mean zero and standard deviation 0.05. We use Adam optimizer [46] 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 [47] backend on a NVIDIA TITAN XP GPU.

3.4 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.

3.4.1 SER results

In section 2.4, 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 rate for the top emotion class is defined as

$$ERR_t = \frac{1}{N_U} \sum_j 1(e_j^{\alpha} = e_j) \tag{17}$$

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 (10). Similarly, emotion recognition rate for the top two emotion classes can be defined as

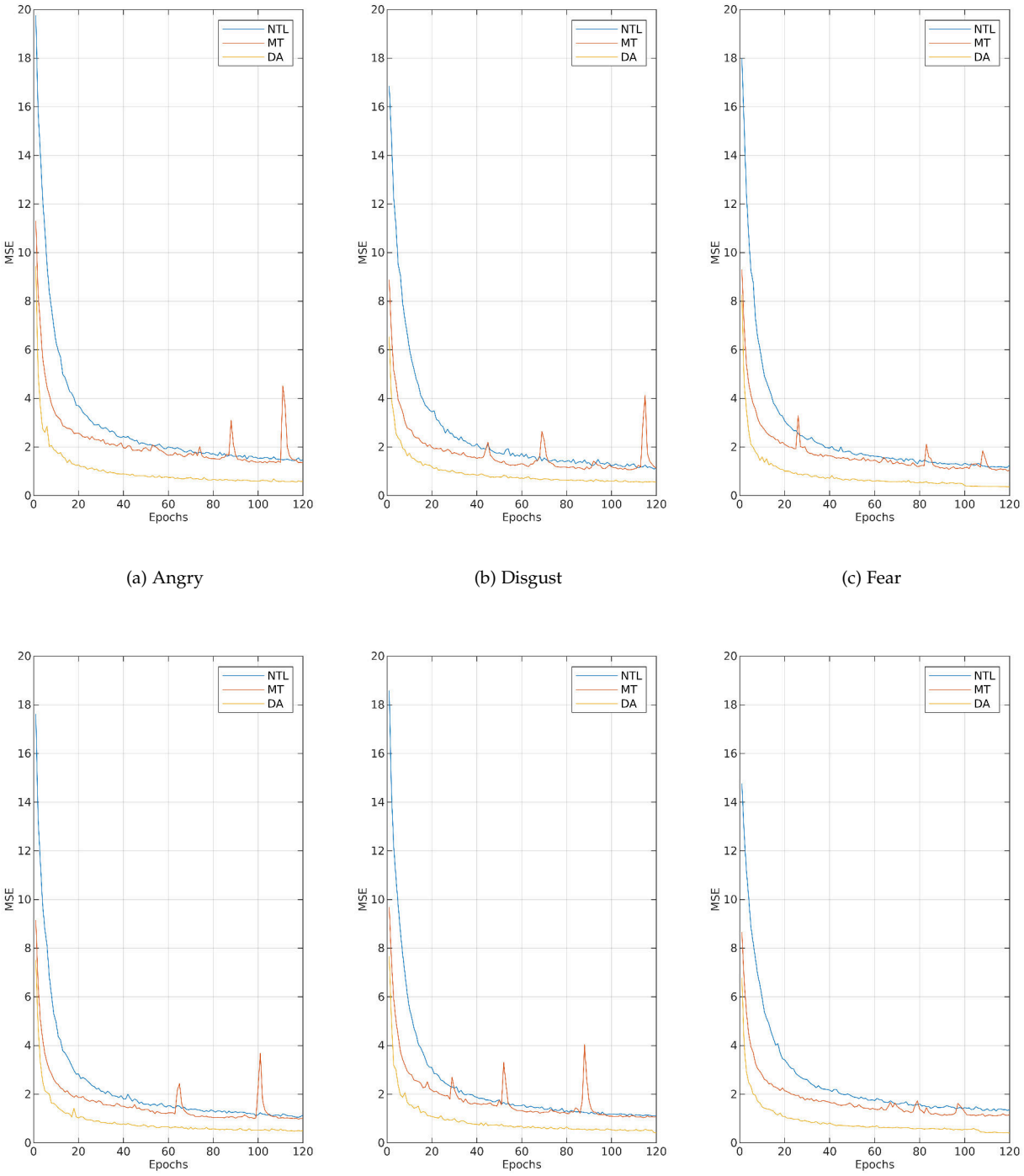
$$ERR_{t2} = \frac{1}{N_U} \sum_j \left[1(e_j^{\alpha} = e_j) + 1(e_j^{\beta} = e_j)\right]. \tag{18}$$

Utterance level emotion recognition rates are reported for each emotion category in Table 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 2.5 can be expected.

TABLE 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

1. https://keras.io/



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

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

Fig. 9: MSE loss curves of the DA2V networks during the learning process for the six emotion categories

3.4.2 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 9 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.

Table 5 provides the MSE loss values over the test set

TABLE 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. [5]	12.90
Asadiabadi et al. [24]	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**

of all the models described in Table 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 [5] 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 and they exhibit lower 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.

3.5 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. Each test session for a participant contains 28 clips for all 4 methods, 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].

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

TABLE 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

Table 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 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 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.

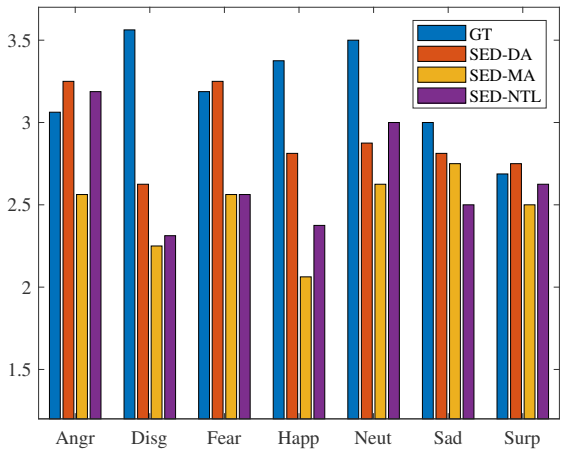


Fig. 10: Mean opinion score (MOS) subjective test results of the lower-face animations of the SED-NTL, SED-MA, SED-DA and ground truth (GT) for each emotion category

TABLE 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 10. 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.

4 Conclusion

We 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. 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.

Acknowledgments

This work was supported in part by the Scientific and Technological Research Council of Turkey (TÜBITAK) under grant number 217E107.

References

[1] T. K. Capin, H. Noser, D. Thalmann, I. Sunday Pandzic, and N. M. Thalmann, "Virtual human representation and communication in vlnet," *IEEE Computer Graphics and Applications*, vol. 17, no. 2, pp. 42–53, 1997.

[2] E. I. Konstantinidis, M. Hitoglou-Antoniadou, A. Luneski, P. D. Bamidis, and M. M. Nikolaidou, "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*, 2009, pp. 1–6.

[3] A. Gaggioli, F. Mantovani, G. Castelnuovo, B. Wiederhold, and G. Riva, "Avatars in clinical psychology: a framework for the clinical use of virtual humans," *Cyberpsychology & behavior*, vol. 6, no. 2, pp. 117–125, 2003.

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

[5] S. Taylor and et all, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 93, 2017.

[6] S. Taylor, A. Kato, I. A. Matthews, and B. P. Milner, "Audio-to-visual speech conversion using deep neural networks." in *Interspeech*, 2016, pp. 1482–1486.

[7] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[8] E. Yamamoto, S. Nakamura, and K. Shikano, "Speech to lip movement synthesis by HMM," in *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*, 1997.

[9] E. Bozkurt, C. E. Erdem, E. Erzin, T. Erdem, and M. Ozkan, "Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation," in *2007 3DTV Conference*. IEEE, 2007, pp. 1–4.

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

[11] L. Xie and Z.-Q. Liu, "A coupled hmm approach to video-realistic speech animation," *Pattern Recognition*, vol. 40, no. 8, pp. 2325–2340, 2007.

[12] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[13] C. Luo, J. Yu, X. Li, and Z. Wang, "Realtime speech-driven facial animation using gaussian mixture models," in *IEEE International Conference on Multimedia and Expo Workshops, ICMEW*. IEEE, 2014, pp. 1–6.

[14] S. Mariooryad and C. Busso, "Generating human-like behaviors using joint, speech-driven models for conversational agents," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2329–2340, 2012.

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

[16] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[17] O. Abdel-Hamid and et al, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[18] W. Song and J. Cai, "End-to-end deep neural network for automatic speech recognition," 2015.

[19] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 801–804.

[20] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "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)*. IEEE, 2016, pp. 5200–5204.

[21] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[22] S. Suwajanakorn, S. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 95, 2017.

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

[24] S. Asadiabadi, R. Sadiq, and E. Erzin, "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)*. IEEE, 2018, pp. 1508–1512.

[25] H. X. Pham, S. Cheung, and V. Pavlovic, "Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach." in *CVPR Workshops*, 2017, pp. 2328–2336.

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

[27] 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 Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 94, 2017.

[28] 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*, vol. 120, no. 5, pp. 2421–2424, 2006.

[29] S. Haq and P. Jackson, "Multimodal emotion recognition," in *Machine audition: principles, algorithms and systems*. IGI Global, 2011, pp. 398–423.
[30] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.
[31] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
[32] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.
[33] E. V. Bonilla, K. M. Chai, and C. Williams, "Multi-task gaussian process prediction," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 153–160.
[34] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 120–128.
[35] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 513–520.
[36] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.
[37] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," *arXiv preprint arXiv:1206.4683*, 2012.
[38] J. T. Zhou, S. J. Pan, I. W. Tsang, and Y. Yan, "Hybrid heterogeneous transfer learning through deep learning," in *Twenty-eighth AAAI Conference on Artificial Intelligence*, 2014.
[39] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.
[40] L. Tóth, "Phone recognition with hierarchical convolutional deep maxout networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–13, 2015.
[41] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.
[42] T. F. Cootes and C. J. Taylor, "Active shape model search using local grey-level models: A quantitative evaluation," in *BMVC*, 1993.
[43] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society B*, vol. 53, no. 2, 1991.
[44] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, pp. 561–580, 2007.
[45] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.
[46] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[47] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A System for Large-scale Machine Learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. Berkeley, CA, USA: USENIX Association, 2016, pp. 265–283.



Rizwan Sadiq (S'17) received his Bechlors in Computer Engineering from COMSATS University Abbottabad, Pakistan, in 2007. From 2007 to 2012, He worked as lab engineer at International Islamic University, Islamabd, Pakistan. In 2011 he obtained his Masters in Electronic Engineering from International Islamic university Islamabad, Pakistan. From 2012 to 2013 he taught as a lecturer at the Department of Electrical engineering at COMSATS University, Islamabad, Abbottabad campus, Pakistan. He is currently pursuing Ph.D. degree in the Electrical Engineering Department of Koc University . His research interest includes computer vision, emotion recognition, machine learning for emotional speech and videos, domain adaptations and transfer learning.



Engin Erzin (S'88-M'96-SM'06) received his Ph.D. degree, M.Sc. degree, and B.Sc. degree from the Bilkent University, Ankara, Turkey, in 1995, 1992 and 1990, respectively, all in Electrical Engineering. During 1995-1996, he was a postdoctoral fellow in Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies in September 1996, and he was with the Consumer Products for one year as a Member of Technical Staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he has been with the Electrical & Electronics Engineering and Computer Engineering Departments of Koc University, Istanbul, Turkey. Engin Erzin is currently a member of the IEEE Speech and Language Processing Technical Committee and Associate Editor for the IEEE Transactions on Multimedia, having previously served as Associate Editor of the IEEE Transactions on Audio, Speech & Language Processing (2010-2014). His research interests include speech-audio-visual signal processing, affective computing, human-computer interaction and machine learning.

View publication stats