2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

Expressive Speech-Driven Lip Movements with Multitask Learning

Najmeh Sadoughi and Carlos Busso
Multimodal Signal Processing (MSP) Laboratory, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
nxs137130@utdallas.edu, busso@utdallas.edu

***Abstract*—The orofacial area conveys a range of information, including speech articulation and emotions. These two factors add constraints to the facial movements, creating non-trivial integrations and interplays. To generate more expressive and naturalistic movements for *conversational agents* (CAs) the relationship between these factors should be carefully modeled. Data-driven models are more appropriate for this task than rule-based systems. This paper provides two deep learning speech-driven structures to integrate speech articulation and emotional cues. The proposed approaches rely on *multitask learning* (MTL) strategies, where related secondary tasks are jointly solved when synthesizing orofacial movements. In particular, we evaluate emotion recognition and viseme recognition as secondary tasks. The approach creates shared representations that generate behaviors that not only are closer to the original orofacial movements, but also are perceived more natural than the results from single task learning.**

***Keywords*-Lip movement driven by speech, expressive lip movements, multitask learning.**

I. Introduction

The orofacial area plays an important role in human interaction, conveying traits associated with personality, gender, and emotion. The information is conveyed at various temporal resolutions creating nontrivial interplays between the communicative goals. In addition, the orofacial area conveys speech articulation information [1], which improves speech intelligibility [2]. Listeners unconsciously decode the information, inferring the verbal and nonverbal cues conveyed by the speaker. This complex process has to be considered to create more believable facial movements for *conversational agents* (CAs).

Several studies have relied on storing predefined parameters for lip movements associated with articulatory units [3]–[5] (e.g., phonemes). A natural extension for expressive speech is to use a predefined set of lip parameters per emotion and per articulation unit (e.g., each phoneme in angry emotion). However, discretizing the expressive lip configurations into a reduced set of emotional categories may result in very caricature-like expressions, overemphasizing the emotional state. Using the same set of parameters may also result in repetitive movements which may affect the naturalness perception of the animation. Furthermore, these models will not be appropriate for subtle or ambiguous emotions, which are common during human interactions. Speech driven systems are an alternative approach to handle these issues, providing a principled method to incorporate expressive speech [6], [7].

The lower face region is affected with articulatory and emotional cues [1], [8]. Although verbal content is the primary channel while speaking, emotional content will modulate the speech articulation process creating expressive speech. Although there are differences between the time resolutions of emotional cues and articulatory cues, the relationship between speech and emotional orofacial movements can be exploited toward automating the process of generating more expressive and naturalistic lip movements from speech. The advantage of using features that are related to the emotional content of the message is that it will automatically broaden the emotional spectrum conveyed by the CA, which is not possible with rules for discrete emotional categories.

This paper provides deep learning solutions to integrate the articulatory and emotional features from the input speech. The intrinsic relationship between speech articulation and emotional content is directly exploited in the model by performing *multitask learning* (MTL). The primary task of predicting lip movements is complemented with two secondary tasks: viseme recognition, and emotion recognition. This model creates shared representations across these related tasks, increasing the robustness and accuracy of the regression task. These secondary tasks are relevant to variations in the orofacial movements. Emotion modulation is different under different phonetic units (e.g., pronouncing /b/ and /a/ with a smile). Therefore, features extracted for viseme recognition and emotion recognition can help in learning orofacial movements. Objective and subjective evaluations of the predicted lip movements demonstrate the advantages of adding these auxiliary tasks to the model (i.e. multitask learning).

II. Related Works

Approaches to generate lip movements can be categorized into unit selection [3], [5], [7], generative model [6], [9], [10], and discriminative model [11].

The unit selection strategy relies on defining units of articulations. For expressive lip motion, the units also consider emotional articulations. One of the early works in this area was conducted by Cao et al. [7]. They defined anime nodes comprising phonemes, emotions, speech features and motion capture features to index their training data. During testing, the input speech is segmented based on its phonetic content, searching the database for the most suitable segments sharing similar emotions. The search minimized the differences between the speech features, while penalizing non-smooth movements. The selected segments are then time warped and concatenated to generate the final facial trajectory. Deng et al. [5] proposed a method to model speech coarticulation. They defined canonical shapes for the lips which were blended with different weights to generate arbitrary configurations. They used motion capture recordings of an actor while speaking, finding the weights associated with each canonical shape during different diphones, and triphones. Xu et al. [3] proposed a similar approach, where they defined canonical shapes, which were combined using artist-produced weights. These approaches require phoneme alignment

0-7695-6381-3/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00066

of the testing utterances. To incorporate expressive lip movements, the weights of the canonical shapes need to be defined for each of the target emotions.

The second approach uses generative models to implicitly model speech coarticulation. Choi et al. [9] proposed to use *hidden Markov models* (HMM) inversion to predict lip movements from speech. Speech and visual features are jointly modeled for each phoneme with three state HMMs. During testing, they relied on the Baum-Welch algorithm to estimate the visual parameters. Xie and Liu [10] proposed to use *coupled HMMs* (CHMMs) to explicitly model the differences and dependencies between audio and visual streams, such as their asynchrony, different number of classes, and their temporal coupling. They use the Baum-Welch algorithm to find the maximum likelihood estimates of the visual features. Anderson et al. [6] proposed to use an extension of HMMs called *cluster adaptive training* (CAT) to model expressive audiovisual speech. In their system, CAT models emotion dependent information by using decision trees to cluster the data, and learning the weights associated with each cluster for the target emotions from the data.

The third approach uses discriminative models which directly learn the mapping between the input audio and lip movements eliminating the need for their joint modeling. For instance, Fan et al. [11] proposed a deep structure with *bidirectional long short-term memory* (BLSTM) units to learn the mapping between triphonemes (i.e., the previous, current, and next phones) and lip movements. Objective and subjective evaluations showed better results for their approach compared with an HMM-based approach. Taylor et al. [12] proposed to use a feedforward neural network, and a sliding window to generate lip movements using *Mel-frequency cepstral coefficients* (MFCCs). Objective and subjective evaluations showed improvements when using their approach compared with the HMM inversion approach proposed by Choi et al. [9]. Li et al. [13] evaluated several structures to generate emotional facial movements from speech using a small emotional corpus. Their best result was achieved with a two-model approach. The first model learns the mapping between neutral speech and facial movements. This model is evaluated with emotional speech. The predictions of this model are concatenated with speech features as the input of the second model that learns emotional facial movements. The approach of Li et al. [13] assumes that emotional labels are available during testing. However, our approach does not have this requirement, since it directly infers this information from speech. Sadoughi and Busso [14] investigated separate versus joint modeling of facial regions using models built with BLSTMs. They used MTL to explore the dependencies across facial features, where the secondary tasks corresponded to facial movements in other facial regions. They achieved better objective performance by jointly modeling the three facial regions. While our proposed approach also relies on MTL, our goal is radically different, creating an important distinction between the studies. Sadoughi and Busso [14] explored the local dependencies across facial features. Our study uses MTL to improve lip-motion by predicting visemes and emotions. Both secondary tasks are crucial for lip appearance. By choosing these secondary tasks, we can create more realistic lip motion sequences.

This paper proposes speech-driven models for expressive lip movements. The model relies on BLSTMs, trained with segment level spectral features (MFCCs), and sentence-level statistics of spectral and prosodic features, which are commonly used in speech emotion recognition. We propose novel structures for adding the emotional features, in a principled manner, leveraging the advances in MTL with deep learning. By considering emotion recognition and viseme recognition as secondary tasks, we incorporate nuances that are important for the face appearance while speaking. This formulation is truly novel, compared with previous studies on this area. To the best of our knowledge, this is the first MTL study focusing on lip-motion generation, which is the key novelty of this study. MTL prevents over-fitting the models by providing data-driven regularization during training. Regularization becomes particularly important in learning expressive facial movements, as the data size for emotional audiovisual corpora is currently limited.

III. Resources

A. Corpus

This study uses the *interactive emotional motion capture* (IEMOCAP) corpus [15]. The IEMOCAP corpus includes audio, video, and motion capture recordings of dyadic interactions between ten actors. The database is recorded with spontaneous and script-based scenarios designed to elicit renditions of different emotions. All the speaking turns are annotated by three annotators in terms of ten emotional categories: anger, disgust, excited, fear, frustrated, happiness, neutral, sadness, surprised, and other. Similar to previous studies on this corpus, we merge the classes excited and happiness [16], [17]. The consensus label per speaking turn is estimated using the majority vote rule. Lip movements conveying emotions are very subject dependent, so limiting the study to a single speaker reduces the need to compensate for idiosyncratic differences. Therefore, this study only considers the recordings from the first female subject (418 speaking turns), where the distribution of emotions is 66 (neutral), 72 (anger), 57 (happiness), 42 (sadness), 0 (disgust), 3 (fear), 68 (frustrated), 4 (surprised), and 0 (other). There are 106 sentences without consensus label. The classes with few samples (i.e., disgust, fear, surprise and other), and sentences without consensus are grouped together. The IEMOCAP corpus also includes the transcripts and phoneme alignments of the recordings. We group the phonemes into 14 viseme categories using the mapping proposed by Lucey et al. [18] listed in Table I. More information about this corpus is given in Busso et al. [15].

B. Audio and Visual Features

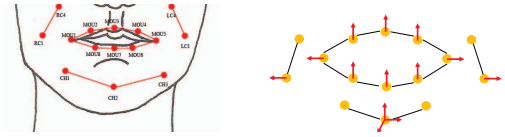
We extract features relevant to speech and emotion production in our speech-driven models. The speech production features are MFCCs. We extract 25 MFCCs, following the results shown by Taylor et al. [12], which demonstrated that 25 MFCCs gives the best result in predicting lip movements. We use Praat with 25ms windows shifted by 8.33ms to get the same frames per second as the motion capture data (i.e., 120).

For emotional features, we use the *extended Geneva minimalistic acoustic parameter set* (eGeMAPS) [19], which was carefully selected as a reduced set for paralinguistic problems. This feature set comprises 88 statistics extracted over 23 *low-level descriptors* (LLDs). We extract these features over each speaking turn, following common practice in emotion recognition (note that emotional labels are also assigned per speaking turn).

Table I
THE PHONEME TO VISEME MAPPING. COUNTS CORRESPOND TO THE FREQUENCY OF OCCURRENCES OF THE VISEMES FOR THE FIRST FEMALE SUBJECT IN THE IEMOCAP CORPUS.

Phoneme	Viseme	Count
B M P	P	13531
F V	F	6225
T TD TH D DD DX DH TS S Z	T	42607
CH SH ZH JH	CH	3202
EY EH AE AW	EY	15553
ER	ER	1163
SIL	SIL	64865

Phoneme	Viseme	Count
IY IH	IY	16984
W R F	W	14144
AO OY IX OW	X	9392
UH UW AXR	UH	10064
AH AY AX	AH	26317
AA	AA	4063
K G N L HH Y NG KD	K	42180



(a) Markers (b) Mapping to Xface

Figure 1. The location of the 15 markers from the IEMOCAP recordings used in this study, and their mapping to FAPs in Xface.

From the motion capture recordings, we focus on 15 markers in the orofacial area, shown in Figure 1(a). We use the (X, Y, Z) values for each marker resulting in a $45D$ feature vector, which our models aim to predict. We z-normalize all the features across the entire data.

C. Data Augmentation

Since we are using only 418 speaking turns for training the models, during training we augment our samples. We consider a sliding window of fixed length, w, for training the models. We shift the window by Δ, creating overlapped windows (see Fig. 2). This process generates new samples for training. We set $\Delta = 4$ frames for training the models.

D. Objective Metrics

The focus of this study is on predicting the location of markers in the orofacial area, so we formulate the task as a regression problem. Previous studies have mostly relied on minimizing the *mean squared error* (MSE) between the predictions and true values [12], [13], [20], or maximizing ρ_c, the *concordance correlation* (CCC) [14], [21]. The definition of ρ_c between two inputs x and y is given in Equation 1, where ρ is the Pearson correlation between x and y, σ_x and σ_y are the standard deviations of x and y, and μ_x, and μ_y are the means of x and y. Since the goal is to maximize ρ_c, we use the loss function $\ell = 1 - \rho_c$ and the goal during training is to minimize ℓ. This loss function has resulted in higher

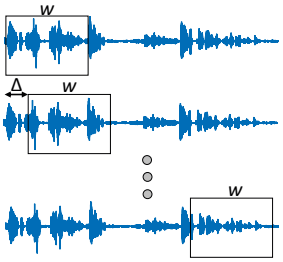


Figure 2. Using overlapped windows to augment the training set.

ρ_c compared to using a loss function based on MSE [21]. Moreover, our preliminary evaluation showed that using a loss function based on ρ_c increases the range of movements for the lips, resulting in more appealing animations. Therefore, our loss function is set to $1 - \rho_c$. However, we report the performances of the models in terms of MSE and ρ_c after concatenating all the predictions.

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

E. Rendering the Animations

We use Xface [22] for rendering the lip movements. Xface animates the face by using *facial action parameters* (FAPs) which are defined over *facial points* (FPs). The FPs in Xface follow the MPEG4 standard. Most of the markers in the IEMOCAP corpus also follow the MPEG4 standard, so finding a mapping is possible. We follow the mapping process proposed by Marioryad and Busso [23]. First, we find the position of the markers for the neutral pose of the actor, and map that to the neutral pose of the face in Xface in term of FAPs. Then, the range of movements of the actor is scaled to the range of movements of FAPs allowed by Xface. Figure 1(b) illustrates the positions of the markers in the orofacial area that are mapped into FAPs.

IV. PROPOSED APPROACH

This section describes the methods used to generate expressive lip movements. We propose two multitask learning structures to integrate the spectral and eGeMAPS features, where our goal is to learn the best shared representation that optimizes the prediction of facial movements. These models are built with BLSTMs, which are suitable to capture the temporal dependencies between speech and movements on the orofacial area.

A. Bidirectional Long-short Term Memory

Recurrent neural networks (RNNs) use temporal connections between consecutive frames for the hidden layers to encode the temporal dependencies in the time continuous signals. However, the RNNs suffer from the problem of vanishing or exploding gradients during training [24]. Hence, extensions of RNNs such as LSTMs are proposed to address this issue [24]. LSTMs utilize cell units to selectively keep track of past content. LSTMs use three gating mechanisms: input, output, and forget gates. The input gate determines the amount of input content being used to update the cell, the forget gate determines the amount of previous content being retained in the cell, and the output gate determines how much of the cell content to be used as the output of the hidden layer.

The LSTMs used in this study are implemented with Keras [25] using Theano as backend.

Bidirectional LSTMs utilize the future frames as well as the previous ones. They use forward and backward paths. At each time frame, the hidden units of the forward and backward paths are concatenated resulting in twice the number of hidden states for the layer. Our proposed frameworks use BLSTMs trained and tested over a fixed window length (Sec. III-C).

B. Multitask Learning for Lip Synthesis

Our approach relies on *multitask learning* (MTL). MTL jointly solves related problems creating shared representations for the tasks. We consider the prediction of orafacial movements as the primary task and the prediction of the triviseme and emotion as the auxiliary tasks. Triviseme recognition consists of three separate problems where we predict the previous, current and next visemes (i.e., three $\times$ 14-class problems). Emotion recognition is a six-class problem that consists of predicting the emotion associated with the speaking turn (anger, happiness, sadness, frustration, neutral, and other). Since the orofacial area is affected by both emotion and speech articulation, we hypothesize that adding these auxiliary tasks helps the network to learn more predictive features to synthesize orofacial movements. From a machine learning perspective, the auxiliary tasks can be considered as regularizes for the network to learn more robust features that generalize better for unseen data.

We build our models by stacking multiple layers including ReLUs and BLSTMs (the specific structures are described later). ReLU is a nonlinear function which facilitates better flow of information, and avoids over-fitting by sparse representation. ReLUs have been successfully used in previous studies for lip movement synthesis [11], [12]. The output of the primary task (i.e., orofacial movement) corresponds to a linear layer, since the goal is to predict continuous variables. The loss function for the primary task is the function ℓ described in Section III-D. For the recognition tasks, the output is a softmax layer of length n, where n is the number of categories (i.e. $n = 6$ for emotions, $n = 14$ for visemes). The objective function is the categorical cross entropy given in Equation 2, where p is the true distribution and q is the predicted distribution.

$$H(p,q) = \sum_{i=1}^{n} p(x_i)\log(q(x_i)) \tag{2}$$

The objective functions for the primary and auxiliary tasks are combined with their corresponding weights to form the loss function. Equation 3 gives the loss function: x_i is the input feature vector for the i^{th} input sample; ℓ^p denotes the loss function of primary task; W^p denotes the weights associated with the primary task; y_i^a is the target output of the primary task for the i^{th} sample; N is the total number of samples; ℓ^a is the loss function of the auxiliary task a; A is the set of all auxiliary tasks, W^a is the weights associated with the auxiliary task a; y_i^a is the target output of the auxiliary task a for the i^{th} sample; f and g^a denote the neural network paths for the primary task and the auxiliary task a; λ^p is the weight considered for the loss function of the primary task; and λ^a is the weight associated with the loss function of the auxiliary task a. We set $\lambda^p = 1$ for all the experiments, and vary the auxiliary weights maximizing performance on the validation set (i.e., λ^{tv} for triviseme recognition and λ^e for emotion recognition).

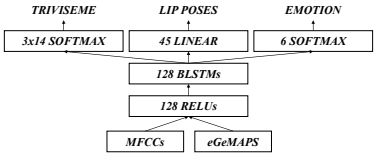


Figure 3. The MTL structure 1. Both hidden layers are shared between primary and secondary tasks.

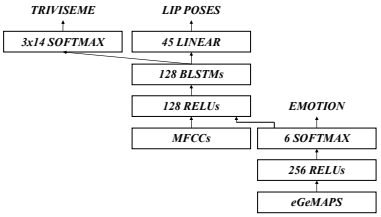


Figure 4. The MTL structure 2. The emotion recognition task is evaluated early, providing six softmax outputs as input of the second structure.

Notice that setting $\lambda_a = 0; \forall a$, reduces the multitask learning to *single task learning* (STL), focusing only on the primary task. We consider this setting as one of our baselines.

$$\ell = \sum_{i=1}^{N} \lambda^p \ell^p \left(y_i^p, f\left(x_i; W^p\right)\right) + \sum_{i=1}^{N} \sum_{a \in A} \lambda^a \ell^a \left(y_i^a, g^a\left(x_i; W^a\right)\right) \tag{3}$$

1) Structure 1: Figure 3 shows the first MTL structure, which we refer to as *structure 1*. This model has a ReLU layer and a BLSTM layer. Both hidden layers are shared between the tasks. The input of the models are the MFCCs and the eGeMAPS features. Note that the eGeMAPS features are extracted per speaking turn. Since the models operate frame-by-frame, we repeat the eGeMAPS vector multiple time so they have the same length as the MFCCs features. This structure assumes that the input features can be fused without any pre-processing step.

2) Structure 2: Figure 4 shows the second model, which addresses the difference in time resolution between MFCCs (frame-by-frame features) and eGeMAPS features (speaking turn level). The model has two connected steps, where the goal is to recognize emotions early, so that the output of its softmax layer can be used as input of the MTL structure. The first part of the model receives the eGeMAPS features as input. It has a hidden layer implemented with ReLU (256 nodes) and a softmax layer as output (six nodes). The output of this structure is concatenated with MFCCs, serving as the input of a second structure with two hidden layers implemented with ReLUs and BLSTMs, respectively. The secondary task of this structure is the triviseme recognition. We refer to this model as *structure 2*. A nice property of this structure is the smaller dimension of the input, reducing the 88 features in the eGeMAPS to six softmax outputs. Having less features prevents the second network from over-fitting. The model is jointly trained with back propagation.

V. Experimental Evaluation

We divided the data into five folds, where three folds are used for training, one fold for validation, and one fold for testing. We initialize the weights with the Glorot approach [26]. All the layers use dropout of 0.2 for regularization. We use a maxnorm

of 2.0 on the ReLU nodes (i.e. the norm of each weight vector connecting all its inputs to the current node is constrained to this value). To optimize the parameters of the model during training we use *stochastic gradient descent* (SGD) with a momentum of 0.95 applied on each batch. For SGD, we set the initial learning rate (η) equal to 0.1, decreasing its value as a function of the number of epochs (Eq. 4, where $decay$ is set to 0.01, and n_e is the number of previous epochs). All these hyper-parameters for SGD are set by using the validation set. We set the fixed window size (w in Section III-C) for LSTMs as 71 frames (591.7ms), since this is the value after which the performance of the model on the validation set saturates. We set the batch size as 256 (i.e., each batch contains 256 samples with 71 frames). We use the validation set for early stopping with a patience of 10 epochs for the primary task, and 3 epochs for each of the auxiliary tasks. We set $\lambda^a = 0$ when we want to stop the training on the auxiliary task a. During testing, we set $\Delta = 1$.

$$\eta = \frac{\eta}{1 + decay * n_e} \tag{4}$$

A. Objective Evaluations

1) Comparison of STL with MTL: This section reports the evaluation with objective metrics. Table II gives the performance of the models for lip movements prediction over the validation and test sets. We perform MTL by varying the weights associated with the auxiliary tasks with a random grid search, which have shown to be more effective than a uniform grid search [27]. We select the coefficients that give the best lip movement prediction on the validation set. Table II shows that we improve performance for MTL over STL methods (higher ρ_c, lower MSE). The results demonstrate the benefit of jointly solving triviseme and emotion recognition while predicting orofacial movements.

The validation set gives the best result for *structure 1* when $\lambda^{tv} = 1.0$, and $\lambda^e = 0.1$, and for *structure 2* when $\lambda^{tv} = 0.3$, and $\lambda^e = 0.1$. With these weights we get improvements on the test set for both models. *Structure 2* obtains better results. However, the relative improvements over the corresponding baselines is higher for *structure 1*. Since the ratio of emotional features to articulatory features is 3.52, *structure 1* is more prone to overfit the emotional features. *Structure 2* extracts more abstract features from emotional features before concatenating them with MFCCs, reducing the chances of over-fitting. We hypothesize that this is the reason that *structure 2* obtains better results, and *structure 1* obtains higher relative improvements over the baselines by using regularization with MTL.

Although our primary task is learning the lip movements from speech, the table also reports the accuracies for the auxiliary tasks over the test set (i.e., emotion recognition and viseme recognition). Viseme recognition accuracy denotes the accuracy for the middle viseme predictions in the triviseme output. Although the weights are optimized to maximize the performance of the primary task, the accuracies for emotion recognition and viseme recognition are all above chances. These results show that the proposed structures are learning features that are also discriminative for the auxiliary tasks.

If we consider the models trained with different weights for the auxiliary tasks as different ensembles, we can also aggregate their predictions by simply averaging their outputs. These results are given in the row "7 Ensembles" in Table II. Combining the models shows improvements for MSE for both the validation and test sets, suggesting that the models are diverse enough that their combinations can boost performance.

2) Comparison with Baselines: We also compare our proposed approach with two previous studies: the studies by Sadoughi and Busso [14], and Taylor et al. [12].

We replicated the Joint2-512 model which was the best framework proposed by Sadoughi and Busso [14]. This model has four layers and learns the facial movements across three facial regions (lower, middle and upper facial regions). The first two layers of the model are shared across the three tasks, and the last two layers are task specific. Similar to their study, we use the data from all the subjects in the IEMOCAP corpus for training this model, excluding the training set from the test set. We implement the training details described in that study. Table III reports the results for the Joint2-512 model. We also trained this model with our training partition, which has a smaller set. We reduced the number of nodes to 64 to avoid over-fitting on the validation set, since the number of samples is approximately five time smaller. We refer to this models as Joint2-64, showing the results in Table III. This model does not benefit from our data augmentation approach (Sec. III-C). We also implemented the framework proposed by Taylor et al. [12], following the provided description. This model takes as input concatenated frames of acoustic features (e.g., MFCCs), providing concatenated frames of lip movements. The average of the output predictions across the frames is considered as the prediction for the middle frame. Similar to their study, we consider an input window of 340ms ($\sim$41 frames), and an output window of 100ms ($\sim$ 13 frames). Table III also reports the results for this model.

The results on Table III show clear improvements by using the proposed MTL approach in terms of MSE and ρ_c. For the subjective evaluations, we focus on our models, comparing the STL and MTL settings, which is the key research question addressed in this paper.

B. Subjective Evaluations

We also evaluate our models with subjective evaluations. We rely on the objective evaluations to limit the conditions for the subjective evaluations, selecting only the best MTL settings (*structure 1*: $\lambda^{tv} = 1.0, \lambda^e = 0.1$; *structure 2*: $\lambda^{tv} = 0.3, \lambda^e = 0.1$). We also evaluate the STL cases for both structures where $\lambda^{tv} = \lambda^e = 0$, and the results from the ensembles. As a reference, we generate movements using the original motion capture recordings. We randomly selected 10 videos from the test set and rendered their animations with Xface. For all the seven conditions, we used the original motion capture recording to animate the upper region of the face (eyebrows and eyelids). The evaluation relies on crowdsourcing using Amazon mechanical turk, where we asked the raters to assess the naturalness of the videos using a 10 point Likert scale from 0 (low naturalness) to 9 (high naturalness). To avoid fatigue, each rater is given five videos in the seven conditions (i.e. 35 videos). The order of the videos is randomized for each rater, presenting one video at a time. To reduce the chance of raters answering the questions without watching the full video, the questionnaire is displayed only after the video is played.

We asked 24 evaluators to rate 35 videos, resulting in 12 annotations for each of the 70 videos (10 videos $\times$ 7 conditions).

Table II
EVALUATION OF THE PROPOSED MODELS WITH STL AND MTL (S1: *structure 1* AND S2: *structure 2*).

S	Mode	λ^{tv}	λ^{e}	Validation		Test		Test	
				ρ_c	MSE	ρ_c	MSE	Viseme Acc.	Emotion Acc.
S1	STL	0	0	0.374	1.326	0.311	1.024	-	-
	MTL	1	0	0.383	1.272	0.323	0.964	0.493	-
		0	1	0.323	1.486	0.273	1.055	-	0.396
		1	1	0.380	1.368	0.328	0.969	0.500	0.414
		1	0.1	0.385	1.240	0.343	0.937	0.501	0.430
		0.5	0.05	0.351	1.342	0.315	0.943	0.514	0.333
		0.3	0.1	0.347	1.406	0.340	0.924	0.505	0.314
	7 Ensembles	-	-	**0.391**	**1.192**	**0.347**	**0.856**	-	-
S2	STL	0	0	0.374	1.266	0.353	0.933	-	-
	MTL	1	0	0.408	1.189	0.361	0.881	0.518	-
		0	1	0.391	1.322	0.315	1.037	-	0.415
		1	1	0.411	1.246	0.322	0.962	0.507	0.385
		1	0.1	0.421	1.172	0.346	0.921	0.525	0.397
		0.5	0.05	0.419	1.172	**0.369**	0.869	0.520	0.384
		0.3	0.1	0.423	1.130	0.357	0.904	0.518	0.366
	7 Ensembles	-	-	**0.427**	**1.112**	0.362	**0.860**	-	-

Table III
THE TABLE COMPARES OUR MTL MODELS WITH PREVIOUS STUDIES.

Model	ρ_c	MSE
Joint2-512 by Sadoughi and Busso [14]	0.350	0.908
Joint2-64 by Sadoughi and Busso [14]	0.194	1.170
Taylor et al. [12]	0.158	0.990
Proposed (Best MTL)	0.357	0.904
Proposed (Best MTL-Ensembles)	**0.362**	**0.860**

The Cronbach's alpha between the raters is 0.5498. To remove annotator and video biases, we normalize the scores given by each rater to each video by dividing the sum of the scores assigned to its seven conditions. Figure 5 gives the average of these normalized ratings across all the evaluators. The *analysis of variance* (ANOVA) shows that the means for conditions are statistically different ($F(6, 833) = 9.5680, p < 1e-9$). We evaluate pairwise comparisons using Tukey's multiple comparisons test asserting significance at p-value=0.05. The color coded asterisks in Figure 5 shows the result (i.e., a bar with an asterisk is significantly higher than the bar indicated by the color of the asterisk). The pairwise comparisons demonstrate higher scores when we use the original movement, as expected. The best result for the generated orofacial movements is with MTL using *structure 1*, which has significantly higher score than its STL version. For *structure 2*, the average of the naturalness scores are higher for the MTL condition, but the difference with the STL condition is not statistically significant. Although the ensembles gave the lowest MSE (Table II), the naturalness scores are not as high, which may suggest that the movements are over smoothed.

VI. CONCLUSIONS

Although several studies have modeled orofacial movements for neutral speech, expressive lip movement synthesis is still challenging. This study proposed a novel multitask framework to generate expressive lip movements for CAs. We explored predictive models with BLSTMs for lip movement, integrating spectral and eGeMAPS features as the input. The secondary tasks in the multitask framework correspond to viseme recognition and emotion recognition. Objective and subjective evaluations demonstrate the advantages of using MTL, obtaining shared representations that

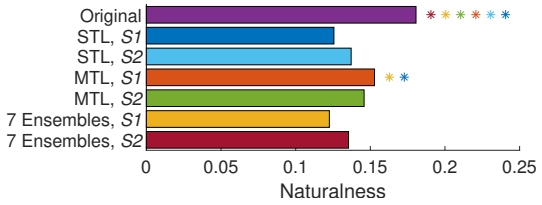


Figure 5. Average of the normalized perceptual evaluations. The color-coded asterisk denotes that a bar with an asterisk is significantly higher than the bar with the color of the asterisk ($p < 0.05$).

generate lip movements closer to the original sequences, increasing the perceived naturalness of the animations.

Using effective regularization in deep learning is especially important when modeling expressive facial movements, as the emotional audiovisual corpora are usually limited in size [13]. In this paper, the secondary tasks are carefully selected to improve the performance of the primary task (e.g., lip movement prediction). An important strength of our framework is that we can train MTL using datasets with partial information, without requiring one dataset to have all the required labels. Therefore, our MTL approach is useful for practical applications.

The evaluation considered data from a single speaker. We are interested in studying idiosyncratic differences between speakers that can be directly added to our models to create personality traits. We are also planing to evaluate whether the emotional content conveyed over the orofacial area is preserved in the generated movements.

ACKNOWLEDGEMENT

This work was funded by US National Science Foundation grant IIS-1718944.

REFERENCES

[1] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, November 2007.

[2] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, March 1954.

[3] Y. Xu, A. W. Feng, S. Marsella, and A. Shapiro, "A practical and configurable lip sync method for games," in *Motion in Games (MIG 2013)*, Dublin, Ireland, November 2013, pp. 131–140.

[4] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Magnenat-Thalmann N., Thalmann D. (Editors), Models and Techniques in Computer Animation, Springer Verlag*, Tokyo, Japan, 1993, pp. 139–156.

[5] Z. Deng, J. Lewis, and U. Neumann, "Synthesizing speech animation by learning compact speech co-articulation models," in *Computer Graphics International (CGI 2005)*, Stony Brook, NY, USA, June 2005, pp. 19–25.

[6] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, "Expressive visual text-to-speech using active appearance models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, Portland, OR, USA, June 2013, pp. 3382–3389.

[7] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Transactions on Graphics*, vol. 24, no. 4, pp. 1283–1302, October 2005.

[8] S. Mariooryad and C. Busso, "Feature and model level compensation of lexical content for facial emotion recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013)*, Shanghai, China, April 2013, pp. 1–6.

[9] K. Choi, Y. Luo, and J. Hwang, "Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system," *The Journal of VLSI Signal Processing*, vol. 29, no. 1-2, pp. 51–61, August 2001.

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

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

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

[13] X. Li, Z. Wu, H. Meng, J. Jia, X. Lou, and L. Cai, "Expressive speech driven talking avatar synthesis with DBLSTM using limited amount of emotional bimodal data," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 1477–1481.

[14] N. Sadoughi and C. Busso, "Joint learning of speech-driven facial motion with bidirectional long-short term memory," in *International Conference on Intelligent Virtual Agents (IVA 2017)*, ser. Lecture Notes in Computer Science, J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, and S. Kopp, Eds. Stockholm, Sweden: Springer Berlin Heidelberg, August 2017, vol. 10498, pp. 389–402.

[15] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[16] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 2474–2477.

[17] S. Mariooryad and C. Busso, "Factorizing speaker, lexical and emotional variabilities observed in facial expressions," in *IEEE International Conference on Image Processing (ICIP 2012)*, Orlando, FL, USA, September-October 2012, pp. 2605–2608.

[18] P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Australian International Conference on Speech Science & Technology (SST 2004)*, Sydney, NSW, Australia, December 2004, pp. 265–270.

[19] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April-June 2016.

[20] K. Haag and H. Shimodaira, "Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis," in *International Conference on Intelligent Virtual Agents (IVA 2016)*, ser. Lecture Notes in Computer Science, D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Scherer, and A. Leuski, Eds. Los Angeles, CA, USA: Springer Berlin Heidelberg, September 2016, vol. 10011, pp. 198–207.

[21] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[22] K. Balci, "Xface: MPEG-4 based open source toolkit for 3D facial animation," in *Conference on Advanced Visual Interfaces (AVI 2004)*, Gallipoli, Italy, May 2004, pp. 399–402.

[23] 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, October 2012.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, November 1997.

[25] F. Chollet, "Keras: Deep learning library for theano and tensorflow," https://keras.io/, April 2017. [Online]. Available: https://github.com/fchollet/keras

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, Sardinia, Italy, May 2010, pp. 249–256.

[27] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, February 2012.