Joint Learning of Speech-Driven Facial Motion with Bidirectional Long-Short Term Memory

Najmeh Sadoughi✉ and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory,
Department of Electrical and Computer Engineering,
The University of Texas at Dallas

{nxs137130 and busso}@utdallas.edu

Abstract. The face conveys a blend of verbal and nonverbal information playing an important role in daily interaction. While speech articulation mostly affects the orofacial areas, emotional behaviors are externalized across the entire face. Considering the relation between verbal and nonverbal behaviors is important to create naturalistic facial movements for *conversational agents* (CAs). Furthermore, facial muscles connect areas across the face, creating principled relationships and dependencies between the movements that have to be taken into account. These relationships are ignored when facial movements across the face are separately generated. This paper proposes to create speech-driven models that jointly capture the relationship not only between speech and facial movements, but also across facial movements. The input to the models are features extracted from speech that convey the verbal and emotional states of the speakers. We build our models with *bidirectional long-short term memory* (BLSTM) units which are shown to be very successful in modeling dependencies for sequential data. The objective and subjective evaluations of the results demonstrate the benefits of joint modeling of facial regions using this framework.

1 Introduction

While spoken language is the primary way of communication, nonverbal information provides important information that enriches speech during face-to-face interaction. Nonverbal information not only complements speech, but also conveys extra information [26]. Humans unconsciously use different channels to express and externalize their thoughts, emotions and intentions. These channels are integrated in a non-trivial manner. However, listeners can easily decode the message, inferring each of these communicative goals. The models should consider these relationships, if we want to design better *conversational agents* (CAs) that express realistic expressive human-like behaviors.

The face is one of the primary channels to express different communicative goals. Different facial muscles contribute in creating speech articulation and facial expression. Previous studies have shown the temporal and spatial interplay

between speech and emotion in the face [7, 24]. In general, the activity in the orofacial area is dominated by speech articulation, and the activity in the upper face area is dominated by emotions. However, the interplay is not trivial. Since humans can easily decode these communication goals, an effective CA should capture this interplay. Likewise, emotional traits associated with an emotion may involve multiple facial movements (e.g., surprise externalized as opening of mouth and raising of eyebrows). Even a single facial muscle may activate different facial regions. For example, the Zygomaticus major, which affects the cheek area, allows us to smile. The activation of the Levator labii superioris affects the lips and the upper facial region. The aforementioned relations not only between speech and facial expression, but also across facial regions suggest that generating realistic behaviors for CAs require careful consideration of these underlying dependencies. In fact, previous studies have demonstrated that joint models for eyebrow and head motion produced more realistic sequences than the ones created with separate models [13, 23].

Speech carries verbal and nonverbal cues, including the externalizations of the affective state of the speaker. Given the strong correlation between speech and facial expressions [6], speech-driven models offer appealing solutions to generate human-like behaviors that preserve the timing relation between modalities. This study proposes to create joint speech-driven models for facial expressions using the latest advances in deep learning. The framework relies on *bidirectional long-short term memory* (LSTM) units to capture (1) the relation between speech and facial expressions, (2) the relation across facial features. We use deep structures that help learning the interplay between the facial movements in different regions of the face in a systematic and principled manner. We achieve this goal by using multitask learning, where predictions for lower, middle and upper facial regions are jointly estimated. Our results demonstrate the benefit of learning the facial regions jointly rather than separately. While other studies used generative models to jointly model facial behaviors [13, 23, 28], this is the first study that solves this problem using multitask learning with deep learning.

2 Related Works

The conventional approach for facial animations is the use of rule-based system. For example, predefined shapes based on the target articulatory unit can be concatenated to generate facial movements [12, 32]. Defining facial trajectories that are tightly coupled with speech is a challenge, especially in the presence of emotion (rhythm, emphasis). Although some studies have considered continuous emotional descriptors (e.g. Albrecht et al. [1]), the most common approach to model emotion is to consider specific models created for prototypical emotional categories [25, 27]. However, defining the facial expressions per emotion reduces the subtle differences that exist between slightly different facial expressions, and makes the animation seem repetitive. As an appealing alternative, data-driven methods are usually better at handling these fine changes, by learning the vari-

ations shown in real recordings [2, 10, 21]. This study focuses on data-driven solutions.

There are several data-driven studies to predict facial movements from speech. Brand [4] proposed to use HMMs to learn the mapping between speech and facial features using entropy minimization. Gutierrez et al. [18] designed a system to synthesize facial movements from speech features. Their system used 12 *perceptual critical band features* (PCBFs), fundamental frequency and energy. The approach predicts lip movements by identifying the 12 *nearest neighbors* (NNs) in the speech feature space The selected segments are concatenated and smoothed by a moving average window. Taylor et al. [30] proposed to use *deep neural network* (DNNs) composed of densely connected *rectified linear units* (RELUs) to predict lip movements from speech. They used 25 *mel frequency cepstral coefficients* (MFCCs) as speech features. They concatenated the acoustic features extracted for each frame over a window, predicting the lip movements, which are smoothed by averaging the estimations over a target window. Subjective and objective evaluations demonstrated better performance over an *HMM inversion* (HMMI) method proposed by Choi et al. [11]. Fan et al. [16] explored different *deep bidirectional LSTMs* (DBLSTMs) for mapping speech or speech plus text into lips movements. When using speech as input, they extracted 13 MFCCs, and their first and second order derivatives (i.e. 39D) as the input. When the text is provided, they concatenate the input with tri-phonemes. Their objective and subjective evaluations showed better results for the DBLSTMs compared with HMMs. All these studies did not directly consider emotional information.

There are also studies that have generated expressive facial movements using data driven models, when the input is text. Cao et al. [10] conducted one of the early data-driven works on generating expressive facial movements. Their approach relied on defining expressive units of articulations. They segmented and stored their recordings into anime nodes indexed by the corresponding phoneme, emotion, prosodic features, and motion capture features. For synthesis, suitable anime nodes are selected, time warped, concatenated and smoothed as dictated by the target requirement. Mana and Pianesi [22] trained different left-to-right HMMs for pairs of visemes and emotions. Anderson et al. [2] proposed a visual text-to-speech system that synthesizes expressive audiovisual speech with a set of continuous weights for emotional categories. They used *cluster adaptive training* (CAT) which is built upon *hidden Markov models* (HMMs) for text-to-speech. The HMM states are modeled by decision trees, and the model learns the appropriate weight vector for each emotion, which is used to find the linear combination of states. They evaluated the synthesized results in terms of the precision of the perceived target emotion. All these studies used text as input.

To the best of our knowledge, there is only one study on expressive facial movement synthesis, using emotional audio features and the target emotion as the input. Li et al. [21] proposed several structures using DBLSTMs to synthesize emotional facial movements based on limited emotional data. They used a neutral corpus with 321 utterances, and an emotional corpus of 44 subjects reading sentences in six different emotional categories. They evaluated their pro-

posed structures in terms of the perceived naturalness and expressiveness of the videos. Their best structure is composed of two models. The first one is trained with neutral data, and is used to predict the movements for emotional inputs. These predictions are concatenated with the input audio features and used as the input to another model which is trained with emotional data to predict the emotional movements. Our study also generates expressive facial behaviors from speech. However, (1) our approach does not require the target emotion of the input speech to be known, rather it captures the relationship between facial movements and emotional features extracted from speech, (2) we rely on a bigger database, and (3) we investigate joint versus separate modeling of facial features from the speech signal using powerful deep learning structures with BLSTMs under multitask framework.

3 Resources

3.1 IEMOCAP Corpus

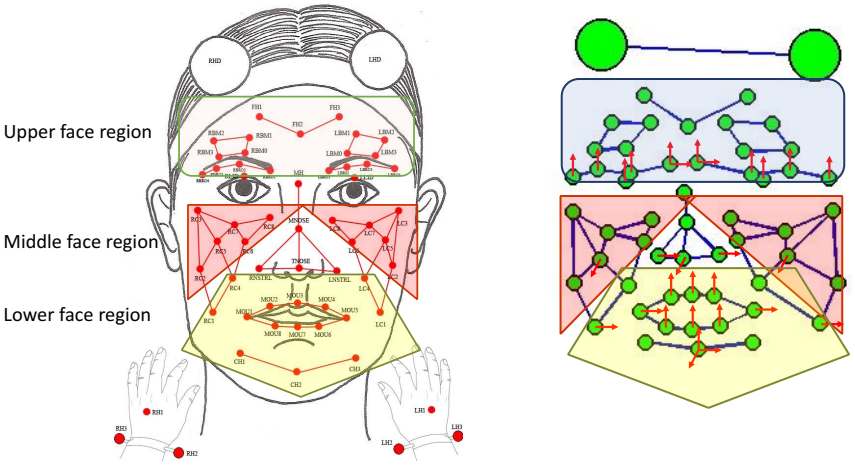
This paper uses the *interactive emotional dyadic motion capture* (IEMOCAP) database [9]. This database is multimodal, comprising audio, video, and motion capture recordings from 10 actors during spontaneous and script-based dyadic interactions. We use the data from all the actors in our experiments. From the motion capture data, we use the position of the facial markers grouped into three regions; upper face region, middle face area, and lower face area (Fig. 1(a)). Note that the three regions here are chosen inspired by the study conducted by Busso et al. [8]. In the extreme case, we can consider each marker as a region. Busso et al. [9] provides more details about this corpus.

3.2 Multimodal Features

From the motion capture recordings, we use 19 markers for the upper facial region ($19 \times 3D$), 12 markers for the middle facial region ($12 \times 3D$), and 15 markers for the lower facial region ($15 \times 3D$). The motion capture data is recorded at 120 *frame per second* (fps). From the speech signal, we extract 25 MFCCs, fundamental frequency, and energy with Praat over $25ms$ windows every $8.3ms$. Eyben et al. [14] proposed the *extended Geneva minimalistic acoustic parameter set* (E-GeMAPS), which is a compact set of features that were carefully selected for paralinguistic tasks. The set has 23 *low level descriptors* (LLDs), where six of them are already included in the features extracted by Praat. Therefore, we add the rest of these features (17 D). These features are extracted over $20ms$ or $60ms$ every $10ms$. We up-sample the speech features using linear interpolation to get 120 fps, matching the sampling rate of the motion capture data. We use Z-normalization per subject for the speech and visual features.

3.3 Rendering the Animations with Xface

For rendering the animations, we use Xface [3]. Xface uses the MPEG4 standard to define *facial points* (FPs). To animate the face, Xface uses *facial action*



(a) Facial Regions (b) Mapping

Fig. 1. Layout and groups of facial markers (a) markers belonging to upper, middle and lower face regions, (b) markers mapped to FAPs in Xface (highlighted with arrows).

parameters (FAPs) which change the position of the FPs. Most of the markers used in the IEMOCAP database were placed following the FPs defined by the MPEG4 standard, facilitating the mapping between markers and FAPs. We follow the same mapping proposed by Marioryad and Busso [23]. Figure 1(b) highlights the markers that are mapped into FAPs in Xface. We use the idle position of the markers for the actors as the neutral pose, and extract the range of movements for each actor. The neutral pose is mapped to the neutral pose of the face in Xface defined by FAPs, and the changes in the position of the markers are scaled to the changes of FAPs allowed by Xface. While there are other more realistic talking heads, the direct map between markers and FAPs facilitate the evaluation of this study.

3.4 Objective Metrics

The models in this paper learn how to derive facial movements. Therefore, the outputs of the models are continuous variables, where previous studies have either minimized the *mean squared error* (MSE) [15,30], or maximized the *concordance correlation* (ρ_c) [31]. If x and y are the target and predicted values, Equation 1 defines ρ_c, where ρ is the Pearson correlation between x and y, and μ_x and μ_y, and σ_x and σ_y are the means and variances of x and y, respectively.

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

Our preliminary experiments showed that using ρ_c as the optimizing criterion generates higher range of movements for the target variable, which looks better

when the trajectories are visualized. Therefore, we relied on minimizing $1 - \rho_c$ for our experiments. However, we report both metrics to assess the performance of the models after concatenating all the test segments.

4 Speech-Driven Models with Deep Learning

Deep learning structures are very powerful to learn complex temporal relationships between modalities, hence, they are a perfect framework for speech-driven models for facial expressions. This study proposes to build joints models that consider the relation not only between speech and facial movements, but also across facial regions. For comparison, we assess models that either separately or jointly generate facial movements for the lower, middle and upper facial regions.

We build our models by stacking multiple non-linear layers where the input corresponds to the 44D speech feature set (Sec. 3.2). The models have densely connected layers with *rectified linear units* (RELUs), BLSTMs, and a linear layer at the top, since the task is to generate the position of the markers.

4.1 Bidirectional Long-Short Term Memory (BLSTM)

We rely on *recurrent neural networks* (RNNs) to capture the temporal dependencies for continuous signals. RNNs use temporal connections between consecutive hidden units at each layer to model the dependencies between time frames. However, as the length of the input signal increases, RNNs are susceptible to the problem of exploding or vanishing gradients [19]. LSTMs are an extension of RNNs, which were introduced to handle this problem [19].

LSTM utilizes a cell to keep track of the useful past content given the input, and previous hidden state. LSTM uses gating mechanisms to capture the long and short term dependencies in the temporal signals. It uses three gates for this goal: input, forget, and output gates. The input gate controls the amount of the current input to be stored in the cell unit. The forget gate controls the amount of the previous cell content being retained in the cell. The output gate modulates the amount of the cell content being used as the output of the hidden state at time t. We use the implementation of LSTM in Keras.

An extension of LSTM is its bidirectional version, BLSTM, which utilizes the previous and future frames to predict the outputs at each time (Fig. 2). The implementation of BLSTM consists of training forward and backward LSTMs, and concatenating their hidden units. The key benefit of BLSTMs is that they generate more smooth movements. Although BLSTMs can be used in real time by using a post-buffer, this study estimates the facial movements off-line using the whole turn sequence. We run BLSTMs on each turn, predicting a sequence of the same length as the output (speech features are up-sampled to 120fps).

4.2 Separate Models

Our baseline models consists of structures that separately generate facial behaviors for the lower, middle and upper face regions. These models independently

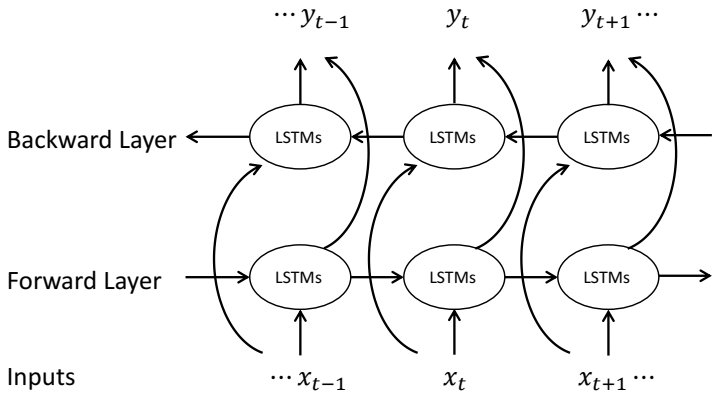
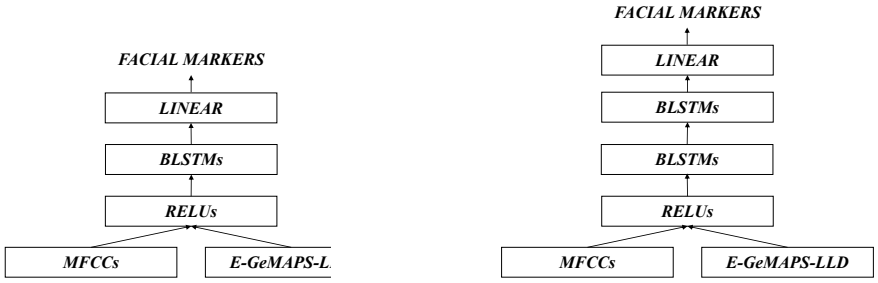


Fig. 2. Illustration of BLSTM composed of forward and backward paths.

create the facial markers trajectories for each region. While local relationships within regions are preserved, the intrinsic relationship across regions are neglected. The underlying assumptions is that these relationships across the three regions are not important. Figure 3 shows two alternative frameworks. *Separate-1* uses one BLSTM layer, whereas *Separate-2* uses two BLSTM layers. We consider



(a) Separate-1 (b) Separate-2

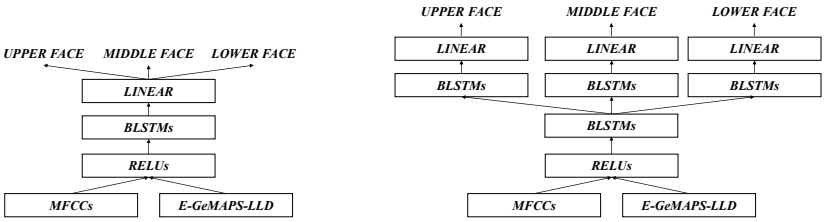
Fig. 3. Baseline speech-driven models, where the facial movements for the lower, middle and upper face regions are separately generated.

4.3 Joint Models

We create the proposed joint models using multitask learning. Multitask learning aims to jointly solve related problems using shared layer representation. In our formulation, we have three related tasks consisting of predicting the movements

of the lower, middle and upper face regions, where part of the neural networks are shared between all the tasks. These models assume that facial movements over different regions have principled relationships. From a learning perspective, when predicting movements in one region, the estimation of the movements for the other two regions can be considered as a systematic regularization that helps the network to learn more robust features with better generalization.

Figure 4 shows the two joint models that we investigate. The model *Joint-1* has the whole network shared between the three tasks except for the linear output layer. This model has shared representation of the three tasks in all the nonlinear layers, regularizing the whole network. This network is equivalent to a model that predicts all the facial movements at once. The model *Joint-2* shares the first two layers between all the tasks. However, the last two layers are task-specific. The task specific layers capture localized facial relationship within regions, while the shared layers preserve relationship across regions.



(a) Joint-1 (b) Joint-2

Fig. 4. Proposed joint speech-driven models for facial movements. The *Joint-1* model has shared layers. The *Joint-2* model has shared and task-specific layers.

5 Experiment & Results

The proposed models are implemented and evaluated using the IEMOCAP corpus, where we used 60% of the data for training, 20% for validation, and 20% for testing. We use Keras with Theano as backend to implement and train the models. We rely on *adaptive moment estimation* (ADAM) [20] for the optimization of the parameters. ADAM keeps track of estimates of first and second moments of gradient during training, and utilizes the ratio between the bias-corrected first moment and the bias-corrected second moment of the gradient to update the parameters. This process helps scaling the update, according to the uncertainty (second moment), and making the step size invariant to the magnitude of the gradient. We use different learning rates ($\backsim \{0.1, 0.01, 0.001, 0.0001\}$), and evaluated the model on the validation set. The results demonstrated that a learning rate of 0.0001 works better. Furthermore, all the layers use dropout of 0.2 to counter overfitting [29]. Our training examples have various lengths. We set a

Table 1. Objective metrics for facial movements generated with joint and separate models for the lower, middle and upper face region.

Model	# nodes per layer	# params	Upper face		Middle face		Lower face	
			ρ_c	MSE	ρ_c	MSE	ρ_c	MSE
Separate-1	512	12.8M	0.140	1.47	0.268	1.36	0.401	1.12
Joint-1	512	4.4M	0.150	1.32	0.274	1.30	0.390	1.26
Separate-1	1024	50,8M	0.149	1.41	0.277	1.16	0.411	1.05
Joint-1	1024	17,1M	0.160	1.40	0.297	1.24	0.413	1.14
Separate-2	512	31.7M	0.135	1.44	0.260	1.24	0.392	1.04
Joint-2	512	23.2M	0.160	1.37	0.307	1.14	0.411	1.06

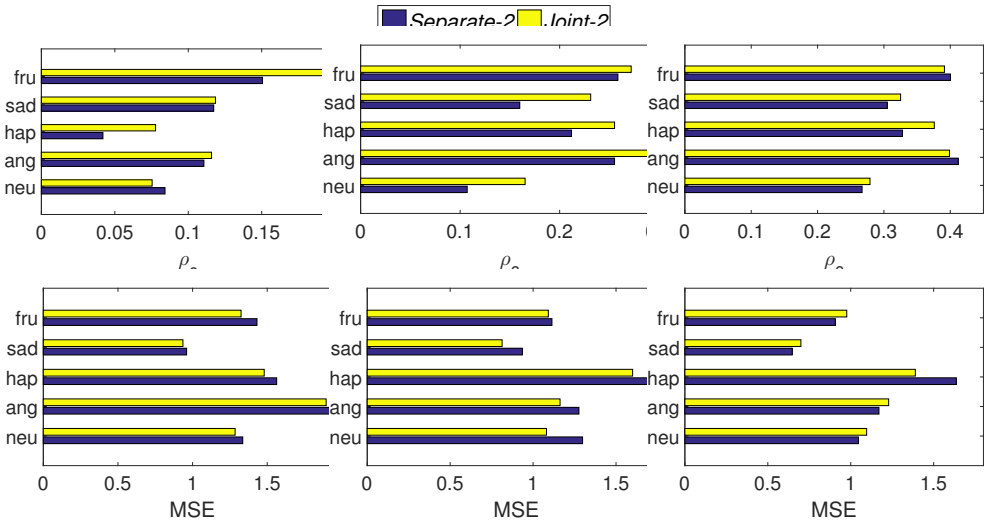
batch size of 4,096, making sure that the total number of frames used in one batch does not exceed this number. As a result, we have different number of sequences in different batches. All the weights are initialized with the approach proposed by Glorot et al. [17]. We train all the models with 50 epochs.

5.1 Objective Evaluation

We train the models with different number of nodes and layers for the joint and separate models. Table 1 summarizes the results. When we compare *Joint-1* and *Separate-1* with 512 nodes, the results show improvements in the joint model for the middle and upper face regions. When we increase the number of nodes to 1,024, we observe improved performance for all the regions, where the joint model achieves higher ρ_c and smaller MSE. The table also compares the *Separate-2* and *Joint-2* models which have the same number of layers. Changing the structure to the *Joint-2* and *Separate-2* models tend to improve the MSE for all the facial regions, compared to the *Joint-1* and *Separate-1* models. Furthermore, the *Joint-2* model achieves better concordance correlation than the *Separate-2* model. Note that the *Separate-1* model requires approximately three times more parameters than the *Joint-1* model. Likewise, the *Separate-2* model has 36.31% more parameters than the *Joint-2* model. The proposed joint structure not only provides better performance, but also requires less parameters which is an advantage due to memory requirements.

Emotional Analysis We compare the performance of *Separate-2* and *joint-2* models for different emotional categories. The IEMOCAP corpus is emotionally annotated at the speaking turn level by three annotators in term of nine emotional categories (neutral, anger, happiness, sadness, fear, frustration, surprise, disgust, and other). We derive a consensus label using the majority vote rule, where turns without consensus are excluded from this analysis. In the test set, we have the following distribution: 113 (neutral), 161 (anger), 86 (happiness), 131 (sadness), 3 (fear), 247 (frustration), 12 (surprise), 0 (disgust), and 2 (other). We only consider emotional classes with more than 50 speaking turns. We concatenate all the speaking turns belonging to a given emotion, estimating ρ_c and MSE.

Figure 5 shows the average ρ_c and MSE for the three facial regions per emotional category. For the upper face area, the *Joint-2* model shows better results across the emotions, except for ρ_c for neutral. Furthermore, for the middle face region, the results show improvements for all the emotions. For the lower face region, ρ_c shows improvements for neutral speech, happiness, and sadness, while MSE is improved only for happiness.



(a) Upper face (b) Middle Face (c) Lower Face

Fig. 5. Comparison of the results achieved for ρ_c and MSE per emotional category using the *Separate-2* and *Joint-2* models.

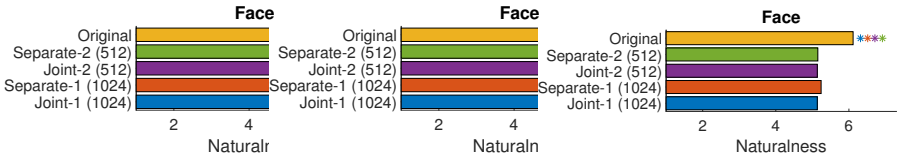
5.2 Subjective Evaluation

Subjective evaluations provide convincing evidences about the performances of the models. Table 1 shows that the *Joint-1* and *Separate-1* models provide better results with 1,024 nodes per layer than with 512 nodes per layer. Therefore, we only include the *Joint-1* and *Separate-1* models trained with 512 nodes per layer. The *Joint-2*, and *Separate-2* models are trained with 512 nodes. The evaluation also includes the animations generated with the *original* motion capture data. Therefore, we have five conditions per speaking turn. We randomly selected 10 turns from the test set, and generated their animations for all of these five conditions using Xface (50 videos). We do not include head motion, so our raters can focus on facial movements. We use the original eyelid and nose markers positions across all the videos.

The evaluation is conducted using crowdsourcing with *Amazon mechanical turk* (AMT). We limit our pool of evaluators to workers who have performed

well in our previous crowdsourcing tasks [5]. We ask each evaluator to rate the *naturalness* of the 50 videos in a likert-like scale from 1 (low naturalness) to 10 (high naturalness). The task requires to annotate the perceived naturalness for the overall animation. In addition, we ask the raters to annotate the naturalness of the eyebrow and lips movements (i.e., three questions per video). Since we have only animated one of the markers in the cheek area (see Fig. 1(b)), we do not ask the annotators for separate ratings for the middle face region. We show one video at a time to the annotators, displaying the questionnaire only after the video is fully played. This approach reduces the chance of annotators providing random answers without even looking at the video. We randomize the order of the videos per evaluator. We recruited 20 subjects for this evaluation.

Figure 6 shows the average scores for the five conditions. The Cronbach's alpha between the annotators is $\alpha = 0.6720$. *One way analysis of variance* (ANOVA) shows statistically different values between the five cases for all three questions ($p < 0.001$). Pairwise comparisons of the results between the five conditions only show that the animations generated with the *original* sequences are significantly better than the animations automatically generated from speech, which is expected. This result reveals that the differences between the animations generated with the joint and separate models were subtle. From the objective metrics, we observe that the middle face region shows the highest improvements when the joint models are used (Sec. 5.1). However, these differences may not be visually perceived due to limitations in Xface. We hypothesize that the difference in facial movements will be more clear if we use a more expressive talking head, which is the focus of our future work. Even with the these results, it is important to highlight that using joint models allows the network to achieve similar performances than separate models using fewer parameters.



(a) Eyebrows (b) Mouth (c) Face

Fig. 6. Perceived naturalness of the animations. The color-coded asterisks indicate that the bar is significantly higher than the bars identified by the asterisks' colors ($p < 0.01$).

6 Conclusions

This paper explored multitask learning architectures to train speech-driven models for facial movements. The framework relied on BLSTMs to capture temporal information, using speech features as the input to predict facial movements. The models jointly learn the relationship not only between speech and facial

expressions, but also across facial regions, capturing intrinsic dependencies. We compared the results with models that separately estimate movements for the lower, middle and upper part of the face, ignoring relations between regions. Objective evaluation of the results showed improvements for the joint models in different facial regions as measured by ρ_c and MSE. The improvement are higher for the *Joint-2* model, which has shared layers and task specific layers. Interesting, by sharing the layers the proposed solutions reduced the number of parameters, which is another advantage of our approach. While subjective evaluations did not reveal any significant difference between the joint and separate models, we believe that this result is due to the lack of expressiveness of Xface to create animations with subtle behaviors. We will explore more sophisticate toolkits to present our results, including photo realistic videos [30]. We will also evaluate generating head motion driven by speech as an extra task in the multi-task learning framework. We expect that the behaviors will be better synthesized with the rest of the facial movements, providing better speech driven solutions.

7 Acknowledgment

This work was funded by NSF grants (IIS: 1352950 and IIS: 1718944).

References

1. I. Albrecht, M. Schröder, J. Haber, and H.-P. Seidel. Mixed feelings: expression of non-basic emotions in a muscle-based talking head. *Virtual Reality*, 8(4):201–212, September 2005.
2. 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)*, pages 3382–3389, Portland, OR, USA, June 2013.
3. K. Balci. Xface: MPEG-4 based open source toolkit for 3D facial animation. In *Conference on Advanced Visual Interfaces (AVI 2004)*, pages 399–402, Gallipoli, Italy, May 2004.
4. M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1999)*, pages 21–28, New York, NY, USA, 1999.
5. A. Burmania, S. Parthasarathy, and C. Busso. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388, October-December 2016.
6. C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1075–1086, March 2007.
7. C. Busso and S. Narayanan. Interplay between linguistic and affective goals in facial expression during emotional utterances. In *7th International Seminar on Speech Production (ISSP 2006)*, pages 549–556, Ubatuba-SP, Brazil, December 2006.
8. 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*, 15(8):2331–2347, November 2007.

9. C. Busso and S. Narayanan. Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database. In *Interspeech 2008 - Eurospeech*, pages 1670–1673, Brisbane, Australia, September 2008.
10. Y. Cao, W. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics*, 24(4):1283–1302, October 2005.
11. 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*, 29(1-2):51–61, August 2001.
12. 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*, pages 139–156, Tokyo, Japan, 1993.
13. Y. Ding, C. Pelachaud, and T. Artieres. Modeling multimodal behaviors from speech prosody. In R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira, editors, *International Conference on Intelligent Virtual Agents (IVA 2013)*, volume 8108 of *Lecture Notes in Computer Science*, pages 198–207. Springer Berlin Heidelberg, Edinburgh, UK, August 2013.
14. 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*, 7(2):190–202, April-June 2016.
15. 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)*, pages 4884–4888, Brisbane, Australia, April 2015.
16. B. Fan, L. Xie, S. Yang, L. Wang, and F. K. Soong. A deep bidirectional LSTM approach for video-realistic talking head. *Multimedia Tools and Applications*, 75(9):5287–5309, May 2016.
17. X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, pages 249–256, Sardinia, Italy, May 2010.
18. R. Gutierrez-Osuna, P. Kakumanu, A. Esposito, O. Garcia, A. Bojorquez, J. Castillo, and I. Rudomin. Speech-driven facial animation with realistic dynamics. *IEEE Transactions on Multimedia*, 7(1):33–42, February 2005.
19. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997.
20. D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–13, San Diego, CA, USA, May 2015.
21. 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*, pages 1477–1481, San Francisco, CA, USA, September 2016.
22. N. Mana and F. Pianesi. HMM-based synthesis of emotional facial expressions during speech in synthetic talking heads. In *International Conference on Multimodal Interfaces (ICMI 2006)*, pages 380–387, Banff, AB, Canada, November 2006.
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*, 20(8):2329–2340, October 2012.
24. 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)*, pages 1–6, Shanghai, China, April 2013.

25. S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. Virtual character performance from speech. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2013)*, pages 25–35, Anaheim, CA, USA, July 2013.
26. A. Mehrabian. Communication without words. In C. Mortensen, editor, *Communication Theory*, pages 193–200. Transaction Publishers, New Brunswick, NJ, USA, December 2007.
27. C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, January 1996.
28. N. Sadoughi, Y. Liu, and C. Busso. Speech-driven animation constrained by appropriate discourse functions. In *International conference on multimodal interaction (ICMI 2014)*, pages 148–155, Istanbul, Turkey, November 2014.
29. N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, June 2014.
30. S. Taylor, A. Kato, I. Matthews, and B. Milner. Audio-to-visual speech conversion using deep neural networks. In *Interspeech 2016*, pages 1482–1486, San Francisco, CA, USA, September 2016.
31. 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)*, pages 5200–5204, Shanghai, China, March 2016.
32. 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)*, pages 131–140, Dublin, Ireland, November 2013.