Expressive Speech-Driven Lip Movements with Multitask Learning

Najmeh Sadoughi and Carlos Busso

Multimodal Signal Processing Lab (MSP)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas - 75080, USA

Motivation

Background:

- Lower facial region is modulated by articulatory and emotional cues
- Modeling these factors can be useful for the generation of expressive lip movements

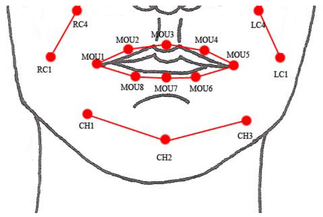
Our Work:

- Deep learning solutions to integrate the articulatory and emotional features from speech
- Using MTL to capture the relationship between speech articulation and emotional content
- The primary task of predicting lip movements is complemented with two secondary tasks:
  - Viseme recognition
  - Emotion recognition
- Adding auxiliary tasks helps the network to learn more predictive features for orofacial movements

Resources

Corpus:

- IEMOCAP (1st female subject)
  - Removing idiosyncratic differences
- Emotion annotations for six emotional categories
  - Three annotators (majority vote)
- 14 viseme categories
- Audio:
  - 25 MFCCs
  - 88 eGeMAPS per turn
- Orofacial region:
  - 15 motion capture markers (3x15D)



Rendering Toolkit:

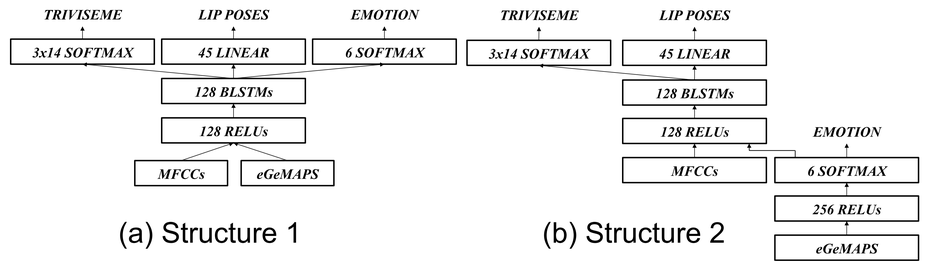
- Xface

Method

Multitask Learning (MTL):

- MTL jointly solves related problems
- Prediction of orafacial movements is the primary task
- Triviseme and emotion recognition are auxiliary tasks
- The auxiliary tasks can be considered as regularizers for the network to learn more and generalizable robust features

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



(a) Structure 1 (b) Structure 2

Results

Objective Evaluation

- Concordance correlation (CC)
- Mean square error (MSE)

	Mode	λ^{tv}	λ^{e}	CC	MSE
Structure 1	STL	0	0	0.311	1.024
	MTL 1	1	0	0.323	0.964
		0	1	0.273	1.055
		1	1	0.328	0.969
		1	0.1	0.343	0.937
		0.5	0.05	0.315	0.943
		0.3	0.1	0.340	0.924
	7 Ensembles	--	--	**0.347**	**0.856**
Structure 2	STL	1	0	0.353	0.933
	MTL 2	0	1	0.361	0.881
		1	1	0.315	1.037
		1	0.1	0.322	0.962
		0.5	0.05	0.346	0.921
		0.3	0.1	**0.369**	0.869
		0	0	0.357	0.904
	7 Ensembles	--	--	0.362	**0.860**

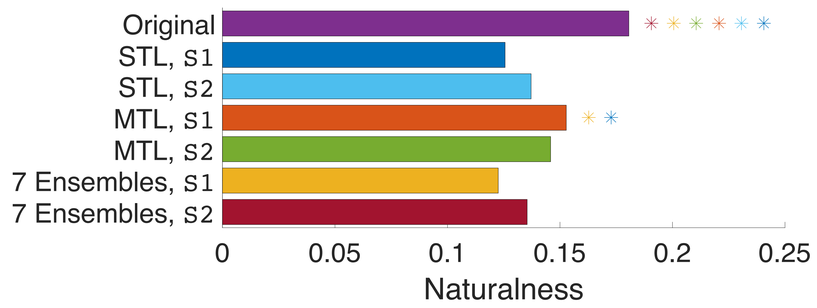
Comparison of MTL and Baselines

Model	CC	MSE
Joint 2- 512, Sadoughi and Busso (2017)*	0.350	0.980
Join 2-64, Sadoughi and Busso (2017)	0.194	1.170
Taylor et al. (2016)	0.158	0.990
Best of MTL (Proposed)	0.357	0.904
Best of MTL-Ensembles (Proposed)	0.362	0.860

*Trained with more data

Subjective Evaluation

- 70 videos
  - 10 videos per model
- 24 evaluators (AMT)
  - Each evaluated 35 videos
  - 12 annotations per video
  - Cronbach's alpha is 0.549



Discussion

Conclusions

- Using effective regularization in deep learning is important for modeling expressive facial movements
- The secondary tasks were carefully selected to improve the performance of the primary task
- An important strength of our framework is that we can train MTL using datasets with partial information

Future Work

- Modeling idiosyncratic differences between speakers that can be directly added to our models to create personality traits
- Evaluating whether the emotional content conveyed over the orofacial area is preserved in the generated movements

This work was funded by NSF (IIS 1718944)