Generating Talking Face Landmarks from Speech

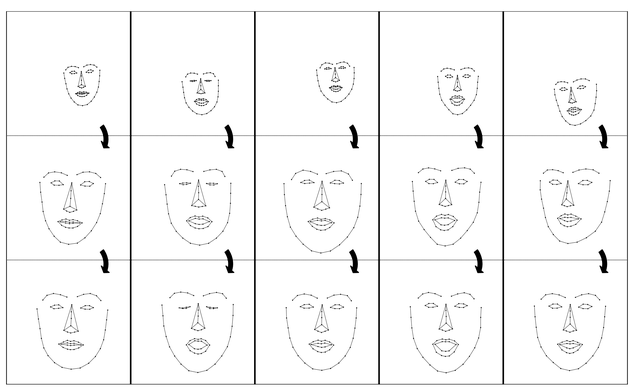
Sefik Emre Eskimez [1], Ross K Maddox [2 3], Chenliang Xu [3], Zhiyao Duan [1]
{eeskimez, rmaddox}@ur.rochester.edu, {chenliang.xu, zhiyao.duan}@rochester.edu
[1]Department of Electrical and Computer Engineering, [2]Department of Biomedical Engineering,
[3]Department of Neuroscience, [4]Department of Computer Science

Abstract

- **Problem:** Human speech comprehension suffers from background noise, channel distortion, reverberation, and hearing impairment.
- **Inspiration:** The presence of visual signals of speech has been shown to significantly improve speech comprehension [1] for ordinary and hearing impaired population.
- **Solution:** Generate a synthetic, natural looking talking face to act as a "visual hearing aid."

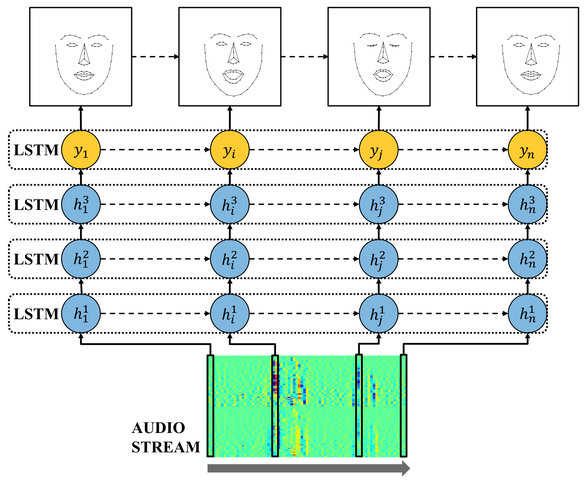
Face Landmark Normalization

We scale, rotate and translate the face landmarks to align them. Then, we remove the identity information by transforming different faces to the mean face.



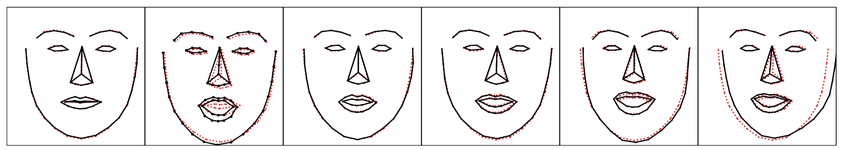
Proposed System Overview

We propose an LSTM network for generating talking face landmarks from speech. We use 40 ms window size without overlap to extract log-mel spectrogram.

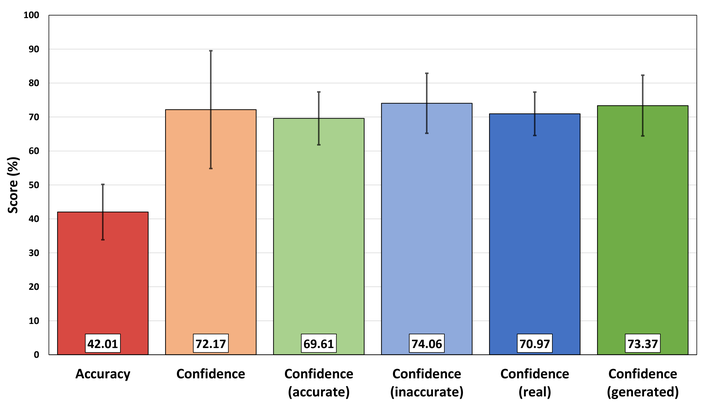


Generation Results

Pair-wise comparison between ground-truth (black solid lines) and generated landmarks (red dotted lines) on unseen talkers and sentences.



Subjective Evaluation Results



Objective Evaluation & Model Selection

We present the objective evaluation results for different system configurations. The models are named according to the amount of delay and contextual information. For example, "D40-C5" describes a model trained with 40 ms delay and 5 frames of context.

	RMSE	RMSE First Diff	RMSE Second Diff
D0-C3	0.0954	0.0045	0.0073
D0-C5	0.0945	0.0042	0.0071
D40-C3	0.0932	0.0039	0.0068
D40-C5	**0.0921**	**0.0032**	**0.0065**
D80-C3	0.0946	0.0044	0.0072
D80-C5	0.0944	0.0043	0.0069

Conclusions

- Proposed an LSTM based method to generate talking face landmarks from speech
- Showed how to normalize landmarks and remove the identity information
- Reported objective and subjective evaluation results that are promising

References

[1] Maddox, Ross K and Atilgan, Huriye and Bizley, Jennifer K and Lee, Adrian KC. Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. eLife 4 (2015)