17

Crossmodal action unit sequence estimation and image-to-image mapping

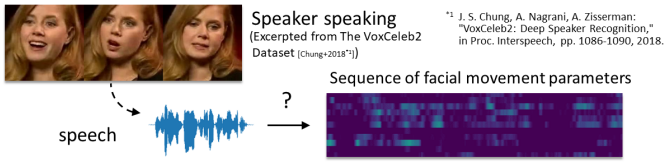
Controlling facial expressions in face image from speech

Abstract

Speech contains not only linguistic information, corresponding to the uttered sentence, but also nonlinguistic information, corresponding to the emotional expression and mood. This information plays an important role in spoken dialogue. This study is the first attempt to estimate the action unit (facial muscle motion parameter) sequence of the speaker from speech alone, assuming that the nonlinguistic information in speech is expressed in the facial expressions of the speaker. Until now, there have been no attempts to estimate action units from speech alone, and how much accuracy could be achieved was not known. This study reveals this for the first time. By combining the action unit sequence estimated from speech with an image-to-image converter, we implemented a system that modifies the facial expression of a still face image in accordance with input speech, making it possible to visualize the expression and mood of speech. Emotional expressions and moods have traditionally been treated symbolically, assigning discrete subjective labels. In contrast, action units are suitable as continuous quantities for expressing emotional expressions and moods, and we have shown that action units can be estimated from speech in this study. In the future, we expect to open up a variety of new applications that simultaneously utilize speech and face images, such as speech synthesis that matches facial expressions and face image generation that matches speech.

Estimating face movement from speech

✓ If face movement can be predicted from speech, ...



*1 J. S. Chung, A. Nagrani, A. Zisserman: "VoxCeleb2: Deep Speaker Recognition," in Proc. Interspeech, pp. 1086-1090, 2018.

- it can be used to visualize nonlinguistic information in speech
- it can be used as useful nonlinguistic-information-related feature for speech synthesis and voice conversion applications

✓ Is this task solvable and how difficult is it?

➡ The aim of this study is to answer these questions

Deep learning approach using speaking face-tracks

✓ As quantities that represent facial movements, we focus on the **facial action units (AUs)**[*2]

*2 **Facial muscular activity units that are related to the contraction or relaxation of specific facial muscles**

✓ Train neural network that predicts AU sequence from speech

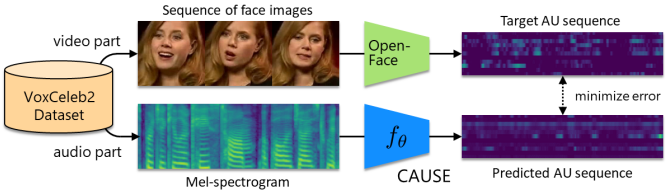
"Crossmodal Action Unit Sequence Estimator (CAUSE)"

Approach

By using many speaking face-tracks, we train CAUSE so that
- AU sequence extracted using OpenFace[*3] from the video part and
- AU sequence predicted by CAUSE from the audio part

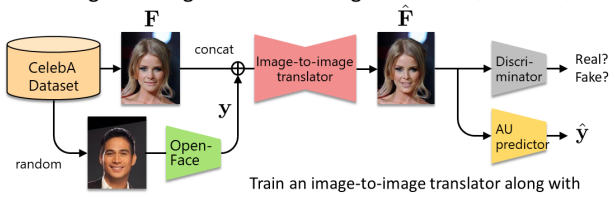
become consistent

*3 Open-source facial behavior analysis toolkit



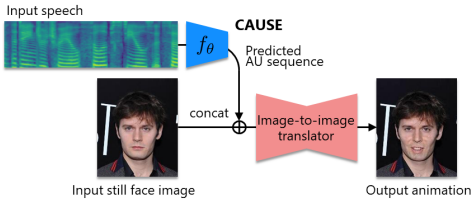
Crossmodal face image control

✓ Train Image-to-image translator using GANimation [Pumarola+2018]



Train an image-to-image translator along with an AU predictor and a discriminator, which scores how realistic a given image is

✓ Convert still face image in accordance with predicted AU sequence



➡ Allows us to control facial expression using speech

※ All the face images are excerpted from The CelebA Dataset [Liu+2015*4]

*4 Z. Liu, P. Luo, X. Wang, X. Tang: "Deep Learning Face Attributes in the Wild," in Proc. ICCV, pp. 3730-3738, 2015.

Face image control experiment

Other examples can be found here:

✓ Examples of animations generated from same speech

✓ Generated animations were more natural when controlled by AUs than when controlled by probability vectors of emotional states (neutral, happiness, surprise, sadness, anger, disgust, fear, contempt)

References

[1] H. Kameoka, T. Kaneko, S. Seki, K. Tanaka, "CAUSE: Crossmodal action unit sequence estimation from speech," submitted to The 23rd Annual Conference of the International Speech Communication Association (Interspeech 2022).

Contact

Hirokazu Kameoka / Recognition Research Group, Media Information Laboratory
Email: cs-openhouse-ml@hco.ntt.co.jp