CONTINUOUS ULTRASOUND BASED TONGUE MOVEMENT VIDEO SYNTHESIS FROM SPEECH

Jianrong Wang,[1] *Yalong Yang*,[2] *Jianguo Wei*,[3*] *Ju Zhang*[1]
[1]School of Computer Science and Technology, Tianjin University
[2]Caulfield School of Information Technology, Monash University
[3]School of Computer Software, Tianjin University

1. Motivation

Visualising tongue movement can:
- contribute to the speech intelligibility
- help learning a second language
- be used in speech therapy

2. Background

- Slient Speech Interface (SSI)
  - Ultrasound devices for acoustics
- Machine Learning (ML)
- Statistical Mapping

3. Framework

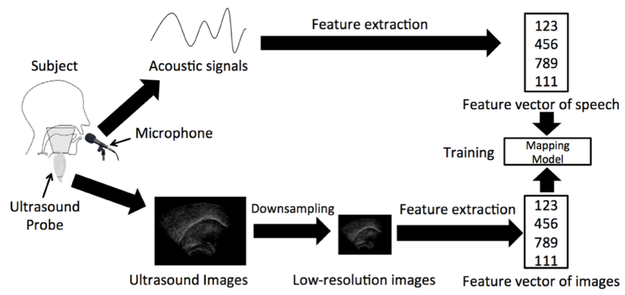


Fig 1. Training mapping model using speech signal and ultrasound images of tongue

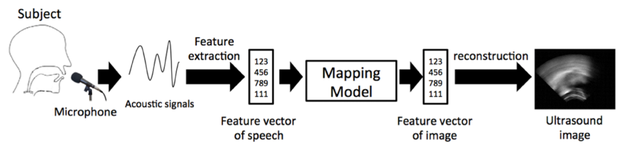


Fig 2. Synthesising tongue movement using speech signal and mapping model

4. Mapping Model

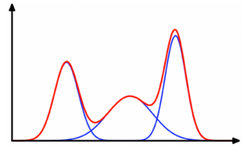
4.1 Vector Quantization (VQ) based

- Build one codebook for each speech class

- For each input frame, using the mean of its top *k* mappings in code-book to generate result

4.2 Gaussian Mixture Model (GMM) based

- Build joint GMM model



- Minimize mean-square error to generate result

5. Data

- Terason T3000 ultrasound system
- a male speaker of Chinese Mandarin
- 44.1K audio
- 90fps 640 × 480 video
- 931 sentences with 6,732s audio and 606,133 ultrasound tongue images.

Result

6. Result

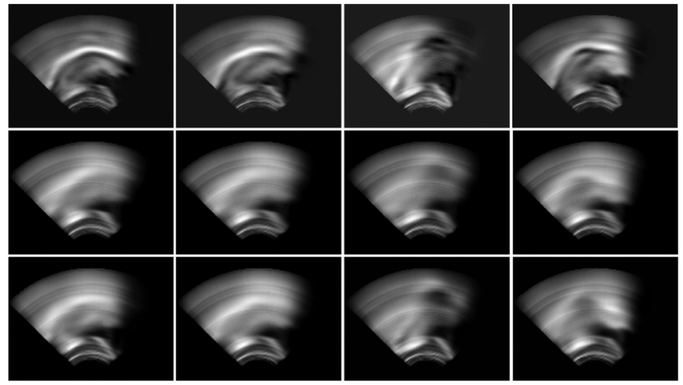


Fig 3. Tongue movement synthesis of one sentence (screenshots);
the firs row is the target; the second row is the result from VQ based method;
third row is the result from GMM based method

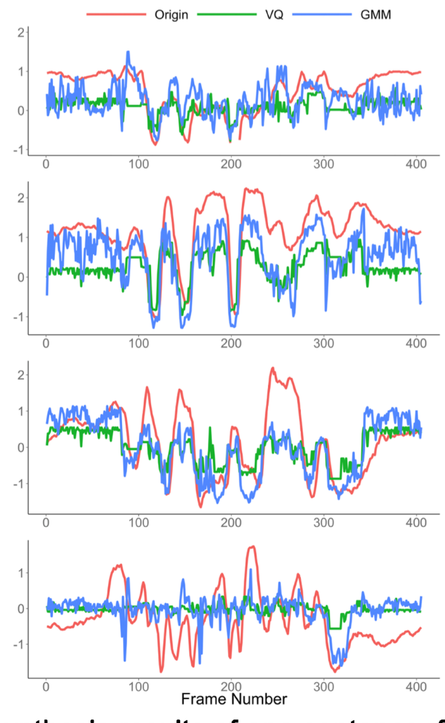


Fig 4. Numeric synthesis results of one sentence, from top to bottom are the results of 1st, 2nd, 3rd, 4th EigenTongue coefficients

Poster Designed by Yalong Yang