How Silent Are Silent Speech Interfaces?

Speech Reconstruction From Whispered and Silent Ultrasound Tongue Images

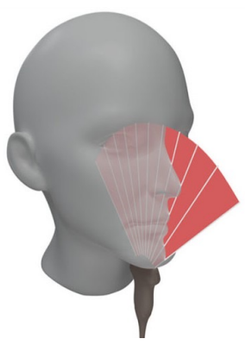
*Gábor Gosztolya, **Ibrahim Ibrahimov**, and Csaba Zainkó*

BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS

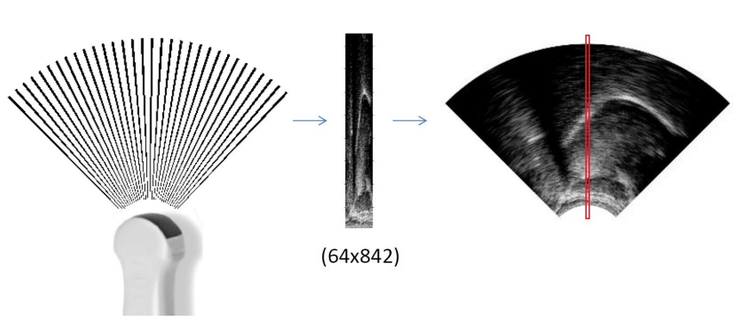
UNIVERSITY OF SZEGED

RGAI

Introduction



Midsaggital ultrasound tongue imaging



Raw vs. wedge representation of ultrasound data

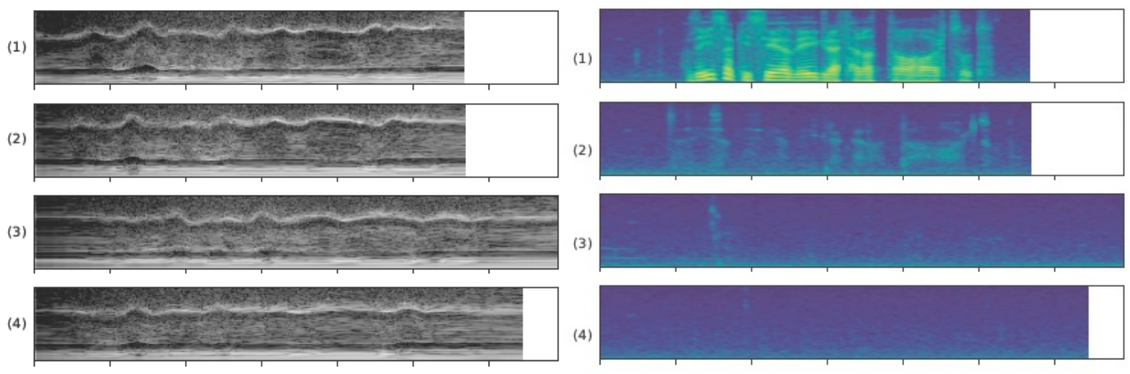
Motivation

- The final goal in articulation-to-speech direct synthesis is to **produce speech** from **silent** articulation
- Yet the recordings typically contain speech and articulatory movements recorded **in parallel**
- But what happens to synthesized speech, if the DNN model is evaluated on **silent** or **whispered** articulation?

Data

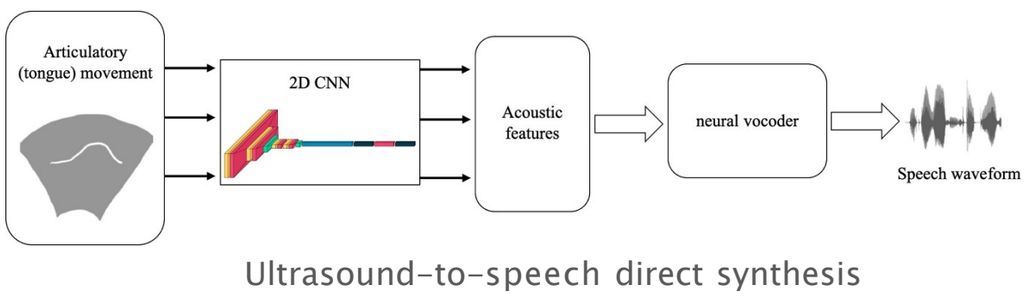
- UTI and parallel speech recordings were used from three speakers (048, 049, 103)
- 200 sentences (~15 minutes) of (Hungarian) speech was used to train the CNN UTI-to-Mel networks with a 190-10 train-dev split
- In addition, the tale `The North Wind and the Sun' (9 sentences) were read in four variations by each speaker (as test):

(1) **Normal**: the sentences were read aloud
(2) **Whispered**: The speakers read the sentences whispering
(3) **Silent (hyperarticulation)**: The sentences were read silently with articulation, moving the tongue and lips in an exaggerated manner.
(4) **Silent (normal articulation)**: The sentences were read silently with articulation. They were asked to retain from hyperarticulation.



Kymograms (left) and Mel-spectrograms (right) for a sentence read by spk103 in four variations

Network structure



Ultrasound-to-speech direct synthesis

- A straightforward 2D-CNN architecture was used
- It processed one 64x128 ultrasound image as input, and produced one 80-dimensional Mel-spectrogram frame:

(1) four convolutional layers (30-60-90-120 filters) (SiLU activation) with max-pooling after every second layer
(2) one fully-connected layer (1000 neurons) (SiLU activation)
(3) the output layer (80 neurons) (linear activation)

- Dropout layers were used after the 2nd, 4th and 5th layers

Evaluation

- We used an **ASR system** to calculate phonetic error rates (PER%) as a proxy for the **intelligibility** of the samples
- A traditional phone-level HMM/DNN system was used with a phone bigram

Phonetic Error Rates

Speaking style	spk048	spk049	spk103
Original recording	21.6%	18.8%	19.5%
Normal (audible)	64.9%	64.8%	51.5%
Whispered	88.8%	86.0%	91.7%
Silent (hyperarticulated)	96.9%	89.6%	96.9%
Silent (normal articulation)	96.7%	96.0%	96.9%

- The **original** recordings show the error of the ASR system
- The **normal** recordings show the error of the 2D-CNN
- The 86-92% PER% shows a mismatch in the movement of the tongue and lips during normal and **whispered** speech
- The 96+% PER% values for the two types of **silent** recordings are even higher – except for spk049, where hyperarticulation led to a similar score as whispering

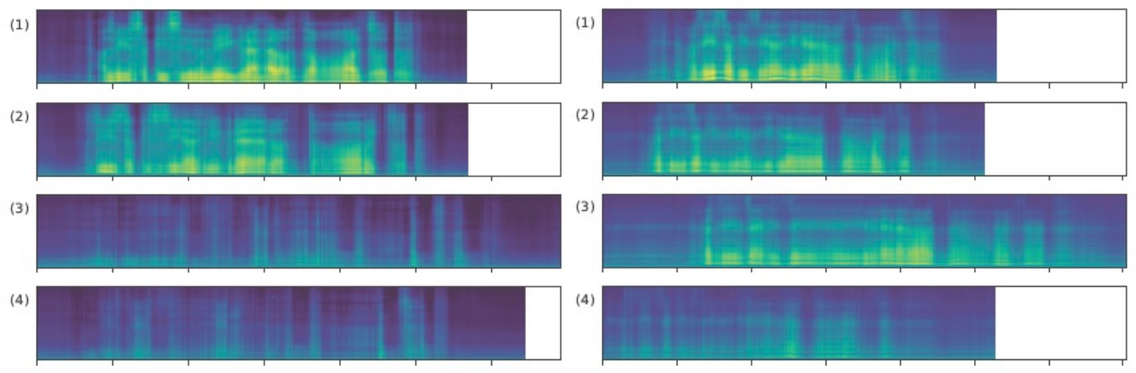
Recognizet phonetic ratios

- There were many deletion errors, probably due to the low volume
- We express the no. of phones recognized / real number of phones

Speaking style	spk048	spk049	spk103
Original recording	100.9%	100.9%	100.3%
Normal (audible)	71.8%	64.8%	83.1%
Whispered	26.6%	32.8%	20.0%
Silent (hyperarticulated)	7.9%	29.5%	4.5%
Silent (normal articulation)	6.4%	11.9%	4.3%

- Even the **normal** recordings contain less phones (i.e. more silence) than the originals do
- The **whispered** ones have even less, i.e. less intense articulatory movements
- The **silent** ones contain almost exclusively silence, except spk049

Recognizet phonetic ratios



The generated Mel-spectrograms for spk103 (left) and spk049 (right)

- The Mel-spectrograms of the generated **whispered** recordings are similar to those generated for the normal recordings, but the timing is different and the formants are more blurred
- The **silent** recordings had quite low intensity for spk103 (for spk048 too) – this led to the ASR system to detect mostly silence…
- For spk049, the generated Mel spectrogram for **hyperarticulated silent** articulation was more similar to the whispered one – and so were the PER% values and the recognized phonetic ratios
- …probably spk049 followed the instruction more precisely…
- The **duration** of the **hyperarticulated** recordings was in general larger than the duration of the other three recording types (for all three speakers)

B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg. Silent speech interfaces. *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

I. Ibrahimov, G. Gosztolya, and T. G. Csapó, Data augmentation methods on ultrasound tongue images for articulation-to-speech synthesis. *Proceedings of SSW*, Grenoble, France, Aug 2023, pp. 230–235.

M. S. Ribeiro, A. Eshky, K. Richmond, and S. Renals. Silent versus modal multi-speaker speech recognition from ultrasound and video. *Proceedings of Interspeech*, Brno, Czech Republic, Sep 2021, pp. 641–645.

M. Wand, M. Janke, and T. Schultz. Investigations on speaking mode discrepancies in EMG-based speech recognition. *Proceedings of Interspeech*, Florence, Italy, Aug 2011, pp. 601–604.

S. Petridis, J. Shen, D. Cetin, and M. Pantic. Visual-only recognition of normal, whispered and silent speech. *Proceedings of ICASSP*, Calgary, AB, Canada, Apr 2018, pp. 6219–6223.

ggabor@inf.u-szeged.hu
HUN-REN-SZTE Research Group on Artificial Intelligence, Szeged, Hungary

ibrahim@tmit.bme.hu
Department of Telecommunications and Media Informatics, BME, Budapest, Hungary

zainko@tmit.bme.hu
Department of Telecommunications and Media Informatics, BME, Budapest, Hungary