TransVoice: Real-Time Voice Conversion for Augmenting Near-Field Speech Communication

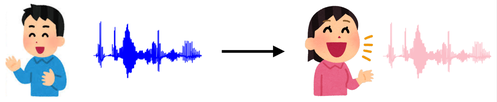
***Riku Arakawa*, Shinnosuke Takamichi, Hiroshi Saruwatari**
The University of Tokyo , arakawa-riku428@g.ecc.u-tokyo.ac.jp, shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
The details of our DNN-based real-time voice conversion algorithm has been shown in 10th ISCA Speech Synthesis Workshop.
→ https://isca-speech.org/archive/SSW_2019/pdfs/SSW10_P_1-10.pdf

Synopsis

- **Identify the problem when we use real-time voice conversion for near-field communication.**
- **Propose a remedy.**

Background and problem

- Deep Neural Network (DNN)-based real-time voice conversion (data-driven speaker conversion) has been established [Arakawa et al., 2019].



voice conversion movie: https://www.youtube.com/watch?v=P9rGqoYnfCg

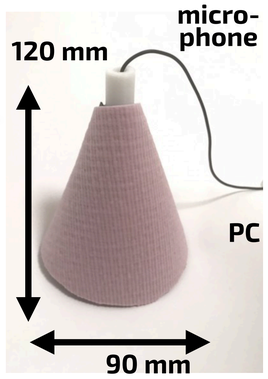
- Lots of applications have been discussed such as film, chat rooms, and gaming environments [Yannis Stylianou, 2009].

However,

- **Speaker's original speech degrades immersion to voice conversion experiences.**
- **Few has focused on using voice conversion for augmenting "near-field speech communication".**

Proposed method

→ Physical mask + Filter + Deep learning



- A physical mask is devised to confine the original voice.
- To ameliorate the conversion quality, a filter is applied to weaken low frequency range amplified by muffling.
- Then a DNN model is trained on filtered speech of a given source - target pair.

Experiment

(1) Conversion quality
→ compare performance of trained DNN that transforms speech features of the source speaker to those of the target speaker.
(2) Soundproof effect
→ compare speech volume at a close point (~1 m).

Result

→ Quality improvement and soundproof effect

(1) Mean Squared Error of predicted mel-cepstral coefficients

without a mask	with a mask without a filter	with a mask with a filter
0.918 ± 6.8e-3	0.960 ± 4.3e-3	**0.928 ± 3.0e-3**

(2) Average Root Mean Square of speaker's original speech

	without a mask	with a mask
at a mask device	63.2	70.8
at a listener	38.3	**14.6**

Future direction

- Explore applications of real-time voice conversion.
- Investigate their effects in speech communication to cognitive reactions.

"Does changing our daily speech to another's influence our personality?"