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

Riku Arakawa
The University of Tokyo
Tokyo, Japan
arakawa@star.rcast.u-tokyo.ac.jp

Shinnosuke Takamichi
The University of Tokyo
Tokyo, Japan
shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

Hiroshi Saruwatari
The University of Tokyo
Tokyo, Japan
hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

ABSTRACT
Despite promising initial studies, a speaker's original voice can cause problems when it comes to the application of real-time voice conversion (data-driven speaker conversion) technology in our daily lives, specifically in our near-field communication, because the overlapping speech degrades the sense of immersion to the converted speech. We present *TransVoice*, a real-time voice conversion system that physically confines original speech with a mask-shaped device. Our preliminary study shows the proposed device can reduce the volume of original speech significantly, while it ameliorates the deteriorated conversion quality of the deep neural network (DNN) thanks to an integrated filter that weakens the low frequency range. We discuss novel applications using TransVoice that can augment our communication.

CCS Concepts
•Human-centered computing → Sound-based input / output;

Author Keywords
speech communication, voice conversion, deep neural network

INTRODUCTION
Speech plays an important role in people's lives and is presumably the most natural method of communication. It conveys not only linguistic information but also emotions, nuances, and speaker identities. Hence, it is expected that our speech communication can be enhanced by converting these non-linguistic parameters to achieve one's desired voice, which is referred to as *voice conversion* [8]. Very recently, Arakawa et al. [1] has established a DNN-based voice conversion that works in real-time, which further expands its possibilities. While some applications have been discussed in film, chat rooms, and gaming environments [7], this technology has not been explored thoroughly in the HCI context, especially toward augmenting near-field speech communication where listeners are close enough to a speaker to hear their original voice. In such situations, voice conversion can expand listeners' perceptive experiences by keeping the sense of the speaker's presence in front of them. For example, it could be used to enable costumed characters at amusement parks to speak freely to children as if they were characters from movies or cartoons.

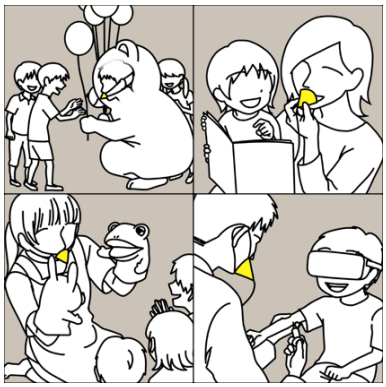

Figure 1. TransVoice facilitates augmented communication with converted speech. Diminishing the original voice can immerse people in the experience.

However, in spite of its promising applications, there is an obstacle to achieving an augmentation of speech communication in this way. The problem is that the contamination of the original speech makes it hard for others to listen, thus crucially degrading immersion in the experiences. In this paper, we propose *TransVoice*, a system featuring a mask-shaped device of soundproof material for use in real-time voice conversion, and conduct a preliminary study on how it prevents the original speech from being heard by nearby listeners. Furthermore, as a physical mask such as this aggravates the quality of the output speech, we consider introducing a filtering technique, whose effectiveness is also assessed. Finally, we discuss its potential applications in the context of near-field speech communication to demonstrate the importance of making the original speech inaudible.

RELATED WORK
The HCI literature has favored approaches called silent speech where output speech is estimated from an input that is silent or extremely small. AlterEgo enables us to converse with a computing device without any voice [4]. Moreover, Toda and Shikano achieved conversion from a non-audible murmur to

UIST '19 Adjunct, October 20-23, 2019, New Orleans, LA, USA.

ACM ISBN 978-1-4503-6817-9/19/10 ...$15.00.
http://dx.doi.org/10.1145/3332167.3357106

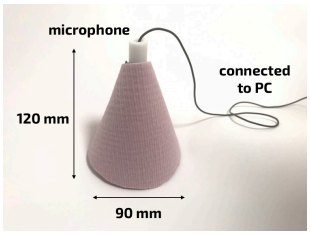

Figure 2. Proposed mask-shaped device.

Table 1. MSE of trained DNNs.

(a) without a mask	(b) with a mask without a filter	(c) with a mask with a filter
$0.918 \pm 6.8 \times 10^{-3}$	$0.960 \pm 4.3 \times 10^{-3}$	$0.928 \pm 3.0 \times 10^{-3}$

Table 2. Average RMS of speaker's original speech.

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

ordinary speech [9], and Kimura et al. achieved it by recognizing ultrasound images of the throat [5]. However, these studies aim to transmit linguistic information, not to provide pitch information to the output speech, and as a result, it is nearly impossible for users to generate a variety of expressions, e.g., those related to emotions and emphasis.

Our proposed approach is unique in that it physically confines the original voice using a mask-shaped device without losing any of this information. Furthermore, in contrast to approaches to suppress one's voice (e.g., [3]), which often require heavy computation, ours has no such computational constraint, thus making it suitable for real-time usage.

METHOD

Mask-shaped device prototype

Figure 2 shows our prototype device. The mask body is covered with conventional polyvinyl chloride material to make it soundproof [6].

Real-time voice conversion

In this work, we follow the implementation of real-time voice conversion proposed by Arakawa et al. [1]. The system is built using pre-recorded voices. We filter these voices to weaken the low frequency region amplified by muffling due to the mask. The filter reduces the frequency band below 100 Hz by 10 dB. Then we train a DNN to convert the filtered speech into a target speech.

PRELIMINARY STUDY

Setting

Conversion quality of DNNs

A multi-layer perceptron composed of 2 $\times$ 1000-unit leaky ReLU hidden layers was used to map a speaker's mel-cepstral coefficients (0th-through-39th) to a target speaker's. To train the DNNs, we used mean squared error (MSE) for the loss function. As a corpus, we used the Voice Actress Corpus [10] (100 utterances, approx. 12 min.) uttered by two men was used, 90 % of which was for training and the remainder for evaluation. The sampling frequency was 16 kHz. The lengths of the frame and the FFT were 400 (25 ms) and 512, respectively. We compared three conditions to examine the filtering effect: (a) without a mask, (b) with a mask and without a filter, and (c) with a mask and with a filter.

Volume of original voice

First, one of the authors got accustomed to converting his voice both with/without the mask, spending around three minutes in a tidy conventional room of about ten square meters. During this process, he could hear the converted speech as well as his own voice. Then we asked him to speak a given sentence in the same manner, but the converted speech was not heard this time. Once we confirmed that he did not significantly change his way of speaking due to listening, we recorded the sound simultaneously with two microphones, one installed in the system itself and one located at the position of an assumed nearby listener. Whether he used the mask or not, the input microphone was about 10 cm away from the front of the mouth and the latter microphone was fixed about 1 meter away from the mouth in the same direction. As a result, we obtained two pairs of about ten seconds of waveform, each of which corresponds to whether he used the mask.

Results

Table 1 summarizes the MSE between the predicted and target mel-cepstral coefficients of the evaluation data. It is clear that the introduced filter recovers the conversion quality.

Next, after removing the silent parts from the recorded speech, we calculated the waveform amplitude in the root mean square (RMS), which is shown in Table 2. These results confirm the significant soundproof effect of the proposed mask.

APPLICATION EXAMPLE

Finally, in order to envisage the future with TransVoice, potential application examples are given in this section (see Figure 1). For example, we can enjoy entertainment with costumed characters (top-left), a picture book (top-right), and a puppet show (bottom-left). Another example is an injection with the VR experience, where patients might feel less pain thanks to being immersed in a virtual reality where they can listen to the doctor's converted speech (a similar effect with HMD has been reported [2]) (bottom-right).

FUTURE WORK

Although a simple filter is a remedy, the conversion quality becomes lower due to the introduction of a mask. This muffled influence can be addressed by designing the filter carefully or by introducing a data augmentation method for the training [1]. In addition, we aim to investigate new application scenarios and are also interested in whether transforming speech using TransVoice can affect listeners' cognitive reactions or the speaker's personality.

ACKNOWLEDGEMENT

This research and development work was supported by the MIC/SCOPE #182103104.

References

[1] Riku Arakawa, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2019. Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device. In *The 10th ISCA Speech Synthesis Workshop (to appear)*.

[2] Rudnick Chad, Sulaiman Emaan, and Orden Jillian. 2018. Effect of virtual reality headset for pediatric fear and pain distraction during immunization. *Pain management* 8, 3 (2018), 175–179.

[3] Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani. 2018. Single channel target speaker extraction and recognition with speaker beam. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Calgary, Canada, 5554–5558.

[4] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces*. ACM, Tokyo, Japan, 43–53.

[5] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow, United Kingdom, 146.

[6] Ji-Zhao Liang and Xing-Hua Jiang. 2012. Soundproofing effect of polypropylene/inorganic particle composites. *Composites Part B: Engineering* 43, 4 (2012), 1995–1998.

[7] Yannis Stylianou. 2009. Voice transformation: a survey. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Taipei, Taiwan, 3585–3588.

[8] Tomoki Toda, Alan W Black, and Keiichi Tokuda. 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 8 (2007), 2222–2235.

[9] Tomoki Toda and Kiyohiro Shikano. 2005. NAM-to-speech conversion with Gaussian mixture models. In *Proceedings of the INTERSPEECH2005 - the 9th European Conference on Speech Communication and Technology*. ISCA, Lisbon, Portugal, 1957–1960.

[10] y_benjo and MagnesiumRibbon. 2017. Voice-Actress Corpus. http://voice-statistics.github.io/. (2017).