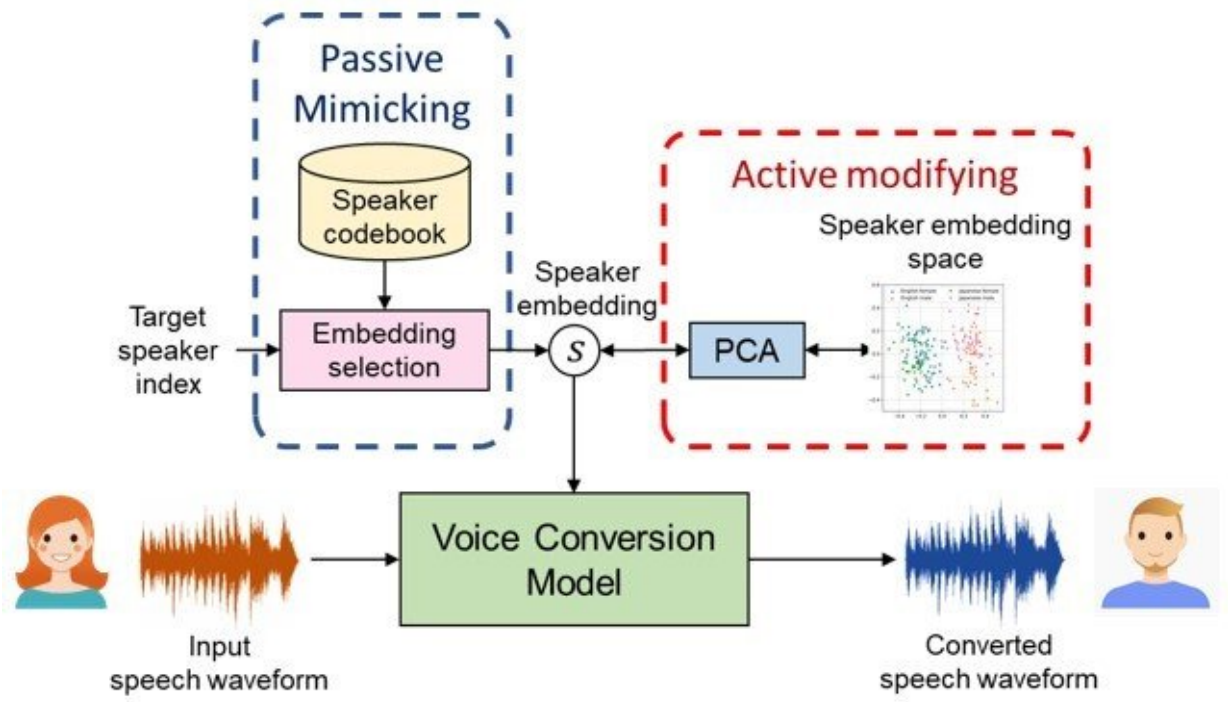
Sounds familiar: A speaker identity-controllable framework for machine speech translation

April 26 2021



Voice conversion is carried out by selecting target speaker embedding from speaker codebook. Voice characteristic can be independently controlled via principal components of speaker embedding. Credit: Masato Akagi

Robots today have come a long way from their early inception as insentient beings meant primarily for mechanical assistance to humans.

Today, they can assist us intellectually and even emotionally, getting ever better at mimicking conscious humans. An integral part of this ability is the use of speech to communicate with the user (smart assistants such as Google Home and Amazon Echo are notable examples). Despite these remarkable developments, they still do not sound very "human."

This is where voice conversion (VC) comes in. A technology used to modify the speaker identity from one to another without altering the linguistic content, VC can make the human-machine communication sound more 'natural' by changing the non-linguistic information, such as adding emotion to speech. "Besides linguistic information, non-linguistic information is also important for natural (human-to-human) communication. In this regard, VC can actually help people be more sociable since they can get more information from speech," explains Prof. Masato Akagi from Japan Advanced Institute of Science and Technology (JAIST), who works on speech perception and speech processing.

Speech, however, can occur in a multitude of languages (for example, on a language-learning platform) and often we might need a machine to act as a speech-to-speech translator. In this case, a conventional VC model experiences several drawbacks, as Prof. Akagi and his doctoral student at JAIST, Tuan Vu Ho, discovered when they tried to apply their monolingual VC model to a "cross-lingual" VC (CLVC) task. For one, changing the speaker identity led to an undesirable modification of linguistic information. Moreover, their model did not account for cross-lingual differences in "F0 contour," which is an important quality for speech perception, with F0 referring to the fundamental frequency at which vocal cords vibrate in voiced sounds. It also did not guarantee the desired speaker identity for the output speech.

Now, in a new study published in *IEEE Access*, the researchers have proposed a new model suitable for CLVC that allows for both voice

mimicking and control of speaker identity of the generated speech, marking a significant improvement over their previous VC model.

Specifically, the new model applies language embedding (mapping natural language text, such as words and phrases, to mathematical representations) to separate languages from speaker individuality and F0 modeling with control over the F0 contour. Additionally, it adopts a deep learning-based training model called a star generative adversarial network, or StarGAN, apart from their previously used variational autoencoder (VAE) model. Roughly put, a VAE model takes in an input, converts it into a smaller and dense representation, and converts it back to the original input, whereas a StarGAN uses two competing networks that push each other to generate improved iterations until the output samples are indistinguishable from natural ones.

The researchers showed that their model could be trained in an end-to-end fashion with direct optimization of language embedding during the training and allowed good control of speaker identity. The F0 conditioning also helped remove language dependence of speaker individuality, which enhanced this controllability.

The results are exciting, and Prof. Akagi envisions several future prospects of their CLVC model. "Our findings have direct applications in protection of speaker's privacy by anonymizing one's identity, adding sense of urgency to speech during an emergency, post-surgery voice restoration, cloning of voices of historical figures, and reducing the production cost of audiobooks by creating different voice characters, to name a few," he comments. He intends to further improve upon the controllability of speaker identity in future research.

Perhaps the day is not far when smart devices start sounding even more like humans.

More information: Tuan Vu Ho et al, Cross-Lingual Voice Conversion With Controllable Speaker Individuality Using Variational Autoencoder and Star Generative Adversarial Network, *IEEE Access* (2021). DOI: 10.1109/ACCESS.2021.3063519