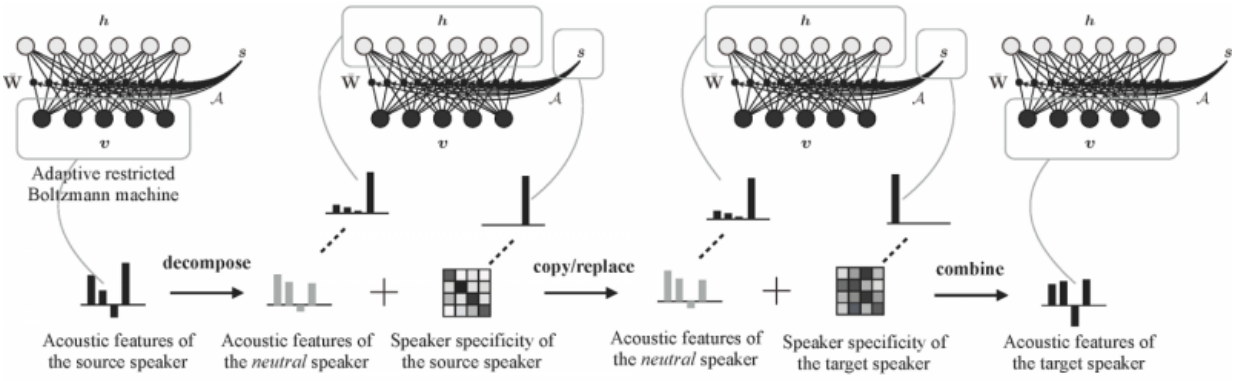
Speech signal processing—enhancing voice conversion models

December 27 2016



Researchers in Japan have created a new voice conversion method using an adaptive restricted Boltzmann machine - a model capable of deconstructing speech and rebuilding it to sound like a different person speaking. Crucially, this model works without the need for parallel data from two speakers for training, meaning target voices can say words and sentences not used in training. Credit: University of Electro-Communications

Altering a person's voice so that it sounds like another person is a useful technique for use in security and privacy, for example. This computational technique, known as voice conversion (VC), usually requires parallel data from two speakers to achieve a natural-sounding conversion. Parallel data requires recordings of two people saying the same sentences, with the necessary vocabulary, which are then time-matched and used to create a new target voice for the original speaker.

However, there are issues surrounding parallel data in speech processing, not least a need for exact matching vocabulary between two speakers, which leads to a lack of corpus for other vocabulary not included in the pre-defined model training. Now, Toru Nakashika at the University of Electro-Communications in Tokyo and co-workers have successfully created a model capable of using non-parallel data to create a target voice - in other words, the target voice can say sentences and vocabulary not used in model training.

Their new VC method is based on the simple premise that the acoustic features of speech are made up of two layers - neutral phonological information belonging to no specific person, and 'speaker identity' features that make words sound like they are coming from a particular speaker. Nakashika's model, called an adaptive restricted Boltzmann machine, helps deconstruct speech, retaining the neutral phonological information but replacing speaker specific information with that of the target speaker.

After training, the model was comparable with existing parallel-trained models with the added advantage that new phonemic sounds can be generated for the target speaker, which enables speech generation of the target speaker with a different language.

More information: Toru Nakashika et al. Non-Parallel Training in Voice Conversion Using an Adaptive Restricted Boltzmann Machine, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2016). DOI: 10.1109/TASLP.2016.2593263

Provided by University of Electro-Communications