Voice Conversion for Stuttered Speech, Instruments, Unseen Languages and Textually Described Voices

Matthew Baas[0000−0003−3001−6292] and Herman Kamper[0000−0003−2980−3475]

MediaLab, Electrical & Electronic Engineering, Stellenbosch University, South Africa
{20786379,kamperh}@sun.ac.za

Abstract. Voice conversion aims to convert source speech into a target voice using recordings of the target speaker as a reference. Newer models are producing increasingly realistic output. But what happens when models are fed with non-standard data, such as speech from a user with a speech impairment? We investigate how a recent voice conversion model performs on non-standard downstream voice conversion tasks. We use a simple but robust approach called k-nearest neighbors voice conversion (kNN-VC). We look at four non-standard applications: stuttered voice conversion, cross-lingual voice conversion, musical instrument conversion, and text-to-voice conversion. The latter involves converting to a target voice specified through a text description, e.g. "a young man with a high-pitched voice". Compared to an established baseline, we find that kNN-VC retains high performance in stuttered and cross-lingual voice conversion. Results are more mixed for the musical instrument and text-to-voice conversion tasks. E.g., kNN-VC works well on some instruments like drums but not on others. Nevertheless, this shows that voice conversion models – and kNN-VC in particular – are increasingly applicable in a range of non-standard downstream tasks. But there are still limitations when samples are very far from the training distribution. Code, samples, trained models: https://rf5.github.io/sacair2023-knnvc-demo/.

Keywords: Voice conversion · Speech processing · Speech synthesis · Instrument conversion · Stuttered speech