Investigating Inter- and Intra-speaker Voice Conversion using Audiobook

Aghilas Sini[1], Damien Lolive[1] , Nelly Barbot[1], Pierre Alain[1]

sini.aghilas@gmail.com, {damien.lolive,nelly.barbot,pierre.alain}@irisa.fr

[1]Univ Rennes, CNRS, IRISA, 22300 Lannion, France

Introduction

Motivation : when reading aloud a book, a speaker applies significant voice modifications between narration and dialog parts.

⇒ Can a Voice Conversion (VC) system reproduce these intra-speaker modifications?

Framework :

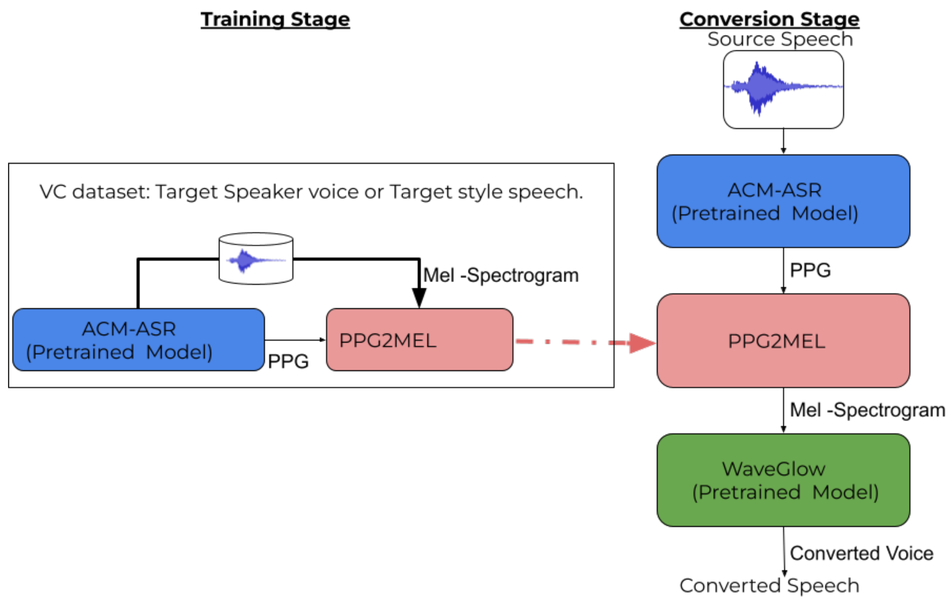
- VC system based on Phonetic Posteriorgrams (PPG), which are temporal representations of phoneme probabilities
- Database composed of audiobooks in french language

1 - Voice Conversion System

Origin : VC system inspired from [1], designed for foreign accent conversion to address the 2020 VC Challenge

Structure : the conversion system is composed of 3 main blocks

1. The ACM-ASR model [2] is a pre-trained TDNN-HMM acoustic model that extracts PPG from input source speech.
2. The PPG-to-Mel model, derived from Tacotron2 [3], predicts mel-spectrogram from PPG. It is trained on target speaker data.
3. The WaveGlow vocoder is a flow-based network generatingg high quality speech from mel-spectrograms [4]. It is fine-tuned for each target speaker.



2 - Experimental Setup

All voices used for experiments are extracted from MUFASA[a]

Inter-Speaker conversion dataset : 3 female and 3 male voices

- 2 audiobooks (*Boule de suif, Petite comtesse*), each read by 3 speakers (female for one, male for the other) in same recording conditions
- train/validation/test sets composed of 1430/30/30 sentences per voice ($\sim$ 1.5 hour)

Intra-Speaker conversion dataset : 1 Direct Speech style voice (DS), 1 Indirect Speech style voice (IS) with a same speaker

- Subcorpus Synpaflex [5]: audiobook *Les misérables* read by a female speaker
- Manual annotated character-based speaking turns (DS) and narrator discourse parts (IS)
- For each style, 1 hour set to train PPG-to-Mel model, 35 sentences to test

[a]http://aghilassini.github.io/demo/mufasa/index.html

3 - Test 1: Speaker conversion speech quality

To evaluate the fine-tuned vocoder and the VC system quality, a CMOS test is done:

- 5-point scale (1 indicates bad and 5 indicates excellent)
- 2 samples with the same linguistic content
  1. *VocTargetVoice*, the target voice sample re-synthesized by the vocoder
  2. *ConvIntraGen* or *ConvInterGen*, provided by the VC system :
     ConvIntraGen: different source and target speakers, with the same gender
     ConvInterGen: source and target speakers have different gender
- 12 french native listeners have annotated the quality of 60 samples each

Results :

Configuration	MOS score
VocTargetVoice	4.03 $\pm$ 0.13
ConvInterGen	2.50 $\pm$ 0.20
ConvIntraGen	2.60 $\pm$ 0.20

VocTargetVoice is a good upper bound to evaluate this VC approach

Converted signal quality is similar to state-of-the-art comparable VC systems

4 - Test 2: Inter-speaker similarity

To assess the similarity between converted and target voices, a MUSHRA test is done:

- *VocTargetVoice* (the vocoded target voice sample) is given as reference
- 5 candidates to evaluate, with the same linguistic content as *VocTargetVoice*
  - 3 converted signals: *ConvInterGen*, *ConvIntraGen* and, as upper bound, *TTC* (the target voice signal is given as input of the VC system)
  - 2 vocoded signals of source speakers different from the target one, as lower bound:
    VocInterGen, source and target speakers with different gender
    VocIntraGen, source and target speakers with the same gender
- 17 french native listeners have annotated 100 samples each

Results :

Configuration	*Conv*	*Voc*
InterGender	54.40 $\pm$ 3.27	18.81 $\pm$ 3.50
IntraGender	50.72 $\pm$ 3.20	25.39 $\pm$ 3.14
TTC	61.82 $\pm$ 3.20	-

As expected, *TTC* gives the best similarity, but its performance is moderate

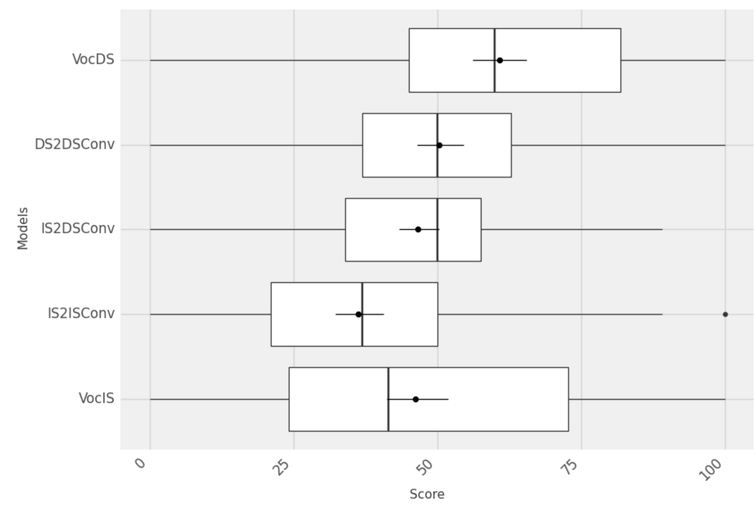
ConvInterGen and *ConvIntraGen* give promising results.

5 - Test 3: Intra-speaker similarity

The ability of VC system to convert IS style to DS one is assessed by a MUSHRA test

- 2 references: vocoded IS and vocoded DS samples
- 5 candidates whose IS/DS style proximity is to be rated, from 0 (IS) to 100 (DS):
  - 3 converted samples in configurations IS-to-IS, DS-to-DS, IS-to-DS
  - 2 vocoded samples: one with IS style, one with DS style

Results :



- The vocoded IS samples seem to be hard to be recognized as IS style samples
- The vocoded DS samples are fairly associated to DS style
- IS-to-IS converted samples are better recognized as IS than vocoded IS samples

Conclusion

- One of rare studies to apply a PPG-based VC system in french language.
- Quality of converted samples similar to ones generated by state of the art VC systems.
- Promising results of similarity listening tests: speaker identity can be changed if speakers are different and preserved otherwise.
- This work is a first attempt to convert speech from indirect style to direct style.

Future works :

- The use of a prosodic measure representing the identity of speaker is necessary [6].
- Additional subjective evaluation would be done to assess intra-speaker conversion.

References

[1] Zhao, G., Ding, S. and Gutierrez-Osuna, R., Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams, Interspeech 2019.

[2] Peddinti, V., Povey, D. and Khudanpur, S., A time delay neural network architecture for efficient modeling of long temporal contexts, Interspeech 2015.

[3] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z. et al., Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, ICASSP 2018.

[4] Prenger, R., Valle, R., Catanzaro, B., Waveglow: A Flow-based Generative Network for Speech Synthesis, ICASSP 2019.

[5] Sini, A., Lolive, D., Vidal, G., Tahon, M. et al., SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis, LREC 2018.

[6] Sini, A., Le Maguer, S., Lolive, D. et al., Introducing Prosodic Speaker Identity for a Better Expressive Speech Synthesis Control, Speech Prosody 2020.

LREC, 2022 — Marseille, France