Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7305–7313
Marseille, 20-25 June 2022

Investigating Inter- and Intra-speaker Voice Conversion using Audiobooks

Aghilas Sini, Damien Lolive, Nelly Barbot, Pierre Alain
Univ Rennes, CNRS, IRISA
6 Rue de Kerampont CS 80518, F-22305 Lannion, France

Abstract

Audiobook readers play with their voices to emphasize some text passages, highlight discourse changes or significant events, or in order to make listening easier and entertaining. A dialog is a central passage in audiobooks where the reader applies significant voice transformation, mainly prosodic modifications, to realize character properties and changes. However, these intra-speaker modifications are hard to reproduce with simple text-to-speech synthesis. The manner of vocalizing characters involved in a given story depends on the text style and differs from one speaker to another. In this work, this problem is investigated through the prism of voice conversion. We propose to explore modifying the narrator's voice to fit the context of the story, such as the character who is speaking, using voice conversion. To this end, two complementary experiments are designed: the first one aims to assess the quality of our Phonetic PosteriorGrams (PPG)-based voice conversion system using parallel data. Subjective evaluations with naive raters are conducted to estimate the quality of the signal generated and the speaker similarity. The second experiment applies an intra-speaker voice conversion, considering narration passages and direct speech passages as two distinct speakers. Data are then non parallel and the dissimilarity between character and narrator is subjectively measured.

Keywords: Voice Conversion, Expressive Speech, Audiobook, PPG

1. Introduction

Audiobook generation is a hard task for which speech synthesis systems need to be able to mimic both the narrator and also different characters, with sometimes very different characteristics. Usually, audiobook readers play with their voices to emphasize some text passages, highlight discourse changes or significant events, or in order to make listening easier and entertaining. For instance, dialogs are central passages in audiobooks where the reader applies significant voice transformation, mainly prosodic modifications, to realize character properties and changes. However, these intra-speaker modifications are hard to reproduce with simple text-to-speech synthesis. The manner of vocalizing characters involved in a given story depends on the text style and differs from one speaker to another. In this work, this problem is investigated through the prism of voice conversion.

Voice Conversion (VC) aims to transform the speech of a source speaker, without changing the linguistic information, such that it seems to be uttered by a target speaker. During the last decade, VC techniques have known a large development due to the flexibility of the recent synthesis technologies based on advanced deep learning methods. They allow the design of new applications requiring voice modification or a large variety of voices.

In 2016, a bi-annual Voice Conversion Challenge (VCC) has been launched to provide a common framework to compare state-of-the-art VC techniques on a same task and a common dataset. The task of VCC 2020 is to perform cross-lingual VC, considering non-parallel training over different languages. In this VCC edition, four voice conversion kinds can be distinguished: a) combination of Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) (Huang et al., 2020), this method uses in cascade an ASR system for transcribing input speech to text and a TTS system to generate speech with target voice; b) PPG-based methods (Sun et al., 2016; Tian et al., 2018; Liu et al., 2018; Liu et al., 2021; Tian et al., 2020; Zheng et al., 2020) that use a temporal representation of phonemes derived from the source speech, c) auto-encoder based approaches (Tobing et al., 2020; Ho and Akagi, 2020) and d) Generative Adversarial Network (GAN)-based approaches (Tobing et al., 2020).

According to the results of subjective evaluations, it turns out that the best conversion systems are those based on Phonetic Posteriorgrams (PPG) (Liu et al., 2020; Zheng et al., 2020). A PPG represents the temporal evolution of the phoneme distribution. Since PPGs are derived from the phonetic-acoustic representation of the input speech (source voice), it makes them speaker-independent and language modeling independent. In the context of voice conversion, computing a PPG removes the speaker dependent information from the signal and keeps only the linguistic content, which can simply be followed by a target speaker dependent speech synthesis model. Contrary to the PPG approaches, the ASR-TTS cascading method strongly depends on the ASR performance. Nevertheless, using text as a pivot also helps to remove the speaker dependent information. For the two last kinds of methods, one can notice that GAN and Auto-encoder systems are hard to be trained and require a substantial quantity of data. This is a serious drawback in the context of voice conversion where usually, the quantity of data from the target speaker is reduced.

In this paper, we propose to explore the modification of the narrator's voice to fit the context of the story,

such as the character who is speaking, using voice conversion. To do so, we investigate the inter and intra-speaker voice conversion. In this context, a PPG-based approach seems to be best suited for several reasons. First, for some characters in audiobooks, data may be available in a small quantity. Second, this kind of approaches does not need source speaker data for the training phase. A multi-speaker PPG extraction model gives good results without needed to re-train a speaker specific model. Thus this kind of approaches is more easily usable in real applications. Third, this is the state-of-the-art method for voice conversion that provides the best results.

The main contributions of this work are three-fold: (1) to tackle the feasibility of the intra-speaker voice conversion, we apply an any-to-one voice conversion model and propose to train models using amateur audiobooks ; (2) two datasets in French are presented and used in this work including a parallel dataset containing six different speakers for studying inter-speaker voice conversion, and a non-parallel dataset of single female speaker to address the intra-speaker voice conversion ; (3) results show that voice conversion methods can be used to convert indirect speech style to direct speech style using subjective evaluations.

This article is structured as follows. Section 2 presents the system details set up for inter- and intra-speaker voice conversion. The materials and the models specifications during the training and inference stages are presented in Section 3. The evaluation protocol is then presented in Section 4 and the results are presented and discussed in Section 5.

2. Voice Conversion System

This work has been mainly inspired from the system presented in (Zhao et al., 2019). The main modifications are the use of the French language and the conversion between two different speakers or different styles in the same language instead of foreign accent conversion. More precisely, we consider two voice conversion tasks in the French language framework: (1) the voice conversion between two different speakers, called inter-speaker case, and (2) the voice conversion between two distinct speech styles, Indirect Speech (IS) and Direct Speech (DS) for a same speaker, called intra-speaker case.

The conversion system contains three main blocks: the first one extracts PPGs from input speech, the second one predicts Mel-spectrogram from these PPGs, and finally, a vocoder generates target speech from Mel-spectrogram.

The global functioning of the system is described in Figure 1 and more detailed in this section.

During the training stage, the PPG-to-Mel model is learnt specifically for the target voice. It is based on PPGs derived from the target speech signals using a pre-trained ACM-ASR model (Peddinti et al., 2015b; Povey et al., 2016; Peddinti et al., 2017) and is independent of the source voice. As for the inference stage, input PPGs are extracted from the source speech and converted to output speech using the PPG-to-Mel model specific to the target voice.

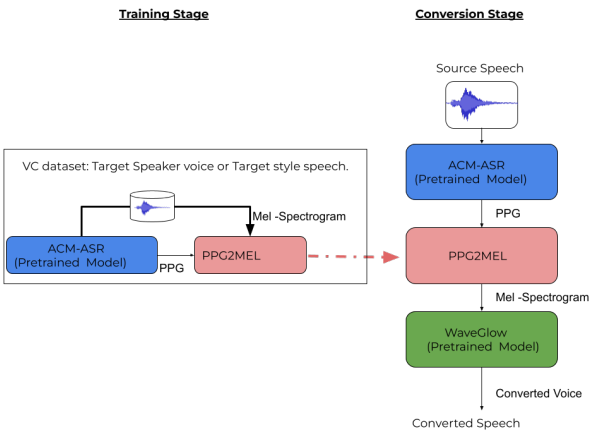


Figure 1: PPG based voice conversion framework

2.1. PPG Extraction

A PPG sequence is a temporal representation of phonemes' probabilities. This representation is less complex to derive from a speech signal than its textual content; it is also easier to generate speech from PPGs than from text. This makes it possible not to worry about the meaning of sentences, especially for homonyms. PPGs also allow to keep temporal information, like the duration of each phoneme. A PPG (Fig. 2 illustrate an example of PPG) can be thought of as an intermediate representation making the link between the phonetic level and the acoustic level. Moreover, PPGs have the advantage of being generally considered speaker-independent, so there is little processing when converting voices.

2.2. PPG to Mel-spectrogram Model

The system presented in Figure 3 has been originally proposed to convert the foreign accent of a speaker so as to be perceived as the accent of a native speaker. Thus, it has been introduced as an accent conversion system, from non-native to native speech. To adapt it to our problem, we can see how the narrator speaks during the direct speech of a character as an accent. The narrator pronounces the same sentence but in a slightly different way. We thus seek, in a way, to convert the native "accent" of the speaker into the "accent" voluntarily put for each character.

The PPG is first extracted from the source speaker signal. It is then passed through the PPG-to-Mel model in order to generate the associated Mel-spectrogram, including the phonetic content as spoken by the source speaker and the voice characteristics of the target speaker. This step depends on the target speaker: the PPG is supposed to be speaker independent, unlike the Mel-spectrogram. Therefore, for each target speaker,

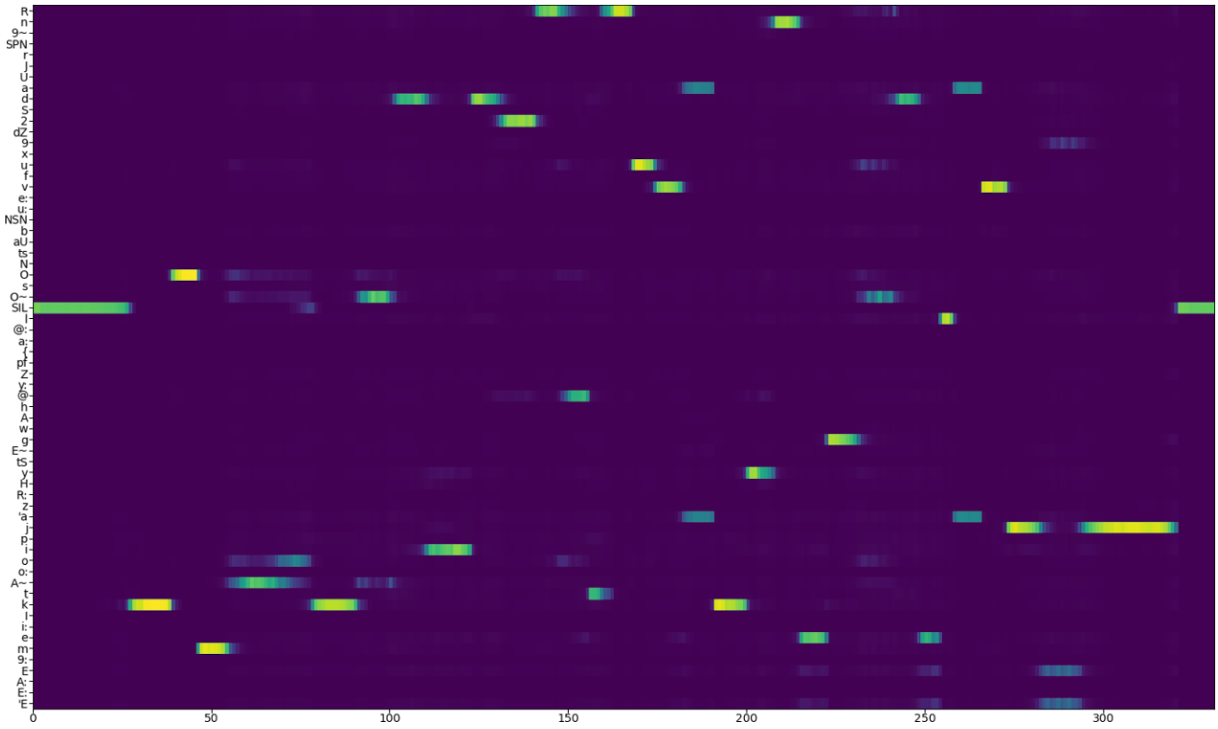


Figure 2: Example of ACM-ASR model output - PPG features, from the matrix we can derive the most likely phonemes sequence which is: /**sil** kɔmã kã did ʁətʁuva kynegɔ̃ d e la vjɛj **sil**/. The reference transcription : "Comment Candide retrouva Cunégonde et la vieille." [in english. How Candide found Cunegonde and the Old Woman]

training a model PPG-to-Mel is necessary. The generated Mel-spectrogram is then passed to the vocoder trained for the target speaker to generate the desired waveform.

The PPG-to-Mel framework is derived from Tacotron2 (Shen et al., 2018), which initially predicts a spectrogram from the input text. Since the input is different (PPG instead of text), the first processing steps are modified. The given input is then passed to a PreNet network and then encoded. An attention mechanism is used to focus the network on the most important parts to predict the Mel-spectrogram and help the system stop.

2.3. Mel-spectrogram to Speech

A WaveGlow vocoder is used to convert the output of the PPG-to-Mel model back into a speech waveform. WaveGlow is a flow-based (Prenger et al., 2019) network capable of generating high quality speech from mel-spectrograms (comparable to WaveNet). It takes samples from a zero mean spherical Gaussian (with variance α) with the same number of dimensions as the desired output and passes those samples through a series of layers that transforms the simple distribution to one that has the desired distribution. In the case of training a vocoder, we use WaveGlow to model the distribution of audio samples conditioned on a mel-spectrogram. WaveGlow can achieve real-time inference speed using only a single neural network, whereas

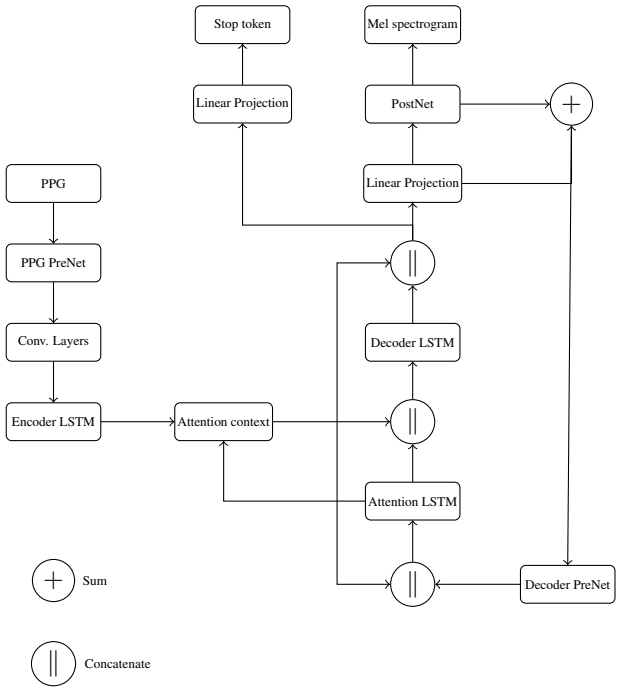


Figure 3: PPG to Mel-spectrogram model proposed in (Zhao et al., 2019)

WaveNet takes a long time to synthesize an utterance due to its auto-regressive nature.

3. Experimental Setup

In this section, we present the experimental setup used to evaluate the voice conversion system in the case of inter- and intra-speaker voice conversion. First, datasets are presented. Then, the model configuration as well as the hyper-parameters are detailed.

3.1. Datasets

For training and inference stages, we used a subset of the MUFASA corpus[1] (Sini, 2020), a multi-speaker dataset collected from two different libraries: LittératureAudio[2] and LibriVox[3]. This dataset is composed of samples of 20 speakers (10 Females/10 Males) encoded as follows: Female (F)/Male (M), FR: French, ID:XXXX. This database contains some novels in French published between the 18th and 20th centuries. All the transcriptions of the corpus are freely available on wikisource[4]. The speech signals are sampled at 22.05 kHz. The meta-data information related to describing the book (speaker identifier, library name) has been removed.

As we want to perform both inter- and intra-speaker voice conversions, we describe the subsets used for both tasks in the following.

3.1.1. Inter-Speaker Dataset

For this task, two audiobooks have been chosen for which three persons read each transcription. Thus, data for six distinct speakers have been collected, with three female voices and three male voices. We have chosen the voices using an informal listening test in order to retain only samples with no audible difference in recording conditions without taking into account differences between speaker voice qualities. The duration and number of samples of the selected voices are given in Table 1.

3.1.2. Intra-Speaker Dataset

For the intra-speaker voice conversion, the Synpaflex corpus, described in (Sini et al., 2018), is used. A subset of this corpus was manually annotated in character-based speaking turns. Indeed, we first aimed to consider the voice conversion between character speech and indirect speech, but due to the small amount of data per character for the training Mel-spectrogram to speech model (Zhao et al., 2019), we have decided to distinguish more generally direct and indirect speeches: the direct speech corpus groups all the character-based speaking turns whereas the indirect speech corpus is composed of the narrator discourse. As mentioned in Table 1, one hour of speech (with a sample duration inferior to 10 sec) per speech (direct or indirect) style has been used to train the Mel-to-Spectrogram in the case of style speech conversion.

3.2. Model Configuration

The different models to learn for the system are (1) the PPG extraction model, (2) the PPG-to-Mel model, and (3) the Mel-spectrogram to speech model. Whereas (2) and (3) are specific to the target speaker, (1) is a pre-trained TDNN-HMM acoustic model (Peddinti et al., 2015a) in this work.

As for the PPG-to-Mel model, Table 2 details the corresponding architecture and hyper-parameters.

Finally, concerning the Mel-spectrogram to speech model, we fine-tuned the NVidia pre-tained Waveglow[5] model during 155 epochs for each target speaker.

4. Evaluation

To assess the achievements of the proposed method, we have conducted three subjective evaluations. The first test aims to evaluate the quality of speech generated using the inter-speaker voice conversion and the vocoder. The second one considers the speaker similarity between the converted sample and target one. As for the last subjective test, its goal is the assessment of the PPG-based voice conversion performance to achieve a discourse style conversion and the capacity of the system to handle and capture the intra-speaker variation. To help the listeners, the two first tests use parallel samples, i.e. having the same linguistic content. For the third test, non-parallel samples are used.

All assessments have been conducted with native French speaking testers aged between 24 and 45. The majority of them have experience with listening tests but are not necessarily experts in the annotation of audio files. Besides, given an experimental configuration, all samples are randomly chosen with respect to the associated protocol described below. In the following, we will present the evaluation protocol and the objectives. These evaluations have been conducted using the FlexEval (Fayet et al., 2020)[6] online platform.

4.1. Speaker Conversion Speech Quality

To evaluate the speech quality of the fine-tuned vocoder and the quality of the implemented voice conversion system, a CMOS (Comparison MOS (Rec, 1996)) test has been set up. For this experience, at each step, the listeners have to rate, on a 5-point scale, where 1 indicates bad and 5 indicates excellent, the quality of two samples with the same linguistic content produced by two distinct systems:

- One sample obtained by the re-synthesis of the target voice sample using the Waveglow vocoder. The set of such samples is called Vocoded Target Voice (VocTargetVoice).

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

[2] http://www.litteratureaudio.com/

[3] https://librivox.org/

[4] https://fr.m.wikisource.org/wiki/Wikisource:Accueil

[5] https://ngc.nvidia.com/catalog/models/nvidia:waveglow_ljs_256channels

[6] https://gitlab.inria.fr/expression/tools/FlexEval

	Speaker or Style Speech	As source	As target	Number of samples			Duration (hours)
				Train	Validation	Test	
Inter-speaker	FFR0009	✓	✓	1430	30	30	1.44
	FFR0012	✓	✓	1430	30	30	1.53
	MFR0015	✓		/	/	30	1.25
	FFR0011	✓		1430	30	30	1.58
	MFR0013	✓	✓	1430	30	30	1.43
	MFR0014	✓	✓	/	/	30	1.72
Intra-speaker	Direct Speech (DS)	✓	✓	1850	50	35	1.12
	Indirect Speech (IS)	✓	✓	2002	63	35	1.12

Table 1: Duration and sample numbers for datasets used for Voice Conversion (VC) as source or target

Module	Parameters
PPG PreNet	Two fully connected (FC) layers; 600 ReLU units; 0.5 dropout rate [41]
Conv. Layers	Three 1-D convolution layers (kernel size 5); batch normalization [42] after each layer
Encoder LSTM	One-layer Bi-LSTM; 300 cells in each direction
Decoder PreNet	Two FC layers; 300 ReLU units; 0.5 dropout rate
Attention LSTM	One-layer LSTM; 300 cells; 0.1 dropout rate
Attention context	dimension of 150, attention location filters of 32 with a kernel size of 31
Decoder LSTM	One-layer LSTM; 300 cells; 0.1 dropout rate
PostNet	Five 1-D conv. layers; 512 channels; kernel size 5

Table 2: Hyper-parameters of ppg-to-mel presented in Figure 3.

- One sample resulting from one of the following Voice conversion configurations:
  - Conversion Intra-Gender (ConvIntraGen); the source and target voice belong to different speakers with the same gender ;
  - Conversion Inter-Gender (ConvInterGen); the source voice and target voice differ in terms of speaker and gender. The score results for the two combinations female to male and male to female are gathered together.

The main goal of this test is to evaluate the performance of the conversion system compared to a gold standard system. This gold standard system is a re-synthesis system and does no conversion at all. Overall, if the signal generation model is badly performing, the overall results of the other listening tests will be impacted by poor audio quality and will not help us to evaluate the benefits of the approach (listeners will poorly judge all samples presented).
For this experiment, 12 listeners have annotated 60 samples each, which permits to derive the confidence interval of opinion score mean. During the test, listeners answered the question "How do you judge the quality of the following samples?".

4.2. Inter-Speaker Similarity

In order to validate our voice conversion system in terms of speaker similarity, a subjective evaluation has been carried out based on the MUSHRA protocol (ITU, 2001). The main goal for this test is to give some clues about the presence or the absence of additional information on the speaker eventually included in the PPG (the gender for example).
At each step, a reference sample stemming from VocTargetVoice is given and five candidates to evaluate are presented in random order.
Among the candidates, a sample resulting from the Target-to-Target Conversion (TTC) configuration is presented. For this configuration, no conversion is done, neither between speakers nor styles: the target voice signal is used as input and consequently, the output can be considered as the upper bound for this voice conversion approach. The TTC system plays the role of an anchor in this MUSHRA test.
We also use two lower bound references as follows:

- The re-synthesis of a source voice sample belonging to the same gender as the target voice, called Vocoded Intra-Gender (VocIntraGen) ;
- The re-synthesis of source voice with different gender named Vocoder Inter-Gender (VocInterGen).

Finally, samples of isolated configurations coming from Conversion Intra-Gender (ConvIntraGen) and Conversion Inter-Gender (ConvInterGen) complement the test. For those last two configurations, the input sample is produced by a speaker different from the target one, but with the same gender in ConvIntraGen contrary to ConvInterGen where the gender is different. For this experience, the duration of each of the 120 samples presented to the listeners varies from 4s to 6s. The ratio of speech breaks present in the selected samples does not exceed the quarter of the total duration of

the sample.
One evaluation instance is composed by 20 steps including all the models presented before, thus evaluating 5 samples at each step. The question asked to testers is "How do you judge the similarity of the following candidates with the reference?". 17 French native speakers completed the evaluation.

4.3. Intra-Speaker Similarity

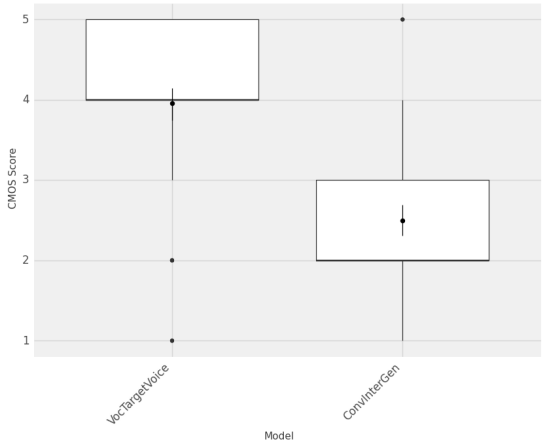
For evaluating the intra-speaker conversion, a subjective assessment has been designed. The objective is to evaluate the ability of the model to transform the Indirect Speech style to the Direct Speech style. Consequently, two samples made from two distinct conditions are presented to the listeners as reference (A) and (B) in a randomized manner. The two conditions respectively refer to Vocoded Indirect Speech (VocIS) and Vocoded Direct Speech (VocDS), the last one corresponding to the re-synthesis of the target speech style.
Then, relying on these references, the listeners have to rate five candidate audio samples using different configurations following the MUSHRA score scale (0 scores a candidate very close to the reference (A) and 100 scores a candidate very close to the reference (B)). We have Indirect Speech-to-Indirect Speech Conversion (IS2ISConv) and Direct Speech-to-Direct Speech Conversion (DS2DSConv) as intermediate anchors to evaluate the impact of conversion processing pipeline when the source and target speech styles are identical.
We also insert the samples coming from the same configuration as the references with different linguistic content, to estimate the ability of the listeners to distinguish the two speaking styles.
Finally, the last candidate is the one doing the conversion we want to evaluate, i.e. with a stimuli coming from the IS-to-DS Conversion (IS2DSConv). 9 listeners have completed this experiment. The number of participants is lower in comparison to the Section 4.2 and Section 4.1 which is due to the fact that the task is harder. We asked listeners to answer the following question: "Between A and B, to which sample the candidate X is the most similar (in terms of intonation, accentuation, tempo, rhythm, pause...)?". The test has 15 steps with 5 samples at each step.

5. Results and Discussion

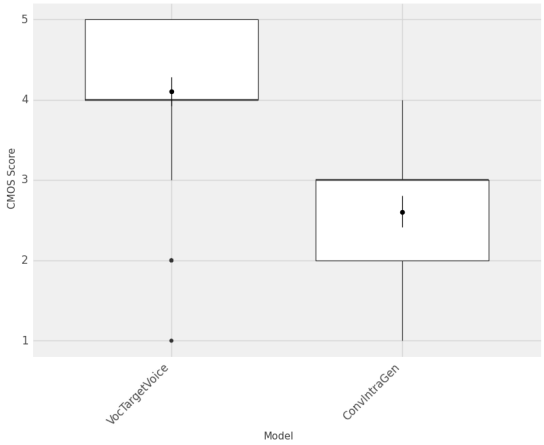
Inter-speaker similarity results are presented in Figure 5 and intra-speaker similarity results in Figure 6. In addition to box plots, the 95% confidence intervals on the average score, derived by a bootstrap method, are also presented.
According to Figure 4, the system VocTargetVoice, which corresponds to the waveglow vocoder, produces a very good quality voice signal with an average score about 4. This result is important since it gives an upper bound of speech quality than can be reached by this approach and permits to compare the different conversion systems only from their PPG management difference.
Results obtained for the two conversion systems ConvInterGen and ConvIntraGen are comparable to state-of-the-art results using comparable systems associating PPG and the Waveglow vocoder as in the VCC 2020 Challenge (Yi et al., 2020). One can notice that changing the vocoder may improve the results, at least for the signal quality evaluation.
The results obtained from the inter-speaker similarity evaluation in Figure 5a reveal that the Target-to-Target Conversion (TTC) system offers the highest similarity to the reference signal (around 60 points). This result could be thought as surprisingly low, given that we cannot achieve a better result with the other systems. The vocoder-based systems, VocIntraGen and VocInterGen, using the wrong target voice have the worst similarities (around 25 points when the vocoded signal belongs to the same gender as the target one and about 20 points otherwise). Moreover, both conversion systems have similarity results a bit above 50 points.
The results obtained by the conversion systems are promising due to their proximity to the TTC result. One slightly strange observation can be done: PPGs seem to achieve a greater similarity with a cross-gender voice conversion than when the gender is the same for the target and source voices. However, if this observation deserves some additional investigations, the difference between the average scores of both systems is not significant which would confirm the hypothesis of a significant independence between speakers and PPGs.
In Figure 5b, we can see the distribution of the ranks (the lower the better) for each system and for four target voices (2 females and 2 males). Ranks are computed from the MUSHRA test results. These ranks correlate the MUSHRA score results with the TTC system often ranked in first place and the vocoding systems (VocIntraGen and VocInterGen) which target the wrong voice often ranked lasts. Moreover, we can see that some voices perform better when they are used to do a gender swap voice conversion task: systems ConvIntraGen and VocIntraGen have better ranks than system TTC and ConvInterGen on average.
The results obtained from the intra-speaker similarity in Figure 6 tend to be more tedious to understand. The value 0 refers to the indirect speech reference and the value 100 refers to the direct speech reference. The indirect speech vocoded samples seem to be hard to be recognized as this system obtains a MUSHRA score around 50 points. On the other hand, the direct speech vocoded samples are fairly recognized as direct speech but the result is not as high as we expected. Surprisingly, the indirect to indirect conversion system samples have been better recognized as indirect style than the vocoded ones.

6. Conclusion

We described in this paper a voice conversion system trained on a large spoken corpus with different speakers. This is one of the rare studies to apply state of the

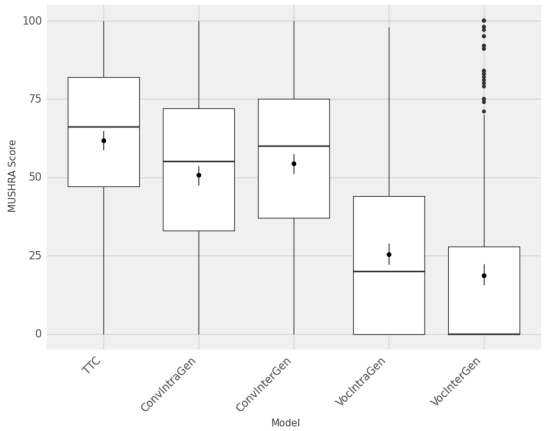


(a) Result of CMOS test - ConvInterGen

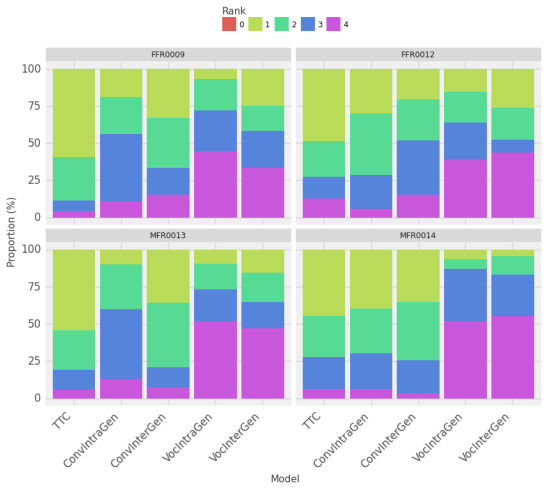


(b) Result of CMOS test - ConvIntraGen

Figure 4: Results of the CMOS listening test.



(a) Results of the MUSHRA listening test related to speaker similarity (protocol details given in Section 4.2)



(b) Distribution of ranks (the lower the better) of the target voices according to the speaker similarity to the target voice using MUSHRA evaluation results. 4 voices are evaluated: 2 female voices (FFR0009, FFR0012) and 2 male voices (MFR0013, MFR0014).

Figure 5: Results of speaker similarity listening test.

art systems to voice conversion in the French language. Moreover, this is the first study to apply and evaluate voice conversion methods to convert speech from indirect style to direct style. In our case, audiobooks read by multiple male and female speakers including spoken dialogues and narrative texts are used. With this dataset, we have been able to address voice conversion in different views : speakers, genders and styles (direct or indirect).

Perceptive experiments have been conducted to evaluate these different conversions. Results are promising and show that voice conversion methods are a track to explore further to convert indirect style to direct style. It also confirms that the identity can be changed if speakers are different and preserved otherwise.

Based on this study, it is necessary to conduct a prosodic study of the speakers as proposed in (Sini et al., 2020), notably concerning the use of a prosodic measure representing the identity of the speaker. Moreover, to clarify the results related to the intra-speaker subjective assessment, we propose to conduct a new evaluation, by simplifying the evaluation task with the use of an AB test instead of a MUSHRA test as presented. Finally, adding prosodic features such as F0 to the input of the PPG-to-Mel-Spectrogram should be investigated to improve the conversion of speaker specific prosodic aspects.

7. Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011011870R1 made by GENCI.

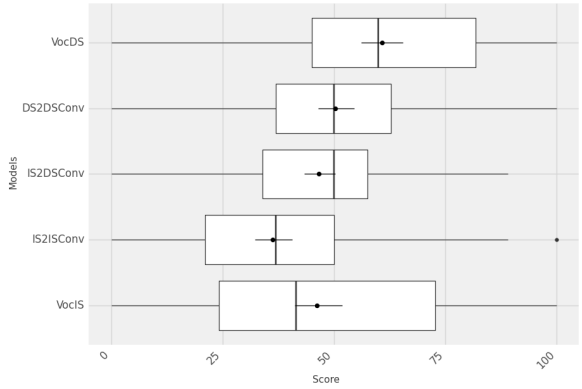


Figure 6: Results of the MUSHRA listening test related to intra-speaker (discourse style conversion, presented in Section 4.3)

8. Bibliographical References

Fayet, C., Blond, A., Coulombel, G., Simon, C., Lolive, D., Lecorvé, G., Chevelu, J., and Le Maguer, S. (2020). FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. In Christophe Benzitoun, et al., editors, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4 : Démonstrations et résumés d'articles internationaux*, pages 22–25, Nancy, France. ATALA.

Ho, T. V. and Akagi, M. (2020). Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*.

Huang, W.-C., Hayashi, T., Watanabe, S., and Toda, T. (2020). The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*.

ITU. (2001). Method for the subjective assessment of intermediate sound quality (MUSHRA). Technical Report P.1534-1, International Telecommunication Union (ITU-R).

Liu, L.-J., Ling, Z.-H., Jiang, Y., Zhou, M., and Dai, L.-R. (2018). WaveNet Vocoder with Limited Training Data for Voice Conversion. In *Proc. of Interspeech*, pages 1983–1987.

Liu, L.-J., Chen, Y.-N., Zhang, J.-X., Jiang, Y., Hu, Y.-J., Ling, Z.-H., and Dai, L.-R. (2020). Non-Parallel Voice Conversion with Autoregressive Conversion Model and Duration Adjustment. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*.

Liu, S., Cao, Y., Wang, D., Wu, X., Liu, X., and Meng, H. (2021). Any-to-many voice conversion with location-relative sequence-to-sequence modeling. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:1717–1728, jan.

Peddinti, V., Povey, D., and Khudanpur, S. (2015a). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proc. of Interspeech*.

Peddinti, V., Povey, D., and Khudanpur, S. (2015b). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*.

Peddinti, V., Wang, Y., Povey, D., and Khudanpur, S. (2017). Low latency acoustic modeling using temporal convolution and lstms. *IEEE Signal Processing Letters*, 25(3):373–377.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755.

Prenger, R., Valle, R., and Catanzaro, B. (2019). Waveglow: A flow-based generative network for speech synthesis. *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621.

Rec, I. (1996). P. 800: Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva*, 22.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Sini, A., Lolive, D., Vidal, G., Tahon, M., and Delais-Roussarie, E. (2018). SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis. In *Proc. of the 11th International Conference on Language Resources and Evaluation (LREC)*.

Sini, A., Maguer, S. L., Lolive, D., and Delais-Roussarie, E. (2020). Introducing Prosodic Speaker Identity for a Better Expressive Speech Synthesis Control. In *Proc. of the 10th International Conference on Speech Prosody*, pages 935–939, Tokyo, Japan, May. ISCA.

Sini, A. (2020). *Characterisation and generation of expressivity in function of speaking styles for audiobook synthesis*. Ph.D. thesis, University of Rennes 1.

Sun, L., Li, K., Wang, H., Kang, S., and Meng, H. (2016). Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Tian, X., Wang, J., Xu, H., Chng, E.-S., and Li, H. (2018). Average Modeling Approach to Voice

Conversion with Non-Parallel Data . In *Proc. of the Speaker and Language Recognition Workshop (Odyssey)*, pages 227–232.

Tian, X., Wang, Z., Yang, S., Zhou, X., Du, H., Zhou, Y., Zhang, M., Zhou, K., Sisman, B., Xie, L., and Li, H. (2020). The NUS & NWPU system for Voice Conversion Challenge 2020. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*.

Tobing, P. L., Wu, Y.-C., and Toda, T. (2020). Baseline System of Voice Conversion Challenge 2020 with Cyclic Variational Autoencoder and Parallel WaveGAN. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*.

Yi, Z., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z.-H., and Toda, T. (2020). Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*.

Zhao, G., Ding, S., and Gutierrez-Osuna, R. (2019). Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *Proc. of Interspeech*.

Zheng, L., Tao, J., Wen, Z., and Zhong, R. (2020). CASIA Voice Conversion System for the Voice Conversion Challenge 2020. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*.