Voice Conversion: State-of-the-Art and Future Work

David Sündermann

Universitat Politècnica de Catalunya, 08034 Barcelona, Spain, e-mail: suendermann@gps.tsc.upc.edu

Introduction

Voice conversion is the adaptation of the characteristics of a source speaker's voice to those of a target speaker. Over the last few years, the interest in voice conversion has risen significantly. This is due to its application to the individualization of text-to-speech systems, whose voices, in general, have to be created in a rather time-consuming way requiring human assistance. In this paper, the most popular applications and solution approaches are itemized. We will see that some applications require text-independent or even cross-language voice conversion. Then, evaluation methods are discussed and, finally, the author's future work is outlined.

Applications

In the literature, a vast number of applications for voice conversion are given. Obviously, only a few of them were ever carried out. Therefore, the author restricts his considerations to the most popular ones and those, he is going to investigate in the future.

- Definitely, the most important application of voice conversion is the use as a module in the text-to-speech synthesis, where we want to change the standard speaker's voice characteristics so that we suppose to hear another speaker, the target speaker [1].
- In embedded environments, voice conversion can serve to manipulate voices in such a way that the original speaker cannot be recognized or the speaker's gender or age is changed. Due to computational limitations, conversion algorithms of embedded systems focus on manipulating only a few parameters as fundamental frequency and vocal tract length [2].
- Cross-language voice conversion has been applied to dubbing tasks in the film industry [3]. As we proposed in [2], in the next future, it is to be used inside a speech-to-speech translation framework, where a translated sentence is uttered with the source speaker's voice.

Approaches

Due to the vast number of approaches proposed in the literature, in this section, only the most recognized ones are taken into account.

- The "inventors" of voice conversion, Abe et al. [4], clustered feature vectors of source and target speaker utterances and tried to find an appropriate mapping between these classes. This vector quantization approach has been extended by Arslan [5] and is still used in state-of-the-art voice conversion.
- The most popular technique is the application of linear transformations to the spectra of pitch-synchronous frames that are coded using mel frequency cepstral coefficients or line spectral frequencies [1].
- Over the last two years, remarkable efforts were taken to describe the voice transformation applying formant and antiformant parameters [6].
- As already argued in the last section, embedded voice conversion requires resource-efficient solutions. One adequate technique that is adapted from speech recognition is the vocal tract length normalization [2]. However, subjective evaluations have shown that this method is not able to imitate a certain target speaker's voice although the result is sufficiently different from the source.
- All the above considered approaches deal with spectral conversions although the prosody also contains a lot of speaker-specific information. Consequently, several prosodic transformations have been proposed, all of them resulting in a higher subjective similarity to the target speaker than pure spectral conversions, cf. e.g. [5].

Text-Independent and Cross-Language Voice Conversion

So far, almost all training procedures used for voice conversion are based on parallel corpora, i.e., training material of source and target speaker uttering the same text. This pre-condition allows the application of dynamic time warping or similar alignment methods in order to find corresponding time frames that are the basis for any parameter training.

In the real world, we often are not provided parallel utterances of both involved speakers. This observation leads to the field of text-independent voice conversion, a brand-new research area. Currently, the following publications are the only ones dealing with text-independent training for voice conversion:

- For text-independent VTLN-based voice conversion, we proposed a technique that clusters the magnitude spectra of pitch-synchronous frames of source and target speaker into artificial phonetic classes and then finds the most appropriate mapping between source and target classes using dynamic frequency warping [2].
- Another approach is based on maximum likelihood constrained adaptation [7]. In addition to the non-parallel corpus for the considered speaker pair, it uses a parallel corpus provided for a different speaker pair, trains the conversion parameters as usual for text-dependent approaches, and then adapts the obtained parameters to the former speaker pair using maximum likelihood estimation techniques.
- Assuming that there is no preexisting data of the source speaker, we can use a speech recognizer to index the utterances of the unknown speaker and assign similar speech frames of the already indexed target speaker frames [8]. The corresponding frame pairs serve as basis of a conventional voice conversion parameter training.

Having a look at the literature dealing with cross-language voice conversion, we note that, up to now, bilingual utterances of at least one of both speakers involved in the transformation process were required. E.g., one used a parallel English corpus of source and target speaker to train the conversion parameters. During conversion, the same source speaker uttered a Japanese sentence that was converted to sound like the target speaker using the above trained parameters [9].

Since certain applications as speech-to-speech translation require cross-language voice conversion based on monolingual speakers, the ability of the above itemized approaches of text-independent training for the application to cross-language tasks should be assessed. First investigations towards this idea were presented in [2].

Evaluation

When evaluating voice conversion technology, generally, we have two questions in mind:

- Does the technique change the speaker identity in the intended way?
- How is the overall sound quality of the converted speech?

The answers can be found applying objective and subjective error criteria. The former expresses the distance between the converted speech and corresponding reference speech of the target speaker. The latter is based on listening tests. In the following, the mainly used objective and subjective criteria are discussed.

Objective Error Measures

In order to compare equivalent segments of the converted and reference speech, the corresponding utterances are time-aligned and result in two parallel frame sequences. For each frame pair, we compute a spectral distance, sum up over all frames, and obtain a global distance. Often, this distance is divided by the global distance between the unconverted source and the target speech that leads to a measure that is 1.0 if we do not convert at all.

In the literature, we find as many distance definitions as authors dealing with voice conversion [10]. The reason for this great number of unrecognized error measures is the following aspects:

- As the author's experience shows, the proposed measures do not sufficiently correlate with the human's perception. I.e., an objective error measure that indicates that the converted speech is closed to the reference does not necessarily mean that the voices are perceived to be similar. This is due to the very complex way in that a listener assesses and recognizes a voice. The basic components a voice is composed of are the vocal tract, the excitation and the prosody, that, so far, have not completely been covered by an objective error measure.
- Listening to examples of state-of-the-art voice conversion, even so called "high quality" voice conversion [3], one notes a lot of distortions and artifacts, although the speaker identity was adequately converted. The question for an evaluation of voice conversion quality using objective error measures has not been investigated at all.

Subjective Error Measures

One could assume that asking human listeners for their opinion about the two questions we asked above should lead to meaningful results that avoid the shortcomings of the subjective evaluation. Unfortunately, we also face a number of difficulties:

- The test must be carefully defined in order to result in the information we want to obtain: For instance, the very popular ABX test asks if the converted utterance (X) is more similar to the corresponding source or the target utterance (either A or B). Kain [1] applied an ABX test to his voice conversion system that resulted in an equipartition: About 50 % of the speech examples were assigned to speaker A or B, respectively. This result seemed to indicate a success of the conversion in half of the cases. However, interviews with the subjects showed that most of them had the impression that a "third" speaker was created. This unfavorable effect led to an extension of the ABX test that includes the possibility of choosing a speaker that is neither A nor B [11].
- The different components that form an individual voice should be separately assessed. E.g., evaluations have been performed that removed prosodic or excitation characteristics [1].
- To obtain statistically relevant results, one should consider as many subjects as possible in the test (in the literature dealing with voice conversion, numbers of participants between 5 and 23 are reported.) Besides, they should represent the clientele of the voice conversion technology, i.e., native speakers of the converted voice's language without particular speech expertise should be preferred.

Outline

The last sections discussed some aspects of state-of-the-art voice conversion and also raised several questions that the author wants to answer in future investigations. Referring to the respective section titles, the most important issues are the following:

- **Applications.** Already former publications, cf. e.g. [2], testified the author's strong interest in applying voice conversion to speech-to-speech translation.
- **Approaches.** According to the author's experience, linear-transformation-based voice conversion with line spectral frequencies results in the best sound quality in comparison with other voice conversion techniques and feature types. Hence, this technology is used as a baseline for current investigations. However, using line spectral frequencies to represent the spectral envelope neglects a lot of spectral details that should be regained by means of an appropriate residual prediction technique [1]. The current research is aimed at improving the already existing sparse residual prediction approaches [11].
- **Text-Independent and Cross-Language Voice Conversion.** The three proposed techniques for text-independent voice conversion are to be compared, and their adaptability to the cross-language conversion task is to be investigated.
- **Evaluation.** Perhaps the most challenging part of the author's future work is the evaluation of voice conversion technology, a subject that, up to now, has not been systematically investigated. The author is in contact with more than ten international research groups that deal with voice conversion and obtained a lot of speech material that is to be assessed in the scope of a large-scale subjective evaluation in order to derive appropriate objective error measures that cover both the ability of voice identity conversion and speech quality.

References

[1] A. Kain, "High Resolution Voice Transformation," Ph.D. dissertation, OGI, Portland, USA, 2001.

[2] D. Sündermann, H. Ney, and H. Höge, "VTLN-Based Cross-Language Voice Conversion," in *Proc. of the ASRU'03*, St. Thomas, USA, 2003.

[3] O. Turk and L. M. Arslan, "Subband Based Voice Conversion," in *Proc. of the ICSLP'02*, Denver, USA, 2002.

[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice Conversion through Vector Quantization," in *Proc. of the ICASSP'88*, New York, USA, 1988.

[5] L. M. Arslan, "Speaker Transformation Algorithm using Segmental Codebooks (STASC)," *Speech Comm.*, vol. 28, 1999.

[6] E. Turajlic, D. Rentzos, S. Vaseghi, and C.-H. Ho, "Evaluation of Methods for Parameteric Formant Transformation in Voice Conversion," in *Proc. of the ICASSP'03*, Hong Kong, China, 2003.

[7] A. Mouchtaris, J. Spiegel, and P. Mueller, "Non-Parallel Training for Voice Conversion by Maximum Likelihood Constrained Adaptation," in *Proc. of the ICASSP'04*, Montreal, Canada, 2004.

[8] H. Ye and S. J. Young, "Voice Conversion for Unknown Speakers," in *Proc. of the ICSLP'04*, Jeju, South Korea, 2004.

[9] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of Cross-Language Voice Conversion Based on GMM and STRAIGHT," in *Proc. of the Eurospeech'01*, Aalborg, Denmark, 2001.

[10] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A First Step Towards Text-Independent Voice Conversion," in *Proc. of the ICSLP'04*, Jeju, South Korea, 2004.

[11] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A Study on Residual Prediction Techniques for Voice Conversion," in *Proc. of the ICASSP'05*, Philadelphia, USA, 2005.