# Text-informed speech inpainting via voice conversion

Pierre Prablanc, Alexey Ozerov, Ngoc Q. K. Duong, Patrick Pérez

# TEXT-INFORMED SPEECH INPAINTING VIA VOICE CONVERSION

*Pierre Prablanc, Alexey Ozerov, Ngoc Q. K. Duong and Patrick Pérez*

Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{pierre.prablanc, alexey.ozerov, quang-khanh-ngoc.duong, patrick.perez}@technicolor.com

**ABSTRACT**

The problem of speech inpainting consists in recovering some parts in a speech signal that are missing for some reasons. To our best knowledge none of the existing methods allows satisfactory inpainting of missing parts of large size such as one second and longer. In this work we address this challenging scenario. Since in the case of such long missing parts entire words can be lost, we assume that the full text uttered in the speech signal is known. This leads to a new concept of *text-informed speech inpainting*. To solve this problem we propose a method that is based on synthesizing the missing speech by a speech synthesizer, on modifying its vocal characteristics via a voice conversion method, and on filling in the missing part with the resulting converted speech sample. We carried subjective listening tests to compare the proposed approach with two baseline methods.

***Index Terms*—** Audio inpainting, speech inpainting, voice conversion, Gaussian mixture model, speech synthesis

## 1. INTRODUCTION

The goal of audio inpainting consists in filling in missing portions of an audio signal. This concept was recently formulated by Adler *et al.* [1] as a general framework covering several existing audio processing problems such as audio declipping [2], clicks removal [3] and bandwidth extension [4]. The term *inpainting* is borrowed from *image inpainting* [5], a similar problem in image processing, where the goal is to fill in missing parts in an image. The difficulty of an audio inpainting problem depends mainly on the nature of the signal (e.g., speech or music) and on the distribution of the missing parts (e.g., tiny holes of few samples or bigger holes of several milliseconds). For example IP packet losses in VoIP systems usually lead to missing intervals in the transmitted speech of length ranging from 5 ms to 60 ms. This problem is often addressed using packet loss concealment (PLC) algorithms [6,7] that are only able to fill in the missing part with a quasi stationary signal. This is achieved either by repeating the last packet received [6] or by more sophisticated autoregressive model-based prediction/interpolation [7]. A more advanced method consisting in smoothly filling-in the missing part with previously seen speech examples was recently proposed by Bahat *et al.* [8]. This method allows producing a more non-stationary and more natural signal in the missing part.

In this paper we address the problem of speech inpainting when the duration of the missing part may be very large (i.e., one or several seconds). Existing approaches such as PLC algorithms [6, 7] or example-based speech inpanting [8] are not designed to handle such long missing areas. Indeed, whole words or big portion of a word may be entirely missing, and it often becomes not even clear what was really said. For example in *"I ... you."* sentence, the missing word (represented by dots) can be *love*, *miss*, *hate*, etc. To make the problem slightly better defined we assume that the text that should be pronounced in the missing part is known. As such, we assume that the text of the whole sentence is available (the text of the observed part may be always transcribed if needed). This leads to a so-called *informed audio inpainting* setup, where by analogy with informed or guided audio source separation [9] some information about the missing signal is assumed known. More specifically, this particular audio inpainting setup is very close in spirit to text-informed audio source separation [10].

A successful text-informed speech inpainting algorithm might be still applied for speech restoration in VoIP transmission, though it is not the most straightforward application. Indeed, first, the approach must operate online in this case and, second, the text needs to be known on the receiver side. However, there are several new applications that become possible. First, this new inpainting strategy may be used in the audio post-production workflows. One important and demanding task in audio post-production is the post-synchronization (a.k.a. additional dialogue recording) where actors must record again their lines in a studio because on-set recordings contain slight text errors or unexpected noise, or because dialogue changes are made *a posteriori*. Such problems could be partially addressed by the proposed technique instead. Similarly, dubbing often requires final edits to just slightly correct small portions of the new speech, a task that could benefit from our technique. Second, it would allow restoring beep censored speech in TV shows or movies by either reproducing the original or a modified speech in the bleeped part. Finally, text-informed speech inpainting could be suitable for various other speech editing needs, including the partial "rewriting" of speech sequences in the more general context of audio-visual content editing, e.g. [11].

We propose a solution for text-informed speech inpainting that is based on speech synthesis [12] and voice conversion [13]. More precisely our approach is based on the following main steps:

1. A speech sample (*source speech*) corresponding to both observed and missing parts is synthesized given the text;
2. A voice conversion mapping is learned from observed parts of the speech to inpaint (*target speech*) and the corresponding parts of the synthesized speech;
3. The resulting voice conversion mapping is applied to the source speech parts corresponding to the missing parts of the targets speech;
4. The missing parts in the target speech are filled in with the obtained converted speech.

This work was partially supported by ANR JCJC program MAD (ANR-14-CE27-0002).

Another option we consider is when the source speech is not synthesized, but more naturally pronounced by a user. This may be possible within a user-assisted speech processing tool. The proposed method is somehow related to an analysis/synthesis-based speech enhancement method [14], though the problem considered in this latter work is speech enhancement, which is very different from speech inpainting considered here.

The rest of the paper is organized as follows. Section 2 is devoted to a description of voice conversion in general and of the particular voice conversion method we used. Understanding voice conversion is necessary to further understand some particularities of the proposed speech inpainting method described in Section 3. Subjective listening tests were carried out to compare the proposed approach with two baselines: the source speech and the converted source speech. The results of subjective tests are presented in Section 4 and some conclusions are drawn in Section 5.

## 2. VOICE CONVERSION

In this section we recall main principles of voice conversion and describe a particular voice conversion system we used in this work.

### 2.1. Generalities

The general goal of voice conversion is to modify some characteristics of a speech signal such as speaker identity, gender, mood, age, accent, etc. [15], while keeping unchanged the other characteristics including the linguistic information. In this work we are interested in speaker identity transfer. As such, the goal of voice conversion we consider here is to modify a speech signal uttered by a so-called *source speaker* so that it sounds as if it was pronounced by another so-called *target speaker*.

Most of voice conversion systems consist of the following two ingredients:

- *Analysis / synthesis*: An analysis system transforms the speech waveform into some other representation that is related to the speech production model, and thus easier to modify some speech characteristics. The corresponding synthesis system resynthesizes back a speech signal waveform from the transformed representation.
- *Voice conversion mapping* is applied to the transformed speech representation so as to modify its characteristics. This mapping is usually learned from some training data.

### 2.2. Analysis and synthesis

As analysis system we use the STRAIGHT-analysis [16] that allows estimating the fundamental frequency $f0$ and a smooth spectrum. We then compute the Mel-frequency cepstral coefficients (MFCCs) [17] from the STRAIGHT-spectrum. As such our transformed representation consists of $f0$ and MFCCs. For synthesis we reconstruct STRAIGHT-spectrum from the MFCCs and then re-synthesize the speech waveform from STRAIGHT-spectrum and $f0$ using STRAIGHT-synthesis [16].

### 2.3. Voice conversion mapping

Many approaches were proposed to build voice conversion mapping including those based on Gaussian mixture models (GMMs) [13,18], nonnegative matrix factorization (NMF) [19], artificial neural networks (ANN) [20] and partial least squares regression [21]. Here we have chosen to follow one of the most popular GMM-based approach proposed by Toda *et al.* [18].

#### *2.3.1. Modeling*

As feature vector to be predicted we consider, as in [18], the MFCCs (*static features*) concatenated with their derivatives (*dynamic features*). We use a so-called joint GMM modeling [15] that models a joint distribution of source and target features [22]. Moreover, we consider GMM with "tri-diagonal" covariance matrices, which are much more efficient to compute as compared to GMM with full covariance matrices.

#### *2.3.2. Training*

Voice conversion mapping is usually trained from a so-called *parallel dataset*, i.e., a set of sentences uttered by both source and target speakers. These sentences are first aligned using, e.g., the dynamic time warping (DTW) [23] applied to the MFCCs. Then, a joint GMM is trained from the set of aligned and concatenated together source and target feature vectors using the expectation-maximization (EM) algorithm [24].

#### *2.3.3. Conversion*

Once the joint GMM is learned, to convert a new source speech, target speech features are first predicted in the minimum mean square error (MMSE) sense, given the source features and the model [18]. Finally, since most likely the predicted MFCCs and their derivatives do not correspond to any original MFCC sequence, the MFCCs are re-estimated in a maximum likelihood sense, thus introducing some temporal smoothness in their trajectories thanks to the derivatives [18].

The $f0$ is often not predicted by the GMM, but via a simpler linear regression in the logarithmic domain, and we follow the same strategy here.

## 3. PROPOSED SPEECH INPAINTING APPROACH

In this section a description of the proposed approach is given throughout the subsections below. First, it is assumed that a speech signal with a missing part is given and the exact location of this part is indicated. It is also assumed that the speech is pronounced by just one speaker. Following voice conversion terminology the speech signal is called *target speech* and the corresponding speaker is called *target speaker*. For the sake of simplicity the description below is given for the case when there is only one missing segment in the target speech signal.

### 3.1. Source speech sample production

Given the uttered text for both observed and missing parts, a *source speech* is produced, inline with [10], either using a speech synthesizer or by a human operator within a user-guided tool. Both strategies have their pros and cons as follows. Within an application where a fully-automated process is needed the synthesis-based strategy is preferable, since it does not require any human intervention. However, the user-guided strategy may provide a source speech of a much higher quality. First, it is not synthetic. Second, in contrast to the synthesis-based approach, it can be much better adapted to the target speech rate, emotion, and other characteristics.

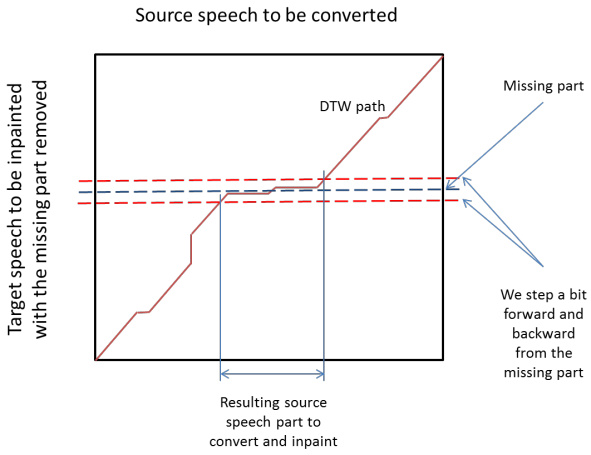



**Fig. 1**: Source vs. target alignment and missing part identification.

### 3.2. Alignment between source and target speech

Source and target speech signals are temporally aligned for the following two reasons:

- First, this alignment is needed to identify which portion of the source signal corresponds to the missing part of the target signal. This portion will be then used for inpainting.
- Second, these aligned signals are then used as a parallel dataset for training voice conversion mapping.

Note that this alignment is not trivial, since there is a missing part in the target speech and we also need to identify the corresponding part in the source speech. To achieve that we propose a simple strategy based on DTW that is visualized on Figure 1 and briefly described as follows. The missing part is simply removed from the target speech, but its location is retained. MFCCs are extracted from both source and target signals and a DTW path based on the distances between MFCC vectors is computed. One might expect that this path should form approximately a straight line (horizontal on Fig. 1) around the region corresponding to the missing part. As such, by introducing a small forward/backward tolerance around the known missing part position within the target signal, one can identify the corresponding source speech part as shown in the figure.

### 3.3. Voice conversion mapping training

Voice conversion mapping, in our case a GMM, is trained from the parallel data, i.e., the aligned source and target speech signals in the regions where the target signal is observed. Note that if some auxiliary target speech data with the corresponding text (transcription) is available, the training parallel dataset may be augmented to improve voice conversion performance.

### 3.4. Voice conversion

A source speech segment to be converted is extracted as follows. This segment must include the region corresponding to the missing part in the target speech (as identified in Section 3.2), but also must contain some signal before and after this region. This precaution of taking a slightly bigger segment is needed to assure a smooth transition during the inpainting. The extracted source speech segment is transformed by the analysis system. The resulting transformed representation, in our case MFCCs and $f0$, is then converted using the pre-trained voice conversion mapping to make its characteristics closer to those of the target speaker.

### 3.5. Speech inpainting in the transformed domain

A transformed representation of the observed target speech parts is computed in its turn by the analysis system. In the transformed representation the converted source speech segment is inserted to fill in the missing part in the target speech in a smooth way via a fade-in fade-out interpolation strategy. This is possible thanks to the fact that the converted source speech segment is bigger than the missing region. Note though that, as for the spectral envelop parameters, we interpolate by fade-in fade-out directly the STRAIGHT-spectrum rather than the MFCCs. The resulting target speech waveform is resynthesized from the obtained inpainted transformed representation via the synthesis system.

## 4. EXPERIMENTS

Since to our best knowledge none of the existing speech inpainting approaches allows handling such long missing segments, we resort to two baselines for comparisons. Namely, we compare the following three methods:

- *"Inpainted"*: the proposed approach;
- *"Converted"*: the entire source speech (not only the missing part) that was converted by voice conversion;
- *"Source"*: the non-processed source speech that was either synthesized or pronounced.

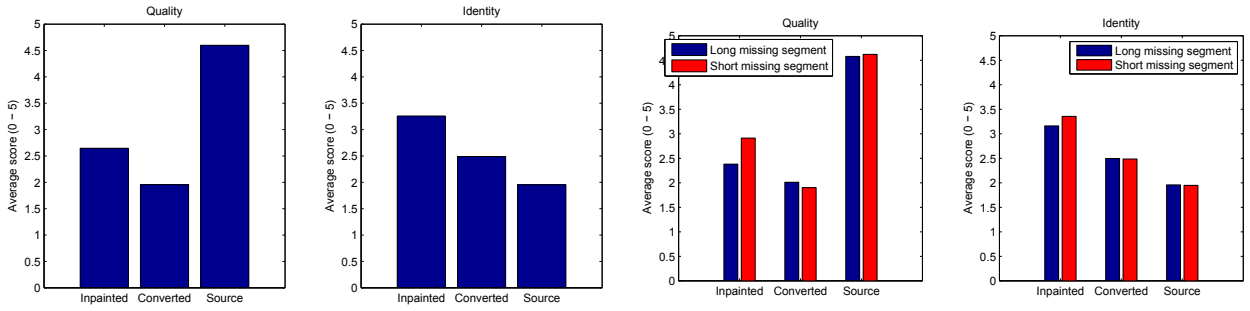
We compare these three approaches using a subjective listening test.

### 4.1. Data

We have created a dataset including natural speech samples of 4 English speakers (2 male and 2 female speakers) from the CMU ARCTIC database [25] and synthetic speech samples of 2 English speakers (1 male and 1 female speaker) synthesized by the IVONA speech synthesizers.[1] One male and one female speaker from the CMU ARCTIC database are always considered as target speakers and two other speakers are always considered as source speakers. Both synthetic IVONA speakers are considered as source speakers. All the signals are sampled at 16 kHz. Each speech inpainting setup is characterized by a pair of different speakers (source and target) and includes:

- *Target speech to inpaint*: one approximately 3 sec long target speech sample with a missing segment that is chosen randomly, while assuring that it is not toot close to the border;
- *Source speech*: one approximately 3 sec long source speech sample uttering the same text as the target;
- *Parallel dataset for training*: an auxiliary parallel dataset of these two speakers for voice conversion training that does not include the above 3 sec long test speech samples;
- *Target speech examples*: two 3 sec long speech samples uttered by the target speaker that are different from the test samples to inpaint and are needed to allow the listening test participants judging the speaker identity preservation as it will be explained below.

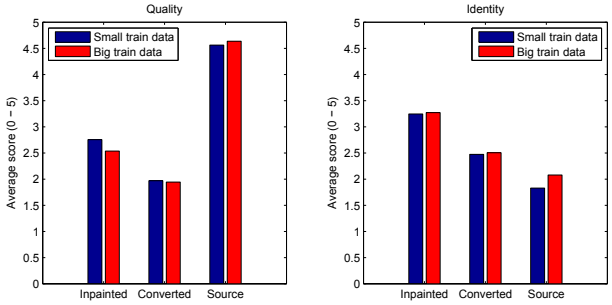
[1] www.ivona.com/en/



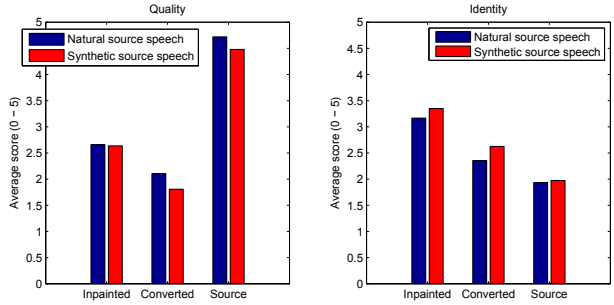

(a) Average results




(b) Long (750 ms) vs. short (250 ms) missing segment




(c) Small (30 sec) vs. big (5 min) training parallel dataset




(d) Natural vs. synthetic source speech

**Fig. 2**: Listening test results averaged over 12 test participants and over different conditions.

We then created the 16 different speech inpainting setups consisting of all possible combinations of the following binary choices:

1. Source and target speakers either both male or both female,
2. Missing segment either long (750 ms) or short (250 ms);
3. Parallel training dataset either small (30 sec) or big (5 min);
4. Source speech either natural or synthetic.

Finally, none of inpainting setup pairs shared the same test sentence.

### 4.2. Parameters and simulations

For DTW alignment (Sec. 3.2) we used only 12 first MFCCs computed directly from the corresponding signals. For MFCCs computed from STRAIGHT-spectrum (Sec. 2.2) we kept all 40 coefficients, which allows quite precise, yet not perfect, STRAIGHT-spectrum reconstruction from MFCCs during the resynthesis step. The joint GMM (Sec. 2.3.1) included 32 Gaussian components.

For each speech inpainting setup we ran the proposed inpainting approach together with voice conversion applied to the whole source speech signal, thus obtaining three speech samples: *inpainted*, *converted* and *source* as described in the beginning of Section 4. It should be noted that when the proposed missing part identification strategy (see Fig. 1) failed in correctly identifying the missing part, it was re-adjusted by hand. This happened for 6 out of 16 sequences.

### 4.3. Listening test and results

A total of 15 persons (2 women and 13 men) participated in the listening test. All the listeners used headphones. For each of the 16 speech inpainting setups presented in a random order, each participant first listened to two target speech examples, to have an idea about the target speaker identity. The participant would then listen to the three speech samples (*inpainted*, *converted* and *source*) presented in a random order, and was asked to note for each sample both the speech audio quality ("Does it sounds natural? Can you hear artefacts or not?") and the identity preservation ("Does the voice in the samples resemble to the voice of the target speaker?") on a 0 to 5 scale (greater is better). The random orders of presentation were different for different participants. The participants were not informed about the nature of the processing of speech samples. We have decided not keeping the results of participants who have noted at least once a natural source speech quality smaller than 4 or smaller than the quality of a processed sample (*inpainted* or *converted*). After such a filtering we have retained the results of 12 participants.

Note that in this experiments we did a deliberate choice not to compare the original speech having missing parts, since it is meaningless to perform such a comparison in terms of the speech quality and the identity preservation. Indeed, for example if the missing part corresponds to just one entire word in the sentence, removing this word from the speech sample would not affect neither the quality nor the identity preservation.

Figure 2 summarizes the test results averaged over 12 participants and over different conditions. As expected the source has always the best quality (Fig. 2 (a)). However, the inpainted speech has a better quality than the converted one. Moreover, the inpainted speech outperforms the two baselines in terms of identity preservation. One can note from Fig. 2 (b) that inpainting long missing segments leads to a quality degradation as compared to inpainting short missing segments. As for other conditions (Figs. 2 (c) and (d)), they do not seem to influence too much the results.

## 5. CONCLUSION

In this paper we have formulated the problem of text-informed speech inpainting, where it becomes potentially possible to perform a satisfactory inpainting of quite long missing parts (several seconds) in a speech signal thanks to the knowledge of the uttered text. This new framework opens a door for new speech editing capacities and can be applied, e.g., for post-sync and dubbing. We have proposed a solution for this problem based on voice conversion. Experimental results have shown that the proposed speech inpainting approach leads to both better speech quality and better speaker identity preservation as compared to using voice conversion alone. Further work will include research towards a complete automation of the inpainting process and the inpainted speech quality improvement. Another interesting research path would be to consider a similar problem in music processing: a score-informed music inpainting.

## 6. ACKNOWLEDGEMENT

The authors would like to thank colleagues from Technicolor who participated in the listening test.

## 7. REFERENCES

[1] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 922–932, 2012.

[2] S. Abel and J.O. Smith III, "Restoring a clipped signal," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1991, p. 17451748.

[3] S. J. Godsill and P. J. Rayner, "A Bayesian approach to the restoration of degraded audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 267–278, 1995.

[4] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for spectral audio signals," in *Proc. Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Grenoble, France, Sep. 2009, pp. 1 – 6.

[5] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *SIGGRAPH'00*, 2000, pp. 417–424.

[6] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *Network, IEEE*, vol. 12, no. 5, pp. 40 – 48, Sep. 1998.

[7] Z. Guoqiang and W.B. Kleijn, "Autoregressive model-based speech packet-loss concealment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, 2008, pp. 4797 – 4800.

[8] Y. Bahat, Y. Y. Schechner, and M. Elad, "Self-content-based audio inpainting," *Signal Processing*, vol. 111, pp. 61–72, 2015.

[9] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[10] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.

[11] Christoph Bregler, Michele Covell, and Malcolm Slaney, "Video rewrite: Driving visual speech with audio," in *SIGGRAPH*, 1997.

[12] W Bastiaan Kleijn and K. K Paliwal, *Speech coding and synthesis*, Elsevier Science Inc., 1995.

[13] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, 1998.

[14] J.L. Carmona, J. Barker, A.M. Gomez, and Ning Ma, "Speech spectral envelope enhancement by HMM-based analysis/resynthesis," *IEEE Signal Processing Letters*, vol. 20, no. 6, pp. 563–566, June 2013.

[15] E. Helander, *Mapping Techniques for Voice Conversion*, Ph.D. thesis, Tampere University of Technology, 2012.

[16] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1997, vol. 2, pp. 1303–1306.

[17] R. Vergin, D. O'shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 525–532, 1999.

[18] T. Toda, A. W Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[19] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–10, 2014.

[20] M Narendranath, Hema A Murthy, S Rajendran, and B Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech communication*, vol. 16, no. 2, pp. 207–216, 1995.

[21] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 912–921, 2010.

[22] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1998, vol. 1, pp. 285–288.

[23] M Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[24] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[25] J. Kominek and A. W Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.