INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Applying Spectral Normalisation and Efficient Envelope Estimation and Statistical Transformation for the Voice Conversion Challenge 2016

Fernando Villavicencio[1], *Junichi Yamagishi*[1], *Jordi Bonada*[2], *Felipe Espic*[3]

[1]National Institute of Informatics (NII), Tokyo, Japan.
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[3]The Centre for Speech Technology Research (CSTR), Edinburgh, United Kingdom.

Abstract

In this work we present our entry for the Voice Conversion Challenge 2016, denoting new features to previous work on GMM-based voice conversion. We incorporate frequency warping and pitch transposition strategies to perform a normalisation of the spectral conditions, with benefits confirmed by objective and perceptual means. Moreover, the results of the challenge showed our entry among the highest performing systems in terms of perceived naturalness while maintaining the target similarity performance of GMM-based conversion.

Index Terms: voice conversion, speech synthesis, statistical spectral transformation, spectral envelope modeling.

1. Introduction

One of the fields of speech synthesis that has received significant attention in the last decade is the one intending to convert the identity of a speaker to another specific target, known as Voice Conversion (VC). Following a number of pioneering works ([1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]), the work of [12] proposing a statistical conversion of spectral features derived from parallel corpora of source and target speakers became a reference for a number of further studies. Among them we highlight prominent contributions such as joint acoustic modeling ([13]), maximum-likelihood and eigenvoices based strategies ([14], [15]), non-parallel data processing ([16]), incorporating frequency warping ([17], [18]), and works as [19] and [20] considering novel conversion frameworks based on deep learning and non-negative matrix factorization respectively, among others.

In previous work we applied accurate spectral envelope estimation to VC with clear benefits on the perceived quality and naturalness of converted speech. More precisely, the technique True-Envelope (TE) ([21], [22]) was used to derive all-pole systems as spectral features of higher accuracy in terms of envelope fitting compared to linear prediction (LPC) or other cepstrum-based techniques ([23]). As a result, the quality of speech and singing-voice converted by following the joint Gaussian Mixture Model (GMM) based approach ([13]) outperformed ([24], [25]). Later, we proposed in [26] an optimised spectral transformation that compensates for limitations of such a probabilistic model to efficiently represent the features space, resulting in a perceived reduction of degradations on the converted speech.

Although a mapping of the main spectral features can be achieved by GMM-based VC, a robust gender conversion effect it is not always observed. This suggests some limitations to robustly reproduce a warping-like transformation on the source speech spectra at inter-gender conversions following well-known differences (in average) of the vocal-tract length conditions. Inspired by works such as [17] and [18] we propose applying a warping factor to perceptually assure a gender conversion effect. Additionally, we study the benefits of applying downwards pitch transposition to female speech to reduce over-estimations of the envelope amplitude on the TE algorithm due to particular spectral conditions at low-frequencies on high-pitched speech, as explained in following sections.

We report in this paper the application of these techniques as gender-dependent pre-processing to normalise the spectral conditions between speakers before GMM-based conversion. By following this strategy we obtained a reduction of the spectral conversion error and improvements on both perceived target similarity and naturalness according to a perceptual evaluation. Moreover, the results obtained at the Voice Conversion Challenge 2016 (VCC2016) with the resulting conversion methodology were among the highest performing systems in terms of naturalness (ranked second overall) while maintaining a target performance comparable to GMM-based conversion.

A summary of previous work and the proposed spectral normalisation in which is based our conversion system for the VCC 2016 are described in Section 2. In Section 3 we report the results of objective and subjective evaluations. The results obtained at the challenge are presented and discussed in Section 4. The paper finishes with conclusions at Section 5.

2. Our methodology: Improved Spectral Processing applied to GMM-VC

2.1. GMM-based differential spectral transformation

Our conversion framework is based on the well-known joint source-target acoustic modeling approach, denoting a mapping of spectral features on a frame-by-frame basis derived by linear regression [13]. As proposed in [25] and [27], we apply this transformation by means of a *transformation filter* $H_k(\omega)$ corresponding to the differences between input and predicted spectral envelopes:

$$H_k(\omega) = |\hat{Y}_k(\omega)| - |X_k(\omega)|, \tag{1}$$

where $X_k(\omega)$ and $\hat{Y}_k(\omega)$ denote the spectral envelopes according to the input (source) feature x_k and the corresponding target prediction $\hat{y}_k$ for frame number k. Note that $H_k(\omega)$ is applied pitch-synchronous following a Wide-Band Harmonic Sinusoidal Modeling (WBHSM) approach in which a phase correction model is considered for spectral amplitude modification (see [28] for further details).

http://dx.doi.org/10.21437/Interspeech.2016-305

2.2. Accurate spectral envelope extraction

Spectral features based on linear prediction (LP) or cepstral coefficients do not generally lead to accurate spectral envelope information ([29]). We exploit the benefits of TE estimation ([30], [21]) which provides efficient envelope fitting and allows an optimisation of the estimation based on the F0 information [31], resulting, according to previous work, in clear benefits in terms of converted speech quality ([24], [32], [25]).

Thus, we perform optimal TE estimation that is mel-scaled before deriving an all-pole model represented as Line Spectral Frequencies (LSF) (our final features). We denote this model mel-based True Envelope All-Pole (mel-TEAP). Given a sample-rate of 16 kHz we found in forty a good compromise as order to closely fit the spectra of male and female speech.

2.3. New feature: spectral conditions normalisation

2.3.1. Reducing over-estimations on high-pitched speech

True Envelope estimation performs an iterative smoothing of a cepstrum-based envelope to achieve a smooth interpolation of the spectral peaks. Considering the harmonic partials as support points, the case of high pitched spectra represent an augmented challenge to this technique since larger amplitude fluctuations may be observed in spectra with a smaller number of harmonics. As a consequence, some over-estimation issues were found at the frequency interval denoted by [0, F0] by the interpolation done by True Envelope ([22]) on spectra showing large amplitude fluctuations among the first harmonics. Although these conditions may not appear systematically nor affect the conversion performance substantially, we propose to reduce the risk of potential issues by applying one-octave downwards pitch transposition to female speech to artificially create an intermediate support point (harmonic partial) at the mentioned interval.

2.3.2. Global gender normalisation by frequency-warping

For inter-gender conversion, VC frameworks based on a statistical mapping of spectral features do not always show a natural transformation of the target speaker gender, suggesting some limitations to producing a spectral warping adjustment that corresponds to a vocal-tract length normalisation. Accordingly, motivated by works as [17] and [18] we apply a gender-dependent warping factor to the source speech to increase the spectral alignment with the target speaker.

The warping break-point function correspond to $[0 \;\; 0; \;\; F_{in} \;\; F_{out}; \;\; Fs \;\; Fs]$, with values $F_{in} = 5kHz$, $F_{out} = 6kHz$ ($Fs = samplerate$) to convert male to female speech and conversely, $F_{in} = 6kHz$, $F_{out} = 5kHz$ for the opposite conversion. These values were defined subjectively by experimentation on voices from different corpora and that although this is not an optimal solution as in the aforementioned works, a global factor strategy requires less computational cost and was found sufficient to produce a perceived gender transformation already on the source speech before conversion.

We remark that both warping and transposition strategies are applied as a pre-processing step according to the conversion case: female to female (labels including 'SF-TF', transposition on both speakers); female to male ('SF-TM', transposition for female, warping for male); male to female ('SM-TF', warping for male, transposition for female). There is no modification for the male to male (SM-TM) since it already represents the most convenient spectral estimation and matching conditions. Note that the number included in the conversion pairs labels showed in the plots represents the speaker identifier.

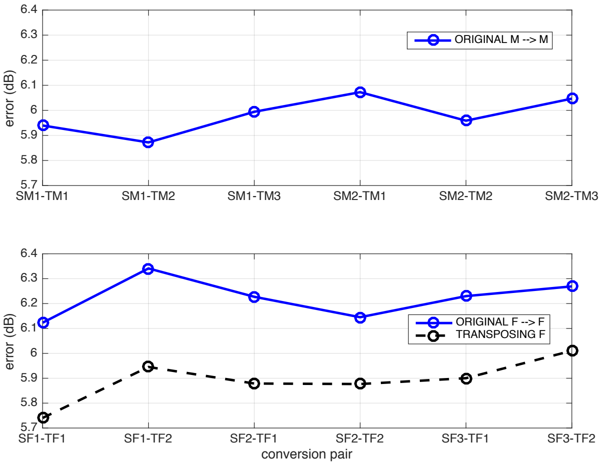

Figure 1: *Spectral conversion error for intra-gender conversion. Top: male to male. Bottom: female to female with (black-dashed) and without (blue) applying pitch transposition.*

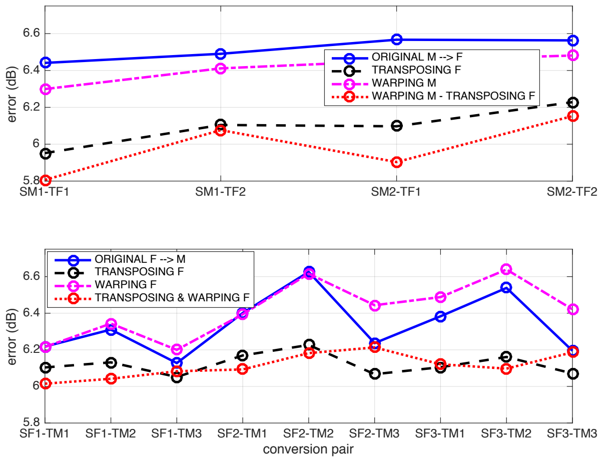

Figure 2: *Spectral conversion error for inter-gender conversion with the original (blue), proposed (red-dotted) and intermediate pre-processing configurations. Top: male to female, bottom: female to male.*

2.4. Statistical modeling error compensation

There exists a modeling error due to limitations of a probabilistic mixture with finite number of components to accurately represent the input features space denoted by x_k. In a GMM-based transformation, this averaging of the information results typically in target features predictions representing over-smoothed spectra. In [26] we proposed to compensate this effect by firstly defining a new transformation filter $Hm_k(\omega)$ in terms of the envelope $X'_k(\omega)$ of the actual feature x'_k *seen* by the mixture:

$$Hm_k(\omega) = |\hat{Y}_k(\omega)| - |X'_k(\omega)|, \tag{2}$$

representing the new predicted envelope $Ym_k(\omega) = X_k(\omega) + Hm_k(\omega)$. Secondly, potential over-emphasized spectral features in $Ym_k(\omega)$ are compensated by applying average amplitude differences between $Ym_k(\omega)$ and $\hat{Y}_k(\omega)$. This strategy proved effective to enhance the converted speech with a perceived reduction of degradations (see [26] for further details).

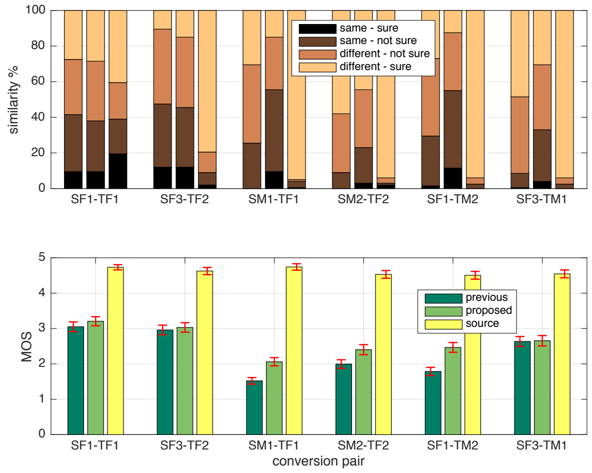

Figure 3: *Target similarity (top) and MOS (bottom) results for six conversion pairs. The three colons per pair corresponding from left to right to our previous conversion method, the proposed pre-processing one, and the original source speech.*

3. Evaluation of the pre-processing configurations

3.1. Speech corpora and training conditions

The data used for the VCC2016 was selected from the DAPS database [33] and down-sampled to 16kHz. It contains five source and five different target speakers, resulting in twenty five conversion pairs, all of them requested by the task of the challenge (see [34] for further information of the VCC2016 task) . The source speakers included three female and two male speakers and conversely for the target ones. The training set consisted of 162 utterances, and 54 additional ones were provided as evaluation set. The mel-TEAP envelope features were extracted from the speech signals also pitch-synchronously, resulting in training sizes within the range $\tilde{[}20,000, 32,000]$ overall. For learning conditions verification, we evaluated the conversion performance using mixtures with 2, 4, 8, 12, and 16 components and found that 12 was the most convenient value in average. The results presented in the following section were therefore obtained using this GMM size with full-covariance matrices.

3.2. Spectral conversion evaluation

As performance measure we computed the average spectral distortion between the mel-scaled spectra given by the target and converted LSFs on a 10-fold cross validation fashion on all the conversion pairs. We evaluated the spectral conversion rates over different pre-processing configurations (the no pre-processing case was labeled as ”ORIGINAL”). The transformation compensation described in section 2.4 was not applied in order to exclusively evaluate the performance of the features mapping for the different spectral conditions on the waveforms.

The results are presented in Fig.1 and Fig.2 for inter and intra gender conversions respectively. For reference, we show in Fig.1 (top) the results for SM-TM conversion although there is no pre-processing considered for this case. Note the reduction of the spectral distortion for the SF-TF conversion (bottom) to a level comparable to the SM-TM conversion when applying the proposed transposition. Similarly, for the SM-TF conversion (Fig.2, top) it can be seen that both pre-processing steps resulted in a reduction of the spectral error. Finally, note that for the female to male conversion (Fig.2, bottom) the warping step only resulted in improved performance in some pairs only after transposing the female speech. The low performance of the warping in this case can be attributed to a lack of optimisation of the warping function and should be investigated deeper.

3.3. Similarity and naturalness evaluation

We firstly evaluated the perceptual impact of the proposed spectral normalisation in terms of target speaker similarity and naturalness on listening tests over 20 listeners. The participants were native english speakers and used high-quality headphones. For simplicity only the three gender combinations involving pre-processing configurations (SF-TF, SM-TF, and SF-TM) were considered. Ten samples of two pairs of each type of these combinations were evaluated, resulting in a total of sixty samples in three different versions: the original recordings of the source speaker and the converted versions with and without pre-processing (both conversions obtained by the compensated transformation previously described, for perceptual evaluation purposes). The different versions were evaluated simultaneously to judge their similarity by comparison with a sample (different utterance) of the target speaker according to four different scores including a certainty level: same-absolutely sure, same-not sure, different-not sure, different-absolutely sure.

The results of the similarity test are shown in Fig. 3 (top). Note that although the performance appears to be highly speakers-pair dependant it shows better scores for the cases involving gender conversion (that we attribute principally to the effect of the frequency warping). For the female to female conversion, the lower conversion error measured objectively does not show a a significant perceptual effect, suggesting somehow a compensation in the spectral mapping process of the observed amplitude over-estimations.

The naturalness test results (Fig.3, bottom) obtained in terms of Mean Opinion Scores (MOS) also show a speakers dependency again and center the benefits of the proposed spectral normalisation on the gender conversions. Note the higher scores compared to the methodology based on previous work (that is reported already as providing quality improvements [26]). Both similarity and naturalness tests were carried out using an interface inspired in MUSHRA tests ([35]) that allows listeners to replay any sample as much as they feel comfortable with their response and to score using a continuous scale with the proposed answers proportionally distributed for each type of test.

4. Results at the Voice Conversion Challenge 2016

We show in Fig. 4 and Fig. 5 the results of the similarity and naturalness tests respectively carried out at the VCC2016 where capital letters represent the entries of the 17 participants (our system using the proposed pre-processing configurations is labeled ’K’, a GMM baseline system as ’Bsl’, and the original source and target speakers as ’Src’ and ’Tar’ respectively). A detailed report of the results can be found in [36] with an extensive analysis of the results. Note that at difference of the tests reported in the previous section the samples were evaluated individually at the challenge (one to one matching for similarity comparison and individual naturalness scoring). This may explain some of the higher scores of our system in the challenge since it appears easier to penalise slight differences or degradations by simultaneously comparing transformed and

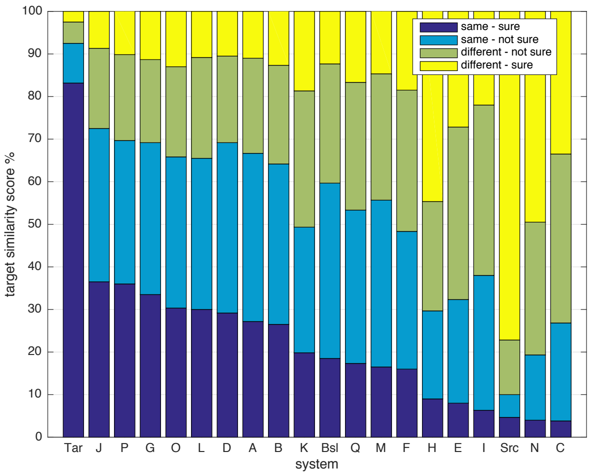

Figure 4: *Target similarity results of the VCC2016 (our system: K). All conversion pairs included.*

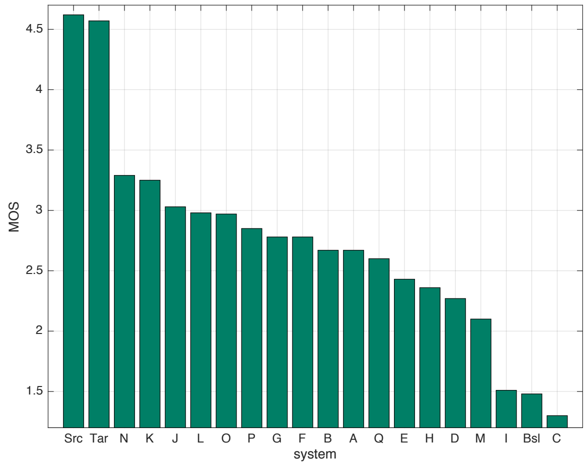

Figure 5: *Naturalness results of the VCC2016 (our system: K). All conversion pairs included.*

non-transformed samples from fixed conversion pairs.

Looking at the percentage of samples judged as absolutely similar to the target (response "same-absolutely sure") shown in Fig. 4 our system shows similar performance to the baseline GMM-based one. While our features conversion process is based on the same framework we expected a slightly higher performance following the incorporation of frequency warping. We assume the highest conversion scores represent systems exploiting recent techniques such as those based on deep learning.

In Fig. 6 we show a comparison per-gender combination case that includes only the baseline, our system, and the best score per case. The scores confirm a comparable performance to that of the baseline system but lower than the most competitive ones. An optimisation of the warping function according to the conversion pair may help to reduce this performance gap. Note however, that the best scores (around 40%) do not yet appear fully satisfactory in terms of robust target similarity.

Concerning the naturalness test (MOS) our scores are among the most competitive ones. Fig. 5 shows that our system ranked in second place and very close to the best system overall ('N'). Note however that this system performs significantly low in terms of target similarity, which suggests a low degree of transformation applied to the waveforms. According to our scores our systems clearly outperforms the majority of entries, denoting the benefits of our methodology as a whole.

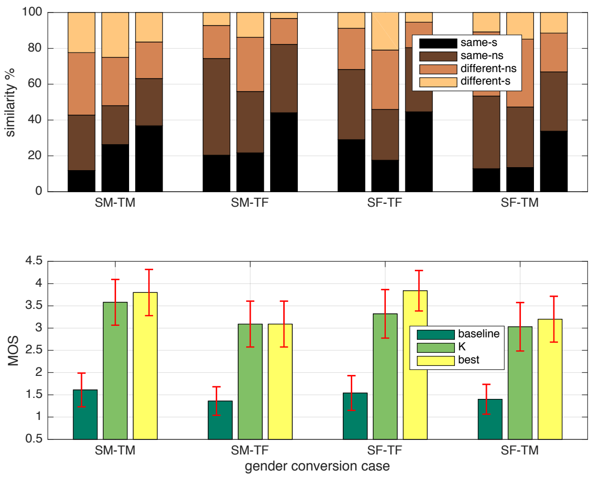

Figure 6: *Target similarity (top) and MOS (bottom) results averaged per gender conversion case. The three colons from left to right correspond to the baseline, our system, and best score.*

Looking at each gender conversion case (Fig. 6) our system performs significantly better than the baseline and very close to the best scores, being the best for male to female conversion (best spectral processing conditions). These findings can be extended and verified in [36].

The results obtained in the VCC2016 allow us to claim benefits overall of applying warping for spectral alignment and efficient spectral envelope processing to reduce the risk of significant degradations on the converted speech due to poor estimated spectral features. Note that this concept refer exclusively to the features extraction task; and therefore, it can be applied on frameworks based on models others than GMM.

5. Conclusions

In this paper, we presented the system that was the basis of our entry for the Voice Conversion Challenge 2016. We incorporated pre-processing configurations to previous work in GMM-based conversion in order to normalise the spectral conditions between speakers. We applied global frequency warping to align the spectral features for gender conversion and pitch transposition on female voices to reduce over-estimations on the spectral envelope information observed on high-pitched speech. This methodology resulted in higher similarity and naturalness rates following objective and subjective evaluations.

At the listening tests conducted for the Voice Conversion Challenge 2016 our system was among the most competitive in terms of naturalness (ranked second overall) while maintaining GMM-based conversion performance, demonstrating the benefits of our methodology to improve converted speech quality.

As future work we will study outperforming features conversion strategies (e.g. deep learning), optimised frequency warping strategies (e.g. [37], and to clarify the benefits of transposing female speech on the envelope extraction by exhaustive evaluation on female voices.

6. References

[1] D. G. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: factors responsible for quality," in *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1985. ICASSP '85.*, 1985, pp. 748–751.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *in Proc. of ICASSP'88*, 1988.

[3] H. Valbret, E. Moulines, and T. J.P., "Voice transformation using psola technique," in *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP '92.*, vol. 1, 1992, pp. 145–146.

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

[5] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: control and conversion," *Speech Communication*, 1995.

[6] W. Verhelst and J. Mertens, "Voice conversion using partitions of spectral feature space," in *Proc. of IEEE-ICASSP'96*, 1996.

[7] M. Hashimoto and N. Higuchi, "Training data selection for voice conversion using speaker selection and vector field smoothing," in *Proc. of ICSLP'96*, 1996.

[8] K. Lee, D. Youn, and I. Cha, "A new voice transformation method based on both linear and non-linear prediction analysis," in *Proc. ICSLP'96*, 1996.

[9] E.-K. Kim, S. Lee, and Y.-H. Oh, "Hidden markov model based voice conversion using dynamic characteristics of speaker," in *In Proceedings of the European Conference on Speech Communication and Technology, 1997, EUROSPEECH '97.*, 1997, pp. 1311–1314.

[10] L. Arslan and D. Talkin, "Speaker transformation using sentence hmm-based alignments and detailed prosody modification," in *Proc. of IEEE-ICASSP'98*, 1998.

[11] L. Schwardt and J. du Preez, "Voice conversion based on static speaker characteristics," in *Proc. of IEEE-COMSIG'98*, 1998.

[12] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE-TASAP*, vol. 6, no. 2, pp. 131–142, 1998.

[13] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *In Proceedings of ICASSP '98.*, vol. 1, 1998, pp. 285–288.

[14] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameters trajectory," *IEEE-TASLP*, vol. 15, no. 8, 2007.

[15] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on gaussian mixture model," in *In Proceedins of the International Conference on Spoken Language Processing, 2006. INTERSPEECH '06*, Pittsburgh, USA, September 2006, pp. 2446–2449.

[16] A. Mouchtaris, J. Van der Spiegel, and P. . Mueller, "Non-parallel training for voice conversion based on a parameter adaptation approach," *IEEE-TASLP*, vol. 14, no. 2, pp. 952–963, 2006.

[17] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE TASLP*, vol. 18, no. 5, pp. 922–931, 2010.

[18] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling for parallel or nonparallel corpora," *IEEE TASLP*, vol. 20, no. 4, pp. 1313–1323, 2012.

[19] L. Chen, Z. Ling, L. Liu, and L. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE-TALSP*, vol. 22, no. 12, 2014.

[20] Z. Wu, T. Virtanen, and E. Siong, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE TASLP*, vol. 22, no. 10, pp. 1506–1521, October 2014.

[21] S. Imai and Y. Abe, "Cepstral synthesis of japanese from cv syllable parameters," in *Proc. of ICASSP'80*, 1980.

[22] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. of DAFx'05*, Spain, 2005.

[23] F. Villavicencio, A. Röbel, and X. Rodet, "Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation," in *proc. of ICASSP*, 2006.

[24] F. Villavicencio, A. Röbel, and X. Rodet, "Applying improved spectral modeling for high-quality voice conversion," in *Proc. of ICASSP*, 2009.

[25] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," in *Proc. of INTERSPEECH*, vol. 1, Tokyo, Japan, 2010, pp. 2162–2165.

[26] F. Villavicencio, J. Bonada, and Y. Hisaminato, "Observation-model error compensation for enhanced spectral envelope transformation in voice conversion," in *Proc. of IEEE-MLSP'15*, 2015.

[27] K. Kobayashi, T. Toda, G. Neubig, and S. Sakti, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. of INTERSPEECH'14*, 2014, pp. 2514–2518.

[28] J. Bonada, "Wide-band harmonic sinusoidal modeling," in *In Proc. of DAFx'08*, Helsinki, Finland, 2008, pp. 265–272.

[29] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[30] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *IEICE (in Japanese)*, vol. 62, no. 4, pp. 10–17, 1979.

[31] A. Röbel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modelling with unknown model order," *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343–1350, 2007.

[32] F. Villavicencio and E. Maestre, "Gmm-pca based speaker-timbre conversion on full-quality speech," in *In Proc. of the 7th Speech Synthesis Workshop (SSW7)*, 2010, pp. 56–61.

[33] M. G.J. (2015) Device and produced speech datdata (daps). [Online]. Available: https://archive.org/details/daps_dataset

[34] T. Toda, L. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in *Proc. of INTERSPEECH*, 2016, (submitted).

[35] [Online]. Available: http://sourceforge.net/projects/matlabmushra/

[36] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results," in *Proc. of INTERSPEECH*, 2016, (submitted).

[37] Y. Agiomyrgiannakis, "Voice morphing that improves tts quality using an optimal dynamic frequency warping-and-weighting transform," in *Proc. of ICASSP*, 2016.