The Voice Conversion Challenge 2016

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

The Voice Conversion Challenge 2016

Tomoki Toda[1], *Ling-Hui Chen*[2], *Daisuke Saito*[3],
Fernando Villavicencio[4], *Mirjam Wester*[5], *Zhizheng Wu*[5], *Junichi Yamagishi*[4,5]

[1] Information Technology Center, Nagoya University, Japan
[2] University of Science and Technology of China, China
[3] The University of Tokyo, Japan
[4] National Institute of Informatics, Japan
[5] The Centre for Speech Technology Research, The University of Edinburgh, UK

vcc2016@vc-challenge.org

Abstract

This paper describes the Voice Conversion Challenge 2016 devised by the authors to better understand different voice conversion (VC) techniques by comparing their performance on a common dataset. The task of the challenge was speaker conversion, i.e., to transform the voice identity of a source speaker into that of a target speaker while preserving the linguistic content. Using a common dataset consisting of 162 utterances for training and 54 utterances for evaluation from each of 5 source and 5 target speakers, 17 groups working in VC around the world developed their own VC systems for every combination of the source and target speakers, i.e., 25 systems in total, and generated voice samples converted by the developed systems. These samples were evaluated in terms of target speaker similarity and naturalness by 200 listeners in a controlled environment. This paper summarizes the design of the challenge, its result, and a future plan to share views about unsolved problems and challenges faced by the current VC techniques.

Index Terms: Voice conversion, speech synthesis, evaluation challenge

1. Introduction

Voice conversion (VC) is a technique to modify a speech waveform which freely converts non-/para-linguistic information while preserving linguistic information. To develop this technique, we need a deep understanding of how to effectively factorize speech acoustics into its individual components such as linguistic, non-linguistic, and para-linguistic information using various technologies, such as speech analysis, speech synthesis, acoustic modeling, and machine learning. Moreover, VC has great potential to develop various applications not only for flexible control of speaker identity of synthetic speech in text-to-speech (TTS) [1] but also as a speaking aid for vocally handicapped people such as dysarthric patients [2] and laryngectomees [3], as a voice changer to flexibly generate various types of emotional [4] and expressive speech [5], for vocal effects to produce more varieties of singing voices [6, 7], for enhanced mobile speech communication using wideband speech [8] and silent speech [9], accent conversion for computer assisted language learning [10], and so on. Therefore, it is worthwhile to study this technique for both scientific purposes and industrial applications.

VC research has a relatively long history from the late 1980s onwards [11]. Originally it was studied to achieve speaker conversion to make it possible to synthesize various speakers' voices in a TTS system, in particular focusing on cross-language VC enabling a user to produce his/her own voice in a different language for speech-to-speech translation [12]. Although a simple conversion function, such as a global linear transformation or frequency warping with a constant warping rate for modifying the spectral envelope, is capable of changing speaker identity, it is insufficient to convert a specific source speaker's voice into another specific target speaker's voice. A more sophisticated conversion function to effectively model a nonlinear mapping between source and target voices needs to be developed to convert speaker identity.

To develop such a nonlinear conversion function, a data-driven approach was applied to VC [1], making it possible to formulate VC as a regression problem [13]. Thanks to this well-formulated approach, VC research has become popular by widely sharing various techniques developed by individual researchers. However, there has been no predefined publicly available protocol for fair scientific comparisons, and therefore, individual researchers have normally conducted their own VC research using individually-owned speech corpora or public corpora developed for other research purposes. As the performance of VC systems developed using the data-driven approach strongly depends on the individual speech corpora used, it is not straightforward to compare across several VC techniques.

The use of a common dataset to evaluate different techniques is very useful for comparing their performance, clarifying existing problems to be addressed, and developing better techniques. This type of evaluation was performed for speech recognition throughout the 1990s. Recently, evaluation activities have become popular in various research fields, such as speaker recognition, machine translation, para-linguistic analysis, and so on. In speech synthesis as well, the Blizzard Challenge has been carried out since 2005 [14], enabling and measuring the improvements of corpus-based TTS technologies. These activities are obviously helpful for developing better research communities and enabling significant technical progress using data-driven approaches.

Inspired by these evaluation activities, the authors launched the Voice Conversion Challenge 2016 (VCC 2016). The objective of the challenge is to better understand different VC techniques by comparing their performance using a freely-available dataset as a common dataset, bringing together different teams to look at a common goal, and to share views about unsolved problems and challenges faced by the current VC techniques.

http://dx.doi.org/10.21437/Interspeech.2016-1066

The VCC 2016 focuses on speaker conversion as the most basic VC task, i.e., to transform the voice identity of a source speaker into that of a target speaker while preserving the linguistic content. Research groups working in VC around the world have been invited to build VC systems and submit converted samples to be evaluated through listening tests.

This paper presents describes the set-up of VCC 2016. After briefly summarizing a basic VC framework for speaker conversion in Section 2, we will explain the task of the challenge, including the guidelines for participants, the details of the common dataset, and how the evaluation of different VC systems was designed, in Section 3. Then, the main results of the challenge will be presented in Section 4, followed by our future plans described in Section 5.

2. Voice Conversion

In this paper, as one of the most popular VC frameworks to achieve speaker conversion, we focus on the VC framework where only a speech signal of the source speaker is given as the input for conversion.

2.1. Basic Framework for Speaker Conversion

As both segment and prosodic features depend on individual speakers, corresponding speech parameters (e.g., spectral envelope and aperiodic parameters as segmental features and F_0 and duration patterns as prosodic features) basically need to be modified to convert speaker identity. However, these speech parameters are also affected by other information, such as linguistic information, which should be kept unchanged. Therefore, it is essential to develop a conversion function to carefully modify these speech parameters to achieve speaker conversion.

The data-driven approach handles this issue by using a parallel speech dataset consisting of utterance pairs of the source and target speakers [1]. A training dataset is developed by performing time frame alignment between the source and target voices in each utterance pair so that each time aligned frame pair shares the same linguistic information. Assuming that the acoustic differences observed between the source and target voices in the time aligned frame pairs are caused by only their speaker difference, they are used as a supervised training dataset to determine a conversion function. The resulting conversion function is used to convert arbitrary utterances of the source speaker without any additional information.

2.2. Speech Parameterization and Waveform Generation

The use of high quality speech analysis/synthesis techniques is important in VC. Various sophisticated techniques, such as harmonic plus noise model (HNM) [15] and STRAIGHT [16], have been often used to extract high quality speech features from a speech waveform, and also to generate a speech waveform from the converted speech features.

Regarding segmental features, the spectral envelope is often parameterized into a low-dimensional representation, such as line spectral pairs (LSPs) [17] or mel-generalized cepstral coefficients [18], which can be easily handled in the conversion function. Recently, a data-driven method to parameterize the spectral envelope has also been proposed [19]. Moreover, aperiodic components [20], phase components [21], or one-pitch waveform shapes [22] may also be parameterized to convert an excitation signal.

As for the prosodic features, F_0 and duration patterns may be parameterized to properly handle their supra-segmental characteristics, which are not well converted within the frame-wise conversion process. Several methods to achieve such a parameterization have been proposed [23, 24, 25] but it is not straightforward to do it without any linguistic information. Consequently, very simple parameters to represent only their static properties, e.g., global mean and variance of log-scaled F_0 values, are often used.

2.3. Conversion Function

To achieve a nonlinear regression mapping, various conversion functions have been proposed. They are mainly grouped into 3 types: 1) a piece-wise linear mapping using probabilistic models, e.g., Gaussian mixture models (GMM) [26, 27], bidirectional associative memories (BAM) [19], and restricted Boltzmann machines (RBM) [19, 28], 2) a nonlinear mapping, e.g., dynamic kernel partial least squares regression [29], Gaussian process regression with kernel functions [30, 31], neural networks (NN) [32], and deep neural networks (DNN) [19], and 3) an exemplar-based mapping, e.g., non-negative matrix factorization (NMF) [33, 34]. Moreover, to produce naturally varying speech parameters, it is essential to model the dynamic properties of speech parameters. In order to allow the conversion process to consider temporal correlations over a speech parameter sequence, several techniques have been proposed, e.g., 1) trajectory-based conversion [27] capable of being widely applied to parametric conversion functions, 2) joint distribution modeling with Gaussian processes [30, 31], and 3) the use of recurrent structure in NN/DNN [35].

In a standard regression problem, the conversion function is usually optimized to minimize a total conversion error between the converted and target speech parameters. However, this optimization framework often causes excessively smoothed speech parameters, making the converted speech sound muffled. To address this oversmoothing problem, there are several methods have been proposed, e.g., 1) a method to model additional features to sensitively capture the oversmoothing effect, such as global variance (GV) [27] and modulation spectrum (MS) [36], 2) a method to keep characteristics of natural speech parameters by partially using the source speech parameters, such as dynamic frequency warping (DFW) [37], and 3) a method to alleviate the averaging process to implement a sparse constraint as in the exemplar-based conversion [33, 34].

3. Voice Conversion Challenge

3.1. Task

The task of the challenge was speaker identity conversion. The dataset of the challenge consisted of parallel corpora (same utterances) of a set of source and target speakers (all different). The participants were asked to develop conversion systems and to produce converted data for all the source-target pairs combinations. Note that phonetic transcriptions were not included in the dataset (only waveforms). A detailed description of the dataset is provided in the following section.

The main guidelines to participate with an entry were as follows:

- Manual editing or system tuning in the conversion step was not allowed. Manual optimisation of individual conversion systems was allowed only in the training stage.
- Manual transcriptions (phoneme or linguistic information) of the training and/or evaluation were not allowed. However, automatic speech recognition systems may be used to generate

this information.

- The use of content from other source and target speakers from the VCC dataset to develop a conversion system for a specific source-target pair was not allowed.
- The transformation of any acoustic features, including supra-segmental and duration features was allowed.
- The use of data other than the VCC 2016 dataset for training purposes was allowed.
- Participants were free to discard content (utterances) of the training set at their convenience.
- Participants were not allowed to submit multiple entries.

The participants were asked to submit their entry (only waveforms) after generating the converted material from the evaluation data and to fill in a questionnaire to obtain information and a description of their conversion system and their main related techniques. Further, the entries were evaluated in terms of target speaker similarity and naturalness using listening tests carried out by the organisers, as described in Section 3.3.

3.2. Dataset

The dataset used in VCC 2016 is based on the DAPS (Data And Production Speech) dataset [38], which was recorded by professional US English speakers in a professional recording studio without significant noise effects and is available online for free[1]. The "clean" version of the original recordings, in which most of the non-speech sounds were removed, was used as the dataset in this challenge. The recorded audio includes about 13 minutes of speech sounds recorded by each of the 20 speakers. The recordings were down-sampled to 16 kHz for this challenge.

10 speakers, including 5 female speakers and 5 male speakers, were select from the 20 speakers in the original dataset for this challenge. The audio files for each speaker were manually segmented into 216 parallel short sentences. 162 sentences were used as training data and were released to registered participants for building and developing their systems. The remaining 54 sentences were left as test data for evaluation and were released to participants about one week before submitting their converted voices. Table 1 shows the details of the VCC 2016 dataset which consists of 5 source speakers and 5 target speakers. The participants were asked to build systems for all the 5×5=25 combinations of source-target pairs. During the evaluation, one female source speaker was removed because of Lombard effects in the recordings, another male target speaker was also removed because of his fast speaking rate. Therefore, 4×4=16 source-target pairs were evaluated in total in the formal evaluation.

Table 1: *Number of source and target speakers, and number of training and evaluation sentences.*

	# of speakers		# of sentences	
	Male	Female	Training	Evaluation
Source	2	3	162	54
Target	3	2	162	n/a

3.3. Evaluation methodology

Subjective listening tests were designed to perceptually evaluate the naturalness and speaker similarity of the converted samples for 16 of the 25 source-target (ST) pairs. A general description of the test is given below, a more detailed description of the listening test design is given in [39].

Naturalness. Subjects were asked to evaluate the naturalness of voice converted samples and natural speech on a scale from 1 (completely unnatural) to 5 (completely natural). The 16 ST pairs were divided into two groups, balanced across gender conditions. Each subject was given one set of 8 ST pairs to rate, which corresponds to 152 stimuli. ((18 participants * 8 ST) + 4 source + 4 target = 152 stimuli)

Similarity. To measure the similarity of VC samples the Same/Different paradigm was used. Subjects were given two samples and were asked the following: *"Do you think these two samples could have been produced by the same speaker? Some of the samples may sound somewhat degraded/distorted. Please try to listen beyond the distortion and concentrate on identifying the voice. Are the two voices the same or different? You have the option to indicate how sure you are of your decision."* The scale for judging was: "Same, absolutely sure", "Same, not sure", "Different, not sure" and "Different, absolutely sure". Each subject rated three ST pairs. The trials consisted of comparisons of VC samples with either the source speaker or the target speaker.

200 subjects participated in the experiment, which took, on average, an hour to complete. The subjects listened to the stimuli over headphones, in sound-treated booths.

4. The results

4.1. Participants

Late 2015, participants from industry and academia were invited to take part in the challenge, 25 sites registered, and 17 submitted their entries early 2016. Table 2 shows the list of participants for the challenge. In the table, team names and their corresponding affiliations are described, and they are listed in random order.

4.2. Baseline system

The baseline system is based on the voice conversion toolkit within the open-source Festvox system[2], as in our previous work [40], we found the toolkit can achieve similar performance to other state-of-the-art voice conversion or speech synthesis adaptation techniques. The toolkit is based on the joint density Gaussian mixture model with maximum likelihood parameter trajectory generation considering global variance as proposed in [27]. The number of Gaussian mixtures was empirically set to 64 without any tuning. The system was trained on the whole training data without using any additional resources.

4.3. Results of listening tests

Figure 1 shows the naturalness MOS results plotted against similarity to the target speaker. The 17 participants are denoted by blue diamonds and the letters A ... Q, the source and target by yellow diamonds and the baseline by a green diamond. For naturalness, systems N and K significantly outperform all other systems. For similarity, there is quite a large cluster of systems that score similarly well (J, P, D, G, A, O, L and B) and outperform the baseline. Further details and analysis of the results can be found in [39].

[1] `https://archive.org/details/daps_dataset`.

[2] Festvox is available at: `http://festvox.org/`

Table 2: *The list of participants of VCC 2016. They are listed in random order.*

Team name	Affiliation
NII	National Institute of Informatics, Tokyo, Japan
NPU-I2R-NTU	Northwestern Polytechnical University, Xi'an, China, Institute for Infocomm Research, Singapore and Nanyang Technological University, Singapore
UTokyo	The University of Tokyo, Japan
USTC-NELSLIP	University of Science and Technology of China, NEL-SLIP, Hefei, Anhui, China
UCL	University College London, UK
Hulk	Speech Lab, Shanghai Jiao Tong University, China
NU-NAIST	Nagoya University and Nara Institute of Science and Technology, Japan
IIT Kharagpur ABSP Lab	Indian Institute of Technology, Kharagpur, India
HCCL-CUHK	Human-Computer Communications Laboratory, The Chinese University of Hong Kong, Hong Kong
CSTR	The Centre for Speech Technology Research, The University of Edinburgh, UK
CASIA-NLPR-Taogroup	National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
AHOLAB	University of the Basque Country, Bilbao, Spain
Team Initiator	Tsinghua University, Beijing, China
DA-IICT	Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India
VoiceKontrol	Center for Spoken Language Understanding (CSLU), Oregon Health & Science University, Portland, OR, USA
AST	Academia Sinica, Taipei, Taiwan
IIIT-H	International Institute of Information Technology, Hyderabad, India

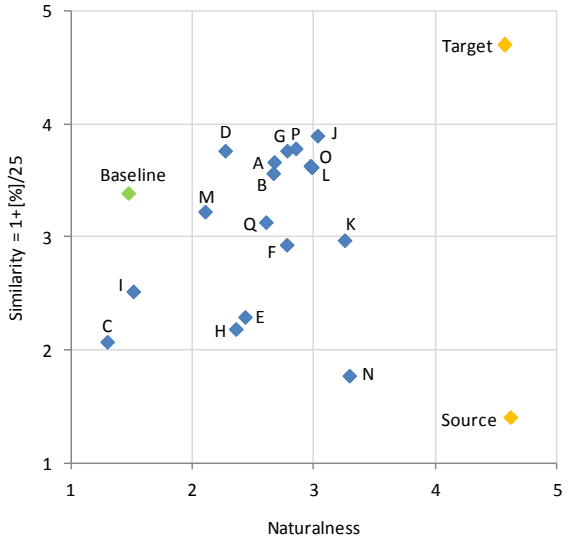

Figure 1: *Overall naturalness MOS versus similarity to target speaker. Figure kindly provided by Daniel Erro (AHOLAB).*

5. Discussion and Future plan

Almost all systems outperform the baseline in terms of naturalness. In terms of similarity, about half the systems obtain better results than the baseline. Hence, we think that the VC community needs to have a more appropriate baseline system for achieving more meaningful experiments in the future. Furthermore, it is clear that achieving good quality and speaker similarity together in a system seems to be a yet unsolved challenge.

Listening tests for the evaluation also have room for improvement. For instance, the majority of listeners who participated in the evaluation this time are British English speakers while the speakers used for the voice conversion are American English. This could mean that the current listeners are less sensitive to prosody differences in the converted speech utterances. Ultimately it would be nice if we can compare spectral and prosody differences of voice converted samples in a controlled way.

Suggestions for the future voice conversion challenges given by participants include fewer or more training utterances, the use of a non-parallel corpus, and the use of speech data recorded in non-ideal acoustic conditions.

The organisers of the voice conversion challenge are also contributing to the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenge. Some of newly built voice conversion methods will be interesting for future ASVspoof challenges. Therefore we plan to organise the next voice conversion challenges in synchronisation with the ASVspoof challenge.

6. Conclusion

This paper has presented the Voice Conversion Challenge 2016 (VCC 2016), which has been a valuable exercise in developing voice conversion (VC) systems using a common dataset. The Challenge has successfully demonstrated performance of the current VC techniques on a speaker conversion task and has helped to share views about unsolved problem. We see the VCC 2016 as the start of a series of the Challenges on VC for not only speaker conversion but also other various applications.

6. Acknowledgements

We are grateful to COLIPS for sponsoring the evaluation of the VCC 2016, and iFLYTEK for supporting the VCC database development. This work was supported in part by the EPSRC under Programme Grant EP/I031022/1 (Natural Speech Technology) and EP/J002526/1 (CAF), and JSPS KAKENHI Grant Number 26280060. The VCC database and listening test results are permanently available at `http://dx.doi.org/10.7488/ds/1430`.

7. References

[1] M. Abe, S. Nakamura, and K. Shikano, "Voice conversion through vector quantization," *The Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[2] A. B. Kain, J. P. Hosom, X. Niu, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743–759, 2007.

[3] H. Doi, T. Toda, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 1, pp. 172–183, 2014.

[4] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken english," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[5] O. Türk and M. Schröder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 18, no. 5, pp. 965–973, 2010.

[6] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," in *Proc. INTERSPEECH*, 2010, pp. 2162–2165.

[7] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, "Voice timbre control based on perceived age in singing voice conversion," *Information and Systems, IEICE Transactions on*, vol. E97-D, no. 6, pp. 1419–1428, 2014.

[8] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.

[9] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2505–2517, 2012.

[10] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920–932, 2009.

[11] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988, pp. 655–658.

[12] M. Abe, K. Shikano, and H. Kuwabara, "Statistical analysis of bilingual speaker's speech for cross-language voice conversion," *The Journal of the Acoustical Society of America*, vol. 90, no. 1, pp. 76–82, 1991.

[13] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using psola technique," *Speech Communication*, vol. 11, no. 2–3, pp. 175–187, 1992.

[14] A. W. Black and K. Tokuda, "The Blizzard Challenge – 2005: evaluating corpus-based speech synthesis on common datasets," in *Proc. INTERSPEECH*, 2005, pp. 77–80.

[15] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 1, pp. 21–29, 2001.

[16] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[17] F. Soong and B. Juang, "Optimal quantization of lsp parameters," *Speech and Audio Processing, IEEE Transactions on*, vol. 1, no. 1, pp. 15–24, 1993.

[18] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," in *Proc. ICSLP*, 1994, pp. 1043–1045.

[19] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 1859–1872, 2014.

[20] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on gmm with straight mixed excitation," in *Proc. INTERSPEECH*, 2006, pp. 2266–2269.

[21] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proc. ICASSP*, 2001, pp. 813–816.

[22] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1301–1312, 2006.

[23] B. Gillett and S. King, "Transforming F_0 contours," in *Proc. INTERSPEECH*, 2003, pp. 101–104.

[24] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 6, pp. 1042–1053, 2015.

[25] K. Yutani, Y. Uto, Y. Nankaku, T. Toda, and K. Tokuda, "Simultaneous conversion of duration and spectrum based on statistical models including time-sequence matching," in *Proc. INTERSPEECH*, 2008, pp. 1072–1075.

[26] 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.

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

[28] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on speaker-dependent restricted boltzmann machines," *Information and Systems, IEICE Transactions on*, vol. E67-D, no. 6, pp. 1403–1410, 2014.

[29] E. Helander, H. Silé, T. Virtanen, and M. M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 806–817, 2012.

[30] N. Pilkington, H. Zen, and M. Gales, "Gaussian process experts for voice conversion," in *Proc. INTERSPEECH*, 2011, pp. 2761–2764.

[31] N. Xu, Y. Tang, J. Bao, A. Jiang, X. Liu, and Z. Yang, "Voice conversion based on gaussian processes by coherent and asymmetric training with limited training data," *Speech Communication*, vol. 58, pp. 124–138, 2014.

[32] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 954–964, 2010.

[33] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion using sparse representation in noisy environments," *Information and Systems, IEICE Transactions on*, vol. E96-A, no. 10, pp. 1946–1953, 2013.

[34] Z. Wu, T. Virtanen, E. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 10, pp. 1506–1521, 2014.

[35] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 4869–4873.

[36] S. Takamichi, T. Toda, A. Black, G. Neubig, S. Sakti, and S. Nakamura, "Post-filters to modify the modulation spectrum for statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 24, no. 4, pp. 757–767, 2016.

[37] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 922–931, 2010.

[38] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? – a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, Aug 2015.

[39] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the Voice Conversion Challenge 2016 evaluation results," in *(submitted to) Interspeech*, 2016.

[40] Z. Wu, P. L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.-H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *Audio, Speech and Language Processing, IEEE/ACM Transactions on*, vol. 24, pp. 768–783, 2016.