Parallel Voice Conversion Based on a Continuous Sinusoidal Model

Mohammed Salah Al-Radhi
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
malradhi@tmit.bme.hu

Tamás Gábor Csapó
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
csapot@tmit.bme.hu

Géza Németh
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
nemeth@tmit.bme.hu

***Abstract*—The main challenge introduced in current voice conversion is the tradeoff between speaker similarity and computational complexity. To tackle the latter problems, this paper introduces a novel sinusoidal model applied for voice conversion (VC) with parallel training data. The conventional source-filter based techniques usually give sound quality and similarity degradation of the converted voice due to parameterization errors and over smoothing, which leads to a mismatch in the converted characteristics. Therefore, we developed a VC method using continuous sinusoidal model (CSM), which decomposes the source voice into harmonic components to improve VC performance. In contrast to current VC approaches, our method is motivated by two observations. Firstly, it allows continuous fundamental frequency (F0) to avoid alignment errors that may happen in voiced and unvoiced segments and can degrade the converted speech, that is important to maintain a high converted speech quality. We secondly compare our model with two high-quality modern (MagPhase and WORLD) vocoders applied for VC, and one with a vocoder-free VC framework based on a differential Gaussian mixture model that was used recently for the Voice Conversion Challenge 2018. Similarity and intelligibility are finally evaluated in objective and subjective measures. Experimental results confirmed that the proposed method obtained higher speaker similarity compared to the conventional methods.**

Keywords—voice conversion, sinusoidal model, continuous F0, neural network

I. INTRODUCTION

Voice conversion (VC), as considered in this paper, aims to modify the speech signal of a source speaker into that of a target speaker. It has great potential in the development of various speech tasks such as Text-to-Speech (TTS) [1], speaking assistance [2], and speech enhancement [3].

Numerous statistical approaches have been employed for mapping the source and the target features. Gaussian mixture model (GMM) [4] is a typical form of VC that requires a source-target alignment for training the conversion models. Some other statistical methods have also been proposed for VC, such as non-negative matrix factorization (NMF) [5], restricted Boltzmann machines [6], variational auto-encoders [7], and maximum likelihood estimation of spectral parameter trajectory [8]. Although these techniques achieve improvements of converting the voice signal into the target one, the naturalness of the sound quality usually deteriorates due to the over-smooth phenomenon or discontinuity problems, which makes the converted speech sound muffled. Recently, deep neural networks (DNNs) have significantly improved the conversion accuracy of statistical VC techniques. Deep belief networks [9], generative adversarial networks [10], deep bidirectional long short-term memory [11] have been recently proposed to preserve the sound quality. Notably, the similarity of the converted voices is still degraded in terms of subjective quality due to model complexity and computational expense. Hence, it is desirable to develop a VC technique to convert speech to more natural-sounding speech with simple DNN models.

Most of the VC systems found in the literature can be built either using a parallel framework in which source and target speakers read out the same set of utterances, or using a non-parallel framework in which the target speaker's utterances are different from those of the source speaker. However, in practice, the subjective experiment results in [12] [13] yield that the average performance of the non-parallel VC system is not outperformed by the parallel VC system. The main reason behind this challenging issue is that it is usually hard to achieve an accurate non-parallel frame alignment between speaker utterances and, therefore, a parallel data-driven approach will be used in this work.

In essence, a well-designed VC system often consists of analysis, conversion, and synthesis modules. The process of parametrizing the input waveform into acoustic features and then synthesizing the converted waveform based on the converted features is one of the major factors that may degrade the performance of VCs. For this, the characteristics of the speech vocoder (analysis/synthesis system) given to the VC are of paramount importance.

Various parametric vocoders, see [14] for comparison, have been used to model the speech signal. In general terms, we can group the state-of-the-art vocoders based VC into three categories. a) Source-filter models: STRAIGHT [15] and mixed excitation [16]; b) Sinusoidal models: Harmonic plus Noise Model [17] is the only model has been found in the literature based VC; c) end-to-end complex models: WaveNet-based waveform generation [18] and Tacotron [19]. In the face of their clear differences, each model has advantages to work reasonably well, for a particular speaker or gender conversion task, which make them attractive to researchers. Nonetheless, such mismatch between the trained, converted, and tested features still exist, which often causes significant quality and similarity degradation. Consequently, simple and uniform vocoders, which would handle all speech sounds and voice qualities (e.g., creaky voice) in a unified way, are still missing in VC.

There seem to be three important factors that should be taken into consideration in the design and development of a VC system. Firstly, the most common feature in most of the above-mentioned VC techniques is the fact that they are based on the spectral envelope (SE). Although SE contains enough information to convert the original speech signal onto that of the target speaker, SE is not enough alone to achieve the desired converted results, for particular applications, even with a better SE estimation method. Secondly, traditional conversion systems focus on the prosodic feature represented by the discontinuous fundamental frequency (F0) assumption that depends on a binary voicing decision. Therefore, modelling of F0 in VC applications is problematic because of the differing nature of F0 observations between voiced and unvoiced speech regions. An alternative solution of increasing the accuracy of the acoustic VC model is using a continuous F0 (contF0) to avoid alignment errors that may happen in voiced and unvoiced segments and can degrade the converted speech. It should be pointed out to the third issue that leads to the degradation of the performance of VC is that most of the existing VC techniques discard or does not typically preserve phase spectrum information. However, the effectiveness of phase information in detecting synthetic speech has recently been proved by [20]. Hence, one possible way of enhancing the accuracy of VC models is to incorporate phase information in order to achieve superior synthesized speech. Therefore, it is still worth to develop advanced vocoder based VC for achieving high-quality converted speech.

To tradeoff between the complexity of the model and conversion accuracy in statistical VC, we propose to use a sinusoidal type synthesis model based on contF0. Moreover, the goal of this paper is to evaluate the performance of a continuous sinusoidal model (CSM) that is suitable for statistical modeling on a voice conversion system. The remainder of the paper is organized as follows. In Section II, we propose the novel idea of using CSM-based voice conversion considering continuous F0 and feed-forward neural network. Datasets, experimental conditions, and baseline VC systems are described in Section III. In Section IV, objective and subjective evaluations are presented. Finally, we summarize this paper in Section V and suggest avenues for future research.

II. PROPOSED METHOD

A. Continuous Sinusoidal Model

Continuous vocoder based sinusoidal model (CSM) was designed to overcome shortcomings of discontinuity in the speech parameters and the computational complexity of modern vocoders. The novelty behind this vocoder is to use harmonic features to facilitate and improve the synthesizing step before speech reconstruction.

By keeping the number of our previous source-filter vocoder parameters unchanged [21] and similarly to [22] [23], the synthesis algorithm implemented in this paper decomposes the speech frames into a lower-band voiced component $s_v(t)$ and an upper-band noise component $s_n(t)$ based on Maximum Voiced Frequency (MVF) values. We define these components here as

$$s(t) = s_v(t) + s_n(t) \tag{1}$$

In order to avoid discontinuities at the frames boundaries, Overlap-add (OLA) technique is used to reconstruct the speech signal from their corresponding parameters estimated from our analysis model in [21]. If the current frame is voiced, the harmonic part can be expressed as:

$$s_v^i(t) = \sum_{k=1}^{K^i} A_k^i(t) \cos\left(w_k^i t + \emptyset_k^i(t)\right) \tag{2}$$

$$w_k^i = 2\pi k\,(\text{contF0})^i \tag{3}$$

where $A_k(t)$ and $\emptyset_k(t)$ are the amplitude and phase at frame i (both are obtained in a similar manner as described in [23]), $t = 0, 1, \ldots, N$ and N is the frame length. K is the time-varying frequency components or harmonics that depends on the contF0 and MVF as:

$$K^i = \begin{cases} round\left(\dfrac{MVF^i}{contF0^i}\right) - 1, & voiced\ frames \\ 0, & unvoiced\ frames \end{cases} \tag{4}$$

The synthetic noise signal $n(t)$ is filtered by a high-pass filter $f_h(t)$ with a cutoff frequency equal to the local MVF, and then modulated by its time-domain envelope e(t) as we described it in our previous study [21]

$$s_n^i(t) = e^i(t)\left[f_h^i(t) \cdot n^i(t)\right] \tag{5}$$

If the current frame is unvoiced (MVF=0), the harmonic part is zero and the synthetic frame is usually equal to the produced noise. Thus, the synthesized speech signal is obtained by adding the harmonic and noise components.

B. Voice Conversion Based on DNN

In [24] [25], the neural network based VC reaches higher performance on conversion than the GMM-based solution. In this work, a feed-forward deep neural network (FF-DNN) is used to model the transformation between source and target speech features as shown in the middle part of Fig. 1. It consists of 6 feed-forward hidden layers, each consisting of 1024 units and performs a non-linear function of the previous layer's representation and a linear activation function at the output layer. We applied a hyperbolic tangent activation function whose outputs lie in the range (-1 to 1) which can yield lower error rates and faster convergence than a logistic sigmoid function. For the first 15 epochs, a fixed learning rate of 0.002 was chosen with a momentum of 0.3. More specifically, after 10 epochs, the momentum was increased to 0.9, and then the learning rate was halved regularly. Thus, input features are propagated forward through the FF-DNN with estimated parameters to produce the corresponding output parameters.

The framework of the proposed VC system is shown in Fig. 1. Feature processing, training and conversion-synthesis steps are performed. MVF, contF0, and MGC parameters are extracted from the source and target voices using the analysis function of the CSM. A learning process based FF-DNN is applied to construct the conversion phase. The purpose of the conversion function is to map the training features of the source speaker $X = \{x_i\}_{i=1}^{I}$ to the corresponding training features of the target speaker $Y = \{y_j\}_{j=1}^{J}$. Here, X and Y vector sequences are time-aligned frame by frame by the

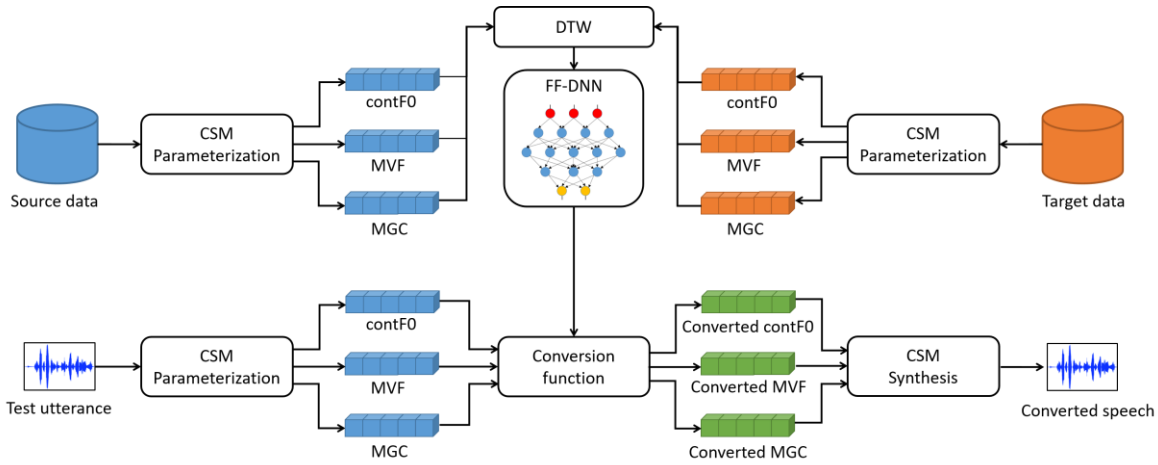


Fig. 1. Voice conversion process with CSM based waveform generation.

Dynamic Time Warping (DTW) algorithm [26] [27]. Then, the time-aligned acoustic feature sequences of both speakers are trained and used for the conversion function in order to predict the target features from the features of the source speaker. Finally, the converted conF0, MVF, and MGC are synthesized to get the converted speech waveform by the synthesis function of the CSM.

III. Experimental Framework

A. Datasets

We used a CMU-ARCTIC database [28] to evaluate the sound quality and speaker identity of the proposed VC framework. The parallel speech data of four speakers are chosen as our corpus, denoted BDL (American English, male), JMK (Canadian English, male), SLT (American English, female), and CLB (US English, female), each one consisting of 1132 sentences. The four speakers read out the same set of sentences. These sentences are divided into training, validation and testing set, each with 1000, 66, 66 sentences, respectively. As sample frequency 16kHz and 16-bit samples are used, and acoustic features were extracted with a 5ms frame shift. We conducted intra-gender and cross-gender pairs. Consequently, the number of combinations of the source and target speaker was 12 pairs. Note that we trained the conversion models for every speaker pair independently. The FF-DNN used in this work was implemented in the open source Merlin toolkit for speech synthesis [29] with some changes are introduced to be able to train our CSM. Besides, the training procedures were conducted on an NVidia Titan X GPU.

B. Baseline Systems

In this experiment, the proposed CSM based VC system was evaluated by comparing it with three systems:

- WORLD: It was found in [30] that the WORLD vocoder outperformed the state-of-the-art vocoders based speech synthesis (e.g., STRAIGHT). Therefore, we used WORLD vocoder based VC as a first baseline to measure our model performance.
- MagPhase: As our CSM followed the sinusoidal concept that contains both amplitude (intensity) and phase information, we chose a recently proposed MagPhase vocoder [31] based VC as a second baseline system.
- Sprocket: It is a vocoder-free VC system based on a differential GMM [32] submitted to the Voice Conversion Challenge 2018 (VCC2018). It will be used as a third baseline system in our study.

To fairly compare all systems mentioned above, we used the same nonlinear conversion function architecture (FF-DNN) as for the proposed system, except baseline 3 that is a linear function based on GMM. Thus, we ran 48 experiments in order to measure the performance of these VC-systems.

IV. Evaluation Results and Discussion

A. Objective Evaluation

Two objective speech quality measures are considered to evaluate the quality of the proposed model. Frequency-weighted segmental signal-to-noise ratio (fwSNRseg) [33] was firstly calculated, defined as

$$\mathrm{fwSNR}_{\mathrm{seg}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}}\left(\frac{\sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{W}_{\mathrm{i,j}} \cdot \log\frac{\mathrm{X}_{\mathrm{i,j}}^2}{\mathrm{X}_{\mathrm{i,j}}^2 - \mathrm{Y}_{\mathrm{i,j}}^2}}{\sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{W}_{\mathrm{i,j}}}\right) \tag{6}$$

where $\mathrm{X}_{\mathrm{i,j}}^2$, $\mathrm{Y}_{\mathrm{i,j}}^2$ are critical-band magnitude spectra in the j^{th} frequency band of the target and converted frame signals respectively, K is the number of bands, and W is a weight vector. Secondly, Log-Likelihood Ratio (LLR) [34] was employed to evaluate the distance between the converted and target speech from their linear prediction coefficients (LPC), which takes the form

$$\mathrm{LLR} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \log\left(\frac{\mathrm{a}_{\mathrm{y,i}}^{\mathrm{T}}\, \mathrm{R}_{\mathrm{x,i}}\, \mathrm{a}_{\mathrm{y,i}}}{\mathrm{a}_{\mathrm{x,i}}^{\mathrm{T}}\, \mathrm{R}_{\mathrm{x,i}}\, \mathrm{a}_{\mathrm{x,i}}}\right) \tag{7}$$

where a_{x}, a_{y}, and R_{x} are the LPC vector of the target signal frame, converted signal frame, and the autocorrelation matrix of the target speech signal, respectively.

TABLE I. AVERAGE SCORES ON CONVERTED SPEECH SIGNAL PER EACH OF THE SPEAKER PAIRS CONVERSION

Model	WORLD		MagPhase		Sprocket		Proposed	
	fwSNRseg	LLR	fwSNRseg	LLR	fwSNRseg	LLR	fwSNRseg	LLR
BDL → JMK	2.19	1.57	**3.21**	**1.37**	2.20	1.48	2.47	1.50
BDL → SLT	1.12	1.72	1.25	1.69	1.04	**1.49**	**2.33**	1.57
BDL → CLB	0.79	1.83	1.65	1.72	0.37	**1.69**	**1.66**	1.74
JMK → BDL	1.31	1.76	**2.49**	**1.56**	1.73	1.63	2.15	1.57
JMK → SLT	0.55	1.74	**1.93**	**1.56**	0.11	1.64	1.54	1.65
JMK → CLB	1.45	1.74	1.75	1.66	0.69	**1.60**	**1.81**	1.67
SLT → BDL	1.65	1.71	1.60	1.70	1.80	1.51	**2.95**	**1.49**
SLT → JMK	2.16	1.61	**2.71**	1.42	0.713	1.56	2.59	**1.39**
SLT → CLB	1.51	1.75	**2.89**	1.59	2.32	1.56	2.51	**1.50**
CLB → BDL	0.97	1.81	1.60	1.70	0.95	1.72	**1.92**	**1.60**
CLB → JMK	2.50	1.49	2.74	1.40	0.98	1.46	**3.00**	**1.30**
CLB → SLT	0.98	1.70	**2.17**	1.53	1.96	1.54	2.12	**1.47**

A more detailed case-by-case analysis by fwSNRseg and LLR are shown in Table 1. The results were averaged over 20 synthesized test utterances for each pair. A calculation is done frame-by-frame, and the best value in each column of Table 1 is bold faced.

First, it could be observed that our proposed method gives significantly better LLR scores than other systems in female-to-male voice conversion. In other words, the CSM can convert voice characteristics more accurately than other methods when a female is a source speaker. Similar observations can be found in male-to-female voice conversions (in particular, BDL-to-SLT, BDL-to-CLB, and JMK-to-CLB), where the fwSNRseg measure tended to have the highest scores in our proposed model. In a sense, there is a tendency to an increased fwSNRseg when considering continuous F0 in the proposed method. Second, for the same-gender speaker pairs, the LLR values in Table 1 indicate that the proposed system obviously outperforms the baseline systems in female-to-female conversions. On the other hand, in terms of male-to-male voice conversions, our proposed system achieves the second highest sound quality. Overall, these findings demonstrate that the CSM can yield a good performance comparable to other systems.

The comparison of the spectral envelope of one speech frame converted by the proposed method is given in Fig. 2a. It may be observed that the converted spectral envelope is more similar in general to the target one than the source one. It can also be seen in Fig. 2b that the converted contF0 trajectories generated from the proposed method follow the same shape of the target confirming the similarity between them and can provide better F0 predictions. Similarly, when looking at Fig. 2c, it makes apparent that the proposed framework produces converted speech with MVF more similar to the target trajectories rather than to the source ones.

As a result, these experiments show that the proposed model with continuous sinusoidal vocoder is competitive for the VC task and superior to the reference WORLD model.

B. Subjective Evaluation

A perceptual listening test was designed to test and evaluate the quality of our proposed model. First, we performed a web-based MUSHRA-like (MUlti-Stimulus test with Hidden Reference and Anchor) listening test [35] to evaluate the speaker identity/similarity of the converted speech to a natural-reference target voice. The listeners had to rate the naturalness and similarity of each stimulus, from 0 to 100. Twelve utterances were randomly chosen and presented in a randomized order. Altogether, 72 utterances were included in the MUSHRA test (6 types x 12 sentences). Twenty listeners (11 males and 9 females) participated in the experiment. On average, the MUSHRA test took 10 minutes to fill. The listening tests samples can be found online[1].

The MUSHRA similarity scores of the listening test are presented in Fig. 3. An interesting note is that the listeners preferred our system compared to others developed earlier. According to Mann-Whitney-Wilcoxon ranksum tests (with a 95% confidence level), all differences are statistically significant. This means that our proposed model has successfully converted the source voice to the target voice on the same-gender and cross-gender cases. Moreover, Fig. 3 shows that the WORLD and Sprocket systems get higher scores in the MUSHRA test for only the JMK-to-SLT, JMK-to-BDL, and CLB-to-SLT speaker conversions, respectively.

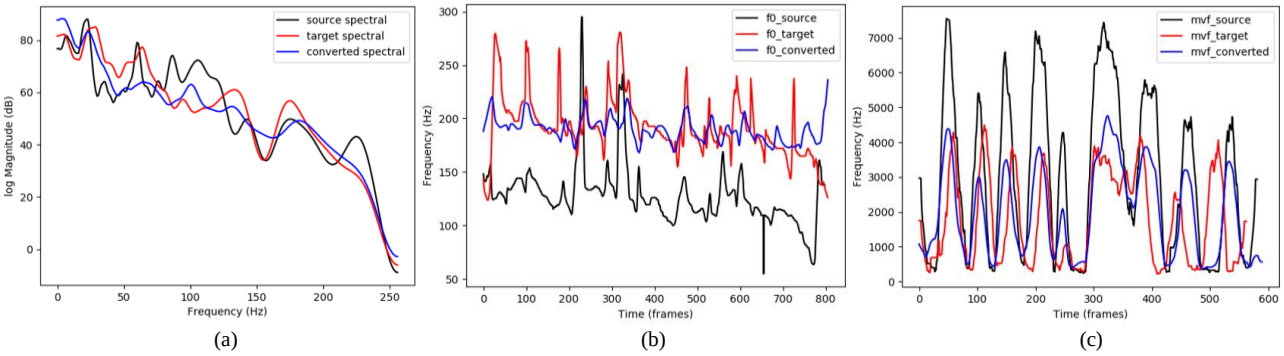
Overall, these results suggest that the best conversion technique for the source-filter based vocoder is the CSM, while the WORLD is also a good option, having the second highest similarity scores.

V. DISCUSSION AND CONCLUSIONS

This work proposed a CSM-based voice conversion framework, and the continuous F0 is our main interest to avoid alignment errors that may happen at voiced-unvoiced boundaries. A number of recently developed VC methods have been applied and compared with the proposed model. The performance of the methods was statistically analyzed with two error metrics and subjectively evaluated by the use of expert opinion. The results discussed in Section IV show the effectiveness of the proposed method in terms of naturalness and speaker similarity. The advantage of the CSM is that it gives the closest results to the target speaker in both objective and similarity tests compared to other approaches.

Future works will aim at improving the quality scores through the use of bidirectional recurrent neural networks, in which many-to-one and one-to-many voice conversion can be achieved.

[1] http://smartlab.tmit.bme.hu/interspeech2019_vc



(a) (b) (c)

Fig. 2. Example of the natural source (black), target (red), and converted (blue) spectral envelope, contF0, and MVF trajectories using the proposed method.

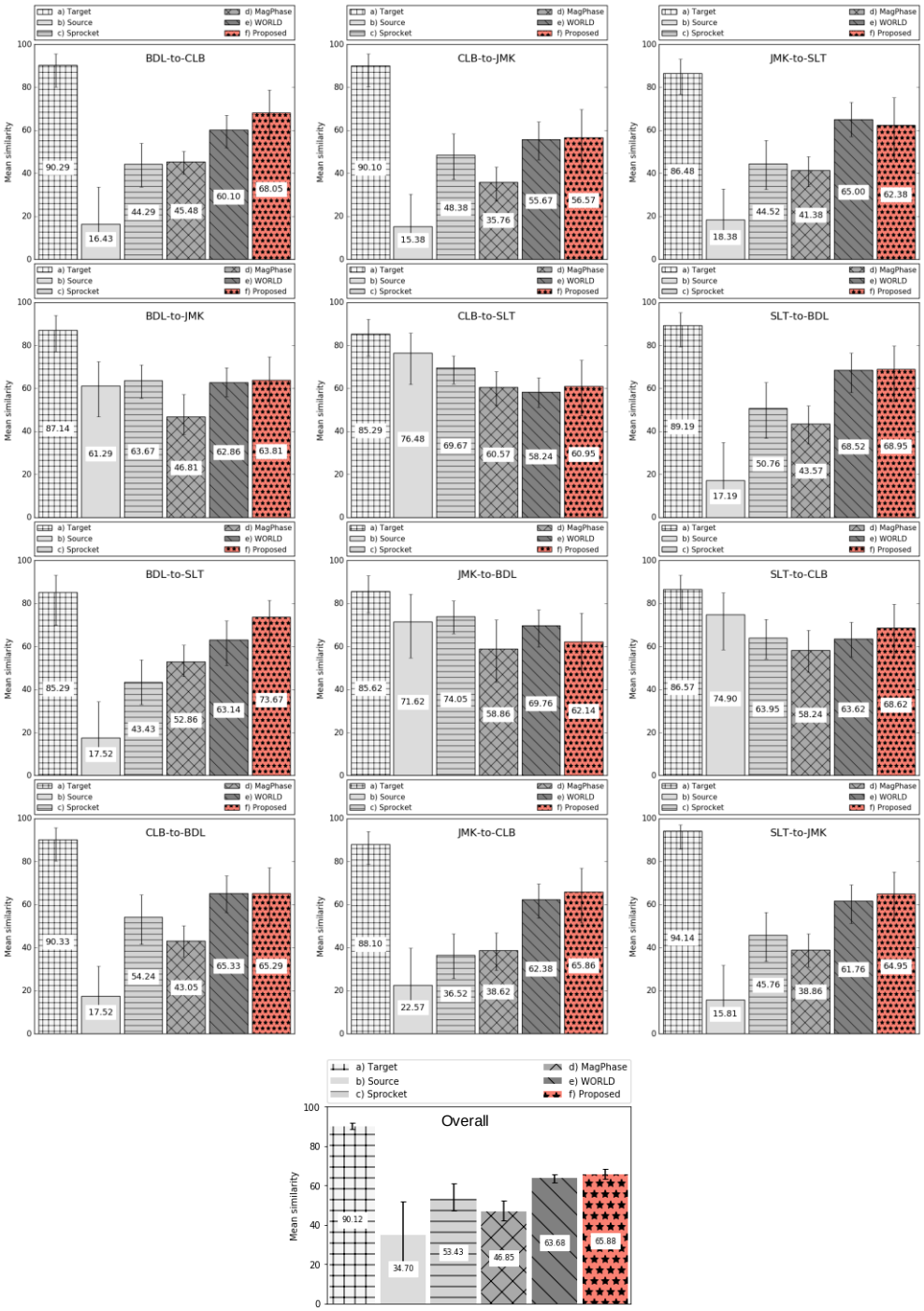


Fig. 3. MUSHRA scores for the similarity question. Higher value means better overall quality. Errorbars show the bootstrapped 95% confidence intervals.

ACKNOWLEDGMENT

The research was partly supported by the AI4EU project and by the National Research, Development and Innovation Office of Hungary (FK 124584). The Titan X GPU used was donated by NVIDIA Corporation. We would like to thank the subjects for participating in the listening test.

REFERENCES

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in Proceedings of the IEEE Acoustics, Speech and Signal Processing, pp. 285-288, 1998.

[2] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," Speech Communication, vol. 54, no. 1, pp. 134-146, 2012.

[3] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 2505-2517, 2012.

[4] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Transactions Speech and Audio Processing, vol. 6, no. 2, pp. 131-142, 1998.

[5] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion using sparse representation in noisy environments," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 96, no. 10, pp. 1946–1953, 2013.

[6] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on speaker-dependent restricted boltzmann machines," IEICE Transactions on Information and Systems, vol. 97, no. 6, pp. 1403-1410, 2014.

[7] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5274-5278, 2018.

[8] 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.

[9] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in Proceedings of the INTERSPEECH, pp. 369-372, 2013.

[10] T. Kaneko, H. Kameoka, and K. Hiramatsu, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in Proceedings of the INTERSPEECH, pp. 1283-1287, 2017.

[11] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4869-4873, 2015.

[12] D. Erro, A. Moreno, and A. Bonafonte, "INCA Algorithm for Training Voice Conversion Systems From Nonparallel Corpora," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 944-953, 2010.

[13] A. Mouchtaris, J. Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 3, pp. 952-963, 2006.

[14] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in Proceedings of the 8th International Speech Communication Association (ISCA) Speech Synthesis Workshop, pp. 155-160, 2013.

[15] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 841-844, 2001.

[16] M. Lenarczyk, "Parametric Speech Coding Framework for Voice Conversion Based on Mixed Excitation Model," in Proceedings of the Text, Speech, and Dialogue, pp. 507-514, 2014.

[17] W. Lifang and Z. Linghua, "A Voice Conversion System Based on the Harmonic Plus Noise Excitation and Gaussian Mixture Model," in Proceedings of the Instrumentation, Measurement, Computer, Communication and Control, pp. 1575-1578, 2012.

[18] B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "Adaptive WaveNet Vocoder for Residual Compensation in GAN-Based Voice Conversion," IEEE Spoken Language Technology Workshop (SLT), 2018.

[19] Y. Wang, et al., "Tacotron: towards end-to-end speech synthesis," in Proceedings of the INTERSPEECH, pp. 4006-4010, 2017.

[20] I. Saratxagaa, J. Sanchez, Z. Wu, and I. Hernaeza, "Synthetic Speech Detection Using Phase Information," Speech Communication, vol. 81, pp. 30-41, 2016.

[21] M.S. Al-Radhi, T.G. Csapó, and G. Németh, "Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis," in Proceedings of the INTERSPEECH, pp. 434-438, 2017.

[22] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in Proceedings of the Eurospeech, pp. 451-454, 1995.

[23] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis," IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 2, pp. 184-194, 2014.

[24] S. Desai, E.V. Raghavendra, B. Yegnanarayana, A.W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3893-3896, 2009.

[25] G. Kotani, D. Saito, and N. Minematsu, "Voice conversion based on deep neural networks for time-variant linear transformations," in Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1259-1262, 2017.

[26] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 655-658, 1988.

[27] A.W. Black, J. Kominek, and C. Bennett, "Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis," in Eurospeech, pp. 313-316, 2003.

[28] J. Kominek and A.W. Black, "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, 2003.

[29] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," in Proceedings of the 9th International Speech Communication Association (ISCA) Speech Synthesis Workshop (SSW9), 2016.

[30] M. Morise and Y. Watanabe, "Sound quality comparison among high-quality vocoders by using re-synthesized speech," Acoustical Science and Technology, vol. 39, no. 3, pp. 263-265, 2018.

[31] F. Espic, C. Valentini-Botinhao, and S. King, "Direct Modelling of Magnitude and Phase Spectra for Statistical Parametric Speech Synthesis," in Proceedings of the INTERSPEECH, 2017.

[32] K. Kobayashi and T. Toda, "sprocket: Open-Source Voice Conversion Software," in Proceedings of the Odyssey: The Speaker and Language Recognition, pp. 203-210, 2018.

[33] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," Acoustical Society of America, vol. 125, no. 5, pp. 3387-3405, 2009.

[34] S. Quackenbush, T. Barnwell, and M. Clements, "Objective Measures of Speech Quality," Englewood Cliffs, NJ: Prentice-Hall, 1988.

[35] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2003.