

Multi-Task WaveRNN With an Integrated Architecture for Cross-Lingual Voice Conversion

Yi Zhou , *Student Member, IEEE*, Xiaohai Tian , *Member, IEEE*, and Haizhou Li , *Fellow, IEEE*

Abstract—Spoken languages are similar phonetically because humans have a common vocal production system. However, each language has a unique phonetic repertoire and phonotactic rule. In cross-lingual voice conversion, source speaker and target speaker speak different languages. The challenge is how to project the speaker identity of the source speaker to that of the target across two different phonetic systems. A typical voice conversion system employs a generator-vocoder pipeline, where the generator is responsible for conversion, and the vocoder is for waveform reconstruction. We propose a novel Multi-Task WaveRNN with an integrated architecture for cross-lingual voice conversion. The WaveRNN is trained on two sets of monolingual data via a two-task learning. The integrated architecture takes linguistic features as input and outputs speech waveform directly. Voice conversion experiments are conducted between English and Mandarin, which confirm the effectiveness of the proposed method in terms of speech quality and speaker similarity.

Index Terms—Cross-lingual voice conversion (xVC), multi-task learning (MTL), WaveRNN, integrated architecture.

I. INTRODUCTION

VOICE conversion (VC) is a technique to convert one's voice (source) to sound like that of another (target) [1], where we change the speaker identity while carrying over the linguistic content. If the source and target speakers speak in the same language, we can construct a source-target parallel database of the same content for model training. However, in cross-lingual voice conversion (xVC), the source and target speakers speak different languages, it is difficult to obtain parallel data of the same content [2]–[4]. Therefore, we are motivated

to study model training with non-parallel data. As we only change the speaker identity and keep the linguistic content, xVC is different from speech translation [5].

A typical VC system consists of two modules, a generator for spectral feature mapping and a vocoder for speech waveform generation. For the generator, early studies on statistical modeling approaches provide great insights into xVC. Some choose to adapt a pre-trained multi-speaker average model towards the target speaker [4], [6]. Others attempt to find the equivalent of speech segments between source and target, e.g., unit selection [7]–[10], iterative frame alignment [11]–[14] and vocal tract length normalization [14]–[16].

Deep learning has opened up new possibilities for xVC. Cycle-consistent generative adversarial networks (CycleGANs) [17] is studied to deal with non-parallel training data problem. However, it generally works for a specific source-target pair and requires a large amount of training data. Another idea is to disentangle the speaker-dependent component (speaker identity) from the speaker-independent component (speech content). In this way, we can convert voice from one another by just changing the speaker-dependent component while keeping the speaker-independent component. Variational autoencoder (VAE) [18] is one of such implementations, which learns a latent space for speaker-independent representation. Phonetic PosteriorGram (PPG) [19], an intermediate result of speech recognition, represents the underlying phonetic content of an utterance [20]. With PPG, we can train a xVC model for multiple speakers [21], and condition on speaker i-vector [22] to control the output speaker identity. For example, one can combine the PPGs of utterance $\mathcal{A}$ with the speaker i-vector of $\mathcal{B}$ to compose a new utterance that sounds like speaker $\mathcal{B}$ speaking what speaker $\mathcal{A}$ speaks [21], [23].

The findings in previous work prompt us to look into the research problem in two aspects. First, a xVC system involves two languages of diverse characteristics [24]. It might not be the best to use one shared neural network for conversions between two languages [25]. Second, many existing xVC implementations employ a generator-vocoder pipeline, where the vocoder is trained on acoustic features of natural speech, but tested on those of converted speech. Therefore, there exists an inherent mismatch [26]–[28].

We propose a Multi-Task WaveRNN with an integrated architecture for xVC for the first time, each task representing one language. By doing so, we do not require a separate vocoder, thus alleviating feature mismatch problem arising from the generator-vocoder pipeline. This paper is organized as follows. Section II presents the related work as a motivation of our proposal. In Section III, we formulate the proposed Multi-Task WaveRNN. In Section IV, we report the experiments. Section V concludes this study.

Manuscript received June 7, 2020; revised July 11, 2020; accepted July 14, 2020. Date of publication July 17, 2020; date of current version August 6, 2020. This work was supported in part by the Human-Robot Interaction Phase 1, National Robotics Programme, National Research Foundation (NRF) Singapore, under Grant 192 25 00054; in part by the AI Speech Lab, AI Singapore Programme, NRF Singapore, under Award AISG-100E-2018-006; and in part by the Human Robot Collaborative AI for AME, NRF Singapore, under Grant A18A2b0046. The work of Y. Zhou was supported by the NUS Research Scholarship. The work of H. Li was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany) The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Odette Scharenborg. (*Corresponding author: Yi Zhou.*)

Yi Zhou and Xiaohai Tian are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: yi.zhou@u.nus.edu; eletia@nus.edu.sg).

Haizhou Li is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, and also with the Faculty of Computer Science and Mathematics, University of Bremen, 28359 Bremen, Germany (e-mail: haizhou.li@nus.edu.sg).

Digital Object Identifier 10.1109/LSP.2020.3010163

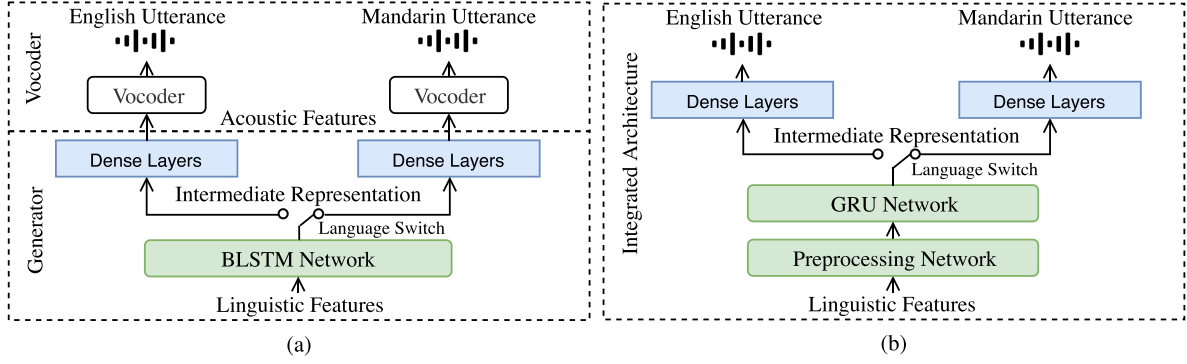


Fig. 1. (a) A generator-vocoder xVC pipeline [25] with a multi-task trained, modularized NN as the generator. The generator takes linguistic features as input and generates acoustic features as output. The vocoder synthesizes the speech waveform from the acoustic features. (b) The proposed Multi-Task WaveRNN with an integrated architecture for xVC, that generates waveform directly without a separate vocoder. Both frameworks employ a *hidden section* shared by two languages and an *output section* that consists of two heads, each head is only trained on speech of native speakers, but used to infer speech for foreign speakers.

II. RELATED WORK

A. Multi-Task Generator-Vocoder Pipeline

Multi-task learning (MTL) is a form of inductive transfer. Training the model with multiple tasks causes the model to prefer hypotheses that explain more than one task [29], [30]. The MTL technique has been studied in applications such as natural language processing [31], speech synthesis [32], [33] and speech recognition [34], [35]. A modularized neural network (NN) with MTL [25] was recently studied for xVC between English and Mandarin. The modularized NN has several bidirectional long short-term memory (BLSTM) layers in the *hidden section*, that are shared between two languages, while the two heads with linear layers in the *output section* are language-dependent, each for one language [25]. The shared *hidden section* serves as the bridge between the two languages, while the *output section* takes care of the language-specific acoustic renderings.

As multi-task learning produces higher quality xVC voice than single-task learning [25], that becomes a source of inspiration for this work. Fig. 1(a) illustrates a generator-vocoder pipeline, where the modularized NN serves as generator. The generator takes linguistic features as input, namely, PPG [36], i-vector [22] and auxiliary features, i.e., voiced/unvoiced flag (vuv), fundamental frequency (F_0), aperiodicity (AP), language identity (L_{id}) from speech samples, and generates acoustic features as output.

During training, the parameters in the *hidden section* are updated for both languages, while those in the *output section* are only updated with its corresponding output language, respectively. At run-time, we can convert an English utterance from an English speaker to a Mandarin speaker, or a Mandarin utterance from a Mandarin speaker to an English speaker. Taking the former as an example, we would like the converted voice to sound like a native English speaker with voice characteristics similar to the Mandarin speaker.

B. Neural Vocoder

The generator-vocoder pipeline showed in Fig. 1(a) requires a vocoder to synthesize speech waveform from acoustic features. One way to improve the vocoding quality is to replace the parametric vocoder with a trainable neural vocoder [27], [37] like WaveNet [38] and WaveRNN [39] since they consistently outperform traditional parametric vocoders.

On the other hand, as neural vocoders are trained in a data-driven manner, it can take different feature representations as input, for example, text token [38], PPG [36] and i-vector [22], for speech waveform generation. Considering the feature mismatch between training and run-time inference in the generator-vocoder pipeline [28], one can expect that a generator-vocoder integrated architecture is a better solution than the generator-vocoder pipeline. There have been some studies to re-purpose a neural vocoder for monolingual voice conversion, that takes linguistic features as input and generates speech waveform directly [28], [40]. We would like to further investigate this idea for xVC by taking WaveRNN [39] as the case study in this work.

III. MULTI-TASK WaverNN WITH AN INTEGRATED ARCHITECTURE

WaveRNN is a high quality speech waveform generator [39], that generally consists of two sections: a *hidden section* that maps the input features to intermediate representations, and an *output section* that relates the intermediate representations to the target waveform. The *hidden section* consists of a preprocessing network and a GRU network [41]. The preprocessing network upsamples input features into the desired resolution. The GRU network is made of one dense layer, two gated recurrent unit (GRU) layers with skip connections, followed by one dense layer. The *output section* consists of two dense layers. Cross-entropy loss is adopted between the predicted waveform and the natural target speech as objective function. During training, the gradients are back-propagated to update both *output section* and *hidden section*.

It is possible to train a single WaveRNN to perform xVC for two languages, i.e., to convert an English utterance from an English speaker to a Mandarin speaker, and to convert a Mandarin utterance from a Mandarin speaker to an English speaker. We call such a network Single-Task WaveRNN, where both *hidden* and *output sections* are shared by two languages.

We propose a Multi-Task WaveRNN as in Fig. 1(b), that consists of a shared *hidden section* trained on speech data of two languages via a two-task learning, and a *output section* of two heads, each for one language. It also features an integrated architecture that doesn't require a dedicated vocoder to avoid any mismatch between training and run-time inference. We will study the contributions of the two ideas, namely, multi-task learning and integrated architecture, in the experiments.

TABLE I
A SUMMARY OF CONFIGURATIONS FOR FOUR EXPERIMENTAL
SYSTEMS IN THE COMPARISON

System	(a) P-WORLD	(b) P-WaveRNN	(c) STWaveRNN	(d) MTWaveRNN
Generator	[BLSTM, BLSTM] [2048, 2048] [Dense, Dense] [512, 127]		Upsampling factor [4, 4, 5] [Dense, GRU, GRU, Dense] [512, 512, 512, 512]	
Vocoder	WORLD	Same WaveRNN as in (c) and (d)	[Dense, Dense] [512, 512]	

At run-time, converting an utterance from speaker $\mathcal{A}$ (source) to speaker $\mathcal{B}$ (target), we first obtain PPG, vuv, AP , L_{id} from speaker $\mathcal{A}$, that represent the linguistic content of the source, and i-vector from target speaker $\mathcal{B}$. They form the input features to the *hidden section*. The target speech waveform is then generated through the output head corresponding to the language of the input speech.

IV. EXPERIMENTS

The proposed Multi-Task WaveRNN and three reference systems are implemented for comparison. We use P to stand for generator-vocoder pipeline systems; MT for multi-task learning, and ST for single-task learning system. The details of system configurations are at Table I.

- (a) P-WORLD: The generator-vocoder pipeline as illustrated in Fig. 1(a), with a modularized NN under multi-task learning [25] as the generator, followed by WORLD vocoder.
- (b) P-WaveRNN: The generator-vocoder pipeline similar to system (a) except that we replace WORLD vocoder with WaveRNN vocoder.
- (c) STWaveRNN: The Single-Task WaveRNN, an integrated architecture that takes linguistic features as input and generates speech waveform as output.
- (d) MTWaveRNN: The Multi-Task WaveRNN, an integrated architecture, as illustrated in Fig. 1(b), with two language-dependent output heads in the *output section*.

BLSTM is used in the generator for the pipeline systems, while GRU is used in the WaveRNN. They are both effective recurrent neural networks for generating sequences [25], [39], [42]. As BLSTM is reported to show similar performance as GRU [43], we consider that it is reasonable to compare the proposed MTWaveRNN (GRU) with the baseline P-WaveRNN (BLSTM + GRU) to observe the effect of integrated versus pipeline architecture.

A. Database and Feature Extraction

We conducted xVC experiments between English and Mandarin speakers (en→cn and cn→en). For example, en→cn refers to the experiments converting an English utterance from an English speaker towards a Mandarin speaker.

For English, we selected 96 speakers from five database: VCTK [44], CMU ARCTIC [45], VCC2016 [46] and VCC2018 [47]. As for Mandarin, 50 speakers in the Average Mandarin Library from DataBaker were used. In total, we had 59,900 utterances for training. During testing, four bilingual speakers (MF4, MF5, MM1, MM2) from the EMIME [48] database were selected, each speaking 20 utterances.

In P-WORLD and P-WaveRNN, 40 dimensional Mel Cepstral Coefficients (MCCs) are computed by the Speech Signal Processing Toolkit¹. Together with their dynamic features, a

TABLE II
A SUMMARY OF MCD AND RMSE FOR EN→CN AND CN→EN EXPERIMENTS,
AND THEIR AVERAGE, DENOTED AS 'AVG'

	MCD (dB)			RMSE (dB)		
	en→cn	cn→en	avg	en→cn	cn→en	avg
Source	8.61	8.65	8.63	17.48	19.96	18.72
(a) P-WORLD	6.21	7.01	6.61	13.71	13.59	13.65
(b) P-WaveRNN	7.58	8.04	7.81	15.26	15.92	15.59
(c) STWaveRNN	7.48	7.90	7.69	14.63	14.35	14.49
(d) MTWaveRNN	6.89	7.39	7.14	13.51	14.28	13.90

frame of acoustic features {vuv, MCC, F_0 , AP } is a 127-dimension vector. As for STWaveRNN and MTWaveRNN with the integrated architecture, such intermediate acoustic features are not required.

All systems take the same input linguistic features, where one frame has 573 dimensions, that includes {PPG, i-vector, vuv, F_0 , AP , L_{id} } with their respective dimensions of [414, 150, 1, 3, 3, 2]. The speech waveform is re-sampled at 16 kHz, and encoded by 9 bits μ -law for systems involving WaveRNN: P-WaveRNN, STWaveRNN and MTWaveRNN. All acoustic features are extracted by WORLD vocoder [49]. PPGs are extracted from an English-Mandarin acoustic model [25] using Kaldi toolkit [50].

B. Results and Discussions

1) *Objective Evaluation*: We report the results in terms of mel-cepstrum distortion (MCD) and root mean square error (RMSE) in Table II.

We observe that MTWaveRNN outperforms P-WaveRNN by a wide margin with the average MCD and RMSE values reduced from 7.81 dB to 7.14 dB and from 15.59 dB to 13.90 dB, respectively. As both follow multi-task learning, the results reflect the effect of the integrated architecture. Also, we find that MTWaveRNN achieves lower average MCD and RMSE than STWaveRNN consistently (7.14 dB vs. 7.69 dB and 13.90 dB vs. 14.49 dB), that confirms the advantage of multi-task learning over single-task learning.

We would like to further examine the contributions of the two proposed ideas, namely, multi-task learning and integrated architecture. We note that MTWaveRNN outperforms P-WaveRNN with MCD and RMSE reduction of 0.67 dB and 1.69 dB, while MTWaveRNN outperforms STWaveRNN with MCD and RMSE reduction of 0.55 dB and 0.59 dB. While both multi-task learning and integrated architecture contribute to the superior performance of MTWaveRNN, the effect of the integrated architecture is more prominent.

We also note that P-WORLD with WORLD vocoder achieves the lowest MCD (6.61 dB) and RMSE (13.65 dB). The traditional parametric vocoders, such as WORLD, typically report lower values in terms of objective metrics [26], [28]. However, such lower values don't translate directly to a better perceptual quality in subjective evaluations. This is consistent with what we observe in this paper.

2) *Subjective Evaluation*: We further conduct four subjective evaluations: AB and MUSHRA [51] tests to evaluate the speech quality; XAB and VCC-like [46], [52] tests to evaluate the speaker similarity. 25 bilingual English and Mandarin listeners were invited to participate in all listening tests. Each listener was presented with 20 utterances randomly selected from 160 converted utterances. All listeners are university students and staff members, who have English as their official language of

¹<https://sourceforge.net/projects/sp-tk/>

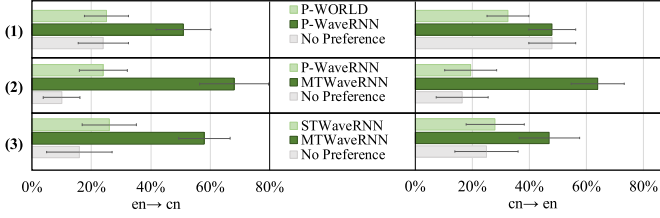


Fig. 2. Three groups of AB preference tests for speech quality with 95% confidence intervals.

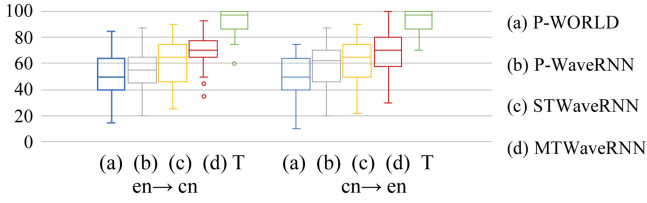


Fig. 3. MUSHRA test scores for speech quality among all systems. Converting an utterance from speaker $\mathcal{A}$ to $\mathcal{B}$, ‘T’ denotes the natural speech from speaker $\mathcal{B}$. The quality of each utterance is rated between 0 and 100, with a higher score indicating a better quality.

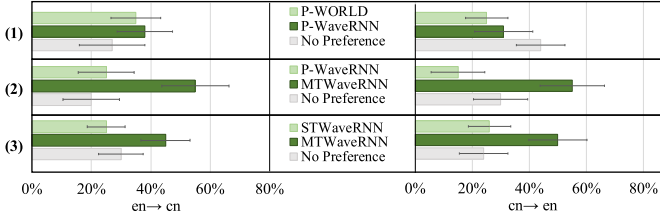


Fig. 4. Three groups of XAB preference tests for speaker similarity with 95% confidence intervals.

teaching, while having Mandarin as their mother tongue. The converted samples can be found at the [demo link](https://vcsamples.github.io/spl2020/)².

- (i) AB Quality Preference Test: A and B are the selected speech samples of the same content from two different systems. Listeners choose their preference in terms of speech quality. Fig. 2 reports the percentage of listeners’ preference.
- (ii) MUSHRA Quality Test: In each query, the reference natural speech and four converted samples of the same content are presented to the listeners in a random order. The listeners are asked to rate the speech quality of each sample between 0 and 100. The results are reported in Fig. 3.
- (iii) XAB Similarity Preference Test: X is the natural target speech. A and B are the selected samples of the same content from two different systems. With reference to X, the listeners need to choose which sample sounds closer to the reference speech in terms of speaker similarity. Fig. 4 demonstrates the percentage of all listeners’ preference.
- (iv) VCC-like Similarity Test: In each query, six speech samples are presented to the listeners including four converted samples from the four experimental systems, source speech, and native speech in a different language

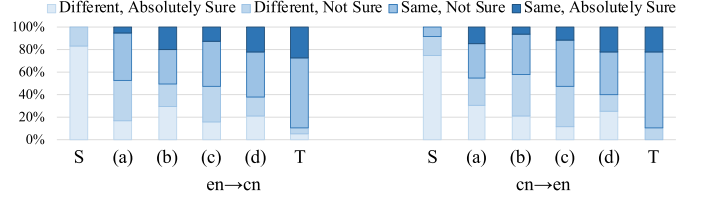


Fig. 5. VCC similarity test results for all systems: (a) P-WORLD, (b) P-WaveRNN, (c) STWaveRNN and (d) MTWaveRNN. Converting an utterance from speaker $\mathcal{A}$ to $\mathcal{B}$, ‘S’ denotes the natural speech from speaker $\mathcal{A}$ in one language, and ‘T’ denotes the natural speech from speaker $\mathcal{B}$ in another language. A higher percentage of ‘Same, Not Sure’ and ‘Same, Absolutely Sure’ together suggests a higher similarity to ‘T’.

from the target speaker. With reference to the target speech, the listeners listen to each sample and decide whether it is uttered by the target speaker. Fig. 5 presents the distributions of their decisions.

We conduct three groups of experiments, where two systems are compared each time in AB and XAB preference tests.

- Two reference pipeline systems, P-WORLD vs. P-WaveRNN: While P-WaveRNN outperforms P-WORLD more for speech quality (Fig. 2(1)), and less for speaker similarity (Fig. 4(1)), it suggests that WaveRNN compares favorably with the traditional parametric vocoder WORLD.
- The proposed MTWaveRNN vs. the reference system P-WaveRNN: We notice that MTWaveRNN significantly outperforms P-WaveRNN for both speech quality (Fig. 2(2)) and speaker similarity (Fig. 4(2)). The results suggest that the integrated architecture has a clear advantage over the generator-vocoder pipeline.
- The proposed MTWaveRNN vs. the reference system STWaveRNN: It is found that MTWaveRNN outperforms STWaveRNN in Fig. 2(3) and 4(3), which confirms that multi-task learning is superior to single-task learning.

We observe a consistent performance trend for all systems in terms of speech quality (Fig. 3) and speaker similarity (Fig. 5), ranked in ascending order (from the worst to the best): P-WORLD, P-WaveRNN, STWaveRNN, MTWaveRNN. The results further confirm the effectiveness of MTWaveRNN. While both the integrated architecture and multi-task learning contribute to the performance gain, we observe that the advantage of integrated architecture is more prominent, that is consistent with the objective evaluation results in Table II.

V. CONCLUSION

We study a Multi-Task WaveRNN for cross-lingual voice conversion, that benefits from the ideas of multi-task learning and integrated architecture. The proposed WaveRNN framework can be extended to other speech generation tasks, such as multi-speaker, multi-accent, or multi-lingual speech synthesis.

REFERENCES

- [1] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.
- [2] M. Abe, K. Shikano, and H. Kuwabara, “Cross-language voice conversion,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 345–348.

²<https://vcsamples.github.io/spl2020/>

- [3] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," *Inf. Process. Soc. Jap. J.*, vol. 43, no. 7, pp. 2177–2185, 2002.
- [4] B. Ramani, M. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "A multi-level GMM-based cross-lingual voice conversion using language-specific mixture weights for polyglot synthesis," *Circuits, Syst., Signal Process.*, vol. 35, no. 4, pp. 1283–1311, 2016.
- [5] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 517–520.
- [6] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," in *Proc. Interspeech*, 2009, pp. 1635–1638.
- [7] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 373–376.
- [8] H. Duxans, D. Erro, J. Pérez, F. Diego, A. Bonafonte, and A. Moreno, "Voice conversion of non-aligned data using unit selection," in *Proc. TC-STAR Workshop Speech-to-Speech Transl.*, 2006, pp. 237–242.
- [9] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, "Text-independent cross-language voice conversion," in *Proc. Interspeech*, 2009, pp. 2262–2265.
- [10] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," presented at the IEEE Int. Conf. Acoust. Speech Signal Process. Proc., Toulouse, France, 14–19 May 2006.
- [11] D. Erro and A. Moreno, "Frame alignment method for cross-lingual voice conversion," in *Proc. Interspeech*, 2007, pp. 1969–1972.
- [12] A. J. Uriz, P. D. Agüero, A. Bonafonte, and J. C. Tulli, "Voice conversion using k-histograms and frame selection," in *Proc. Interspeech*, 2009, pp. 1639–1642.
- [13] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 944–953, Jul. 2010.
- [14] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 5120–5123.
- [15] D. Sündermann, H. Ney, and H. Höge, "VTNLN-based cross-language voice conversion," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2003, pp. 676–681.
- [16] S. Desai, B. Yegnanarayana, and K. Prahallad, "A framework for cross-lingual voice conversion using artificial neural networks," in *Proc. 7th Int. Conf. Natural Lang. Process.*, 2009, pp. 59–66.
- [17] B. Sisman, M. Zhang, M. Dong, and H. Li, "On the study of generative adversarial networks for cross-lingual voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 144–151.
- [18] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.
- [19] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2009, pp. 421–426.
- [20] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2016, pp. 1–6.
- [21] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6790–6794.
- [22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
- [23] J. Wu, Z. Wu, and L. Xie, "On the use of i-vectors and average voice model for voice conversion without parallel data," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.
- [24] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proc. IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013.
- [25] Y. Zhou, X. Tian, E. Yilmaz, R. K. Das, and H. Li, "A modularized neural network with language-specific output layers for cross-lingual voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 160–167.
- [26] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.
- [27] Y.-C. Wu, K. Kobayashi, T. Hayashi, P. L. Tobing, and T. Toda, "Collapsed speech segment detection and suppression for WaveNet vocoder," in *Proc. Interspeech*, 2018, pp. 1988–1992.
- [28] X. Tian, E. S. Chng, and H. Li, "A speaker-dependent WaveNet for voice conversion with non-parallel data," in *Proc. Interspeech*, 2019, pp. 201–205.
- [29] R. Caruna, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
- [30] Y. Zhang and Q. Yang, "A survey on multi-task learning," 2017, [arXiv:1707.08114](https://arxiv.org/abs/1707.08114).
- [31] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.
- [32] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4460–4464.
- [33] Q. Hu, Z. Wu, K. Richmond, J. Yamagishi, Y. Stylianou, and R. Maia, "Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning," in *Proc. Interspeech*, 2015, pp. 854–858.
- [34] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6965–6969.
- [35] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, [arXiv:1706.05098](https://arxiv.org/abs/1706.05098).
- [36] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4869–4873.
- [37] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. Interspeech*, 2017, pp. 1138–1142.
- [38] A. V. D. Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, [arXiv:1609.03499](https://arxiv.org/abs/1609.03499).
- [39] N. Kalchbrenner *et al.*, "Efficient neural audio synthesis," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2415–2424.
- [40] H. Lu, Z. Wu, R. Li, S. Kang, J. Jia, and H. Meng, "A compact framework for voice conversion using WaveNet conditioned on phonetic posteriorgrams," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6810–6814.
- [41] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A comparison of recent neural vocoders for speech signal reconstruction," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 7–12.
- [42] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual tts using phonetic posteriorgrams," in *Proc. Interspeech*, 2016, pp. 322–326.
- [43] M. V. Ramos, "Voice conversion with deep learning," M.S. thesis, Dept. Electr. Comput. Eng., Técnico Lisboa, Lisbon, Portugal, 2016.
- [44] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version0.92)," Univ. Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.
- [45] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Proc. ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.
- [46] T. Toda *et al.*, "The voice conversion challenge 2016," in *Proc. Interspeech*, 2016, pp. 1632–1636.
- [47] J. Lorenzo-Trueba *et al.*, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel method," 2018, [arXiv:1804.04262](https://arxiv.org/abs/1804.04262).
- [48] M. Wester and H. Liang, "The EMIME Mandarin bilingual database," Univ. Edinburgh, Edinburgh, U.K., Tech. Rep. EDI-INF-RR-1396, 2011.
- [49] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [50] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. Workshop Autom. Speech Recognit. Understanding*, 2011, no. EPFL-CONF-192584.
- [51] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate sound quality," 2003.
- [52] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.