IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 October 2022, accepted 27 October 2022, date of publication 1 December 2022,
date of current version 29 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3226350

RESEARCH ARTICLE

Jointly Trained Conversion Model With LPCNet for Any-to-One Voice Conversion Using Speaker-Independent Linguistic Features

IVAN HIMAWAN[1], RUIZHE WANG[1], SRIDHA SRIDHARAN[2], (Life Senior Member, IEEE), AND CLINTON FOOKES[2], (Senior Member, IEEE)

[1]ObEN, Pasadena, CA 91103, USA
[2]Queensland University of Technology, Brisbane, QLD 4000, Australia

Corresponding author: Ivan Himawan (ivan@oben.com)

ABSTRACT We propose a joint training scheme of an any-to-one voice conversion (VC) system with LPCNet to improve the speech naturalness, speaker similarity, and intelligibility of the converted speech. Recent advancements in neural-based vocoders, such as LPCNet, have enabled the production of more natural and clear speech. However, other components in typical VC systems are often designed independently, such as the conversion model. Hence, separate training strategies are used for each component that is not in direct correlation to the training objective of the vocoder preventing exploitation of the full potential of LPCNet. This problem is addressed by proposing a jointly trained conversion model and LPCNet. To accurately capture the linguistic contents of the given utterance, we use speaker-independent (SI) features derived from an automatic speech recognition (ASR) model trained using a mixed-language speech corpus. Subsequently, a conversion model maps the SI features to the acoustic representations used as input features to LPCNet. The possibility to synthesize cross-language speech using the proposed approach is also explored in this paper. Experimental results show that the proposed model can achieve real-time VC, unlocking the full potential of LPCNet and outperforming the state of the art.

INDEX TERMS Automatic speech recognition, conversion model, joint training, neural vocoder, voice conversion.

I. INTRODUCTION

Voice conversion (VC) is a technique for converting paralinguistic information of a source speaker's speech without changing the linguistic content. The objective of the VC system is to learn a mapping function from the source to the target speech. For a given utterance from the source speaker, the standard VC pipeline decomposes the speech signals into feature vectors, and the mapping module changes them towards the target speaker. Time-domain speech waveforms are then reconstructed using a vocoder [1], [2]. There are many applications that can benefit from VC research such as personalized text-to-speech synthesis systems, voice dubbing, and speaking-aid devices [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Ganesh Naik.

Statistical approaches to VC such as Gaussian Mixture Model (GMM)-based mapping have been among the successful methods in the past [4], [5], [6]. These methods for VC generally require a training set containing parallel data, where speech of the same linguistic content is available from both the source and the target speakers to learn the spectral mapping. However, it is not always possible to record the audio of the same sentences from different speakers. In recent years, neural network methods with more powerful regression capabilities have become popular means to solve the conversion problem. In particular, VC research with non-parallel training data has benefited greatly from deep learning techniques where a mapping function can be learned effectively [7], [8], [9]. For example, [7] proposed a sequence-to-sequence (Seq2Seq) VC in order to distangle the linguistic representations and the speaker identity components. The

model learns linguistic representations from acoustic features using the output of a text encoder as the reference. Thus, the training process requires phoneme transcriptions from audio samples. At the conversion time, a Seq2Seq decoder is employed to reconstruct the acoustic features using the target speaker representation. Moreover, techniques that do not require transcriptions such as CyleGAN-VC [10], StarGAN-VC [11], and VAW-GAN [12] have incorporated generative adversarial networks that improve voice quality and similarity to the target speaker when a large amount of speech data is employed.

One of the popular research directions towards non-parallel VC is the use of linguistic-related features derived from the automatic speech recognition (ASR) model, such as bottleneck features or Phonetic Posteriorgrams (PPGs) [13], [14], [15], [16], [17], [18]. In these approaches, the ASR module trained for phoneme classification is used to extract speech embeddings as intermediate phonetic representations [13], [19], [20]. Typically, features derived from the ASR system trained by using a large multi-speaker corpus is considered to be speaker-independent (SI). In this framework, a conversion model is typically employed to convert PPGs extracted from the source speech into spectral features of the target speaker. Also, the transcriptions from audios are not necessary in order to train the conversion model. Finally, a vocoder is used to synthesize the speech waveforms of the target speaker from the converted features [21], [22].

One main drawback of a VC system that uses intermediate representations, such as the spectral features (i.e., Mel-cepstrum), is the separate training process involved when building the neural networks for the conversion process. The vocoders are usually designed independently where the majority of the neural vocoders use Mel-cepstrum computed from short-time Fourier transform (STFT) as input. This causes artifacts to be produced in the converted speech, as the conversion model and the vocoder are two separate modules. Although a traditional parametric vocoder can be used, the quality of the synthesized speech is lower compared to the neural vocoder. One of the most successful implementations of a neural vocoder is WaveNet [23]. WaveNet is an autoregressive (AR) generative model that can produce high-fidelity audio. The AR structure improves the continuity of the generated waveform. However, it is too slow for real-time synthesis because it generates the waveform sampling points one by one. As an alternative to WaveNet, the WaveRNN [24] has been proposed to match the quality of the WaveNet model. The WaveRNN model uses a sparse gated recurrent unit (GRU) layer instead of the dilated causal convolutions used in WaveNet.

Recently, an efficient neural vocoder based on WaveRNN, called LPCNet is introduced [25]. LPCNet exploits linear predictive coding (LPC) to model the vocal tract response and applies linear prediction techniques to WaveRNN, which reduces the complexity of generating the raw speech waveform. LPCNet can synthesize higher quality speech than WaveRNN for the same network size. Furthermore, LPCNet inference has been implemented to run faster than real-time on a single CPU core with the use of efficient vectorization. Since it was proposed, LPCNet has been one of the popular choices for the speech synthesis task. Therefore, many techniques have been proposed to accelerate the inference speed of LPCNet [26], [27], [28]. High-fidelity neural vocoders based on the use of generative adversarial networks (GANs) have also attracted great interest due to their lightweight architectures and fast speech generation [29], [30], [31]. However, these vocoders can be difficult to train and may produce audible artifacts such as pitch error and periodicity artifacts due to their non-AR structures [32].

In this paper, we propose a jointly trained conversion model and LPCNet vocoder for *any-to-one* VC, to convert an arbitrary speaker's voice, including speakers who were unseen during the training, to the voice of a known speaker. We show that the proposed joint training scheme can achieve high-quality conversion in terms of speech naturalness and speaker similarity of the converted speech. We also explore whether the proposed framework can synthesize speech in a specific language that the target speaker does not speak [33]. The latter task is possible since the SI features derived from ASR trained using a mixed-language speech corpus can be assumed to be language independent [34]. The rest of the paper is organized as follows. Section II discusses related work. Section III presents the proposed method, followed by experiments in Section IV. We present and discuss the results in Section V, and conclude the study in Section VI.

II. RELATED WORK

A recent work in any-to-one VC is an auto-regressive voice conversion (ARVC), a technique based on sequence-to-sequence (Seq2Seq) VC [7], [8], [9] that translates the PPGs to acoustic features [16]. In Seq2Seq VC, the encoder-decoder architecture is typically used to learn mapping between a source and target feature sequences, which are often of different lengths by capturing and using the long-range dependencies. The system in [16], [17], and [18] contains a CBHG (1-D convolution bank + highway network + bidirectional GRU) module [35] as an encoder that consists of a bank of 1-D convolutional filters, followed by highway networks and a bidirectional gated GRU. For the decoder, it contains an attention layer, a long short-term memory (LSTM) and a pre-net. The encoder-decoder model translates the PPGs to acoustic features. LPCNet is used as a vocoder, which is trained independently. It is not possible to use PPGs alone as the conditional features to LPCNet since LPCNet requires linear predictive coefficients to be computed explicitly from acoustic features during inference [36]. While the techniques in [16], [17], and [18] can achieve high quality converted speech, our goal is not to report one-to-one performance comparison with those approaches, but rather focus on unlocking the potential of LPCNet for VC. By conducting subjective and objective voice quality tests of our proposed approach and through comparison with relevant baseline

systems, we have established that our proposed approach provides state-of-the-art voice conversion quality.

Our approach is similar to a jointly trained conversion model and WaveNet proposed in [37]. However, the technique in [37] is a more complicated conversion framework because of the use of WaveNet. For example, the conversion model proposed in [37] takes PPG features, voiced/unvoiced flag (VUV), and log fundamental frequency (F0) as inputs, predicts Mel-spectrograms as the target features, and outputs the bottleneck features. The PPGs are then concatenated with the bottleneck features and need to be upsampled to match the time resolution of the speech waveform as input to the WaveNet in the training stage. These additional steps in [37] of predicting Mel-spectrograms and outputting bottleneck features, and the upsampling procedure afterwards would not be necessary to perform the speech conversion in the proposed framework. Moreover, in our method, we use speaker embeddings extracted from the speaker encoder network trained for classifying many speakers as auxiliary features to better capture the characteristics of the target speakers. LPCNet offers significant advantage in terms of model complexity. Compared to LPCNet, WaveNet is a significantly more complex model (i.e., more neurons). Thus, it typically needs a larger amount of training data to achieve high-quality speech.

III. PROPOSED METHOD

The proposed VC framework consists of three main components: (1) Linguistic content extractor, (2) Conversion model, and (3) LPCNet vocoder. We use the frame-level SI linguistic features as the input. The conversion model maps these features into acoustic features. Finally, the LPCNet transforms the predicted acoustic features into speech waveforms.

A. LINGUISTIC CONTENT EXTRACTOR

An ASR model is used for the linguistic content extraction. The acoustic model for the ASR employs LSTMs which are trained by using the frame-level cross entropy criterion. The LSTMs structure is used to address the vanishing gradient problem encountered when training deep neural networks (DNNs) [38]. The model is trained from ObEN's [1] one thousand hour mixed-language corpus (including English and Mandarin) using 42-dimensional features (39-dimensional MFCC plus pitch features). The acoustic features are extracted from 16 kHz speech waveforms with a 25 ms frame length.

The model is trained to estimate posterior probabilities of roughly 5K tied-state (senone) targets. The model uses a three-layer LSTM architecture with 512 units in each layer. The output of the LSTM layer is connected to a fully connected (bottleneck) layer. The output of the bottleneck layer is connected to a fully connected layer to predict the frame-level labels. The Kaldi toolkit [39] is used to obtain the state alignments from the GMM/HMM system for training the

[1] http://www.oben.me

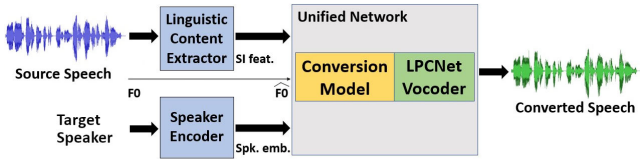

FIGURE 1. At run-time conversion, the SI features and F0 are extracted from the source speech. A linear transformation is applied on F0 to match the statistics of the target speaker. These features along with the speaker embeddings are inputted into the unified network to generate the converted speech.

LSTMs. The outputs from the bottleneck layer are utilized as the frame-level linguistic features. We consider the features to be speaker-independent when they are derived from the ASR system trained by using a large multi-speaker corpus.

B. CONVERSION MODEL

The conversion model is constructed using stacked bidirectional LSTMs (BiLSTMs) [40]. We connect the output of the BiLSTM layer with a residual network to predict the Bark-scale frequency cepstral coefficient (BFCC) features, pitch period, and pitch correlation parameters. The primary task of the conversion model is feature-mapping, which is to minimize the distance between the predicted and ground-truth of the target speaker's acoustic features. In our experiments, we use mean squared error (MSE) loss as the objective function to optimize the parameters of the network.

C. LPCNet VOCODER

We use the LPCNet [25] to synthesize high-quality speech. LPCNet is an efficient vocoder based on WaveRNN [24] with two key components: the frame rate network and the sample rate network. The input acoustic features for LPCNet comprise the 18-dimensional BFCCs, one dimensional pitch period, and one dimensional pitch correlation for a sampling frequency of 16 kHz. The frame rate network extracts embedded representations from the input acoustic features. The sample rate network consists of two gated recurrent units (GRU_A and GRU_B) and one dual fully connected layer, followed by a softmax layer to model the probability distribution of the excitation signal e_t. In order to generate the audio sample s_t, e_t is sampled from this distribution and combined with the prediction p_t from the LPC filter [25].

D. MODEL TRAINING

First, we train both the conversion model and the LPCNet vocoder independently. In this phase, we initialize the parameters of the two networks since the random weight initialization is unlikely to be effective due to the size of the network. In the second step, a joint network is built by concatenating a conversion model and the LPCNet. The conversion model tries to reconstruct the acoustic features which are inputted to the LPCNet. In the jointly trained system, the LPCNet parameters are learned together with the conversion model parameters through back-propagation. The unified network is

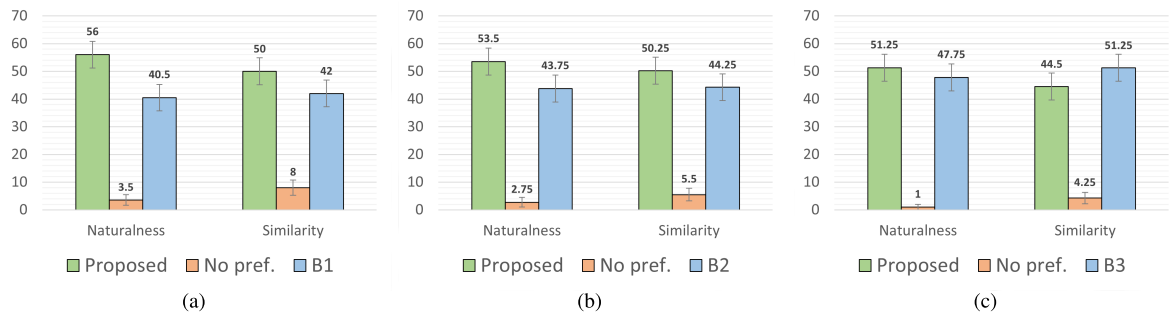

FIGURE 2. Testing results on two male target speakers on naturalness and similarity between the proposed method and (a) baseline system 1 (B1), (b) baseline system 2 (B2), and (c) baseline system 3 (B3). The error bars represent 95% confidence intervals.

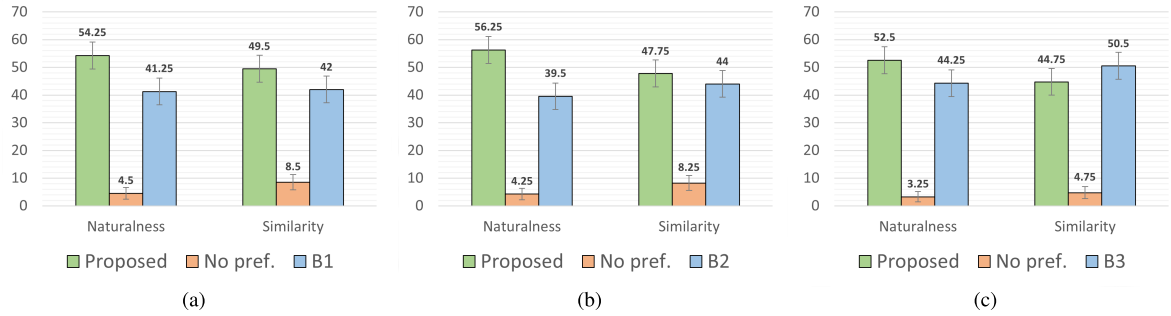

FIGURE 3. Testing results on two female target speakers on naturalness and similarity between the proposed method and (a) baseline system 1 (B1), (b) baseline system 2 (B2), and (c) baseline system 3 (B3). The error bars represent 95% confidence intervals.

trained using a weighted combination of a LPCNet loss and the acoustic features reconstruction loss, with a total loss,

$$\mathcal{L}_{total} = \mathcal{L}_{LPCNet} + \beta \mathcal{L}_{reconstruction}, \tag{1}$$

where the reconstruction loss is computed using the MSE between the predicted and the ground-truth acoustic features. β defines the weight of the reconstruction loss.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

We use both public and our proprietary datasets for VC experiments. One female English speaker is taken from [43]. The other three speakers (a male English speaker, a male Mandarin speaker, and a female Mandarin speaker) are obtained from the studio recordings by voice actors, originally sampled at 48 kHz. The VC tasks often assume that there is limited data from the target speaker. However, the neural vocoders typically employ a large amount of training data to achieve high-quality synthesized speech. Therefore, we use four speakers as target speakers in our experiments: Two English speakers (a male and a female) with 2 hours of data and two Mandarin speakers (a male and a female) with 1 hour of data. We build four VC systems, one for each target speaker. To evaluate the proposed system, we sample 20 random speakers (10 males and 10 females) from the VCTK corpus [41] as source English speakers. For source Mandarin speakers, 20 random speakers (10 males and 10 females) with neutral utterances from the CSLT-ESDB corpus [42] are selected.

B. IMPLEMENTATION DETAILS AND INFERENCE SPEED

The input of the conversion model is a sequence of SI features. We also use speaker embeddings as auxiliary features to better represent different aspects of speaker characteristics. This is not detrimental to the performance of any-to-one VC systems. The speaker embeddings are extracted from a deep neural network (DNN) trained to classify many speakers [44]. The RAPT algorithm [45] is used to extract the F0. The conversion model is constructed using a stacked four-layer BiLSTM with 256 hidden units for each layer. Dropout is set to 0.2. The network is trained using the Adam optimizer with a learning rate of 1×10^{-3}.

The LPCNet operates at 16 kHz sampling rate and a frame rate network that processes 10 ms frames (160 samples). We use 18 Bark-scale frequency cepstral coefficients with 10 ms shift size and 320 window size. We set GRU_A with 256 units and GRU_B with 16 units. Other network parameters follow the original LPCNet implementation. The LPCNet is trained for 140 epochs and 120 epochs for English speaker and Mandarin speaker models, respectively. The batch size is 32, and a learning rate is set to 1×10^{-3}.

For the unified network training, a pre-trained conversion model is concatenated with a pre-trained LPCNet. In this second stage of training, we reduce the learning rate to

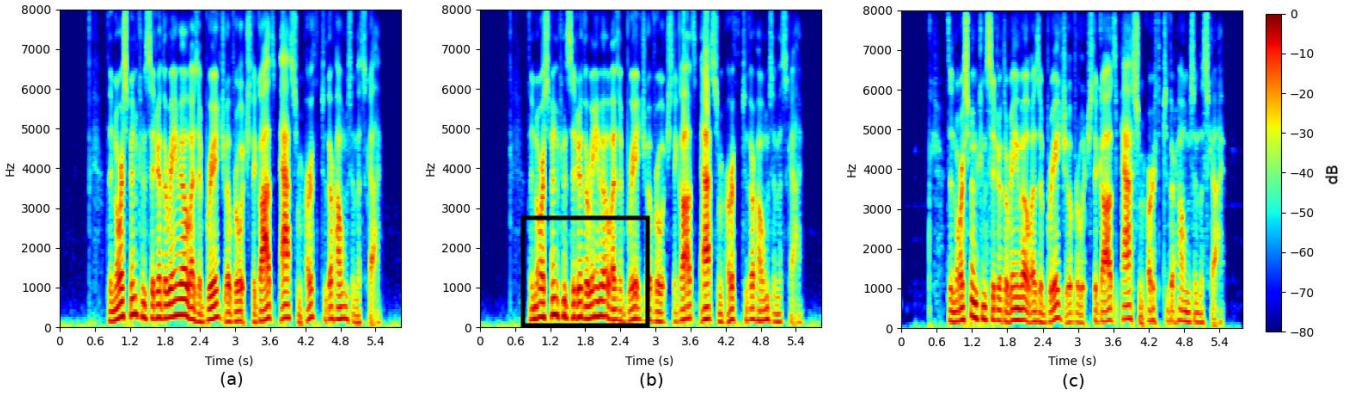

FIGURE 4. Spectrogram of the converted speech from (a) the proposed system, (b) the B1 system, (c) the B3 system.

TABLE 1. Mean opinion score (MOS) results with 95% confidence intervals.

Scenario	Naturalness		Similarity	
	Male speakers	Female speakers	Male speakers	Female speakers
Proposed	3.32 ± 0.10	3.46 ± 0.10	3.33 ± 0.11	3.48 ± 0.10
B1	2.93 ± 0.11	3.33 ± 0.10	3.18 ± 0.11	3.35 ± 0.10
B2	2.91 ± 0.11	3.29 ± 0.10	3.15 ± 0.11	3.38 ± 0.10
B3	3.33 ± 0.10	3.39 ± 0.10	3.47 ± 0.10	3.37 ± 0.10
Target Voice	3.92 ± 0.10	3.72 ± 0.09		

TABLE 2. Average cosine similarity values.

Target	Source-Target	Proposed	B3
Male speaker 1	0.318	0.650	0.695
Male speaker 2	0.219	0.521	0.539
Female speaker 1	0.313	0.584	0.617
Female speaker 2	0.242	0.436	0.372

5×10^{-4}. The β is set empirically (i.e., 0.001). To generate converted speech, we extract the SI features and F0 from the source speech. We apply linear transformation on F0 to match the statistics of the target speaker. To compute the speaker embeddings for the target speakers, we select an utterance that is long enough in the training data as a reference utterance. These features along with the speaker embeddings are inputted into the unified network. This process is shown in Fig. 1.

We measure the synthesis of 16 kHz speech waveform on a laptop device with an Intel Core i7-8550U 1.80 GHz CPU. At run-time conversion, the whole source sequence is used as input that we wish to convert to another person's voice. For the whole system, we obtain 0.64 real-time factor (RTF), that is the time (in seconds) to synthesize a one second waveform. This measurement yields 0.09, 0.1, and 0.45 RTFs for feature extraction, acoustic features reconstruction, and LPCNet synthesis, respectively. Hence, the proposed method can perform voice conversion faster than real-time.

C. BASELINE SYSTEMS

We implement three systems for the baseline comparison. The first baseline system (B1) uses an independently trained conversion model and LPCNet [17]. The second baseline system (B2) uses LPCNet fine-tuned using outputs of the conversion model. In this case, only the parameters of LPCNet model are updated during training. We use a different vocoder for the third baseline system (B3). The vocoder is based on MelGAN [29]. The generator follows MelGAN architecture but we increase the receptive field by deepening the ResStack layers. Each ResStack has 4 layers with dilation 1, 3, 9 and 27 with kernel-size 3. We also add additional discriminators, a discriminator on the Mel-spectrogram and an ensemble of random window discriminators (RWD) used in GAN-TTS [46]. An esemble of RWD combines outputs from 5 unconditional and 5 conditional discriminators which operate on randomly sub-sampled portions of the real or generated waveforms. There are five window sizes (i.e., 240, 480, 960, 1920, 3600 samples) obtained by downsampling the input raw waveform to a constant temporal dimension. This GAN vocoder is selected since our goal is to synthesize speech in real-time. For the B3 system, we pre-train the vocoder with a multi-speaker corpus and then fine-tune the model towards the target speaker. We directly input the SI features to a vocoder for inference.

D. TEST PROCEDURES

We conduct speech naturalness and speaker similarity for subjective evaluations. The Amazon Mechanical Turk

TABLE 3. P-values produced by binomial test.

	Male target speakers		Female target speakers	
	Naturalness	Similarity	Naturalness	Similarity
Proposed vs. B1	0.0018	0.1059	0.0089	0.1294
Proposed vs. B2	0.0538	0.2367	0.0007	0.4649
Proposed vs. B3	0.5136	0.1839	0.1036	0.2596

platform is used to perform the listening tests. Each utterance is assessed by ten random human workers, and each worker can answer at most five hits in a single experiment.

We perform an AB preference test to evaluate the naturalness of the converted speech. In the AB test, pairs of samples, consisting of converted speech obtained by the proposed approach and baseline, are presented to the listeners in random order. The listeners are asked to judge which sounded more natural (A or B) or choose no preference. In the speaker similarity ABX test, a real speech from the target speaker is played first, followed by each converted utterance from the two systems we would like to evaluate, played in random order. The listeners are asked to judge which utterance is closer in speaker identity to the target speaker's speech or to choose no preference when they could not tell the difference.

In addition, we also conduct mean opinion score (MOS) study. For the naturalness test, a real speech sample from the target speaker and the four conversion methods (proposed, B1, B2, and B3) are played one at a time in random order, and the listener is asked to give a 5-scale opinion score (1 for the completely unnatural speech and 5 for the completely natural speech). In the speaker similarity MOS test, a real speech utterance from the target speaker is played first as a reference, and then each of the converted speech from the four conversion models are played in random order. The listener is asked to give a 5-scale opinion score whether the speech is produced by the same speaker in a reference clip (1 for unlikely produced by the same speaker and 5 for the definitely produced by the same speaker). The participants can replay the audio clips before submitting their scores.

To measure the speech intelligibility of the converted speech, we use an off-the-self ASR system [2] and measure the word error rate (WER) for English sentences and the character error rate (CER) for Mandarin.

V. RESULTS

The subjective evaluation results of pairwise system comparison for the target male English speakers are illustrated in Fig. 2. Overall, we can see in Fig. 2(a), Fig. 2(b), and Fig. 2(c) that the proposed approach outperforms the baseline systems in terms of naturalness and speaker similarity of the converted speech. However, the B3 system is preferred over the proposed system in terms of similarity. Similar trends are observed for the female speakers in Fig. 3. We find that the proposed system performs better than the baseline systems,

TABLE 4. Word error rate (WER) (%) of the converted speech for English speakers and Character error rate (CER) (%) for Mandarin speakers.

	Source	Proposed	B1	B2	B3
English target	7.6%	**27.4%**	31.2%	46.0%	36.2%
Mandarin target	4.9%	**29.7%**	37.2%	52.2%	35.6%

except for the B3 system in terms of speaker similarity. Table 1 shows MOS with 95% confidence intervals. The same performance trends are observed using MOS test to the subjective results in Fig. 2 and Fig. 3, in which the proposed approach has higher naturalness and similarity scores than B1 and B2.

To analyze speaker similarity of the converted speech from the B3 system, we compute cosine similarity between embedding vectors of audio samples and the target speaker embedding vectors before and after conversion. A pre-trained neural speaker embedding model from a deep speaker system is used to extract the embedding vectors .[3] It is expected that the cosine similarity between each pair of embedding vectors of the same speaker would be higher than for any pair of vectors of different speakers. We can see from Table 2 that the cosine similarity values are higher after the conversion regardless of techniques. However, B3 produces higher scores compared to the proposed approach except for female speaker 2. One reason is because we use a multi-speaker corpus to pre-train the B3 vocoder. Even though the SI features are assumed to be speaker independent, they may still contain speaker-dependent information that can degrade the similarity of the converted speech. When we do not pre-train the vocoder, for example, a male speaker 1, we obtain cosine similarity value of 0.506 that is lower than 0.695. Therefore, to improve the generalization towards a new speaker which is not part of the training data, we use dataset with many speakers to pre-train MelGAN. Another reason for pre-training is to improve the synthesis quality of the vocoder. In our experiments, LPCNet has better synthesis quality compared to MelGAN using 1-2 hours of speech. The reason may be that the past waveforms are used for auto-regressive structure in LPCNet in such a way that it reduces the data requirement [47].

We also perform binomial test where the null hypothesis is that the two categories (proposed vs. baseline) are equally preferred. The p-values for each experiment are listed in Table 3. The test results suggest that the proposed technique

[2]https://github.com/watson-developer-cloud/speech-to-text-nodejs

[3]https://github.com/philipperemy/deep-speaker

is significantly better than B1 and B2 systems in terms of naturalness (p-values < 0.05), except for proposed vs. B2 for male target speakers for which p-value is 0.0538. This suggests that it is vital to re-estimate the weights of the conversion model during joint training in order to improve the quality of the converted speech. Fig. 4 shows improvements through visualization by comparing the converted speech of an English male target between the proposed, B1, and B3 systems. We observe that our jointly trained system removes amplitude artifacts, and for the B1 system, the reconstructed speech produces artifacts (as shown inside the bounding box in Fig. 4(b)). For the B3 system, audible artifacts sometimes appear in the non-speech segments.

As shown in Table 4, the proposed method performs the best in terms of intelligibility. Although the B3 system (our modified MelGAN-based vocoder) performs competitive in terms of subjective measures, it does not perform well in terms of the intelligibility measure. A better MelGAN model may generate better voices, however, it can be prone to overfitting due to its complex discriminators [48].

A. SYNTHESIZING CROSS-LANGUAGE SPEECH

We further explore the language independent characteristics of the SI features by synthesizing speech in a language not spoken by the target speaker. Hence, we convert 20 speakers from the VCTK to the female Mandarin speaker. We also convert 20 speakers from the CSLT- ESDB to a male English speaker. We notice a degradation of 6% absolute WER when synthesizing English speech using the female Mandarin speaker. On the other hand, the 5.7% absolute CER degradation is measured when synthesizing Mandarin using the male English speaker. This degradation in the intelligibility suggests that the error produced by the ASR could propagate to the acoustic modeling process in the later stages. Despite the degradation in speech intelligibility, the SI features extracted from multilingual ASR model are able to capture the phonetic patterns for synthesizing cross-language speech when the training data contain both the source and target languages.

In our experiments, we notice that using F0 as input features to automatic speech recognition (ASR) model and conversion model improve the quality of the converted speech in Mandarin. Since the proposed method depends on using SI features to convey the linguistic contents from source to target, the robustness of ASR model has a great influence on the quality of SI features' phonetic contents. In the future, we will investigate the data requirement for Mandarin and other languages when training ASR system and conversion model. Speech samples from voice conversion experiments, including cross-language samples are available online at https://oben-ssw10.github.io/lpcnet_vc/.

VI. CONCLUSION

In this work, we propose a jointly trained voice conversion model with LPCNet to convert a given utterance faster than real-time. The subjective experiments show that the jointly trained system produces high-quality converted speech in terms of naturalness and similarity to the target speaker's voice when compared to the state of the art. We use four target speakers with widely different speech characteristics, and observe speech quality improvement in the converted voice when compared to the independently trained system. In particular, the male English speaker's voice is difficult to model with many occurrences of vocal fry in the training sentences, and it benefits greatly from the jointly trained network. We also demonstrate that the proposed framework can synthesize speech in a language that the target speaker does not speak by leveraging a multilingual ASR training. Future research includes finding a method to effectively disentangle speech content and speaker identity from utterances.

ACKNOWLEDGMENT

This work is performed while the author Ivan Himawan is at ObEN.

REFERENCES

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017.

[2] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 132–157, 2021.

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

[4] 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.

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

[6] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 912–921, Jul. 2010.

[7] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 631–644, Mar. 2019.

[8] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "ATTS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6805–6809.

[9] H. Kameoka, W.-C. Huang, K. Tanaka, T. Kaneko, N. Hojo, and T. Toda, "Many-to-many voice transformer network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 656–670, 2021.

[10] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2100–2104.

[11] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 266–273.

[12] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," in *Proc. Interspeech*, Aug. 2017, pp. 3364–3368.

[13] 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 (ICME)*, Jul. 2016, pp. 1–5.
[14] X. Tian, E. S. Chng, and H. Li, "A speaker-dependent WaveNet for voice conversion with non-parallel data," in *Proc. Interspeech*, Sep. 2019, pp. 201–205.
[15] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 540–552, 2020.
[16] Z. Lian, Z. Wen, X. Zhou, S. Pu, S. Zhang, and J. Tao, "ARVC: An auto-regressive voice conversion system without parallel training data," in *Proc. Interspeech*, Oct. 2020, pp. 4706–4710.
[17] L. Zheng, J. Tao, Z. Wen, and R. Zhong, "CASIA voice conversion system for the voice conversion challenge 2020," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, Oct. 2020, pp. 136–139.
[18] Z. Lian, R. Zhong, Z. Wen, B. Liu, and J. Tao, "Towards fine-grained prosody control for voice conversion," in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Jan. 2021, pp. 1–5.
[19] X. Tian, J. Wang, H. Xu, E.-S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, Jun. 2018, pp. 227–232.
[20] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and D-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5274–5278.
[21] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *Proc. Interspeech*, Aug. 2017, pp. 1138–1142.
[22] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet vocoder with limited training data for voice conversion," in *Proc. Interspeech*, Sep. 2018, pp. 1983–1987.
[23] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
[24] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2410–2419.
[25] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5891–5895.
[26] H. Kanagawa and Y. Ijima, "Lightweight LPCNet-based neural vocoder with tensor decomposition," in *Proc. Interspeech*, Oct. 2020, pp. 205–209.
[27] R. Vipperla, S. Park, K. Choo, S. Ishtiaq, K. Min, S. Bhattacharya, A. Mehrotra, A. G. C. P. Ramos, and N. D. Lane, "Bunched LPCNet: Vocoder for low-cost neural text-to-speech systems," in *Proc. Interspeech*, Oct. 2020, pp. 3565–3569.
[28] V. Popov, M. Kudinov, and T. Sadekova, "Gaussian LPCNet for multi-sample speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6204–6208.
[29] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14910–14921.
[30] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6199–6203.
[31] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17022–17033.
[32] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive GAN for conditional waveform synthesis," in *Proc. ICLR*, 2022.
[33] Z. Yang, W. Zhang, Y. Liu, and X. Xing, "Cross-lingual voice conversion with disentangled universal linguistic representations," in *Proc. Interspeech*, Aug. 2021, pp. 1604–1608.
[34] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2012, pp. 5891–5895.
[35] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 4006–4010.
[36] K. Subramani, J.-M. Valin, U. Isik, P. Smaragdis, and A. Krishnaswamy, "End-to-end LPCNet: A neural vocoder with fully-differentiable LPC estimation," in *Proc. Interspeech*, Sep. 2022.
[37] S. Liu, Y. Cao, X. Wu, L. Sun, X. Liu, and H. Meng, "Jointly trained conversion model and WaveNet vocoder for non-parallel voice conversion using mel-spectrograms and phonetic posteriorgrams," in *Proc. Interspeech*, Sep. 2019, pp. 714–718.
[38] T. He and J. Droppo, "Exploiting LSTM structure in deep neural networks for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5445–5449.
[39] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Jun. 2011, pp. 1–4.
[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[41] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," Univ. Edinburgh, Centre Speech Technol. Res. (CSTR), Edinburgh, U.K., Tech. Rep., 2019.
[42] F. Bie, D. Wang, T. F. Zheng, J. Tejedor, and R. Chen, "Emotional adaptive training for speaker verification," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Oct. 2013, pp. 1–4.
[43] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-Fi multi-speaker English TTS dataset," in *Proc. Interspeech*, Aug. 2021, pp. 2776–2780.
[44] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4480–4490.
[45] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., Amsterdam, The Netherlands: Elsevier, 1995.
[46] M. Binkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *Proc. ICLR*, 2020, pp. 1–17.
[47] K. Matsubara, T. Okamoto, R. Takashima, T. Takiguchi, T. Toda, Y. Shiga, and H. Kawai, "Investigation of training data size for real-time neural vocoders on CPUs," *Acoust. Sci. Technol.*, vol. 42, no. 1, pp. 65–68, 2021.
[48] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12104–12114.

IVAN HIMAWAN received the B.E. degree in electrical and computer engineering and the Ph.D. degree with the research concentration from the Signal Processing, Artificial Intelligence and Vision Technologies (SAIVT), Queensland University of Technology (QUT), Brisbane, QLD, Australia, including a period of research at the CSTR, The University of Edinburgh, U.K. In 2010, he joined the Mobile Innovation Laboratory, QUT, and worked on a number of applied research projects in the areas of human–computer interactions on mobile devices. In 2014, he joined the IDIAP Research Institute, Switzerland. While at IDIAP, he worked in the field of automatic speech recognition. He worked as a Researcher with SAIVT and Ecoacoustics Groups, QUT, from 2015 to 2018. In 2018, he joined ObEN Inc., to develop and commercialize voice conversion and text-to-speech technology. He has coauthored and published more than 35 peer-reviewed papers at top journals and conferences in the areas of speech and signal processing.

RUIZHE WANG received the B.S. degree from Tsinghua University, the M.S. degree from the Caltech, and the Ph.D. degree from the Department of Computer Science, University of Southern California. He is currently a Principal Research Scientist at ObEN, an artificial intelligence company creating intelligent avatars that look, talk and behave like you and are authenticated and secured on the blockchain. Much of his work has been focusing on developing algorithms to digitize human visual appearance by only using commodity level hardwares. As of his work at ObEN Inc., he is developing a 3D avatar system which allows regular users to easily create and customize their corresponding 3D full body avatars from a single selfie taken with a smartphone. During his Ph.D. study, he worked on various projects, including reconstructing dynamic textured surfaces from three or four handheld depth sensors, body scanning and animation from a single fixed RGB-D camera, object detection from RGB-D images, and home monitoring of patients with Parkinson's disease using depth sensors. He has coauthored and published more than 15 research papers at top computer vision and computer graphics conferences, including CVPR, ECCV, SIGGRAPH, and VR. He also holds multiple patents directly deriving from his research. His research interests include computer vision, computer graphics, and machine learning.

SRIDHA SRIDHARAN (Life Senior Member, IEEE) received the M.Sc. degree in communication engineering from The University of Manchester, U.K., and the Ph.D. degree from the University of New South Wales, Australia. He is currently a Professor with the School of Electrical Engineering and Robotics, Queensland University of Technology (QUT). He is also a Research Leader of the Research Program in Signal Processing, Artificial Intelligence and Vision Technologies (SAIVT), QUT. He has published more than 600 papers consisting of publications in journals and in refereed international conferences in the areas of signal processing, computer vision, and machine learning, from 1990 to 2022. During this period, he has graduated over 80 Ph.D. students in these areas. He has also received a number of research grants from various funding bodies, including Commonwealth Competitive funding schemes, such as the Australian Research Council (ARC) and the National Security Science and Technology (NSST) Unit. Several of his research outcomes have been commercialized.

CLINTON FOOKES (Senior Member, IEEE) received the B.Eng. (Aero/Av), M.B.A., and Ph.D. degrees in computer vision. He is currently a Professor of vision and signal processing with the Queensland University of Technology and the Signal Processing, Artificial Intelligence and Vision Technologies (SAIVT) Laboratory. His research interests include computer vision, machine learning, signal processing, and pattern recognition. He is a Senior Member of an Australian Institute of Policy and Science Young Tall Poppy and an Australian Museum Eureka Prize Winner, and a Senior Fulbright Scholar. He serves on the Editorial Boards for the IEEE TRANSACTIONS ON IMAGE PROCESSING, *Pattern Recognition*, and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.