An Approach to Cross-Lingual Voice Conversion

by

sai sirisha Rallabandi, Suryakanth V Gangashetty

in

International Joint Conference on Neural Network
(*IJCNN-2019*)

Budapest, Hungary

Report No: IIIT/TR/2019/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
July 2019

An Approach to Cross-Lingual Voice Conversion

Sai Sirisha Rallabandi and Suryakanth V Gangashetty
Speech Processing Laboratory
International Institute of Information Technology, Hyderabad, India.
sirisha.rallabandi@research.iiit.ac.in, svg@iiit.ac.in

***Abstract*—The most prevalent multilingual Text-to-Speech (TTS) synthesis systems encounter an unnatural speaker shift at the language boundaries. This is observed when they are employed for code-mixed TTS synthesis. For the very fact that the collection of polyglot speech is non-trivial, many alternative approaches have been in focus. Cross-Lingual Voice Conversion (CLVC) has been one of those to generate speech with desired speaker and language identities. Our aim in this paper is to design a light-weighted CLVC framework between a pair of Mandarin-English speakers. CLVC is challenging when compared to traditional Voice Conversion (VC) because of its nature of accommodating unaligned corpus from the source and target speakers. We thus focus on generating a parallel corpus for CLVC and bridging the gap between speakers and languages. We perform a text-independent voice conversion with a three-layered conventional Neural Network (NN) for this purpose. The main contributions include i) Source similarity in both training and conversion stages of CLVC, ii) generation of a parallel corpus and iii) text independent and transcription free CLVC. We exploit two variants of a Neural Network in the proposed framework, i) an autoencoder to enable the source similarity and generation of parallel corpus, ii) a traditional DNN for feature mapping between the source and target. The subjective and objective evaluations show that the proposed method is indeed capable of performing a CLVC with an auto-encoded speech.**

***Index Terms*—Deep Neural Networks, Cross-Lingual Voice Conversion, Scaled Exponential Linear Units, Mel Generalised Cepstral Coefficients, Auto-encoded speech.**

I. Introduction

The need for globalization and improved computing abilities have made once prescient technological advancements achievable. Today's computers, not only emulate human speech but also do it more naturally. Neural generative models have surpassed the then widespread parametric and concatenative approaches in speech generation [1] [2]. However, our speech synthesizers are yet to impersonate the subjective traits of humans. One among them is to code-switch while in a conversation. This is more frequently observed in multilingual countries like India and Singapore. Code-mixing (or code-switching) is mainly observed in bilinguals/multilinguals who mix or borrow words from another language. This can be for various reasons such as i) to refer named entities [3], ii) ease of recall [4], iii) speech with respect to the target audience. Lately, multilingual speech synthesizers have been developed for such mixed language synthesis. These multilingual synthesizers utilized the speech collected from speakers of different languages. Thus resulting in multiple speaker shifts at the language boundaries. This peculiarity in generated speech can be addressed using a polyglot speech synthesizer. A polyglot speech synthesizer is a TTS system that can produce speech in a single voice for multiple languages [5]. Hence, the generated speech does not create any sort of perceptual unpleasantness at the language boundaries.

Significant work has been done on building polyglot speech synthesis systems [5]–[8]. The main approaches to it are: i) synthesis from polyglot speech corpus, ii) speaker adaptation techniques [9], iii) mapping techniques [10] [11], and iv) cross-lingual voice conversion [12]. We specifically focus on and discuss cross-lingual voice conversion in our work. Cross-lingual Voice Conversion (CLVC) is to transform the source speaker's characteristics in a way that his/her voice is perceived as if spoken by the target speaker who is oblivious of the language [5] [13]–[15]. Research on CLVC can be broadly divided into 2 categories: 1) Text-dependent CLVC 2) Text-independent CLVC. Under Text-dependent CLVC, bilingual speakers and the corresponding parallel data is available for a conventional frame-based mapping [13] [16] [17]. Nevertheless, this would restrict the system performance to specific speakers and languages. Also, the data collection from bilingual speakers for all the desired languages is not feasible. Accordingly, text-independent CLVC systems have been developed [14] [15]. Of late, Phonetic Posteriorgrams (PPG) based cross-lingual TTS systems have achieved much attention [18], [19]. However, the language-specific systems involved in such frameworks might hinder the performance of a cross-language generation. PPGs were also used in intralingual voice conversion for unaligned corpus [20] [21]. Here, the speech transcriptions obtained from a Speaker Independent Automatic Speech Recognition system (SI-ASR) were exploited for non-parallel VC. Our aim in this work is to propose a text independent and transcription free CLVC.

In this paper, we explore a pair of conventional Neural Networks with their two different variants for the task of CLVC. Our work is inspired by the previous studies that use a bilingual source for CLVC. We intend to create a source similarity from two monolingual speaker's speech. We achieve it using an auto-encoder as the cepstrum generator. The generated speaker and language normalized cepstrum is further treated as the source for CLVC. To the best of our knowledge, there is no work on text free CLVC so far. The languages that we experimented on are: English and Mandarin (Chinese). We refer to Mandarin as Chinese throughout our experiments. The proposed framework was also tested for a within language voice conversion with the unaligned corpus.

The main ideas of this paper are given below.

- We reduce the CLVC task into a monolingual one by generating artificial parallel corpus.
- We propose to use an auto-encoder to learn a language and speaker independent representation from both source (target language) and the target (target voice) data.
- Transcription free and text independent voice conversion framework is being proposed for CLVC.
- There is no dynamic time warping or EHMM forced alignment involved in the proposed framework thus avoiding any alignment issues.

The organization of this paper is as follows: In Section II, we provide a detailed explanation of the proposed framework followed by the extension of work to Non-parallel conversion in Section III. In Section IV, the experimental setup is discussed and the evaluation of the same is done in Section V. Sections VI and VII have the conclusions and future work respectively. Acknowledgements are provided in Section VIII.

II. Proposed Framework

We perform a CLVC between the languages English and Chinese in our proposed framework. English speaker and the corresponding linguistic information are considered as the source speaker and target language respectively. Accordingly, the Chinese voice is treated as target speaker. The research was carried out on investigating the conversion of English by English speaker to English by a Chinese speaker. Figure 1 displays the entire framework being proposed. The proposed framework has two training stages. Training stage 1 is the main contribution of this work. It renders both source similarity and parallel corpus for the conversion of speech to non-native speaker's voice. Figure 1 presents a detailed implementation of the proposed framework. The choice of models for each of the training stages along with the framework for a run-time conversion are explained in this section.

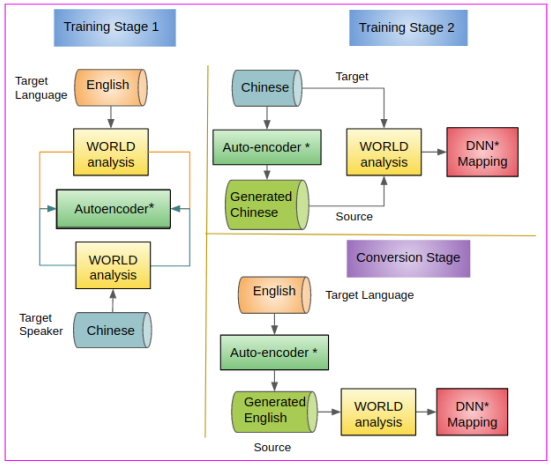

Fig. 1. Proposed framework for cross-lingual voice conversion. (*) indicates that the models are same.

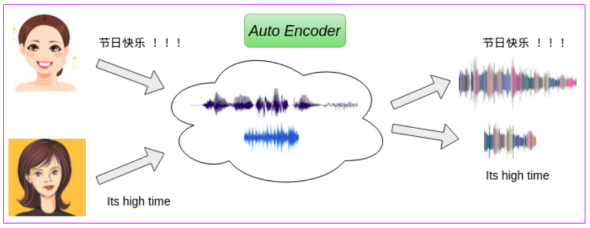

Fig. 2. Pictographic representation of the proposed framework

A. Training an auto-encoder for source similarity and parallel data

The apparent reason for CLVC to be challenging as opposed to conventional VC is the lack of parallel corpus. Studies show that a pair of bilingual speakers or a bilingual source can mitigate the difficulty in dealing with CLVC. Motivated by the idea of the bilingual source, we intend to create a source similarity and parallel corpus for CLVC from different speaker's speech. Based on some preliminary experiments and observations we presume to exploit an auto-encoder to bridge the gap between the speakers and languages. Auto-encoders are generally preferred for dimensionality reduction and the output of these networks is almost never used [22]. However, they are designed to capture the important information present in the data and not to exactly reproduce the input. Consequently, we employ an auto-encoder as a parallel corpus generator in our work. Figure 2 is a simple representation of using an autoencoder for speech generation. A similar idea of generating parallel corpus has been implemented in [23] for non-parallel voice conversion. However, our work differs from [23] in text independent parallel corpus generation i.e., we operate solely on speech throughout our work. Hence, our systems do not undergo any text-speech alignment problems. Our auto-encoder is initially fed with the data from 2 different speakers. The speakers deliver varied linguistic information in Chinese and English. Diverse data that is provided to the network is then projected onto a common feature space at the encoder output. These latent variables created by the encoder are later remapped to their respective representations. Hence, the remapped features have the characteristics of both the speakers and their languages. Chinese features are represented as a set of values in X and English features are in set Y. The feature transformation function is given in Equation (1).

$$\begin{aligned} Z(X,Y) &= F(X,Y) \\ G(\hat{X},\hat{Y}) &= Z(X,Y) \end{aligned} \tag{1}$$

In the above equation, Z refers to the hidden latent representation obtained from the encoder of the network. G is the decoder output reconstructed from the latent information present in Z. $\hat{X}$ and $\hat{Y}$ are the generated Chinese and English features respectively. Additionally, we inspect the effect of varied latent dimensions; bottleneck nodes for a high-quality speech generation. The auto-encoded speech thus generated has provided us with a source similarity for CLVC in addition to the parallel corpus generation.

B. Training a Cross-Lingual Voice Conversion model

From the subjective evaluations conducted on auto-encoder generated speech, we observe that the network with bottleneck layer containing 256 nodes has better feature representation and speech quality. Accordingly, the corresponding speech was used as a source in the training stage 2 and Conversion stage. In training stage 2, a parallel conversion is carried out in Chinese. The speech samples of generated and natural Chinese are of equal length. Thus, we do not employ any alignment approaches in our framework. The DNN model trained on Chinese is tested with the auto-encoded speech in English for CLVC.

C. Run-time conversion

Implementation of the proposed framework for a run-time conversion is provided in this section. The target speech in a non-target voice will be provided to the trained auto-encoder which generates a speaker and language normalized features. Thus generated features are passed through the DNN trained in stage 2 of the proposed framework. The transformed features along with their converted f0 and corresponding band aperiodicities are synthesized for speech using a WORLD vocoder [24]. The diagrammatic representation of run-time conversion is displayed in the Figure 3.

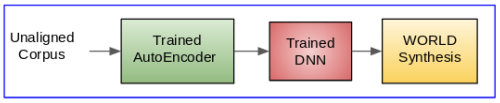

Fig. 3. A run-time conversion for proposed framework.

III. Mono-Lingual voice conversion with unaligned Corpus

Our framework for CLVC was further extended to intralingual voice conversion. The complexity of dealing with the unaligned corpus is also existent in within language speaker conversion. Under intralingual conversion, even though the source and target converse in the same language the linguistic content they deliver is different. Hence a frame-to-frame mapping of their acoustic features is unattainable. Accordingly, the autoencoder generated speech serves as a bridge and enables the parallel conversion between unaligned speech samples of the speakers.

IV. Experimental Setup

A. Baseline

A standard frame-based parallel voice conversion framework was developed as a baseline [25] [26] [27] [28] [29] [30]. A typical VC framework is shown in Figure 4. Two gender dependent experiments were carried out initially with a 4 layered DNN network. The number of nodes present in each layer of DNN are [60, 512, 512, 512, 512, 60]. The converted speech was validated with both subjective and objective tests. Motivated by the observations, we employ same DNN architecture in the proposed framework for frame-to-frame mapping. We further elaborate on the datasets used in each of the experiments. CMU arctic database was used in the implementation of the baseline systems [31]. The male-to-male conversion was done between speakers bdl (US male), rms (US male) and female-to-female conversion was between the speakers slt (US female), clb (US female). We train the model on 900 speech samples and test it on 100 in both the systems. All the speech samples undergo a signal analysis by the WORLD vocoder [24]. The obtained features of source and target are then aligned using a Dynamic Time Warping algorithm for a frame-to-frame mapping. Finally, the converted features are synthesized using the WORLD vocoder [24].

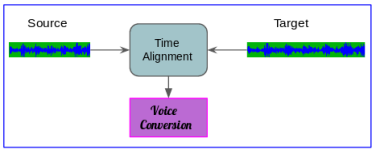

Fig. 4. Training phase of a traditional Parallel VC framework.

B. Neural Architectures in the Proposed Framework

The NN architectures employed for the auto-encoder in training stage 1 and the DNN based frame mapping in stage 2 are provided here. The inputs and outputs of both the networks are the Mel Generalized Cepstral Coefficients (MGCs) which constitute a 60-dimensional feature vector.

1) Autoencoder: An auto-encoder network has been extensively investigated to understand its reliability for parallel corpus generation. 4 different experiments were performed with varied bottleneck nodes: 10, 50, 256 and no bottleneck. Hence, the architecture of the auto-encoder is [60, 512, B, 512, 60] (B-bottleneck). Learning rate for bottleneck layers 10 and 50 was set to 0.01. For experiments that contained 256 bottleneck nodes and no bottleneck nodes, it was 0.1. The decay rate was set to $1e^{-6}$ for all the experiments.

2) DNN in Training stage 2: For a frame-based speaker mapping in stage 2, we use a four-layered DNN. This is same as the one used in building the baseline. Thus the architecture utilized is [60, 512, 512, 512, 512, 60]. Learning rate was set to 0.1 with a decay rate of $1e^{-6}$. Training samples were randomized and fed to the networks in batches of size 32.

C. Activations and Loss Functions

Since we were dealing with two different languages and speakers, we required our models to capture the feature representations projected onto a common space. Recently, Self-Normalising Neural Networks (SNNs) have been developed to map the means and variances of data within the network in a layerwise fashion. We thus exploited SELU activations in all the layers throughout our experiments [32]. The activation function is detailed in Equation (2).

$$SELU(input) = \lambda \left\{ \begin{array}{ll} \text{input} & \text{if input} > 0 \\ \alpha \mathrm{e}^{input} - \alpha & \text{if input} \leqslant 0 \end{array} \right\} \quad (2)$$

where α and λ regulate the means and variances of the output distribution of a layer. The values of α, λ have to be nearly 1.67326 and 1.0507 respectively.

In our proposed framework, we consider 4 possibilities of errors occurring in the predictions. Those errors are expected between 1) English to English, 2) Chinese to Chinese, 3) English to English due to Chinese, and 4) Chinese to Chinese due to English. All the four prediction errors are summed together for the error calculation and are back propagated. The anticipated errors are given below:

$$\delta 1 = G(\hat{X}) - X \tag{3}$$

Error $\delta 1$ in Equation (3) is calculated between the predicted Chinese $G(\hat{X})$ and original Chinese X. This error is due to the additional noise introduced by the neural network during predictions.

$$\delta 2 = G(\hat{Y}) - Y \tag{4}$$

Error $\delta 2$ in Equation (4) is calculated between the predicted English $G(\hat{Y})$ and original English Y. Thus, $\delta 2$ is also a network introduced error.

$$\delta 3 = G(\hat{X}, \hat{Y}) - Y \tag{5}$$

Error $\delta 3$ is calculated between the predicted English and original English Y. However, here the predictions are influenced by the Chinese speaker's speech. Hence, in Equation (5), $\delta 3$ is the error calculated between Chinese influenced English $G(\hat{X}, \hat{Y})$ and the natural one.

$$\delta 4 = G(\hat{X}, \hat{Y}) - X \tag{6}$$

Error $\delta 4$ is calculated between the predicted Chinese and original Chinese X. The Chinese predictions are influenced by the English speaker's speech. Hence, in Equation (6), $\delta 4$ is the error calculated when the predicted Chinese is under the impact of the English speaker's speech.

Thus, the total error is computed from sum of all the above errors for each of the derived predictions. The mathematical approach used for estimating total conversion loss is given in Equation (7).

$$Summed\_loss = \sum_{1}^{N} [(\delta 1)^2 + (\delta 2)^2 + (\delta 3)^2 + (\delta 4)^2] \tag{7}$$

Mean Squared Error (MSE) was computed as the loss function over the predicted MGCs. The MSE metric reports a frame-level error between the predicted and the target features. NN models in both the training stages are trained through a back propagation algorithm to learn the desired characteristics from the pair of speakers and languages [33]. Stochastic Gradient Descent (SGD) optimization is carried out for minimizing the MSE error [34], [35]. The momentum factor used for training the networks was 0.2. Our models are trained for 30 epochs in all our experiments. Both the networks were built in Keras [1].

D. Datasets used in the proposed framework

We present our proposed framework for two tasks where the availability of parallel corpus is not feasible: i) CLVC, and ii) Non-parallel VC. The datasets used for CLVC experiments are Blizzard Challenge 2010 [36] for Chinese (Chinese female) and SLT (US female) from CMU Arctic database [31] for English. The two languages chosen belong to diverse language origins. The type of speech is monolingual in nature for both the speakers; we do not have the instants where Chinese speaker speaks English and vice versa. We divide the entire corpus into three different sets namely: Train, Development, and Test. A detailed description of data split for the same is presented in Table I.

TABLE I
SUMMARY OF SPEECH SAMPLES USED FOR THE PROPOSED METHOD.

Dataset	Train	Development	Test
Chinese	624 (54 Min)	905 (1Hr:20Min)	100(9Min)
English	1092 (54 Min)	-	40 (2Min)

For training the auto-encoder we consider an equal amount of data from both the speakers to avoid any misinterpretations that might arise due to the speaker dominance. Generated speech (development data in stage 1) being the source in training stage 2, we leveraged the larger database for parallel conversion.

Under Non-parallel VC experiments, we used two unaligned Chinese datasets. They were Blizzard Challenge 2010 [36] (female) and THCHS30 [37] (female) released by Center for Speech and Language Technology (CSLT), Tsinghua University. We considered THCHS30 as Target speech and Blizzard Chinese speaker to be the Target voice. Similar to CLVC setup we utilize equal amount of data from both the datasets in training stage 1 and operate on larger datasets for parallel conversion in training stage 2.

E. Feature Extraction

Throughout our experiments, we employ the WORLD vocoder for speech signal analysis and synthesis [24]. We obtain the raw speech features: spectrum (sp) of 513 dimensions, fundamental frequency (f0) of 1 dimension and aperiodicity information (ap) of 513 dimensions upon WORLD analysis on the speech samples. Speech Signal Processing Toolkit (SPTK)[2] is utilized for the conversion of vocoder parameters to their corresponding Mel Generalised Cepstral coefficients (MGCs) of 60 dimensions, log f0 (lf0) of 1 dimension and band aperiodicities (bap) of 5 dimensions. All the speech samples were sampled at 16 kHz before feature extraction.

[1]https://github.com/SaiSirishaR/VC_Scripts/blob/master/train_DNN.py
[2]https://sourceforge.net/projects/sp-tk/

F. Speech Signal Reconstruction

In the proposed framework, a frame based mapping is carried out for MGCs using NNs. Further, the fundamental frequency is converted using a linear transformation calculated on the logarithmic means and variances of source and target speakers [38]. The mathematical representation for the linear transformation of f0 is given in Equation (8).

$$f0_{tgt} = \frac{\sigma_{tgt}}{\sigma_{src}} * (f0_{src} - \mu_{src}) + \mu_{tgt} \tag{8}$$

In the above expression, σ, μ represent the standard deviation and the mean respectively calculated on the logarithmic f0 of source and target speaker's speech. The band aperiodicities (bap) of source speaker were used as is for signal reconstruction. Finally, speech was reconstructed by the WORLD Vocoder from the vocoder parameters: spectrum (sp), frequency (f0) and aperiodicities (ap) [24].

V. Evaluation and Observations

We perform both subjective and objective tests for the proposed framework and baseline systems.

A. Objective Evaluation

Mel Cepstral Distortion (MCD) was considered as the objective metric for evaluation of the systems. MCD was obtained based on the distance between the converted and original features. Hence, lower MCD scores result in acceptable conversion of speaker identity; the predicted features are close to that of the target [39]. MCD scores are calculated over the dimensions of MGCs. The mathematical representation implemented for MCD calculation is given in the Equation (9).

$$MCD = \frac{10}{ln10}\sqrt{2\sum_{p=1}^{60}(mgc_p - mgc'_p)^2} \tag{9}$$

where MCD-Mel Cepstral Distortion, mgc-Mel Generalized Cepstrum, mgc and $mgc^{'}$ are the natural and converted feature matrices evaluated frame-wise at p^{th} dimension. The MCD scores for the baseline method are provided in Table II.

TABLE II
MCD scores for the Baseline

System	MCD scores
Female-Female	4.22
Male-Male	4.52

MCD scores obtained from both the systems are not too far from the target. Hence, motivated by these observations we have carried out the frame-to-frame mapping in CLVC experiments with the same DNN.

1) MCD calculation for auto-encoder generated output: The auto-encoded features generated through the network in training stage 1 are evaluated under objective tests. Multiple auto-encoder experiments were carried out to understand the optimal number of bottleneck nodes that are suitable for the task. We thus compute the MCD scores for each of them. Figure 5 depicts the MCD scores calculated between the generated and natural features. A simplified representation for each of feature categories being compared is presented in the Figure. We indicate generated Chinese MGCs as GC and natural Chinese as NC. Similarly, GE for generated English and NE for Natural English. The scores in each category are calculated on 100 frames from each of Chinese and English for the objective evaluation.

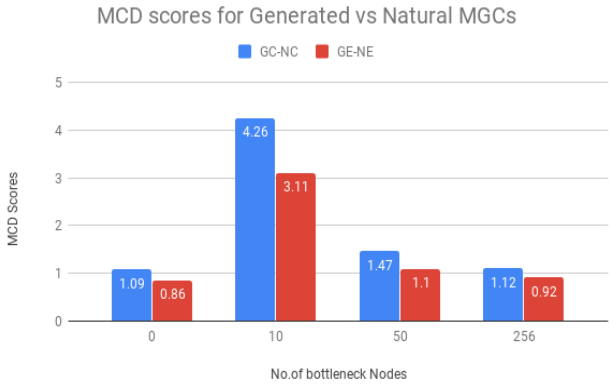

Fig. 5. Objective Evaluation over auto-encoded features.

We obtain unusually lower MCD scores in our proposed method because of multiple reasons. Firstly, our model is an auto-encoder where the input and the output nodes are fed with the same speaker information. Secondly, the model is a 3 layered DNN trained on large amounts of data. Thirdly, the auto-encoder is trained on data from both the speakers. As a result, the model has learnt the common feature representation of the speakers and their languages. From the MCD scores, we observe that the generated MGCs of Chinese and English are almost equidistant from each of natural Chinese and English MGCs respectively.

B. Subjective Evaluation

The reliability of TTS and VC systems is highly dependent on the human evaluation of generated speech. Thus, we conduct a subjective evaluation of the proposed and baseline systems. The systems tested were 1) baseline: female-to-female conversion, 2) baseline: male-to-male conversion, 3) proposed method: parallel conversion from validation data, 4) proposed method: non-parallel conversion, and 5) proposed method: cross-lingual conversion. We ensured that the listeners involved in the evaluation process belonged to one or more of these categories: speech experts, non-speech experts, native speakers. All the subjects participated in the listening tests were students. 10 subjects were invited for this task and were

provided with 5 examples from each of the systems built. Thus, a total of 250 subjective scores were collected for the overall systems built.

We conduct two different tests under subjective evaluation: Mean Opinion Scores (MOS) test and Speaker Similarity test. Under MOS tests, the samples have to be rated on a 5-point scale such that 1-poor, 2-fair, 3-good, 4-very good, and 5-best. The ratings have to be provided based on the speech quality and naturalness upon hearing the samples. Under speaker similarity test, the subjects rate the samples as per how close they perceive the converted speaker's voice to that of the target (target voice). The rating is again on a 5-point scale with 1-completely different, 2-a bit different, 3-not so close, 4-close, and 5-very similar. The results of the MOS and similarity tests for baseline are provided in Table III. The subjective evaluation results for the proposed framework with 95% Confidence intervals are presented in Figure 6. Speech samples for CLVC experiments can be found at [3]

TABLE III
SUBJECTIVE SCORES FOR THE BASELINE METHOD

System	MOS	Similarity
Female-Female	3.76	3.70
Male-Male	3.58	3.55

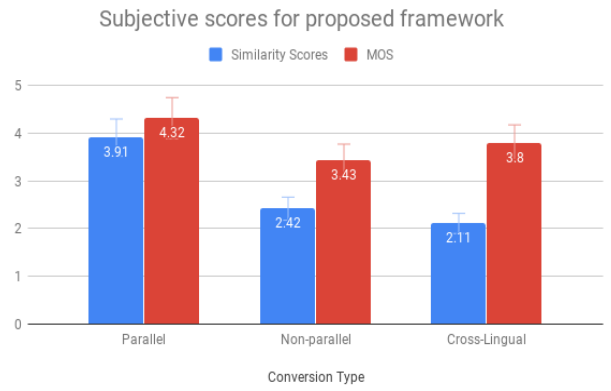

Fig. 6. Subjective scores of speaker similarity, naturalness and speech quality evaluated for the proposed framework.

Both the subjective scores indicate that our proposed framework can render a good quality speech. It is also evident that the system is capable of performing a speaker conversion irrespective of languages. Thus, one of the approaches to dealing with unaligned corpus in VC has been discussed.

C. Observations

From our experiments, we observe that the self-normalizing nature of SELU units has benefited us with better feature predictions compared to commonly preferred RELUs. We also notice that the speech quality is highly acceptable even though the model was trained on minimum data (compared to TTS data). The obvious reasons are 1) use of auto-encoder in stage 1, 2) models in stages 1 and 2 worked as error reduction networks. Hence, we assert that an error reduction network not only improves the speech quality but also addresses the issues associated with unaligned data sets.

[3]https://saisirishar.github.io/Cross-Lingual_Voice_Conversion/

VI. SUMMARY AND CONCLUSIONS

In this work, we exploit two different variants of Neural Networks to bridge the gap between the speakers and languages in CLVC. The auto-encoder used in the proposed framework alleviates the challenges of a CLVC. The challenges are addressed through i) source similarity from two monolingual speakers speech and ii) parallel corpus generation. An auto-encoder is utilized as a cepstrum generator by training the network with both source and target's speech in training stage 1. The generated cepstrum has a common representation for both the speakers and their languages. The reconstructed speech in one language is therefore treated as a source for parallel conversion in training stage 2. Finally, the conversion for cross-language is performed in the target language during the conversion stage. We also perform a non-parallel voice conversion with the proposed framework and report that the system is certainly capable of speaker conversion.

VII. PLAN FOR FUTURE WORK

We further extended our experiments to Indian languages namely, Telugu, Hindi, and Marathi. The languages we chose differ in their family origins. We found that the conversion between the languages Telugu (target language) and Chinese (target speaker) had the speaker similarity to some extent with that of the target speaker. However, the experiments with Hindi and Marathi as the target languages did not have much speaker similarity. This could be because of various factors like nature of languages, their recording equipment, speakers itself and many more. Thus, there is a wide scope to explore different languages and speakers in the future. We also intend to improve the performance of the existing framework with the use of variational auto-encoders in place of auto-encoders and leverage the variational inference in the cepstrum generation.

VIII. ACKNOWLEDGMENTS

We would like to thank Minghui Dong for providing the Blizzard 2010 Chinese dataset for our experiments. Authors would like to thank Berrak Sisman and Mingyang Zhang for their valuable discussions and suggestions. The first author was partially funded by the grants R-263-000-C35-731 and R-263-000-C35-133 under the Voice Morphing project at the National University of Singapore for this work. We would like to thank all the listeners for their participation in the subjective evaluation.

REFERENCES

[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, 2016.

[2] Y. Wang, R. 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*, 2017.

[3] K. C. Raghavi, S. K. Rallabandi, S. Sitaram, and A. W. Black, "Speech synthesis for mixed-language navigation instructions," in *Proc. INTERSPEECH*, 2017.

[4] R. Heredia and J. Altarriba, "Bilingual language mixing: Why do bilinguals code-switch?" *Current Directions in Psychological Science - CURR DIRECTIONS PSYCHOL SCI*, pp. 164–168, 2001.

[5] B. Ramani, M. P. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "Cross-lingual voice conversion-based polyglot speech synthesizer for Indian languages." in *Proc. INTERSPEECH*, 2014.

[6] B. Ramani, M. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "Voice conversion-based multilingual to polyglot speech synthesizer for Indian languages," in *Proc. TENCON IEEE Region 10 Conference (31194)*, 2013.

[7] B. Ramani, M. P. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "A multi-level GMM-based cross-lingual voice conversion using language-specific mixture weights for polyglot synthesis," *CSSP*, 2016.

[8] P. Vijayalakshmi, B. Ramani, M. P. Jeeva, and T. Nagarajan, "A multilingual to polyglot speech synthesizer for indian languages using a voice-converted polyglot speech corpus," *CSSP*, 2017.

[9] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an hmm-based speaker adaptable synthesizer," *Speech Communication*, 2006.

[10] L. Badino, C. Barolo, and S. Quazza, "Language independent phoneme mapping for foreign TTS," in *SSW*, 2004.

[11] Y. Qian, H. Liang, and F. K. Soong, "A Cross-Language State Sharing and Mapping Approach to Bilingual (Mandarin-English) TTS," *IEEE Transactions on Audio, Speech & Language Processing*, 2009.

[12] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on GMM and straight," in *Proc. INTERSPEECH*, 2001.

[13] M. Abe, K. Shikano, and H. Kuwabara, "Statistical analysis of bilingual speakers speech for cross-language voice conversion," *The Journal of the Acoustical Society of America*, 1991.

[14] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.

[15] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation." in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[16] M. Mashimo, T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion using bilingual and non-bilingual databases," in *Proc. Seventh International Conference on Spoken Language Processing*, 2002.

[17] M. Abe, K. Shikano, and H. Kuwabara, "Cross-language voice conversion," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1990.

[18] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, Cross-Lingual TTS using Phonetic Posteriorgrams," in *Proc. INTERSPEECH*, 2016.

[19] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN approach to cross-lingual TTS," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*, 2016.

[20] L. Sun, K. Li, H. Wang, S. Kang, and H. M. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. IEEE International Conference on Multimedia and Expo, ICME*, 2016.

[21] F.-L. Xie, F. K. Soong, and H. Li, "A KL Divergence and DNN-Based Approach to Voice Conversion without Parallel Training Sentences," in *Proc. INTERSPEECH*, 2016.

[22] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and helmholtz free energy," in *Proc. Advances in Neural Information Processing Systems 6*, 1994.

[23] Y. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The NU non-parallel voice conversion system for the voice conversion challenge 2018," in *Proc. Odyssey*, 2018.

[24] M. MORISE, F. YOKOMORI, and K. OZAWA, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, 2016.

[25] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[26] Desai Srinivas , W. Black Alan, Yegnanarayana B and Prahallad Kishore, "Spectral Mapping using Artificial Neural Networks for Voice Conversion," *IEEE Transactions on Audio, Speech, and Language Processing,*, 2010.

[27] L. Sun, S. Kang, K. Li, and H. M. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," *In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[28] Xie Feng-Long, Qian Yuang, Fan Y, Soong Frank and Li Haifeng, "Sequence Error (SE) minimization training of neural network for voice conversion," *in Proc. INTERSPEECH*, 2014.

[29] C. E. Siong, W. Zhizheng, and L. Haizhou, "Conditional restricted boltzmann machine for voice conversion," in *Proc. International Conference on Signal and Information Processing (ChinaSIP),*, 2013.

[30] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, 1990.

[31] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004.

[32] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-Normalizing Neural Networks," in *Proc. Advances in Neural Information Processing Systems 30*, 2017.

[33] H. G. E. Rumelhart David E. and W. R. J., "Learning Representations by Back-propagating Errors," *Nature: International Journal of Science*, 1988.

[34] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, 1951.

[35] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, 1952.

[36] S. King and V. Karaiskos, "The Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, 2010.

[37] Z. Z. Dong Wang, Xuewei Zhang, "THCHS-30 : A Free Chinese Speech Corpus," 2015. [Online]. Available: http://arxiv.org/abs/1512.01882

[38] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," in *Proc. Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, 2007.

[39] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993.