Interspeech 2018
2-6 September 2018, Hyderabad

Unsupervised Vocal Tract Length Warped Posterior Features for Non-Parallel Voice Conversion

Nirmesh J. Shah[1], *Maulik C. Madhavi*[2] *and Hemant A. Patil*[1]

[1]Speech Research Lab, DA-IICT, Gandhinagar-382007, India
[2]Electrical and Computer Engineering Department, National University of Singapore, Singapore
{nirmesh88_shah and hemant_patil}@daiict.ac.in,elemaul@nus.edu.sg

Abstract

In the non-parallel Voice Conversion (VC) with the **I**terative combination of **N**earest Neighbor search step and **C**onversion step **A**lignment (INCA) algorithm, the occurrence of one-to-many and many-to-one pairs in the training data will deteriorate the performance of the stand-alone VC system. The work on handling these pairs during the training is less explored. In this paper, we establish the relationship via intermediate speaker-independent posteriorgram representation, instead of directly mapping the source spectrum to the target spectrum. To that effect, a Deep Neural Network (DNN) is used to map the source spectrum to posteriorgram representation and another DNN is used to map this posteriorgram representation to the target speaker's spectrum. In this paper, we propose to use unsupervised Vocal Tract Length Normalization (VTLN)-based warped Gaussian posteriorgram features as the speaker-independent representations. We performed experiments on a small subset of publicly available Voice Conversion Challenge (VCC) 2016 database. We obtain the lower Mel Cepstral Distortion (MCD) values with the proposed approach compared to the baseline as well as the supervised phonetic posteriorgram feature-based speaker-independent representations. Furthermore, subjective evaluation gave relative improvement of 13.3 % with the proposed approach in terms of Speaker Similarity (SS).

Index Terms: Vocal Tract Length Normalization, Posteriorgram, Deep Neural Network, Voice Conversion.

1. Introduction

Voice Conversion (VC) is a technique that modifies the speech signal uttered by a source speaker in such a way that it perceives as if it was uttered by a particular target speaker [1]. The spectral features from both the source and target speakers need to be aligned during training, due to the speaking rate variation across the speakers (i.e., interspeaker variations) and speech rate variations within the speaker (i.e., intraspeaker variations), in order to apply stand-alone VC techniques [1]. Dynamic Time Warping (DTW) [1] and the unsupervised **I**terative combination of **N**earest Neighbor search step and **C**onversion step **A**lignment (INCA) [2,3] algorithms are popular, in the area of parallel (i.e., both speakers have spoken same utterances) and non-parallel (i.e., both speakers have spoken different utterances) VC, respectively. Both the alignment techniques will generate one-to-many and many-to-one feature pairs as studied in [2, 4]. In addition, if the word spoken by a source speaker is repeated by the target speaker with different variations, it will generate such pairs. Furthermore, if the same word is repeated for several times, it will result in the different speech pattern. This also generates such kind of pairs. Directly learning the relationship in the presence of such pairs is very challenging. These one-to-many and many-to-one pairs will affect the learning of the mapping function and results in the muffling and oversmoothing effect in VC [5].

The earlier approaches used context-dependent information to overcome this issue [5]. Recently, equalizing formant locations using Dynamic Frequency Warping (DFW) was proposed to tackle these issue [6]. In addition, some of the approaches proposed to filter out such pairs from the training [7, 8]. However, loosing number of pairs will not be useful in the case where the amount of training data is small. There is also an attempt in past to use pre-stored speakers parallel data to train initial model in the case of non-parallel [9]. Furthermore, adaptation [10, 11] and the model-based [12–15] approaches have been proposed to avoid alignment step, which also help to solve one-to-many pairs-related issues. Recently, Phonetic Posteriorgram (PPG) (which are believed to be speaker-independent representations) have been proposed that consider two-stage mapping [16, 17]. However, it requires a huge amount of labeled speech data to train the Automatic Speech Recognition (ASR) systems for estimating PPG. Since the training data is small in most of the applications of non-parallel VC. Hence, we propose to exploit unsupervised technique for computing speaker-independent posterior features.

We propose to avoid alignment step by using the two separate Deep Neural Networks (DNNs) where one DNN will map source speaker's spectral features to the speaker-independent representations and the another DNN will map these speaker-independent representations to the particular target speaker's spectral features. Here, we propose to use unsupervised Vocal Tract Length Normalization (VTLN)-based warped Gaussian Posteriorgram (GP) features as the speaker-independent representations. Earlier stand-alone VTLN and frequency warping-based techniques were used in the VC [18, 19]. We performed experiments on a small subset of publicly available first Voice Conversion Challenge (VCC) 2016 database [20]. We found that in non-parallel scenarios, our proposed approach is working better or comparable in terms of the objective and subjective evaluations for the developed VC systems.

2. Speaker-Independent Posterior Representations

Earlier Gaussian Mixture Model (GMM) posteriorgram and PPG were used as features for phoneme classification and template matching-based ASR [21–24]. In this paper, three different types of speaker-independent posterior features have been considered, namely, phonetic posteriorgram, GMM posteriorgram, and Vocal Tract Length (VTL)-warped posteriorgram.

10.21437/Interspeech.2018-1712

2.1. Phone Posteriorgram (PPG)

PPG contains the posterior probability for each phonetic class obtained for a given speech signal [16]. In this paper, we used Brno University of Technology (BUT) phoneme recognizer tool, which is Split-Temporal Context (STC) neural network-based phoneme recognition system [25]. In particular, BUT system trained on the English data from TIMIT corpus have been used to extract PPG. The trained speech recognizers models may not be available for all the languages. Hence, it is necessary to develop the unsupervised speaker-independent posteriorgram.

2.2. Gaussian Posteriorgram (GP)

Gaussian posteriorgram has been extensively used for Query-by-Example Spoken Term Detection (QbE-STD) task [26]. In particular, the problem of detecting the presence of query within the spoken utterance is known as QbE-STD [27, 28]. The posterior probability $P(C_k|\mathbf{o}_t)$ of the current frame $\mathbf{o}_t$ (for k^{th} cluster C_k, and t^{th} feature vector) of GP can be computed as follows:

$$P(C_k|\mathbf{o}_t) = \frac{\omega^k_{init}\mathcal{N}(\mathbf{o}_t;\mu^k_{init},\Sigma^k_{init})}{\sum_{j=1}^{N_p}\omega^j_{init}\mathcal{N}(\mathbf{o}_t;\mu^j_{init},\Sigma^j_{init})}, \quad (1)$$

where N_p is the number of GMM components, ω^k_{init}, μ^k_{init} and Σ^k_{init} are the weights, mean vectors and covariance matrices, respectively for each k^{th} Gaussian components ($1 \leq k \leq N_p$). The GMM parameters are estimated using Expectation-Maximization (EM) algorithm [29]. The GMM parameters are trained from large number of speakers and the procedure for training the GMM is described in the following sub-section 2.3, step 2.

2.3. Proposed VTL-Warped GP

It has been observed that the formants for uniform vocal tract are *inversely* proportional to the length of vocal tract [30]. Thus, VTL variations is a well-known contributing factor to speaker-related spectral variability for having different speaker characteristics. To obtain the speaker-independent posterior features, we applied VTLN technique to achieve speaker-independent VTL-warped posterior features. The conventional approaches for VTLN warping factor estimation require an acoustic model and phonetic transcription. The maximum likelihood search is performed to obtain suitable speaker-specific VTLN warping factor [31, 32]. The procedure of VTL-warped Gaussian posteriorgram is as follows [28], [33]:

1. *Feature Extraction*: Compute warped feature vector sequence, i.e., $X^\alpha := \{x^\alpha_1, x^\alpha_2, \ldots x^\alpha_T\}$ that carry information from the different warping factors, namely, $\alpha = 0.88, 0.90, \ldots; 1.12$. In this paper, we used Perceptual Linear Prediction (PLP) cepstral feature vectors [34]. Human VTL varies from nearly 13 cm (for adult female) to 18 cm (for adult male) [31]. Due to this length variations, formant frequencies can deviate by 25 % among different speakers. To incorporate this deviation, the VTLN warping factors are considered in the range from 0.88 to 1.12 [31].
2. *Initial speaker-independent GMM Training*:
   The unwarped feature vectors are pulled to model the gender-independent characteristics for acoustic features. The unwarped features, i.e., the feature vector with VTLN warping factor $\alpha = 1$, are taken from the large number of speakers and thus, can be assumed to have speaker-independent characteristics. Let this trained gender-independent GMM be λ having the weight parameters ω_i, mean vectors μ_i, Covariance matrix Σ_i for i^{th} components of GMM.
   $$\lambda_{init} = \{\omega^i_{init}, \mu^i_{init}, \Sigma^i_{init}\}_{i=1}^{N_c}, \quad (2)$$
   where N_c is the number of components. In a practical scenario, covariance matrix Σ_i is considered to have only diagonal elements for computational simplicity. We used this speaker-independent GMM to compute the Gaussian posteriorgram as described in sub-Section 2.2.
3. *VTLN warping factor estimation*: The formants shifting w.r.t. speaker-independent GMM can be captured by VTLN warping factor. In order to estimate the VTLN warping factor, we follow the approach as suggested in our previous work [28], [33]. VTLN warping factor is estimated from the sets of different warped feature vector sequences $\mathbf{X}^\alpha$, with different values of α. The likelihood values of $\mathbf{X}^\alpha$ are computed against the speaker-independent GMM, λ_{init} and MLE criteria is considered to estimate VTLN warping factor.
   $$\hat{\alpha} = \underset{0.88\leq\alpha\leq1.12}{\arg\max} \; P(\mathbf{X}^\alpha|\lambda_{init}). \quad (3)$$
4. *Retraining of GMM*: The VTL warped features can be combined together from large number of speakers to train GMM. The objective is to achieve further speaker-independent model by utilizing the acoustic features after spectral scaling compensation. GMM is further retrained with warped features, X^α. This new model $\lambda_r \sim (\mu_r, \Sigma_r, \omega_r)$ is different from the earlier GMM model, λ_{init} and expected to have more speaker-independent characteristics.
5. Set $\lambda_{init} = \lambda_r$. Run steps 3 to steps 5 for five times [28], [33].
6. *Computation of Posteriorgram*: Now based on the estimated warping factors and trained GMM, Gaussian posteriorgrams are computed. Thus, the Eq. (1) of Gaussian posteriorgram is modified as follows by considering new speaker-independent model λ_r:
   $$P(C_k|\mathbf{o}_t) = \frac{\omega^k_r\mathcal{N}(\mathbf{o}_t;\mu^k_r,\Sigma^k_r)}{\sum_{j=1}^{N_p}\omega^j_r\mathcal{N}(\mathbf{o}_t;\mu^j_r,\Sigma^j_r)}. \quad (4)$$

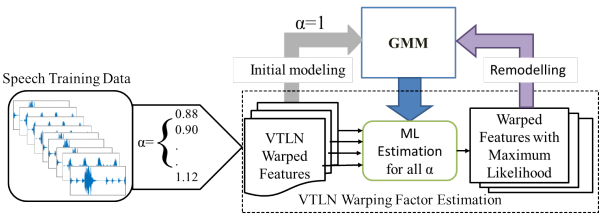

Figure 1: *Schematic block diagram of extracting VTL-warped posterior features. After [28], [33].*

Figure 1 shows the block diagram of warped Gaussian posteriorgram computation using VTLN warping factor and speaker-independent GMM. In particular, as described VTL warped features are extracted with different warping factor α. Initial GMM model is created as Speaker-independent GMM using unwarped acoustic features. Speaker-independent GMM is used to estimate VTLN warping factor for each utterance using MLE criteria. Using estimated VTLN warping factor, VTLN

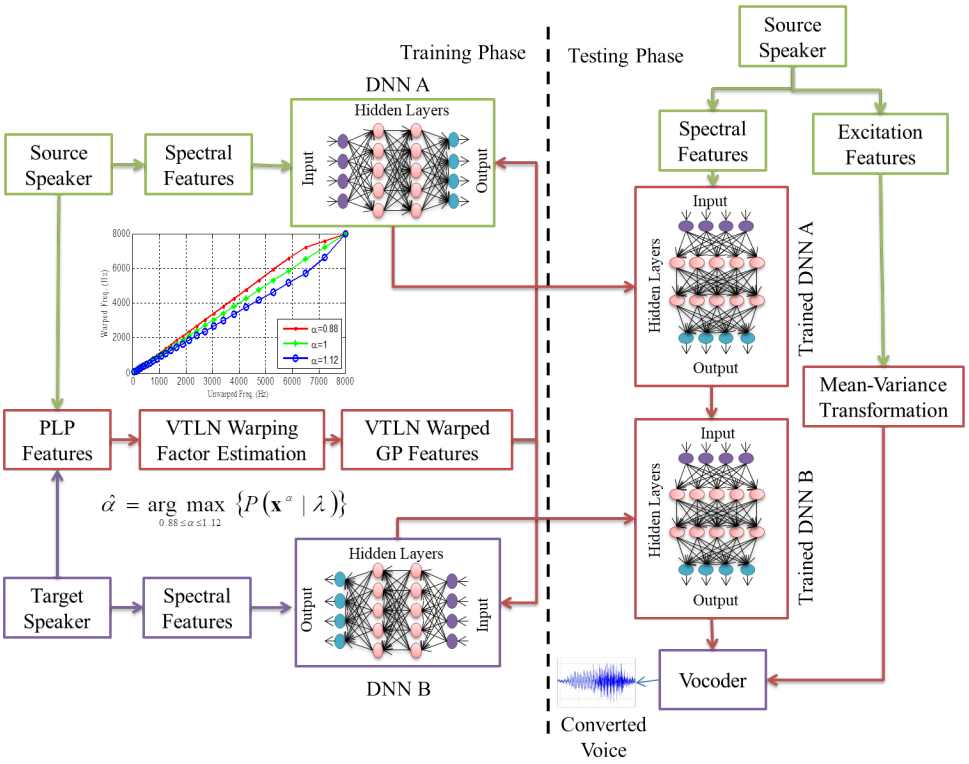

Figure 2: *The system architecture for proposed VTL-warped GP for non-parallel VC.*

warped acoustic features are used to retrain the GMM parameter to characterize more speaker invariance property in GMM. This new speaker-independent GMM is further used to estimate VTLN warping factor and warped acoustic features are again used for re-training. This procedure is iteratively executed for five iterations (as suggested in [28], [33]). Speaker-pairs data from the VCC 2016 database is used to train speaker-independent GMM for the VTLN posteriorgram.

3. DNN-based VC System Architecture

The relation between the spectral feature vectors $\mathbf{X}$, and $\mathbf{Y}$ are obtained using the DNN, which consists of $k > 2$ multiple layers, where k is the total number of layers. The first, last, and the middle layers of the DNN are called as an input, output, and the hidden layers, respectively [35]. Here, each layer performs either nonlinear or linear transformation. The transformation at i^{th} layer is given by [36]:

$$\mathbf{h}_{i+1} = f(\mathbf{W}_i^T \mathbf{h}_i + \mathbf{b}_i), \tag{5}$$

where $\mathbf{h}_i$, $\mathbf{h}_{i+1}$, $\mathbf{W}_i$, $\mathbf{b}_i$ are called as input, output, weights and bias of i^{th} layers, and f is an activation function that is either nonlinear (such as, tanh, sigmoid, ReLU) or linear. $\mathbf{h}_1 = \mathbf{X}$ and $\mathbf{h}_{K+1} = \mathbf{Y}$ are the input, and output layers of DNN. Due to higher number of layers, DNN captures more complex relationship between the source and target speakers' spectral features. The adaptive moment estimation-based optimization, i.e., Adam optimization is used to train the weights and biases of the DNN such that Mean Squared Error (MSE), i.e., $E = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ is minimized, where $\hat{\mathbf{Y}}$ is the predicted output. For baseline system, we use the single DNN trained with the aligned pairs obtained using the INCA algorithm.

Figure 2 shows the architecture of the proposed VC system for the non-parallel case. Here, the VTLN warping factor is estimated first for a given speaker-pair. This warping factor is then used to estimate the VTL-warped GP. One DNN (i.e., DNN-A) is trained by taking source speaker's spectral features as an input and the VTL-warped posteriorgram features as an output. The other DNN (i.e., DNN-B) is trained by considering target speaker's VTL-warped posteriorgram features and spectral features as an input and output, respectively. During the testing phase, the source speaker's spectral features are first converted using DNN-A and then this predicted VTL-warped posteriorgram is given as an input to the DNN-B. Converted spectral features predicted by DNN-B is then applied to the vocoder and converted into the speech signal. The proposed framework does not need the aligned source and target speakers' features. Thus, the proposed framework behaves exactly the same way for parallel VC case.

4. Experimental Results

4.1. Experimental Setup

In this paper, a subset of the VCC 2016 database, which contains speakers SF1, SM1, TF1, and TM1 has been used to build VC systems [20]. Out of 162 training utterances given for each speaker in the VCC 2016 database, we have selected 40 non-parallel utterances from the source and target speakers. In particular, utterance 1 to 40 are selected from the source speakers. On the other hand, 41 to 80 utterances are taken from the target speaker. Thus, both the set have non-overlapping utterances. *25*-D Mel Cepstral Coefficients (MCCs) (including the 0^{th} coefficient) and *39*-D PLP features (including Δ and $\Delta\Delta$) as suggested in [28], [33]. *1*-D F_0 for each frame (with 25 ms frame duration, and *5* ms frame shift) have been extracted. *120*-D phonetic posterior features have been extracted using BUT decoder. Similarly, *120*-D GMM and VTL-warped posteriorgram is extracted in order to compare its performance w.r.t. the *120*-D PPG. The number of mixture components were ini-

tially set to 128 with Vector Quantization (VQ) initialization. To obtain *120*-D posterior features, we perform the iterative approach to merge the two closest centroids till we obtain 120 centroids. We have built in total 16 non-parallel VC systems for all the four speaker-pairs. In particular, we developed proposed system with phonetic, GMM and VTL-warped posteriorgram-based features for all the four speaker-pairs. We used three hidden layered DNN-A with $25, 512, 512, 512, 120$ number of neurons in each input, hidden and the output layers, respectively. Similarly, DNN-B contains the $120, 512, 512, 512, 25$ number of neurons in each layers, respectively. Baseline system contains $25, 512, 512, 512, 25$ number of neurons in each layer, respectively. We used Rectifier Linear Unit (ReLU) as nonlinear activation function and Adam optimization with the $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to train the network [37]. Mean-variance (MV) transformation is used for the F_0 transformation [38]. The AHOCODER is used for the analysis-synthesis [39].

4.2. Objective Evaluation

The state-of-the-art Mel Cepstral Distortion (MCD) is used for objective evaluation [38]. It can be observed that proposed VTLN posteriorgram-based system is performing better than the baseline system in all the VC systems. In particular, on an average VTLN posteriorgram-based VC system is performing better than the GMM, and phonetic posteriorgram-based VC system. The higher value of the MCD for phonetic posteriorgram is possibly due to the fact that the BUT speech recognizer did not use the training data from VCC 2016 database. In addition, development of speech recognizer from 40 utterances of VCC is difficult. Hence, the development of unsupervised speaker-independent posterior features (i.e., proposed VTL-warped posterior) is indeed helpful in the VC task where the less number of training utterances are available from the target speaker in the most practical scenarios.

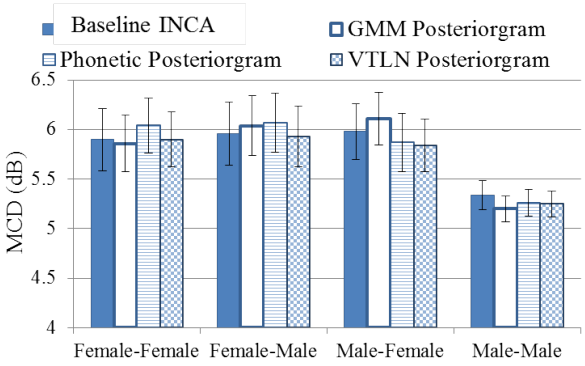

Figure 3: *The MCD analysis of the VC systems with 95 % confidence interval.*

4.3. Subjective Evaluation

Mean Opinion Score (MOS) test have been performed for evaluating the speech quality and the Speaker Similarity (SS) of converted voice. The MOS of total 192 samples from the 12 non-native English subjects (3 females and 9 males with no known hearing impairments, and with the age variations between 19 to 29 years) were taken. In the MOS test, subjects were asked to evaluate the randomly played utterances for the speech quality and SS. The subjects were asked to rate the converted voices on the scale of 1 (i.e., very bad) to 5 (i.e., very good) for speech quality (i.e., naturalness). On the other hand, subjects were asked to rate the converted voice in terms of SS on the scale of 1 (not at similar) to 5 (exactly similar) w.r.t. the target speaker. The result of the MOS test is shown in the Figure 4 with 95 % confidence interval. We observed 13.3 % relative improvement over the baseline system in MOS test for SS. On the other hand, there is a relative reduction by 3.3 % in MOS for speech quality. It is possibly due to the fact that the DNN trained with aligned pairs obtained using nearest neighbor-based technique is trying to learn the mapping function, which is nearer to the identity mapping. This leads to the issue where the converted spectrum is nearer to the source spectrum than the actual target spectrum. Hence, the quality is more preserved than the SS in the converted voice.

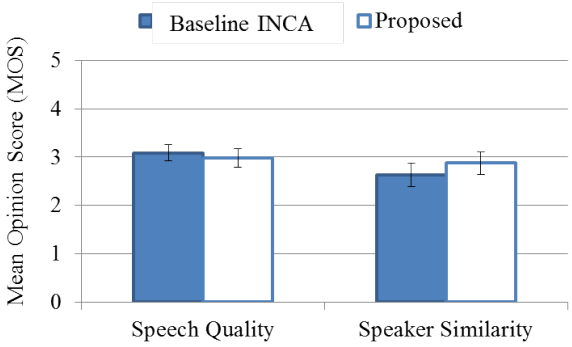

Figure 4: *The MOS analysis for the developed VC systems with 95 % confidence interval.*

5. Summary and Conclusions

In this study, we presented the use of unsupervised VTL-warped Gaussian posteriorgram representation to establish the mapping between the source and the target speakers' spectrum. The direct alignment procedure between the source and target spectrum results into many-to-one or one-to-many correspondences, which deteriorate the performance of the VC system. The proposed VTL-warped Gaussian posteriorgram representation is mapped from and to the source and target speaker spectral features, respectively, using two different DNNs trained with each speakers data. The speaker-independent feature set is obtained by training a GMM with warped vocal tract length (VTL) cepstral features extracted from the data of the pair speakers. The warping factor is found iteratively with GMM training such that maximum likelihood is obtained under a discrete set of warped values and given the speaker-pair non-parallel acoustic data. The final feature set is the Gaussian posteriograms computed from the resulting GMM. Proposed VTL-warped Gaussian posteriorgram gave lower MCD scores and higher speaker similarity for non-parallel VC systems. The improvement in SS might be due to the use of VTLN between source and target. In future, we plan to explore the possible use of additional mapping in order to avoid the mismatch between source and target posteriorgram representation.

6. Acknowledgments

Authors would like to thank the authorities of DA-IICT, Gandhinagar and MeitY, Govt. of India for their kind support to carryout this work.

7. References

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, no. 3, pp. 65–82, 2017.

[2] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech and Lang. Process.*, vol. 18, no. 5, pp. 944–953, 2010.

[3] N. J. Shah and H. A. Patil, "On the convergence of INCA algorithm," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, Kuala Lumpur, Malaysia, 2017, pp. 559–562.

[4] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," in *INTERSPEECH*, Brisbane, Australia, 2008, pp. 1453–1456.

[5] E. Godoy, O. Rosec, and T. Chonavel, "Alleviating the one-to-many mapping problem in voice conversion with context-dependent modelling," in *INTERSPEECH*, Brighton, United Kingdom, 2009, pp. 1627–1630.

[6] S. H. Mohammadi, "Reducing one-to-many problem in voice conversion by equalizing the formant locations using dynamic frequency warping," *arXiv preprint arXiv:1510.04205*, 2015.

[7] O. Turk and L. M. Arslan, "Robust processing techniques for voice conversion," *Computer Speech & Language*, vol. 20, no. 4, pp. 441–467, 2006.

[8] S. V. Rao, N. J. Shah, and H. A. Patil, "Novel pre-processing using outlier removal in voice conversion," in 9^{th} *ISCA Speech Synthesis Workshop (SSW)*, Sunnyvale, CA, USA, 2016, pp. 147–152.

[9] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Non-parallel training for many-to-many eigenvoice conversion," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010, pp. 4822–4825.

[10] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using *i*-vector PLDA: Towards unifying speaker verification and transformation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5535–5539.

[11] C. H. Lee and C. H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *INTERSPEECH*, Pittsburgh, USA, 2006, pp. 2254–2257.

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

[13] 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 *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3364–3368.

[14] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial network," to appear *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018.

[15] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversariel network," to appear *in International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018.

[16] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *IEEE International Conference on Multimedia and Expo (ICME)*, Seattle, USA, 2016, pp. 1–6.

[17] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1268–1272.

[18] N. J. Shah and H. A. Patil, "Novel amplitude scaling method for bilinear frequency warping based voice conversion," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5520–5524.

[19] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, U.S., 2003, pp. 676–681.

[20] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *INTERSPEECH*, San Fransisco, USA, 2016, pp. 1–5.

[21] J. Pinto and H. Hermansky, "Combining evidence from a generative and a discriminative model in phoneme recognition," in *INTERSPEECH*, Brisbane, Australia, 2008, pp. 2414–2417.

[22] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition," in *INTERSPEECH-ICSLP*, Pittsburgh, USA, 2006, pp. 2570–2573.

[23] J. Pinto, G. S. V. S. Sivaram, M. Magimai-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP-based hierarchical phoneme posterior probability estimator," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 2, pp. 225–241, 2011.

[24] S. Thomas, S. Ganapathy, A. Jansen, and H. Hermansky, "Data-driven posterior features for low resource speech recognition applications," in *INTERSPEECH 2012*, Portland, Oregon, USA, 2012, pp. 791–794.

[25] P. Schwarz, P. Matějka, and J. Černocký, *Towards lower error rates in phoneme recognition*. Berlin, Heidelberg: Springer, 2004, pp. 465–472.

[26] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Merano, Italy, 2009, pp. 398–403.

[27] A. Mandal, K. P. Kumar, and P. Mitra, "Recent developments in spoken term detection: a survey," *International Journal of Speech Technology (IJST), Springer*, vol. 17, no. 2, pp. 183–198, 2014.

[28] M. Madhavi, "Design of QbE-STD system: Audio representation and matching perspective," Ph.D. Thesis, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India, 2017.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[30] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, 3^{rd} ed. Pearson Education, 2009.

[31] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. on Speech and Audio Process.*, vol. 6, no. 1, pp. 49–60, 1998.

[32] Y. Kim and J. Smith, "A speech feature based on Bark frequency warping-the non-uniform linear prediction (NLP) cepstrum," in *IEEE Workshop on Applications of Signal Process. to Audio & Acoust.*, New Paltz, NY, 1999, pp. 131–134.

[33] M. C. Madhavi and H. A. Patil, "VTLN-warped Gaussian posteriorgram for QbE-STD," in *European Signal Processing Conference (EUSIPCO)*, Kos island, Greece, 2017, pp. 593–597.

[34] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the J. of the Acoust. Soc. of Amer. (JASA)*, vol. 87, no. 4, pp. 1738–1752, 1990.

[35] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, First Edition, 2016.

[36] B. Yegnanarayana, *Artificial Neural Networks*, 1^{st} ed. PHI Learning Pvt. Ltd., 2009.

[37] D. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *International Confernece on Learning Representation (ICLR)*, San Diego, USA, 2015, pp. 1–15.

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

[39] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved HNM-based vocoder for statistical synthesizers," in *INTERSPEECH*, Florence, Italy, 2011, pp. 1809–1812.