Novel Inter Mixture Weighted GMM Posteriorgram for DNN and GAN-based Voice Conversion

Nirmesh J. Shah, Sreeraj R., Neil Shah and Hemant A. Patil
Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India
E-mail: {nirmesh88_shah,sreeraj_r,neil_shah,hemant_patil}@daiict.ac.in

***Abstract*—Voice Conversion (VC) requires an alignment of the spectral features before learning the mapping function, due to the speaking rate variations across the source and target speakers. To address this issue, the idea of training two parallel networks with the use of speaker-independent representation was proposed. In this paper, we explore the unsupervised Gaussian Mixture Model (GMM) posteriorgram as a speaker-independent representation. However, in the GMM posteriorgram, the same phonetic information gets spread across more than one component due to the speaking style variations across the speakers. In particular, this spread is limited to a group of neighboring components for a given phone. We propose to share the posterior probability of each component with the limited number of neighboring components that are sorted based on the Kullback-Leibler (KL) divergence. We propose to employ a Deep Neural Network (DNN) and a Generative Adversarial Network (GAN)-based framework to measure the effectiveness of the proposed Inter Mixture Weighted GMM (IMW GMM) posteriorgram on the Voice Conversion Challenge (VCC) 2016 database. The relative improvement of 13.73 %, and 5.25 % is obtained with the proposed IMW GMM posteriorgram w.r.t. the GMM posteriorgram for the speech quality and the speaker similarity of the converted voices, respectively.**

Index Terms: IMW GMM Posteriorgram, generative adversarial network, voice conversion.

I. Introduction

Voice Conversion (VC) is a technique that maps perceived speaker identity presents in the speech signal uttered by a source speaker to a particular target speaker without changing the message content of the signal [1], [2]. Conventional approaches in the VC, require the aligned spectral features from both the source and the target speakers, due to the speaking rate variations across the speakers (i.e., interspeaker variations) and speech rate variations within the speaker (i.e., intraspeaker variations). In both parallel and non-parallel VC tasks, it has been shown that the accuracy of the alignment between source and target speakers' data, impacts the quality of the converted voices [3]–[11]. To avoid the issues related to the alignment, several adaptation [12], [13] and the generation model-based VC techniques have been proposed [14]–[19].

Recently, Generative Adversarial Network (GAN)-based architectures have shown notable improvements in the area of, Speech Enhancement (SE) [20]–[24], VC [14], [16]–[18], [25], and cross-domain speech conversion techniques [26], [27]. Such a VC framework maps the spectral representations from the source speaker to the speaker-independent latent representations and this representation is further used to generate the more realistic target spectral features using the GAN [14]. Earlier approaches employ the variational autoencoded features, or the Phonetic Posteriorgram (PPG) obtained via developing Automatic Speech Recognition (ASR) [14], [28], [29]. However, developing the ASR requires a large amount of transcribed speech data. In addition, such ASR will be more robust if it also includes training data from both the source and the target speakers, which is difficult to obtain in realistic VC scenarios.

In this work, we propose to explore unsupervised techniques to learn the speaker-independent representations. In particular, unsupervised Gaussian Mixture Model posteriorgram (GMM-PG) are very popular speaker-independent representations in the area of Query-by-Example Spoken Term Detection (QbE-STD) [30]–[34]. However, the key issue with the conventional GMM-PG is that the same phone gets spread across more than one component (due to the speaking style variations across the speakers) [35]. In particular, we observe that this spread is limited to a group of neighboring components for a given phone. Hence, we propose Inter Mixture Weighted GMM-PG (i.e., IMW GMM-PG), that shares the posterior probability of each mixture in GMM-PG with the limited number of neighboring mixture that are *sorted* based on the Kullback-Leibler (KL) divergence. In this paper, we measure the effectiveness of the proposed unsupervised speaker-independent feature representation, namely, IMW GMM-PG over the conventional GMM-PG with the two-stage Deep Neural Network (DNN) and the GAN-based VC framework. The experiments are performed on the publicly available Voice Conversion Challenge (VCC) 2016 database. The detailed subjective and objective evaluations have also been presented for the developed VC systems.

II. Speaker-Independent Representations

A. Analysis of speaker-independent features

Posteriorgram, such as Gaussian posteriorgram, when trained on multispeaker data, are speaker-independent to a certain extent [30], [35]. The posterior probability $P(C_k|\mathbf{o}_t)$ (for k^{th} cluster C_k, and $\mathbf{o}_t$ feature frame) of GMM-PG can be computed as follows [30], [36]:

$$P(C_k|\mathbf{o}_t) = \frac{\omega_k \mathcal{N}(\mathbf{o}_t; \mu_k, \Sigma_k)}{\sum_{j=1}^{N} \omega_j \mathcal{N}(\mathbf{o}_t; \mu_j, \Sigma_j)}, \quad (1)$$

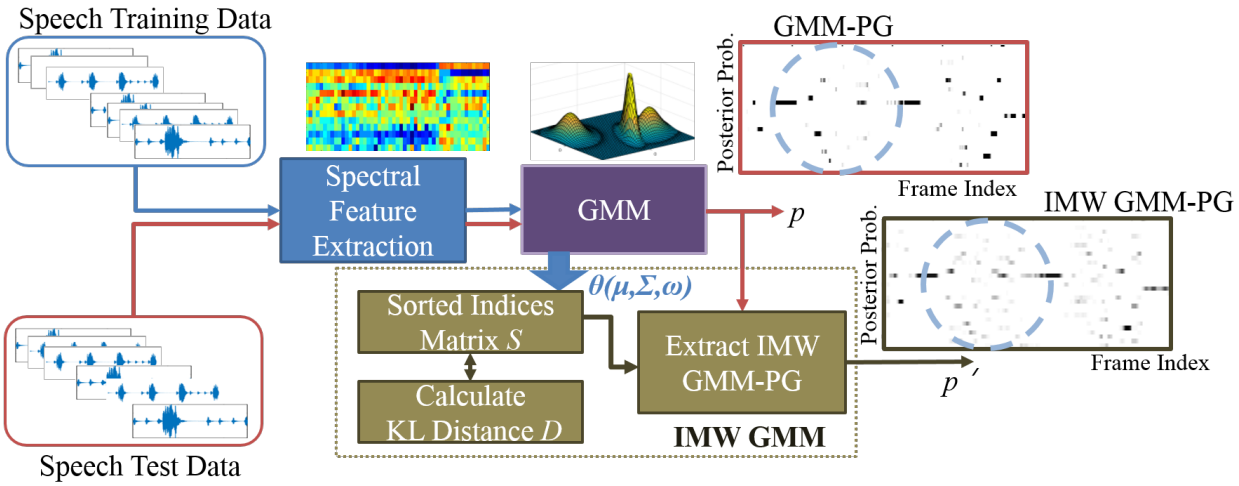

Fig. 1: Functional block diagram of proposed IMW GMM method for VC. It can be inferred from the circled regions that the probability is shared across components in IMW GMM-PG than in GMM-PG.

where N is the number of GMM components, ω_k, μ_k and Σ_k are the weights, mean vectors and covariance matrices, respectively, for the k^{th} Gaussian components ($1 \leq k \leq N$). The GMM parameters are estimated using Expectation-Maximization (EM) algorithm [37]. In this paper, entire TIMIT database is used to train the GMM. In particular, we have analyzed frame-level GMM-PG features for three randomly selected speakers from the TIMIT database. For selected phones, we present in Table I the indices of those components, that are having higher probabilities in the decreasing order from the top, for a specific speaker. An ideal posterior speaker-independent representation should contain distinct phonetic information in each component, irrespective of the speaker. However, in the case of GMM-PG, the phonetic information gets spread across the components.

From Table I, it can be observed that the components that share the frame posteriorgram values are almost the same for one user and for one phone. In case of phones, for example, 'aa ', the found component labels $46, 53, 26, 40$ are representing one particular phone, across the speakers. However, while observing the parameters (i.e., $mean$ and $variance$) of these components, we observed the distance calculated by KL-divergence is lesser w.r.t. each other. This means that the feature vectors for a phone lies closer across the speakers even though they cannot be clustered to the same component. Furthermore, it should be noted that across the different speakers (as shown in Table I), prominent (i.e., highest posterior probability) components remains similar at most of the cases. Though they are similar, order of prominence varies. This explains why the prominent component of a phone, for example, */iy/* is 25^{th} component for *male* speaker, while for *female* speaker, it is 29^{th} component. While using conventional GMM-PG matching methods, this may lead to many false recognition or increased mismatch [35]. To address this issue, we propose IMW GMM-PG, a modification to the GMM-PG, and is discussed in detail in the next sub-Section.

TABLE I: Prominent component variation in GMM posteriorgram, 64 dimensions, for selected phones and speakers from TIMIT database

Speaker	Selected Phonetic Classes									
	aa	**iy**	**ow**	**l**	**m**	**dh**	**f**	**sh**	**p**	**t**
Male1	46	25	44	51	3	54	14	1	7	15
	53	58	56	44	19	63	35	35	15	28
	26	29	51	3	23	31	1	9	9	7
Female1	46	29	37	44	3	15	13	14	4	7
	26	22	18	32	23	39	35	35	9	4
	40	25	41	20	19	63	4	4	15	9

B. Proposed Inter Mixture Weighted (IMW) GMM Posteriorgram

IMW GMM is a post-processing method obtained by sorting components based on the distance calculated among the components of GMM using KL divergence. The neighbor components of a given component, are assigned a fraction of its probability value along with their posterior probability value. This in effect helps feature being represented by a set of components than a single component. The block diagram to extract IMW GMM-PG is shown in Fig. 1.

After extracting GMM-PG for a frame, the component having the highest probability is selected and its neighbor components are identified in a sorting order. Depending on the order, the probability value is shared with all the components, in such a way that the nearest neighbor components get the highest share, while the farthest components get the lowest

share. This results in spreading the posterior probability across the neighbor components which can be seen in the circled region in Fig. 2. The details of the IMW GMM-PG extraction are given in the following Algorithm 1.

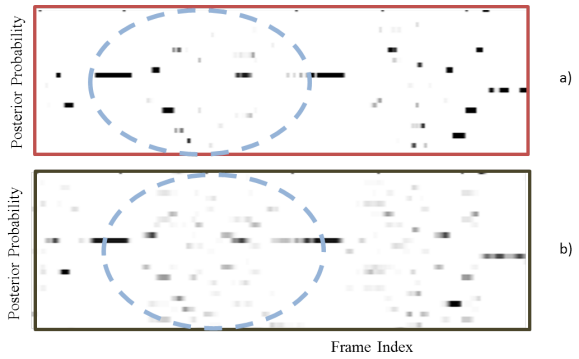

Fig. 2: Posteriorgram comparison for the query '$intelligence$', a) GMM-PG and b) IMW GMM-PG.

To visualize the effectiveness of IMW GMM-PG over GMM-PG, we calculated the distance among posterior features across different phonetic classes on entire TIMIT database. The broad phone classes considered includes vow: vowels, svw: semi vowels, nas: nasals, fri: voiced fricatives, ufr: unvoiced fricatives and plo: plosives. Fig. 3 clearly shows that in the case of obstruents (fricatives, affricates, and plosives), there is a more ambiguity with the GMM-PG features than the IMW GMM-PG features. Vowels and semivowels broad classes are easily distinguished within in case of IMW GMM-PG. However, it can also be noticed that in GMM-PG the vowel and semivowels can be very clearly distinguished from the plosive, fricatives broad phone classes, while the distinction between the broad classes decreased in case of IMW GMM-PG.

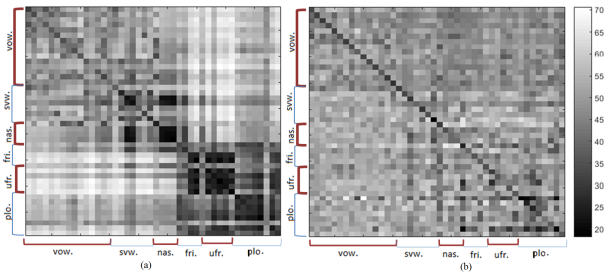

Fig. 3: Distance matrix with 44 phones for (a) GMM and (b) proposed IMW GMM approach.

III. Proposed VC System Architecture

We develop the two-stage DNN-based VC architecture. In particular, the first stage maps the cepstral features of a source speaker to the corresponding source speaker's IMW GMM-PGs using DNN and the second network maps the target speaker's IMW GMM-PGs to the corresponding cepstral features of the target speaker using DNN or GAN as shown in Fig. 4 and Fig. 5. Both the networks are trained simultaneously.

Algorithm 1 Proposed IMW GMM Algorithm

extractIMWGMMPost()

extractIMWGMMPost returns p' IMW GMM-PG which is obtained from GMM-PG p after applying IMW GMM logic.

Input: $N \leftarrow$ Number of components in GMM.
  $n \leftarrow$ Dimension of feature vector.
  $\theta \leftarrow$ Model parameters μ, σ, ω.
  $p(i) \leftarrow$ Posterior probability of a i^{th} component for a frame calculated using GMM ($1 \times N$).
$S \leftarrow$ ***IMWGMM(n, N, θ)***
$k \leftarrow \arg\max_i p(i)$, index of component with maximum probability in p.
$i \leftarrow 1$
while $i \leq N$**:**
  $M \leftarrow$ position of i in $S[k,:]$. (To find how closer i^{th} component is from k)
  $p'(i) \leftarrow p(i) + \frac{p(k)}{2^M}$
end
return $p' \leftarrow$ normalize p'

function *IMWGMM(n,*N*,*θ*)*

IMWGMM returns a matrix S which includes the indices of components sorted in ascending order of KL distance from each component.

$i \leftarrow 1$
$S \leftarrow$ zero initialized matrix of dimension $N \times N$,
  (Stores indices of nearest component $\forall$ i)
$D \leftarrow$ zero initialized array of dimension $1 \times N$
  (Stores distance array of a component.)
while i $\leq N$**:**
  $D \leftarrow$ *calcDist(i θ)*
  $i \leftarrow i+1$
  S[i,:] $\leftarrow$ Indices of D after sorting in ascending order.
end
return S

function *calcDist(i,*θ*)*

calcDist returns array d : KL distance of i^{th} component from all N components given by θ.

Input: $i \leftarrow$ Component under consideration.
for j **from 1 to** N **:**
  $P \leftarrow \theta(i)$
  $Q \leftarrow \theta(j)$
  $d[j] \leftarrow D_{KL}(P||Q)$
end
d[i] $\leftarrow$ high value, to avoid the same component to be detected as neighbor.
return d

During the time of conversion, the posterior features are predicted from the source speaker's cepstral features using the first network. These predicted posterior features are then

passed through the second network to predict the target speaker's cepstral features. Since the GAN is able to produce natural realistic samples, we also propose to use GAN at the second stage for synthesis of the target cepstral features from the posterior features, instead of a simple DNN-based network. The block diagram of the GAN-based network is shown in Fig. 5.

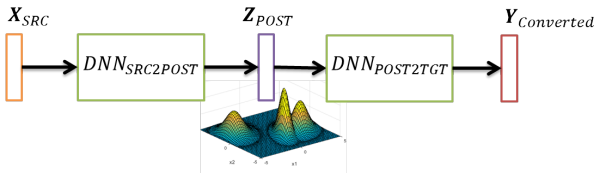

Fig. 4: Schematic representation of the proposed two stage DNN VC framework.

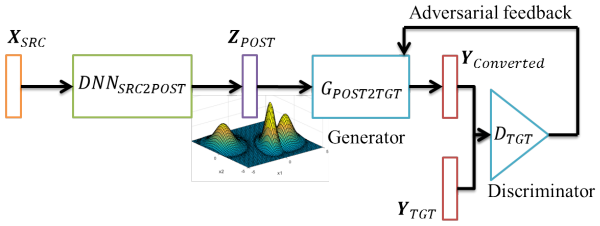

Fig. 5: Schematic representation of the proposed hybrid DNN-GAN VC framework.

GANs are the generative model that do not explicitly define a likelihood function, rather generates the samples by establishing a min-max game between a generator (G), and a discriminator (D). The G network learns to map samples x from some prior distribution $\mathcal{X}$ to samples y belonging to the data distribution $\mathcal{Y}$. The G network aims to minimize the distributional divergence between the model distribution $\hat{\mathcal{Y}}$ and the data distribution $\mathcal{Y}$. The D network is a binary classifier with input as real samples (y) or generated samples (output of G network). The G network is trained to maximally confuse the D network, whereas the D network is trained to maximize its classification accuracy [38], [39]. As the training proceeds, the G network generates the samples closely following $\mathcal{Y}$, and maximally confuses the D network in differentiating between $\mathcal{Y}$ and $\hat{\mathcal{Y}}$.

The vanilla GAN architecture when employed for the speech technology related applications, may sometimes fail in learning the accurate spectral representation, given the input representation. The architecture fails in preserving the speech quality and improving the speech intelligibility [20]. The G network may generate the samples resembling $\mathcal{Y}$, that may not correspond to the input speech-cepstral frames. To address this issue, MMSE regularization to the G network's adversarial loss reduces the numerical difference between the input-output cepstral feature-pair, in addition to minimizing the distributional divergence (adversarial training) [20]. The regularized adversarial objective function can be mathematically formulated as [20]:

$$\min_D V(D) = -\mathbb{E}_{y\sim\mathcal{Y}}[\log D(y)] - \mathbb{E}_{x\sim\mathcal{X}}[1 - \log(D(G(x)))], \tag{2}$$

$$\min_G V(G) = -\mathbb{E}_{x\sim\mathcal{X}}[\log(D(G(x)))] + \frac{1}{2}\mathbb{E}_{y\sim\mathcal{Y},x\sim\mathcal{X}}[\log(y) - \log(G(x))]^2, \tag{3}$$

where $\mathbb{E}_{y\sim\mathcal{Y}}$ denotes the expectation over all the samples y coming from the distribution $\mathcal{Y}$.

IV. Experimental Results

The VCC 2016 database consists of training utterances from 5 source and 5 target speakers [40]. In this paper, we develop *25* VC systems using each method among the available speaker-pairs. AHOCODER have been used for analysis-synthesis framework [41]. We extract 25-dimensional (d) Mel Cepstral Coefficients features (MCC) over a 25 ms window duration with the 5 ms frame shift. For each speaker in VCC 2016 database, *64*-d GMM and SGMM posteriors are extracted based on the model trained on the TIMIT database on *39*-d MFCC (including *13*-d static + Δ + $\Delta\Delta$ features.

The G network in the MMSE-GAN has three hidden layers, with 512 hidden units. Each layer is followed by batch normalization [42] and sigmoid activation. The output layer has 25 units to predict the target cepstral features, with linear activation. The D network also has three hidden layers, with 512 hidden units and each followed by batch normalization and *tanh* activation. The last layer uses the sigmoid activation in the D network. Dropout with 0.3 drop probability is selected for all the hidden layers in the G and D networks. The network is trained for 250 epochs, with an effective batch size of 1000. The network parameters are updated through Adam optimization [43], with a suitable learning rate of 0.001 [20]. Once the network is trained, the model with the least Minimum Square Error (MSE) on the validation set is selected and the testing is performed.

A. Subjective Evaluation

In this paper, two Mean Opinion Score (MOS) tests have been performed to evaluate the developed VC systems, based on the speech quality and the Speaker Similarity (SS) of the converted voices. 26 subjects (4 females and 22 males without any hearing impairments, and with the age variations between 18 to 22 years) participated in both the tests. Subjects evaluated the randomly played utterances for the speech quality on *5*-point scale. In particular, the subjects rated the converted voices on the scale of 1 (i.e., very bad) to 5 (i.e., very good) for speech quality. Fig. 6 shows the MOS analysis (obtained from total 384 samples) for the developed VC systems along with their 95 % confidence interval to quote the statistical significance of the results. Effectiveness of the proposed IMW GMM-PG over GMM-PG is ubiquitous in both the architecture in the context of speech quality of the converted voices. In particular, we obtained on an average 19.52 % and 7.94 % relative improvement in MOS for speech quality with the DNN

and the GAN-based VC systems, respectively. Similarly, in another MOS test, subjects rated the converted voices in terms of SS w.r.t. the target speaker on *5*-point scale. In the *5*-point scale, 1 means totally different to the target speaker, and 5 means exactly similar to the target speaker for the SS. Fig. 7 shows the MOS for the SS of the developed VC systems along with their 95 % confidence interval. In particular, we obtained on an average 5.25 % of relative improvement in terms of MOS for SS with the proposed IMW GMM-PG features compared to the GMM-PG features. The lack of large number of training examples for the adversarial training, results in lower performance of the GAN w.r.t. the DNN-based VC system as shown in Fig. 6 and Fig. 7 for speech quality and SS, respectively. However, the proposed IMW GMM-PG features clearly outperform the conventional GMM-PG.

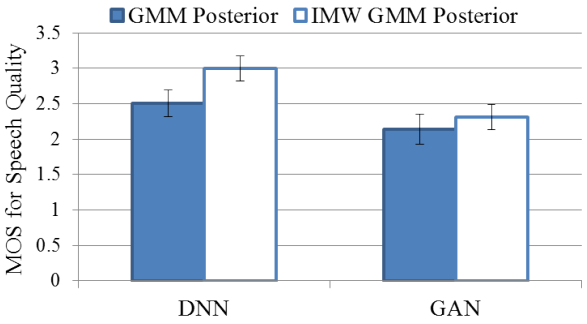

Fig. 6: MOS scores w.r.t. the speech quality of the developed systems along with the 95 % confidence interval.

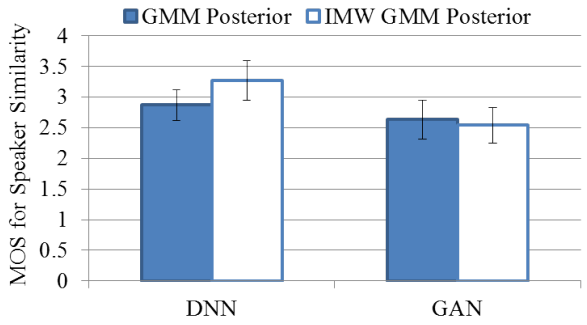

Fig. 7: MOS scores w.r.t. the speaker similarity of the developed systems along with the 95 % confidence interval.

B. Objective Evaluation

In this paper, traditional objective measure, namely, Mel Cepstral Distortion (MCD) (in dB) has been used for the objective evaluation [44]. The systems having lower MCD values can be considered as the better compared to the system having higher values of MCD. We obtain 0.2 dB of absolute reduction in the MCD with the proposed IMW GMM-PG w.r.t. to the GMM-PG in the DNN-based VC systems as shown in Fig. 8.

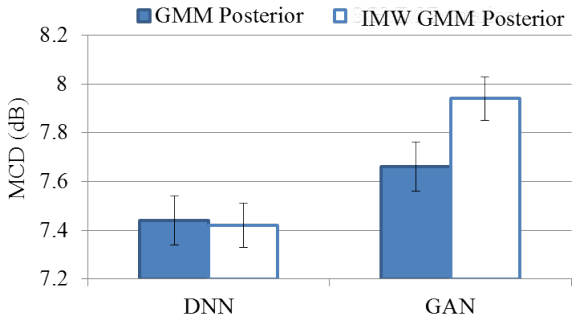

Fig. 8: MCD scores of the developed systems along with the 95 % confidence interval.

The MCD computes the scores on the basis of the numerical similarity between the cepstral features corresponding to the converted and the target speaker's data. However, the adversarial optimization minimizes the distributional divergence and do not optimize the numerical difference between the converted and the target speakers' cepstral features. Hence, we observe increment in the MCD in the case of GAN-based VC systems.

V. Summary and Conclusions

In this paper, we propose a novel unsupervised speaker-independent IMW GMM-PG features for the case of two-stage DNN as well as GAN-based VC framework. The key idea of IMW GMM-PG feature is to share the probability values of the current component with its neighbor, to spread the posterior probability across the components. The effectiveness of the proposed IMW GMM-PG features over the GMM-PG features can be clearly observed in the context of VC systems, developed on the VCC 2016 database. In particular, the relative improvement of 13.73 %, and 5.25 % is obtained with the proposed IMW GMM-PG w.r.t. the GMM-PG for the speech quality and the speaker similarity of the converted voices, respectively. In future, we plan to extend this work by adapting different strategies for sharing the probabilities across the components to obtain better speaker-independent representations.

VI. Acknowledgments

We would like to thank authorities of DA-IICT, Gandhinagar and Ministry of Electronics, and Information Technology (MeitY), Govt. of India, for their kind support to carryout this research work. We thank all the subjects who took part in the subjective evaluations.

References

[1] Y. Stylianou, "Voice transformation: a survey," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 3585–3588.

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

[3] 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. 1–5.

[4] 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*, Sunnyvale, CA, USA, 2016, pp. 147–152.

[5] N. J. Shah and H. A. Patil, *Analysis of features and metrics for alignment in text-dependent voice conversion*. B. Uma Shankar et. al. (Eds), PReMI, Lecture Notes in Computer Science (LNCS), Springer, vol. 10597, pp. 299–307, 2017.

[6] 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.

[7] Y. Agiomyrgiannakis, "The matching-minimization algorithm, the INCA algorithm and a mathematical framework for voice conversion with unaligned corpora," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 5645–5649.

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

[9] N. J. Shah and H. A. Patil, "Effectiveness of dynamic features in INCA and Temporal Context-INCA," in *INTERSPEECH*, Hyderabad, India, 2018.

[10] N. J. Shah, B. B. Vachhani, H. B. Sailor, and H. A. Patil, "Effectiveness of PLP-based phonetic segmentation for speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 270–274.

[11] M. Zaki, N. J. Shah, and H. A. Patil, "Effectiveness of multiscale fractal dimension-based phonetic segmentation in speech synthesis for low resource language," in *International Conference on Asian Language Processing (IALP)*, Kuching, Borneo Malaysia, 2014, pp. 103–106.

[12] 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.

[13] 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.

[14] 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.

[15] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational autoencoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, Jeju, Korea, 2016, pp. 1–6.

[16] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1283–1287.

[17] 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.

[18] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 1, pp. 84–96, 2018.

[19] N. J. Shah, M. Madhavi, and H. A. Patil, "Unsupervised vocal tract length warped posterior features for non-parallel voice conversion," in *INTERSPEECH*, Hyderabad, India, 2018.

[20] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," to appear in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018.

[21] S. Pascual, A. Bonafonte, and J. Serr, "SEGAN: Speech enhancement generative adversarial network," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3642–3646.

[22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, 2016, pp. 1–16.

[23] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3642–3646.

[24] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1993–1997.

[25] 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.

[26] N. J. Shah, M. Parmar, N. Shah, and H. A. Patil, "Novel mmse discogan for cross-domain whisper-to-speech conversion," in *Machine Learning in Speech and Language Processing (MLSLP) Workshop*, Google Office, Hyderabad, India, 2018.

[27] N. Shah, N. J. Shah, and H. A. Patil, "Effectiveness of generative adversarial network for non-audible murmur-to-whisper speech conversion," in *INTERSPEECH*, Hyderabad, India, 2018.

[28] 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, Queensland, Australia, 2015, pp. 4869–4873.

[29] 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.

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

[31] 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.

[32] M. C. Madhavi and H. A. Patil, "Partial matching and search space reduction for QbE-STD," *Comput. Speech Lang.*, vol. 45, no. 3, pp. 58–82, Sep. 2017.

[33] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition," in *INTERSPEECH*, Piitsburgh,USA, 2006, pp. 1–5.

[34] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Posterior features applied to speech recognition tasks with user-defined vocabulary," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 3809–3812.

[35] 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.

[36] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[37] 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.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2014, pp. 2672–2680.

[39] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[40] 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.

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

[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *in The International conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 448–456.

[43] 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.

[44] 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.