Interspeech 2018
2-6 September 2018, Hyderabad

Effectiveness of Dynamic Features in INCA and Temporal Context-INCA

Nirmesh J. Shah and Hemant A. Patil

Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),
Gandhinagar, India-382007

{nirmesh88_shah and hemant_patil}@daiict.ac.in

Abstract

Non-parallel Voice Conversion (VC) has gained significant attention since last one decade. Obtaining corresponding speech frames from both the source and target speakers before learning the mapping function in the non-parallel VC is a key step in the standalone VC task. Obtaining such corresponding pairs, is more challenging due to the fact that both the speakers may have uttered different utterances from same or the different languages. Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment (INCA) and its variant Temporal Context (TC)-INCA are popular unsupervised alignment algorithms. The INCA and TC-INCA iteratively learn the mapping function after getting the Nearest Neighbor (NN) aligned pairs from the intermediate converted and the target spectral features. In this paper, we propose to use dynamic features along with static features to calculate the NN aligned pairs in both the INCA and TC-INCA algorithms (since the dynamic features are known to play a key role to differentiate major phonetic categories). We obtained on an average relative improvement of 13.75 % and 5.39 % with our proposed Dynamic INCA and Dynamic TC-INCA, respectively. This improvement is also positively reflected in the quality of converted voices.

Index Terms: INCA, Temporal Context (TC)-INCA, dynamic features, alignment, Voice Conversion.

1. Introduction

Voice Conversion (VC) is a technique that maps the perceived speaker identity from a source speaker to a given target speaker without changing the message contained in a speech signal [1]. VC broadly can be categorized into parallel (if both the speakers have spoken the same utterances) and non-parallel cases (if both the speakers have spoken different utterances from a same language or different language). Stand-alone VC techniques that are based on Gaussian Mixture Model (GMM) [2,3], frequency warping (FW) [4, 5], exemplar [6] and Deep Neural Network (DNN) [7–9] requires the aligned spectral features before learning the mapping function. In the VC literature, it has been shown that the alignment accuracy clearly affects the quality of converted speech signal [10–12]. Hence, the accurate aligned spectral features from both the source and the target speakers' training speech database are required. On the other hand, recently proposed VC techniques that are based on adaptation [13] and generative model [14] avoided the need for such an alignment. However, in order to apply standalone VC systems, alignment is an unavoidable task. The alignment is more challenging in text-independent case (i.e., non-parallel data), since both the speakers have uttered different utterances, which is the most obvious realistic scenario.

Dynamic Time Warping (DTW) algorithm is used for the alignment task in the parallel VC task [15]. If the text utterances corresponding to the training speech data are available, the phoneme boundaries can be estimated using the forced Viterbi [16] or Spectral Transition Measure (STM)-based segmentation algorithm [17]. Recently, the text information have been used to generate speaker-independent phoneme posterior probability features, and used for non-parallel VC [18, 19]. However, it requires a separate training of Automatic Speech Recognition (ASR). Furthermore, developing robust ASR requires a huge amount of transcribed speech data from both the source and target speakers. In addition, publicly available ASR can be used for this task. However, it may not work in the cases where the low resource language is involved. Among various alignment techniques reported in the literature, the unsupervised alignment algorithms for the non-parallel data case is the **I**terative combination of a **N**earest Neighbor search step and a **C**onversion step **A**lignment (INCA) algorithm [20, 21].

The unsupervised INCA iteratively learns the mapping function that uses the nearest neighbor (NN) aligned features between the intermediate converted spectral features and the target spectral features [20, 21]. The % phonetic accuracy (PA) reported in the literature for the CMU-ARCTIC database for the non-parallel case is around 10 %, which is very less [22]. To overcome this issue, Temporal-Context (TC) INCA algorithm was proposed [22], which tries to incorporate the contextual information. Furthermore, since speech is a sequential data, extracting the contextual features from the speech, captures the local features (including coarticulation) and preserves the crucial harmonics [23, 24]. It is well known in the speech literature that the surrounding acoustic context affects the human speech perception [25–27]. Recently, researchers have tried to identify the underlying representations in the primary auditory cortex and secondary auditory cortex, and have examined the information modulated by varying the context in the area of the neuroscience of speech perception [28].

In this paper, we propose to use the dynamic features that incorporates the contextual information in the INCA. In particular, it extends non-parallel VC using relatively moderate modifications of existing frame-alignment algorithms. Furthermore, it has been found in speech perception literature that the dynamic features also play an important role to differentiate various phonetic categories, such as vowels, nasals, fricatives, stops, etc. [29–32]. In this paper, we propose to use dynamic features along with the static features for calculating the NN in the INCA and TC-INCA. In addition, we have also discussed the convergence behavior of our proposed algorithm. We have done relative analysis of % PA with the proposed dynamic features in INCA and TC-INCA. Furthermore, we have developed VC systems on the CMU-ARCTIC database using the aligned spectral feature-pairs obtained via different alignment algorithms [33].

10.21437/Interspeech.2018-1538

2. Proposed Dynamic Features for INCA and TC-INCA

2.1. INCA Algorithm

INCA iteratively performs three steps, namely, a nearest neighbor search step, training of mapping function using Joint Density Gaussian Mixture Model (JDGMM)-based VC and the transformation until convergence. Let $X = \{x_k\}_{k=1}^{N_x}$, $Y = \{y_j\}_{j=1}^{N_y} \in \mathbb{R}^d$ be the spectral features related to non-parallel corpus from the source and the target speakers, respectively. The alignment procedure is given below using *asymmetric-1* variant of the INCA algorithm since it is considered relatively the best among all other variants of INCA algorithm [20].

1. **Initialization:** At t^{th} iteration, auxiliary vector set, i.e., $(\mathcal{F}_{t-1}(\{x_k\}) = \{x_k^{'}\})$ represents an intermediate acoustic space of converted spectral features of previous iteration. The mapping function is initialized as $\mathcal{F}_0(x) = \{x_k\}$, which is called trivial initialization.

2. **NN search:** At each iteration for each vector $x_k^{'}$, the index of its corresponding NN vector in Y is estimated and stored in p(k). Similarly, for each vector y_j, its corresponding NN vector is found from $\{x_k^{'}\}$ and stored its index in $q(j)$.

$$\begin{aligned} p_t(k) &= \arg\min_j d(\mathcal{F}_{t-1}(x_k), y_j), \\ q_t(j) &= \arg\min_k d(y_j, \mathcal{F}_{t-1}(x_k)), \end{aligned} \tag{1}$$

   where $d(.)$ is the Euclidean distance.

3. **Training:** The spectral feature vectors given by $\{x_k, y_{p(k)}\}$ and $\{x_{q(j)}, y_j\}$ are concatenated and trained using the JDGMM-based method [2] and the mapping function $\mathcal{F}_t(.)$ is obtained using the GMM-based technique with MMSE-based conversion [2].

4. **Transformation:** The auxiliary vector set $X^{'}$ is updated after applying the mapping function $\mathcal{F}_t()$, i.e.,

$$x_k^{'} = \mathcal{F}_t(x_k), \forall k. \tag{2}$$

5. **Convergence Checking:** If the converted spectral features are very near to the target spectral features in mean square error (MSE) sense then the convergence is achieved, otherwise go to the *step 2*. The MSE between intermediate converted vectors and the target vectors is given by [20]:

$$\begin{aligned} E_T = \frac{1}{N_x + N_y} \Bigg(\sum_{k=1}^{N_x} \|\mathcal{F}_t(x_k) - y_{p_t(k)}\|^2 \\ + \sum_{j=1}^{N_y} \|y_j - \mathcal{F}_t(x_k)_{q_t(j)}\|^2 \Bigg), \end{aligned} \tag{3}$$

   where $||.||^2 = \sum_{<n>} |x(n)|^2$ (i.e., square of l^2 norm). The empirical and theoretical convergence of the E_t was shown in [20], [21], respectively.

2.2. Proposed Dynamic Features for INCA

Speech signal consists of various basic speech sound units, which are called as phonemes. These sounds and their features differ both in time and spectral characteristics [29]. The dynamic features, such as the change of distribution of spectral energy and temporal characteristics will play a vital role to discriminate major phonetic categories, such as nasals, stops, vowels, fricatives, etc. [32, 34]. In this paper, we propose to use the dynamic features to capture the contextual information that is present across the frames by taking longer contextual frames as given in [22]. The dynamic feature is given by [34]:

$$\Delta x_k = \frac{\sum_{i=1}^{T/2} i(x_{k+i} - x_{k-i})}{2\sum_{i=1}^{T/2} i^2}, \tag{4}$$

where T is even and contextual window length, $W_d = T + 1$, which is taken at the current frame, i.e., x_k by considering $T/2$ frames from the left and right side. Dynamic features are calculated and concatenated along with the static features. Hence, the new set of feature vectors is defined as $X_k = [x_k^T, \Delta x_k^T]^T$ and $Y_k = [y_k^T, \Delta y_k^T]^T$. Let $X = \{X_k\}_{k=1}^{\hat{N}_x}$, $Y = \{Y_j\}_{j=1}^{\hat{N}_y} \in \mathbb{R}^d$, where $\hat{N}_x \leq N_x$, $\hat{N}_y \leq N_y$ are the number of feature vectors from source and target speakers, respectively. Here, the cost function which is given by eq. (3) is modified from [20, 22] for our new set of feature vectors and it is given by:

$$\begin{aligned} \mathcal{E}_t = \frac{1}{\hat{N}_x + \hat{N}_y} \Bigg(\sum_{k=1}^{\hat{N}_x} \|\mathcal{F}_t(X_k) - Y_{p_t(k)}\|^2 \\ + \sum_{j=1}^{\hat{N}_y} \|Y_j - \mathcal{F}_t(X_k)_{q_t(j)}\|^2 \Bigg), \end{aligned} \tag{5}$$

where $\mathcal{F}(X_k)$ is the transformation function [2] and the JDGMM is trained for joint feature vectors, which is obtained by concatenating the static and the dynamic features from both the source and target speakers. The cost function given by eq. (5) cannot be solved using the gradient descent algorithm due to its dependency on the mapping function, i.e., $\mathcal{F}(.)$ [35]. In such a scenario, alternating minimization technique is used. It is well known optimization technique that iteratively minimizes the cost function depending on more than one variables [36, 37]. Hence, dynamic INCA algorithm can be defined as an optimization problem, aiming to minimize the following cost:

$$\{p^*, q^*, \mathcal{F}^*\} = \arg\min_{\{\mathcal{F}, p, q\}} \mathcal{E}(p, q, \mathcal{F}), \tag{6}$$

where p, q are the warping paths obtained after NN step 2 in the INCA. This joint optimization can be split into two separate minimization problems, which will be solved iteratively for $t = 1, 2, ...$ Hence, at iteration t, the algorithm (i.e., dynamic features in INCA (D-INCA) algorithm) is given by:

$$\{p_t, q_t\} = \arg\min_{\{p,q\}} \mathcal{E}(p, q, \mathcal{F}_{t-1}), \tag{7}$$

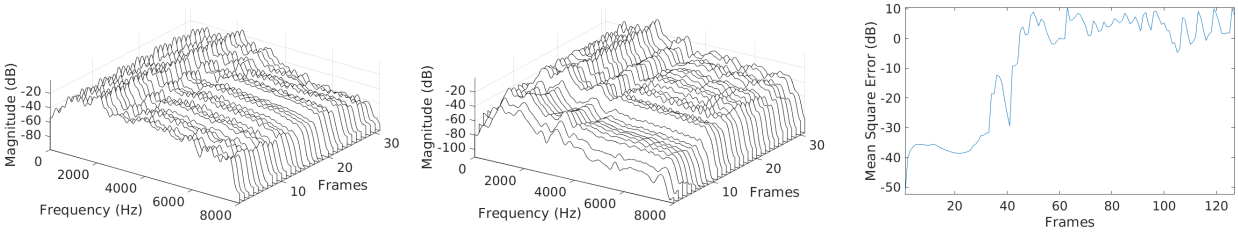
$$\mathcal{F}_t = \arg\min_{\mathcal{F}} \mathcal{E}(p_t, q_t, \mathcal{F}). \tag{8}$$

2.3. Convergence of Proposed D-INCA Algorithm

Due to alternating minimization nature of our training approach which is given by eq. (7) and eq. (8), it is clearly seen that the following inequality will always hold:

$$\begin{aligned} \mathcal{E}_t = \mathcal{E}(p_t, q_t, \mathcal{F}_t) \leq \mathcal{E}(p_t, q_t, \mathcal{F}_{t-1}) \\ \leq \mathcal{E}(p_{t-1}, q_{t-1}, \mathcal{F}_{t-1}) = \mathcal{E}_{t-1}, \forall t. \end{aligned} \tag{9}$$

Here, the cost function $\mathcal{E}_t$ is nothing but the Mean Square Error (MSE) and the transformation function is also given by the



(a) (b) (c)

Figure 1: *Waterfall plot of spectrum of a female twin-pair (both at the age of 27 years) uttering Hindi word /achanak/ (i.e., "Suddenly") (a) source (b) target speaker, and (c) Mean Square Error (MSE).*

MMSE criteria as given in [2]. Hence, the above mentioned inequality will be non-increasing and bounded below, the subsequence $\mathcal{E}_t$ must converge monotonically as per the Bolzano-Weierstrass theorem [38]. However, convergence to the global minimum is not guaranteed as our cost function is nonconvex. Hence, it will converge to a local minimum. In particular, convergence is in MSE sense than pointwise since source speaker can not be able to exactly match the spectral representation of target shape. In this context, we have taken the same utterance from two identical female twins speakers (a twin speaker-pairs) who are having almost identical speaker characteristics (as they look and sound perceptually very similar) [39]. It is observed from Figure 1 that even if they perceptually sound similar, their time-varying spectral representations is different and hence, the MSE in between their spectrum is not zero at every point as shown in Figure 1 (c).

2.4. Proposed Dynamic TC-INCA Algorithm

Similar to D-INCA, we propose to extend the idea of using dynamic features in the TC-INCA. The TC-INCA tries to use the Temporal Context information for finding NN aligned pairs in the INCA algorithm [22]. In particular, this is achieved by concatenating the current spectral feature vector with the (T/2) successive feature vectors in both the sides, i.e., $X_k = \{x_{k-T/2}^T, ..., x_k, ..., x_{k+T/2}^T\}$, $Y_k = \{y_{k-T/2}^T, ..., y_k, ..., y_{k+T/2}^T\}$ for a given contextual window length (i.e., $W_c = T + 1$).

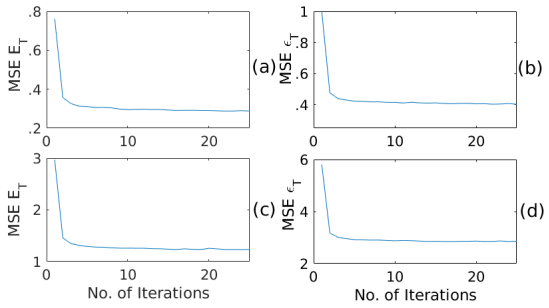


Figure 2: *Empirical convergence analysis for (a) INCA (b) D-INCA (c) TC-INCA, and (d) D-TC-INCA.*

The details of TC-INCA is given in [22]. Similarly, the spectral features will be considered along with its dynamic features and the context for finding the NN feature pairs for the proposed D-TC-INCA. In particular, the cost function given by eq. (5) is the same except, $X_k = [x_{k-T/2}^T, \Delta x_{k-T/2}^T, ..., x_k^T, \Delta x_k^T, ..., x_{k+T/2}^T, \Delta x_{k+T/2}^T]^T$, $Y_k = [y_{k-T/2}^T, \Delta y_{k-T/2}^T, ..., y_k^T, \Delta y_k^T, ..., y_{k+T/2}^T, \Delta y_{k+T/2}^T]^T$ and the transformation functions can be given by:

$$\mathcal{F}_t(X_k) = [\mathcal{F}(x_{k-T/2})^T, \mathcal{F}(\Delta x_{k-T/2})^T, ..., \mathcal{F}(x_k)^T, \mathcal{F}(\Delta x_k)^T, ..., \mathcal{F}(x_{k+T/2})^T, \mathcal{F}(\Delta x_{k+T/2})^T]^T, \quad (10)$$

where Δx_k is calculated over the contextual window length W_D using eq. (4) and the TC is taken over the window length W_C. The convergence for D-TC-INCA can be easily adapted from the convergence characteristics of D-INCA algorithm as discussed above. Empirical convergence also observed in D-INCA and D-TC-INCA for all the speaker-pairs. Among which the empirical convergence for one of the randomly selected speaker-pairs is shown in Figure 2. Cost function in all the proposed variants of INCA is different and hence, range of MSE will be different in all the cases. Still monotonically decrement in MSE sequence is clearly visible in all the cases.

2.5. Analysis of Phonetic Accuracies

In this paper, we converted the ground truth labeling, which is at phone-level to the frame-level labeling for the CMU-ARCTIC database [33]. The ground truth for the CMU-ARCTIC database is developed by training the speaker-dependent HMM model over 1132 utterances [33]. After alignment, using INCA algorithm and the proposed algorithm (i.e., D-INCA), if the aligned pairs are coming from the same phone label then it is considered as hit and if not then false. From this, % Phone Accuracy (PA) is defined as [20]:

$$\%\ \textit{Phone Accuracy} = \frac{\textit{Total no. Hits}}{\textit{Total no. Frames}} \times 100, \quad (11)$$

where *Total no. Frames* = *Total no. Hits* + *Total no. Falses*. Table 1 shows the % PA obtained using 40 non-parallel utterances from the CMU-ARCTIC database for each speaker-pairs (namely, BDL-RMS (male-male), BDL-SLT (male-female), CLB-RMS (female-male) and CLB-SLT (female-female)) using eq. (11). We have considered various different contextual length for calculating dynamic features in D-INCA algorithm, such as $W_D = \{3, 5, 7, 9, 11\}$. It is observed from Table 1 that the D-INCA (where dynamic features are calculated across three frames, i.e., W_{D3}) performs better than the INCA. Thus, dynamic features obtained over the W_{D3} is used for calculating the D-TC-INCA. We obtained clear improvement in % PA as the W_C increases in both TC-INCA and D-TC-INCA. In particular, we have observed on an average relative improvement of 5.39 % with our proposed D-TC-INCA over TC-INCA. In addition, it is clear from the Table 1 that for each considered contextual window length (namely, $W_C = \{3, 5, 7, 9, 11\}$), there is a clear improvement in % PA with D-TC-INCA over the TC-INCA. In all the cases, the best performing system (in terms of % PA) is selected for the further development of VC system.

Table 1: *% PA analysis after alignment for different VC systems w.r.t. the different contextual window length*

Speaker-Pair	INCA	D-INCA					TC-INCA					D-TC-INCA				
		$\mathbf{W}_{D3}$	$\mathbf{W}_{D5}$	$\mathbf{W}_{D7}$	$\mathbf{W}_{D9}$	$\mathbf{W}_{D11}$	$\mathbf{W}_{C3}$	$\mathbf{W}_{C5}$	$\mathbf{W}_{C7}$	$\mathbf{W}_{C9}$	$\mathbf{W}_{C11}$	$\mathbf{W}_{C3}$	$\mathbf{W}_{C5}$	$\mathbf{W}_{C7}$	$\mathbf{W}_{C9}$	$\mathbf{W}_{C11}$
M-M	25.87	**28.94**	23.70	17.74	10.74	6.10	28.42	28.80	30.71	31.76	**35.31**	30.90	31.85	32.51	34.33	**35.31**
M-F	20.66	**24.86**	22.89	16.59	11.50	8.11	23.75	26.66	29.18	28.55	**31.32**	27.09	28.32	29.33	30.98	**33.16**
F-M	19.24	23.06	**25.37**	19.78	15.32	8.94	24.40	26.0	27.53	27.66	**29.38**	23.51	26.16	27.49	29.56	**31.01**
F-F	32.46	**36.00**	28.32	20.07	14.85	11.20	36.58	39.34	40.57	43.17	**44.7**	38.2	39.92	41.26	43.99	**44.26**

3. Experimental Results

In this paper, various VC systems have been developed using the aligned features obtained by INCA and proposed dynamic INCA (D-INCA). 40 non-parallel utterances for each speaker-pairs from the CMU-ARCTIC database have been used. The state-of-the-art methods, namely, Joint Density (JD) GMM-based VC has been selected among the available various VC techniques, since it uses the conditional expectation, which is the best minimum mean square error (MMSE) estimator [2,40]. Hence, it leads to the minimum error between converted and the target spectral features. *25*-D Mel Cepstral Coefficients (MCC) (including the 0^{th} coefficient) and *1*-D F_0 per frame (with *25* ms frame duration and *5* ms frame shift) have been used. The number of mixture components has been varied, for example, *m=8,16, 32, 64, and 128*. The system having optimum Mel Cepstral Distortion (MCD), is selected for the subjective evaluation. Here, F_0 contour is transformed using Mean-Variance (MV) transform method [3].

3.1. Results

We have selected Mean Opinion Score (MOS) and XAB tests for subjective evaluations of speech quality and speaker similarity (SS) of converted voice, respectively. Both the subjective tests are taken from the *19* subjects (with no known hearing impairments with the age between 23 to 30 years), *6* females and *13* males from total *608* samples. In MOS test, subjects were asked to evaluate randomly played utterances for the speech quality (i.e., how natural is the converted voice ?) on the scale of 1 to 5 (1 very bad to 5 very good). Figure 3 shows the detailed MOS analysis for the developed VC systems along with 95 % confidence interval. On an average, effectiveness of the proposed D-INCA over INCA and the proposed D-TC-INCA over the TC-INCA is visible in Figure 3. The poor performance of the proposed algorithm for F-F case may be due to spectral resolution problem associated with female speech [41].

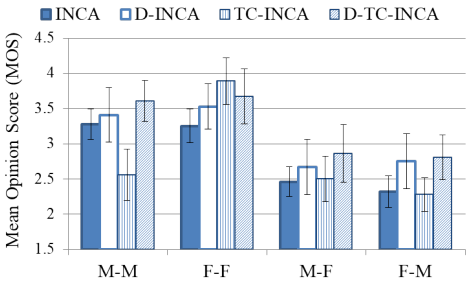


Figure 3: *MOS analysis for VC systems.*

In XAB test, the listeners were asked to select from the randomly played *A* and *B* samples (generated with INCA and TC-INCA, and the proposed D-INCA, and D-TC-INCA) based on the SS with reference to the actual target speaker's speech signal X. We found equal preference for both the systems as subjects were unable to distinguish at all, which system is performing better in terms of SS. This result indicates important observation that the accurate alignment may not lead to the better converted voice in terms of SS. However, it will definitely lead to the better speech quality of converted voice.

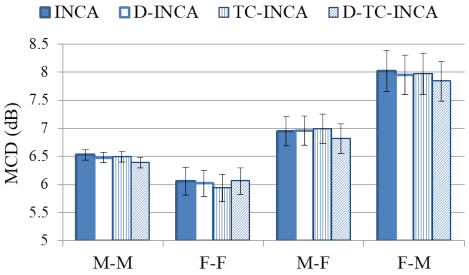


Figure 4: *MCD analysis of VC systems.*

The traditional Mel Cepstral Distortion (MCD) is used for the objective evaluations of various VC systems [3]. Our proposed D-TC-INCA and the D-INCA are performing better (i.e., relatively lesser MCD) compared to the TC-INCA and the INCA as shown in Figure 4.

Table 2: *PCC between % PA and MCD with the subjective score*

PCC	MOS	MCD
% **PA**	0.36	-0.8

Table 2 presents the analysis of Pearson Correlation Coefficient of % PA with the MOS and the MCD. We obtained 0.36 and -0.8 correlation of % PA with the MOS and the SS, respectively.

4. Summary and Conclusions

In this paper, we proposed to use the dynamic features in the INCA. We formulated the updated cost function for the proposed D-INCA. The dynamic features calculated using lesser context (in particular, with frame context length W_{D3}) performs better compared to the longer context (i.e., with $w = 5, 7, 9, 11$). In addition, we also discussed the convergence of D-INCA. Moreover, we also propose to use this dynamic features along with the TC-INCA by considering the different contextual window length W_C. We obtained on an average relative improvement of 13.75 % and 5.39 % with our proposed Dynamic INCA and Dynamic TC-INCA (w.r.t. the INCA and TC-INCA), respectively. This better performance of proposed approach may be due to its ability to exploit representation of speaking style via local *vs.* global coarticulations, that is captured using localized dynamic features and several contextual frames, respectively. In all the cases, the best performing systems in terms of % PA is selected for the further development of VC system. In addition, it has been observed that the VC systems that are developed using D-INCA and D-TC-INCA perform better than the INCA and the TC-INCA in terms of quality of converted voice from the subjective and objective evaluations. Since similar to INCA, our proposed algorithm does not require phonetic information, and hence, in the future, we would like to extend this work to the cross-lingual VC systems.

5. Acknowledgments

We thank authorities of DA-IICT, Gandhinagar and MeitY, Govt. of India for their kind support

6. References

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

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, 1998, pp. 285–288.

[3] 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.

[4] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 21, no. 3, pp. 556–566, 2013.

[5] 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.

[6] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[7] L. H. Chen, Z. H. Ling, L. J. Liu, and L. R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[8] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Spoken Language Technology Workshop (SLT)*, Nevada, USA, 2014, pp. 19–23.

[9] S. H. Mohammadi and A.Kain, "Semi-supervised training of a voice conversion mapping function using a joint-autoencoder," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 1–5.

[10] 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.

[11] 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.

[12] 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), Lecture Notes in Computer Science (LNCS), Springer, PReMI, vol. 10597, pp. 299–307, 2017.

[13] T. Kinnunen *et al.*, "Non-parallel voice conversion using i-vector PLDA: Towards unifying speaker verification and transformation," in *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] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustics, Speech and Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.

[16] F. Jelinek, *Statistical Methods for Speech Recognition*, 1^{st} ed. MIT Press, 1997.

[17] N. J. Shah, B. B. Vachhani, H. B. Sailor, and H. A. Patil, "Effectiveness of PLP-based phonetic segmentation for speech synthesis," in *ICASSP*, Florence, Italy, 2014, pp. 270–274.

[18] L. Sun *et al.*, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *International Conference on Multimedia and Expo*, Seattle, USA, 2016, pp. 1–6.

[19] 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.

[20] 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.

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

[22] H. Benisty, D. Malah, and K. Crammer, "Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion," in *ICASSP*, Florence, Italy, 2014, pp. 7909–7913.

[23] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Springer Science & Business Media, First Edition, 2012.

[24] N. J. Shah, M. Zaki, and H. A. Patil, "Influence of various asymmetrical contextual factors for TTS in a low resource language," in *International Conference on Asian Language Processing (IALP)*, Singapore, 2014, pp. 107–110.

[25] M. H. Davis, M. A. Ford, F. Kherif, and I. S. Johnsrude, "Does semantic context benefit speech understanding through top–down processes? evidence from time-resolved sparse fMRI," *Journal of Cognitive Neuroscience*, vol. 23, no. 12, pp. 3914–3932, 2011.

[26] L. L. Holt and A. J. Lotto, "Speech perception within an auditory cognitive science framework," *Current Directions in Psychological Science*, vol. 17, no. 1, pp. 42–46, 2008.

[27] M. Chait, D. Poeppel, A. De Cheveigné, and J. Z. Simon, "Processing asymmetry of transitions between order and disorder in human auditory cortex," *Journal of Neuroscience*, vol. 27, no. 19, pp. 5207–5214, 2007.

[28] M. K. Leonard and E. F. Chang, "Dynamic speech representations in the human temporal lobe," *Trends in Cognitive Sciences*, vol. 18, no. 9, pp. 472–479, 2014.

[29] G. Fant, *Speech Sounds and Features*. The MIT Press, 1973.

[30] B. Delgutte, "Auditory neural processing of speech," *The Handbook of Phonetic Sciences*, pp. 507–538, 1997.

[31] J. W. Schnupp, I. Nelken, and A. J. King, *Auditory Neuroscience: Making Sense of Sound*. The MIT Press, First Edition, 2012.

[32] S. Furui, "On the role of spectral transition for speech perception," *The J. of the Acoust. Soc. of Amer. (JASA)*, vol. 80, no. 4, pp. 1016–1025, 1986.

[33] J. Kominek and A. W. Black, "The CMU-ARCTIC speech databases," in *ISCA Speech Synthesis Workshop*, Pittsburgh, USA, 2004.

[34] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. 34, no. 1, pp. 52–59, 1986.

[35] K. P. Murphy, *Machine Learning : A Probabilistic Perspective*. The MIT Press, First Edition, 2013.

[36] P. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Patern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[37] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *Journal of Machine Learning Research*, vol. 6, pp. 2049–2073, 2005.

[38] R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*, 1^{st} ed. Wiley New York, 1992.

[39] H. A. Patil and T. K. Basu, "Detection of bilingual twins by teager energy based features," in *International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, 2004, pp. 32–36.

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

[41] A. Rajpal, N. J. Shah, M. Zaki, and H. A. Patil, "Quality assessment of voice converted speech using articulatory features," in *ICASSP*, New Orleans, USA, 2017, pp. 5515–5519.