NOVEL METRIC LEARNING FOR NON-PARALLEL VOICE CONVERSION

Nirmesh J. Shah and Hemant A. Patil

Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar
Email: {nirmesh88_shah and hemant_patil}@daiict.ac.in

ABSTRACT

Obtaining aligned spectral pairs in case of non-parallel data for stand-alone Voice Conversion (VC) technique is a challenging research problem. Unsupervised alignment algorithm, namely, an Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment (INCA) iteratively tries to align the spectral features by minimizing the Euclidean distance metric between the intermediate converted and the target spectral feature vectors. However, the Euclidean distance may not correlate well with the perceptual distance between the two (sound or visual) patterns in a given feature space. In this paper, we propose to learn distance metric using Large Margin Nearest Neighbor (LMNN) technique that gives a minimum distance for the same phoneme uttered by the different speakers and more distance for the different set of phonemes. This learned metric is then used for finding the NN pairs in the INCA. Furthermore, we propose to use this learned metric only for the first iteration in the INCA, since the intermediate converted features (which are not the actual acoustic features) may not behave well w.r.t. the learned metric. We obtained on an average 7.93 % relative improvement in Phonetic Accuracy (PA). This is reflected positively in subjective and objective evaluations.

Index Terms— VC, INCA, Metric Learning, LMNN.

1. INTRODUCTION

Non-parallel Voice Conversion (VC) has been a focus of research for last one decade. Alignment is one of the key research issues in non-parallel VC [1]. Though adaptation and generation model-based techniques avoid the alignment step in non-parallel VC [2–8], aligned spectral feature pairs are still required to apply standalone VC techniques for non-parallel data [1]. Among available alignment approaches for the non-parallel VC, the state-of-the-art unsupervised algorithm is the **I**terative combination of a **N**earest Neighbor search step and a **C**onversion step **A**lignment (INCA), which iteratively computes the mapping function that uses the Nearest Neighbor (NN) aligned feature pairs [9, 10].

The key issue in the INCA is that it tries to minimize the Euclidean metric among acoustic features for the time alignment. However, the same phoneme uttered by the two speakers may not have the minimum Euclidean distance [11–14]. Recently, in the area of keyword search, the possibilities of exploring distance metric learning is proposed to use instead of standard distance metric in the DTW [15, 16]. In this paper, we propose to learn the metric that gives a minimum distance for the same phoneme and maximum distance for the different phonemes uttered by the two different speakers. In particular, we propose to use this learned metric instead of the Euclidean for finding the NN pairs in the INCA. In this paper, we globally learn the metric on the TIMIT database due to the availability of phone annotations.

Furthermore, the INCA is sensitive to the initialization due to an alternating minimization nature of the algorithm [10, 17]. Except for *iteration 1* in the INCA, NN is obtained between intermediate converted and the target spectral feature vectors. These feature vectors may not behave like the actual spectral features. Hence, we propose to use this learned metric only for the initial iteration, where the spectral features are derived from both the source and target speakers. Among various available metric learning techniques [18], we used the state-of-the-art Large Margin Nearest Neighbor (LMNN) technique [19, 20]. These aligned feature pairs that are obtained using the Euclidean metric, and the learned metric, are further used to develop various VC systems.

2. METRIC LEARNING FOR ALIGNMENT TASK

2.1. Motivation for Metric Learning in VC

INCA consists of three steps, namely, initialization, NN search and transformation step [9]. These steps are repeated until the convergence. In the literature, lower Phonetic Accuracy (PA) is reported after the alignment step [17, 21]. To further investigate the possible reasons behind this low PA, we apply *t*-stochastic neighbor embedding (t-SNE) visualization technique to the acoustic space of the source and target speakers [22]. We have taken one of the available speaker-pairs (namely, BDL-RMS (M-M)) from the CMU-ARCTIC database [23]. The acoustic space for a vowel, stop, nasal and fricative is shown in Fig. 1. We can clearly see that the

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

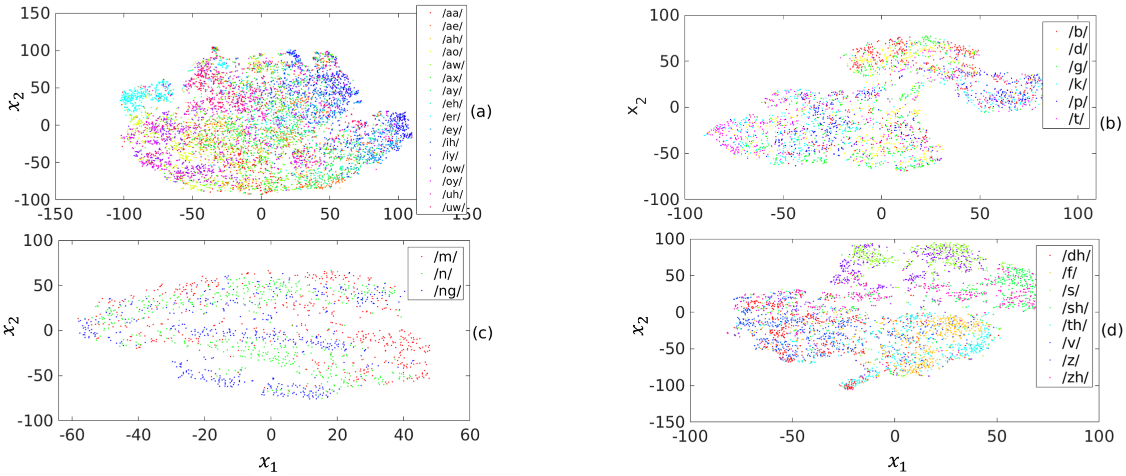


Fig. 1: Acoustic features space visualization in 2-D using t-SNE for different speech sound classes, such as (a) vowel, (b) stop, (c) nasal, and (d) fricative.

same phoneme uttered by the two speakers does not lie in the neighborhood in Euclidean space, rather they are spread across the 2-D acoustic space. This is primarily due to the difference in vocal tract system (i.e., size and shape) and excitation source (difference in size of the glottis, vocal fold mass, tension in the vocal folds and hence, the manner in which glottis opens or closes, i.e., the glottal activity) across the speakers (Chapter 3, pp. 59) [24]. This motivated the authors to define new metric that represents the acoustic feature space. In this paper, we propose to use the learned metric for finding the NN pairs in the iteration II of the INCA [9].

2.2. Metric Learning

The metric learning is concerned with the learning of a distance function w.r.t. a particular task. Metric learning has been shown to be extremely useful when used along with the NN methods [18]. The metric learning techniques can be broadly classified into linear (which uses Mahalanobis distance) *vs.* nonlinear approaches [18]. Let $X = [x_1, x_2, ..., x_n]$ be the matrix of all the data points. The mapping $d : X \times X \rightarrow \mathbb{R}$ is called a *metric* if it satisfies following four conditions [25]:

1. $d(x_i, x_j) \geq 0$ (non-negativity),
2. $d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$ (identity of indiscernible),
3. $d(x_i, x_j) = d(x_j, x_i)$ (symmetry),
4. $d(x_i, x_j) \leq d(x_i, x_r) + d(x_r, x_j)$, where $\forall x_i, x_,x_r \in X$ (triangle inequality).

If condition (2) is dropped then the mapping is called as *pseudo* metric [25]. In particular, distance metric is defined through inner product space. For example, $d^2(x, y) = ||\langle x - y, x - y\rangle|| = x^T y$. Hence, in general a distance metric is defined as:

$$d_A(x, y) = (x - y)^T A(x - y). \tag{1}$$

If $A = \Sigma^{-1}$ then distance is called as the Mahalanobis distance [26]. Here, Σ is covariance matrix of the data. In most of the cases, true covariance is unknown and hence, the sample covariance is used. Here, A must be positive-semidefinite (PSD) (i.e., $A \succeq 0$, where $\succeq$ notation is used to indicate positive semidefinite) to satisfy the metric definition. Furthermore, if A is PSD then it can be factorized as $A = G^T G$ that leads to $d_A(x, y) = ||Gx - Gy||_2^2$ (where $||\cdot||$ is the L^2 norm). Hence, the idea behind learning the metric can be considered as the learning of global linear transformation. The idea behind learning Mahalanobis metric was proposed by Xing *et. al.* in [27]. Here, the desired metric should give minimum squared distance for the pairs $(x_i, x_j) \in \mathcal{S}$ (where S is a set of similar pairs) with constraint $\sum_{(x_i,x_j)\in D} d_A(x_i, x_j) \geq 1$, where D is set of dissimilar pairs. Here, the objective function is given by [27]:

$$\underset{A}{\arg\min} \sum_{(x_i,x_j)\in\mathcal{S}} ||x_i - x_j||_A^2, \tag{2}$$

$$\text{subject to} \sum_{(x_i,x_j)\in\mathcal{D}} ||x_i - x_j||_A^2 \geq 1, \quad A \succeq 0. \tag{3}$$

The above mentioned objective function is linear and both the constraints are convex and hence, the convex optimization algorithms can be applied to get global optimal solution. In particular, the gradient descent and the idea of iterative projections can be used to solve the above mentioned convex optimization problem [27]. Weinberger *et. al.* proposed the LMNN technique that uses the relative distance constraints, which is one of the most popular and state-of-the-art metric learning technique in the literature [19, 20]. The main aim of LMNN technique is that the given feature should have the same label as its neighbors, while the features that are having different labels should be distant apart from the given feature. The key idea behind the LMNN is illustrated in Figure 2.

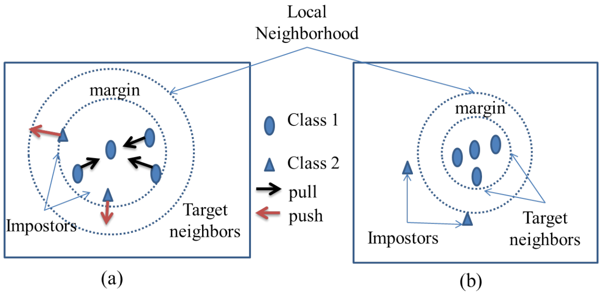


Fig. 2: Schematic representation of LMNN technique (a) before and (b) after applying the LMNN technique. Adapted from [19].

Here, the target neighbors refer to the features that have similar label and impostor is also the neighbor feature vector. However, it is having different label. The goal of LMNN technique is to minimize the number of impostors via relative distance constraint. The objective function is given by [19] :

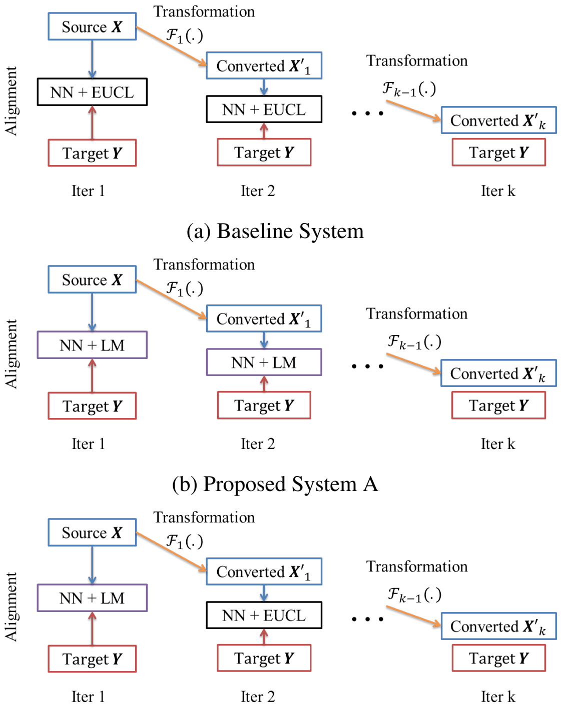
$$\begin{aligned}\arg\min_{A\succeq 0} &\sum_{(i,j)\in\mathcal{S}} d_A(x_i,x_j) \\ &+\lambda \sum_{(i,j,k)\in\mathcal{R}} [1+d_A(x_i,x_j)-d_A(x_i,x_k)],\end{aligned} \tag{4}$$

where $\mathcal{R}$ is the set of all triplets (i,j,k) such that x_i and x_j are the target neighbors and x_k is the impostor.

In this paper, we have used TIMIT (American English) database for estimating the learned metric as the manual phone-annotations are available, which is obtained from the highly trained human annotators [28]. We used the full phone set label. We randomly selected a small subset of the database to learn the metric using LMNN technique. We extracted *25*-D Mel Cepstral Coefficients (MCC) per frame (with *25* ms frame duration, and *5* ms frameshift). In this paper, we globally learn the metric for the spectral features and use this learned metric for calculating NN feature pairs. We considered three possible approaches to use the learned metric in the INCA. Schematic representations of the various approaches are given in Figure 3. Proposed system A uses the learned metric in each iteration of the baseline INCA as shown in Figure 3 (a) and (b). On the other hand, the proposed system C uses the learned metric only at the iteration I in the INCA as shown in Figure 3 (c). Proposed system B first applies the global transformation that is learned via metric learning to the spectral features obtained from both the speakers and then the baseline INCA is applied to the transformed features.

3. EXPERIMENTAL RESULTS

We have used CMU-ARCTIC database due to the availability of the phone annotations that is obtained using speaker-dependent hidden Markov model (HMM) trained over 1132 utterances [23]. Here, we converted the reference phone-annotations to the frame-level labeling. In this paper, 40 non-parallel utterances from each speaker-pair have been used to develop VC system using the aligned features obtained via the baseline and proposed techniques. The state-of-the-art methods, namely, Joint Density Gaussian Mixture Model (JDGMM)-based VC has been selected among the available various VC techniques [29]. The JDGMM-based method is selected since it uses conditional expectation, which is the best minimum mean square error (MMSE) estimator [30]. Hence, it leads to the minimum error between converted and the target spectral features. *25*D MCC and *1*-D F_0 per frame (with *25* ms frame duration, and *5* ms frameshift) have been extracted using AHOCODER. 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, Mean-Variance (MV) transformation has been used for F_0 conversion [31].



(a) Baseline System

(b) Proposed System A

(c) Proposed System C

Fig. 3: Schematic representation of (a) baseline, (b) proposed system A, and (c) proposed system C. Proposed system B is not shown here, since it applies the baseline technique to the transformed features obtained via the LM, and hence, similar to (a). EUCL: Euclidean metric, LM: Learned metric.

3.1. Analysis of Phonetic Accuracies

In the context of VC, if the aligned pair contains features from the same phoneme then it is considered as hit and if not then false. From this, Phonetic Accuracy (PA) is defined as [13]:

$$PA \text{ (in \%)} = \frac{Total\ no.\ of\ Hits}{Total\ no.\ of\ Frames} \times 100, \tag{5}$$

$$\textit{Total no. of Frames} = \textit{Total no. of Hits} + \textit{Total no. of Falses}.$$

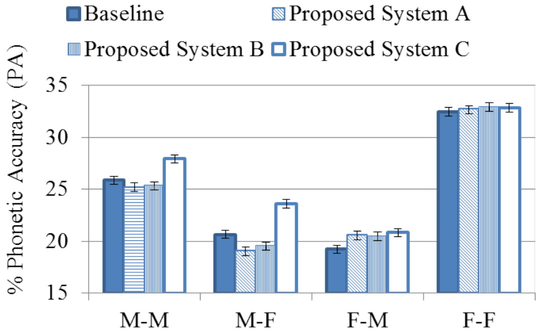


Fig. 4: PA of different initialization techniques for non-parallel VC systems.

Fig. 4 shows the accuracy of alignment obtained using three proposed techniques. It is observed that there is on an average 0.71 % relative reduction and 0.07 % relative improvement (in the PA) w.r.t. the baseline using the proposed method A and B, respectively. It is possibly due to fact that this metric is globally learned for entire TIMIT. The broad phonetic classes, such as vowel, stops, fricatives, nasal, etc. will behave very much differently in acoustic space due to the different manner of articulation required to produce these sounds [24]. Since the metric is learned for the true acoustic features and not for the intermediate converted acoustic features, we propose technique C, which is performing consistently better (with on an average 7.93 % relative improvement in PA) than the INCA.

3.2. Subjective and Objective Evaluations

Mean Opinion Score (MOS) and ABX tests have been performed for measuring speech quality, and speaker similarity (SS) of the converted voices, respectively. Both the subjective tests are taken from the *16* subjects (*5* females and *11* males with no known hearing impairments and with the age variations between 18 to 29 years) from the total of *288* samples. The result of *5*-point (1 (very bad) to 5 (very good)) MOS test are shown in Table 1 along with the 95 % confidence intervals. It can be seen from Table 1 that the proposed system C is more preferred than the baseline in terms of speech quality (i.e., naturalness) for the VC (except in the case of F-M).

Table 1: MOS analysis for the naturalness of converted voices. Number in the bracket indicates a margin of error corresponding to the 95 % confidence intervals for VC systems

	M-M	**M-F**	**F-M**	**F-F**
Baseline	3.06 (0.27)	2.41 (0.29)	**2.66** **(0.28)**	3.5 (0.26)
Proposed System C	**3.31** **(0.29)**	**2.81** **(0.22)**	2.53 (0.21)	**3.5** **(0.25)**

In ABX test for SS, the listeners were asked to select from the randomly played *A* and *B* samples (generated with the baseline and the proposed system C) based on the SS with reference to the actual target speaker's speech signal X. Eight samples for ABX test were taken from both the approaches. All the subjects have given *equal* preference to both the systems. This result indicates that accurate alignment may not lead to the better converted voice in terms of SS. However, it will lead to the better speech quality of converted voice.

Table 2: MCD analysis. Number in bracket indicates the margin of error corresponding to the 95 % confidence intervals

	M-M	**M-F**	**F-M**	**F-F**
Baseline	6.53 (0.34)	6.95 (1)	8.02 (1.29)	6.06 (0.93)
Proposed System C	**6.41** **(0.09)**	**6.76** **(0.26)**	**7.85** **(0.34)**	**6.02** **(0.24)**

The traditional Mel Cepstral Distortion (MCD) is used here [31]. It can be seen from Table 2 that our proposed system C is performing better than the baseline in all the cases. Table 3 presents the analysis of Pearson Correlation Coefficient (PCC) of PA and MCD with the MOS and the SS. It is clear from the Table 3 that the PCC between PA and MOS is 0.96, i.e., PA is having more correlation with the MOS. This clearly indicates that better alignment will lead to better speech quality. On the other hand, PCC between PA and SS is less compared to the PCC between PA and MOS. For the case of MCD, PCC ideally should be -1 since lesser value of MCD means that the system is performing better than the given systems (that are having higher values of MCD). It is clearly seen that the traditional MCD is *not* correlating well with the MOS and the SS. Less correlation between MCD and the subjective scores have also been reported in the VC literature [32].

Table 3: PCC of % PA and MCD with the subjective score

PCC	**MOS**	**SS**
PA	0.96	0.37
MCD	-0.3	0.10

4. SUMMARY AND CONCLUSIONS

In this study, we proposed to exploit metric learning technique for finding NN in the INCA than state-of-the-art Euclidean distance. Furthermore, we also proposed to use our learned metric only for the initial iteration of INCA since the metric is learned for the actual acoustic features. Therefore, during other iterations in the INCA, intermediate converted features may not represent the true acoustic features. We compare our proposed system C with the baseline INCA and found that our proposed system performs better (in terms of PA) than the baseline INCA. Moreover, subjective as well as objective evaluations also confirm that the proposed system C performs better w.r.t. the baseline system. In particular, improvement (in terms of PA) obtained due to system C is clearly reflected in the MOS scores with the PCC of 0.96. In the future, we plan to apply local learning of the metric in order to capture local metric for each broad phonetic classes.

5. REFERENCES

[1] Seyed Hamidreza Mohammadi and Alexander Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 8, no. 4, pp. 65–82, 2017.

[2] Tomi Kinnunen et al., "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.

[3] Toru Nakashika, "Cab: An energy-based speaker clustering model for rapid adaptation in non-parallel voice conversion," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3369–3373.

[4] Chin-Cheng Hsu et al., "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3364–3368.

[5] Fuming Fang et al., "High quality nonparallel voice conversion based on cycle-consistent adversarial network," in *ICASSP*, Calgary, Canada, 2018, pp. 5279–5283.

[6] Nirmesh J. Shah, Maulik C Madhavi, and Hemant A. Patil, "Unsupervised vocal tract length warped posterior features for non-parallel voice conversion," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 1968–1972.

[7] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *ICASSP*, Calgary, Canada, 2018, pp. 5274–5278.

[8] Nirmesh J Shah, Sreereaj R., Neil Shah, and Hemant A. Patil, "Novel inter mixture weighted GMM posteriorgram for DNN and GAN-based voice conversion," in *Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*, Hawaii, USA, 2018, IEEE, pp. 1776–1781.

[9] D. Erro et al., "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 18, no. 5, pp. 944–953, 2010.

[10] Nirmesh J. Shah and Hemant A. Patil, "On the convergence of INCA algorithm," in *APSIPA ASC*, Kuala Lumpur, Malaysia, 2017, IEEE, pp. 559–562.

[11] Jui-Ting Huang et al., "Kernel metric learning for phonetic classification," in *Workshop on Automatic Speech Recognition & Understanding (ASRU)*, Merano, Italy, 2009, pp. 141–145.

[12] Xiong Xiao et al., "Distance metric learning for kernel density-based acoustic model under limited training data conditions," in *APSIPA ASC*, Hong Kong, 2015, pp. 54–58.

[13] Nirmesh J. Shah and Hemant 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.

[14] Julian Martin Fernandez and Bart Farell, "Is perceptual space inherently non-Euclidean?," *Journal of Mathematical Psychology*, vol. 53, no. 2, pp. 86–91, 2009.

[15] Batuhan Gündoğdu and Murat Saraçlar, "Distance metric learning for posteriorgram based keyword search," in *ICASSP*, New Orleans, USA, 2017, pp. 5660–5664.

[16] Batuhan Gündoğdu et al., "Joint learning of distance metric and query model for posteriorgram based keyword search," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1318–1328, 2017.

[17] Nirmesh J. Shah and Hemant A. Patil, "Effectiveness of dynamic features in INCA and temporal context-INCA," in *INTERSPEECH*, Hyderabad, India, 2018, pp. 711–715.

[18] Brian Kulis et al., "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, pp. 287–364, 2013.

[19] Kilian Q Weinberger and Lawrence K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[20] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2005, pp. 1473–1480.

[21] 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.

[22] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[23] John Kominek and Alan W Black, "The CMU-ARCTIC speech databases," in 5^{th} *ISCA Workshop on Speech Synthesis*, Pittsburgh, USA, 2004, pp. 223–224.

[24] Thomas F Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* 1^{st} *Edition,*, Pearson Education India, 2006.

[25] E. Kreyszig, *Introductory Functional Analysis with Applications*, vol. 81, Wiley New York, 1^{st} edition, 1989.

[26] PC Mahalanobis, "Mahalanobis distance," in *Proceedings National Institute of Science of India*, 1936, vol. 49, pp. 234–256.

[27] Eric P Xing et al., "Distance metric learning with application to clustering with side-information," in *NIPS*, Vancouver, Canada, 2002, vol. 15, p. 12.

[28] John S Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST),USA*, vol. 15, pp. 29–50, 1988.

[29] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP*, Seattle, WA, USA, 1998, pp. 285–288.

[30] Steven M. Kay, "Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory," *Upper Saddle River, New Jersey: Prentic Hall*, 1998.

[31] 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.

[32] Avni Rajpal, Nirmesh J. Shah, Mohammadi Zaki, and Hemant A. Patil, "Quality assessment of voice converted speech using articulatory features," in *ICASSP*, New Orleans, USA, 2017, pp. 5515–5519.