INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Phone Aware Nearest Neighbor Technique using Spectral Transition Measure 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, India

Email: {nirmesh88_shah,hemant_patil}@daiict.ac.in

Abstract

Nearest Neighbor (NN)-based alignment techniques are popular in non-parallel Voice Conversion (VC). The performance of NN-based alignment improves with the information about phone boundary. However, estimating the exact phone boundary is a challenging task. If text corresponding to the utterance is available, the Hidden Markov Model (HMM) can be used to identify the phone boundaries. However, it requires a large amount of training data that is difficult to collect in realistic VC scenarios. Hence, we propose to exploit a Spectral Transition Measure (STM)-based alignment technique that does not require apriori training data. The idea behind STM is that neurons in the auditory or visual cortex respond strongly to the *transitional* stimuli compared to the steady-state stimuli. The phone boundaries estimated using the STM algorithm are then applied to the NN technique to obtain the aligned spectral features of the source and target speakers. Proposed STM+NN alignment technique is giving on an average 13.67% relative improvement in phonetic accuracy (PA) compared to the NN-based alignment technique. The improvement in %PA after alignment has positively reflected in the better performance in terms of speech quality and speaker similarity (in particular, a relative improvement of 13.63% and 13.26% , respectively) of the converted voice.

Index Terms: Voice Conversion, Spectral Transition Measure, INCA.

1. Introduction

Though generation model-based techniques, such as Generative Adversarial Network (GAN), Variational AutoEncoder (VAE), etc. avoid an alignment step [1–5], still the alignment is a key step in order to apply stand-alone Voice Conversion (VC) techniques in case of a non-parallel training data [6]. Obtaining the aligned spectral features in non-parallel VC is more challenging due to the fact that both the source and target speakers have spoken different utterances. Among various available alignment approaches, the most popular alignment techniques are based on Nearest Neighbor (NN), for example, the state-of-the-art **I**terative combination of a **N**earest Neighbor search step and a **C**onversion step **A**lignment (INCA) [7, 8] and its variants [9–11]. However, lower % Phonetic Accuracy (PA) has been reported after the NN-based alignment techniques [7]. Hence, if the phone boundaries corresponding to the non-parallel data are available, then the NN can be applied among the features corresponding to the same phones. Estimating the phone boundaries is more challenging due to the coarticulation phenomenon of speech sound, which leads to splitting or disappearing of current sound due to merging with or interference of the adjacent sounds (primarily due to local *vs.* global coarticulation) [12]. If the text corresponding to the training data is available, Hidden Markov Model (HMM) can be used to identify the correct phone boundaries or recently proposed the speaker-independent phone posterior probability features can also be used [13–15]. However, these methods require a large amount of training data to train HMM or develop Automatic Speech Recognition (ASR), which is difficult due to the unavailability of a large amount of transcribed training utterances from the target speaker in most of the real-world VC applications.

In this paper, we propose to exploit computationally simple Spectral Transition Measure (STM)-based alignment technique that does not require any apriori training data for estimating the phone boundaries. The STM is derived from the information pertaining to linear regression coefficients. These regression coefficients have large values due to the rapidly varying cepstral information resulting into large STM values in the vicinity of spectral transition. Such high transitions, estimated via STM algorithm, are hypothesized as the phone boundaries. The earlier studies have also shown the relation between the maximum spectral transition positions (i.e., the location of peaks in the STM contour), and the perceptual critical points (PCPs) for syllable perception [16]. Later, the important relationship between the manual phone boundaries and the PCPs have been investigated [17, 18]. The STM-based alignment techniques have also been used for identifying the syllable and phone boundaries for the low resource languages [18–23].

In this paper, we propose a simple and practical way of utilizing the phone boundary information via STM algorithm for the alignment task in non-parallel VC. The phone boundaries estimated using the STM algorithm are then applied to the NN method (i.e., phone-aware NN) to find the alignment between the source and target speakers' spectral features. We compare the effectiveness of the phone-aware NN-based alignment technique with the NN-based method. In addition, subjective and objective evaluations of the developed VC systems using the proposed approach are also presented.

2. Proposed STM-based Alignment

Figure 2 shows an example of manual phone boundaries obtained for a CMU-ARCTIC data (i.e., word 'author') spoken by the two speakers. Even if both the speakers have spoken the utterances with different speaking rate, still it can be observed that the phone boundaries occurs at the spectral transition locations (as shown in Figure 2). It is often observed that two human annotators can never identify exactly the same phone boundaries. In addition, obtaining the manual phonetic segmentation on the given speech corpus is extremely tedious and time con-

http://dx.doi.org/10.21437/Interspeech.2019-1504

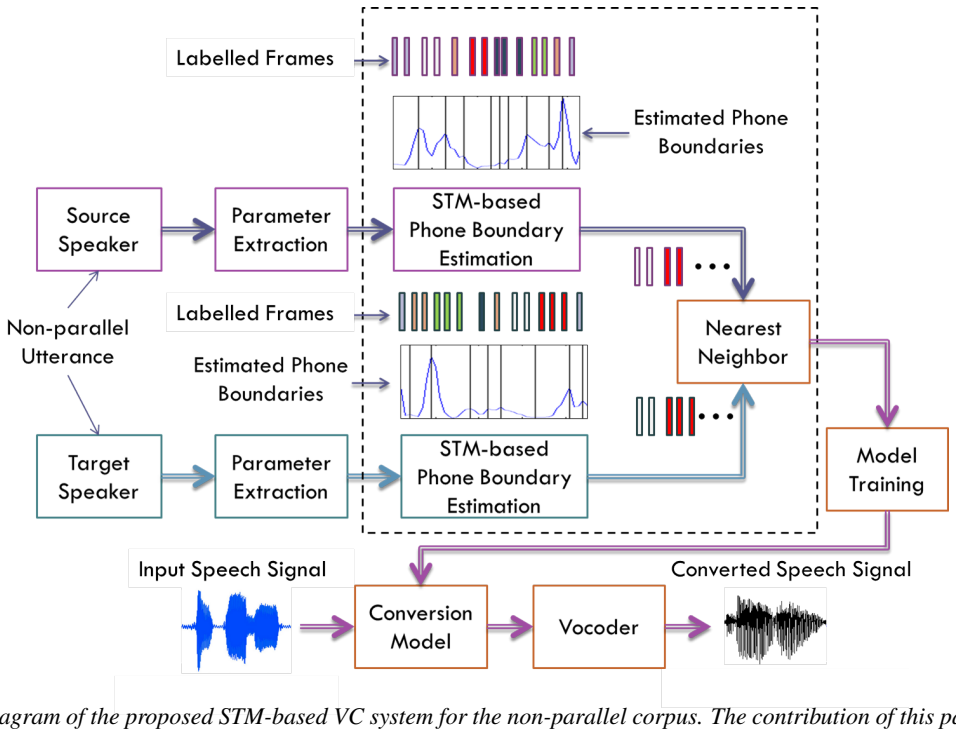

Figure 1: *Block diagram of the proposed STM-based VC system for the non-parallel corpus. The contribution of this paper is indicated via dotted box.*

suming. Furthermore, it requires highly trained human annotators, which makes this process very costly. In real-time VC, it is impossible to do manual segmentation as soon as one gets the training data from the target speaker. Hence, there is a need for automatic speech segmentation algorithm for the alignment task of VC.

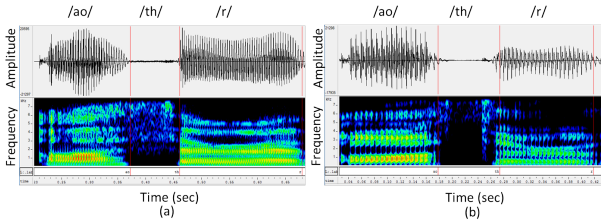

Figure 2: *Speech signal, and its corresponding spectrogram for English word, namely, "author" from CMU-ARCTIC database from (a) male, and (b) female speaker. After [19].*

Speech signal consists of various basic speech sound units called as *phonemes*. These sounds and their features differ both in time and spectral characteristics. The dynamic features, such as temporal envelope characteristics and the change of distribution of spectral energy will play a key role to distinguish major phonetic categories, such as vowels, nasals, stops, fricatives, etc. [24–27]. These rapid changes in amplitude and spectrum are apparently represented in the discharge pattern of auditory-nerve fibers (ANFs) [28]. This spatio-temporal pattern of auditory-nerve activity has shown to contain pointers to the regions of rapid changes that captures the important phonetic (transitional) information [28, 29]. In addition, the earlier studies reported that the neurons present in the auditory cortex poorly respond to the steady-state stimuli, whereas they have high auditory sensitivity for the *transitional* sounds [30]. Thus, STM exploits spectral variations to detect the phone boundaries.

Figure 1 shows the block diagram of a proposed approach.

Here, Mel Frequency Cepstral Coefficients (MFCC) have been used for capturing these spectral transitions across the consecutive phones. First, the speech signal is pre-processed using the silence removal technique. In order to estimate the phone boundary, we have followed the same experimental setup as suggested in [17, 19]. *10*-dimensional (D) MFCC features (including 0^{th} coefficient) are first extracted (with 30 ms window and 10 ms frame rate) as suggested in [17, 19, 23]. After computing the MFCC, STM is used to capture the spectral variations between the two consecutive phones. The STM, at the i^{th} frame, can be computed as [16]:

$$C(i) = \frac{\sum_{l=1}^{L} a_l^2(i)}{D}, \tag{1}$$

where $C(i)$ is the STM calculated at a given frame i, D is the dimension of the spectral features, a_l's are the regression coefficients, which are the rate of change of spectral features. The a_l is given by [16]:

$$a_l(i) = \frac{\sum_{j=-I}^{I} MFCC_l(j+i) \cdot j}{\sum_{j=-I}^{I} j^2}, \tag{2}$$

where j is the frame index, and I indicates the number of frames (on both side of the current frame) used to compute the linear regression coefficients. Finally, peak detection algorithm is used to estimate the phone boundaries. A number of estimated boundaries may not be equal to the number of phones in a given utterance and hence, the proposed algorithm is used to get the exact number of estimated boundaries equal to the number of phones. Here, we have taken $I = 2$ for 10 ms frame shift. Hence, the value of C in eq. (1) is computed over an interval of

40 ms. A larger interval results in the missing of some phone boundaries, whereas the shorter interval results in a false estimate of the phone boundaries. Previous studies presented the effectiveness of the algorithm in the context of a tolerance interval [18, 19, 23]. It means that if detected phone boundary is within the tolerance interval (namely, 5 ms, 10 ms or 20 ms) of the ground truth, then it is considered as a true estimated boundary else detected boundary is considered as a false.

In the context of VC task, we consider 0 ms tolerance interval, i.e., frames are aligned based on these estimated phone boundaries. The exact locations of the phone boundaries will determine the corresponding pairs among which NN technique will be applied. Hence, we converted the phone-level labeling to the frame-level labeling, and compared it with the corresponding frame-level labelling of ground truth. If both the estimated and the ground truth frame-level label are found to be the same, it is considered as hit and if not then false. From this, % Phonetic Accuracy (PA) is defined as [11, 31]:

$$\%\ \textit{Phonetic Accuracy (PA)} = \frac{\textit{Total no. of Hits}}{\textit{Total no. Frames}} \times 100, \quad (3)$$

where *Total no. of Frames = Total no. of Hits + Total no. of Falses*. Table 1 shows the average % Phonetic Accuracy (PA) for TIMIT and CMU-ARCTIC database (BDL, CLB, RMS, and SLT speakers) using eq. (3). The ground truth for the TIMIT database is developed by highly trained annotators [32]. On the other hand, reference phone annotations for the CMU-ARCTIC database are obtained via training of speaker-dependent HMM model over 1132 utterances [33].

Table 1: *% Phonetic Accuracy (PA) of STM algorithm*

	TIMIT Database	**CMU-ARCTIC Database**
% PA	27.53	31.63

2.1. STM with Nearest Neighbor (NN)

The proposed STM algorithm is shown in Algorithm 1. Once the phone boundaries are estimated for source and target speakers' training data, the nearest neighbor (NN) technique is applied to find correspondence between spectral features of source and target speakers among the same labels of frames estimated using the STM algorithm. For the baseline method, we selected the INCA that performs iteratively three steps, namely, a nearest neighbor search step, training of a mapping function using JDGMM-based VC and transformation step using the JDGMM-based until the convergence [7, 8].

Figure 3 shows the % Phonetic Accuracy (PA) calculated using the proposed technique and the baseline algorithm for four different CMU-ARCTIC database pairs, such as BDL-RMS (male-male), CLB-SLT (female-female), BDL-SLT (male-female), and CLB-RMS (female-male). From Figure 3, it is clear that proposed STM+NN technique is giving better performance compared to the baseline algorithm in all the four cases. The proposed STM+NN algorithm is having on an average 13.67 % relative improvement in % PA compared to the baseline algorithm.

Algorithm 1 Proposed STM-based Algorithm for VC Task

Input: Speech wav file and the corresponding utterance from both the source and target speakers.
Preprocessing of silence removal from speech file.
MFCC feature extraction from the audio file.
Extraction of a STM contour at frame-level.
NE: Number of estimated STM boundaries using peak detection.
NG: Number of ground truth boundaries from a given utterance.
NI: Number of insertions.
ND: Number of deletions.
if $NE < NG$ **then**
    $NI = NG - NE$.
    while $NI \neq 0$ **do**
        find two furthest neighbors estimated boundaries.
        Insert boundary based on average phone duration.
        $NI = NI - 1$
    done
else if $NE > NG$ **then**
    $ND = NE - NG$.
    while $NI \neq 0$ **do**
        Find two nearest neighbors estimated boundaries.
        Merge them by selecting either left or right.
        ND=ND-1
    done
end
Estimate boundaries for both source and target speakers' training speech utterances.
Apply Nearest Neighbor (NN) for a given phoneme.
Estimate the unique aligned pairs between source and target speakers from the NN path.
Train the mapping function using the obtained aligned pairs.

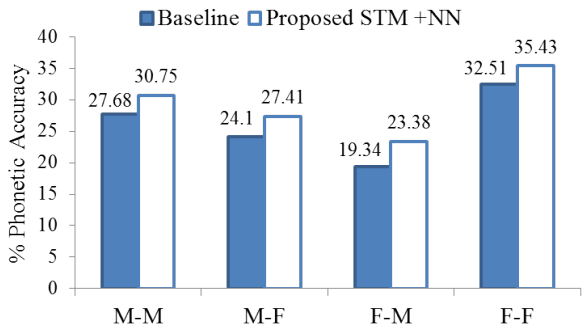

Figure 3: *% PA obtained for different non-parallel VC systems using the baseline and the proposed alignment techniques.*

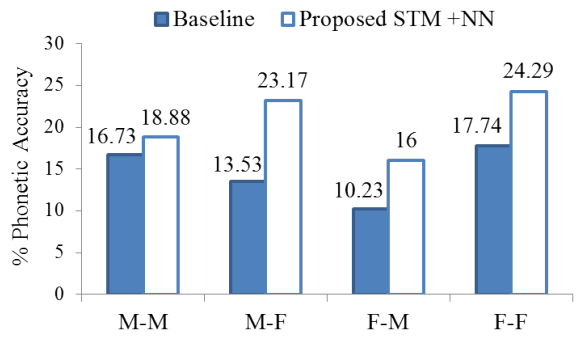

Figure 4: *% PA obtained for vowels for different non-parallel VC systems using the baseline and the proposed alignment techniques.*

Figure 4 shows the % PA for vowel sounds using the baseline and proposed alignment method. The effectiveness of the STM to detect the vowel sounds are indeed observed from the Figure 4. In particular, we obtained on an average 44.36 % absolute increment in correct vowel detection. In the context of VC, it has been reported in the literature that quality of converted voice primarily depend on how accurately voiced (In particular, vowels) sounds are converted [34, 35]. Hence, obtaining better alignment in the vowels with the proposed technique should lead to the high quality converted voices. In addition, the baseline algorithm takes more time compared to the proposed alignment technique due to the iterative computation (approximately, number of iterations times the time taken by the proposed STM+NN algorithm).

3. Experimental Results

In this paper, various VC systems have been developed to measure the effectiveness of the proposed alignment task. We have used 40 non-parallel utterances for each speaker-pairs (whose details are presented in Section 2.1) from the CMU-ARCTIC database. Among the available VC techniques, the state-of-the-art methods, namely, Joint Density (JD) GMM-based VC [36] and BiLinear Frequency Warping plus Amplitude Scaling (BLFW+AS) [37] have been selected. The JDGMM-based method is selected, since it uses conditional expectation, which is the best minimum mean square error (MMSE) estimator. Hence, it leads to the minimum error between converted and the target spectral features. In addition, BLFW+AS has been selected due to its available parametric formulation [37]. *25*-D Mel Cepstral Coefficients (MCC) and *1*-D F_0 per frame (with *25* ms frame duration, and *5* ms frame shift) have been extracted. The number of mixture components have been optimized from the set *m=8,16, 32, 64, 128*. System having an optimum MCD, is selected for subjective evaluation. Here, fundamental frequency (i.e., F_0) contour is transformed using Mean-Variance (MV) transform method [38]. The AHOCODER is used for the analysis-synthesis [39].

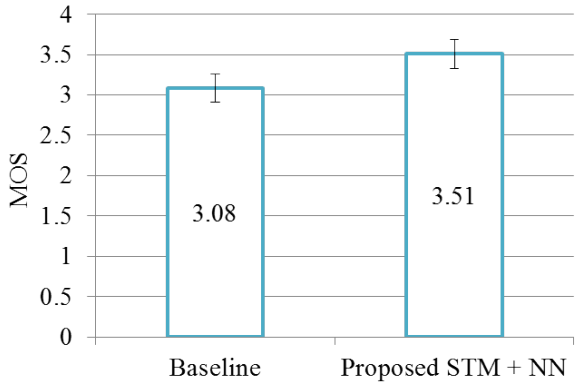

Figure 5: *MOS analysis. Here, we obtained 0.175 margin of error corresponding to the 95 % confidence interval.*

The Mean Opinion Score (MOS) and XAB tests have been selected to evaluate speech quality (i.e., naturalness) and Speaker Similarity (SS) of converted voice, respectively. Since the key goal is to show the effectiveness of the proposed alignment technique over the baseline INCA, we presented average results of the different VC systems w.r.t. the proposed and the baseline alignment strategies. Total *17* subjects (*4* females and *13* males without any hearing impairments with the age between 17 to 28 years) took part in both the tests. Subjects were asked to evaluate the randomly played utterances from both the approaches for the speech quality on the scale of 1 (very bad) to 5 (very good). It can be seen from Figure 5 that proposed STM+NN is more preferred than the baseline in terms of *speech quality*. The result clearly indicates that accurate estimation of phone boundary (especially for the vowel sounds) is indeed helping the NN-based technique to obtain high quality converted voice, which is in line with the study reported in [34].

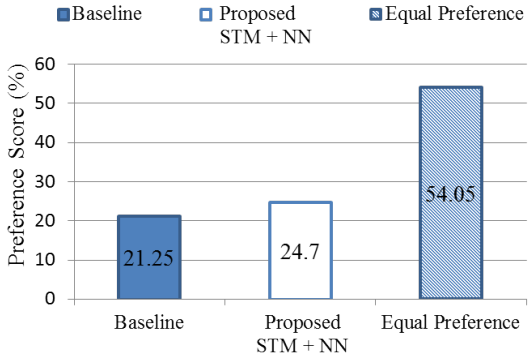

Figure 6: *XAB test analysis for speaker similarity.*

In XAB test, subjects were asked to select from the randomly played *A* and *B* samples (generated with the baseline and the proposed STM+NN algorithm) which one is more similar with the target speaker in terms of *speaker similarity* (SS) with reference to the actual target sample X. In addition, the subjects can select equal preference for the cases, where the samples are perceptually similar in terms of speaker similarity. Samples for XAB test, were taken from both the approaches. It is observed from Figure 6 that the proposed alignment technique is 3.45 % times absolutely more preferred for the speaker similarity of the converted voice. However, 54.05 % times subjects have given equal preference to both the approaches.

4. Summary and Conclusions

In this paper, we exploit the phonetic information via computationally simple STM-based algorithm in NN-based alignment techniques for the non-parallel VC. In particular, we proposed the novel STM+NN-based algorithm for the task of alignment in the case of text-independent VC. The key advantage of the proposed method is that it does not require any a priori training data to estimate the phone boundaries. The % PA obtained after alignment using STM algorithm is found to be better compared to the baseline NN-based alignment technique in all the cases. In particular, % PA obtained for the vowel speech sound class is more using the proposed STM algorithm. The better performance in the alignment task has resulted positively in the context of subjective test for the developed VC systems using proposed approach. Our future work will be directed towards extending this work in the cross-lingual VC task.

5. Acknowledgments

We would like to thank authorities of DA-IICT Gandhinagar, India and Ministry of Electronics and Information Technology (MeitY), Govt. of India for their kind support to carryout this research work.

6. References

[1] 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.

[2] Fuming Fang et al., "High quality nonparallel voice conversion based on cycle-consistent adversarial network," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5279–5283.

[3] Nirmesh 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.

[4] Yuki Saito et al., "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *ICASSP*, Calgary, Canada, 2018, pp. 5274–5278.

[5] 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.

[6] S. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 8, pp. 65–82, 2017.

[7] 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.

[8] Nirmesh J. Shah and Hemant A. Patil, "On the convergence of INCA algorithm," in *APSIPA-ASC*, 2017, pp. 559–562.

[9] 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.

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

[11] Nirmesh J. Shah and Hemant A. Patil, "Novel metric learning for non-parallel voice conversion," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[12] Prasanta Kumar Ghosh and Shrikanth S. Narayanan, "Closure duration analysis of incomplete stop consonants due to stop-stop interaction," *JASA*, vol. 126, no. 1, pp. 1–7, 2009.

[13] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *ICME*, Seattle, USA, 2016, pp. 1–6.

[14] Hiroyuki Miyoshi, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1268–1272.

[15] Jing-Xuan Zhang, Zhen-Hua Ling, Yuan Jiang, Li-Juan Liu, Chen Liang, and Li-Rong Dai, "Improving sequence-to-sequence acoustic modeling by adding text-supervision," in *ICASSP*, Brighton, UK, 2019.

[16] Sadaoki Furui, "On the role of spectral transition for speech perception," *JASA*, vol. 80, no. 4, pp. 1016–1025, 1986.

[17] Sorin Dusan and Lawrence R. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries.," in *INTERSPEECH*, Pittsburgh, USA, 2006, pp. 17–21.

[18] Bhavik B. Vachhani and Hemant A Patil, "Use of PLP cepstral features for phonetic segmentation," in *IALP*, Urumqi, China, 2013, pp. 143–146.

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

[20] Hemant A. Patil et al., "Algorithms for speech segmentation at syllable-level for text-to-speech synthesis system in gujarati," in *Oriental COCOSDA*, New Delhi, India, 2013, pp. 1–7.

[21] Bhavik B Vachhani, Chitralekha Bhat, and Sunil Kopparapu, "Robust phonetic segmentation using multi-taper spectral estimation for noisy and clipped speech," in *EUSIPCO*, Budapest, Hungary, 2016, pp. 1343–1347.

[22] Bhavik B. Vachhani, C. Bhat, and Sunil Kopparapu, *Phonetic Segmentation Using Knowledge from Visual and Perceptual Domain*, Ekštein et. al. (Eds), Lecture Notes in Computer Science (LNCS), Springer, Text, Speech, and Dialogue (TSD), vol. 10415, pp. 393–401, Prague, Czech Republic, 2017.

[23] Maulik C Madhavi, Hemant A Patil, and Bhavik B. Vachhani, "Spectral transition measure for detection of obstruents," in *EUSIPCO*, Nice, France, 2015, pp. 330–334.

[24] Gunnar Fant, *Speech Sounds and Features*, The MIT Press, 1973.

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

[26] Bahar Khalighinejad, Guilherme Cruzatto da Silva, and Nima Mesgarani, "Dynamic encoding of acoustic features in neural responses to continuous speech," *Journal of Neuroscience*, vol. 37, no. 8, pp. 2176–2185, 2017.

[27] Diane Kewley-Port, David B Pisoni, and Michael Studdert-Kennedy, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *The J. of the Acoust. Soc. of Amer.*, vol. 73, no. 5, pp. 1779–1793, 1983.

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

[29] Erich Seifritz, Fabrizio Esposito, Franciszek Hennel, Henrietta Mustovic, et al., "Spatiotemporal pattern of neural processing in the human auditory cortex," *Science*, vol. 297, no. 5587, pp. 1706, 2002.

[30] Aage Moller, *Auditory Physiology*, Elsevier, first edition, 2012.

[31] 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.

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

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

[34] D Childers, B Yegnanarayana, and Ke Wu, "Voice conversion: Factors responsible for quality," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florida, USA, 1985, pp. 748–751.

[35] 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.

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

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

[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.