QUALITY ASSESSMENT OF VOICE CONVERTED SPEECH USING ARTICULATORY FEATURES

Avni Rajpal[1], Nirmesh J. Shah[1], Mohammadi Zaki[2] and Hemant A. Patil[1]

[1]DA-IICT, Gandhinagar-382007, India and [2]IISc, Bangalore 560012, India
[1]{avni_rajpal, nirmesh88_shah and hemant_patil}@daiict.ac.in and [2]zaki@ece.iisc.ernet.in

ABSTRACT

We propose a novel application of the *acoustic- to-articulatory inversion (AAI)* towards a quality assessment of the *voice converted* speech. The ability of humans to speak effortlessly requires the coordinated movements of various articulators, muscles, etc. This effortless movement contributes towards a *naturalness*, *intelligibility* and *speaker's identity* (which is partially present in voice converted speech). Hence, during voice conversion (VC), the information related to the speech production is lost. In this paper, this loss is quantified for a male voice, by showing an increase in *RMSE error* (up to ***12.7 %*** in tongue tip) for voice converted speech followed by showing a decrease in *mutual information (I)* (by ***8.7 %***). Similar results are obtained in the case of a female voice. This observation is extended by showing that the articulatory features can be used as an *objective measure*. The effectiveness of the proposed measure over MCD is illustrated by comparing their correlation with a *Mean Opinion Score* (MOS). Moreover, the preference score of MCD contradicted ABX test by *100 %*, whereas the proposed measure supported ABX test by *45.8 %* and *16.7 %* in the case of female-to-male and male-to-female VC, respectively.

Index Terms**—** voice conversion, acoustic-to-articulatory inversion, articulatory features.

1. INTRODUCTION

Voice Conversion (VC) modifies the perceived speaker's identity from a source-to-target speaker in a given speech utterance [1-3]. During VC, some of the important details in the speech signal are lost due to inaccurate spectral mapping and statistical averaging (i.e., *oversmoothing*) of acoustic speech sound units. Investigating the evaluation measure that truly quantifies the naturalness and speaker similarity of a voice converted speech is still an open research problem [2]. Subjective measures are time-consuming, expensive and their accuracy highly depends on the *cognitive* factors (such as alertness) of the listener [4]. Objective measures, on the other hand, often lack the intuitiveness as well as do not account for the perceptual quality [5].

Machine generated speech, i.e., any computational way of producing the speech signal can never match the way humans articulate to produce speech [6], [7]. In addition, the quality and intelligibility of a voice converted speech are governed mainly by the accurate production of vowels, dynamic or transitional sounds (such as diphthongs, liquids, glides and stops) [8]. Thus, the study of an articulatory parameters (those which are critical in the production of these sounds) could be useful in the voice quality measurement [9], [10], [11]. This idea motivated authors to investigate the difference between a voice converted speech and a natural speech in terms of articulatory parameters. To the best of authors' knowledge, this is in contrast to the previous objective measures which measure the quality in terms of information loss in the *spectral characteristics* during VC [3], [12-14]. The effectiveness of articulatory features has been shown in various applications such as visual aids for training speech [15], speaker recognition [16], speech recognition [17], accent conversion [18], etc. The VC is an another application where the possibility of using articulatory parameters have been explored. However, it appeared that the use of articulatory parameters was not straightforward for improving VC [19].

In this paper, we investigate the novel application of articulatory features for the quality assessment of a voice converted speech. This study investigates the following questions: 1) whether the articulatory information is lost during the VC process? and if so, 2) how can one quantify the information loss? To address this, we propose a novel *Estimation Error (EE),* an articulatory features-based *objective measure*. The subjective score, i.e., Mean Opinion Score (MOS) was taken to evaluate the VC systems. The high correlation coefficient between *EE* and MOS showed the effectiveness of the proposed measure over state-of-the-art *Mel Cepstral Distance* (MCD) measure [10]. Moreover, the preference scores also showed that *EE* is more reliable than the MCD. In particular, MCD contradicted ABX test whereas *EE* supported it to the large extent.

2. EXPERIMENTAL SETUP

This Section briefly discusses the state-of-the-art techniques which are used to develop VC and acoustic-to-articulatory inversion (AAI) systems.

The authors would like to thank Dept. of Electronics and Information Technology (DeitY), Govt. of India, for sponsored project, Development of Text-to-Speech (TTS) System in Indian Languages (Phase-II) and the authorities of DA-IICT, Gandhinagar, India.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

2.1. MOCHA Database

The Multichannel Articulatory (MOCHA) database [20] consists of a simultaneously recorded (*460* phonetically diverse British English TIMIT sentences) acoustic and articulatory data obtained from one male and one female speaker. The audio signal is sampled at *16* kHz and Electromagnetic Articulography (EMA) data is sampled at *500* Hz. The EMA data consists of X and Y coordinates of *9* receiver sensor coils attached to *9* points along the midsaggital plane, namely, the lower incisor or the jaw (*li_x, li_y*), upper lip (*ul_x, ul_y),* lower lip (*ll_x, ll_y*), tongue tip (*tt_x, tt_y*), tongue body (*tb_x, tb_y*), tongue dorsum (*td_x,td_y*), velum (*v_x, v_y*), upper incisor (*ui_x, ui_y*) and bridge of the nose (*bn_x, bn_y*). The upper incisor and bridge of the nose are used as a reference coils. The articulatory data obtained from *14* channels corresponding to first seven coils except the reference coils are used as the articulatory features in our experiments.

2.2. Voice Conversion (VC) System

Among the various available VC techniques [1-3], [21-23] here, a GMM-based [3] and a BiLinear Frequency Warping plus Amplitude Scaling (BLFW+AS) [21] methods were used for transforming the spectral parameters. In a GMM-based VC, joint source and target spectral feature vectors were modeled using a GMM and then conversion was performed using a maximum likelihood estimation (MLE) [3]. On the other hand, nonlinear BLFW technique transforms the frequency-axis of the source-to-target speaker's vocal-tract spectrum and AS method was used to transform the relative amplitude of the spectrum for spectral parameter conversion [21]. The excitation source parameter (i.e., F_0) was transformed using a mean-variance method in the log-domain [24].

2.3. Acoustic-to-Articulatory Inversion (AAI) System

Among the various available AAI techniques [15], [25-26] here, Generalized Smoothness Criterion (GSC)-based, AAI system is used for the articulatory parameterization of a voice converted speech [26]. The estimated trajectories obtained using GSC were *optimal* in the sense that a) the estimated trajectories have minimum energy in the high-frequency region and b) the weighted difference between estimated and original trajectories was minimum. GSC has the advantage that it imposes the articulator-specific constraints which gives a better estimation over methods using a fixed smoothness constraints [26].

3. PROPOSED OBJECTIVE MEASURE

The experiments were conducted to verify and quantify the possible loss of an articulatory information after VC. For this, a GMM-based VC system with *400* training utterances and *64* mixture components was used. Let the target and the voice converted acoustic vector be given by $\mathbf{X_t}$ and $\mathbf{X_{tv}}$, respectively. Furthermore, let EMA vector of the target be $\mathbf{Y_t}$ and estimated EMA vector from $\mathbf{X_t}$ and $\mathbf{X_{tv}}$ be $\mathbf{Z_t}$ and $\mathbf{Z_{tv}}$, respectively.

In order to verify the loss in speech production information after VC, mutual information (*I*) was computed [27]. Since $\mathbf{X_t}$, $\mathbf{Y_t}$ and $\mathbf{X_{tv}}$ are discrete, their probability distributions are calculated by quantizing the acoustic and the articulatory spaces using K-means clustering algorithm (*K=64*) [28]. Mutual Information *(I)* calculated between $(Q(\mathbf{X}_t), Q(\mathbf{Y}_t))$ and $(Q(\mathbf{X}_{tv}), Q(\mathbf{Y}_t))$ is shown in Table 1. Here, $Q(\mathbf{X}_t), Q(\mathbf{Y}_t)$ are quantized acoustic and articulatory spaces, respectively, and $Q(\mathbf{X}_{tv})$ is quantized voice converted acoustic space. Table 1 shows that the information related to the articulators in acoustic vector reduces after VC both for male (i.e., the target is a male) and female (i.e., the target is a female) voice converted speech.

Table 1: Comparison of Mutual Information Before and After VC

I (in bits)	Male Voice	Female Voice
$I(Q(\mathbf{X}_t), Q(\mathbf{Y}_t))$	1.402	1.504
$I(Q(\mathbf{X}_{tv}), Q(\mathbf{Y}_t))$	**1.28**	**1.389**

The following steps were used to estimate the articulatory parameters of a voice converted speech (which is illustrated in Fig. 1) in order to quantify above mentioned loss.

- $\mathbf{Z_{tv}}$ and $\mathbf{Z_t}$ were estimated using GSC-based technique.
- $\mathbf{Z_{tv}}$, $\mathbf{Z_t}$ and $\mathbf{Y_t}$ were time-normalized (by applying DTW on $\mathbf{X_{tv}}$ and $\mathbf{X_t}$) to obtain $\mathbf{DZ_{tv}}$, $\mathbf{DZ_t}$ and $\mathbf{DY_t}$, respectively.
- The estimation accuracy for each articulator position was compared by computing % Δ given by :

$$\text{\% change } (\Delta) = \frac{RMSE_{tv} - RMSE_{tt}}{RMSE_{tt}} \times 100, \qquad (1)$$

where RMSE_{tt} is the average root mean square error (RMSE) calculated between $\mathbf{DY_t}$ and $\mathbf{DZ_t}$ and RMSE_{tv} is an average RMSE between $\mathbf{DY_t}$ and $\mathbf{DZ_{tv}}$.

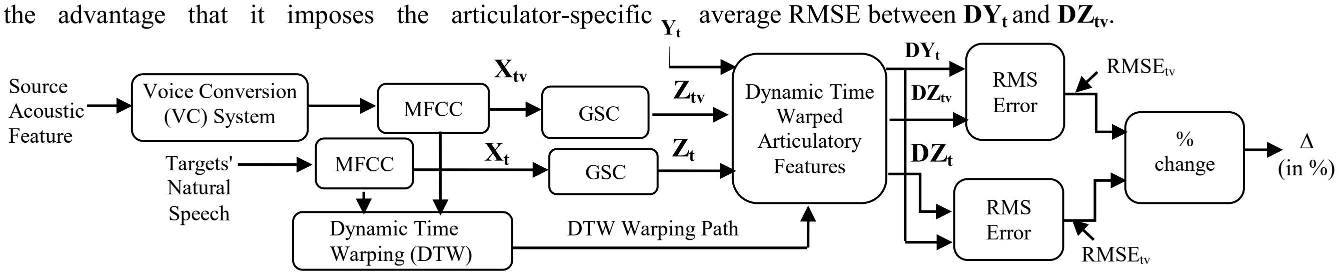


Fig. 1: Proposed system architecture for estimating articulatory features from voice conversion (VC) system.

Table 2: Comparison of an average RMSE in mm (along with standard deviation (SD) of RMSE is shown in the bracket). The dotted box indicates maximum % Δ (i.e., tongue tip is not estimated accurately compared to all other articulators)

	Articulators	li_x	li_y	ul_x	ul_y	ll_x	ll_y	tt_x	tt_y	tb_x	tb_y	td_x	td_y	v_x	v_y
Male Voice	$RMSE_{tt}$ (SD)	0.6 (0.1)	1.11 (0.3)	0.77 (0.2)	1.36 (0.2)	1.28 (0.3)	2.09 (0.4)	2.83 (0.8)	3.5 (0.8)	2.66 (0.6)	2.56 (0.5)	2.35 (0.6)	2.67 (0.5)	0.56 (0.2)	1.19 (0.5)
	$RMSE_{tv}$ (SD)	0.63 (0.1)	1.21 (0.2)	0.81 (0.2)	1.47 (0.3)	1.39 (0.3)	2.35 (0.4)	3.19 (1)	3.87 (0.7)	2.91 (0.7)	2.86 (0.6)	2.58 (0.7)	2.94 (0.6)	0.62 (0.2)	1.29 (0.5)
	%Δ	5	9	5.2	8.1	8.6	12.4	12.7	10.6	9.4	11.7	9.8	10.1	10.7	8.4
Female Voice	$RMSE_{tt}$ (SD)	0.87 (0.2)	1.36 (0.3)	1.01 (0.4)	1.36 (0.3)	1.32 (0.3)	2.92 (0.6)	2.72 (0.6)	2.89 (0.6)	2.49 (0.5)	2.61 (0.5)	2.29 (0.5)	2.7 (0.5)	0.45 (0.2)	0.49 (0.2)
	$RMSE_{tv}$ (SD)	0.93 (0.2)	1.5 (0.3)	1.1 (0.4)	1.41 (0.3)	1.42 (0.3)	3.22 (0.7)	3.2 (0.7)	3.36 (0.6)	2.88 (0.6)	2.99 (0.5)	2.61 (0.6)	2.94 (0.4)	0.52 (0.2)	0.54 (0.2)
	%Δ	6.9	10.3	8.9	3.7	7.6	10.3	17.6	16.3	15.7	14.6	14	8.9	15.6	10.2

Table 2 shows that $RMSE_{tv}>RMSE_{tt}$ for both male and female voice converted speeches, which is indicated by positive % Δ for all the articulators. In particular, among all the articulators, tongue tip (known to be critical for the speech production [7]) shows highest *%* Δ.

The results indicate that the AAI system poorly estimates the articulatory trajectories of a voice converted speech. The difference in the estimation accuracy is utilized to propose the *Estimation Error (EE)*, as an objective measure. The *EE* measures the distance between articulatory trajectories of voice converted speech and the target speech. Estimation error (*EE*) (in *mm*), is defined as:

$$EE = \frac{1}{N}\left(\sum_{n=1}^{N}\sqrt{\sum_{d=1}^{M}\left(\mathbf{DZ}_{\mathbf{tv}_d}^{n} - \mathbf{DY}_{\mathbf{t}_d}^{n}\right)^2}\right), \qquad (2)$$

where for n^{th} frame, $\mathbf{DY}_{\mathbf{t}_d}^{n}$ and $\mathbf{DZ}_{\mathbf{tv}_d}^{n}$ are the time–aligned d^{th}-dimensional measured and estimated trajectory, respectively. In addition, *N* is the length and *M* is the dimensionality of the articulator trajectory.

4. EXPERIMENTAL RESULTS

4.1. Details of VC and AAI system

The female speech is well known to have a *spectral* resolution problem (due to the serious interaction of high pitch (F_0) source harmonics with vocal tract spectrum) [30]. Furthermore, the female speech has relatively small pitch period (T_0) due to the lesser mass of vocal folds than the male counterpart (in the range of *4-5 ms*). As a result, the female speakers are possibly not able to produce as much glottal activity (such as manner in which vocal folds open or close) as compared to the male speaker. Hence, it is known that the male-to-female (M-F) VC is more difficult [29-30]. Here, the VC systems based on GMM and BLFW+AS were built for both M-F and female-to-male (F-M) cases. For this, the number of training utterances (i.e., *10, 25, 50, 200* and *400)* and the number of mixtures in GMM (*i.e., m=8, 16, 32, 64)* were varied. *24-D* Mel Generalized Cepstral (MGC) coefficients were extracted from the speech signals over *25 ms* window with *5 ms* shift for both VC approaches. The training sentences were selected based on maximum diphone coverage [3]. For AAI, out of *400* (from 460 MOCHA-TIMIT) sentences used for training of VC system, *368* sentences for the development set and *55* for test set were used. *14-D* MFCC was calculated per frame (of *20 ms* window with a frameshift of *10 ms*) for inversion. AAI systems were built for both male and female voices.

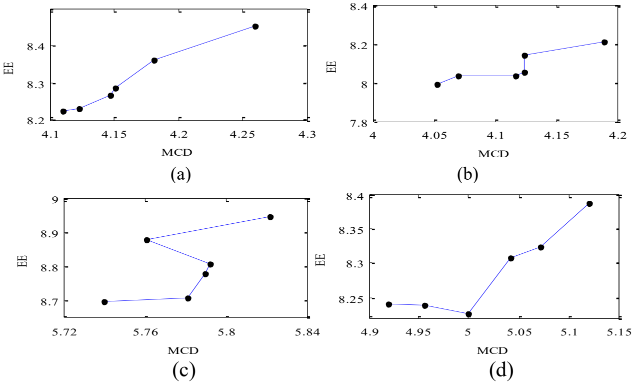
The accuracy of AAI system is measured by calculating an average RMSE and an average correlation coefficient (CC) [26]. Our AAI system shows the lowest estimation accuracy for *ll_y* (average RMSE=*2.92*, average CC=*0.74*) and highest for *v_x* (average RMSE=*0.45*, average CC= *0.70*) in case of a female. For a male, the estimation accuracy is the lowest for *tt_y* (average RMSE=*3.5*, average CC=*0.70*) and highest for *v_x* (average RMSE=*0.56*, average CC=*0.64*).

4.2. Evaluation of VC systems

For a given training utterance set, the one showing the least MCD for different values of *m* was selected for subjective evaluation. This was carried out for GMM and BLFW+AS-based M-F and F-M VC systems. The following sub-Section discusses the analysis of subjective and objective measures

4.2.1. Correlation of EE with objective measure

Fig. 2 shows plot between EE and MCD for the selected systems. These plots indicate that EE and MCD are *partially* correlated. In particular, Fig. 2 (a)-(b) show that *EE* and MCD correlate well for GMM–based VC as compared to BLFW+AS-based VC (as shown in Fig. 2(c)-(d)).



(a) (b) (c) (d)

Fig.2: MCD *vs.* plot for selected systems (a)-(b) M-F and F-M GMM-based VC and (c)-(d) M-F and F-M BLFW+AS-based VC.

One of the possible reasons for such a high correlation could be that articulatory parameters were estimated from acoustic features itself. However, two sounds that are closer in cepstral or acoustic-domain may not be close in articulatory-domain, because AAI is a non-unique and nonlinear [15], [26]. Moreover, it is known that MCD may not always correlate well with subjective scores [31-32]. Therefore, the differences in articulatory-domain were exploited for determining the quality of the VC systems. To verify this, the Pearson correlation coefficient (CC) of MCD, *EE* with subjective measures were calculated.

4.2.2. Comparison of *EE* with subjective measures

For a subjective measure, MOS from *15* subjects (*9* male and *6* female with age group of *21-25* years) was taken for absolute rating [33]. In this test, we randomly played *4* sentences from each VC system (selected for evaluation). The subjects were asked to score them on a *5*-point scale based on the naturalness of speech signal. CC of MCD and *EE* with MOS score was calculated using the Pearson Correlation Coefficient. MOS score, MCD and *EE* for selected systems along with their CC are shown in Table 3 and Table 4, respectively.

Table 3: Subjective and objective scores of various VC systems

Approach	Systems*	M-F VC			F-M VC		
		MOS	MCD	EE	MOS	MCD	EE
BLFW+AS	10_64	2.45	5.66	7.60	2.35	4.87	8.05
	25_64	2.65	5.65	7.68	2.45	4.84	7.72
	50_64	2.53	5.71	7.59	2.33	4.97	7.90
	100_64	2.63	5.99	7.96	2.68	5.36	8.0
	200_64	2.4	6.09	8.17	2.63	5.26	8.29
	400_64	2.33	5.89	8.11	2.6	5.12	8.03
GMM	10_32	2.48	3.97	7.76	2.1	3.98	7.28
	25_32	2.3	4.04	7.29	2.2	3.92	6.92
	50_64	2.53	3.80	7.42	2.15	3.93	7.12
	100_64	2.53	4.24	7.61	2.18	4.16	7.03
	200_64	2.23	4.08	7.76	2.3	4.09	7.36
	400_64	2.35	4.235	7.438	2.225	4.09	7.04

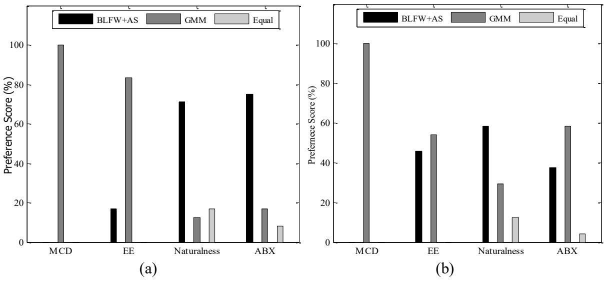
*Systems: Number of training utterances_mixture components

Table 4: Correlation coefficients of MCD and *EE* with MOS

ObjectiveMeasure	GMM		BLFW+AS	
	M-F	F-M	M-F	F-M
MCD	-0.16	0.41	-0.33	0.87
EE	**-0.7**	**0.16**	**-0.5**	**0.46**

Ideally, subjective and objective scores should have a negative correlation. Since higher the MOS better is the quality of speech, as opposed to MCD and *EE* where higher the score lesser is the quality. Table 4 shows that for M-F MCD and *EE* are showing negative CC for both VC techniques. It can be seen from Table 4 that *EE* is more negatively correlated with MOS in the case of M-F. On the other hand, in the case of F-M, *EE* is relatively less positively correlated compared to MCD. Hence, it is found that the interpretation of a quality given by *EE* was more preferable over MCD measure. While conducting the MOS test, it was observed that there were minute perceptual differences within VC systems used for evaluation and subjects had difficulty in giving MOS scores. This is evident from Table III, which indicates a very small change in MOS scores. In order to avoid this ambiguity, ABX test was conducted with the same *15* subjects where *24* utterances were played randomly from both approaches of VC. In this test, the subjects were asked to choose between A and B based on the naturalness and similarity of the utterance compared to the target sample X. The scores of this test are indicated as ABX and naturalness in Figure 3 (a)-(b). On the similar lines, we also calculated the preference scores of these utterances using MCD and *EE*. Out of A and B that gave least value of MCD and *EE*, was preferred and the preference score (in %) are shown in Figure 3(a)-(b).

From Figure 3(a)-(b), it can be seen that for both M-F and F-M VC, MCD gave *100 %* preference to GMM-based VC method. However, unlike MCD which gave *0 %* preference to BLFW+AS VC system, *EE* gave *45.8 %* preference score in case of F-M and *16.67 %* preference score in the case of M-F. Hence, *EE* is relatively more reliable than MCD, which completely nullifies the possibility of BLFW+AS to be better in any case. Thus, we proposed the *EE* as an objective measures for assessing the quality of VC.



(a)

(b)

Fig. 3: Preference score based on MCD, *EE*, naturalness and ABX test for GMM and BLFW VC systems (a) M-F (b) F-M. Equal means, subjects could not judge and give equal preference score.

5. SUMMARY AND CONCLUSIONS

This study investigated the objective measure which is based on the articulatory parameters. In particular, after VC, the articulatory parameters related information is lost which is quantified by a proposed objective measure, namely, *EE*. Though MCD and *EE* were found to be partially correlated and gave almost a similar kind of interpretation, *EE* had more correlation with MOS. The experiments showed that in the case of preference score, where MCD was *100 %* contradicting subjective measure, which is highly unlikely. On the other hand, *EE* supported subjective measure *45.8 %* and *16.67 %* for F-M and M-F VC, respectively. Hence, the proposed measure *EE* is a reliable objective measure for measuring the quality of a voice converted speech. Our future research efforts will be directed towards investigating the articulators that are more responsible for capturing the voice quality of VC speech.

6. REFERENCES

[1] Y. Stylianou, O. Cappe and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. on Speech and Audio Proc.*, vol. 6, no. 2, pp. 131-142, 1998.

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

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

[4] H. B. Sailor, and H. A. Patil, “Fusion of magnitude and phase-based features for objective evaluation of TTS voice,” in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Singapore, pp. 521-525, 2014.

[5] T. Ganchev, A. Lazaridis, I. Mporas, and N. Fakotakis, “Performance evaluation for voice conversion systems,” in *Text, Speech and Dialogue,* Berlin, pp. 317-324, 2008.

[6] S. Aryal, and R. G. Osuna. “Articulatory inversion and synthesis: towards articulatory-based modification of speech,” in *International Conference on Acoustics, Speech and Signal Processing*, Canada, 2013, pp. 7952-7956.

[7] A. W. Black, et al. “Articulatory features for expressive speech synthesis,” in *International Conference on Acoustics, Speech and Signal Processing*, Japan, 2012, pp. 4005-4008.

[8] D. G. Childers, B. Yegnanarayana, and K. Wu, “Voice conversion: Factors responsible for quality,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1985, vol. 1, pp. 748–751.

[9] J. Wang, J. R. Green, and A. Samal, “Individual articulator’s contribution to phoneme production”, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, pp. 7785-7789, 2013.

[10] P. J. Jackson and V. D. Singampalli, “Statistical identification of articulation constraints in the production of speech,” *Speech Communication*, vol. 51, no. 8, pp. 695 – 710, 2009.

[11] P. Ladefoged and K. Johnson, *A Course in Phonetics, Independence*, K Y Cengage Learning, *6th* Edition, 2011.

[12] R. F. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Victoria, British Columbia, Canada, pp. 125-128, 1993.

[13] O. Tu¨rk and M. Schro¨der, “A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis,” in *INTERSPEECH*, Brisbane, Australia, pp. 2282–2285, 2008.

[14] P. Lanchantin and X. Rodet, “Objective evaluation of the dynamic model selection method for spectral voice conversion,” in in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Czech Republic, pp.5232-5135, 2011.

[15] K. Richmond, “Estimating Articulatory Parameters from the Acoustic Speech Signal,” Ph.D. Dissertation, Edinburgh Univ., Centre Speech Technol. Res., Edinburgh, U.K., 2002.

[16] M. Li, J. Kim, P. Ghosh, V. Ramanarayanan, and S. Narayanan, “Speaker verification based on fusion of acoustic and articulatory information,” in *INTERSPEECH*, Lyon, France, pp. 1614– 1618, 2013.

[17] P. K. Ghosh and S. Narayanan, “Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion,” *J. Acoust. Soc. Amer.*, vol. 130, no. 4, pp. 251-257, 2011.

[18] S. Aryal and R. Osuna, “Articulatory-based conversion of foreign accents with deep neural networks,” in *INTERSPEECH*, Dresden, Germany, pp. 3385-3389, 2015.

[19] A. Toth and A. Black, “Using articulatory position data in voice transformation,” in *ISCA Speech Synthesis Workshop (SSW6)*, Bonn , Germany, pp. 182-187, 2007.

[20] A. Wrench and W. Hardcastle, “A multichannel articulatory speech database and its application for automatic speech recognition,” in *Proc. 5th Seminar of Speech Production (SSP)*, Kloster Seeon, Germany, pp. 305–308, 2000.

[21] D. Erro, E. Navas, and I. Hernaez, “Parametric voice conversion based on bilinear frequency warping plus amplitude scaling,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556-566, 2013.

[22] E. Godoy, R. Olivier, and C. Thierry. “Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora,” *IEEE Trans. on Audio, Speech, and Lang. Proces.* vol. 20, no. 4, pp. 1313-1323.

[23] N. J. Shah and H. A. Patil, “Novel amplitude scaling method for bilinear frequency warping-based voice conversion”, accepted in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017.

[24] D. T. Chappell and J. Hansen. “Speaker-specific pitch contour modeling and modification,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, WA, vol. 2, pp. 885-888, 1998.

[25] A. Rajpal and H. A. Patil, “Jerk Minimization-Based Acoustic-to-Articulatory Inversion,” in *ISCA Speech Synthesis Workshop (SSW9)*, USA, 2016, pp. 87-92.

[26] P. K. Ghosh and S. Narayanan, “A generalized smoothness criterion for acoustic-to-articulatory inversion,” *J. Acoust. Soc. Amer.*, vol. 128, no. 4, pp. 2162–2172, 2010.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, *1st* Edition, 1991.

[28] A. Kraskov, H. Stögbauerand P. Grassberger, “Estimating mutual information,” *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip*. Top., vol. 69, no. 6, pp. 1-16, 2004.

[29] S. Nakagawa, K. Shikano, and Y. Tohkura, “Speech, hearing and neural network models,” *IOS press*, *1st* Edition, 1995.

[30] H. A. Patil, P. K. Dutta and T. K. Basu. “On the Investigation of Spectral Resolution Problem for Identification of Female Speakers in Bengali,” *International Conference on Industrial Technology (ICIT)*, Mumbai, India, pp. 375-380, 2006.

[31] D. Sündermann. "Voice conversion: state-of-the-art and future work," *Fortschritte der Akustik*, pp. 735-736, 2005.

[32] A. Machado and M. Queiroz, “Voice conversion: A critical survey,” *in Proc. Sound and Music Computing*, pp.1-8, 2010.

[33] Int. Telecom Union, “A method for subjective performance assessment of the quality of speech voice output devices,” ITU-T Rec., P.85, 1994.