QUALITY ASSESSMENT OF VOICE CONVERTED SPEECH USING ARTICULATORY FEATURES

Avni Rajpal, Nirmesh J. Shah, Mohammadi Zaki and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar- 382007, India
{avni_rajpal, nirmesh_shah, mohammadi_zaki, hemant_patil}@daiict.ac.in

ABSTRACT

We propose a novel application based on acoustic-to-articulatory inversion (AAI) towards quality assessment of voice converted speech. The ability of humans to speak effortlessly requires coordinated movements of various articulators, muscles, etc. This effortless movement contributes towards naturalness, intelligibility and speaker's identity (which is partially present in voice converted speech). Hence, during voice conversion, the information related to speech production is lost. In this paper, this loss is quantified for male voice, by showing increase in RMSE error (upto 12.7% in tongue tip) for voice converted speech followed by showing decrease in mutual information (I) by 8.7%. Similar results are obtained in case of female voice.

This observation is extended by showing that articulatory features can be used as an objective measure. The effectiveness of proposed measure over MCD is illustrated by comparing their correlation with Mean Opinion Score (MOS). Moreover, preference score of MCD contradicted ABX test by 100%, whereas proposed measure supported ABX test by 45.8% and 16.7% in case of female-to-male and male-to-female voice conversion systems, respectively.

Index Terms— Voice Conversion, Acoustic-to-Articulatory Inversion, Subjective and Objective Measure.

1. INTRODUCTION

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

Machine generated speech, i.e., any computational way of producing speech signal can never match the way humans articulate to produce speech. In addition, quality and intelligibility of voice converted speech is governed mainly by accurate production of vowels, dynamic or transitional sounds (such as diphthongs, liquids, glides and stops) [6]. Thus, study of articulatory parameters (those are critical in production of these sounds) could be useful in voice quality measurement [7], [8], [9]. This idea motivates us to investigate the difference between voice converted speech and natural speech in terms of articulatory parameters. To the best of authors' knowledge, this is in contrast to the previous objective measures which measure quality in terms of information loss in spectral characteristics during VC [3], [10], [11], [12]. Effectiveness of articulatory features have been known in various applications such as visual aids for training speech [13], speaker recognition [14], speech recognition [15], accent conversion [16], etc. In this paper, we investigate the novel application of articulatory features for quality assessment of voice converted speech.

The experiments are presented to investigate following questions: 1) Does articulatory information is lost during the voice conversion process? and if so, 2) How is this loss in information can be quantified? To address this, we propose novel Estimation Error (EE) - an articulatory features-based objective measure. The subjective score, i.e., Mean Opinion Score (MOS) was taken to evaluate individual VC systems. High correlation coefficient between EE and MOS showed effectiveness of proposed measure over state-of-the-art Mel Cepstral Distance (MCD) [10]. Moreover, preference scores also showed that EE is more reliable than MCD. In particular, MCD contradicted ABX test whereas EE supported it to the large extent.

2. EXPERIMENTAL SETUP

This Section briefly discusses about basic framework and techniques used for experiments presented in this paper.

2.1. MOCHA Database

The Multichannel Articulatory (MOCHA) database [17] consists of simultaneously recorded acoustic and articulatory data obtained from one male and one female

Authors would like to acknowledge Department of Electronics and Information Technology (DeitY), New Delhi, Govt. of India for their kind support to carry out this research work. They also thank all participants who took part in subjective evaluation.
speaker. The corpus consists of 460 phonetically diverse British English TIMIT sentences, audio signal sampled at 16 kHz and Electromagnetic Articulography (EMA) data sampled at 500 Hz. EMA data consists of X and Y coordinates of 9 receiver sensor coils attached to 9 points along the mid-sagittal 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 reference coils. Articulatory data obtained from 14 channels corresponding to first seven coils except the reference coils are used as articulatory features in our experiments.

2.2. Voice Conversion (VC) System

Voice conversion technique finds the mapping function between spectral parameters as well as excitation parameters obtained from aligned source and target speech database [3]. In this paper, GMM-based voice conversion and BiLinear Frequency Warping plus Amplitude Scaling (BLFW+AS) methods were used for transforming spectral parameters. In GMM-based VC, joint source and target spectral feature vectors were modeled using GMM and then conversion was performed using maximum likelihood estimation (MLE) [3]. On the other hand, non-linear 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 from source-to-target spectral parameter conversion [18]. Excitation source parameter (i.e., F0) was transformed using z-score (i.e., mean variance transform) mapping in the log-domain [19].

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

In order to have articulatory parameterization of voice converted speech, generalized smoothness criterion (GSC) - based AAI system is used here [20]. The estimated trajectories obtained using GSC were optimal in the sense that a) the estimated trajectories have minimum energy in high frequency region and b) the weighted difference between estimated and original trajectories was minimum. GSC has the advantage that it imposes articulator-specific constraint which gives better estimation over methods using fixed smoothness constraints [20].

3. PROPOSED OBJECTIVE MEASURE

Experiments were conducted to verify and quantify the possible loss of articulatory information after VC. For this, GMM-based VC system with 400 training utterances and 128 mixture components was used. Let target and voice converted acoustic vector be given by X_t and X_n, respectively. Furthermore, let EMA vector of target be Y_t and estimated EMA vector from X_t and X_n be Z_t and Z_n, respectively.

In order to verify the loss in speech production information after VC, mutual information (I) was computed [21]. Since X_t, Y_t and X_n are discrete, their probability distribution is calculated by quantizing acoustic and articulatory spaces using K-means (K=64) clustering algorithm [22]. Mutual Information (I) calculated between (Q(X_t), Q(Y_t)) and (Q(X_n), Q(Y_t)) is shown in Table 1. Here, Q(X_t), Q(Y_t) are quantized acoustic and articulatory spaces, respectively and Q(X_n) is quantized voice converted acoustic space. Table 1 shows that information related to articulators in acoustic vector reduces after VC both for male (i.e., target is male) and female (i.e., target is female) voice converted speech.

Table 1: Comparison of Mutual Information before and after VC.

| I (in bits) | Male Voice | Female Voice |
|------------|------------|--------------|
| I(Q(X_t), Q(Y_t)) | 1.402 | 1.504 |
| I(Q(X_n), Q(Y_t)) | 1.28 | 1.389 |

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

- Z_t and Z_n were estimated using GSC-based AAI system.
- Z_t, Z_t and Y_t were time-normalized using warping path to obtain DZ_t, DZ_t and DY_t, respectively. Warping path was obtained by applying DTW on X_n and X_t.
- The estimation accuracy of estimated features for each articulator position was compared by computing % change given by:

\[
\% \text{ change} \ (\Delta) = \frac{\text{RMSE}_{\text{a}} - \text{RMSE}_{\text{d}}}{\text{RMSE}_{\text{a}}} \times 100
\]

where RMSE_a is average RMSE calculated between DY_t and DZ_t and RMSE_d is average RMSE between DY_t and DZ_n.

![Fig.1: Proposed system architecture for estimating articulatory features from voice conversion (VC) system.](image-url)
4. Experimental Results

4.1. Details of VC and AAI System

Female speech is well known to have spectral resolution problem (due to serious interaction of high pitch source harmonics with vocal tract spectrum). Furthermore, female speech has relatively less pitch period (in the range of 4-5 ms), as a result female speakers are possibly not able to produce as much glottal activity and glottal closure as compared to male speaker. Hence, it is known that male-to-female (M-F) VC is more difficult [23], [24], [25]. Thus both M-F and female-to-male (F-M) VC systems based on GMM and BLFW+AS were built. For this, number of training utterances (i.e., 10, 25, 50, 200 and 400) and number of mixtures in GMM (i.e., \( m = 8, 16, 32, 64 \)) were varied. 24-D Mel Generalized Cepstral (MGC) coefficients were extracted from speech signal over 25 ms window with 5 ms shift for both VC approaches. The training sentences were selected based on maximum dipphone coverage [3].

For AAI, out of 400 (from 460 MOCHA-TIMIT) sentences used for training of VC system, 368 sentences were used as training set. From the remaining sentences, 37 for development set and 55 for test set were used. 14-D MFCC was calculated using 20 ms Hamming window with shift of 10 ms for inversion. AAI systems were built for male voice and female voice.

Accuracy of AAI system: The accuracy of AAI system is measured by calculating average root mean square error (RMSE) and average correlation coefficient (CC) [20]. Our AAI system shows least estimation accuracy for \( l_y \) (average RMSE=2.92, average CC=0.74) and highest for \( v_x \) (average RMSE=0.45, average CC=0.70) in case of female. For male, estimation accuracy is least for \( t_y \) (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 least MCD results 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. Both subjective and objective measures were used to evaluate the selected systems which are discussed in the following sub Section.

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

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

\[
EE = \frac{1}{N} \sum_{n=1}^{N} \left( \sum_{d=1}^{M} DZ_{n} - DY_{n} \right)^2 ,
\]

where for \( n^{th} \) frame, \( DY_{n} \) and \( DZ_{n} \) are the time-aligned \( d^{th} \)-dimensional measured and estimated trajectory, respectively. In addition, \( N \) is the length and \( M \) is the dimensionality of articulator trajectory.

Table 2: Comparison of average RMSE (SD) in mm. Dotted box indicates maximum % \( \Delta \) (i.e., tongue tip is not estimated accurately compared to all other articulators).

| Articulators | li_x | li_y | ul_x | ul_y | li_x | li_y | tt_x | tt_y | tb_x | tb_y | td_x | td_y | v_x | v_y |
|--------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Male Voice   |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| RMSE_{tv} (SD) | (0.1) | (0.1) | (0.2) | (0.2) | (0.3) | (0.3) | (0.4) | (0.4) | (0.5) | (0.5) | (0.5) | (0.5) | (0.5) | (0.5) |
| RMSE_{tv} (SD) | (0.6) | (0.6) | (0.7) | (0.7) | (0.8) | (0.8) | (0.9) | (0.9) | (1.0) | (1.0) | (1.0) | (1.0) | (1.0) | (1.0) |
| %\( \Delta \) | 5.2  | 5.9  | 8.1  | 8.6  | 12.4 | 12.7 | 10.6 | 10.5 | 9.4  | 9.7  | 10.1 | 10.7 | 8.4  | 8.4  |
| Female Voice |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| RMSE_{tv} (SD) | (0.2) | (0.2) | (0.3) | (0.3) | (0.3) | (0.3) | (0.6) | (0.6) | (0.6) | (0.6) | (0.6) | (0.6) | (0.6) | (0.6) |
| RMSE_{tv} (SD) | (0.9) | (0.9) | (1.1) | (1.1) | (1.2) | (1.2) | (1.3) | (1.3) | (1.4) | (1.4) | (1.4) | (1.4) | (1.4) | (1.4) |
| %\( \Delta \) | 6.9  | 10.3 | 8.9  | 7.6  | 10.3 | 10.3 | 9.0  | 9.0  | 8.2  | 8.2  | 7.9  | 7.9  | 6.4  | 6.4  |

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 high correlation could be that articulatory parameters were estimated from acoustic features itself. However, two sounds that are closer in cepstral/acoustic-domain may not be close in articulatory-domain, because AAI is non-unique and non-linear problem [13], [20]. Moreover, it is known that MCD may not always correlate well with subjective scores [26], [27]. Therefore, the differences in articulatory-domain were exploited for determining the quality of the VC systems. To verify this, correlation coefficient (CC) of MCD, EE with subjective measures were calculated.

4.2.2. Comparison of EE with subjective measures
For subjective measure, MOS [28] from 10 subjects (7 male and 3 female with 21-25 years of age) was taken for absolute rating. In this test, we randomly played 4 sentences from each VC system (selected for evaluation). The subjects were asked to score them on 5-point scale based on the naturalness of speech signal. CC of MCD and EE with MOS score was calculated using 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 2.6 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 |

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

| Objective Measure | GMM M-F | F-M M-F | BLFW+AS M-F |
|-------------------|---------|--------|------------|
| MCD               | -0.16   | 0.41   | -0.33      |
| EE                | -0.7    | 0.16   | -0.5       |

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 couldn’t judge and gave equal preference score.

5. SUMMARY AND CONCLUSIONS

This paper investigated articulatory parameters-based objective measure which can be used for quality assessment of voice converted speech. In particular, after VC, articulatory parameters related information is lost. This causes error in estimation of articulatory parameters from voice converted speech. This error was used as the measure to examine the quality of voice converted speech. Though MCD and EE were found to be partially correlated and gave almost similar kind of interpretation, EE had more correlation with MOS. The experiments showed that in case of preference score, where MCD was 100 % contradicting subjective measure which is highly unlikely. On the other hand, our proposed method supported subjective measure 45.8 % in case of F-M VC system and 16.67 % M-F VC system. All these results indicate the possibility of using EE as a reliable objective measure for measuring quality of voice converted speech. Our future research efforts will be directed for investigating articulators that are more responsible for capturing the voice quality.
6. REFERENCES

[1] Y. Stylianou, O. Cappe and E. Moulou, "Continuous probabilistic transform for voice conversion," IEEE Trans. on Speech and Audio Process., 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. Sailor, and H. 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 (TSD): Springer, Berlin Heidelberg, pp. 317-324, 2008.

[6] D. G. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," in Proc. ICASSP, 1985, vol. 1, pp. 748–751

[7] J. Wang, J. R. Green, and A. Samal, "Individual articulator’s contribution to phoneme production," Proc. IEEE ICASSP, Vancouver, Canada, pp. 7785-7789, 2013.

[8] 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.

[9] P. Ladefoged and K. Johnson, A course in phonetics (6th Ed.). Independence, KY: Cengage Learning, 2011.

[10] 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.

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

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

[13] K. Richrmond, “Estimating articulatory parameters from the acoustic speech signal,” Ph.D. dissertation, Edinburgh Univ., Centre Speech Technol. Res., Edinburgh, U.K., 2002.

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

[15] P. K. Ghosh and S. 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.

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

[17] 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, pp. 305–308, 2000.

[18] 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.

[19] D. T. Chappell and J. Hansen. "Speaker-specific pitch contour modeling and modification." In International Conference on Acoust., Speech and Signal Processing, Seattle, WA, vol. 2, pp. 885-888, 1998.

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

[21] T. M. Cover and J. A. Thomas, Elements of Information Theory. New York: Wiley, 1991.

[22] A. Kraskov, H. Stögbauer and P. Grassberger, “Estimating mutual information,” Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top. , vol. 69, no. 6, pp. 066138 (1-16), 2004.

[23] S. Nakagawa, K. Shikano, and Y. Tohkura, "Speech, hearing and neural network models", IOS press, 1995.

[24] 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.

[25] A. E. Rosenberg, "Automatic speaker verification: A review." Proc. of IEEE vol. 64, no. 4, pp. 475-487, 1976.

[26] D. Sündermann. "Voice conversion: state-of-the-art and future work," Fortschritte der Akustik, vol. 31, no. 2, pp. 735-736, 2005.

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

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