NOVEL AMPLITUDE SCALING METHOD FOR BILINEAR FREQUENCY WARPING-BASED VOICE CONVERSION

Nirmesh J. Shah and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar

{nirmesh88_shah and hemant_patil}@daiict.ac.in

ABSTRACT

In Frequency Warping (FW)-based Voice Conversion (VC), the source spectrum is modified to match the frequency-axis of the target spectrum followed by an Amplitude Scaling (AS) to compensate the amplitude differences between the warped spectrum and the actual target spectrum. In this paper, we propose a novel AS technique which linearly transfers the amplitude of the frequency-warped spectrum using the knowledge of a Gaussian Mixture Model (GMM)-based converted spectrum without adding any spurious peaks. The novelty of the proposed approach lies in avoiding a perceptual impression of wrong formant location (due to perfect match assumption between the warped spectrum and the actual target spectrum in state-of-the-art AS method) leading to deterioration in converted voice quality. From subjective analysis, it is evident that the proposed system has been preferred *33.81 %* and *12.37 %* times more compared to the GMM and state-of-the-art AS method for voice quality, respectively. Similar to the quality conversion trade-offs observed by other studies in the literature, speaker identity conversion was *0.73 %* times more and *9.09 %* times less preferred over GMM and state-of-the-art AS-based method, respectively.

Index Terms— Voice conversion, frequency warping, amplitude scaling, Gaussian mixture model.

1. INTRODUCTION

Voice Conversion (VC) is a technique to transfer the perceived speaker identity from a source speaker to a particular target speaker for a given speech utterance [1]. Capturing the target speaker's identity and maintaining the high quality in the converted speech signal should be the aim of any VC system. Gaussian mixture model (GMM)-based VC is the state-of-the-art statistical parametric technique [2–4]. Recently, preprocessing using an outlier removal has been proposed to further improve the performance of this method [5]. The GMM-based VC method transforms the overall gross spectral characteristics very well. However, the finer details are *not* well transformed due to a statistical averaging leading to the deterioration of a voice quality which is called the *over-smoothing* in VC [6]. To overcome this, the use of dynamic features and global variance (GV) enhancement methods were proposed [4]. Apart from this, an exemplar-based non-parametric techniques were also proposed which directly uses the target speech exemplars to synthesize the converted speech and hence, keep more spectral details [7–11].

Apart from this, there are Frequency Warping (FW)-based methods in which the source spectrum is modified to match the frequency-axis of the target spectrum. Several different types of FW-based approaches have been proposed, namely, Dynamic Frequency Warping (DFW) [12], [13] and its very recent extensions such as Optimal DFW and Weighting transform (ODFWW) which simultaneously and optimally estimates the frequency warping and frequency weighting parameters [14], Vocal Tract Length Normalization (VTLN) [15], Weighted Frequency Warping (WFW) [16], Bilinear Frequency Warping (BLFW) [17], Correlation-based Frequency Warping (CFW) [18], etc. Among the various FW-based methods, here we have selected the BLFW method. As discussed in [17], the BLFW-based VC can be formulated in the parametric-domain. In addition, the BLFW do not show a locally irregular behavior compared to the piecewise learning-based FW methods. Furthermore, the number of parameters to be learnt is smaller which makes it suitable in the context of *overfitting*.

Since FW-based methods do not remove any spectral details, it produces a high quality speech after conversion. However, they do not modify the relative magnitude of the spectrum. Hence, the speaker similarity (SS) after conversion is not as successful as in the GMM-based VC systems. To overcome this problem, FW-based method is complemented with amplitude scaling (AS) or residual spectrum compensation [6], [17], [19]. The AS modifies the vertical-axis of the warped spectrum. The AS operation in the state-of-the-art BLFW+AS method assumes the perfect match between the warped and target formant structures which is not possible in practice [17]. As a result, the AS vector not only contains information related to the amplitude of the spectrum but also some information related the position of the formant (which

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.

will add spurious peaks in the warped spectrum). Hence, the quality of a converted speech will be degraded.

To eliminate the spurious peaks, we propose a novel AS technique at a spectrum-level which is free from the above mentioned assumption. The proposed AS transfers the spectral range of the warped-only spectrum to the spectral range of GMM-based spectrum. Several attempts have been made to combine the two state-of-the-art methods, namely, GMM and FW-based methods, in order to exploit the advantages of both the methods [6], [20], [21]. Similarly, our proposed AS method also combines the knowledge of this two state-of-the-art methods to obtain a better quality compared to the BLFW+AS method. In this paper, we have used the Voice Conversion challenge database [22]. Analysis of subjective and objective evaluations have also been presented.

2. JOINT DENSITY GMM-BASED VC

The Joint Density (JD) GMM-based VC finds a mapping function between the source and target speakers' spectral feature vectors. Let $X = [x_1, x_2, ..., x_N]$ and $Y = [y_1, y_2, ..., y_K]$ are spectral features of the source and target speakers', respectively. Here, $x_n \in \mathbb{R}^d$ and $y_n \in \mathbb{R}^d$. The joint vector, $Z = [z_1, z_2, ..., z_r, ..., z_T]$, is formed after aligning the spectral features using dynamic time warping (DTW) algorithm and modeled by a GMM, (where $z_r = [x_n^T, y_m^T]^T \in \mathbb{R}^{2d}$) as follows:

$$p(Z) = \sum_{m=1}^{M} \omega_m^{(z)} \mathcal{N}(z|\mu_m^{(z)}, \Sigma_m^{(z)}), \quad (1)$$

where $\mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}$, $\Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}$ and ω_m (with $\sum_{m=1}^{M} \omega_m^{(z)} = 1$ constraint for total probability) are the mean vector, the covariance matrix and weight of the m^{th} mixture component, respectively. The model parameters $\lambda^{(z)} = \{w_m^{(z)}, \mu_m^{(z)}, \Sigma_m^{(z)}\}$ are estimated using the expectation maximization (EM) algorithm [23]. During conversion, the predicted feature vector $\hat{y}$, is given by using the MMSE-based criteria [2]:

$$\hat{y} = \sum_{m=1}^{M} p_m(x)(\mu_m^{(y)} + \Sigma_m^{(yx)}(\Sigma_m^{(xx)})^{-1}(x - \mu_m^{(x)})), \quad (2)$$

where $p_m(x) = \frac{\omega_m \mathcal{N}(x|\mu_m^x, \Sigma_m^{xx}))}{\sum_{k=1}^{M} \omega_k \mathcal{N}(x|\mu_k^x, \Sigma_k^{xx}))}$ is the posterior probability of the source vector *x* for the m^{th} Gaussian component.

3. BLFW+AS AND PROPOSED AS

3.1. BLFW-based VC

The allpass transform is given by [17]:

$$Q(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad (3)$$

where $|\alpha| < 1$. For a given *d*-dimensional cepstral vector *x*, its frequency-warped cepstral vector *y* is given by:

$$y = W_\alpha x, \quad (4)$$

$$W_\alpha = \begin{bmatrix} 1-\alpha^2 & 2\alpha - 2\alpha^3 & \dots \\ -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}, \quad (5)$$

where W_α (called a warping matrix) has been expressed without considering the 0^{th} cepstral coefficient. The relation between the warped frequency and the original frequency is given by:

$$\omega_\alpha = tan^{-1}\left[\frac{(1-\alpha^2)sin\omega}{(1+\alpha^2)cos\omega - 2\alpha}\right]. \quad (6)$$

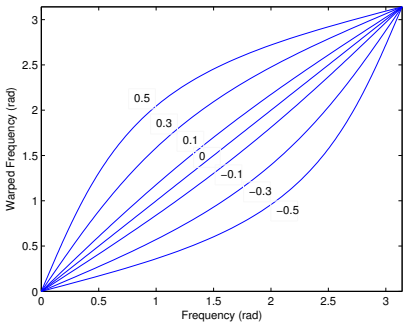

Fig. 1: *Shape of a BLFW function for different values of* α.

Fig. 1 shows the BLFW curve for different values of α. From Fig. 1, it can be observed that the positive values of α move the warped frequencies (i.e., possible formant) to the higher frequencies (as in case of a male-to-female conversion), and similarly, negative values of α move the formants to the lower frequencies (as in case of a female-to-male conversion). Thus, it maintains the *inverse* relationship between the vocal tract length and the formant frequencies. GMM is modeled on training database of a source speaker (i.e., θ). FW factor α_k and AS vector s_k associated with each components of GMM, the conversion function is given by [17]:

$$y = W_{\alpha(x,\theta)}x + s(x,\theta), \quad (7)$$

where $\alpha(x,\theta)$ and $s(x,\theta)$ are the result of combining the basis warping factors and the scaling vectors of all the components of θ, respectively, which is given by:

$$\alpha(x,\theta) = \sum_{k=1}^{m} p_k^{(\theta)}(x)\alpha_k, \quad s(x,\theta) = \sum_{k=1}^{m} p_k^{(\theta)}(x)s_k, \quad (8)$$

where $p_k^\theta(x)$ is the probability that x belongs to k^{th} mixture component of θ. Given the aligned source and target feature vectors and GMM trained on the source speaker data, i.e., θ, the warping factor α_k is first estimated by minimizing the error of warping only conversion which is given by:

$$\epsilon^{(\alpha)} = \sum_n ||y_n - W_{\alpha(x_n,\theta)}x_n||^2. \quad (9)$$

The iterative procedure proposed in [17] for a calculating a set of $\{\alpha_k\}$ for minimizing the eq. (9) is used here.

3.2. State-of-the-art Amplitude Scaling (AS)

Once $\{\alpha_k\}$ are estimated, the $\{s_k\}$ that minimizes the error between the warped and target vectors is given by [17]:

$$\epsilon^{(b)} = \sum_n ||r_n - s(x_n, \theta)||^2, \tag{10}$$

where $r_n = y_n - W_\alpha(x_n)$. This means that calculating the least square solutions of system, i.e., $P \cdot S = R$, where

$$P_{N\times m} = \begin{bmatrix} p_1^{(\theta)}(x_1) & \dots & p_m^{(\theta)}(x_1) \\ \vdots & \ddots & \vdots \\ p_1^{(\theta)}(x_N) & \dots & p_m^{(\theta)}(x_N) \end{bmatrix}, \tag{11}$$

$$\text{and } S_{m\times 1} = [s_1, s_2, \dots, s_m]^T, R_{N\times 1} = [r_1, r_2, \dots, r_N]^T. \tag{12}$$

The least square solution via l^2 norm minimization is given by:

$$S_{opt} = (P^T P)^{-1} P^T R. \tag{13}$$

The AS vector should compensate for the different formant amplitudes. In some of the cases where the warped formants does not coincide with the actual target formants, AS vector is expected to capture the mixed information about the intensity as well as the location of the formant which is potentially harmful to the voice quality of a converted voice.

3.3. Proposed AS

The AS operation in the above mentioned method assumes that there will be a perfect match between warped and target formant structures which is not possible in practice. Hence, the AS operation will induce spurious peaks, giving the perceptual impression of a wrong formant locations leading to a deterioration of speech quality in the converted speech signal.

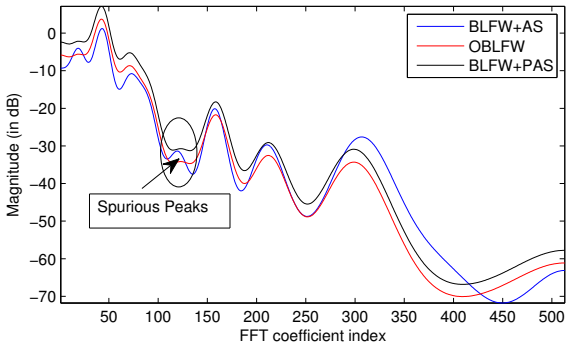

Fig. 2: *Converted spectrum using various VC methods.*

It can be seen from Fig. 2 that BLFW+AS method adds spurious peaks in the only BLFW warped spectrum (OBLFW). Essentially, AS operation should alter only the amplitudes of the warped spectrum (i.e., intensity of the formant) which is not the case. Therefore, we propose the following linear transformation at the spectrum-level,

$$\hat{y}_t(e^{j\omega}) = \frac{(m_3 - m_4)}{(m_1 - m_2)}(\hat{x}_t(e^{j\omega}) - m_2) + m_4, \tag{14}$$

where $\hat{x}_t(e^{j\omega})$ is the warped only spectrum,

$$\begin{aligned} &m_1 = max(\hat{x}_t(e^{j\omega})), \quad m_2 = min(\hat{x}_t(e^{j\omega})), \\ &m_3 = max(\hat{x}_{tgmm}(e^{j\omega})), m_4 = min(\hat{x}_{tgmm}(e^{j\omega})), \end{aligned} \tag{15}$$

where $max()$ and $min()$ will find the maximum and minimum value of a spectrum. $\hat{x}_{tgmm}(e^{j\omega})$ is the converted spectrum using JDGMM method. Here, the proposed AS technique transforms the spectral range of OBLFW spectrum to the spectral range of GMM-based converted spectrum. Since GMM-based VC transfers well the gross spectral characteristics, spectral range of converted spectrum obtained using GMM-based VC will be helpful to compensate the amplitude difference between warping-based spectrum and true target spectrum. As the proposed method uses the spectral range information instead of a finer details of GMM-based converted spectrum, it is free from the issue of oversmoothing.

It can be seen from Fig. 2 that proposed AS, i.e., (BLFW+PAS) will not add any spurious peaks and will compensate only for the amplitude difference without affecting the quality of a converted speech. Here, we would like to show the effectiveness of proposed AS over the state-of-the-art AS on the BLFW-based warped spectrum. Hence, the GMM-based spectrum and the actual target spectrum is not shown in Fig. 2. Similar spurious peaks are observed for the state-of-the-art AS methods for most of the frames.

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Experimental Setup

VC challenge database contains parallel training utterances *5* source and *5* target speakers' [22]. Each speaker's data contains a total of *162* utterances, out of which, *150* utterances have been taken for training and the remaining *12* were taken as a development set. We have built a total *25* systems for each source-target speaker pair using JDGMM-based method, BLFW+AS method and the proposed method (i.e., BLFW+PAS). *25*-D Mel cepstral coefficients (MCEPs) (including the 0^{th} coefficient) and *1*-D F_0 per frame (with *25* ms frame duration and *5* ms frame shift) have been used. The Dynamic Time Warping (DTW) algorithm has been used to align parallel training corpora [24]. For JDGMM-based system and for training of source GMM in the case of BLFW, we have taken different values of number of mixture components. For example, *m=16, 32, 64, 128* and selected the one which leads to the optimum MCD. We used a mean-variance (MV) transform method for F_0 transformation. AHOCODER has been used for the analysis-synthesis framework [25].

4.2. Experimental Results

For the subjective evaluation, comparative subjective test, namely, XAB test has been selected. Subjects were asked to prefer from the randomly played *A* and *B* samples (generated from two different approaches) which is having better voice quality and *speaker similarity* (SS) with reference to actual target sample X. In addition, the subjects can select equal preference in the case of samples that are perceptually similar. XAB test was performed separately between JDGMM and BLFW+PAS (i.e., proposed) and between BLFW+AS and BLFW+PAS.

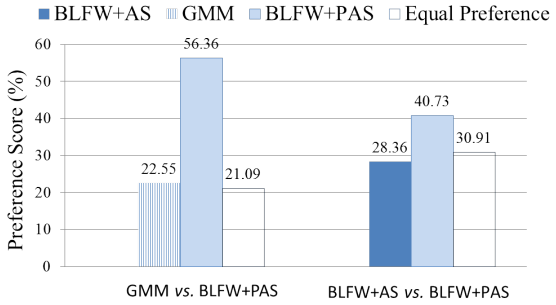

Fig. 3: *XAB test analysis for voice quality with 95 % confidence interval (margin of error: 0.048 for GMM vs. BLFW+PAS and 0.05 for BLFW+AS vs. BLFW+PAS).*

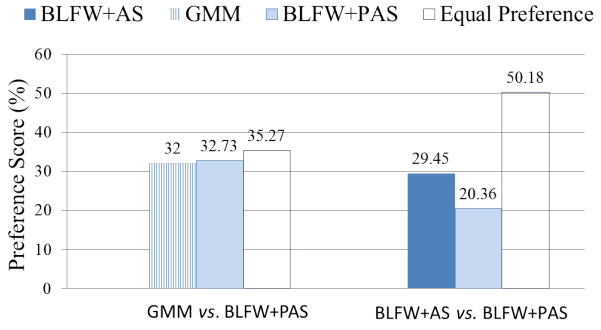

Fig. 4: *XAB test analysis for speaker similarity with 95 % confidence interval (margin of error: 0.05 for the both cases).*

Fig. 3 and Fig. 4 show the MOS obtained from *15* subjects (*5* females and *10* males) from total *375* samples for voice quality and SS, respectively. It is clear from the results that in terms of voice quality, the proposed AS system is preferred *56.36 %* times whereas the GMM-based system is preferred *22.55 %* times by the subjects. Similarly, BLFW+PAS is preferred *40.73 %* times whereas BLFW+AS is preferred *28.36 %* times by the subjects.

The speaker identity conversion was *0.73 %* times more preferred over GMM-based method. Though the proposed system *9.09 %* times less preferred over BLFW+AS system, *50.18 %* times subjects have given an equal preference to the proposed system and BLFW+AS. The less preference for speaker similarity of the proposed system compared to the state-of-the-art BLFW+AS clearly indicates that actual shape of the spectral trajectory also matters for better speaker identity conversion in addition to the formant locations and its amplitude [26]. However, modifying the spectral details will affect the voice quality. Hence, there is a quality conversion trade-offs. Similar trade-offs were observed by other studies in the literature [16], [19], [27]. For objective measure, the traditional MCD is used here which is given by [4]:

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{i=2}^{25}(m_i^t - m_i^c)^2}\ , \qquad (16)$$

where m_i^t and m_i^c are the i^{th} coefficient of MCEP of target and converted speech utterance, respectively. It can be seen from Fig. 5 that the proposed method gives the higher MCD values compared to the GMM-based VC. The BLFW method moves the formants towards their image in the target speaker's spectrum. Thereafter, the proposed AS will modify the amplitude of warped spectrum instead of matching the actual target spectral details. Hence, it will not get less MCD scores compared to the GMM-based VC (as shown in Fig. 5). In addition, it has been observed in the literature that MCD does not correlate well with subjective score for FW-based VC [16], [17], [19], [28]. However, MCD is used here for comparing relative performance of same type of VC for selecting optimum number of mixture components.

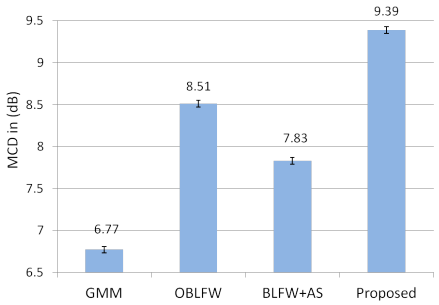

Fig. 5: *MCD analysis for various systems with 95 % confidence interval (margin of error: 0.04 for all the systems).*

5. SUMMARY AND CONCLUSIONS

This study proposed the AS technique which linearly transfers the amplitude of the frequency-warped spectrum using knowledge of the GMM-based converted spectrum without adding any spurious peaks. Hence, the proposed AS is found to have better voice quality compared to traditional BLFW+AS. Since, spectral details are maintained well for a high quality synthesis, the proposed system is found to perform less successful in terms of SS after conversion. Hence, there is indeed trade-offs between the quality and the SS. However, the proposed AS is still able to achieve *50.18 %* times equal preference in SS after conversion which makes it more suitable to use in real-world applications of VC.

6. REFERENCES

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

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

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] S. V. Rao, N. J. Shah, and H. A. Patil, "Novel pre-processing using outlier removal in voice conversion," in 9^{th} *ISCA Speech Synthesis Workshop*, San Fransisco, USA, pp. 134–139.

[6] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Salt Lake City, USA, 2001, pp. 841–844.

[7] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *IEEE Spoken Language Technology Workshop (SLT)*, Miami, Florida, USA, 2012, pp. 313–317.

[8] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *Proc.* 8^{th} *ISCA Speech Synthesis Workshop*, Barcelona, Spain, 2013, pp. 201–206.

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

[10] Z. Jin, A. Finkelstein, S. DiVerdi, J. Lu, and G. J. Mysore, "Cute: A concatenative method for voice conversion using exemplar-based unit selection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5660–5664.

[11] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5175–5179.

[12] N. Maeda, H. Banno, S. Kajita, K. Takeda, and F. Itakura, "Speaker conversion through nonlinear frequency warping of straight spectrum," in *EUROSPEECH*, Budapest, Hungary, 1999, pp. 1–4.

[13] H. Valbret, E. Moulines, and J.-P. Tubach, "Voice transformation using PSOLA technique," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, San Francisco, USA, 1992, pp. 145–148.

[14] Y. Agiomyrgiannakis and Z. Roupakia, "Voice morphing that improves TTS quality using an optimal dynamic frequency warping-and-weighting transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5650–5654.

[15] D. Sundermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, U.S., 2003, pp. 676–681.

[16] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

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

[18] X. Tian, Z. Wu, S. W. Lee, and E. S. Chng, "Correlation-based frequency warping for voice conversion," in 9^{th} *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Singapore, 2014, pp. 211–215.

[19] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[20] T.-C. Zorilă, D. Erro, and I. Hernaez, "Improving the quality of standard GMM-based voice conversion systems by considering physically motivated linear transformations," in *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2012, pp. 30–39.

[21] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, M. Dong, and E. S. Chng, "System fusion for high-performance voice conversion," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2759–2763.

[22] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *INTERSPEECH*, San Fransisco, USA, 2016, pp. 1–5.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[24] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[25] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved HNM-based vocoder for statistical synthesizers." in *INTERSPEECH*, Florence, Italy, 2011, pp. 1809–1812.

[26] H. Kuwabara and Y. Sagisak, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Communication*, vol. 16, no. 2, pp. 165–173, 1995.

[27] D. Mostefa, O. Hamon, N. Moreau, and K. Choukri, "Evaluation Report Deliverable D30 of the EU funded projects TC-STAR, 2007," http://tcstar.org/pages/ThirdEvaluationCampaign.htm, {Last Accessed: August 24, 2016 }.

[28] A. Rajpal, N. J. Shah, M. Zaki, and H. A. Patil, "Quality assessment of voice converted speech using articulatory features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 1–5.