International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 03, March-2015

Cepstral Domain Voice Conversion Based on Constrained Transformations

Dr. G. Indumathi,
Mepco Schlenk Engineering College,
Sivaksi, India.

V. S. Hewitt,
Mepco Schlenk Engineering College,
Sivakasi,India.

V. Rajavel,
Mepco Schlenk Engineering College,
Sivakasi,India.

***Abstract*—This paper proposes a method for voice conversion in cepstral domain. The method basically involves two steps: Bilinear Frequency Warping and Amplitude Scaling. Frequency warping is done so that spectrum of source speaker is moved towards their image in the target speaker's spectrum. Amplitude scaling is done to compensate for the warping inaccuracy. The use of bilinear function is to warp the signal without any significant decrease in the quality scores. Fuzzy logic is applied to the amplitude scaling process in order to improve the perceptional quality. This method despite its simplicity, achieves similar performance scores to the previously available methods based on Gaussian Mixture Model.**

***Keywords*—Voice Conversion, Gaussian Mixture Model, bilinear function, frequency warping, fuzzy amplitude scaling.**

I. INTRODUCTION

Voice Conversion is the process of modifying the characteristics of source speaker in such a way that it is perceived as the voice of target speaker [1]. Till now, voice conversion systems have mainly focused on spectral characteristics and some others operate on prosodic level too. A voice conversion system involves two phases, training phase and conversion phase. During the training phase, the voice conversion system learns a function to transform the source speaker's acoustic data to target speaker's acoustic data. Usually this involves database of involved speaker's speech signals. Analyzing the training data, the parameters corresponding to speaker identity are extracted from source as well as target speech. During the conversion phase, the function learned is used to transform any new input utterances from the source speaker i.e. the source acoustic data to be mapped to the target acoustic data. Finally the converted speech i.e. source message in the voice of target speaker is synthesized [2]. The applications of Voice Conversion is in the entertainment industries mainly dubbed movies, gaming, karaoke, voice masking for chat rooms, customization of speaking devices [3]etc.

At present, most of the method to perform voice conversion mainly uses Gaussian mixture model. Gaussian mixture models are used for statistical modeling of speech data [4]-[6]. This method uses Gaussian mixture model tosegregate the speech data into components and determine the posterior probability of each components of data. Frequency domain transformation is preferred over time domain transforms because the frequency transform does not remove any part of source spectrum thereby preserving the quality of the converted speech. In the line of work of our method, first frequency warping based on piecewise linear frequency warping function along with energy conversion filter is used [7]. In [8], for instance, the lowest-distortion FW path was calculated from discretized spectra via dynamic programming, which is known as dynamic FW [9]. Further improvement lead to the usage of amplitude scaling instead of energy conversion filter that resulted in improved quality scores. At the next level Bilinear frequency warping functions are used which ensures computational simplicity. In our proposed method, fuzzy logic is applied to the amplitude scaling to improve the overall conversion performance [10]. The general block diagram of voice conversion is given in Fig.1.

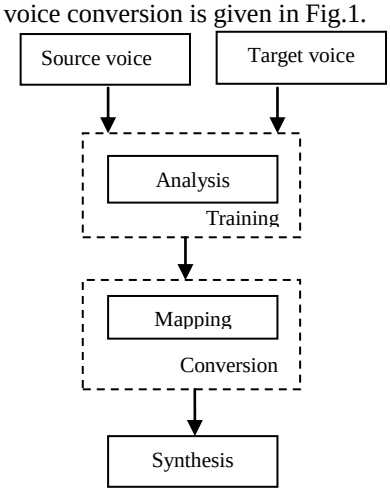

Fig. 1. General block diagram of voice conversion

II.DESCRIPTION OF THE METHOD

A. *Bilinear Warping Functions*

Bilinear functions are parametric Frequency Warping functions are applied to perform vocal tract length normalization (VTLN), used in both speech recognition [11] and conversion. Bilinear functions depends on one single parameter α

$$z_{\alpha}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \qquad |\alpha|<1 \qquad (1)$$

Given a *p*-dimensional cepstral vector x, it has been proven [12]-[14] that the cepstral vector y that corresponds to the frequency warped version of the spectrum represented by x is given by

$$y = W_{\alpha} x, \mathrm{W}_{\alpha} = \begin{pmatrix} 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots \\ -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \qquad (2)$$

The dependence between the warping matrix $\mathbf{W}_{\alpha}$ warping factor α is strongly nonlinear. However, it was observed in [14] that when α is sufficiently closed to zero the higher powers of α can be neglected and this dependence becomes linear.

$$W_{\alpha} = \begin{pmatrix} 1 & 2\alpha & 0 & 0 & \cdots \\ -\alpha & 1 & 3\alpha & 0 & \cdots \\ 0 & -2\alpha & 1 & 4\alpha & \cdots \\ 0 & 0 & -3\alpha & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \qquad (3)$$

Then, the warping operation is equivalent to

$$y = x + \alpha . d(x) \qquad (4)$$

where d(x) is the vector whose *i*th element is given by

$$d(x)[\mathrm{i}] = (\mathrm{i}+1).\mathrm{x}[\mathrm{i}+1] - (\mathrm{i}-1).\mathrm{x}[\mathrm{i}-1], \mathrm{i} = 1.....\mathrm{p} \qquad (5)$$

The block diagram for bilinear frequency warping and fuzzy amplitude scaling process are given in Fig. 2

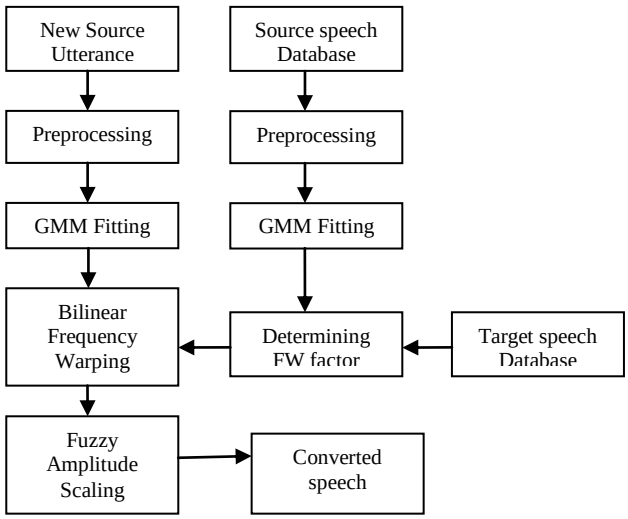

Fig 2. Block diagram of Bilinear Frequency Warping and Fuzzy Amplitude Scaling.

B. *Frequency Warping Factor*

The frequency warping factor is used to determine the warping matrix used in the conversion phase. In practice the value of α_k may not be sufficiently close to zero and then the approximation in eqn 3 is not valid. This happens in case of gross-gender voice conversion, where α value is not sufficiently small. Therefore an iterative process that yields an increasingly accurate solution in any gender to gender conversion is considered.

Step 1: initialize $\alpha_k=0$ for all *k*.

Step 2: for the current α_k, calculate a set of warped vectors $\{z_n\}, z_n = W_{\alpha(x_n,\theta)} x_n$ where the warping matrix is given by expression (2)

Step 3: calculate the differential warping factors $\{\alpha_k\}$ needed to make the warped source vector (z_n) closer to target vector (y_n). The differential warping vector can be obtained from the distance matrix (D) and error vector (e) as explained in [16]-[17]

$$\underset{Np\times m}{D} = \begin{pmatrix} p_1^{(\theta)}(x_1)d(z_1) & \cdots & p_m^{(\theta)}(x_1)d(z_1) \\ \vdots & \ddots & \vdots \\ p_1^{(\theta)}(x_N)d(z_N) & \cdots & p_m^{(\theta)}(x_N)d(z_N) \end{pmatrix}$$

$$\underset{m\times 1}{\alpha} = [\Delta\alpha_1 \quad \cdots \quad \Delta\alpha_m]^T$$

$$\underset{Np\times 1}{e} = [(y_1 - z_1)^T \quad \cdots \quad (y_N - z_N)^T]^T \qquad (6)$$

From the distance matrix and error vector the optimal value of warping factor would be

$$\alpha_{opt} = (\mathrm{D}^T\mathrm{D})^{-1}\mathrm{D}^T e \qquad (7)$$

Step 4: accumulate $\{\Delta\alpha_k\}$ into the current $\{\alpha_k\}$. According to [38], this can be done as follows.

$$\alpha_k^{(\text{updated})} = \frac{\alpha_k + \Delta\alpha_k}{1 + \alpha_k . \Delta\alpha_k} \qquad (8)$$

Step 5: if the updated α_k value in the previous step did not show insignificant change (i.e $|\Delta\alpha_k|<0.001$ for all k), exit. Otherwise go to step 2.

Using this iterative method between any pair of speakers the conversion error is minimized after each iteration until the process is converged. During the conversion phase, from the obtained warping factors $\{\alpha_k\}$, the precise bilinear frequency warping matrix is used.

C.*Conversion Phase*

After obtaining the frequency warping factor $\alpha_k$for each of the Gaussian component of θ, the expression of conversion function as,

$$y = W_{\alpha(x,\theta)} x \qquad (9)$$

where the warping matrix W is given by expression (2), and α(x,θ) the result of combining the basic warping factors of all the components of θ, is given by

$$\alpha(x,\theta) = \sum_{k=1}^{m} p_k^{\theta}(x)\alpha_k \qquad (10)$$

C.Amplitude Scaling Vectors

After the warping factor $\{\alpha_k\}$ is determined, the value of amplitude scaling vector is calculated in such a way that the error between warped and target vectors are minimized as in [2].

$$\varepsilon^{(b)} = \sum_n \| r_n - s(x_n,\theta) \|^2,\ r_n = y_n - W_{\alpha(x_n,\theta)} x_n \quad (11)$$

This means calculating the least squares solution of the system **P.S= R.** where,

$$\underset{N\times m}{\mathrm{P}} = \begin{pmatrix} p_1^{(\theta)}(x_1) & \cdots & p_m^{(\theta)}(x_1) \\ \vdots & \ddots & \vdots \\ p_1^{(\theta)}(x_N) & \cdots & p_m^{(\theta)}(x_N) \end{pmatrix}$$

$$\underset{m\times p}{S} = [s_1 \quad \cdots \quad s_m]^T$$

$$\underset{N\times p}{R} = [r_1 \quad \cdots \quad r_N]^T \quad (12)$$

The solution of the system is the optimal value of the amplitude scaling vector.

$$\mathrm{S}_{opt} = (\mathrm{P}^T\mathrm{P})^{-1}\mathrm{P}^T\mathrm{R} \quad (13)$$

C. Fuzzy Amplitude Scaling

Fuzzy rule applied to the amplitude scaling process mainly involves three steps, Fuzzification, Rule base engine and Defuzzification. The warped source signal and the frequency response of the target signal are the inputs to the fuzzy amplitude scaling system.

i.*Fuzzification*

Fuzzification refers to the process of converting real value to fuzzy value. The warped source signal is assumed to have three membership functions, 'low', 'medium' and 'high' amplitude ranges. The frequency response of the target signal is assumed to have three membership functions, 'low', 'medium' and 'high' frequency ranges. Trapezoidal membership functions are considered for the inputs of fuzzy system. The fuzzy mapping function is shown in Fig. 3

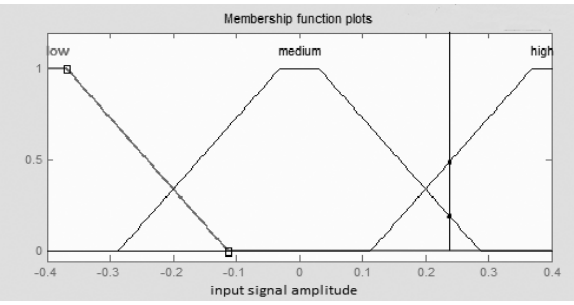

Fig. 3 The input signal amplitude is mapped as functions of degree of truth values. For example, amplitude of 0.24 has a degree of truth value of 0.22 and 0.48 in the medium and high range membership function.

ii.*Rule base engine*

Rule base engine refers to the decision matrix of fuzzy knowledge base composed of expert IF<antecedents>THEN<conclusions> rule. It takes into account the range of amplitude values and frequency response and decides the range of the output amplitude values [8]. All the possibilities are taken into account and the corresponding output amplitude ranges are defined as in Table 1.

Table.1 Decision matrix

Amp / Freq.	Low	Medium	High
Low	Low	High	Low
Medium	Medium	Low	High
High	Medium	Low	low

iii.Defuzzification:

Defuzzification involves the process of transposing the fuzzy outputs to real outputs. The output amplitude values are determined by reverse mapping the truth values of the membership function to the corresponding amplitude values.

III.RESULTS AND DISCUSSIONS

Experimental Procedure

This section accounts various results that expose the performance aspects of Bilinear Frequency Warping and Fuzzy Amplitude Scaling method. The speech data used in the experiment were created by Zabaware Text-to-Speech software considering different source and target speakers. Equal number of training sentences is used in the training phase. The sampling frequency of the signals is 11.025 kHz. The Gaussian Mixture Models used in our experiments had 5 mixtures with full covariance matrices. The parameters of the model used in the EM algorithm are initialized to get consistent result.

Accuracy of Frequency Warping Factor

At first 25 parallel training sentences are given to the conversion system. The obtained value of frequency warping factor α is in the order of 10^{-5}. With the value in that range sometimes it is not feasible to extract time domain signal from cepstral coefficients. Then the system is trained with 50 parallel sentences. The warping factor reduces to the range of 10^{-7}. With this range it is possible to recover the signal but the amplitude range gets disrupted. When the system is trained with 100 parallel data, the value of warping factor falls to the range of 10^{-8}. With this value the time domain signal is easily extracted and suitable for further processing.

Accuracy of Fuzzy Amplitude scaling

The performance of Fuzzy Amplitude Scaling depends on the defined range and shape of the membership function. Different shapes results in varied results. Trapezoid membership function is found to have better performance in comparison with the triangular membership function.

IV. CONCLUSION

This paper has proposed a voice conversion method based on bilinear frequency warping and fuzzy amplitude scaling. The method can be implemented in cepstral domain. The conversion function parameters are trained to the most accurate level by an iterative process from a few parallel data. This method achieves more computational efficiency. Moreover, the average conversion performance is good in comparison with the state-of-the-art statistical methods. The quality of the converted speech is increased without worsening the conversion performance. Subjective evaluation shows that there is a good trade-off between quality and simplicity.

REFERENCES

[1] E.Moulines and Y. Sagisaka, "Voive conversion: State of the art and perspectives," *Speech Commun. Special Issue*, vol. 16, no. 2, 1995.

[2] Daniel Erro, Eva Navas, and Inma Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 21, no. 3, pp. 556-565, Mar. 2013.

[3] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech syntesis," in *Proc. ICASSP*, 1998,vol. 1,pp. 285-288.

[4] Y. Stylianuou, O.Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans, Speech Audio Process.*, vol. 6, no. 2, pp.131-142, Mar.1998.

[5] A. Kain, "High resolution voice transformation," Ph.D. dissertation, Oregon Health & Science Uniov., Portland, 2001.

[6] 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, Nov. 2007.

[7] D. Erro, A. Moreno, and A. Bonafonte,"Voice conversion based on weighted frequency warping," *IEEE Trans. Audio, Speech, Lang. Process.*, vol, 18, no. 5,pp,922-931, Jul, 2010.

[8] H. Valbret, E. Moulines,and J.P. Tubach, "Voice transformation using PSOLA techniques," *Speech Commun.*, vol. 1, pp. 145-148, 1992.

[9] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or non-parallel corpora," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp,1313-1323, May 2012.

[10] Rafael Alcala, Jesus Alcala-Fdez, Maria Jose Gacto and Francisco Herrera,"Genetic Lateral and Amplitude Tuningof Membership Functions for Fuzzy Systems,"International Conference on Machine Intelligence, Tozeur – Tunisia, pp.589-595, 2005

[11] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," 1997, CMU Computer Science Tech. Rep.,

[12] J. McDonough and W. Byrne, "Speaker adaptation with all-pass transforms," in *Proc. ICASSP*, 1999, pp.757-760.

[13] M. Pitz and H.ney, "Vocal tract normalization equals linear transformation in cepstral space." *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5,pp. 930-944, Sep. 2005.

[14] T. Emori and K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," in *Proc. Eurospeech*, 2001, pp. 1649-1652.