Novel Amplitude Scaling Method for Bilinear Frequency Warping-based Voice Conversion

Nirmesh J. Shah and Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India-382007

Introduction

- Bilinear frequency warping-based Voice Conversion (VC) [1].
- A novel proposed Amplitude Scaling (AS).
- Limitation of state-of-the-art AS.
- Effectiveness of proposed AS.
- VC Challenge database [4].
- Subjective and objective evaluation of VC systems.

BLFW-based VC

- The allpass transform is given by [1]:

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

where $|\alpha| < 1$.

- Frequency warping in cepstral domain.

$$y = W_\alpha x, \tag{2}$$

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

where W_α is a warping matrix (without 0^{th} cepstral coefficient).

- Conversion function for BLFW+AS:

$$y = W_{\alpha(x,\theta)} x + s(x, \theta), \tag{4}$$

where $\alpha(x, \theta)$ and $s(x, \theta)$ are given by,

$$\alpha(x,\theta) = \sum_{k=1}^{m} p_k^{(\theta)}(x)\alpha_k, \quad s(x,\theta) = \sum_{k=1}^{m} p_k^{(\theta)}(x) s_k. \tag{5}$$

- Alignment of source and target feature vectors
- GMM trained on source speaker data, i.e., θ.
- The iterative procedure proposed in [1] for calculating set of $\{\alpha_k\}$ (warping factors) for minimizing the following eq. (6):

$$\epsilon^{(\alpha)} = \sum_n ||y_n - W_{\alpha(x_n,\theta)} x_n||^2. \tag{6}$$

State-of-the-art AS Technique

- The $\{s_k\}$ that minimizes the error between warped and target vectors which is given by

$$\epsilon^{(b)} = \sum_n ||r_n - s(x_n, \theta)||^2, \tag{7}$$

where $r_n = y_n - W_\alpha(x_n)$.

- The least square solution of $P \cdot S = R$ is given by

$$S_{opt} = (P^T P)^{-1} P^T R. \tag{8}$$

where

$$P_{N \times m} = \begin{bmatrix} p_1^{(\theta)}(x_1) & \dots & p_m^{(\theta)}(x_1) \\ \vdots & \ddots & \vdots \\ p_1^{(\theta)}(x_N) & \dots & p_m^{(\theta)}(x_N) \end{bmatrix}, \tag{9}$$

and $S_{m\times 1} = [s_1 \dots s_m]^T, R_{N \times 1} = [r_1 \dots r_N]^T$. (10)

Proposed AS Technique

- Perfect match assumption of state-of-the-art AS.
- Induce spurious peaks.
- Perceptual impression of wrong formant locations.
- Use of GMM-based converted spectrum [3].

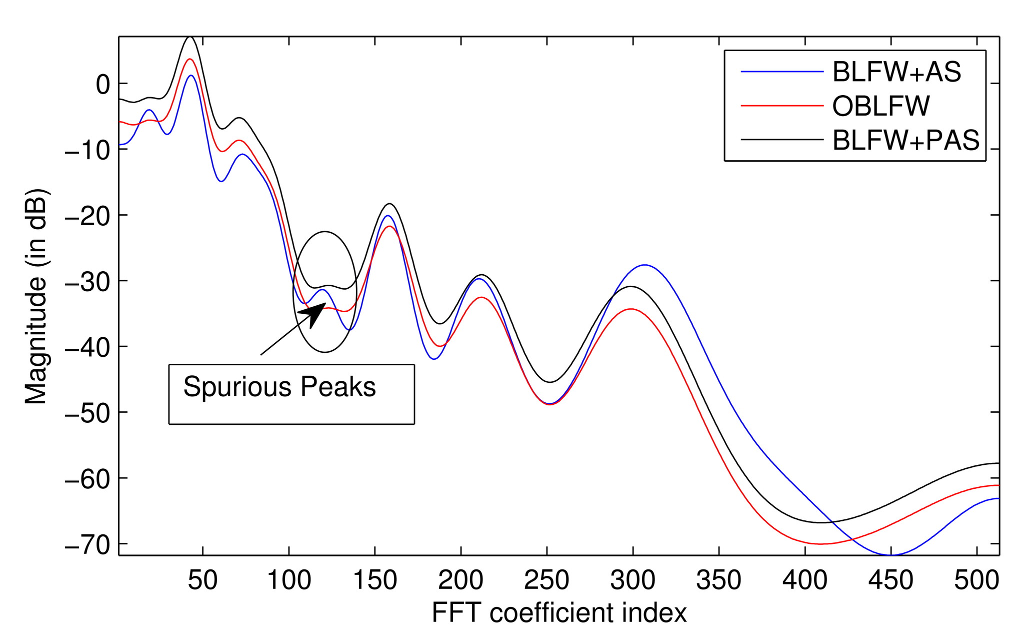

Figure 1: *Converted spectrum using VC methods.*

- Propose linear transformation at spectrum-level,

$$\hat{y}_t(e^{j\omega}) = \frac{(m_3 - m_4)}{(m_1 - m_2)}(\hat{x}_t(e^{j\omega}) - m_2) + m_4, \tag{11}$$

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

$$m_1 = max(\hat{x}_t(e^{j\omega})), \quad m_2 = min(\hat{x}_t(e^{j\omega})),$$
$$m_3 = max(\hat{x}_{tgmm}(e^{j\omega})), m_4 = min(\hat{x}_{tgmm}(e^{j\omega})), \tag{12}$$

Experimental Setups

- Total *5* source and *5* target speakers' parallel training data.
- Training set *150* utterances and development set *12*.
- Total *25* VC systems for each source-target speaker pair using JDGMM-based method [3], BLFW+AS method and proposed method.

Experimental Results

- XAB test from *375* samples (*15* subjects: *5* females and *10* males).

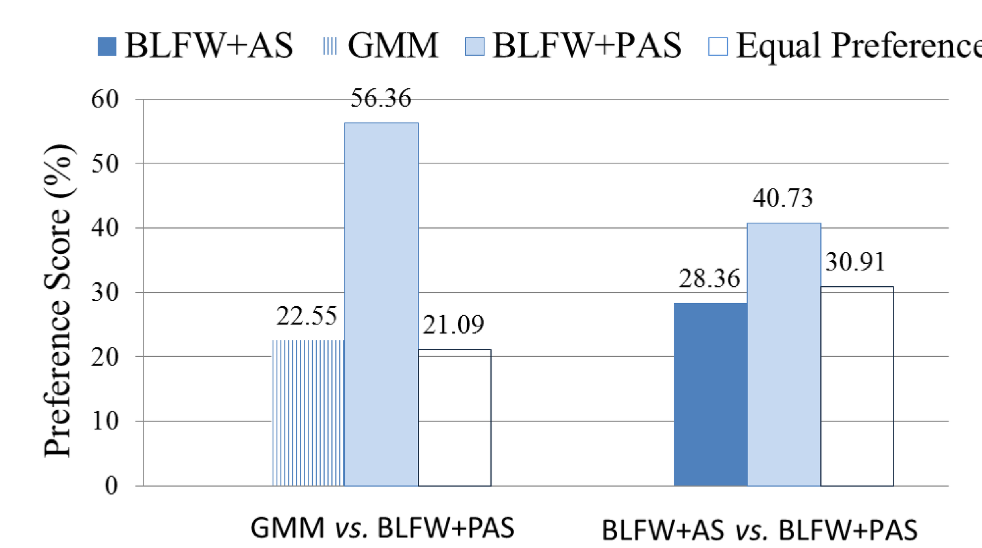

Figure 2: *XAB test analysis for voice quality with 95 % confidence interval (margin of error: 0.048 for GMM vs. BLFW+PAS and 0.05 for BLFW+AS vs. BLFW+PAS).*

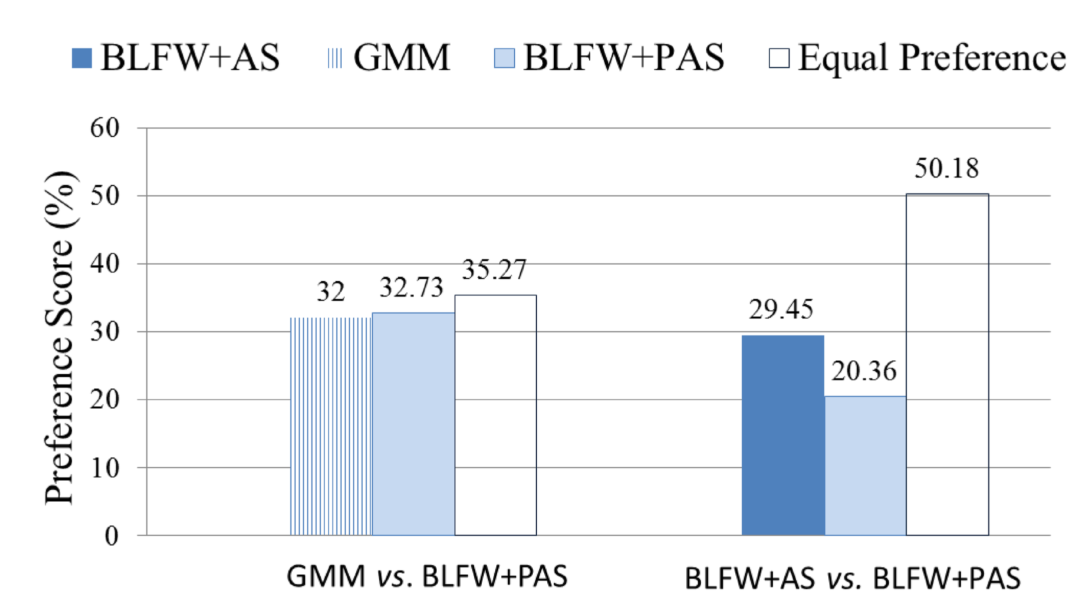

Figure 3: *XAB test analysis for speaker similarity with 95 % confidence interval (margin of error: 0.05 for the both cases).*

Objective Evaluations

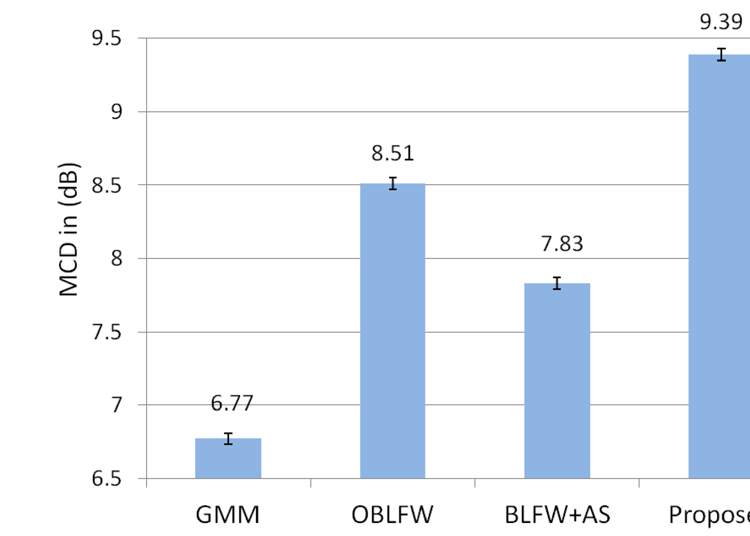

Figure 4: *MCD analysis for various systems with 95 % confidence interval (margin of error: 0.04 for all the systems).*

Conclusion

- The proposed AS is found to have better voice quality compared to traditional BLFW+AS.
- The proposed system is found to perform less successful in terms of SS after conversion.
- Trade-offs between the quality and the SS.

Selected References

[1] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conver- sion based on bilinear frequency warping plus amplitude scal- ing," in *IEEE Transactions on Audio, Speech and Language Proc.*, vol. 21, no. 3, pp. 556 - 566, 2013.

[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] 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.

[4] 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-

Acknowledgements

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.

Presented by Nirmesh J. Shah at 42^{nd} IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, USA, March 05-09, 2017.