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]Speech Research Lab, DA-IICT, Gandhinagar-382007, India and [2]IISc, Bangalore 560012, India

Introduction

- Novel application of the Acoustic-to-Articulatory Inversion (AAI).
- The ability of humans to speak effortlessly requires the coordinated movements of various articulators.
- This effortless movement contributes towards a naturalness, intelligibility and speaker's identity.
- It is partially present in voice converted speech.
- Whether the articulatory information is lost during the VC process?
- How can this information loss quantified?
- Factors responsible for quality of VC

Proposed Objective Measure

- Uses MOCHA database [1].
- Gaussian mixture model (GMM)-based VC [2].
- Bilinear frequency warping plus amplitude scaling-based (BLFW+AS) VC [3].
- AAI: Generalized Smoothness Criterion (GSC) [4].
- Target and voice converted acoustic vector be given by and $\boldsymbol{X_{tv}}$, respectively.
- Electromagnetic Articulography (EMA) vector of the target be $\boldsymbol{Y_t}$
- Estimated EMA vector from $\boldsymbol{X_t}$ and $\boldsymbol{X_{tv}}$ be $\boldsymbol{Z_t}$ and $\boldsymbol{Z_{tv}}$, respectively.

Table 1: Comparison of mutual information before & after VC

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

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

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

- where $RMSE_{tt}$ is calculated between $\boldsymbol{DY_t}$ and $\boldsymbol{DZ_t}$ and $RMSE_{tv}$ is an average RMSE between $\boldsymbol{DY_t}$ and $\boldsymbol{DZ_{tv}}$.
- The *Estimation Error (EE)* (in *mm*), measures the distance between articulatory trajectories of voice converted speech.

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

N is the length and *M* is the dimensionality of the articulator trajectory.

Proposed System Architecture

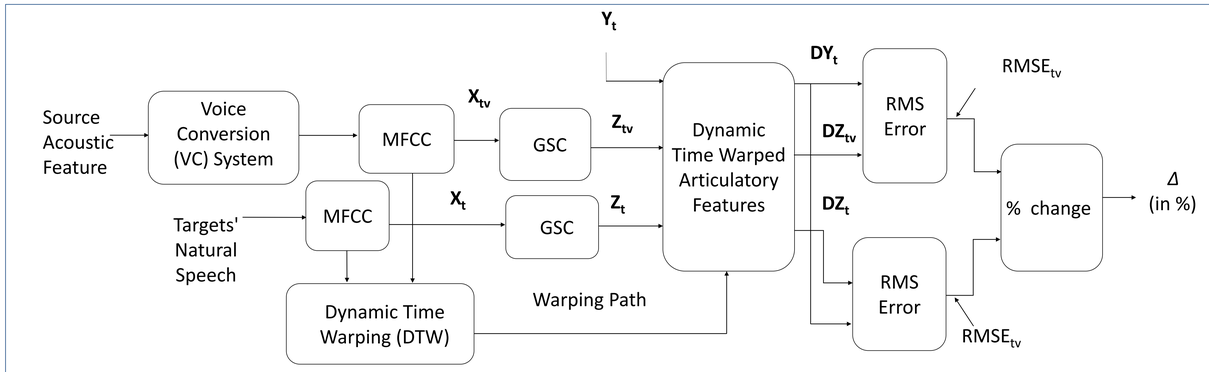

Fig. 1: Proposed system architecture for estimating articulatory features from voice conversion (VC) system.

Experimental Results

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

- For a subjective measure, MOS of *360* samples, from *15* subjects (*9* male and *6* females with *21-25* years of age).

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

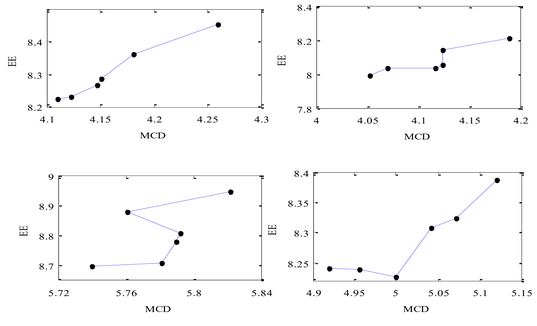

Fig. 2 : MCD *vs.* plot for selected systems (a)-(b) M-F and F-M GMM-based VC and (c)-(d) M-F and F-M BLFW+AS-based VC.

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**

Correlation with Subjective Scores

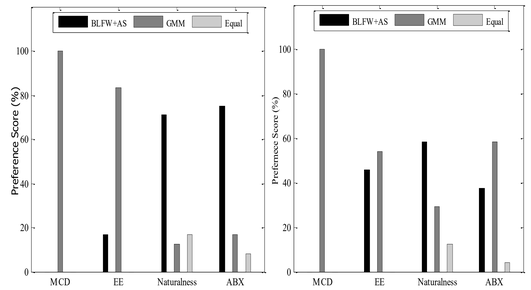

Fig. 3: Preference score based on MCD, *EE*, naturalness and ABX test for GMM and BLFW VC systems (a) M-F (b) F-M. Equal means, subjects could not judge and give equal preference score.

Conclusion

- Among all the articulators, tongue tip (known to be critical for the speech production) shows highest % Δ.
- The AAI system poorly estimates the articulatory trajectories of a voice converted speech.
- After VC articulatory parameters related information is lost.
- MCD and EE are found be partially correlated.
- EE has more correlation with MOS.
- In preference test, MCD 100% contradicted preference test.
- EE supported subjective measure 45.8 % and 16.67 % for F-M and M-F VC, respectively.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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.
- We also thanks all the participants who took part in subjective evaluation.

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