昆山杜克大学 DUKE KUNSHAN UNIVERSITY

F0 CONTOUR ESTIMATION USING PHONETIC FEATURE IN ELECTROLARYNGEAL SPEECH ENHANCEMENT

Zexin Cai[1]*, Zhicheng Xu*[2]*, Ming Li*[1]

[1]Data Science Research Center, Duke Kunshan University, Kunshan, China

[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong

Introduction

Credit: https://www.practicalslpinfo.com/the-electrolarynx.html

Electrolaryngeal (EL) speech enhancement aims to improve the naturalness and intelligibility of speech produced by Electrolarynx. In this way, people without larynx could retrieve the capability to produce a normal voice.

EL speech does not sound like human-produced speech in three ways:

- The sound quality degrades due to the noise caused by the continuous vibration of the EL;
- EL speech sounds unnatural because it is generated by the mechanical excitation signals;
- The intelligibility is limited since the EL produces monotonous speech.

The intelligibility of the converted EL speech is constrained by fundamental frequency ($F0$) contour, especially in a tone based language like Mandarin. Hence, our proposed framework aims to predict the $F0$ contour using additional linguistic information.

Proposed Conversion Framework

Training Phase

Joint density Gaussian mixture model (JDGMM) is used in our systems for estimating the converted Mel Cepstral Coefficients (CVMCC) and $F0$.

GMM Training:

$$P(X_t, Y_t|\lambda) = \sum_{m=1}^{M} \omega_m \mathcal{N}([X_t^\intercal, Y_t^\intercal]^\intercal; \mu_m^{(X,Y)}, \Sigma_m^{(X,Y)})$$

$$\mu_m^{(X,Y)} = \begin{bmatrix} \mu_m^{(X)} \\ \mu_m^{(Y)} \end{bmatrix}, \Sigma_m^{(X,Y)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix}$$

Acoustic model training:

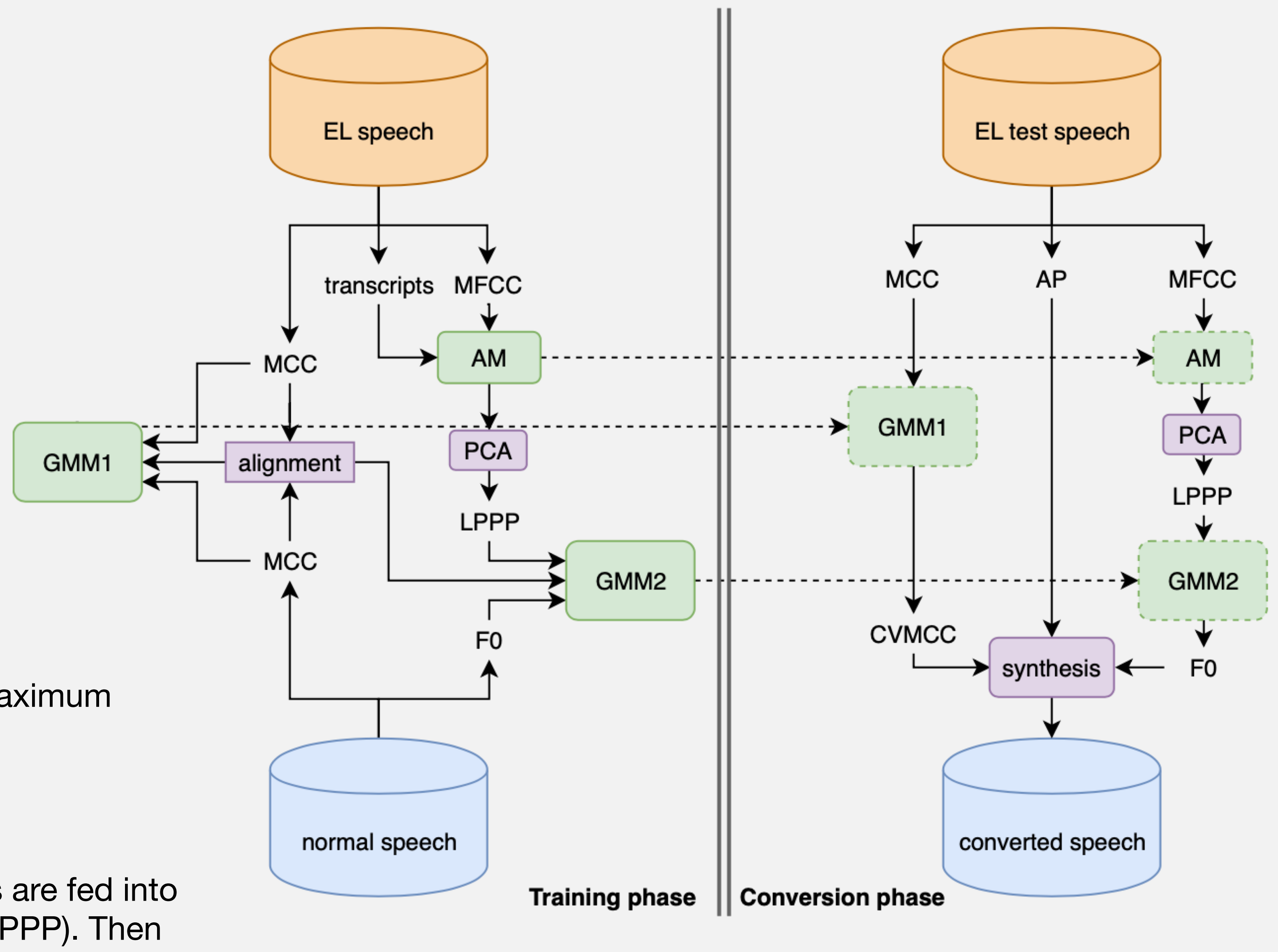
1. Train Hidden Markov Model – GMM automatic speech recognition system for phoneme recognition;
2. Based on the alignment produced by HMM-GMM, train neural network acoustic model.

Conversion Phase

The converted speech parameters are estimated by employing maximum likelihood estimation upon the trained GMM.

$$\hat{y} = \underset{y}{argmax}\, P(Y|X, \lambda)$$
$$subject\ to\ Y = Wy$$

The Mel-frequency cepstral coefficient (MFCC) feature sequences are fed into the acoustic model to obtain the phonetic posterior probabilities (PPP). Then the PCA matrix is applied to reduce the dimension of PPP vectors to get LPPP. The $F0$ contour is estimated by GMM2 given the input LPPP sequences.



Experimental Setup

Data set

5 hours of parallel EL speech and normal speech, respectively.
3206 mandarin pairs recorded by one Chinese female speaker. The recording is sampling at 16kHz.
2669 utterance pairs for training set, 310 utterance of pairs for evaluation set.

Conversion toolkits

- The Adobe Audition is used for noise reduction.
- The WORLD analysis and synthesis vocoder for speech analysis and speech synthesis.
- The SPROCKET for joint density Gaussian mixture model (JDGMM) training and converted speech parameters estimation.
- The KALDI toolkit for acoustic model training and phonetic feature extraction.

Investigated Systems

- LPPP: use joint vectors of low-dimension phonetic posterior probabilities (LPPP) and $F0$ sequence for GMM2 training.
- CVMCC: use joint vectors of the converted Mel Cepstral Coefficients (CVMCC) and $F0$ sequence for GMM2 training.

Evaluation and Result

Objective Evaluations

- Mel-Cepstral Distortion

$$MCD[dB] = \frac{1}{T}\sum_{t=1}^{T}\frac{10\sqrt{2\sum_{i=1}^{24}(c_i - c_i^{cov})^2}}{\ln 10}$$

- Voicing Decision Error

$$VDE = \frac{N_{V\to U} + N_{U\to V}}{N} \times 100\%$$

- Gross Pitch Error

$$GPE = \frac{N_{F0E}}{N_{VV}} \times 100\%$$

N_{F0E} denotes the number of frames for which

$$\left|\frac{F0_{i,estimated}}{F0_{i,reference}} - 1\right| > 20\%$$

- $F0$ correlation coefficient

Table. 1 The objective result of the converted systems

	EL	LPPP	CVMCC
MCD	11.56dB	8.0dB	
VDE	-	0.1225	0.1219
GPE	-	**0.6514**	0.8045
$F0$ correlation coefficient	0.0088	**0.606**	0.4606

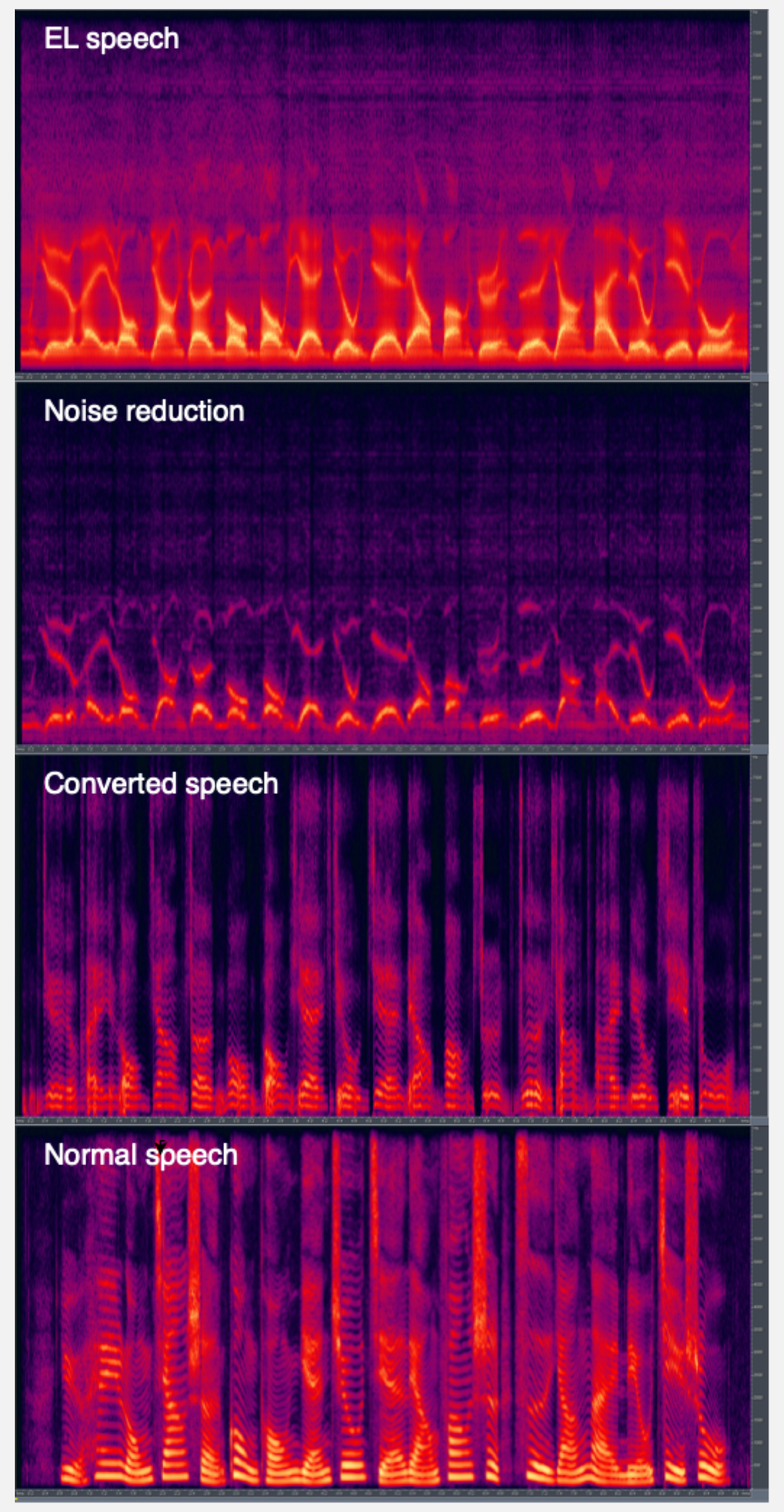


Fig. 1 Spectrograms of four speeches

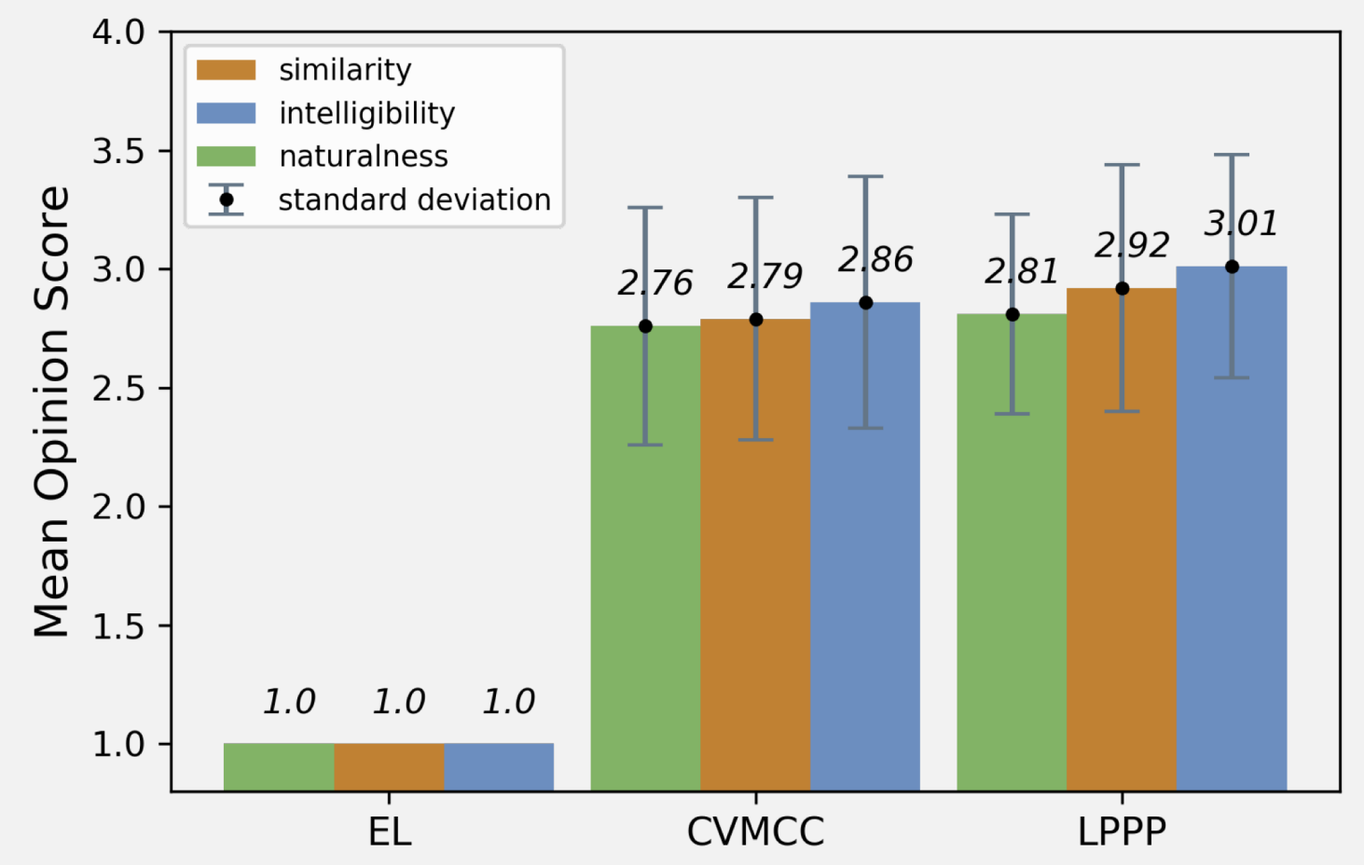


Fig. 2 The mean opinion score of subjective evaluations

Subjective Evaluations

22 native mandarin speakers scored the converted speech regarding naturalness, intelligibility and similarity.

Results

- The LPPP system outperforms significantly in comparison to the CVMCC system regarding the $F0$ correlation coefficient.
- The intelligibility and the similarity of the speech converted by LPPP system outperform those converted by CVMCC system.
- The phonetic feature rather than acoustic feature is useful in $F0$ contour estimation for EL speech enhancement.

ICASSP 2019