Enhanced Virtual Singers Generation by Incorporating Singing Dynamics to Personalized Text-To-Speech-To-Singing

Kantapon Kaewtip , Fernando Villavicencio, Fang-Yu Kuo, Mark Harvilla, Iris Ouyang, Pierre Lanchantin
ObEN Inc., Pasadena, California, U.S.A.

Motivation

- Text-To-Speech systems can already generate high-quality speech content.
- Speech-To-Singing (STS) refers to techniques generating singing voice from spoken one [1].
- However, STS-based singing may observe limitations in the perceived quality due to the use of speech content presenting "weak" or pathological voice conditions (e.g. low-energy, vocal-fry, breathiness, hoarseness).
- We propose to apply actual singing voice dynamics to a Template-based Text-To-Speech-To-Singing (TTSing) schema to generate "Virtual Singers" after voice model adaption.
- A perceived quality enhancement was achieved by following this strategy according to a subjective evaluation on Mandarin singing.

TTSing: Personalized Template-based Text-to-singing

Baseline system

- Our baseline TTSing system comprises the modules show in Fig. 1 outside the shaded region.
- Based in [2], TTS **voice personalization** is done by adapting a pre-trained model using 1 hour of speech.
- We use Spectral Autocorrelation (SAC) [3] for pitch/voicing extraction and **WORLD vocoding** [4] (using MGC and BAP features).
- Similarly to **template-based schemas** as in [5] an acappella recording is used to extract melody and timing information of the target singing content.
- for waveform reconstruction TTS-generated features are aligned in pitch, duration and energy to match the template (blocks A, E).

Enhancing TTSing By Incorporating Singing Dynamics

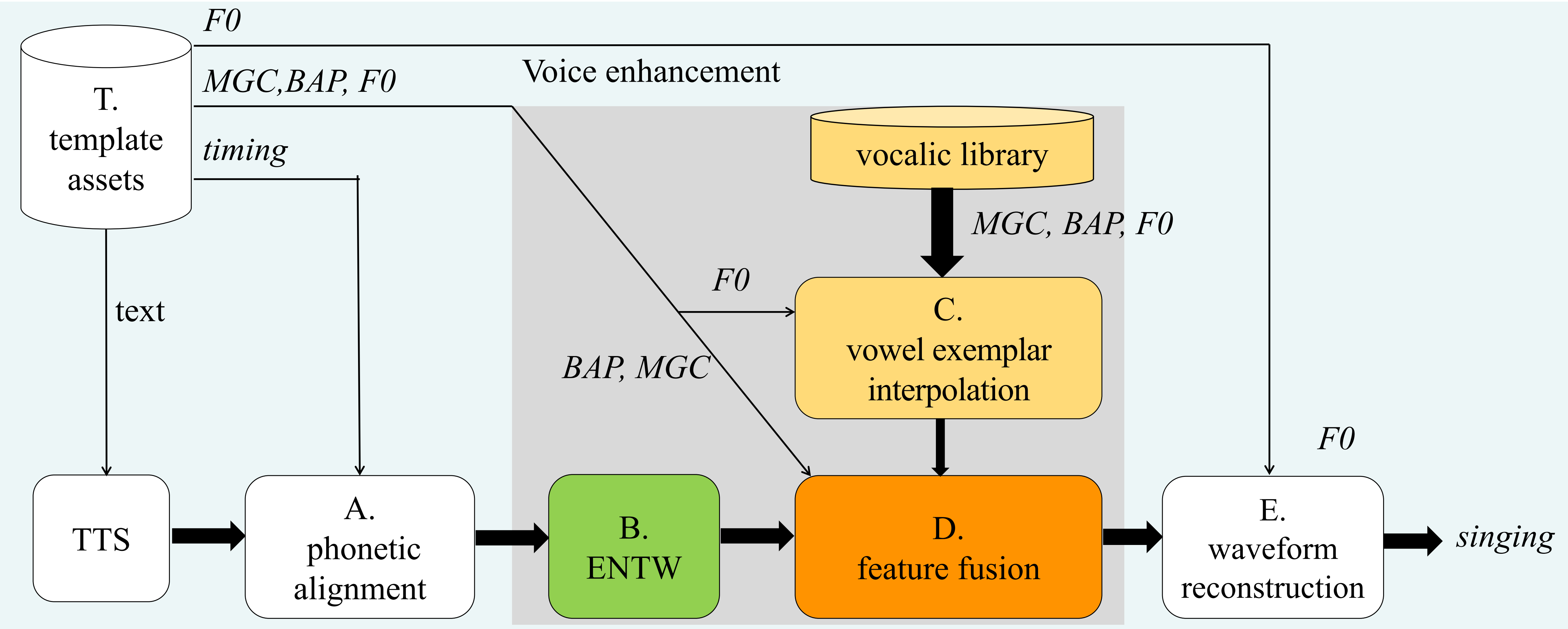

Fig. 1: TTSing system schema. The components shown in the shaded region denote the proposed enhancement strategy.

Energy-based Nonlinear Time Warping (ENTW)

- ENTW (block B) applies a **non-linear time-warping** function *d(n)* to MGC and BAP features so that the **beginning and central part** of vowels are mainly used to generate a consistent singing stream.
- For a given vowel segment, let N_1 be the first frame and N_2 be the last frame of the segment, our warping function is defined as

$$d(n) = N_1 + \frac{\sum_{m=N_1+1}^{n} e^{X_0(m,0)}}{\sum_{m=N_1+1}^{N_2} e^{X_0(m,0)}}(N_2 - N_1)$$

- Note that the total length remains the same as $d(N_1) = N_1$ and $d(N_2) = N_2$.
- The effect of ENTW is illustrated in Fig. 2 and in Fig. 3.

Acoustic fusion at phoneme transitions

•We use a ramp function between vowel/non-vowel transitions to **avoid abrupt changes** at the boundaries when applying the MGCs interpolation.

•Energy and spectral slope related features (C0, C1) are also adjusted using the template information to **ensure a smoother and natural progression of them.**

•A short-term **amplitude normalization** is applied to the reconstructed waveform using the template.

•A particular processing is applied at sonorant/obstruent transitions to **avoid amplitude instabilities** at stop, fricative or affricate sounds.

F0-driven timbre interpolation

• The spectral envelope **shows a progression** on its characteristics **across the singing pitch range** of a voice.

•A *vocalic library* of exemplars is built from short recordings of vowels sung at different music keys.

• A sequence of MGC features is obtained by linear interpolation from the exemplars based in the target pitch. The high cepstral dimensions are used for interpolation **to incorporate natural dynamics** to the fine spectral information of the warped TTS features.

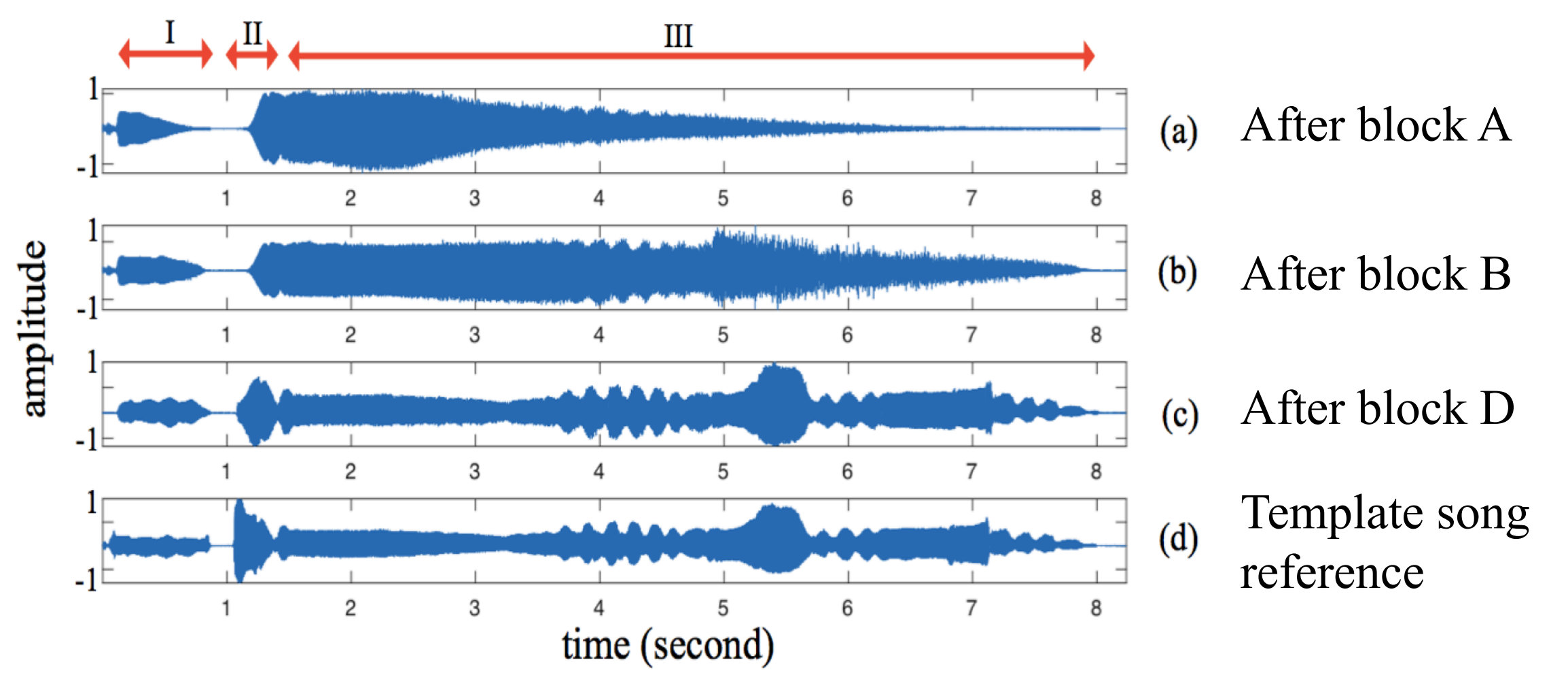

Fig. 2: Intermediate outputs in time domain: (a) - (c) are intermediate outputs and (d) is the template reference.

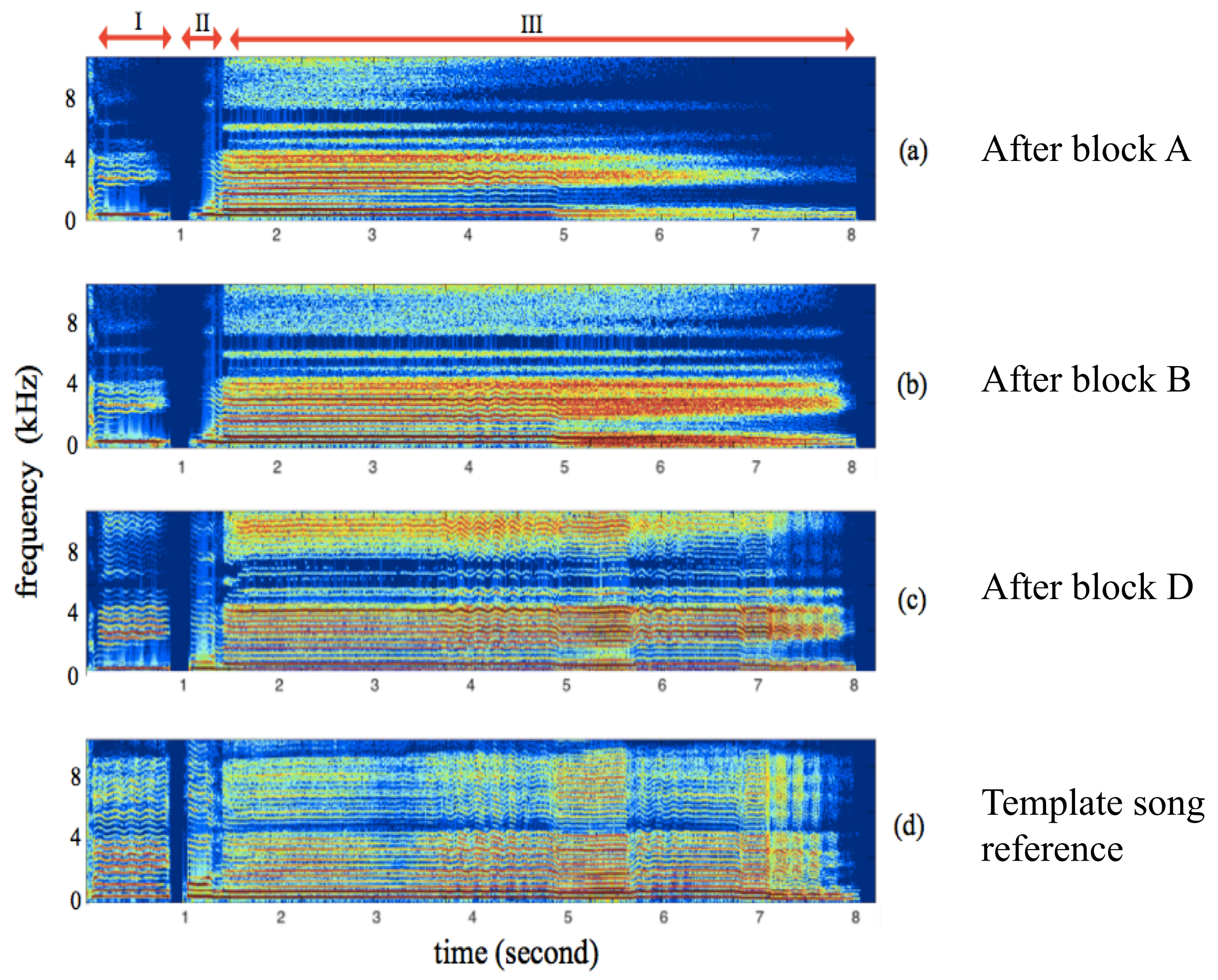

Fig. 3: Intermediate outputs in frequency domain: (a) - (c) are intermediate outputs and (d) is the template reference.

[1] T. Saitou et al., "Vocal conversion from speaking voice to singing voice using STRAIGHT," IS2007.
[2] Z. Wu et al., "Merlin: An open source neural network speech synthesis system," 9th SSW, 2016.
[3] F. Villavicencio et al., "Efficient pitch estimation on natural opera singing by a Spectral Correlation Strategy", IEICE's IPSJ-SIG 2013.
[4] M. Morise et al., "World: a vocoder- based high-quality speech synthesis system for real-time applications," IEICE's TIS 2016.
[5] L. Cen et al., "Template-based personalized singing voice synthesis," *ICASSP* 2012.

Subjective Evaluation

Evaluation Framework

•A listening test using the Comparsion Mean Opinion Score (CMOS) approach to evaluate three different methods against one another:

1. Base (baseline TTSing),
2. Enh-SD (proposed enhancement, vocalic library by the virtual singer),
3. Enh-SI (proposed enhancement, vocalic library by the opposite gender of the TTS voice).

- The purpose of our subjective evaluation is to determine:

◆ how *cleaner and healthier* (without noticeable hoarseness or lack of energy) approaches are perceived compared to the baseline

◆ if there is a significant difference between speaker-dependent (SD) and speaker-independent (SI) approaches

- 12 short singing excerpts were used to generate samples with the 3 methods. 22 native speakers of Chinese listened pairs of audio clips and were asked to compare how clean or healthy the two clips are relative to each other on a 7-point scale.

	Method effect		Gender effect	
	t-value	p-value	t-value	p-value
Base vs. Enh-SD	2.079	0.0387 *	0.180	0.8571
Base vs. Enh-SI	3.035	0.00266 **	-0.267	0.78949
Enh-SD vs. Enh-SI	0.773	0.4404	-0.303	0.7619

Table 1: Significance tests. The '*' symbol indicates significant results.

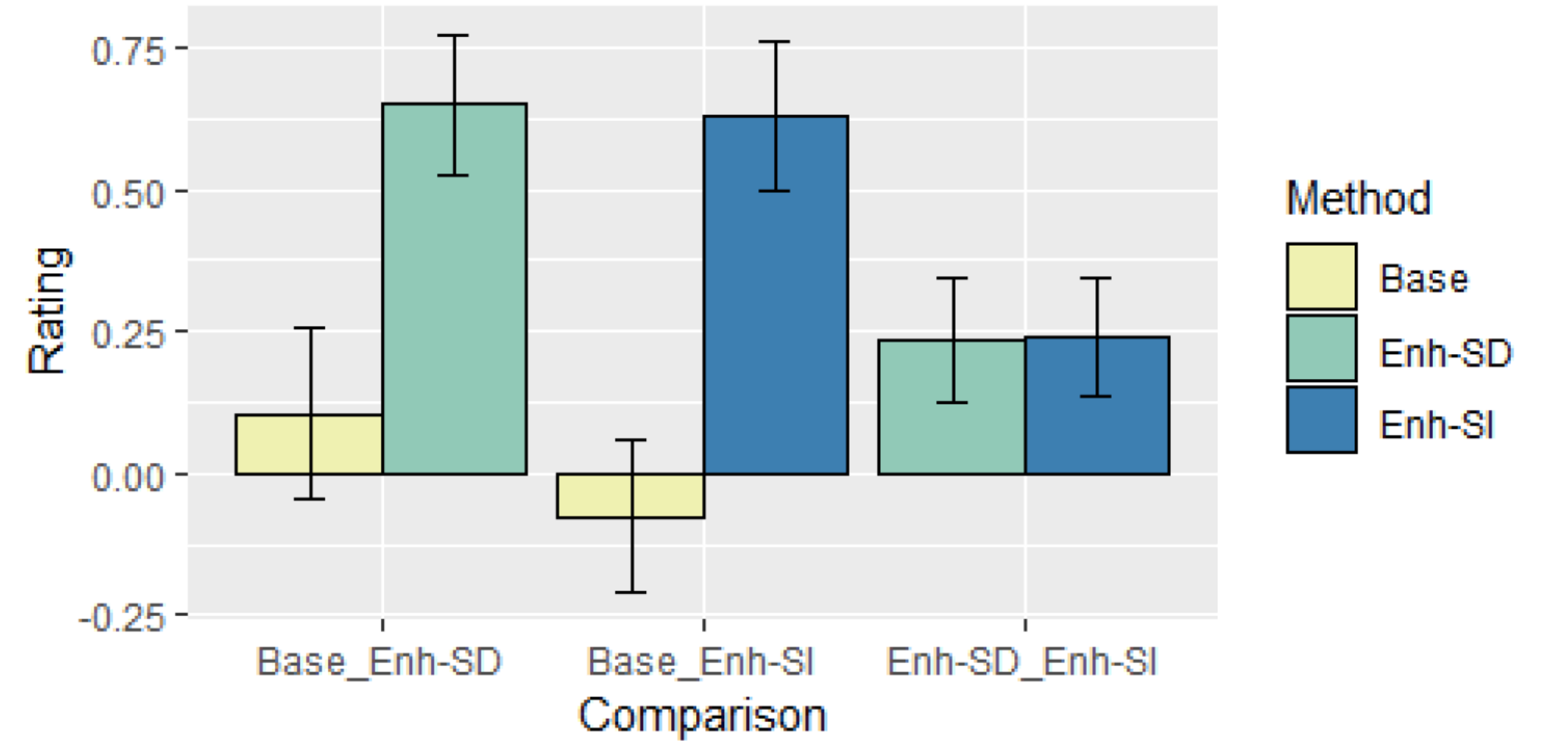

Fig. 5. Error bars show standard errors of the mean.

✓ The **proposed** method **sounds significantly better** than the baseline.
✓ This pattern **holds** for both speaker-dependent and speaker-independent methods **regardless of gender**
✓ The performance of the **speaker-independent method** (Enh-SD) was found totally **comparable** to the speaker-independent method (Eng-SI)

Conclusions

➢ We have presented a TTS-based singing framework as well as techniques to enhance the singing voice output. Key components include non-linear time warping, applying real singing dynamics and enhanced phonemes transitions processing.
➢ The listening test validates that the enhanced singing was perceived with higher quality than a baseline framework.
➢ The method is speaker and gender independent.