Generating Singing Voice Expression Contours Based On Unit Selection

Martí Umbert, Jordi Bonada & Merlijn Blaauw

{marti.umbert, jordi.bonada, merlijn.blaauw}@upf.edu

Music Technology Group

A. Introduction

Expression is a key factor to generate natural sounding singing voice synthesis performances which evoke emotions, a singing style or singer characteristics.

Motivation

- To simplify singing voice synthesis edition.
- Automatically generate control parameters from an expression DB for an input symbolic score.
- Model a singer's expression in a given style.

Interaction with Vocaloid synthesizer

To edit a song, the user:

1. Manually inputs the lyrics and notes.
2. Manually tunes the control parameters.

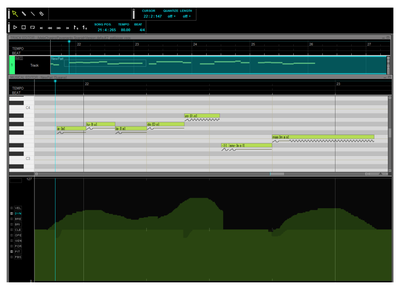

Figure 1: Interface: Notes, lyrics and control [1, 2].

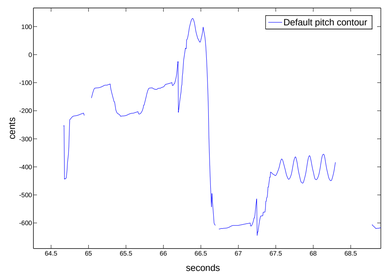

Figure 2: Generated pitch contour [3].

Overview

- Focus on local pitch and dynamics contexts.
- Concatenative approach based on unit selection, with units from recorded singing voice.
- Preservation of control parameter details of the recorded songs.

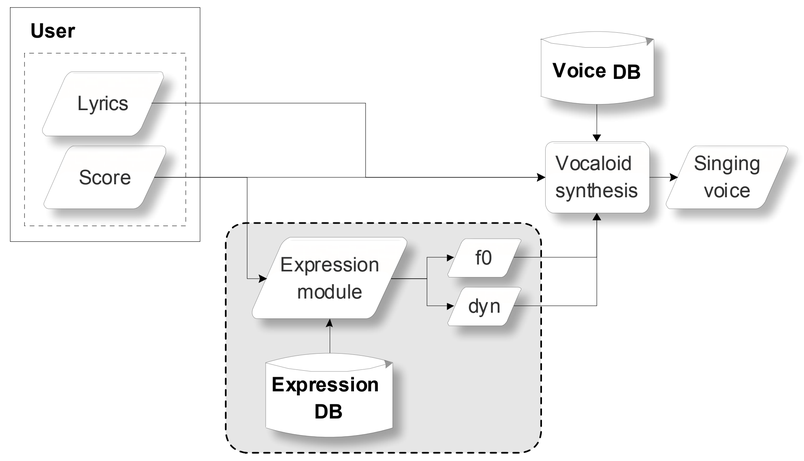

Figure 3: System interaction with the synthesis engine.

B. Proposed system

Block Diagram

The generation of the expression contours involves:

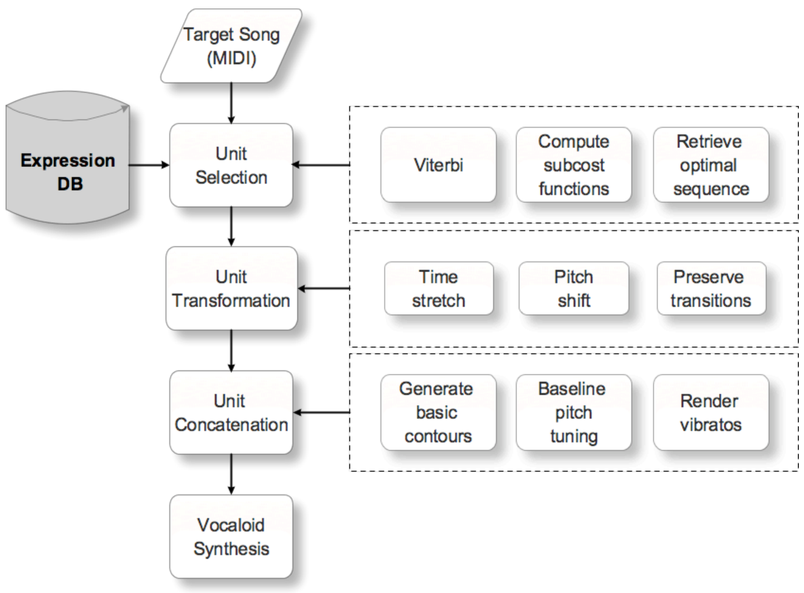

Figure 4: Proposed workflow

Expression database creation

A small set of songs is recorded in order to retrieve expression contours from shorter excerpts (units). The basic steps to create the expression database are:

1. Use /ua-i-a-i/ lyrics to avoid microprosody effect in pitch and dynamics.
2. Segment notes using GMM clustering and regression followed by manual check.
3. Estimate pitch and dynamics.
4. Estimate note transition segments.
5. Estimate vibrato parameters:

$$\tilde{F0}(n) = \bar{F0}(n) + d(n)\left[sin(\varphi(n) + \varphi_{sign}) + err(n)\right] \quad (1)$$

$$\varphi(n) = \sum_{k=0}^{n-1} 2\pi r(k)\Delta_t \quad (2)$$

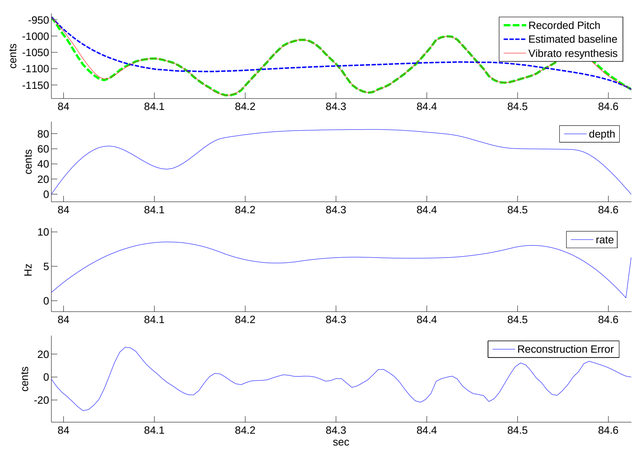

Figure 5: Vibrato resynthesis and parameters: depth, rate, reconstruction error and baseline pitch.

Unit selection

Get minimum cost sequence considering subcosts for:

- Concatenation, alternative scores, continuity.
- Transformation cost:

$$C^t(t_i, u_i) = \frac{1}{2}\left(C^t_{ts}(t_i, u_i) + C^t_{ps}(t_i, u_i)\right) \quad (3)$$

$$C^t_{ts}(t_i, u_i) = \sum_{n=1}^{3} \omega_{ts}(n) log2\left(\frac{dur(u_i(n))}{dur(t_i(n))}\right) \quad (4)$$

$$C^t_{ps}(t_i, u_i) = \sum_{n=1}^{2} \omega_{ps}(n) log2\left(\frac{int(u_i(n))}{int(t_i(n))}\right) \quad (5)$$

Unit transformation

Transform selected units to match target notes while preserving note transitions.

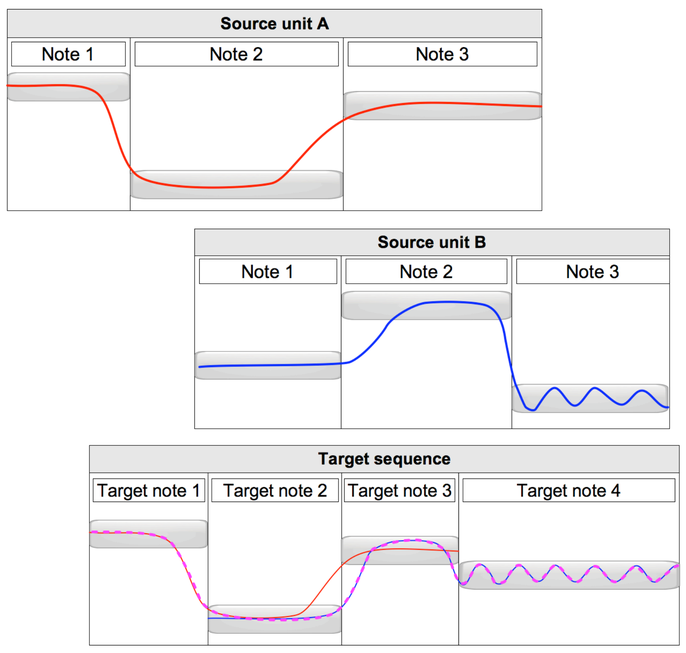

Figure 6: Transformation: pitch shift and time stretch.

Unit concatenation

Final expression contours are rendered by:

- overlapping transformed unit contours.
- synthesizing vibratos.

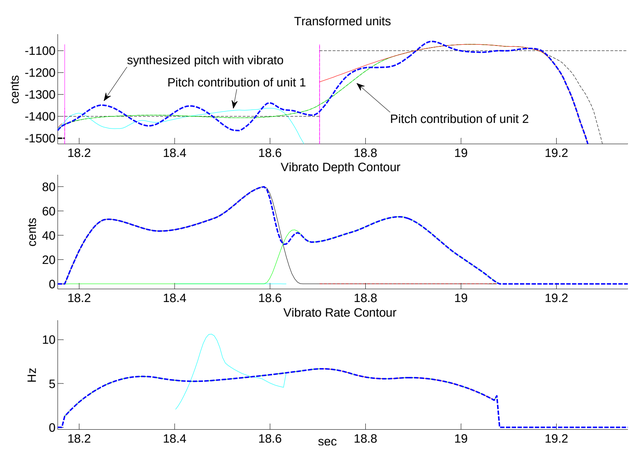

Figure 7: Transformed unit pitches and vibrato control contours concatenation.

C. Evaluation

Experimental setup

- MOS test (1-5 ratings), 16 participants.
- Compare default, manual, proposed system.
- Expression DB: four songs, six minutes.
- `http://www.dtic.upf.edu/~mumbert/smac2013/`

Results

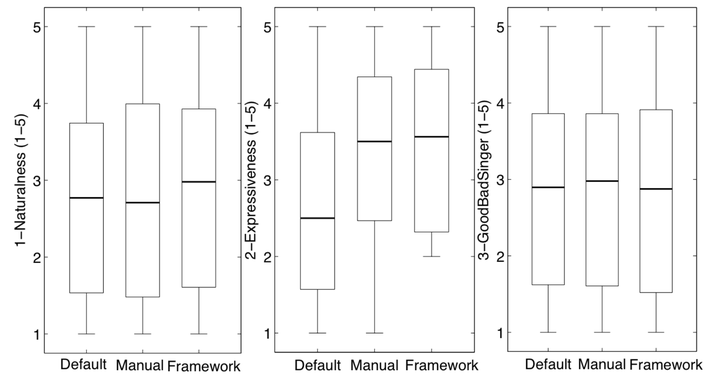

Figure 8: Results of the listening tests.

D. Discussion

Conclusions

- Statistically significant differences in expression perception.
- The system does not rely on statistical models.
- It preserves the recorded expression details.
- Singer expression and style are modeled.

Future work

- Systematic recording scripts instead of songs.
- Include richer contexts: note figure and strength, lyrics, timing deviations.
- Include tremolo in vibrato model.
- Improve cost functions.
- Compare performance with SOTA [4], [5].

D. References

[1] J. Bonada, X. Serra. Synthesis of the singing voice by performance sampling and spectral models In *IEEE Signal Processing Magazine*, March 2007: 67-79.

[2] H. Kenmochi, H. Ohshita. VOCALOID - commercial singing synthesizer based on sample concatenation In *Proc. Interspeech*, 2007: 4009-4010.

[3] J. Bonada. Voice Processing and Synthesis by Performance Sampling and Spectral Models PhD Thesis, 2008: 194 -199.

[4] K. Saino, M. Tachibana, H. Kenmochi. A Singing Style Modeling System for Singing Voice Synthesizers. In *Proc. Interspeech*. Chiba, 2010. 2894-2897.

[5] T. Nakano, M. Goto. Vocalistener: A singing-to-singing synthesis system based on iterative parameter estimation. In *Proceedings of the 6th Sound and Music Computing Conference*. Porto, 2009. 343-348.