VOCALOID – Commercial singing synthesizer based on sample concatenation

Hideki Kenmochi, *Hayato Ohshita*

Center for Advanced Sound Technologies, Yamaha Corporation, Japan
hideki_kenmochi@gmx.yamaha.com, hayato_ohshita@gmx.yamaha.com

Abstract

The song submitted here to the "Synthesis of Singing Challenge" is synthesized by the latest version of the singing synthesizer "Vocaloid", which is commercially available now. In this paper, we would like to present the overview of Vocaloid, its product lineups, description of each component, and the synthesis technique used in Vocaloid.
Index Terms: singing synthesis

1. Introduction

Vocaloid is a singing synthesizer developed by Yamaha Corporation. It is one of the few singing synthesizers that are available to end-users, and is the most widely used in the world currently. It provides end-users not merely a synthesis engine but an integrated environment in which the user can generate singing voice easily and use it for music production.

Vocaloid consists of three parts as shown in the system diagram in Figure1: (A) Score Editor, (B) Singer Library, and (C) Synthesis Engine. Each part is described in detail later.

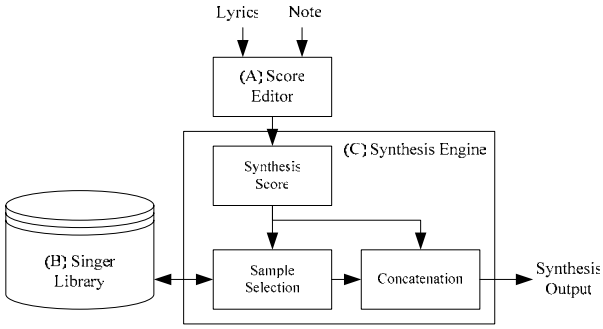

Figure 1: System Diagram

Vocaloid is not released as Yamaha's product. Yamaha licenses the technology and software to third-party companies. Those companies develop and release their own singer library bundled with the Vocaloid software.

Since 2004, five products have been released with Vocaloid version 1 so far: "Leon", "Lola", and "Miriam" from Zero-G Limited, UK, and "Meiko" and "Kaito from Crypton Future Media, Japan (Figure 2).

In 2007, Vocaloid version 2 is released. The song we submitted here is generated by Vocaloid version 2. Three or four products by Vocaloid version2 will have already been released before August 2007: "Prima" from Zero-G Limited, UK, and "Big-AL" and "Sweet Ann" from PowerFX AB, Sweden.

Leon, Lola, Miriam (version 1)

Meiko, Kaito (version 2) Prima (Version 2)

Figure 2: Vocaloid Products

2. Score Editor

The Score Editor (Editor hereafter) provides an environment in which the user can input notes, lyrics, and optionally some expressions. The Editor is designed specially for Vocaloid. The user can type-in lyrics in normal writing and the Editor automatically converts the lyrics into phonetic symbols by looking into a built-in pronunciation dictionary. If the word consists of two or more syllables, the Editor automatically decomposes it into syllables. The user can easily add vibrato in the Editor.

A screenshot of the Score editor is shown in Figure 3.

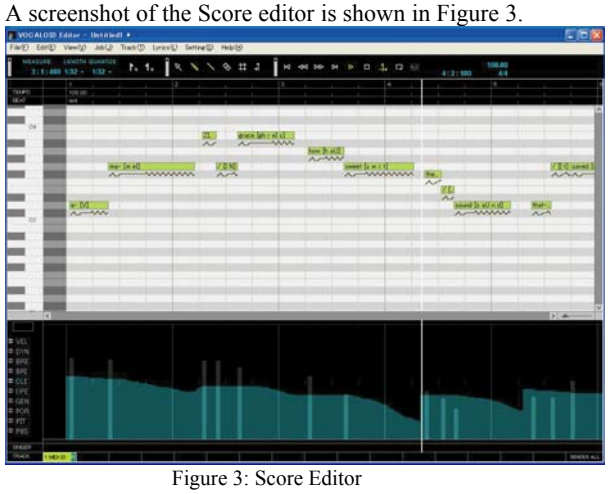

Figure 3: Score Editor

In the latest version of Vocaloid, the user can alternatively use a MIDI keyboard for inputting notes so that the user can "play" Vocaloid with pre-defined lyrics.

3. Singer Library

The Singer Library (Library hereafter) is a database of samples (mostly diphones) extracted from real people's singing.

The samples must include all possible combinations of phonemes of the target language. In English case, all possible combinations of C-V, V-C, V-V are recorded, processed, and put into the Library. The developer can optionally add polyphones with more than two phonemes.

Sustained vowels are also put into the Library. They are used to reproduce the behavior of sustained vowels, which are essential in singing synthesis.

The number of samples is approximately 2000 per one pitch. In order to record these samples effectively, a special script is designed. After a recording is done, the recorded wave files are semi-automatically segmented and necessary segments are extracted.

The Library used for the song we submitted includes samples in three pitch ranges.

4. Synthesis Engine

The Synthesis Engine receives score information, selects necessary samples from Singer Library and concatenates them. In Figure 4, a block diagram of synthesis engine is shown.

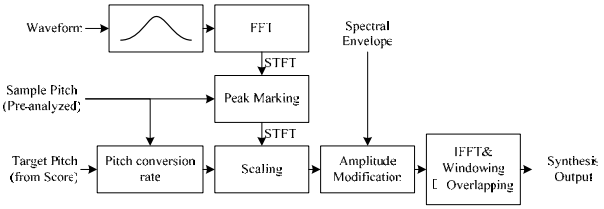

Figure 4: Block Diagram of Synthesis Engine

The problem in concatenating samples is that the samples are recorded in different pitches and different phonetic contexts, i.e. the pitch must be converted to a desired one, and the timbre must be "smoothed" around the junction of samples. In the synthesis engine, the pitch conversion and the timbre manipulation are done in frequency domain.

The pitch conversion is done by "scaling" spectrum. After getting STFT of a sample waveform, the power spectrum is divided into several regions. For each region, the spectrum is scaled so that the scaling factor corresponds to the pitch conversion. The local spectral shape near each harmonic is kept as it is (hence non-linear scaling). The timbre manipulation is done by changing an amplitude of each harmonic.

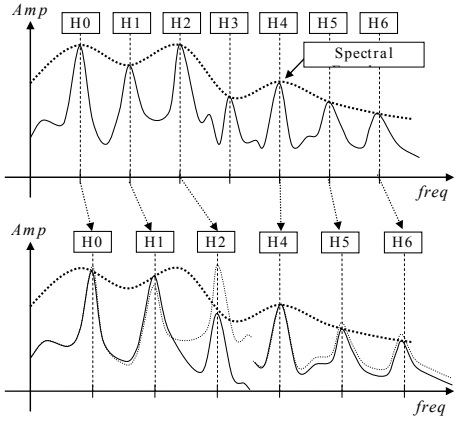

Figure 5: Pitch Conversion and Timbre Manipulation

In changing the pitch, the phase needs to be corrected. Assuming perfect harmonic, the following compensation value is added to the phase for *i*th harmonic.

$$\Delta\varphi_i = 2\pi f_0 (i+1)(T-1)\Delta t \quad (1)$$

where T is a pitch conversion ratio of f_{0T} (after conversion) and f_0 (original pitch), Δt is a frame duration.

The timbre of a sustained vowel is generated by interpolating spectral envelopes of the surrounding samples. For example, if you would like to concatenate a sequence s-e, e, e-t (which is a part of [set]), the spectral envelope of sustained [e] at each frame is generated by interpolating [e] in the end of [s-e] and [e] in the beginning of [e-t]. By doing these processes, there is theoretically no timbre gap in concatenation.

Sample timing is automatically arranged so that the vowel onset of a syllable should be strictly on the "Note-On" position.

5. Acknowledgements

The basic signal processing technique used in Vocaloid is developed though a joint research project between Yamaha Corporation and Music Technology Group (MTG), Universitari Pompeu Fabra, Barcelona. The authors would like to thank the staff at MTG, especially to Mr. Jordi Bonada and Dr. Alex Loscos, for their contribution to Vocaloid.

6. References

[1] Bonada, Loscos, Kenmochi, "Sample-based Singing-voice Synthesizer by Spectral Concatenation", Proc. of SMAC 03, 439-442, 2003.

[2] Bonada et al., "Spectral Approach to the Modeling of the Singing Voice", Proc. of the 11th AES Convention, 2001.

[3] Bonada et al., "Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models", Proc. of ICMC, 2001.Smith, J. O. and Abel, J. S., "Bark and ERB Bilinear Transforms", IEEE Trans. Speech and Audio Proc., 7(6):697-708, 1999.

[4] http://www.zero-g.co.uk/

[5] http://www.crypton.co.jp/

[6] http://www.powerfx.com