SINGING SYNTHESIS AS A NEW MUSICAL INSTRUMENT

Hideki Kenmochi

yamaha+ Division, Yamaha Corporation, Japan

ABSTRACT

Singing synthesis technology is changing musical situation in Japan recently. There are a lot of original compositions using singing synthesis softwares such as Vocaloid in a video site *Niko Niko Douga* or YouTube. Hit songs in the video sites are remastered and published as CDs from major record companies. There are also various types of derivative products and services. The author would like to discuss this phenomenon in detail.

Index Terms— singing synthesis, music, vocaloid

1. INTRODUCTION

Singing synthesis dates back to a research by Kelly et al. [1] in 1962. There have been various singing synthesizers proposed since then, whether they are commercial or non-commercial ([2][3][4][5][6]). However, those singing synthesizers have not been widely used in musical creation, although some musicians experimentally utilized them in their music.

Since 2007, there has been a big boom of musical contents using singing synthesis software especially in Japan. There can be found a lot of original musical compositions in the video sites such as YouTube or *Niko Niko Douga*. In most cases, amateur musicians use singing synthesis software to create their original compositions. Vocaloid2 is sold more than 120,000 units in total mainly in Japanese market. The author would like to briefly introduce Vocaloid as a representative of singing synthesis software.

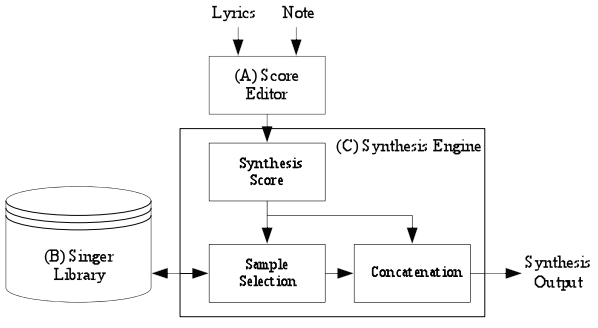

Figure 1: System Diagram of Vocaloid

Vocaloid is a commercial singing synthesis software product based on sample concatenation [7]. It should be noted that a concatenation-based singing synthesizer was already proposed by Macon et al. [8] in 1992. The uniqueness of Vocaloid is its signal processing and integrated environment for singing synthesis to end-users.

As in Fugure 1, There are three major blocks in Vocaloid system: (a) Score Editor, where users can input lyrics and notes additionally with some musical expressions, (b) Singer Database, which includes diphone and sustained vowels of a human singer, and (c) Synthesis Engine, which receives inputs from Editor (a) and select necessary segments from Singer Database (b) and concatenates them.

The first version was released in 2004, and the second version "Vocaloid2" was released in 2007. The latest version, Vocaloid3, is released in October 2011. Yamaha licenses the technology to third-party companies and those companies develop and release their own singer library.

The most popular Vocaloid library is Hatusne Miku released by Crypton Future Media. It has sold over 55,000 units since its release in August 2007, having recorded a great hit in the musical creation software market in Japan.

Besides Hatsune Miku, there are 26 Vocaloid products already released at the point of October 2011, from Zero-G Limited (UK) [9], Crypton Future Media Inc. (Japan) [10], PowerFX Systems AB (Sweden) [11], Internet Co. Ltd. (Japan) [12], AHS Co. Ltd. (Japan) [13], bplats Inc. (Japan) [14], Ki/oon Records (Japan)[15], and Yamaha Music Communications (Japan) [16].

In this paper, technical details of the singing synthesis software Vocaloid will be shown, with introduction of its social impacts in Japan. Discussion on why people like to create and listen to music with singing synthesis follows.

2. VOCALOID SYNTHESIS SYSTEM

2.1. SCORE EDITOR

The Score Editor (a) in Figure 1 provides an environment in which the user can input notes, lyrics, and optionally some expressions. The user puts notes and lyrics in a piano-roll style screen called "Musical Editor Window". At the point of October 2011, Vocaloid can handle three languages: Japanese, English and Korean (Chinese and Spanish will follow.) The user can type-in lyrics in hiragana or katakana (Japanese phonetic writing system) in Japanese case, in normal orthography in English, or in Hangeul characters in Korean. The lyrics are automatically converted to phonetic symbols. In English case, if the word has more than one

syllables, the Editor automatically decomposes the word into syllables. This is done by looking into a built-in pronunciation+syllabification dictionary with more than 120,000 words. The dictionary consists of an entry of word with syllabification, pronunciation in phonetic symbols with syllabification which is shown in Figure 2.

Word Entry	**Pronunciation**
sep-tem-bers	e p - t e m – b @r
oc-to-ber	Q k – t @U – b @r
no-vem-ber	n @u - v e m – b @r

"-" stands for syllable boundary

Figure 2: Sample of pronunciation + syllabification dictionary

The user can add vibrato to notes by operating the vibrato icon shown near the note by mouse. The user can also draw several synthesis parameters (such as "Velocity", "Pitch Bend", etc.) in the bottom of the Musical Editor Window.

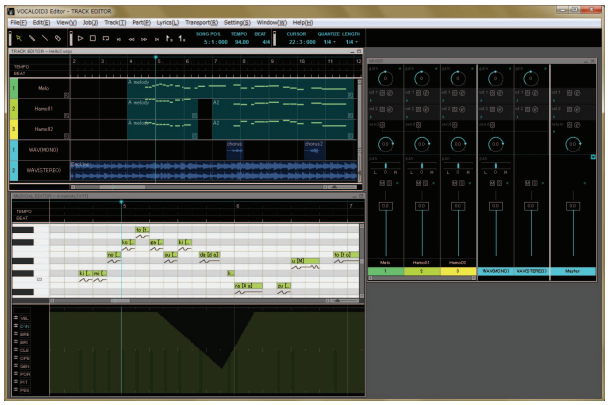

Figure 3: Score Editor (Vocaloid Editor)

Figure 3 is a screenshot of the Score Editor. The Score Editor also has "Track Window" and the user can include up to 16 tracks to make "chorus" in the software. It also has a functionality to add audio accompaniments. The user should use Musical-Part-Editor Window to input notes and lyrics. The user can use VST (Virtual Studio Technology) plugin to add effects (reverb, compressor, etc.) to the synthesis output. The user can finalize his/her composition in the software.

2.2 SINGER LIBRARY

The Singer Library (b) consists of diphones, sustained vowels, and optionally triphones. The diphones should include all possible combinations of phonemes (C-V, V-C, V-V, C-C, where C and V stand for consonant and vowel respectively) of the target language. In Japanese case, approximately 500 diphones are necessary, while in English the number of necessary diphones is 2,500. This figure can be calculated by summing up all necessary combination of phonemes in the target language. The number of phonemes and syllable structure of the target languages make the difference. In addition to diphones, sustained vowels are also stored in the Singer Library. They will be used to reproduce the behavior of sustained notes (fluctuation of pitch, dynamics, etc.), which is essential in singing synthesis. Triphones are mostly V-C-V, and can be used optionally. Triphones are effective when C in V-C-V is a consonant which is easily affected by the surrounding vowels (for instance, [h] becomes from unvoiced to voiced if it is surrounded by vowels).

In the recording, a special script to efficiently cover those diphones and sustained vowels is used. The voice donor (in most cases, professional singer or voice actor/actress) "sings" the script for several pitches, which, in most cases, are two or three in an octave, depending on the skill of the donor.

The recorded material is phonetically segmented semi-automatically to extract diphones, triphone and sustained vowels. Pitch and spectral envelope of each frame is also detected or analyzed [17][18] and stored in the Library aligned with the sample waveform.

The semi-automatic processing is checked by human operator finally, but this manual checking takes more than two months per one Singer Library in Japanese language.

2.3. SYNTHESIS ENGINE

The Synthesis Engine (c) receives notes and lyrics from the Editor, selects necessary samples from the Singer Library and concatenates them. Figure 4 is a block diagram of the Synthesis Engine.

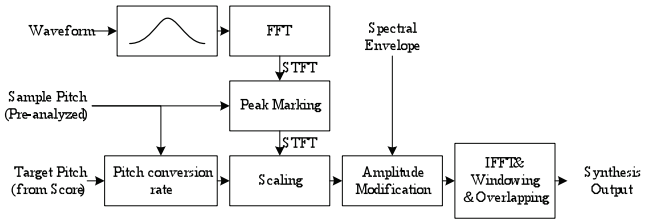

Figure 4: Block Diagram of Synthesis Engine

In concatenation, the timing of the samples is adjusted so that the vowel onset is aligned to the note-on position. This is done by having an internal "score" shown in Figure 5 and adjusting the timing of V in C-V or the second V in V-C-V to the specified note. In the internal score, synthesis parameters (such as pitch) are drawn as well. The Synthesis Engine knows which samples to be selected and transposed to which pitch (using other synthesis parameter values) at each point of time.

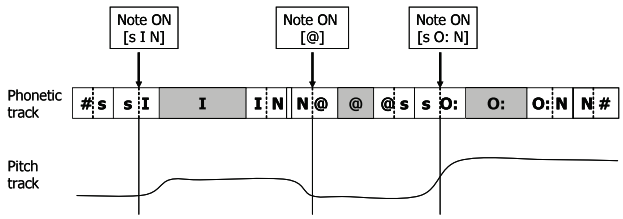

Figure 5: Sample timing adjustment and pitch curve ("Sing a song" [sIN @ sO: N])

The samples to concatenate usually have different pitches from the target one specified by the musical score. Moreover, they are recorded in a phonetic context from the desired lyrics. Therefore the pitch must be transposed to the desired one specified by the note, and the timbre must be smoothed at the boundary of the samples. The pitch transposition and timbre manipulation are done in the frequency domain.

The pitch conversion is done by "scaling" spectrum. After getting STFT of a sample waveform, the spectrum is scaled so that the scaling factor corresponds to the pitch conversion. The fine structure of spectrum near each harmonic is kept as it is, but the one between the harmonics is stretched (hence non-linear scaling). The timbre manipulation is done by changing amplitude of each harmonic.

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

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

where T is a pitch conversion ratio of $f_0 T$ (after conversion) and f_0 (original pitch), Δt is a frame duration.

Smoothing timbre is done by interpolating spectral envelope of two surrounding samples (C-V and V-C, C-V and V-C-V, etc.), in the area of the sustained vowel (V) between the two samples. In other words, the original timbre of a sustained vowel is not used, but generated by interpolation of the surrounding samples' timbres. In Figure 6, spectral envelope of [I] in "sing"[sIN] is generated by interpolating the last frame of the preceding sample [s-I] and the first frame of the following sample [I-N]. Sudden change of timbre will never happen in principle by doing this.

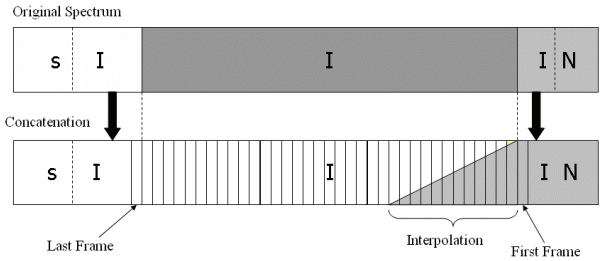

Figure 6: Interpolation of Spectral Envelope

3. VOCALOID PHENOMENON

3.1. BREAK THROUGH IN *NIKO NIKO DOUGA*

After Vocaloid2 Hatsune Miku was released in August 2007, there was a big change in the Japanese video site *Niko Niko Douga*. People began to post their original composition (with video) to the site, and some of them got quite popular. People who saw those videos and thought that they could also do the same thing went to buy Hatsune Miku, and started to make their original composition. Positive feedback took place. It sold more than 55,000 units, a record in the area of software synthesizer.

This movement led to the boom of "Vocalo-music" (music with Vocaloid). Creators who use Vocaloid in their composition are called "Vocalo-P", where "P" stands for "producer". At the point of October 2011, approximately 100 "Vocalo-music" contents in *Niko Niko Douga* have gained more than 1 million playbacks. Vocalo-P's are competing literally every night to create their own composition and make their music more popular.

It is not limited to a simple musical creation. A person who liked one of those original compositions began to add a different video to his/her favorite music voluntarily and re-posted it. Another person who liked the music changed the lyrics, re-mixed the music and re-posted it. Another person who liked the music sung the song and re-posted it. Another person makes a weekly ranking video using the original videos. People use the contents there to make a new original content, and repeat this for many times. This phenomenon can be called "Nth fanfiction" [19].

For example, the most popular music in *Niko Niko Douga* is "*Miku miku ni shiteageru*", which has more than 9 million playbacks at the point of October 2011. The number of its derivative contents as a result of the Nth fanfiction is not less than 1,500.

The reason why such "Nth fanfiction" is possible is that it has been regarded that the content in Niko Niko Douga can be freely re-used so long as the original creator does not explicitly prohibit this.

Nowadays, *Niko Niko Douga* is full of Vocaloid compositions, and the number of Vocaloid-related videos (videos that have a "VOCALOID" tag) is over 160,000, although not all of them are original compositions using Vocaloid.

3.2. REAL WORLD

"Vocalo-music" in the virtual world did not remain there. Hit songs are re-mastered and published as CD from major CD companies, and those CD's sometimes become a hit in the real world as well.

For example, there are two CD's from a major record label which topped the ORICON chart: "Vocalogenesis feat. Hatsune Miku", which was released on the 19th May 2010 [20], and "Vocalonexus feat. Hatsune Miku", which was

released on the 19th January 2011. There are some CD's from major CD record companies that were ranked in the top 10. A CD titled "supercell", released on 4th March 2009 has more than 100,000 cumulative sales.

3.3. DERIVATIVE PRODUCTS AND SERVICES

There have been a lot of derivative products or services launched since the release of Hatsune Miku. Here in this article, only a few examples of such derivative products/services are shown.

(1) Hatsune Miku features a cute anime-style character. No wonder it leads to a project to make a real figure (doll) of Hatsune Miku. At least 15 types of figure products of Hatsune Miku have already been on sale at the point of August 2010. Some of them sold more than 100,000 units.

(2) There is a PSP (PlayStation Portable) game featuring Hatsune Miku titled "Project DIVA" launched in July 2009 from SEGA. It is a musical game including the original songs posted to Niko Niko Douga. It is said that it sold approximately 200,000 units. Needless to say that a sequel titled "Project DIVA 2nd" was released in a very short period of time.

(3) JOYSOUND, one of major karaoke providers in Japan, distributes karaoke of the original Vocaloid compositions. Approximately 1300 original compositions are being distributed at the point of October 2011 and still counting. According to the yearly ranking of JOYSOUND in 2010, 6 of the top 10 are "Vocaloid" songs.

4. WHY SINGING SYNTHESIS?

You may want to ask a question to a songwriter who uses singing synthesis software. "Why do you use singing synthesis such as Vocaloid instead of real singer?" You may also want to ask a question to a person who loves Vocalo-music. "Why do you like to listen to synthetic song rather than human song?"

From the creator's point of view, the merit of using singing synthesis can be "We can make musical tracks whenever you want, even in the midnight or early in the morning without any complaints." But this cannot explain the fact that such creators who became popular still continue to use Vocaloid. The author assumes the following hypothesis:

- The synthetic voice has less emotion, so the creator needs to put his/her emotion to the song.
- As a result, the music has a direct reflection of the creator's emotion, bypassing the singer's emotion.
- Listeners can directly perceive the creators emotion through the synthetic song.
- However, we are not familiar with such direct musical communication. Therefore we need a symbol that connects creator and listener.
- The listener feels that the song is sung by the symbol, but also perceives the direct emotion of the creator.

In other words, singing synthesis software is a new musical instrument that enables songwriters to transmit his/her emotion more directly to the listeners. In this context, it can be said that singing synthesis software such as Vocaloid should not be a substitution of human singer, but something that can yield new values to music.

5. CONCLUSION

In this paper, Vocaloid is introduced with some technical details as an example of singing synthesis software. Social impact of the singing synthesis software and the reason why people use it is discussed. It will give the musical creators an opportunity to create new types of music.

7. REFERENCES

[1] Kelly, Lochbaum, "Speech Synthesis," Proceedings of the Fourth International Congress on Acoustics, pp 1-4 (1962).

[2] Cook, "SPASM: a Real-Time Vocal Tract Physical Model Editor/Controller and Singer: the Companion Software Synthesis System," Computer Music Journal, 17(1) pp 30-44 (1992).

[3] Carlson, Neovius, "Implementations of Syntheis Models for Speech and Singing,"STL-Quarterly Progress and Status Report, KTH (1990).

[4] Rodet, "Time-Domain Formant-Wave-Function Synthesis," Computer Music Journal 8(3) pp 9-14 (1984)

[5] http://www.kaelabs.com/

[6] http://www.ntt.co.jp/news/news00/0009/000907.html

[7] Kenmochi, Ohshita, "VOCALOID - commercial singing synthesizer based on sample concatenation", In INTERSPEECH-2007, 4011-4010, 2007.

[8] Macon, Jensen-Link, Oliverio, Clements, Geourge, "A singing voice synthesis system based on sinusoidal modeling," Proceedings of ICASSP 97, pp. 435-438 (1997)

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

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

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

[12] http://www.ssw.co.jp/

[13] http://www.ah-soft.com/

[14] http://www.bplats.co.jp/

[15] http://www.kioon.com/

[16] http://www.yamahamusic.co.jp/

[17] Bonada, Loscos, Cano, Serra, Kenmochi "Spectral Approach to the Modeling of the Singing Voice", Proc. of the 11th AES Convention, 2001.

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

[19] Hamano, "A note concerning Hatsune Miku or market, organization and history" (Japanese) EUREKA, vol40-15, 125-131, 2008

[20] http://www.oricon.co.jp/news/rankmusic/76554/full