VOCALOID and Hatsune Miku phenomenon in Japan

Hideki Kenmochi

Corporate Research & Development Center, Yamaha Corporation, Japan

kenmochi@beat.yamaha.co.jp

Abstract

Recently, there are a lot of original compositions using Vocaloid in a Japanese video site Niko Niko Douga. Such hit songs in the video site 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 introduce this phenomenon in detail and discuss how we can do to encourage the movement in the future.

1. Introduction ~ What is Vocaloid?

Vocaloid is a commercial singing synthesis software developed by Yamaha based on diphone concatenation. 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.

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.

The most popular Vocaloid product is Hatsune Miku. It has sold over 50,000 units since its release in August 2007, having recorded a great hit in the musical-related software market in Japan.

Table 1: List of Vocaloid products

<table>
<thead>
<tr>
<th>Product Name</th>
<th>Language</th>
<th>Gender</th>
<th>Release</th>
<th>Vendor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leon</td>
<td>English</td>
<td>M</td>
<td>2004</td>
<td>Zero-G</td>
</tr>
<tr>
<td>Lola</td>
<td>English</td>
<td>F</td>
<td>2004</td>
<td>Zero-G</td>
</tr>
<tr>
<td>Miriam</td>
<td>English</td>
<td>F</td>
<td>2004</td>
<td>Zero-G</td>
</tr>
<tr>
<td>Meiko</td>
<td>Japanese</td>
<td>F</td>
<td>2004</td>
<td>Crypton</td>
</tr>
<tr>
<td>Kaito</td>
<td>Japanese</td>
<td>M</td>
<td>2006</td>
<td>Crypton</td>
</tr>
<tr>
<td>Sweet Ann</td>
<td>English</td>
<td>F</td>
<td>2007</td>
<td>PowerFX</td>
</tr>
<tr>
<td>Hatsune Miku</td>
<td>Japanese</td>
<td>F</td>
<td>2007</td>
<td>Crypton</td>
</tr>
<tr>
<td>Kagamine Lin/Ren</td>
<td>Japanese</td>
<td>F/M</td>
<td>2007</td>
<td>Crypton</td>
</tr>
<tr>
<td>Prima</td>
<td>English</td>
<td>F</td>
<td>2008</td>
<td>Zero-G</td>
</tr>
<tr>
<td>Gackpoid</td>
<td>Japanese</td>
<td>M</td>
<td>2008</td>
<td>Internet</td>
</tr>
<tr>
<td>Megurine Luka</td>
<td>Japanese</td>
<td>F</td>
<td>2009</td>
<td>Crypton</td>
</tr>
<tr>
<td>Megpoid</td>
<td>Japanese</td>
<td>F</td>
<td>2009</td>
<td>Internet</td>
</tr>
<tr>
<td>Sonika</td>
<td>English</td>
<td>F</td>
<td>2009</td>
<td>Zero-G</td>
</tr>
<tr>
<td>Hiyama Kiyoteru</td>
<td>Japanese</td>
<td>M</td>
<td>2009</td>
<td>AHS</td>
</tr>
<tr>
<td>Kaai Yuki</td>
<td>Japanese</td>
<td>F</td>
<td>2009</td>
<td>AHS</td>
</tr>
<tr>
<td>SF-A2 miki</td>
<td>Japanese</td>
<td>F</td>
<td>2009</td>
<td>AHS</td>
</tr>
<tr>
<td>Big-AL</td>
<td>English</td>
<td>M</td>
<td>2009</td>
<td>PowerFX</td>
</tr>
<tr>
<td>Tonio</td>
<td>English</td>
<td>M</td>
<td>2010</td>
<td>Zero-G</td>
</tr>
<tr>
<td>Lily</td>
<td>Japanese</td>
<td>F</td>
<td>2010</td>
<td>Internet</td>
</tr>
<tr>
<td>VV1</td>
<td>Japanese</td>
<td>F</td>
<td>2010</td>
<td>blpls</td>
</tr>
<tr>
<td>Gachapoid</td>
<td>Japanese</td>
<td>M?</td>
<td>2010</td>
<td>Internet</td>
</tr>
</tbody>
</table>

Zero-G: Zero-G Limited (UK)
PowerFX: PowerFX Systems AB (Sweden)
Crypton: Crypton Future Media Inc. (Japan)
Internet: Internet Co. Ltd. (Japan)
AHS: AHS Co. Ltd. (Japan)
blpls: blpls Inc. (Japan)

Most of the Vocaloid users in Japan use Vocaloid as just a vocal track generator. They put lyrics and notes
with some expressions to some of the tracks and get synthesis results instantly. Normally they export the tracks to wav files and arrange them into the digital audio workstation (DAW), or synchronize the tracks with backing music on the DAW using ReWire™. They add effects such as reverb and compressors to the vocal tracks to finalize them.

2. Vocaloid Synthesis System

2.1. Score Editor

The Score Editor (a) 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. The user can type-in lyrics in hiragana or katakana (Japanese phonetic writing system) in Japanese case, or in normal orthography in English case. The lyrics are automatically converted to phonetic symbols. In English case, it the word has more than one syllables, the Vocaloid Editor automatically decomposes the word into syllables. This is done by looking into a built-in 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 in the bottom of the Vocaloid Editor.

Below is a screenshot of Vocaloid Editor.

![Figure 2: Vocaloid Editor](image)

2.2. Singer Library

The Singer Library (b) consists of diphones and sustained vowels. The diphones should include all possible combinations of phonemes (C-V, V-C, V-V, C-C) of the target language. In Japanese case, approximately 500 diphones are necessary, while in English the number of necessary diphones is 2,500. This is because of the number of phonemes and syllable structure difference of the both languages. In addition to diphones, sustained vowels are also stored in the Singer Library. They will be used to reproduce the behavior of sustained notes, which is essential in singing synthesis.

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.

The recorded material is phonetically segmented semi-automatically to extract diphones. Pitch and spectral envelope of each frame is also detected or analyzed 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.

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. Fig. 3 is a block diagram of the Synthesis Engine.

![Figure 3: Block Diagram of Synthesis Engine](image)

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 Fig. 4 and adjusting the timing of V in C-V to the specified note. In the internal score, synthesis parameters (such as pitch) are drawn as well. The Synthesis Engine can know which samples to be selected and transposed to which pitch (using other synthesis parameter values) at each point of time.

![Figure 4: Sample timing adjustment and pitch curve](image)

The samples to concatenate usually have different pitches and are recorded in different phonetic context. 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 transposition is done by scaling the spectrum. In scaling the spectrum, fine structure near each harmonic is preserved and the rest of the spectrum is stretched or compressed along the frequency axis. Hence non-liner scaling.
Smoothing timbre is done by interpolating spectral envelope of two diphones (C-V and V-C), in the area of the sustained vowel (V) between the two diphones. In other words, the original timbre of a sustained vowel is not used, but generated by interpolation of the surrounding samples’ timbres. In Fig. 5, spectral envelope of [I] in “sing”[sIN] is generated by interpolating the last frame of the preceding diphone [s-I] and the first frame of the following diphone [I-N]. Sudden change of timbre will never happen in principle by doing this.

Figure 5: Interpolation of Spectral Envelope

3. Vocaloid Phenomenon

3.1. Break through in Niko Niko Douga

After 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, and Hatsune Miku is still selling very well. It sold more than 50,000 units, a record in the area of software synthesizer.

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 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”. Apart from music, since the package of Hatsune Miku features a cute anime-style character, some people used the character to draw their original picture. It can be said that Hatsune Miku has high affinity with doujin (coterie) culture. “Nth fanfiction” took place in this context as well.

The most popular music in Niko Niko Douga is “Miku miku ni shiteageru”, which has nearly 8 million playbacks at the point of 27-August 2010. 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.

Anyway, Niko Niko Douga is full of Vocaloid compositions, and the number of Vocaloid-related videos (videos that have a “VOCALOID” tag) is over 100,000.

3.2. Real World

Vocaloid songs 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.

A CD titled “EXIT TUNES PRESENTS Vocalogenesis feat. Hatsune Miku”, which was released on 19th May 2010 became a big hit and topped the ORICON chart. 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. Shown below is one of those figure products. Of course the figure products are not limited to Hatsune Miku.

Figure 6: Nendoroid Hatsune Miku © Good Smile Company

(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 planned and released in a very short period of time. There is also an arcade game of Project DIVA.

(3) JOYSOUND, one of major karaoke providers in Japan, distributes karaoke of the original Vocaloid compositions. Approximately 250 original compositions are being distributed at the point of August 2010 and still counting. It should be noted that not only the Otaku,
but normal youngsters are enjoying the music of Vocaloid-origin in a Karaoke studio.

### 4. Why Vocaloid?

You may want to ask a question to a songwriter who uses Vocaloid. “Why do you use Vocaloid instead of real singer?” You may also want to ask a question to a person who loves Vocaloid song. “Why do you like to listen to Vocaloid song rather than human song?” The fact that a lot of people are still using Vocaloid and a lot of people (especially younger generations) like to listen to Vocaloid songs implies that its synthetic voice provides more than the human singer can provide.

From the creator’s point of view, the merit of using Vocaloid 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 the creators who became popular because of their Vocaloid music still continue to use Vocaloid. The author assumes the following hypothesis:

- The synthetic voice has less emotion, so the creator needs to put 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 Vocaloid 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, Vocaloid is a new musical instrument that enables songwriters to transmit his emotion more directly to the listeners. In this context, it can be said that Vocaloid (and other singing synthesizers) should not be a substitution of human singer, but something that can provide merits that human singers cannot.

### 5. Conclusion

In this paper, Vocaloid is introduced with some technical details. Brief history of Vocaloid after the release of Hatsune Miku is presented. Social impact of Vocaloid and the reason why people use Vocaloid is discussed. The author would like to improve Vocaloid both in terms of synthesis quality and GUI.

### 6. Acknowledgements

The author would like to thank to Mr. Wataru Sasaki from Crypton Future Media for his suggestion and discussion.

### 7. References