

# SPEAKER RECOGNIZABILITY EVALUATION OF A VOICEFONT-BASED TEXT-TO-SPEECH SYSTEM

*Masaharu Sakamoto, Takashi Saito*

IBM Research, Tokyo Research Laboratory, IBM Japan Ltd.  
1623-14, Shimotsuruma, Yamato-shi, kanagawa-ken, 242-8502, Japan  
[sakamoto@jp.ibm.com](mailto:sakamoto@jp.ibm.com), [saito@jp.ibm.com](mailto:saito@jp.ibm.com)

## ABSTRACT

We have developed a new text-to-speech system based on the VoiceFont technology. A VoiceFont is a voice dictionary for speech synthesis that holds the acoustic and prosodic characteristics extracted from the voice corpus of a speaker. The text-to-speech system using a VoiceFont is able to synthetically mimic the voice of the donor speaker. In this paper, we evaluated speaker recognizability of the synthetic speech, which means whether the synthetic speech can be recognized as the donor speaker's voice. We conducted a subjective evaluation for five VoiceFonts and here report on the evaluation results. The results show that our text-to-speech system based on VoiceFonts can retain the acoustic and prosodic characteristics of the donor speaker and the synthetic speech can be recognized as the donor speaker's voice. Furthermore, we report on how much the spectral characteristics, phoneme duration, and pitch frequency affect speaker recognizability.

## 1. INTRODUCTION

In recent years, text-to-speech systems are expected to not only read aloud orthographic transcriptions with good intelligibility, but also to synthesize a high quality mimicked voice of a donor speaker. The mimicking voice function could extend the application field of text-to-speech systems. Some examples are:

- The application scenario of a voice response system will can be changed at any time, since new recording would not be necessary.
- The voice response system could respond with a famous actor' s voice.
- For man-machine communications, it could be used in order to give personality to robots and CG characters.

Several corpus-based waveform concatenation speech synthesizers have been reported [1, 2, 3, 4, 5]. These systems have the potential of mimicked speech synthesis. In [1, 2], a new text-to-speech system based on the VoiceFont technology was introduced. This text-to-speech system is able to synthetically mimic the voice of the donor speaker. We describe the outline of the VoiceFont technology later. In [3], a corpus-based waveform concatenation speech synthesizer named CHATR was introduced. It uses a database of recorded speech and a unit selection algorithm that selects the segments that best match the utterance to be synthesized. In [4],

statistical training techniques applied to a large corpus are used to make decisions about predicted speech events and selected speech inventory units. In [5], a limited domain synthesis technique was proposed. In that study, a speech database close to the target domain of the speech application was introduced. Since the small and limited domain speech database is a guarantee of reliable unit selection, it can generate very high quality synthetic speech. However, it needs a speech database for each application.

Although these methods have the potential of mimicking voices, they have not investigated speaker recognizability. In this paper, we conducted a listening test to evaluate the speaker recognizability of the synthetic speech using a VoiceFont. Speaker recognizability means whether the synthetic speech can be recognized as the voice of the donor speaker. In this listening test, the respondents were asked to evaluate the synthesized voices on a scale from "Very different" to "Very similar". In each test, they were comparing one synthetic voice to one original recording. As a result, we reached the conclusion that the system produced recognizable voices.

This paper is organized as follows. In Section 2, we describe the outline of the VoiceFont system. Section 3 describes the speech corpora and the profiles of the five VoiceFonts. In Section 4, we report on the subjective evaluations. Section 5 discusses the subjective evaluation results. In the last section, we give a brief summary of this study.

## 2. VOICEFONT TECHNOLOGY

### 2.1. Concept

Our VoiceFont technology is a method of creating a speaker's voice database to synthesize a mimicked voice in a text-to-speech system. The voice features we intend to extract are of two kinds: segmental (phonetic) and suprasegmental (prosodic). The segmental features as well as the prosodic features are simultaneously extracted for direct use in the same speech synthesizer. However, the automatic voice feature extraction is not perfect, so an easy-to-use checking and editing is essential for the creation process. As an efficient tool for database creation, we made a VoiceFont creation system, "**VoiceFont Builder**" [2]. It is designed to make the creation process easier and more effective. Customization functions like registering

the application's vocabulary are essential for a text-to-speech system to make it adaptable to a wide range of applications.

Figure 1 shows the configuration of the system. The input of the system is a speech corpus of a new speaker, which consists of speech data and the corresponding Japanese orthographic text in mixed kanji/kana form. The output is the speaker's VoiceFont, a voice dictionary for that speaker. VoiceFont Builder has three functions: voice feature extraction, data checking/editing, and dictionary generation. A typical procedure of VoiceFont creation using the system is as follows:

## 2.2. Script Checking

First, text-to-phoneme conversion of the utterance scripts of a speech corpus is carried out to check for reading errors from the text-to-speech synthesizer. To correct errors in the reading, all unknown words must be registered in the word dictionary of the synthesizer. This checking process is required to run the following feature extraction successfully, since it uses reference templates provided by the synthesizer for phonemic alignment.

## 2.3. Voice Feature Extraction

The voice features are automatically extracted for the speech corpus using the word dictionary prepared for the script set. Through this procedure, input speech utterances are decomposed into multi-layered segments: breath groups, accent phrases, phonemes, and pitch marks. The approach we take here to the automatic extraction of these features is based on a simple phonemic alignment of the input utterance with reference templates generated by the text-to-speech synthesizer. Our strategy is very simple, but has several notable merits. First, phonetic and prosodic boundaries from the alignment are basically well suited and consistent for use as synthesis control parameters, since the reference templates are generated by the same criteria necessary for speech synthesis. Second, the accuracy of phonetic segmentation is very high, mainly because the spectral context dependency expressed by the synthetic templates is fairly appropriate, and as a result, we can use DTW to align between the mel-scale cepstral coefficients and the delta coefficients of the recorded and synthesized waveforms [1]. The details of this procedure are described in [2]. The extracted features are as follows:

- Phoneme duration information (prediction parameter).
- F0 unit (the target prediction parameter and F0 pattern).
- Waveform unit (phoneme context dependent CV unit).

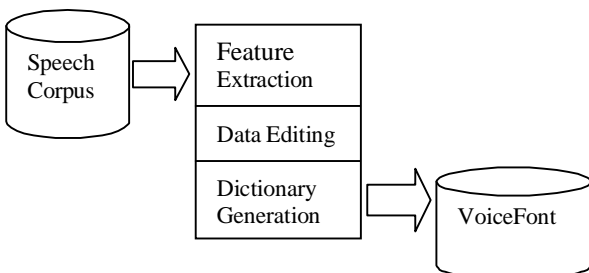


Figure 1: VoiceFont building system.

## 2.4. Data Checking and Dictionary Generation

If segmentation errors or pitch marking errors are found, these portions can be checked, and corrected or replaced by using the data editor. Figure 2 shows the output of extracted features from the data editor. All the parameters for the VoiceFont creation can be checked in the GUI environment. Finally, a VoiceFont for the given corpus is generated by compiling the segmented speech data into three kinds of VoiceFont information: an acoustic unit inventory, an F0 unit inventory, and a duration parameter set.

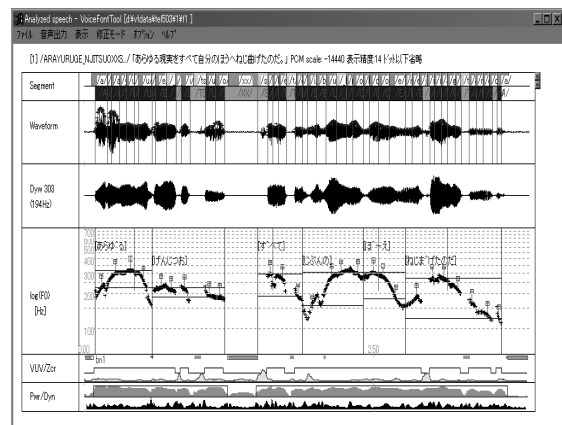


Figure 2: Output of voice features using the data editor.

## 3. VOICEFONTS

### 3.1. Speech Corpora

We recorded four professional radio actors reading aloud some scripts with their special speaking styles or characters. As a result, we obtained five speech corpora (because one female speaker did readings as two different characters). Table 1 shows the profiles of the speech materials, the speaking styles and the scripts of each speaker. For the purposes of comparison, we prepared two small waveform dictionaries that are made from the M0 and F0 corpora, respectively. Since M0 and F0 contain only 1576 words, their waveform dictionaries consist of a single waveform unit for each phoneme context. In order to distinguish these from a VoiceFont, we call them waveform dictionaries. For the waveform dictionaries, the speech synthesizer engine gave the pitch pattern and phoneme duration by following rules.

The utterances were recorded on DAT tapes with 16 bit resolution, at a 48 kHz sampling rate by a professional recording engineer at a recording studio. After that, the recorded data on the DAT tapes was transferred to a PC and simultaneously downsampled to a 22.05 kHz sampling rate.

### 3.2. VoiceFonts

We created five VoiceFonts from the five speech corpora. The file size, numbers of waveform units and F0 units of each VoiceFont are shown in Table 2. The file size of the waveform dictionaries of M0 and F0 are also shown. For 22 kHz and 16-

bit PCM, the waveform units occupy about 90% of the total file size.

Gender	Age	ID	Character	Text materials
Male	40	M1		name, company name, address, phone number, voice response sentence, phonetically balanced sentences
Female	30	F2		The same as above
Female	50	F2B	Young boy	phonetically balanced sentences/words, monologue
Female	50	F2G	Young girl	phonetically balanced words, monologue
Female	40	F3B	Young boy	story, phonetically balanced sentences, conversational sentence, address, phone number, name, company name
Male	20	M0		phonetically balanced words
Female	30	F0		phonetically balanced words

Table 1: Speech corpora

ID	file size [MB]	# of waveform units	# of F0 units
M1	296	34931	5495
F1	300	35650	7448
F2B	240	28433	6360
F2G	140	13502	3190
F3B	375	48814	10603
M0	18	2280	
F0	23	2241	

Table 2: The sizes of the VoiceFont information

## 4. EVALUATIONS

### 4.1. Subjective Evaluations

We conducted a listening test that evaluated whether the synthetic speech preserved the speaker recognizability. Sixteen listeners rated the speaker individuality of 16 synthetic speeches on a scale from -3 (very different) to 3 (very similar). When synthesizing the test speech, in order to remove the influence of text analysis, reading errors and accent errors were corrected. The evaluation objects were the five VoiceFonts and two waveform dictionaries. In the experiment, the sound was output with a loudspeaker. In the subjective experiment, we consider the following points:

- The respondents were asked to evaluate the synthesized voices on a scale from "Very different" to "Very similar".

In each test, they were comparing one synthetic voice to one original recording.

- In order to calibrate the evaluation measure, first we played his or her recorded speech and another speaker's recorded speech.
- The listeners could hear the sample recorded speech at any time.
- The respondents were asked to ignore any qualities other than speaker recognizability, such as spectrum discontinuity.

Evaluation	Rate
Very different	-3
Fairly different	-2
Little different	-1
Fair	0
Little similar	1
Fairly similar	2
Very similar	3

Table 3: Evaluations and scoring

	VoiceFont	M0, F0
her/his recorded voice	2	2
other recorded voice	1	1
Synthetic speech	5	5
Waveform units substitution	2	
F0 units substitution	2	
Phoneme duration substitution	2	
Synthetic speech by rule	2	
Total	16	8

Table 4: The numbers of experimental trials

Table 3 shows the evaluations and the scoring used. Table 4 shows the number of trials for each test category. The trials include several synthetic voices synthesized by substituting a part of a VoiceFont, such as replacing the waveform units, F0 units, or phoneme durations with those from another donor speaker's VoiceFont. Also, the trials include several synthetic voices that were synthesized by using a prosody rule, such as the Fujisaki model. The total number of sentences for each VoiceFont was 16 and for each waveform dictionary was 8. The total number of trials for each respondent was 96 and the time required was about 20 minutes. The sentences that were used for the trials were not used in the creation of the VoiceFonts.

### 4.2. Result

The evaluation results for each VoiceFont and each waveform dictionary are shown in Table 5. The scores in the table are mean scores over all respondents. Table 6 shows the evaluation results for VoiceFont involving substitutions.

## 5. DISCUSSION

The mean scores in Table 5 show that the synthetic speeches using the VoiceFonts are similar to the donor speaker's voice. Meanwhile, the synthetic speeches using the single instance

waveform dictionaries that are made from M0 and F0, respectively, are different from the donor speaker's voice. The results also show that the respondents distinguished a donor's recorded voice from another speaker's recorded voice. The frequency distributions of the evaluation scores of each VoiceFont and the two waveform dictionaries are shown in Figure 3. Although the score varies for each respondent, most respondents answered that the synthetic speeches using VoiceFonts are similar to the donor speaker's voice, and the synthetic speeches using the single instance waveform dictionaries are different. The respondents were asked to ignore any qualities other than speaker recognizability, such as spectrum discontinuity. However, because ignoring such qualities was difficult for some respondents, the score varied for each of them. From comparison of Table 2 and Table 5, we can conclude that the file size of VoiceFont is not especially significant in determining the degree of speaker recognizability. Table 6 shows that the most significant factor for speaker recognizability is the waveform units. The next most significant factor is the F0 patterns. The influence of phoneme duration is comparatively small. These results support the conclusions of [7].

## 6. CONCLUSIONS

We conducted a listening test to evaluate the speaker recognizability of the synthetic speech using VoiceFonts. The result shows that our text-to-speech system based on VoiceFonts can retain the acoustic and prosodic characteristics of the donor speakers and the synthetic speech can be recognized as the donor speaker's voice. The result also shows the most significant factor for speaker recognizability is the waveform units. The next most significant factor is the F0 patterns. The influence of phoneme duration is comparatively small.

ID	Other voice	VoiceFont	Her/his voice
M1	-2.93	1.29	2.97
F1	-3.00	0.87	2.77
F2B	-2.93	1.17	2.70
F2G	-2.87	1.12	2.83
F3B	-2.93	1.28	2.73
M0	-2.80	-1.55	2.93
F0	-3.00	-1.33	2.87

Table 5: The evaluation results for each VoiceFont. A positive score shows that it is similar to the original recording and a negative score shows that it is different from the original recording.

ID	Waveform units	F0 units	Phoneme Duration	Rule
F1 - F2B	-2.57	0.30	0.67	0.33
M1 - F2B	-2.80	-0.83	1.17	1.00
F2B - F1	-2.13	-0.03	0.53	0.77
F2G - F1	-2.37	-1.07	0.30	0.97
F3B - F1	-2.47	-1.63	-0.57	0.87

Table 6: The evaluation results for substituted VoiceFont components. F1-F2B shows that the waveform units,

phoneme durations, or waveform units of F1 were replaced with one from F2B.

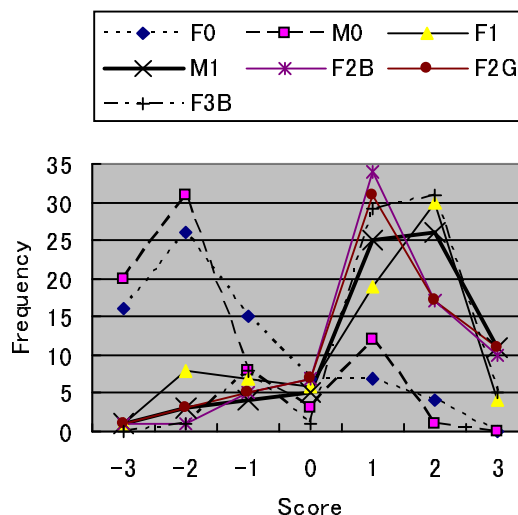


Figure 3: The frequency distribution of the evaluation scores

## 7. ACKNOWLEDGEMENT

I thank the respondents who cooperated in this experiment.

## 8. REFERENCES

- [1] Saito, T., "A method for registering new voices in a text-to-speech synthesizer", Proc. of 3rd ASA & ASJ Joint Meeting, 1057-1060, 1996.
- [2] Saito, T., Sakamoto, M., "A method of creating a new speaker's voicefont in a text-to-speech system", Proc. of ICSLP 2000 Beijing, 2000: 771-774, 2000.
- [3] Campbell, W. N., "CHATR: A high-definition speech re-sequencing system", Proc. of 3rd ASA & ASJ Joint Meeting, 1223-1228, 1996.
- [4] Syrdal, K. A., Wightman, C. W., Conkie, A., Stylianou, Y., Beutnagel, M., Schroeter, J., Strom, V., Makashay, M., Lee, K., "Corpus-based techniques in the AT&T NextGen synthesis system", Proc. of ICSLP 2000, 3: 410-415, Beijing 2000.
- [5] Black, A. W., Lenzo, K. A., "Limited Domain Synthesis System", Proc. of ICSLP 2000 Beijing, 411-414, 2000.
- [6] Sakamoto, M., Saito, T., "An automatic pitch-marking method using wavelet transform", Proc. of ICSLP 2000, Beijing, 2000, 3: 650-653, 2000.
- [7] Kuwabara, H., "A perceptual experiment on voice individuality by altering pitch and formant frequencies", Proc. of 3rd ASA & ASJ Joint Meeting, 829-832, 1996.