3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

GENERATION OF NON-ENTRY WORDS FROM ENTRIES OF THE NATURAL SPEECH DATABASE

Yasuhiko Arai†, Toshimitsu Minowa†, Hiroko Yoshida†,
Hirofumi Nishimura†, Hiroyuki Kamata‡ and Takashi Honda‡

†Matsushita Communication Industrial Co., Ltd.
600 Saedo-cho, Midori-ku, Yokohama, 226 Japan
‡School of Science and Technology, Meiji University
1-1-1 Higashi-mita, Tama-ku, Kawasaki, 214 Japan

ABSTRACT

In this paper, we describe a method to generate non-entry words from entries of the natural speech database which an automatic public announcing system is possessed of. Thereby, it becomes unnecessary to record new voices by a narrator.

Non-entry words are generated by means of the waveform editing, that is, by the method of segmental speech sound concatenation. In case that there is no need to change the pitch pattern at editing, quality of the generated words is maintained to the level of natural speech sound.

In case that the pitch pattern must be changed at editing, the zero-phased pitch waveform superposing method is used for pitch modification. In order to extract raw pitch waveforms, various windows including the Hanning and the Blackman-Harris whose length are proportional to the pitch period are tested. And, following results are obtained: (1)The Hanning window whose length is twice the pitch period is slightly superior to the Blackman-Harris windows. (2)Quality degradation of the generated words is a little bit.

I. INTRODUCTION

Automatic public announcing systems are currently used in subway stations, airports and other places. Those sysytems present information on arraivals, departures and so on, to the general public by speech. The systems form one-way communication to the general public. In order to make thorough understanding, a clear and natural speech sound is required. Therefore, in the current systems, the PCM-recorded voices from which messages are automatically compiled by means of the word-concatenation are utilized.

The PCM-recorded voices compose a natural speech database containing some sentence patterns and a set of words. Some words in a sentence are replaced to other words by compilation software commands. In order to obtain a clear and natural speech sound, on the occasion of speech recording, a narrator arranges a tone and an articulation of each word so as to meet those of the sentence to be embedded in.

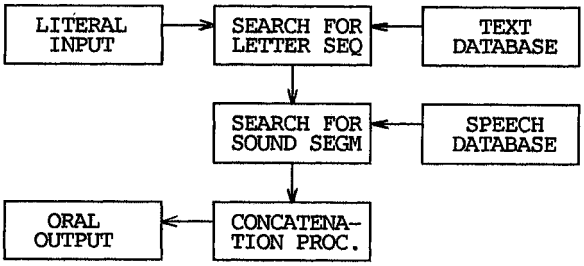


Fig.1 A system configuration for non-entry word generation.

Therefore, in the current systems there has been the inherent problem that flexibility of sentence generation is restricted depending on narrator's vocalization. Further, during long-term operations, the system sometimes required additional words. Accordingly, there has arisen the new problem that voice quality of additional records mismatches with that of old records, because voice quality changes as the narrator advances in age.

The problems above mentioned would be solved, if non-entry words could be generated from the entry words in the natural speech database by means of the segmental speech sound concatenation. The current automatic announcing systems are possessed of the natural speech database which contains speech sound of 3-to-15-mitute long altogether. It is considered that generation of the non-entry words is possible from the database of that size.

II. A VOICE EDITING SYSTEM

Figure 1 shows a configuration of a voice editing system which generates non-entry words from entries of the natural speech database. Letters of a non-entry word are put in through the literal input. And, candidate letter sequences for new word generation, which include at least a portion of letter sequence same as the input word, are found out in the text database. The candidate letter sequences may be sentences, phrases and/or words. Here, a set of candidate letter sequences of which coincident portions collectively complete the letter sequence same as the input word, is selected.

The text database and the speech database form a pair so that it is able to find out segmental speech sounds corresponding to partly coincident letter sequences. The sound waveforms in the natural speech database must have marked frames which point out phoneme boundaries for reading out segmental speech sounds. But phoneme boundaries may be roughly defined so as to indicate phonetically steady portions of the speech sounds.

Thus, the read-out speech sound segments are concatenated to complete the phonetic word corresponding to the input letter sequence. In concatenation processing, the accent type, the rhythm and the intensity are modified so that the generated word sounds naturally.

III. CONCATENATION PROCESSING

3.1 Selection of concatenation units

The Concatenation units which are the partly coincident letter sequences, are extracted from the narrowed-down candidate letter sequenses. The candidate letter sequences are narrowed down by the following strategies[1]:

(1)keeping CV-interaction,
(2)keeping interaction between voiced sounds,
(3)including longer concatenation unit, and
(4)including more overlapping letters between concatenation units.

Finaly, a set of concatenation units which complete collectively the same letter sequence as the input word is selected so that the segmental speech sounds corresponding to those concatenation units are possibly joined together by sharing same speech sounds at joining points respectively. However, connections by diphthongs, semivowels, contracted sounds such as pya, rya, kyu and so forth(Japanese yôon), nasal sounds and flapped sounds are avoided, because joint sounds are sometimes phonetically debased.

3.2 Concatenation of segmental speech sounds

The segmental speech sounds are read out from the speech database. And, those segments are connected together by the following method[2] depending on the sharing speech sound:

(1)In case of unvoiced fricative sounds and an affricate(ts-sound), connecting at the place where the low frequency power is minimum.

(2)In case of unvoiced plosive sounds and an affricate(tʃ-sound), connecting at the place where the full range power is minimum.

(3)In case of voiced sounds, connecting at the place where the spectral distance is minimum.

Signal processing for the concatenation is performed on the short frame of the speech waveforms. One of 3 methods above mentioned is applied to the overlapping portion of consecutive two speech sound segments. And, a pair of joint frames are determined. That pair of frames are merged into one frame by using a pair of tapered windows. Thus, selected speech sound segments are all joined togeter in succession.

At that time, amplitude of the following side frame is multiplied by a modification coefficient so that power of the following side frame becomes equal to that of the preceding side frame. The modification coefficient is gradually returned to 1 whthin a following few frames.

3.3 Phase adjustment

In case that speech sound segments are connected at the voiced sound, the phase of the fundamental frequency must be adjusted between two joint frames in order to avoide phase distortion. Phase adjustment is performed by shifting back the following side frame by the delay time where the correlation value between the frames becomes maximum.

The phonetic words of the speech database are all recorded in similar tone, therefore natural sounding non-entry words are easily generated by power modification and phase adjustment when the pitch modification is not required.

3.4 Change of the pitch pattern

In case of that the pitch pattern must be changed on concatenating speech sound segments, one of the pitch patterns, that is suitable for the new word, is borrowed from the database. And, the pitch waveforms are re-superposed according to the borrowed pitch pattern. Here, in order to avoid distortion caused by waveform discontinuity, the zero-phased pitch waveforms are put to use.

3.5 Adjustment of the phoneme durations

Finaly, the phoneme durations are adjusted by means of the PICOLA(Pointer Interval Control OverLap and Add) method[3] to obtain rhythmically natural sounding speech.

IV. EXTRACTION OF PITCH WAVEFORMS

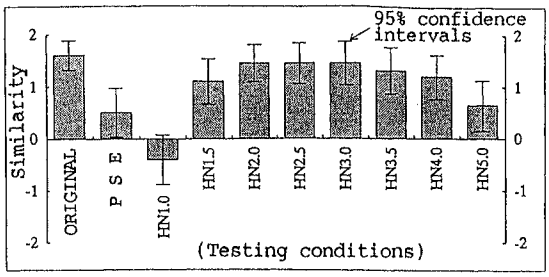
4.1 Window length

As well-known, the Hanning window is used for extracting pitch waveforms in pitch-synchronous waveform processing[4]. The window length is usually 2 to 4 times the pitch period.

The authors studied on the optimum window length by the subjective evaluation method. The female voice *"atami"* was digitized in 12 bits at the sampling rate of 10 kilo-samples/s. Pitch waveforms were extracted by using the various Hanning windows whose lengths were 1 to 5 times the pitch period, and transformed to the zero-phase impulse responses, namely zero-phased pitch waveforms, by means of zero-padded 1024-point FFT operations. Zero-phased waveforms were re-superposed according to the original pitch periods to yield synthesized speech.

Synthesized speech sounds were evaluated concerning similarity to the original speech sound on the 5-point rating scale by ten male subjects. The original sound, the PSE(Power Spectrum Envelop[5])-synthesized sound and above synthesized sounds were recorded in the audio tape in the order of testing conditions as shown in fig.2 and in the reverse order. Subjects heard recorded speech sounds through headphones and judged the similarity. Results are shown in fig.2.

Figure 2 shows that it is possible to yield very similar sounds to the original sound by



+2:same as original sound,
+1:hard to distinguish,
0:slightly different,
-1:fairly different and
-2:considerably different.

Fig.2 Subjective evaluation on the window length (in case of the Hanning window).

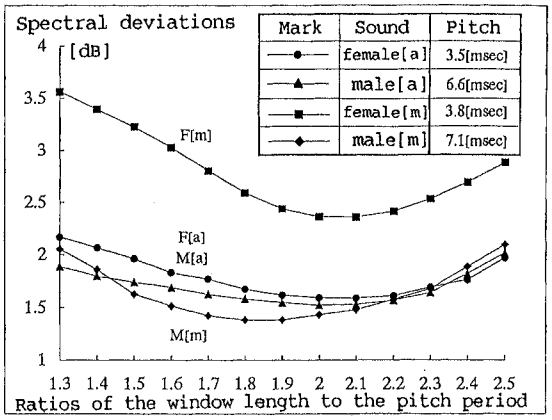


Fig.3 Objective evaluation on the window length by using a speech signal model (in case of the Hanning window).

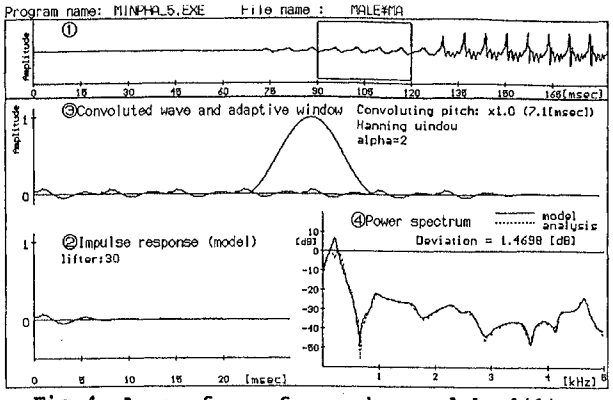


Fig.4 A waveform of speech sound [ma](1), an impulse response(2), a speech signal model and an adaptive window(3), and comparative spectra (4).

using the Hanning windows whose lengths are 2 to 3 times the pitch period. Taking account of pitch frequency control, narrower window is better for avoiding sound quality degradation caused by the original pitch frequency. Therefore, conventional waveform extraction by using the Hanning window whose length is twice the pitch period is quite adequate.

The authors also studied on the Blackman-Harris windows, resulting that the windows whose lengths were 2.5 to 4 times the pitch period yielded similar sounds to the original sound. This result meant that both optimum windows had the identical equivalent-bandwidth. However, in case of the Blackman-Harris window, slightly dark voices were heard in comparison with the Hannimg window. This meant that waveform deformation by windowing caused slight degradation in speech sound quality. Therefore, there was possibility to be obtained more excellent speech sound by using a flat top window in time domain.

4.2 Verification of the optimum window length

The authors verified the optimum window lengths objectively by using a speech signal model. For the first, a minimum phase impulse response was derived from an extracted pitch waveform by means of the FFT operation. And, the ARMA system with 20 poles and 20 zeros was identified from the minimum phase impulse response in order to obtain a linear signal model. For the second, the impulse train with the same period as the original speech was put in to the ARMA system, yielding the linear signal model. Finally, the spectrum of the pitch waveform extracted from the speech signal model was compared with the frequency responce of the ARMA system.

Single sounds [m] and [a] in single syllabic sounds [ma] of female and male voices were used for identification of the ARMA system . Spectral deviation was calculated for each of the Hanning windows whose length were 1.3 to 2.5 times the pitch period. Results are shown in fig.3. Consequently, it is confirmed that the window length of twice the pitch period is approximately the optimum value which minimizes the maximum spectral deviation down to 2.5 dB as for the Hanning window. Figure 4 shows a waveform of sound [ma], a speech signal model, an impulse response of the ARMA system and comparative spectra in case of male voice. In the same manner as for the Blackman-Harris window, it is confirmed that the window length of 2.9 times the pitch period is approximately the optimum value.

V. EXAMPLES

5.1 In case that the pitch pattern unchanges

In case that the change of pitch pattern is not necessary at editing a new non-entry word, concatenation processing is performed first, and if necessary, duration modification is excuted.

Example 1: Concatenation by vowels [a].

n a g o y a y u k i g a
h i r a t s u k a h o o m e n
↓
n a g o y a h o o m e n

In this case, the pitch pattern of generated word was not necessary to be changed. And, neither phonetic degradation nor joint distortion were perceived.

Example 2: Concatenation by syllables [zi].

t s u z i d o o
f u z i s a w a
↓
t s u z i s a w a

In this case, both speech sounds were joined at the vowel portion where the spectral distance became minimum. However, the duration of vowel sound [i] became longer, then the duration was manually adjusted to yield rhythmically natural sounding speech.

5.2 In case that the pitch pattern changes

In case that the pitch pattern must be changed at editing, prosody modification is performed on voiced portions by using the zero-phased pitch waveform superposing method, while unvoiced portions are retained as they are.

Example 3: Concatenation by syllables [ka] and [ta].

s h i z u o k a
y a m a k i t a y u k i n o
↓
s h i z u o k a y u k i n o

In this case, both words are connected at the phoneme boundaries, namely voiceless intervals between stop consonants and succeeding vowels. And, the pitch pattern must be altered.

Quality evaluation as for the example 3 was performed on four testing speech sounds synthesized under the following conditions:

(1) Concatenation processing only (no prosody modification).
(2) Modifing the pitch pattern by using that of natural speech uttered "*shizuokayukino*".
(3) Adjusting the duration by 50ms shorter between [o] and [k] sounds.
(4) Modifing the pitch pattern by using that of "*yamakitayukino*" which has the same pattern as "*shizuokayukino*", as well as adjusting the duration.

Four pairs of a natural speech and a synthesized speech were recorded in the audio tape in random order. Ten male subjects heard the recorded speech through headphones for three times, and judged accentual naturalness and rhythmical naturalness on the 7-point rating scale for each time. Results are shown in fig.5.

Figure 5 shows that generation of a new word with high quality is possible by applying prosody modification. Further, it is suggested that there exist mutual relation between the pitch accent and the rhythm. Therefore, balanced control is important to yield natural sounding speech. Our final purpose is to generate new non-entry words under the 4th condition. However, speech sound synthesized under the 4th condition is evaluated slightly lower than that synthesized under the 3rd condition. This suggests that further study on prosody modification is required to make the syntesized speech more natural.

VI. CONCLUSIONS

A method to generate non-entry words from the entries of the natural speech database of the automatic public announcing system is discussed.

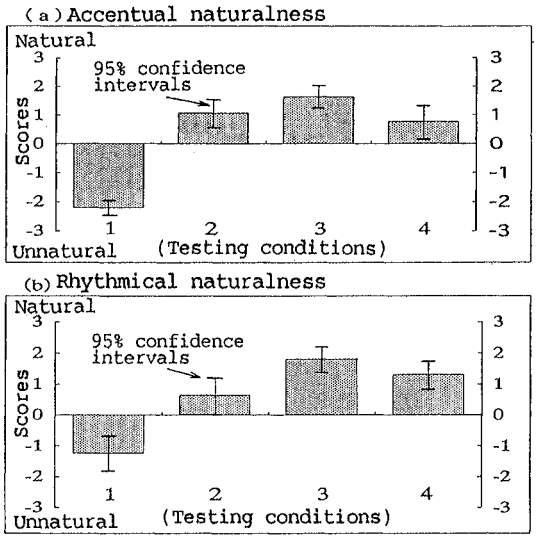


Fig.5 Subjective evaluation of naturalness of the phonetic words generated under the four conditions.

By this method, it has become possible to add new words to the database without additional recording by a narrator.

As for the window for extracting pitch waveforms, the Hanning window is slightly better than the Blackman-Harris window. However, it has been suggested that a flat top window in time domain has possibility to produce much better rsults.

In case that there is no need to change the pitch pattern, new words with exellent quality are generated. However, in case that pitch pattern must be changed, naturalness is slightly debased. In the latter case, further improvement in prosody modification is desired.

Recently, an interactive sysytem for editing new words has been constructed. Making the editing process automatic, and making the prosody modification advanced are problems to be solved hereafter.

REFERENCES

[1]Y.Sagisaka, "Speech Synthesis by Rule using an Optimal Selection of Non-uniform Synthesis Units," Proc. IEEE ICASP 88, S14.8, pp.679-682, April, 1988.

[2]K.Abe, K.Takeda and Y.Sagisaka, "On the Concatenation of Speech Synthesis Units According to Unit Extraction Context," IEICE Technical Report, SP89-66, pp.17-22, Nov., 1989(in Japanese).

[3]N.Morita and F.Itakura, "Time-Scale Modification Algorithm for Speech by Use of Pointer Interval Control OverLap and Add(PICOLA) and Its Evaluation," ASJ Autumn Meeting, 1-4-14, pp.149-150,Oct., 1986(in Japanese).

[4]F.Charpentier and E.Moulines, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones," Proceedings Eurospeech'89, pp.13-19, 1989.

[5]T.Nakajima and T.Suzuki, "Pitch Pair Synchronous PSE Analysis Based on a Non-steady State Wave Spectral Model," JASJ, Vol.44, No.12, pp900-908, 1988(in Japanese).