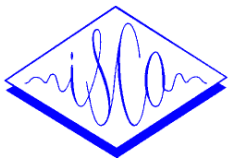


First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

PROPOSAL AND EVALUATION OF A NEW TYPE OF TERMINAL ANALOG SPEECH SYNTHESIZER

Hiroya Fujisaki, Keikichi Hirose and Yasuharu Asano

Dept. of Electronic Engineering, Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT

Because of simplification in the realization of both the vocal tract transfer functions and the excitation sources, certain limitations exist in the quality of speech synthesized by conventional terminal analog speech synthesizers. In order to realize high quality synthesis of speech, a new type of terminal analog synthesizer has been developed consisting of four paths of cascade connection of pole/zero filters and three types of source waveform generators. The four separate paths simulate the vocal tract transfer functions of the four different speech categories: the vowels and vowel-like sounds, the nasal murmur and the buzz bar, the frication and the plosion. Configuration of each path has been decided based on the results of analysis of natural utterances. The generators produce voicing source waveforms, white Gaussian noise waveforms, and impulse-like waveforms. The quality of synthesized speech, especially stop consonants, has indicated the advantage of the proposed synthesizer over the conventional ones.

1. INTRODUCTION

In spite of many significant contributions made to the technology of speech synthesis, the quality of speech for the conventional text-to-speech conversion systems still needs to be improved. Various factors can be listed up which are responsible for rather poor quality in the prosodic and segmental features of synthetic speech. As for the segmental features, speech synthesizers used in the systems have oversimplified configurations based on rather rude approximation of the process of speech production, and hence fail to produce close approximations to the actual speech.

Since the terminal analog method simulates the acoustic process of speech production, it is considered to be one of the best approach to achieve a high quality of synthesized speech in the text-to-speech conversion systems. Moreover it requires only a small sized memory for the storage of control parameters to be realized using computers. Therefore, several versions of terminal analog speech synthesizers have been proposed and constructed. Certain limitations, however, have been existing in the quality of speech synthesized with these synthesizers, because of the simplification and approximation of the vocal tract transfer functions and of the source characteristics. For the purpose of overcoming the shortcomings of conventional speech synthesizers, we have developed a terminal analog speech synthesizer with a new configuration.

In this paper, after a brief survey of the conventional terminal analog speech synthesizers, we propose a new one consisting of four paths of cascade connection of pole/zero filters and of three source waveform generators. Its detailed configuration is then explained and, finally, its validity is shown by the evaluation of the synthesized speech.

2. OVERVIEW OF CONVENTIONAL TERMINAL ANALOG SYNTHESIZERS AND THEIR SHORTCOMINGS

In terminal analog speech synthesizers, the transfer functions of vocal tract can be approximated either by cascade connection of pole/zero filters[1] or parallel connection of pole filters[2]. Although the cascade connection has advantages over the parallel connection in that it corresponds directly to the actual vocal tract transfer function and can be controlled without individual gain control, it requires accurate data of pole and zero frequencies, bandwidths, and other parameters to obtain a high quality speech. In the absence of such accurate data, the parallel connection can provide a practical means for spectral approximation. As a compromise, a hybrid configuration, which possesses cascade path for synthesizing vowels and vowel-like sounds, and parallel path for voiceless consonants, has been widely used[3][4]. Most of the existing synthesizers, however, share the following shortcomings which seem to restrict the quality of synthetic speech: 1) oversimplification for the vocal tract transfer functions of certain classes of speech sounds such as the nasalized vowels, the nasal murmur, and the buzz bar, and 2) oversimplification for (or omission of) the excitation sources such as the glottal source and the source for the plosion.

3. CONFIGURATION OF PROPOSED SYNTHESIZER

3.1 Overview

Based on the considerations in the preview chapter, a new type of terminal analog synthesizer has been developed for the synthesis of high quality speech[5]. As for the vocal tract transfer func-

tions, it possesses four separate paths of cascade connection of pole and zero filters. Vowel, nasal, fricative and stop paths are respectively for the synthesis for vowel and vowel-like sound, the nasal murmur and the buzz bar, the frication and the plosion. As for the excitation sources, it possesses three waveform generators: 1) a glottal waveform generator with up to six control parameters, 2) a white Gaussian noise generator for the turbulent noise source, and 3) an impulse generator for the source of plosion. In order to produce the sounds of aspiration, the turbulent noise source is supplied also to the vowel path.

3.2 Excitation Sources

Most of the conventional terminal analog synthesizers have only two types of excitation sources, one for glottal waveform and another for the turbulent noise. The former is used to synthesize voiced sounds, and the later is used to synthesize voiceless consonants and aspirations. When pronouncing the plosion, however, the excitation source is generated by the abrupt release of the vocal tract closure and its acoustic features are quite different from those of the turbulent noise source. Therefore, if the noise source is used to synthesize the plosion, high quality synthesis can not be achieved. We have introduced an impulse waveform generator as the source for the synthesis of plosion.

On the other hand, most of the conventional synthesizers adopted the pulse train model as the glottal source waveform. But, the actual waveform of the glottal vibration is much more complex. Because of this simplification, the quality of the synthesized speech is limited. So, we adopted a polinominal model which can yield a better approximation to the actual waveform[6]. The model generated waveform is illustrated in Fig. 1 together with the formula.

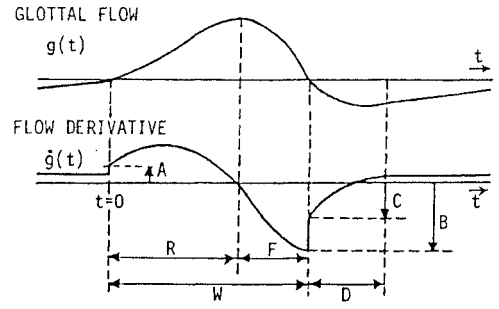
As for the noise source for synthesizing frication and aspiration, the random noise generator is used.

Since the radiation characteristics can be approximated by the differentiation of time, the input sources to the filter paths are differentiated beforehand.

3.3 Paths for Representing Vocal Tract Transfer Function

In order to decide the configuration of pole/zero filters for each path, we need rough data for the transfer function which each path should represent. As for vowels, the transfer functions can be represented by several poles and their characteristics have already been clarified. As for the consonants, however, most of the transfer functions are represented by not only poles but also zeros, and besides, they usually have rapidly varying features. Therefore, the analysis of these consonants is much more difficult than that of vowels. The following shows the results of preliminary analyses of natural speech conducted to decide the configuration of each cascade path.

In constructing the synthesizer, we fixed the sampling frequency at 10kHz, and thus only took poles and zeros under 5kHz into account. As for the speech samples, Japanese CV and VCV syllables



GLOTTAL PARAMETERS

W - PULSE WIDTH $(R+F)$ — A - SLOPE AT GLOTTAL OPENING
S - PULSE SKEW $(R+F)/(R-F)$ — B - SLOPE PRIOR TO CLOSURE
D - GLOTTAL CLOSURE TIMEING — C - SLOPE FOLLOWING CLOSURE

$$g(t) = \begin{cases} A - \frac{2A+R\alpha}{R}t + \frac{A+R\alpha}{R}t^2, & 0 < t \le R, \\ \alpha(t-R) + \frac{3B-2F\alpha}{F^2}(t-R)^2 - \frac{2B-F\alpha}{F^3}(t-R)^3, & R < t \le W, \\ C - \frac{2(C-\beta)}{D}(t-W) + \frac{C-\beta}{D^2}(t-W)^2, & W < t \le W+D, \\ \beta, & W+D < t \le T, \end{cases}$$

where $\alpha = \frac{4AR-6FB}{F^2-2R^2}$ and $\beta = \frac{CD}{D-3(T-W)}$,
T = fundamental period.

Fig.1. Waveform and formula of the glottal waveform model adopted for the synthesizer.

pronounced by one Japanese male speaker were used. And these samples were low-pass filtered at 4.5kHz, and were digitized at a rate of 10kHz and 12 quantum bits.

3.3.1 Configuration of Vowel Path

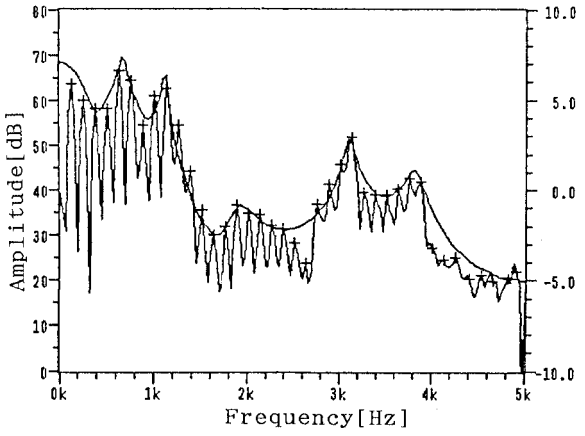
It is well known that the transfer functions (below 5kHz) of the vowel sounds can be represented by five poles (formants) in almost all the cases. The back vowels preceded by nasal consonants, however, have tendencies to be nasalized. In order to simulate this phenomenon, extra pole-zero pairs are necessary. We analyzed the utterances of Japanese back vowels /a/ preceded by nasal consonants /m/, /n/, / / in order to obtain the detailed features of the nasalization. Figure 2 shows the spectrum together with the pole/zero frequencies and bandwidths extracted by the method of analysis-by-synthesis (henceforth A-b-S method) for the nasalized vowel of /ma/. This figure indicates the presence of two pole-zero pairs below first formant and between second and third formants.

The result indicates that five poles and two pole-zero pairs are necessary for vowels and vowel-like sounds. When synthesizing vowels without nasalization, the effect of pole-zero pairs can be canceled by controlling the pairing pole and zero to have the same frequency and bandwidth.

Since the spectral envelopes of aspirations have structures similar to those of vowels, we can synthesize the aspirations using the vowel path by driving with the noise source instead of the glottal source.

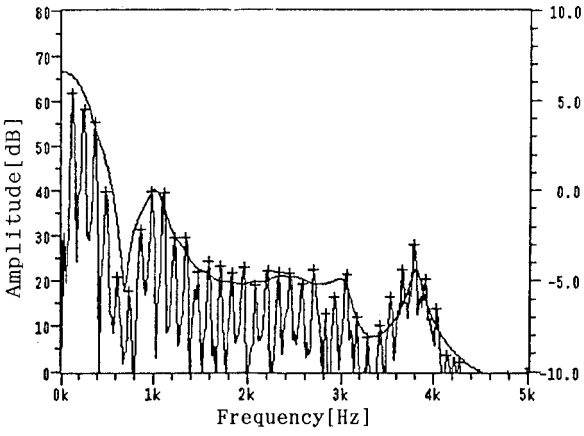
3.3.2 Configuration of Nasal Path

When pronouncing the nasal murmur, the vocal tract has a nasal branch and its transfer function is known to have zeros. In order to decide the



FP1	666Hz	BP1	84Hz
FP2	1116Hz	BP2	50Hz
FP3	1899Hz	BP3	275Hz
FP4	3100Hz	BP4	109Hz
FP5	3816Hz	BP5	184Hz
FZ1	500Hz	BZ1	1275Hz
FZ2	1683Hz	BZ2	359Hz

Fig.2. Spectrum of the nasalized vowel of /ma/ together with into pole/zero frequencies and bandwidths extracted by A-b-S method.



FP1	383Hz	BP1	267Hz
FP2	1000Hz	BP2	184Hz
FP3	2431Hz	BP3	600Hz
FP4	3016Hz	BP4	209Hz
FP5	3799Hz	BP5	117Hz
FZ1	683Hz	BZ1	42Hz
FZ2	3216Hz	BZ2	400Hz

Fig.3. Spectrum of the nasal murmur of /ma/ together with into pole/zero frequencies and bandwidths extracted by A-b-S method.

number of zeros necessary for the path, three types of nasal consonants of Japanese /m/, /n/, /ŋ/ were analyzed. The spectrum of nasal murmur of /ma/ is shown in Fig. 3 together with its pole/zero frequencies and bandwidths obtained by A-b-S method. According to these results, two is enough for number of zeros to represent the transfer functions of nasal consonants. Based on the analysis, two pole-zero pairs and three poles were prepared for the nasal path.

3.3.3 Configuration of Fricative Path

As for the voiceless fricative consonants /s/, /ʃ/, detailed analyses have already been conducted. According to reference [7], a good approximation for the transfer function of fricative consonants can be obtained by two poles and one zero. The analysis of the fricative part of voiceless stops, however, indicated the necessity of one additional zero. Therefore, the fricative path is constructed by three poles and two zeros.

According to reference [8] on the fricative consonant /h/, the spectra of /h/ followed by /a/, /e/, /o/ are very similar to those of vowels. Thus, the sound of /h/ is synthesized by the vowel path driving with the noise source.

3.3.4 Configuration of Stop Path

A voiceless stop consonant generally consists of burst, frication, aspiration and vowel transition. Analysis of high time-resolution, say below one pitch period, is necessary in this case.

Stop consonants /p/, /t/, /k/ followed by /a/ are analyzed and the results indicate that four poles and two zeros are necessary for synthesizing the burst.

When synthesizing the stop consonants, firstly, the stop path is driven by the impulse source, secondly, the fricative path is driven by the noise source, and lastly, the vowel path is used to synthesize aspiration and vowel transition.

3.4 Configuration of Synthesizer

Based on the above considerations, we developed a new type of terminal analog speech synthesizer shown in Fig. 4. The control parameters for the synthesizer are listed in Table 1.

The control parameters for the synthesizer vary as functions of time. The updating of the control parameters for segmental features is conducted as follows:

1) At the voiceless part, the parameters are updated with the same interval of 5ms (default value).
2) At the voiced part, the updating is conducted pitch-synchronously. Since the glottal pulse is generated periodically at the voiced part, the quick change in the control parameters leads to the instability of the circuits or the occurrence of click sounds.

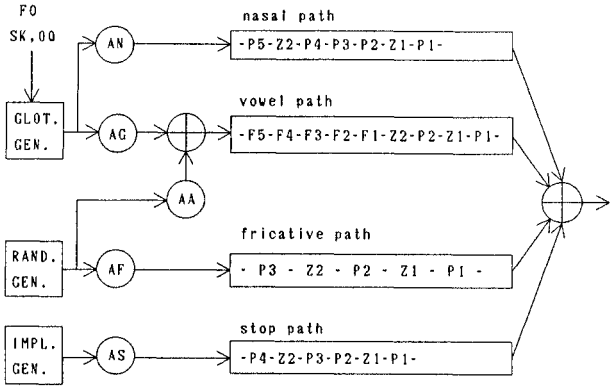
The voiced part and the voiceless part of speech are indicated by the value of AG and AN. The parts where both AG and AN are equal to 0 are voiceless, and the other parts are voiced.

4.EVALUATION OF SYNTHESIZED SPEECH

We have made the CV syllable templates of control parameters to realize the rule based speech synthesis of Japanese sentences. The phone intelligibility of the speech synthesized by the synthesizer is currently about 74% yielding a great improvement from 61% obtained by the synthesizer formerly configurated (a modified version of Klattalk[3]). The improvement is mainly due to the following two causes.

1) Detailed analysis of each syllable.

In making the control parameters for this synthesizer, we conducted a detailed analysis of



GLOT.GEN.: GLOTTAL WAVEFORM GENERATOR
RAND.GEN.: RANDOM NOISE GENERATOR
IMPL.GEN.: IMPULSE GENERATOR

Fig.4. Configuration of the proposed terminal analog speech synthesizer.

Table 1. Control parameters for the proposed synthesizer.

vowel path	F1-F5, B1-B5(formant frequency and bandwidth) P1F, P2F, Z1F, Z2F, P1B, P2B, Z1B, Z2B (pole-zero pairs for nasalized vowels) AG(gain from glottal source to vowel path) AA(gain from fricative source to vowel path)
nasal path	NP1F-NP4F, NZ1F, NZ2F(pole and zero frequency) NP1B-NP4B, NZ1B, NZ2B(pole and zero bandwidth) AN(gain from glottal source to nasal path)
fricative path	FP1F-FP3F, FZ1F, FZ2F(pole and zero frequency) FP1B-FP3B, FZ1B, FZ2B(pole and zero bandwidth) AF(gain from fricative source to fricative path)
stop path	SP1F-SP4F, SZ1F, SZ2F(pole and zero frequency) SP1B-SP4B, SZ1B, SZ2B(pole and zero bandwidth) AS(gain from impulse source to stop path)
glottal control	F0(fundamental frequency) SK, QQ(skew and open quatient of glottal source)

natural utterance using various techniques including pole-zero analysis by the A-b-S method. This cause is responsible mainly for the improvement in vowel-like sounds which are synthesized using the vowel path of the similar configuration to the previous synthesizer.

2) Configuration of the synthesizer.

This cause consists of two sub causes. One is the use of new filter circuit and waveform generator, viz., the stop path and the impulse generator. This cause improved the quality of stops in a great deal. Another sub cause is the change in the pole/zero configuration of the path: This leads to an easier control of the parameters. This cause is especially effective for the fricatives. This sounds are synthesized by the cascade circuit in this synthesizer, while by the circuit of parallel configuration of poles in the previous synthesizer.

Figure 5 shows the waveforms of (a) a natural utterance and (b) synthetic speech for /ka/. When synthesizing /ka/, stop, fricative and vowel paths are necessary. The result indicates the validity of the proposed synthesizer.

5.CONCLUSION

In this paper, based on the considerations on the shortcomings of the conventional terminal analog speech synthesizers, a new configuration was proposed. The proposed synthesizer has four separate paths, which are respectively used to synthesize vowels and vowel-like sounds, the nasal murmurs and the buzz bars, the sounds of frication and the sounds of plosion. It also has three excitation source generators for glottal waveform, noise source and impulse source.

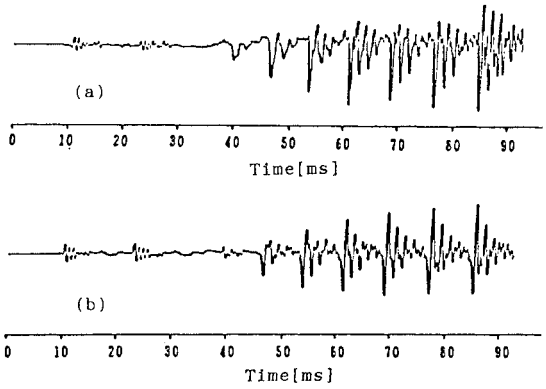


Fig.5. Waveforms of (a) natural utterance and (b) synthetic speech for /ka/

The quality of speech produced by this synthesizer is very high, especially that of stop consonants. This is due to the inclusion of the stop path and the impulse source to the synthesizer. This result indicated the advantage of this proposed synthesizer over the conventional ones, and convinced us of the validity of this synthesizer.

This work was supported by a Grand-in-Aid for Scientific Research (No. 01608003) from the Ministry of Education.

REFERENCES

[1]G.Fant, "Acoustic analysis and synthesis of speech with applications to Swedish," Ericsson Technics, No.1, 3-108 (1959).
[2]W.Lawrence, "The Synthesis of Speech from Signal which have a Low Information Rate," in Communication Theory, edited by W. Jachson, Butterworths, London, England, pp.460-469 (1953).
[3]D.H.Klatt, "Software for Cascade/Parallel Formant Synthesizer," J. Acoust. Soc. Am., 67 3, pp.971-995 (1980).
[4]D.H.Klatt and L.C.Klatt, "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talker," J. Acoust. Soc. Am., 87 2, pp.820-857 (1990).
[5]H.Fujisaki, K.Hirose and Y.Asano, "Terminal Analog Speech Synthesizer for High Quality Speech Synthesis," IEICE Technical Report, SP90-1 (1990).
[6]H.Fujisaki and M.Ljungqvist, "Proposal and Evaluation of Models for the Glottal Source Waveform," Proc. ICASSP, 31.2, pp.1605-1608 (1986).
[7]H.Fujisaki, O.Kunisaki and S.Shibayama, "Analysis, Synthesis and Perception of Voiceless Fricative Consonants in Japanese," Trans. Committee on Speech Research, Acoust. Soc. Jpn., S75-06 (1975).
[8]O.Kunisaki, T.Suehiro and H.Fujisaki, "Acoustical Feature of Voiceless Fricative Consonants /h/," Trans. Committee on Speech Research, Acoust. Soc. Jpn., EA77-47 (1978).