Ann.Bull.RILP(1988)No.22, 153-164

Advanced Techniques for Speech Synthesis
-A Research Plan on Techniques for Synthesizing Segmental Features-

Shigeru Kiritani and Satoshi Imaizumi

A nation-wide speech research project, "Advanced Man-Machine Interface Through Spoken Language", headed by Prof. H. Fujisaki, University of Tokyo, was started in Japan in 1987. It is supported by Grant-in-Aid for Scientific Research on Priority Areas, the Ministry of Education, Science and Culture, Japan. As one of the core groups of this project, a research group on "Advanced Techniques for Speech Synthesis" was organized. The main members of the group are Profs. H. Fujisaki and K. Hirose (Faculty of Engineering, University of Tokyo) and the present authors. The aim of the group is to develop new synthesis techniques both for segmental features and prosodic features, and the present authors are mainly concerned with the former problem. This report presents part of the research plan of this group related mainly to the synthesis of segmental features.

1. Introduction

Speech is the most fundamental means of human communication. It is widely admitted that communication through speech can provide the most efficient and flexible means for the human-machine interface. However, in spite of remarkable progress in recent years, the quality of synthetic speech still has to be improved in several respects to realize an "Advanced Man-Machine Interface through Spoken Language". The major limitations of the current technique of speech synthesis by rule can be summarized as follows.

1) The synthesis of the acoustic correspondences of high-level linguistic information, such as syntactic and discourse structures, is incomplete.
2) The synthesis of variations in voice quality and segmental features related to utterance conditions is also incomplete.

The first problem is mainly related to the prosodic features of speech. In this connection, it has been demonstrated that the the F_0 contour generation model of Fujisaki et. al.[1,2] can generate close approximations to observed F_0 contours. However, in order to generate natural prosodic patterns for various types of sentences, further studies are needed to establish a means of automatically deriving necessary linguistic information and to establish a quantitative model of the correspondence between linguistic information and prosodic patterns covering a wide variety of sentences.

As for the second problem, there are two important factors in improving the naturalness of synthetic speech quality. One is the more precise control of variations in voice source

characteristics over different utterance conditions[3]. The other is the generation of natural patterns of temporal trajectories for formants at a fast, natural conversational speaking rate[4,5].

Based on the above considerations, the followings were selected as the immediate objectives of our core project "Advanced Techniques for Speech Synthesis".

1) Development of an effective model of the voice source signal to generate various voice qualities.
2) Development of rules for synthesizing natural patterns of formant trajectories with special reference to fast conversational speech.
3) Development of rules for synthesizing natural prosodic patterns of speech based on the syntactic, semantic and discourse information of texts.
4) Development of an effective method of text analysis to derive the linguistic information necessary for speech synthesis.

2. Overview of Problems in Speech Synthesis by Rule

Speech synthesis by rule can be divided into two major categories. One is speech synthesis from a concept, the other is speech synthesis from a text. In both cases, output speech is synthesized from underlying linguistic information. Linguistic information for speech synthesis consists of information on lexical items and information on the syntactic, semantic and discourse structures of sentences. In speech synthesis from a concept, this information is generated from the structural representation of the concept to be transmitted by speech. In speech synthesis from a text, this information has to be derived through a linguistic analysis of the input text. This means that the process of speech synthesis from a text requires an understanding of the given text.

Figure 1 shows a block diagram of the text-to-speech conversion system. The system can be regarded as consisting of three parts, viz., the linguistic processor, the phonetic processor and the acoustic processor. The linguistic processor performs morphemic, syntactic, semantic and discourse analyses of the text. The function of the linguistic processor is to derive the linguistic information necessary for subsequent stages of the speech synthesis.

Based on the output from the linguistic processor, the phonetic processor generates a sequence of sound symbols and prosodic symbols. Using information from a dictionary, the string of lexical items is converted into a sequence of sound symbols. At the same time, several symbols necessary for generating the prosodic pattern of speech are derived from the linguistic information.

At present, the F_0 contour generation model developed by

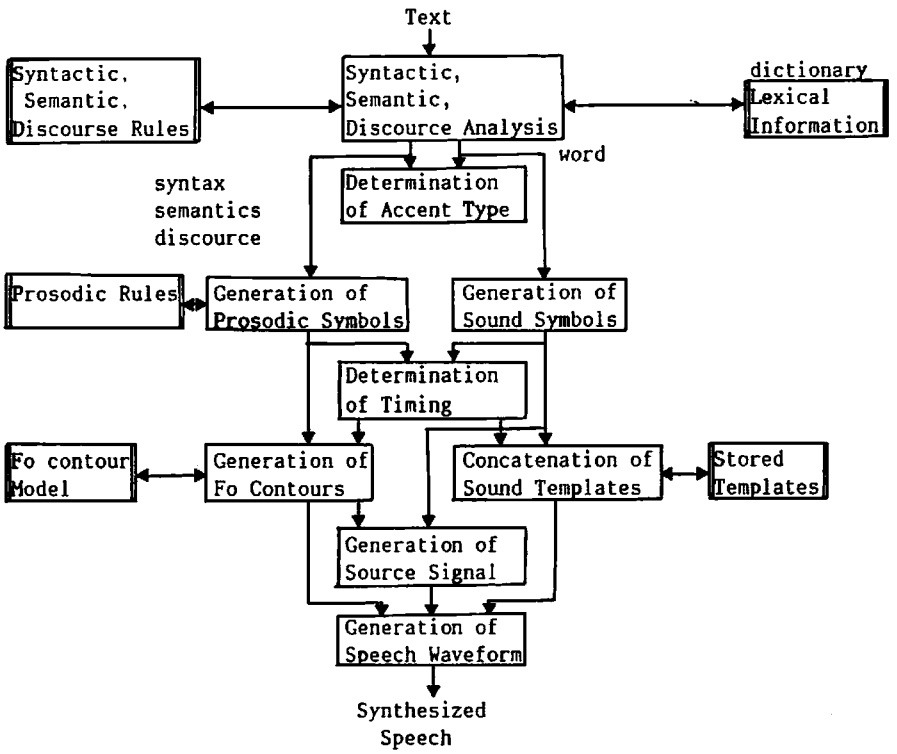

Fig. 1. A system for text-to-speech conversion.

Fujisaki et. al.[1,2] appears to be the most effective model for generating natural F_0 contours from a small set of parameters. It is based on a quantitative formulation of the process whereby the logarithmic fundamental frequency is controlled in proportion to the sum of two components corresponding to the effects of phrase and accent. In this model, the symbols for controlling the prosodic pattern of speech are pause, phrase and accent. The pause symbols represent the duration of pauses. On the other hand, the phrase symbols and the accent symbols, respectively, represent the magnitude of the phrase commands and the amplitude of the accent commands in the F_0 contour model.

As for the segmental features of the speech signal, the acoustic processor in the system produces a continuous speech signal by concatenating stored templates of the spectrum patterns of the basic units (formant patterns of CV syllables in the present system) and generating appropriate source signals. The F_0 contours of sentences are produced using prosodic symbols. In recent years, progress has been achieved in these techniques which has contributed to improvements in the quality of synthesized speech. However, there still remain several problems

to be solved in order to to obtain a fully natural sound quality.

At present, the synthesis system is not flexible enough to produce the various types of voice quality occurring in natural speech. An effective model has to be developed to control natural variations in voice source characteristics which depend on differences among speakers (including male-female differences), various modes of phonation and various prosodic patterns. It is also important to realize a more detailed control of the influence of consonant gestures upon voice source signals.

As for the concatenation rules for the spectrum pattern, fairly natural results can now be obtained for slow or normal speaking rates. However, for a fast, natural conversational speaking rate in which the duration of vowels tends to be very short, there still exists difficulty in generating natural spectrum patterns. More flexible concatenation rules should be developed which include a more precise formulation of the effects of speaking rate.

3. Studies on Voice Source Characteristics

3.1. Aims

In recent years, several articulatory speech synthesizers[5,6,7] have been developed which incorporate an interactive model of the glottal source and generate natural voice quality. However, these synthesizers have difficulties in generating the control parameters representing the articulatory dynamics and glottal adjustments and also require too much in computation time. On the other hand, formant synthesizers have difficulty in including a good model of the glottal source and fail to generate natural voice quality. From a practical point of view, therefore, it is crucial to construct a functional model of the voice source in the time domain for a formant synthesizer.

Our purpose here is to contruct a functional model of the voice source for a formant synthesizer by which we can control voice source characteristics in various utterance conditions with special respect to the following three items.

1) Various voice qualities for different speakers.
2) Effects of intonation and emphasis.
3) Interaction with the vocal tract.
4) Interactions with consonantal gestures.

The main problems here are, we believe, the modeling of variations in the voice source waveforms including waveform fluctuations or noise, and the modeling of the interaction between the voice source and the vocal tract load under the various conditions mentioned above. Our plans for developing these models are the following.

1) Voice source characteristics will be measured under the

various conditions mentioned above via an inverse filtering of speech waveforms and volume-velocity waveforms measured at the mouth.

2) The articulatory model of speech production shown in Fig. 2 will be used in the simulation of voice source characteristics. In this model, the glottal area function and the vocal tract area function are specified as the input. The input impedance and the transfer function of the vocal tract is calculated from the F-matrix in the frequency domain. The interaction between the glottal source and the vocal tract load is simulated in a way similar to that suggested by Sondhi and Schroeter[5].

3) Comparing the results of actual measurements and of the simulation, the principal factors in the voice source affecting synthetic voice quality will be extracted and implemented in a functional model adapted to the formant synthesizer.

3.2. Preliminary Studies

3.2.1. Speech Materials and Subjects

The following speech materials have been recorded and analyzed.

1) Sustained vowels and vowel sequences.
   /a/ /i/ /u/ /e/ /o/ /aiueo/ /uoaei/
2) Three sentences consisting of only vowels and semi-vowels.
   /aoi ei o ou/. (Somebody drives a blue ray away.)
   /yayoi wa ayu o ou/. (Yayoi follows a sweetfish.)
   /iwao wa yayu o yuu/. (Iwao banters.)
3) Twelve sentences including /b,d,g/. For example:
   /korewa abiba desu/. (This is "abiba".)
   /korewa agiga desu/. (This is "agiga".)
4) Twelve sentences with various accents. For example:
   /korewa umi desu/. (This is the sea.)
   /korewa umi desu/. (This is the discharge.)

The sustained vowels have been uttered at three loudness levels--soft, normal, and loud--each for three pitch levels, low, normal, and high. The other materials were pronounced at a comfortable level by each subject.

The subjects have been five males and five females.

3.2.2. Procedure

For half of the tokens, the EGG (Electro-glottogram) signal and speech waveform have been are recorded simultaneously. For the other half, the volume velocity waveform at the mouth has been measured using Rothenberg's mask simultaneously with the EGG signal.

The inverse filtering is being carried out based mainly on the short term LPC analysis. For the short term analysis, the EGG signal is used to estimate the closed and open periods of the glottis. The volume velocity waveform at the mouth is used to

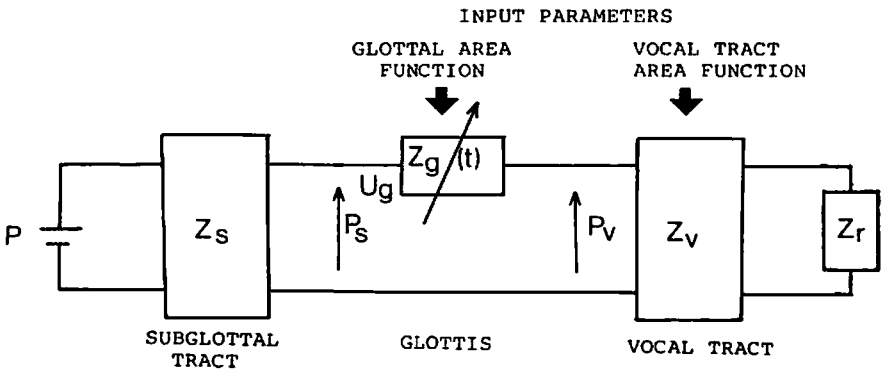

Fig. 2. An articulatory model of speech production.

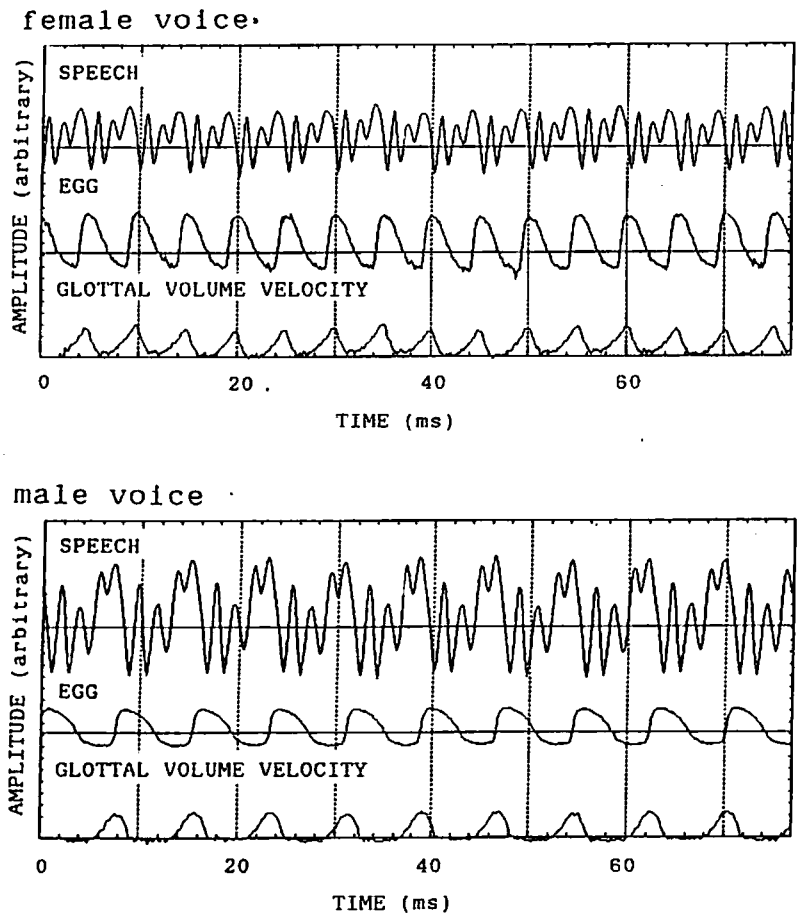

Fig. 3. Waveforms for the sustained vowel /a/, the EGG signal and the glottal volume velocity for a male and female subject.

estimate the glottal source, including the DC component.

For some utterances, the tongue configuration in the mid-sagittal and frontal planes during utterances is observed using the ultrasonic technique reported by Niimi and et al.[8]. The tongue configuration and the EGG signal will be used as reference data to determine the input signals for the articulatory model, that is, the vocal tract area function and the glottal area function.

3.2.3. Some Preliminary Results

Figure 3 shows waveforms for the sustained vowel /a/, with the EGG signal and the glottal volume velocity estimated through inverse filtering for a male subject and a female subject. Preliminary inspection shows that the glottal volume velocity waveform for the female speaker tends to have a larger DC offset, a larger amount of noise and a slower closing speed than that for the male speaker. These results coincide with the experimental findings obtained from a physical model of the glottis with imperfect closure[9,10]. Imperfect closure of the glottis and the resulting noise and slower closing speed might be important factors in female voice. Data collection and analyses along these lines are now being performed to develop voice source models for various types of utterances.

4. Studies on Formant Trajectories

4.1. Aims

In order to synthesize natural speech quality, rules for generating the temporal patterns of the formant trajectories under various speaking rates must be constructed. There are still numerous differences among the conclusions of studies on the effects of speaking rate[4,11-19]. Some papers[11,12,19] claim that increased rates of speech result in systematic deviation in the obtained formant values from their putative targets, that is, "vowel reduction". Others[13,14] insist that such "vowel reduction" does not always occur at fast speaking rates. Still others[17,18] claim that adjustments in speaking rate are achieved by strategies which differ among speakers.

The differences mentioned above might be due to differences in the speech materials and speaking rates being analyzed. For instance, Lindblom[11] used eight Swedish vowels in the contexts /b-b/, /d-d/ and /g-g/ which were embedded in four sentences with different stress patterns. Lindblom found that the degree of vowel reduction could be represented as a function of rate. He used only voiced stop consonants as contexts. The duration of /a/, for instance, varied between about 100 and 250 ms with the speaking rate. A shortening of about 60% in vowel segment duration was observed. This was also valid for other vowels.

On the other hand, Gay[14] used nine American vowels in a /pVp/ context, /i,a,u/ in a /bVp/ context, and /i,a/ in a

/kVlpV2p/ context with two stress patterns. Gay stated that the midpoint frequencies of the vowels did not vary as a function of rate. He mainly used the unvoiced bilabial stop consonant /p/. The average duration of /a/, for instance, was 115 ms for the fast rate and 145 ms for the slow rate. A shortening of only 21% occurred on average. For the vowel /i/, the shortest duration was about 60 ms, and the longest was about 115 ms. A shortening of about 48% was observed. In this case, however, Gay pooled the data for /pip/ and /bip/ and almost all the data points scattered between 60 ms and 90 ms, which were quite short compared to Lindblom's data.

The vowel durations, and thus the speaking rates, are rather different for Lindblom's data and Gay's data. This might be due to differences between Ameriacan English and Swedish, or it might be due to the speaking modes of their subjects. The present study aims at more precise analyses of speech data collected under various conditions from a number of speakers and the construction of synthesis rules describing how speaking rates affect formant trajectories.

4.2. Preliminary Studies

4.2.1. Materials and Subjects

The following speech materials are being analyzed.

1) The vowels V_2 in nonsense words of the form /$V_1CV_2CV_1$/, where V_1, V_2 are either /a,i/, and C is one of /b,d,g/.

All possible combinations result in 12 nonsense words. The phoneme /d/ is uttered as /dz/ when it preceds /i/. Each of the 12 words is embedded in a carrier sentence / korewa --- desu/, which means /this is ---/, and is repeated five times in random order.

The subjects are several adult native speakers of the Tokyo dialect of Japanese who speak in styles different from each other. Three speaking rates are being studied: a slow, normal and fast rate. Each speaking rate is based on the subject's own appraisal of his basic comfortable rate.

4.2.2. Method

The speech, EGG and intra-oral pressure signals are recorded. Using the ultrasonic technique, the tongue configuration in the mid-sagittal and frontal planes during utterances is also observed.

The formant trajectories are being extracted from an LPC analysis based on the autocorrelation method at present. However, we are planning to use a short-term LPC analysis to extract the rapid transition of the formant frequencies. The EGG signal will be used to estimate the closed and open periods of the glottis, which are useful for the short term LPC analysis. The

intra-oral pressure waveform will be used to estimate the closure for the stop consonants. Data on the tongue shape will be used to interpret the variations in formant trajectories due to the speaking rates with reference to the articulatory movements.

To construct a numerical model of the formant trajectories, the following characteristics will be measured.

1) The onset formant frequencies.
2) The transition rates of the formant frequencies at onset.
3) The midpoint formant frequencies.

4.2.3. Some Preliminary Results

Figure 4 shows the formant trajectories extracted by the autocorrelation LPC method for the utterances /agiga/, /igagi/. /abiba/ and /ibabi/. Even during the stop consonants, poles were extracted. Therefore, these pole frequencies were connected smoothly. The thin lines indicate the five utterances at the slow speaking rate, and the thick lines those at the fast speaking rate. Each utterance was aligned with reference to the peaks of the dynamic measure of the spectral change calculated from the LPC cepstrum[20]

The following phenomena can be seen in Fig. 4.

1) Variations in the formant trajectories due to the speaking rate seem dependent on the vowels and consonants. Comparing /agiga/ and /igagi/, it was found that the midpoint formant frequencies for /a/ varial more between the two speaking rates than those for /i/. While, for /abiba/ and /ibabi/, it can be seen that the degrees of variation due to the speaking rate seem less significant. The consonant /b/ permits a rather free tongue configuration, while /g/ requires a specific tongue configuration which is more different from that of /a/ than that of /i/. Therefore, the magnitude of the vowel reduction can be interpreted as a function of the transition distance from consonant to vowel, and the degree of the freedom in the tongue articulation during the production of the consonant.

2) The transition rates seemed faster when the speaking rate was faster, except in one case. For the second and third formants for /iba/ in /abiba/, the transition rates were slower when the speaking rate was faster.

3) The onset frequencies did not seem to change much between the two speaking rates.

These data suggest that in order to generate natural formant trajectories, more detailed rules which take into account the dynamic characteristics of the production of individual sounds should be developed.

5. Summary

Problems with current techniques of speech synthesis by rule have been reviewed. With respect to the synthesis of segmental features, the importance of 1) a more detailed control of the

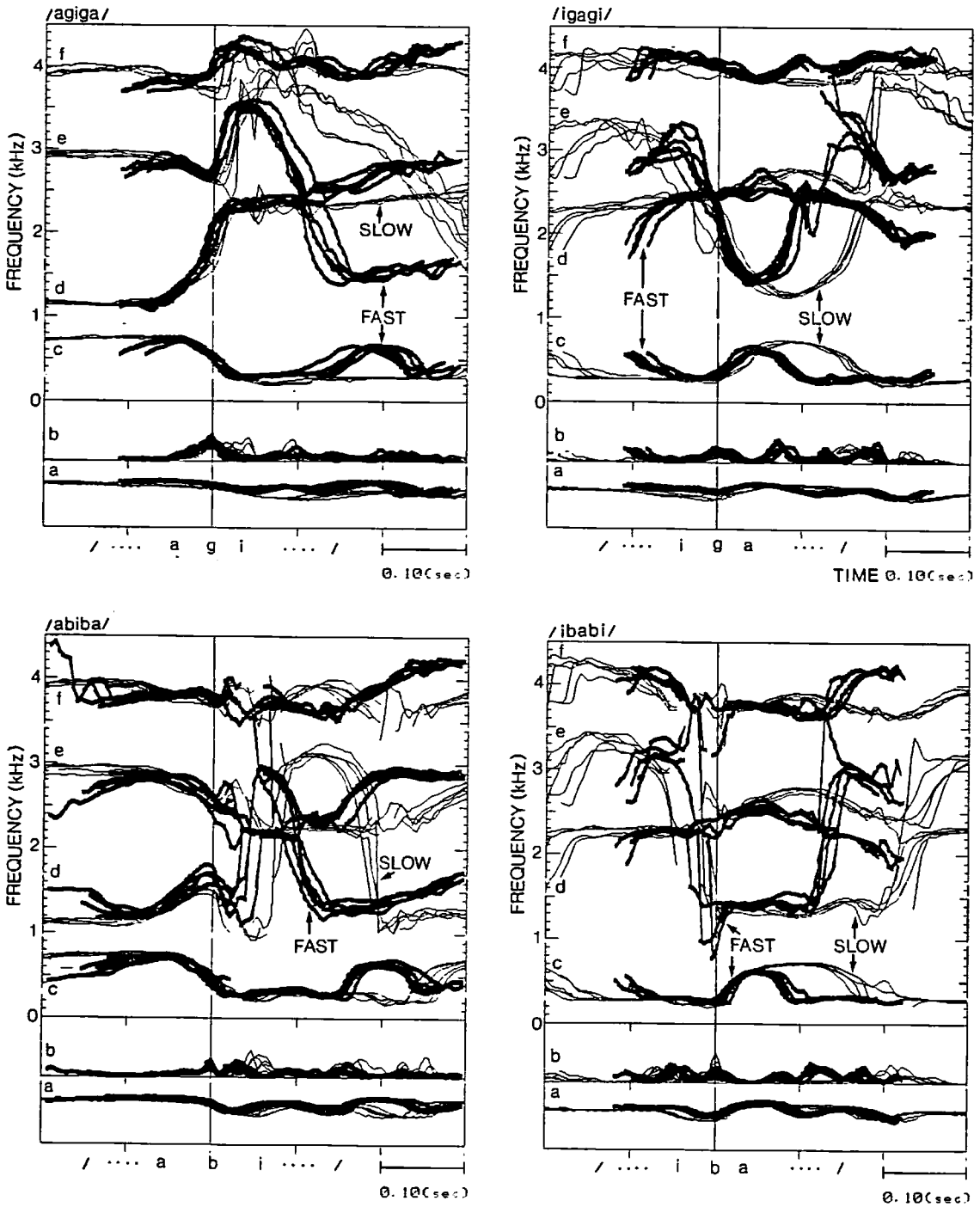

Fig. 4. Formant trajectories in five slow utterances and five fast ones for four test sentences.
a: power of speech signal. b: dynamic measure calculated from LPC cepstrum. c-f: 1st-4th fromant trajectories.

source waveform; and 2) a more precise generation of the temporal pattern of the formant change with special reference to fast, conversational speech were pointed out. Research plans were discussed on how to develop improved techniques for overcoming these problems.

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas, "Advanced Man-Machine Interface Through Spoken Language," the Ministry of Education, Science and Culture, Japan.

References

1) H. Fujisaki and K. Hirose: Analysis of voice fundamental frequency contours for declarative sentences of Japanese, J. Acoust. Soc. Jpn.(E), 5, 233-242, 1984.
2) K. Hirose, H. Fujisaki and H. Kawai: Generation of prosodic symbols for rule-synthesis of connected speech of Japanese, ICASSP-86, Tokyo, 45,4, 1986.
3) For instance, Theme Issue; Voice Acoustics and Dysphonia, J. Phonetics, 14, 3/4, 1986.
4) J. L. Miller: Effects of speaking rate on segmental distinctions. in Perspectives on the study of speech. P.D. Eimas and J.L. Miller (Eds.), Lawrence Erlbaum Associates, New Jersey, 39-74, 1981.
5) M. M. Sondhi and J. Schroeter: A hybrid time-frequency domain articulatory speech synthesizer, IEEE Trans. on ASSP, ASSP-35, 7, 955-967, 1987.
6) K. Ishizaka and J. L. Flanagan: Synthesis of voiced sounds from a two-mass model for the vocal cords, Bell Syst. Tech. J., 51,6, 1233-1268, 1972.
7) J.L. Flanagan, K. Ishizaka and L. Shipley: Synthesis of speech from a dynamic model of the vocal cords and vocal tract, Bell Syst. Tech. J., 45,3, 485-506, 1975.
8) S. Niimi, S. Kiritani and H. Hirose: Ultrasonic observation of the tongue with reference to palatal configuration. Ann. Bull. RILP, 19, 21-27, 1985.
9) S. Imaizumi and S. Kiritani: A preliminary study on the generation of pathological voice qualities. Proceedings of the Vth Conference on Vocal Fold Physiology, Tokyo, Jan., 1987.
10) S. Kiritani, H. Fukawa, H. Imagawa, S. Imaizumi and S. Saito: Measurment of air flow pattern through a mechanically driven oscillating slit - a preliminary report - . Ann. Bull. RILP, 21, 1-8, 1987.
11) B. Lindblom: Spectrographic study of vowel reduction. J. Acoust. Soc. Am., 35(11),1773-1781, 1963.
12) T. Gay: Effect of speaking rate on dipthong formant movements. J. Acoust. Soc. Am., 44, 1570-1573, 1968.
13) R. R. Rerbrugge and D. Shankweiler: Prosodic information for vowel identity. J. Acoust. Soc. Am., 61, S39, 1977.
14) T. Gay: Effect of speaking rate on vowel formant movements. J. Acoust. Soc. Am., 63(1), 223-230, 1978.
15) D. O'Shaughnessy: The effects of speaking rate on formant transitions in French synthesis-by-rule. Proc. 1986 IEEE-IECEJ-ASJ, Tokyo, 2027-2030, 1986.

16) J. Miller and T. Baer: Some effects of speaking rate on the production of /b/ and /w/. J. Acoust. Soc. Am., 73(5), 1751-1755, 1983.
17) D. P. Kuehn and K. L. Moll: A cineradiographic study of VC and CV articulatory velocities. J. Phonetics, 4, 303-320, 1976.
18) Y. Sonoda: Effect of speaking rate on articulatory dynamics and motors event. J. Phonetics, 15, 145-156, 1987.
19) S. Nakajima: Rules of vowel formant modification in accordance with speech rate. Trans. IEICE Committee on Speech Research, 17-24, 1986.
20) S. Sagayama and F. Itakura: On individuality in a dynamic measure of speech, Proc. Annual Meeting of ASJ, 589-590, June 1979.