Ann. Bull. RILP, No. 9, 119-122 (1975)

AUTOMATIC RECOGNITION OF SEMIVOWELS IN SPOKEN WORDS

Hiroya Fujisaki, Yasuo Sato*, Yoshiro Noguchi* and Takao Yamakura*

Automatic recognition of phonemes in connected speech has to take into account the variability of their acoustic characteristics due to speaker idiosyncracy and coarticulation. Based on an approximate formulation of the coarticulatory process at the acoustic level and a method for adaptation to individual differences, a scheme for reliable segmentation and recognition of connected vowels has already been established. [1,2] The present report describes an extension of the scheme for recognition of semivowels in word context. Following the description of a model of the coarticulatory process, a scheme for recognition of vowels, semivowels and their sequences based on formant targets and command duration is proposed and tested experimentally using meaningful and nonsense words uttered by four speakers.

Formulation of the Coarticulatory Process in the Formant Frequency Domain

As shown in our previous paper, [1] it is possible to approximate the process of coarticulation between successive phonemes by a hypothetical linear system, which accepts a set of target formant frequencies of each phoneme as input command and converts them into observed formant patterns. In the case of connected vowels and semivowels, in particular, the system can be approximated by a critically-damped second-order linear system. The approximation allows one to calculate the trajectory of the n th formant frequency as

$$f_n(t) = F_{n,1} + \sum_{j=2}^{m} (F_{n,j} - F_{n,j-1}) \left\{ 1 - \left\{ 1 + \alpha_{n,j}(t - \tau_j) \right\} \exp\left\{ -\alpha_{n,j}(t - \tau_j) \right\} \right\} u_{-1}(t - \tau_j), \quad n = 1, 2, 3, \qquad (1)$$

where $F_{n,j}$ is the target value of the n th formant frequency, and τ_j is the instant of onset of the command for the j th vowel. The numbering of formants is in the ascending order of frequency for back vowels, but the order should be reversed for the second and the third formants of the front vowels /i/ and /e/, based on considerations of continuity of resonance modes in the vocal tract.

Because of coupling between resonance modes of the vocal tract, however, frequencies of two formants never coincide nor even come very close to each other in real speech. The effect of such coupling can also be simulated by a model of coupled resonance circuits, and is used to modify the trajectory of Eq. (1) into a more realistic one. Using this model, Analysis-by-Synthesis of formant trajectories of connected vowels and semivowels

* Dept. of Electrical Engineering, Faculty of Engineering, Univ. of Tokyo

makes it possible to estimate the instant of command onset and the target formant frequencies of each phoneme as well as the rate of transition between successive phonemes.

Principle of a Scheme for Recognition of Connected Vowels and Semivowels

The model of the preceding section has been applied to preliminary analysis of a number of connected utterances of vowels and semivowels. The results indicated that target formant frequencies of the semivowels /j/ and /w/ were respectively quite similar to those of the vowels /i/ and /u/, and the rate of transition between a semivowel and a vowel was not significantly different from that between two vowels.[4] Consequently, reliable recognition of vowels and semivowels has to be based on utilization of the command duration, in order to separate semivowels from vowels that possess corresponding target formant frequencies.

Thus a scheme is proposed for the recognition of connected vowels, semivowels and their sequences in two stages.[5] At the first stage, the input speech is segmented by Analysis-by-Synthesis of formant trajectories into intervals that possess a set of target formant frequencies corresponding to the five Japanese vowels /a/, /i/, /u/, /e/ and /o/, under the condition that the target formant frequencies of the semivowels /j/ and /w/ are respectively identical to those of the vowels /i/ and /u/. The vowels /a/, /e/ and /o/ can be uniquely recognized at this stage, but the intervals which possess target formant frequencies of the vowels /i/ and /u/ require further processing for their ultimate recognition.

At the second stage, these intervals are classfied into the semivowels /j/, /w/, the vowels /i/, /u/ and their sequences /ij/, /uw/ on the basis of their durations and information concerning the speech rate.

Recognition Experiment on Connected Vowels and Semivowels

In order to test the validity of the proposed scheme, a recognition experiment was performed on a total of 300 utterances of both meaningful and nonsense words containing vowels, semivowels and their sequences. These speech materials were pronounced by four male speakers, sampled at 10kHz with an accuracy of 10 bits/sample. Frequencies of the first three formants were extracted pitch-synchronously by Analysis-by-Synthesis of short-time spectra,[6] and were then interpolated at every 10 msec. At the first stage of recognition, instants of command onset were extracted from the formant trajectories, and the intervals thus segmented were temporarily classified as one of the five vowels. Examples of the results of this stage are illustrated by Fig. 1, which shows extracted formant trajectories (empty circles) and their best approximations by the model (solid curves) for the voiced intervals in utterances of /saju/, /saiu/ and /saiju/, all segmented into three intervals and classified as /aiu/, but with the durations of the medial segment differing significantly from one utterance to the other. Segments corresponding to /j/, /i/ and /ij/ were then classified by using optimum linear discriminant functions in a two-dimensional plane of segment duration versus average phoneme duration. The same technique was also applied for the separation of /w/, /u/ and /uw/. For the present speech materials, percentages of correct recognition of the semivowels /j/, /w/, the vowels /i/, /u/ and their sequences /ij/, /uw/ were 100%, 96% and 90%, respectively.

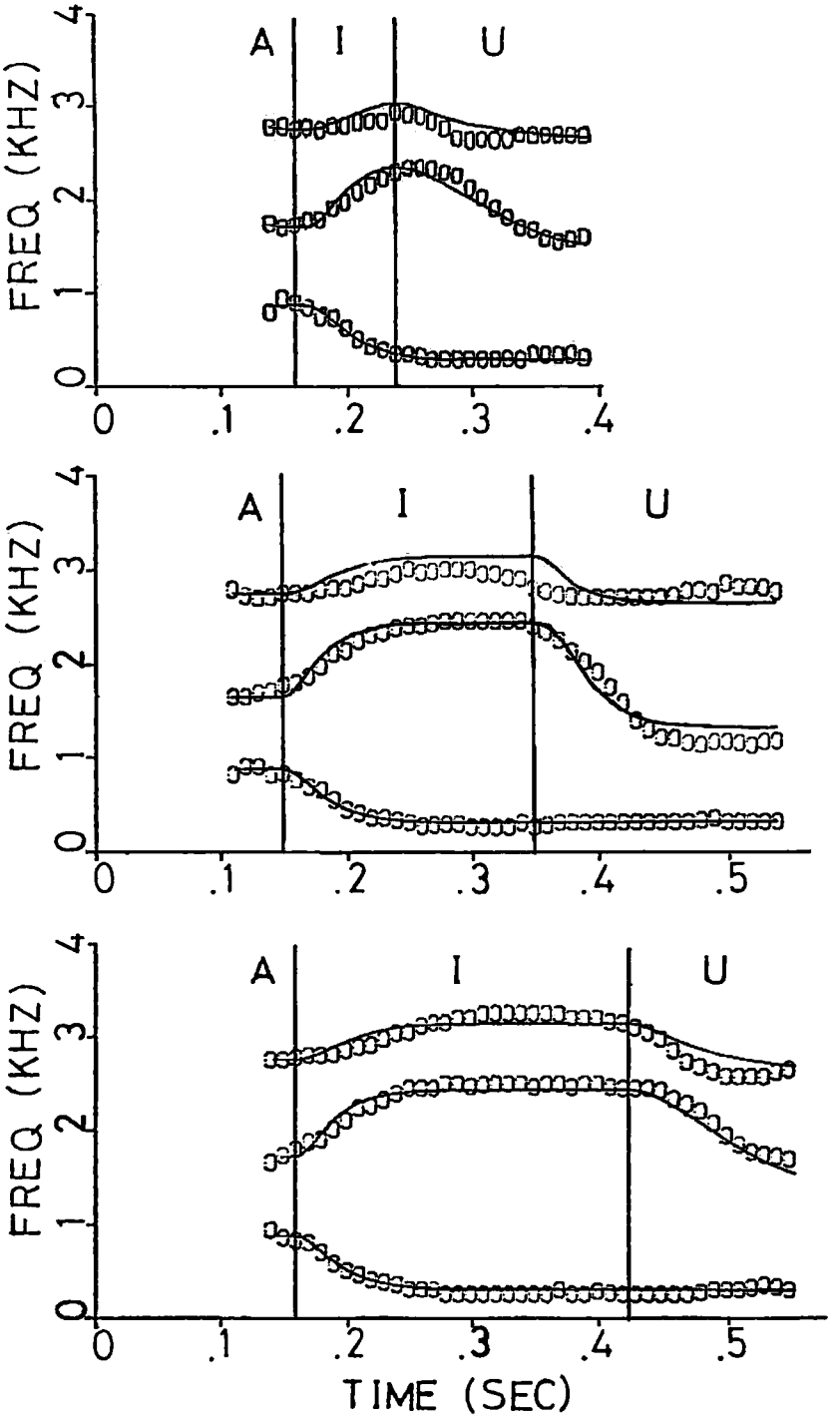

Fig. 1: Examples of segmentation and classification of /j/, /i/ and /ij/ in the same /sa-u/ context at the first stage of the proposed recognition scheme.

References

1. Fujisaki, H., M. Yoshida, Y. Sato, and Y. Tanabe (1973), "Automatic recognition of connected vowels using a functional model of the co-articulatory process," J. Acoust. Soc. Japan, 29, 636-638.

2. Fujisaki, H., M. Yoshida and Y. Sato (1974), "Segmentation and recognition of vowels in connected speech based on a model of coarticulation," 8th ICA, London.

3. Fujisaki, H. M. Yoshida and Y. Sato (1974), "Recognition of vowels and semivowels based on a model of coarticulation," Nat. Conv. REC., Inst. Elect. Comm. Eng. Japan, 1437.

4. Fujisaki, H. and Y. Sato (1974), "Recogntion of semivowels based on a model of coarticulation," Reports of Autumn Meeting, Acoust. Soc. Japan, 245-246.

5. Fujisaki, H., Y. Sato and T. Yamakura(1975), "Recognition of semivowels of a number of speakers based on a model of coarticulation," Reports of Spring Meeting, Acoust. Soc. Japan, 269-270.

6. Fujisaki, H. and K. Yoshimune (1971), "Estimation of short-time frequency spectrum of quasi-periodic waveforms with applications to pitch-synchronous analysis of speech," Report of Spring Meeting, Acoust. Soc. Japan, 21-22.