FORMULATION OF THE COARTICULATORY PROCESS IN THE FORMANT FREQUENCY DOMAIN AND ITS APPLICATION TO AUTOMATIC RECOGNITION OF CONNECTED VOWELS

H. Fujisaki, M. Yoshida** and Y. Sato**

Abstract

Speech recognition in its most general sense, where utterances of unspecified speakers have to be converted into phoneme strings, has to cope with individual differences of speakers and mutual interferences of phonemes. A novel approach is presented to overcome these difficulties in the case of connected vowels. Techniques for accurate extraction of formant frequencies and for reliable recognition of isolated vowels are combined with a model of the coarticulatory process in the formant frequency domain to perform unambiguous segmentation and recognition of successive vowels by the method of 'Analysis-by-Synthesis.' The validity of the approach is demonstrated by experiments in which all the vowels in two- and three-vowel words are successfully recognized.

1. Introduction [1]

Realization of speech recognition in its ultimate sense, whose objective is to accept, without restrictions on the vocabulary and the speaker, connected discourses of a particular dialect of a language and identify each of the linguistic units and their interrelationships, and thus extract and process the semantic information therein contained, is a problem covering a wide range of academic fields, both established and inter-disciplinary. It is superfluous to say that a vast number of previous efforts aiming at automatic speech recognition have certainly contributed to clarification of the extent and depth of the problem, but have hardly achieved anything more than limited success.

With increased awareness of the difficulties involved in the achievement of the ultimate goal, recent efforts in speech recognition seem to have been concentrated mostly on placing heavy restrictions either on vocabulary or on speakers or on both, and have tested the validity of specific methods for problems of heavily limited scopes. In this way, a number of less drastic but none the less important achievements have been steadily accumulated in recent years. It is apparent, however, that these achievements cannot by themselves constitute a solution to the ultimate goal; rather a judicious combination of individual methods of approach is required with full understanding of their limitations, along with development of new methods for dealing with intricacies not encountered in solving individual problems.

---

* Paper to be presented at Speech Communication Seminar, Stockholm, Aug. 1-3, 1974.

** Faculty of Engineering; University of Tokyo.

The problem of speech recognition in the most general sense can be roughly divided into three stages:

(1) extraction of an effective set of parameters from a continuous speech signal.
(2) conversion of extracted parameters into a sequence of discrete linguistic units.
(3) recognition of larger linguistic units and their interrelationships based on lexical, syntactic and semantic structure of a language.

The present work is concerned with the second stage, which is usually called segmentation and recognition of phonemes, and has to cope with two major difficulties: variability due to individual differences of speakers and intersymbol interference between successive phonemes, commonly called coarticulation. It proposes a novel approach to overcome both of these difficulties in the case of connected vowels. A method of segmentation based on the 'Analysis-by-Synthesis' of formant trajectories is combined with a method of recognition of isolated vowels, achieving both unique segmentation and recognition of connected vowels at the same time. Although the proposed scheme is applicable to recognition of vowels and vowel-like sounds in more generalized contexts, the scope of the present paper is restricted to recognition of connected vowels.

2. Extraction of parameters of connected vowels [2, 3]

Since accurate extraction of acoustic characteristics of speech is considered to be essential for reliable recognition by the proposed scheme, formant frequencies are extracted using techniques that have been proved to be most accurate and reliable. Formant frequencies are preferred to predictor coefficients because of better separation of linguistic from non-linguistic information.[4] In order to follow their temporal variations, pitch-synchronous frequency analysis is adopted. That is, short-term frequency analysis is performed on the input speech signal with a time window approximately equal to one fundamental period, centered at successive local maxima of the absolute value of the speech signal. An example of such pitch-synchronous analysis is illustrated in Fig. 1, which indicates reliable and accurate detection of the spectral envelope in spite of the extremely short analysis window.

Formant frequencies are extracted from the spectral envelope by the method of 'Analysis-by-Synthesis' in the frequency domain. The model for synthesis of the spectral envelope over the frequency range of 0-5 kHz consists of factors representing contributions of the vocal tract transfer function, glottal source and radiation, and an additional factor for miscellaneous frequency characteristics that are not accounted for by other components. Frequencies of the first four formants are extracted by the method of successive approximation, minimizing the mean squared error between the observed and the synthesized spectra expressed in dB scale, and are determined with a quantum step of 50 Hz. Figure 2 shows an example of formant trajectories extracted from an utterance of connected vowels /aoi/.

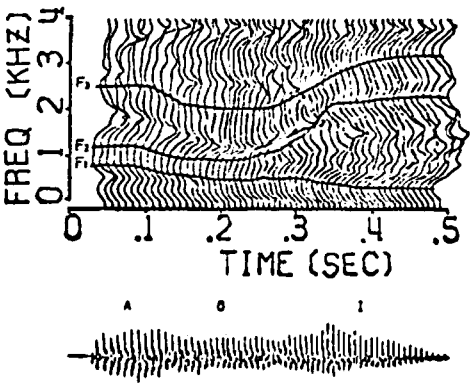

Fig. 1 An example of pitch-synchronous short-term spectral analysis of speech.

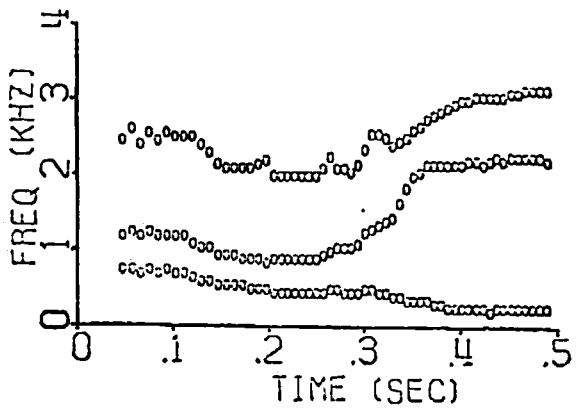

Fig. 2 An example of extracted formant trajectories in an utterance of /aoi/, obtained by analysis-by-synthesis in the frequency domain.

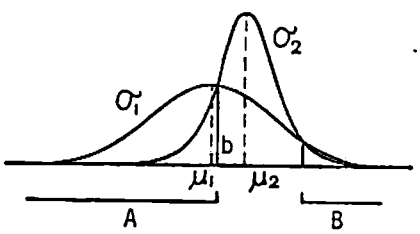

Fig. 3 An example of determination of decision boundary b for a pair of normally distributed pattern classes.

3. Recognition of isolated vowels [4]

Due to coarticulatory interferences, most of the vowels in connected discourse are hardly stationary, and their formant frequencies are not identical to those of isolated sustained vowels even at points where they appear to be nearly stationary. The initial vowel is an exception, however, and possesses a quasi-stationary portion which may be recognized by the technique that has already proved to be effective for recognition of isolated vowels.

A sample of a stationary vowel can be represented by a vector in a multidimensional space of formant frequencies, and samples belonging to a particular vowel are statistically distributed in the space. Thus optimum classification of vowels requires partition of the space into subspaces bounded by complex surfaces. For practical reasons, however, these boundaries may be approximated by a set of hyperplanes called linear discriminant functions.

A linear discriminant function is defined by a vector $\mathbf{a}$ and a scalar quantity b as

$$F(\mathbf{x}) = \mathbf{a}\mathbf{x} - b \qquad (1)$$

and $F(\mathbf{x}) = 0$ is a hyperplane which dissects the feature space. Though there are many ways to approximate optimum boundaries between pattern classes by a set of such hyperplanes, we first calculate the linear discriminant functions for the optimum separation of each pair of vowels, and then utilize them sequentially to construct the optimum decision tree for the classification of all the vowels with the minimum error rate.

The vector $\mathbf{a}$ is determined for each pair of vowels such that the separability ρ, defined as the ratio of inter-class variation to the sum of intra-class variations, should be maximized when samples of the pair of vowels are projected on $\mathbf{a}$. The maximum separability ρ between vowel classes V_i and V_j is given by

$$\rho = \frac{m_i m_j}{m_i + m_j} (\mathbf{g}_i - \mathbf{g}_j)^t (\mathbf{C}_i + \mathbf{C}_j)^{-1} (\mathbf{g}_i - \mathbf{g}_j) \qquad (2)$$

where m_i, $\mathbf{g}_i$ and $\mathbf{C}_i$ are the number of samples, the mean vector, and the covariance matrix of the vowel V_i. The mapping vector $\mathbf{a}$ is then given by

$$\mathbf{a} = \beta (\mathbf{C}_i + \mathbf{C}_j) (\mathbf{g}_i - \mathbf{g}_j) \qquad (3)$$

The decision boundary b is determined to minimize the error rate when Bayes' decision rule is adopted for the discrimination of a pair of normally distributed patterns, and is given by

$$b = \left\{ \sigma_j^2 \mu_i - \sigma_i^2 \mu_j + \sigma_i \sigma_j \sqrt{(\mu_i - \mu_j)^2 + 2(\sigma_j^2 - \sigma_i^2) \log(\sigma_j / \sigma_i)} \right\} \Big/ (\sigma_j^2 - \sigma_i^2) \quad , \qquad (4)$$

where μ_i and σ_i respectively denote the mean and the standard deviation of the samples of V_i projected on $\mathbf{a}$, and μ_j is assumed to be greater than μ_i (see Figure 3).

Figure 4 shows the optimum decision tree constructed to minimize the rate of recognition error using a linear discriminant function for each pair of vowels. Calculation of the decision function is based on isolated utterances of five Japanese vowels collected from a total of 60 speakers, including male and female adults and children differing widely in their ages. [5] The scores of correct recognition are 89.0% and 96.3% for the F_1-F_2 space and the F_1-F_2-F_3 space respectively, and indicate that highly reliable recognition is possible for isolated vowels by utilizing the first three formant frequencies. The result also suggests that information concerning the individual idiosyncrasies of speakers, contained in the set of three formant frequencies but discarded in the recognition of isolated vowels, may prove to be useful in the recognition of succeeding vowels in connected speech.

4. A model of the coarticulatory process in the formant frequency domain[6]

As shown in Fig. 5, the information content of an utterance is coded into phoneme strings at the center of speech, which are then converted into motor commands at the motor speech center and are sent out in parallel to various articulators. Although the commands for the realization of speech may be assumed to be stepwise signals, the resultant motions of the articulators are smoothed due to their inherent physiological and physical characteristics. Consequently, the area function of the vocal tract, as well as its formant frequencies, undergo complex processes of smoothing, i.e. coarticulation or concatenation. Since an exact formulation of this multi-stage intersymbol interference is beyond our present knowledge about mechanisms of speech production, an attempt is made here to approximate the entire process by a hypothetical linear system which accepts target formant frequencies of each vowel as input and converts them into the observed formant trajectories.

In the case of connected vowels and semivowels, results of analysis indicate that a formant transition from one phoneme to another is usually characterized by a gradual increase of the slope at the initial part and a non-oscillatory approach to its final or asymptotic value. The simplest mathematical expression satisfying these requirements is a step response of a critically-damped second-order linear system and is given by

$$f(t) = 1 - (1 + \alpha t) \exp (- \alpha t) \quad . \tag{5}$$

Although there is no proof that the approximation by a critically-damped system is optimum, and most of the results of analysis indicate that a better approximation could be obtained by an over-dampted system, the former is adopted here because it meets the above-mentioned requirements with the least number of parameters and yet is capable of approximating the major part of a formant transition with a fair degree of accuracy. The validity of such an approximation has already been proved in experiments on speech synthesis by rules. [7]

When a sequence of m vowels is pronounced in succession, therefore, the observed trajectory of the n-th formant frequency $f_n(t)$ can be approximated by

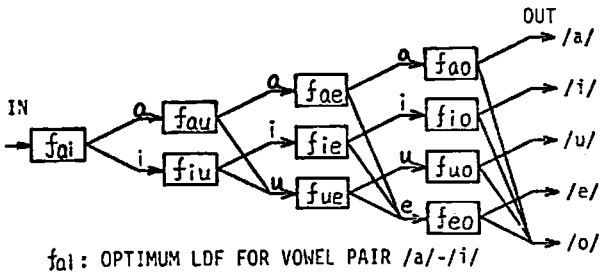

Fig. 4 Optimum decision tree for recognition of five Japanese vowels by linear discriminant functions (LDF).

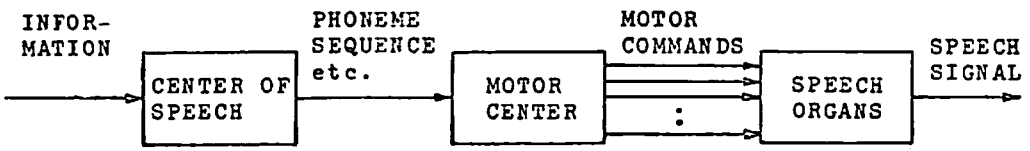

Fig. 5 Processes involved in speech production.

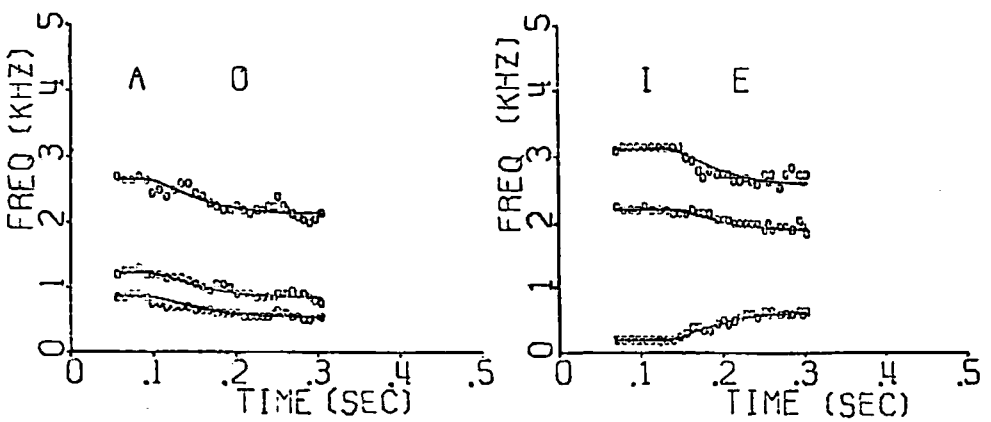

Fig. 6 Extracted formant trajectories and their best approximations based on a model of the coarticulatory process for /ao/ (left) and /ie/ (right).

$$f_n(t)=F_{n,1}+\sum_{j=2}^{m}(F_{n,j}-F_{n,j-1})\left[1-\left\{1+\alpha_{n,j}(t-\tau_j)\right\}\right]\exp\left\{-\alpha_{n,j}(t-\tau_j)\right\}U_{-1}(t-\tau_j)\ , \qquad (6)$$

where $F_{n,j}$ and τ_j respectively denote the target value of the n-th formant frequency and the time of onset of the command for the j-th vowel.

Figure 6 shows examples of comparison between the observed formand trajectories (in empty circles) and their best approximations (in solid lines) for connected vowels /ao/ (left) and /ie/ (right).

5. Principles of a scheme for recognition of connected vowels [8]

On the basis of techniques and results described earlier, a scheme is proposed for the recognition of connected vowels, as shown by the flow chart of Fig. 7. The input data are three formant frequencies extracted from pitch-synchronous short-term spectra by the method of 'Analysis-by-Synthesis' and the instantaneous power within each analysis frame.

Since initial vowels are comparatively free from coarticulatory influences, the method of linear discriminant functions described in Section 3 for the recognition of isolated vowels can be applied to the mean formant frequencies averaged over the initial portion of an utterance.

Once the initial vowel is recognized, its formant frequencies can be utilized for estimating target formant frequencies of other vowels of the same speaker, since the primary cause of individual variations in formant frequencies is known to be the difference in vocal tract length.

The detection and recognition of succeeding vowels are performed by iterating the following procedures. Unless the termination of an utterance is detected by a threshold of the instantaneous power, the match between measured formant patterns and the outputs of coarticulatory model is evaluated by the running average of squared relative errors in the three formant frequencies. A threshold value is selected to decide whether the same vowel is being sustained or a new vowel is initiated. When the latter is detected, an 'Analysis-by-Synthesis' of formant trajectories is performed using estimated target formant frequencies of other vowels of the speaker. One set of target formant frequencies among the four possible candidates as well as the instant of initiation of the transition is determined on the basis of the best match to measured formant transitions, thus accomplishing unambiguous segmentation and recognition of the succeeding vowel. The procedures are iterated to recognize a string of connected vowels until the end of an utterance is detected.

6. Experimental procedure and results

Recognition experiments were performed to test the validity of the proposed scheme. The speech material consisted of all possible combinations of two and three vowels of Japanese uttered by two male adult speakers, sampled at 10 kHz with an accuracy of 11 bits/sample, and stored in the magnetic tape unit of a digital computer. Each of these utterances was then analyzed by the method of Section 2, and the first three formant frequencies were extracted along with the instantaneous power within each analysis frame, the latter being utilized for the determination of the speech intervals.

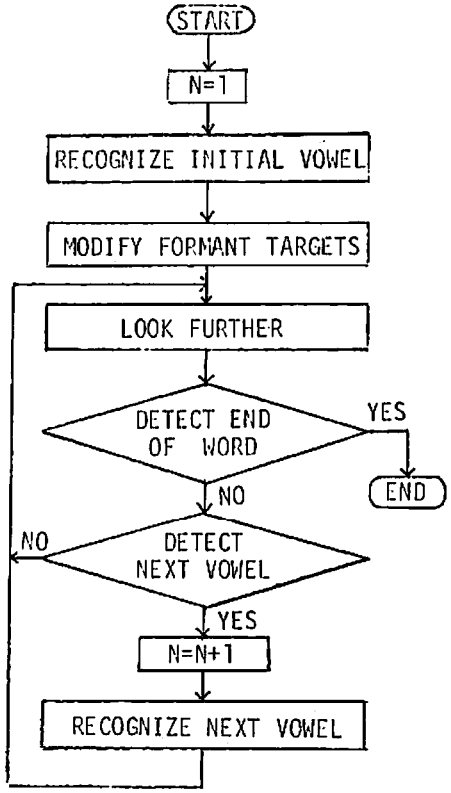

Fig. 7 A scheme for recognition of connected vowels.

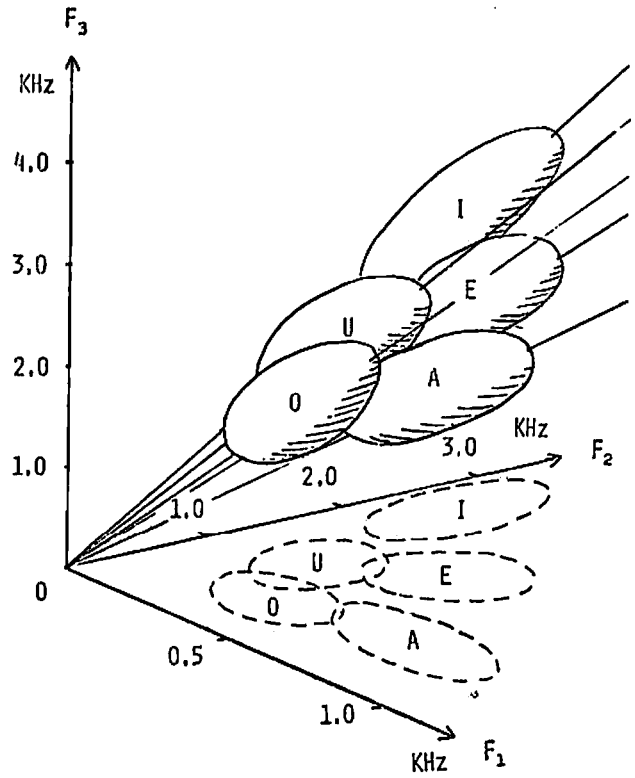

Fig. 8 Distribution of vowel samples in three-dimensional formant space.

As mentioned above, the initial portion of approximately 30 msec of connected vowels can be regarded to be prior to the onset of commands for the second vowel, so that the method for recognition of isolated sustained vowels can be applied to the mean formant frequencies of the initial portion For the speech materials under study, a recognition score of 100% was obtained.

Using formant frequencies of the initial vowel thus recognized, target or steady-state formant frequencies of other vowels of the same speaker are estimated by the following procedure. Since the primary cause of variations in formant frequencies is the individual difference in vocal tract length, formant frequencies of all the vowels are distributed mainly along the radial dimension in the space constructed by three formant frequencies, as illustrated by Fig. 8. Based on the knowledge of typical formant frequencies of all the five steady-state vowels, and those of one vowel of a particular speaker, therefore, it is possible to obtain a proportional coefficient, by which the typical formant frequencies of the other four vowels can be converted into fairly good estimates for those of the other four vowels of the same speaker. Namely, let the typical n-th formant frequency of a vowel V_i be $F_{n,i}$ and the measured n-th formant frequency of the same vowel of a particular speaker be $f_{n,i}$. Then the formant frequency of another vowel V_j of the same speaker can be estimated as

$$f_{n,j}=F_{n,j}\sqrt{\sum_{n=1}^{3} f_{n,i}^2 \Big/ \sum_{n=1}^{3} F_{n,i}^2} \qquad (7)$$

For reliability of estimation, the weighted mean of the above estimate and the observed formant frequency of the same vowel at the initial position, whenever it is available, is adopted as the true estimate for the target formant frequency.

When the initial vowel is recognized, the output of the coarticulatory model sustains its formant frequencies, and the overall match between measured and predicted formant trajectories is evaluated at every 10 msec by taking the running average of squared relative errors in three formant frequencies over the immediately preceding interval of 40 msec duration. The comparison is repeated until the mismatch exceeds a certain threshold, indicating the existence of a succeeding vowel.

Once the mismatch exceeds the predetermined threshold, a search for the optimum set of target formant frequencies among the four candidate sets and for the instant of onset of transition is initiated, based on comparison of measured and predicted formant trajectories over an interval starting at the instant 50 msec prior to the instant of detection of the succeeding vowel and extending approximately over the duration of a mora, which is set equal to 250 msec in the present study.

After the second vowel is thus recognized, the match between measured and predicted formant trajectories is again evaluated at every 10 msec to see whether the same vowel is being sustained or a third vowel is initiated. In this case, however, the target formant frequencies can be readjusted within a narrow range around the originally estimated values to improve the estimation. These procedures are repeated for the sequential

segmentation and recognition of each of the connected vowels until the end of the final vowel.

An example of analysis and recognition of connected vowels /aoi/ is shown in Fig. 9, where empty circles indicate formant frequencies, and dotted curves in the left figure indicate formant trajectories obtained by 'Analysis-by-Synthesis,' while the straight lines in the right figure indicate extracted target formant frequencies as inputs to the coarticulatory model. The crossing of trajectories of the second and the third formant frequencies during transition from the back vowel /o/ to the front vowel /i/ is based on considerations of continuity of resonance modes in the vocal tract. If, however, we adhere to the conventional numbering of formants in the ascending order of their frequencies, the apparent instants of onset for the second and the third formants may differ considerably in transitions between a back vowel and a front vowel, in spite of the fact that an articulatory transition should cause simultaneous onset of all the formant transitions.

Due to the existence of coupling between resonance modes, frequencies of two formants never coincide in the actual speech signal, but always tend to separate, as shown by the extracted results. The solid curves in the left figure indicate outputs of an improved model of coarticulation which takes into account the separation of formant frequencies due to coupling between resonance modes, with the degree of coupling being estimated from measured formant trajectories.

Except for speech materials with insufficient recording levels, all the vowels in 35 samples of two-vowel words by two speakers and in 71 samples of three-vowel words by one speaker were correctly recognized. Although the proposed method has to be tested both for a greater number of speakers as well as for a wider range of talking rates, preliminary results of recognition of some five-vowel words indicate that a highly reliable recognition will be possible without limitation on the length of utterances to be recognized.

7. Conclusions

A new method for recognition of connected vowels has been proposed to cope with both individual differences of speakers and mutual interferences of phonemes due to coarticulation. On the basis of techniques for accurate extraction of formant frequencies and for reliable recognition of isolated vowels, both previously developed by the authors, the process of coarticulation between vowels has been formulated in the formant frequency domain and utilized as a model in the 'Analysis-by-Synthesis' of formant trajectories, thus accomplishing unambiguous segmentation and recognition of successive vowels. The validity of the method has been confirmed by recognition experiments on a number of two- and three-vowel words. Further work is in progress to test its applicability to a greater number of speakers and a wider range of talking rates, as well as to recognition of vowels in more generalized contexts than in connected vowels.

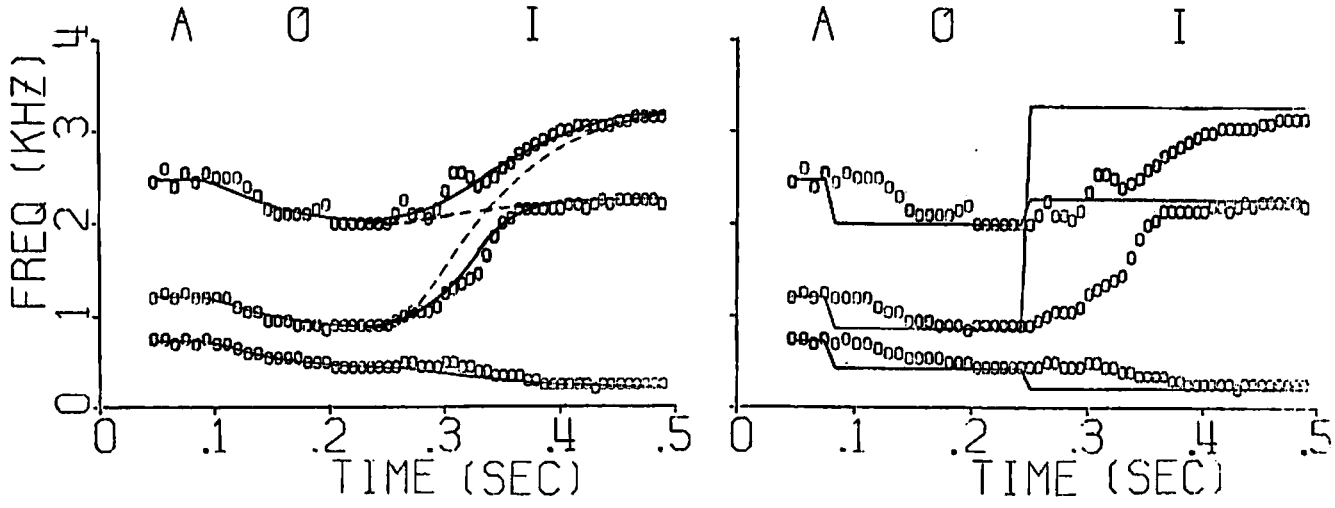

Fig. 9 Analysis-by-Synthesis of formant trajectories (left) and their target values (right) of vowels in an utterance /aoi/ based on a coarticulatory model in the frequency domain.

References

1. Fujisaki, H. : Current problems in speech recognition, J. Acoust. Soc. Japan, 28 (1972) 33-41.
2. Fujisaki, H. and Yoshimune, K. : Estimation of short-time frequency spectrum of quasi-periodic waveforms with applications to pitch-synchronous analysis of speech, Rec. Spring Meeting, Acoust. Soc. Japan, (May 1971) 1-1-11.
3. Bell, C. G., Fujisaki, H. et al. : Reduction of speech spectra by Analysis-by-Synthesis techniques, J. Acoust. Soc. Am., 33 (1961) 1725-1736.
4. Fujisaki, H. and Sato, Y. : Evaluation and comparison of feature spaces for speech recognition, Rec. Spring Meeting, Acoust. Soc. Japan, (May 1973) 2-1-4.
5. Fujisaki, H. and Nakamura, N. : Normalization and recognition of vowels, Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo, 28 (1969) 61-66.
6. Fujisaki, H. and Tanabe, Y. : Coarticulatory model in frequency domain and its application to recognition of connected speech, Rec. Spring Meeting, Acoust. Soc. Japan, (May 1973) 2-1-2.
7. Rabiner, L. R. : Speech synthesis by rules: An acoustic approach, Bell Syst. Tech. J., 47 (1968) 17-37.
8. Fujisaki, H. and Yoshida, M. : An experiment on recognition of connected formant frequency domain, Rec. Spring Meeting, Acoust. Soc. Japan, (May 1973) 2-1-3.