STUDIA PHONOLOGICA IV (1966)

A Procedure for Formant Domain Extraction

Toshiyuki SAKAI, Shuji DOSHITA and Yasuhisa NIIMI

Summary

This paper describes a procedure for extracting the formant domains, in which there are the formants of a vowel, from a measured vowel spectrum by means of a digital computer, KDC-I, as a step preceding an accurate extraction of the formants. The formant domains are expected to be obtained as the frequency regions, where a spectral value is larger than a certain threshold value. A locally weighted mean of spectral values was adopted as the threshold value. The spectral data used for the test of this procedure were obtained from the connected vowel sounds of the words spoken by two male speakers. The formant domains may be said to be extracted fairly well for the tested utterances, although they are small in number.

1. Introduction

When speech sounds are analyzed by means of a frequency analyzer, some peaks appear in their spectra. Especially in vowel sounds, four peaks are dominant in the frequency range lower than about 4 kc. The frequencies at these peak points are called formant frequency and they are named first formant, second formant, etc., in the order they occur in the frequency scale and nearly corresponding to the resonance frequencies of the vocal tract cavity system; that is, the acoustic tube from lips to the glottis. The resonance frequencies can change only as a result of an articulatory change affecting the dimensions of the various parts of the vocal tract system and thus formant frequencies.

It is believed that the lowest two formants contribute primarily to the discrimination among vowels and also the third formant in front vowels, and that other higher formants as well as the pitch, the fundamental frequency of the source excitation, mainly contribute to the naturality and personality of vowels. At an attempt to recognize the speech sounds automatically, therefore, the first two formants play the important roles in vowel sounds. Although a number of automatic formant extraction methods have been studied by means of digital computers as well as various analog techniques, the method is not discoverd by which the accurate formants are extracted automatically and in real time.

Toshiyuki SAKAI, Ph.D. (坂井利之): Professor of Electrical Engineering, Kyoto University.
Shuji DOSHITA (堂下修司): Assistant Professor of Electrical Engineering, Kyoto University.
Yasuhisa NIIMI (新美康永): Assistant of Electrical Engineering, Kyoto University.

As an approach to the goal, in the present paper, a procedure was tested for extracting the frequency domains, where the formants exist, rather than the accurate formant frequencies. Although the procedure was performed by using a digital computer, it could be easily realized with analog circuits.

2. Procedure

As mentioned in the preceding section, the formant frequency is frequency at which the dominant peak appears in a vowel spectrum. If the formant levels, or the levels of the dominant peaks in a spectrum, are nearly equal, the formant domains are expected to be extracted as the frequency regions where spectral values are larger than a certain threshold value. In general the formant levels, however, are not equal in a vowel spectrum. Especially the second formant level of a Japanese vowel /u/ is often lower than other formant levels. The change of the input level of a vowel sound and of the gain of the frequency analyzing system brings the variation of a spectral level. In extracting the formant domains by the threshold method, therefore, it is not suitable to use the constant threshold value all over the frequency range of interest. The threshold value satisfying the following requirements is desirable;

(1) it is large or small depending on the local magnitude of a spectral value,

(2) it is relatively invarient to the variation of a spectral level, that is, the formant domains to be extracted are invariant to it.

A locally weighted mean was adopted as the threshold value, since it satisfies these requirements as shown blow. The threshold curves were computed for two different spectral representations; for a power spectrum in decibel scale and in linear scale. For each case, the formant domains were extracted as the frequency regions where the difference between the spectral value and its threshold value was positive.

Case (1), where a power spectrum is expressed in decible scale. The threshold value $T(i)$ for the i-th channel of the frequency analyzer used is given by

$$T(i)=\sum_{j=c_1}^{c_2} w_{ij}S(j) \qquad (i=1,\ 2,\ \cdots\cdots n) \tag{1}$$

where $S(j)$ is the j-th channel output of the analyzer in decibel scale, c_1 the larger integer of zero and $i-r$, c_2 the smaller integer of n and $i+r$, n the number of channels of the analyzer, forty five in this case, and r a given positive integer. The w_{ij}'s are weighting factors of the threshold value in the i-th channel and are expressed as follows;

$$w_{ij}=\begin{cases} a_i \Big/ \left\{1+\dfrac{(i-j)^2}{k^2}\right\} & (i\neq j) \\ 0 & (i=j) \end{cases} \tag{2}$$

where a_i's and k are parameters. The w_{ii}'s were equated to zero so that the difference between $S(i)$ and $T(i)$ might be more prominent in the formant domains. When the parameter k is small, a threshold curve draws near its original spectrum.

Since the threshold curve is quite coincident with its original spectrum at the limit as k tends to zero, it is no more useful as a threshold curve. On the other hand, the threshold curve becomes gradually flat, as k increases. Since the weighting factors are equal at the limit as k tends to infinite, the threshold curve dose not satisfy the first requirement mentioned above. Although it is sure from these considerations that the optimum value of the parameter k exists, it is difficult to find it theoretically. So the parameter k was set three from the result of the test computation of the threshold curves for some spectra in the stationary and the transient parts of connected vowel sounds. The integer r was set twenty also from the result of the test computation. The a_i's are selected such that $\sum_{j=c_1}^{c_2} w_{ij}=1$.

In the present case, the variation of the spectral level means the addition of a constant A to $S(j)$. The difference between a channel output and its threshold value is

$$\{S(i)+A\}-\sum_{j=c_1}^{c_2} w_{ij}\{S(j)+A\}=S(i)-\sum_{j=c_1}^{c_2} w_{ij}S(j)+A\{1-\sum_{j=c_1}^{c_2} w_{ij}\}$$

The third term of the right side of the above identity is equal to zero according to the above condition on w_{ij}'s. The difference between a channel output and its threshold value is turned to be invariant to the variation of the spectral level; in other words, the formant domains to be found can be invariant since they are extracted depending only on the sign of the difference. Of course, $T(i)$, locally weighted mean of $S(j)$'s, is large or small depending on the local magnitude of the spectral value. Thus it is known that $T(i)$ defined by the equation (1) satisfies the two requirements mentioned above.

Case (2), where a power spectrum is expressed in linear scale. The threshold value $T'(i)$ for the i-th channel of the frequency analyzer used is given by

$$T'(i)=10\ \log\sum_{j=c_1}^{c_2} w_{ij}^2P(j) \qquad (i=1,\ 2,\ \cdots\cdots n) \qquad (3)$$

where $P(j)$ is the value of the j-th channel of a spectrum expressed in linear scale and is related to $S(j)$ by the equation,

$$P(j)=10^{\frac{S(j)}{10}} \qquad (j=1,\ 2,\ \cdots\cdots n) \qquad (4)$$

and w_{ij}'s are given by the equation (2). The other notations have the same meanings as those in the case (1). A channel output obtained from the device is expressed in decibel scale as mentioned in the next section, and so $P(j)$'s must be reproduced in a computer following the equation (4). This is, however, easily performed by a table look up method in a computer. The parameter k has the same relation with the threshold curve as that in the case (1). The k and r were set three and fifteen, respectively, from the result of the test computation. The a_i's were selected such that $\sum_{j=c_1}^{c_2} w_i^2=1$, except for the i's lower than five and higher than forty. It follows that the threshold curve for a flat spectrum is also flat and has the same level as the original flat spectrum. The larger is the sum of w_{ij}^2's selected, the higher level has the threshold curve. In general the spectrum of a vowel sound has the high energy level in the low frequency region, not a formant domain, because of the pitch frequency components of the source excitation itself, and has the low energy level and the unwanted fluctuations in level in the high frequency region. The a_i's for these exceptional i's were selected such that $\sum_{j=c_1}^{c_2} w_{ij}^2 > 1$, in order to keep from extracting a wider domain including the redundant low frequency region as a formant domain, and from being influenced by the unwanted fluctuations in the spectral level in the high frequency region.

The variation of the spectral level means the addition of a constant A to $S(j)$ as stated in the case (1). In the present case, it corresponds to the multiplication of $P(j)$ by a constant factor B related by $A=10 \log B$. It is easily shown by substituting $BP(j)$ into the equation (3) that the change of the threshold value $T'(i)$ is equal to that of $S(j)$, or A, in spite of the selection of the values of a_i's. In other words, the formant domains, which were defined as the frequency regions where the difference, $S(i)-T'(i)$, was positive, can be invariant to the variation of the spectral level.

3. Data

In order to obtain the spectral data for testing this procedure, spectral analysis was performed through a 45 channel filter bank, each having a half power bandwidth of 100 cps. The center frequency of the lowest filter was 50 cps, and the spacing was uniformly 100 cps. Thus the filter bank could cover the frequency range lower than 4.5 kc. The output the filter bank was rectified and sampled by a multiplexor at a rate of 100 samples per second and converted to decibel. A sampled spectrum was represented on a cathode ray tube and photographed. Each channel output was read with a decibel step and punched on a paper tape for computer input. The spectral data were obtained from successive vowel parts in fourteen utterances produced by two male speakers, each of which have about thirty spectral samples.

4. Results and Discussions

The difference spectra, $S(i)-T(i)$ for the case (1) and $S(i)-T'(i)$ for the case (2), were computed. The frequency regions in which they have positive value were extrated as the formant domains. The difference spectrum was quantized with a 4 dB step in the formant domains and was deemed as zero where it had negative value. Corresponding to the spectrum pattern of the original speech sound, the difference pattern was printed out by the computer. An appropriate letter was assigned for each level; for channels in which the differences were negative, no letter were printed out.

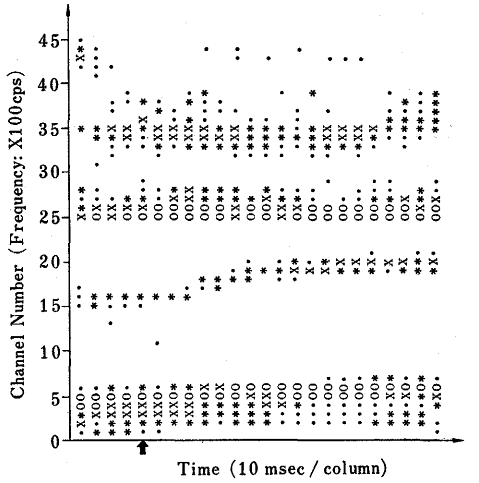

Figure 1. Typical result in the case (1) where k=3 and r=20. The spectral data are obtained from the vowel part of a Japanese word /sue/. The symbol "0" denotes the quantized level higher than 12 dB, "X" from 12 dB to 8 dB, "*" from 8 dB to 4 dB and "•" from 4 dB to 0 dB.

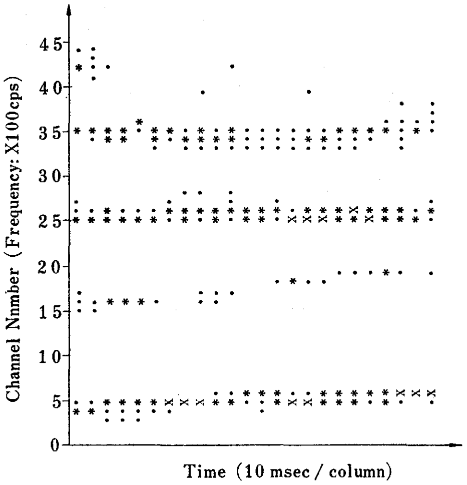

Figure 2. Typical result in the case (2) where k=3 and r=15. The spectral data and the meaning of symbols are the same in Figure 1.

The typical results of this procedure are shown in Fig. 1 and Fig. 2; Fig. 1 shows the result of the case (1) and Fig. 2 that of the case (2). The spectral data in these figures are obtained from the vowel part of a Japanese word /sue/. A spectrum and threshold curves at a sampling point marked with an arrow in Fig. 1 are shown in Fig. 3. In Fig. 1, the first and the fourth formant domain are wider and especially in the fourth formant domain, the effect of local peaks which are not formant is dominant. In Fig. 2, these points are corrected, but the

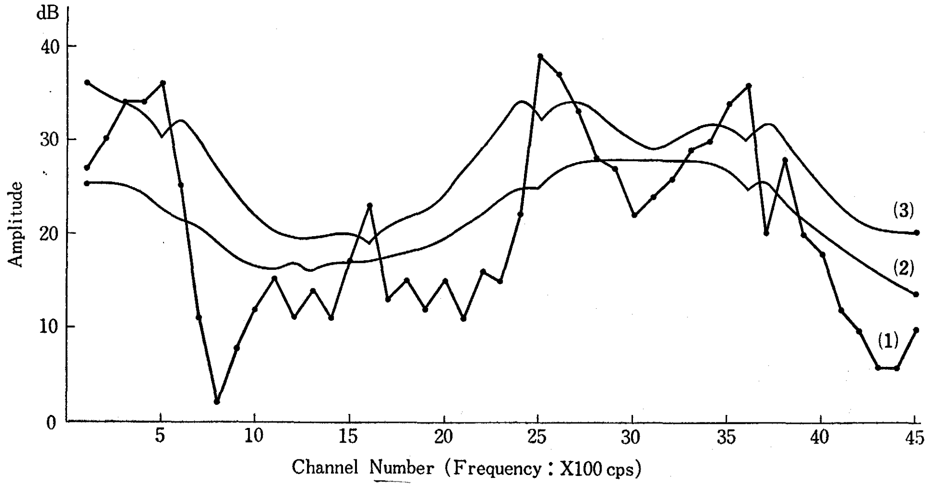

Figure 3. A spectrum and threshold curves at the sampling point marked with an arrow in Figure 1 ; curve (1) shows a spectrum and curve (2) a threshold curve in the case (1) and curve (3) in the case (2).

extraction of the second formant domain fails at some sampling points. As showed in Fig. 3, the formant domain in which a spectral level is small as compared with those in other formant domains are considerably extracted. However, there are some utterances in which the first and the second formant, closed in frequency as in vowel /a/, can not be separated into two domains. The movement of formants is apparent as much as in the pattern of Sonagram. The formant domains may be said to be extracted fairly well for tested utterances, although they are small in number. It is necessary to test this procedure for much more utterances and to study the effects of varing weighting factors on the extraction accuracy.

The formant domains obtained by this procedure may be used effectively as the first approximations for the accurate formant extraction method, say, Analysis-by-Synthesis method(1)(2)(3) which has been proposed and developed by Stevens, et al. Because of not requiring so much time, this method may be useful and substituted for Sonagram if it is desired to deal with great number of data and is not required high accuracy in extracting formants. Since the procedure could be realized with simple analog circuits, it might be expected to play an important role as a part of a real time processing system of speech sounds.

The procedure was hinted from the consideration of masking with multitones. Masking with a pure tone or a band limited noise has been long studies since the basic experiment by Wegel and Lane in 1924.(4) However, only a few studies of masking with multitones have been done. It is known that beat oscilations between maskers as well as markers themselves may seem to influence on a maskee.(5) Neglecting the influence of these beats, it was considered how the influence of each masker on the maskee was superposed. Assuming that the influence of each masker be idealized as the weighting factor, w_{ij}, mentioned in section (2) and be superposed additively as the power sum, the equation (3) was obtained as the amount of masking with multitones. In order to get the closer analogy of this procedure to the analysis in the auditory system, the procedure must performed in logarythmic or mel scale rather than in linear scale. To do this, it is necessary to analyze the speech sounds by the filters with band widths equal in logarythmic or mel scale, and to use the expression of the weighting factors in logarythmic or mel scale.

Reference

1) Stevens, K.N. : Toward a Model for Speech Recognition. J. Acoust. Soc. Am. 32 ; 47. 1960

2) Bell, C.G. et al. : Reduction of Speech Spectra by Analysis-by-Synthesis Techniques. J. Acoust. Soc. Am. 33 ; 1725. 1961

3) Paul, A.P. et al. : Automatic Reduction of Vowel Spectra : An Analysis-by-Synthesis Method and Its Evaluation. J. Acoust. Soc. Am. 36 ; 303. 1964

4) Wegel, P.L. and Lane, C.E. : The Auditory Masking of One Pure Tone by Another and Its Probable Relation to the Dynamics of the Inner Ear. Phys. Rev. 23 ; 266. 1924

5) Green, D.M. : Masking with Two Tones. J. Acoust. Soc. Am. 37 ; 802. 1965