STUDIA PHONOLOGICA XIX (1985)

Environmental Noise Reduction Method Based on Formant Analysis-Synthesis Principle

Masaaki NAGATA, Yasuo ARIKI, and Toshiyuki SAKAI

Summary

A method of noise reduction based on formant analysis-synthesis principle is described. In this method, pitch and formants are extracted from noise superimposed speech, and noise reduced speech is synthesized from extracted parameters. To improve the accuracy of parameter extraction, the extraction procedure consists of three steps. First, vowel and consonant intervals are detected and rough estimates of the parameters in each interval are calculated. Secondly, the extracted parameters are checked to see if they are consistent in view of both global information such as parameter continuity and local information such as phonetic formant structure in order to make a plan for detailed reinvestigation. Lastly, the data are investigated again according to the plan in order to make a final decision of the value of the parameters.

Experiments based on this method show that, in case of white noise superimposed speech, although the intelligibility of speech is rather decreased, the back ground noise is reduced and the easiness of hearing is fairly improved.

1. Introduction

Recent advance in the research on speech recognition and synthesis has made audio input units and audio response units practical. It has shown, however, that the recognition rate and synthesis quality of speech by these units are degraded considerably in the presence of environmental noise. To meet the increasing demand of improving S/N ratio of speech in environmental noise, a method of noise reduction employing formant analysis-synthesis is established and described in this paper.

Speech information are said to be mainly associated with pitch and formant. Pitch, which is the fundamental vocal frequency, has the information of vocal excitation. Formants, which are the vocal resonances, have the spectrum envelope information. Vowels are characterized by the first three or four formants. It is said that the relative location of formants are most closely assoicated with proper

Masaaki NAGATA (永田昌明): Student, Department of Information Science, Kyoto University.
Yasuo ARIKI (有木康雄): Assistant Professor, Department of Information Science, Kyoto University.
Toshiyuki SAKAI (坂井利之): Professor, Department of Information Science, Kyoto University.

vowel identification. Consonants do not have formants. It is said, however, that the initial formant transition are related to proper identification of the consonant preceding the vowel, and the final formant transition are related to proper identification of the consonant following the vowel. For these reason, it is considered to be important to obtain the accurate data of formants during both the transition state and steady state of vowel intervals.

Formant analysis-synthesis method is based on the speech generation model. It is assumed that, if accurate pitch and formants are extracted from noisy speech, noise reduction can be performed by synthesizing voice from the extracted parameters. But there is a problem in this method that it is difficult to get accurate values of parameters due to the effect of superimposed noise.

To improve the accuracy, global information as well as local information are used in this method. Global information means the outline of parameter transition which indicates the rough estimates and the tendency of transition. Local information means the data obtained by the signal or acoustic processing within a short time analysis interval or 'frame'. In formant extraction, extracted parameters are also checked to see if they are consistent to the phonetic formant structure, which is represented by the relative location of the formants in each phoneme.

2. FORMANT ANALYSIS-SYNTHESIS METHOD

2.1 Principles of the method

The feature of the method described in this paper is that it is based on parametric method and that it considers the reliability of the results of extraction.

Noise reduction methods are classified into two groups: parametric method and nonparametric method. The former is based on the speech generation model, whereas the latter is not.

Nonparametric method is mainly based on signal processing. The periodicity of speech, or the absence of correlation between speech and noise are used for noise reduction. Nonparametric method often results in the loss of naturalness and articulation, because it does not consider the spectrum structure of speech.

Parametric method, on the contrary, can preserve spectrum structure, because it is based on the speech generation model. It is expected that parametric method has a ability to restore spectrum structure damaged by superimposed noise and to improve the naturalness and articulation of noise reduced speech. In conclusion, parametric method is more powerful in noise reduction than nonparametric method.

The parameters extracted from each short-time analysis frame are not equally reliable and accurate. According to the experiment, in noisy environment, extraction accuracy in steady part of vowel is better than in the other portions such as consonant or transition part of vowel. The reason for this is that the S/N ratio in the steady part of vowel is the highest in its neighborhood, because the total power of the speech is the maximum there.

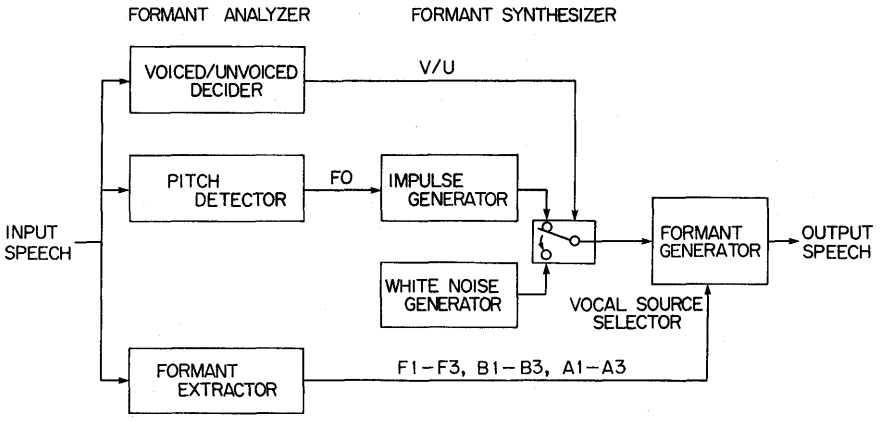


Fig. 2.1. Formant analysis-synthesis system

Table 2.1. Parameters of formant analysis-synthesis system

Vocal excitation parameters	V/U	Voiced or unvoiced
	FO	Pitch frequency
Vocal tract parameters	F1–F3	Central frequency of the first three formant
	B1–B3	Band width of the first three formant
	A1–A3	Amplitude of the first three formant

It is expected that the method considering the reliability of the extraction parameters is more powerful than the method ignoring the reliabiilty and dealing with every extracted parameter equally. It is also expected that the analysis method starting from the most reliable intervals and proceeding both forward and backward improve the accuracy of parameters.

2.2 Formant Analysis-Synthesis System

Formant analysis-synthesis system consists of formant analyzer and formant synthesizer. The block diagram of formant analysis-synthesis system is shown in fig 2.1 and the parameters which connect analyzer and synthesizer is shown in table 2.1.

2.2.1 Formant Analyzer[1)]

Formant analyzer analyzes input speech and extracts the information of vocal excitation and vocal tract as the parameters shown in table 2.1.

Vocal excitation is analyzed by voiced/unvoiced decider and pitch detector. Voiced/unvoiced decision is made by voiced/unvoiced decider and the result is indicated by the paraemter V/U. Pitch detection is performed by pitch detector and pitch frequency F0 is extracted, if exists.

Vocal tract is analyzed by formant extractor. Central frequencies F1–F3, band widths B1–B3, and amplitudes A1–A3 of the first three formants are ex-

tracted from the spectrum envelope of the speech.

2.2.2 Formant Synthesizer[2)]

Formant synthesizer synthesizes speech by generating vocal excitation and approximating the transmission properties of the vocal tract.

Vocal excitation is generated by impulse generator, white noise generator, and vocal source selector. Impulse generator generates impulse train whose period equals the reciprocal of pitch frequency F0 as the vocal source of voiced sound. White noise generator generates gaussian white noise as the vocal source of unvoiced sound. Vocal source selector selects these vocal sources according to voiced/unvoiced decision V/U.

The filter whose transmission property approixmates that of vocal tracts is made by formant generator. It changes its transmission property according to the formant data F1–F3, B1–B3, and A1–A3.

3. PITCH AND FORMANT EXTRACTION IN NOISY ENVIRONMENT

3.1 Pitch and Formant Extraction based on Cepstrum Method

In formant analyzer, the information of vocal excitation and vocal tract is extracted in separate form by the cepstrum method. Cepstrum is the power spectrum of the logarithm spectrum. It gives us a method for pitch and voiced/unvoiced detection as well as a method for obtaining spectrum envelope in the following manner.

Consider speech generation model such that it consists of vocal excitation and vocal tract. Vocal excitation is approximated by the combination of impulse generator, white noise generator, and selector. Vocal tract is approximated by filter.

The speech $x(t)$ equals the convolution of the vocal excitation $u(t)$ and the impulse response of the vocal tract $h(t)$,

$$x(t)=h(t)*u(t) \tag{3.1}$$

where * indicates the convolution. By taking the Fourier transform, the relationship that the spectrum of speech $X(\omega)$ equals the product of the spectra of the vocal excitation $U(\omega)$ and of the vocal tract $H(\omega)$ is introduced.

$$X(\omega)=H(\omega)\cdot U(\omega) \tag{3.2}$$

The information of vocal excitation is separated from that of vocal tract by taking the logarithm of the speech spectrum.

$$\log|X(\omega)|=\log|H(\omega)|+\log|U(\omega)| \tag{3.3}$$

By taking the inverse Fourier transform of this expression, the spectrum of the logarithm spectrum or 'cepstrum' of the speech $x(\tau)$ is expressed by the sum of the cepstra of the vocal excitation $h(\tau)$ and of the vocal tract $u(\tau)$.

$$x(\tau)=h(\tau)+u(\tau) \tag{3.4}$$

The independent variable in the cepstrum is called 'quefrency' which has the diminsion of time.

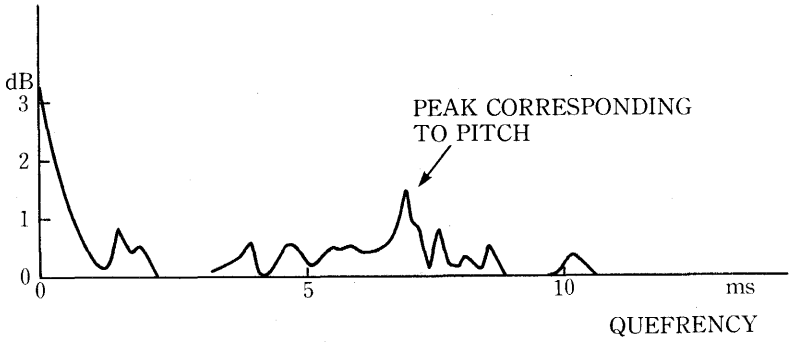
The quasiperiodic repetition of the waveform in a voiced interval of speech causes periodic ripples in the speech spectrum. The frequency spacing between the peaks of these ripples equals the fundamental frequency of the speech. As the formants are usually widely separated in the spectrum, the cepstrum of speech consists of high quefrency components caused by the vocal excitation and of low quefrency components caused by the formant structure.

Pitch detection and voiced/unvoiced decision are possible based on the presence or absence of a ceptsrum peak in high quefrency. Spectrum envelope, on the other hand, can be obtained by the Fourier transform of the low quefrency components. Formant extraction is achieved by selecting the first three or four major peaks in the spectrum envelope.

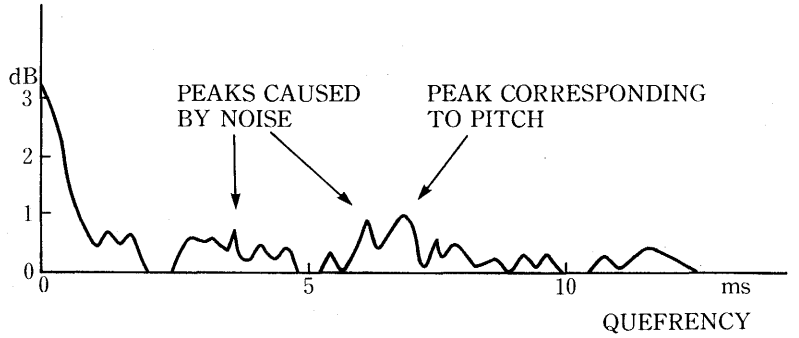
3.2 Influence of Noise

Pitch and formant extraction are based on the peak picking, but this method does not work well in noisy environment because of spurious peaks caused by the superimposed noise.

In the cepstrum of noise superimposed speech, the conspicuous peak in high quefrency of voiced interval which corresponds to pitch becomes indistinct and

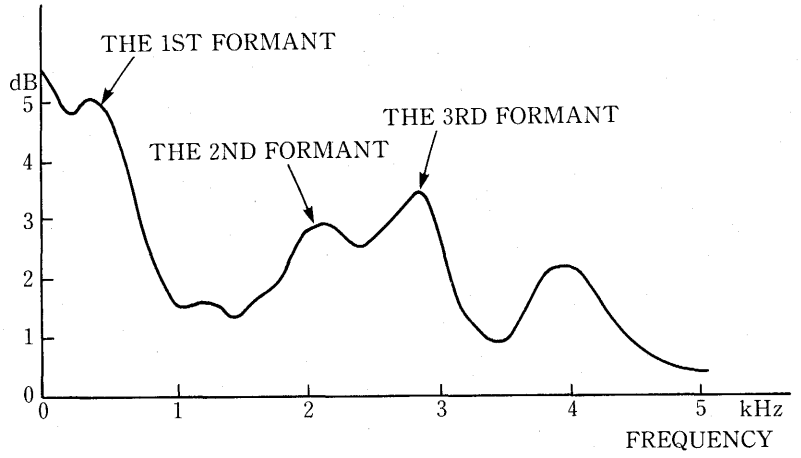


(a) Cepstrum in a voiced interval (noise free)

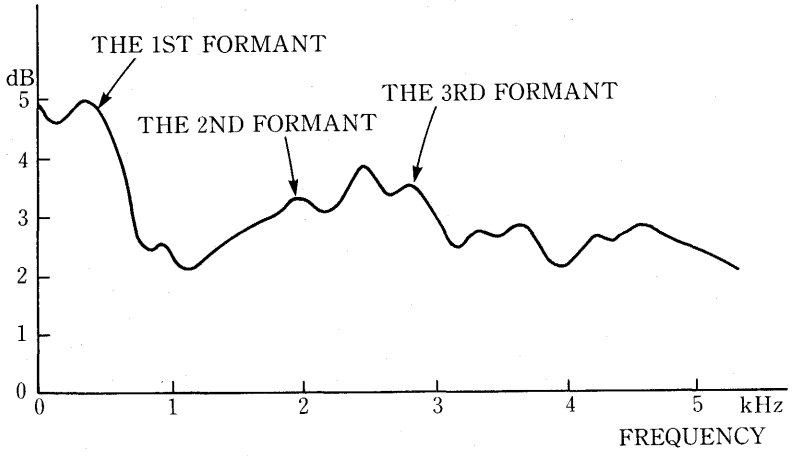


(b) Cepstrum in a voiced interval (noise superimposed)

Fig 3.1 Cepstrum in noisy environment.



(a) Spectrum in a vowel interval (noise free)



(b) Spectrum in a vowel interval (noise superimposed)

Fig 3.2 Spectrum in noisy environment.

many spurious peaks irrelevant to pitch appear as shown in fig 3.1. This makes the pitch detection and voiced/unvoiced decision very difficult.

In the spectrum envelope of noise superimposed speech, similarly the major peaks correspond to formant become indistinct and many spurious peaks irrelevant to formant appear as shown in fig 3.2. This makes the formant extraction very difficult.

3.3 Structural Analysis to Improve Extraction Accuracy

Parameters extracted in noisy environment by the information within a frame often have errors due to spurious peaks. To improve the accuracy of extracted parameters, the method described in this paper uses global information.

As spurious peaks in noise superimposed speech appears irregularly, it is assumed that they will be cancelled if several frames are averaged. A rough estimate or tendency of transition is obtained from this average. However, as a result,

detailed informations such as the locus of formant transition are lost, which is said to be related to consonant identification.

The pitch and formant extraction procedure is consists of three steps. The first step is to obtain rough estimates or tendencies by averaging the parameters extracted from each frame. The second step is to make a plan for detailed investigation based on the data obtained in the previous step. The third step is to make a close reinvestigation to extract the more accurate parameters.

4. Pitch and Formant Extraction Algorithm

4.1 Detection of Vowel and Consonant Intervals

The input speech data is segmented into vowel, consonant, and silent intervals at the first stage of the processing.

The segmentation of pronounced interval and silent interval is based on the level of the total power of the speech. In pronounced interval, the segmentation of vowel interval and consonant interval is based on the level of the power of speech in the frequency from 200 Hz to 1000 Hz, as the first formant usually appears in this frequency band in vowel interval.

4.2 Pitch Extraction and Voiced/Unvoiced Decision

4.2.1 Pitch Extraction

The pitch extraction algorithm is based on the cepstrum method, but it is modified to improve accuracy in noisy environment. The pitch extraction proceudre consists of three steps.

First, a frame n such that its power is the maximum is searched in a vowel interval, where n is the frame number. In the preceding frame, the frame, and the following frame, the maximum cepstral peak which probably corresponds to pitch is selected to obtain its quefrency $\tau(n-1)$, $\tau(n)$, $\tau(n+1)$, respectively.

Secondly, mean value is calcuulated, which serves as a reliable estimate of pitch period.

$$\bar{\tau}=1/3(\tau(n+1)+\tau(n)+\tau(n-1)) \tag{4.1}$$

The quefrency band where pitch is likely to appear is established as follows,

$$[(1-\delta)\bar{\tau},\ (1+\delta)\bar{\tau}] \tag{4.2}$$

where δ is a parameter which decides the width of this band. The value we used is $\delta=0.3$.

Lastly, all frames in the vowel interval are looked over again to find the maximum cepstral peak within the above band. This peak is judged to correspond pitch and pitch preiod is obtained from its quefrency.

In consonant interval, the band set up in the following vowel interval is used for pitch detection, because pitch period in a voiced consonant and in the following vowel are assumed to be continuous. The ratio of the cepstrum amplitude of maximum peak and that of zero quefrency is used to decide presence or absence

of pitch. As the cepstrum amplitude of zreo quefrency represents the average power, The ratio represents the normalized hieght of the peak in the frame. Therefore, it can be used to judge whether the peak is conspicuous or not.

4.2.2 Voiced/Unvoiced Decision

The decision of voiced or unvoiced is basically performed by the presence or absence of pitch. Pitch extraction, however, often fails in noisy environment. So the voiced/unvoiced decision is performed by considering both the result of vowel-consonant segmentation and the results of pitch extraction.

Vowel interval is unconditionally decided to be voiced, but consonant interval is decided to be voiced only if the half number of its frames are judged to have pitch.

4.3 Formant Extraction

The simplest way to extract formants is to pick up three major peaks in a frame, but the formant extraction algorithm described in this paper is modified to improve accuracy in noisy environment.

4.3.1 Formant Candidate Extraction based on Formant Structure

At the first stage of formant extraction, formant candidates are chosen in a frame using only the information obtained from the frame with the knowledge of formant structure.

Vowel is characterized by the relative position of formants and each vowel have a different formant structure. Fig. 4.1 shows an example of a man's central frequency of the first three formants. Consider frequency bands as shown in fig 4.1 such that each band represents the area where a formant is likely to appear. The first band is the area where the first formant appears. The second and the third band are the area where the second formant of /a/, /o/, /u/ and of /i/ appear, respectively. The fourth and the fifth band is the area where the third formants of /e/, /a/, /o/, /u/ and of /i/ appear, respectively. Let us call the first band as the first formant band, and the second and the third band as the second formant band, and the fourth and the fifth band is the third formant band, respectively, according to the area where each formant appears. We defined these bands as follows based on the results of experiments.

The 1st band	200Hz–1000Hz
The 2nd band	1000Hz–1800Hz
The 3rd band	1700Hz–2600Hz
The 4th band	2400Hz–3300Hz
The 5th band	2900Hz–3600Hz

There are some rules or restrictions between the peaks extracted as the formant candidate in each band. For example, the third formant appears in the fourth band if the second formant appears in the second band, and the third formant appears in either the forth or fifth band if the second formant appears in the third band.

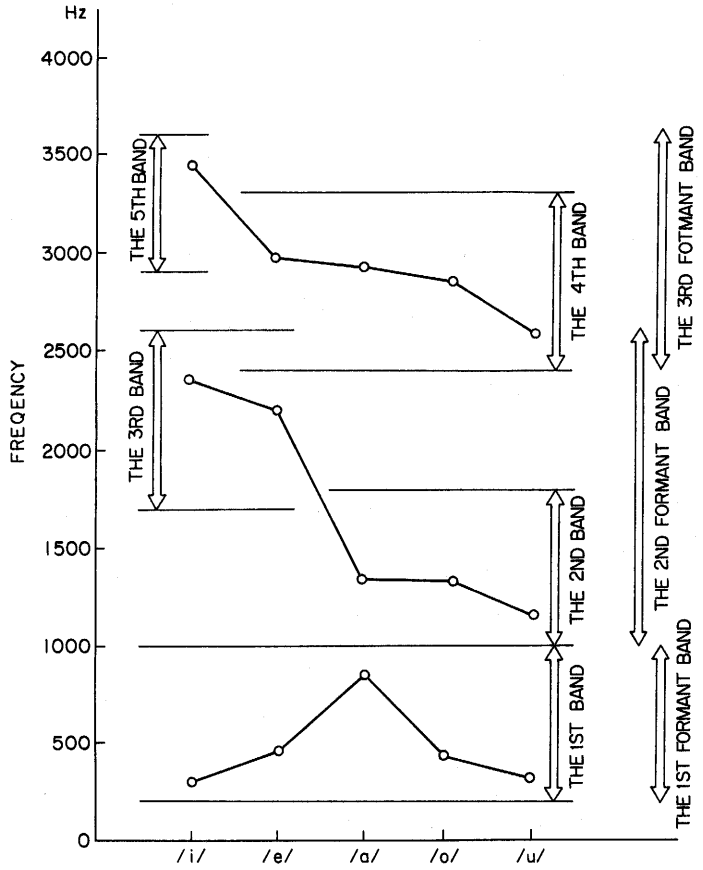


Fig. 4.1. Formant frequency of five vowels and formant bands

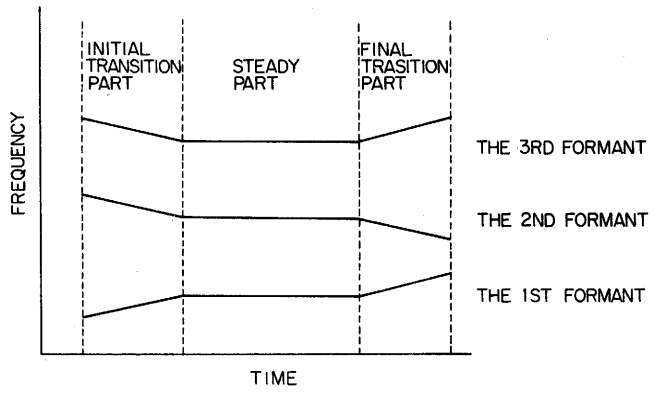


Fig. 4.2. Simple model of formant locus

These formant bands are established to avoid choosing spurious peaks as formant candidates. In the extraction of formant candidates, a maximum spectrum peak is chosen in a band, and three peaks are chosen from these five peaks according to the restriction between the bands.

4.3.2 Simple Model of Formant Locus

A simple model of formant locus in a vowel interval is established as shown in fig 4.2 to represent the outline of the formant transition.

A vowel interval is divided into three parts: transition from the preceding phoneme, steady-state of the vowel, and transition to the following phoneme. Let us call them initial transition prat, steady part, and final transition part, respectively. Formant locus in steady part is approximated by a horizontal line, which means the formant frequency is fixed in this part. The formant locus in initial and final transition part is approximated by a inclined line, which means the formant locus is approximated by a first order polynomial, namely, either upward or downward transition. These line segments must be connected according to the continuity of formants.

4.3.3 Outline Extraction

The second stage of formant extraction is to extract the outline of formant locus by applying the model described in sec 4.3.2 to the results of formant candidate extraction.

First, a frame n where its power is the maximum is searched in a vowel interval. The preceding frame, this frame, and the following frame are treated as the steady part. The mean value of the i-th formant candidate in the steady part is computed and the outline of the i-th formant is fixed to this value.

$$m_i=\frac{1}{3}\sum_{j=n-1}^{n+1}F_i(j) \tag{4.3}$$

$$L_i=m_i \quad (i=1, 2, 3) \tag{4.4}$$

where m_i is the mean value, and L_i is the outline of the i-th formant.

Secondly, the frames from the beginning of the vowel interval to the previous frame of the steady part are treated as the initial transition part. The mean value of the i-th formant candidate in the initial transition part is computed and the outline of the i-th formant is defined as a line connecting between the two points: $(n/2, m_i)$ and $(n-1, F_i(n-1))$.

$$m_i=\frac{1}{n-1}\sum_{j=1}^{n-1}F_i(j) \tag{4.5}$$

$$L_i=F_i+\frac{2}{n}(F_i(n-1)-m_i)(i-n+1) \quad (i=1, 2, 3) \tag{4.6}$$

where m_i is the mean value and L_i is the outline of the i-th formant.

The outline of formants in the final transition part is defined the same manner as the initial transition part.

4.3.4 Search Area for Reinvestigation and Decision of Formant

At the last stage of the formant extraction, the search area is defined in the neighborhood of the outline of formant locus in order to look up again in this area. If there is a peak which is large enough to be a formant, it is judged to be a for-

mant.

The search area is settled both sides of the outline whose width equals the standard deviation of the i-th formant candidate.

$$I_i=[L_i-\sigma_i, L_i+\sigma_i] \quad (i=1, 2, 3) \tag{4.7}$$

Where I_i is search area, L_i is the outline of formant locus, and σ_i is the standard deviation of the i-th formant.

The standard deviation of the formant candidate is used as the width of the search area. The reason for this is to loosen the restriction of the outline of formant locus in proportion to the reliability of the formant candidate.

5. Evaluation Experiments and Results

An ideal method for evaluating formant analyzer is to reconstruct speech from the extracted parameters by formant synthesizer. The characteristics of the formant synthesizer, however, is not fully known. For this reason, two kinds of evaluation test are done.

The first test is based upon comparing the output of formant analyzer to formant data derived by human from sound spectrograms of the input speech. The second test is based upon comparing the monosyllabic intelligibility of outputs of the formant synthesizer to that of the input speech.

5.1 Formant Extraction Error in Noisy Environment

5.1.1 Test Procedure and Scoring the Data[3)]

The accuracy of formant extraction by the analyzer was evaluated in case of original speech and noise superimposed speech. 50 bisyllabic nonsense words which is the combination of 100 monosyllables appearing in Japanese was constructed for speech material. Nonsense words were used to minimize nonuniform stress patterns resulting from speakers varying familiarity to uttering words. Furthermore, noise superimposed speech was made by superimposing gaussian white noise on the 50 bisyllabic nonsense words. The noise level was 3.0dB.

Both materials were fed into formant analyzer and the outputs were compared with the data obtained by interpreting the sound spectrograms. Error-count criteria was established for measuring the discrepancies between the two outputs. It represents the approximate limits of formant accuracy for correct identification of vowel. The criteria we used as follows.

Tolerance of the 1st formant.....150Hz
Tolerance of the 2nd formant.....200Hz
Tolerance of the 3rd formant.....300Hz

If an output of formant analyzer differed from the formant of sound spectrogram for the specified tolerance, an error was counted.

5.1.2 Results

The results of error measurement is shown in table 5.1(a). The number of

Table 5.1. Formant extraction error rate

(a) Our extraction method

	Extraction error rate of the 1st formant			Extraction error rate of the 2nd formant			Extraction error rate of the 3nd formant		
	first	steady	final	first	steady	final	first	steady	final
Original	4.0%	1.3%	2.0%	7.0%	4.3%	9.0%	3.0%	6.0%	8.0%
Original +Noise	2.0%	2.2%	5.0%	37.0%	29.2%	36.0%	18.0%	21.9%	34.0%

(b) Conventionl extraction method

	Extraction error rate of the 1st formant			Extraction error rate of the 2nd formant			Extraction error rate of the 3nd formant		
	first	steady	final	first	steady	final	first	steady	final
Original	3.0%	1.3%	2.0%	12.0%	7.3%	9.0%	16.0%	15.5%	14.0%
Original +Noise	2.0%	1.3%	4.0%	42.0%	28.8%	23.0%	43.0%	51.9%	52.0%

errors with regard to each formant and time segment were summed and expressed as percentages of the total number of the frames. For comparison, the results in the case of cenventional method is shown in table 5.1(b). The previous method in this usage means the one which does not use the global information such as the outline of formant locus.

The results show that the extraction error rate of the second and third formant is high particularly in the presence of noise in the conventional method, whereas those in our method is comparatively low. Moreover, the extraction rate of the initial and final formant transition in vowel interval is considerably improved. These facts prove that the method employing global information and making reinvestigation plan has the effect of improving the reliabiilty and accuracy of parameter extraction in noisy environment. The extraction error rate, however, is still high enough to make the proper identification of synthesized speech difficult as described in the following section. More effort has to be made to improve the quality of the synthesized speech.

5.2 Monosyllabic Intelligibility of the Noise Reduced Speech

5.2.1 Test Procedure and Scoring the Data[4)]

The monosyllabic intelligibility of the synthesized speech was evaluated as to noise superimposed speech and noise reduced speech. The results indicates the ability of noise reduction in this system. Discrete monosyllabic utterances of five vowels /a/, /i/, /u/, /e/, and /o/ were used for basic speech material. Noise superimposed speech was made by superimposing gaussian white noise on this material. Noise superimposed speech was fed into formant analysis-synthesis system and the output was used for noise reduced speech. The noise level was varied from –3dB

Table 5.2. Monosyllabic intelligibility of noise superimposed speech and noise reduced speech

S/N	Noise superimposed speech	Noise reduced speech
+3dB	84%	62%
0dB	72%	32%
−3dB	68%	26%

to 3dB in S/N ratio.

Both materials were presented in random order to five listeners for identification. Each utterance was repeated four times and the listeners were instructed to answer a vowel to the phoneme presented.

5.2.2 Results

The results of monosyllabic intelligibility test is shown in table 5.2. The number of correct responses with regard to noise level and speech material is expressed as percentges of the number of total presented vowel. The data of table 5.2 show that the intelligibility of noise reduced speech is lower than that of noise superimposed speech. This tendency is particularly true when noise level is high.

Although the intelligibility of noise reduced speech is lowered, the easiness of hearing is improved as the back ground noise is reduced. Furthermore, it is recognized that the naturalness of noise reduced speech is improved compared with noise superimposed speech. This is probably because the speech-like structure of spectrum envelope is restored by the formant analysis-synthesis method which is based on the speech generation model.

6. Conclusion

The technique and procedure for noise reduction by formant analysis-synthesis in the presence of noise is established and described in this paper. The subject of this study is to improve the accuracy of pitch and formant extracted from noisy speech. Global information of formant transition is used in the form of the outline of formant locus, and as local information, the knowledge about the relative location of formant is used in the selection of formant candidate.

The evaluation experiments show that the accuracy of the second and the third formant and that of transition interval are improved by this method. But the intelligibility of noise reduced speech is still lower than that of noise imposed speech. However, easiness of hearing and naturalness of noise reduced speech is improved compared with noise superimposed speech.

At the present stage, no other parameter than pitch and formants are used for the synthesis of speech. These parameters are based on the speech generation model of vowels. Therefore, parameters and methods considering consonants must be developed to improve the quality of the synthesized speech.

REFERENCES

1) J. L. Flanagan: "Automatic extraction of formant frequencies from continuous speech", JASA, Vol. 28, No. 1, Jan. 1956.
2) P. H. Klatt: "Software for a cascade/parallel formant synthesizer", JASA, 67(3), Mar. 1980.
3) J. L. Flanagan: "Evaluation of two formant extracting devices", JASA, Vol. 28, No. 1, Jan. 1956.
4) J. L. Flanagan: "Development and testing of a formant-coding speech compression system", JASA, Vol. 28, No. 6, Jan. 1956.

(Aug. 31, 1985, received)