STUDIA PHONOLOGICA XVIII (1984)

Environmental Acoustic Noise Cancelling based on Formant Enhancement

Tooru HASEGAWA, Yasuo ARIKI and Toshiyuki SAKAI

SUMMARY

Non-parametric noise cancelling methods such as spectral subtraction sometimes destroy the formant structure of speech after noise cancelling. To solve this problem, we developed a system which can reduce the noise and also increase the intelligibility by enhancing the formant structure of speech. The first, second and third formants are extracted by cepstrum analysis from the white-noisy speech. Speech spectrum is estimated on the basis of these extracted formants. By this method, 6.4% increasement of vowel intelligibility is obtained.

1. INTRODUCTION

Acoustic noise cancelling methods may be classified into two groups. One is a parametric method which estimates parameters based on a speech production model. The other is a non-parametric method which cancels acoustic noise by using speech property.

As the parametric method, α parameter estimation in the AR (Auto Regressive) or ARMA (Auto Regressive Moving Average) process can be mentioned[1),2)]. Applicability of these methods, however, is restricted to noises which can be modeled by the parameters.

The non-parametric methods may be classified into two groups. One is based on the speech periodicity as comb filtering[3)]. The other is based on non-correlation[4)] between speech and noise as spectral subtraction[5)]. In the comb filtering, noise reduction accuracy is problem because the speech periodicity is disturbed by noise. In the spectral subtraction, formant structure or smooth spectral envelope is sometimes destroyed by subtracting noise so that intelligibility is often decreased.

Conventional researches based on non-parametric method seldom keep the formant structure of speech or enhance it which is important cue for perception[6)]. In this study, we propose the speech spectral estimation and enhancement by keeping the formants based on the non-parametric method[7)].

To extract the formants accurately, acoustic analysis is performed on the basis

Tooru HASEGAWA (長谷川享), 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.

of the cepstrum analysis which can separate the excitation and vocal tract in the frequency domain. At present, white noise is under consideration.

2. WHITE NOISE CANCELLATION BY CEPSTRUM ANALYSIS-SYNTHESIS

2.1 Noise Cancellation by Formant Enhancement

Speech spectrum consists of spectral envelope presenting the resonance response at vocal tract and spectral fine structure (fundamental frequency and its harmonics) presenting the excitation response. White noise makes the peak-valley structure of the speech unclear in the spectral envelope and disturbes the periodicity in the spectral fine structure.

The aim of this study is to increase the speech intelligibility by enhancing the formant information extracted from spectral envelope. Spectral fine structure is also enhanced by clearing the fundamental frequency.

2.2 Cepstrum Analysis and Noise Cancelling System

Cepstrum analysis linearly separates the excitation response, vocal tract response and radiation response on the spectrum and presents the speech as a function of time (quefrency). Due to this linear separation, low quefrency corresponds to the spectral envelope as the vocal tract response and high quefrency corresponds to the

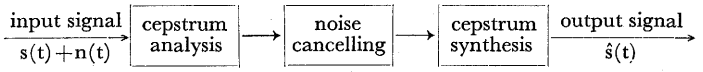

Fig. 1. Noise cancelling system.

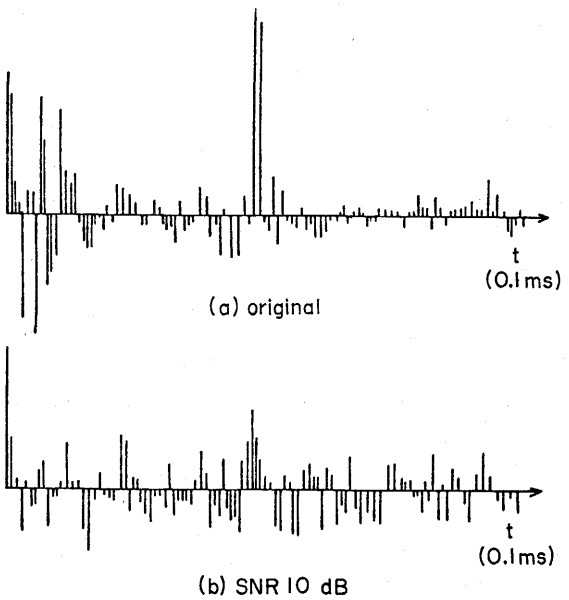

Fig. 2. Cepstrum of speech /a/.

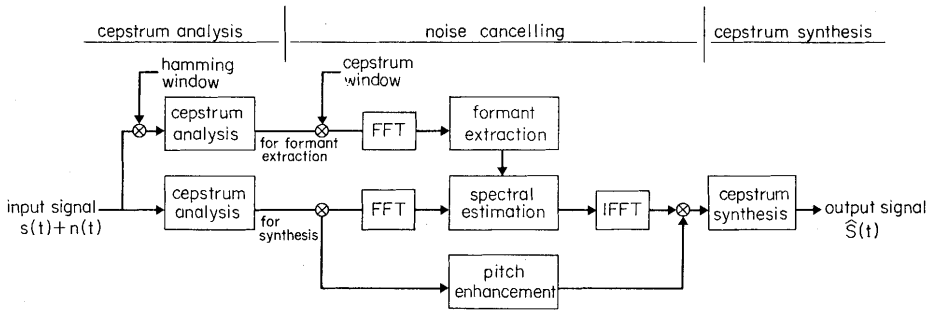

Fig. 3. Detailed block diagram of noise cancelling system.

spectral fine structure as the excitation response.

To cancel the white noise, a schema shown in Fig. 1 is employed. It consists of three parts: namely, cepstrum analysis, noise cancelling, cepstrum synthesis block. In the noise cancelling block, speech spectral envelope is estimated from low quefrency and fundamental frequency is enhanced on the high quefrency. In the cepstrum synthesis block, speech is produced from processed cepstrum in the noise cancelling block.

Fig. 2(a) shows an example of the cepstrum of an original speech /a/. Acoustic analysis is performed with 10 kHz sampling frequency, 10 bits accuracy, 25.6 ms frame length. Fig. 2(b) shows the cepstrum of the same speech /a/ with white noise of 10 dB SNR (Signal to Noise Ratio). The purpose in the noise cancelling block is to obtaine the cepstrum like Fig. 2(a) from the observed cepstrum like Fig. 2(b).

Fig. 3 is the detailed organization of the noise cancelling schema of Fig. 1. In the noise cancelling block, formant information is extracted from low quefrency by Fast Fourier Transformation (FFT) and noise cancelled speech spectrum is estimated with keeping the formant structure. Low quefrency after noise cancelling is obtained by Inverse Fast Fourier Transformation (IFFT) of the estimated spectrum. Two types of cepstrum are produced for the noise cancelling. One is for formant extraction with hamming window. The other is for spectral estimation and cepstrum synthesis with no hamming window. As for high quefrency, the quefrency around the pitch period is enhanced and the other quefrency is suppressed. From these obtained low and high quefrency, estimated speech is obtained by cepstrum analysis.

3. Formant Extraction

3.1 Formant Extraction Method

Three typical methods may be mentioned as the formant extraction techniques. They are A-b-S methods, moment method, local peak picking method.

A-b-S method determines the i-th formant frequency F_i in the manner of sequential approximation with spectral envelope model of speech. The problem of

the method is the time consumption due to sequential approximation of the F_i.

Moment method determines the formant frequency as the ratio of the first order moment to the zero-th order moment in the each possible frequency range. This method has the problem in the accuracy of the formant extraction when white noise is superimposed on speech because spectral slope is flattened by white noise.

Peak picking method determines the formant frequency as the local peak on the spectral power envelope. In this method, formant decision rule is required to determine the true formants among the local peaks especially when white noise is superimposed. Human auditory system is sensitive to the local peaks on the spectrum so that this local peak method seems to be prospective for the formant extraction from the noisy speech.

3.2 Formant Extraction by Peak Picking Method

We extracted formants according to the schema depicted in Fig. 3 by directly applying the peak picking method to noisy speech. Hamming window size is 25.6 ms and cepstrum window size (lifter) is the following.

$$h_m=\begin{cases} 1 & m=0\sim19 \\ 1+\cos\{2\pi(m+12)/128\} & m=20\sim39 \\ 0 & m=40\sim128 \end{cases} \quad (1)$$

Speech data used for experiment were 67 mono-syllables spoken by two males. White noise superimposed on the speech data is produced on a computer by gaussian random number. SNR was set as the averaged power on the steady part of the each mono-syllable to the averaged white noise.

The results of formant extraction at -3 dB and 0 dB of SNR are shown in Table 1 for each vowel. Extraction rate in the Table is the ratio of the truly extracted formants to the true formants. Error rate means the ratio of the frequency difference between the extracted formant and true formant to the true formant.

Table 1. Formant extraction rate

	(%)	−3dB			0dB		
		F_1	F_2	F_3	F_1	F_2	F_3
/a/	extraction rate	67.0	67.0	44.0	78.1	71.6	48.3
	error rate	3.1	3.4	3.7	2.9	2.9	3.1
/i/	extraction rate	67.9	41.1	74.3	73.0	68.5	78.9
	error rate	7.7	3.32	3.27	5.6	2.7	2.4
/u/	extraction rate	97.1	51.1	41.3	99.0	60.5	43.1
	error rate	6.8	5.5	5.1	5.8	5.1	5.0
/e/	extraction rate	86.8	56.4	57.8	95.2	69.2	62.3
	error rate	4.9	2.9	3.4	4.3	2.5	2.67
/o/	extraction rate	89.4	60.6	40.3	94.3	73.1	48.0
	error rate	4.8	4.5	5.1	4.0	3.3	5.1
average	extraction rate	81.6	55.2	51.4	88.0	68.6	56.1

error rate $\frac{\Delta F_i}{F_i}$ (%)
ΔF···error frequency

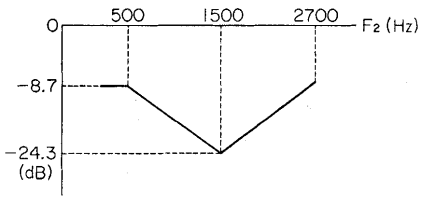

Fig. 4. Relation between the first formant and second formant.
(ratio of the amplitude of the second formant to that of the first formant)

Table 1 shows that the second formant extraction rate decreases when SNR decreases. Formant extraction rate is also low for the vowels /i/ and /e/. The reason of above low extraction rate is as follows.

(1) When SNR decreases, many false peaks are observed on the second formant frequency range 900–2500 Hz.

(2) White noise flattens the power spectrum so that the rule used in this formant extraction algorithm to verify the first and second formant like Fig. 4 does not become applicable.

To solve this problem spectral slope adjustment is required.

3.3 Peak Picking Method Using Spectral Slope Adjustment

Spectral slope is obtained as the slope of the following line Y(i) which has the minimum square error to the power spectral envelope S(i).

$$Y(i)=A*i+B \qquad (i=0,\ldots,128) \tag{2}$$

Spectral slope adjustment is done by subtracting the Y(i) from S(i), namely adjusted spectral slope Z(i) becomes as follows.

$$Z(i)=S(i)-Y(i) \qquad (i=0,\ldots,128) \tag{3}$$

First formant dominates extremely so that spectral slope adjustment is performed after first formant extraction. The process to extract the formants is as follows.

(1) First formant extraction
(2) Spectral slope adjustment
(3) The second formant extraction
(4) The third formant extraction

Detailed algorithm is the modification of the formant extraction algorithm[8] in the previous section. Table 2 shows the results of formant extraction by spectral slope adjustment to the same speech data as Table 1. Compared to the Table 1, the extraction rate of the first formant is almost same, however, the second and third formant extraction rate increases for vowels /a/ and /e/. Table 3 shows the increasement of the formant extraction rate for each vowel. Extraction rate of the first, second and third formant at −3 dB SNR was 84.8%, 65.4%, 63.3% respectively on average. Extraction rate increased due to spectral slope adjustment, by 3.2%, 10.2%, 11.9% to each formants.

Table 2. Formant extraction by spectral slope adjustment

	(%)	−3dB			0dB		
		F_1	F_2	F_3	F_1	F_2	F_3
/a/	extraction rate	83.0	91.4	63.8	91.4	97.1	72.3
	error rate	3.3	3.5	3.7	3.0	2.9	3.3
/i/	extraction rate	67.9	50.0	73.1	73.0	70.4	82.1
	error rate	7.6	3.26	3.1	5.5	2.7	2.4
/u/	extraction rate	99.0	50.4	50.4	99.0	62.1	50.4
	error rate	6.7	5.8	5.1	5.8	5.1	5.2
/e/	extraction rate	86.8	73.7	71.9	95.6	93.1	83.3
	error rate	4.9	2.7	3.4	4.2	2.4	2.9
/o/	extraction rate	87.5	61.3	57.3	94.3	72.1	52.0
	error rate	4.7	4.7	4.89	3.9	3.5	4.7
average	extraction rate	84.8	65.4	63.3	90.6	79.0	68.0

Table 3. Increasement of the formant extraction rate

(%)	−3dB			0dB		
	F_1	F_2	F_3	F_1	F_2	F_3
/a/	16.0	24.4	19.8	13.3	25.5	24.0
/i/	0.0	8.9	−1.2	0.0	1.9	3.2
/u/	1.9	−0.7	9.1	0.0	1.6	7.3
/e/	0.0	17.3	14.1	0.4	23.9	21.0
/o/	−1.9	0.7	17.0	0.0	−1.0	4.0
average	3.2	10.2	11.9	2.6	10.4	11.9

4. NOISE CANCELLING

Speech spectrum is estimated based on the first, second and third formants extracted from noisy speech.

4.1 Speech Spectrum Estimation based on Formant Information

Speech spectrum estimation consists of two blocks. The first one is the estimation within the formant bandwidth. In this frequency range, spectral slope is enhanced to remove the flattening by white noise. The second is the estimation in the frequency range between two formant bandwidth. In this frequency range, spectral envelope is approximated by smoothly connected second order curve.

(1) Spectrum estimation within the formant bandwidth

As shown in Fig. 5, flattend spectral slope A is enhanced to $\bar{A}$ which is the averaged spectrum slope of the speech according to the following expression.

$$Z(i)=S(i)-(A-\bar{A})*F_j/FT \quad (j=1,2,3) \qquad (4)$$
$$[(F_j-100)/FT]\leq i\leq[(F_j+100)/FT]$$

S(i): observed spectrum

Z(i): estimated spectrum

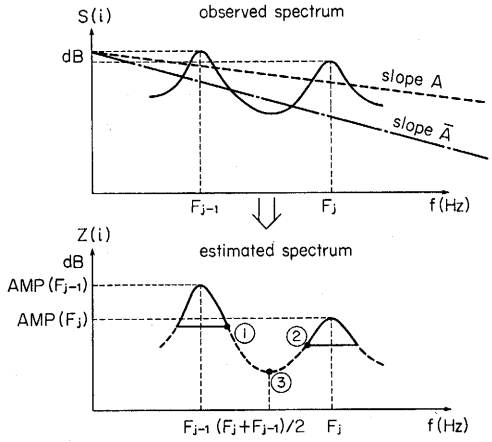

Fig. 5. Estimation of speech spectrum.

$\bar{A}$: averaged spectral slope of speech
$(=-0.27$ dB/FT)

A: observed spectral slope

FT: frequency interval on the spectrum

F_j: extracted formant frequency

(2) Spectrum estimation between the formant bandwidth

As shown in Fig. 5, spectrum is estimated between the formant F_{j-1} and F_j using the following second order curve expression.

$$Z(i)=B_j*i^2+C_j*i+D_j \quad (j=1,2,3) \qquad (5)$$
$$[(F_{j-1}+100)/FT] \leq i \leq [(F_j-100)/FT]$$

This curve passes the following three points as shown in Fig. 5.

(1) The high frequency point in the F_{j-1} formants bandwidth

(2) The Low frequency point in the F_j formant bandwidth

(3) Middle frequency point between F_i and F_{i-1} with the amplitude of $\alpha*$ $\{AMP(F_{j-1})+AMP(F_j)\}$. Constant α is set to 0.6–0.8 by examining the natural speech spectrum.

The spectrum above the third formant remains unprocessed.

4.2 Pitch Enhancement

The highest peak within the 4 ms—10 ms on the cepstrum is determined as the pitch period. Pitch enhancement is carried out by the following expression. Where C_i is the cepstrum, C_p is the cepstrum coefficient which corresponds to pitch period.

$$\begin{aligned} C_i &= (1+\beta)*C_i \quad (i=p-1,\ p,\ p+1) \\ &\gamma*C_i \quad (i \neq p-1,\ p,\ p+1) \\ &\beta>0,\ 1>\gamma>0 \end{aligned} \qquad (6)$$

Cepstrum coefficient corresponding to pitch period and both side cepstrum coefficients are enhanced by $1+\beta$. The remaining cepstrum coefficient is suppressed

by $\gamma<1$. The parameters β, γ are set to 0.6 and 0.7 respectively by experiment.

5. EXPERIMENT AND RESULT

Subjective SNR improvement and mono-syllable intelligibility test was carried out to estimate the noise cancelling method.

5.1 Subjective SNR Improvement Test

Speech data is sentence with 5 second duration spoken by one male. Noise cancelling was carried out for the −1 dB SNR speech with white noise. This processed speech data was compared with twelve same speech data with different SNR from −3 dB to 8 dB by 1 dB change. Fourteen males judged the easiness to listen by hearing the 24 pairs of processed and unprocessed speech (12 pairs and their exchanged pair). Table 4 shows the results and about 1–2 dB improvement is shown by noise cancelling.

5.2 Mono-syllable Intelligibility Test

White noise with 0 dB SNR was superimposed on the 67 mono-syllable spoken by one male. Ten males judged mono-syllables by hearing, three times, 134 mono-syllables including 67 processed and 67 unprocessed mono-syllables. Table 5 shows the results of intelligibility for each vowel and consonant.

From Table 5, intelligibility increased by 6.4% for vowel and 1.7% for consonant and 3% in total. This means that this method is useful for the vowel which shows the definit formant structure.

5.3 Effectness of Pitch enhancement

Mono-syllable intelligibility test is carried out after pitch enhancement for the speech with 0 dB SNR white noise. Table 6 shows the result of intelligibility for vowels and consonants. Intelligibility increased by 13.4% for the vowels, however,

Table 4. Subjective SNR improvement

SNR −1 dB 14 males × twice

subjective SNR improvement (dB)	preference of processing	same	preference of unprocessing
−2	18	5	5
−1	18	6	4
0	9	12	7
1	12	10	6
2	7	6	15
3	7	6	15
4	3	7	18
5	6	4	18
6	3	7	18
7	4	5	19
8	1	4	23
9	1	2	25

Table 5. Intelligibility

vowel SNR 0 dB (%)

	/a/	/i/	/u/	/e/	/o/	total
before	88	44	68	56	24	56
after	92	32	84	72	32	62.4

consonant

	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/	total	vowel+consonant total
before	23	10	16	46	10	53	61.6	10	85	21	18	30	15	22	30.0	36.8
after	25	5	19	42	6	49	68.3	16	85	16	16	50	23	23	31.7	39.8

Table 6. Intelligibility by pitch enhancement

vowel SNR 0 dB (%)

pitch enhancement	/a/	/i/	/u/	/e/	/0/	total
not	92	16	72	52	12	48.8
done	92	32	84	72	32	62.4

⋮

pitch enhancement	/k/	/s/	/t/	/h/	/p/	total
not	17	5	28	10	37	19.4
done	25	5	19	6	23	15.6

⋮

pitch enhancement	/n/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	total
not	44	55	76.7	14	75	16	15	50	19	40.5
done	42	49	68.3	16	85	16	16	50	23	40.6

decreased by 3.8% for the unvoiced consonants and same for voiced consonant. This means the pitch enhancement is useful for the voiced sound.

6. Concluding Remarks

We described the method to cancel the white noise from the noisy speech by estimating the speech spectrum based on the extracted formant information. The features of this method are as follows.

(1) Cepstrum analysis and synthesis can separate vocal responce and excitation response so that not only noise cancelling on the spectrum but also the pitch enchancement becomes possible.

(2) The method to estimate the speech spectrum based on the extracted formants has the good association with the human auditory system.

(3) Spectral slope adjustment can remove the slope variation so that colored

noise also can be cancelled.

The future works remained are as follows.

(1) Spectral estimation between formant bandwidth should be improved in stead of second order curve to keep the speech naturality.

(2) This method is useful for voiced sound, but not useful for the unvoiced sound so that another noise cancelling method for unvoiced sound is required.

REFERENCES

1) K. Ozeki: Maximum Likelihood Estimation of Predictor Coefficients from Noisy Speech Signals, Technical Report of Professional Group on Speech of ASJ, S79–44, 1979.
2) H. Morikawa and H. Fujisaki: Spectral Estimation of Speech in Noisy Environments, Technical Report of Professional Group on Speech of ASJ, S81–13, 1981.
3) H. Nagabuchi: Suppression of Random Noise in Speech based on the Comb Filtering Technique, Report of Professional Group on Speech of ASJ, S79–48, 1979.
4) T. Takasugi, J. Suzuki and R. Tanaka: Function of SPAC (Speech Processing System by use of AutoCorrelation function) and Fundamental Characteristics, Trans. IECE Japan, Vol. J62–A, No. 3, pp. 175–182, 1979.
5) S. F. Boll: Suppression of Acoustic Noise in Speech Using Spectral Subtraction, IEEE Trans., ASSP-27, 2, April 1979.
6) J. S. Lim and A. V. Oppenheim: Enhancement and Bandwidth Compression of Noisy Speech, Proc. IEEE, 67, 12, 1979.
7) Y. Ariki and T. Sakai: Cancelling of Environmental Acoustic Noise Superimposed on Speech, Report of Professional Group on Speech of ASJ, S84–09, 1984.
8) Schafer and Rabiner: System for Automatic Analysis for Voiced Speech JASA, Vol. 47, No. 2, Feb., 1970.

(Aug. 31, 1984, received)