다단계 보상 기능을 갖는 통계적 방법에 의한 음소 분할

A Statistical Approach to Phoneme Segmentation through Multi-step Compensation

김 홍 국* 이 황 수* 은 종 관*

(Hong Kook Kim, Hwang-Soo Lee, Chong Kwan Un)

요 약

본 논문에서는 통계적방법에 의한 음소의 자동분할에 관한 알고리즘을 제안하였다.

우선 음성신호를 AR모델로 모델링한 후 스펙트럼이 변화하기 전과 변화한 후의 모델에 대해서 likelihood ratio와 mutual information을 고려한 test statistics로 부터 모델 계수가 변화하는 곳을 예측해 내고 이 곳을 음소의 경계로 판단한다. 이 경우 검파되지 못하는 대부분의 음소는 짧은 자음이었으며 Signed Front-to-Back maximum area Ratio(SFBR)을 이용하여 개선하였다. 또한 false alarm error를 줄이기 위해 두 segment 사이의 distortion으로 부터 smoothing을 하였다.

3명의 화자에 대한 실험 결과 non-detection error는 10%, false alarm error는 20% 정도로 나타났지만 화자간에 알고리즘의 성능 변화가 거의 없으며 특히 분할된 경계치 분포는 전체 음소의 90% 이상이 이 30ms이내에 위치하였다.

ABSTRACT

A statistical approach to automatic phoneme segmentation is presented in this paper. The proposed segmentation algorithm is an extension of the divergence test using the test statistics by which we can detect abrupt changes in speech signal that are considered as phonemic boundaries. In order to reduce the errors in phoneme boundary detection, some compensation techniques, such as turbulence noise detection and smoothing using a distortion measure, are incorporated in the proposed segementation algorithm, thus resulting in reduction of non-detection and false alarm error.

Computer simulation is done to test the performance of the proposed algorithm for speaker independent speech recognition. Error rates of about 10% and 20% are obtained for non-detection errors and false alarm errors, respectively.

I. INTRODUCTION

Continuous speech recognition systems can be classified as the two categorites[1]. One begins with subword units and successively combines them to larger lingusitic units. This is called the bottom-up system. The top-down system reverses the process by predicting sentences and successively hyperthesizing phrases, words and phonemes that make up the sentences, and then comparing the predicted patterns with input patterns. In the bottom-up system, the performance of an acoustic-phonetic processor which encodes input speech signal into a string of discrete subword units is critical to the whole performance of the speech recognition sys-

*한국과학기술원 전기및 전자공학과

tem. Several approaches have been proposed to segment input speech into phoneme units and can be classified in two calsses[1]~[2]. The first one performs jointly both labeling and segmentation using the acoustic cues extracted from the input speech. The other one segments the input speech into phonemes prior to the labeling, based on some test statistics or other methods such as using heuristics, knowledge bases, and so on. One useful statistical approach to phoneme segmentation is the divergence test[2]~[3]. This test can detect the abrupt change between two AR models, by examining a distorition between these models. When we simply consider these abrupt changes as phoneme boundaries, there exist many errors in boundar detection, such as omission or oversegmentation.

Our approach to speech segmentation is an extension of the divergence test based on test statisties. Prior to the divergence test, preprocessing of input speech is done by using a coarse vowel/nasal segmentation algorithm. After the divergence test, we employ a postprocessing stage for turbulence noise detection and smoothing with a distortion measure. By incorporating the preprocessing, postprocessing, and smoothing, we can improve the performance of the proposed speech segmentation algorithm by reducing the non-detection and false alarm errors.

Ⅱ. SEGMENTATION BY THE DIVERGENCE TEST

The divergence test assumes an AR model of speech signal when the speech signal can be deseribed by a string of homogencous units. When we assume the parameters of an AR model change abruptly at some unknown time θ, the observed scalar signal $\{y_n\}$ may be represented as

$$y_n=\sum_{i=1}^{p} a_i^{(n)}\, y_{n-i}+e_n \qquad (1)$$
$$\sigma_n^2=var(e_n)$$

where

$$a_i^{(n)}=a_i^0,\ 1\leqq i\leqq p$$
$$\sigma_n^2=\sigma_0^2 \quad \text{for } n<\theta \qquad (2)$$

and

$$a_i^{(n)}=a_i^1,\ 1\leqq i\leqq p$$
$$\sigma_n^2=\sigma_1^2 \quad \text{for } n<\theta \qquad (3)$$

$\{e_n\}$ is the white noise sequence or the innovation process of AR model. The AR parameter vectors of the models 0 and 1 will be denoted by

$$A^j=(a_1^j,\cdots,a_p^j),\ j=0,1 \qquad (4)$$

and the past observations up to n-1 by

$$Y^{n-1}=(y_{n-1},\cdots,y_1) \qquad (5)$$

In this case, if the paramenters of the models as indicated in Fig. 1, are supposed to be known both before and after the model change, the only unknown variable is the time θ of the model change.

In this case, if the paramenters of the models as indicated in Fig. 1, are supposed to be known both before and after the model change, the only unknown variable is the time θ of the model change. In real situation, however, identifcation of AR model parameters, and estimation and detection of the time θ of the model change should be carried out simultaneously. Therfore, the observed signal $\{y_n\}$ is filtered through identified AR filters and changes in the innovation sequences $\{e_n^0\}$ and $\{e_n^1\}$ are used for a proper cumulative sum test to detect the model boundary.

Let the signal $\{y_n\}$ be described by the conditional densities $g^0(y_n|Y^{n-1})$ and $g^1(y_n|Y^{n-1})$ before and after the model change, respectively. And let us

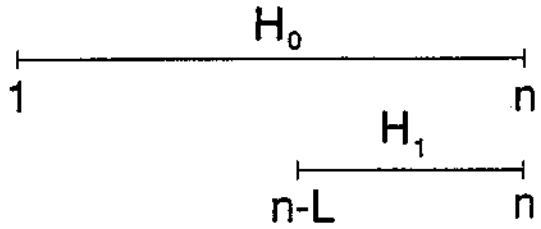

Fig.1. Analysis windows of models H_0 and H_1 for the divergence test. The parameters of the model H_0 are calculated at each sampling instant i=1 to n where n is the present observation time instant. The parameters of the model H_1 are obtained from the observations $\{y_{n-L+1}, \cdots, y_n\}$ by block analysis.

consider the following cumulative sum test by taking into account of not only the mutual entropy between the two models but also their self-entropies[3].

A test statistic U_n is defined as

$$U_n=\sum_{i=1}^{n} T_i \tag{6}$$

where T_i is a test statistic with information and divergence between the two models and defined as

$$T_i=\int g^0(y|Y^{i-1})\log\frac{g^1(y|Y^{i-1})}{g^0(y|Y^{i-1})}\,dy \;-\log\frac{g^1(y_i|Y^{i-1})}{g^0(y_i|Y^{i-1})} \tag{7}$$

In the Gaussian case, T_i can be written as

$$T_i=\frac{1}{2}\left[2\frac{e_i^0 e_i^1}{\sigma_1^2}-(1+\sigma_0^2\div\sigma_1^2)\frac{e_i^{0)2}}{\sigma_0^2}-\left(\frac{\sigma_0^2}{\sigma_1^2}-1\right)\right] \tag{8}$$

Since the conditional drifts of U_n before and after the change are zero and negative, respectively, the time θ of the model change can be detected when U_n becomes negative. In real implementation, θ is detected at $U_n<-\lambda$ as indicated in Fig 2(a), allowing some delay for detection. The delay can be compensated with the aid of Hinkely's stopping rule as shown in Fig. 2(b)[3].

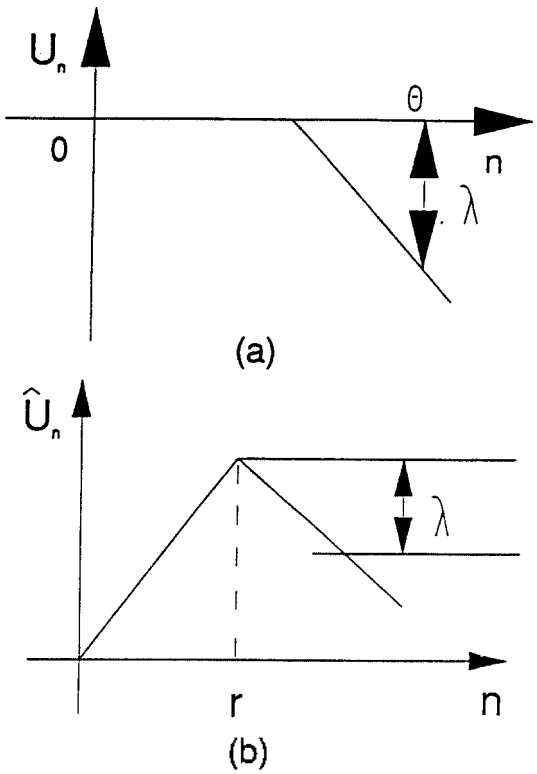

Fig.2. Characteristics of U_n and $\hat{U}_n$. (a) Determination of the estimated time of model change n=θ based on variation of U_n. (a) Detezmination of the estimated time of model change n=r based on variation of $\hat{U}$

1. Divergence Test Algorithm

1) Initialization

The prewindowed recursive least square(PRLS) algorithm is applied to obtain the stable model parameters of H_0 for n=1, ···, L and the model change within the first L points can not be detected.

2) Calculate test statistics T_i and $\hat{U}_n$.

Calculate test statisties $\{T_i\}$ and $\{\hat{U}_n\}$, a cumulative sum with Hinkely's stopping rule deseribed by

$$\hat{U}_n=\sum_{i=1}^{n}(T_i+\delta) \tag{9}$$

where δ is fixed drift determined a prior

3) Detect the time r of the model change.

Choose the time r satisfying the equation(Fig.

2(b))

$$\max \hat{U}_r - \hat{U}_n > \lambda \qquad (10)$$

The parameters of the model H_0 is identifien by the PRLS algorithm. The model H_1 is obtained from speech signal of a fixed window size L by LPC autocorrelation method. The e_i^0 and σ_0^2 are the forward prediction error and cost function divided by n, respectively. The e_i^1 and σ_1^2 are obtained from prediction error and gain divided by L in the LPC analysis, respectively.

II. Performance of The Divergence Test

2.1. Data Base

Among phonetically balanced 100 word vocabularies spoken by 3 male speakers, 40 words are selected for this simple experiment. The input speech is sampled at 10kHz and end-point-detection is done by using zero-crossing rate(ZCR) and energy parameters.

2.2. Performance Criteria

Performances of the divergence test are obtained for false alarm errors, non-dtection errors and boundary alignment errors.

1) False alarm errors occur when a phoneme is segmented into two or more ones.

2) Non-detection errors occur when a phoneme boundary is not detected but really exists.

3) Boundary aligment errors represent the difference between the detected boundary and the actual boundary of the phoneme.

2.3. Experimental Results

In this simple experiment, we use L=200(20ms), δ=0.5, λ=60 and the 16-th order AR model. Fig. 3 and 4 show examples of the segmentation results by using the divergence test. The waveform in Fig. 3(a) shows the word /onl/, the Korean for "today". As shown in Fig. 3(b), both the false alarm error and the non-detection error do not occur in spite of the large boundary alignment error. On the contrary, as shown in Fig. 4(b), corresponding to $\hat{U}_n$ of the word /jəngu/, the Korean for "research", the false alarm error in /n/ and the non-detection error in /g/ occur.

For the vocied plosive /g/, we can observe that $\hat{U}_n$ has more drift, however, the variation,

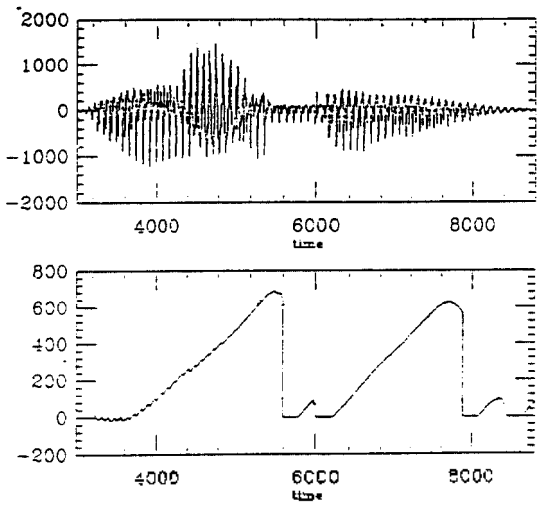

Fig.3. Results of the divergence test on the word /onl/. (a) Speech waveform. (b) Plot of the cumulative sum $\hat{U}_n$.

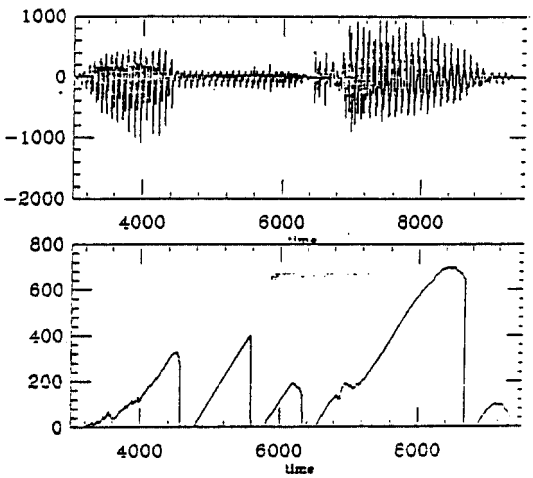

Fig.4. Result of the divergence test on the word /y'ngu/. (a) Speech waveform. (b) Plot of the cumulative sum $\hat{U}_n$.

of $\tilde{U}_n$ is below the threshold λ. The similar phenomena are observed, on the whole, for consonants. This can be compensated with the signed front-to-back max-imum area ratio(SFBR) feature[4]. Also, false alarm error in nasal /n/ can be improved with the aid of smoothing using a distortion measure between adjacent segments.

The first row of Table 1 shows the results of segmentation by only applying the divergence test. The error rate of 10.59% among the total 24.11% error rate for the non-detection error is the system error which can not be recovered because the phonemes, whose length is less than 20ms, can not be detected.

Table 1. Comparison of Divergence test and two Compensations.

Method	No.of Phonemes	Non-detection error	False alarm error
Divergence test	141	34(24.11%)	14(9.92%)
Turbulence noise detection	141	24(17.02%)	20(14.18)
Smoothing	141	24(17.02%)	11(7.80%)

Ⅲ. TURBULENCE NOISE DETECTION

In order to reduce the non-detection error, the SFBR parameter, early proposed by Wakita[4], is introduced. SFBR is defincd as

$$SFBR=sign(k_1)\cdot FBR \tag{11}$$

where $FBR=\frac{\max\{A_1,\cdots,A_p\}}{\max\{A_{p+1},\cdots,A_M\}}$ and A_i(i=0, ···, M+1 : $A_0=\infty$) is the area of the i-th section of the vocal tract in acoustic tube modeling of speech signal and k_1 is the first reflection coefficient of the acoustic tube model.

The back vowels and turbulence noises have large FBRs. The turbulence noise has $k_1>0$ since the frequency characteristics of that represent more energy in the high frequency region than others. For the back vowels, however, $k_1<0$. The flow chart for turbulence noise detection is shown in Fig. 5 . After this compensation, the non-detection error rate is reduced to 17.02% and the false alarm error rate increases to 14.18%. Therefore, it is necessary to adopt a new compensation technique to reduce the false alarm error.

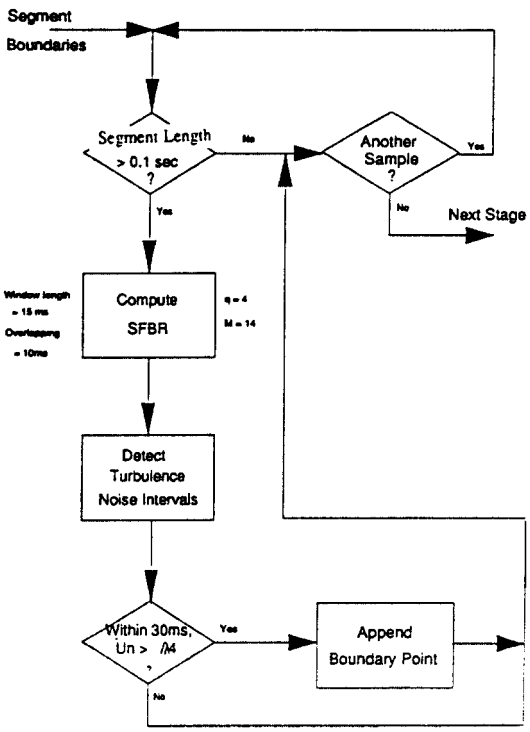

Fig.5. Flow chart of the turbulence noise detection algorithm

Ⅳ. SMOOTHING BASED ON A DISTORTION MEASURE

This smoothing is used to decrease the false alarm errors. A distortion between adjacent segments obtained from so far. The adjacent segments are then combined into one if the distortion between them is below a certain threshold.

In our smoothing method, we select gain-normalized Itakura-Saito distortion measure(d_{IS}) as the distortion measure between speech segments.

However, since d_{N} is asymmetric, it is not suitable for our purose. Therfore, we use $d_{N}=\frac{d_{N1}+d_{N2}}{2}$ where d_{N1} and d_{N2} are calculated in reverse order. When $d_{N}<0.85$ for adjacent segments, these segments are smoothed to make a segment, and then this process is repeated until there are no adjacent segments satisfying $d_{N}<0.85$. Using this smoothing method, the non-detection error rate is not changed, on the other hand, the false alarm error rate is reduced to 7.8%. This result reveals the improvements of 2% and 8% comparing with those of the divergence test and the turbulence noise detection, respectively.

V. COARSE VOWEL / NASAL SEGMENTATION

In the divergence test, the choise of the parameter values of δ and λ is important to improve the performance of phoneme segmentation. It is perferable to choose different values for vowel / nasal segment and others. The vowel / nasal segmentation uses the total energy, ZCR, and four frequency band energies as the feature sets[5]. Here we choose $\delta=0.2$, $\lambda=40$ for vowel / nasal sounds and $\delta=0.8$, $\lambda=80$ for others.

VI. COMIPUTER SIMULATION AND DISCUSSION

A block diagram for our phoneme segmentation method is shown in Fig.6. Computer simulation is done to test the speaker-independence of our algorithm using speech data extracted from 3 male speakers. The segmentation results on this experiment is shown in Table 2. From the simulation results, the two-step compensation can reduce 2 % and 10% of non-detection errors and false alarm errors, respectively. Table 3 shows that, considering the number of short segments, the non-detection error rate is really nothing but 4.96%. The

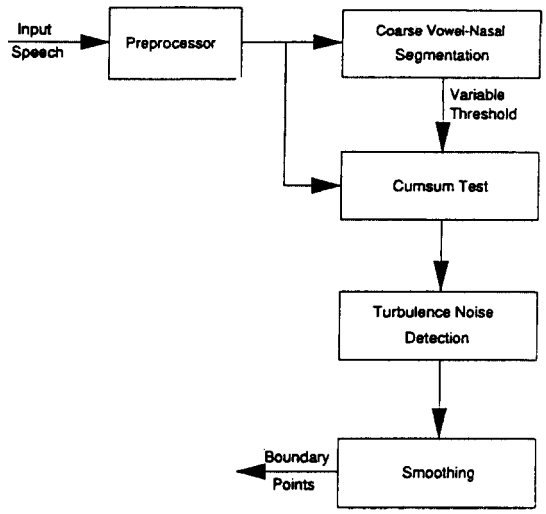

Fig.6. Block diagram of the proposed phoneme segmentation algorithm

Table 2. Results on the phoneme segmentation.

Speaker	No.of phonemes	Divergence test		Turbulence noise detection			
		ND	FA	ND	FA	ND	FA
A	136	15	43	13	45	14	25
B	133	16	38	9	43	10	30
C	134	16	30	15	34	16	24
Total	303	47 (11.66%)	111 (27.54%)	37 (9.18%)	122 (30.27%)	40 (9.93%)	79 (19.6%)

ND : Non-detection error
FA : False alarm error

Table 3. The number of short segments and oversegments in false alarm error.

Speaker	No.of short phonemes	Divergence test	Turbulence noise detection	Smoothing
A	10	104 / 43	113 / 38	56 / 30
B	8	93 / 45	113 / 43	64 / 34
C	9	77 / 25	84 / 30	51 / 24

average number of segments in case of oversegmentation when false alarm errors occur is reduced from 2.47 to 2.16. Finally, we observe the boundary alignment errors. Fig.7(a) shows the histogram

of the segmentation boundary errors versus time (ms) for speaker A. For example, a bar on the time axis -10ms indicates that the probability the time diference of an actual boundary from the segmented boundary is within 10ms, is 0.43. If we allow the difference within±30ms, the correct segmentation rate becomes 90% and the rate becomes 99% as shown in Fig.7(b) if we allow± 50ms difference.

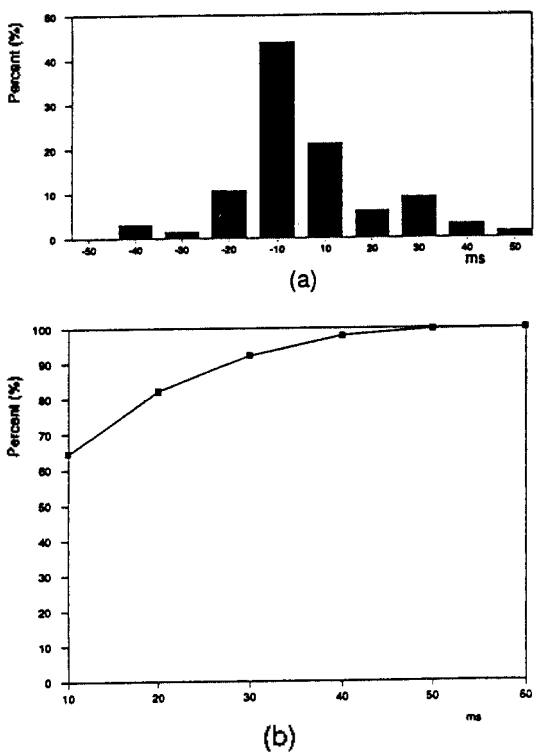

Fig.7. Distribution of the boundary alignment errors. (a) Histogram of the correct segmentation boundaries according to the time of the boundary diference between the actual boundary and segmented boundary for speaker A. (b) Cumulative correction rate of the proposed phoneme segementation system with respect to the time of the boundary diference.

Ⅶ. CONCLUSIONS

In this paper, we propose an automatic phoneme segmentation method based on the divergence test. Two step compensation techniques, turbulence noise detection and smoothing, are used for the postprocessing of the divergence test. Also, a coarse vowel / nasal segmentation algorithm is used for preprocessing to give different thresholds to the divergence test. This effort imporves segmentation results in that non-detection errors and false alarm errors are reduced compared with the segmentation system using the divergence test alone. From the computer simulation results of the proposed segmentation method, the error rates of 10% and 20% are obtained for the non-detection errors and the false alarm errors, respectively. And this experiment reveals that performance improvements of 2% and 10% for the non-detection errors and for the false alarm errors are obtained.

References

1. W.A. Lea, *Trends in Speech Recognition*, Englewood Cliffs, N.J., Prentice-Hall, 1980.
2. R. Andre-Obrecht, "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals", IEEE Trans. on ASSP, Vol. ASSP-36, No. 1, pp.29-40, 1988.
3. M. Basseville and A. Benveniste, "Sequential detection of abrupt changes in spectral characteristics of digital signals", IEEE Trans. on Inform. Theory, Vol. IT-29, pp.709-723, 1983.
4. M. Kasuya and H. Wakita, "An approach to segmenting speech into vowel and nonvowel-like intervals." IEEE Trans. on ASSP, Vol. ASSP-27, No. 4, pp319-327, 1979.
5. F.R. Chen, "Acoustic-Phonetic Constraints in Continuous Speech Recognition : A Case Study Using the Digit Vocabulary", MIT, June, 1985.

▲ Hong Kook Kim was born in Seoul, Korea, on Oct. 24, 1965. He rceeived the B.S. degree in control and instrumentation engineering from Seoul National University, Seoul, in 1988, and the M.S. degree in cleeirical engineering from the Korea Advanced Institute ot Seience and Technologly (KAIST), Seoul, in 1990. Since 1990, he has been with Samsung Electronics as an engineer in the area of speech signal proccss·ing and he is currently working toward the Ph.D. at KAIST. His research interests include speech reeognition, speech coding, speaker recognition and verification, and speech processing using ne·ural networks.

▲이황수 : 한국과학기술원 전기및 전자공학과
(제 6 권 3 호 참조)

▲은종관 : 한국과학기술원 전기및 전자공학과
(제 9 권 4 호 참조)