論文 89-26-4-20

A Segmentation Algorithm of the Connected Word Speech by Statistical Method

(統計的인 方法에 依한 連結音의 音素分割 알고리듬)

曺 政 鎬*, 洪 再 根*, 金 秀 重*

(Jeong Ho Cho, Jae Keun Hong and Soo Joong Kim)

要 約

本 論文에서는 音聲信號의 音素分割을 위한 統計的인 方法을 研究하였다. 이 方法은 3個의 AR 모델을 使用하며, 이 中 2個의 모델은 音聲의 스펙트럼 變化前 및 變化後의 安定된 部分에서 求해지고 이들 間의 距離가 커지면 音素가 바뀐 것으로 간주된다. 다른 한 모델은 두 固定 모델 사이에 位置하며 音素間의 境界를 推定하는데 使用된다. 이 音素分割 알고리듬을 連結音에 對해 試驗해 본 結果, 從來의 方法에 비해 音素의 境界點을 좀더 正確히 찾을 수 있고, 또한 過度分割 誤謬도 줄일 수 있었다.

Abstract

A statistical approach for the segmentation of speech signals is described in this paper. The main idea of this algorithm is the use of three AR models. Two fixed models are identified at the stationary parts of the signal before and after the spectral change. Changes are detected when the distance between these two models is high. Another model is located between two fixed models and is used to estimate spectral change time. This segmentation algorithm has been tested with connected words and compared to classical methods. The results showed that it can provide more accurate locations of boundaries of segments and can reduce the amount of oversegmentation.

I. Introduction

Connected word recognition is one of the most interesting problems at the present stage of speech recognition research since isolated word recognition techniques have been improved up to a practically high level. There have been two kinds of approaches to recognize the connected words. The first approach is a segmentation-free method, which presumes that all frames might be word boundaries. The basic strategy employed is the technique of Dynamic Time Warping (DTW). A set of concatenated reference patterns is matched to the unknown word string. The concatenated word which yields the best fit determines the

*正會員, 慶北大學校 電子工學科
(Dept. of Elec. Eng., Kyungpook Nat'l Univ.)
接受日字 : 1988年 11月 21日

recognition result and implicitly the word boundaries [1]-[4]. This approach has shown satisfactory results in the case of relatively small and limited vocabulary. But there are still many unsolved problems. Both the processing cost and massive storage make this approach unfeasible for large vocabulary speech recognition applications.

A completely different approach which gives some solutions to the above problems is to segment the word string explicitly before recognition of segmented units is performed [5]-[8]. Because of the size of the vocabulary, basic recognition units are usually phonetic in nature, such as diphones, phones, allophones. The key problem in this approach is the segmentation of word strings into basic acoustic units. One useful approach for the segmentation of speech signals is the statistically based method. There are three basic methods of segmentation.

1) Brandt's Generalized Likelihood Ratio (GLR) Test [9]: This is an efficient simplified realization of the GLR method of Willsky.
2) The Divergence Test proposed in [10]: This procedure uses the J-divergence of conditional distribution as a measure of distance between models.
3) The Original Pulse Method [11]: This method is derived from the previous one by taking into account the special feature of the glottal excitation in voiced speech.

An analysis of the behavior of speech signal reveals that spectral change locations always correspond to articulary or acoustic changes. The above three methods may be powerful for detection of abrupt changes in spectral characteristics, however, very often segmentation procedures are ambiguous. In other words, they provide erroneous boundaries or oversegmentation in gradual spectral changes. Another failing of the above methods is due to the fact that the locations of models are sometimes chosen at the nonstationary part of the speech signal. The work of Andre-Obrecht shows that the three basic methods give similiar results [11]. But from a computational point of view, the divergence test gives the least computational cost, so it is the best method to preprocess the on-line speech signal in a recognition system.

The purpose of this paper is to give a presentation of some segmentation algorithms based on the divergence test, including new one, and to compare their performance when applied to connected words. Our approach presented here uses three AR models: Two models are located at the adjacent stationary parts of speech signal, another model is located between them. By using a convenient distance measure for comparing them, the segment boundaries are detected. The resulting algorithm turns out to a clearly less computation-consuming procedure than other methods, while providing a better estimate of the segment boundary and reducing the amount of oversegmentation.

II. Segmentation Based on Statistical AR Models

One general approach for the detection of model changes in signals is shown in Fig. 1(a). The observed signal (y_n) is transformed into an innovation sequence (e_n) which plays the role of a cumulative error sequence. If no change occurs in the underlying model structure, (e_n) resembles a Brownian motion process with zero mean. When the model changes, the mean of (e_n^2) monotonically increases after the change. These properties of (e_n) enable one to construct test statistics of the cumulative sum type which is affected by the model change via a change in drift. The use of cumulative sum techniques based on the innovations e_n or the squared innovations e_n^2 is a standard approach for the detection of change in AR models. Among the several approaches based on the cumulative sum, we will describe two algorithms applied to speech signals and then propose a new algorithm.

For three methods we will discuss the observed speech signal (y_n) is assumed to be described by a string of homogeneous units, each of which is characterized by an auto-regressive (AR), or all-pole model, whose parameters may change at some unknown time θ, i.e.,

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

where (e_n) is a zero mean white noise sequence, and $/A=(a_1^{(n)}, \cdots, a_p^{(n)}$ and σ_n^2 are the parameters of the model. Before the model change, these parameters are denoted as

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

After the model change, they are denoted as

$$a_i^{(n)} = a_i^1, \quad 1 \leq i \leq p, \quad \text{for } n \geq \theta;$$
$$\sigma_n^2 = \sigma_1^2, \quad \text{for } n \geq \theta \qquad (2b)$$

Most often in practice the true parameter values of the AR models before and after the change are unknown. Moreover, the structure of the true underlying models may be unknown, and then AR models are used as a tool for the segmentation. In order to detect the change in the model parameters, several models are estimated at different time locations in signal, and their similarity is evaluated by suitable test statistics.

In this section, we will describe: the divergence test proposed in [10], [11] (section II-A); the forward-backward divergence test proposed in [11] (section II-B); and, the proposed method (section II-C).

A. The Divergence Test

This method uses two AR models as indicated in Fig.1 (b). An AR model M_0 is adjusted in a global window and used as a reference model. Another model M_1 is estimated in a moving window with a fixed size. When the two models differ too much from each other the segmentation is done.

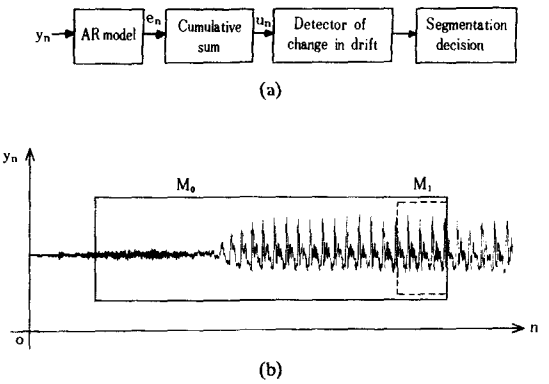


Fig.1. (a) Segmentation based on AR modes. (b) Locations of the windows used in the divergence test.

The test is based on the monitoring of the cumulative sum

$$U_n = \sum_{i=1}^{n} T_i \qquad (3)$$

where

$$T_i = \int g^0(y \mid \mathbf{Y}^{i-1}) \log \frac{g^1(y \mid \mathbf{Y}^{i-1})}{g^0(y \mid \mathbf{Y}^{i-1})} dy - \log \frac{g^1(y_i \mid \mathbf{Y}^{i-1})}{g^0(y_i \mid \mathbf{Y}^{i-1})} \qquad (4)$$

$$\mathbf{Y}^{i-1} = (y_{i-1}, y_{i-2}, \cdots\cdots, y_2, y_1)^T \qquad (5)$$

In the above $g^0(y_i|\mathbf{Y}^{i-1})$, and $g^1(y_i|\mathbf{Y}^{i-1})$ are the two conditional densities corresponding to the models of Fig. 1(b). The Eq. (4) is a distance measure which involves the cross entropy between the conditional distribution of these two models, which is in the Gaussian case given by

$$Tn = \frac{1}{2}\left[2\frac{e_{0,n}\, e_{1,n}}{\sigma_1^2} - \left(1 + \frac{\sigma_0^2}{\sigma_1^2}\right)\frac{e_{0,n}^2}{\sigma_0^2} + \left(1 - \frac{\sigma_0^2}{\sigma_1^2}\right)\right] \qquad (6)$$

where

$$e_{q,n} = y_n - \sum_{i=1}^{p} a_i^q\, y_{n-i}, \quad q = 0, 1$$

are the prediction errors corresponding to the two different models. Before the spectral change i.e., $n < \theta$, the statistics U_n has a zero conditional drift, while after the spectral change i.e., $n \geq \theta$, its conditional drift is negative. A change detection occurs when the global and moving models disagree in the sense of the cumulative sum statistics (3). The statistics (U_n) has, for real implementation, the behavior indicated in Fig. 2(a). When observing its crossing of a threshold λ, the delay for detection may be large if the threshold is too low. In order to reduce this delay and obtain a good estimate of the change time, one can apply Hinkley's cumulative sum test to detect a change in the drift of U_n [12]. It consists of adding a positive bias δ to T_n, and to observe the deviation of the new cumulative sum U_n' with respect to its past maximum. The estimated change time θ is the time at which the past maximum is reached. The detailed procedure is discussed in [10], [11].

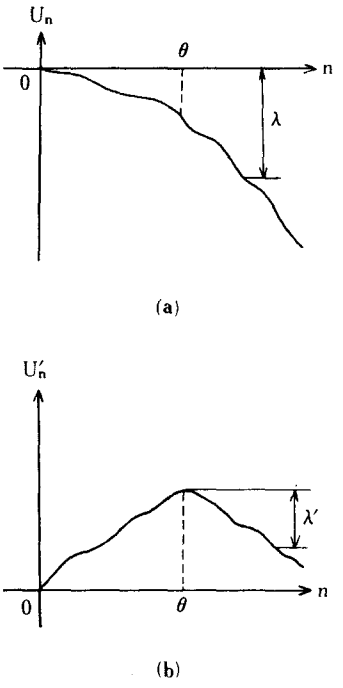


Fig.2. (a) Behavior of the statistics (U_n). (b) Hinkley's test.

The Practical Implementation

The global model M_0 is sequentially identified by a lattice method using the Burg algorithm, and the moving model M_1 by the auto correlation method [13]. Each of them has 16 LPC coefficients. The length of the moving model is 160 samples, corresponding to 20 ms.

Choice of bias is the key point. It is convenient to choose different biases for voiced and unvoiced segments. For this purpose, a coarse voiced/unvoiced decision is performed using the energy of the signal, and the test is performed with

$(\delta, \lambda')=(0.4, 30)$ for voiced segment,
$(\delta, \lambda')=(1.0, 60)$ for unoviced segment.

Discussion: The results of the experiment show the following facts.

- While stationary segments of speech are well located, the nonstationary segments are often oversegmented.
- Where the spectral properties change slowly, there are some omissions of segment boundaries, such as / ju / in Fig.4.

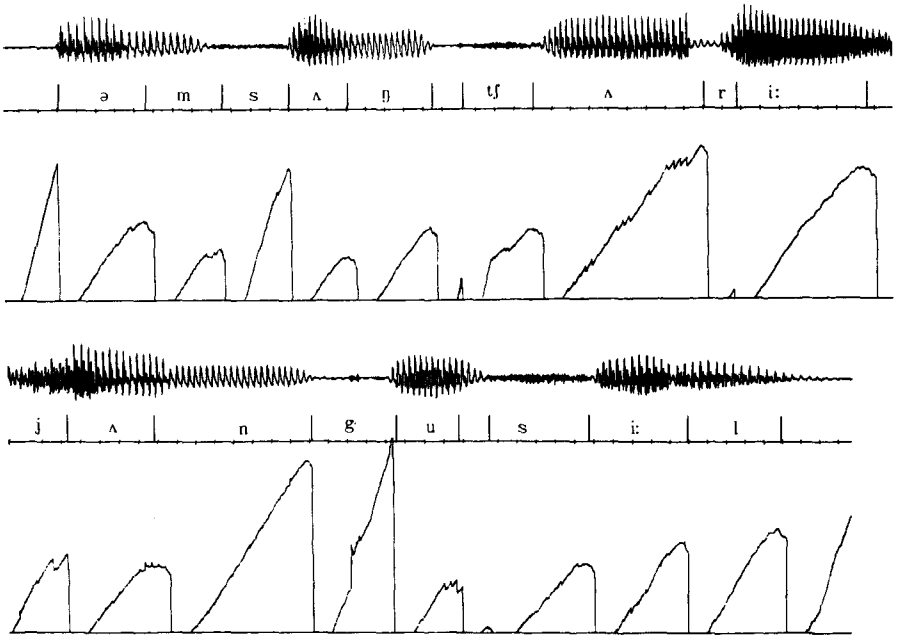


Fig.3. The divergence test on the word: "음성처리 연구실/əm sʌŋ tʃʌ ri: jʌn gu si:l/."

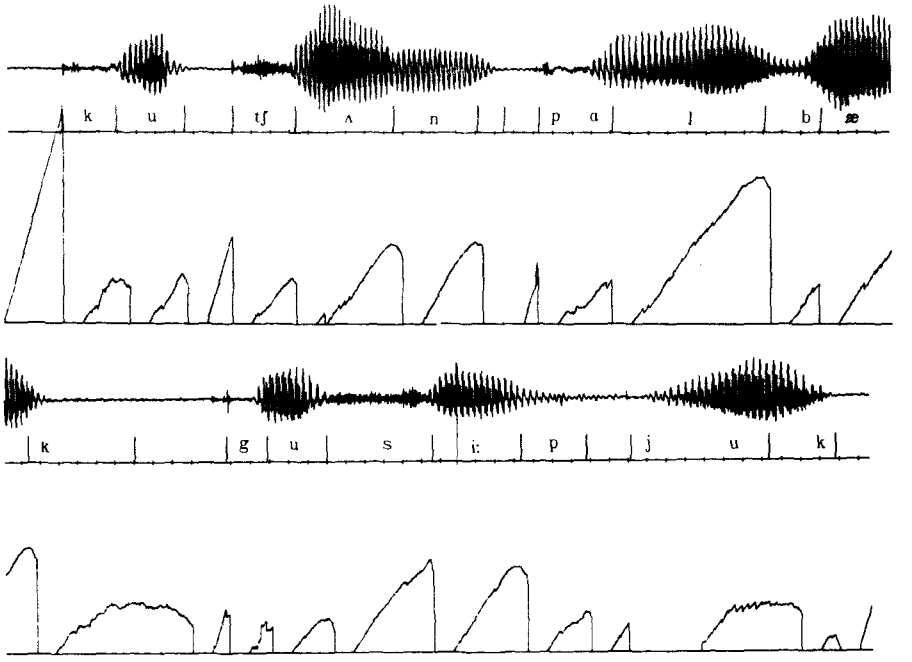


Fig.4. [The divergence test on the word: "구천팔백구십육/ku tʃʌn pal bæk gu si:p juk/."

- On nonstationary segments where formantic structure vanishes, there are some oversegmentation or badly located segments.
- The voiced stop consonants, such as/b/, /d/, and /g/, are often omitted or badly located.
- A boundary of the stop consonant which appears at the end of vowel is often omitted or badly located.

B. The Forward-Backward Divergence Test

The model set, the test statistics, and the procedures are the same as for the divergence method except that a voiced segment, which is judged too long, is processed backward with the same statistics.

Introduce the minimal allowed length for two voiced segments, and denote it by L_{min}; assume the signal $(y_n)_{n \geq 1}$ is processed forward sample-by-sample; a spectral change is detected at time θ. When the segment is voiced and $\theta > L_{min}$, the backward procedure is fired as follows.

Step 1: Process backward, with the divergence method, the whole sample

$$\{y_k\}_{1 \leq k \leq \theta}$$

If still no change is detected, the forward procedure is fired again from the instant θ. Otherwise, we proceed with the second step.

Step 2: Let n_b be the last change location detected by the backward processing (i.e., the location corresponding to the smallest time index k), then

$$\{y_k\}_{1 \leq k \leq n_b}$$

is accepted as the new segment, and the processing goes on forward again from the sample y_{n_b}.

Discussion: The results of experiment showed that the behavior of the forward-backward divergence test is similiar to that for the divergence test but small difference is as follows:

- Some boundaries which have been omitted in the divergence test are detected, such as /ju/ in Fig. 6.
- Some short stops are detected.

C. The Proposed Method

1) Estimation of the Spectral Change Time

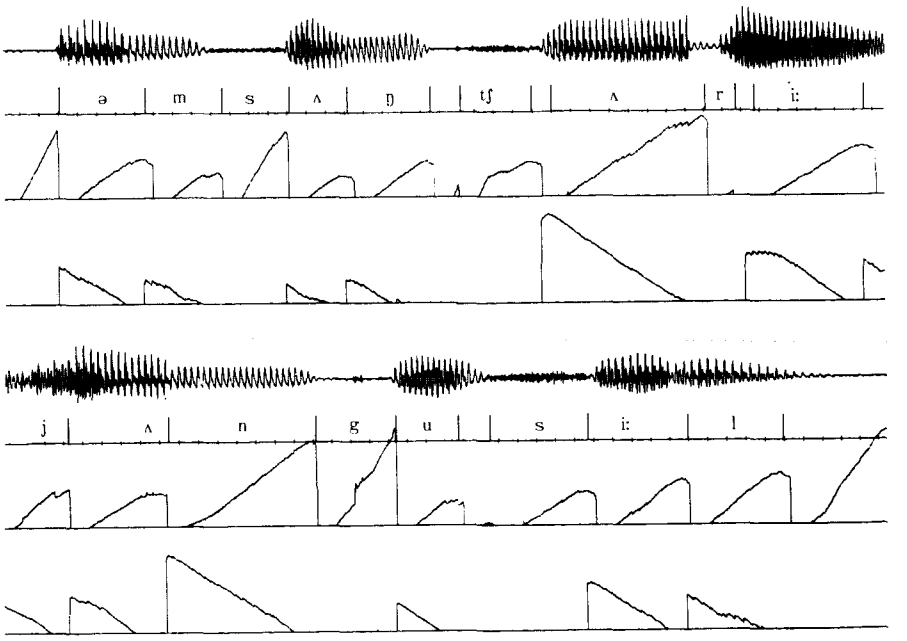


Fig.5. Results of the forward-backward divergence test on the word: "음성처리연구실/ əm sʌŋ tʃʌ ri:jʌn gu si:l/."

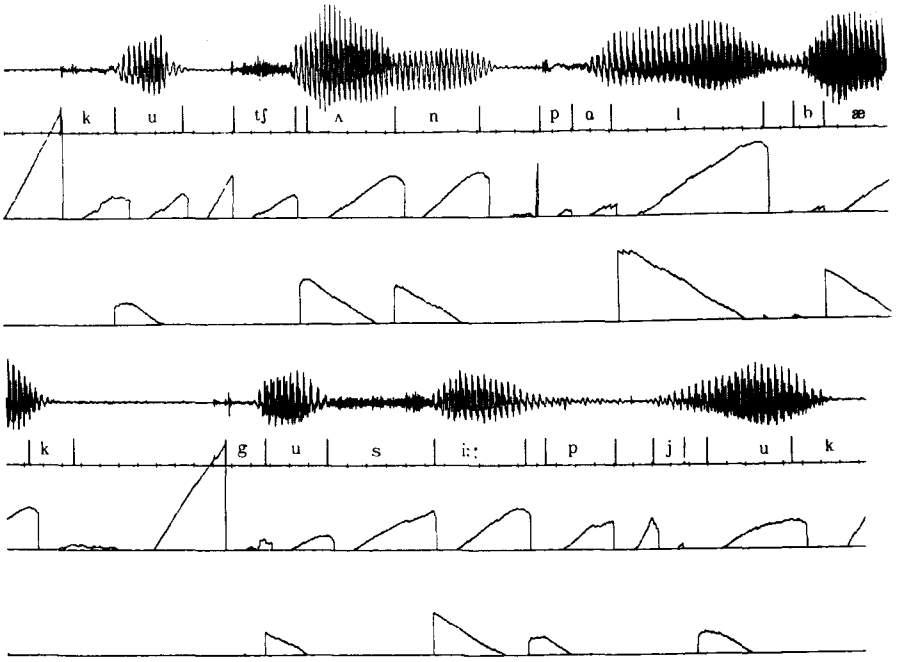


Fig.6. Results of the forward-backward divergence test on the word: "구천팔백구십육/ ku tʃʌn pal bæk gu si:p juk/."

As it has been previously outlined, this method uses three AR models as indicated in Fig.7. The two fixed models M_0 and M_1 are located at the quasi-stationary parts of the signal and another model M_s is a moving model shifted along the time axis from n_0 to n_1. The problem of the choice of these locations will be discussed later.

The basis for the following discussion is the expectation

$$SC(n)=E\left[\log\frac{g^1(y_n\mid Y^S)}{g^0(y_n\mid Y^S)}\right] \tag{7}$$

where

$$Y^S=[y_{n-N},\cdots,y_{n-1},y_{n+1},\cdots,y_{n+N}] \tag{8}$$

$g^0(y_n\backslash Y^S)$ and $g^1(y_n\backslash Y^S)$ denote the two conditional densities corresponding to the two fixed models M_0 and M_1, and SC(n) is an abbreviation of "spectral change". The above is nothing but the log-likelihood ratio between the two joint probability laws $p^1(y_{n-N},\cdots,y_{n-1},y_{n+1},\cdots,y_{n+N})$ and $p^0(y_{n-N},\cdots,y_{n-1},y_{n+1},\cdots,y_{n+N})$. In the AR(p) Gaussian case the Eq. (7) has the following form.

$$SC(n)=\frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2}+\frac{\alpha_0}{2\sigma_0^2}-\frac{\alpha_1}{2\sigma_1^2} \tag{9}$$

where

$$\alpha_0=E[(e_n^0)^2]$$
$$\alpha_1=E[(e_n^1)^2] \tag{10}$$

$(e_n^0)_n$ and $(e_1^1)_n$ are the sequence of innovations of models M_0 and M_1, respectively. A more convenient form is obtained by using the Itakura-Saito distorition measure d_{IS}[14],

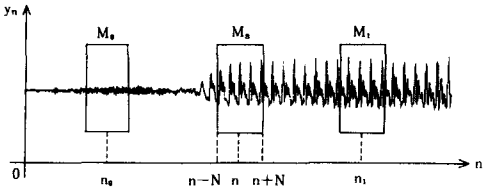


Fig.7. Locations of the three windows for estimation of segment boundaries.

$$SC(n)=\frac{1}{2}[d_{IS}(|X(z)|^2;|G_0(z)|^2)-d_{IS}(|X(z)|^2;|G_1(z)|^2)] \tag{11}$$

Here X(z) is the z-transform of the windowed speech in model M_S. $G_0(z)$ and $G_1(z)$ are all-pole filters corresponding to the two models M_0 and M_1, respectively. Viewed in Eq. (11), the first term is the Itakura-Satio distortion between M_S and M_0. The second is the Itakura-Saito distortion between M_S and M_1. When $n=n_0$, the model M_S is identical to model M_0. As n goes to n_1, M_S becomes more similar to M_1 and less similar to M_0. Thus at last, when $n=n_1$, M_S is identical to M_1. Therefore, as n goest from η_0 to n_1, SC(n) varies from negative to positive. Especially when SC(n)=0, the two Itakura-Saito distortions are equal, i.e., moving model Ms has equal similarity to two models, M_0 and M_1. The analysis of the speech signal in this point from several experiments shows it is reasonable to assume that the spectral change time is the time when M_S has equal similarity to M_0 and M_1 i.e., the absolute value of SC(n) is minimum. Thus the spectral change time estimate θ is given by

$$\theta=\underset{n}{\operatorname{argmin}}\,|SC(n)|,\quad \text{for } n_0<n<n_1 \tag{12}$$

The Itakura-Saito distortion is too sensitive to the gain of model. So, when the gain of the two models are greatly different, it is difficult to estimate the change time accurately. Hence we modify Eq. (11) using the gain insentive version of Itakura-Saito distortion, i.e., the gain normalized Itakura-Saito distortin d_{GN}. In this case, Eq. (11) becomes

$$SC'(n)=\frac{1}{2}[d_{GN}(|X(z)|^2;|G_0(z)|^2)-d_{GN}(|X(z)|^2;|G_1(z))|^2] \tag{13}$$

and the spectral change time estimate is

$$\theta'=\underset{n}{\operatorname{argmin}}\,|SC'(n)| \tag{14}$$

Fi.g 8 illustrates the procedure for estimating the spectral change time.

2) The Locations of the Three AR Models

The two fixed models used here must lie on the

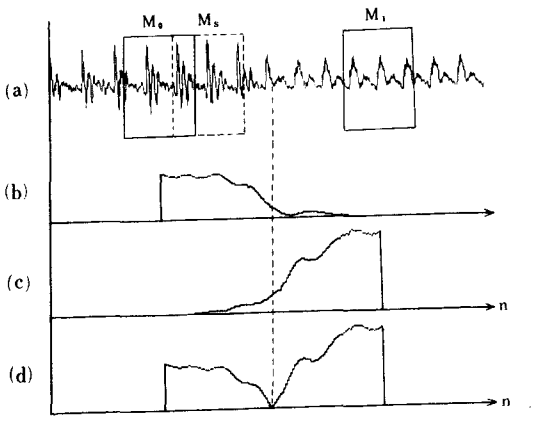


Fig.8. Estimation of the spectral change time.
(a) speech signal.
(b) $d_{GN}(|X(z)|^2; G_1(z)|^2)$.
(c) $d_{GN}(|X(z)|^2; Go(z)|^2)$.
(d) SC''(n).

quasi-stationary parts of the signal. Fig.9 describes the procedure for detecting the quasi-stationary parts of the signal. The signal is processed frame-by-frame. Assume that n denotes the index of the current frame and n' is a constant. The distance $D_{adj}(n)$ between (n-n') th frame and (n+n') th frame

$$D_{adj}(n) = d_{GN}(|X_{n-n'}(z)|^2; |X_{n+n'}(z)|^2) \quad (15)$$

is computed. Where $X_n(z)$ is the z-transform of the windowed speech in nth frame. When $X_{n-n'}(z)$ and $X_{n+n'}(z)$ are in stationary parts of the signal, $D_{adj}(n)$ has lower distance, while in nonstationary parts it has higher distance. Hence, we regard the portions where $D_{adj}(n)$ is locally minimum as quasi-stationary parts and use them to identify two fixed models. The moving model is located between these two fixed models. For easy finding of the local minimums, we smoothed $D_{adj}(n)$ by using a simple mean filter. Fig. 10 illustrates an example of this procedure.

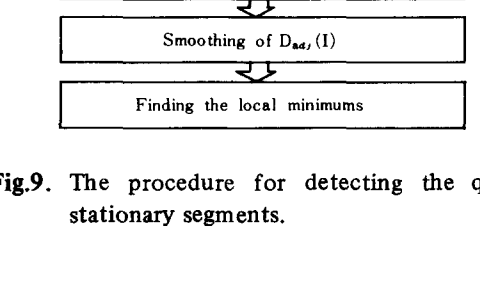


Fig.9. The procedure for detecting the quasi-stationary segments.

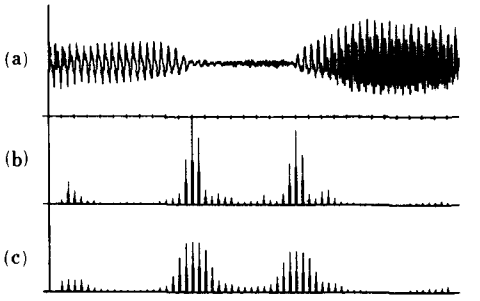


Fig.10. Detection of the quasi-stationary segments in signal. (a) speech signal. (b) $D_{adj}(n)$: the distance between (n-n') th frame and (n+n') th frame. (c) smoothed $D_{adj}(n)$.

3) The Pratical Implementation

Given two fixed models located at the adjacent stationry parts of the signal, the distance between M_0 and M_1 is calculated to test whether a significant spectral change occurs or not. Though the gain normalized Itakura-Saito distortion is suitable for this purpose, it has a drawback of asymmetricity, i.e.,

$$d_{GN}(|G_0(z)|^2; |G_1(z)|^2) \neq d_{GN}(|G_1(z)|^2; |G_0(z)|^2) \quad (16)$$

This sometimes gives a failing in detecting the spectral change, hence, we used the following distance measure d_{SC} for this purpose.

$$d_{SC} = \max[d_{GN}(|G_0(z)|^2; |G_1(z)|^2), d_{GN}(|G_1(z)|^2; |G_0(z)|^2)] \quad (17)$$

The practical procedure for segmentation is as follows.

Step 0: A model M_0 is adjusted in location where the first local minimum of $D_{adj}(n)$ occurs and the other model M_1 is in the second local minimum.

Step 1: Calculate the distance d_{SC} between M_0 and M_1. When the distance is lower than a certain threshold λ, i.e.,

$$d_{SC} < \lambda , \tag{18}$$

then go to step 3. Otherwise i.e., when the two models differ too much, process step 2.

Step 2: Let n_0 and n_1 denote the locations of the two models, M_0 and M_1, respectively. Calculate $SC'(n)$ is Eq.(13) where n goes from n_0 to n_1 in step n_s. Then, the time index θ' where $SC'(n)$ has minimum absolute value is regarded as an estimation of the spectral change time.

Step 3: The model M_1 becomes M_0 and the model M_1 is adjusted in the next local minimum. Return to step 1.

LPC Analysis: Hamming window of 160 samples (20ms) is applied to the speech signal. The window is advanced in step of 40 samples (5ms) to get the next frame. For each frame, the 16 LPC coefficients are computed using the autocorrelation method [13]. The three models have the same length of windows of 160 samples and each of them has 16 LPC coefficients.

Choice of the λ, n' and n_s: From several preliminary experiments, we choose them as follows:

λ=1.0,
η'=2 frames,
η_s=8 samples (1ms)

Discussion: Fig. 11 and 12 show the results of the proposed method. Following is noted.

- This reduces the amount of oversegmentation in nonstationary segments.
- The boundaries of segments are more accurate compared to those of the previous two methos.
- The boundaries of the short stops are sometimes omitted.

III. Experimental Results

The segmentation schemes described earlier have been implemented on a IBM-PC/AT. The block diagram of the system is shown in Fig. 13. The speech signals were sampled at 8 KHz rate.

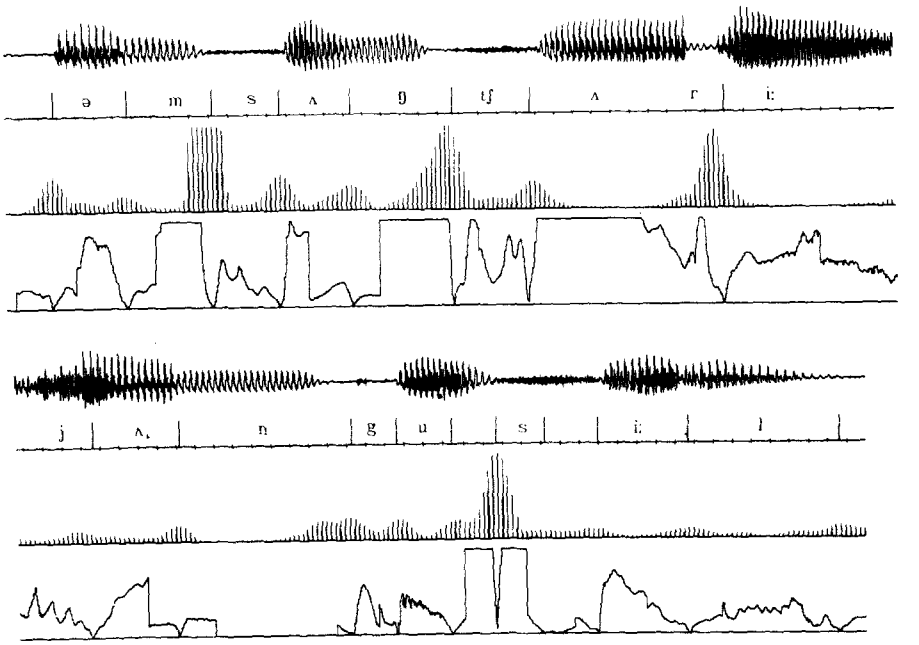


Fig.11. Results of the proposed method on the word: "음성처리 연구실/əm sʌŋ tʃʌ ri: jʌn gu si:l/."

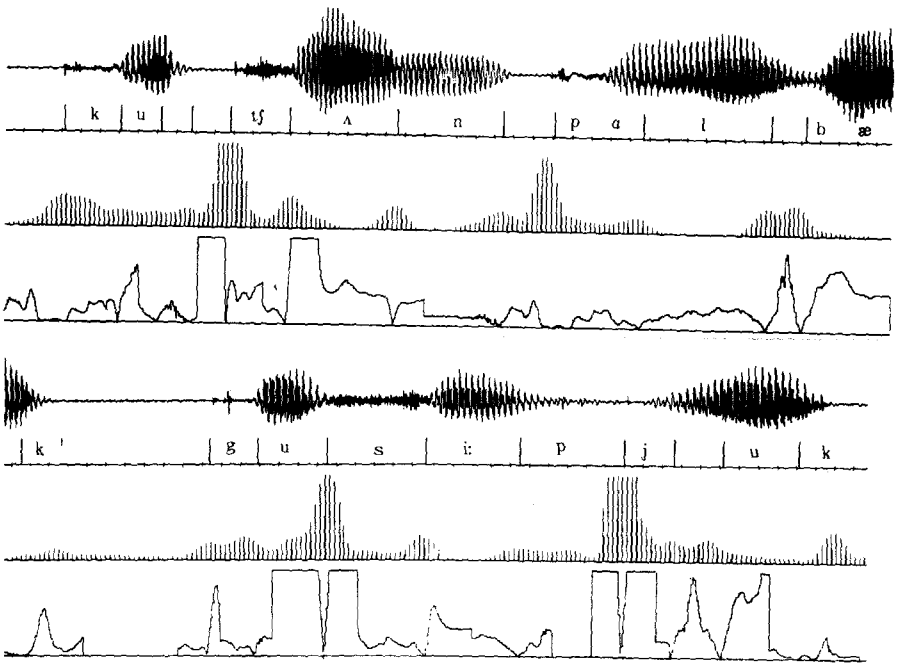


Fig.12. Results of the proposed method on the word: "구천팔백구십육/ku tʃʌn pal bæk gu si:p juk/."

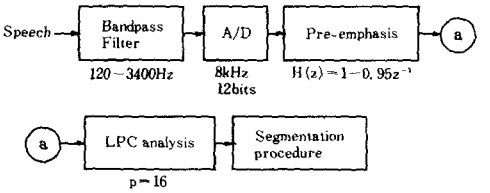


Fig.13. Block diagram of overall system.

digitized to 12 bits, and LPC analyzed using the 16 pole model.

The data base of 10 phonetically balanced words, in Korean, spoken by 10 different adult speakers (5 male, 5 female) was used. A total of 1390 phonemes was contained in this data base.

A number of experiments have been carried out which concentrate on two main topics. The first is a performance comparison of the three methods described in the previous section and the second topic is an analysis of segmentation errors.

1. Performance Comparison of the Three Methods

For objective evaluation, we assumed that two phonemes must be separated by a boundary, even if their realization can be only one transient unit. The three performance criteria of the segmentation algorithms for comparison are as follows:

1) Amount of the oversegmented segments
2) Amount of the omitted segments
3) Amount of the badly located boundaries (when a detected boundary is apart from its true boundary more than 10 msec, it is regarded as a badly located boundary).

The results of experiments are summarized in Table I. It shows that the proposed one is the most preferred algorithm in its performance. One of the key features of the proposed algorithm is its ability to eliminate spurious segments in the transient portions of the signal. For 1390 phonemes in data base, it detected 1493 segments which is much less than those of other methods. A second point to note is that the proposed algorithm can give more accurate boundaries compared to other methods. As for the amount of omissions, all the three algorithms gave the similar results in this data base.

Table 1. Comparison of three methods.

	No. of phonemes	No. of detected segments	No. of bad locations	No. of omission
Divergence method method	1390	1691	329	164
Forward-backward divergence method	1390	2010	206	138
Proposed method	1390	1493	57	145

2. Analysis of Segmentation Errors

We describe omissions, oversegmentation, and bad location of segments

Omission of a segment: It arises on the short segments such as the stop consonants. The voiced stop consonants /b/, /d/, and /g/ are transient, noncontinuant sounds which are produced by building up pressure behind a total constriction somewhere in the oral tract, and suddenly releasing the pressure. They are highly influenced or coarticulated by the vowel which follows the stop consonant. Therefore the previous assumption that the speech signal is described by a string of homogeneous units is not valid, thus the three method, in most cases, give omissions of these consonants. The unoviced consonants /p/, /t/, and /k/ which appear at the end of a word are are produced by total closure of the vocal tract, thus following the period of closure, there is a period of aspiration. The waveforms of them give little distinguishing features. Therefore, all the three algorithms which are based on the detection of the spectral change do not work well on these consonants. When a slow spectral change occurs on the transition portion of two vowels, the divergence test sometimes gives omission of a boundary between them. The forward-backward divergence test removes some of these omissions, but there still exist some omissions.

Oversgementation: It often occurs on the portions as follows:

- Both ends of a voiced phonemes
- The portions where gradual spectral change arise.
- Noisy segments
- Unvoiced segments.

When using Hinkley's test, the value of δ influenced greatly on the amount of oversegmentation. When δ is too low it increases the oversegmentation, while a too high value of δ gives many omissions of segments.

Bad location of a segment: The results of the divergence test and the forward-backward divergence test show many bad locations. This is due to an understimation of moving model M_1 in the neighborhood of the transition. While using the proposed method, there are only a few bad locations.

IV. Conclusions

In this paper, we have presented a speaker-indpendent segmentation of speech signals based on three statistical AR models. Two fixed models are located at the quasi-stationary parts of signals and between them another moving model is located. Changes are detected when a distance between these two fixed models exceeds a preset threshold and then the procedure for estimating the change time is executed. The proposed method has been compared to classical methods via experiments, and the improvement in performance was shown. The main features of the algorithm are summarized as follows:

- The algorithm uses robust statistical models. It yields a very reliable and speaker-independent segmentation.
- Two fixed models are identified at the quasi-stationary parts of the signal before and after spectral change. Thus precise identification of the parameters is possible. This gives a more accurate detection of spectral change.
- Since the relative distances, one from model M_S and M_O, the other from model M_S and M_1, are used to estimate the spectral change time, there is no need to set a threshold value. This is also suited for a speech signal where a gradual spectral change arise, and can reduce the amount of oversegmentation.

The segmentation is performed on overlapping frames. When a spectral change is detected, the frame is shifted in less smaller step for more accurate detection of the change time. This substantially reduces the computational cost while maintaining the accuracy of the spectral change time.

References

[1] H. Sakoe, "Two-Level DP-matching-A dynamic programming-based pattern matching algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 6, pp. 588-595, Dec. 1979.

[2] L.R. Rabiner and C.E. Schmidt, "Application of dynamic time warping to connected digit recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 377-388, Aug. 1980.

[3] C.S. Myers and L.R. Rabiner, "A level building dynamic time warping algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 2, pp. 284-297, Apr. 1981.

[4] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 2, pp. 263-271, Apr. 1984.

[5] M.R. Sambur and L.R. Rabiner, "A statistical decision approach to the recognition of connected digits," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 6, pp. 550-558, Dec. 1976.

[6] R. Zelinski and F. Class, "A segmentation algorithm for connected word recognition based on estimation principles," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, no. 4, pp. 818-827, Aug. 1983.

[7] T. Ukita, T. Nitta, and S. Watanabe, "A speaker independent recognition algorithm for connected word using word boundary hypothesizer," *ICASSP*, Tokyo, 1986.

[8] M. Cravero, R. Pieraccini, and F. Raineri, "Definition and evaluation of phonetic units for speech recognition by hidden Markov models," *ICASSP*, Tokyo, 1986.

[9] A. Von Brandt, "Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio test," in proc. *ICASSP*, Boston, 1983.

[10] M. Basseville and A. Benveniste, "Sequential detection of abrupt changes in spectral characteristics of digital signals," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, pp. 709-724, Sept. 1983.

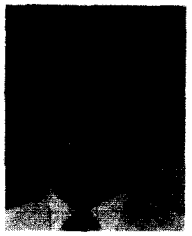
[11] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, no. 1, pp. 29-40, Jan. 1988.

[12] M. Basseville, "Edge detection using sequential methods for change in level. Part II: Sequential detection of change in mean," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 1, pp. 32-50, Feb. 1981.

[13] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[14] R.M. Gray, A. Buzo, A.H. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 367-376, Aug. 1980. *

著 者 紹 介



曺 政 鎭(正會員)

1959年 1月 3日生. 1981年 2月 경북대학교 전자공학과 졸업. 1986年 2月 경북대학교 대학원 전자공학과 졸업 공학석사학위 취득. 1986年 3月～현재 경북대학교 대학원 전자공학과 박사과정 재학중. 주관심분야는 음성인식 및 음성합성 등임.

●

金 秀 重 (正會員) 第25卷 第7號 參照
현재 경북대학교 전자공학과 교수

洪 再 根(正會員)

1951年 1月 25日生. 1975年 2月 경북대학교 전자공학과 졸업. 1979年 2月 경북대학교 대학원 전자공학과 졸업. 1979年～1983年 경북공업 전문대학 전자과 조교수. 1983年～현재 경북대학교 전자공학과 조교수.