Blind speech segmentation과 에너지 가중치를 이용한 문장 종속형 화자인식기의 성능 향상

김정곤(부산대), 김형순(부산대)

<차 례>

<Abstract>

Performance improvement of text-dependent speaker verification system using blind speech segmentation and energy weight

Jung-Gon Kim, Hyung Soon Kim

We propose a new method of generating client models for HMM based text-dependent speaker verification system with only a small amount of training data. To make a client model, statistical methods such as segmental K-means algorithm are widely used, but they do not guarantee the quality or reliability of a model when only limited data are avaliable. In this paper, we propose a blind speech segmentation based on level building DTW algorithm as an alternative method to make a client model with limited data. In addition, considering the fact that voiced sounds have much more speaker-specific information than unvoiced sounds and energy of the former is higher than that of the latter, we also propose a new score evaluation method using the observation probability raised to the power of weighting factor estimated from the normalized log energy. Our experiment shows that the proposed methods are superior to conventional HMM based speaker verification system.

* Keywords: Text dependent speaker verification, blind speech segmentation

1. 서 론

일반적으로 화자확인 시스템은 사용자가 지정한 문장 또는 단어를 사용하는 문장 종속형 시스템과 내용에 제한이 없는 문장 독립형 시스템으로 구분된다[1]. 문장 종속형의 경우 미리 정의한 문장에 대해서만 시스템을 훈련하면 되므로 적은 훈련 데이터를 이용해서 좋은 성능을 나타낸다. 반면에 문장 독립 시스템의 경우 입력 음성에 제한이 없는 반면 사용자의 많은 훈련 데이터가 필요하다는 문제점과, 아직까지 인증 성능이 문장 종속형에 비해서 떨어지는 단점이 있다.

일반적으로 문장 종속형 화자인식에 사용되는 방법으로는 패턴 정합법과 Hidden Markov Model (HMM)등이 있다. 첫 번째로, 패턴 정합법(template matching)은 dynamic time warping (DTW)알고리즘[2] 등을 사용하여 입력 패턴을 미리 정해진 참조 패턴(reference pattern)과 비교하여 유사성을 판단하는 방법이다. 이 방법은 길이가 다른 두 패턴을 비선형적으로 정합하는 방법이다. 그러나 화자내의 변이를 수용할 수 있는 참조 패턴의 작성이 어려우며, 사용자의 참조 패턴에 해당하는 특징 파라미터를 모두 저장하고 있어야 하는 단점이 있다. 두 번째로, HMM을 이용한 방법이 있다[3]. 이 방법은 확률 모델로써 현재 음성인식에서 가장 성공적인 방법으로 알려져 있으며, 화자인식에서도 좋은 결과를 보이고 있다. HMM은 학습 기능을 이용하여 화자내의 변이를 흡수할 수 있으며, 입력 패턴의 비선형 정합을 수행하는 특성이 있다. 이 방법은 모델의 구성 형태에 따라 문장 종속형이나 문장 독립형 화자인식 시스템의 구현이 가능하고, 충분한 학습 자료의 이용이 가능하다면 신뢰성 있는 사용자 모델의 구성이 가능하지만, 그렇지 않은 경우 훈련 과정에서 신뢰성 있는 모델의 생성이 힘들다는 단점이 있다. 따라서 화자인식기의 신뢰성 향상을 위해서는 충분한 양의 학습 자료를 사용하는 것이 좋지만, 사용자의 편의성 측면에서는 사용자에게 적은 양의 학습 자료를 요구하는 것이 바람직하다.

본 논문에서는 사용자의 편의를 고려하여 적은 양의 학습 자료만을 사용하여 HMM 기반의 문장 종속형 화자인식기를 위한 사용자 모델을 생성하는 방법을 제안한다. 이와 함께 추가적인 성능 향상을 위하여 입력음성에 존재하는 묵음과 무성음 같은 화자간의 분별력이 떨어지는 구간보다 화자의 정보를 많이 포함하고 있는 유성음 부분을 강조하기 위하여 관측 확률에 정규화된 log 에너지 기반의 가중치를 주는 방법을 함께 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 사용자 모델 구성을 위한 음성 분할 방법에 대해서 설명하고, 3장에서는 관측 확률에 에너지 가중치를 주는 방법에 대해서 살펴본다. 4장에서는 기존의 방법과 제안한 방법의 실험 결과를 살펴보고, 5장에서 결론을 맺는다.

2. 사용자 모델 생성을 위한 음성 분할

현재 사용되고 있는 대부분의 음성 합성, 음성 인식 그리고 화자인식 시스템들은 음성을 여러 개의 상태로 나누어서 모델링하고 있기 때문에 음성 분할은 이들 시스템을 구현하는데 필수적으로 요구되는 사항이다. 음성 분할은 전문가들이 직접 수작업으로 하는 방법이 가장 좋은 성능을 보이고 있지만 이는 시간과 노력이 많이 소요되는 작업이어서 대량의 음성 데이터를 분할하는 데는 적합하지 않다. 더욱이 화자인식의 경우는 사용자의 등록 음성을 미리 수집해서 저장하지 않는 경우가 대부분이어서 수작업을 통한 음성 분할이 거의 불가능하다. 따라서 대부분의 음성인식과 화자인식 시스템에서는 자동으로 음성을 분할하는 방법을 사용하고 있다. 이들 자동 음성 분할 알고리즘들은 음성이 몇 개의 구간으로 분할 될 것인지를 미리 결정해야 하는데, 특히 사용자가 암호를 선택하는 문장 종속형 화자인식 시스템의 경우처럼 발성 내용을 알지 못하는 상태에서 음성을 분할하는 방법을 blind speech segmentation이라고 한다[4].

2.1.절에서는 blind speech segmentation을 위한 분할 개수의 추정 방법에 대해서 살펴보고, 2.2.절에서 기존에 HMM에서 사용되고 있는 음성 분할 방법인 segmental K means 알고리즘[5]과 화자인식을 위한 HMM 모델 생성 시의 문제점에 대해서 살펴본 후, 2.3.절에서는 본 논문에서 사용하고 있는 level building DTW 기반의 음성 분할 방법[4]에 대해서 살펴본다.

2.1. 음성 분할 개수의 추정

본 논문에서는 음성 스펙트럼의 변화를 측정하는 스펙트럼 변이 함수 (Spectral Variation Function)으로서 식 (1) 과 같은 delta cepstrum의 유클리디언 거리를 사용하였다.

$$SVF_{\Delta cep}(n) = \left[\sum_{m=1}^{p} [\Delta C_n(m)]^2\right]^{1/2} \tag{1}$$

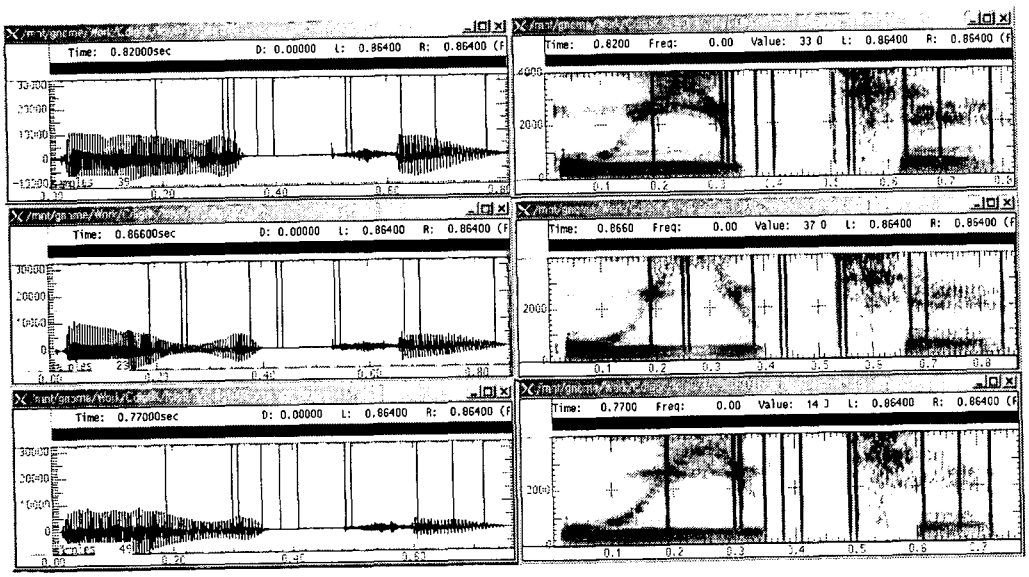
여기서 n은 프레임 인덱스이고 $\triangle C_n(m)$은 n번째 프레임의 m번째 음성 특징 벡터의 delta cepstrum을 나타낸다. Cepstrum이 한 프레임 내에서의 log 스펙트럼 정보를 표현하는데 비해서 그것의 시간차 성분인 delta cepstrum은 시간에 따른 음성 스펙트럼의 변화를 표현하게 되므로 스펙트럼의 변화가 큰 구간 즉 음가가 변하는 구간에서 상대적으로 큰 값을 가지고 스펙트럼의 변화가 작은 구간에서는 작은 값을 가지게 되는 특성이 있다. 본 논문에서는 스펙트럼 변이 함수의 국부적인 정점 개수를 분할해야 할 구간의 수로 사용하였다.

2.2. segmental K means 방법

Segmental K means 알고리즘은 HMM 기반의 음성인식을 위한 sub word 단위의 모델 생성을 위해서 음성 데이터를 분할하는데 널리 사용되는 방법이다.

이 방법은 초기에 모든 구간들이 동일한 길이를 가지도록 음성 데이터를 분할한 다음 그 분할 정보를 이용해서 HMM 모델을 생성한다. 이렇게 만들어진 HMM 모델을 이용해서 Viterbi decoding을 수행하고 나서 역추적(backtracking) 과정을 거치게 되면 새로운 경계 정보를 얻을 수 있다. 이렇게 얻어진 경계 정보는 이전의 경계 정보에 비해서 좀 더 음성 특징들이 유사한 프레임들이 하나의 구간을 이루게 된다. 따라서 HMM 모델을 생성과 Viterbi decoding 과정을 계속 반복하게 되면 음성 특징이 동일한 프레임들이 하나의 구간(segment)를 이루게 된다.

이 방법은 HMM이라는 통계적인 방법에 기반을 두고 있어서 음성인식의 경우처럼 데이터가 많은 경우 일관성 있는 음성 분할이 가능하기 때문에, 신뢰성 있는 HMM 모델을 만들어 낼 수 있다. 하지만 문장 종속형 화자인식 시스템의 경우, 사용자의 편의성을 고려하여 등록 시 같은 문장을 3~4회 정도만 발성하도록 요구하는 것이 대부분이어서 데이터 부족으로 인해서 모델 생성과 Viterbi decoding 과정을 계속 반복하더라도 경계 정보가 수렴하지 않거나, 수렴 결과의 신뢰성이 떨어지게 된다. 아래의 <그림 1>은 3개의 훈련 데이터를 이용해서 segmental K means 방법으로 음성 분할을 했을 때의 결과를 그림으로 보여주고 있다.



<그림 1> Segmental K means 방법을 이용한 음성 분할 결과

2.3. Level building DTW 기반의 음성 분할 방법

Level building DTW기반의 음성 분할 방법은 통계적인 방법에 기반을 둔 segmental K means 알고리즘과는 달리 음성 자체의 스펙트럼 특징을 이용해서 음성을 분할하는 방법이다. 알고리즘은 다음과 같다.

먼저 L을 전체 세그먼트의 수, l을 세그먼트 카운터, N을 분할할 음성의 전체 프레임 수라고 하고 $d_l(i,n)$을 l 번째 레벨에서 i 번째 프레임부터 n 번째 프레임들 사이의 국부적인 거리(local distance), $D_l(n)$을 l 번째 레벨에서 n 번째 프레임까지의 최소 누적 거리(minimum accumulated distance), $B_l(n)$을 그 때까지의 경로라고 했을 때, 첫 번째 $l=1$ 단계에서는 식 (2)로 초기화한다.

$$\begin{aligned} D_l(n) &= d_l(1,n) \qquad n=1 \ \ to \ \ (N-L+1) \\ B_l(n) &= 1 \end{aligned} \tag{2}$$

두 번째 $l=2$ 인 단계부터 $l=L-1$인 단계에서는 아래와 같이 갱신시킨다.

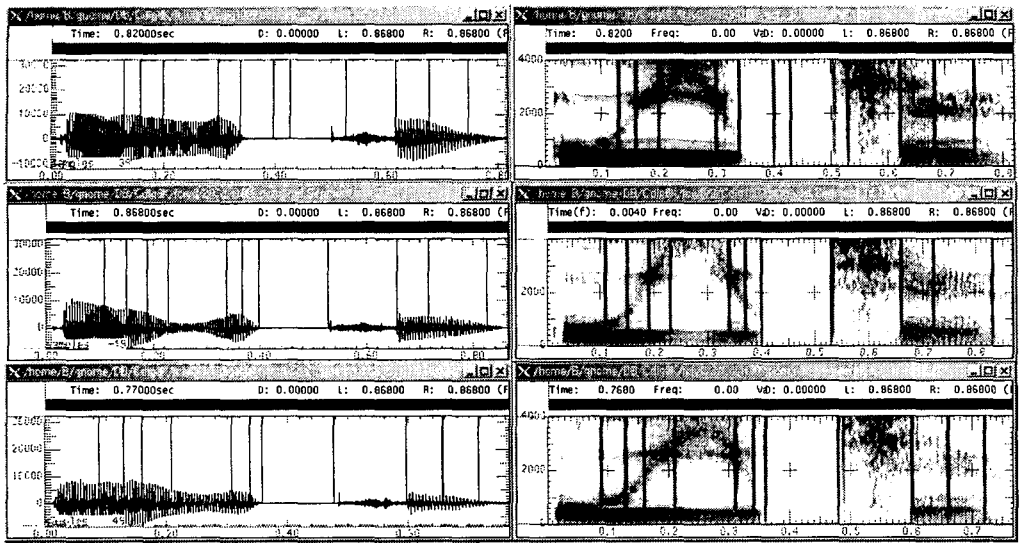
$$\begin{aligned} D_l(n) &= \min_i\{d_l(i+1,n)+D_{l-1}(i)\} \quad (i<n,\ n=l \text{ to } (N-L+l)) \\ B_l(n) &= \arg\min_i\{d_l(i+1,n)+D_{l-1}(i)\} \end{aligned} \tag{3}$$

마지막 $l=L$ 단계에서는

$$\begin{aligned} D_l(n) &= \min_i\{d_l(i+1,n)+D_{l-1}(i)\} \qquad (i<N) \\ B_l(n) &= \arg\min_i\{d_l(i+1,n)+D_{l-1}(i)\} \end{aligned} \tag{4}$$

의 과정을 거친 후 $B_l(n)$을 이용해서 그 때까지의 최적 경로를 역추적해서 분할 정보를 얻게 된다.

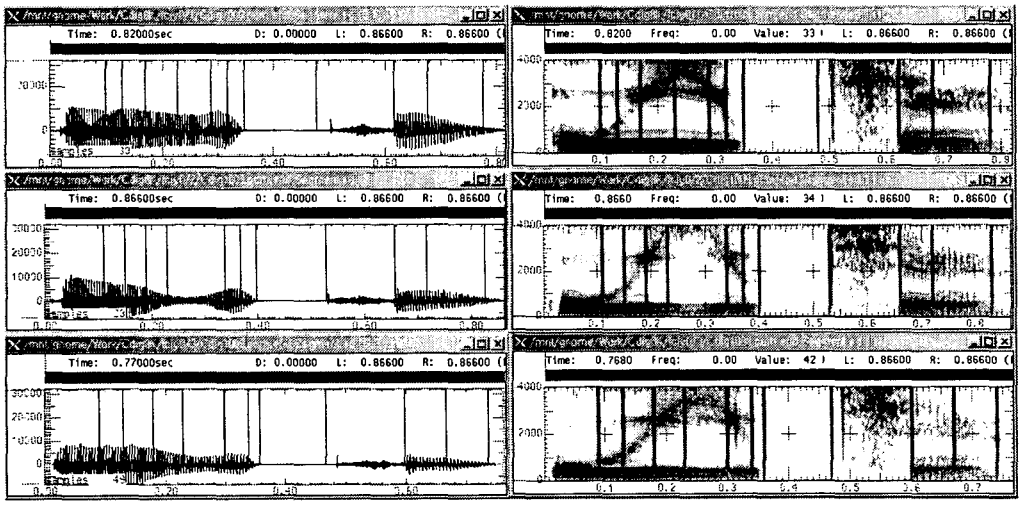
앞에서 언급한 바와 같이 level building DTW를 이용한 음성 분할 방법의 경우 통계적인 방법에 기반을 둔 segmental K means 알고리즘과는 달리 개별 음성별의 스펙트럼 정보들을 이용해서 구간 경계를 찾기 때문에 음성 데이터를 분할하는데 있어서 통계치 추정을 위한 데이터 부족으로 인한 분할의 신뢰성 저하 문제는 발생하지 않는다. 하지만 같은 발성 내용이라 하더라도 개별 음성의 스펙트럼 특징은 동일하지 않기 때문에 <그림 2>에서 보는 것처럼 각 음성별 분할 정보가 일치하지 않는 경우가 빈번하게 발생하는 단점이 있다.



<그림 2> Leve building DTW 기반의 음성 분할 방법을 이용한 음성 분할 결과

본 논문에서는 level building DTW기반의 음성 분할 방법에서 나타나는 등록 음성별 분할 정보의 불일치 문제점을 극복하기 위해서 음성 분할 이전에 DTW를 먼저 수행하여 등록 음성들 사이의 프레임별 대응 관계와 평균 거리를 계산한 후 등록 음성 중 다른 등록 음성들과의 평균 거리가 가장 가까운 1개의 음성에 대해서만 음성을 분할하고, 나머지 음성들에 대해서는 DTW를 통해서 얻은 프레임별 대응 관계를 이용해서 분할하는 방법을 사용하였다.

<그림 3>은 제안한 방법을 사용했을 때의 음성 분할 결과이다.



<그림 3> 제안한 방법을 사용한 음성 분할 결과

<그림 3>을 <그림 1>, <그림 2>와 비교해 보았을 때 제안한 방법이 보다 일관성 있는 음성 분할이 가능함을 보여주고 있다.

3. 에너지 가중치를 이용한 관측 확률

일반적으로 사용자의 등록 음성에는 화자의 정보와는 상관없는 묵음 구간과 유성음에 비해서 화자 간의 분별력이 떨어지는 무성음 구간이 존재한다. Viterbi decoding의 경우 화자 간의 분별력이 떨어지는 묵음과 무성음 구간에서의 확률값들이 화자 간의 분별력이 높은 유성음 부분에서의 확률값들과 동일한 정도의 영향력을 가지게 된다. 하지만 음성 인식과 달리 화자인식에서는 입력 음성이 가지고 있는 발성 내용보다는 입력 음성이 내포하고 있는 화자 정보가 더욱 중요함을 고려할 때, 무성음과 유성음에 대한 가중치를 동일하게 주는 것보다는 유성음 부분을 좀더 강조해 주는 것이 화자인식을 성능을 높이는 데 도움을 줄 수 있다.

본 논문에서는 유성음 구간의 에너지가 무성음과 묵음 구간에 비해서 상대적으로 크다는 점을 이용하여 유성음에 좀 더 많은 가중치를 주는 방법으로서 (0, 1) 사이로 정규화 된 log 에너지를 관측 확률의 가중치로 사용하는 방법을 제안한다. t인 시간의 j 번째 상태에서의 관측 확률 $B_j(O_t)$라고 할 때 가중치를 적용한 식은 다음과 같다.

$$B'_j(O_t) = B_j(O_t)^{W_t} \tag{5}$$

여기서 W'는 $W_t = \begin{cases} 1 & E_t \geq 0.5 \\ 0.5 & E_t < 0.5 \end{cases}$ 이고, E_t는 시간 t에서의 (0, 1) 사이로 정규화된 log 에너지이다.

이 방법은 관측 확률에 지수승의 가중치를 주기 때문에 같은 가중치를 주더라도 각 상태별 관측 확률에 미치는 영향이 다르게 된다. 따라서 $B'_j(O_t)$가 최적 경로를 찾아나간다는 보장을 하지 못한다. 따라서 관측 확률 $B_j(O_t)$는 최적 경로를 찾는데 사용하고, 실제 테스트 과정에서 사용할 스코어는 $B'_j(O_t)$를 사용해서 얻게 된다. 제안한 방식은 정규화된 log 에너지만을 사용해서 가중치를 주므로 적은 계산량만으로 에너지가 큰 유성음 부분에 효과적으로 가중치를 줄 수 있는 장점이 있다.

4. 실험 및 결과

4.1. Database

본 논문에서는 문장 종속형 화자확인 시스템의 평가를 위해 국어공학센터의 한국어 4연 숫자 음성 데이터베이스 중에서 남성 21명과 여성 19명의 음성만을 사용하였다[6]. 원음은 방음 부스에서 Senheizer HMD224X를 사용하여 녹음되었으며, A/D 변환에는 KAY CSL 4300B가 사용되었다. 그리고 16 kHz로 샘플링되고 16 Bits로 양자화 되어있다. 발성에 참여한 화자는 대부분 서울 지역의 화자로 구성되어 있다. 화자확인 실험을 위해 이 음성 데이터를 8KHz로 다운 샘플링하고 에너지를 이용하여 자동으로 끝점을 검출한 후 사용하였다.

실험에 사용한 데이터는 화자가 20개의 4연 숫자를 4회씩 발성한 음성 데이터로써 각각의 4연 숫자에 대한 화자종속 모델을 만들기 위해 3회의 음성 데이터는 훈련에 사용하고 나머지 1회의 음성 데이터는 화자확인에 사용하도록 했다.

실험에 사용하는 음성 특징 파라미터로는 10차 LPC cepstrum과 정규화된 log 에너지를 파라미터를 사용한다.

화자확인 실험에 사용된 HMM방법은 whole word 모델을 사용하였고 모델의 상태 수와 분포는 상태별 분포를 얻기 위한 음성 분할은 2절에서 설명한 방법을 사용하였다. 화자확인 대상 어휘 당 3번의 훈련 데이터만을 사용하는 관계로 분산 추정이 잘 되지 않으므로 각 상태별 분산을 따로 구하지 않고 입력 음성 전체의 분산을 공통으로 사용하였다. <표 1>은 segmental K means 방법과 본 논문에서 사용한 level building DTW에 기반한 음성 분할 방법을 사용하여 생성한 사용자 모델을 사용하였을 때의 Equal Error Rate (EER)과 각각의 사용자 모델을 이용하여 Viterbi decoding을 했을 때 사용자 음성의 평균 log likelihood를 나타낸 것이다.

<표 1> HMM 기반의 화자확인 시스템의 EER(%)

	EER(%)	사용자 음성의 평균 log likelihood
HMM (Segmental K means 사용)	1.92	-16.14
HMM (Level Building DTW사용)	2.20	-16.29
HMM (Level Building DTW, DTW사용)	1.44	-15.98
HMM (LB DTW, DTW, 에너지 가중치 사용)	1.35	N/A

<표 1>에서 보는 바와 같이 segmental K means 알고리즘을 이용한 HMM 방법보다는 본 논문에서 사용한 level building DTW 알고리즘을 이용한 HMM 방식이

더 좋은 성능을 보였다. 이와 함께 사용자 음성의 평균 log likelihood 값에서 보듯이 훈련 과정에서 제한된 데이터만을 사용할 때는 통계적인 방법을 이용하는 segmental K means 알고리즘보다는 개별 음성의 특징 벡터를 이용한 분할 방법이 좀 더 사용자의 음성을 잘 표현해 줄 수 있음을 보여준다. 마지막으로 관측 확률에 에너지 기반의 가중치를 사용해 화자에 대한 정보가 많이 존재하는 유성음 구간에서의 확률값을 강조해 줌으로써 화자 간의 분별력을 조금 더 높일 수 있었다.

5. 결 론

본 논문에서는 적은 훈련 데이터로도 높은 성능을 얻을 수 있도록 level building DTW 기반의 화자인식 알고리즘을 제안하였다. 이의 검증을 위해 국어공학센터의 한국어 4연 숫자 음성 DB를 사용하여 컴퓨터 시뮬레이션을 수행하였으며, 3번의 훈련 데이터만을 이용하여 화자종속 모델을 만들고 테스트하였다. 적은 훈련 데이터를 사용하는 경우 HMM 모델의 생성을 위한 음성 분할 방법으로 기존의 segmental K means 알고리즘보다 level building DTW를 이용한 음성 분할이 더 효과적임을 확인하였고, 정규화된 에너지를 이용하여 화자의 정보가 많은 유성음 구간을 강조함으로써 화자 간의 분별력이 조금 더 향상되었다. 앞으로 보다 나은 성능을 얻기 위해서 음성 분할 시 최적의 구간 개수를 추정하는 방법에 대한 연구와 더불어서 일관성 있는 분할을 위한 거리 측정 방법에 대한 연구가 필요하다고 판단된다.

참 고 문 헌

[1] S. Furui, "An overview of speaker recognition technology", *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification,* pp.1-9, Apr. 1994.

[2] H. Sakoe, S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE, Trans. Acoust, Speech, Signal Processing,* Vol. ASSP-26, No. 1, PP.91-97, Feb. 1978.

[3] L. R. Rabiner, B. H. Juang et al., "Recognition of isolated digits using hidden markov models with continuous mixture densities", *AT & T Technical Journal,* Vol. 64, No. 6, pp.1211-1234, July-August 1985.

[4] M. Sharma, R. Mammone, "Blind speech segmentation: automatic segmentation of speech without linguistic knowledge," *Proc. ICSLP,* Vol. 2, 1996.

[5] L. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition,* New Jersey: Prentice-Hall International Inc., 1993.

[6] *Korean speech data base CD-ROM,* 국어공학센터, 1998.

접수일자: 2003년 8월 25일
게재결정: 2003년 9월 17일

▶ 김정곤(Jung-Gon Kim)
주소: 609-735 부산광역시 금정구 장전동 산 30번지 부산대학교
소속: 부산대학교 공과대학 전자공학과 음성통신 연구실
전화: 051) 510-1704
E-mail: gnome@pusan.ac.kr

▶ 김형순(Hyung Soon Kim)
주소: 609-735 부산광역시 금정구 장전동 산 30번지 부산대학교
소속: 부산대학교 공과대학 전자공학과 음성통신 연구실
전화: 051) 510-2452
E-mail: kimhs@pusan.ac.kr