음소 특성 정규화를 통한 화자 변화 검출*

김형순(부산대), 박혜영(부산대), 박선영(부산대)

<차 례>

<Abstract>

Speaker Change Detection by Normalization of Phonetic Characteristics

Hyung Soon Kim, Hae Young Park, Sun Young Park

Speaker change detection is to detect automatically a point of time at which speaker was replaced. Since feature parameters used for speaker change detection depend not only on speaker characteristics but also on phonetic characteristics, spoken contents included in the feature parameters inevitably causes performance degradation of speaker change detection. In this paper, to alleviate this problem, a method to normalize phonetic variations in speech feature parameters is proposed for emphasizing changes due to speaker characteristics. Experimental results show that the proposed method improves the performance of speaker change detection.

* Keywords: speaker change detection, speaker characteristics, normalization of phonetic characteristics

* 본 논문은 한국과학재단 목적기초연구(R01-2000-000-00275-0) 지원으로 수행되었음.

1. 서 론

정보화의 진전과 더불어 다루어야 할 멀티미디어 데이터의 규모가 엄청나게 증가하고 있으며, 이러한 많은 데이터 중에서 원하는 정보를 효과적으로 추출하기 위해 자동화된 데이터 인덱싱(indexing) 시스템과 검색 시스템의 필요성이 증대되고 있다[1]. 음성 데이터로부터 원하는 정보를 추출하는 일은 문자 데이터의 경우와 비교할 때 훨씬 더 어려운 일이다. 굳이 예를 들지 않더라도 발성 내용을 인식하거나 특정 화자의 발성을 검출하는 것이 텍스트 문서에서 단어 열이나 이름을 검출하는 것보다 훨씬 더 어렵다. 그리고 텍스트 문서를 읽는 것보다 음성 데이터 전체를 듣는 것이 더 많은 시간이 걸린다. 따라서 음성 정보의 효율적인 검색을 위해서는 음성 데이터의 원하는 부분에 직접 접근 할 수 있는 방법이 필요하다.

화자 기반 인덱싱 시스템의 목적은 누가, 언제 발성했는지를 색인화하는 것이다. 즉 입력 음성 신호를 단일 화자의 발성으로 구성되도록 분할하고 이를 각 화자별로 분류한다. 이러한 인덱싱 시스템은 뉴스 자막 시스템[2], 자동 음성 메시지 관리[3], 화자 추적[4] 등 많은 분야에 이용된다. 방송 뉴스 자막 시스템에서는 다양한 화자들의 발성으로 훈련된 모델을 사용하여 발성 내용을 인식하기 때문에 인식 성능이 상대적으로 저조하며, 인식 성능 향상을 위해서는 개별 화자에 적응된 모델을 사용할 필요가 있다. 화자 기반 인덱싱 시스템은 화자 적응을 위해 입력 음성 신호를 개별 화자의 발성으로 분할하고 분류하는 역할을 한다.

화자 기반 인덱싱 시스템은 크게 두 부분으로 구성된다. 먼저 한 화자가 발성한 부분만이 포함되도록 화자 변화 부분을 검출하여 입력 음성 신호를 분할하는 화자 기반 분할 단계이다. 그 다음은 분할된 부분들을 동일 화자가 발성한 부분으로 화자별로 묶어 주는 화자 기반 분류 단계이다. 본 논문에서는 그 첫 번째 단계인 화자 변화 구간 검출에 대해 검토하였다.

화자 변화 구간 검출에 사용되는 알고리즘은 Generalized Likelihood Ratio (GLR) 기반 검출 방법[5], Bayesian Information Criterion (BIC) 기반 검출 방법[1], 이들 두 방법을 접목시킨 DISTBIC 기반 검출 방법[6], 그리고 Kullback-Leibler (KL) 거리 기반 검출 방법[7] 등이 있다. 그러나 이들 방법들에 사용되는 음성 특징 파라미터들이 화자 특성 이외에도 발성 내용, 즉 음성학적 특성에도 영향을 받기 때문에 발성 내용 차이에 의한 변이가 화자 변화 검출의 성능을 저하시킨다. 본 논문에서는 이 문제의 해결을 위해서 발성 내용에 포함된 음소 특성을 정규화시킴으로써 각 화자의 개별 특성을 강조하는 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 화자 변화 검출을 위해 기존에 사용되는 알고리즘에 대해 설명하고, 3장에서는 제안된 방법을 기술한다. 4장에서 실험에 사용한 데이터베이스 및 실험 결과를 언급하고, 마지막으로 5장에서 결론을 맺는다.

2. 화자 변화 구간 검출 알고리즘

화자 변화 구간 검출은 음성 데이터에서 발성 화자가 바뀌는 시점을 자동적으로 찾아내는 것이다. 일반적으로 화자 변화 검출에 사용되는 거리 기반의 화자 변화 검출의 원리는 다음과 같다. 먼저 화자 변화 여부를 판단하고자 하는 시점의 양쪽으로 인접한 두 분석 윈도우를 설정하여 각각의 분포 특성을 모델링한다. 그리고 그 모델들 간의 거리를 측정하여 그 값이 특정 문턱치 이상이면 그 지점을 화자 변화 구간 후보로 검출하게 된다. 여기서 분석 윈도우는 신호에서 프레임별로 특징 벡터를 구한 후, 일정한 수의 특징 벡터를 포함하도록 나눈 부분을 말한다. 분석 윈도우의 크기가 크면 윈도우 내에 포함된 프레임 수가 많아서 통계적 추정에 유리하나, 한 화자의 발성 길이 등을 고려하여 적절하게 조절해야 한다. 이하에 본 논문에서 검토한 거리 기반의 대표적인 화자 변화 검출 방식들에 대해 간략히 설명한다.

2.1. GLR 기반 검출 방법

GLR 방법은 두 개의 분석 윈도우 각각을 가우시안(Gaussian) 분포로 모델링한 것과 두 분석 윈도우를 합한 전체를 하나의 가우시안 분포로 모델링한 것의 비를 이용하는 방법이다[5]. χ_i는 i번째 프레임의 특징 파라미터 벡터라고 할 때, 두 분석 윈도우 $\chi_1=\{x_1,\cdots,x_i\}$, $\chi_2=\{x_{1+1},\cdots,x_N\}$가 있다면 시간 i에서 화자 변화가 일어났는지를 확인하기 위해 다음과 같은 두 가지 가설을 세운다.

가설 1: 두 분석 윈도우는 한 화자에 의해 발성된 것으로, 하나의 가우시안 분포로 표현된다.

가설 2: 두 분석 윈도우는 서로 다른 화자에 의해 발성되었으며, 각각 다른 가우시안 분포로 표현된다.

가설 1, 2를 검증하기 위한 GLR 값은 가설 1과 가설 2의 우도(likelihood)의 비로 다음 식과 같이 정의된다.

$$R=\frac{L(\chi,N(\mu_\chi,\Sigma_\chi))}{L(\chi_1,N(\mu_{\chi_1},\Sigma_{\chi_1}))L(\chi_2,N(\mu_{\chi_2},\Sigma_{\chi_2}))} \tag{1}$$

이 때, R값이 크면 가설 1에 가깝고, R값이 작으면 가설 2에 가깝게 된다. 최종적으로 GLR 거리는 GLR 값의 로그 값으로 계산된다.

$$d_{GLR} = -\log R \tag{2}$$

일정 크기의 인접한 두 분석 윈도우를 시간에 따라 이동시키면서 d_{GLR}을 계산한다. GLR 기반 검출 방법에서는 GLR값이 GLR 거리 곡선의 국소 최대값이면서 주위의 국소 최소값과의 차이가 전체 GLR 거리에서 구한 표준편차의 일정 비율보다 크면 그 시점을 화자 변화가 일어난 지점으로 검출한다.

2.2. BIC 기반 검출 방법

BIC 기반 검출도 GLR을 이용한 검출 방법과 유사하며, 인접한 두 분석 윈도우를 모델링해서 분포의 차이를 비교한다[1]. GLR 방법과의 가장 큰 차이는 BIC 기반 방법에서는 적용 모델의 복잡성에 의한 가중치가 적용된다는 점이다.

$\chi_1=\{x_1, \cdots, x_i\}$와 $\chi_2=\{x_{1+1}, \cdots, x_N\}$가 두 분석 윈도우이고, 각각의 프레임 수를 N_{χ_1}, N_{χ_2}라고 하면 가설 1과 가설 2 사이의 우도 비는 다음과 같다.

$$R(i) = \frac{N_\chi}{2}\log|\Sigma_\chi| - \frac{N_{\chi_1}}{2}\log|\Sigma_{\chi_1}| - \frac{N_{\chi_2}}{2}\log|\Sigma_{\chi_2}| \tag{3}$$

가설 1과 가설 2에 해당하는 모델들 사이의 BIC 값의 차이는 다음 식 (4)와 같이 주어진다.

$$\Delta BIC(i) = -R(i) + \lambda P \tag{4}$$

여기서, 적용모델의 복잡성 P는 다음 식과 같이 표현되며,

$$P = 0.5(p + 0.5p(p+1))\log N_\chi \tag{5}$$

이때 p는 특징 파라미타의 차수이다. BIC 방법은 세 단계로 이루어지는데, 1 단계에서는 큰 분석 윈도우를 이용하여 후보 지점을 찾고, 2 단계에서는 그 후보 지역을 중심으로 더 작은 분석 윈도우를 이용해 검출의 정밀도를 높이며, 마지막 3 단계에서는 2 단계에서 검출된 후보들을 검증하여 화자 변화 구간을 검출한다. 구체적으로 1 단계에서는 화자 변화의 대략적인 위치를 결정하기 위해 두 인

접한 분석 윈도우 사이의 경계를 옮기면서 구한 $\triangle BIC$의 값 중에 최대값이 음수 값을 가지면 화자 변화의 대략적인 위치로 결정된다. 2 단계는 1 단계에서 나온 후보 지점을 중심으로 1 단계보다 작은 분석 윈도우를 이용해 동일한 방식으로 후보 지점을 검출한다. 3 단계는 1, 2 단계를 통해 구한 후보들을 확인해서 조건에 만족하지 않는 후보들을 버리는 단계로 후보들 사이를 분석 윈도우로 잡고 $\triangle BIC$ 값을 구하여 이 값이 음수이면 i를 화자 변화 시점으로 검출한다.

2.3. DISTBIC 기반 검출 방법

BIC 알고리즘은 추정에 사용되는 특징 벡터의 열이 짧으면 이용할 수 있는 정보가 작아서 가우시안 모델의 올바른 추정이 어렵다. DISTBIC 기반 검출 방법에서는 길이에 의존적이지 않은 GLR 방법을 먼저 수행해 화자 변화 후보들을 정하고, 다음 단계로 BIC 기반의 3단계를 통해 화자 변화 구간을 검출한다[6].

2.4. KL 거리 기반 검출 방법

KL 거리는 분포 특성의 차이를 뜻하며 두 확률밀도 함수 사이의 거리를 나타낸다. 두 확률밀도 함수를 각각 $P_A(x)$와 $P_B(x)$라고 하면 KL 거리는 아래와 같다.

$$KL(A;B) = \int P_A(x) \log \frac{P_A(x)}{P_B(x)} dx \tag{6}$$

식 (6)으로 표현되는 KL 거리는 비대칭이라서 거리 표현으로 부적합하기 때문에 대칭적인 거리 KL2를 다음과 같이 정의한다.

$$KL2(A;B) = KL(A;B) + KL(B;A) \tag{7}$$

이 때, $P_A(x)$와 $P_B(x)$가 가우시안 분포이면 KL2는 다음과 같이 정리될 수 있다.

$$KL2(A;B) = \frac{1}{2}(\mu_A - \mu_B)^T(\Sigma_A^{-1} + \Sigma_B^{-1})(\mu_A - \mu_B) + \frac{1}{2} tr(\Sigma_A^{-1}\Sigma_B + \Sigma_B^{-1}\Sigma_A - 2I) \tag{8}$$

여기서 μ_A, μ_B는 분포 $P_A(x)$ 및 $P_B(x)$의 평균 벡터이고 $\sum_A$, $\sum_B$는 이

들의 공분산 행렬이다. 만약 $P_A(x)$와 $P_B(x)$가 동일한 확률밀도 함수이면 KL 거리가 0이 된다. 동일 길이를 가지는 인접한 두 분석 윈도우를 오디오 신호의 모든 점을 경계로 옮기면서 KL 거리를 구한다. 그리고 일정 크기의 탐색 윈도우의 내에서 KL2 거리가 최대값을 가지는 곳을 분할 구간의 경계로 잡는다.

CMU에서 개발한 화자 분할 시스템에서는 KL 기반 방법으로 후보 지점을 검출하고, 화자들의 발성은 대부분 묵음을 중심으로 구분되므로 후보 지점을 중심으로 일정 크기 내에서 묵음 구간을 찾고 이 구간을 화자 변화 지점으로 검출한다 [7].

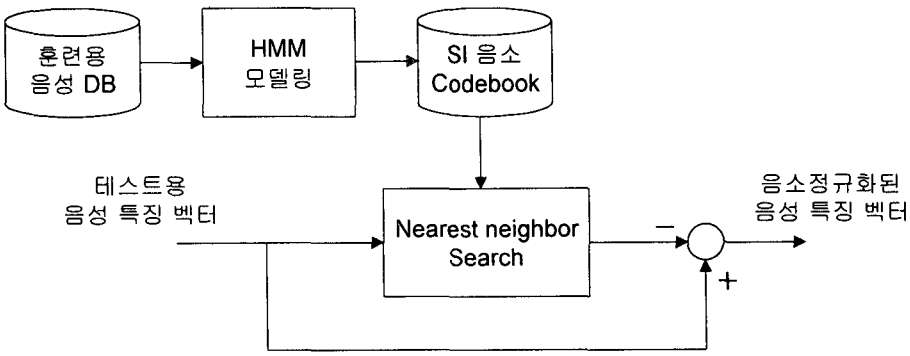
3. 음소 특성 정규화를 통한 화자 변화 검출

화자 변화 구간을 검출하기 위해서는 화자의 개별성에 의한 차이를 추출해야 하므로 화자들이 동일한 내용을 발성할 때 가장 좋은 결과를 보일 수 있다. 그러나 실제 환경에서는 화자들이 동일한 내용을 발성하지 않으므로 다른 발성 내용에 의한 정보가 특징 벡터에 포함되어 화자 변화 구간 검출 성능을 저하시킨다. 그러므로 음소 특성을 정규화하여 화자 개별성 정보를 강조하는 것이다. 본 논문에서는 이러한 음소 특성 정규화를 위해, 미리 많은 화자의 발성 데이터로부터 모델링하여 음소 특성을 대표할 수 있는 화자 독립적인 음소별 평균 벡터를 사용한다. 특정한 화자의 음성 특징 벡터는 그 프레임에 해당하는 음소의 화자 독립적인 평균 벡터와 비교할 때 화자 특성에 따른 변화량을 가지게 되므로, 특징 벡터에서 화자 독립적인 해당 음소의 평균 벡터를 빼줌으로써 화자별 특성을 강조하도록 하였다.

먼저, 음소 특성을 대표하는 화자 독립 음소별 평균 벡터를 구하기 위해, 우리말 유사 음소(phone-like units) 45 개에 묵음을 포함하여 총 46개 문맥 독립형 음소 모델을 사용하였다. 여러 화자로 구성된 훈련 DB를 이용하여 연속확률분포를 가지는 HMM (Hidden Markov Model)으로 화자 독립 모델을 훈련하였다. 본 논문에서는 음소 모델링을 위해 Baum-Welch reestimation으로 모델링한 다음 segmental K-means로 다시 모델링하였다. 실험 결과, 이 방법이 Baum-Welch reestimation만으로 모델링 한 방법보다 음소 특성 정규화 성능이 우수하였다.

본 논문에서 제안한 음소 특성 정규화 방법은 <그림 1>과 같다. 먼저, 화자 독립적인 음소 모델링에 의해 구해진 각 상태당 평균 벡터를 벡터 양자화 codebook의 원소(codeword)로 삼는다. 실제 테스트 음성이 들어오면 각 프레임별로 음성의 특징 벡터와 미리 구해진 각 codeword들 사이의 Euclidian distance를 구하여 가장 가까운 codeword를 구한다. 음소인식 성능이 완벽할 수 없으나, 이렇게 구해진 codeword를 각 특징 벡터의 화자 독립적인 음소 평균 벡터로 간주한다. 그리고 매

프레임별 음성 특징 벡터로부터 해당 음소 평균 벡터를 빼줌으로써 음소 특성 정규화를 수행한다.



<그림 1> 특징 벡터에서 음소 특성을 정규화하는 방법

4. 실험 및 결과

본 논문에서는 성능 평가를 위한 DB로 원광대 국어공학센터에서 구축한 PBS (Phonetically Balanced Sentence) DB와 자체적으로 공중파 방송 뉴스를 녹음하여 구축한 DB를 사용하였다. PBS DB중에서 남성화자 5명과 여성화자 5명이 발성한 201개 문장을 연결해서 화자 변화 구간 검출 실험에 사용하였다. 이 음성 데이터의 총 길이는 약 20분으로 한 화자의 연결 발성 길이는 평균 약 6초이다. 그리고 또 하나의 DB는 2001년 9월에 방송된 실제 공중파 방송 뉴스를 녹음하였으며, 이 데이터의 길이는 약 58분이고 한 화자의 연결 발성 길이는 평균 약 16.5초이다. 모든 음성 데이터를 16kHz로 샘플링하여 16bit로 양자화하였으며, 특징 파라미타로는 12차 MFCC를 사용하였다.

화자 변화 검출의 성능은 오검출률(False Alarm Rate (FAR))과 미검출률(Missed Detection Rate (MDR))로 나타내며, 이들은 다음과 같이 정의된다[9].

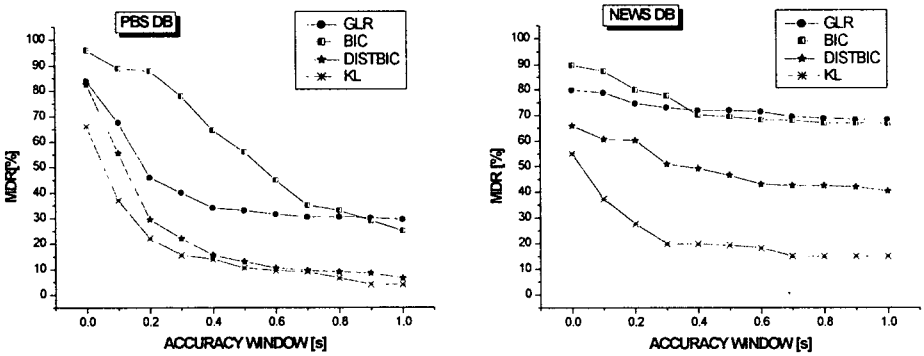
$$FAR = \frac{FA}{N + FA} \times 100 \quad (\%) \tag{9}$$

$$MDR = \frac{MD}{N} \times 100 \quad (\%) \tag{10}$$

여기서 N은 실제 화자 변화의 수이고, FA는 존재하지 않는 화자 변화를 잘못 검출한 경우의 수이며, MD는 실제 화자 변화를 검출하지 못한 경우의 수를 말한다. 실제로 화자 기반 인덱싱 시스템에서는 화자 분할 단계 이후에 화자 분류 단계를 통해 잘못 나누어진 구간들을 통합함으로써 오검출률을 낮출 수 있다. 따라서 본 논문의 연구 주제인 화자 분할 단계에서는 오검출률을 낮추는 것보다 미검출률을 낮추는 것이 더욱 중요하다.

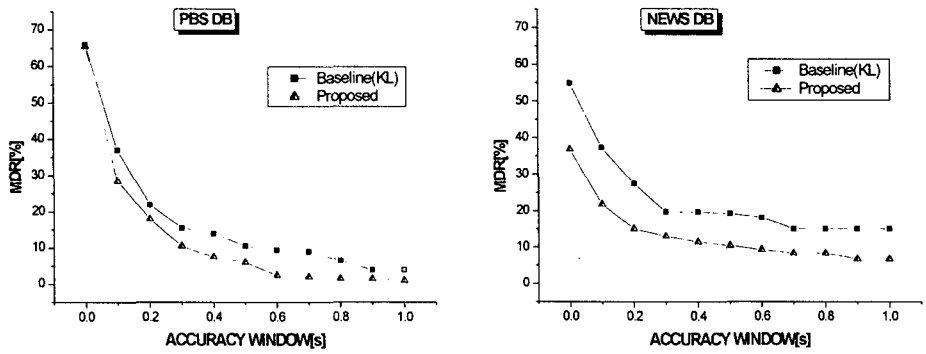
성능 평가는 사람이 청취를 통해 검출한 화자 변화 지점을 기준으로 하여 검출여부를 결정한다. 그러나, 동일한 데이터에 대해서 각 사람마다 다른 기준에 의해 화자 변화 구간을 검출할 수 있다. 이에 따라 사람에 의한 화자 변화 구간 검출에도 어느 정도의 차이가 생길 수 있으므로, 화자 변화 구간과 검출함 화자 변화 시점의 차이에 대해 어느 정도의 오차는 용납하고 성능을 평가할 필요가 있다. 그리고 일정한 오차 범위 안에 화자 변화가 얼마나 검출되는가 하는 것도 의미있는 정보이므로 실제 화자 변화 구간을 중심으로 일정 크기의 윈도우를 씌워 그 크기 내에서 검출되면 화자 변화가 검출된 것으로 판단한다. 이 일정 크기의 윈도우를 정확성 윈도우(accuracy window)라고 부른다[10].

<그림 2>는 기존의 여러 가지 화자 변화 검출 방법들을 PBS DB와 뉴스 DB를 사용하여 오검출률(FAR)을 40%로 고정시킨 상태에서, 정확성 윈도우의 크기를 변화하면서 미검출률(MDR)을 비교 실험한 결과이다. 그림에서 기존의 여러 방법 중에서 KL 방법의 성능이 가장 좋은 것을 알 수 있다. 이에 따라 음소 특성 정규화에 의한 화자 변화 검출 실험은 KL 방법에 대해서만 적용하기로 하였다.



<그림 2> 기존의 화자 변화 검출 방법들의 성능 비교

음소 특성 정규화에 사용하기 위한 화자 독립 음소 평균 벡터를 구하기 위해서, PBS DB중에 남성 10명과 여성 10명이 발성한 3,156 문장을 훈련 데이터로 사용하였다. <그림 3>에 기존의 KL 방법과 음소 특성을 정규화한 특징 벡터를 이용한 KL 방법의 성능 비교 결과가 나타나 있다. 음소 특성을 정규화한 제안된 방법이 음소 특성을 정규화하지 않은 기존 방법에 비해 화자의 개별 특성이 상대적으로 강조되어 미검출률(MDR)이 감소하였음을 확인할 수 있다.



<그림 3> 음소정보 정규화를 통한 성능향상

5. 결 론

본 논문에서는 화자 변화 구간의 검출 성능을 높이기 위해서 음성 특징 벡터로부터 음소 특성을 정규화하여 화자 특성을 강조하는 방법을 제안하였다. 본 논문에서 매우 단순한 음소 특성 정규화 방법을 사용했음에도 불구하고, 실험 결과 기존의 방법에 비해 성능이 향상되었음을 확인할 수 있었다. 앞으로 음소 특성 정규화 방법의 정확성을 향상시킬 수 있는 방법에 대한 추가적인 연구가 필요하다고 판단되며, 화자 변화 검출 결과를 바탕으로 단일 화자별 분류 방법에 대한 연구도 계속될 예정이다.

참 고 문 헌

[1] S. S. Chen, P. S. Gopalakrishnan, "Speaker, environment, and channel change detection and clustering via the Bayesian information criterion", *in Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, pp.127-132, 1998.

[2] P. C. Woodland, M. J. F. Gales, "The development of the 1996 HTK broadcast news transcription system", *DARPA Speech Recognition Workshop*, pp.73-78, 1997.

[3] D. Reynolds, E. Singer, "Blind clustering of speech utterances based on speaker and language characteristics", *Proc. of Int. Conf. on Spoken Language Processing*, pp.3193-3196, 1998.

[4] A. E. Rosenberg, Q. Huang., "Speaker detection in broadcast speech databases", *Proc. of Int. Conf. on Spoken Language Processing*, Vol. 4, pp.1339-1342, 1998.

[5] P. H. Gish, M. H. Siu, "Segregation of speakers for speech recognition and speaker identification", *Proc. of Int. Conf. Acoustics, Speech, and Signal Processing*, Vol. 2, pp.873-876, 1991.

[6] P. Delacourt, C. J. Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing", *Speech Communication,* Vol. 32, pp.111-126, Sep. 2000.

[7] M. A. Siegler, U. Jain, "Automatic segmentation, classification and clustering of broadcast news audio", *Proc. of the DARPA Speech Recognition Workshop*, pp.97-99, 1997.

[8] J. W. Hung, H. M. Wang, L.-S. Lee, "Automatic metric-based speech segmentation for broadcast news via principal component analysis", *Int. Conf. on Spoken Language Processing*, Vol. 4, pp.1347-1350, 1998.

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

[10] D. Liu, F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing", *Proc. of Eurospeech*, Vol. 3, pp.1031-1034, 1999.

접수일자: 2003년 8월 25일
게재결정: 2003년 9월 17일

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

▶ 박혜영(Hae Young Park)
주소: 609-735 부산광역시 금정구 장전동 산 30번지 부산대학교
소속: 부산대학교 전자공학과 음성통신 실험실
전화: 051) 510-1704
E-mail: phyoe@pusan.ac.kr

▶ 박선영(Sun Young Park)
주소: 609-735 부산광역시 금정구 장전동 산 30번지 부산대학교
소속: 부산대학교 전자공학과 음성통신 실험실
전화: 051) 510-1704
E-mail: sunypark@pusan.ac.kr