음성학적 정보의 제거를 통한 화자변화 구간 검출

박선영, 김형순
부산대학교 전자공학과

Speaker Change Detection by Removing Phonetic Information

Sun Young Park, Hyung Soon Kim
Dept. of Electronics Engineering, Pusan National University
E-mail:{sunypark, kimhs}@hyowon.pusan.ac.kr

요 약

본 논문에서는 음성 신호에서 발성 화자가 바뀌는 시점을 자동적으로 찾아내는 화자변화 구간 검출에 대하여 연구하였다. 화자변화 검출을 위해서는 음성 신호에 나타나는 화자 개별성에 의한 차이만 비교해야 하는데 실제 환경에서는 화자들이 동일한 내용의 발성을 하지 않으므로 다른 발성내용에 의한 정보가 포함되어 검출 성능을 저하시킨다. 그러므로 각 화자의 개별특성만 강조되도록 발성내용에 포함된 음성학적 정보의 영향을 제거하는 방법을 통해 검출 성능을 향상시켰다.

1. 서 론

정보통신의 발전과 인터넷의 보급으로 인해 멀티미디어 정보가 증가함에 따라 텍스트 정보와 영상 정보, 오디오 정보 등의 미디어별로 빠르고 효과적으로 정보를 추출하는 멀티미디어 정보 검색 시스템의 필요성이 증대되고 있다. 이 중 오디오 정보 검색의 전처리부로서 특정화자의 발성내용을 찾거나 내용검색부에서 화자적응을 위한 화자별 발성 내용을 분류하는 데 화자변화 구간 검출이 사용된다.

화자변화 검출에 사용되는 알고리즘은 Generalized Likelihood Ratio(GLR) 기반 검출방법[1], Baysian Information Criterion(BIC) 기반 검출방법[2], 그리고 이들 두 방법을 접목시킨 DISTBIC기반 검출방법[3], Kullback-Leibler(KL) 거리 기반 검출방법[4]이 있다. 그리고 Carnegie Mellon University(CMU)에서 개발한 화자기반 분할 시스템에서는 KL 거리 기반 방법으로 후보를 검출하고 후보지점을 중심으로 silence를 검출하여 silence 구간을 화자변화 구간으로 검출하였다[5].

본 논문에서는 화자변화 검출에 사용되는 기존 알고리즘의 성능을 비교하였고, 성능 향상을 위해 각 화자의 개별특성만 강조되도록 발성내용에 포함된 음성학적 정보의 영향을 제거하는 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 2절에서는 화자변화 구간 알고리즘으로 기존에 사용되는 알고리즘에 대해 설명하고 3절에서는 본 논문에서 제안한 음성학적 정보의 영향을 제거하는 방법에 대해 설명한다. 4절에서는 실험결과를 언급하고 5절에서 결론을 맺는다.

2. 화자변화 구간 검출 알고리즘

화자변화 구간 검출은 음성신호에서 발성화자가 바뀌는 시점을 자동적으로 찾아내는 것이다. 화자변화 검출의 원리는 그림 1과 같이 음성 신호를 따라서 변화하는 두 인접한 분석 윈도우들 사이의 분포 특성의 차이가 최대가 되는 지점을 찾는 것이다. 분석 윈도우 $\chi_1=\{x_1,\ldots,x_i\}, \chi_2=\{x_{i+1},\ldots,x_N\}$ 가 있다면 시간 i에서 화자 변화가 일어났는지를 확인하기 위해 두 가지 가설을 세운다.

가설 1: 두 분석 윈도우는 한 화자에 의해 발성되었다. 그래서 하나의 가우시안 모델로 모델링된다.

$$\chi=\chi_1\cup\chi_2 \sim N(\mu_\chi,\Sigma_\chi)$$

가설 2: 두 분석 윈도우가 다른 화자에 의해 발성되었다. 그래서 각각 다른 가우시안 모델로 모델링된다.

$$\chi_1 \sim N(\mu_{\chi_1},\Sigma_{\chi_1}),\ \chi_2 \sim N(\mu_{\chi_2},\Sigma_{\chi_2})$$

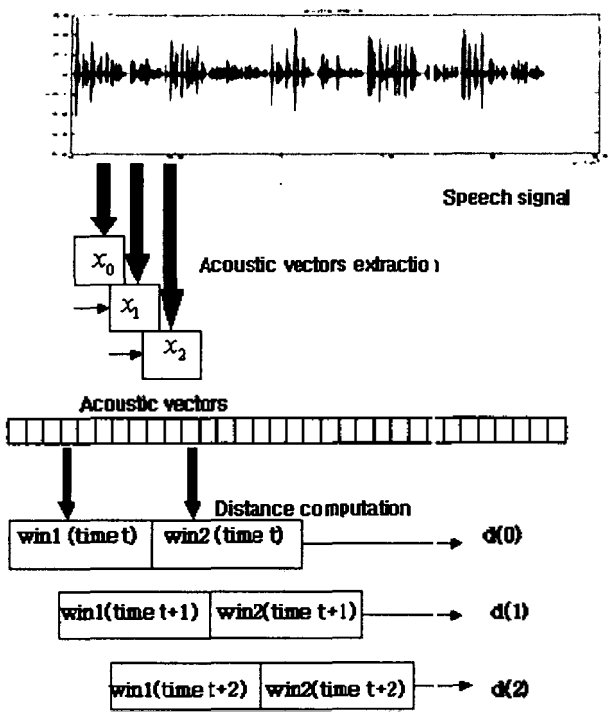

그림 1. 거리기반 화자변화 검출알고리즘의 기본구성

2. 1. GLR기반 검출 방법[1]

가설 1, 2를 검정하기 위해서 GLR은 가설 1과 가설 2의 likelihood의 비로 다음 식과 같이 정의된다.

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

GLR의 거리는 GLR의 로그 값으로 계산된다.

$$d_{GLR}=-\log R \quad (2)$$

일정크기의 분석윈도우를 음성 신호를 따라 이동시키면서 d_{GLR}을 계산한다. 높은 R값은 가설 1에 적합하고 낮은 R값은 가설 2에 적합해서 GLR 거리 곡선의 국소 최대값이면서 주위의 국소 최소값과의 차이가 전체 GLR 거리에서 구한 표준편차의 일정비율보다 크면 화자변화가 일어난 지점으로 검출된다[1].

2. 2. BIC 기반 검출 방법[2]

BIC 기반 방법에서는 적용 모델 복잡성에 의한 가중치가 적용된다. $\chi_1=\{x_1,\ldots,x_i\}$와 $\chi_2=\{x_{i+1},\ldots,x_N\}$는 분석윈도우이고, 프레임 수를 N_{χ_1},N_{χ_2}라고 하면 가설1과 가설2 사이의 likelihood 비는 다음과 같다.

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

가설1과 가설2에 해당하는 모델들 사이의 BIC 값의 차이는 다음 식으로 주어진다.

$$\Delta BIC(i)=-R(i)+\lambda P \quad (4)$$

여기서, 적용모델의 복잡성은 $P=0.5(p+0.5p(p+1))\log N_\chi$이고 p는 특징 파라미타의 차수이다. BIC 방법은 세 단계로 이루어지는데 1단계에서는 큰 분석 윈도우를 이용하여 후보지점을 찾고 2단계에서는 그 후보지역을 중심으로 더 작은 분석윈도우를 이용해 검출하며 3단계에서는 2단계의 후보들을 확인해서 화자변화구간을 검출한다. 구체적으로 1단계에서는 화자변화의 대략적인 위치를 결정하기 위해 두 인접한 분석 윈도우 사이의 경계를 옮기면서 구한 ΔBIC의 값 중에 최대값이 음수 값을 가지면 화자변화의 대략적인 위치로 결정된다. 2단계는 1단계에서 나온 후보지점을 중심으로 1단계보다 작은 분석 윈도우를 이용해 동일한 방식으로 후보지점을 검출한다. 3단계는 1, 2단계를 통해 구한 후보들을 확인해서 조건에 만족하지 않는 후보들을 버리는 단계로 후보들 사이를 분석 윈도우로 잡고 ΔBIC 값을 구하여 음수이면 i에서 화자변화가 일어났음이 확인하고 화자변화 구간으로 검출한다[2].

2.3.DISTBIC 기반 검출방법[3]

BIC 알고리즘은 추정에 사용되는 특징벡터의 열이 짧으면 이용할 수 있는 정보가 작아서 가우시안 모델의 올바른 추정이 어렵다. DISTBIC 기반 검출방법에서는 길이에 의존적이지 않은 GLR 방법을 먼저 수행 해 화자변화 후보들을 정하고, 다음 단계로 BIC 기반의 3단계를 통해 화자변화 구간을 검출한다.

2. 4. KL 거리 기반 검출방법[4]

KL 거리는 분포특성의 차이를 뜻하며 두 확률밀도 함수 사이의 거리를 나타낸다. 두 확률밀도 함수 각각을 P_A와 P_B라고 하면 KL 거리는 아래와 같다.

$$KL(A;B)=\int P_A(u)\log\frac{P_A(u)}{P_B(u)}du \quad (5)$$

식(5)는 비대칭이라서 거리의 표현이 될 수 없으므로, KL2를 다음과 같이 정의한다.

$$KL2(A;B)=KL(A;B)+KL(B;A) \quad (6)$$

P_A와 P_B가 가우시안 분포이면 KL2는 다음과 같다.

$$KL2(A;B)=\frac{\sigma_A^2}{\sigma_B^2}+\frac{\sigma_B^2}{\sigma_A^2}+(\mu_A-\mu_B)^2(\frac{1}{\sigma_A^2}+\frac{1}{\sigma_B^2}) \quad (7)$$

여기서 μ_A,μ_B는 분포 P_A및 P_B의 평균이고 σ_A,σ_B는 이들의 표준편차를 말한다. 만약 KL 거리가 0이면 P_A와 P_B는 동일한 확률밀도 함수임을 뜻한다. 그림 1처럼 두 분석 윈도우를 오디오 신호의 모든 점을 경계로 옮기면서 KL 거리를 구한다. 그리고 KL 거리 곡선에서 각 점을 중심으로 탐색 윈도우를 씌워서 그 크기 내에서 그 지점의 KL 거리가 최대값을 가지면 그 지점을 화자변화 구간으로 검출한다.

2.5. CMU 화자기반 분할 시스템[4]

CMU에서 개발한 화자기반 분할 시스템에서는 KL 기반 방법으로 후보지점을 검출하고, 화자들의 발성은 대부분 silence를 중심으로 구분되므로 후보지점을 중심으로 일정크기 내에서 silence 구간을 찾고 이 구간을 화자변화 지점으로 검출한다. 그림 2는 silence 검출을 위한 탐색구간과 윈도우를 그림으로 나타낸 것이다. KL 방법에서 검출한 후보를 중심으로 일정크기의 탐색 구간을 정하고 탐색 구간 내에 각 점을 중심으로 inner window와 outer window를 정한다. Silence 구간은 에너지의 변화가 작고 음성 구간보다 에너지가 작은 구간이므로 inner window의 에너지 변화가 일정 크기보다 작고, 음성 구간으로 추측되는 outer window의 최대에너지보다 inner window의 평균에너지가 일정 크기만큼 작으면 silence 구간으로 검출한다.

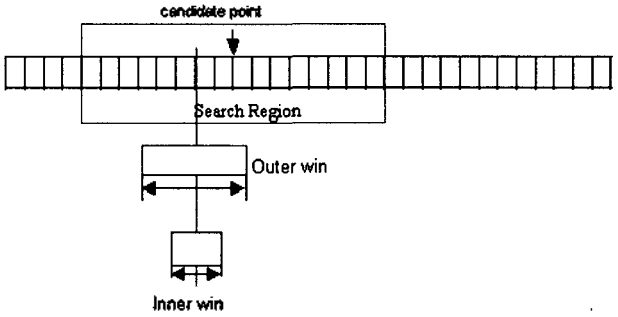

그림 2. CMU 시스템에서의 silence 검출 방법

3.발성 내용 정보를 고려한 특징 파라미터

화자변화 구간 검출에서 두 분석윈도우의 발성내용이 다르면 화자 개별성 뿐만 아니라 원하지 않는 음소 특성에 의해 분포 특성이 달라진다. 그러므로 음소 특성을 제거하여 화자 개별성 정보를 강조하는 것이 바람직하다. 예를 들어 그림 3의 한 분석 윈도우 내에 '학'이라는 발성이 들어 있고 다른 분석 윈도우 내에 '함'이라는 발성이 있다면 음소 구성의 차이가 분포 특성에 영향을 끼친다. 이러한 음소 특성을 제거하기 위해 여러 화자의 발성데이타로 음소 특성을 대표할 수 있는 음소별 평균벡터를 구한다. 한 화자의 데이터로 모델링한 음소 평균벡터는 여러 화자의 데이터로 모델링된 음소 평균벡터에서 화자별로 변화량을 가지므로 특징 파라미터에서 여러 화자의 데이터로 모델링된 음소 평균벡터를 제거하여 화자별 특성을 강조하였다. 실험에서는 음소별 평균벡터를 구하기 위해 PBS DB의 20명분을 사용하여 monophone 단위로 3 state의 GMM으로 모델링하였다. 테스트 DB의 특징 파라미터별로 가장 likelihood가 높은 모델을 구해 그 모델의 평균 벡터를 특징 파라미터에서 제거하였다. 이 방법에서는 테스트 DB의 특징 파라미터가 해당 모델의 음소가 아닌 데 그 모델의 likelihood가 높게 나와 그 모델의 평균 벡터를 제거할 수 있는 위험이 있다.

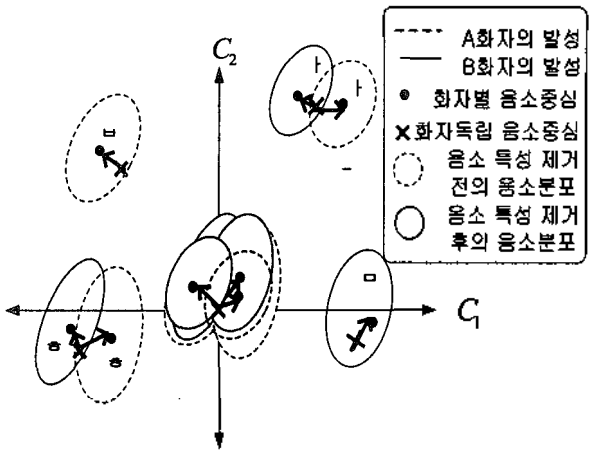

그림 3. 음소 특성 제거 전후의 음소 분포도의 예

4.실험 및 결과

본 논문에서는 성능평가를 위한 DB로 원광대 국어공학센터에서 구축한 PBS(Phonetically Balanced Sentence) DB와 공중파 방송을 녹음하여 사용하였다. PBS DB는 남성화자 5명과 여성화자 5명분의 201개 문장을 연결해서 화자변화 지점이 200개로 구성되었다. 이 음성 데이터의 총길이는 약 20분으로 한 화자의 연결발성 길이는 평균 약 6초이다. 그리고 실제 공중파 방송으로 2001년 9월 KBS뉴스를 녹음하여 사용하였다. 이 데이터의 길이는 약 58분으로 한 화자의 연결발성 길이는 평균 약 16.5초이다. 16kHz로 샘플링되었고 16bit로 양자화되었으며 특징 파라미타로 12차 MFCC와 에너지를 사용했다.

화자변화 검출의 성능은 False Alarm Rate(FAR) 과 Missed Detection Rate(MDR)로 나타낸다. False Alarm (FA)은 존재하지 않는 화자변화를 검출하는 것이고 Missed Detection(MD)은 존재하는 화자변화를 검출하지 못하는 것을 말한다[5].

$$FAR = 100 \times \frac{\text{FA의 수}}{\text{실제 화자변화의 수} + \text{FA의 수}} (\%) \quad (7)$$

$$MDR = 100 \times \frac{\text{MD 의 수}}{\text{실제 화자변화의 수}} (\%) \quad (8)$$

성능평가는 사람이 청취를 통해 검출한 화자변화시점을 기준으로 하여 검출여부를 결정한다. 그러나, 사람에 의한 화자변화 구간 검출에도 어느 정도의 오차가 발생할 수 있으므로, 화자변화 구간과 검출한 화자변화 시점의 차이에 대해 어느 정도의 오차는 용납하고 성능을 평가해야 한다. 그리고 일정한 오차 범위 안에 화자변화가 얼마나 검출되는가 하는 것도 의미 있는 정보이므로 실제 화자변화 구간을 중심으로 일정 크기의 윈도우를 씌워 그 크기 내에서 검출되면 화자변화가 검출된 것으로 판단하고, 이 일정 크기의 윈도우를 정확성 윈도우(accuracy window)라고 한다.

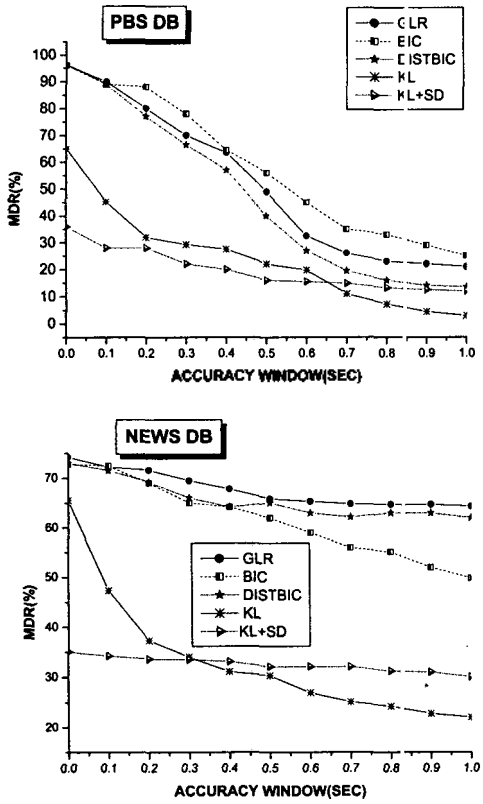

그림 4. 기존 방법의 성능 비교

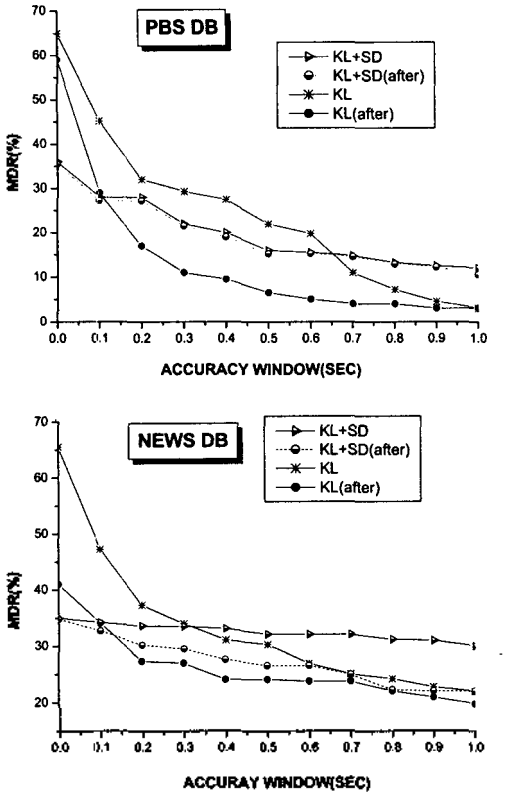

그림 5. 음소특성 제거 전후의 성능 비교

그림 4는 기존방법들을 PBS DB와 뉴스 DB를 사용하여 FAR을 50%로 고정하고, 정확성 윈도우의 크기를 변화하면서 실험한 결과이다. 그림에서 보는 바와 같이 기존의 방법 중에서 정확성 윈도우의 크기가 작을 때는 CMU시스템(그림에서 KL+SD), 클 때는 KL방법이 가장 좋은 성능을 보였다. 정확성 윈도우가 클 때 CMU시스템이 KL방법보다 좋지않은 결과를 보이는 것은 silence 검출 성능이 좋지않은 이유로 추정된다. 음소 특성을 제거한 특징 파라미터를 이용한 방법은 이들 두 방법에 대해서만 적용해 보았다. 그림 5에서 두 방법에 대해서 음소 특성을 제거하기 전보다 제거 후에 화자의 개별특성이 강조되어 MDR이 감소하였음을 확인할 수 있다.

5.결 론

본 논문에서는 음성 신호에서 화자변화 구간 검출에 대해서 연구하였다. 기존의 방식 중에서 정확성 윈도우의 크기가 작을 때는 CMU시스템, 클 때는 KL방법이 가장 성능이 좋은 결과를 보였으며, 본 논문에서 제안된 발성내용에 포함된 음성학적 정보의 영향을 제거하는 방법을 도입함으로써 화자 개별성이 강조되어 기존의 방식보다 성능이 개선되었다. 하지만 특징 파라미터가 해당하는 음소 모델을 잘못 찾아 잘못된 음소특성을 제거하는 오류를 범할 수 있으므로 향후 이 문제를 보완하기 위한 방법에 대해 추가적인 연구가 필요하다.

본 논문은 한국 과학 재단의 목적 기초 연구 수행 결과의 일부입니다.

참고문헌

[1] P. H. Gish and M. H. Siu, "Segregation of speakers for speech recognition and speaker identification," *in Proc. IEEE ICASSP*, pp. 873-876, 1991.

[2] S. S. Chen and P. S. Gopalakrishnan, "Speaker environment and channel change detection and clustering via the Bayesian information criterion," *in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

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

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

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

[6] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall Inc., 1993.