통계적 기법을 이용한 화자변화 검출 실험

A Speaker Change Detection Experiment that Uses a Statistical Method

이 경 록* · 김 진 영*
Kyong Rok Lee · Jin Young Kim

ABSTRACT

In this paper, we experimented with speaker change detection that uses a statistical method for NOD (News On Demand) service. A specified speaker's change can find out content of each data in speech if analysed because it means change of data contents in news data. Speaker change detection acts as preprocessor that divide input speech by speaker. This is an important preprocessor phase for speaker tracking. We detected speaker change using GLR(generalized likelihood ratio) distance base division and BIC (Bayesian information criterion) base division among matrix method. An experiment verified speaker change point using BIC base division after divide by speaker unit using GLR distance base method first. In the experimental result, FAR (False Alarm Rate) was 63.29 in high noise environment and FAR was 54.28 in low noise environment in MDR (Missed Detection Rate) 15% neighborhood.

Keyword: Speaker Change Detection, Matrix Method, GLR, BIC, NOD

1. 서 론

NOD 서비스는 사용자가 원하는 컨텐트의 뉴스 정보를 검색하여 제공하는 서비스를 말한다. 뉴스 데이터를 컨텐트별로 분할하고 이를 인덱싱하는 기술에는 비디오 인덱싱과 오디오 인덱싱 기술이 있다. 이 중 오디오 인덱싱은 음성이 뉴스 데이터에서 정보전달의 중요한 매개체라고 가정하고 이를 분석하여 컨텐트를 분류한다[1]. 뉴스 데이터는 특성상 각 컨텐트에 대한 정보가 특정화자(앵커, 앵커우먼 등)에 의해서 미리 제시되어진다. 이를 이용해서 입력 음성을 화자별로 분할하고 이 중 특정 화자의 음성만을 분석하여 컨텐트를 분석한다[2, 3]. 즉, 정보전달의 핵심부분인 앵커의 음성부분을 정확하게 분석한다면, 이어지는 리포터 음성부분의 컨텐트를 결정할 수 있다.

화자변화 검출은 다양한 형태의 뉴스 데이터로부터 유효한 결과값을 도출하기 위해서 다음과 같은 전제 조건을 갖는다. 첫째, 입력음성 신호에서의 총 화자 수를 사전에 알 수 없다. 이는 화자변화 검출이 출현 화자의 수에 독립적이어야 한다는 것을 의미한다. 둘째, 화자에

* 전남대학교 전자공학과, RRC HECS

대한 사전정보를 필요로 하지 않아야 한다. 이는 출현 화자에 대한 확률모델이나 성별 등에 대한 아무런 사전정보가 없다는 것을 의미하며, 화자 종속적인 방법론을 적용하는 게 곤란하다는 것을 뜻한다. 셋째는 실시간 분석을 필요로 하지 않는다. 이는 입력음성 신호를 오디오 인덱싱하기 위해서는 여러 가지 전처리 단계와 핵심어 인식, 컨텐트 결정 등을 필요로 한다. 이러한 이유로, 실시간 지원을 고려하지 않는다는 것이다.

화자변화에 기반하여 음성신호를 자동분할하는 화자 기반 분할 알고리듬에는 다음과 같은 3 가지 방법이 있다. 첫 번째 디코더 기반 분할은 입력음성을 연속음성 인식기 등을 이용하여 인식하고, 그 결과를 분석하여 분할하는 방법이다. 두 번째, 모델 기반 분할은 사전에 출현 화자들의 모델을 구축하고, 이를 이용해서 음성신호를 모델들의 수열로 변환하는 것이다. 세 번째, 메트릭스 기반 분할은 인접한 두 분석 윈도우간의 비유사도가 국소 최대가 되는 것을 검출하여 분할하는 것이다. 본 연구에서는 세 번째 메트릭스 기반 분할을 사용하였다[4, 5, 6].

2. 데이터베이스

데이터베이스는 훈련 데이터베이스와 평가 데이터베이스로 분류된다. 훈련/평가 데이터베이스는 각각 50여 분의 뉴스방송 1 회분으로 구성되었다. 뉴스는 중앙방송과 지방방송, 그리고 스포츠로 구성되었다. 각 뉴스 데이터는 음성부만을 추출한 다음 묵음구간(2 sec 이상)을 기준으로 하여 분할하였다.

데이터베이스의 분석은 화자의 성별과 역할을 기준으로 하여 실시하였다. 화자들의 정보는 차후 화자 추적에 유용하게 사용되기 때문에 화자역할별로 조사하였다. 표 1은 화자분류를 위한 인덱싱 기준을 나타낸 것이다. 화자분류는 화자의 성별을 기준으로 1 차 분류를 하고 각각의 역할에 의해서 2 차 분류를 실시하였다. 가장 중요한 것은 컨텐트의 변화를 주도하는 역할인 아나운서이다. 동시발성(TS)으로 분류되는 경우는 복수 화자 발성시에 각각의 언의를 파악 가능할 정도일 때로 한정하였다. 표 2는 훈련/평가용 데이터베이스의 화자 종류별 인원 및 출현 횟수를 조사한 것이다. 우리가 목적하고 있는 아나운서는 훈련/평가 데이터베이스에서 공통적으로 4 명이 출현하고 있다. 가장 높은 출현 횟수를 보이는 것은 리포터들로서 인원 면에서는 32%, 출현 횟수 면에서는 39%를 차지한다.

표 1. 데이터베이스의 화자 분류를 위한 인덱싱 기준

구 분	아나운서	리포터	시민	복수화자
남성(Male)	MA	MR	MC	TS
여성(Female)	FA	FR	FC	

표 2. 데이터베이스의 화자종류별 분석

1구분	MA		MR		MC		FA		FR		FC		TS		계
	인원	횟수	인원	횟수	인원	횟수	인원	횟수	인원	횟수	인원	횟수	인원	횟수	
훈련용	3	30	26	57	28	29	1	18	1	1	8	8	10	10	19377/153
평가용	3	41	17	63	25	31	1	11	2	6	9	13	9	10	66/175

3. 화자 기반 분할 알고리듬

화자 기반 분할의 목적은 음성 데이터를 발성 화자가 변화할 때마다 분할하여 단일 화자만으로 구성된 동일한 성질의 클러스터를 만드는 것이다. 화자 기반 분할 알고리듬은 인접한 일정한 크기의 윈도우에서 검출한 파라미터들의 유사도에 의해서 화자변화를 검출한다. 이때 적용된 가정은 다음과 같다[7].

가정 1. 인접한 두 개의 분석 윈도우가 하나의 화자에 의해 발성되었다고 가정할 경우이다. 통합된 하나의 윈도우에는 가우시안(Gaussian) 모델을 적용하였다.

이때, $N(\mu_x, \sum_x)$는 가우시안 모델을 뜻한다. $X=(x_1, x_2 \ldots x_{N_x})$, $X_1=(x_1, x_2 \ldots x_i)$, $X_2=(x_{i+1}, x_{i+2} \ldots x_{N_x})$이다.

$$X= X_1 \cup X_2 \sim N(\mu_x, \sum_x)$$

가정 2. 인접한 두 개의 분석 윈도우가 다른 화자에 의해 발성되었다고 가정할 경우이다. 이때는 두 분석 윈도우의 독립적인 가우시안 확률을 계산하고 이를 비교한다. 각각의 분석 윈도우에는 가우시안 모델을 적용하였다.

$$X_1 \sim N(\mu_{x_1}, \sum_{x_1}),\ X_2 \sim N(\mu_{x_2}, \sum_{x_2})$$

가정 3. 실제 방송에서와 마찬가지로 화자에 대한 선험적인 지식이 없는 상태라고 가정한다. 이는 화자에 대한 모델이나 화자의 성별, 발성 시간, 발성구간 등에 대한 사전정보가 없다는 것이다. 이 때문에 화자 독립적인 화자변화 검출 알고리듬만이 고려대상이 된다.

3.1 거리 기반 분할 방법

거리 기반 분할 방법은 인접한 두 분석 윈도우의 파라미터들간의 비교를 통해 화자변화 여부를 결정한다[7, 8]. 이 방법은 두 개의 분석 윈도우로 구성된 화자변화 검출 윈도우를 시간 축을 따라 이동시키면서 두 분석 윈도우의 비유사도가 최대가 되는 지점을 검출한다. 유사도가 최소가 되는 부분은 두 분석 윈도우를 발성한 화자가 각각 다르다는 것을 의미한다.

검출된 지점들은 검증과정을 거쳐서 화자변화 지점으로 결정된다.

실험에서는 GLR(general likelihood ratio) 거리 기반 분할 알고리듬을 적용하였다. 사용된 GLR 식은 다음과 같다.

$$GLR = \frac{L(X, N(\mu_x, \sum_x))}{L(X, N(\mu_{x_1}, \sum_{x_1})) \cdot L(X, N(\mu_{x_2}, \sum_{x_2}))} \quad (1)$$

이때, 다차원 가우시안 처리 (multi-dimensional Gaussian process) $N(\mu_x, \sum_x)$는 주어진 특징 파라미터 X의 수열에 대한 가우시안 확률값이다. 식 1의 분모는 분석윈도우 1, 2가 각각 다른 화자에 의해서 발성되었다는 가정 2에 대한 확률값이고, 분자는 분석윈도우 1, 2가 동일 화자에 의해서 발성되었다는 가정 1에 대한 확률값이다.

두 클러스터 사이의 거리를 보다 명확히 하기 위해서 GLR에 음의 로그를 취하였다.

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

이때, 높은 값의 DGLR은 가정 2, 두 윈도우가 각기 다른 화자에 의해서 발성된 상황에 적합하고, 낮은 값의 DGLR은 가정 1, 두 윈도우가 동일한 화자에 의해서 발성된 상황에 적합하다. DGLR은 다음과 같은 과정에 의해서 계산된다. 일정 길이의 두 분석 윈도우를 0.1 초씩 이동시켜 가면서 분석 윈도우간의 GLR 거리를 계산하였다(그림 1 참조). 분석 윈도우는 음성 데이터에서 화자의 특성을 나타낼 수 있는 최소 정보의 크기인 2 초로 정의하였다. 두 분석 윈도우의 경계점을 화자의 특성이 변화하는 부분(화자변화 지점)이라고 가정하고 분석 윈도우간의 성질을 비교한다.

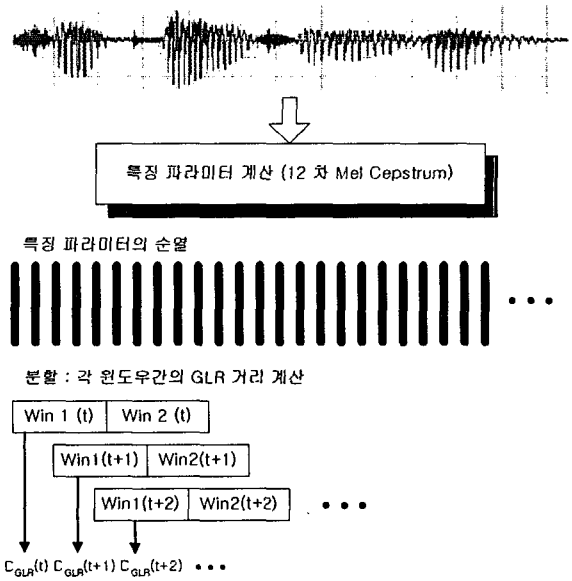

그림 1. GLR 거리 기반 화자 분할

계산된 D_{GLR} 중에서 가장 화자변화지점에 근사한 것을 찾아내기 위해서 국소 최대값과 국소 최소값의 비교를 이용하였다. D_{GLR}의 국소 최대값과 국소 최대값 좌우의 국소 최소값을 검출하여 그 차이가 기준값을 초과하면 국소 최대값 부분에서 화자변화가 일어났다고 정의하였다. 이때 기준값은 음성신호의 성질이나 녹음환경, 배경소음 등에 강하게 의존하는 경향이 있다.

국소값 결정 범위(Config_2)는 국소 최대값 결정을 하기 위한 분석 윈도우의 범위를 결정한다. 이는 지정된 범위 안에서 존재하는 유효한 국소 최대값을 검출하는데 사용된다. 범위 안의 국소 최대값 중 국소 최소값에 둘러싸인 국소 최대값만을 유효한 국소 최대값 후보라고 정의하였다. 검출된 국소 최대값 후보는 국소 최대값 좌우의 국소 최소값들과의 차이를 비교하여 문턱치를 넘는 것만을 인정하게 된다. 이러한 과정을 그림 2로 나타내었다.

본 실험에서는 유효 국소 최대값 결정 문턱치를 이용하여 국소 최대값 결정 범위 안에서 검출된 국소 최대값 중에서 국소 최대값과 좌우의 국소 최소값 사이의 차이가 문턱치 이상인 것만을 유효한 국소 최대값으로 인정하였다. 식 3에서와 같이 두 조건을 모두 만족하는 경우만을 유효한 국소 최대값으로 인정하였다. 그림 2는 αA을 이용하여 유효한 국소 최대값을 검출하는 것을 나타낸 것이다. 이때 A는 L_{Max}값과 좌·우측 L_{Min}값간 차의 표준편차이다.

$$\begin{aligned} L_{\max} - L_{\min\ left} &> \alpha A \\ L_{\max} - L_{\min\ right} &> \alpha A \end{aligned} \tag{3}$$

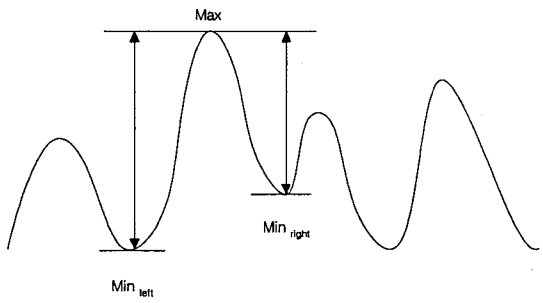

그림 2. 국소 최대값 검출

3.2 BIC 기반 분할

BIC 기반의 분할은 GLR과 비슷한 방법을 사용한다. BIC 기반 분할은 적용모델의 복잡성에 의해 가중치가 적용된다. BIC의 값은 아래의 식에 의해 구해진다[9].

$$BIC(m) = \log L(X,M) - \lambda \frac{m}{2} \log N_x \tag{4}$$

$L(X,M)$은 확률 모델 M에서의 X의 유사도(likelihood ratio)이다. m은 확률 모델 M의 파라미터들의 수이고, λ는 가중치 벡터이다. N_x는 입력음성신호의 전체 프레임 수이다.

가정 1과 가정 2의 최대유사도(maximum likelihood ratio)는 다음과 같이 정의된다.

$$R=-\frac{N_x}{2}\log|\Sigma_x|-\frac{N_{x_1}}{2}\log|\Sigma_{x_1}|-\frac{N_{x_2}}{2}\log|\Sigma_{x_2}| \quad (5)$$

Σ_x, Σ_{x_1}, Σ_{x_2}는 각 윈도우의 공분산 매트릭스이다. N_x, N_{x_1}, N_{x_2}는 입력신호의 프레임 수이다. 가정1과 가정 2의 가우시안 모델들 간의 $\triangle BIC$는 다음과 같이 계산하였다.

$$\triangle BIC(i)=-R(i)+\lambda P \quad (6)$$

$$P=\frac{1}{2}\left(p+\frac{1}{2}p(p+1)\right)\log N_x \quad (7)$$

p는 특징 파라미터의 차수, λ는 패널티 벡터이다. 양수의 $\triangle BIC$값은 가정 1에 적합하고, 음수의 $\triangle BIC$값은 가정 2에 적합하여 화자변화가 일어났음을 의미한다.

이때 패널티 벡터 λ에 의해서 BIC 기반 분할의 성능이 좌우된다. 참고논문에서는 실험적으로 보통 0.5의 값을 적용하였다. 본 실험에서는 결정범위를 0.5에서 1.5까지 조정하면서 최적화하였다.

3.3 거리 기반 분할과 BIC 기반 분할의 통합

GLR을 이용한 거리 기반 분할은 짧은 간격의 화자변화나 배경음향의 변화에 민감하게 반응하여 FAR이 높은 경향을 보인다. BIC 기반 분할은 분석 윈도우가 일정한 길이(3 초 이상)를 만족할 때에만 좋은 성능을 나타내기 때문에 짧은 화자변화에는 둔감한 단점이 있다.

두 가지 알고리듬을 상호 보완시켜서 다음과 같이 적용하였다. 먼저 GLR을 이용한 거리 기반 분할을 적용하여 화자변화 예상지점을 검출하고, BIC 기반 분할을 이용하여 검출된 화자변화 예상지점을 검증하였다.

MDR이 FAR보다 시스템의 성능에 큰 영향을 끼치므로 MDR을 최소로 하기 위해서 입력신호의 성질변화에 민감한 GLR 거리 기반 화자변화 검출기로 느슨하게 예상 화자변화지점을 검출한 다음, 긴 간격의 화자변화지점에 대해서 성능이 좋은 BIC 기반 화자변화 검출기로 이를 검증하였다.

GLR 거리 기반 화자변화 검출과 BIC 기반 화자변화 검출은 다음과 같은 알고리듬으로 통합된다. 먼저 GLR 거리 기반 화자변화 검출에 의해서 정해진 예상 화자변화 지점들을 기준으로 하여 분석 윈도우의 크기를 이와 매칭시키면서 BIC 기반 화자변화 검출을 실행하였다.

BIC 기반 화자변화 검출은 GLR 거리 기반 분할의 결과값을 바탕으로 분석 윈도우의 크기를 결정하기 때문에 서로 다른 크기의 인접한 두 개의 분석 윈도우를 비교하여야 한다.

화자변화 예상구간에 대한 검증 판단은 $\triangle$BIC의 값을 이용한다. $\triangle$BIC의 값이 0보다 크면 해당하는 예상 화자변화 지점을 인정하지 않고, 0보다 작으면 화자변화 지점으로 인정한다.

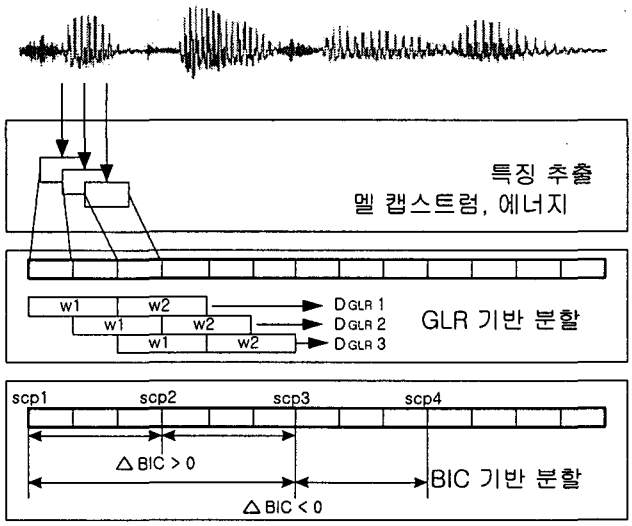

그림 3. GLR과 BIC의 통합적용 과정

4. 실험 결과

화자변화 검출의 목적은 동일한 성질을 가지는 클러스터를 생성하는 것이다. 즉, 하나의 클러스터에 동일한 성질의 단일 화자의 음성만이 존재하도록 입력신호를 분할하는 것이다. 본 논문에서는 화자변화를 검출하기 위해서 GLR 거리 기반 분할과 BIC 기반 분할을 사용하였다. 추정된 화자변화 지점이 실제 화자변화 지점과 2 초 이내의 범위에 위치하는 것만을 화자변화 검출로 인정하였다. 이는 2 초 이내의 오차 내에서는 멀티미디어 데이터의 컨텐트가 크게 영향을 받지 않는다는 가정 하에서 결정되었다. 기본 실험 시스템은 GLR 거리 기반 화자변화 검출기이고, 성능향상을 위하여 배경잡음의 특성을 고려하여 화자변화 검출기를 최적화하였다. 또한 사람의 음성 특징을 고려하여 유성음/무성음에 대한 패널티를 적용하였다.

○ FAR (false alarm rate): 화자변화가 일어나지 않은 지점을 화자변화 지점으로 검출하는 경우

$$FAR = \frac{\textit{False Alarm}\text{의 수}}{\text{검출된 화자변화의 수} + \textit{False Alarm}\text{의 수}} \tag{8}$$

○ MDR (missed detect rate): 화자변화 지점을 검출하는 못하는 경우

$$MDR = \frac{\textit{Missed Detection}\text{의 수}}{\text{실제 화자변화의 수}} \tag{9}$$

○ SR (shift rate) : 실제 화자변화 지점과 검출된 화자변화 지점간의 시간적 오차

$$SR = \frac{|\text{실제 화자변화시간} - \text{검출 화자변화시간}|}{\text{실제 화자변화의 수}} \quad (10)$$

4.1 고소음/저소음 구분

기존의 화자변화 검출 시스템은 특정업무에 한정되었기 때문에 환경잡음 특성의 변화가 적었다. 그러나, 본 논문에서는 뉴스 데이터의 특성상 다양한 환경잡음이 존재하여 시스템 성능에 큰 영향을 준다. 이를 해결하기 위해서 환경잡음을 고소음과 저소음으로 구분하고 각 구분 항목별로 화자변화 검출기 모델을 최적화하였다.

먼저, 고소음/저소음 환경을 구분하기 위해서 청취 테스트를 통해서 배경소음이 많은 스포츠 뉴스를 기준으로 하여 이와 유사한 배경소음 정도를 가지는 파일들을 편집하여 고소음 환경을 가진 약 4 분 길이의 파일을 만들었다. 이와 마찬가지로 저소음 환경에서의 앵커음성 등의 파일들을 편집하여 저소음 환경을 가진 약 4 분 길이의 파일을 만들었다. 고소음과 저소음을 구분하기 위한 특징 파라미터로는 영교차와 정규화 로그 에너지가 고려되었다. 실험결과 정규화 로그 에너지가 더 효과적이었다.

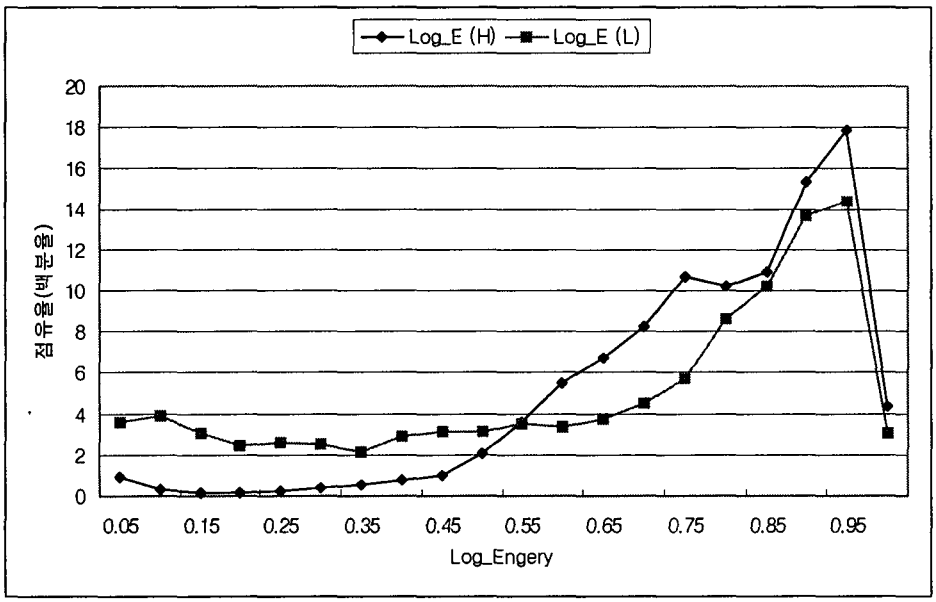

그림 4. 고소음/저소음 환경하에서의 정규화 로그 에너지의 분포

위의 그림을 참조하면 0.45의 값에서 저소음 환경과 고소음 환경에서의 정규화 로그 에너지의 값의 분포가 분명하게 구분된다. 그래서 고소음 환경과 저소음 환경의 구분을 위한 정규화 로그 에너지 문턱치를 0.45로 정하였다. 실험에서는 정규화 로그 에너지가 0.45보다 낮은 에너지를 가지는 것들의 점유율을 비교함으로써 고소음 환경과 저소음 환경을 구분하였다.

먼저 고소음 환경은 정규화 로그 에너지의 문턱치보다 낮은 프레임이 전체 23065 프레임 중에서 1037 프레임으로 0.0449%, 저소음 환경에서는 전체 23534 프레임 중에서 6159 프레임

으로 0.2617%의 점유율을 나타냈다. 이를 바탕으로 문턱치를 조절하면서 실험한 결과 0.2일 때 가장 양호한 결과를 나타내었다.

실제 트레인 DB에서는 총 46 개의 파일에 대해서 고소음 24 개 파일, 저소음 22 개 파일로 분류할 수 있었다.

4.2 D_{GLR}의 저역 통과 필터링

식 2에 의해서 계산된 D_{GLR}은 고주파 성분에 의한 섭동을 방지하기 위해 헤밍 윈도우를 사용하여 저역통과 필터링을 실시하였다. 이때, 실험결과는 저역통과 필터링의 차수에 크게 영향을 받는다. 이에 실험에서는 저역통과 필터의 차수(Config_1)를 조정하면서 시스템을 최적화하였다.

저역통과 필터는 음성처리 분야에서 많이 사용되고 있는 헤밍 윈도우를 이용하였다. 사용된 헤밍 윈도우는 식 11과 같다. 이때 N의 크기를 5~15 프레임까지 변경하면서 최적화하였다. 이때 n에는 D_{GLR} 값이 대입된다.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{11}$$

4.3 유성음과 무성음의 분리

사람의 음성은 무성음과 유성음이 유사 주기적으로 출현하는 경향을 보인다. 이 중 무성음의 경우에는 음성의 특징에 대한 정보가 유성음에 비해서 극히 적다. 이를 감안하여 무성음 부분에 대해서 패널티를 적용함으로써 정보가 상대적으로 집중된 유성음 부분과 차별하였다.

일반적으로 무성음과 유성음을 구분하기 위해서 에너지와 영교차율을 사용한다. 본 논문에서는 영교차율을 이용하였다. 실험 결과 무성음의 영교차율 평균값은 0.63, 유성음의 영교차율 평균값은 0.28이었다. 아래의 그림은 실제 뉴스 데이터에서의 무성음과 유성음의 영교차 변화를 나타낸 것이다. 무성음 구간과 발성간 휴지기에서 영교차가 급격히 상승하는 것을 관찰할 수 있다. 이를 이용하여 입력음성에서 무성음과 유성음 부분을 분리하고 이 중 유성음 부분만을 처리하면 성능이 향상될 것이다. 이를 위해서 무성음 부분에 패널티를 부여하는 영교차율에 대한 패널티 윈도우를 설계하였다.

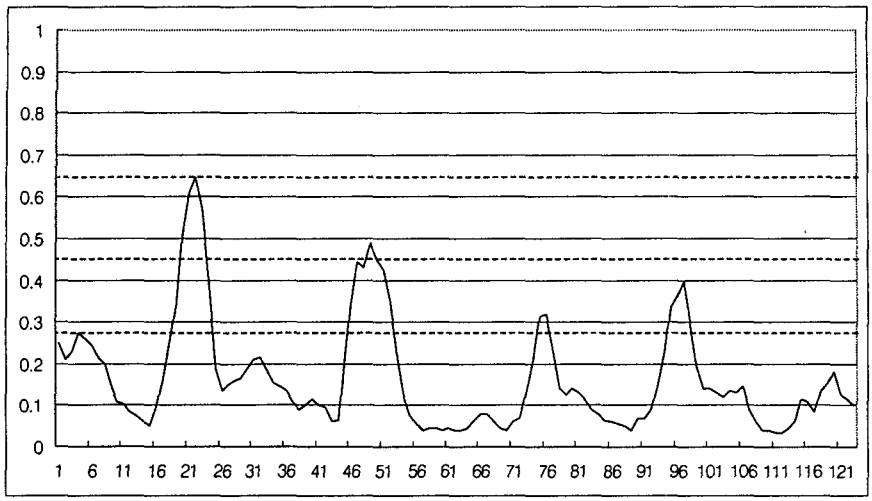

그림 5. 무성음과 유성음의 영교차 비교

영교차율을 이용한 유/무성음에 대한 패널티 부여는 해밍 윈도우를 변형하여 유성음 부분만을 강조한 패널티 윈도우를 실험하였다. 영교차율에 대한 패널티 윈도우는 그림 6과 같은 특성을 가진다. 먼저 패널티 윈도우의 중심은 유성음의 영교차율의 평균값인 0.28로 고정하였다. 패널티 윈도우의 범위(Config_4)를 가변시켜 가면서 성능을 최적화하였다. 이때 w(n)의 최대값은 1로 제한하였으며, N은 1 초부터 26 초까지 변형하면서 실험하였다.

$$w(n)=\left(h(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right)\right)\times 2 \qquad \text{for} 0\le n\le N-1 \tag{12}$$

실험을 위해서 녹음실 수준의 음성부분만을 추출하여 3 분 55 초 길이의 총 13 회 화자변화가 존재하는 테스트 파일을 별도로 만들었다. 실험결과 기존의 GLR 거리 기반 화자분할 방법에 비해서 성능이 향상되었다.

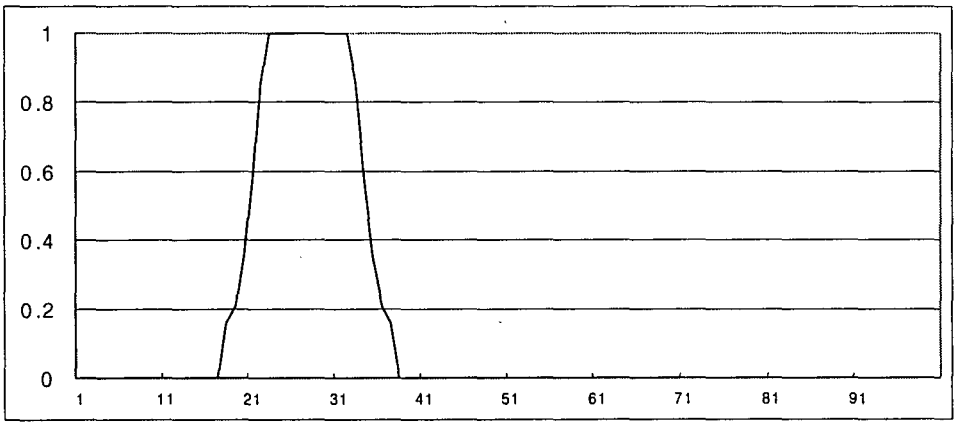

그림 6. 영교차율을 이용한 유/무성음 패널티 윈도우

표 3. 패널티 윈도우 적용에 의한 시스템의 성능변화

구 분	패널티 윈도우 적용 전	패널티 윈도우 적용 후
MDR	0	0
FAR	80.59	78.88

4.4 GLR 거리 기반 화자변화 검출기 최적화

실험은 아래와 같은 조건 하에서 실시되었다. 각각의 조건들은 위에서 제시되었던 것으로서 결정범위에서 수치를 변경해 가면서 성능을 평가하였다. 실험은 4.1에서 언급한 바와 같이 고소음 환경과 저소음 환경을 별도로 실시하였다.

실험 결과를 MDR 12% 부근을 기준으로 하여 살펴보면, 고소음 환경에서의 FAR은 82.77%이고, 저소음 환경에서의 FAR은 87.78%이다.

표 4. GLR 거리 기반 화자변화 검출을 위한 실험조건

구 분	결 정 범 위	명 칭
저역 통과 필터	0.5 ~ 1.5 (초)	Config_1
국소값 결정 범위	0.5 ~ 4.0 (초)	Config_2
유효 국소 최대값 결정 문턱치	0.1 ~ 2.0	Config_3
ZCR penalty Window	1.0 ~ 26.0 (초)	Config_4
GMM의 mixture	1 ~ 9	Config_5

표 5. 고소음 환경에서의 GLR 거리 기반 화자변화 검출 결과

구 분		1	2	3	4	5	6	7	8	9	10
구성	Config_1	0.5	1.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	Config_2	2.5	2.0	2.0	2.5	1.5	0.5	3.0	3.0	2.5	2.5
	Config_3	0.2	0.1	0.3	0.3	0.3	0.3	0.3	0.2	0.3	0.2
	Config_4	26	17	24	23	24	24	26	22	23	26
	Config_5	1	8	1	7	1	1	1	1	1	7
MDR (%)		6.12	6.12	7.14	8.16	8.16	8.16	10.20	11.22	11.22	12.24
FAR (%)		85.52	86.64	85.89	84.21	86.91	87.40	83.72	83.69	85.15	82.77
SR (sec)		0.90	0.75	0.84	0.80	0.84	0.80	0.94	0.95	0.91	85.23

표 6. 저소음 환경에서의 GLR 거리 기반 화자변화 검출 결과

구분		1	2	3	4	5	6	7	8	9	10
구성	Config_1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	Config_2	3.0	3.0	2.0	2.0	3.5	3.5	3.0	3.0	3.5	3.5
	Config_3	0.1	0.2	0.1	0.1	0.2	0.1	0.1	0.2	0.2	0.1
	Config_4	16	16	16	16	22	16	15	16	21	16
	Config_5	2	4	7	1	2	7	9	2	2	5
MDR (%)		6.25	6.25	6.25	6.25	9.37	9.37	9.37	9.37	12.50	15.62
FAR (%)		89.04	89.22	91.13	91.23	87.20	87.73	88.27	89.22	87.78	87.69
SR (sec)		0.85	0.75	0.50	0.54	1.09	0.62	0.65	0.76	0.98	0.80

4.5 BIC 기반 화자변화 검증

3.3 절에서 언급한 바와 같이 화자변화에 민감한 GLR 거리 기반 분할을 이용하여 화자변화 지점을 검출하고 BIC 기반 분할을 이용하여 검증을 실시하였다. 실험에 사용된 조건은 표 7과 같다. BIC 패널티 1은 GLR 거리 기반 화자변화 검출결과에 대하여 1차 BIC 기반 화자변화 검증에 사용되었고, BIC 패널티 2는 1차 BIC 기반 화자변화 검증결과에 대하여 적용되었다. 실험결과를 MDR 15% 부근을 기준으로 하여 살펴보면 고소음 환경에서의 FAR은 63.29%이고, 저소음 환경에서의 FAR은 54.28%이다.

표 7. BIC 기반 화자변화 검증을 위한 실험조건

구 분	결 정 범 위	명 칭
BIC 패널티 1	0.5 ~ 1.5	Config_6
BIC 패털티 2	0.5 ~ 1.5	Config_7

표 8. 고소음 환경에서의 BIC 기반 화자변화 검증 결과

구분		1	2	3	4	5	6	7	8	9	10	11	12
구성	Config_1	0.5	1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	Config_2	2.5	2.0	2.0	2.0	2.5	1.5	2.0	2.5	2.0	2.0	2.0	2.5
	Config_3	0.3	0.1	0.1	0.1	0.2	0.3	0.3	0.2	0.3	0.3	0.3	0.2
	Config_4	23	17	17	17	26	24	24	26	24	24	24	26
	Config_5	7	8	8	8	1	1	1	1	1	1	1	1
	Config_6	1.3	1.2	1.2	1.2	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.6
	Config_7	1.3	1.4	1.9	1.7	0.5	0.5	0.5	0.6	0.7	0.6	0.5	0.6
MDR (%)		10.20	10.20	11.22	11.22	12.24	12.24	13.26	15.30	17.34	17.34	17.34	17.34
FAR (%)		80.93	83.63	79.95	81.47	67.97	68.18	67.11	63.29	56.05	59.83	60.32	61.11
SR (sec)		0.83	0.79	0.80	0.79	0.91	0.86	0.86	0.90	0.90	0.89	0.89	0.91

표 9. 저소음 환경에서의 BIC 기반 화자변화 검증 결과

구분		1	2	3	4	5	6	7	8	9	10	11	12
구성	Config_1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	Config_2	2.0	2.0	2.0	2.0	2.0	2.0	3.0	3.0	2.0	2.0	2.0	3.0
	Config_3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2
	Config_4	16	16	16	16	16	16	17	17	16	16	16	16
	Config_5	1	1	1	1	1	1	5	5	1	1	1	2
	Config_6	0.5	0.5	0.5	0.5	0.6	0.6	0.5	0.5	0.7	0.7	0.6	0.5
	Config_7	0.5	0.4	0.6	0.7	0.6	0.5	0.4	0.5	0.7	0.6	0.7	1.0
MDR (%)		6.25	6.25	9.37	15.62	18.75	18.75	18.75	18.75	25.00	25.00	25.00	25.00
FAR (%)		67.67	68.62	57.33	54.28	56.16	57.33	87.25	87.25	44.82	45.76	50.00	74.60
SR (sec)		0.54	0.54	0.55	0.55	0.60	0.55	0.81	0.81	0.57	0.57	0.55	0.86

5. 결 론

본 논문에서는 통계적 기법을 이용한 화자변화 검출에 대하여 실험하였다. 화자변화 검출은 NOD 서비스를 위한 중요한 전처리 과정이다. 화자변화 검출은 입력음성을 화자별로 분할하여 동일화자만으로 구성된 음성 클러스터를 생성한다. 화자변화 검출 알고리듬은 매트릭스 방법 중 GLR 거리 기반 화자분할과 BIC 기반 화자분할을 이용하였다. GLR과 BIC의 장단점을 취합하여 먼저, GLR 거리 기반 화자분할을 이용하여 화자변화 예상지점을 검출하고 BIC 기반 화자분할을 이용하여 이를 검증하였다. 실험결과를 MDR 15% 부근을 기준으로 하여 살펴보면 고소음 환경에서의 FAR은 63.29%이고, 저소음 환경에서의 FAR은 54.28%이다.

참 고 문 헌

[1] 이경록, 서봉수, 김진영. 2000. "음성/음악 분류기를 위한 특징 비교," *한국음향학회 하계학술발표대회.*

[2] I. Magrin-Changnolleau et al. 1999. "Detection of target speakers in audio databases." *ICASSP.*

[3] A. E. Rosenberg et al. 1998. "Speaker detection in broad- cast speech databases," in *ICSLP.*

[4] 이경록, 서봉수, 김진영. 2001. "오디오 인덱싱을 위한 음성/음악 분류 특징 비교." *한국음향학회지,* 20(2), 10-15.

[5] Beigi, H. & S. Maes. 1998. "Speaker, channel and environment change detection." *World congress of automation.*

[6] M. A. Siegler et al. 1997. "Automatic segmentation, classification, and clustering of broadcast news audio." *DARPA speech recognition workshop.*

[7] Delacourt, P., David Kryze. & Christian J. Wellekens. 1999. "Detection of speaker

changes in an audio document." *Eurospeech99*.
[8] Delacourt, P., Delacourt, D. Kryze. & C. J. Wellekens. 1999. "Speaker based segmentation for audio data indexing." *ESCA workshop*: accessing information in audio data.
[9] Chen, S. & P. Gopalakrishnan. 1998. "Speaker, environment and channel change detection and clustering via the Bayesian information criterion." *DARPA speech recognition workshop*.

접수일자: 2001. 10. 20.
게재결정: 2001. 12. 8.

▲ 이경록
광주광역시 북구 용봉동 300번지 (우: 300-757)
전남대학교 전자공학과 멀티미디어 신호처리 실험실
Tel: +82-62-530-0472 Fax: +82-62-530-0472
E-mail: krlee@dsp.chonnam.ac.kr

▲ 김진영
광주광역시 북구 용봉동 300번지 (우: 300-757)
전남대학교 전자공학과, RRC HECS
Tel: +82-62-530-1757 Fax: +82-62-530-1757
E-mail: kimjin@dsp.chonnam.ac.kr