# 음성 구간 검출기의 실시간 적응화를 위한 특징 벡터의 차원 축소 방법

김평환, 한학용, 김창근, 고시영*, 허강인
동아대학교 전자공학과, 경일대학교 전자정보통신공학과*

# Dimension Reduction Method of Feature Vector for Real-Time Adaptation of Voice Activity Detection

Pyoung-Hwan Kim, Hag-Yong Han, Chang-Keun Kim, Si-Young Koh*, Kang-In Hur
Dept. of Electronic Engineering, Dong-A University
Dept. of Electronic Information and Communication Engineering, Kyung-il University*
E-mail : redbook7@donga.ac.kr

## 요 약

본 논문은 잡음 환경하에서 특징 벡터의 차원 축소를 통한 음성 구간 검출에 관한 연구이다. 음성/비음성 분류는 통계적 모델을 이용한 분류-기반 방법을 사용한다. 검출기에서 실시간 적응화를 위해 우도-기반의 특징 벡터에 대한 차원 축소 방법을 제안한다. 이 방법은 음성/비음성 클래스에 대한 가우시안 확률 밀도 함수에 의한 비선형적 우도값을 새로운 특징으로 취하는 방법이다. 음성/비음성 결정은 우도비 검증(Likelihood Ratio Test)의 방법을 이용하며, LDA(Linear Discriminant Analys)에 의한 축소 결과와 성능을 비교한다. 실험 결과 제안된 차원 축소 방법을 통하여 2차원으로 축소된 특징 벡터가 고차원에서의 결과와 대등함을 확인하였다.

## I. 서 론

자동 음성 인식(ASR) 시스템은 많은 발전을 이루었음에도 불구하고 잡음 환경에서의 음성 인식 시스템의 구축에는 많은 어려움을 겪고 있다. 음성 구간 검출은 음성이 포함된 어떤 신호에서 음성 구간의 시작과 끝을 찾아내는 전처리 과정을 말한다. 음성 구간 검출 기술은 음성인식의 전처리에서 뿐만 아니라 음성 통신에서 음성 신호의 유무를 판별하여 평균 전송률을 높이기 위한 가변 전송률 음성 부호화기에서의 핵심 기술이기도 하다.

지금까지 음성 구간 검출에 관한 여러 방법들이 제안되어져 왔는데, 이를 크게 분류하면 규칙-기반 방법과 분류-기반 방법으로 나눌 수 있다. 규칙-기반 방법은 휴리스틱하게 얻은 몇 가지 특징을 척도로 사용하여 음성 구간 검출을 위한 규칙을 유도하여 이용하는 방법으로, 에너지의 변화, 영 교차율, 피치 등을 이용하는 방법이 대표적인 방법이다. 규칙들이 특징 파라메터에 고정되므로 새로운 특징에 대하여 새로운 규칙이 만들어져야하는 단점이 있다.

분류-기반 방법은 음성/비음성 이벤트를 통계적으로 모델링하여 이진 분류의 문제로 음성 구간 검출 문제를 다룬다. 일반적으로 모델은 단일 혹은 혼합 가우시안 함수로 통계적인 모수로 표현된다. 이 방법은 규칙-기반의 방법보다 성능이 대체로 우수하다고 알려져 있지만, 모델의 학습 환경에 고정되는 단점이 있다. 그러므로 모델에 대한 실시간 적응화가 필요하다. 적응화를 행할 경우, 음성 인식에서 사용하는 특징 벡터가 다차원이므로 통계적 모델의 모수를 적응시키기 위해서 많은 계산 시간을 요구한다. 음성 구간 검출이 음성인식 등을 위한 전 처리 과정임을 볼 때 시간 지연은 실시간 적용에 문제점이 된다.

본 논문에서는 분류-기반 음성 구간 검출기의 실시간 적응화가 가능하기 위한 선결 과제인 특징 벡터의 효과적인 차원 축소 방법에 관한 연구로 비교적 간단하면서 최적의 분류률을 가지는 우도에 기반한 특징 벡터의 비선형적 차원 축소 방법을 제안한다.

## 2. 본 론

모델에 대한 적응화의 경우, 널리 사용되는 적응화 방법은 MAP(Maximum A Posteriori)와 확장된 MAP (Extended MAP) 그리고, MLLR(Maximum Likelihood Linear Regression)과 같은 ML 적응 방법이 있다. 고차원 특징에 대한 MAP와 ML 적응 방법은 데이터 상에서 다중 패스로 인하여 많은 계산량을 요구한다. 이는 실시간 음성 구간 검출을 어렵게 하는 원인이 되며 이를 해결하기 위해 특징 벡터의 차원 축소하여 계산량을 줄일 필요가 있다. 그리고, 차원 축소를 할 때 음성/비음성이 최대한 분별 가능한 특징 정보를 보유하면서 차원을 축소하는 비선형 차원 축소방법이 필요하다.

### 2.1 비선형 우도에 의한 차원 축소

음성/비음성 클래스 우도를 이용한 비선형 차원 축소 방법은 $N$개의 클래스를 구별하려고 할 때, 특징 벡터는 모델의 개수가 $N$개일 경우, $N$차원 공간상으로 비선형적인 사영값을 얻을 수 있다. 이때, 각 차원은 개별 클래스에 대한 확률값으로 단조 함수이며 보통 로그 함수를 사용한다. 결론적으로 $D$차원 벡터 $X$는 $N$차원 벡터 $Y$로 축소될 수 있다.

$$Y=[log(P(X|C_1))\,log(P(X|C_2))\cdots log(P(X|C_N))] =[y_1\ y_2\ \cdots\ y_N] \quad (1)$$

여기서 $log(P(X|C_i)$는 클래스 $C_i$의 확률 밀도 함수에 의한 벡터 $X$의 로그 우도이다. 이것이 새로운 특징 $Y$의 $i$번째 요소인 $y_i$를 구성한다. 식(1)을 통하여 새로운 $N$차원의 우도 공간으로 사영된다고 할 수 있다.

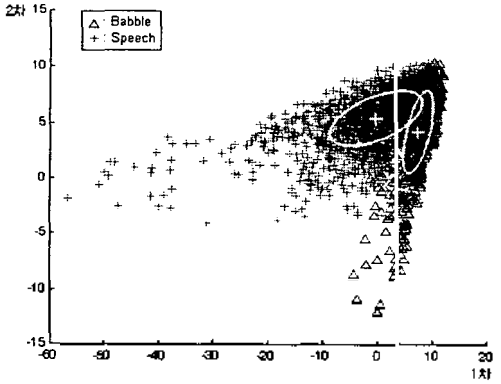


그림 1. 우도에 의해 2차로 축소된 특징 분포도 (Babble Noise, 5dB)

음성/비음성 분류에 이 축소 방법을 적용할 경우 그림 2와 같은 2가지 방법으로 2차원 특징 벡터로 축소가 가능하다.

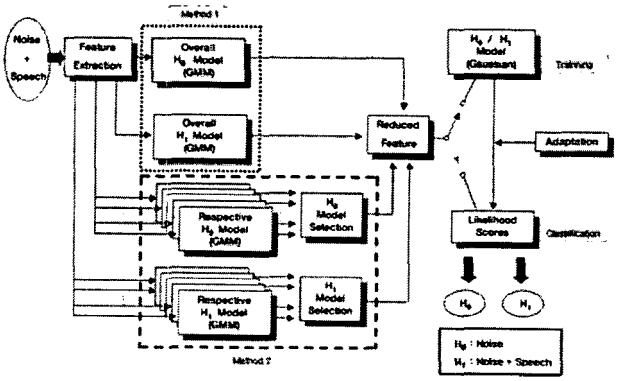


그림 2. VAD using Likelihood-Based Dimension Reduction

첫 번째 방법은 전체 노이즈 환경의 음성/비음성에 대한 혼합 밀도 함수로부터 우도값을 취해 축소하는 방법이고, 두 번째 방법은 각각의 노이즈 환경에 대한 혼합 밀도 함수로부터 우도값을 취해 축소한다. 축소한 특징 벡터를 이용한 분류는 우도비 검증의 방법을 사용한다.

우도비 검증의 방법은 축소된 특징 벡터 $y$가 어느 클래스에 속하는가를 결정하는 것인데, 이를 위해서 음성($c_1$)/비음성($c_2$) 클래스의 사후확률 $P(c_i|y)$를 계산하고 가장 큰 사후확률 값을 가지는 클래스를 결정한다.

즉, $P(c_1|y) > P(c_2|y)$ 라면 $c_1$을 선택하고 그렇지 않다면 $c_2$를 선택한다.

### 2.2 선형 판별 분석(LDA)에 의한 차원 축소

선형 판별 분석은 원 특징 벡터($x$)에 포함된 분류 정보가 충분한 학습 샘플을 통해서 추정된 선형 변환 행렬($\Theta$)에 의하여 축소된 특징 벡터($y$)로 완전히 서술될 수 있다고 가정한다.

$$y=\Theta^T x \quad (2)$$

클래스 종속LDA의 경우 테스트 데이터에 대한 분류는 변환 행렬이 각 클래스 마다 존재하므로 테스트 데이터를 각각 변환 행렬로 축소하여 각 클래스의 중심간의 거리에 의하여 판별하는 방법을 적용하고, 클래스 독립

LDA의 경우에는 단일한 변환 행렬을 통하여 특징 벡터를 2차원으로 축소한 후, 우도비 검증의 방법을 적용 한다.

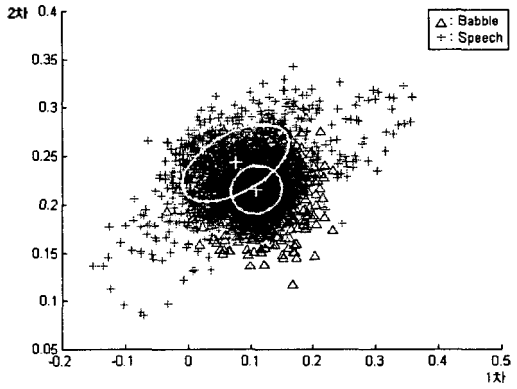


그림 3. 클래스 독립 LDA에 의해 2차로 축소된 특징 분포도 (Babble Noise, 5dB)

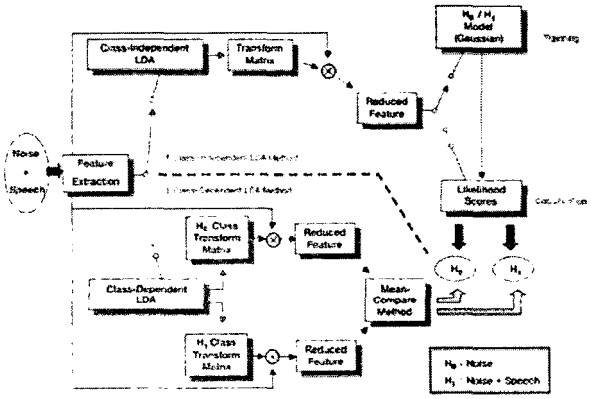


그림 4. VAD using LDA-Based Dimension Reduction

## 3. 실험 및 결과

### 3.1 특징 파라메터

ASR 시스템에서 사용되는 특징 벡터들은 신호의 단구간 퓨리에 변환으로부터 유도된 스펙트럴에 기반한 특징과 이러한 기본 특징의 변화량인 델타 성분 등을 추가하여 사용되고 있다. 분류-기반의 음성구간 검출을 위한 특징 벡터는 ASR 시스템에서 사용되는 특징인 MFCC(Mel-Frequency Cepstrum Coefficient)를 기준 벡터로 사용하는 것이 적절한 것으로 사료된다.

### 3.2 실험 조건 및 DB

실험에 사용된 음성 데이터는 ETRI 한국어 중가마이크 음성인식용 낭독체 문장 데이터를 사용하였고, 잡음 데이터는 NOISEX-92 잡음 데이터를 사용하였다. 이 잡음은 19.98kHz, 16bit의 anti-aliasing 필터링 된 데이터로 본 실험에서는 16kHz, 16bit로 변환하였고, 잡음 환경 구현을 위해 5dB, 15dB, 25dB의 레벨로 음성 데이터와 잡음 데이터를 섞어서 학습 데이터 및 테스트 데이터를 만들었다.

학습에는 3명이 각각 1문장씩(문장 당 약 7초) 3문장으로 학습을 하였고, 테스트에는 학습자 외의 10명의 10문장으로 음성 구간 검출 실험을 하였다.

표 1. NOISEX-92

| Index | Noise Type |
|---|---|
| N1 | Speech babble noise |
| N2 | Jet cockpit noise |
| N3 | Destroyer engine room noise |
| N4 | Destroyer operators room noise |
| N5 | F-16 cockpit noise |
| N6 | Factory floor noise1 |
| N7 | Factory floor noise2 |
| N8 | HF channel noise |
| N9 | Military vehicle noise |
| N10 | Tank noise |
| N11 | Machine gun noise |
| N12 | Pink noise |
| N13 | Car interior noise |
| N14 | White noise |

### 3.3 실험 결과

음성 구간 검출에 대한 실험은 ❶전체 잡음 모델을 사용한 GMM Ⅰ과 ❷개별 잡음 모델에 의한 GMM Ⅱ의 경우에는 모두 특징 파라메터를 10차 MFCC를 사용하여 GMM을 이용하여 분류하며, ❸Method Ⅰ은 전체 잡음에 대하여 제안하는 우도-기반 차원 축소 방법에 의해 2차원으로 축소를 한 경우이고, ❹Method Ⅱ는 개별 잡음에 대하여 제안하는 차원 축소 방법으로 차원을 2차원으로 축소한 경우이다. ❺Class-Indep.LDA는 개별 잡음에 대하여 클래스 독립LDA로 2차원으로 축소한 경우이다. ❶ ~ ❺의 방법은 축소된 특징에 대하여 우도비 검증의 방법으로 분류 실험을 하였다. 마지막으로 ❻Class-Dep.LDA는 개별 잡음에 대하여 클래스 종속LDA로 2차원으로 축소한 경우인데, 이 방법은 변환 행렬이 클래스마다 각각 존재하여 우도비 검증의 방법을 적용할 수 없으므로 최근접 중심점에 의한 방법으로 분류 실험을 하였다.

표 2. Result of Test Data (25dB 단위 : %

| | | | | | | |
|---|---|---|---|---|---|---|
| N1 | 82.98 | 93.08 | 83.76 | 93.07 | 85.19 | 43.22 |
| N2 | 87.01 | 95.11 | 86.65 | 95.16 | 93.41 | 85.38 |
| N3 | 93.37 | 96.11 | 92.68 | 96.14 | 85.95 | 89.77 |
| N4 | 91.10 | 94.68 | 90.42 | 94.54 | 91.91 | 86.14 |
| N5 | 91.96 | 96.37 | 91.26 | 96.33 | 95.11 | 89.21 |
| N6 | 90.09 | 95.59 | 89.05 | 95.36 | 95.86 | 79.61 |
| N7 | 90.47 | 95.48 | 91.15 | 95.35 | 86.20 | 81.37 |
| N8 | 92.85 | 96.86 | 92.26 | 97.14 | 96.68 | 69.20 |
| N9 | 91.59 | 93.65 | 91.80 | 94.11 | 92.16 | 82.43 |
| N10 | 92.04 | 94.10 | 91.63 | 94.68 | 93.35 | 86.51 |
| N11 | 90.06 | 92.00 | 90.16 | 92.09 | 78.92 | 83.19 |
| N12 | 89.61 | 96.07 | 88.71 | 96.12 | 93.48 | 90.40 |
| N13 | 92.48 | 92.99 | 92.12 | 94.44 | 88.96 | 87.20 |
| N14 | 84.36 | 94.83 | 83.51 | 95.29 | 90.46 | 94.04 |

표 3. Result of Test Data (15dB) 단위 : %

| | | | | | | |
|---|---|---|---|---|---|---|
| N1 | 74.33 | 85.16 | 73.93 | 84.45 | 78.48 | 73.90 |
| N2 | 65.88 | 91.64 | 63.37 | 91.85 | 87.08 | 86.45 |
| N3 | 88.36 | 92.71 | 86.87 | 92.79 | 82.87 | 90.78 |
| N4 | 84.59 | 89.26 | 82.43 | 90.08 | 81.30 | 56.09 |
| N5 | 84.97 | 91.70 | 82.35 | 92.05 | 87.14 | 77.73 |
| N6 | 80.33 | 89.69 | 77.10 | 90.05 | 73.84 | 41.66 |
| N7 | 78.22 | 92.44 | 82.68 | 92.21 | 84.94 | 75.72 |
| N8 | 85.72 | 94.89 | 84.43 | 94.95 | 93.04 | 88.08 |
| N9 | 87.25 | 92.34 | 86.41 | 93.23 | 86.39 | 83.63 |
| N10 | 84.82 | 90.09 | 84.73 | 90.42 | 87.64 | 82.50 |
| N11 | 88.70 | 88.93 | 88.11 | 89.70 | 53.64 | 62.61 |
| N12 | 76.75 | 91.00 | 72.81 | 91.52 | 87.77 | 83.06 |
| N13 | 89.95 | 94.51 | 89.24 | 94.41 | 85.57 | 84.82 |
| N14 | 63.79 | 92.96 | 62.22 | 92.94 | 88.58 | 79.17 |

표 4. Result of Test Data (5dB) 단위 : %

| | | | | | | |
|---|---|---|---|---|---|---|
| N1 | 71.81 | 69.55 | 69.61 | 66.76 | 66.31 | 71.52 |
| N2 | 51.63 | 81.59 | 42.69 | 81.32 | 74.15 | 67.19 |
| N3 | 82.44 | 85.28 | 79.51 | 85.27 | 69.01 | 75.60 |
| N4 | 67.77 | 72.88 | 63.31 | 71.88 | 56.96 | 50.70 |
| N5 | 52.33 | 79.33 | 46.87 | 78.75 | 62.11 | 77.92 |
| N6 | 58.38 | 70.87 | 49.67 | 70.03 | 58.66 | 45.36 |
| N7 | 80.13 | 85.19 | 79.63 | 84.85 | 67.94 | 35.70 |
| N8 | 80.39 | 87.99 | 77.88 | 88.35 | 78.23 | 46.55 |
| N9 | 83.53 | 88.98 | 83.09 | 89.42 | 80.43 | 62.92 |
| N10 | 73.82 | 77.53 | 71.75 | 77.38 | 68.26 | 66.62 |
| N11 | 87.12 | 87.23 | 86.67 | 85.81 | 77.60 | 78.98 |
| N12 | 57.25 | 80.15 | 45.97 | 79.95 | 69.26 | 77.42 |
| N13 | 88.22 | 93.12 | 87.58 | 93.30 | 72.33 | 36.39 |
| N14 | 45.21 | 85.81 | 36.19 | 85.55 | 83.75 | 84.63 |

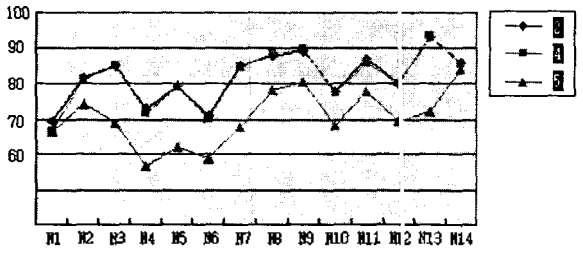


그림 5.Results of Test Data ( , 5dB)

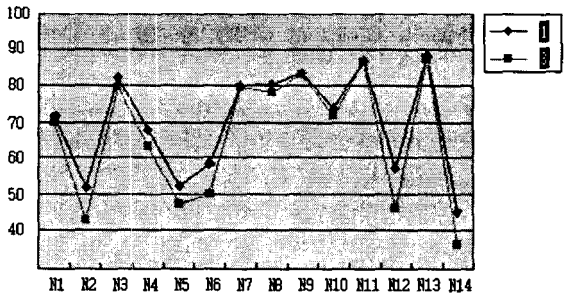


그림 6. Results of Test Data ( , 5dB)

## 5. 결 론

본 논문에서는 잡음 환경에서의 통계적 모델에 대한 실시간 적응화를 위한 고차원의 특징 벡터의 차원 축소 방법을 제안하였다. 실험결과 제안된 특징 축소 방법이 원래 특징 벡터인 MFCC 10차의 결과와 거의 대등한 성능을 얻을 수 있음을 확인하였다. 반면에 LDA에 의한 차원 축소 방법은 잡음에 따라 큰 폭의 편차를 보여줌을 확인하였다. 따라서 제안된 방법으로 2차원으로 차원을 축소할 경우, 음성/비음성 분류를 위한 통계적 모델에 대한 빠른 적응화가 가능하여 실시간 적용이 가능할 것으로 사료된다.

## 참고문헌


1. Rabiner, L.R. and Sambur, M.R., "An Algorithm for Determining the Endpoints of Isolated Utterances", The Bell System Technical Journal, Vol. 54, No. 2, pp. 297-315, February 1975

2. Jean-Claude Junqua, Brian Mak and Ben Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, Vol. 2, No. 3, pp. 406-412, July 1997

3. M.H. Savoji, "Endpointing of Speech Signals", Speech Communication,Vol. 8,No. 1, pp.46-60, March 1989

4. J. Sohn and W.Sung,"A voice activity detector employing soft decision based noise spectrum adaptation," In Proc. Ont. Conf. Acoust. Speech, and Signal Processing, pp.365-3, 1998

5. 김창근, 박정원, 권호민, 허강인 "음성인식기 구현을 위한 잡음에 강인한 음성구간 검출기법" 한국 신호처리 시스템학회 논문집, 제4권 2호 pp.18-24, 2003년 4월