음성구간 검출기의 실시간 적응화를 위한 음성 특징벡터의 차원 축소 방법

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

박 진 영*, 이 광 석**, 허 강 인*

Jin-young Park*, Kwang-seok Lee**, Kang-in Hur*

요 약

본 논문에서는 다양한 잡음환경에서의 실시간 적응화 기법을 적용하기 위한 선결 과제로 다차원 음성 특징 벡터를 저차원으로 축소하는 방법을 제안한다. 제안된 방법은 특징 벡터를 확률 우도 값으로 매핑시켜 비선형적으로 축소하는 방법으로 음성/비음성의 분류는 우도비 검증(Likelihood Ratio Test ; LRT)을 이용하여 분류하였다. 실험 결과 고차원 특징 벡터를 이용하여 분류한 결과와 대등하게 분류됨을 확인할 수 있었다. 그리고, 제안된 방법에 의해 검출된 음성 데이터를 이용한 음성인식 실험에서도 10차 MFCC(Mel-Frequency Cepstral Coefficient)를 사용하여 분류한 경우와 대등한 인식률을 보여주었다.

Abstract

In this paper, we propose the dimension reduction method of multi-dimension speech feature vector for real-time adaptation procedure in various noisy environments. This method which reduces dimensions non-linearly to map the likelihood of speech feature vector and noise feature vector. The LRT(Likelihood Ratio Test) is used for classifying speech and non-speech. The results of implementation are similar to multi-dimensional speech feature vector. The results of speech recognition implementation of detected speech data are also similar to multi-dimensional(10-order dimensional MFCC(Mel-Frequency Cepstral Coefficient)) speech feature vector.

Keywords : VAD(Voice Activity Detector), ASR(Automatic Speech Recognition), Dimension Reduction, Noisy Environments

I. 서 론

음성 인식 시스템을 구성함에 있어 가장 고려되는 사항은 잡음환경에서의 시스템 성능이라 할 수 있을 것이다. 일반적으로 잡음환경에서의 음성 인식 시스템은 인식 성능의 현저한 저하를 나타낸다. 이를 개선하기 위한 다양한 방법이 제시되고 있으며 그 중 음성 구간을 정확하게 검출해내는 전처리 과정을 이용한 음성 인식 시스템의 성능 향상 방법에 관한 연구가 진행되고 있다. 또한 음성 구간 검출 기술은 음성 통신의 부호화기에서 평균 전송률을 높이기 위한 핵심적인 기술로 사용되기도 한다.

지금까지 음성 구간 검출에 관한 여러 방법들이 제안되어져 왔는데, 이를 크게 구분하면 규칙-기반 방법과 분류-기반 방법으로 나눌 수 있다.

규칙-기반 방법은 경험으로 얻은 몇 가지 특징을 척도로 사용하여 음성 구간 검출을 위한 규칙을 유도하여 이용하는 방법으로 에너지의 변화, 영 교차율, 피치 등을 이용하는 방법들이 대표적인 방법이다. 이 방법들은 규칙

*동아대학교 전자공학과 **진주산업대학교 전자공학과
논문 번호 : 2006-3-6 접수 일자 : 2006. 5. 26
심사 완료 : 2006. 7. 6
※본 논문은 2004년도 동아대학교 학술연구비(공모과제) 지원에 의하여 연구되었음.

들이 특징 파라미터에 고정되므로 새로운 환경에 대하여 새로운 규칙이 만들어져야 한다는 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 환경에 따라 실시간으로 모델을 적응시킬 필요가 있다. 적응화를 하는 경우, 음성 인식에서 사용하는 특징 파라미터가 10차 이상의 다차원이므로 통계적 모델의 모수를 적응시키는데 많은 계산 시간이 요구된다. 음성구간 검출이 음성 인식 등을 위한 전처리 과정임을 볼 때 시간 지연은 실시간 적용에 문제점이 된다.

본 논문의 구성은 다음과 같다. 2장에서는 제안한 차원 축소 방법에 대해 설명하고, 3장에서는 제안된 방법으로 검출된 음성신호를 사용하여 음성인식 실험 및 결과를 나타내었고, 4장에서 결론을 맺는다.

II. 제안방법

음성구간은 에너지의 강도, 에너지 변이 패턴, 스펙트럼 패턴 등의 특징이 비음성 구간과는 큰 차이를 보인다. 음성인식에 적용되는 음성구간 검출기의 특징벡터는 음성구간 검출과 음성인식과정의 계산 중복을 피하기 위해 동일한 특징 파라미터를 사용함이 일반적이다.

잡음환경에 적응하는 음성구간 검출기의 적응화 방법으로는 MAP(Maximum A Posteriori)와 ML(Maximum Likelihood) 적응 방법이 일반적으로 사용되나 이를 적용하기 위해서는 많은 데이터가 필요하기 때문에 실시간 적응화가 어렵게 된다. 그래서 고차원 공간상 야기되는 문제들을 LDA와 같은 차원 축소 방법을 사용하여 데이터 공간을 축소하게 되는데 이 또한 축소공간의 차원이 작을수록 원시 데이터의 정보손실이 커지게 된다. 그러므로 분류 성능의 저하를 최소화하는 비선형 차원 축소 방법이 필요하다.

2.1 선형 판별 분석(LDA)

선형 판별 분석은 원 특징 파라미터(X)에 포함된 분류 정보가 충분한 학습 샘플을 통해서 추정된 선형 변환 행렬(Θ)을 찾고, 선형 변환 행렬에 의하여 특징 벡터(Y)로 축소할 수 있다고 가정하고 있다. 식으로 나타내면 식 (1)과 같다.

$$Y=\Theta^{T}X \tag{1}$$

선형 변환 행렬을 통하여 특징 벡터를 2차원으로 축소 한 후, 우도비 검증의 방법으로 축소된 특징 벡터 Y가 어느 클래스에 속하는가를 결정한다.

선형 판별 분석(LDA)은 음성구간 검출 과정에서 제안하는 파라미터와의 비교를 위해 사용되었으며 그림 1은 LDA에 의해 버블 노이즈가 5dB의 레벨로 섞인 음성 데이터의 10차 MFCC 특징 파라미터를 2차 특징 파라미터로 축소한 후의 분포도이다.

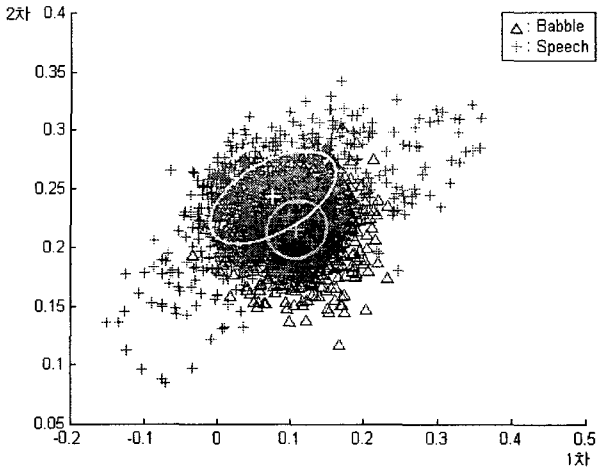


그림 99. LDA를 이용하여 축소한 MFCC 분포도 (버블 노이즈, 5dB)

Fig. 1. Scatter plot of Reduced MFCC using LDA (Babble noise, 5dB)

2.2 제안한 차원 축소 방법

음성/비음성 클래스 우도를 이용한 비선형 차원 축소 방법은 N개의 클래스를 구별하려고 할 때, N차원 공간상으로 비선형적인 사영 값을 얻을 수 있다. 이때, 각 차원은 개별 클래스에 대한 확률 값으로 단조함수이며 보통 로그 함수를 사용한다. 결론적으로 식 (2)와 같이 D차원 특징 파라미터 X는 N차원 특징 파라미터 Y로 축소된다.

$$\begin{aligned} Y&=[\log(P(X|C_1))\ \log(P(X|C_2))\cdots\log(P(X|C_N))] \\ &=[y_1 y_2\cdots y_N] \end{aligned} \tag{2}$$

여기서 $\log(P(X|C_i))$는 클래스 C_i의 확률 밀도 함수에 의한 특징 파라미터 X의 로그 우도이다. 이것이 새로운 특징 파라미터 Y의 i번째 요소인 y_i를 구성한다. 식(2)을 통하여 새로운 N차원의 우도 공간으로 사영됨을 알 수 있다.

그림 2는 우도를 이용하여 버블 노이즈가 5dB의 레벨로 섞인 음성 데이터의 10차 MFCC 특징 파라미터를 2차 특징 파라미터로 축소한 후의 분포도이다. 음성구간과 노이즈 구간인 비음성구간이 서로 분류 가능하게 축소 되었음을 분포도를 통해 알 수 있다. 그림 1과 비교할 때 음성구

간과 비음성구간의 겹침 정도가 더 작은 것을 분포도를 통해 알 수 있다.

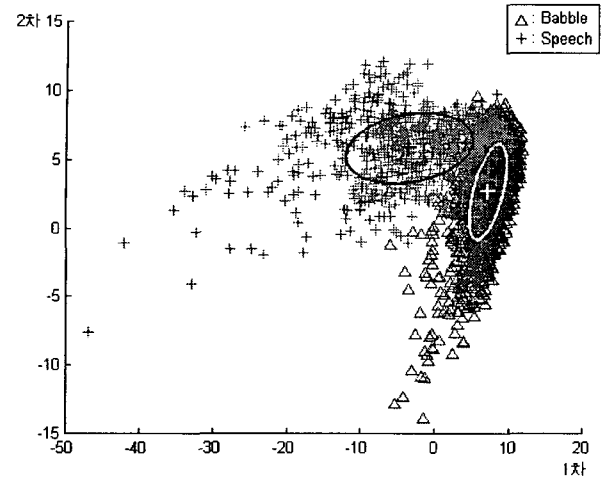


그림 100. 우도를 이용하여 축소한 MFCC 분포도 (버블 노이즈, 5dB)

Fig. 2. Scatter plot of Reduced MFCC using Likelihood (Babble noise, 5dB)

각각의 잡음 환경에 대한 혼합 밀도 함수로부터 우도값을 취해 축소한다. 축소한 특징 파라미터를 이용한 분류는 우도비 검증의 방법을 사용한다.

우도비 검증의 방법은 축소된 특징 파라미터 Y가 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_1을 선택하고 그렇지 않다면 c_2를 선택한다.

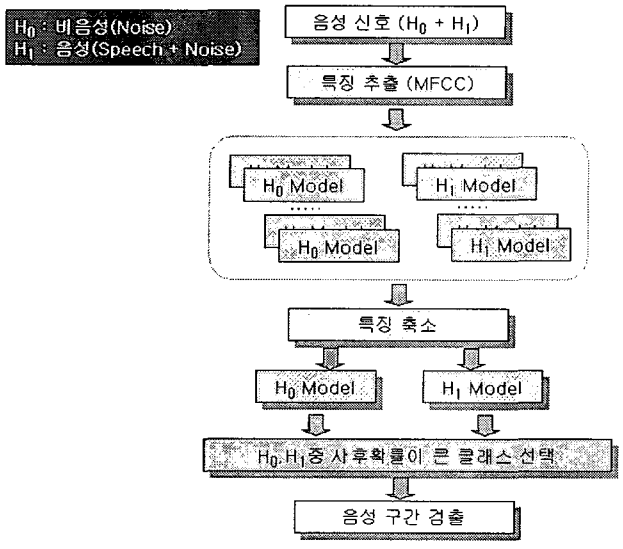


그림 101. 우도-기반 차원축소를 이용한 음성구간 검출기

Fig. 3. VAD using Likelihood-Based Dimension Reduction

음성/비음성 분류에 이 축소 방법을 적용할 경우 그림 3과 같은 방법으로 2차원 특징 파라미터로 축소가 가능하다.

Ⅲ. 실험 및 결과

3.1 실험 데이터

1) 음성구간 검출 실험 데이터

표 1. 음성구간 검출 실험 데이터

Table 1. Voice Activity Detection Experiment Data

학습 데이터	종류	ETRI 한국어 중가 마이크 음성인식용 낭독체 문장
	학습자 수	3명
	문장 수	3문장
실험 데이터	종류	ETRI 한국어 중가 마이크 음성인식용 낭독체 문장
	학습자 수	10명 (학습자 제외)
	문장 수	10문장

2) 음성 인식 실험 데이터

표 2. 음성 인식 실험 데이터

Table 2. Speech Recognition Experiment Data

학습 데이터	종류	ETRI 샘돌이 데이터
	학습자 수	10명 3회 발성
	문장 수	숫자음 10개
실험 데이터	종류	ETRI 샘돌이 데이터
	학습자 수	20명 4회 발성 (학습자 포함)
	문장 수	숫자음 10개

3) 음성 데이터의 분석조건

표 3. 음성 데이터의 분석 조건

Table 3. Analysis Condition of Speech Data

Speech Format	PCM Raw data
A/D Conversion	16㎑, 16bit
Frame Size	320 samples (20ms)
Overlap Size	160 samples (10ms)
Pre-emphasis	0.97
FFT Size	512 (zero padding)
Mel Filter Bank Number	20
MFCC Order	10(except power)

4) 잡음 데이터

잡음 데이터는 NOISEX-92를 사용하였고 표 4와 같다. 이 잡음은 19.98㎑, 16bit의 anti-aliasing 필터링 된 데이터로 본 실험에서는 16㎑, 16bit로 변환하였고, 잡음 환경 구현을 위해 5dB, 15dB, 25dB의 레벨로 음성 데이터와 잡음 데이터를 섞어서 학습 데이터 및 테스트 데이터를 만

들었다.

표 4. 노이즈엑스-92 모델
Table 4. NOISEX-92 Model

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

5) 음성 특징 파라미터

앞서 언급한 바와 같이 분류-기반의 음성 구간 검출을 위한 특징 파라미터는 ASR 시스템에서 사용되는 특징 파라미터인 MFCC를 기준 특징 파라미터로 사용하는 것이 적절한 것으로 사료된다.

3.2 음성구간 검출 실험

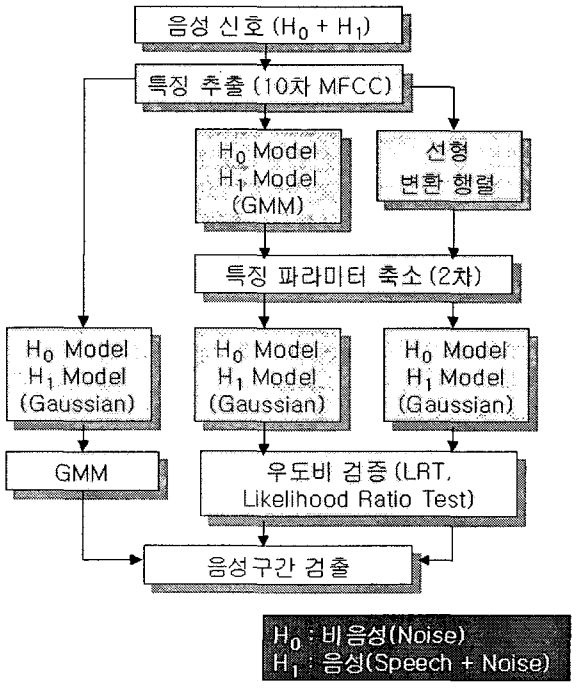


그림 102. 음성구간 검출 실험 과정
Fig. 4. Voice Activity Detection Experiment Procedure

음성구간 검출 시스템의 구성은 그림 4와 같다. 그림 4의 음성구간 검출에 대한 실험은 세 가지로 나누어 실험을 하였는데, 첫 번째 방법은 ①특징 파라미터를 축소하지 않고 10차 MFCC를 사용하여 GMM(Gaussian Mixture Model)을 이용하여 분류한 경우이고, 두 번째는 ②제안한 우도-기반 차원 축소 방법에 의해 2차원으로 축소한 특징 파라미터를 우도비 검증의 방법으로 분류한 경우이다. 세 번째는 ③LDA에 의해 2차원으로 축소한 특징 파라미터를 우도비 검증의 방법으로 분류한 경우이다.

음성구간 검출에 사용된 데이터와 분석 조건은 표 1과 표 3과 같다. 그리고 NOISEX-92 잡음 데이터를 25dB, 15dB, 5 dB의 세 가지 레벨로 섞어 잡음환경을 구현하였다.

표 5 ~ 표 7은 25dB, 15dB, 5dB의 잡음레벨에 따른 음성구간 검출율이다. 검출율은 프레임 별로 음성/비음성 분류된 결과를 잡음이 없는 클린 스피치에서의 음성/비음성 결과와 비교하여 일치하는 정도를 나타낸다.

표 5. 음성구간 검출 결과 (25dB)
Table 5. Result of VAD (25dB)

	①	②	③		①	②	③
N1	93.08	93.07	85.19	N8	96.86	97.14	96.68
N2	95.11	95.16	93.41	N9	93.65	94.11	92.16
N3	96.11	96.14	85.95	N10	94.10	94.68	93.35
N4	94.68	94.64	91.91	N11	92.00	92.09	78.92
N5	96.37	96.63	95.11	N12	96.07	96.22	93.48
N6	95.59	95.66	95.86	N13	92.99	94.44	88.96
N7	95.48	95.55	86.20	N14	94.83	95.29	90.46

표 6. 음성구간 검출 결과 (15dB)
Table 6. Result of VAD (15dB)

	①	②	③		①	②	③
N1	85.16	84.45	78.48	N8	94.89	94.96	93.04
N2	91.64	91.85	87.08	N9	92.34	93.28	86.39
N3	92.71	92.79	82.87	N10	90.09	90.42	87.64
N4	89.26	90.08	81.30	N11	88.93	89.70	80.42
N5	91.70	92.05	87.14	N12	91.00	91.52	87.77
N6	89.69	90.06	73.84	N13	94.51	94.41	85.57
N7	92.44	92.24	84.94	N14	92.96	92.94	88.58

표 7. 음성구간 검출 결과 (5dB)
Table 7. Result of VAD (5dB)

	①	②	③		①	②	③
N1	69.55	66.76	66.31	N8	87.99	88.35	78.23
N2	81.59	81.32	74.15	N9	88.98	89.42	80.43
N3	85.28	85.27	69.01	N10	77.53	77.38	68.26
N4	72.88	71.88	56.96	N11	87.23	85.81	77.60
N5	79.33	78.75	62.11	N12	80.15	79.95	69.26
N6	70.87	70.03	58.66	N13	93.12	93.30	72.33
N7	85.19	84.85	67.94	N14	85.81	85.55	83.75

제안된 축소 방법을 이용한 음성구간 검출 실험을 통해

10차 MFCC를 사용한 경우와 2차로 축소한 경우의 음성구간 검출 성능에 큰 차이가 없음을 알 수 있다.

3.4 음성 인식 실험

제안된 방법을 사용하여 음성 구간을 검출하고 음성 인식을 하는 실험을 수행하였다. 실험에 사용된 음성 데이터와 분석조건은 표 2와 표 3과 같다. 그리고 음성 구간 검출 실험과 동일한 NOISEX-92 잡음 데이터를 사용하여 잡음 환경을 구현하였다.

음성 인식 실험을 위한 음성구간 검출에 대한 실험은 두 가지로 나누어 실험을 하였다. 아래 표 8 ~ 표 10은 각각의 잡음 레벨에 따른 음성 구간 검출율을 나타낸다.

표 8. 음성구간 검출 결과 (25dB, 샘돌이 데이터)
Table 8. Results of VAD (25dB, Samdori Data)

	④	⑤		④	⑤
N1	98.14	97.00	N8	97.12	96.88
N2	98.86	98.54	N9	99.00	99.25
N3	98.87	98.72	N10	98.72	98.86
N4	98.36	97.86	N11	99.12	98.54
N5	97.54	97.10	N12	98.88	98.10
N6	98.90	98.72	N13	99.24	99.00
N7	98.25	98.50	N14	98.04	97.37

표 9. 음성구간 검출 결과 (15dB, 샘돌이 데이터)
Table 9. Result of VAD (15dB, Samdori Data)

	④	⑤		④	⑤
N1	95.14	96.30	N8	96.42	95.30
N2	97.35	98.02	N9	97.50	97.47
N3	98.20	97.41	N10	98.00	97.80
N4	97.87	96.45	N11	97.52	96.05
N5	96.32	96.25	N12	95.20	94.10
N6	98.00	97.22	N13	99.00	98.57
N7	96.30	96.72	N14	96.04	95.64

표 10. 음성구간 검출 결과 (5dB, 샘돌이 데이터)
Table 10. Result of VAD (15dB, Samdori DATA)

	④	⑤		④	⑤
N1	93.27	92.58	N8	95.70	96.00
N2	96.15	95.44	N9	92.17	94.02
N3	94.45	93.60	N10	93.35	94.27
N4	94.37	92.33	N11	93.02	92.05
N5	95.12	94.75	N12	92.66	91.15
N6	93.15	91.12	N13	97.58	97.69
N7	94.50	90.80	N14	94.27	93.24

음성구간 검출 결과 실험의 첫 번째 방법은 ④특징 파라미터를 축소하지 않고 10차 MFCC를 사용하여 GMM을 이용하여 분류한 경우이고, 두 번째는 ⑤제안한 우도-기반 차원 축소 방법에 의해 2차원으로 축소한 특징 파라미터를 우도비 검증의 방법으로 분류한 경우이다.

음성 인식 실험을 위해서 HMM(Hidden Markov Model) 알고리즘을 사용하였다. 그림 5 ~ 그림 7은 음성인식 결과이다.

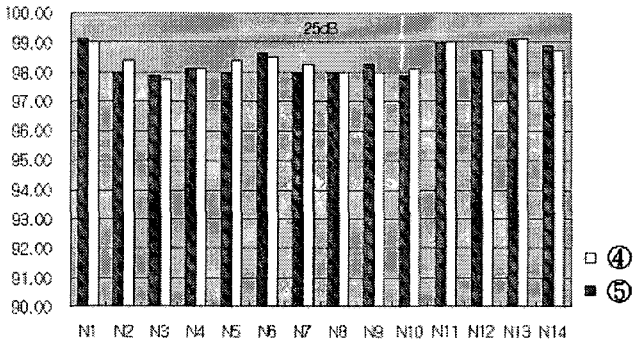


그림 103. 인식 결과 (25dB, 샘돌이 데이터)
Fig. 5. Results of Recognition (25dB, Samdori Data)

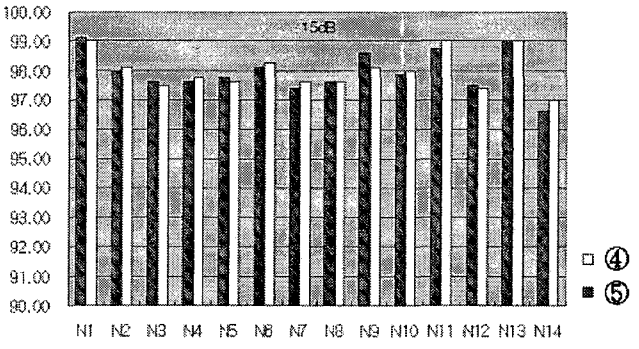


그림 6. 인식 결과 (15dB, 샘돌이 데이터)
Fig. 6. Result of Recognition (15dB, Samdori Data)

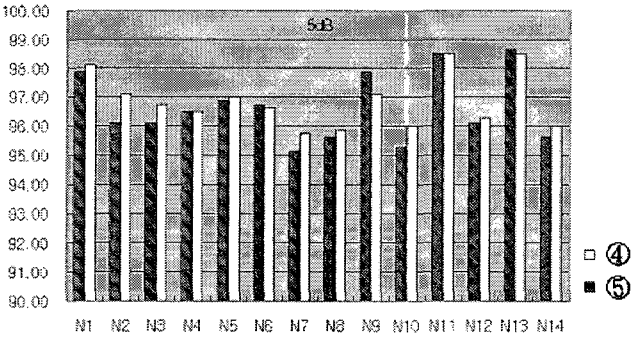


그림 7. 인식 결과 (5dB, 샘돌이 데이터)
Fig. 7. Result of Recognition (5dB, Samdori Data)

Ⅳ. 결 론

본 논문에서는 잡음 환경에서의 통계적 모델에 대한 실시간 적응화를 위한 고차원의 특징 파라미터의 차원 축소 방법을 제안하였다. ETRI 음성 인식용 낭독체 문장 데이터를

사용한 음성 구간 검출 실험 결과(표 5 ~ 표 7)를 참조하면 제안된 특징 축소 방법(②)이 원래 특징 파라미터인 MFCC 10차의 결과(①)와 비교해서 노이즈 별로 분류율의 차이는 있었지만 거의 대등한 성능을 얻을 수 있음을 확인할 수 있었다. 반면에 LDA에 의한 차원 축소 방법(③)은 낮은 분류율과 잡음에 따라 큰 폭의 편차를 보여줌을 확인할 수 있었다. 따라서 2차원으로 차원을 축소할 경우 LDA에 의한 방법이 아닌 제안된 방법을 사용하는 것이 더 우수한 성능을 보여줌을 알 수 있었다.

두 번째 실험인 ETRI 샘돌이 숫자음 데이터를 사용한 음성 구간 검출(표 8 ~ 표 10) 및 음성 인식 실험 결과(그림 5 ~ 그림 7)를 통해 제안된 방법(⑤)을 이용한 음성 인식에도 유효한지 확인하였다. 실험 결과 음성 인식률에서 10차 MFCC를 사용하여 음성 구간 검출한 결과(④)와 대등함을 확인할 수 있었다.

따라서 제안된 방법으로 2차원으로 차원을 축소할 경우, 계산 시간 단축으로 인해 음성/비음성 분류를 위한 통계적 모델에 대한 빠른 적응화가 가능하여 다양한 잡음 환경에서의 실시간 음성 인식기를 위한 음성구간 검출기에 적용이 가능할 것으로 생각된다.

참 고 문 헌

[1] Bhiksha Raj and Rita Singh, "Classifier-based nonlinear projection for adaptive endpointing of continuous speech," YCSLA 204 Article in Press, 12 May 2002

[2] Jean-Claude Junqua, Brian Mak and Ben Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise," IEEE Trans. Speech and Audio Processing, Vol.2 No.3, pp.406-412, July 1997

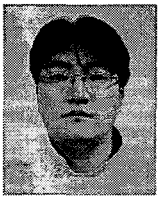
[3] M.H. Savoji, "Endpointing of Speech Signals," Speech Communication, Vol.8 No.1, pp.46-60, March 1989

[4] Nikos Doukas, Patrick Naylor and Tania Stathaki, "Voice Activity Dectection Using Source Separation Techniques," Signal Processing Section, Proc. Eurospeech, 1997.

[5] 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.

[6] L. F. Lemel, "An improved endpoint detection for isolated word recognition," IEEE Trans. Acoust. Speech and Signal Processing, Vol.2, No.3, pp.406-412, 1994.

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



박 진 영(Jin-young Park)
2002년 2월 동아대학교 전자공학과(공학사)
2004년 2월 동아대학교 전자공학과(공학석사)
2006년 2월 동아대학교 전자공학과(박사과정)
관심분야 : 음성구간검출, 음성인식, 화자인식, DSP 응용 분야

이 광 석(Kwang-seok Lee)
1983년 2월 동아대학교 전자공학과(공학사)
1985년 2월 동아대학교 전자공학과(공학석사)
1992년 2월 동아대학교 전자공학과(공학박사)
2004년 2월~2005년 1월 미국 애리조나 주립대학 객원교수
1995년~현재 진주 산업대학교 전자공학과 부교수
관심분야: 음성신호처리 및 인식, 생체 신호처리, 지능화 기술

허 강 인(Kang-in Hur)
1980년 2월 동아대학교 전자공학과(공학사)
1982년 2월 동아대학교 전자공학과(공학석사)
1990년 8월 경희대학교 전자공학과(공학박사)
1984년~현재 동아대학교 전자공학과 교수
1988년 9월 ~ 1989년 8월 : 일본 객원연구원
1992년 9월 ~ 1993년 8월 : 일본 객원연구원
관심분야: DSP, 음성인식, 합성, 신경회로망