성대마이크를 이용한 ASR 시스템 개발을 위한 인식기 최적화

Recognizer Optimization for a Isolated-word Recognition system using Throat Microphone

정영규, YoungGiu Jung*, 한문성, MunSung Han**, 이상조, SangJo Lee***
*한국전자통신연구원, **한국전자통신연원, ***경북대학교

요약 성대마이크는 디바이스의 특성상 환경 잡음을 최소화하는 장점이 있다. 그러나 고주파정보의 손실과 부분적인 포먼트 정보의 손실 때문에, 성대마이크를 이용한 명령어 인식기는 표준마이크를 이용한 명령어 인식기보다 낮은 성능을 보인다. 본 논문은 한국어 음운자질의 특성을 적용한 특징추출 알고리즘과 최적화된 인식모델을 이용하여 높은 성능을 갖는 명령어 인식시스템을 제안한다. 성대 울림 특성이 한국어 내의 분포 분석하여 성대 울림 정보만으로 명령어 인식기 개발이 가능함을 보이고 음성인식에 높은 성능을 보이는 Time Delay Neural Network(TDNN)[1]을 성대신호 명령어 인식에 최적화한 구조를 제안한다. 실험을 통해 찾은 최적 TDNN 구조를 성대신호에 적용한 했을 때 약 87%의 높은 성능을 보였다.

핵심어: 성대마이크, TDNN, 파라미터 최적화, 성대신호 명령어 인식기, 성대신호분석

1. 서론

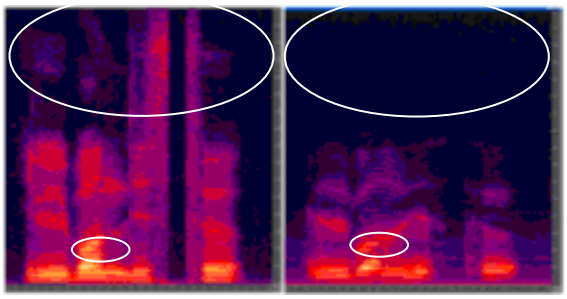
현재 음성인식 기술은 잡음이 많은 환경에서는 적정 수준의 성능을 내지 못하는 문제점이 있다. 이러한 문제점을 극복하기 위해서 두 가지 방법으로 연구가 진행되고 있다. 하나는 잡음처리 알고리즘에 의한 접근 방법이고, 다른 하나는 환경 잡음의 영향을 적은 디바이스-골도마이크, 이어마이크, 성대마이크- 를 이용한 음성인식기 기술에 관한 연구이다.

잡음처리 알고리즘을 이용한 잡음 처리 방법은 많은 연구가 진행되고 있으나 아직 만족할 만한 성과를 보이고 있지는 못한다. 실제적으로 대형컨퍼런스내에서 명령어 인식기조차 환경 잡음으로 인해 잘 활용되고 있지 못하다. 다른 한편으로 진행되고 있는 Noise-free한 디바이스를 이용한 음성인식기 개발은 디바이스 장점 때문에 Noise-free 마이크만을 이용한 Automatic Speech Recognition(ASR) 시스템 개발 연구가 진행되었으나 마이크의 한계로 인하여 현재 대부분 음성신호의 보조정보로 사용하는 연구가 주를 이룬다. 그러나 현재 음성인식기술에서의 노이즈 제거 기술의 한계를 볼 때 Noise-free 마이크만을 이용한 음성인식 기술의 개발은 음성인식 시스템의 상용화에 큰 기여를 할 것으로 보인다.

ASR 시스템 개발에 가장 중요한 모듈은 특징추출 모듈과 인식 모듈이다. 특징 추출 모듈에서 가장 중요한 것은 해당 신호를 충분히 모델링할 수 있는 특징 추출 알고리즘을 선택하는 것이다. 본 논문은 성대신호의 특징을 스펙트로그램을 이용하여 설명하고 한국어 내 성대 울림자질을 갖는 음운의 분포를 분석하여 성대마이크만을 이용한 명령어 인식기 개발이 가능함을 설명한다. 그리고 최적 특징추출 알고리즘을 실험을 통해 보인다.

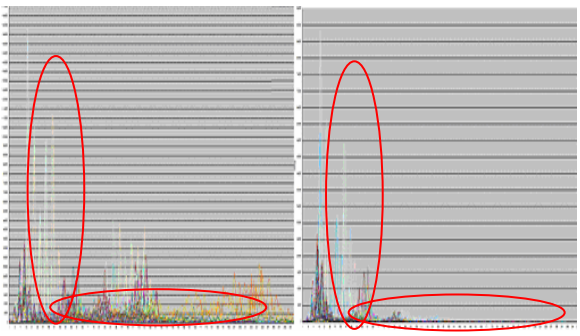
두 번째로 중요한 모듈인 인식 모듈은 입력신호에 따라 다양한 구조를 갖는다. 본 논문은 성대신호에 최적화된 인식 모델을 제안하다. 음성 ASR 시스템에 널리 사용되고 높은 성능을 갖는 TDNN을 다양한 실험을 통해 성대신호 분석에 적합하도록 최적화한다.

그림 1은 성대마이크와 일반 마이크의 주파수 정보를 비교한 것이다. 그림1에서 보면 성대신호는 고주파가 거의 없으며, 저주파에서 부분적으로 포먼트의 손실이 있는 것을 알 수 있다. 따라서 성대마이크를 이용한 명령어 인식기 개발에서의 가장 큰 어려움은 손실된 정보에서 어떻게 유용한 정보를 추출하는가 이다.



(a) 표준 마이크출력 (b) 성대마이크 출력
그림1 성대마이크와 표준마이크의 스펙트럼 비교

그림 2는 성대신호를 실제 음성 ASR시스템에 널리 사용되는 MFCC를 통과한 결과를 비교한 것이다.



(a) 표준 마이크출력 (b) 성대마이크 출력

그림2 MFCC를 통과한 성대신호와 음성신호 비교

그림 2에서 가로축은 주파수영역을 256개로 나눈 인덱스이고 세로축은 주파수영역에 포함된 에너지 값이다. 그리고 다양한 색은 개별 프래임을 나타낸다. 그림 2에서 보듯이 2kHz이하에서는 두 신호의 에너지 분포가 비슷한 분포를 보인다. 그러나 2kHz에서부터 성대신호의 정보량이 감소하기 시작되고, 4kHhz이상에서는 거의 정보가 없음을 알 수 있다. 또한 저주파 영역에서도 많은 정보의 손실이 발생됨을 알 수 있다.

그림1과 그림2를 통해 알수 있듯이 성대 신호의 정보량은 음성신호에 비해 현저히 낮음을 알 수 있다.

2. 관련연구

Noise-free device를 사용한 인식 기술의 연구는 일본과 미국 등에서 부분적으로 이루어지고 있다. Nakajima et al.[3]는 non-audible murmur(NAM) 인식을 위해 피부에 붙이는 형태의 새로운 입력 인터페이스를 제안하였다. 이러한 Stethoscopic microphone은 흡입 디스크와 폴리에스테르 판으로 구성된다. NAM sampling을 사용하여 학습 시킨 결과 자음은 거의 인식되지 않았으며, 모음은 잘 인식되었다. 이러한 문제를 해결하기 위해 자음, 모음에 대해 균형있는 power ratio를 제공하는 최적의 센싱 위치를 제시하였다.

Jou et al.[4]은 성대 마이크를 사용하여 녹음된 soft whisper를 인식하기 위한 다양한 adaptation 기술들을 비교 설명한다. 본 논문에서 사용된 Adaptation들은 Maximum likelihood linear regression, feature-space adaptation 그리고 downsampling, sigmoidal low-pass filter나 linear multivariate regression을 통한 재학습 방법들이 있다. 또한 Zheng et al.[5]은 bone-conductive microphone과 regular air-conductive microphone, In-ear microphone, Throat microphone들을 결합하여 높은 노이즈 환경에서 음성 검출 및 speech enhancement 와 인식률 향상에 대한 연구를 진행하였다.

Dupont, et al. [6]은 성대 마이크와 일반 마이크의 입력 신호를 동시에 받은 다음 각각의 acoustic models에 의해 제공되는 확률벡터를 결합하여 음성인식의 성능을 향상 시키는 방법을 제안한다. Graciarena et al. [7]은 성대마이크를 일반 마이크의 보완 센서로 사용하여 두 마이크로부터 들어온 입력신호에서 노이즈 Mel-cepstral feature를 추출하여 clean standard microphone mel-cepstral features로 변환하는 알고리즘을 제안하였다. 그리고 정영규[8] 는 성대마이크를 이용한 ASR 인식 시스템에서 성대신호를 대상으로 한 특징에 Zero-Crossing 과 Peak가 유용함을 실험을 통해 보였다.

Noise-free 디바이스를 이용한 ASR 시스템 개발에 있어서 특징벡터 선택의 중요성과 함께 최적 인식기의 개발도 매우 중요하다. Masahide Sugiyama et al.[9]는 음성 인식이 널리 사용되는 TDNN의 구조를 설명하고, 일본어 음운들과 절에 최적화된 TDNN을 제안하였다. 그리고 이를얀ㅊㄱㄷㅅㄷ hmm, Continuous HMM, LVQ, LVQ+HMM, Fuzzy LVQ+HMM와 성능을 비교 측정하였다. Ashouri[10]는 고차원 통계적 방법과 TDNN을 기반으로 하는 두 개의 숫자음 고립어 인식기를 비교 설명하고, 숫자음 인식에 높은 성능을 보이는 TDNN을 제안하였다.

다음절에서는 성대마이크만을 이용한 명령어 인식기의 개발 가능성을 확인하기 위해 한국어내 성대울림 자질을 분석한다. 그리고 성대신호의 특성을 반영한 최적의 TDNN 구조를 제안한다.

3. 인식기 최적화

3.1 한국어음운자질과 성대신호와의 관계

발성기관은 폐, 기관, 후두, 인두, 코, 입, 입술등이 있는데, 이들이 일체가 되어 폐에서 입술로 이어지는 복잡한 관강을 형성한다. 인간이 이러한 발성기관을 통해 의미를 갖는 말을 생성하기 위해서는 그 나라의 문자가 갖는 음운적 특징을 오랜 시간 훈련을 해야 가능하다. 이렇듯 문자의 음운적 특징은 특징 분석의 가이드 라인을 제시해줄 수 있다.

한국어는 음소 문자로써 자음과 모음으로 이루어져 있으며 이를 음절단위로 조합해서 글자를 나타낸다. 모음은 총 21개로 모두 유성성의 특징을 갖는다. 자음의 경우 총 19자인데 형태와 위치에 따라 유성음이 되기도 하고 무성음이 되기도 한다. 한국어가 음절을 이루는 원리는 자음+모음+자음 또는 자음+모음, 모음+자음, 모음들 중에 한가지 경우이다. 그리고 이러한 음절은 그 차체로 음운자질을 갖거나 발성할 때 음운자질을 갖게 된다.

음운자질이란 하나의 음운을 다른 음운과 구별해주는 성질이다. 한국어를 위한 음운자질은 총 14개로 구성된다. 본 논문은 14개의 음운자질 중에서 성대 진동과 관련된 자질을 자음/모음과 연관시켜서 설명한다.[11] 이를 통해 한국어를 대상으로 성대마이크만을 이용하여 ASR시스템을 개발할 수 있음을 보인다.

음운자질은 음향적, 조음적 관점에서 공명성(sonorant), 자음성(consonantal), 성절성(syllabic) 분류된다. 공명성은 성대의 울림자질을 나타내는 것이다. 한국어는 모든 모음이 공명성을 가지며 일부 공명 자음이 존재한다. 표 1은 한국어 자음의 성대울림에 관련된 음운자질 표이다.

표 1 한국어 자음의 성대울림 특성을 갖는 음운자질표[12]

자음	ㅂ	ㅃ	ㅍ	ㄷ	ㄸ	ㅌ	ㄱ	ㄲ	ㅋ	ㅅ
공명성	-	-	-	-	-	-	-	-	-	-
긴장성	-	+	+	-	+	+	-	+	+	-
기식성	-	-	+	-	-	+	-	-	+	-
자음	ㅆ	ㅈ	ㅉ	ㅊ	ㅁ	ㄴ	ㅇ	ㄹ	ㅎ	
공명성	-	-	-	-	+	+	+	+	-	
긴장성	+	-	+	+	-	-	-	-	-	
기식성	-	-	-	+	-	-	-	-	+	

공명 자음은 “ㅁ”, “ㄴ”, “ㅇ”, “ㄹ” 이다. 그 외 나머지 자음은 공명성을 갖지 않는다. 그리고 나머지 자음은 두 개의 자질을 이용하여 분류한다. 그것은 긴장성과 기식성이다. 긴장성은 평음(ㅂ,ㄷ,ㄱ,ㅅ)을 경음(ㅃ,ㄸ,ㅉ,ㄲ)/격음(ㅍ,ㅌ,ㅊ,ㅋ)과 구분하는 자질로써, 성대의 긴장을 동반한다. 경음과 격음이 긴장성 자질을 가진다. 그러나 긴장성 자질을 갖는 경음과 격음은 기식성의 차이를 보인다. 기식성은 성대의 진동을 어렵게 만드는 자질이다. 따라서 경음인 “ㅃ”, “ㄸ”, “ㅉ”, “ㅆ”, “ㄲ”은 성대 마이크로 검출 가능한 음운이다.

다음으로 중요한 특징은 음운 변화이다. 중요한 음운 현상으로 비음화와, 유성음화가 있다. 비음화는 [+자음성, -공명성] 을 갖는 (ㅂ,ㄷ,ㄱ,ㅍ,ㅌ,ㅋ,ㄲ,ㅅ,ㅆ,ㅈ,ㅊ,ㅎ) 이 비음인 (ㅁ,ㄴ,ㅇ) 을 만나 비음으로 바뀌는 현상이다. 유성음화는 공명성과 공명성의 자질을 갖는 음운들 사이에서 장애음이 유성음으로 변화는 현상이다. 한국어 자음에서는 (ㄱ,ㄷ,ㅂ,ㅊ,ㅎ)이 유성음과 유성음 사이에서 유성음화(intersonorant obstruent voicing) 된다. 이렇듯 19개의 자음 중에서 9개의 자음이 유성성을 가지며 나머지 10개의 자음도 비음화와 유성음화 현상으로 인해 유성음화 되는 현상이 단어 내에서 빈번히 나타난다. 따라서 성대 신호와 같이 부분적 정보 손실이 있는 신호라 할 지라도 이러한 언어적 특징을 본다면 유성음을 정확히 특징으로 모델링할 수 있는 특징추출 알고리즘을 사용한다면 성대마이크만을 사용한 ASR 시스템이 개발 가능함을 예상할 수 있다.[8]

3.2 성대신호에 최적화된 TDNN의 구조

인식기 개발에 있어서 가장 중요한 개발 모듈은 끝점 검출과 특징추출 그리고 인식모듈이다. 성대신호를 대상으로 할 때 보다 많은 연구가 필요한 부분이 특징추출 모듈과 인식 모듈이다. 성대신호 분석을 위한 특징추출 알고리즘은 선행 연구에서 밝힌 바와 같이 Zero-Crossing with Peak Amplitude(ZCPA)이 가장 적합하다[14].

ZCPA 알고리즘은 음성샘플을 주파수 정보로 변환하는 band pass cochlear 필터모듈, band별 Zero-crossing(ZC) 검출모듈, ZC내에 Peak 검출 모듈, 그리고 윈도우내 Peak 검출모듈로 구성된다. 따라서 ZCPA는 ZC와 Peak를 이용하여 시간 정보와 세기 정보를 반영하여 성대신호와 같이 낮은 정보량을 갖는 신호도 적합하게 모델링 할 수 있다. 수식 1은 시간 t에서 ZCPA의 출력이다.

$$y(t:i) = \sum_{channel} \sum_{k=1} \delta_{ij_k} f(A_k), \quad 1 \le i \le N \quad (1)$$

k는 각 channel에서 upward zero-crossing의 수이고, N은 frequency bin의 수이다. 그리고 j_k 는 k번째와 (k+1)번째 zero crossing을 이용하여 계산된 frequency bin의 인덱스이고 A_k 는 peak amplitude이다. 그리고 f는 log 함수가 사용된다. 마지막으로 δ_{ij} 는 Kronecker delta이다.

ZCPA를 통과하여 생성된 특징 벡터은 본 논문에서 제안한 TDNN모델을 거쳐 인식된다. TDNN은 음성인식 시스템에서 높은 성능을 보여왔다. 많은 Neural Network들에 사용된 기본 단위는 입력 노드와 weight의 곱들의 합을 계산 한 후 nonlinear 함수를 통과한 결과를 사용한다. 이때 nonlinear 함수로 threshold 나 sigmoid 함수를 사용한다. 그림 2는 TDNN unit의 그림이다.

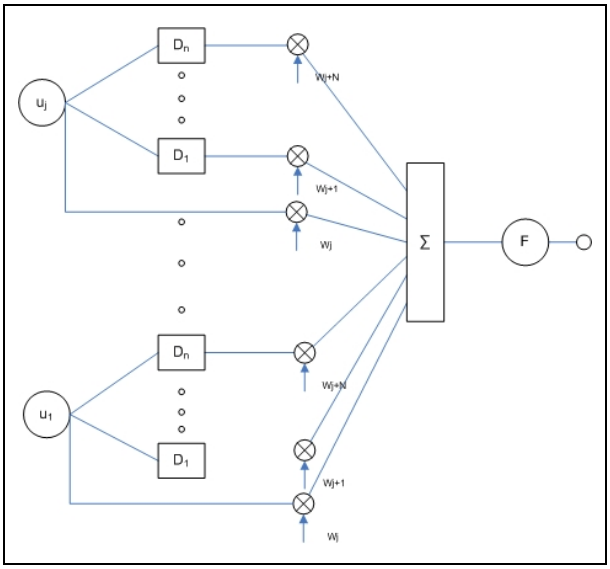


그림 2. TDNN의 unit

제안된 TDNN은 세 개의 레이어로 구성된다. 64 프레임으로 구성된 입력레이어와 3개의 입력프레임을 중첩하여 한 개의 프레임값을 결정하는 첫번째 은닉 레이어, 그리고 5개의 프레임을 중첩하여 한 개의 프레임값을 결정하는 두번째 은닉 레이어, 그리고 인식할 단어 수를 나타내는 출력 레이어로 구성된다.

각 입력 프레임은 전체 16개의 노드로 구성된다. 이는 주파수별 다른 윈도우 길이를 사용하는 ZCPA 알고리즘의 출력이다. 첫번째 은닉 레이어의 각 프레임도 16개의 노드로 구성된다. 은닉 레이어의 각 노드는 입력 레이어의 3개 프레임 16개 노드들과 완전 연결을 이룬다. 두번째 은닉노드는 각 프레임이 4개의 노드로 구성된다. 은닉 레이어의 각 프레

임에 각 노드는 첫번째 은닉 레이어의 5 프레임 16개 노드들과 완전 연결을 이룬다. 마지막으로 출력노드는 두번째 은닉 레이어의 각 프레임에 중첩없이 완전 연결을 이룬다. 그림 3은 성대명령어 인식기를 위해 제안된 TDNN의 구조이다.

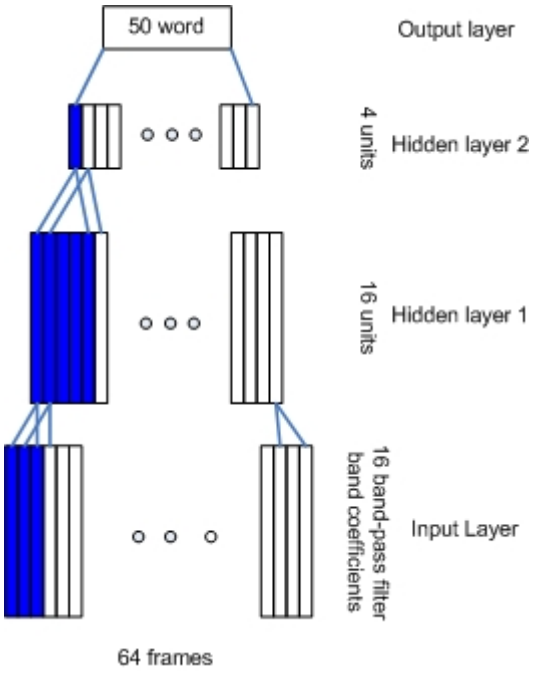


그림3. 성대 명령어 인식기에 최적화된 TDNN 의 구조

4. 실험 결과

본 장에서는 특징추출 알고리즘 따른 성능차이와 제안된 TDNN 모델에 따른 성능을 측정한다. 실험환경은 동일한 화자에 대해 일반 마이크와 성대 마이크를 통해서 녹음한 100명분 50단어를 학습데이터로 사용하고 25명분 50단어를 테스트데이터로 사용한다. 발성된 데이터는 16kHz-16bit으로 sampling된 PCM데이터를 사용한다.

인식기의 구조에 따른 성능 측정에 앞서 특징추출 알고리즘에 따른 성능을 비교한다. 표 2는 기존의 연구와 본 시스템의 특징벡터에 따른 성능이다. 이연철[15]은 화자 종속으로 성대마이크를 이용한 고립명령어 인식기를 개발하였다.

표 2. 알고리즘에 따른 성능 비교

	적용된 알고리즘	인식률(%)
이연철 시스템[5]	PLP Cepstrem	52.3
	MFCC + CMS	75.0
제안된 ASR	MFCC + CMS	74.8
	ZCPA	83.63
	ZCPA + RASTA	85.16

인식기 구조는 각 레이어의 프레임수, 프레임내에 노드수, 중첩 윈도우수에 따라 달라진다. 본 논문은 가장 높은 성능을 갖는 TDNN의 구조를 찾기 위해 레이어의 수와 레이어 내에 있는 노드의 수 그리고 중첩 윈도우 수를 다양하게 변화 시켜가며 성능을 측정하였다. 최초의 안정적 성능을 보이는 TDNN의 구조와 최적의 성능을 보이는 구조는 표 3과 같다.

표 3. 성대신호 ASR 시스템에 적용된 TDNN 의 구조

	레이어	프레임수	노드수	윈도우수
수정된 TDNN	입력	64	16	3
	은닉1	62	8	5
	은닉2	58	4	1
최적 TDNN	입력	64	16	3
	은닉1	62	16	5
	은닉2	58	4	1

표4에 제시된 두 개의 TDNN에 따른 인식률을 비교한 것이다. 본 실험에 적용된 특징추출 알고리즘은 ZCPA를 사용하여 채널 노이즈 제거를 위해 RelAtive SpecTrAl(RASTA)[16] 필터를 사용하였다.

표 4. TDNN 구조에 따른 인식률 비교

적용된 알고리즘	인식률(%)
수정된 TDNN	85.16
16-16-4 노드	87.1

5. 결론

기존의 연구에서 성대마이크를 이용한 ASR시스템의 개발은 성대신호가 갖는 단점(고주파정보의 부재와 부분적인 포먼트 정보의 손실) 때문에 표준 마이크 신호의 보조적인 정보로 사용되어왔다. 그러나 본 논문은 한국어 내 성대 울림 자질을 분석하여 성대마이크만으로 명령어 인식기 개발이 가능함을 보이고 성대 신호에 대한 특징추출 알고리즘으로 ZCPA 알고리즘을 제안하였다. 그리고 높은 성능의 인식기 개발을 위해 기존의 TDNN 구조를 다양한 실험을 통해 성대신호에 대해 높은 성능을 내도록 수정된 TDNN구조를 제안하였다.

제안된 TDNN은 세개의 레이어로 구성되며 입력레이어는 64개의 프레임에 개별 16개의 노드, 3개의 중첩 윈도우를 가진다. 첫번째 은닉레이어는 62개의 프레임에 개별 16개의 노

드, 5개의 중첩 윈도우를 가지며, 두번째 은닉레이어는 58개의 프레임에 개별 4개의 노드, 1개의 중첩 윈도우로 구성된다. 제안된 성대신호 화자 독립 명령어 인식기는 높은 노이즈 환경에서도 87%의 높은 성능을 보인다.

참고문헌

[1] A. Waibel, Phoneme Recognition Using Time-Delay Neural Networks, Report of Speech Committee, SP 87-100, pp. 19-24 Dec. 1987

[2] Fang Zheng, Guoliang Zhang and Zhanjiang Song, "Comparison of Different Implementations of MFCC," J. Computer Science & Technology, 16(6): 582-589, Sept. 2001.

[3]Nakajima. Y, Kashioka. H, Shikano. K and Campbel. N, "Non-audible Murmur Recognition Input Interface using Stethoscopic Microphone Attached to the Skin," ICASSP'03, volume 5, pp 708-11, 2003.

[4] S.C. Jou, T. Schultz, and A. Waibel, "Adaptation for Soft Whisper Recognition Using a Throat Microphone," in Proc. ICSLP, Jeju Island, Korea, Oct 2004.

[5] Zhengyoun Zhang, Zicheng Liu, Sinclair. M, A, Li Deng, Droppo. J, Xuedong Huang, Yanli Zheng, "Multi-sensory Microphones for Robust Speech Detection, Enhancement and Recognition," ICASSP'04, page:iii-781-6 vol3, May 2004.

[6] S. Dupont, C. Ris, "Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise," proc. of Robust 2004 (Workshop (ITRW) on Robustness Issues in Conversational Interaction), Norwich, Aug. 2004.

[7] M. Graciarena, H. Franco, K. Sonmez, H Bratt, "Combining Standard and Throat Microphones for Robust Speech Recognition," in IEEE Signal Processing Letters, Vol. 10 No.3, pp. 72-74, March 2003.

[8] 정영규,한문성,조관현, "성대신호 명령어 인식기를 위한 음운자질에 기반한 성대신호 연구," HCI2006 학술대회, Page 565-570, 2 월 2006.

[9] M. Sugiyama, H. Sawai, and Alex Waibel, "Review of TDNN(Time Delay Neural Network) Architectures for Speech Recognition," Proceedings of the ISCAS(International Conference on Circuits and System)'91, Singapore, Malaysia, June, 1996.

[10]M.R. Ashouri, "Isolated word recognition using a high-order statistic and time delay neural network," Proc of the 1997 IEEE Signal Processing Workshop, Page: 57-61, July 1997.

[11]Hyun-Oak Gu, *Understanding of Korean Phonology*, The Institute of Language Culture, 1999.

[12]구현옥, *한국 음운의 이해*, 한국문화사, 1999.

[13]신지영, 차재은, *우리말 소리의 체계*, 한국문화사, 2003

[14]Doh-Suk Kim, Soo-Young Lee, Rhee M. Kil, "Auditory Processing of Speech Signals for Robust Speech Recognition in Real-Word Noisy Environments," IEEE Tran. Speech and Audio Processing, vol., 7 No. 1, Jan., 1999.

[15]이연철, 이상훈, 홍훈섭, 한문성, 마평수, "성대마이크 입력 신호를 위한 음성인식 연구," 18 회 한국정보과학회, Vol9, PP, 747-750, 11 월, 2002.

[16] H. Hermansky, N. Morgan, "RASTA Processing of Speech," IEEE Trans. Speech Audio, 2(4), pp 578-589, 1994