자동 음성분할 및 레이블링 시스템의 구현

Implementation of the Automatic Segmentation and Labeling System

성 종 모*, 김 형 순*
(Jongmo Sung*, Hyung Soon Kim*)

※본 논문은 한국전자통신연구원 음성언어연구실 위탁연구과제 결과의 일부로서 부산대학교 정보통신연구소에서 수행한 것입니다.

요 약

본 논문에서는 한국어 음성 데이터베이스 구축을 위하여 자동으로 음소경계를 추출하는 자동 음성분할 및 레이블링 시스템을 구현하였다. 기존의 음성분할 및 레이블링 기술을 근간으로 본 시스템을 구현하였으며, 또한 사용자가 자동분할된 음소경계를 확인하여 그 경계를 쉽게 수정할 수 있도록 한글 모티프 환경에서 그래픽 사용자 인터페이스를 개발하였다. 개발된 시스템은 16kHz로 샘플링된 음성을 대상으로 하고 있으며, 레이블링 단위는 45개의 유사음소와 하나의 묵음으로 구성하였다. 그리고 언어학적 정보의 입력방식으로는 음소표기와 철자표기를 사용하였으며, 패턴매칭 방법으로는 hidden Markov model(HMM)을 이용하였다.

개발된 시스템의 각 음소 모델은 수작업에 의해서 음소단위로 분할한 음성학적으로 균형잡힌 445 단어 데이터베이스를 이용해서 훈련되었다. 그리고 본 시스템의 성능평가를 위해 훈련에 사용되지 않은 문장 데이터베이스에 대해서 자동 음성분할 실험을 수행하였다. 실험결과, 수작업에 의해서 분할된 음소경계위치와의 오차가 20ms 이내인 것이 74.7%였으며, 40ms이내에는 92.8%가 포함되었다.

ABSTRACT

In this paper, we implement an automatic speech segmentation and labeling system which marks phone boundaries automatically for constructing the Korean speech database. We specify and implement the system based on conventional speech segmentation and labeling techniques, and also develop the graphic user interface(GUI) on Hangul Motif™ environment for the users to examine the automatic alignment boundaries and to refine them easily. The developed system is applied to 16kHz sampled speech, and the labeling unit is composed of 46 phoneme-like units(PLUs) and silence. The system uses both of the phonetic and orthographic transcription as input methods of linguistic information. For pattern-matching method, hidden Markov models(HMM) is employed.

Each phoneme model is trained using the manually segmented 445 phonetically balanced word (PBW) database. In order to evaluate the performance of the system, we test it using another database consisting of sentence-type speech. According to our experiment, 74.7% of phoneme boundaries are within 20ms of the true boundary and 92.8% are within 40ms.

I. 서 론

음성인식 기술은 음성합성 기술과 함께 인간의 가장 편리한 의사전달 수단인 음성을 통해 인간이 컴퓨터와 대화할 수 있도록 해주는 도구로서 정보화의 진전과 더불어 그 필요성이 더욱 증대되고 있다[1]. 미국, 일본, 유럽 등 선진 각국에서는 1970년대 이전부터 음성인식에 대한 연구를 추진해 왔으며, 특히 국가 주도 형태의 대규모 프로젝트를 통해 많은 기술적 진보를 가져왔다. 그 결과 수십 단어 정도의 어휘를 대상으로 불특정 화자의 음성을 인식하거나, 화자적응을 통해 수만 단어를 인식할 수 있는 상용 시스템들이 등장했으며, 최근 미국에서는 전화 사업자가 음성인식에 의한 다이얼링 서비스를 시작하기에 이르렀다. 그러나, 자연스럽게 발음한 연속음성의 인식기술은 아직까지 성능면에서 크게 뒤떨어져 있으며, 이에 따라 최근의 음성인식연구는 주로 코퍼스(corpus) 기반의 접근 방식에 의한 대용량 어휘의 연속음성인식에 촛점이 모아지고 있다.

이러한 연속음성인식을 위해서는 우수한 성능을 가지

*부산대학교 전자공학과
접수일자:1997년 3월 19일

는 음성학적 모델과 연속음성에 적용가능한 언어모델의 개발, 그리고 운율정보의 효과적인 사용 등이 관건이다. 그 중에서도 음성학적 모델을 개선시키기 위해서는 많은 양의 데이터베이스를 수집하는 것만으로는 충분하지 않으며, 음성 데이터를 음소와 같은 음성의 기본단위로 분할하고 레이블링하여 그 다음 단계의 통계적 처리가 가능하도록 가공하는 작업이 필수적으로 요구된다. 실제로 미국 등 선진국에서는 TIMIT 데이터베이스와 같이 잘 가공된 음성 데이터를 구성하고 보급함으로써 많은 연구 그룹들이 양질의 동일한 음성 데이터를 토대로 한 경쟁적인 연구를 하게 되었고, 그 결과 음성인식기술에 많은 발전을 가져왔다. 따라서 한국어 음성처리기술의 발전을 위해서도 대용량의 음소분할 및 레이블링된 음성 데이터베이스를 구축하는 것이 시급한 과제이며, 구축작업의 원활한 진행을 위해서는 한국어 자동 음성분할 및 레이블링 시스템의 개발이 중요한 역할을 할 것으로 판단된다.

음성신호의 음성분할 및 레이블링 작업은 사람이 수작업에 의해서 직접 수행할 수 있지만, 수작업에 의한 음성분할 및 레이블링을 할 경우 다음과 같은 문제점을 지닌다[2]. 첫째로 스펙트로그램(spectrogram) 판독 및 반복되는 청취평가를 통해서 이루어지므로 매우 지루한 작업일 뿐만 아니라 많은 시간이 소요되게 된다. 둘째로 수작업에 의한 음성분할은 높은 수준의 음성학적 지식을 요하며, 소수의 음성학 전문가에게 의존할 수 밖에 없다. 세째로 음성경계 선정을 위한 구체적인 판단기준을 미리 정해놓더라도 상당 부분 주관적인 판단을 피할 수 없으며, 이로 인해 음성경계 선정과정에서 일관성이 보장되지 못한다[3]. 따라서 서로 다른 음성학 전문가들이 동일한 음성 데이터를 분할할 경우는 물론이고, 동일한 사람이 동일한 음성을 분할하더라도 추출된 음성경계에는 차이가 생기게 된다. 그 밖에도 지루한 작업이 계속됨에 따른 판단 오류도 문제가 될 수 있다.

이러한 문제들은 음성분할 및 레이블링 작업이 자동으로 수행될 수 있다면 어느 정도 해결될 수 있으며, 수작업에 의한 업무량과 비용을 크게 줄일 수 있을 것이다. 이에 따라 자동 음성분할 및 레이블링을 위한 여러 가지 방법들이 개발되어 왔으며[4]-[9], 영어에 대해서는 상품화된 시스템이 나오기도 했다[10]. 본 논문에서는 기존의 자동 음성분할 및 레이블링 기술들에 대한 검토를 토대로 한국어를 대상으로 한 자동 음성분할과 레이블링 시스템을 구현하였다. 그리고 음성분할 및 레이블링을 자동으로 수행하는데는 성능면에서 한계가 있으므로, 보다 신뢰도가 높은 결과를 얻기 위해서는 수작업에 의한 음성경계 오류를 수정하는 작업이 필수적으로 요청된다. 본 논문에서 개발한 시스템은 이러한 작업이 용이하도록 한글 모티프 환경에서 그래픽 사용자 인터페이스를 구현하였다.

본 논문의 구성은 다음과 같다. 2장에서 자동 음성분할 및 레이블링 기술에 대해서 설명하고, 3장에서는 음성분할 및 레이블링 시스템의 구성에 대해서 기술하고, 4장에서는 실제 구현된 시스템의 동작과 사용자 인터페이스에 관하여 설명한다. 그리고 5장에서는 시스템의 성능평가 실험과 그 결과에 대한 검토를 하고, 마지막으로 6장에서 결론을 맺는다.

Ⅱ. 자동 음성분할 및 레이블링 기술

지금까지 다양한 형태로 개발되어 온 음성분할 및 레이블링 기술은 음성분할 과정에서 언어학적 정보를 사용하지 않는 방식과 사용하는 방식의 두 부류로 크게 나눌 수 있다. 음소표기(phonetic transcription)와 같은 언어학적 정보가 주어지지 않은 상황에서 음성을 분할하는 방법들은 입력음성에 몇 개의 음소들이 포함되어 있는지도 모르는 상황에서 단지 음성신호에 포함된 음향학적 정보들 만으로 음성을 분할하게 되며, 일반적으로 음향학적 분할(acoustic segmentation)방법이라고 불리운다. 음향학적 분할방법으로는 음성신호의 스펙트럼 변화특성을 이용하는 방법[4], 스펙트로그램 상에서 다단(multi-level) 분할하는 방법[5], temporal decomposition 방법[6], 그리고 constrained clustering vector quantization 방법[4] 등이 알려져 있다. 이와 같은 음향학적 분할방법에 의해 음소경계 후보들이 선정되면 각각의 후보 음소부분들에 대해 패턴매칭 과정을 통해 레이블링을 하게 된다. 그러나, 현재의 기술 수준으로는 화자독립 음소인식 성능이 70% 정도밖에 되지 않음을 감안할 때 이 방식에 의한 자동 음성분할 및 레이블링의 성능은 언어학적 정보가 제공되는 방식에 비해 크게 뒤떨어질 수 밖에 없다.

이에 반해서, 언어학적 정보가 제공되는 상황에서의 자동 음성분할 및 레이블링은 입력음성에 포함된 음소열에 대한 대부분의 정보를 알고있는 경우이므로, 음성분할과 레이블링 과정이 단일 과정으로 수행된다. 즉, 훈련과정을 통해 각각의 음소들에 대한 대표 패턴들 또는 통계적 모델들을 미리 구성한 다음, 입력음성과 이에 해당하는 음소열에 대한 정보가 들어오면 구성가능한 음소열에 의해 각 음소들의 대표패턴들 또는 모델들을 연결시켜서 이들과 입력음성을 매칭시키는 과정에서 음소경계가 자동적으로 추출된다. 이러한 패턴 또는 모델 매칭 방법으로는 Dynamic Time Warping(DTW) 방법과 Hidden Markov Model(HMM) 방법이 사용될 수 있다. 그러나, DTW 방법에서는 복수 개의 대표패턴들만으로 음성신호에 내재된 변화요인들을 대처하는 데에 한계가 있기 때문에, 이러한 변화요인들을 확률모델 형태로 다루는 HMM 방법이 널리 사용되고 있다[7]-[9].

자동 음성분할 및 레이블링을 위해 제공되는 언어학적 정보는 음소표기(phonetic transcription)와 철자표기(orthographic transcription)로 나눌 수 있다. 음소표기가 제공되는 경우는 음소 레이블링이 이미 이루어진 상황이므로 음소사이의 묵음구간 검출과 더불어 각각의 음소들에 대한 경계위치만 찾아내면 된다. 이에 반해서 철자표기

가 제공되는 경우는 철자표기를 발음되는 형태의 음소표기로 자동변환해 주는 과정이 추가로 필요하게 되며, 동일한 단어라도 다양한 형태로 발음될 수 있기 때문에 경우에 따라서는 복수 개의 발음사전을 사용하게 된다[8].

그림 1은 언어학적 정보가 제공되는 경우의 HMM 방식에 의한 자동 음성분할 및 레이블링 시스템의 일반적인 구성도이다. 음성신호가 들어오면 음성특징 분석과정에서 음성특징계수들을 추출한다. 그 다음 미리 구성된 음소 HMM 모델들을 음소표기에 따라 연결하여 추출된 음성특징계수 시퀀스와 Viterbi 알고리즘에 의해 최적 time alignment를 수행하는 과정에서 각각의 음소간의 경계위치가 얻어진다. 이 때, 음소표기 정보는 사람의 수작업에 의해 직접 제공되거나, 철자표기 정보를 발음표기변환 프로그램에 의해 자동변환한 결과가 사용된다. 이 시스템의 출력은 각각의 음소 레이블과 그 경계위치들이며, 후처리 과정에서 수작업에 의한 수정작업이 이루어질 수 있다. 이와 같이 언어정보가 제공된 상황에서의 자동 음성분할 및 레이블링 작업은 미지의 음소 시퀀스를 식별해 내는 음성인식보다는 용이한 작업이다.

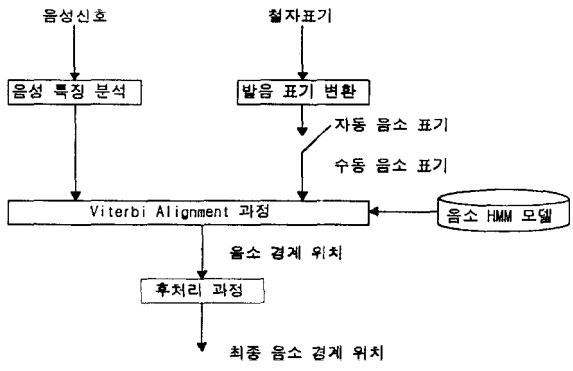


그림 1. HMM을 이용한 자동 음성분할 및 레이블링 시스템의 구성도

Fig. 1 Block diagram of the automatic segmentation and labeling system using HMM.

외국의 경우를 보면, 미국의 AT&T에서는 별도의 지속시간 모델을 포함한 문맥종속(context-dependent) 음소모델을 이용한 음성분할 및 레이블링 실험을 수행하였으며, TIMIT 음성 데이터베이스에 대한 실험결과에 따르면 자동분할에 의한 음소경계 위치 중 80% 정도가 음성학 전문가에 의한 음소경계 위치와 15ms 이내에 들어왔다고 보고되고 있다[7]. Texas Instruments에서는 철자 표기만을 이용한 음성분할 및 레이블링 실험을 보다 난이도가 높은 음성 데이터베이스인 SWITCHBOARD 데이터베이스를 대상으로 수행하였으나, 이 데이터베이스의 경우 수작업에 의한 음소분할 정보를 가지고 있지 않으므로 정확한 성능은 평가하지 못하였다[8]. 그리고 Entropic사는 Cambridge 대학의 연구 결과를 토대로 Aligner라는 영어 음성분할 시스템을 상품화하였는데, 이 제품은 TIMIT 음성 데이터베이스에 대한 실험 결과 자동분할 경계와 수동분할 경계와 차이가 16ms 이내에 들어오는 경우가 71%, 32ms 이내에 90%, 그리고 64ms 이내에 들어오는 경우가 97%인 것으로 발표하고 있다[10]. Brugnara 등도 TIMIT 데이터베이스를 이용한 음성분할 및 레이블링 실험을 수행하였다[8]. 이들은 음성분할 성능평가와 관련하여 훈련과정에 대해서는 수동 음소표기 및 수작업에 의해 음소분할이 이루어진 경우, 수동 음소표기만 되고 수작업에 의한 음소분할을 하지 않은 경우, 그리고 철자표기로부터 음소표기로 자동변환하고 수작업에 의한 음소분할도 하지 않은 경우의 세가지 부류로 나누고, 인식과정에서는 음소표기를 수동 및 자동으로 한 두 부류로 나누어 모두 여섯 가지 경우에 대해 성능평가 결과를 얻었으며 그 결과는 표 1에 나타나 있다. 표에서 보는 바와 같이 가능한 모든 작업에 전문가의 수작업을 동원한 경우가 모든 작업을 자동화한 경우에 비해 우수한 결과를 나타내고 있다. 이는 자동 음성분할 및 레이블링의 성능을 높이기 위해서는 수작업에 의한 분할 및 레이블링을 한 음성 데이터베이스가 충분히 제공될 필요가 있음을 시사한다.

표 1. TIMIT 데이터베이스에 대한 자동 음성분할 및 레이블링 성능 예[9]

Table 1. The Performance of the automatic segmentation and labeling for the TIMIT database[9].

훈 련		실 험	음소경계 검출성능 (%)		
레이블링	음소분할	레이블링	≤ 10ms	≤ 20ms	≤ 40ms
수동	수동	수동	74.6	88.7	97.0
수동	수동	자동	68.4	83.6	94.0
수동	자동	수동	55.4	77.6	92.3
수동	자동	자동	50.4	73.2	89.5
자동	자동	수동	50.9	73.4	88.1
자동	자동	자동	48.0	71.3	86.7

Ⅲ. 자동 음성분할 및 레이블링 시스템의 구성

본 장에서는 본 논문에서 구현한 한국어 자동 음성분할 및 레이블링 시스템의 구성과 관련한 기술적인 사항들을 정리하였다. 이 과정에서 영어에 대한 상용 음성분할 및 레이블링 시스템인 Entropic사의 Aligner 사양을 참고로 하였다[10].

3.1 음성신호 전처리 과정

입력음성의 샘플링 주파수는 Aligner 시스템의 경우, 8kHz와 16kHz의 두 가지를 지원하고 있으며, 그 밖의 다른 샘플링 주파수를 가지는 음성신호에 대해서는 주파수 변환 프로그램에 의해서 위의 두 가지로 변환해서 사용하도록 하고 있다. 본 논문에서 구현한 자동 음성분할 및 레이블링 시스템은 일차적으로 16kHz의 음성신호만을 대상으로 하였다.

일반적으로 음성인식에서는 매 10ms마다 20ms 구간의 음성신호로부터 음성특징을 추출하는 방식이 널리 사용되고 있다. 그러나 정교한 음성분할 및 레이블링 시스템에서는 음소경계 검출의 정밀도가 10ms 수준인 것은 바람직하지 않으며, 보다 세밀한 분석 시간단위를 필요로 한다. 참고로, TIMIT 음성 데이터베이스 구축에 사용된 자동 음성분할에서는 2.5ms의 시간단위가 사용된 것으로 알려져 있다[7]. 따라서, 시간단위가 5ms를 넘지 않도록 설정하는 것이 바람직하나, 이를 위해서는 충분히 정교한 음소 모델이 구성되어야 한다. 현재 훈련 대상으로 하고 있는 음성 데이터베이스로는 음소경계 오차범위를 5ms 이내로 하는 것이 어렵다는 판단하에서 본 연구에서는 음성분석 시간단위를 10ms로 정하였다.

음성인식을 위한 음성특징분석 파라미터는 음소간의 변별력이 뛰어나면서 음성학적으로는 중요하지 않은 변화요인들에는 둔감한 특징을 가질 것이 요구된다. 지금까지 음성인식에서 효과적으로 사용되어 온 음성특징분석 파라미터로는 LPCC(Linear Predictive Cepstral Coefficient)와 MFCC(Mel-Frequency Cepstral Coefficient), 그리고 이들의 시간축 미분값들이다[12]. 그 외에도 음성의 단구간 에너지 및 그 미분치도 중요한 정보로 활용될 수 있다. 본 논문에서는 12차의 LPCC, 그 미분치, 그리고 정

표 2. 유사음소 및 묵음 기호 목록
Table 2. List of the set of the PLUs and silence

번 호	기 호	설 명	번 호	기 호	설 명
1	B	ㅃ	24	j	ㅣ
2	D	ㄸ	25	ja	ㅑ
3	E	ㅐ + ㅔ	26	je	ㅒ + ㅖ
4	G	ㄲ	27	jo	ㅛ
5	N	ㅇ	28	ju	ㅠ
6	S	ㅆ	29	jv	ㅕ
7	SIL	묵음	30	k	ㅋ
8	U	ㅡ	31	l	[r]의 [l]되기
9	Wi	ㅢ	32	m	ㅁ
10	Z	ㅉ	33	n	ㄴ
11	a	ㅏ	34	o	ㅗ
12	b	ㅂ	35	p	ㅍ
13	b+	유성음화 ㅂ	36	r	ㄹ
14	b'	불파음화 ㅂ	37	s	ㅅ
15	c	ㅊ	38	t	ㅌ
16	d	ㄷ	39	u	ㅜ
17	d+	유성음화 ㄷ	40	v	ㅓ
18	d'	불파음화 ㄷ	41	wE	ㅚ + ㅙ + ㅞ
19	g	ㄱ	42	wa	ㅘ
20	g+	유성음화 ㄱ	43	wi	ㅟ
21	g'	불파음화 ㄱ	44	wv	ㅝ
22	h	ㅎ	45	z	ㅈ
23	h+	유성음화 ㅎ	46	z+	유성음화 ㅈ

규화된 단구간 에너지 및 그 미분치로 구성된 26차원 벡터를 음성특징 파라미터로 선정하였다.

3.2 레이블링 단위의 선정

음성을 분할하고 레이블링하는 기본단위로는 음소, 유사음소(PLU, phoneme-like unit), 변이음 등이 사용될 수 있다. 본 논문에서는 유사음소를 기본단위로 정하였다. 한국어 기본 음소단위에 유성음화, 불파음화, [r]의 [l]되기 등 3가지 규칙을 적용하였으며[13], 'ㅖ'와 'ㅒ', 'ㅔ'와 'ㅐ', 그리고 'ㅢ', 'ㅟ', 'ㅙ'를 각각 하나로 묶어서 45개의 유사음소 set을 정하였다. 이 유사음소 set에 묵음을 포함하여 총 46개를 음성분할 단위로 선정하였다. 본 논문에서 구현한 자동 음서분할 및 레이블링 시스템에 사용되는 유사음소 목록과 기호는 표 2에 나타나 있다.

3.3 음소 모델의 구성

HMM에 의해 음소모델을 구성하기 위해서는 다음과 같은 사항을 결정해야 한다. 먼저 HMM에서 관찰확률분포를 이산분포, 연속분포 또는 반연속분포 중에서 선정해야 하며, 상태 수와 천이방식 등 HMM topology의 길이, 즉 모델의 시작에서 끝으로 가는 동안 거쳐야 하는 최소 상태수는 매우 중요한 역할을 하며 음소의 최소 지속기간보다 작아야 한다. 그러나 적절한 음소 지속기간에 대한 모델없이 topology 길이가 너무 짧으면 인식할 때 삽입이 많아지고 음소분할에도 바람직하지 않다[9]. 음소 그룹에 따라 서로 다른 topology를 가지는 방법도 검토할 수 있다. 본 연구에서는 일차적으로 그림 2에서 보여지는 3개의 상태와 8개의 천이를 가지는 단순한 형태의 모델을 사용하였다. 관찰확률분포는 시작(B)과 중간(M), 그리고 끝(E)의 세 가지로 tying시켰으며, 각각의 확률분포는 연속확률분포를 사용하였다.

3.4 언어학적 정보 입력방식

이미 언급한 바와 같이 자동 음성분할 및 레이블링을 위해 제공되는 언어학적 정보로는 음소표기와 철자표기가 있다. 전문가에 의한 음소표기가 제공되는 경우 철자표기를 자동적으로 음소표기로 변환시키는 방법에 비해 성능면에서 우수하다[8]. 그러나, 음소표기를 수동으로 하는 작업은 음소 경계를 수동으로 하는 작업에는 비할 바

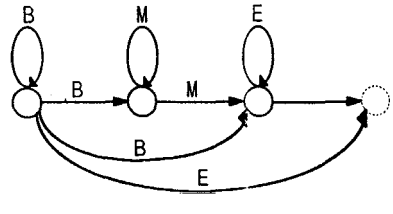


그림 2. 유사음소 모델링을 위한 HMM 구조
Fig. 2 HMM topology of the PLUs.

는 못되나 마찬가지로 많은 시간이 소요되는 일이다. 따라서 이들 두 가지 방식 중에서 어떤 것을 택할 것인지는 성능과 소요비용 사이의 trade-off에 의해 결정될 문제이다. 이와 더불어 철자표기를 발음되는 형태의 음소표기로 자동변환해주는 과정의 성능을 개선하기 위한 시도들도 계속되어야 할 것이며, 다양한 발음형태를 고려한 단어들의 발음사전 구성과 단어 경계에서의 상호 조음현상 처리 방안, 그리고 문장 내에 언제든지 삽입될 수 있는 묵음에 대한 처리 방안들이 그 대상이다.

Aligner의 경우 입력음성에 대한 언어정보는 일차적으로 철자표기 형태로 제공하도록 되어 있다. 따라서, 시스템 내부에 단어들의 발음사전을 가지고 있으며 많은 경우 복수의 발음형태를 포함시키고 있다. 그리고 발음사전에 없는 단어나 사전에 있는 발음과 상이한 발음을 가지는 단어를 위해서 사용자 정의의 발음사전을 추가시킬 수 있도록 되어 있다. 그 외에도 대화체 음성 등에서 빈번히 나타나는 형태인 발음되다가 중단된 단어나 잘못 발음된 단어, 그리고 발음상의 축약이 일어나는 경우 등에 대처하기 위해서 음소표기도 함께 사용할 수 있도록 하였다.

본 논문에서 구현하는 음성분할 및 레이블링 시스템에서도 음소표기를 입력받을 수 있을 뿐 아니라, 한국전자통신연구소의 발음표기변환 프로그램[13]을 이용하여 철자표기도 함께 사용할 수 있도록 하였다.

Ⅳ. 자동 음성분할 및 레이블링 시스템의 동작

음성분할 및 레이블링 작업을 자동으로 수행하는 데에는 성능면에서 한계가 있으므로, 보다 신뢰도가 높은 결과를 얻기 위해서는 음성학 전문가가 자동분할된 음소경계위치와 레이블링된 결과를 점검하여 오류를 수정하는 작업이 요구된다[14]. 따라서 본 장에서는 이러한 작업의 편의를 위해 제공되는 사용자 인터페이스 환경과 시스템의 전반적인 동작에 대해서 간략하게 설명한다.

4.1 동작 모드

개발된 시스템은 interactive 모드와 batch 모드 두 가지의 실행 모드를 가지고 있다. 첫 번째로 interactive 모드는 수 초에서 수십 초의 정도의 비교적 짧은 음성 구간을 대상으로 하며 사용자가 뒤에서 설명할 그래픽 사용자 인터페이스 환경에서 입력음성 파일을 열고 입력음성을 들어본 다음 직접 해당 언어정보를 입력하는 방식으로 수행된다. 그리고 자동 음성분할 및 레이블링 작업을 수행하고 나면 화면에 그 결과가 나타나는데, 필요할 경우 사용자가 직접 구간 청취나 스펙트로그램 등의 경계정보를 가지는 음성특징과 비교하면서 수정작업을 할 수 있다.

반면, batch 모드는 interactive 모드에서 대상으로 하는 음성의 길이보다 큰 대량의 음성신호에 대해서 작업을 수행하는 방식이다. 사용자가 미리 알고 있는 해당 언어

정보를 텍스트 파일 형태로 입력하며, 자동 음성분할 및 레이블링 된 결과 역시 파일의 형태로 출력된다. 이 경우에도 수정작업이 필요할 경우 위에서 설명한 interactive 모드에서 음성신호와 분할된 결과를 읽어서 교정 작업을 할 수 있다.

4.2 입력음성 파일 형식

개발된 시스템에서 사용되는 음성 데이터 파일은 기본적으로 헤더가 없는 바이너리 형태를 가지며 샘플링 주파수는 16kHz이고 한 샘플은 16 비트로 이루어진다. 다른 샘플링 주파수를 가지는 데이터에 대해서는 샘플링 주파수 변환 과정을 통해 16kHz로 바꾸어서 사용하도록 한다.

4.3 그래픽 사용자 인터페이스

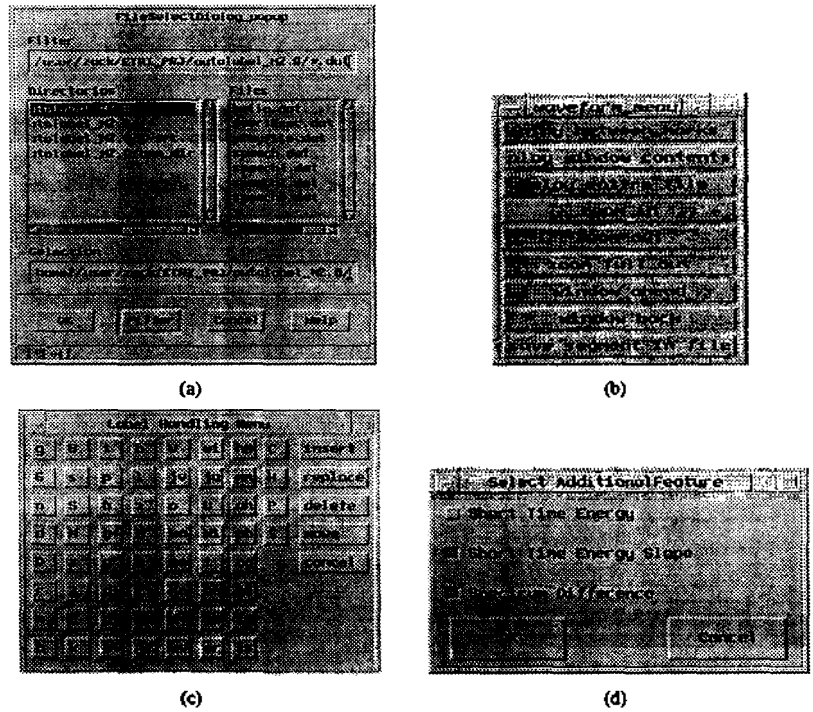
앞에서 언급한 바와 같이 자동 음성분할 및 레이블링 된 결과는 성능에 한계를 가지므로, 사용자에 의한 수정작업이 필요로 하게 된다. 개발된 시스템은 이러한 작업을 보다 편리한 환경에서 할 수 있도록 그래픽 사용자 환경을 제공한다. 전체 화면 구성은 크게 제어패널, 파형 윈도우, 레이블 윈도우, 스펙트로그램 윈도우, 그리고 부가적인 음성특징을 위한 윈도우로 구성된다. 본 논문에서 구현된 그래픽 사용자 인터페이스는 삼보 한글 모티프 2.0 하에서 이루어졌다[15][16].

4.3.1 제어패널

제어패널은 자동 음성분할 및 레이블링 시스템의 interactive 모드에서의 전반적인 기능을 제어하는 패널로서 그림 3에서와 같이 [Load], [AutoLabel], [WaveformHandlingMenu], [LabelHandlingMenu], [AdditionalFeature], [Continue] 및 [Exit]의 7개의 버튼으로 이루어져 있다.

해당 버튼에 대한 기능은 다음과 같다. 먼저 [Load] 버

그림 3. 제어패널
Fig. 3 Control panel



(a) (b) (c) (d)

그림 4. 자동 음성분할 및 레이블링 시스템에서 사용되는 메뉴들
(a)파일 선택 대화상자 (b)파형 메뉴
(c)레이블 메뉴 (d)부가적인 음성특징 선택 윈도우
Fig. 4 Menus for the automatic speech segmentation and labeling system
(a)File selection dialog box (b)Waveform menu
(c)Label menu (d)Additional feature selection window

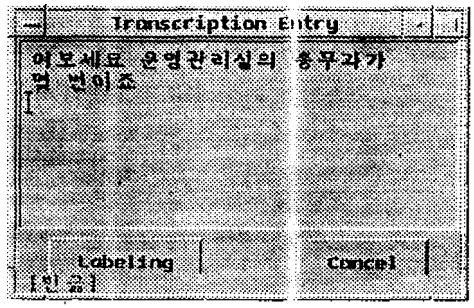


그림 5. 트랜스크립션 입력
Fig. 5 Transcription entry

턴은 음성분할 및 레이블링 작업을 위한 초기단계로 음성 파일을 불러오기 위한 것인데, 이 버튼을 누르게 되면 그림 4-(a)와 같은 음성 파일을 선택하기 위한 대화상자가 나타난다. [AutoLabel] 버튼은 음성 파일을 불러온 후, 그 음성에 대해서 자동분할 및 레이블링 작업을 수행하도록 한다. 버튼을 누르면 그림 5와 같은 음성의 언어정보를 입력하기 위한 트랜스크립션 입력이 화면에 나타난다. 이 에디터에 철자표기나 음소표기와 같은 언어정보를 입력하고 아래에 있는 [Labeling] 버튼을 누르면 자동 음성분할 및 레이블링 작업을 수행하게 된다. [WaveformHandlingMenu]을 누르면 파형 윈도우에서 사용하는 여러가지 기능을 실행하기 위한 메뉴가 화면에 나타나는데, 이 메뉴는 그림 4-(b)에 나타내었다. [LabelHandlingMenu]는 레이블 윈도우를 위한 메뉴를 화면에 띄우기 위한 버튼으로 사용자가 교정작업을 원활히 수행할 수 있도록 하기 위해서 음소테이블과 레이블 윈도우의 상태전환 버튼등으로 구성되어 있다(그림 4-(c)). [AdditionalFeature] 버튼은 음성 파형과 스펙트로그램 외에 사용자가 필요로 할 경우 다른 음소경계 정보를 가지는 음성특징을 선택하기 위한 버튼으로, 이 버튼을 누르면 그림 4-(d)에서 보는 바와 같이 부가적인 음성특징 선택 윈도우가 나타난다. [Continue] 버튼은 커맨드 라인 상에서 자동 음성분할 및 레이블링 시스템을 구동시킬 때 하나 이상의 음성 파일을 지정한 경우 순차적으로 분할 및 레이블링 작업을 할 수 있는데 이 경우 순차적으로 지정된 음성 중에서 하나의 작업을 끝내고 다음 음성으로 넘어갈 때 사용된다. 마지막으로 [Exit] 버튼은 모든 프로그램을 종료하고 빠져나갈 때 사용된다. [Continue]와 [Exit] 버튼을 누르면 수정된 음소경계위치는 자동적으로 저장된다.

4.3.2 파형 윈도우

파형 윈도우는 시간축 상에서 음성신호 자체의 진폭변화를 화면에 나타낸다(그림 6-(a)). 이 윈도우의 제일 상단에는 현재 작업 중인 음성의 파일명을, 그리고 왼쪽 상단에는 현재 포인터가 가르키는 위치의 시간과 그때의 음성신호진폭을 보여 주고 있다. 이 파형 윈도우에서 입력음성의 출력 기능, zoom 기능, 선택된 블럭을 새로운 파일로 저장하는 기능 등은 파형 메뉴를 이용해서 실행할 수 있다.

이후에 설명할 음성특징 윈도우와 레이블 윈도우는 파형 윈도우의 시간축상의 변화에 따라 자동적으로 정렬된다. 즉, 모든 윈도우는 항상 파형 윈도우와 동일한 시간축을 가지며, 또한 각 윈도우 내에서 마우스 포인터의 위치는 모든 윈도우에서 동일한 시간을 가리킨다.

4.3.3 레이블 윈도우

레이블 윈도우는 두 부분으로 나뉘는데 상단은 음소단위로 분할 및 레이블링된 결과를 나타내고, 하단은 어절단위로 분할 및 레이블링된 결과를 나타낸다. 왼쪽 상단에는 현재 레이블 윈도우의 상태를 표시하고 있다(그림 6-(b)). 자동 음성분할 및 레이블링 작업이 끝난 후 수작업에 의한 교정작업은 주로 이 레이블 윈도우에서 이루어진다. 이러한 작업은 레이블 메뉴를 적절히 이용함으로써 수행되는데, 레이블 메뉴는 음소 테이블과 모드 선택 버튼으로 구성된다. 레이블 윈도우에서 모드는 새로운 음소나 어절의 삽입을 위한 insert, 선택된 음소나 어절을 다른 것으로 바꾸는 replace, 레이블 윈도우에 있는 특정 음소나 어절을 삭제하는 delete, 음소나 어절의 경계위치를 옮기기 위한 move 모드 등이 있다.

4.3.4 스펙트로그램 윈도우

스펙트로그램 윈도우는 입력 음성신호의 광역(wideband) 스펙트로그램을 화면에 나타낸다. 입력 음성에 매 2ms마다 8ms의 해밍 윈도우를 씌운 각 프레임을 1024 포인트 FFT를 한 후 각 포인터에 대해서 크기를 계산하였다. 그리고 그 값들의 최대값과 최소값 사이를 256개의 단계로 나누어서 gray-scale로 할당하였다(그림 6-(c)).

4.3.5 부가적인 음성특징 윈도우

구현된 시스템은 기존적인 음성특징인 음성의 파형과 스펙트로그램이외의 다른 세 가지의 음성 특징을 부가적인 음성특징으로 제공하는데, 각각은 단구간 에너지, 단구가 에너지의 기울기, 프레임간의 스펙트럼 차이이다. 단구간 에너지의 경우 20ms의 해밍 윈도우를 씌워서 2ms씩 옮겨가면서 각 프레임에 대한 에너지를 계산한다. 단구간 에너지의 기울기는 단구간 에너지로부터 기준 프레임에서 앞뒤로 각각 3 프레임씩 떨어진 프레임간의 에너지 차이를 구한다. 마지막으로 스펙트럼 차이는 역시 20ms의 해밍 윈도우를 씌워서 2ms씩 옮겨가며 12차의 LPCC를 구한 다음, 기준 프레임으로부터 앞뒤로 2 프레임 사이의 켑스트럼 계수들 간의 거리를 구해서 사용한다(그림 6-(d)).

4.4 음성분할 및 레이블링 작업 예

본 절에서는 개발된 시스템을 이용한 음성분할 및 레이블링 작업에 전형적인 예를 보여준다. 먼저 'speech.

dat'라는 16kHz로 샘플링된 음성 데이터 파일이 준비되어 있다고 가정하면, 다음과 같은 작업순서로 진행된다.

(1 단계) 커맨드 라인에서 음성분할 및 레이블링 시스템 구동한다.

(2 단계) 제어패널에서 [load] 버튼을 누른 다음, 파일선택 대화상자에서 분할할 음성 파일을 선택함으로써 데이터를 불러온다.

(3 단계) 파형 윈도우의 출력기능을 이용하여 해당 음성의 언어정보를 얻은 다음, 제어패널에서 [Auto-Label] 버튼을 누르면 생기는 트랜스크립션 입력에 언어정보를 입력하고 [Label] 버튼을 누르면 자동 음성분할 및 레이블링 수행된다.

(4 단계) 기본적인 음성특징이외의 부가적인 음성특징이 필요한 경우, 제어패널에서 [Additional-Feature] 버튼을 눌러서 원하는 부가적인 음성특징을 화면에 디스플레이한다.

(5 단계) 자동 음성분할 및 레이블링 된 결과를 파형, 스펙트로그램, 부가적인 음성특징 등에 나타난 음소경계위치 정보와 청취평가를 통해서 수작업에 의한 교정 작업을 수행한다.

(6 단계) 수정 작업을 끝내고 제어패널에서 [Exit] 버튼을 누르면, 수정된 결과를 저장하고 마친다.

이러한 작업순서에 따라서 자동 음성분할 및 레이블링 과정을 거친 후의 전체 화면을 그림 6에 나타내었다.

4.5 자동 음성분할 및 레이블링 결과

자동 음성분할 기능의 실행 결과는 음소단위 분할결과와 어절단위 분할결과 파일을 생성한다. 음소단위 분할결과 파일의 포맷은 [시간] [음소기호]이고, 어절단위 분

(a)	(b)
0.370000 SIL	0.370000 SIL
0.540000 jv	1.050000 여보세요
0.590000 b+	1.390000 SIL
0.630000 o	2.260000 운영관리실의
(중략)	2.830000 홍무과가
3.410000 z+	3.060000 멋
3.570000 o	3.570000 빈이죠
4.100000 SIL	4.100000 SIL

그림 7. 자동 음성분할 및 레이블링 시스템의 음성분할 결과 예
(a)음소단위 분할 결과 (b)어절단위 분할 결과

Fig. 7 Example of the results of the automatic segmentation and labeling process
(a)Phoneme-level segmentation result
(b)Word-level segmentation result

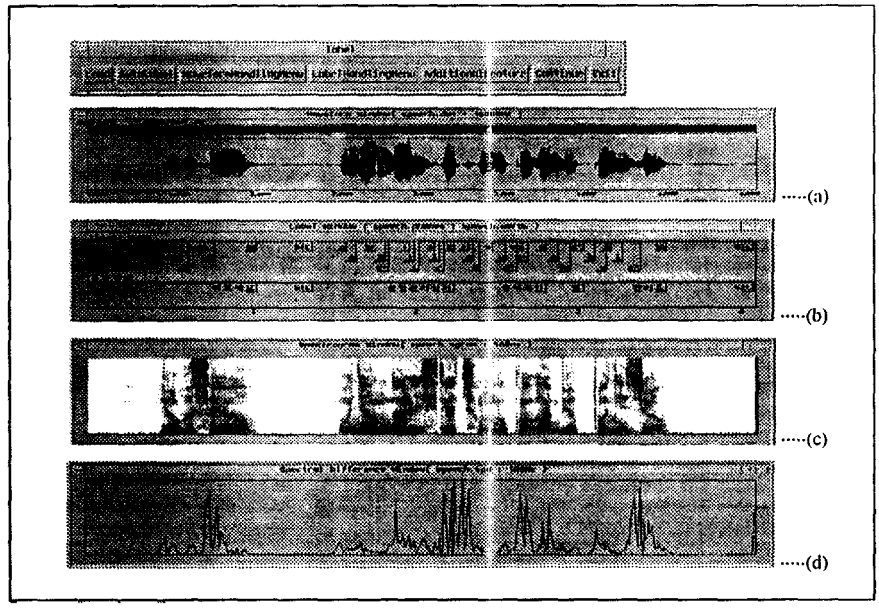


그림 6. 자동 음성분할 및 레이블링 작업을 거친 후의 전체 화면 예
(a)파형 윈도우 (b)레이블 윈도우
(c)스펙트로그램 윈도우 (d)부가적인 음성특징 윈도우

Fig. 6 Example of the screen after automatic speech segmentation and labeling processing
(a)Waveform window (b)Label window
(c)Spectrogram window (d)Additional feature window

한결과는 [시간] [어절]이다. 이때, [시간]은 초 단위이며 해당 [음소기호] 및 [어절]이 끝나는 시점을 나타내는 시간이다. 그림 7은 자동 음성분할 및 레이블링 시스템의 음성분할 결과의 한 예를 보여주고 있다. "여보세요, 운영관리실의 총무과가 몇 번이죠?"라는 입력음성에 대한 음성 데이터와 언어정보가 사용되었다.

V. 실험 결과 및 검토

본 논문에서 개발된 시스템의 성능평가를 위한 실험을 수행하였다. 각각의 유사음소의 훈련을 위해서 한국전자통신연구소의 음소적으로 균형이 잡힌(phonetically balanced) 445 단어 데이터베이스에서 남성화자 20명의 음성 데이터를 사용하였다[17]. 그리고 성능평가를 위해 훈련에 사용된 445 단어 데이터베이스와 훈련에 사용되지 않은 한국전자통신연구소의 부서명 문장 데이터베이스[17]에 대해서 각각 음성분할 및 레이블링 실험을 수행하였다. 훈련을 위해 사용된 445 데이터베이스와 성능평가를 위해 추출된 부서명 데이터베이스는 수작업에 의해서 음소단위로 분할되었다.

자동 음성분할 및 레이블링 실험결과는 표 3에 나타내었다. 표에 나타난 성능평가 결과는 수작업에 의해서 분할된 경계위치와 자동 음성분할 및 레이블링 시스템의 의해서 자동 분할된 경계위치 간의 차이가 주어진 오차범위에 해당하는 갯수와 전체 문장에 포함된 유사음소의 총 갯수의 비를 백분율로 표시한 것이다. 여기서 오차범위는 10ms이내, 20ms이내, 30ms이내, 40ms이내, 그리고 50ms이상이 사용되었다.

표 3에서 closed test는 훈련용 음성 데이터베이스의 일부로 음성분할 및 레이블링 실험을 수행한 결과이며, 훈련에 사용된 445 데이터베이스 중에서 4명의 음성 데이터를 추출하여 실험한 것이다. 사용된 총 유사음소의 개수는 11475개이고, 실험 결과는 표 3의 closed test 부분에 나타내었다. 또한, 표 3에서 open test는 부서명 문장 데이터베이스에서 남성 화자 44명으로부터 각 1문장씩을 선택적으로 추출하여 사용하였다. 실험에 사용된 총 유사음소의 개수는 1503개이고, 훈련용 음성 데이터베이스와 실험용 음성 데이터베이스의 내용 및 화자가 다른 경우의 결과이다. 이 표로부터 closed test와 open test의 성능에 별다른 차이가 없음을 알 수 있다.

표 3. 자동 음성분할 및 레이블링 시스템의 성능평가 결과
Table 3. The Performance evaluation of the automatic segmentation and labeling system

경계오차 범위	음소경계 검출 성능	
	closed set(445 DB)	open set(부서명 DB)
≤ 10ms	57.4 % (6599)	50.4 % (758)
≤ 20ms	77.9 % (8953)	74.7 % (1122)
≤ 30ms	86.5 % (9940)	87.2 % (1310)
≤ 40ms	91.3 % (10488)	92.8 % (1394)
≤ 50ms	94.1 % (10806)	96.0 % (1443)

VI. 결 론

본 논문에서는 한국어 음성 데이터베이스의 효율적인 구축을 위한 자동 음성분할 및 레이블링 시스템을 구현하였다. 이를 위하여 기존의 음성분할 및 레이블링 방식 분석을 토대로 한국어 자동 음성분할 및 레이블링 시스템의 규격을 정하고, 또한 사용자가 자동 분할된 결과를 수작업에 의해서 경계위치의 수정이 용이하도록 한글 모티프 환경에서 그래픽 사용자 인터페이스를 개발하였다. 개발된 시스템은 16kHz의 샘플링 주파수를 가지는 음성을 대상으로 하였으며, 레이블링 단위는 45개의 유사음소와 하나의 묵음으로 정하였으며, 언어정보의 입력방식으로는 음소표기와 철자표기를 둘 다 사용할 수 있도록 하였다. 그리고 패턴매칭 방법으로는 HMM을 사용하였다.

시스템의 성능평가를 위해 한국전자통신연구소의 445 단어 데이터베이스로 훈련과정을 수행한 다음, 한국전자통신연구소의 부서명 문장 데이터베이스에 적용하였다. 그 결과, 20ms이내에 74.65%가 포함되었고 40ms이내에는 92.7%가 포함되었다. 이 결과는 충분한 훈련용 데이터베이스가 확보되어 있는 외국의 연구결과에 비해서는 떨어지는 성능으로[9], 이를 개선하기 위해서는 보다 정교한 음소 모델링을 위한 데이터베이스가 필수적으로 요구된다.

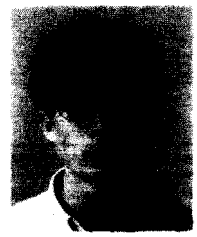
본 논문에서 구현한 자동 음성분할 및 레이블링 시스템은 일관성있는 판단기준에 의해 음성신호를 유사음소 단위로 자동 분할해 줌으로써, 음성학 전문가에 의한 수동 분할의 업무량을 현저하게 절감시킬 수 있다. 따라서, 본 시스템은 방대한 양의 한국어 음성 데이터베이스를 구축하는 데 기여할 수 있을 것으로 기대되며, 그 결과로 한국어 음성인식 시스템의 인식 성능을 향상시키는데 유용한 도구로 활용될 수 있을 것이다.

참 고 문 헌

1. D. B. Roe and J. G. Wilpon, *Voice Communication between Humans and Machines*, National Academy Press, 1994.
2. H. C. Leung and V. Zue, "A procedure for automatic alignment of phonetic transcription with continuous speech," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.429-432, Apr. 1984.
3. B. Eisen, H. Tillmann and C. Draxler, "Consistency of judgements in manual labeling of phonetic segments: the distribution between clear and unclear cases," in Proc. Int. Conf. Spoken Language Processing, pp.871-874, Oct. 1992.
4. T. Svendsen and F. K. Soong, "On the automatic segmentation of speech signals," in Proc. IEEE Int. Conf. Acoust.,

Speech, Signal Processing, pp.77-80, Apr. 1987.
5. J. R. Glass and V. W. Zue, "Multilevel acoustic segmentation of continuous speech," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.429-432, Apr. 1988.
6. F. Bimbot, G. Chollet, P. Deleclise and C. Montacie, "Temporal decomposition and acoustic-phonetic decoding of speech," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.445-448, Apr. 1988.
7. A. Ljolie and M. D. Riley, "Automatic segmentation and labeling of speech," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.473-476, Apr. 1991.
8. B. Wheatley, G. Doddington, C. Hemphill and J. Godfrey, "Robust automatic time alignment of orthographic transcription with unconstrained speech," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.1-553-556, Apr. 1992.
9. F. Brugnara, D. Falavigna and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," Speech Communication, vol.12, no.4, pp. 357-370, Aug. 1993.
10. *The Aligner: A System for Automatic Time Alignment of English Text and Speech*, Entropic Research Laboratory, Inc., 1994.
11. K. Silverman et al., "TOBI: a standard for labeling English prosody," in Proc. Int. Conf. Spoken Language Processing, pp.867-870, Oct. 1992.
12. L. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
13. 유재원, 연속음성인식을 위한 음성 단위 발음사전 구성방법 연구, 위탁과제 최종 연구보고서, 한국전자통신연구소, 1995년 1월.
14. S. Fujiwara, Y. Komori and M. Sugiyama, "A phoneme labeling workbench using HMM and spectrogram reading knowledge," in Proc. Int. Conf. Spoken Language Processing, pp.791-794, Oct. 1992.
15. *The X Window System & OSF/MOTIF, TGxwindow X11R5/Motif 2.0* 한글 환경 설명서, 삼보바이크로 시스템.
16. D. A. Young, *The X Window System™ Programming and Applications with Xt*, 2nd Edition, Prentice-Hall, 1994.
17. 이영직, 류준영, 김상훈, 황규웅, "ETRI의 음성 데이터베이스 구축 현황," 제12회 음성 통신 및 신호처리 워크샵 논문집, pp.256-267, 1995년 6월.

▲성 종 모(Jongmo Sung) 1971년 12월 26일생



1995년 2월 : 부산대학교 전자공학과 학사
1997년 2월 : 부산대학교 전자공학과 공학석사
1997년~현재 : LG전자 평택연구소 연구원
※주관심분야 : 음성인식, 음성합성, 음성/영상 부호화, 시리얼 통신, 인터넷 보안

▲김 형 순(Hyung Soon Kim)

1983년 2월 : 서울대학교 전자공학과 (공학사)
1984년 2월 : 한국과학기술원 전기 및 전자공학과(박사과정 조기진학)
1989년 2월 : 한국과학기술원 전기 및 전자공학과(공학박사)
1987년 1월~1992년 6월 : 디지콤 정보통신연구소 선임연구원, 연구부장
1992년 7월~현재 : 부산대학교 전자공학과 조교수 부산대학교 정보통신연구소 연구원