한국음향학회지 제25권 제5호 pp. 213~221 (2006)

HMM기반 자동음소분할기의 음소분할 오류 유형 분석

The Error Pattern Analysis of the HMM-Based Automatic Phoneme Segmentation

김 민 제*, 이 정 철*, 김 종 진**
(Min-Je Kim*, Jung-Chul Lee*, Jong-Jin Kim**)

*울산대학교 컴퓨터 정보통신 공학부, **한국전자통신연구원
(접수일자: 2006년 4월 7일; 수정일자: 2006년 6월7일; 채택일자: 2006년 7월 20일)

합성음의 음질을 향상시키기 위하여 분할된 corpora로부터 합성유닛을 선택하여 사용하는 연속음성합성에서 정확한 음소분할은 매우 중요하다. 일반적으로 음소분할은 사람에 의해 수행 되지만 많은 작업량으로 인한 시간적 지연, 일관성 유지 어려움 등 많은 문제가 발생한다. 이에 따라 음성인식에서 도입된 HMM기반의 자동음소분할이 음성인식, 음성합성에서 널리 사용되어지고 있지만 음성전문가의 수작업 결과와 비교할 때 HMM기반 자동음소분할은 오류가 있고, 이는 합성음 품질의 열화의 주요 원인이 되고 있다.

본 논문에서는 HMM기반의 자동음소분할기를 사용하여 나타난 자동음소분할 결과와 수작업에 의한 음소분할 결과를 비교하고 유형별로 분석함으로써 음성합성의 성능향상을 위해 개선해야 할 문제점들을 제시한다. 실험에서는 ETRI의 표준형 한국어 공통 음성 DB을 사용하였고, 오차의 범위가 20ms를 벗어난 경우를 분절 오류로 간주하였다.

실험 결과 여성화자의 경우 파열음+모음, 파찰음+모음, 모음+유음 음소쌍에서는 각각 약 99%, 99.5%, 99%의 높은 정확률을 보인 반면, 폐쇄음+비음, 폐쇄음+유음, 비음+유음 음소쌍에서는 44.89%, 50%, 55%의 낮은 정확률을 보였으며, 남성화자에 대한 실험결과에서도 유사한 경향을 보였다.

핵심용어: HMM, 자동음소분할, 음성데이터베이스

투고분야: 음성처리분야 (2.4)

Phone segmentation of speech waveform is especially important for concatenative text to speech synthesis which uses segmented corpora for the construction of synthetic units, because the quality of synthesized speech depends critically on the accuracy of the segmentation. In the beginning, the phone segmentation was manually performed, but it brings the huge effort and the large time delay. HMM-based approaches adopted from automatic speech recognition are most widely used for automatic segmentation in speech synthesis, providing a consistent and accurate phone labeling scheme. Even the HMM-based approach has been successful, it may locate a phone boundary at a different position than expected.

In this paper, we categorized adjacent phoneme pairs and analyzed the mismatches between hand-labeled transcriptions and HMM-based labels. Then we described the dominant error patterns that must be improved for the speech synthesis. For the experiment, hand labeled standard Korean speech DB from ETRI was used as a reference DB. Time difference larger than 20ms between hand-labeled phoneme boundary and auto-aligned boundary is treated as an automatic segmentation error. Our experimental results from female speaker revealed that plosive-vowel, affricate-vowel and vowel-liquid pairs showed high accuracies, 99%, 99.5% and 99% respectively. But stop-nasal, stop-liquid and nasal-liquid pairs showed very low accuracies, 45%, 50% and 55%. And these from male speaker revealed similar tendency.

Key words: HMM, Automatic segmentation, Speech database

ASK subject classification: Speech signal processing (2.4)

I. 서 론

최근 고품질 음성 합성 기술은 대용량 음성 데이터로부터 분절된 음편을 접합하는 TTS 방식이 주류를 이루고 있다 [1]. 이와 같은 합성 방식에서는 음성 데이터의

책임저자: 김 민 제 (ink1978@mail.ulsan.ac.kr)
680-749 울산광역시 남구 무거2동 울산대학교 컴퓨터, 정보통신공학
(전화: 052-259-1435)

용량뿐만 아니라 음성 데이터의 분절표기 정확도가 합성음의 음질을 결정짓는 주요 요인이 되고 있다 [2].

또한 임의의 어휘를 대상으로 하는 연속음성인식은 발성자에 따른 개인차는 물론이고 전후에 발성되는 음소의 영향에 의한 조음 결합에 따라 그 특성이 크게 변화한다. 이러한 개인차 및 조음결합의 현상을 분석하기 위해서는 많은 사람이 발성한 다양한 음성데이터를 수집한 후, 음성 데이터를 음소와 같은 음성의 기본단위로 분할하고 레이블링하여 그 다음 단계에서 통계적 처리가 가능하도록 가공하는 작업이 필수적으로 요구된다 [3][4].

하지만 음성 데이터베이스를 구축하는 과정 중, 음소 분할 및 레이블링 작업은 일반적으로 수작업에 의해서 행해지며, 여기에는 여러 가지 문제점이 있다 [5]. 이러한 문제점들을 해결하기 위해 자동음소분할에 관한 연구가 국내외적으로 활발히 진행되고 있다.

자동음소분할의 방법중 모델 매칭방법인 HMM방식은 음소 분할에서 높은 일관성과 정확성을 확보하고 있어 음성 인식분야에서 널리 사용되고 있다 [6]. 그러나 HMM기반의 자동 음소분할에서 사용되는 비터 (Viterbi) 정렬은 인접 음소간의 최적의 경계점을 찾는 것이 아니라 음소열과 HMM파라미터 열이 주어졌을 때 최적의 HMM열을 찾는 것을 목표로 하기 때문에, 자동 음소분할 결과는 음성전문가가 수작업으로 분할한 결과와 괴리가 발생하는 것을 피할 수 없다 [7]. 따라서 정밀도가 떨어지는 부분의 자동 탐지 및 이를 정제할 수 있는 후처리 기술이 필요하고 어떤 후처리 기술을 적용할 것인가를 판단하기 위해서는 오류가 발생하는 부분에 대한 분석이 요구된다.

기존의 자동음소분할 성능 향상을 위한 몇 가지 방법을 살펴보면 첫 번째는 음소를 여러 클래스로 나누고 각 클래스별로 각각 다른 길이의 시간 창 (time window)을 적용함으로써 경계점의 오류를 최소화 하였지만, 시간 창의 길이를 넘어서는 큰 오류에 대해서는 보정할 수 없는 단점이 있었다. 두 번째는 오류가 빈번히 나타나는 묵음+음소와 음소+묵음의 경계를 보정하기 위해 에너지 기반의 후처리를 사용하였으며, 세번째는 기존의 fuzzy -logic based system의 단점을 보완하기 위해 단일 신경회로망을 사용하여 전체적인 부분에서의 성능향상을 보였고, 이를 바탕으로 각각의 클래스별로 특화된 여러 개의 신경회로망을 사용하는 방법도 제안되어 성능의 향상을 가져왔다. 하지만 화자가 다른 경우에는 잘 적용되지 않았다 [7][8][9].

본 논문은 이러한 연구의 일환으로, HMM 기반의 자동음소분할기를 사용하여 나타난 결과를 정제할 수 있는 후처리기 개발을 위한 전단계로서, 자동음소 분할 결과와 수작업에 의한 음소 분할 결과를 유형별로 비교하여 그에 따른 오류들을 정리하였다.

본 논문의 구성은 다음과 같다. 2절에서는 자동음소분할 시스템의 구성 및 음성 데이터베이스에 대한 실험 결과를 살펴보고, 3절에서는 유형별 오류의 분포를 살펴본다. 그리고 4절에서는 결론 및 향후 연구계획에 대해서 언급한다.

II. 자동음소분할 시스템의 구성

자동음소분할은 소량의 수작업된 데이터 (bootstrap data)를 이용하여 초기화하는 방법과 초기 모든 HMM을 동일하게 구성하는 방법이 있다. 그렇지만 초기 모든

표 1. 자동음소분할 시스템의 구성
Table 1. The composition of the automatic phoneme segmantation system.

음소모델	모노폰(monophone)
HMM형태	3상태(start/end 상태제외) 7천이, left-right 형태
특징 파라미터	12차 MFCC+로그에너지+1차미분계수 +2차미분계수 = 39차
mixture갯수	3개
음성 분석 구간	15ms 구간을 10ms단위로 이동

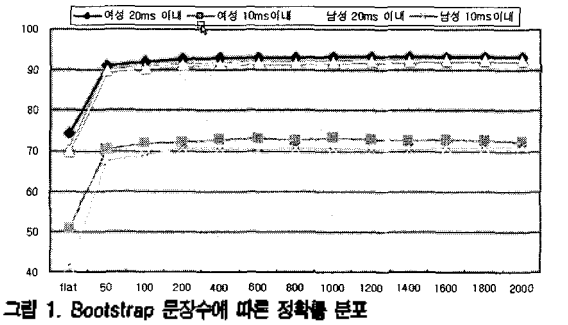

그림 1. Bootstrap 문장수에 따른 정확률 분포
Figure 1. The distribution of the accuracy for the number of Bootstrap sentence.

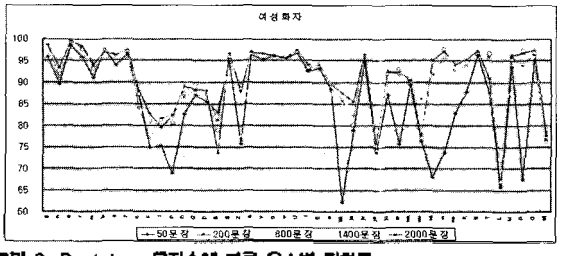

그림 2. Bootstrap 문장수에 따른 음소별 정확률
Figure 2. The accuracy of the phoneme for the number of bootstrap sentence.

HMM을 동일하게 구성할 경우 자동음소분할의 성능은 bootstrap 방법에 비하여 20%이상 성능이 저하 된다. 따라서 본 논문에서는 소량의 bootstrap data를 사용하여 자동음소분할을 수행하였다.

표 1은 자동음소분할 시스템의 파라미터 구성을 나타내며, 실험을 통하여 최적의 파라미터를 선정하였다.

HMM훈련 과정에서 2000문장 중 200문장 bootstrap data로 사용한 이유는 그림1에서 볼 수 있는 것과 같이 200문장 이상을 HMM훈련에 사용하더라도 정확률의 향상이 눈에 띄게 나타나지 않았으며, 그림2에서 볼 수 있는 것처럼 음소별 정확률에 대해서도 큰 향상을 보이지 않았기 때문이다.

표 2. 음성 분할 단위 및 기호
Table 2. The unit and symbol of phoneme segmentation.

	기호	설명		기호	설명		기호	설명
초성	g	ㄱ	중성	a	ㅏ	종성	K	ㄱ
	n	ㄴ		v	ㅓ		N	ㄴ
	d	ㄷ		o	ㅗ		T	ㄷ
	r	ㄹ		u	ㅜ		L	ㄹ
	m	ㅁ		U	ㅡ		M	ㅁ
	b	ㅂ		i	ㅣ		P	ㅂ
	s	ㅅ		E	ㅐ		O	ㅇ
	z	ㅈ		e	ㅔ		sil	묵음
	c	ㅊ		Wi	ㅢ			
	k	ㅋ		ja	ㅑ			
	t	ㅌ		jv	ㅕ			
	p	ㅍ		jo	ㅛ			
	h	ㅎ		ju	ㅠ			
	G	ㄲ		je	ㅖ			
	D	ㄸ		wa	ㅘ			
	B	ㅃ		wv	ㅝ			
	S	ㅆ		wi	ㅟ			
	Z	ㅉ		wE	ㅙ			
				we	ㅞ			

2.1. 음소 분할 단위 선정

본 연구에서 사용한 음소셋은 한국어 초성 18개, 중성 19개, 종성 7개와 묵음을 포함한 45개의 음소로 구성되었다. 한국어 모음 중에서 'ㅐ'와 'ㅔ', 'ㅙ'와 'ㅞ'는 발음이 유사하여 동일한 그룹으로 분류할 수 있지만 [10] 실험에서는 ETRI DB의 발음 전사 기준을 사용하였다. 표 2에 본 논문에서 사용한 음소 목록과 각각의 기호를 나타내었다.

III. 유형별 오류 분포

2절에서 구성된 시스템을 사용해서 자동음소분할을 수행하였다. 실험에 사용된 음성 데이터는 한국전자통신연구원에서 배포하는 음성합성용 낭독체 문장 데이터베이스 (남성 1인 2000문장, 여성 1인 2000문장, 16kHz 샘플링, 16bit 양자화)를 사용하였다. HMM 훈련에는 200문장의 음소분할 및 레이블링 정보를 이용하였고, 훈련된 HMM을 이용한 2000문장의 자동음소분할 결과를 음성전문가에 의해 수작업된 결과와 비교할 때, 오차의 범위가 20ms를 벗어난 경우를 분절 오류로 간주하였다. 자동음소분할의 평균 정확률은 남성화자, 여성화자 각각 91.22%, 92.73%를 보였다.

자동음소분할의 오류 분석은 크게 모음+모음, 자음+모음, 모음+자음, 종성+초성의 4가지 유형으로 수행하였으며, 초성과 종성은 조음특성에 따라 파열음, 폐쇄음, 마찰음, 파찰음, 비음, 유음의 형태로 구분하였다. 모음은 단모음과 이중모음을 모두 포함하며, 자음의 경우에는 종성이라는 언급이 없으면 초성을 의미한다.

3.1. 모음+모음

모음+모음의 형태는 평균적으로 약 71.5%의 분할 성공률을 가지고 있다. 표 3에서 여성화자에 대한 모음+모음 음소쌍의 정확률, 출현빈도, 오류빈도를 나타내었다.

유형들에 대해 전체적인 정확률 (오차범위가 20ms이내인 개수/유형별 출현 빈도)을 살펴보면, 가장 높은 빈도를 보이는 o+i의 형태를 포함해서 15가지의 음소쌍에서 95% 이상의 분할 성공률을 보이고 있다. 이중 11가지의 형태에서는 100%의 분할 성공률을 보이는 반면에, 나머지의 경우에서는 높게는 92%에서 낮은 경우는 20% 미만에 이르고 있다. 표 3을 바탕으로 몇 가지 공통적인 상황이 나타난다.

표 3. 모음+모음 형태의 빈도수별 정리표 (female)
Table 3. The frequency of the Vowel-Vowel class (female).

음소쌍	출현빈도	오류	정확률	음소쌍	출현빈도	오류	정확률
o+i	261	4	98.47	v+i	53	4	92.45
e+i	204	83	59.31	E+i	52	20	61.54
a+i	193	14	92.75	a+o	50	30	40.00
i+e	160	41	74.38	o+a	44	19	56.82
i+a	146	18	87.67	i+jo	43	15	65.12
i+v	121	11	90.91	E+e	42	32	23.81
j+i	114	22	80.70	j+o	40	13	67.50
u+i	75	6	92.00	we+e	39	25	35.90
U+e	71	34	52.11	a+v	39	30	23.08
a+e	70	29	58.57	a+wa	39	12	69.23
i+jv	60	6	90.00	e+jo	37	24	35.14
o+e	59	8	86.44	j+wi	37	2	94.59
v+u	57	25	56.14	E+ja	36	11	69.44
v+jo	55	38	30.91	a+U	36	7	80.56
u+e	54	11	79.63	a+a	35	18	48.57
a+u	54	14	74.07	U+i	35	4	88.57
a+jo	53	28	47.17	wi+e	35	12	65.71

첫 번째는 같은 모음이 연속해서 나타나는 상황에서는 분할 성공률이 낮다는 것이다. i+i와 같이 같은 모음이 연속적으로 나타나는 경우는 정확률이 67%정도이다. 그림3은 i+i에서의 오류의 예이다.

두 번째는 we+e와 같이 이중 모음 후에 비슷한 모음이 나타나는 경우로서 32%의 매우 낮은 정확률을 보인다. 그림4는 이러한 유형 중에서 한 가지를 나타낸 것이다.

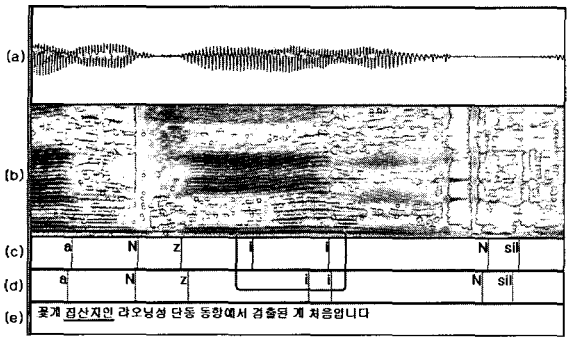

그림 3. i+i 음소쌍의 오류 (오차 약75ms)
(a)음성파형, (b)스펙트로그램, (c)수작업에 의한 음소분할, (d)자동음소분할 결과, (e)철자전사

Figure 3. The error pattern of the pair of i-i (The segmentation error is about 75ms). (a)waveform (b)spectrogram (c)hand-labeled segmentation (d)automatic labeling and segmentation (e)transcription.

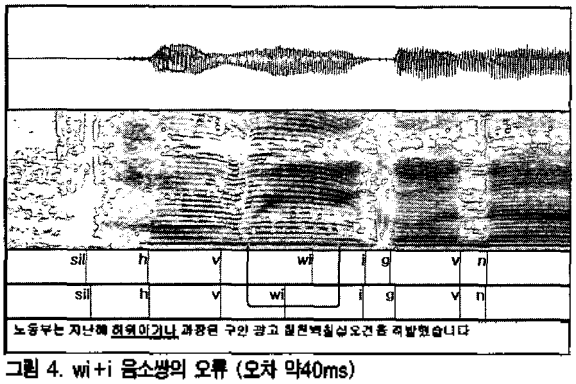

그림 4. wi+i 음소쌍의 오류 (오차 약40ms)

Figure 4. The error pattern of the pair of wi-i (The segmentation error is about 40ms).

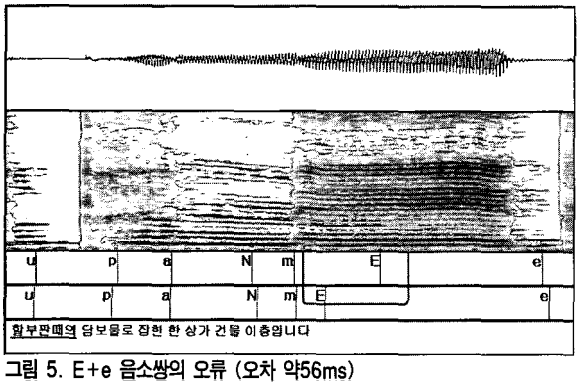

그림 5. E+e 음소쌍의 오류 (오차 약56ms)

Figure 5. The error pattern of the pair of E-e (The segmentation error is about 56ms).

세 번째는 한국어의 ㅞ, ㅙ, 발음과 ㅔ, ㅖ와 같이 발음의 구분이 불분명한 음소들이 연속해서 나오는 경우로서 정확률이 23% 정도로 매우 낮다. 그림5는 E+e 형태에서의 오류의 한 예이다.

이상의 경우들은 모두 음성 파형상이나 스펙트로그램상에서 보아도 아주 유사한 형태가 나타나기 때문에 사람의 눈으로도 확실한 경계를 구분하기 어렵고, 음성을 듣고서 분할을 하더라도 정확한 경계를 구분 짓기는 힘들기 때문에 자동 분할시에도 성공률이 낮은 것으로 판단된다.

3.2. 자음+모음

한국어 초성자음 18개를 조음적 특징에 따라 파열음, 파찰음, 마찰음, 유음, 비음의 5가지로 나누어 비교하였고, 종성의 경우는 하나의 분류로 나타내었다.

3.2.1. 파열음+모음

'ㅂ', 'ㅃ', 'ㅍ', 'ㄷ', 'ㄸ', 'ㅌ', 'ㄱ', 'ㄲ', 'ㅋ'의 파열음과 모음으로 구성된 음소쌍에서는 남성화자, 여성화자에서 각각 약 99%, 98.95%의 높은 정확률을 보였다.

3.2.2. 마찰음+모음

'ㅅ', 'ㅆ', 'ㅎ'의 마찰음과 모음으로 구성된 음소쌍은 여성화자의 경우 94.36%, 남성화자의 경우 92.28%의 정확률을 보였다. 나타나는 오류의 유형을 표 4에 나타내었다.

표 4. 마찰음+모음의 오류 유형

Table 4. The error pattern of the class of the fricative-vowel.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
마찰음+모음	7893	7898	445	609	94.36	92.28
h+모음	2506	3209	340	474	86.43	85.22
s+모음	3623	3577	76	114	97.90	96.81
S+모음	1764	1112	29	21	98.35	98.11

s+모음, S+모음에서는 98%의 정확률을 보이지만, h+모음에서는 86%의 낮은 정확률이 나타났다. 오류 중에서 h+모음의 경우가 여성화자에서는 76%, 남성화자에서 78%를 차지해 마찰음+모음 유형에서 대부분의 오류가 h+모음의 형태에서 난다고 볼 수 있다. 이 오류를 살펴보면 대부분은 'ㅎ'의 유성화로 인하여 발생하였고, 뒤의 모음이 무성화 되어 발생하는 오류도 있었다.

3.2.3. 파찰음+모음

파찰음은 파열음과 마찰음의 소리특성을 동시에 가지는 소리이며, 한국어에서는 'ㅈ', 'ㅉ', 'ㅊ'가 여기에 속한다. 파찰음+모음 음소쌍에 대한 정확률은 여성화자, 남성화자 각각 99%, 98%을 보이고 있다. 표 5에 나타난 파찰음+모음의 오류 분포를 살펴보면, Z+모음에서는 100%에 가까운 정확률을 보였다.

표 5. 파찰음+모음의 오류 유형
Table 5. The error pattern of the class of the affricate-vowel.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
파찰음+모음	4863	4864	24	57	99.50	98.82
z+모음	3129	3121	19	25	99.39	99.2
Z+모음	467	474	0	2	100	99.58
c+모음	1267	1269	5	30	99.60	97.63

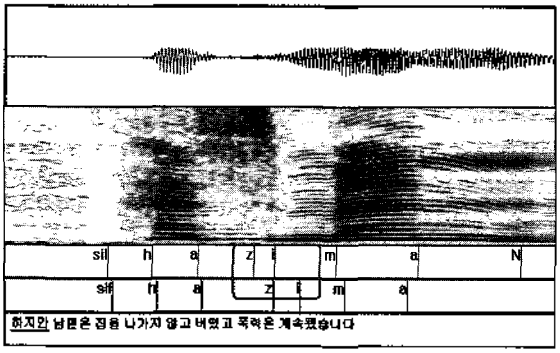

그림 6. z+i 음소쌍의 오류 (오차범위 약22ms)
Figure 6. The error pattern of the pair of z-i (The segmentation error is about 22ms).

파찰음+모음에서 나타나는 오류의 대부분은 파찰음에 의해서 뒤에 나타나는 모음이 무성화된 경우이다. c+u, z+i, z+a 경우에서처럼 모음 앞의 c와 z에 의해 후에 나타나는 모음이 무성화 되었기 때문에 나타난 오류라고 판단 할 수 있다. 그림 6은 이러한 경우의 예를 보여준다.

3.2.4. 비음+모음

한국어의 비음에는 'ㄴ', 'ㅁ', 'ㅇ'가 있지만 여기서는 'ㅇ'을 제외한 'ㄴ', 'ㅁ'+모음의 경우만 생각한다. 'ㅇ'는 비음에 하나이지만 초성에서는 나타나지 않고 종성에서만 나타나므로 후에 종성+모음의 형태에 포함시켜 비교하기로 한다. 비음+모음에 대해서 여성화자, 남성화자 각각 95%, 98%의 정확률을 보였다.

표 6을 보면 여성화자의 경우, m+모음의 정확률은 99%를 보이는 반면에 n+모음은 93%이고 비음+모음 오류의 92%가 n+모음에서 발생하고 있다. 그리고 n+모음에서 나타나는 오류의 92%는 그림 7에서 볼 수 있는 것처럼 n+i에서 발생한다. 이 오류의 원인은 문장의 마지

표 6. 비음+모음의 오류 유형
Table 6. The error pattern of the class of the nasal-vowel.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
비음+모음	7063	7056	364	97	94.86	98.62
n+모음	4848	4842	338	70	93.02	98.55
m+모음	2215	2214	26	27	98.82	98.78

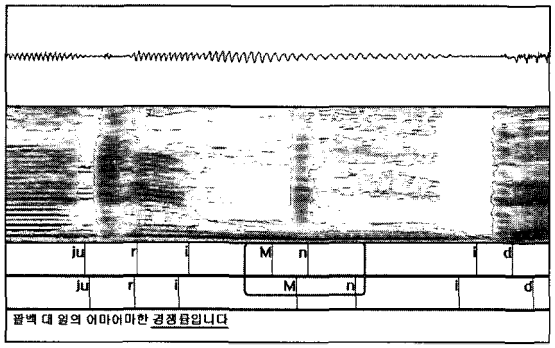

그림 7. n+i 음소쌍의 오류 (오차범위 약32ms)
Figure 7. The error pattern of the pair of n-i (The segmentation error is about 32ms).

막에서 발생하는 "니다"나 "니까"의 발성이 불명확하기 때문이다.

3.2.5. 유음+모음

한국어에서 유음은 'ㄹ'이다. 유음+모음에서는 여성화자과 남성화자에서 각각 96.11%, 90.25%의 정확률을 보였다. 자동음소분할된 결과를 살펴보면, 수작업된 결과에 비하여 'ㄹ'의 구간이 길게 나타남을 볼 수 있었다.

3.2.6. 종성+모음

한국어에서 실제 발음에 나타나는 종성은 'ㄱ', 'ㄴ', 'ㄷ', 'ㄹ', 'ㅁ', 'ㅂ', 'ㅇ'의 7가지이다. 종성+모음의 경우에는 여성화자과 남성화자에서 각각 90.27%, 82.78%의 정확률을 보이고 있다. 표 7에 유형들을 정리하였다.

이러한 오류를 분석해 보면 단어 마지막의 종성 뒤에 전사파일에서는 존재하지 않지만 발생되는 짧은 묵음이

표 7. 종성+모음의 오류 유형
Table 7. The error pattern of the class of the final consonant-vowel.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
종성+모음	1686	1580	164	272	90.27	82.78
K+모음	79	73	21	8	73.41	89.04
N+모음	366	280	39	44	89.34	84.28
T+모음	11	9	2	1	81.81	88.89
L+모음	187	186	41	38	78.07	79.57
M+모음	68	61	5	12	92.64	80.32
P+모음	45	41	10	6	77.78	85.36
O+모음	930	930	46	163	95.05	82.47

포함 되어져 있을 경우 수작업에서는 짧은 묵음을 모음에 두었지만, 자동음소결과에서는 그렇지 못하였다.

3.3. 모음+자음

3.3.1. 모음+파열음

모음+파열음에서는 여성화자과 남성화자에 대해서 각각 94.14%, 91.09% 의 정확률을 보이고 있다. 표 8에서와 같이 모음+b, 모음+d, 모음+g 의 경우에는 96%이상의 높은 정확률을 보이는 반면에 모음+B, 모음+t, 모음+D 의 경우에는 두 화자모두 80%미만의 정확률을 보였다. 이 경우 대부분은 모음 발화 후 발생하는 closer를 수작업에서는 중간부분에 경계를 취하였지만, 자동음소분할에서는 모음이 끝나는 부분에 경계를 두는 현상을 확인하였다.

표 8. 모음+파열음의 오류 유형
Table 8. The error pattern of the class of the vowel-plosive.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
모음+파열음	9258	9128	542	813	94.14	91.09
모음+b	1184	1139	44	61	96.28	94.64
모음+B	98	92	38	24	61.22	73.91
모음+p	474	466	86	120	81.85	74.25
모음+d	3111	3082	24	66	99.22	97.86
모음+D	157	155	44	40	71.97	74.19
모음+t	512	503	106	102	79.29	79.72
모음+g	2916	2888	58	296	98.01	89.75
모음+G	303	303	52	44	82.83	85.48
모음+k	503	500	90	60	82.10	88.00

3.3.2. 모음+마찰음

모음+마찰음에서는 여성화자과 남성화자에 대해서 각각 92.78%, 91.94%의 정확률을 보이고 있다.

표 9. 모음+마찰음의 오류 유형
Table 9. The error pattern of the class of the vowel-fricative.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
모음+마찰음	4268	4233	308	341	92.78	91.94
모음+s	1979	1952	53	22	97.32	98.87
모음+S	1132	1152	21	22	98.14	98.09
모음+h	1157	1129	234	297	79.77	73.69

표 9에서 볼수 있듯이, 모음+마찰음의 경우 여성화자, 남성화자 모두 모음 뒤에 s, S, h가 나타나는 빈도는 비슷하지만, 모음+s에서는 97%, 모음+S에서는 98%의 높은 정확률을 보이는 반면에 모음+h의 경우는 80%정도이다. 모음+h 이후에 모음이 연이어서 나타나는 경우에는 무성음인 h가 유성음처럼 발음되는 상황이 발생하기 때문에 모음과 모음 사이에 나타나는 h에서는 오류가 발생한다.

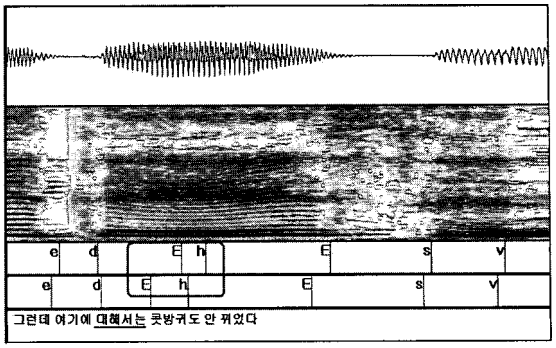

그림 8. E+h 음소쌍의 오류 (오차범위 약35ms)
Figure 8. The error pattern of the pair of E-h (The segmentation error is about 35ms).

그림 8은 이러한 경우의 예를 보여 준다. 이전과 이후에 나타나는 모음의 영향으로 파형상에서나 스펙트로그램상에서 무성음인 h의 특성을 찾아볼 수 없다.

3.3.3. 모음+파찰음

모음+파찰음 음소쌍에 대해서 여성화자, 남성화자 각각 93.19%, 94.99%의 분할 성공률을 보였다. 파찰음들에 대한 분할 성공률을 살펴보면 여성화자, 남성화자에 대해 각각 모음+z에서 95%와 98%, 모음+Z에서 84%와 84%, 모음+c에서 87.5%와 93.6%를 보였으며, 이는 모음과 파찰음 사이에 존재하는 closer부분에 대한 경계오류였다.

3.3.4 모음+종성

모음+종성 음소쌍의 경우 여성화자, 남성화자 각각 90%, 93%의 정확률을 보였다. 표 10에 나타난 모음+종성 유형의 정확률을 살펴보면 대부분 93% 이상의 정확률을 나타내는 반면에 모음+L의 경우 여성화자은 67%로 매우 낮으며, 남성화자은 83%로 낮은 편이다.

표 11과 표 12에서 모음+L에서 나타나는 오류의 형태를 빈도수 별로 정리하였다. 출현빈도는 (오차 범위 20ms 이상인 수/전체 출현 빈도)의 형태로 나타내었다. 전반적으로 정확률이 낮고 특히 그림 9에서 볼 수 있는

표 10. 모음+종성의 오류 유형
Table 10. The error pattern of the class of the vowel-final consonant.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
모음+종성	12610	12586	1283	836	89.82	93.36
모음+K	1151	1149	70	68	93.91	94.08
모음+N	4945	4927	139	214	97.18	95.66
모음+T	507	507	21	27	95.85	94.67
모음+L	2745	2748	905	455	67.03	83.44
모음+M	2676	2676	112	43	95.81	98.39
모음+P	586	579	36	29	93.85	94.99

표 11. 모음+L의 오류 유형 (female)
Table 11. The error pattern of the class of the vowel-L (female).

음소쌍	출현빈도	정확률	음소쌍	출현빈도	정확률
U+L	230/1010	77.23	jv+L	7/142	95.07
i+L	64/387	83.46	e+L	7/43	83.72
a+L	28/465	93.98	wv+L	6/41	85.37
u+L	61/240	74.58	we+L	3/23	86.96
o+L	9/167	94.61	wa+L	3/34	91.18
v+L	29/149	80.54	ju+L	6/15	60.00

표 12. 모음+L의 오류 유형 (male)
Table 12. The error pattern of the class of the vowel-L (male).

음소쌍	출현빈도	정확률	음소쌍	출현빈도	정확률
U+L	431/1009	57.28	jv+L	28/143	80.42
i+L	123/386	68.13	e+L	16/43	62.79
a+L	113/465	75.70	wv+L	8/40	80.00
u+L	69/240	71.25	we+L	8/23	65.22
o+L	47/163	71.17	wa+L	7/34	79.14
v+L	34/149	77.18	wi+L	7/11	36.36

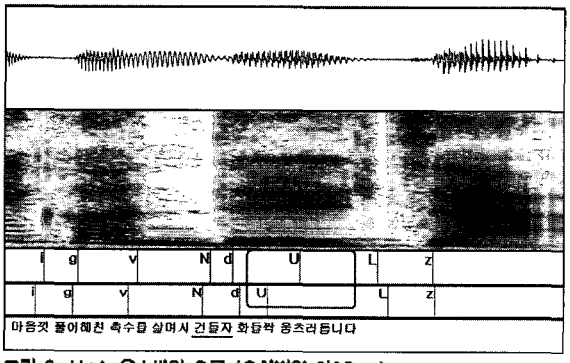

그림 9. U+L 음소쌍의 오류 (오차범위 약35ms)
Figure 9. The error pattern of the pair of U-L (The segmentation error is about 35ms).

것처럼 모음+L형태의 오류중 대부분을 차지하는 U+L의 형태에서는 여성화자의 정확률이 60%에도 미치지 못하고 있다. 종성 "ㄹ"의 경우 모음과의 구별이 불분명하여 수작업에서도 그 경계의 구분이 상당히 힘들다. 그리고 이 경우 그림 9처럼 자동음소분할된 결과를 보면 수작업된 결과에 비하여 모음쪽으로 경계가 많이 이동한 것을 확인할 수 있었다.

3.4. 종성+초성

7가지의 종성을 다시 폐쇄음 (K, P, T), 비음 (N, M, O), 유음 (L)으로 나누어 출현빈도와 오류분포를 표 13, 표 14, 표 15에 정리하였다.

표 13에서 나타난 폐쇄음+자음의 오류 유형을 살펴보았을 때, 전반적으로 정확률이 낮은 것을 볼 수 있다. 특히 폐쇄음+비음의 경우는 약52%의 아주 낮은 정확률을 보였다. 폐쇄음+비음에서 나타나는 오류의 70%는 K+m에서 나는 오류이다. K+m에서 오류가 나타나는 두 가지

표 13. 폐쇄음+자음의 오류 유형
Table 13. The error pattern of the class of the stop-consonant.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
폐쇄음+파열음	1107	1084	94	134	91.23	87.64
폐쇄음+마찰음	497	495	67	66	86.51	86.67
폐쇄음+파찰음	386	387	68	61	82.38	84.24
폐쇄음+비음	55	49	26	27	52.72	44.89
폐쇄음+유음	6	4	2	2	66.66	50.00

표 14. 비음+자음의 오류 유형
Table 14. The error pattern of the class of the nasal-consonant.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
비음+폐쇄음	3244	3110	67	104	97.93	96.65
비음+마찰음	1923	1826	89	72	95.37	96.05
비음+파찰음	1197	1169	26	39	98.07	96.66
비음+비음	2557	2514	425	266	83.37	89.41
비음+유음	20	20	8	9	60.0	55.0

표 15. 유음+자음의 오류 유형
Table 15. The error pattern of the class of the liquid-consonant.

	출현빈도		20ms이상		정확률(%)	
	female	male	female	male	female	male
유음+폐쇄음	971	959	24	64	97.52	93.32
유음+마찰음	483	479	36	47	92.54	90.18
유음+파찰음	354	344	12	19	96.61	94.48
유음+비음	224	226	16	21	92.85	90.71
유음+유음	480	473	31	79	93.54	83.30

예를 그림10과 그림11에 나타나 있다. 그림 10은 실제 발성시에 K와 m이 휴지 (pause)구간 없이 연이어서 발음되어 종성K가 종성O로 발음된다. 이는 수작업 과정에서 오류라고 판단할 수 있다. 그림11는 K와 m사이에 짧거나 혹은 긴 휴지구간을 포함 한 예로서 짧은 휴지구간 (short pause)을 수작업 분할할 때 일관성 결여가 원인인 것으로 판단된다.

비음+자음의 경우에서는 표 14에 나타난 것처럼, 비음+비음과 비음+유음의 경우에 상대적으로 낮은 정확률을 보였다. 특히 M+n의 음소쌍에서 가장 많은 오류가 발생하였다. 여성화자의 경우 오류의 약 30%가 M+n에

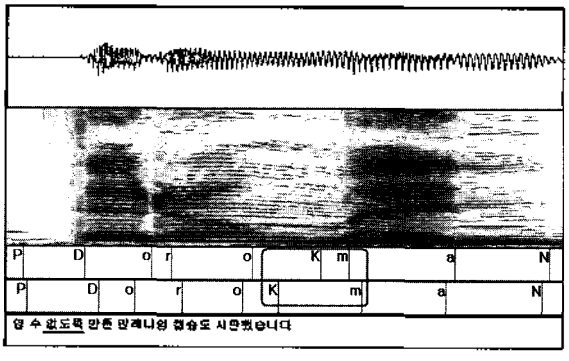

그림 10. K+m 음소쌍의 오류 (오차범위 약35ms)
Figure 10. The error pattern of the pair of K-m (The segmentation error is about 35ms).

서 발생하며 문장의 끝에서 발생하는 경우가 대부분이다. 그림 12는 비음+유음의 오류 형태에서 많이 나타나는 M+n 오류의 한 예이다. 이 경우 "습니다"와 같이 문장의 마지막에서 발생하였다.

표 15에 나타난 유음+모음의 경우는 대부분 90%이상의 정확률을 보였다. 유음+마찰음일 경우에는 오류의 대부분이 L+h에서 나타났다 이런 결과의 원인중 하나는 그림 13에서 볼 수 있는 것처럼 수작업에 의한 분절 시에는 L과 h의 경계를 나누었지만, 이 경우에는 L+h에서 실제 발음시에 h가 발음되지 않기 때문이다.

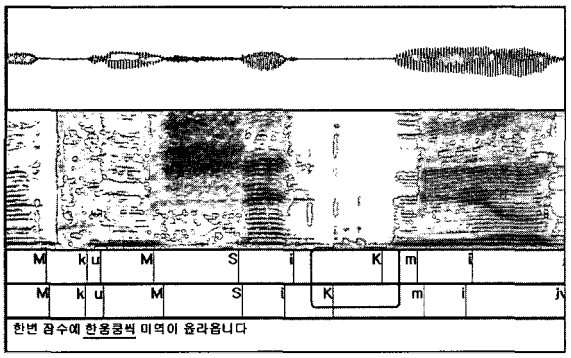

그림 11. K+m 음소쌍의 오류 (오차범위 약80ms)
Figure 11. The error pattern of the pair of K-m (The segmentation error is about 80ms).

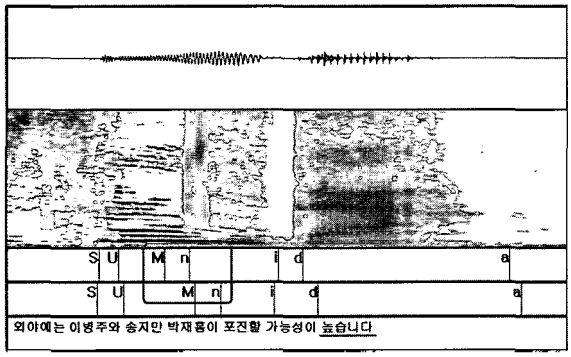

그림 12. M+n 음소쌍의 오류 (오차범위 약 38ms)
Figure 12. The error pattern of the pair of M-n (The segmentation error is about 38ms).

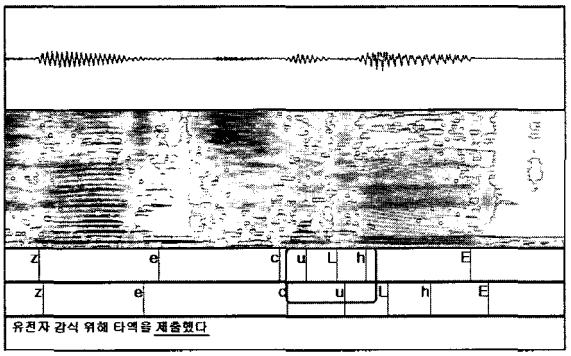

그림 13. L+h 음소쌍의 오류(L의 오차범위는 약 63ms, h의 오차범위는 약 82ms)
Figure 13. The error pattern of the pair of L-h (L's error range is about 63ms, h's error range is about 82ms)

IV. 결론 및 향후 연구 계획

본 논문은 후처리기 개발시, HMM기반의 자동음소분할의 결과, 정확률이 낮은 유형들을 모색하여 그 부분에 대한 정확률을 개선하기 위한 기본 데이터를 제공하는데 그 목적이 있다. 유성음이 무성음화 되는 경우, 혹은 무성음이 유성음화 되는 부분에서는 대부분의 경우 오류가 나타났으며, 폐쇄구간 (stops)을 포함하고 있는 폐쇄음, 마찰음이 연속해서 나올 경우에도 자동 분할 성공률이 낮음은 확인할 수 있었다. 또한 대부분이 문장의 끝에서 나타나는 M (종성ㅁ)+n (초성ㄴ), n+i (ㅣ) 음소쌍에서도 오류의 빈도가 높았다. 그리고 유음이 나타날 경우 (종성 ㄹ (음소기호 L))에는 정확률이 현저히 떨어지는 경향을 보이므로 이에 대한 연구가 필요하다.

앞으로 정리한 유형들을 바탕으로 폐쇄구간을 통계적으로 분절할 수 있는 방안과 경계를 보정할 수 있는 방안에 대해 연구가 진행될 예정이다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구센터 육성지원사업의 연구결과로 수행되었습니다.

참 고 문 헌

1. 김상훈, 이정철, 강동규, 이영직, "대용량 운율 음성데이타를 이용한 자동 합성방식" 제 15회 음성통신 및 신호처리 워크샵, **15** (1) 87-92, 1998
2. Sethy, A., Narayanan, S., "Refined Speech Segmentation for Concatenative Speech Synthesis", Proc. ICSLP 2002, Denver, 2002, 149 - 152
3. A. Komatsu, A. Ichikwa, D. Nakota, Y. Asakawa, H. Matsuzaka, "Phoneme recognition in the continuous speech", in Proc. ICASSP 1982, 883-886
4. 박순철, 김봉완, 이용주, "문맥종속 반음소단위에 의한 음운 자동 레이블링 시스템의 성능 개선", 말소리 **37** (6), 23 - 48, 1999
5. Leung, H. C. and W. ZUE, "A procedure for automatic alignment of phonetic transcription with continuous speech", in Proc. ICASSP 1984, Apr., 429~432
6. Matthew J. Makashay, Colin W. Wightman, Ann K. Syrdal, and Alistair Conkie, "Perceptual Evaluation of Automatic Segmentation in Text-to-Speech Synthesis", in Proc. ICSLP 2000, Beijing, 2000, 431-434.
7. Yeon-Jun Kim, A.Conkie, "Automatic segmentation Combining an HMM-Based Approach and Spectral Boundary Correction", in Proc. ICSLP 2002 Sept. 145~148

8. 박혜영, 김형순, "자동 음성 분할을 위한 음향 모델 에너지 기반 후처리", 말소리 43 (6) 대한음성학회, 137-149, 2002
9. D.T. Toledano, "Neural network boundary refining for automatic speech segmentation", Porc. ICASSP 2000, 3438~3441
10. 신지영, "우리말 소리의 이해", 대한음성학회 창립25주년기념학술대회 논문집, 15-23, 2002

저자 약력

• 김 민 제 (Min-Je Kim)

2004년: 울산대학교 컴퓨터.정보통신공학부 (학사)
2004년~2006: 울산대학교 컴퓨터공학과 대학원 (석사)
※주관심분야: 음성인식, 자동음소분할

• 이 정 철 (Jung-Chul Lee)

1984년: 서울대학교 전자공학과 (학사)
1988년: 서울대학교 전자공학과 (석사)
1998년: 서울대학교 전자공학과 (박사)
1985~2000년: 한국전자통신연구원 책임연구원
2000년: L&H Korea 전문위원
2001년: (주) 보이스텍 전문위원
2002년: (주) 코난테크놀로지 책임연구원
2002년~현재: 울산대학교 교수
※주관심분야: 음성합성, 음성코딩

• 김 종 진 (Jong-Jin Kim)

한국음향학회지 제 24권 2호 참조