문맥종속 반음소단위 모델을 이용한 자동 음소분할 및 레이블링 시스템의 구현

김태환*, 김봉완*, 박순철*, 이용주*

The Implementation of Automatic Segmentation and Labelling System Using Context-dependent Demi-phone

(Tae-Hwan Kim*, Bong-Wan Kim*, Soon-Cheol Park*, Yong-Ju Lee*)
*Dept., of Computer Eng., Wonkwang Univ.

요 약

음소 단위로 레이블링된 데이터베이스는 음성연구에 있어 매우 중요하다. 그러나 수작업에 의한 음소분할 및 레이블링 작업은 많은 시간과 노력이 필요하기 때문에 자동 음소분할 및 레이블링 시스템에 대한 많은 연구가 진행되고 있다.

본 논문에서는 monophone과 triphone의 장점을 포함하는 문맥 종속 반음소 단위 모델을 이용한 자동 음소분할 및 레이블링 시스템을 구현하였다. 레이블링 단위로는 68개의 유사음소와 묵음 등 총 69개로 정하였으며, 음소 모델링은 연속 HMM을 사용하였다. 기존의 subword 단위모델과 본 논문에서 제안한 문맥종속 반음소 모델을 이용한 자동 음소분할 및 레이블링 시스템의 성능비교 결과 음소경계오차가 10ms이내인 경우 각각 60.17%, 66.32%를 포함하여 6.15%의 향상을 보이고, 40ms 이내인 경우 90.36%, 94.27%를 포함하여 3.92%의 성능향상을 보였다.

I. 서 론

음성의 과학적 연구를 위해서는 음소 단위로 분할되고 레이블링된 대량의 음성 데이터베이스의 구축이 필수적이다. 수작업에 의한 음성분할은 많은 시간이 소요되는 작업이며, 일관성이 보장되지 않는 문제점[1]을 안고 있기 때문에, 자동 음성분할 기술이 다양하게 연구되어 왔다[2,3,4,5].

음성을 분할하기 위한 기본 단위로는 monophone, biphone, triphone 등이 사용될 수 있다. monophone의 경우 훈련할 모델의 수가 적기 때문에 훈련하기 쉬운 반면, 전·후 음소에 의한 조음효과를 표현하지 못하는 단점이 있다. 이러한 단점을 극복하기 위해 triphone과 같은 문맥종속 단위를 사용한다. 그러나, triphone의 경우 인접한 음소들에 의하여 음향학적 특성이 변화되는 각 음소들마다 별도의 모델링을 하기 때문에 모델의 수가 크게 증가하게 되고, 이로 인해서 학습데이터 양이 부족하게 되는 단점이 있다.

따라서, 본 논문에서는 triphone에 비해 모델의 수를 훨씬 줄이면서도, 전·후 음소에 대한 조음효과를 반영하는 문맥종속 반음소단위를 기본 모델로 사용하여 자동 음소분할 시스템을 구현하였다.

2절에서는 본 논문에서 사용한 반음소에 대해서 기술하고, 3절에서는 구현된 자동 음소분할 시스템의 구성에 대해서, 그리고 4절에서 실험 및 결과에 대해서 기술한 후, 마지막으로 결론을 맺는다.

Ⅱ. 반음소(Demiphone)

1. 반음소의 정의

반음소는 음소 또는 변이음을 그것의 전·후 음소에 영향을 받지 않는 정상상태 시점의 중점을 경계로 양분함으로써 얻어지는 음성단위이다. 반음소는 선행음소에 의한 조음효과를 포함하는 전반음소(left-demiphone)와 후행음소에 의한 조음효과를 포함하는 후반음소(right-demiphone) 두 부분으로 이루어져 있다[6,7].

이와 같이 음소를 정상상태 시점의 중점을 경계로 나누어 생각할 수 있는 것은 대부분의 음소가 상이한 특성을 가진 두 부분으로 이루어져 있으며, 선행음소에 의한 조음효과는 전반음소에 국한되고 후행음소에 의한 조음효과는 후반음소에 제한되기 때문이다[6].

2. 반음소의 장점

음성 인식에서 반음소를 인식의 단위로 사용함으로써 얻을 수 있는 장점은 아래와 같다[7].

첫째, 훈련시 발생하지 않은 음소에 대해서 인식시 새로운 모델을 생성하기가 유리하다. 그림 1에 그 예를 보이고 있다. 그림 1.a에서 음소 'a'는 'g-a a+z'로 훈련되고 1.b에서 'n-a a+b'로 훈련되었을 때, 인식시 반음소 'g-a'와 'a+b'를 가지는 새로운 'a'에 대해서 'g-a a+b'로 구성하므로써 쉽게 생성해 낼 수 있다. 그림 1에서 # 기호는 묵음(silence)를 의미하고, '-'기호는 전반음소를 '+'기호는 후반음소를 의미한다.

둘째, triphone 모델이나 병렬적인 biphone 모델을 이용하지 않고도 전·후 음소에 의한 조음현상을 반영하는 모델을 생성할 수 있다.

셋째, triphone 모델에 비해 훨씬 적은 모델수로 전·후 조음현상을 반영할 수 있고, 같은 양의 데이터를 사용할 경우 각 모델에 대해 더 많은 양의 훈련을 시킬 수 있다. 만약 10개의 음소가 어떠한 제약조건도 없이 출현할 수 있는 상황에서 triphone 단위 모델을 훈련시키려 한다면, 총 1000개(10×10×10)의 triphone을 훈련시켜야 한다. 반면에 demiphone의 경우 전반음소 100개(10×10)와 후반음소 100개(10×10)를 포함하여 총 200개의 모델을 훈련시키기만 하면 된다. 또한 학습량에 있어서 하나의 음소에서 전반음소와 후반음소를 동시에 훈련시킬 수 있으므로, 같은 양의 학습데이터에서 biphone모델을 훈련시키는 것에 비해 2배의 효과를 볼 수 있다.

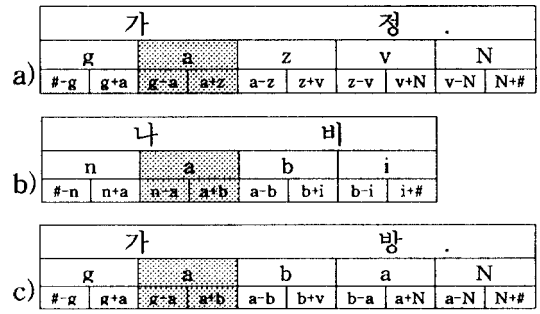

그림 1. 훈련되지 않은 음소에 대한 반음소 모델 예.
a, b) 훈련용 데이터
c) 테스트 데이터

Ⅲ. 자동 음소분할 시스템의 구현

본 절에서는 본 논문에서 구현한 문맥종속 반음소단위의 자동 음소분할 시스템과 평가를 위해서 구현한 두 개의 자동 음소분할 시스템에 대해서 기술한다.

1. 레이블링 단위 선정

구현된 레이블링 시스템에서는 레이블링의 기본단위로 PBW 452 어절에 사용된 바 있는 유사음소를 선정하였다[8,9].

유사음소의 목록은 표 1~4에 나타낸 바와 같이 한국어 기본 음소단위에 파열음, 파찰음, 그리고 유음의 경우 폐쇄구간과 파열/마찰 구간으로 나누고 나뉘어진 각 구간의 유성음화를 고려하였다. 공명음화, 파열음·파찰음에서의 불파음화, 그리고 유음의 탄설음화의 경우 단일 구간으로 취급하였다. 마찰음의 경우 유성음화와 공명음화를

고려하였으며, 이중모음인 'ㅖ' 와 'ㅒ', 'ㅚ' 와 'ㅞ'를 하나의 음소로 취급하였다. 이렇게 89개의 유사음소와 1개의 묵음을 포함하여 총 90개의 레이블링 단위를 선정하였다.

	음소	폐쇄구간	폐쇄구간의 유성음화	파열/마찰구간	파열/마찰구간의 유성음화	불파음화	공명음화
파열음	ㄱ	g	gV	gH	gHV	gC	gR
	ㄲ	G	GV	GH	GHV		
	ㅋ	k	kV	kH	kHV		
	ㄷ	d	dV	dH	dHV	dC	dR
	ㄸ	D	DV	DH	DHV		
	ㅌ	t	tV	tH	tHV		
	ㅂ	b	bV	bH	bHV	bC	bR
	ㅃ	B	BV	BH	BHV		
	ㅍ	p	pV	pH	pHV		
파찰음	ㅈ	z	zV	zH	zHV		zR
	ㅉ	Z	ZV	ZH	ZHV		
	ㅊ	c	cV	cH	cHV		

표 1. 파열음과 파찰음의 기호 목록

	음소	마찰성분	유성음화	공명음화		음소	기호
마찰음	ㅅ	s	sV		비음	ㅁ	m
	ㅆ	S				ㄴ	n
	ㅎ	h	hV	hR		ㅇ	N

표 2. 마찰음과 비음의 기호 목록

	음소	폐쇄구간	폐쇄구간의 유성음화	기식음	기식음의 유성음화	공명음화	탄설음화
유음	ㄹ	r	rV	rH	rHV	rR	l

표 3. 유음의 기호 목록

	음소	기호	음소	기호	음소	기호	음소	기호
모음	ㅏ	a	ㅓ	i	ㅑ	ja	ㅝ	wv
	ㅓ	v	ㅔ	e	ㅕ	jv	ㅞ	wE
	ㅗ	o	ㅐ	E	ㅛ	jo	ㅟ	wi
	ㅜ	u	ㅢ	Ui	ㅠ	ju	ㅙ, ㅞ	we
	ㅡ	U			ㅒ, ㅖ	je	ㅘ	wa
묵음	C							

표 4. 모음과 묵음의 기호 목록

2. 분할을 위한 음소단위 구성

평가를 위한 유사음소단위 자동 음소분할 시스템은 레이블링 단위에서 유음의 폐쇄음화, 마찰음의 유성음화, 파열음의 불파음화와 'ㄱ, ㄷ, ㅂ'을 제외한 파열음 폐쇄구간과 기식구간의 유성음화를 제외한 68개의 유사음소와 1개의 묵음을 포함하여 총 69개의 음소단위를 사용하였다.

본 논문에서 구현한 문맥종속 반음소단위 자동 음소분할 시스템은 평가용 시스템에서 사용한 68개의 유사음소를 문맥종속 반음소단위를 구성하였다.

일반적으로 반음소라 하면, 앞 절에서 정의한 바와 같이 음소의 정상상태를 기점으로 나눈다. 그러나, 이렇게 레이블링된 데이터를 가지고 있지 않은 상황이므로 각 유사음소의 중점을 기준으로 나누고, 표 5 에서처럼 음소를 17개의 전·후 문맥정보로 분류하여 문맥종속 반음소단위를 사용하였다.

음소분류	기호	음소분류	기호
파열음의 폐쇄구간	sS	파찰음의 폐쇄구간	sA
파열음 폐쇄구간의 유성음화	sVS	파찰음의 폐쇄구간의 유성음화	sVA
파열음의 파열구간	bS	파찰음의 마찰구간	bA
파열음의 파열구간의 유성음화	bVS	파찰음의 마찰구간의 유성음화	bVA
마찰음	F	비음	N
성문마찰음	hF	모음	V
유음	L	이중모음	yV
유음의 공명음화	RL	묵음	sil
유음의 탄설음화	lL		

표 5 문맥정보를 위한 음소 분류

3. 시스템의 훈련

훈련에 사용한 음성 데이터베이스는 국어공학센터의 음소적으로 균형이 잡힌 452단어(PBW : Phonetically Balanced Word) 데이터베이스[9]에서 레이블링된 남성화자 4명분을 사용하였다. 국어공학센터의 PBW 음성 데이터베이스는 방음 부스에서 Senheizer HMD224X를 사용하여 녹음되었으며, DAT(Digital Audio Tape)에 저장되었다. AD/DA 변환은 KAY CSL 4300B를 이용하여 16kHz로 Sampling하고 16Bits로 양자화되었다.

음성 분석은 10ms의 Hamming 윈도우를 5ms 간격으로 이동시키면서 분석하고, 특징파라미터로는 Mel-frequency cepstrum과 이들의 시간축 미분값, 그리고 정규화된 에너지와 그 미분치를 사용하였다.

음소 모델의 확률분포모델로는 연속확률

분포를 사용하였고, 모델의 형태는 평가용 시스템의 경우 5상태 7개의 천이를 가지는 모델과 8상태 13개의 천이를 가지는 left-right 모델로 구성하였으며, 문맥종속 반음소단위 시스템의 경우 5상태와 7개의 천이를 가지는 평가용 시스템과 동일한 모델을 사용하였다.

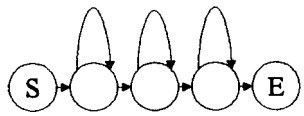

그림 3. 5상태 7천이를 가지는 left-right모델

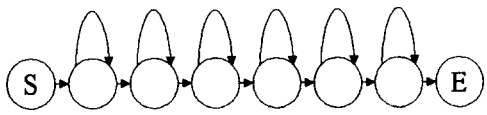

그림 4. 8상태 13천이를 가지는 left-right 모델

평가를 위해서 2개의 모델을 사용한 이유는 다음과 같다. 첫째, 반음소 모델과 동일한 topology로 모델화된 것을 비교한다. 둘째, 그림 5에서 보이는 것과 같이 반음소단위 모델의 경우 전반음소로 시작하여 후반음소로 종료함으로써 하나의 유사음소를 표시하므로, 반음소 모델의 2배가 되는 상태를 가지고 모델화된 것을 비교한다.

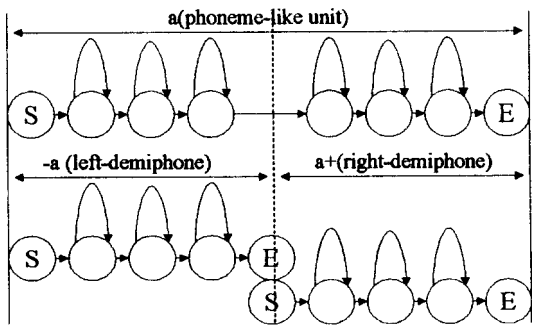

그림 5. 반음소단위 모델과 평가용 단위모델 비교

Ⅳ. 실험 및 결과

실험은 PBW 452어절 DB에서 훈련에 사용하지 않은 1명의 화자의 데이터를 이용하여 수행하였다.

본 시스템은 음성 입력에 대한 언어정보로 철자 표기와 음소 표기를 지원한다. 철자표기는 ETRI에서 개발한 한국어 발음 표기변환 프로그램을 사용하여 발음표기를 생성한 후, 이를 유사음소 또는 문맥종속 반음소 표기로 변환시켜 사용하였다.

평가에서는 언어정보로 철자표기를 사용하였으며, 수작업으로 레이블한 결과 음소(발성 내용)와 본 시스템에서 레이블한 음소가 다른 경우에는 서로 대응하는 음소들을 비교함으로써 평가하였다.

수행한 결과는 수작업으로 레이블한 결과와 비교하여 표 6에 보이는 것처럼 경계의 위치가 벗어난 정도를 5ms에서 50ms까지 5ms씩 증가시켜가며 표시하였다.

오차율	유사음소단위 모델 (5상태 7 천이)	유사음소단위 모델 (8상태 13천이)	반음소단위 모델 (5상태 7천이)
<=5	38.61%	39.37%	44.49%
<=10	59.10%	60.17%	66.32%
<=15	70.17%	71.73%	76.82%
<=20	77.16%	77.03%	83.77%
<=25	81.14%	81.77%	88.07%
<=30	84.10%	85.43%	91.34%
<=35	86.30%	88.58%	93.21%
<=40	88.00%	90.36%	94.27%
<=45	90.35%	92.10%	95.51%
<=50	91.48%	93.67%	96.32%
총 음소개수 : 3857			

표 6. 자동 음소분할 시스템의 성능평가 결과

평가 결과 유사음소단위 모델보다 반음소단위를 사용한 모델의 경계오차가 10ms인 경우 6.15%, 40ms인 경우 3.92%의 성능향상을 가져왔다.

반음소단위 모델에서 각 음소별 전이에서의 성능평가를 위해서 음소를 모음, 파열음의 폐쇄/폐쇄유성, 파열음의 파열/파열유성, 파찰음의 폐쇄/폐쇄유성, 파찰음의 마찰/마찰유성, 마찰음, 성문마찰음, 유음, 비음, 묵음의 총 10개의 그룹으로 나누어 평가하였다. 각 음소그룹의 기호는 표 7에 나타내었고, 전이에서의 성능평가는 표 8에 나

타내었다.

표 8을 살펴보면, 경계 오차는 대부분 모음과 모음, 모음과 성문마찰음, 성문마찰음과 모음, 파열/파찰음의 폐쇄구간이 연속된 부분에서 발생하였다.

파열/파찰음의 연속된 부분, 비음이 연속된 부분, 포만트 주파수가 비슷한 모음이 연속된 부분은 레이블링시 규칙에 의해 분할되었기 때문에 정확한 평가가 이루어졌다고 볼 수 없다[8]. 모음의 경우 성문 마찰음이 연속된 부분에서는 성문 마찰음이 약화되어 나타나지 않는 경우에 문제가 발생하였다. 또한 파열음/파찰음/마찰음 뒤에 모음이 나올 경우, 모음이 앞 음소의 영향을 받아 무성음화되어 뒤에 오는 음소와 경계 오류를 발생시켰다.

음소분류	기호	음소분류	기호
파열음의 폐쇄/폐쇄유성	sS	파찰음의 폐쇄구간	sA
파열음의 파열/파열유성	bS	파찰음의 마찰/마찰유성	bA
마찰음	F	비음	N
성문마찰음	hF	모음/이중모음	V
유음	L	묵음	sil

표 7 음소그룹별 기호

전	후	Total	<=10	<=20	<=40	전	후	Total	<=10	<=20	<=40
sil	bA	47	93.62%	97.87%	100.00%	bA	V	172	98.84%	99.42%	100.00%
sil	bS	152	84.21%	96.05%	100.00%	bS	V	604	84.92%	95.83%	99.21%
sil	F	31	61.29%	90.32%	100.00%	F	V	105	97.17%	100.00%	100.00%
sil	hF	26	80.77%	92.31%	96.15%	hF	V	68	48.53%	72.06%	79.41%
sil	L	4	25.00%	75.00%	100.00%	L	F	3	66.67%	100.00%	100.00%
sil	N	42	80.95%	97.62%	97.62%	L	hF	3	33.33%	66.67%	100.00%
sil	V	149	79.19%	98.66%	97.32%	L	N	3	33.33%	66.67%	66.67%
sS	bS	315	88.25%	95.56%	99.68%	L	bA	5	40.00%	40.00%	80.00%
sS	F	14	57.14%	100.00%	100.00%	L	sil	39	43.59%	74.36%	100.00%
sS	bA	12	33.33%	66.67%	100.00%	L	sS	18	61.11%	83.33%	94.44%
sS	sS	24	8.33%	54.17%	75.00%	L	V	76	72.37%	94.74%	98.68%
V	F	55	90.91%	100.00%	100.00%	N	F	14	92.86%	92.86%	100.00%
V	hF	35	22.86%	51.43%	71.43%	N	hF	22	90.00%	63.64%	86.36%
V	L	174	63.79%	79.31%	94.25%	N	N	22	45.45%	59.09%	90.91%
V	N	319	58.31%	90.88%	95.30%	N	sA	24	33.33%	70.83%	100.00%
V	sA	90	52.22%	83.33%	97.78%	N	sil	80	36.25%	71.25%	95.00%
V	sil	312	35.26%	62.82%	84.62%	N	sS	56	46.43%	85.71%	100.00%
V	sS	331	69.75%	87.01%	98.19%	N	V	169	63.91%	84.02%	92.31%
V	V	247	25.51%	46.96%	69.64%	bA	bA	94	92.55%	98.94%	100.00%
유사음소 개수 : 총 38577개											

표 8. 음소군별 천이에서의 성능평가

V. 결 론

본 연구에서는 음성을 분할하기 위해 문맥종속 반음소단위를 기본으로 하는 자동 음소분할 시스템을 구현하였다. 음성분할의 기본단위로 사용한 반음소는 전·후 음소에 의한 상호조음정보를 포함하면서도, 기존의 Triphone보다 훈련하기 쉬운 장점을 가진다.

실험결과 유사음소단위 모델보다 반음소 단위를 사용한 모델의 경계오차가 10ms인 경우 6.15%, 40ms인 경우 3.92%의 성능향상을 가져왔다. 경계 오차는 많은 부분이 파열음/마찰음/파찰음 뒤에 무성음화된 모음이 올 경우, 무성음화된 모음과 그 뒤에 나오는 음소와의 경계에서 발생하였다.

앞으로 모음이 무성음화된 부분에 대한 처리에 대해서 검토해야 할 것이며, 더 나은 성능을 위해서는 레이블링된 충분한 음성 데이터베이스를 가지고 모델을 훈련시킬 필요가 있다고 판단된다.

< 참 고 문 헌 >

[1] B. Eisen, H. G. Tillman, and C. Draxler, Consistency of judgments in manual labeling of phonetic segments: The distinction between clear and unclear cases, Proc of the ICSLP (Banff), 1992, pp. 871-874.

[2] T. Svendsen and Frank K. Soong, "On the Automatic Segmentation of Speech Signals," Proc. ICASSP, pp.77-80, 1987.

[3] O.Mella, D.Fohr, "Semi-Automatic Phonetic Labelling of Large Corpora," Eurospeech97, pp.1732, 1997

[4] Andreas Kipp, Maria-Barbara Wesenick, Florian Schiel,"Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora,"

[5] Ryszard Gubrynowics, Adan Wrzoskowics, "Labeller -A System of Automatic Labelling of Speech Continuous Signal," Eurospeech '93 pp.297, 1993.

[6] José B. Mariño, Albino Nogueiras, Antonio Bonafonte, "The Demi-phone: An efficient subword unit for continuous speech recognition".

[7] 이종락, "반음소 : 새로운 음성합성 및 인식단위", 제 10회 음성통신 및 신호처리 워크샵 논문집, pp. 208-212, 1993.

[8] 김종진, 이용주 외, "음성 DB 구축을 위한 한국어 레이블링 기준에 관한 연구," 제 13회 음성통신 및 신호처리 워크샵, pp.250-255, 1996.6

[9] 김봉완, 김종진, 김선태, 이용주, "공동이용을 위한 음성DB의 설계 및 구축에 관한 연구," 한국음향학회지, 16권 4호, 1997.5.