음운 자동 레이블링 시스템의 처리단위에 의한 성능비교

박순철*, 김태환*, 김봉완*, 이용주*
{scpark, babydol, joe, yjlee}@speech.wonkwang.ac.kr, 원광대학교 컴퓨터공학과

Performance compare by the processing unit of the automatic phoneme labelling system

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

<요 약>

본 논문에서는 레이블링 시스템에서 기본단위로 새롭게 제안된바 있는 demiphone의[1] 성능을 평가하기 위하여 monophone과 triphone, demiphone을 단위로 하는 레이블링 시스템을 구축하여 demiphone의 성능을 평가 하였다. 음성 데이터 베이스는 PBW 452단어를 대상으로 남자 30명분의 데이터를 훈련에 사용하였으며, 훈련에 사용하지 않는 남자 4명분의 데이터를 시스템의 평가에 사용하였다.

평가결과 demiphone을 사용한 경우 경계오차가 20ms 이하의 경우에는 monophone에 비하여 6.31%, triphone에 비해 6.21%로 성능이 우수하다. 그리고, 40ms 이하의 경우에는 각각 4.33% 와 3.68%의 성능 향상을 가져왔다.

1. 서론

음성연구를 수행하기 위해서는 음소와 같은 기본 단위로 분할되고 레이블링된 대량의 음성데이터베이스의 구축이 필수적으로 요구된다. 음소와 같은 기본단위로 분할 및 레이블링하는 작업은 사람이 직접 수행할 수 있지만, 수작업에 의한 음성데이터베이스 구축은 시간이 많이 소요되는 작업이며, 소수의 음성 전문가에 의존할 수 밖에 없고, 구체적인 판단기준을 미리 정해놓더라도 상당부분 개인의 주관적인 판단에 의존해야 하기 때문에 일관성이 결여되는 문제가 있다[2,3].

이와같은 문제를 해결하기 위하여, 음성을 자동분할 및 레이블링하는 기술이 다양하게 연구되어 왔다[4,5,6,7].

대용량 음성데이터베이스를 구축하기 위해 고려되어할 분야중의 하나가 레이블링시스템의 인식단위에 대한 문제이다. 음성을 분할하기 위한 기본단위로는 문장, 음절 등과 단위에서 현재는 subword-unit인 monophone, biphone, triphone등의 단위가 많이 사용되고 있다. monophone의 경우 사용하는 모델의 수가 대략 40~50개 정도로 적기 때문에 훈련데이터수가 적어 훈련에 쉬운 반면, 화자별 다양한 발성속도나 습관, 전후음소에 의한 음소의 조음효과를 표현하지 못하는 단점이 있다. 이러한 단점을 극복하기 위하여 biphone이나 triphone과 같은 문맥종속단위를 사용한다. 그러나, triphone의 경우 각 음소들을 인접한 음소들에 따라 별도로 모델링하기 때문에 음소의 전후음소에 대한 조음효과를 잘 표현한다. 반면, 모델의 수가 크게 증가하기 때문에 훈련 데이터의 양이 부족하게 되는 단점이 있다.

따라서, 본 논문에서는 triphone에 비해 모델의 수를 훨씬 줄이면서, triphone과 같이 전후음소에 대한 조음효과를 잘 반영할수 있는 demiphone의 유용성과 효용성을 평가하기 위하여 monophone과 triphone, demiphone을 비교 평가하였다.

2절에서는 본 논문에 사용된 demiphone를 정의하고, 3절에서는 자동음소분할 시스템의 구성에 대해서 기술한다. 4절에서 monophoe, triphone, demiphone의 성능을 분석한 후 마지막으로 결론을 맺는다.

2. demiphone

일반적으로 음소는 정상시점을 기준으로 선행음소의 영향을 받는 전반부와, 후행음소의 영향을 받는 후반부

로 분류할 수 있다.

이와같이, 음소를 두 부분으로 나누어 생각해볼 수 있는 이유는, 많은 경우 선행음소와 후행음소가 미치는 영향이 음소의 전반부 및 후반부에 국한된다는 점에서 [그림 1]. 의 (d)와 같이 음소를 성질이 서로 다른 두 부분으로 구분 지울 수 있다[8,9].

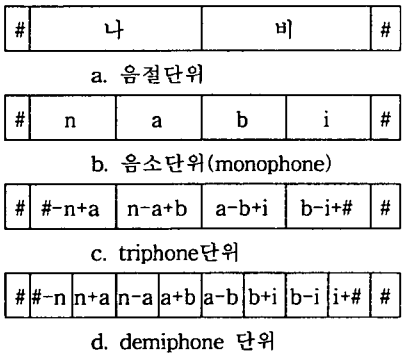

그림1. demiphone의 경계

demiphone은 음소 또는 변이음을 그것의 전후 음소의 영향을 받지 않는 정상상태시점을 중점을 경계로 하여 음소를 양분함으로써 얻어지는 음성단위이다. 이렇게 나누어진 demiphone는 선행음소에 의한 조음효과를 포함하는 left-demiphone과 후행음소에 의한 조음효과를 포함하는 right-demiphone의 두부분으로 나누어 진다. 즉, 음소의 경계와 diphone의 경계를 동시에 가지는 단위라고 할 수 있다.

만약 30개의 음소가 어떠한 제약조건도 없이 출현할 수 있는 상황에서 triphone 단위 모델을 훈련시키려 한다면, 총 21000개(30×30×30)의 triphone을 훈련시켜야 한다. 반면에 demiphone의 경우 left-demiphone 900개(30×30)와 right-demiphone 900개(30×30)를 포함하여 총 1800개의 모델을 훈련시키면 되므로 triphone에 비해서 훈련시킬 모델의 수가 훨씬 감소하게 된다.

또한, [그림1].에서 보는 것과 같이 demiphone은 선행음소의 영향을 받는 left-part와 후행음소의 영향을 받는 right-part로 나누어지기 때문에 전반음소와 후반음소를 적당히 결합하면 음소나 triphone이나 biphone, diphone을 생성할 수 있다.

이와 같이, triphone과 같이 문맥의 조음효과를 잘 표현하면서 triphone보다 모델의 수를 훨씬 줄일수 있어서 메모리를 절감할 수 있다. 또한 데이터 훈련량에 있어서 하나의 음소에서 전반음소와 후반음소를 동시에 훈련시킬 수 있으므로, 같은 양의 훈련데이터에서 biphone모델을 훈련시키는 것에 비해 2배의 훈련 효과를 볼 수 있다.

3. 시스템의 구현

본절에서는 monophone, triphone, demiphone을 비교평가 하기 위하여 구축한 자동 음소분할 레이블링 시스템에 대해서 기술한다.

[그림 2]는 HMM 모델의 일반적인 생성절차이다. HMM 모델을 생성하기 위해서는 먼저 음성을 분할하기 위한 단위와 레이블링 단위를 결정하여야 한다. 그리고, HMM 모델의 형태와 HMM에서 사용할 특징 파라미터를 결정한 후, HMM 모델을 초기화 하게 된다. 초기화된 HMM 모델은 반복적인 훈련과정을 거쳐 자동 음소분할 및 레이블링 시스템의 입력으로 사용될 최종적인 HMM 모델을 생성하게 된다.

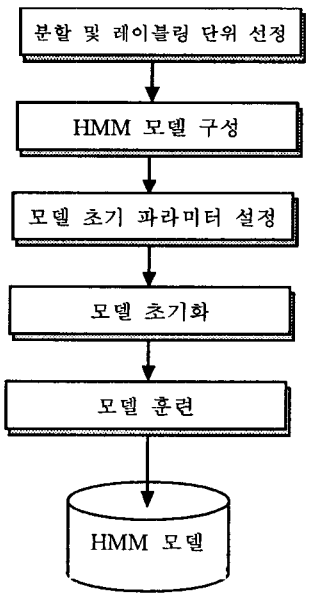

그림 2. HMM 모델 생성 절차

[그림 3]은 HMM 모델을 이용한 자동 음소분할 및 레

이블링 시스템의 일반적인 구성을 보이고 있다.

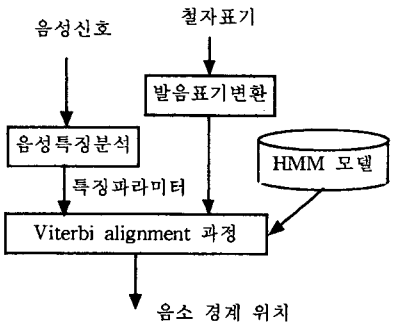

그림 3. HMM을 이용한 자동 음소분할 및 레이블링 시스템의 구성도

3.1 레이블링 단위 선정

구현된 레이블링 시스템에서는 레이블링의 기본단위로 당 연구실에서 "국어 정보베이스"사업의 일환으로 제작한 음소적으로 균형이 잡힌 PBW(Phonetically Balanced Word) 452 단어에 사용된 유사음소를 선정하여, triphone과 demiphone으로 확장하여 사용하였다.

레이블링을 위한 유사음소를 선정하기 위해 먼저 레이블된 PBW 452 단어의 음성 데이터베이스를 분석하였다. 데이터의 분석은 [표 2]에서 나타낸 유사음소만을 대상으로 하였으며, 그 결과는 각각의 유사음소에 대해 음소, 출현 빈도, 평균 지속시간과 표준편차로 구하였다. 거의 출현하지 않거나 전혀 출현하지 않는 음소들은 레이블링 단위에서 제거하였다.

한국어 기본 음소단위를 기준으로 하여, 파열음, 파찰음, 그리고 유음의 폐쇄구간과 파열/마찰 구간으로 나누고, 나뉘어진 각 구간의 유성음화를 고려하였다. 공명음화, 파열음·파찰음에서의 불파음화, 그리고 유음의 탄설음화의 경우 단일 구간으로 취급하였다.

분석을 통해서 선택된 레이블링 단위는 PBW 452어절 음성 데이터베이스에 사용한 유사음소 94개 중 [표 1~4]에 나타낸 바와 같이 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.

최종적으로 선택된 레이블링의 기본 단위는 [표 3]에 나타낸 것처럼 89개의 유사음소와 1개의 묵음을 포함하여 총 90개이다.

	음소	폐쇄구간	폐쇄구간의 유성음화	파열/마찰구간	파열/마찰 구간의 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	ㅣ	i	ㅑ	ja	ㅝ	wv
	ㅓ	v	ㅐ	e	ㅕ	jv	ㅙ	wE
	ㅗ	o	ㅒ	E	ㅛ	jo	ㅟ	wi
	ㅜ	u	ㅢ	Ui	ㅠ	ju	ㅚ, ㅞ	we
	ㅡ	U			ㅖ, ㅒ	je	ㅘ	wa
묵음	C							

표 4. 모음과 묵음의 기호 목록

3.2 음성 분할을 위한 단위 선정

레이블링 단위로 선정된 유사음소의 목록중 monophone단위의 자동 음소분할 시스템은 유음의 폐쇄음화, 마찰음의 유성음화, 파열음의 불파음화와 'ㄱ, ㄷ, ㅂ'을 제외한 파열음 폐쇄구간, 기식구간의 유성음화를

제외한 68개의 유사음소 그리고, 1개의 묵음을 포함하여 총 69개의 음소단위를 사용하였다.

본 논문에서 구현한 triphone과 demiphone단위 자동 음소분할 시스템의 레이블링 단위는 monophone 시스템에서 사용한 68개의 유사음소를 각각 triphone과 demiphone으로 확장하여 구현하였다. 앞에서 기술한 바와 같이 demiphone은 전후반 음소로 나누어진다. 그러나, demiphone단위로 레이블링된 데이터를 가지고 있지 않은 상황이므로 각 유사음소의 중점을 기준으로 나누어서 사용하였다.

triphone과 demiphone의 경우 출현빈도수가 낮은 음소의 경우 훈련량이 충분하지 않아 훈련에 어려움이 있으므로, [표 5.]에서처럼 음소를 17개의 전·후 문맥정보로 클러스트링하여 각각의 단위로 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

3.3 시스템의 훈련

시스템의 훈련에 사용한 음성 데이터베이스는 PBW 452 단어 데이터베이스에서 레이블링된 남성화자 30명분을 사용하였다. 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 10ms단위의 Hamming 윈도우를 5ms 간격으로 이동시키면서 분석하고, 특징 파라미터로는 12차의 MFCC (Mel-frequency cepstrum Coefficient)과 MFCC의 시간축 미분값, 그리고 정규화된 에너지와 그 미분치를 사용하였으며, 각 특징 파라미터들에 가중치를 주고, 각기 독립적인 벡터로 사용하였다. 특징파라미터에 부여한 가중비율은 MFCC와 MFCC의 시간축 미분값, 정규화된 에너지를 각각, 5, 3, 2의 비율로 적용하였다.

음소 모델의 확률분포모델로는 연속확률 분포를 사용하였고, 모델의 형태는 [그림 4.]와 같은 도약경로가 존재하지 않는 5상태7천이를 가지는 left-right 모델로 각 상태당 3개의 mixture를 사용하였다.

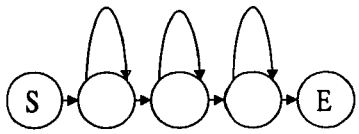

그림 4. 5상태 7천이를 가지는 left-right모델

HMM 모델을 초기화하기 위하여 Viterbi 알고리즘을 이용하여 HMM 모델을 초기화한후 Baum-Welch 알고리즘을 이용하여 초기화된 HMM 모델을 훈련하였다

4. 실험 및 결과

본 논문의 실험에서는 monophone, triphone, demiphone에서 각각의 훈련 및 인식 알고리즘과 특징 파라미터를 동일하게 정한 환경에서 각 단위의 인식률을 비교하였다.

전체적인 실험은 PBW 452 단어의 데이터베이스에서 남자 30명분의 데이터를 훈련에 사용하였으며, 훈련에 사용하지 않은 4명분의 데이터로 레이블링 시스템의 평가를 수행하였다.

철자표기는 ETRI에서 개발한 한국어 발음 표기변환 프로그램을 사용하여 발음표기를 생성한 후, 이를 monophone 또는 triphone, demiphone 단위의 표기로 확장하여 사용하였다.

시스템의 평가를 위해서 언어정보로 철자표기를 사용하였으며, 수작업으로 레이블링한 음소(발성 내용)와 본

시스템에 의해 자동으로 레이블링된 음소가 다른 경우에는 서로 대응하는 음소들을 비교함으로써 평가하였다. 만약 "소프트웨어"라는 단어가 수작업으로 레이블링한 데이터내에는 "S o pV pH u t tH we v"로 레이블링 되어 있고, 자동 레이블러는 "s o p pH U t tH we v"로 레이블링 하였다면, 각각의 대응하는 음소와의 경계를 비교하여 평가한다. 평가된 결과는 수작업으로 레이블한 결과와 비교하여 경계의 위치가 벗어난 정도를 10ms에서 50ms까지 10ms씩 증가시켜가며 경계 인식률을 비교하였다.

평가 결과 [표 6.]과 같이 demiphone을 사용한 경우 경계오차가 20ms 이하의 경우에는 monophone에 비하여 6.31% triphone에 비해 6.21%로 성능 우수하고, 40ms 이하의 경우에는 각각 4.33% 와 3.68%의 성능 향상을 가져왔다.

경계오차 (ms)	monophone	triphone	demiphone
10 이하	60.74 %	62.06 %	67.44 %
20 이하	77.99 %	78.09 %	84.30 %
30 이하	84.05 %	85.34 %	89.98 %
40 이하	88.77 %	89.42 %	93.10 %
50 이하	91.37 %	92.34 %	95.29 %

표 6. 자동음소분할 시스템의 인식률 비교

5. 결론

본 연구에서는 음성 분할 및 레이블링 시스템에서 음성을 분할하기 위해 단위를 평가하기 위하여, monophone, triphone, demiphone을 각각 단위로 하는 음성분할 시스템을 구축한후 각 시스템을 비교평가 하였다. 실험결과 demiphone 단위는 전·후 음소에 의한 상호조음정보를 포함하면서도, 기존의 triphone보다 훈련하기 쉬운 장점을 가진다.

demiphone을 사용한 경우 경계오차가 20ms 이하의 경우에는 monophone에 비하여 6.31% triphone에 비해 6.21%로 성능 우수하고, 40ms 이하의 경우에는 각각 4.33% 와 3.68%의 성능 향상을 가져왔다.

본 논문에서 구현된 demiphone단위의 음소분할 시스템의 경우 평가에 일관성을 기하기 위해 일반적인 모델을 사용하였다. 향후에 demiphone의 특성에 맞는 시스템을 구축할 경우 demiphone의 성능이 더욱 향상될 것으로 기대된다.

< 참 고 문 헌 >

[1] 김태환, 문맥종속 반음소 단위를 이용한 자동 음소 분할 및 레이블링 시스템의 구축, 원광대학교, 컴퓨터공학과 석사논문, 1998

[2] 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.

[3] 김종진, 김봉완, 이용주, 한국어 음성데이터베이스 구축을 위한 한국어 레이블링 기준에 관한 연구, 제 13 회 음성통신 및 신호처리 워크샵 논문집, KSCSP '96 13권 1호, PP. 250-255., 1996.8.

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

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

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

[7] 성종모, 김형순 외 2, 한국 음성 데이터베이스 구축을 위한 반자동 음성분할 및 레이블링 시스템 구현, 제 13 회 음성통신 및 신호처리 워크샵 논문집, KSCSP '96 13권 1호, PP. 161-166., 1996.8.

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

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