시간동기형 Viterbi 알고리즘과 HMM에 기반한 음성의 자동 세그멘테이션

°오세진*, 황철준**, 김범국**, 정호열*, 정현열*

*영남대학교 전자정보공학부, **대구과학대학 정보전자통신계열
osj@speech.yeungnam.ac.kr,{hcj, kbk}@mail.taegu-c.ac.kr, {hoyoul, hychung}@yu.ac.kr

Auto-Segmentation of Unsegmented Speech based on HMM and Time-Synchronous Viterbi Algorithm

°Se-Jin Oh*, Cheol-Jun Hwang**, Bum-Koog Kim**, Ho-Youl Jung*, Hyun-Yeol Chung*

*School of Electrical Eng., & Computer Science, Yeungnam University
**Informational Electronics & Communication Div., Taegu Science College

요 약

본 연구에서는 음성인식에 있어서 음향모델의 고정도화를 위해 통계적 방법인 HMM과 시간 동기형 Viterbi 알고리즘을 기반으로 한 세그멘트되지 않은 음성의 자동 세그멘테이션에 관한 연구를 수행하였다. 본 연구에서는 소량의 세그멘트된 음성에 대해 연속분포형 HMM 기본모델을 작성한 후 이를 표준패턴으로 사용하고, 세그멘트되지 않은 입력음성의 특징 파라미터에 대해 시간동기형 Viterbi 알고리즘의 누적확률이 프레임마다 최대가 되는 지점을 최적경계로 설정하고, 앞에서 구한 최적 경계 정보와 언어학적 지식인 발음사전 정보를 이용하여 음성을 세그멘테이션 하는 것이다. 본 연구와의 비교를 위해 HTK를 이용하여 위와 동일한 과정을 수행하였다. 이렇게 구한 음성의 세그멘테이션 정보를 이용하여 연속분포형 HMM 기본모델과 HTK의 CHMM 기본모델을 각각 작성한 후, 국어공학센터(KLE) 단어 데이터에 대해 단어인식 성능을 평가하였다. 실험결과, KLE 452 남성과 여성에 대해, 본 연구실 인식 시스템은 화자독립 단어인식률 89.4%, 85.1%, HTK의 화자독립 단어인식률 85.1%, 81.9%를 각각 얻었다.

1. 서 론

대어휘 연속음성인식을 위해서는 보다 정밀한 음향모델과 연속음성에 적용할 수 있는 언어모델의 개발, 그리고 운율정보의 사용과 잡음처리 등이 중요한 관점이라 할 수 있다. 그 중에서도 음향모델의 정도를 향상시키기 위해서는 많은 양의 데이터베이스를 수집하는 것만으로는 충분하지 않으며, 음성데이터를 음소와 같이 기본단위로 분할하고 세그멘테이션을 수행하여 통계적 처리가 가능하도록 만들어 주는 작업이 필수적으로 요구된다[1].

수집한 음성의 세그멘테이션 작업은 사람에 의해 수작업으로 직접 수행할 수 있지만, 수작업에 의해 음성의 세그멘테이션을 수행할 경우 다음과 같은 문제점을 지닌다. 우선, 음성학적 지식이 풍부한 소수의 전문가에 의존할 수밖에 없다는 점과 스펙트로그램 판독 및 청취평가가 반복적으로 이루어지기 때문에 매우 지루한 작업일 뿐만 아니라 많은 시간이 소요된다는 점을 들 수 있으며, 마지막으로 음성 경계를 결정하기 위한 구체적인 판단기준을 미리 정해 두더라도 거의 대부분이 주관적인 판단을 피할 수 없으며, 이로 인해 음성 경계를 결정하는 과정에서 일관성이 보장되지 못한다는 점을 들 수 있다[2]. 따라서 서로 다른 전문 음성학자들이 동일한 음성을 분할할 경우는 물론이고, 동일한 사람이 동일한 음성을 분할하더라도 추출된 음성 경계에는 차이가 생기게 된다. 또한 지루한 작업이 계속됨에 따른 판단의 오류도 발생하게 된다. 이러한 문제들은 음성의 세그멘테이션이 자동으로 수행될 수 있다면 어느 정도 해결될 수 있으며, 수작업에 의한 지루한 작업과 시간과 비용을 많이 줄일 수 있을 것이다.

따라서 본 연구에서는 통계적 접근방법으로 소량의 세그멘트된 음성에 대해 연속분포형 HMM 기본모델을 작성한 후 이를 표준패턴으로 사용하고, 세그멘트되지 않은 입력음성의 특징 파라미터에 대해 시간동기형 Viterbi 알고리즘[7]의 누적확률이 프레임마다 최대가 되는 지점을 최적경계로 설정하고, 앞에서 구한 최적 경계 정보와 언어학적 지식에 기반한 음소표기식의 발음사전 정보를 이용하여 음성의 자동 세그멘테이션을 수행하였다.

이를 위해 세그멘트된 한국전자통신연구원(ETRI)의 445 단어 데이터베이스로 작성한 CHMM 기본모델과 발음사전을 이용하여, 세그먼트 되지 않은 국어공학센터(KLE)의 452 단어 음성 데이터베이스에 대해 자동 세그멘테이션을 수행하고, 세그먼트 정보의 정확도를 확인하기 위해 단어음성인식 실험을 수행한 후, 그 유효성을 검토하고자 한다.

2. 음성 데이터베이스

2.1 음성 데이터 및 전처리

표 1에 나타낸 것과 같이 음성 데이터베이스는 한국전자통신연구원의 445 단어 데이터베이스 중 14명과 국어공학센터의 남자와 여자의 452 단어 데이터베이스를 이용하였다.

사용한 음성 데이터는 모두 방음부스에 녹음되었으며, 16khz로 샘플링하고 16bit로 양자화 되었다. 본 연구실의 전처리 과정에서는 16ms 해밍 윈도우를 곱하여 5ms 단위로 중첩하면서 14차의 LPC 분석을 통하여 10차의 멜-켑스트럼 계수를 추출하고, 1차 차분 성분인 10차의 회귀계수를 음성특징 파라미터로 사용한다[5]. 또한 HTK[6]를 이용할 경우에는 25ms 해밍 윈도우를 곱하여 10ms 단위로 중첩하면서 에너지 성분을 제외한 12차의 MFCC와 1차 차분 성분인 13차의 MFCC, 2차 차분 성분인 13차의 MFCC로서 총 38차원의 특징 파라미터를 사용한다. 사용된 특징 파라미터의 분석조건을 표 2에 나타내었다.

[표 1] 음성 데이터베이스

화자	데이터 명	발성	세그먼트 정보	사 용
14명	ETRI445	1회	수작업, frame 단위	기본모델 학습
38명	KLE452 A	2회	unsegmented	segmentation
32명	KLE452 B	2회	unsegmented	segmentation

[표 2] 음성의 분석조건

조 건	본 연구실	HTK
해밍윈도우	16ms	25ms
프레임주기	5ms	10ms
특징 파라미터	10차 MFCC + △ 10차 MFCC = 20차원	12차 MFCC+ △ 13차 MFCC+ △△ 13차 MFCC = 38차원

[표 3] 48개 음소기호의 정의

모 음	aa /아/	axr /어/	ao /오/	uh /우/
	U /으/	ih /이/	ae /애/	eh /에/
	ja /야/	jv /여/	jo /요/	ju /유/
	wa /와/	wv /워/	wE/외/	we/웨,왜/
	wi /위/	je /예/	jE /얘/	Wi /의/
자 음	b~ /ㅂ/	d~ /ㄷ/	g~ /ㄱ/	z~ /ㅈ/
	bb /ㅃ/	dd /ㄸ/	gg /ㄲ/	zz /ㅉ/
	p /ㅍ/	t /ㅌ/	k /ㅋ/	ch /ㅊ/
	s /ㅅ/	ss /ㅆ/	hh~/ㅎ/	r /ㄹ/
	n /ㄴ/	m /ㅁ/		
첫음절	b /ㅂ/	d /ㄷ/	g /ㄱ/	z /ㅈ/
	hh /ㅎ/			
종 성	b+ /ㅂ/	d+ /ㄷ/	g+ /ㄱ/	l /ㄹ/
묵 음	SIL			

2.2 발음사전

자동 세그멘테이션을 위해 제공되는 언어학적 정보는 음소표기식(phonetic transcription)과 철자표기식(orthographic transcription)으로 나눌 수 있다[3]. 음소표기식이 제공되는 경우는 음소 세그멘테이션이 이미 이루어진 상황이므로 음소사이의 묵음구간 검출과 더불어 각각의 음소들에 대한 경계위치만 찾아내면 된다. 하지만 철자표기식이 제공되는 경우에는 발음되는 형태의 음소표기식으로 자동 변환해 주는 과정에 필요하게 되며, 동일한 단어라도 다양한 형태로 발음될 수 있기 때문에, 경우에 따라 여러 개의 발음사전을 사용하게 된다.

본 연구에서는 시간과 비용을 감소시키기 위해, 일반적으로 사람들이 발성할 때 나타날 수 있는 국어 음성학적 지식을 부가하여 발성 리스트의 발음사전을 음소표기식에 기반하여 한가지로 제한하였으며, 사전지식을 획득하기 위해 한국어의 음운학적 특징을 부가시킨 KPS tool을 이용하였다[4].

본 연구에서 언어학적 지식으로 사용되는 발음사전은 48개의 음소기호로 정의한 유사음소단위(PLUs)를 이용하였으며, 이를 표 3에 나타내었다.

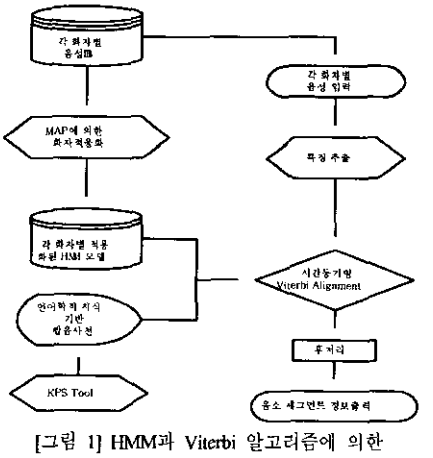
3. 자동 세그멘테이션

3.1 연속분포형 HMM

본 연구에서 본 연구실의 기본모델과 HTK의 기본모델은 표 3에 나타낸 48개의 유사음소단위로 모노폰(monophone) 형태의 4상태 3출력의 left-right형 연속분포 HMM을 사용하였다. 또한 수작업에 의해 세그먼트 된 음성과 세그먼트 되지 않은 음성사이의 발성환경, 발성형태, 화자 특성 등에 다른 상관성이 저하되기 쉬운 점을 고려하여 보다 효율적인 음성의 세그먼트 정보를 획득하기 위해 ETRI 445로 작성한 기본모델에 세그먼트를 수행할 화자의 음성 데이터를 최대사후확률추정법(MAP)[5]으로 화자적응화를 통하여 적응화된 화자의 모델에 대해 자동 세그먼트를 수행하였다.

3.2 시간동기형 Viterbi 알고리즘

그림 1은 언어학적 정보가 제공되는 경우의 HMM 방식에 의한 자동 세그멘테이션 시스템을 나타낸 것이다. 음성신호가 들어오면 음성특징 분석과정에서 음성특징 계수를 추출한다. 그 다음 세그먼트 된 음성으로 작성한 HMM 기본모델과 음소표기에 따라 연결한 사전정보에 따라 입력 음성의 특징 파라미터와 시간 동기형 Viterbi 알고리즘[7]에 의해 누적확률이 최대가 되는 경계를 저장한 후 백-트레킹 과정에 의해 각각의 음소간의 최적 경계 정보가 얻어진다. 본 연구에서 사용된 음소표기는 국어 음성학적 특징을 부여하여 자동으로 획득한 정보를 이용하였다. 이 시스템의 출력은 단어번호, 각 단어의 음소기호와 프레임 단위의 최적 경계 정보이며, 출력된 결과는 수작업에 의해 후처리 작업을 수행할 수 있다. 이와 같이 언어학적인 정보로 음소표기가 주어진 경우, 미지의 음소열을 찾아내는 음성인식보다는 용이하다고 할 수 있다. 또한 대용량 음성 데이터베이스를 이용한 음성인식 시스템을 구축할 때 시간 및 비용 절감에 보다 효과적이라고 할 수 있다.



[그림 1] HMM과 Viterbi 알고리즘에 의한 자동 세그먼트 시스템의 구성도

4. 실험 및 고찰

자동 세그멘테이션 시스템에서 출력한 세그먼트 정보의 정확도를 평가하기 위하여 단어인식실험을 수행하여 그 유효성을 검토하고자 한다. 기본 인식 시스템은 CHMM 기반의 본 연구실 음성인식기[5]와 HTK[6]를 이용하였다.

자동 세그먼트 된 국어공학센터의 남성과 여성 화자의 단어음성 발성 중, 남성의 경우 35명 1, 2회를, 여성의 경우 29명의 1, 2회를 대상으로 본 연구실의 HMM 학습기를 이용하여 20차원의 4상태 3출력 1혼합수를 가지는 모노폰 형태의 CHMM 초기 음향모델을 각각 학습하였다. 또한 HTK 학습기를 사용하여 위와 동일한 화자수를 대상으로 38차원의 5상태 3출력 1혼합수를 가지는 모노폰 형태의 CHMM 초기 음향모델을 각각 학습하였다.

성능평가를 위해 사용된 단어는 학습에 참가하지 않은 남성 3명과 여성 3명의 1, 2회를 이용하였다. 인식방법은 시간동기형 One-Pass DP

방법을 이용하였으며, 인식에 사용된 문법은 유한상태 네트워크(FSN) 형태의 문맥자유문법(CFG)이며 어휘 수는 452개이다.

우선, 본 연구실 인식기(A)를 이용한 인식실험 결과로서 표 4와 5에 남성과 여성에 대한 학습률과 화자독립 단어인식률을 각각 나타내었다. 그리고 HTK(B)를 이용한 인식실험 결과로서 표 6과 7에 남성과 여성에 대한 학습률과 화자독립 단어인식률을 각각 나타내었다.

[표 4] A 인식기를 이용한 화자 학습률

성별	20차, mix1	발성회수	학습률(%)
남성	35명	1	89.3
		2	89.6
여성	29명	1	89.9
		2	89.0

[표 5] A 인식기를 이용한 화자독립 인식률

성별	20차, mix1	화 자			평균(%)
남성	1회	A1	B1	C1	
		91.6	85.6	90.9	89.4
	2회	A2	B2	C2	
		92.9	85.4	89.8	89.4
여성	1회	D1	E1	F1	
		87.6	77.4	92.0	85.7
	2회	D2	E2	F2	
		87.2	74.3	92.0	84.5

[표 6] B 인식기를 이용한 화자 학습률

성별	38차, mix1	발성회수	학습률(%)
남성	35명	1	84.5
		2	84.9
여성	29명	1	86.2
		2	85.3

[표 7] B 인식기를 이용한 화자독립 인식률

성별	38차, mix1	화 자			평균(%)
남성	1회	A1	B1	C1	
		88.5	80.6	86.1	85.1
	2회	A2	B2	C2	
		87.2	80.5	87.4	85.0
여성	1회	D1	E1	F1	
		83.4	76.3	88.3	82.7
	2회	D2	E2	F2	
		81.4	74.1	87.4	81.0

[표 8] HTK에 의한 세그먼트 정보를 이용한 A 인식기에 의한 화자독립인식률

성별	20차, mix1	화 자			평균(%)
남성	1회	A1	B1	C1	
		92.9	85.4	91.6	90.0
	2회	A2	B2	C2	
		92.7	86.1	92.0	90.3
여성	1회	D1	E1	F1	
		88.3	78.3	92.7	86.4
	2회	D2	E2	F2	
		86.7	75.2	92.5	84.8

또한 국어공학센터 남성과 여성의 단어음성을 대상으로 HTK 학습기에 의해 혼합수 9의 CHMM 기본모델을 작성한 후 forced alignment[6]를 수행한 후 얻어진 음성의 세그먼트 정보를 본 연구실의 HMM 학습기로 이상의 방법과 같이 학습한 후 평가한 결과를 표 8에 나타내었다.

표 4, 5, 6, 7에서 본 연구실의 자동 세그멘테이션 시스템이 HTK와 비교하여 음성의 분석조건 및 특징 파라미터의 차원 등에 따라 보다 음성의 세그먼트 정보가 우수함을 확인할 수 있었다. 하지만 표 8에서 HTK를 이용하여 음성을 자동 세그멘트할 경우 혼합수가 증가할수록 음성의 세그먼트 정보가 본 연구실의 시스템보다 세그먼트 정보가 조금 향상됨을 확인할 수 있었다. 따라서 향후 본 연구실의 HMM 학습기에서도 혼합수를 증가시킬 경우 성능을 향상시킬 수 있을 것으로 기대된다.

이상의 결과로부터 본 연구실에서 구축한 시간 동기형 Viterbi 알고리즘과 연속분포 HMM을 이용한 음성의 자동 세그멘테이션 방법의 유효성을 확인할 수 있었다.

5. 결 론

본 연구에서는 음성인식에 있어서 음향모델의 고정도화를 위해 통계적 방법인 HMM과 시간 동기형 Viterbi 알고리즘을 기반으로 한 세그멘트되지 않은 음성의 자동 세그멘테이션에 관한 연구를 수행하였다. 소량의 세그멘트된 음성에 대해 연속분포형 HMM 기본모델을 작성한 후 이를 표준패턴으로 사용하고, 세그멘트되지 않은 입력음성의 특징 파라미터에 대해 시간동기형 Viterbi 알고리즘의 누적확률이 프레임마다 최대가 되는 지점을 최적경계로 설정하고, 앞에서 구한 최적경계 정보와 언어학적 정보인 발음사전 정보를 이용하여 음성의 자동 세그멘테이션을 수행하였다. 본 연구와의 비교를 위해 HTK를 이용하여 위와 동일한 연구를 수행한 결과, 본 연구에서 이용한 방법의 유효성을 확인하였으며, 혼합수의 증가에 따라 정밀한 음향모델의 사용이 요구됨을 확인하였다. 이상의 결과로부터 다양한 분야의 음성인식을 적용할 경우 시간과 비용을 많이 줄일 수 있을 것으로 기대된다.

향후에는, 화자의 특성에 따른 다양한 음소표기에 의한 발음사전과 triphone과 같은 보다 정밀한 음향모델을 사용하여 연구를 계속 수행할 계획이다.

6. 참고문헌

[1] 성종모, 김형순, "자동 음성분할 및 레이블링 시스템의 구현," 한국음향학회지 16권 5호, pp. 50-59, 1997.

[2] B. Eisen, H. Tillmann, C. Draxler, "Consistency of judgements in manual labeling of phonetic segments: the distribution between clear and unclear cases," Proc. of ICSLP'92, pp. 871-874, 1992.

[3] B. Wheatley, G. Doddington, C. Hemphill, J. Godfrey, "Robust automatic time alignment of orthographic transcription with unconstrained speech," Proc. of ICASSP'92, Vol. 1, pp. 553-556, 1992.

[4] 이상호, 오영환, 서정연, "한국어 문서 음성변환 시스템을 위한 문서 분석기," 한국음향학회지 15권 3호, pp. 50-59, 1996.

[5] 김득수, 황철준, 정현열, "음성인식 기능을 가진 주소입력 시스템의 개발과 평가," 한국음향학회지 18권 2호, pp. 3-10, 1999.

[6] S. Young, "HTK Book," 1999.

[7] Rabiner, Juang, "Fundamentals of speech recognition," Prentice-Hall Int'l, Inc, 1993.