대용량 음성 D/B 구축을 위한 AUTO-SEGMENTATION에 관한 연구

이병순*, 장성욱*, 양성일*, 권영헌**
*한양대학교 전자전기제어계측공학부 , **한양대학교 물리학과

A study on Auto-Segmentation Improvement for a Large Speech DB

Byong-soon Lee*, Sungwook Chang*, Sung-il Yang*, Y. Kwon**
*School of Electrical and Computer Engineering,
**Department of Physics Hanyang University
E-mail: aleph@netian.com

요약

본 논문은 음성인식에 필요한 대용량 음성 D/B 구축을 위한 auto-segmentation의 향상에 관한 논문이다. 50개의 우리말 음소(잡음, 묵음 포함)를 정하고 음성특징으로 MFCC(Mel Frequency Cepstral Coefficients), ΔMFCC, ΔΔMFCC, 39차를 추출한 다음 HMM 훈련과 CCS(Constrained Clustering Segmentation) 알고리즘[1]을 사용하여auto-segmentation을 수행하였다. 이 과정에서 대부분의 음소는 오류범위(±25ms) 안에서 분절이 이루어지지만, 짧은 묵음, 모음+유성자음('ㅁ', 'ㄴ', 'ㄹ', 'ㅇ') 등에서 자주 오류범위를 넘어 분절이 발생하였다. 이러한 음운환경에 따른 경계의 오류를 구간별로 Wavelet 변환 신호의 MLR(Maximum Likelihood Ratio) 값을 이용, 기존 문제점을 보완하여 오류의 범위를 줄임으로서 auto-segmentation의 성능 향상을 얻을 수 있었다.

1. 서론

음소 단위 분절은 인식과 합성에 커다란 영향을 끼치므로 정확한 segmentation이 필요하다. 그 방법으로는 수작업을 이용한 hand-segmentation과 자동 분절을 이용한 auto-segmentation이 있다.

Hand-segmentation의 경우 정확성은 있지만 소수 음성 전문가에 의존할 수밖에 없고, 매우 긴 시간이 소요되며, 구체적인 판단기준을 미리 정해 놓더라도 상당 부분 주관적 판단에 의존하므로 일관성의 유지가 어렵다는 단점 때문에 대용량 음성 D/B를 구축할 경우 auto-segmentation의 사용은 거의 필수적이다.

본 논문에서는 한국어 연속음성 인식을 위해 초성, 모음, 종성, 잡음, 묵음 등으로 나눈 50개의 음소로 지정하고 MFCC를 이용하여 특징파라미터를 추출한 후 연속음 인식에 적합한 HMM 모델을 이용하여 segmentation을 하였다. 그 결과 대부분의 음소에서 25ms 오차 범위 내에서 segmentation이 이루어졌지만, 몇 가지 문제가 발생했다. 그중, 첫째로 짧은 묵음구간을 탈락시키는 오류가 많이 발생했고, 둘째로는 음운특성이 비슷한 모음+유성자음('ㅁ', 'ㄴ', 'ㅇ' ,'ㄹ')에서 오류가 많이 발생했다.

이러한 점을 보완하기 위해 시간 및 주파수 영역에서 동시에 음성신호의 국부적 특성을 잘 반영하는 웨이블렛 변환을 이용하여, 변환된 신호의 스펙트럼상의 변

화 검출에 적합한 MLR 값과 무성음 검출에 용이한 LCR(Level Crossing Rate)을 이용 묵음구간을 검출하여 HMM으로 segmentation 하기 전, 미리 연속음으로 이루어진 문장을 나눠줌으로써 짧은 묵음 구간에 따른 오류를 방지하였다. 모음+유성자음의 검출 오류에 대해서는 모음과 유성자음의 각 음소에 따른 MLR값의 특징을 이용, 변화가 크게 발생하는 부분을 segmentation 부분으로 다시 지정해 주었다.

본 논문의 구성은 다음과 같다. 2절에서는 특징 파라미터인 MFCC의 추출 및 HMM 훈련에 대해 설명하고 3절에서는 Segmentation 방법에 대해 설명하며, 4절에서 실험과 결과에 대해 서술하였고, 5절에서 결론을 맺었다.

2. 특징 파라미터 추출 및 HMM훈련

2.1 MFCC 파라미터 추출

특징파라미터 추출은 입력으로 얻어진 음성 데이터로부터 인식에 필요한 특징을 뽑아 내는 과정인데 여기서는 비교적 잡음에 강인하다고 알려진 MFCC을 사용하였다. <그림 1>은 MFCC가 생성되는 과정이다..

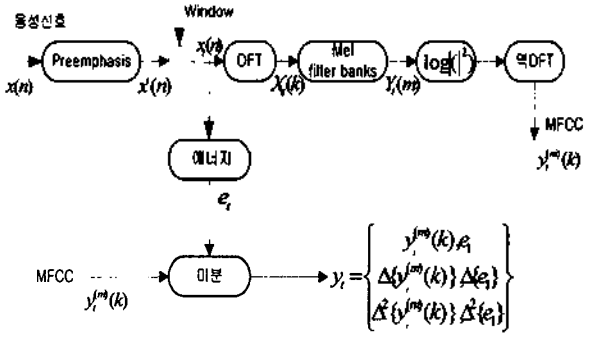

그림 1. MFCC 연산의 개략도

먼저 고주파 영역의 formant는 저주파 영역의 formant보다 매우 작은 크기를 갖는다. 따라서 모든 주파수 대역에서 유사한 크기의 formant를 갖게 하기 위해 preemphasis과정이 필요하다. 이와 같은 처리는 다음과 같은 1차의 FIR 필터로 간단히 처리할 수 있다.

$$H(Z)=1-a\cdot z^{-1} \quad 0\le a\le 1 \tag{1}$$

$$x'(n)=x(n)-ax(n-1) \tag{2}$$

다음으로 windowing과정이다. 음성 신호의 주파수 영역 분석에 있어서 신호의 stationary는 필수적인 요소이다. 이 과정을 위해 음성이 stationary하다고 가정할 수 있는 매우 짧은 시간동안의 신호만을 분석할 수 있도록 음성 신호 x'(n)을 프레임이라 불리우는 연속된 windowed sequences $x_t(n)$로 나누게 된다.

$$x_t(n)\equiv x'(n-t\cdot Q),\ 0\le n<N \quad 1\le t\le T \tag{3}$$

$$x_t(n)\equiv w(n)\cdot x_t(n) \tag{4}$$

이렇게 구한 음성 신호 $x_t(n)$의 성도의 특성은 DFT의 크기의 제곱 즉, $|X_t(k)|^2$으로 쉽게 추정할 수 있으며 인간의 청각구조에 근접한 24개의 band-pass filters로 구성된 Mel scale filter를 이용 차수만큼의 대역별 에너지를 얻은 다음 계수들의 크기의 제곱의 logarithm을 계산한다. 마지막으로 식(5)에 의해 역 DFT를 취해주면 MFCC를 구할수 있다.

$$y_t^{(m)}(k)=\sum_{m=1}^{M}\log\{|Y_t|\}\cos\left(k\left(m-\frac{1}{2}\right)\frac{\pi}{M}\right),\quad k=0,1,\cdots,L \tag{5}$$

실험에서 사용한 특징 파라미터는 12차 MFCC와 에너지(1차)에 음성의 시변 특성을 고려하기 위한 1차 또는 2차의 계수간의 차이 값을 포함시켜서 모두 39(13*3)차를 사용하였다.

$$\Delta^i\{u_t\}=\Delta^{i-1}\{u_{t+1}\}-\Delta^{i-1}\{u_{t-1}\},\quad \Delta^0\{u_t\}=u_t \tag{6}$$

2.2 HMM 훈련

HMM에서의 확률 추정은 전향 절차, 후향 절차에 의해서 이루어지고 관측열이 j 상태에 있을 때 관측열 y를 발견할 확률 $b_j(y)$는 식(7)과 같은 확률 밀도 함수로 구할 수 있다. 여기서 μ_j는 mean 벡터이고 U_j는 covariance 행렬이다.

$$b_j(y)=\frac{1}{\sqrt{(2\pi)^D\det U_j}}\exp\left\{-\frac{1}{2}(y-\mu_j)^T U_j(y-\mu_j)\right\} \quad j=1,\ldots,N \tag{7}$$

하지만 식(7)의 경우는 다중화자 시스템에는 부적합하므로 다음 식(8)과 같이 M mixture의 확률밀도 함수를 사용한다.

$$b_j(y)=\sum_{m=1}^{M}c_j^m b_j^m(y) \quad \sum_{m=1}^{M}c_j^m=1 \quad c_j^m\le 1 \tag{8}$$

HMM의 훈련은 이전의 음성 모델 Θ가 있을 때 주어진 관측 벡터 Y와 해당 상태 S의 더욱 최적화된 likelihood값 $P(Y|S,\Theta)$을 찾는 과정이다. 이를 위해 <그림 2>의 Baum-Welch 알고리즘은 각 HMM의 파라미터들의 likelihood를 최대화시키는 과정을 수행한다.

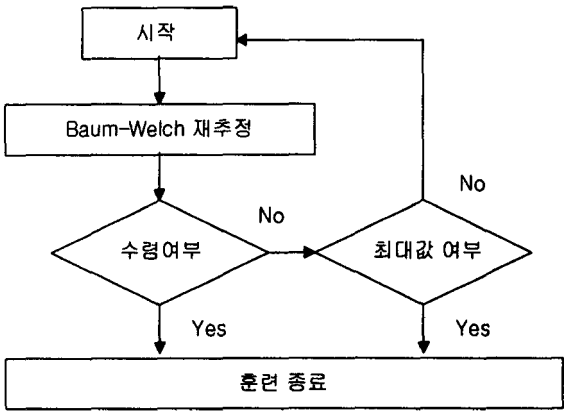

그림 2. Baum-Welch 재추정법

3. Segmentation 알고리즘

3.1 HMM에서의 CCS 알고리즘

Segmentation 알고리즘은 stationary 상태들과 음소들(phonemes)의 경계를 추정하는 알고리즘이다. 흔히 사용되고 있는 알고리즘으로는 Constrained Clustering Segmentation(CCS) 알고리즘이 있다. 이 알고리즘을 이용하여 음성 신호는 각 section안에서의 cepstral의 변화는 무시할 수 있을 정도로 여러개의 section들로 나누어진다. 각 section들은 각 section에 해당하는 관측열들의 mean 벡터 m(generalized centroid)으로 표현된다. CSS 알고리즘은 distortion measure를 최소화하는 알고리즘으로 식(9)와 같이 정의된다.

$$\|y_j - m\|^2 = [y_j - m]^T \cdot [y_j - m] \qquad (9)$$

3.2 Wavelet변환 신호의 MLR값을 이용한 알고리즘

Wavelet은 시간 및 주파수 영역에서 동시에 음성신호의 국부적인 특성을 잘 반영하므로 음성 신호의 분석에 많이 이용되고 있다.[2] 본 논문에서는 Coiflet 24차를 이용하여 Wavelet으로 변환된 신호를 스펙트럼상의 변화 검출에 적합한 MLR 값으로 구해서 앞에서 제기한 HMM으로 segmentation 했을 경우 나타나는 문제점을 보완하려한다.

먼저, 음성을 프레임별로 나누어 식(10)의 MLR을 구한다. 식(10)에서 σ^2와 σ^2_{noise}는 각각 해당 프레임에 대하여 웨이블렛 변환한 신호의 분산과 묵음구간의 분산을 나타낸 것이다.

$$MLR = \frac{1}{2}\left| \ln\left(\frac{\sigma^2}{\sigma^2_{noise}}\right) - \frac{\sigma^2}{\sigma^2_{noise}} \right| \qquad (10)$$

MLR은 특성상 무성음과 묵음이 판별되지 않으므로 무성음의 판별에 용이한 ZCR(Zero Crossing Rate)을 같이 적용하여야 한다. 여기서는 잡음의 특성을 고려해 LCR을 사용하였다.[3]

$$L_n = \frac{1}{2}\sum_{m=0}^{N-1} |sgn[x(n-m) - TH] - sgn[x(n-m-1) - TH]| \qquad (11)$$

$$sgn[s(n)] = \begin{cases} 1, s(n) \geq 0 \\ -1, otherwise \end{cases}$$

묵음구간 검출 조건: [프레임 크기=100 sample]
1. MLR $< \lambda_1$: 연속되는 2개의 프레임
2. LCR $< \lambda_2$: 연속되는 2개의 프레임

λ_1, λ_2 는 실험에 의한 값으로 정해지는 문턱 값이다.

모음과 유성자음의 경우 서로 음성적인 특징이 비슷하므로 hand-segmentation으로도 정확히 나누기가 쉽지 않다. 여기서 제안하는 방법은 MLR값의 크기 변화에 의한 분류이다. 이 방법으로 모든 음소를 분류할 수는 없지만 특정한 음소(모음과 유성자음)에서의 규칙성을 가진다. 대부분 모음과 유성자음의 사이, valley에서 분절이 일어나며, 만약 이 부분이 없다면 모음의 peak에서 분절이 일어난다. <그림 3>은 HMM과 MLR 값을 이용 segmentation한 것을 비교한 것이다.

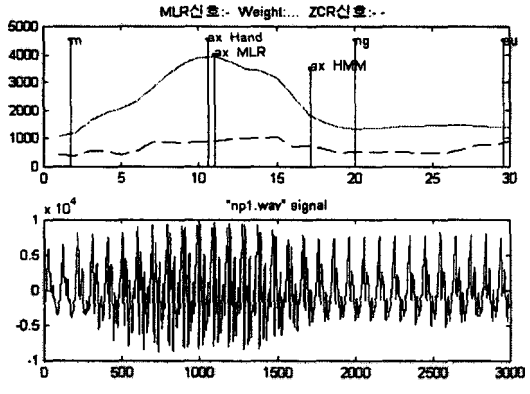

그림 3. MLR 값의 특성에 따른 모음+유성자음의 분절

각 음소가 표기된 부분은 해당 음소의 끝을 나타낸다. 망(m+ax+ng)이라는 발음에서 모음(ax)과 유성자음(ng)의 segmentation을 살펴 보았을 때 모음(ax)이 나눠지는 부분의 sample은, hand-segmentation로 나눈것은 1060이지만 HMM으로 segmentation한 경우 1720에서 나눠졌다. 하지만 MLR 값 변화의 특성을 이용하면 1100에서 나누어지므로 오류를 수정할 수 있다.

4. 실험 및 결과

실험에 사용한 음성데이터는 남성 화자가 신문내용을 낭독한 것으로 5737개의 음소를 가지고 실험했다.

음성 녹음은 16kHz sampling rate, 16bit resolution으로 실험실 환경에서 녹음했으며, 특징파라미트 추출을 위한 window size는 32ms, overlap은 22ms로 정해주었다. <표 1>은 앞에서 언급한 50음소의 분류이다.

표 1. 50개로 분류한 음소 기호

분류	기호	음소	분류	기호	음소
파열음	g	ㄱ 초성	유성자음	n	ㄴ
	gx	ㄱ 종성		m	ㅁ
	g+	ㄱ 유음사이		ng	ㅇ 종성
	G	ㄲ	모음	a	ㅏ 장음
	k	ㅋ		ax	ㅏ 단음
	d	ㄷ 초성		ya	ㅑ
	dx	ㄷ 종성		eo	ㅓ
	d+	ㄷ 유음사이		yeo	ㅕ
	D	ㄸ		o	ㅗ 장음
	t	ㅌ		ox	ㅗ 단음
	b	ㅂ 초성		yo	ㅛ
	bx	ㅂ 종성		u	ㅜ
	b+	ㅂ 유음사이		yu	ㅠ
	B	ㅃ		eu	ㅡ
	p	ㅍ		i	ㅣ 장음
마찰음	s	ㅅ		ix	ㅣ 단음
	S	ㅆ		eui	ㅢ
	h	ㅎ 초성		wa	ㅘ
	h+	ㅎ 유음사이		wi	ㅟ
파찰음	z	ㅈ		weo	ㅝ
	Z	ㅉ		e	ㅐ,ㅔ
	c	ㅊ		ye	ㅒ,ㅖ
유성자음	r	ㄹ 초성		we	ㅚ,ㅙ,ㅞ
	rx	ㄹ 종성	묵음	sil	묵음
	r+	ㄹ 유음사이	잡음	#	잡음

Hand-segmentation한 데이터의 부족으로 훈련에 참가한 데이터를 실험에 같이 참가시켰다.

실험은 MLR과 LCR을 이용 묵음 구간을 검출한 다음, HMM으로 segmentation을 실행한 후 MLR값의 변화를 가지고 모음과 유성자음구간을 다시 수정해 주었다. <표 2>는 오차범위에 따른 auto-segmentation의 오류를 나타낸 것이다. 15~25ms범위 내에는 파열음, 모음, 파찰음의 오류가 많이 나타났는데 보통 음소의 길이가 짧은 발음에서 많이 나타났다. 25ms범위를 벗어난 것은 모음과 유성자음에서 많이 나타났다.

표 2. auto-segmentation의 오차범위에 따른 분포 (%)

분류	음소수	HMM을 이용한 auto-segmentation		MLR로 수정한 auto-segmentation	
		15~25ms	25ms이상	15~25ms	25ms이상
파열음	1004	82(8.2)	27(2.7)	82(8.2)	27(2.7)
마찰음	313	9(2.9)	5(1.6)	9(2.9)	5(1.6)
파찰음	279	14(5.0)	2(0.7)	14(5.0)	2(0.7)
유성자음	1171	46(3.9)	40(3.4)	46(3.9)	40(3.4)
모음	2261	128(5.7)	77(3.4)	93(4.1)	52(2.3)
묵음	631	18(2.9)	9(1.4)	8(1.3)	4(0.6)
잡음	78	2(0.03)	1(1.3)	2(0.03)	1(1.3)
합계	5,737	299(5.2)	161(2.8)	254(4.4)	131(2.3)

5. 결론

본 논문에서는 대용량 음성 D/B를 구축하기 위한 auto-segmentation을 HMM을 통해 구현하고 그 결과에 따른 문제점을 보완하기 위해 wavelet으로 변환된 신호의 MLR값을 이용하여 오차구간 15~25ms, 25ms이상에서 오차를 0.8%, 0.5% 줄였다. 실험 과정에서 hand-segmentation의 label 표기 오류를 많이 수정해 주었는데 이 같은 오류는 auto-segmentation에 치명적인 영향을 미쳤다. 앞으로 보다 많은 음성 데이터를 실험하고 segmentation한 결과를 음성 인식에 직접 적용해서 분석해 볼 필요가 있다.

6. 참고 문헌

[1] Claudio Becchetti and Lucio Prina Ricotti, "Speech Recognition" John Wiley & Sons, 1999

[2] C. S. Burrus, R. A. Gopinath, H. Guo, *Introduction to Wavelets and Wavelet Transforms*, Prentice Hall, 1998.

[3] Lawrence Rabiner Biing-Hwang Juang, "Fundamentals of Speech Recognition", 1993

[4] 박은영, 김상훈, "합성단위 자동생성을 위한 자동 음소 분할기 후처리에 대한 연구", 1998

[5] E. Wesfreid, V. Wickerhauser, "Vocal command signal segmentation and phonemes classification", 1999.