화자확인 시스템을 위한 분절 알고리즘

A Blind Segmentation Algorithm for Speaker Verification System

김 지 운*, 김 유 진*, 민 홍 기**, 정 재 호*
(Ji Un Kim*, Yu Jin Kim*, Hong Ki Min**, Jae Ho Chung*)

※ 본 논문은 99년도 인하대학교 연구비 지원에 의하여 수행되었습니다.

요 약

본 논문에서는 하위단어에 기반한 전화선 채널에서의 어구 종속 화자 확인 시스템을 위한 음성 분할 알고리즘인, 파라미트릭 필터링 에 기반한 델타 에너지를 제안한다. 제안한 알고리즘은 특정 밴드의 주파수를 기준으로 대역폭을 변화시키며 필터링한 후 델타 에너지를 이용하는 방법으로 다른 알고리즘에 비해 주변환경에 강인한 것으로 나타났다. 이를 이용해 음성을 하위단어로 분할하고, 각 하위단어를 이용해 화자의 성문을 모델링하였다. 제안한 알고리즘의 성능 평가를 위해 EER(Equal Error Rate)를 사용한다. 그 결과 단일 모델의 EER이 약 6.1%, 하위 단어 모델의 EER이 약 4.0%로 본 논문에서 제안한 알고리즘을 사용했을 때 약 2%의 성능이 향상되었다.
핵심용어: 화자 확인 시스템, 분절 알고리즘, 델타 에너지, 전화선상

ABSTRACT

This paper proposes a delta energy method based on Parameter Filtering(PF), which is a speech segmentation algorithm for text dependent speaker verification system over telephone line. Our parametric filter bank adopts a variable bandwidth along with a fixed center frequency. Comparing with other methods, the proposed method turns out very robust to channel noise and background noise. Using this method, we segment an utterance into consecutive subword units, and make models using each subword unit. In terms of EER, the speaker verification system based on whole word model represents 6.1%, whereas the speaker verification system based on subword model represents 4.0%, improving about 2% in EER.
Key words: Speacker verification, Segmentation algorithm, Delta energy, Over telephone line
투고분야: 음성처리(2.5)

I. 서 론

화자 확인 시스템은 사용하는 어구의 종류에 따라 어구 종속(text-dependent), 어구 독립(text-independent), 어구 지시(text-prompted) 시스템으로 구분된다. 이중, 어구 종속 시스템에서는 암호, 카드 번호, PIN (Personal Identity Number)와 같은 특별한 어구를 발성하게 하여 화자의 신분을 확인한다[1]. 최근들어, 여러 논문에서 subword에 기반한 화자 확인 시스템의 성능이 whole word에 기반한 시스템의 성능보다 우수하다고 발표되었다[2][3]. 이는, subword에 기반한 시스템이 많은 어휘와 다양한 화자간의 변이를 잘 모델링하기 때문이라고 사료된다. 그러나, 실제 인식 시스템을 구현할 경우 모든 화자에 대해서 일일이 수동으로 음성을 분할한다는 것은 불가능하기 때문에, 입력된 음성데이터를 음운환경의 변화에 따라서 어떠한 언어학적인 정보 없이 자동으로 분할해야 한다. 이러한 새로운 분할 알고리즘을 blind segmentation이라 부른다. Blind segmentation은 입력된 음성을 어떠한 언어학적인 정보(orthograhic or phonetic transcription 등)도 없이 단지 음향학적인 정보만을 이용하여 subword로 분할해야만 한다. 그러므로, blind segmentation은 실제로 분할하기 전에 가장 적절한 subword 개수를 결정하는 방법까지 포함하고 있어야 한다[2]. 이상적인 경우 이 새로운 분할 알고리즘은 음향학적인 정보만으로 음소단위로 분할이 가능하나 전화선 채널에 의한 감쇄효과나 부족한 훈련 데이터 등을 고려하여 본 논문에서는 음절 단위 분할을 사용한다.

Blind segmentation은 어떤 종류의 파라미터를 사용하는가에 따라 크게 model based method와 model free method로 분류된다. Model based method는 특정 파라미터 모델(주로 LPC 모델)에 기반한 방법으로 delta cepstrum 을 이용한 방법, cepstrum계수에 의한 maximum likelihood를 이용한 방법 등이 있고, model free(non-parametric) method는 Parametric Filtering(PF) method, Maximum Likelihood Ratio (MLR) method 등이 있다[4][5].

All pole model을 이용하는 LPC 파라미터는 비음(nasal sound)이나 파찰음(fricative)등의 변화가 적절히 반

* 인하대학교 전자공학과 디지탈신호처리연구실
** 인천대학교 정보통신공학과
접수일자 : 1999년 10월 20일

영되지 않기 때문에 LPC 파라미터를 이용하는 model based method는 적절하지 않다[4]. 따라서, 본 논문에서는 blind segmentation 방법 중, model free(nonparametric) method를 이용한다. 특히, 여러 model free method중 가장 좋은 화자 확인 성능을 나타내는 Parametric Filtering method의 성능을 개선시킨 PF에 기반한 delta energy를 이용한 새로운 분할 알고리즘을 제안한다. 제안한 알고리즘은 음성을 동질적인(homogeneous) 부분으로 분할하기 위해 parametric filter bank와 filtering된 음성신호의 정규화 된 에너지를 함께 사용함으로써 전화선 등과 같은 채널잡음이나 부가적인 잡음에 강인함을 나타내었다.

II. Segmentation algorithm

2.1. Parametric Filtering

PF method는 잘 구현된 filter bank 출력의 통계적 특성에 의해 신호의 상관관계 구조가 특징지어진다는 사실에 근거한다.

$\{x_t\}$는 $P_k = E\{X_{t+k}X_t\}/E\{X_t^2\}$의 자기상관관계식을 갖고 평균이 0인 실수의 정상 신호라고 가정한다. 아래와 같은 IIR filter, $H(z^{-1};\alpha)$를 고려해보자.

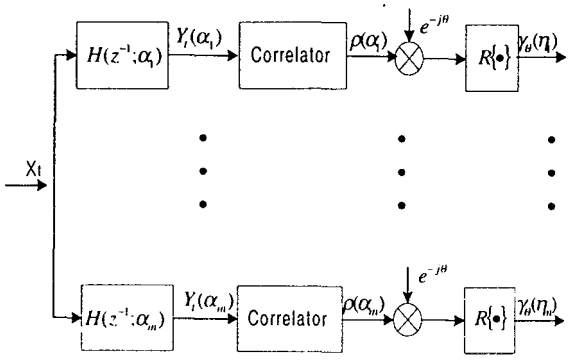


그림 1. Filter bank 의block diagram [4]
Fig. 1. Block diagram of filter bank.

$$Y_t(\alpha) = \sum_{l=0}^{\infty} \bar{\alpha}^l X_{t-l} = \bar{\alpha} Y_{t-1}(\alpha) + X_t \quad (1)$$

여기서, $\alpha = \eta e^{-j\theta}$ 이고 $|\eta|<1$인 복소수이고, $\bar{\alpha}$는 α의 켤레복소수 관계를 나타낸다. $P(\alpha)$는 $\{Y_t(\alpha)\}$의 1차 자기상관함수(lag-one autocorrelation)로 식 (2)와 같이 정의한다.

$$P(\alpha) = \frac{E\{Y_{t+1}(\alpha)\overline{Y_t(\alpha)}\}}{E\{|Y_t(\alpha)|^2\}} \quad (2)$$

이를 이용해 고정된 θ에 대해 $\{Y_t(\alpha)\}$의demodulated lag-one autocorrelation을 정의한다.

$$\gamma_\theta = R\{e^{-j\theta}P(\alpha)\} \quad (3)$$

여기서 $R\{\cdot\}$은 복소수의 실수부분을 나타낸다. $\{X_t\}$의 상관관계를 나타내기 위해 Fourier spectrum에 상응하는 새로운 특성 함수(characterization function)로서 $\gamma_\theta(\eta)$를 사용한다

특성 함수, $\gamma_\theta(\eta)$는 그림 1과 같은 filter bank로 쉽게 구해진다. 여기서 η_k는 $[\eta_a, \eta_b] \subset (-1,1)$사이의 일정한 간격을 두고 택할 수 있다[4]. 또한, 중심주파수 θ를 변화시키면서 분석할 수 있다. 즉, θ를 채널잡음이나 부가적인 잡음에 영향을 덜 받은 대역으로 택함으로써 주변환경에 더 강인해질 수 있다.

특성 함수, $\gamma_\theta(\eta)$는 음성신호를 분석하는데 있어 많은 흥미있는 특성을 가지고 있다. PF method를 뒷받침할 몇 가지 특성을 요약해 볼 수 있는데, 첫 번째로 $\gamma_\theta(\eta)$는 단조특성을 가지고 있다. 대부분의 정상 신호에서 $\gamma_\theta(\eta)$는 η가 증가함에 따라 증가함을 볼 수 있다. 이는 비정상신호의 상관관계를 도식적으로 해석할 수 있음을 의미한다.

둘째로, $H(z^{-1};\alpha)$는 중심주파수가 θ이고 대역폭이 $(\eta-1)/\sqrt{\eta}$인 bandpass filter의 특성을 가지고 있다. (만일 η가 음수이면 filter의 중심 주파수는 $\pi-\theta$가 된다.) θ가 변함에 따라 filter는 대역폭이 일정한 가운데 전체 주파수 영역에 대해 분석할 수 있으며, η가 변함에 따라 중심주파수가 일정한 가운데 대역폭을 변화시키며 분석 할 수 있다. 이 특징은 신호의 광대역과 협대역을 함께 평활화 시키는 경향이 있기 때문에 주변환경에 강인함을 지닌다. 이러한 유연성은 다른 filter bank와 구분되는 Parametric Filtering method의 특징이다[4].

2.2. Delta energy 의 이용

PF method에서 $\gamma_\theta(\eta)$는 음성의 스펙트럼을 다른 방법으로 표현함으로써 음성을 음향학적 subword로 분할하는데 이용한다. 그러나 시간 축에서 변화를 고려하지 않기 때문에 전화선 채널의 경우 채널 잡음이나 부가적인 잡음에 의한 스펙트럼의 변화에 너무나 민감하고 상대적으로 음성의 변화에 둔감해진다는 문제점이 있다. 따라서, 시간에 의한 음향학적 변화를 고려하고 채널 잡음이나 부가적인 잡음에 둔감하기 위해 delta energy를 이용한다.

Parametric Filt된 음성 $Y_t(\alpha(\eta))$을 η에 대해 적분함으로써 식 (4)와 같이 새로운 energy ering $E_t(\alpha(\eta))$을 정의하고 이를 이용해 delta energy를 구한다.

$$E_t(\alpha(\eta)) = \int_{\eta_a}^{\eta_b} |Y_t(\alpha(\eta))|^2 d\eta \quad (4)$$

$$d_t = E_t(\alpha(\eta)) - E_{t-1}(\alpha(\eta)) \quad (5)$$

Parametric Flitering 이후 delta energy를 이용함으로써 특성함수, $\gamma_\theta(\eta)$가 가지고 있는 고유의 특성들이 delta energy에 상속된다. 즉, 제안된 delta energy 역시 단조특성을 지니고 있으며, 신호의 협대역과 광대역을 함께 평활화 시키는 경향을 띠게 된다. 그러나, delta energy는 $\gamma_\theta(\eta)$보다 채널잡음이나 부가적인 잡음에 둔감한 특성을 보이기 때문에 주변 환경에 더욱 강인해 진다. 제안한 delta energy가 일반적인 delta energy와 다른 점은 Parametric Filtering의 특성을 그대로 상속 받는다는 것과, 이후 설명하겠지만, 잡음에 의한 영향을 적게 받은 대역의 정규화된 energy를 사용한다는 것이다. 특히, 제안한 delta energy는 model free에 기반한 방법이므로 model-based 방법에 의해 발생할 수 있는 modeling 오류에 대해서도 강인함(robustness)을 얻을 수 있다.

2.3. Center frequency, θ의 결정

Parametric Filtering을 이용한 방법은 중심주파수, θ에 따라 화자 확인 성능에 영향을 많이 끼친다. 즉, 채널 잡음이나 배경 잡음에 영향을 적게 받고, 입력된 음성을 음향학적 subword로 분할하기에 적합한 중심주파수 θ를 결정하는 것이 매우 중요하다.

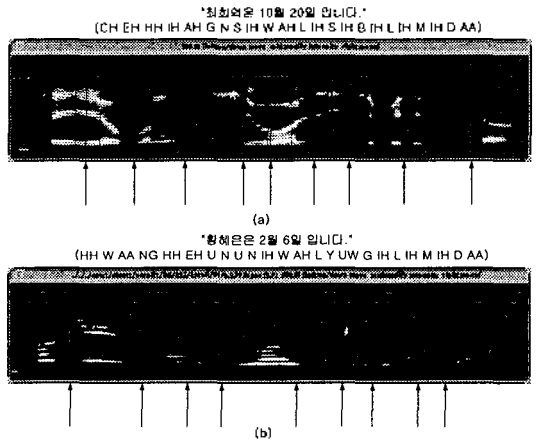


그림 2. 분할 된 음성의 예
(a) 채널의 영향을 적게 받은 음성
(b)채널의 영향을 많이 받은 음성

Fig. 2. Example of segmented utterances;
(a) Less effected by channel noisc,
(b) Severely effected by channel noise.

본 논문에서는 FFT 알고리즘을 이용해 중심주파수 θ를 결정하였다. 즉, 음성의 변이가 가장 심한 부분을 선택하기 위해 FFT알고리즘을 적용한 후 가장 에너지가 큰 band를 θ로 결정하였다. 음성의 변이가 심한 밴드를 택하기 위해 분산을 이용할 수도 있다. 두 가지 방법 모두 이용 가능하며 화자 확인 성능에 역시 큰 차이가 없는 것으로 나타났다.

그림 2에 위에서 설명한 Parametric Filtering에 기반한 delta energy를 이용하여 음성을 분할한 예를 보인다. 사용된 어구와 이에 해당하는 발음기호를 나타내었다. 발음기호는 CMU(Carnegie Mellon Univ.)에서 발표한 표준 발음 표기를 따랐다. 그림 2(a)는 채널의 영향을 심하게 받지 않은 음성이고, 그림 2(b)는 채널에 의해 심하게 감쇄된 음성이다. 두 가지 음성 모두에서 알고리즘은 음향학적으로 변화하는 부분에서 음성을 분할하고 있다.

III. 실험방법

본 논문에서 사용한 음성 데이터 베이스는 남자 53명, 여자 53명, 전체 106명에 대해 수집하였다. 이중 남자 22과 여자 20명은 동일한 음성, "범일정보통신입니다."(SP1)를 발성하였고, 남자 26명, 여자 33명은 다른 음성, "000(이름)은 00월 00일(몇월 몇일) 입니다."(SP2)를 발성하였다. 또한, 시간에 따른 화자의 발성 변화를 포함하도록 날짜별로, 요일별로, 오전, 오후, 저녁으로 나누어 3개월간 수집하였다 (오전 : 9시~11시, 오후 : 2시~5시, 저녁 : 7시~10시). 각각의 어구에 대해 화자별로 훈련용 데이터는 6 session으로 구성되어 있고, 각 session당 6개의 훈련 데이터가 있다. 테스트 데이터는 화자별로 최대 20 session으로 구성되어 있고, 각 session당 4회의 테스트 데이터가 있다.

음성 데이터를 각 화자가 있는 직장이나 가정에서 전화기(sender)를 이용해 전화선을 통하여 보내면 서울의 범일정보통신 본사에 있는 시스템(receiver)에서 수집 하였다. 단, 이동전화나 공중전화는 사용하지 않도록 하였다. Receiver에서 수집된 데이터의 형태는 8KHz 8bit μ-law 데이터이다.

DB를 구성하고 있는 총 6개의 훈련 세트 중 2개는 객관적인 분석을 하기에는 시료의 개수가 부족하다고 판단되어 실험에서 제외시켰다. 즉, 6 session 중 4 session을 사용하였다. 각 session당 훈련데이터는 끝점 검출 오류 등을 고려하여 4개만을 사용하였다. 테스트 데이터로는 모델에 해당하는 화자의 테스트 session을 모두 사용하고, 모델에 해당하는 화자를 제외한 화자의 테스트 데이터에서는 한 화자당 1개의 데이터만을 랜덤하게 추출하여 사용하였다.

수집된 데이터는 8KHz, μ-law 데이터이므로 실험을 위하여 μ-law 테이블을 이용하여 8KHz, 16 bit linear 데이터로 변환하였다. 음성 신호로부터 특징을 추출하기 위해서 20ms의 길이의 Hamming 윈도우를 사용하였고 10ms씩 이동하면서 분석하였다. 전화선 채널은 300Hz~3400Hz의 대역폭을 가지므로 불필요한 채널의 잡음을 줄이기 위해 특징 벡터를 추출하기 전에 BPF(Band Pass Filtering)을 하였다. 특징 벡터로는 잡음환경에 강인하다고 알려진

MFCC(Mel Frequency Cepstral Coefficient)와 delta 계수를 사용하였고, 모델링 과정과 인식 과정에서 조금씩 변하는 채널의 왜곡성분을 켑스트럼 영역에서 제거하기 위해 CMS(Cepstral Mean Subtration)을 사용하였다.

본 논문에서는 HTK를 이용하여 whole word model과 subword model의 화자 확인 성능을 비교하였다. 이때 각 subword는 3 states, 3 mixtures를 사용했고, whole word는 subword의 평균 개수가 7~8개 임을 감안하여 24 states (3 states 8), 3 mixtures를 사용했다. 성능 평가의 척도는 의뢰인을 거부하는 오인 거부율(FR, false rejection)과 사칭자를 수락하는 오인 수락율(FA, false acceptance)이 같아지는 지점인 EER(equal error rate)을 사용하였다. 본 논문에서는 EER을 화자별로 구하였으며, 남자와 여자, 그리고, SP1, SP2로 구분하여 평균 EER을 비교한다.

3.1. DTW를 이용한 패턴정합

같은 화자가 같은 단어를 발성할 경우라도, 발성음의 길이는 매시간 비선형적으로 전개와 수축하면서 변화한다. 뿐만 아니라 부가적인 잡음까지 포함이 되기도 하기 때문에 segmentation 알고리즘만으로는 각 발성들을 일관적으로 분할하기란 매우 어려운 일이다. 또한, 이런 subword 경계의 불일치는 화자 확인 성능에 치명적인 영향을 미친다. 따라서 각 발성간에 정확한 subword경계를 구하기 위해 dynamic programming기술을 이용해 warping function을 구한 뒤 각 발성간의 subword 경계를 정합시킨다. 즉, 기준 패턴 R(w(n))과 입력 패턴 T(n)의 거리 D가 최소가 되는 w(n)을 찾는다[6].

$$D = \min_{w(n)} \sum_{n=0}^{T} d(T(k), R(w(k))) \qquad (6)$$

3.2. 프레임 길이에 대한 정규화

사용자가 음성을 발성할 때, 각각의 음성은 길이가 다르므로 프레임 길이에 대한 정규화 과정이 필요하다. Subword model의 경우 정규화 방법은 전체 음성의 길이에 대해 정규화하는 방법과 각 subword의 길이에 대해 정규화 하는 방법이 있다.

표 1. 두 가지 정규화 방법의 비교
Table 1. Comparison of two normalization methods.

	Name	Whole	Sub
의뢰인	A	-2.75	-2.50
	B	-1.03	-2.51
	C	-3.83	-2.57
사칭자	A	-0.21	-3.00
	B	-0.23	-4.51
	C	-8.64	-3.00

다음 표 1은 한 화자의 성문 모델에 사칭자의 음성과 의뢰인의 음성을 인식한 후 두가지 정규화 방법을 적용한 결과를 보여준다. Whole은 음성 전체길이에 대한 정규화 방법이고 Sub는 각 subword의 길이에 대한 정규화 방법이다. 의뢰인의 경우에는 subword의 길이로 정규화 할 때 whole word의 길이로 정규화 할 때와 큰 차이가 없으나 사칭자의 경우 효과적으로 작용한다. 즉, 사칭자에 대해 프레임의 길이가 짧아 확률값이 클 경우(B), whole word의 길이로 정규화 하면 여전히 확률값이 큰 반면 (-0.23), subword의 길이로 정규화하면 평균확률보다 작아지기 때문이다(-4.51).

IV. 실험결과

실험 결과를 분석하기 위해 다음과 같이 몇가지로 분류하여 비교하였다. 먼저 whole word model 및 알고리즘에 의해 분할한 subword model의 성능을 비교하고, 알고리즘에 의한 분할한 것과 손으로 직접 분할 한 것의 성능 비교, 마지막으로 프레임 길이 정규화에 의한 성능을 비교하였다. 각 훈련세트의 결과는 3개의 테스트 세트로 나타난 결과의 평균치이다.

표 2. Whole word model 및 알고리즘에 의한 subword model의 EER
Table 2. EER's of whole word model and subword model.

			PF_	PF_D	WHOLE
Enrollment Set 1	M	SP1	6.22	5.61	7.19
		SP2	2.88	2.50	4.05
	F	SP1	6.95	6.66	7.49
		SP2	5.05	4.09	4.97
Enrollment Set 2	M	SP1	6.79	5.55	7.28
		SP2	2.75	2.15	3.19
	F	SP1	7.32	7.23	8.19
		SP2	4.42	3.49	5.40
Enrollment Set 3	M	SP1	4.75	4.04	6.92
		SP2	2.73	2.25	3.75
	F	SP1	8.35	8.05	9.45
		SP2	4.24	3.79	5.10
Enrollment Set 4	M	SP1	6.52	6.33	6.25
		SP2	3.30	1.94	3.74
	F	SP1	7.12	6.67	7.37
		SP2	4.04	3.72	4.65
평균			4.87	4.36	5.61

표 2에는 whole word model 및 각 알고리즘을 사용하여 분할한 subword model의 성능을 비교하였다. 본 논문에서 제안하는 Parametric Filtering에 기반한 delta energy method의 성능을 비교하기 위해 여러 가지 알고리즘들 중 가장 좋은 성능을 나타내는 특성 함수, $\gamma_\theta(\eta)$를 이용한 방법과 비교한다. 표 2에서 PF_γ는 특성 함수를 이용한 방법을 나타내고 PF_D는 delta energy를 이용한 방법을 나타낸다. 알고리즘에 의한 성능만을 비교하기 위해 프레임 길이에 대한 정규화는 whole word model과 동일하게 적용하였다. PF_D를 사용하였을 때가 평균 EER이 4.36%(남자 3.92%, 여자 5.55%)로 가장 좋은 성능을 보였다. PF를 기반으로한 방법 중 본 논문에서 제안한 delta energy를 이용한 방법이 characterization function을 이용

한 방법보다 성능이 우수한 이유는 음성의 작은 변화나 부가적인 잡음, 채널잡음에 둔감하기 때문이라고 사료된다.

다음 표 3은 직접 손으로 음성을 분할한 것과 알고리즘을 이용해 분할한 것의 성능을 비교하였다. 24명의 화자를 무작위로 선택하여 손으로 분할하였다. 화자 확인 시스템에서는 적은 양의 데이터를 사용하기 때문에 음소단위로 분할하면 각 음소가 잘 모델링되지 않으므로 적당하지 않다고 판단되어 음절 단위로 분할하였다.

표 3. Hand label 과 알고리즘에 의한 subword model의 EER
Table 3. Performance comparison of hand label and PF_D.

			Hand Label	PF_D
Enrollment Set 1	M	SP1	5.19	4.89
		SP2	2.68	3.24
	F	SP1	10.02	7.79
		SP2	3.43	4.38
Enrollment Set 2	M	SP1	9.94	9.05
		SP2	2.74	4.34
	F	SP1	11.08	10.75
		SP2	3.11	3.12
Enrollment Set 3	M	SP1	7.24	7.72
		SP2	2.14	3.29
	F	SP1	11.69	11.62
		SP2	2.70	3.30
Enrollment Set 4	M	SP1	7.34	7.37
		SP2	1.47	1.61
	F	SP1	12.00	11.89
		SP2	2.66	2.34
평균			5.75	5.85

손으로 분할 했을 때(5.75%)가 알고리즘으로 분할했을 때(5.85%) 보다 EER이 좋게 나타났다. 이유는 끝점 검출에 의한 오류 및 분할 개수의 차이 등으로 추론할 수 있다. 손으로 분할 시 끝점 검출에 의한 오류는 모두 삭제하였으나 알고리즘으로 분할 시 삭제할 수 없으므로 이에 의한 오류가 첨가된다. 특히, 단어 사이에 묵음이 존재할 경우 이는 화자 확인 성능에 더욱 치명적인 영향을 끼친다. 이는 좀 더 잡음에 강인한 끝점 검출 알고리즘을 사용한다거나 keyword sporting 기술을 이용해 정확한 끝점을 검출한다면 더 좋은 성능을 얻을 수 있다고 사료된다. 또, 채널의 영향을 많이 받은 화자나 음성의 변화가 많지 않은 화자의 경우 분할 개수가 일반적으로 적기 때문에 분할 개수가 더 많은 손에 의한 분할 보다 성능이 나빠진다.

다음 표 4 는 프레임 길이 정규화 방법에 따른 성능 비교를 나타냈다. 프레임의 전체 길이(whole)로 정규화 할 때보다 subword의 길이(sub)로 정규화 할 때가 성능이 더 우수한 것으로 나타났다. 이것은 의회인의 경우에는 두 가지 방법이 큰 차이가 없으나 사칭자의 경우 프레임 길이가 짧아 확률 값이 상대적으로 높은 모델에 대해 subword의 길이로 정규화하는 것이 더 효과적이기 때문이다.

표 4. 프레임 길이 정규화 방법에 따른 EER
Table 4. EER and comparative results for different normalization method by frame length.

		Enrollment Set 1		Enrollment Set 2		Enrollment Set 3	
		Whole	Sub	Whole	Sub	Whole	Sub
M	SP1	5.61	5.26	5.55	5.39	5.04	5.15
	SP2	2.50	1.81	2.15	1.83	2.25	1.67
F	SP1	6.68	6.93	7.23	7.14	8.35	8.61
	SP2	4.09	3.69	3.49	3.29	3.79	2.96
평균		4.72	4.01	4.61	3.94	4.85	4.07

다음 표 5는 whole word model과 본 논문에서 제안하는 알고리즘으로 분할한 subword model의 성능을 비교한다. 본 논문에서는 PF에 기반한 delta energy를 이용해 음성을 분할 하였다. 중심 주파수를 결정하기 위하여 512 point FFT알고리즘을 이용했으며, 각 subword의 프레임 길이로 정규화 하였다.

표 5. Whole word model과 subword model의 EER
Table 5. Average EERs with whole word model and subword model.

		Enrollment Set 1		Enrollment Set 2		Enrollment Set 3	
		Whole Word	Sub Word	Whole Word	Sub Word	Whole Word	Sub Word
M	SP1	7.22	5.26	8.10	5.39	7.57	5.15
	SP2	4.20	1.81	3.45	1.83	4.00	1.67
F	SP1	8.30	6.93	9.25	7.14	9.62	8.61
	SP2	4.97	3.69	5.59	3.29	5.40	2.96
평균		5.72	4.01	6.17	3.94	6.27	4.07

본 논문에서 제안한 알고리즘으로 분할한 subword model (약 4.00%)의 EER이 whole word model의 EER(약 6.05%)보다 약 2% 향상되었다.

V. 결 론

본 논문에서는 PF에 기반한 delta energy를 이용해 subword로 분할하는 방법을 제안하였다. 제안한 알고리즘은 특정 밴드의 주파수를 중심으로 다양한 대역폭을 갖는 filter bank를 이용해 음성을 분석한다. 중심주파수를 선택하기 위해 FFT 알고리즘을 사용했으며, 각 발성 간에 subword의 정확한 matching을 위해 DTW알고리즘을 사용했다. 그 결과 whole word model의 EER이 약 6.1%, subword model의 EER이 약 4.0%로 본 논문에서 제안한 알고리즘을 사용했을 때 약 2%의 성능이 향상되었다.

끝점 검출의 오류에 의해 묵음이 첨가되거나 음성사이의 묵음이 길어질 경우 이에 영향을 받은 subword model은 화자 확인 성능에 나쁜 영향을 미침을 볼 수 있다. 향후, 이와 같이 첨가된 묵음을 subword의 경계를 정합하는 과

정에서 제거하여 보다 정확한 음성의 경계를 구할 것이다.

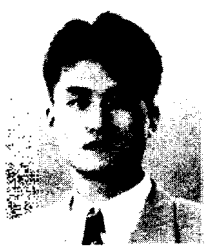
참 고 문 헌

1. Aaron E. Rosenberg. "Chin-Hui Lee, and Sedat Gokcen, Connected Word Talker Verification Using Whole Word Hidden Markov Models," *Proc. ICASSP* 1991, pp. 381 384, 1991.
2. 김유진, 김지운, 장재호, "SVAPI 1.0 환경에서의 어구 종속 화자 확인 시스템," 제 15회 음성통신 및 신호처리 워크샵, pp. 401-405, 1998.
3. Manish Sharma, and Richard mammane, "Subword-based Text-dependent Speaker Verification System with User-selectable Password," *Proc. ICASSP* 1996, pp. 93-96, 1996.
4. S. Euler, R. Langlitz, and J. Zinke, "Comparison of Whole Word and Subword Modeling Techniques for Speaker Verification with Limited Training Data," *ICASSP* 1997, pp. 1079-1082, 1997.
5. Ta-Hsin Li, and Jerry D. Gibson, "Speech Analysis and Segmentation by Parametric Filtering," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 3, May 1996.
6. Torbjorn Svendsen, and Frank K. Soong, "On the Automatic Segmentation of Speech Signals," ICASSP 1987, pp. 77-80, 1987.
7. Hiroaki Sakoe, and Seibi Chiba, "Dynamic Programming Alogrithm Optimization for Spoken Word Recognition," *IEEE Trans. on ASSP*, Vol. 26, No.1, pp. 43-49, Feb.1978.
8. Aaron E. Rosenberg, Chin-Hui Lee, and Sedat "Gokcen, Connected Word Talker Verification Using Whole Word Hidden Markov Models," *ICASSP* 1991, pp. 381-384, 1991.

▲김 지 운(Ji Un Kim)

1998년 2월 : 인하대학교 전자공학과 공학사
2000년 2월 : 인하대학교 전자공학과 공학석사
2000년 3월 ~ 현재 : 인하대학교 전자공학과 박사과정
※주관심분야 : 음성인식, 발화인증, 화자인식

▲김 유 진(Yu Jin Kim)



1995년 2월 : 인하대학교 전자공학과 공학사
1997년 2월 : 인하대학교 전자공학과 공학석사
1996년 8월 ~ 1997년12월 : 한국전자통신연구소(ETRI) 음성언어처리연구실 위촉연구원
1997년 2월 ~ 1998년 5월 : LG반도체 SD연구소(System Device R&D Lab.) 연구원
1998년 9월 ~ 현재 : 인하대학교 전자공학과 박사과정
※주관심분야 : 패턴인식, 음성인식, 발화인증, 화자인식

▲민 홍 기(Hong Ki Min)

1979년 2월 : 인하대학교 전자공학과 공학사
1981년 8월 : 인하대학교 전자공학과 공학석사
1990년 8월 : 인하대학교 전자공학과 공학박사
1985년 10월 ~ 1991년 7월 : 한국과학기술연구원 선임연구원
1993년 8월 ~ 1994년 7월 : 미국Delaware대학 방문교수
1991년 8월 ~ 현재 : 인천대학교 정보통신공학과 부교수
※주관심분야 : 신호처리, 재활공학, AAC

▲정 재 호(Jae Ho Chung)
1982년 : University of Maryland(BSEE)
1984년 : University of Maryland(MSEE)
1990년 : Georgia Institute of Technology(Ph.D.)
1984년 ~ 1985년 : 미국 국방성 산하 해군 연구소, 신호처리실, 연구원
1991년 ~ 1992년 : AT&T Bell Laboratories, 음성신호처리 연구실, 연구원(MTS)
1992년 ~ 현재 : 인하대학교 공과대학 전자공학과, (현)부교수