논문 2004-41SP-4-13

화자 정규화를 위한 새로운 파워 스펙트럼 Warping 방법

(A New Power Spectrum Warping Approach to Speaker Warping)

유 일 수*, 김 동 주*, 노 용 완*, 홍 광 석**

(Il-Soo Yu, Dong-Joo Kim, Yong-Wan Rho, and Kwang-Seok Hong)

요 약

화자 정규화 방법은 화자 독립 음성인식 시스템에서 음성 인식의 정확성를 높이기 위한 성공적인 방법으로 알려져 왔다. 널리 사용되는 화자 정규화 방법은 maximum likelihood 기반의 주파수 warping 방법이다. 본 논문은 주파수 warping 보다 더 좋은 화자 정규화의 성능 개선을 위해 새로운 파워 스펙트럼 warping 방법을 제안한다. 파워 스펙트럼 warping은 멜 주파수 켑스트럼 분석(MFCC) 방법을 이용하며, MFCC 처리 단계에서 필터 뱅크의 파워 스펙트럼을 조절함으로써 화자 정규화를 수행하는 간단한 메커니즘으로 갖는다. 또한 본 논문은 파워 스펙트럼 warping과 주파수 warping 방법을 서로 결합한 hybrid VTN 방법을 제안한다. 본 논문의 실험은 baseline 시스템에 각 화자 정규화 방법을 적용하여 SKKU PBW DB에서 인식 성능을 비교 분석하였다. 실험 결과를 보면 baseline 시스템의 단어 인식 성능을 기준으로 주파수 wapring은 2.06%, 파워 스펙트럼 warping은 3.05%, 그리고 hybrid VTN은 4.07%의 단어 에러 율의 감소를 보였다.

Abstract

The method of speaker normalization has been known as the successful method for improving the accuracy of speech recognition at speaker independent speech recognition system. A frequency warping approach is widely used method based on maximum likelihood for speaker normalization. This paper propose a new power spectrum warping approach to making improvement of speaker normalization better than a frequency warping. Th power spectrum warping uses Mel-frequency cepstrum analysis(MFCC) and is a simple mechanism to performing speaker normalization by modifying the power spectrum of Mel filter bank in MFCC. Also, this paper propose the hybrid VTN combined the power spectrum warping and a frequency warping. Experiment of this paper did a comparative analysis about the recognition performance of the SKKU PBW DB applied each speaker normalization approach on baseline system. The experiment results have shown that a frequency warping is 2.06%, the power spectrum is 3.06%, and hybrid VTN is 4.07% word error rate reduction as of word recognition performance of baseline system.

Keywords : Speaker Normalization, Power spectrum warping, Frequency warping, MFCC

I. 서 론

일반적인 음성인식시스템은 화자 독립형 시스템을 지향하고 있으며, 한정된 훈련 모델에 의존한다. 이 경우 화자들 사이의 성도 모양의 변이를 모두 고려해 줄 수 없기 때문에 음성인식의 성능 저하를 보이게 된다. 이 문제점을 보완하기 위해 화자 정규화 방법이 고안되었고 여러 논문에서 다양한 방법들이 소개되었으며, 특히 HMM 기반 음성인식시스템의 성능 개선을 보여주고 있다.[1,2,3] 성도 모양의 변이를 정규화 하는 것을 성도 정규화(VTN; Vocal Tract Normalization)라고 하며, 화자 사이의 성도 길이 변이와 밀접한 관련이 있다. 특히, 성별에 따라 성도 길이의 큰 차이를 보인다. 조사된 자료에 의하면, 성인의 화자 별 성도 길이의 범위는 13cm에서 18cm 정도의 변이를 보이는 것으로 나타났으며, 이것은 화자 사이에서 포먼트 중심 주파수의 변이가 25% 정도 차이를 보이는 것이라고 해석할 수 있다.[1] 성도 정규화를 위해 널리 사용되는 방법은 음성 스펙트럼의 주파수 warping으로 여러 논문에서 소개되어 왔다. 이 방법은 멜 주파수 켑스트럼(MFCC) 특징 분석에서 멜 필터 뱅크(MFB; Mel Filter Bank)의 주파수 축의

* 학생회원, ** 정회원, 성균관대학교
(Sungkyunkwan University)
접수일자: 2003년10월11일, 수정완료일: 2004년6월4일

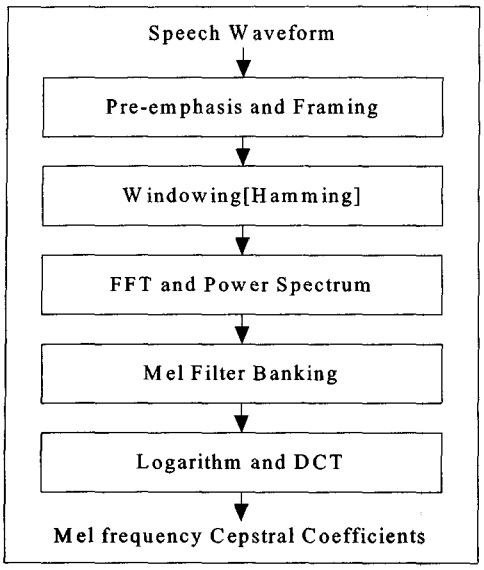

그림 1. 일반적인 특징 추출(MFCC)의 블럭도
Fig. 1. A block diagram of the MFCC.

선형 warping을 통해 쉽게 구현 될 수 있다.[1,2]

본 논문은 기존의 화자 정규화 방법보다 더 나은 인식성능의 개선을 위해 파워 스펙트럼 warping 방법을 새롭게 제안한다. 제안하는 파워 스펙트럼 warping은 기존의 주파수 warping 방법과 비슷하게 MFCC 특징 분석에서 MFB를 조절함으로써 쉽게 구현 될 수 있다. 기존의 주파수 warping 방법은 MFB의 주파수 축을 warping 하는 반면, 제안하는 파워 스펙트럼 warping은 MFB의 파워 스펙트럼 축을 warping 한다. 또한 또한 본 논문은 파워 스펙트럼 warping과 주파수 warping 방법을 서로 결합한 hybrid VTN 방법을 제안한다.

기존의 주파수 warping 방법에서 중요하게 다뤄진 부분은 크게 두 가지로 요약할 수 있다. 첫 번째는 warping factor를 estimation 하는 것이고, 두 번째는 이것을 효율적으로 인식과정에서 처리하는 단계(recog-nition pro- cedure)이다. 제안하는 파워 스펙트럼 warping과 hybrid VTN 방법도 이 두 가지 요소를 중요하게 고려하였다. 파워 스펙트럼 warping factor를 estimation 하기 위해 주파수 warping에서 널리 사용되는 maximum likelihood 방법을 사용하였다. 그리고 인식 처리를 수행하기 위해 주파수 warping에서 널리 사용되는 multiple-pass 처리 방법을 사용하였다. 마지막으로 우리가 제안하는 화자 정규화의 인식성능을 평가하기 위해, 한국어 단어 단위 음성 DB(SKKU PBW)를 이용하였고, baseline 시스템과 주파수 warping의 인식성능을 바탕으로 파워 스펙트럼 warping과 hybrid VTN에 대해 각각의 인식 성능을 비교 분석하여 각 화자 정규화 방법의 성능을 평가하였다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 Li Lee외[1]와 L. Welling 외[2]의 주파수 warping 방법을 소개한다. Ⅲ장에서는 본 논문에서 제안하는 화자 정규화 방법으로 파워 스펙트럼 warping 방법을 다룬다. Ⅳ장에서는 Ⅱ장과 Ⅲ장에서 다룬 성도 정규화 방법을 결합한 hybrid VTN을 다룬다. Ⅴ장에서는 SKKU PBW의 한국어 단어 단위 음성 DB를 사용하여, baseline 시스템, 주파수 warping, 파워 스펙트럼 warping, 그리고 hy-brid VTN의 인식 실험을 통하여 각각의 인식 성능을 비교 분석한다. 마지막으로 Ⅵ장에서는 논문의 전반적인 내용을 검토하고, 결론을 맺는다.

Ⅱ. 주파수 Warping

본 장에서는 화자 정규화를 위해 널리 사용되고 있는 주파 warping 방법에 대해 Li Lee 외[1]와 L. Welling 외[2] 의해 소개된 내용을 다룬다. 주파수 warping 방법은 화자 사이의 음성 신호 변이를 줄이기 위해 성도 길이의 변이를 정규화 하는 대표적인 방법이라고 할 수 있다.

1. 주파수 warping을 위한 MFB

일반적으로 성도 정규화는 front end에서 특징 벡터의 변형을 통해 수행된다. 특히, 주파수 warping 방법은 스펙트럼 분석 기반 음성 특징 분석 방법으로 널리 사용되는 그림1의 MFCC 처리를 이용하여 쉽게 구현될 수 있다.[1,2] 주파수 warping은 주파수 축을 warping 함으로써 성도 길이의 변이를 정규화 하는 방법이다. 이 방법은 MFCC 처리 과정에서 MFB 부분을 조절함으로써 주파수 축의 warping이 이루어진다. 주파수 war-ping factor에 따라 멜 주파수가 warping된 모습은 그림 2와 같이 표현된다. warping factor α 값이 1보다 작은 경우는 스펙트럼의 주파수 영역을 확장하는 것과 같고, 1보다 큰 경우는 스펙트럼의 주파수 영역을 압축하는 것과 같다.

그림 2와 같이 멜 주파수를 warping 한다면, 주파수 영역을 확장시키는 경우에는 MFB 크가 원래 스펙트럼의 영역에서 벗어난 부분까지 미치는 문제가 발생한다. 그리고 주파수 영역을 압축시키는 경우에는 필터가 원래 스펙트럼의 영역 내의 정보를 모두 포함하지 못하는 문제가 발생한다. 이 문제를 해결하기 위해 구분적

(piecewise) 선형 warping 방법이 고안되었고, 일반적인 선형 warping 보다 더 강인한 인식 성능을 보여주는 것으로 알려져 있다.[8] 구분적 선형 warping 방법이 적용된 멜 주파수는 그림 3과 같이 나타난다. 본 논문에서는 주파수 warping 처리를 위해 구분적 선형 warping을 사용하였으며, 다음과 같이 표현된다.

$$\mathcal{F} = 2595\log\left(1+\frac{f}{700}\right)$$

$$Mel^{\alpha}(f) = \alpha \cdot \mathcal{F} \quad ,0 \le f \le f_0$$

$$Mel^{\alpha}(f) = \frac{\mathcal{F}_{\max} - \alpha \cdot \mathcal{F}_0}{\mathcal{F}_{\max} - \mathcal{F}_0}(\mathcal{F} - \mathcal{F}_0) + \alpha \cdot \mathcal{F}_0 \quad ,f_0 \le f \le f_{\max} \tag{1}$$

여기서 f_0 는 음성의 포먼트의 중심 주파수가 미치는 가장 바깥쪽 주파수를 의미하며, f_0를 3.6kHz 로 설정하였다. 그리고 $f_{\max}$는 Nyquist 주파수를 말한다. 식(1)이 적용된 구분적 선형 warping 멜 주파수는 그림 3에 나타내었다. 그림 1의 일반적인 MFCC 처리 과정에서 MFB의 처리 부분은 다음과 같이 표현 할 수 있다.

$$\mathcal{S}[l] = \sum_{k=0}^{k=K/2} S[k] \cdot M_l[k]$$

$$l = 0, 1, \ldots, L-1 \tag{2}$$

여기서 $S[k]$는 파워 스펙트럼을, $M_l[k]$는 멜 삼각 필터 뱅크를, L는 멜 삼각 band-pass 필터의 개수를, K는 FFT의 resolution의 값이다. 식(2)에 식(1)의 구분적 선형 warping을 적용한 식은 다음과 같이 표현한다.

$$\mathcal{S}^{\alpha}[l] = \sum_{k=0}^{k=K/2} S[k] \cdot M_l^{\alpha}[k]$$

$$l = 0, 1, \ldots, L-1 \tag{3}$$

2. 주파수 warping factor의 estimation

주파수 warping 방법의 중요한 부분 중에 하나는 특정 화자에 대해 어느 정도의 성도 길이를 정규화 해야 하는지를 결정하는, 주파수 warping factor를 구하는 것이다. 주파수 warping factor의 estimation은 성도 길이 정규화 factor를 결정하는 작업이라 할 수 있다. 주파수 warping factor는 기준 성도 길이(기준 HMM 음향 모델)와 특정 화자의 성도 길이 사이의 비율로 나타낼 수 있다. 일반적으로 성도 길이는 주파수 warping factor와 반비례의 관계를 갖는다.

주파수 warping factor의 estimation을 수학적 표기

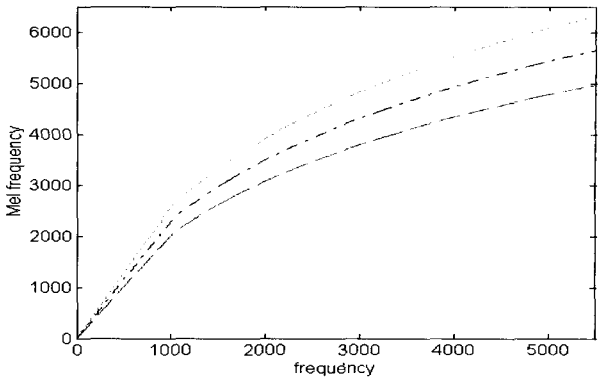

그림 2. 선형 주파수 대 선형 warping 멜 주파수(점선; 최대 warping 멜(α=1.12), 점-대시선: 기준 멜(α=1.00), 대시선: 최소 warping 멜(α=0.88))

Fig. 2. Linear frequency vs. linear warping Mel frequency (dot line: maximum warping Mel (α=1.12), dot-dash line: base Mel(α=1.00), dash line: mini-mum warping Mel(α=0.88)).

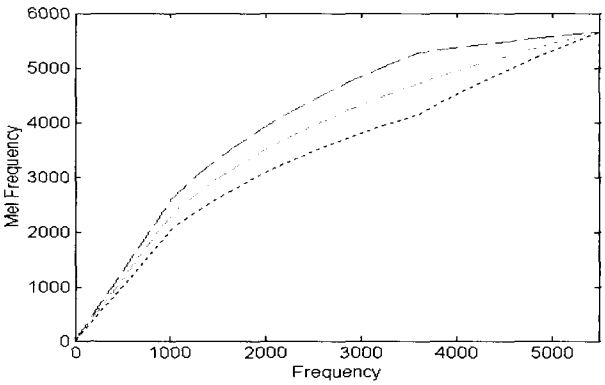

그림 3. 선형 주파수 대 구분적(piecewise) 선형 warping 멜 주파수(점선; 최대 warping 멜(α=1.12), 점-대시선: 기준 멜(α=1.00), 대시선: 최소 warping 멜(α=0.88))

Fig. 3. inear frequency vs. piecewise linear warping Mel frequency (dot line: maximum warping Mel(α=1.12), dot-dash line: base Mel(α=1.00), dash line: mini-mum warping Mel(α=0.88)).

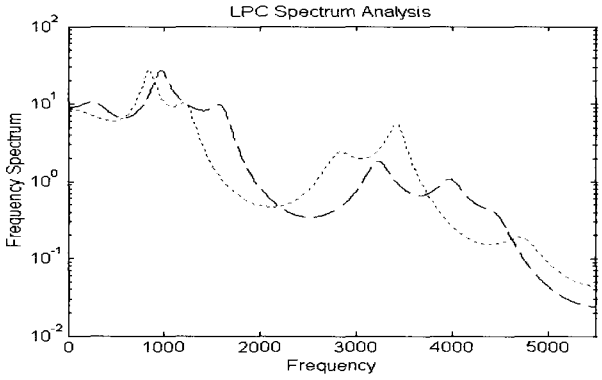

그림 4. 발성 음성 /아/의 남자와 여자 LPC 스페트럼 분석(대시선: 여자, 점선: 남자)

Fig. 4. Male and female LPC spectrum analysis of utterance /a/ (dash line: female, dot line: male).

로 나타내보면, 먼저 주파수 warping factor가 적용되지 않은 baseline 시스템의 발성 음성에 대한 특징벡터는

$$X_i = \{x_i(0), \ldots, x_i(T)\} \tag{4}$$

이다. i는 발성 화자를 의미하며, T는 특징 벡터의 전체 개수이다. 주파수 warping이 적용된 특징 벡터는 2.1절에서 다룬 식(3)에 의해 수행된다. 주파수 warping factor가 적용된 특징 벡터는

$$X_i^{\alpha} = \{x_i^{\alpha}(0), \ldots, x_i^{\alpha}(T)\} \tag{5}$$

이다. 그리고 추가적으로 전체 인식 후보 단어(transcription set)의 표기는 다음과 같다.

$$W = \{w_1, w_2, \ldots, w_N\} \tag{6}$$

화자 i의 최적 주파수 warping factor는 다음과 같이 HMM decoding 단계에서 maximum likelihood 방법을 적용하여 얻을 수 있다.

$$\hat{\alpha} = \arg\max P(X_i^{\alpha} | \lambda, W) \quad , \text{for } \alpha \tag{7}$$

즉, 최적 주파수 warping factor $\hat{\alpha}$는 HMM 음향 모델 λ와 전체 인식 후보 단어 W에 대해, warping된 발성 음성의 특징 벡터 X_i^{α}와의 likelihood가 최대로 되는 α로 정의 할 수 있다.

최적 주파수 warping factor $\hat{\alpha}$를 구하기 위한 α의 탐색 범위는 성인의 성도 길이의 변이가 25%의 차이를 갖는다는 점을 이용하여 다음과 같이 정의 할 수 있다.

$$0.88 \le \alpha \le 1.12 \quad , spaced \ 0.02 \tag{8}$$

3. 주파수 warping의 인식 처리

앞 절에서는 최적 주파수 warping factor를 estima-tion하는 방법에 대해 소개했다. 이 절에서는 화자 i의 발성 음성으로부터 최적 주파수 warping factor를 적용하여, 음성 인식 처리를 수행하는 부분에 대해 다룬다. 이 부분은 인식 성능과 처리 시간에 많은 영향을 주는 부분이므로 효율적인 처리 방법이 요구된다.

본 논문은 인식 처리를 효율적으로 수행하기 위해 multiple-pass 처리 방법을 고려하였다. 인식 처리의 처리 속도를 개선하기 위해 mixture 기반의 warping factor의 estimation 방법이 여러 논문에서 소개되었지만, multiple-pass 처리 방법을 사용하였을 때 보다 낮은 인식 성능이 보였다.[1,2] 주파수 warping이 적용된 multiple -pass 처리는 다음과 같이 세 단계로 구성된다.

1) warping되지 않은 발성 음성의 특징 벡터 X_i에 대해 HMM decoding을 수행하여 가장 score가 높은 후보 단어 $\tilde{w}$를 얻는다.

2) 각각의 주파수 warping factor α가 적용된 특징 벡터 X_i^{α}에 대해, 식(7)을 이용하여 최적 주파수 warping factor $\hat{\alpha}$를 estimation한다.

$$\hat{\alpha} = \arg\max P(X_i^{\alpha} | \lambda, \tilde{w}) \quad , \text{for } \alpha \tag{9}$$

3) 최적 주파수 warping factor $\hat{\alpha}$가 적용된 특징 벡터 $X_i^{\hat{\alpha}}$ 에 대해, 1)의 단계를 다시 수행하여, 최종 인식 단어 $\hat{w}$를 결정한다.

$$\hat{w} = \arg\max P(X_i^{\hat{\alpha}} | \lambda, w) \quad , \text{for } w \tag{10}$$

III. 파워 스펙트럼 Warping

본 장에서는 화자 정규화의 성능 개선을 위해 논문에서 새롭게 제안하는 파워 스펙트럼 warping 방법에 대해 다룬다. 파워 스펙트럼 warping은 기존의 주파수 warping 방법과 비슷한 메카니즘을 가지며, MFCC에서 MFB의 조정에 의해 쉽게 구현될 수 있다. 기존의 주파수 warping 방법은 MFCC에서 MFB의 주파수 축의 warping을 수행하는 반면, 파워 스펙트럼 warping은 MFB의 파워 스펙트럼 축을 warping 한다.

일반적으로 동일한 단어를 발성 할 경우, 화자에 따라 성도 모양의 변이로 인해 포먼트 위치와 스펙트럼의 포락 정보가 서로 다르게 나타나는 것을 쉽게 확인 할 수 있다. 특히 그림 4와 같이 남성과 여성 화자 사이의 포먼트 위치와 스펙트럼 포락 정보가 크게 다르다. 기존의 주파수 warping 방법에서는 화자 사이의 성도 길이의 정규화를 위해 음성 스펙트럼의 포먼트 위치 정보에 중점을 두고 주파수 축의 스펙트럼 정보를 warping 하였다. 반면 파워 스펙트럼 warping 방법은 멜 스케일이 적용된 음성 스펙트럼의 파워 스펙트럼 축을 war-ping 함으로써, 화자 사이의 스펙트럼 포락 정보와 포먼트 위치까지 정규화를 수행한다.

1. 파워 스펙트럼 warping을 위한 MFB

파워 스펙트럼 warping은 MFCC에서 각 멜 필터 뱅크(MFB)의 파워 스펙트럼을 warping 함으로써 성도 변이를 정규화 할 수 있다. 파워 스펙트럼 warping

factor β에 따라 MFB를 선형 warping에 적용되는 함수의 결과 그래프는 그림 5와 같다. 주파수 warping과 대응해 보면, β 값이 1보다 작은 경우는 주파수 축을 압축하는 것과 같은 효과를 얻을 수 있고, 1보다 큰 경우는 주파수 축을 확장시키는 것과 같은 효과를 얻을 수 있다. 파워 스펙트럼 warping의 선형 warping 함수 $w_\beta(l)$를 표현하면 다음과 같다.

$$w_\beta(l)=\frac{\beta-1}{L}\cdot(l+1)+1$$
$$l=0,1,\ldots,L-1 \tag{11}$$

MFCC의 DCT 처리 부분에서 파워 스펙트럼 warping을 수행하기 위해 식(11)의 선형 warping 함수를 대입하면 다음과 같다.

$$c[j]=\sum_{l=0}^{L-1}\ln\left(\Im[l]^{w_\beta(l)}\right)\cdot\cos\left[\frac{\pi}{L}(j+0.5)\right]$$
$$j=0,1,\ldots,C-1 \tag{12}$$

효율적인 처리를 위해 식(12)에서 선형 warping 함수 $w_\beta(l)$를 로그 밖으로 빼면, 다음과 같이 다시 쓸 수 있다.

2. 파워 스펙트럼 warping factor의 estimation

$$c[j]=\sum_{l=0}^{L-1}w_\beta(l)\cdot\ln(\Im[l])\cdot\cos\left[\frac{\pi}{L}(j+0.5)\right]$$
$$j=0,1,\ldots,C-1 \tag{13}$$

주파수 warping에서 성도 길이에 따른 warping factor의 비율이 반비례의 성질을 갖는 것과는 달리, 파워 스펙트럼 warping은 성도 길이에 따라 warping factor가 비례 관계를 갖는다. 이것은 주파수 warping factor의 분포(그림6)와 파워 스펙트럼 warping factor의 분포(그림7)가 서로 상반되어 나타나기 때문이다. 파워 스펙트럼 warping factor를 estimation하는 과정은 주파수 warping과 동일하다. 파워 스펙트럼 warping이 적용된 특징 벡터 X_i^β는

$$X_i^\beta=\{x_i^\beta(0),\ldots,x_i^\beta(T)\} \tag{14}$$

이다. 화자 i의 최적 파워 스펙트럼 warping factor $\hat{\beta}$는 주파수 warping의 식(7)와 같이, HMM decoding에서 maximum likelihood 방법을 이용하여 얻을 수 있으며, 다음과 같다.

$$\hat{\beta}=\arg\max P(X_i^\beta|\lambda,W_i)\ ,\ \text{for}\ \beta \tag{15}$$

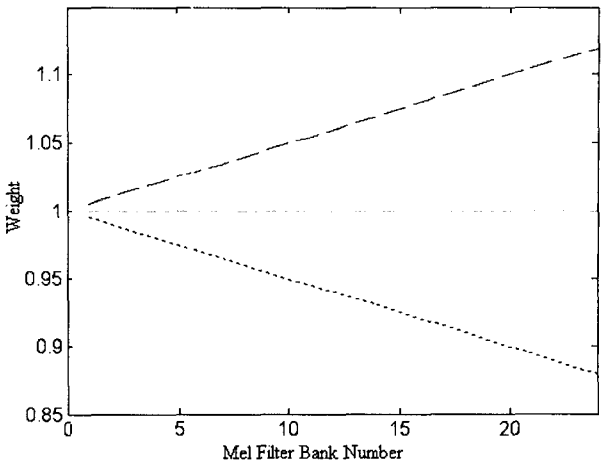

그림 5. 멜 필터 뱅크 순번 대 파워 스펙트럼의 가중치 값 (대시선: 최대 warping 가중치 값(β=1.12), 점-대시선: 기준 가중치 값(β=1.00), 점선: 최소 warping 가중치 값(β=0.88), 필터 뱅크의 개수 =24)

Fig. 5. Mel filter bank number vs. weight value of power spectrum(dash line: maximum warping weight value(β=1.12), dot-dash line: base weight value(β=1.00), dot line: minimum warping weight value(β=0.88), the number of filter bank = 24).

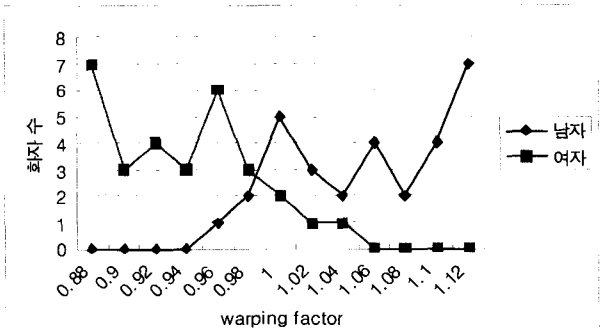

그림 6. 주파수 warping의 최적warping factor 분포

Fig. 6. Optical warping factor distribution of a frequency warping.

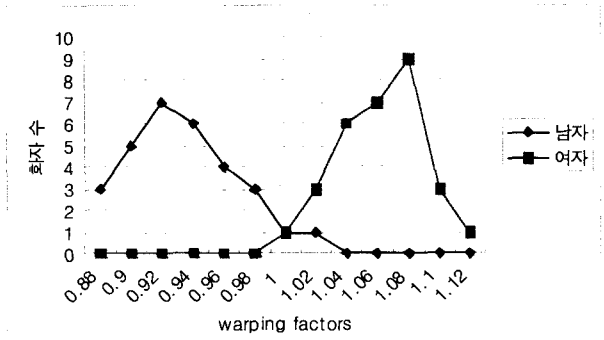

그림 7. 파워 스펙트럼 warping의 최적warping factor 분포

Fig. 7. Optical warping factor distribution of the power spectrum warping.

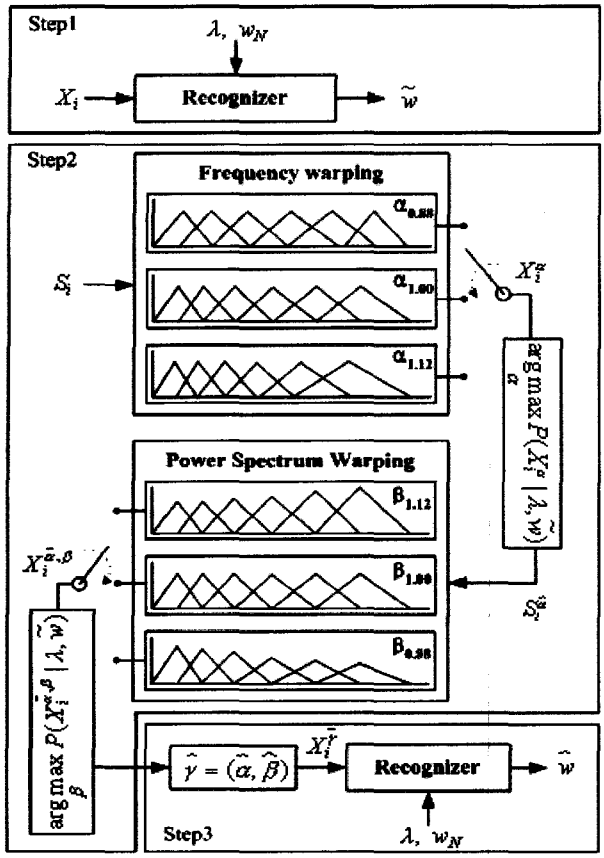

그림 8. Hybrid VTN을 위한 HMM 인식 처리
Fig. 8. HMM recognition procedures for hybrid VTN.

즉, 최적 주파수 warping factor $\tilde{\beta}$는 HMM 음향 모델 λ와 전체 인식 후보 단어 W에 대해, warping된 발성 음성과의 likelihood를 최대로 하는 β로 정의할 수 있다.

최적 파워 스펙트럼 warping factor를 구하기 한 탐색 범위는 다음과 같이 주파수 warping과 동일한 탐색 범위 값을 갖는다.

$$0.88 \le \beta \le 1.12 \quad , spaced \ 0.02 \tag{16}$$

3. 파워 스펙트럼 warping의 인식 처리

이 절에서는 화자 i의 발성 음성으로부터 앞 절에서 다룬 최적 파워 스펙트럼 warping factor $\tilde{\beta}$를 적용하여, 인식 처리를 수행하는 부분에 대해 다룬다. 이 부분은 2.3절의 주파수 warping의 음성인식 처리 부와 동일하게 multiple-pass 처리를 적용하였다. 파워 스펙트럼 warping이 적용된 multiple -pass 처리는 다음과 같이 세 단계로 구성된다.

1) warping되지 않은 발성 음성의 특징 벡터 X_i에 대해 HMM decoding을 수행하여 가장 score가 높은 후보 단어 $\tilde{w}$를 얻는다.

2) 각각의 파워 스펙트럼 warping factor β가 적용된 특징 벡터 X_i^{β}에 대해, 식(15)을 이용하여 최적 파워 스펙트럼 warping factor $\tilde{\beta}$를 estimation한다.

$$\tilde{\beta} = \arg max P(X_i^{\beta} | \lambda, \tilde{w}) \quad , \text{for } \beta \tag{17}$$

3) 최적 파워 스펙트럼 warping factor $\tilde{\beta}$가 적용된 특징 벡터 $X_i^{\tilde{\beta}}$ 에 대해, 1)의 단계를 다시 수행하여, 최종 인식 단어 $\hat{w}$를 결정한다.

$$\hat{w} = \arg max P(X_i^{\tilde{\beta}} | \lambda, w) \quad , \text{for } w \tag{18}$$

IV. Hybrid VTN

본 장에서는 화자 정규화의 성능 개선을 위해 본 논문에서 두 번째로 제안하는 화자 정규화 방법으로, 2장의 주파수 warping 방법과 3장의 파워 스펙트럼 warping을 결합한 hybrid VTN에 대해 다룬다.

1. Hybrid VTN warping factor의 estimation

Hybrid VTN을 수행하기 위한 MFCC의 MFB의 warping 처리는 2.1절의 주파수 warping과 3.1절의 파워 스펙트럼 warping에서 정의한 것과 동일하다. hybrid VTN은 이 두 가지의 방법을 고려한 최적 hybrid VTN warping factor $\hat{\gamma}$를 estimation 해야 한다. 이 두 가지의 최적 warping factor의 estimation 방법은 2.2절과 3.2절에서 다룬 내용과 동일하다. 하지만 이 두 가지의 warping factor를 동시에 estimation하는 것은 상당한 계산량이 요구되므로, 본 논문은 순차적으로 두 가지의 최적 warping factor($\hat{\alpha}, \tilde{\beta}$)를 estimation 한다.

Hybrid VTN warping factor γ의 estimation은 먼저 주파수 warping의 최적 warping factor를 estimation한다. 그 다음 파워 스펙트럼 warping의 최적 warping factor를 estimation 하도록 순차적적으로 수행한다. 이렇게 순서를 정한 이유는 주파수 warping이 성도 정규화를 위해 성도 길이(포먼트 중심 주파수)의 warping에만 국한되는 반면, 파워 스펙트럼 warping은 성도 길이뿐만 아니라 스펙트럼 포락 정보에 대한 정규화까지 고려해 줄 수 있기 때문이다.

화자 i의 최적 hybrid VTN warping factor $\hat{\gamma}$는 다음과 같이 표현할 수 있다.

$$\hat{\alpha} = \arg\max P(X_i^{\alpha}|\lambda, W_i) \quad \text{, for } \alpha$$

$$\hat{\beta} = \arg\max P(X_i^{\beta}|\lambda, W_i) \quad \text{, for } \beta$$

$$\hat{\gamma} = (\hat{\alpha}, \hat{\beta}) \tag{19}$$

2. Hybrid VTN의 인식 처리

앞 절에서는 최적 hybrid VTN warping factor $\hat{\gamma}$를 estimation하는 방법에 대해 다루었다. 이 절에서는 화자 i로부터 최적 hybrid VTN warping factor $\hat{\gamma}$를 적용하여, 인식 처리를 수행하는 부분에 대해 설명한다. Hybrid VTN의 인식 처리는 2.3절과 3.3절의 인식 처리부와 동일하게 multiple-pass 처리를 사용하였다.

hybrid VTN warping factor의 estimation이 순차적으로 수행되기 때문에 두 개의 multiple-pass 처리가 직렬로 연결되어 처리된다. 그림 8에 hybrid VTN의 인식 처리 과정의 전체 블록도를 보여주고 있다. Hybrid VTN이 적용된 multiple -pass 처리는 다음과 같이 세 단계로 구성된다.

1) Warping되지 않은 발성 음성의 특징 벡터 X_i 에 대해 HMM decoding을 수행하여 가장 score가 높은 후보 단어 $\tilde{w}$ 를 얻는다.

2) 먼저 각각의 주파수 warping factor α를 적용한 특징 벡터 X_i^{α} 에 대해 식(20a)를 적용하여 최적 주파수 warping factor $\hat{\alpha}$를 estimation한다. 그 다음 식(20a)로 얻은 최적 주파수 warping factor $\hat{\alpha}$를 적용한 특징 벡터 $X_i^{\hat{\alpha}}$ 에 각각의 파워 스펙트럼 warping factor β를 적용한다. 이것이 적용된 특징 벡터 $X_i^{\hat{\alpha}\cdot\beta}$ 에 대해 식(20b)를 적용하여 최적 파워 스펙트럼 warping factor $\hat{\beta}$를 estimation한다. 최종적으로 식(20a)와 식(20b)를 적용한 후, 최적 hybrid VTN warping factor $\hat{\gamma}$를 정의한다.

$$\hat{\alpha} = \arg\max P(X_i^{\alpha}|\lambda, \tilde{w}) \quad \text{, for } \alpha \tag{20a}$$

$$\hat{\beta} = \arg\max P(X_i^{\hat{\alpha}\cdot\beta}|\lambda, \tilde{w}) \quad \text{, for } \beta \tag{20b}$$

$$\hat{\gamma} = (\hat{\alpha}, \hat{\beta}) \tag{20c}$$

3) 최적 hybrid VTN warping factor $\hat{\gamma}$가 적용된 특징 벡터 $X_i^{\hat{\gamma}}$ 에 대해, 1)의 단계를 다시 수행하여, 최종 인식 단어 $\hat{w}$를 결정한다.

$$\hat{w} = \arg\max P(X_i^{\hat{\gamma}}|\lambda, w_i) \quad \text{, for } w \tag{21}$$

V. 실험 및 결과

본 장에서는 baseline 시스템을 기반으로 II장의 주파수 warping, III장의 파워 스펙트럼 warping, 그리고 IV장의 hybrid VTN에 대해, 각각의 인식 실험 결과를 비교 분석한다. 본 논문에서 사용한 baseline 시스템은 다음과 같이 구성되었다.

먼저 front end는 12차 MFCC 특징 벡터와 에너지 특징을 포함하여 기본 특징 벡터로 구성하였다. 또한 음성 신호의 dynamic한 특성을 고려하기 위해 1차와 2차 미분계수를 적용하여, 총 39차의 특징 벡터를 사용했다. 다음으로 벡터 양자화기(VQ;Vector Quantization)의 codebook 생성은 일반적인 LBG 알고리즘을 사용하였다.[5] VQ의 codebook 크기와 화자 정규화의 관련성을 고려해 주기 위해 VQ codebook의 codeword의 개수는 256과 512을 고려하였다. 즉, HMM 모델의 출력 확률의 component density가 256과 512를 갖는 음향 모델을 각각 구성하였다.

HMM 음향 모델과 training을 위해 한국어 단어 음성 DB, SKKU PBW을 사용했다. SKKU PBW(Phonetical Balanced Word) DB는 남자 60명, 여자 60명이 발성한 1001개의 단어로 구성된다. 표1은 training을 위해 사용된 각 DB의 내용이다. 모든 SKKU PBW DB는 11.025kHz로 샘플링 되었다. 전체 음성 DB의 절반은 training을 위해 사용되었고, 나머지 절반은 인식 실험

표 1. Training을 위한 한국어 음성DB
Table 1. Korean speech DB for training.

	Speakers (남자/여자)	Utterances (단어)	Training (남자/여자)
SKKU PBW	60/60	1001	30/15

표 2. 인식 실험을 위한 한국어 음성 DB - Training 음성 DB는 포함하지 않음
Table 2. Korean speech DB for recongition testing - No contains training speech DB.

	Speakers (남자/여자)	Utterances (단어)	Testing (남자/여자)
SKKU PBW	60/60	1001	30/30

표 3. 화자 정규화 작업의 성능 (단어 에러 율, SKKU PRW
Table 3. Performance of speaker normalization procedures.(word error rates(%), SKKU PRW).

Codebook Size	Baseline	α	β	γ
256	15.40	12.00	10.00	8.70
512	10.07	8.01	7.02	6.00

을 위해 사용하였다.

인식 성능 실험을 위해 사용한 SKKU PBW DB는 표2에 나타내었다. 본 논문에서 소개한 세 가지의 화자 정규화 방법에 대한 인식 실험 결과는 표3에 나타내었다. 표3의 인식 실험 결과를 보면, 크게 VQ의 codebook 크기에 따라 단어 에러 율의 차이를 보였다. 512 codebook을 기준으로 화자 정규화의 인식 성능을 분석해 보면, 화자 정규화가 적용되지 않은 baseline 시스템은 10.07%로 가장 높은 단어 에러 율을 보였고, 성도 정규화에 널리 사용되는 기존의 주파수 warping은 8.01%의 단어 에러 율을 보였다. 그리고 제안한 파워 스펙트럼 warping은 기존의 주파수 warping 보다 약 1.00% 정도 낮은 7.02%의 단어 에러 율을 보였다. 따라서 제안한 파워 스펙트럼 warping은 화자 정규화 방법에 있어 기존의 성도 정규화에 사용되는 주파수 warping 보다 성능이 우수한 특성을 보였다.

두 번째로 제안한 Hybrid VTN의 경우는 baseline 시스템 보다 4.07%의 단어 에러 율을 줄였고, 기존의 주파수 warping 보다 2.01%의 단어 에러 율이 줄었다. 그리고 파워 스펙트럼 warping 보다도 1.02%의 단어 에러 율을 줄였다. 이 결과는 hybrid VTN이 주파수 warping과 파워 스펙트럼 warping을 같이 사용함으로써 좀 더 강인하고 우수한 화자 정규화 특성을 보였다.

따라서 본 논문이 제안한 파워 스펙트럼 warping은 기존의 주파수 warping 보다 화자 정규화의 성능이 더 우수한 것으로 나타났으며, 두 방법을 결합한 hybrid VTN은 다른 화자 정규화 방법보다 더 강인한 화자 정규화 특성을 나타냈다.

VI. 결 론

본 논문은 현재 화자 정규화를 위해 널리 사용되고 있는 주파수 warping과 새롭게 제안한 파워 스펙트럼 warping, 그리고 이 두 가지의 방법을 결합한 hybrid VTN을 소개하였고, 한국어 음성 DB(SKKU PBW)로 각 화자 정규화 방법의 인식 성능을 비교하여 화자 정규화 특성을 분석하였다.

인식 실험 결과에 따르면, 제안한 파워 스펙트럼 warping은 기존의 주파수 warping 보다 약 1.00% 정도 낮은 단어 에러 율을 보여, 화자 정규화의 성능이 우수한 것으로 나타났다.

또한 hybrid VTN은 훨씬 강인한 화자 정규화 특성을 보였으며, 특히 baseline 시스템 보다 4.07%, 기존의 주파수 warping 보다 2.01% 낮은 단어 에러 율을 보여, 다른 화자 정규화 방법보다 성능이 우수한 것으로 나타났다. 하지만 hybrid VTN은 두 개의 화자 정규화 방법을 결합하여 사용하기 때문에 처리 시간이 종전보다 1.5배정도 더 소용되는 문제점이 있다. 향후 이 부분의 처리를 위한 fast hybrid VTN의 연구가 필요하다.

참 고 문 헌

[1] Li Lee, Richard Rose, "A Frequency Warping Approach to Speaker Normalization", IEEE Tra-nsactions on Speech and Audio Processing, Vol 6. No. 1, January 1998.

[2] L. Welling, H. Ney, S. Kanthak, "Speaker Adap-tive Modeling by Vocal Tract Normalization", IEEE Transaction on Speech and Audio Proce-ssing, Vol. 10, No. 6, September 2002.

[3] A. Andreou, T. Kam, and J. Cohen, "Experi-ments in Vocal Tract Normalization", in Proc. CAIP Workshop: Frontiers in Speech Recognition II, 1994.

[4] Michael Seltzer, "SPHINX III Signal Processing Front End Specification", CMU Speech Group, August 1999.

[5] Y. Linde, A. Duzo, R. M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Transaction on COM., Vol. 28, January 1980.

[6] J.S. Youn, K.W. Chung and K.S. Hong, "A Con-tinuous Digit Speech Recognition Applied Vowel Sequence and VCCV Unit HMM", Pro-ceeding of the Acoustical Society of Korea, Vol. 20, No. 2, 2001.

[7] T.D. Rossing, P. Wheeler and F.R. Moore, "The Science of Sound", Addition Wesley, 2002.

[8] R. Roth et al, "Dragon systems' 1994 Large Vocabulary Continuous Speech Recognizer", in Proc. Spoken Language Systems Technology Workshop, 1995.

저 자 소 개

유 일 수(학생회원)
2002년 강릉대학교 제어계측공학과 학사 졸업. (공학사)
2004년 성균관대학교 전기전자및컴퓨터공학과 석사 졸업. (공학석사).
현재 인피닉스(주) 근무
<주관심분야: 음성인식, 영상처리, SoC >

김 동 주(학생회원)
1998년 충북대학교 전파공학과 졸업. (공학사)
2000년 충북대학교 전파공학과 졸업. (공학석사)
2001년-현재 성균관대학교 정보통신공학부 박사과정
<주관심분야: 음성인식, 음성코딩, 신호처리>

노 용 완(학생회원)
2001년 남서울대학교 정보통신공학과 졸업. (공학사)
2003년 성균관대학교 정보통신공학부 졸업. (공학석사)
2003년-현재 성균관대학교 정보통신공학부 박사과정.
<주관심분야: 음성인식, 음성이해, 신호처리>

홍 광 석(정회원)
1985년 성균관대학교 전자공학과 졸업. (공학사)
1988년 성균관대학교 전자공학과 졸업. (공학석사).
1992년 성균관대학교 전자공학과 박사 졸업. (공학박사)
1990년~1993년 서울보건전문대학 전산정보처리과 전임강사
1993년~1995년 제주대학교 정보공학과 전임강사
1995년~현재 성균관대학교 정보통신공학부 교수
<주관심분야: 음성인식 및 합성, HCI>