파워 스펙트럼 warping을 이용한 성도 정규화

Vocal Tract Normalization Using The Power Spectrum Warping

유 일 수, 김 동 주, 노 용 완, 홍 광 석

성균관대학교 정보통신공학부(전화:(031)290-7196, 팩스:(031)290-7191, E-mail : gildda@nate.com)

Abstract : The method of vocal tract normalization has been known as a successful method for improving the accuracy of speech recognition. A frequency warping procedure based low complexity and maximum likelihood has been generally applied for vocal tract normalization. In this paper, we propose a new power spectrum warping procedure that can be improve on vocal tract normalization performance than a frequency warping procedure. A mechanism for implementing this method can be simply achieved by modifying the power spectrum of filter bank in Mel-frequency cepstrum feature(MFCC) analysis. Experimental study compared our proposal method with the well-known frequency warping method. The results have shown that the power spectrum warping is better 50% about the recognition performance than the frequency warping.

Keywords : MFCC, Speaker Normalization, Vocal Tract Normalization, The Power Spectrum Warping

I. 서 론

일반적인 음성인식시스템은 화자 독립형 시스템을 지향하고 있으며, 한정된 훈련 모델에 의존한다. 이 경우 화자들 사이의 성도 모양의 변이로 인해 음성인식시스템의 인식 성능 저하에 많은 영향을 받는다. 이 문제점을 보완하기 위해 화자 정규화 방법들이 여러 논문에서 소개되었으며, HMM 기반 음성인식시스템의 성능 개선을 보여주고 있다.[1,2,3] 성도 모양의 변이를 정규화 하는 것을 성도 정규화(VTN; Vocal Tract Normalization)라고 하며, 화자들 사이의 성도 길이의 변이와 밀접한 관련이 있다. 특히, 성별에 따라 성도 길이의 차이가 크게 나타난다. 조사된 결과에 의하면, 성인의 화자 별 성도 길이의 범위는 13cm에서 18cm 정도의 변이를 보인다. 이것은 화자들 사이에서 포먼트 중심 주파수의 변이가 25% 정도 차이를 보이는 것과 같다.[1] 성도 정규화를 위한 방법으로 스펙트럼 분석내의 주파수 warping 방법이 여러 논문에서 소개되었다. 이 방법은 멜 주파수 켑스트럼(MFCC) 특징 분석에서 멜 필터 뱅크(MFB; Mel Filter Bank)의 주파수 축의 선형 warping을 통하여 간단히 구현된다.[1,2]

본 논문에서는 성도 정규화 방법으로 주파수 warping 방법보다 개선된 새로운 파워 스펙트럼 warping 방법을 제안한다. 제안하는 파워 스펙트럼 warping은 기존의 주파수 warping 방법과 비슷하게 MFCC 특징 분석에서 MFB의 조절에 의해 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.

기존의 주파수 warping 방법에서 중요하게 다루어 졌던 부분은 크게 두 가지로 요약할 수 있다. 첫 번째는 warping factor를 estimation 하는 것이며, 두 번째는 이것을 효율적으로 인식과정에서 처리하는 것이다. 제안하는 파워 스펙트럼 warping방법도 이 두 가지 요소를 중요하게 고려하였다.

마지막으로 우리가 제안하는 화자 정규화의 성능 평가를 위해, 한국어 단어 단위 음성 DB(SKKU PBW DB)에 대해, baseline 시스템과 주파수 warping의 인식 성능을 바탕으로 파워 스펙트럼 warping에 대한 각 인식 성능을 비교 분석하여 화자 정규화의 성능을 평가하였다.

II. 주파수 warping

본 장에서는 화자 정규화를 위해 널리 사용되고 있는 주파수 warping 접근 방법에 대해 Li Lee 외[1]와 L. Welling 외[2] 의해 소개된 내용을 다룬다.

2.1 주파수 warping을 위한 MFB

일반적으로 성도 정규화는 front end에서 특징 벡터의 변형으로 수행된다. 특히, 주파수 warping 방법은 스펙트럼 분석 기반 음성 특징 분석 방법으로 널리 사용되는 MFCC[4] 처리 과정을 이용하여 쉽게 구현될 수 있다[1]. 주파수 warping은 주파수 축을 warping 함으로서 성도 길이의 변이를 정규화 하는 방법이다. 이것은 MFCC 처리 과정에서 MFB 부분을 수정함으로써 주파수 축의 warping이 이루어진다. 일반적으로 주파수 warping은 선형 warping을 사용하지만, 주파수 waripng의 비선형적 특성

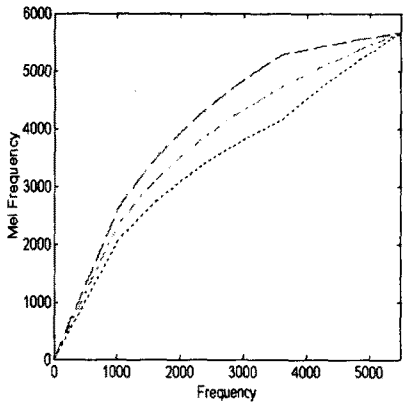

그림 1. 선형 주파수 대 구분적(piecewise) 선형 warping 멜 주파수(점선; 최대 warping 멜(α=1.12), 점-대시선: 기준 멜(α=1.00), 대시선: 최소 warping 멜(α=0.88))

을 고려해주기 위해 구분적(piecewise) 선형 warping 방법이 고안되었고, 일반적인 선형 warping 보다 더 robust 한 인식 성능을 보여주는 것으로 알려져 있다.[8] 구분적 선형 warping 방법이 적용된 멜 주파수는 그림 1과 같이 나타난다. 본 논문에서는 주파수 warping 처리를 위해 구분적 선형 warping을 사용하였으며, 다음과 같이 표현된다.

$$\begin{aligned} \tilde{f} &= 2595\log\left(1+\frac{f}{700}\right) \\ Mel^{\alpha}(f) &= \alpha\cdot\tilde{f} \quad ,0\le f\le f_0 \\ Mel^{\alpha}(f) &= \frac{\tilde{f}_{\max}-\alpha\cdot\tilde{f}_0}{\tilde{f}_{\max}-\tilde{f}_0}(\tilde{f}-\tilde{f}_0) \\ &\quad +\alpha\cdot\tilde{f}_0 \quad ,f_0\le f\le f_{\max} \end{aligned} \tag{1}$$

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

$$\begin{aligned} \tilde{S}^{\alpha}[l] &= \sum_{k=0}^{k=K/2} S[k]\cdot M_l^{\alpha}[k] \\ l &= 0,1,\ldots,L-1 \end{aligned} \tag{2}$$

여기서 $S[k]$는 파워 스펙트럼을, $M_l[k]$는 멜 삼각 필터 뱅크를, α는 주파수 warping factor를, L은 멜 삼각 band-pass 필터의 개수를, K는 FFT의 resolution의 값이다.

2.2 주파수 warping factor의 estimation

주파수 warping factor의 estimation은 성도 길이 정규화의 최적 주파수 warping 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.

화자 i의 최적 주파수 warping factor $\hat{\alpha}$는 다음과 같이 HMM decoding 단계에서 maximum likelihood 방법을 적용하여 얻을 수 있다.

$$\hat{\alpha} = \arg\max P(X_i^{\alpha}|\lambda, W) \quad ,\text{for } \alpha \tag{3}$$

여기서 X_i^{α} 는 주파수 warping factor가 적용된 특징 벡터를, W 은 전체 인식 후보 단어(transcription set)를, λ는 HMM 음향 모델 을 말한다.

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

$$0.88\le\alpha\le1.12 \quad ,spaced\ 0.02 \tag{4}$$

2.3 주파수 warping의 인식 처리

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 이 부분을 수행하기 위해 주파수 warping방법에서 많이 사용되는 multiple-pass 처리 방법을 고려하였다. 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(3)을 이용하여 최적 주파수 warping factor $\hat{\alpha}$를 estimation한다.

$$\hat{\alpha} = \arg\max P(X_i^{\alpha}|\lambda, \tilde{w}) \quad ,\text{for } \alpha \tag{5}$$

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.

$$\hat{w} = \arg\max P(X_i^{\hat{\alpha}}|\lambda, W) \quad ,\text{for } w \tag{6}$$

III. 파워 스펙트럼 warping

본 장에서는 본 논문에서 새롭게 제안하는 파워 스펙트럼 warping 방법에 대해 다룬다. 파워 스펙트럼 warping은 기존의 주파수 warping 방법과 비슷하게 MFCC 처리에서 MFB의 조정에 의해 쉽게 구현된다. 기존의 주파수 warping 방법은 MFB의 주파수 축을 warping하는 반면

파워 스펙트럼 warping은 MFB의 파워 스펙트럼 축을 warping한다.

제안하는 파워 스펙트럼 warping 방법은 MFB의 파워 스펙트럼 축을 warping 함으로써, 성도 모양의 정규화가 가능하다. 즉, 화자 사이의 포먼트 특성뿐만 아니라 스펙트럼 포락 정보까지 정규화가 가능하다.

3.1 파워 스펙트럼 warping을 위한 MFB

파워 스펙트럼 warping은 MFCC 처리 과정에서 MFB 처리 부분의 각 band-pass 필터의 파워 스펙트럼 축을 선형 warping 함으로써 이루어진다. 파워 스펙트럼 factor β에 따라 MFB를 선형 warping 함수의 그래프는 그림 2와 같고, 선형 warping 함수를 수식적으로 표현하면 다음과 같다.

$$w_\beta(l) = \frac{\beta - 1}{L} \cdot (l+1) + 1 \qquad l = 0, 1, \ldots, L-1 \tag{7}$$

여기서 β는 파워 스펙트럼 warping factor이다. 파워 스펙트럼 warping은 MFCC의 DCT(Discrete Cosine Transform)부분에서 적용되며, 선형 warping 함수 $w_\beta(l)$를 대입하면 다음과 같다.

$$c[j] = \sum_{i=0}^{L-1} \ln\left(S[l]^{w_\beta(l)} \right) \cdot \cos\left[\frac{\pi}{L}(j+0.5) \right] \qquad j = 0, 1, \ldots, C-1 \tag{8}$$

3.2 파워 스펙트럼 warping factor의 estimation

파워 스펙트럼 warping factor는 기준 성도 모양(기준 HMM 음향 모델)과 특정 화자의 성도 모양 사이의 비율로 나타낼 수 있다. 실험 결과에 의하면 파워 스펙트럼 warping factor는 성도 길이와 비례 관계를 갖는다.

최적 warping factor의 estimation은 2.2절의 주파수 warping과 동일하게 적용되며, 화자 i의 최적 파워 스펙트럼 warping factor $\hat{\beta}$는 다음과 같다.

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

여기서 X_i^β 는 파워 스펙트럼 warping factor가 적용된 특징 벡터이다. 최적 파워 스펙트럼 warping factor를 구하기 위한 탐색 범위는 주파수 warping과 동일하다.

$$0.88 \le \beta \le 1.12 \quad , spaced\ 0.02 \tag{10}$$

3.3 파워 스펙트럼 warping의 인식 처리

이 절에서는 화자 i의 발성 음성으로부터 앞 절에서 다룬 최적 파워 스펙트럼 warping factor $\hat{\beta}$를 적용하여, 인식 처리를 수행하는 부분에 대해 다룬다. 이 부분은 2.3절의 주파수 warping의 음성인식 처리 부분과 동일하게 multiple-pass 처리를 사용하였다.

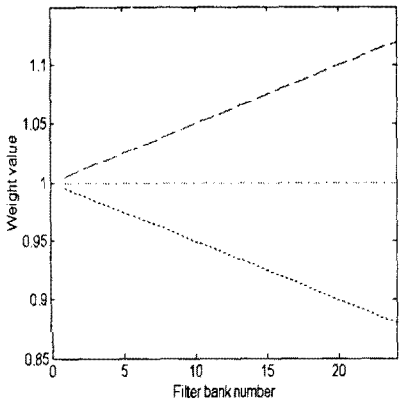

그림 2. 멜 필터 뱅크 순번 대 파워 스펙트럼의 가중치 값 (대시선: 최대 warping 가중치 값(β=1.12), 점-대시선: 기준 가중치 값(β=1.00), 점선: 최소 warping 가중치 값(β=0.88), 필터 뱅크의 개수=24)

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^β에 대해, 식(9)을 이용하여 최적 파워 스펙트럼 warping factor $\hat{\beta}$를 estimation한다.

$$\hat{\beta} = \arg\max P(X_i^\beta | \lambda, \tilde{w}) \quad , \text{for } \beta \tag{11}$$

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

$$\hat{w} = \arg\max P(X_i^{\hat{\beta}} | \lambda, W) \quad , \text{for } w \tag{12}$$

IV. 실험 및 결과

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] 생성된 VQ codebook의 code-word의 개수는 256과 512을 고려하였다.

HMM 음향 모델과 training을 위해 한국어 단어 음

표 1. Training과 인식 실험을 위한 한국어 음성 DB (SKKU PBW DB)

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

표 2. SKKU PRW DB에 성도 정규화가 적용된 인식 실험 결과(단어 에러 율(%))

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

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

본 논문에서 소개한 두 가지의 성도 정규화 방법에 대한 인식 실험 결과는 표2에 나타내었다. 표2의 인식 실험 결과를 보면, 크게 VQ의 codebook 크기에 따라 단어 에러 율의 큰 차이를 보였다. 여기서 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 warping보다 0.99% 낮은 7.02%의 단어 에러 율을 보였다.

V. 결 과

화자 정규화를 위한 기존의 성도 정규화 방법으로 널리 사용되는 주파수 warping과 새롭게 제안하는 파워 스펙트럼 warping을 소개하였고, 한국어 음성 DB(SKKU PBW DB)로 인식 성능을 비교 평가하였다.

인식 실험 결과에 따르면 제안한 파워 스펙트럼 warping은 기존의 주파수 warping 방법의 8.01% 보다 최소 1.00%정도 더 낮은 7.02% 단어 에러 율을 보였다. 또한 처리 속도 면에서도 주파수 Warping보다 20%정도 더 빠른 실험 결과를 보였다.

따라서 제안한 파워 스펙트럼 warping 방법은 기존의 주파수 warping 방법 보다 50% 정도의 성도 정규화의 성능 개선 가져왔다. 이와 같은 성능 개선은 제안한 파워 스펙트럼 warping이 성도 모양의 정규화를 수행한 결과로 보인다.

<감사의 글>

본 연구는 한국과학재단 목적기초연구(R05-2002-000-01007-0)지원으로 수행되었음.

참 고 문 헌

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

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

[3] A. Andreou, T. Kam, and J. Cohen, "Experiments 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 Continuous Digit Speech Recognition Applied Vowel Sequence and VCCV Unit HMM", Proceeding 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.