2001年度 第14回 信號處理合同學術大會 論文集 第14卷 1號

VTN을 이용한 화자 정규화에 관한 연구

손 창 희, 손 종 목, 배 건 성
경북대학교 전자·전기공학부

A Study on Speaker Normalization using VTN

Chang Hee Son, Jong Mok Son, Keun Sung Bae
School of the Electronic & Electrical Engineering, Kyungpook National University
chson@mir.knu.ac.kr

요 약

본 연구에서는 화자에 따라 서로 다른 성도의 길이에 의해 발생하는 음성인식 시스템의 성능 저하를 줄이기 위하여, VTN(Vocal Tract Normalization)을 음성인식 시스템에 적용하고, 주소 인식 실험을 통하여 인식 성능을 평가하였다. 또, VTN을 CMN과 동시에 적용하여 인식 실험을 하였다. 실험에서는 화자간 성도 길이의 차이를 반영하기 위하여 13개의 Warping 계수에 대해 필터 뱅크를 이용한 선형 Warping 방법을 적용하였다. 실험결과, Baseline 인식 시스템에 비하여 VTN을 적용하면, WER(Word Error Rate)이 1.24% 감소하였고, CMN과 VTN을 동시에 적용한 실험에서는 Baseline 인식 시스템과 비교하여 WER이 0.33% 감소하였지만 VTN을 적용한 실험결과와 비교하면 오히려 0.91% 증가하였다.

I. 서 론

화자, 문맥, 그리고 환경변화는 음성인식 시스템의 성능을 변화시키는 주요 요인이다. 이들 요인 중에서 화자간의 차이는 화자의 감정과 성도의 길이, 그리고 강세 및 억양 등에 의해 발생한다. VTN은 화자에 따라 서로 다른 성도의 길이를 정규화 함으로써 음성 인식 시스템의 화자 의존적인 성질을 줄이기 위하여 적용되는 기법이다[1].

VTN을 음성인식에 적용할 때 고려해야 할 사항은 성도 길이의 차이를 반영하는 Warping 함수 및 계수의 설정, 화자에 따른 적절한 Warping 계수의 추정, 그리고 추정된 값을 이용한 Warping 방법 등이 있다[2][3].

새로운 화자의 주파수 스펙트럼을 기준 스펙트럼으로 Warping 하는 함수에는 선형 또는 비선형 함수가 사용될 수 있다. 선형 함수는 구현이 용이한 반면에 대역폭 부정합(Bandwidth mismatch) 문제를 일으키는 단점이 있으며, 부분선형(Piecewise linear) 및 양선형(Bilinear) 등의 비선형(Nonlinear) 함수는 계산과정이 복잡해진다[1].

훈련 및 인식 과정에서 Warping 계수의 추정은 HMM 모델을 이용하는 방법과 Gaussian Mixture 모델을 이용하는 방법이 있다. HMM 모델을 이용한 추정 방법은 구현이 간단하고 인식 성능이 우수 하지만, 응답시간이 길다. 반면, Gaussian Mixture 모델을 이용한 추정 방법은 응답시간은 짧지만, 인식 성능이 떨어진다. Warping 방법에는 Warping 함수를 이용하여 음성 신호 스펙트럼의 주파수 축을 직접 Warping하는 방법과 신호의 스펙트럼은 고정한 채 필터 뱅크를 구성하는 각 필터의 대역폭을 조정하여 Warping 하는 방법이 있다. 필터 뱅크를 이용한 Warping 방법은 주파수 Warping에 비해 비교적 구현이 간단하고 효과적이다[2].

본 연구에서는 HMM 모델을 이용한 Warping 계수 추정 방법과 필터 뱅크를 이용한 Warping 방법을 이용하여 VTN을 음성인식 시스템에 적용하여 보았다. 그리고, 일반적으로 채널특성에 의해 발생하는 잡음 성분을 제거하고, 화자간의 변이를 줄이기 위하여 사용되는 CMN을 VTN과 동시에 적용하여 주소 인식 실험을 수행하고, 그 결과를 비교하였다.

본 논문의 구성은 다음과 같다. I 장의 서론에 이어, II장에서는 필터 뱅크를 이용한 Warping 방법과

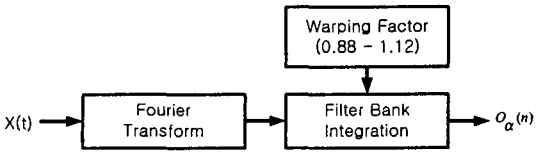

그림 1. 필터 뱅크를 이용한 Warping

실험에 적용한 훈련 및 인식 알고리듬에 대해 설명한다. 그리고, Ⅲ장에서는 Baseline 인식 시스템과 실험에서 사용한 DB에 대해 설명하고, 실험 내용 및 실험 결과를 제시하며, Ⅳ장에서 결론을 맺는다.

Ⅱ. VTN(Vocal Tract Normalization)

성도는 성문(Glottis)으로부터 입술까지이며, 그 길이는 사람에 따라 다르다. 일반적으로, 성도의 길이는 성인 남성의 경우 약 18cm이고, 성인 여성의 경우에는 약 13cm이다. 따라서, 성도의 길이에 반비례하는 포만트 주파수는 화자에 따라 최대 약 25%의 차이가 난다[3]. 동일한 발성에 대해 화자마다 다르게 나타나는 포만트 주파수를 Warping 함수에 의해 Warping된 주파수 축에 재배열하여, 기준 포만트 주파수와의 차이를 줄임으로써 성도 길이의 차이에 의해 발생하는 화자간 변이를 줄일 수 있다. 본 논문에서는 Warping 계수 α를 $0.88 \le \alpha \le 1.12$ 사이에서 0.02씩 증가시키면서 전체 13개의 값으로 설정하였다. 또, Warping 계수 α를 이용한 주파수 축 Warping은 그림 1과 같이 필터 뱅크를 이용한 Warping 방법을 적용하였다[1]. 입력 신호 $x(t)$는 푸리에 변환(Fourier transform) 후, 각 필터의 주파수 대역이 각각의 Warping 계수에 의해 재배열된 필터 뱅크를 통과한다. 필터 뱅크의 출력을 Warping 계수 α에 대해 $O_\alpha(n)$으로 표시하면, $O_\alpha(n)$은 식 (1)과 같이 주어진다.

$$O_\alpha(n) = \sum_{w=l_\alpha(n)}^{w=h_\alpha(n)} T_n(\omega)X(\omega), \quad 0 \le n \le N-1 \qquad (1)$$

여기서, $T_n(\omega)$는 필터 뱅크의 n번째 필터의 전달 함수이고, $X(\omega)$는 입력신호의 주파수 스펙트럼이다. $h_\alpha(n)$, $l_\alpha(n)$은 Warping 계수 α로 스케일 된 필터 뱅크의 n번째 필터의 상하 차단 주파수를 나타내며, N은 필터 뱅크를 구성하는 필터의 수를 나타낸다.

본 연구에서는 식 (2)로 구한 멜 스케일의 필터 뱅크를 사용하였고, 특정 화자에 대한 멜 스케일은 식 (3)을 이용하여 구하였다. 식 (3)에서 α_0는 1400π이며, α는 각 화자의 Warping 계수를 나타낸다. Warping 계수 α에 대한 멜 스케일 $M_\alpha(\omega)$는 식 (2)의 멜 스케일 $M(\omega)$를 Warping 계수 α로 선형 Warping 하여 구해진다. 즉, $M_\alpha(\omega) = M(\alpha\omega)$와 같다.

$$M(\omega) = 2595 \log_{10}(1 + \omega/\alpha_0) \qquad (2)$$

$$M_\alpha(\omega) = 2595 \log_{10}(1 + (\omega\alpha)/\alpha_0) \qquad (3)$$

식 (3)에서 구한 Warping 된 멜 스케일을 이용하여, 식 (1)의 필터 뱅크를 구성하는 각 필터의 상하 차단 주파수, $h_\alpha(n)$, $l_\alpha(n)$를 구하였으며, 필터 뱅크의 각 필터는 구형 필터를 이용하였고, 각 필터에서의 출력을 대역폭으로 나누어 정규화 하였다. 특징 파라미터는 화자의 동일한 입력 음성에 대해 각각의 Warping 계수에 따라 추출하였다[2][3].

Warping 계수의 추정은 비교적 구현이 간단하고, 인식 성능이 좋은 HMM 모델을 기반으로 하였다. HMM 모델 기반의 Warping 계수 추정 방법은 주어진 입력 음성과 주어진 모델에 대해 계수별 확률 값을 구하여, 최대 확률 값을 갖는 계수를 그 화자에 대한 Warping 계수로 설정한다. 훈련 과정은 그림 2와 같다. ETRI 445DB의 각 화자에 대한 Warping 계수를 추정하기 위하여 전체 445 단어에 대하여, 정규화 되지 않은 초기 HMM모델과 Transcription 정보를 이용하여 Warping 된 각각의 관측 벡터 열에 대해 확률 값을 구한 후, 13개의 Warping 계수에 대해 최대 확률 값을 갖는 계수를 그 화자에 대한 Warping 계수로 설정하였다. 그리고, 화자에 따라 구해진 Warping 계수로 Warping 한 데이터를 이용하여 모델을 훈련하고, HMM 모델을 갱신하였다. 이렇게 새롭게 만들어진 모델을 이용하여, Warping 계수를 추정하고, 모델을 갱신하는 과정을 5번 반복 수행하였다.

그림 3에 VTN을 적용한 인식과정을 나타내었다. 먼저, 입력 음성에 대하여 초기 Warping 계수를 1로 두고 훈련과정에서 정규화된 HMM 모델을 이용하여 초기 인식과정을 거쳐, 입력 문장을 구성하는 단어의 Transcription과 상태 정보를 구한다. 그리고, 구해진 Transcription과 상태 정보, 정규화된 HMM 모델을 이용하여 Warping 계수에 대해 각 단어별 확률 값을 구하고, 단어별 확률 값을 더하여 문장에 대한 전체 확률 값을 구한다. 이때, 각 Warping 계수와 문장에 대해 최대 확률 값을 갖는 계수를 그 화자에 대한 Warping 계수로 설정하였다. 또, 구해진 Warping 계수를 이용하여, 입력 신호를 Warping 하고 정규화된 HMM 모델을 이용하여 재인식한다.

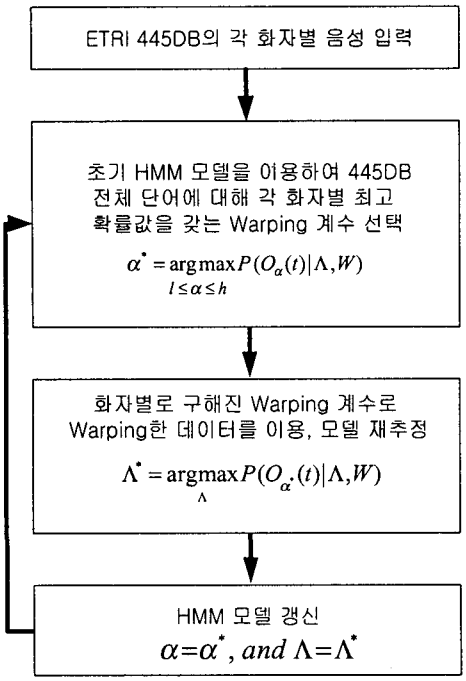

그림 2. VTN을 적용하기 위한 훈련 과정

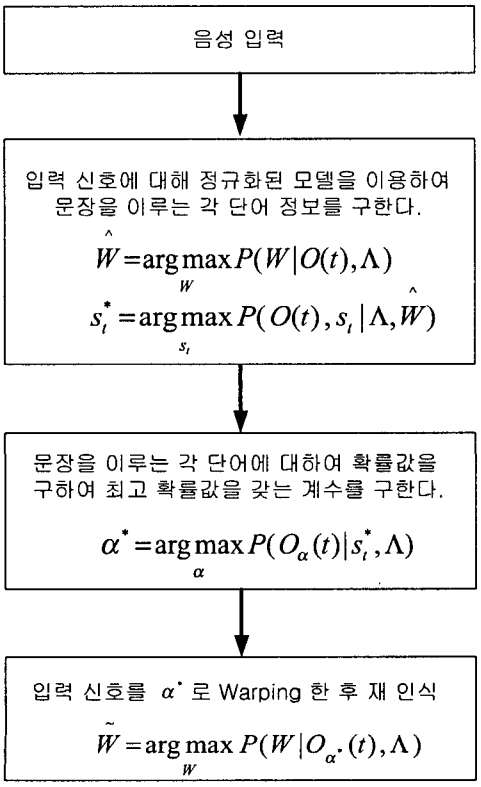

그림 3. VTN을 적용한 인식 과정

Ⅲ. 실험 결과 및 검토

본 논문에서 사용한 음성인식 시스템은 연속 HMM을 이용한 음소 단위 가변어휘 인식 엔진이다. 입력 음성에 대해 프리엠프시스 계수 0.97로 전처리 한 후, 20ms 길이의 해밍 윈도우를 10ms 간격으로 이동하며 구간단위로 분석하였다. 19차의 멜 필터 뱅크를 사용하였고, 각각 13차의 MFCC, ΔMFCC, ΔΔMFCC를 구하여, 전체 39차의 특징 파라미터를 사용하였다. 인식 단위는 50개의 유사음소단위(Phoneme likely unit)이며, HMM은 3상태 Bakis 모델을 기반으로 하였다. 그리고, Mixture의 수를 줄이기 위하여, PTM(Phonetic-Tied Mixture)을 적용하였다[4].

훈련데이터는 ETRI 445DB의 남녀 각각 16, 14명이 발성한 445 × 30 단어를 이용하였다. 그리고, 인식 실험에 사용된 데이터는, 6개 광역시의 각 5개 구, 4개 동을 무작위로 추출하여 그림 4의 예와 같이 구성된 120 문장을 남녀 10명의 화자가 연구실 환경에서 8kHz, 16bit로 녹음한 데이터를 사용하였다.

예) 대구광역시 북구 산격동

그림 4. 실험 데이터 예

실험은 CMN과 VTN을 모두 적용하지 않은 인식 시스템을 기본 Baseline으로 하고, Baseline 인식 시스템에 VTN을 적용한 인식 실험을 하였다. 그리고, VTN을 CMN과 동시에 적용하여 동일한 주소 DB에 대한 인식 실험을 하였다.

표 1 의 남녀 10명이 발성한 1200문장에 대한 실험 결과를 보면, Baseline 인식 시스템에 VTN을 적용하면, WER이 1.24% 감소하였고, CMN과 VTN을 동시에 적용한 실험에서는, VTN만 적용한 결과보다 오히려 WER이 0.91% 증가하였다. 일반적인 기대와는 달리 환경변화 및 잡음의 영향을 줄이기 위해 주로 사용되는 CMN 기법과 화자 정규화에 적합한 것으로 알려진 VTN 기법의 두 가지 알고리듬을 동시에 적용한 실험 결과가 VTN만 적용한 실험결과보다 인식율이 나쁘게

표 1. 실험 결과

	Baseline	VTN	VTN & CMN
WER(%)	8.83	7.59	8.50

나타났다. 또한, 실험에서 HMM 모델을 기반으로 한 Warping 계수 추정 방법을 이용하여 VTN을 적용하였는데, Baseline 인식 시스템에 비해 인식 성능은 향상되었지만 응답 시간이 길어 실시간 구현이 어려웠다.

Ⅳ. 결 론

본 연구에서는 화자에 따른 성도 길이의 차이에 의해 발생하는 음성인식 시스템의 성능 저하를 줄이기 위하여, VTN을 음성인식 시스템에 적용하고, 주소 인식 실험을 통하여 인식 성능을 평가하였다.

HMM 모델을 이용한 Warping 계수 추정 방법을 적용하였고, 필터 뱅크를 이용한 선형 Warping 방법을 사용하였다. 남녀 10명의 화자가 연구실 환경에서 발성한 1200문장에 대해 인식 실험한 결과, Baseline 인식 시스템에 비하여, VTN을 적용하면 WER이 1.24% 감소하였고, CMN과 VTN을 동시에 적용한 실험에서는 Baseline 인식 시스템에 비하여, WER이 0.33% 감소하였으며, VTN을 적용한 실험결과에 비해서는 0.91% WER이 증가하였다. 앞으로, VTN을 음성인식 시스템에 적용하기 위해서 인식 성능 향상을 위한 화자 적응 모델 훈련 기법의 적용[5][6]과 VTN의 실시간 적용[7][8]에 대한 연구와 더불어 앞서 언급한 문제점에 대한 연구를 계속하고자 한다.

참고 문헌

[1] Puming Zhan, A. Waibel, "Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition", School of Computer Science Carnegie Mellon Univ. May 1997.

[2] L. Lee, R. Rose, "A frequency Warping approach to speaker normalization", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, pp.49-60, Jan. 1998.

[3] L. Lee R. Rose, "Speaker normalization using efficient frequency Warping procedures", in Proc. ICASSP Vol. 1, pp.353-356, Atlanta, GA, May 1996.

[4] Akinobu Lee, Tatsuya Kawahara, Kazuya Takeda and Kiyohiro Shikano, "A New Phonetic Tied-Mixture Model for Efficient Decoding," International Conference on Acoustic, Speech and Signal Processing, vol. 3, pp. 1269-1272, 2000.

[5] S. Molau, S, Kanthak, H.Ney, "Efficient Vocal Tract Normalization in Automatic Speech Recognition"

[6] L. Welling, R. Haeb-Umbach, X. Aubert, N. Haberland, "A study on speaker normalization using vocal tract normalization and speaker adaptive training", in proc. ICASSP, vol.2, pp. 797 - 800, 1998.

[7] L .Welling, S. Kanthak, H. Ney, "Improved Methods For Vocal Tract Normalization", Proc. ICASSP, Vol. 2 pp.761-764, Jan. 1999.

[8] A. Sixtus, S. Molau, S. Kanthak, R. Schluter, H. Ney, "Recent improvements of the RWTH large vocabulary speech recognition system on spontaneous speech", proc. ICASSP, vol. 3, pp. 1671 -1674, 2000.