2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

강인한 화자 확인 시스템을 위한 World 모델을 이용한 켑스트럼 정규화 연구

김유진, 정재호
인하대학교 전자공학과, 디지털 신호 처리 연구실

A Study of Cepstrum Normalization Using World Model for Robust Speaker Verification

Yu-Jin Kim, Jae-Ho Chung
Digital Signal Processing Lab., Dept. of Electronic Engineering, INHA Univ.
g1982832@inhavision.ac.kr, jhchung@inha.ac.kr
Phone: +82-32-860-7420, Fax:+82-32-868-3654

Abstract

본 논문에서는 화자 확인 시스템와 등록과 확인 과정의 채널 환경 불일치로 성능이 저하되는 문제를 해결하기 위한 새로운 정규화 방법에 대해 설명한다.

제안된 방법은 첫째, 입력 음성으로부터 효과적으로 채널을 추정·보상하고 둘째, 스코어 정규화 과정에서 사칭자 모델로서 사용되는 world모델과의 차이를 채널 추정 및 화자 모델 생성에 효과적으로 사용하는 것을 목표로 한다. 이를 위해 입력 음성의 켑스트럼과 HMM world 모델의 파라메터인 평균 켑스트럼과의 차이를 통해 음소열에 종속적인 채널 켑스트럼인 Phone-Dependent Difference Cepstrum을 추정한다. 한편 입력 음성의 음소열은 world 모델의 스코어를 얻는 과정에서 함께 얻어질 수 있다.

채널 추정 실험 결과를 통해서 가장 일반적인 채널 정규화 방법인 CMS에 의해 추정된 채널에 비해 실제 채널과 유사하며 화자 고유의 특성을 왜곡시키지 않는 채널 추정이 가능함을 확인할 수 있었다.

1. 서론

환경 변이에 대한 강인함과 화자내(intraspeaker), 화자간(interspeaker) 변이에 따른 스코어 정규화에 대한 연구는 화자 확인 기술의 여러 가지 도전 분야 가운데 두드러진 2가지이다.

환경 변이는 시간 축에서 가산되는 배경 잡음과 마이크, 전화선 환경의 변화에 따라 주파수축에서 곱해지는 convolutive noise로 나눌 수 있다. 환경 변이를 정규화하기 위해 각각의 잡음 형태를 가산된 형태로 변환하고 차감하는 spectral subtraction, cepstral mean subtraction 알고리즘들과 soft decision, pole-filtered cepstrum subtraction 등의 개선된 알고리즘들이 제안되었다.[1,2] 또한 2가지 형태의 잡음을 함께 모델링하여 음성 특징 - quefrency - 도메인에서 보상해주는 방법들도 제안되었다.[3] 특히 전화선 환경에서는 음성을 켑스트럼으로 표현할 경우 선형 변환을 거친 음성과 채널의 형태로 분리되고 이러한 특성을 이용하여 적용 환경과 동일한 stereo data가 있을 경우 매우 정확한 채널의 추정이 가능하다. 따라서 어구 종속 화자 확인 시스템에서 등록 시에 사용된 음성을 테스트 시에 입력된 음성과 비교하여 등록과 확인에서의 채널 불일치(channel mismatch)를 보상하는 방법이 제안되기도 했다.[4] 하지만 실제 적용 환경을 잘 모델링한 stereo data의 구축이 매우 어려우므로 궁극적으로 채널에 대한 사전 정보 없이 채널을 정확하게 추정하고 효과적으로 보상해줄 수 있는 기술이 요구된다.

한편 HMM과 같은 stochastic process로서 음성 특징을 모델링할 경우, PDF의 특성에 따라 화자내의 변이는 정규 분포로 표현된다. 하지만 음성 인식과는 달리 강조되어야 하는 화자간의 변이는 각 화자 모델에서 적절히 표현될 수 없다. 따라서 화자내의 변이를 수용하고 동시에 화자간의 변이로서 사칭자를 구별하기 위해 일반적으로 cohort 모델이나 world 모델의 스코어로서 정규화 하는 방법을 사용한다. 특히 world 모델은 확인 과정에서 연산량, 계산 과정의 복잡도 그리고 저장 공간의 측면에서 장점을 가진다.[5] 그러나 입력되는 음성

과 다른 환경으로부터 생성된 world 모델을 사용할 경우 화자와 사칭자의 차이를 적절하게 얻지 못하고 채널 등의 영향으로 인한 차이가 두드러지게 되는 문제점을 명백하게 갖는다.

본 논문에서는 이상에서 언급한 2가지의 문제를 효과적으로 해결할 수 있는 채널 추정 및 스코어 정규화 방법으로서 world model을 이용한 켑스트럼 정규화 방법을 제안한다.

2절에서는 기존의 방법과 제안된 방법에 대해서 간략히 설명하고 3절에서 제안된 방법에 대한 실험 방법과 결과를 설명한다. 마지막으로 결론 및 앞으로의 연구 방향에 대해 기술한다.

2. 본론

2.1. 켑스트럼을 이용한 채널 보상

입력 음성 $y[n]$의 i번째 프레임에 대한 short real 켑스트럼은 다음과 같이 정의된다.[6]

$$\begin{aligned} c_{i,y}[n] &= F^{-1}(\log|F\{y_i(n)\}|) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\log|Y_i(\omega)|e^{i\omega n}d\omega \ , i=1,2,\ldots,N \end{aligned} \quad (1)$$

켑스트럼은 주파수축에서 곱의 형태로 나타나는 성도의 주파수 응답과 여기 신호의 주파수 응답에 대해 선형 도메인인 quefrency축에서 합의 형태로 나타내는 특징을 가진다. 이러한 기본적인 특징을 이용하여 채널을 거친 음성을 시간 축과 주파수축에서 표현하면 다음과 같다.

$$\begin{aligned} y[n] &= h[n]*x[n] \\ C_{i,y}(\omega) = Q_*^{real}\{Y_i(\omega)\} &= \log|Y_i(\omega)| \\ &= \log|X_i(\omega)H(\omega)| \\ &= \log|X_i(\omega)| + \log|H(\omega)| \\ &= C_{i,x}(\omega) + C_h(\omega) \end{aligned} \quad (2)$$

결국 채널 성분이 첨가된 켑스트럼은 다음과 같이 표현된다.

$$C_i(n) = F^{-1}(\log|X_i(\omega)| + \log|H(\omega)|) \quad (3)$$

이때 적절한 저역 통과 필터를 통해 켑스트럼의 낮은 quefrency성분만을 취하면 상대적으로 느리게 변화하는 성도의 주파수 응답을 얻을 수 있으며 일반적으로 음성인식, 화자인식에서 주요한 특징으로 사용되고 있다. 한편 N프레임 길이의 신호에 대한 평균 켑스트럼(Mean Cepstrum)은 다음과 같이 표현할 수 있다.

$$\begin{aligned} \overline{C}(n) = {} & F^{-1}(\log|H(\omega)|) + \\ & \frac{1}{N}\sum_{i=1}^{N}F^{-1}(\log|X_i(\omega)|) \end{aligned} \quad (4)$$

이때 채널 성분은 시간에 따라 불변이고 장구간에 대한 음성 켑스트럼의 평균이 0이라고 가정하면 평균 켑스트럼은 채널 성분만을 포함하게 된다. 이렇게 얻어진 채널 성분을 전체 음성에서 차감함으로써 채널 성분을 보상하는 방법을 Cepstral Mean Subtraction이라고 한다. 그러나 짧은 발성의 경우 음성 켑스트럼의 평균이 0이 되지 않으므로 화자 확인과 같은 응용에서는 상대적으로 화자의 성도 정보가 왜곡되는 단점이 발생하게 된다. 이러한 단점을 극복하고자 우세 극점을 낮춤으로써 상대적으로 성도 정보의 왜곡을 감소시킨 방법인 pole-filtered CMS 등이 제안되었다.[2]

만약 채널을 거치기 전의 음성을 알고 있다면 각 프레임에서의 켑스트럼의 차이를 통해서 비교적 정확하게 채널을 추정할 수 있다.

$$\begin{aligned} C_i(n) = {} & F^{-1}(\log|H(\omega)|) + F^{-1}(\log|X_i(\omega)|) \\ & - F^{-1}(\log|S_i(\omega)|) \end{aligned} \quad (5)$$

이때 채널을 거치기 전의 음성인 $s[n]$의 i번째 프레임의 음성 스펙트럼 S_i가 채널을 거친 음성 성분인 X_i의 스펙트럼과 동일하다면 거의 정확한 채널 성분 H_i를 얻을 수 있다. 각 프레임의 S_i와 X_i이 동일하다는 것은 성도의 주파수 응답이 일치하는 것을 의미하고 이는 동일 화자의 동일 음소 환경으로 생각할 수 있다. 채널 성분이 시간에 대해 불변하다는 가정을 이용하여 얻어지는 최종 채널 성분 켑스트럼은

$$\overline{C}(n) = \frac{1}{N}\sum_{i=1}^{N}F^{-1}(\log|H_i(\omega)|) \quad (6)$$

와 같이 산술 평균으로 얻을 수 있다.

그러나 명확하게 이 방법은 채널을 거치기 전의 음성을 얻을 수 있어야 하므로 현실적으로 구현하기 어렵다.

2.2. 월드 모델을 이용한 스코어 정규화

HMM을 기반으로 한 화자 확인 시스템의 스코어는 화자 모델에 대한 입력 음성의 우도(likelihood)로 나타나며 그 분포는 감정, 노화, 발성 어구 등에 따른 화자 내 변이와 화자 고유의 특성 차이로 인한 화자간 변이에 따라 다르게 나타난다. 따라서 스코어 정규화의 궁극적인 목적은 화자내의 변이는 최소화하고 화자간의 변이를 최대화하여 화자와 사칭자의 차이를 강조하는 우도 측정 방법을 정의하는 것이라고 할 수 있다. 일반적으로 이러한 화자간의 우도의 변이는 pseudo-imposter 모델을 이용한 정규화 방법으로 해결하며 pseudo-imposter 모델은 cohort set을 이용하는 competition based 방법과 world model을 이용하는 qualifier based 방법으로 근사 될 수 있다. 또한 이러한 측정 방법을 절충한 형태도 연구되었다.[7]

본 연구에서 선택한 world 모델을 이용한 스코어 정규화 방법은, 화자 S의 음성 패스워드에 의해 화자 성분 모델 λ_S가 등록되고 청구된 화자의 음성으로부터 추출된 관찰벡터를

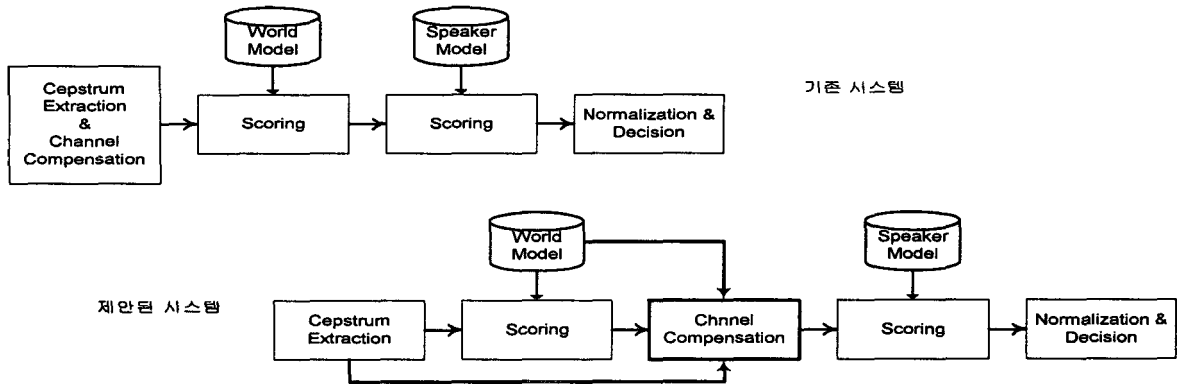

그림 1. 기존의 시스템과 제안된 정규화 방법을 적용한 시스템

O라고 할 때 얻어지는 정규화된 스코어를(likelihood ratio) world 모델 λ_W을 이용하여 다음과 같이 정의한다.

$$R=\frac{\log[P(O|\lambda_S)]-\log[P(O|\lambda_W)]}{|\log P(O|\lambda_W)|} \quad (7)$$

한편, 일반적으로 world 모델의 생성을 위해 화자 확인에 사용되는 어구에 종속적이면서 화자에는 비 종속적인 DB를 사용하며, 어구 종속 화자 확인 시스템과 같이 사용되는 어구가 제한되지 않은 경우에는 음성 인식의 음소 단위 모델 생성을 위한 DB를 사용한다.

이때 world 모델은 적용 환경과 유사한 환경에서 수집된 데이터를 이용하여 생성될 때 최적의 효과를 얻을 수 있다. 이는 발성 환경과 다른 환경에서 훈련된 world모델을 사용할 경우 인식 오류에 따라 상대적으로 화자의 고유 특징에 의한 차이가 작아지므로 사칭자에 대한 정확한 정규화 결과를 얻을 수 없기 때문이다.

2.3. 제안된 켑스트럼 정규화 방법

제안된 방법은 다음과 같은 과정으로 구성된다. (그림1)

1. **음소 단위 world 모델 생성**: 깨끗한 환경에서 수집된 화자 독립 대용량 어휘 음성 인식을 위한 음성 데이터를 통해 음소 단위 world 모델을 훈련시킨다.
2. **음소 상태열 인식**: World 모델을 통해 입력된 음성을 인식하여 각 프레임의 음소와 해당 음소 모델의 상태열을 얻는다.
3. **켑스트럼 정규화**: 각 프레임에서의 음소 모델 상태열에 따라 HMM 음소 모델의 파라메터와 입력 음성 켑스트럼 차이를 구한다. 음소에 종속적인 켑스트럼 차를 입력 음성에 대한 보상 켑스트럼으로 간주하여 새로운 특징 벡터를 구성한다.
4. **모델 등록 및 스코어 정규화**: 보상된 특징 벡터를 통해 화자 모델을 등록하거나 확인 과정에서 스코어를 얻고 world 모델에 대한 스코어와 함께 정규화된 스코어를 얻는다.

3번 과정을 통해 추정된 각 프레임에서의 보상 켑스트럼은 다음과 같이 표현된다.

$$\begin{aligned} C_i(n)=F^{-1}(\log|H(\omega)|) &+F^{-1}(\log|X_i(\omega)|) \\ &-F^{-1}(\log|P_i(\omega)|) \end{aligned} \quad (8)$$

이때 P_i는 world 모델에 대해 입력 음성 i번째 프레임의 인식된 음소이며, 해당 음소의 켑스트럼은 다음과 같이 정의된다.

$$F^{-1}(\log|P_i(\omega)|)\equiv C_{i,P}(n)=\sum_{k=1}^{M} c_{jk}\,\mu_{jk},\ 1\le j\le N \quad (9)$$

위에서 c_{jk}, μ_{jk}는 각각 j상태를 구성하는 k번째 정규분포를 규정하는 가중치 및 평균 파라메터이다. 이들 파라메터는 미리 훈련된 HMM모델로부터 쉽게 얻을 수 있다. 또한 상태열에 대한 정보는 스코어링 과정에서 얻을 수 있다.

기존의 켑스트럼 정규화 방법은 입력된 음성에 대한 언어정보 없이 일반적인 음성의 음향학적인 장구간 특성 또는 단구간 특성만을 이용하여 채널을 추정하였다. 이에 반해 제안된 방법은 음성에 대한 언어정보를 이용하여 각 음소에 종속적인 켑스트럼의 차이를 통해 보다 효과적으로 채널 성분을 추정할 수 있게 된다.

하지만 식(8)에서 각 프레임의 $X_i(\omega)$와 $P_i(\omega)$가 동일할 수 없다. 이는 $P_i(\omega)$가 입력 음성의 화자에 대해서 사칭자인 화자들의 음성을 통해 생성된 world 모델을 통해서 인식된 결과이고 따라서 사칭자의 켑스트럼 분포를 나타내기 때문이다.

결국 보상 켑스트럼은 world 모델 환경에 대한 입력 환경의 차이에 의한 영향과 함께 사칭자와 화자의 차이도 함께 포함하게 된다. 그러나 시간에 대해 불변인 채널 성분의 특성을

고려하여 시간평균 켑스트럼인 $\overline{C}(n)=\frac{1}{N}\sum_{i=1}^{N}C_i(n)$ 을 사용함으로써 화자간 변이에 대한 영향을 최소화 될 수 있다.

최종적으로 수정된 특징벡터를 화자 성문 모델 λ_S와 world 모델 λ_W에 의해 스코어를 얻고 이를 식 (7)에 의해 정규화 하여 문턱값과 비교함으로써 화자를 확인한다.

3. 실험 및 결과

채널 추정 효과를 실험하기 위해 10명의 화자가 발성한 3쌍의 2연 숫자음 데이터베이스를 전화선 채널을 거쳐 다시 녹음한 stereo 데이터베이스를 사용하였다. 한편 입력 음성의 음소 및 상태열을 인식하기 위한 world 모델은 한국 전자 통신 연구소가 제공한 PBW(Phoneme Balanced Word) DB를 사용하였다.

채널을 거치기 전의 음성과 채널을 거친 음성으로부터 각각 MC(Mean Cepstrum), DC(Difference Cepstrum) 그리고 제안된 방법으로 얻어진 PDDC(Phone Dependent Difference Cepstrum)을 구하고 푸리에 변환을 통해 추정한 채널을 비교하였다. 그림 2에서 여자 화자가 발성한 "35-37-36"의 발성에 대해 각각의 방법으로 추정된 채널을 도시하였다. 그림에서 볼 수 있듯이 MC에 의해 추정된 채널은 발성된 음소들의 우세한 포먼트 정보를 그대로 포함하므로 CMS를 통해 차감했을 경우 화자의 고유한 특징을 지나치게 제거할 것으로 예상된다. 반면 거의 정확한 채널 주파수 응답이라고 할 수 있는 DC와 PDDC를 비교했을 경우 상당히 안정적이고 유사한 형태를 나타냄을 확인할 수 있었다.

하지만 전체 비교 실험에서 첫째, 동일화자의 동일 세션에서 DC에 의해 추정된 채널이 동일하지 않고 둘째, 일부 화자의 경우 PDDC에 의해 추정된 채널의 주파수 응답이 DC에 의해 추정된 채널과 상당한 차이를 보이는 문제점을 관찰할 수 있었다.

이는 사용된 데이터가 상대적으로 발생 음소가 제한된 발성이고 더욱이 켑스토럼이 특정 주파수의 통과 대역에 대한 정보만을 강조하여 표현하므로 푸리에 변환을 통해서 정확한 채널을 얻기에는 미흡한 것으로 사료된다.

4. 결론

제안된 방법은 전화선 환경의 입력 음성과 깨끗한 환경의 동일한 가진 음소열을 가진 음성과의 켑스트럼 차이를 통해 채널을 추정하고 이를 보상함으로써 일반적인 화자 확인 시스템에서 발생하는 환경 불일치에 따른 문제를 해결하고자 하였다.

특히 동일한 음향학적 특성을 가정하기 위해 필요한 음소열은 화자 확인 시스템의 world모델을 사용함으로써 스코어 정규화를 위한 스코어와 동시에 얻을 수 있는 장점을 가진다.

간단한 채널 추정 실험 결과 CMS에서 사용되는 평균 켑스트럼에 비해 안정적이고 stereo데이터의 켑스트럼 차이에 의해 얻어진 채널과 근사한 효과를 제안된 방법으로 얻을 수 있음을 확인할 수 있었다. 앞으로 실제 화자 확인 시스템에서의 효과를 보다 정량적으로 측정하기 위해 전화선을 통해 수집된 어구 종속 화자 확인 데이터베이스를 이용한 실험을 실시하고자 한다.

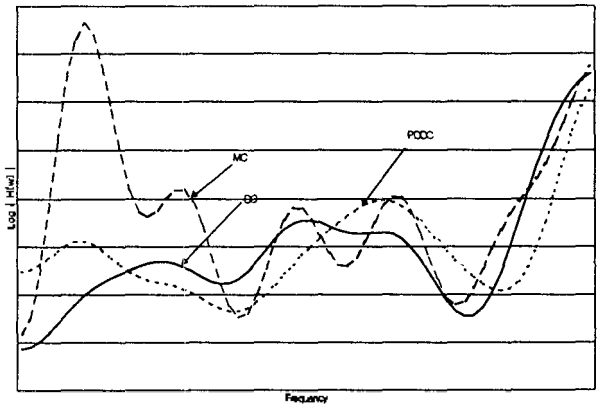

그림 2. MC, DC, PDDC의 채널 추정 비교

참고 문헌

[1] Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. on ASSP*, vol. 29, April 1979.

[2] Devang Naik, "Pole-Filtered Cepstral Mean Subtraction", *Proc. ICASSP*, pp. 157-160, 1995

[3] Hema A. Murthy, Francoise Beaufays, Larry P. Heck, Mitchel Weintraub, "Robust Text-Independent Speaker Identification over Telephone Channels", *IEEE Trans. on ASSP*, vol. 7, No. 5, Sep. 1999

[4] T.F. Lo, M.W. Mak and K.K. Yiu, "A New Cepstrum-Based Channel Compensation Method For Speaker Verification", *Proc. Eurospeech*, pp. 775-778, 1999

[5] Cedric JABOULET, Johan KOOLWAAIJ, Johan LINDBERG, Jean-Benoit PIERROT, Frederic BIMBOT, "The CAVE-WP4 Generic Speaker Verification System," *Proc. RLA2C*, Speaker Recognition and its Commercial and Forenisic Application, pp.202-205, 1998

[6] John R. Deller, Jr., John G. Proakis, John H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, 1993

[7] Yong Gu and Trevor Thomas, "A Hybrid Score Measurement For HMM-Based Speaker Verification," *Proc. of ICASSP*, pp.317-320, 1999