묵음 구간의 평균 켑스트럼 차감법을 이용한 채널 보상 기법[1)]

우승옥° 윤영선
한남대학교 정보통신공학과

Channel Compensation technique using silence cepstral mean subtraction

Seung-Ok Woo°, Young-Sun Yun
Department of Information and Communication Engineering, Hannam University

{mos11,ysyun}@hannam.ac.kr

Abstract

Cepstral Mean Subtraction (CMS) makes effectively compensation for a channel distortion, but there are some shortcomings such as distortions of feature parameters, waiting for the whole speech sentence. By assuming that the silence parts have the channel characteristics, we consider the channel normalization using subtraction of cepstral means which are only obtained in the silence areas. If the considered techniques are successfully used for the channel compensation, the proposed method can be used for real time processing environments or time important areas. In the experiment result, however, the performance of our method is not good as CMS technique. From the analysis of the results, we found potentiality of the proposed method and will try to find the technique reducing the gap between CMS and ours method.

1) 본 연구는 한국과학재단 지역대학 우수과학자 지원사업(R05-2003-000-11398-0)의 지원으로 수행되었습니다

Ⅰ. 서론

최근 음성인식 기술의 발달과 함께 이를 이용한 다양한 상품과 서비스가 제공되고 있다. 특히 전화를 이용한 서비스가 실용화되어, 기차표 예매, 전화 안내 서비스, 신원조회, 정보 검색 등의 서비스를 이용할 수 있게 되었다. 그러나 전화망에서 음성인식을 할 경우 학습 환경과 동작 환경이 다르기 때문에 인식률이 크게 떨어진다. 따라서 인식률을 향상시키고 더 많은 서비스를 제공하기 위해서는 효과적인 잡음 제거 기술이 필요하다[1].

잡음의 종류는 크게 가산 잡음과 채널 왜곡 두 가지로 나누어진다. 가산 잡음은 문을 두드리는 소리나 백색 잡음 등이 신호에 더해지는 것을 의미하고, 채널 왜곡은 전화나 마이크 등 선형시스템을 통과할 때 생기는 신호의 왜곡이다. 채널 왜곡은 시간 영역에서는 신호와의 컨벌루션, 주파수 공간에서는 곱으로 이루어지지만 로그 영역에서는 덧셈으로 볼 수 있다.

잡음에 의한 왜곡을 보상하는 방법으로는 크게 세 가지로 구분할 수 있다. 첫 번째로, 잡음에 강인한 특성 추출 방법이다. 이 방법은 잡음의 특성을 모르더라도 효과적인 성능을 보여주는 반면, 잡음의 특징을 알더

라도 그 특성을 적용할 수 없다는 단점이 있다. 잡음을 보상하는 두 번째 방법으로는 음질 개선(speech Enhancement) 방법이 있다. 이것은 잡음 음성을 학습 환경으로 변환하고 학습 환경에서 학습된 그 시스템으로 잡음 음성을 인식한다. 세 번째로 모델 보상(model compensation)방법이 있다. 이는 참조환경에서 생성된 음성 모델들을 잡음 환경에 맞추어 변환하여 잡음 음성을 인식하는 방법이다[2]. 이 중 본문에서 다룰 켑스트럼 차감법 (CMS; Cepstral Mean Subtraction)과 그 변형인 전역 묵음 켑스트럼 차감법(GSCMS; Global Silence Cepstral Mean Subtraction), 국부 묵음 켑스트럼 차감법 (LSCMS; Local Silence Cepstral Mean Subtraction)의 잡음 제거 방식은 잡음에 강인한 특징 추출 방법 또는 전처리 보상 방법에 속한다.

본문에서는 채널 왜곡보상 방법인 CMS에 대해서 살펴본 후 새로 제안한 방법인 묵음 구간의 평균 차감법을 이용한 채널 보상기법에 대하여 알아보겠다. 이는 학습 데이터의 전체 묵음 구간의 평균 켑스트럼을 원래 음성에서 빼는 GSCMS와 각 학습 음성의 묵음 구간에서의 평균 켑스트럼의 차를 이용하는 LSCMS로 구분하여 실험하였다. 실험결과, 기존의 CMS 성능에는 미치지 못하지만, CMS의 특성을 반영하여 CMS의 단점을 보완할 수 있다는 가능성을 보였다.

II장에서는 CMS, SGCMS, SLCMS에 대하여 기술한 후 III장에서는 실험 및 결과를 보이고, IV장에서 결론을 내린다.

II. CMS 채널보상기법

1. CMS

CMS는 채널 왜곡을 보상하는 방법으로서 켑스트럼 전체 구간에서 구한 평균을 각각의 켑스트럼에서 차감해준다. 전화망을 통과하는 음성신호는 채널 왜곡이 생기는데 이는 주파수 영역에서 신호와 잡음의 곱으로 표현할 수 있다.

$$Y(z) = S(z)H(z)$$

S(z)는 순수한 음성 H(z)는 전화선 채널, Y(z)는 왜곡된 음성을 의미한다. 이를 로그 영역으로 나타내면

$$\log Y(z) = \log S(z) + \log H(z)$$

즉, 켑스트럼 영역에서는 채널 왜곡이 순수한 음성에 부가적 형태로 나타난다.

구현적 측면에서 볼 때 우선 T개의 전체 켑스트럼에 대하여 평균을 구한다.

$$x_{CMS} = \frac{1}{T}\sum_{t=1}^{T} x_t$$

CMS는 각 켑스트럼 벡터에서 평균을 제거해주므로, 다음 식과 같다.

$$x_t' = x_t - x_{CMS}$$

CMS가 정확히 채널 잡음을 제거하기 위해서는 순수한 음성의 켑스트럼에 대해 장구간 평균이 0이어야 하는 조건이 있다. 이 조건이 만족되려면 음성구간에서 유성음, 무성음, 파열음이 음향학적으로 균형을 이루어야 한다. 하지만 이 가정들은 실제로는 불가능하므로 단구간의 음성에서 CMS를 적용했을 때에 음성 정보까지 왜곡할 수 있다[3][4].

3. 묵음 구간 켑스트럼 평균 차감법

음성 인식을 실제 환경에서 적용할 경우, CMS의 전제조건을 충족시키기 어렵다는 어려움이 있다. 따라서 이 조건들을 만족시키지 않고도 효율적으로 채널 특성을 모델링하는 방법으로 묵음(silence)구간의 켑스트럼 평균을 구하여 차감하는 방법을 알아보겠다

전화음성의 묵음구간은 채널 특성만을 담고 있다고 가정을 한다. 따라서 묵음 구간에서의 평균을 제거한다면 CMS의 전제조건과 관계없이 채널 정보만을 제거할 수 있다. 묵음 구간 평균 차감법의 또 다른 장점은 준 실시간 인식이 가능하다는 점이다. 음성이 입력되었을 때 시작 부분에 최소 20 frame의 묵음 구간이 존재한다고 가정하면 약 200 msec의 묵음 구간에서 켑스트럼 평균을 구할 수 있고 이를 이용하여 채널 특성을 보상할 수 있다면, 특징 추출에 약 200 msec의 지연만을 요구하기 때문에 준 실시간으로 음성을 처리할 수 있을 것이다.

학습 자료의 전체 묵음 구간에서의 켑스트럼 평균을 구하는 방법과 각각의 음성 발화에서 묵음 구간 켑스트럼 평균을 구하는 방법으로 나누어 모델링할 수 있다. 학습 시에는 전체 묵음 구간 또는 개개의 발화에서의 묵음 구간을 추출하여 평균을 구했고, 평가 시에는 채널 보상을 하지 않는 학습 모델을 이용한 인식 결과에서 묵음 구간 정보를 추출하여 사용했다. 준 실시간성을 만족시키기 위해서는 끝점 검출이 일관성 있게 이뤄진다면 음성의 시작부분에서 일정 프레임을 묵음구간으로 가정한 후 평가 실험을 할 수 있다.

3.1. GSCMS

일차적으로 고려한 모델은 전체 음성의 묵음 구간에서 켑스트럼 평균을 추출하여 각 프레임에서 제거한다. 음성의 묵음 구간에서의 켑스트럼 평균이 채널 특

성을 반영한다고 가정하였을 경우, 순수한 채널 특성은 동일한 환경에서 녹음된 음성에 고루 분포되어 있다고 가정하고 그 평균을 이용한다.

$$x_{GSCMS} = \frac{1}{N}\sum_{l=1}^{L}\sum_{t\in q_s} x_t^{(l)}$$

N : 학습 자료의 묵음 프레임의 총 수

L : 문장의 수

q_s : 묵음으로 판명된 프레임 집합

3.2. LSCMS

전체 음성에서의 묵음 구간 켑스트럼 평균 값은 마이크나 전화음성의 채널 특성이 발화시점에서 변하지 않는다고 가정을 하였다. 그러나 동일한 마이크라도 발화과정에서 전기적인 신호를 받아 그 특성이 바뀔 수 있으므로 각 발화에서 묵음 구간 켑스트럼 평균을 이용하여 각 프레임에서 차감하는 방법을 취했다.

$$x_{LSCMS} = \frac{1}{N}\sum_{t\in q_s} x_t$$

N : 한 문장내의 묵음 프레임 수

Ⅲ. 실험 및 결과

1. 실험 환경

채널 보상 기법의 성능 차이를 극대화하기 위하여 학습 자료는 조용한 환경에서 녹음된 DB를, 평가 자료는 전화 환경에서 녹음된 DB를 사용하였다.

훈련에 사용된 음성 DB는 원광대에서 제작된 한국어 4연 숫자음성으로서 공, 영, 일, 이, 삼, 사, 오, 육, 륙, 칠, 팔, 구의 12개의 숫자로 구성되었으며, 서울(경기), 충청, 호남, 영남의 사람들 400명이 발성했다. 조용한 사무실 환경에서 녹음되었고, 16kHz로 샘플링되어 16bits로 양자화되었다. 평가 음성 DB는 역시 원광대에서 제작되었고, 200명이 발성하였으며, 8kHz로 샘플링되었다. 녹음 환경은 유선/ 무선/ Cellular/ PCS로 이루어졌다. 학습 DB와 평가 DB를 비슷한 조건에서 실험하기 위하여 훈련 DB를 8kHz로 down sampling 한 후 300~3400Hz에서 필터 뱅크를 구하여 실험하였다.

2. 인식시스템

시스템의 성능 차이에 의한 오류를 최소화하기 위하여 Cambridge 대학의 HTK v3.2로 실험하였다. 실험에 사용된 특징 차수는 12차의 MFCC(Mel-Frequency Cepstral Coefficients)와 log Energy, delta, delta-delta 계수로 구성된 총 39차를 사용하였다. 0.97의 pre-emphasis를 사용한 뒤, 20 msec의 해밍윈도우를 이용하여 10 msec씩 이동하였다. 모노폰 모델과 트라이폰 모델을 모두 사용했으며, 상태 수와 mixture 수는 각각 5개와 9개로 실험하였다.

3. 결과

39차 MFCC 특징벡터를 사용하여 clean DB에서 훈련시켜 전화 잡음 DB에서 테스트 한 것을 Baseline으로삼고 CMS를 적용한 결과를 비교 결과로 하였다

표 1 모노폰 인식 성능 비교

	monophone	
	Sentence	WER
Baseline	17.49	70.72
CMS	50.51	87.23
GSCMS	29.93	76.64
LSCMS	37.12	81.06

표 2 트라이폰 인식 성능 비교

	triphone	
	Sentence	WER
Baseline	34.09	81.09
CMS	69.00	92.92
GSCMS	42.95	82.77
LSCMS	50.79	86.32

<표 1>과 <표 2>에서 문장 인식률을 비교하면 모노폰은 GSCMS가 12.44%, LSCMS가 19.63% , 트라이폰은 GSCMS가 8.86% LSCMS가 16.7%로 채널 특성을 반영하지 않은 baseline보다는 성능이 향상되었지만 CMS보다는 성능이 저하됨을 알 수 있었다.

GSCMS와 LSCMS가 채널 보상 효과를 보이지만, 널리 사용되는 CMS의 성능에는 미치지 못하여, CMS와 GSCMS, LSCMS의 각 특성을 비교해보았다.

<그림 1>은 잡음이 섞이지 않은 음성에서 CMS와 GSCMS, LSCMS의 각 cepstral 차수 별 평균을 나타낸다. 그림에서 볼 수 있듯이 CMS는 낮은 차수의 계수를 억제하고 높은 차수의 계수를 강조하는 High Pass Filter의 특성을 보인다. 기존의 연구에서는 이러한 특성을 반영하여 CMS 대신 High Pass Filter를 적용하기도 하였다[5].

<그림 2>는 전화음성에서 CMS, LSCMS의 각 차수별 평균 켑스트럼을 나타내고 있다. 인식과정에서 전체 묵

음 구간을 구할 수 없기 때문에 각각의 발화에 대해 묵음 구간 켑스트럼 평균을 구한다.

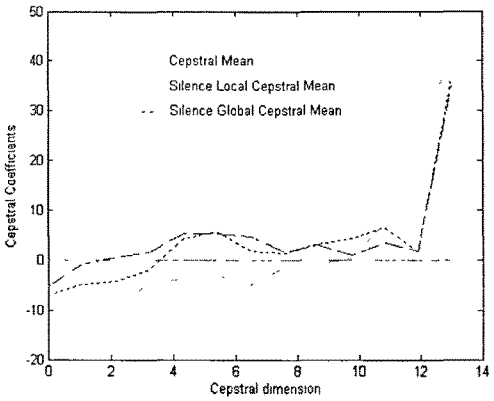

그림 1 조용한 환경에서의 켑스트럼 평균

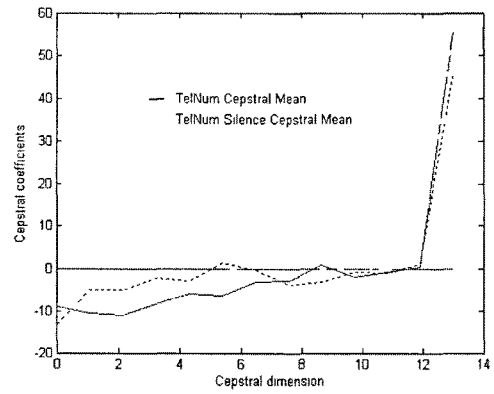

그림 2 전화 음성에서의 켑스트럼 평균

위 특성에서 살펴볼 수 있듯이 조용한 환경이나 전화 환경에서의 켑스트럼 평균은 많은 차이를 보이고 있지 않으나, 묵음 구간에서의 켑스트럼 평균은 많은 차이를 보이고 있다. 결국, 조용한 환경에서 묵음 구간의 평균 스펙트럼을 차감한 경우의 음성과 전화 음성에서의 묵음 구간 평균 스펙트럼을 차감했을 때의 음성의 왜곡 정도가 달라 성능 저하를 가져옴을 알 수 있다. 즉, 채널 특성은 음성 전반 또는 특정 주파수 대역에서의 음성 왜곡을 가져오기 때문에, 조용한 환경과 전화 환경과 같이 채널 특성이 상이한 환경에서는 채널 특성 보상이 결국 음성의 왜곡을 가져왔다고 생각할 수 있다. 그러나 조용한 음성에서의 묵음 구간에서의 켑스트럼 포락과 전화 음성에서의 묵음 구간의 켑스트럼 포락을 살펴보면 유사점이 많아, 성능 개선의 여지는 충분하다고 판단한다. 이러한 점에 초점을 맞춰 채널 특성을 모델링 한다면 성능 향상을 가져올 수 있을 것이다.

Ⅳ. 결론

채널 보상 기법으로 널리 사용되는 CMS는 그 뛰어난 성능에도 불구하고 음성의 왜곡을 가져오며, 전체 음성이 발화된 이후에 처리가 가능하다는 단점을 안고 있다. 따라서 본 연구에서는 CMS의 약점으로 지적되는 음성의 왜곡과 실시간성을 해결하기 위하여 묵음 구간에서의 켑스트럼 평균을 이용한 채널 보상 기법을 고려해 보았다. 성능 평가 결과 LSCMS나 GSCMS 모두 CMS의 성능에 미치지는 못했지만, CMS의 특성과 묵음 구간에서의 켑스트럼 평균 특성을 어느 정도 파악할 수 있었다. 향후 연구과제로는 인식 성능이 CMS에 버금가도록 묵음 구간에서의 켑스트럼 평균 또는 주파수 특성을 찾아 채널 특성을 보상하는 연구가 지속되어야 할 것이다.

참고문헌

[1] 정성윤, 손종목, 김민성, 배건성, "한국어 숫자음 전화음성의 채널왜곡에 따른 특징파라미터의 변이 분석 및 인식실험" 말소리 제 43호

[2] 오영환, "음성언어정보처리", 홍릉과학출판사 pp 142-160 1998.

[3] 김상진, 서영주, 한민수, "LCMS를 이용한 한국어 연속 숫자인식에 관한 연구," 한국음향학회 학술발표대회 논문집 제20권, 2001.

[4] Thomas F.Quatieri Massachusetts Institute of Technology Lincoln Laboratory "Discrete-Time Speech Signal Processing Principles and practice" pp.671-672

[5] Huang, Xuedong Acero, Alex Hon, Hsiao-Wuen "Spoken Language Processing", Prentice Hall