전화망에서의 음성인식을 위한 전처리 연구

Front-End Processing for Speech Recognition in the Telephone Network

전 원 석*, 신 원 호*, 양 태 영*, 김 원 구**, 윤 대 희*
(Won-Suk Jun*, Won-Ho Shin*, Tae-Young Yang*, Weon-Goo Kim**, Dae-Hee Youn*)

※본 논문은 한국통신 연구개발본부의 1996년도 수탁과제연구 지원에 의해 수행되었습니다.

요 약

본 논문에서는 다양한 전화선 채널에서 수집된 한국통신(KT)의 데이터베이스를 이용하여 인식 시스템의 성능을 향상시키기 위한 효율적인 특징벡터 및 전처리방법을 연구하였다.

먼저 잡음 및 주변 환경 변화에 강인한 것으로 알려져 있는 특징벡터들을 이용한 인식 성능을 비교하고, 가중 켑스트랄 거리측정 방법을 이용하여 인식시스템의 성능 향상을 검증하였다. 실험 결과, KT의 인식 시스템에서 이용하는 LPC 켑스트럼의 경우에 비하여 PLP(Perceptual Linear Prediction)과 MFCC(Mel Frequency Cepstral Coefficient)등에 대하여 인식률이 향상되었다. 켑스트럼간의 거리측정에 있어서는 RPS(Root Power Sums)와 BPL(Band Pass Lifter)과 같은 가중 켑스트랄 거리측정 함수들이 인식성능 향상에 도움을 주었다.

스펙트럼 차감법(Spectral Subtraction)의 적용은 왜곡에 의한 효과가 커서 인식률이 저하되었지만, RASTA(RelAtive SpecTrAl) 처리 방법, CMS(Cepstral Mean Subtraction), SBR(Signal Bias Removal)의 적용시에는 인식 성능 향상을 보였다. 특히, CMS 방법은 간편하면서도 높은 인식 성능 향상을 보였다.

마지막으로, CMS의 실시간 구현을 위한 방법들의 인식 성능을 비교하고, 인식 성능 저하를 막기 위한 개선책을 제시하였다.

ABSTRACT

In this paper, we study the efficient feature vector extraction method and front-end processing to improve the performance of the speech recognition system using KT(Korea Telecommunication) database collected through various telephone channels.

First of all, we compare the recognition performances of the feature vectors known to be robust to noise and environmental variation and verify the performance enhancement of the recognition system using weighted cepstral distance measure methods. The experiment result shows that the recognition rate is increased by using both PLP(Perceptual Linear Prediction) and MFCC(Mel Frequency Cepstral Coefficient) in comparison with LPC cepstrum used in KT recognition system. In cepstral distance measure, the weighted cepstral distance measure functions such as RPS(Root Power Sums) and BPL(Band-Pass Lifter) help the recognition enhancement.

The application of the spectral subtraction method decreases the recognition rate because of the effect of distortion. However, RASTA(RelAtive SpecTrAl) processing, CMS(Cepstral Mean Subtraction) and SBR(Signal Bias Removal) enhance the recognition performance. Especially, the CMS method is simple but shows high recognition enhancement.

Finally, the performances of the modified methods for the real-time implementation of CMS are compared and the improved method is suggested to prevent the performance degradation.

Ⅰ. 서 론

잡음이 섞인 음성 신호를 개선하기 위한 잡음 제거(noise subtraction) 기술은 음성 신호처리의 중요한 연구 분야중 하나이다. 특히 최근에는 음성 인식 시스템의 실용화가 늘어나면서 잡음 환경에서의 음성 인식에 관한 연구가 활발히 진행되고 있다. 잡음 및 주위 환경 변화에 적

*연세대학교 전자공학과
**군산대학교 전기공학과

접수일자 : 1997년 2월 25일

응할 수 있는 음성 인식 시스템 구현은 실용적인 음성 인식 시스템을 개발하는데 있어서 고려해야할 중요한 문제들 중의 하나이다.

잡음 및 주변 환경에 강인한 음성 인식에 관한 연구는 그 동안 실험실 환경에서 연구되어온 음성 인식 결과의 실용 가능성을 평가하기 위하여 다양한 잡음 환경에서의 인식 평가가 이루어지고 있다. 그러나 이러한 방법들이 실제 상황의 잡음이나 환경에서도 우수한 성능을 나타낸다고 볼 수는 없다. 따라서 이들 방법을 검증하고 다양한 형태의 잡음과 환경 변화에 강인한 방법에 관한 연구가 요구되어지고 있다.

본 논문에서는 전화 채널을 통한 음성 인식 시스템 성능 개선을 위한 효율적인 특징벡터 및 전처리 기술의 개발을 목표로 하였다. 이러한 목표를 위하여 한국통신(KT)에서 구축한 음성 인식 시스템 및 데이터베이스를 기반으로 잡음에 강인한 특징 벡터 및 가중 켑스트랄 거리 측정 방법, RASTA 처리, 스펙트럼 차감법, 켑스트럼 평균 차감법, 신호 편의 제거 방법 등을 적용하여 인식 실험을 하였다. 그 결과, 스펙트럼 차감법 이외의 방법들이 모두 인식 성능 향상에 도움을 주었는데, 켑스트럼 평균 차감법을 적용한 경우가 가장 뛰어난 인식 성능을 보였다.

본 논문의 구성은 제 2장에서 전화망 및 잡음 환경에 적용하기 위한 잡음 처리 기술에 대하여 소개하고, 제 3장에서 이에 따른 실험 및 결과를 고찰하였고, 제 4장에서는 실시간 시스템에 적용하기 위한 방법을 제시하였으며, 제 5장에서 결론을 맺었다.

Ⅱ. 잡음 처리 기술

2.1 잡음에 강인한 특징 벡터와 거리 측정 방법

잡음에 강인한 것으로 알려진 특징벡터들 중에서 대표적인 것으로는 멜 켑스트럼 계수, PLP(Perceptually Linear Prediction) 계수 등이 있다. 멜 켑스트럼 분석 방법은 인간의 청각 특성을 이용한 것으로, 실제 물리적인 주파수와 인지된 주파수 사이의 대응 관계를 이용하여 켑스트럼 계수로 표현한 것이다[1].

PLP 분석 방법은 1982년 Hermansky에 의해 제안되었는데, 이는 음성 신호의 파워 스펙트럼을 변화시켜 인간이 실제 감지하는 소리와 유사한 저차의 스펙트럼을 얻게 한다. PLP 분석 방법은 음성 인식에 적용되어 좋은 성능을 보여주었다[2].

스펙트럼간의 거리 측정으로 잡음에 강인한 특성을 갖는 가중 켑스트랄 거리 측정 방법이 많이 이용된다[4-10]. 이는 다음과 같이 정의되는데,

$$d(c, c') = \sum_{k=1}^{P} w_k^2 (c_k - c'_k))^2 \quad (1)$$

여기서 w≡(w1, … wP)는 켑스트랄 리프터(cepstral lifter)인 가중 함수이고 c≡(c1, …, cP)와 c'≡(c1', …, cP')는 켑스트럼 계수이다. 음성 인식에 사용되어 좋은 성능을 나타낸 가중 함수들로는, 스펙트럼 기울기에 기초를 둔 RPS(Root Power Sums), 가우시안(gaussian) 형태로 스무딩된 선형 리프터(smoothed linear lifter:SLL), 지수 함수 리프터(general exponential lifter:GEL), 켑스트럼 계수의 높은 차수와 낮은 차수의 바람직하지 못한 변화를 제거하기 위한 밴드 패스 리프터(band pass lifter:BPL), 켑스트럼 계수의 통계적인 분포에 따라 가중 함수를 결정한 것 등이 있다.

2.2 RASTA(RelAtive SpecTrAl) Processing

음성 신호를 분석하는 방법 중 시간에 따라 천천히 변하는 성분(steady-state factor)에는 영향을 덜 받는 음성 신호의 특징을 추출하는 방법으로 RASTA 분석 방법이 제안되었다[12-14]. RASTA 분석 방법에서는 일반적인 단구간 스펙트럼(short-term absolute spectrum)을 사용하는 대신 스펙트럼 성분 중 시간에 따라 천천히 변화하는 성분을 배제하는 대역 통과 스펙트럼(band-pass filtered spectrum)을 사용한다. 필터링 블록은 각 주파수 대역을 IIR 필터를 사용하여 대역 통과 필터링(bandpass filtering)하는 것과 같다.

2.3 잡음 제거 및 보상 방법

2.3.1 스펙트럼 차감법(Spectral Subtraction method)

스펙트럼 차감법은 주변 잡음에 의해 손상된 음성 스펙트럼에서 잡음 스펙트럼의 크기 성분만을 제거하는 방법이다[11]. 스펙트럼 차감법은 배경잡음의 스펙트럼 형태를 미리 알고 있거나, 잡음의 스펙트럼을 추정하기에 충분한 묵음 구간(약 300ms)이 주어있어야 한다. 또한, 배경잡음은 최소한 부분적으로 안정(stationary)한 특성을 가져야 하며, 통계적 특성이 서서히 변화하는 환경에서는 음성이 존재하는 구간과 잡음만이 존재하는 구간을 검출할 수 있는 방법이 필요하다.

2.3.2 켑스트랄 평균 차감법(CMS:Cepstral Mean Subtraction)

켑스트랄 평균 차감법은 전체 구간에 대하여 켑스트럼의 평균을 구하고, 이를 차감하여 채널의 효과를 제거하는 방법이다[15].

전화망을 통한 음성 신호는 채널의 필터링 영향에 의해 선형 왜곡이 일어난다. 이를 간단히 표현하면, $T(z) = S(z)G(z)$으로 표현되는데, 여기서 $S(z)$는 순수한 음성신호, $G(z)$는 전화 채널, $T(z)$는 필터링된 음성을 의미한다.

이를 로그 영역으로 나타내면,

$$\log T(z) = \log S(z) + \log G(z) \quad (2)$$

즉, 채널의 영향은 순수한 음성의 켑스트럼에 대해 부가적인 성분으로 나타나게 된다. 이 때, 순수한 음성의 켑

스트럼에 대해 장구간 평균이 0이라고 가정하면, 채널 켑스트럼의 추정치는 필터링된 음성의 켑스트럼들을 평균하여 구할 수 있다. 채널 효과를 보상하기 위해서는 추정된 채널 켑스트럼을 제거하는 것이 켑스트럼 평균 차감법이다. 그러므로, 채널의 영향이 보상된 켑스트럼은 다음과 같이 구해진다.

$$c_{t,\,CMS} = c_t - E[c_t] \tag{3}$$

$$E[c_t] = \frac{1}{T}\sum_{t=1}^{T} c_t \tag{4}$$

여기서, $E(c_t)$는 채널 켑스트럼의 평균값, c_t는 t 번째 프레임의 켑스트럼, T는 전체 프레임 수이다.

2.3.3 신호 편의 제거(SBR : Signal Bias Removal)

신호 편의 제거(SBR) 방법은 여러 가지 환경에 의하여 오염된 입력신호에서 음성신호와 바이어스를 분리하여, 이를 제거함으로써 채널왜곡이나 잡음에 의한 영향 등을 효과적으로 제거해 준다[17].

입력 신호의 특징벡터열 $X=\{x_1, x_2, \cdots, x_t, \cdots, x_T\}$와 특징벡터에 대한 대표 모델 $\Lambda=\{\lambda_i, i=1, 2, ..., M\}$ (λ_i는 Markov 모델, t는 프레임 수)에 대하여 λ_i가 동일한 확률을 갖는다면, likelihood는 벡터 양자화(VQ)와 동일하게 된다. 여기서 Λ는 바이어스가 없는 모델이라고 가정하면, λ_i는 VQ 코드북에서 코드워드(혹은 centroid)가 된다. 이 코드워드들이 바이어스가 없는 신호를 대표하므로, 입력 신호의 특징벡터와 코드워드들 사이의 차이를 바이어스라고 할 수 있다. 입력신호의 전체구간에 대하여 각 바이어스의 평균을 구하고 이를 차감하는 과정을 반복적으로 수행함으로써 바이어스를 제거한다.

SBR 방법은 CMS와는 달리 바이어스를 제거한 후에도 기존 모델에 대한 의존성이 유지되기 때문에, 모델을 새로 훈련하지 않고 기존의 모델을 사용하면서 인식 단계에서만 처리해 주면 되는 장점이 있다.

Ⅲ. 실험 및 결과

3.1 Database 및 잡음 분석

실험에 이용한 데이터는 KT에서 제공한 전화망에서 녹음된 음성 데이터를 이용하였다. 인식에 이용된 단어는 214단어로 구성되어 있다. 데이터들은 전화 다이얼링 예약어로 사용될 수 있는 단어들로 10대에서 50대사이의 남녀가 발음한 것이다.

실험에 사용한 음성신호들을 살펴보면, 많은 부분에서 약 200Hz근처에서 크기와 주파수의 차이는 있지만, 순음잡음(tonal noise)가 존재하고 있다. 이것을 스피커를 통해 들어보면 흔히 전화를 사용할 때 들을 수 있는 '부-우' 하는 잡음임을 알 수 있다. 그림 1의 입력 음성 신호에 대한 스펙트로그램을 살펴보면, 저주파 영역에 삼각형의

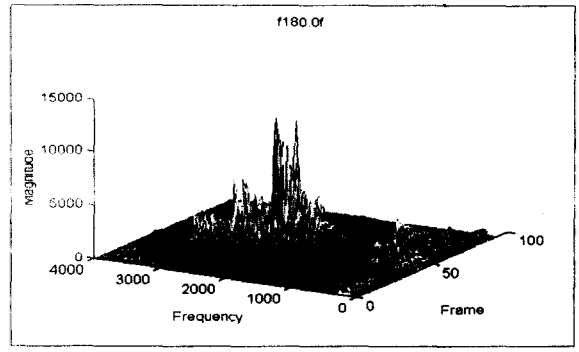

그림 1. 잡음에 오염된 신호의 스펙트로그램

모양의 잡음이 전체 구간에 걸쳐 일정하게 나타남을 볼 수 있는 데, 이것은 전화 채널에 의하여 발생하는 잡음으로 부가 잡음의 성격을 띠고 있다.

이것 이외에 여러 가지 배경 잡음들이 섞여 있는 데, 이는 전화 도중 주위 사람들이 떠드는 소리 등과 같은 주변 소음과 일부 데이터에는 입력 신호의 앞부분이나 뒷부분에 '삐이'하는 소리가 첨가되어 있다. 입력 신호의 앞부분부터 바로 음성 신호가 시작하는 경우도 있어 이러한 경우에는 스펙트럼 차감법 적용시, 전단의 잡음 또는 채널 특성의 파악이 불가능한 경우들이 존재하였다. 또한 분해능 부족에 의한 양자화 잡음에 의하여 크기가 작은 경우 일부가 0으로 나타났으며 계단형의 파형이 나타나기도 하였다.

실험에서 이용된 KT의 데이터 베이스는 주변 잡음이나 주변 화자의 발성 등 불안정한(non-stationary) 잡음이 포함되어 있으며 순음 잡음 등이 포함되어 있어 좋은 학습 모델을 생성하기에 부적합한 형태로 구성되어 있다. 화자의 발성 특성이나 신호의 크기 등도 부적합한 것들이 포함되어 있다. 이러한 데이터베이스의 특성이 실제 데이터베이스의 특성을 그대로 반영한다고 하여도 그 변화의 폭이 워낙 크고 신호에 포함된 불규칙적인 잡음이 인식 성능의 향상에 기여한다고는 볼 수 없다. 따라서 인식 성능을 향상시키기 위한 좋은 학습 모델을 구성하기 위해서는 보다 선별된 데이터 베이스를 이용하는 것이 적합할 것으로 보인다.

전체 데이터베이스는 VDS, VDS95, VDSfromREAL1의 3가지로 구성되어 있다. 이들 중에서 VDS의 경우에는 비교적 토널 노이즈가 적고, 다른 배경 잡음이나 양자화 잡음의 영향도 별로 없는 데이터들이다. VDS95의 경우 노이즈의 영향이 적은 데이타들도 있지만, 토널 노이즈에 의하여 심하게 오염된 경우, 주변 잡음이 많이 들어 있는 경우, 양자화 잡음이 심한 경우 등 여러 가지로 구성되어 있다. VDSfromREAL1은 다양한 상황에서 녹음된 것들이 임의적으로 섞여있는 실제와 가장 비슷한 데이타들이다.

훈련 데이타들로는 VDS 데이터 2,791개와 VDS95 데이

터 4,928개로 구성되었고, 인식 데이터들은 훈련에 사용되지 않은 VDS 데이터 595개, VDS95 데이터 796개, VDSfromREAL1 데이터 458개로 이루어져 있다.

본 논문에서는 편의상 VDS를 DB1, VDS95는 DB2, VDSfromREAL1은 DB3로 표기한다.

3.2 특징 추출 및 인식 시스템 구성

신호의 분석 구간은 20ms에 해당하는 160 sample을 이용하였으며 10ms씩 이동하면서 해밍 윈도우를 이용하여 분석하였다. 각 분석 구간으로부터 얻은 LPC 켑스트럼 또는 멜 켑스트럼 분석을 이용하여 12차의 켑스트럼 계수를 구하였다. 이들의 동적인 특성을 반영하기 위하여 1차와 2차의 차분치를 취하여 이를 각각 특징 벡터로 이용하였다. 에너지에 대하여서는 각 구간으로부터 구한 데이터 에너지와 1차 및 2차 차분값을 에너지 특징 벡터의 성분으로 이용하였다. 이들을 특징 벡터로 각 특징 벡터에 대한 코드북을 생성하였다. 각 코드북의 크기는 256개로 하였으며 에너지 관련 벡터에 대하여서는 64개의 코드크기를 갖도록 하였다. PLP 분석 등과 같은 경우에는 특징 벡터의 차수를 12차보다 작은 값으로 이용하였으나 기타의 경우에는 12차의 차수를 갖도록 하여 전체적으로 동일한 형태를 갖도록 하였다. 스펙트럼 차감법이나 RASTA 처리 등과 같은 경우, 전처리 과정을 통하여 특징 벡터를 추출하도록 하였다.

인식 시스템은 이산 HMM 모델을 이용하였으며, 학습 과정은 Baum-Welch 알고리즘을 이용하였고 테스트 시에는 Beam Search 방법을 이용하였다.

3.3 각 처리 방법에 따른 실험 및 결과

3.3.1 잡음에 강한 특징벡터 및 가중 켑스트랄 거리 측정

인식 시스템의 학습 및 인식은 LPC 켑스트럼 방법을 기준으로, 멜 주파수 영역에 의한 켑스트럼 추출방법(MFCC)과 PLP 분석에 의한 특징 추출 방법을 비교하여 보았다. PLP의 경우 켑스트럼 차수를 10차까지 그리고 RPS의 가중함수를 주어 실험을 하였다.

표 1의 결과를 살펴보면, 두 가지 경우 모두 약간의 인식률 향상을 보이고 있다. 특히, MFCC의 경우 DB1에 대하여 높은 인식률 향상을 보이고 있고, DB2에 대해서는 인식률이 약간 하락한다. 반면에, PLP의 경우에서는 DB1에서는 베이스라인과 비슷한 인식률을 나타내지만, DB2에서는 인식률이 조금 상승한다.

위의 특징벡터 실험에 의하여 인식성능이 좋은 MFCC에 대하여 가중함수를 가하여 이에 따른 인식성능의 변화를 표 2에 나타내었다. RPS와 BPL의 경우 인식률이 향상되는 데, BPL의 경우가 RPS의 경우보다 근소하게 앞선다.

3.3.2 스펙트럼 차감법에 대한 실험

앞에서 잡음에 대한 특성을 살펴보았는데, 이들 중에서 채널에 의한 토널 잡음이 존재하는 경우 인식 성능이 저하됨을 볼 수 있었다. 이의 영향을 제거하기 위하여 스펙트럼 차감법을 이용하였다.

표 1. 특징벡터에 따른 인식 실험

recog. rate (%)	DB1	DB2	DB3	Total
baseline(LPC)	70.08	72.00	60.70	68.67
MFCC	76.81	71.20	60.92	70.38
PLP	69.75	73.26	60.92	69.19

표 2. 특징벡터(MFCC)의 가중 켑스트랄 거리측정에 대한 인식 실험

recog. rate (%)	DB1	DB2	DB3	Total
MFCC	76.81	71.20	60.92	70.38
MFCC_RPS	76.13	74.85	63.54	72.51
MFCC_BPL	75.63	74.51	64.85	72.72

표 3. 스펙트럼 차감법에 의한 인식 실험

recog. rate (%)	DB1	DB2	DB3	Total
baseline(LPC)	70.08	72.00	60.70	68.67
SUB_LPC	64.20	69.60	58.95	65.40

표 4. LPC에 대한 SBR과 CMS의 인식 실험

recog. rate (%)	DB1	DB2	DB3	Total
baseline(LPC)	70.08	72.00	60.70	68.67
LPC_CMS	80.50	74.63	70.31	75.31
LPC_SBR	73.11	73.14	62.45	70.64

표 5. MFCC에 대한 SBR과 CMS의 인식 실험

recog. rate (%)	DB1	DB2	DB3	Total
MFCC	76.81	71.20	60.92	70.38
MFCC_CMS	82.02	75.09	70.74	76.24
MFCC_SBR	78.82	73.83	65.28	73.32

표 6. RASTA 처리 방법에 의한 인식 실험

recog. rate (%)	DB1	DB2	DB3	Total
MFCC	76.81	71.20	60.92	70.38
RASTA	77.65	73.14	64.35	71.73
RASTA_CMS	79.33	75.20	66.16	74.33

표 7. CMS의 실시간 구현을 위한 인식 실험

recog. rate (%)	DB1	DB2	DB3	Total
MFCC_CMS	82.02	75.09	70.74	76.24
MFCC_LCMS	80.17	71.89	65.07	72.82
MFCC_SCMS	78.82	73.49	64.85	73.03
MFCC_D_SCMS	79.50	74.74	67.03	74.38

그러나, 표 3의 결과를 살펴보면, 스펙트럼 차감법을 이용한 경우 전체적으로 성능이 떨어지고 있음을 알 수 있다. 즉, 스펙트럼 차감의 효과로 잡음을 제거해 주기는 하지만, 그 부수적인 영향으로 왜곡을 일으키게 되고, 결과적으로 성능 향상에는 도움이 되지 못한다고 볼 수 있다. 학습 데이터에 토널 잡음이 존재하지 않는다면 스펙트럼 차감법에 의한 효과를 기대할 수 있으나 학습 데이터에도 토널 잡음이 부분적으로 존재하므로 이의 효과를 기대하기 힘들다. 따라서 학습 데이터의 선정 시에는 토널 잡음의 특성이 포함되지 않은 데이터를 이용하는 것이 바람직 할 것으로 보인다.

왜곡의 영향을 감소시키기 위하여 차감하는 크기를 줄인 경우, 추정 잡음을 완전히 제거했을 때보다 인식 성능이 약간 향상되었으나 기준 시스템에는 미치지 못하였다.

잡음이 매우 작은 주파수 영역에서 일어나는 왜곡을 방지하고 순음 잡음(tonal noise)만을 효과적으로 제거하

기 위하여, 잡음 스펙트럼 중 에너지가 상대적으로 큰 주파수 영역만을 차감하는 방법을 이용하여 실험해 보았다. 이 방법을 적용했을 때, LPC 캡스트럼에 대한 실험의 경우 근소하게 나마 인식률 상승이 일어나지만, MFCC의 경우에서는 인식률이 하락한다. 그 내용을 분석해 본 결과, 토널 잡음이 큰 경우에는 의도대로 인식률이 상승하지만, 작은 경우는 역시 왜곡에 의해 인식률에 나쁜 영향을 끼친다.

결과적으로, 전체 데이터에 잡음이 일정 수준 이상 크지 않은 경우 스펙트럼 차감법은 그리 유효하지 않다고 볼 수 있다.

3.3.3 SBR과 CMS에 대한 인식 실험

LPC 및 멜 캡스트럼에 대하여 SBR과 CMS를 이용하여 인식실험을 하였다. SBR의 경우 훈련 데이터의 특징벡터들로부터 만들어진 코드북을 이용하여 각 실험 데이터의 바이어스를 추정하고, 이를 제거한 것으로 인식 실험을 하였다. 반면에 CMS의 경우 각 데이터의 전체구간에 대하여 캡스트럼의 평균을 구한 후, 이것을 차감하여 구한 데이타들로부터 훈련 및 인식 실험을 하였다. 표 4를 보면, SBR의 경우 바이어스 제거 효과에 의하여 각 데이타베이스에 대하여 고른 인식률 향상을 보이고 있지만, 그 향상 폭은 그리 크지 않다. CMS의 경우 전체적으로 큰 인식률 향상을 보이고 있으며, 특히 실제적인 상황과 가까운 DB3의 신호에 대해 큰 상승폭을 나타낸다.

MFCC에 대하여서도 같은 SBR과 CMS의 실험을 하였다. 표 5의 결과를 살펴보면, LPC에서의 경우와 같은 양상을 띠고 있다. SBR과 CMS에 의하여 모두 인식률이 상승하는데, 그 상승폭은 CMS에서의 경우가 더 크다.

3.3.4 RASTA에 대한 실험

멜 주파수 영역에 의하여 스펙트럼을 구하여 각 채널마다 RASTA 필터링한 후, 캡스트럼을 구한 것을 특징벡터로 이용하였다. 표 6의 결과를 보면 RASTA 방법이 어느 정도 유효함을 알 수 있다. 그러나, CMS를 적용한 방법과 비교하여 볼 때 인식 성능이 다소 떨어지며, CMS를 결합한 경우에도 CMS 방법에 미치지 못하였다.

Ⅳ. CMS의 실시간 구현을 위한 실험

CMS 방법은 음성의 전체 구간에 대하여 평균 캡스트럼을 계산해야 하므로, 실시간 처리가 불가능하다. 이를 해결하기 위하여 LCMS(Local CMS)와 SCMS(Sequential CMS) 방법 및 이들의 문제점을 보완한 D_SCMS(Delay Sequential CMS) 방법을 제안하여 인식 실험을 수행하였다.

전자의 두 방법 모두 주어진 입력 신호로부터 캡스트럼 평균에 의해 채널 캡스트럼의 추정치를 구하여 이를 차감하는데, 캡스트럼 평균을 구하는 방법에 차이가 있다.

LCMS 방법은 채널 캡스트럼의 추정치를 현재 프레임과 그 이전에 입력된 일정한 갯수의 프레임들에 대한 캡스트럼 평균으로 구한다[16]. 즉,

$$c_{t,\,\mathrm{LCMS}} = c_t - \frac{1}{L}\sum_{i=0}^{L-1} c_{t-i} \qquad t = 1, 2, \ldots, T \tag{5}$$

여기서, c_t는 t번째 프레임의 캡스트럼, T는 전체 프레임 수이다. L는 채널을 추정하는 단구간 프레임 수이다.

실험에서는 캡스트럼 평균을 구하는 단구간의 길이는 50 프레임으로 하고, 음성이 시작되는 부분에서는 첫 번째 프레임이 그 이전에 계속 존재하였다고 가정하고 단구간 평균을 구하였다.

SCMS는 첫 프레임부터 현재 프레임까지 입력된 신호의 캡스트럼에 대한 평균을 구하여 차감하는 방법이다[17]. 즉,

$$c_{t,\,\mathrm{SCMS}} = c_t - \frac{1}{t}\sum_{i=1}^{t} c_i = c_t - \frac{1}{t}\left[(t-1)\cdot c_{t-1,\,\mathrm{SCMS}} + c_t\right]$$

$$t = 1, 2, \ldots, T \tag{6}$$

표 7에서 위의 두 방법은 비슷한 정도의 인식 성능 향상을 보이고 있는 데, 전체 구간을 이용하는 CMS의 경우에는 크게 미치지 못한다. LCMS의 경우 캡스트럼의 평균을 구하는 구간이 작아 채널 캡스트럼이 잘 추정되지 못하며, SCMS의 경우 각 입력 음성의 앞부분에서는 평균을 구하는 구간이 적어서 CMS의 효과를 제대로 발휘하지 못하기 때문이다.

그런데, 빔 탐색(Beam search)과 같은 인식 방법을 이용하는 경우 입력 음성의 처음 부분이 인식에 중요한 영향을 미치게 된다. 따라서, 이 구간에 대하여 보다 정확한 평균 캡스트럼을 추정하는 방법이 필요하다.

이를 위해 음성 입력시 처음 얼마 동안의 캡스트럼 평균을 구하고, 지연된 구간에 대하여 처음 프레임부터 다시 평균 캡스트럼을 차감한다. 그리고, 그 이후 구간에 대하여는 SCMS와 동일한 방법으로 캡스트럼을 구하게 된다. 이를 D_SCMS라 표기하고 식으로 나타내면,

$$c_{t,\,\mathrm{D\_SCMS}} = \begin{cases} c_t - \dfrac{1}{D}\displaystyle\sum_{i=1}^{D} c_i & 1 \le t \le D \\ c_t - \dfrac{1}{t}\,c_i = c_t - \dfrac{1}{t}\left[(t-1)\cdot c_{t-1,\,\mathrm{D\_SCMS}} + c_t\right] & D+1 \le t \le T \end{cases} \tag{7}$$

와 같다.

즉, 처음 부분에 대하여 지연 프레임 동안의 평균 캡스트럼을 차감하므로 채널의 특성을 보다 정확하게 추정하여 제거해 주며, 지연 프레임 이후로는 매 프레임마다 SCMS와 동일한 캡스트럼을 얻게 된다.

지연 구간 동안에 각 프레임의 캡스트럼 및 그 평균을 계산해 놓고 있으면, 지연 구간이 지난 후에 처음 프레임

에서부터 구해 놓은 켑스트럼 평균을 차감하기만 하면 되므로 이에 대한 연산량은 아주 적다. 또한 실시간 구현이 가능한 시스템에서는 프레임 처리 속도가 음성 신호 입력 속도보다 빠르다. 그러므로, 초기 입력에서의 처리 지연 시간의 영향은 이후 시간이 지날수록 감소하게 되어, 어느 정도 긴 음성 입력을 처리하는 경우에는 초기 지연 시간의 영향을 거의 받지 않게 된다. 하지만, 지연 시간 동안의 음성 입력에 대한 켑스트럼을 저장하고 있을 만큼의 메모리가 더 필요하게 되는 단점이 있다.

실험에서 지연 프레임 수를 50개로 설정하였는데, 이 경우 각 프레임은 시간상 10ms 간격이 있으므로 지연 시간은 500ms가 된다. 표 7의 인식 결과를 보면 74.38%로 CMS를 제외하면 가장 좋은 성능을 보이고 있다.

프레임당 처리 속도와 전체 입력 음성 길이, 메모리 크기 등을 고려하여 적절한 지연 프레임수를 설정한다면, 시간 지연을 감수하지 않으면서도 성능을 높일 수 있으므로 실시간 처리에 크게 도움이 될 것이다.

V. 결 론

본 연구에서는 KT에서 제공한 전화망 데이터 베이스를 이용하여 공중 전화망에서 음성 인식 시스템 성능을 향상시키기 위한 특징 추출 및 전처리 방법을 연구하였다.

여러 가지 잡음 처리 기술을 이용하여 실험한 결과 다음과 같은 결론을 얻을 수 있었다.

1) 멜 켑스트럼을 이용한 특징 벡터를 이용한 결과 기존의 LPC 켑스트럼을 이용한 경우나 다른 특징 벡터들을 사용한 것보다 우수한 인식 성능을 나타냈다.

2) RPS와 BPL과 같은 가중 켑스트랄 거리 측정 함수들이 인식률 향상에 도움이 됨을 확인하였다.

3) 스펙트럼 차감법의 경우에는 노이즈 제거 효과보다 왜곡의 영향이 커서 인식률 향상에는 도움을 주지 못하였다.

4) RASTA 처리, CMS, SBR을 이용하여 인식 성능을 향상시킬 수 있었다. 특히, CMS는 단순하면서도 가장 뛰어난 인식 성능을 보였다.

5) CMS의 실시간 구현을 위해서 LCMS와 SCMS를 적용한 결과, CMS에 비하여 성능이 다소 저하되었다. 이를 개선하기 위하여 약간의 지연 시간을 두는 D_SCMS를 적용하였을 때, 실시간 처리에 방해를 받지 않으면서도 CMS와 비슷한 성능을 얻을 수 있었다.

참 고 문 헌

1. L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall Inc., 1993.
2. H. Hermansky, B. A. Hanson and H. Wakita, "Perceptually Based Linear Predictive Analysis of Speech," *in Proc. ICASSP*, pp. 509-512, March 1985.
3. A. Acero and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition," *in Proc. ICASSP*, pp. 849-852, April 1990.
4. F. Itakura and T. Umezaki, "Distance Measure for Speech Recognition based on the Smoothed Group Delay Spectrum," *in Proc. ICASSP*, pp. 1257-1260, April 1987.
5. J. Junqua and H. Wakita, "A Comparative Study of Cepstral Lifters and Distance Measures for All Pole Models of Speech in Noise," *in Proc. ICASSP*, pp. 476-479, May 1989.
6. B. A. Hanson and H. Wakita, "Spectral Slope Distance Measure with Linear Prediction Analysis for Word Recognition in Noise," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-35, No. 7, pp. 968-973, July 1987.
7. B. H. Juang, L. R. Rabiner and J. G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-35, No. 7, pp. 947-954, July 1987.
8. A. H. Gray and Jr., J. D. Markel, "Distance Measures for Speech Processing," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-24, No. 5, pp. 380-391, Oct. 1976.
9. N. Nocerino, F. K. Soong, L. R. Rabiner and D. H. Klatt, "Comparative Study of Several Distance Measures for Speech Recognition," *in Proc. ICASSP*, pp. 25-28, March 1985.
10. Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-35, No. 10, pp. 1414-1422, Oct. 1987.
11. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, No 2, pp. 113-120, April 1979.
12. H. Hermansky, N. Morgan, H. G. Hirsch, "Recognition of Speech in Additive and Convolutional Noise based RASTA Spectral Processing", *in Proc. ICASSP*, pp. 83-86, 1993.
13. J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, G. Tong, "Integrating RASTA-PLP into Speech Recognition", *in Proc. ICASSP*, pp. 421-424, 1994.
14. H. Hermansky, N. Morgan, A. Bayya, P. Kohn, "Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech(RASTA-PLP)", *in Proc. EUROSPEECH*, vol. 3, pp. 1367-1370, Sep. 1991.
15. Richard J. Mammone, Xiaoyu Zhang, Ravi P. Ramachandran, "Robust Speaker Recognition-A Feature-based Approach", *in IEEE Signal Processing Mag.*, pp. 58-87, Sep. 1996
16. Aaron E. Rosenberg, Chin-Hui Lee, Frank K. Soong, "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", *in Proc. ICSLP*, pp 1835-1838, 1994.
17. Mazin. G. Rahim, Biing-Hwang Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", *IEEE Trans. Speech & Audio Processing*, vol. 4, No. 1, pp. 19-30, 1996.

▲전 원 석(Won-Suk Jun) 1971년 6월 16일생

1996년 8월 : 연세대학교 전자공학과 졸업(공학사)
1996년 9월~현재 : 연세대학교 대학원 전자공학과 석사과정
※주관심분야 : 음성인식, 잡음처리

▲신 원 호(Won-Ho Shin) : 1996년 15권 5호 참조

▲양 태 영(Tae-Young Yang) : 1996년 15권 5호 참조

▲김 원 구(Weon-Goo Kim) : 1994년 13권 1호 참조

▲윤 대 희(Dae-Hee Youn) : 1994년 13권 1호 참조