전화 음성 인식을 위한 특징 추출 방법 비교

Comparison of Feature Extraction Methods for the Telephone Speech Recognition

전 원 석*, 신 원 호*, 김 원 구**, 이 충 용*, 윤 대 희*
(Won Suk Jun*, Won Ho Shin*, Weon Goo Kim**, Chung Yong Lee*, Dae Hee Youn*)

*본 논문은 한국통신 연구개발본부의 1997년도 수탁과제연구 지원에 의한 결과입니다.

요 약

본 논문에서는 전화망 환경에서 음성 인식 성능을 개선하기 위한 특징 벡터 추출 단계에서의 처리 방법들을 연구하였다. 먼저, 고립 단어 인식 시스템에서 채널 왜곡 보상 방법들을 단어 모델과 문맥 독립 음소 모델에 대하여 인식 실험을 하였다. 켑스트럼 평균 차감법, RASTA 처리, 켑스트럼-시간 행렬을 실험하였으며, 인식 모델에 따른 각 알고리즘의 성능을 비교하였다.

둘째로, 문맥 독립 음소 모델을 이용한 인식 시스템의 성능 향상을 위하여 정적 특징 벡터에 대하여 주성분 분석 방법(principal component analysis)과 선형 판별 분석(linear discriminant analysis)과 같은 선형 변환 방법을 적용하여 분별력이 높은 벡터 공간으로 변환함으로써 인식 성능을 향상시켰다. 또한 선형 변환 방법을 켑스트럼 평균 차감법과 결합하여 더욱 뛰어난 성능을 보여주었다.

ABSTRACT

In this paper, the feature processing methods are studied to improve the speech recognition performance over the telephone lines. First, the recognition experiments of the channel compensation methods for the isolated word recognition are carried out using word models and context-independent phoneme models. Experiments are performed for cepstral mean subtraction(CMS), RelAtive SpecTrAl(RASTA) processing and cepstral-time matrix(CTM), and the performance of the algorithms is discussed in view of recognition model.

Secondly, the linear transformation methods, such as principal component analysis(PCA) and linear discriminant analysis (LDA), are applied to static feature to enhance the accuracy of the recognition system using context-independent phoneme model. These methods are shown to successfully improve recognition performance by transforming static feature vectors to the more discriminant vectors. In addition, linear transformation method is combined with cepstral mean subtraction and it is shown that the approach gives better performance.

I. 서 론

음성 인식 기술은 실제 응용 분야에서 다양한 환경에 적용하려는 시도가 계속되고 있다. 특히, 최근에는 정보 통신의 발달로 전화망 환경에서 널리 응용되고 있는 데, 전화 음성은 여러 가지 요인들에 의하여 인식하기에 더 어렵다. 전화선에 의한 성능 저하 요인으로는 전화선 자체의 대역 제한, 임펄스 잡음, 에코, 채널 응답, 저주파 순음 잡음, 순소리 등이 있다[1]. 이 때, 전화선에서 발생하는 신호 왜곡은 크게 두 가지로 분류될 수 있는데, 그것은 배경 잡음이나 전기적인 잡음과 같이 스펙트럼 영역에서 부가적인 성분으로 나타나는 부가 잡음과 전화선이나 송수화기의 주파수 응답 특성에 의한 채널 왜곡이다[5]. 일반적으로 전화 음성의 경우 마이크(송화기) 가까이에서 발음을 하게 되므로 신호대 잡음비가 비교적 크게 되고, 따라서 부가 잡음보다는 채널 왜곡의 영향을 더 크게 받는다. 그러므로, 전화 환경에서는 일반적으로 채널 왜곡에 대한 보상이 더 중요하다.

채널 왜곡을 보상하기 위한 방법은 켑스트럼 영역에서의 채널 바이어스 제거 방법과 음성의 시간 정보(temporal

* 연세대학교 전자공학과
** 군산대학교 전기공학과
접수일자 : 1998년 6월 26일

information)를 이용하는 방법으로 분류할 수 있다. 그런데, 시간 정보를 이용하는 경우에는 현재 프레임 전후의 정보, 즉, 주위의 문맥(context)에 의한 영향을 크게 받게 되므로, 문맥 독립(context independent) 음소로 모델링하기에는 어려움이 있다. 따라서, 채널 보상 방법들의 성능이 인식 모델에 따라 달라지게 되므로, 여러 가지 모델에 따른 성능 평가가 요구된다.

본 논문에서는 단어 모델과 문맥 독립 음소 모델의 두 가지 경우로 나누어, 대표적인 채널 왜곡 보상 방법들에 대한 성능을 비교하였다. 또한, 문맥 독립 음소 모델을 이용하는 인식 시스템의 성능 향상을 위하여 시간 정보를 이용하지 않는 정적(static) 특징 벡터의 변환 방법을 적용하였다.

본 논문의 구성은 다음과 같다. 본 장에 이어서 2장에서는 채널 왜곡 보상 방법에 대하여 살펴보고, 3장에서는 각 방법들에 대한 변조 주파수 영역에서의 응답 특성을 비교하였으며, 4장에서는 잡음에 강인한 몇 가지 전형 변환 방법에 대하여 기술하였다. 5장에서는 실험을 위한 데이터베이스와 인식 시스템에 대하여 설명하고, 6장에서 실험 및 결과를 비교, 분석하였으며, 7장에서 결론을 맺었다.

II. 채널 왜곡 보상 방법

채널 특성을 알 수 없는 전화망에서의 채널 왜곡 보상을 위한 두 가지 접근 방법에 대하여 설명하고 대표적인 방법들을 알아본다.

2.1 켑스트럼 영역에서의 채널 바이어스 제거 방법

채널 왜곡은 스펙트럼 영역에서 곱으로 나타나고, 이는 켑스트럼(로그 스펙트럼) 영역에서는 부가적인 성분으로 변환되며 시간적으로 일정하게 유지된다. 따라서, 채널 왜곡은 켑스트럼 영역에서는 바이어스로 표현되며, 이를 추정하여 제거함으로써 채널 보상을 할 수 있다.

켑스트럼 평균 차감법(CMS: Cepstral Mean Subtraction)은 대표적인 바이어스 제거 방법으로 전화 음성 인식뿐만 아니라 화자 인식에서도 효과적임이 알려져 있다[2,3]. Carnegie-Mellon 대학에서는 잡음 및 채널에 의해 발생되는 학습 데이터와 인식 데이터사이의 환경 불일치 문제를 효과적으로 해결하기 위하여 CDCN(Codeword-Dependent Cepstrum Normalization)을 비롯한 여러 가지 켑스트럼 정규화 방법들을 개발하였으며, 환경 변화에 강인한 성능을 보여주었다[1,4]. Rahim과 Juang은 학습 데이터로부터 얻은 코드북을 기반으로 반복적으로 바이어스를 추정하여 제거함으로써 인식 데이터와 학습 데이터의 환경을 일치시키도록 하는 신호 바이어스 제거(SBR: Signal Bias Removal) 방법을 제안하였으며 전화망 환경에서 좋은 성능을 보여주었다[5]. 이 외에도 학습 과정에서 생성된 모델을 기반으로 최대 우도(maximum likelihood) 혹은 최대 사후 확률(maximum a posteriori probability) 방법에 의하여 바이어스를 추정하는 방법들이 시도되었다[6,7].

위의 방법들은 각각 여러 가지 장, 단점이 있지만 대개 켑스트럼 평균 차감법을 제외하면 비교적 연산량이 많고, 실험 환경에 따라 약간씩 성능이 다르게 나타난다. 일반적으로 켑스트럼 평균 차감법이 성능 평가를 위한 기준이 되므로, 본 논문에서는 채널 바이어스 제거 방법에 대한 평가를 위하여 켑스트럼 평균 차감법을 실험하였다.

켑스트럼 평균 차감법은 순수한 음성의 켑스트럼에 대해 장구간 평균이 0이라고 가정하며, 채널 바이어스는 왜곡된 음성의 켑스트럼들을 평균함으로써 추정할 수 있다. 그러므로, 전체 구간에 대하여 켑스트럼의 평균을 구하고, 이를 차감하여 채널에 의한 왜곡을 제거한다. 이 방법은 채널 왜곡에서 뿐 만 아니라, 화자들간의 변화를 정규화하는 데에도 효과적이다.

모든 채널 바이어스 제거 방법들은 전체 음성 구간에 대하여 바이어스를 구하기 때문에 실시간 처리가 불가능하다. 실시간 처리가 요구되는 시스템에서는 다른 변형된 기법들이 요구된다. 켑스트럼 평균 차감법의 실시간 처리를 위해서는 이전의 시간에 대하여 순차적으로 평균을 구하는 SCMS(Sequential CMS), 단구간에 대하여 평균을 구하는 LCMS(Local CMS), 어느 정도의 지연 구간을 두어 평균을 구하고 순차적으로 갱신하는 DSCMS(Delay-Sequential CMS) 등이 있다[3].

2.2 시간적인 정보를 이용하는 방법

음성은 일정한 범위의 속도로 변화하는 데 비하여 채널 왜곡은 시간적으로 일정하거나 아주 천천히 변화한다. 따라서, 시간적으로 천천히 변화하는 정보를 제거함으로써 채널의 영향을 받지 않는 특징 벡터를 얻을 수 있다. 원하지 않는 정보를 제거하는 방법으로는 프레임간의 시간 궤적(time trajectory)에 대한 필터링 기법이 주로 이용되는데, 필터들은 다양한 형태와 방법으로 구현될 수 있다. 로그 스펙트럼의 서브밴드(subband) 에너지 영역이나 켑스트럼 영역에서의 고역 통과 필터와 대역 통과 필터들이 효과적으로 적용되었다[9]. 특히, 대역 통과 필터는 채널 성분뿐만이 아니라 빠르게 변화하는 성분도 함께 제거하므로 잡음에 강인한 특성을 갖는다. 필터링 방법 중에서 RASTA(RelAtive SpecTrAl) 처리[10]가 가장 널리 이용되는데, 각 주파수 대역을 IIR(Infinite Impulse Response) 필터를 사용하여 대역 통과 필터링하게 되며, 이 대역 통과 필터의 전달 함수는 다음과 같다.

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - \alpha \cdot z^{-1})} \qquad (1)$$

여기서, α는 저역 차단 주파수를 결정하는 상수로 $0<\alpha<1$의 조건을 만족해야 하며, 일반적으로 0.9에서 0.98사이의 값을 사용한다.

Vaseghi와 Milner[11]는 스펙트럼의 시간적인 변화를 고려하는 방법으로 정적 특징 벡터와 그의 시간축 미분치를 이용하는 대신 이산 코사인 변환(DCT: Discrete Cosine Transform)에 의하여 얻어지는 켑스트럼-시간 행렬(CTM:

Cepstral-Time Matrix)을 사용하는 방식을 제안하였다. 켑스트럼-시간 행렬은 로그 스펙트럼의 시간축 시퀀스(sequence)로 이루어진 행렬에 2차원 DCT를 적용하거나 켑스트럼들의 시간축 시퀀스로 이루어진 행렬의 각 행에 대해서 아래 식과 같이 1차원 DCT를 적용함으로써 구해질 수 있다.

$$C_t(m,n)=\sum_{k=0}^{M-1} c_{t+k}(n)\cos\left(\frac{(2k+1)m\pi}{2M}\right) \qquad (2)$$

여기서, $c_{t+k}(n)$은 t번째 프레임의 n차 켑스트럼 계수이고, M은 시간축 DCT를 위한 프레임 수이며, $C_t(m,n)$은 t번째 프레임의 켑스트럼-시간 행렬의 m번째 열과 n번째 행에 해당하는 계수를 가리킨다.

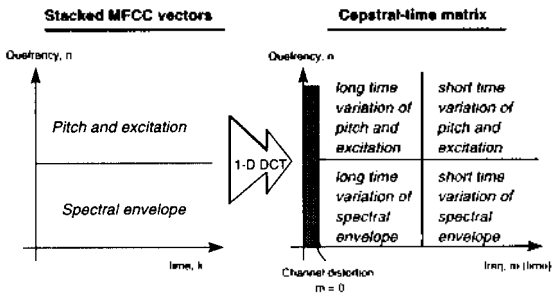


그림 1. 켑스트럼-시간 행렬의 영역

그림 1은 켑스트럼을 1차원 DCT에 의하여 켑스트럼-시간 행렬로 변환할 때의 여러 가지 영역을 나타낸 것이다. 여기서 변환된 후의 왼쪽 아래 영역은 스펙트럼 개형(envelope)의 긴 시간 변화량을 나타내고, 인식에 있어서 가장 중요한 의미를 가지므로 이외의 영역들을 모두 버린다. 또한 0차 열에는 채널왜곡에 대한 성분이 집중되어 있으므로 이를 제거해 주어야 한다. 이렇게 하여 인식 과정에서 중요한 정보만을 담고 있는 부분 행렬을 얻게 되고, 이를 특징 파라미터로 사용한다. 이 때, 각 열은 서로 다른 시간 정보를 포함하고 있으므로 다른 방법들과는 달리 시간 미분을 사용하지 않는다.

본 논문에서 시간적인 정보를 이용하는 방법으로서, 여러 필터링 기법 중에서 로그 스펙트럼 영역에서의 RASTA 처리와 일반적인 인식 시스템과 다른 특징 벡터 형태를 갖는 켑스트럼-시간 행렬의 두 가지를 선택하여 실험하였다.

III. 변조 주파수 응답 특성

서브밴드 에너지나 켑스트럼의 시간적인 변화에 대한 주파수 응답 특성을 변조 주파수 응답(modulation frequency response)[9]이라고 하는 데, 채널 왜곡을 제거하기 위한 방법들의 변조 주파수 응답 특성은 매우 낮은 변조 주파수 성분을 제거하는 역할을 한다. 본 장에서는 2장에서 제시한 채널 왜곡 보상 방법들에 대한 변조 주파수 응답 특성을 비교한다.

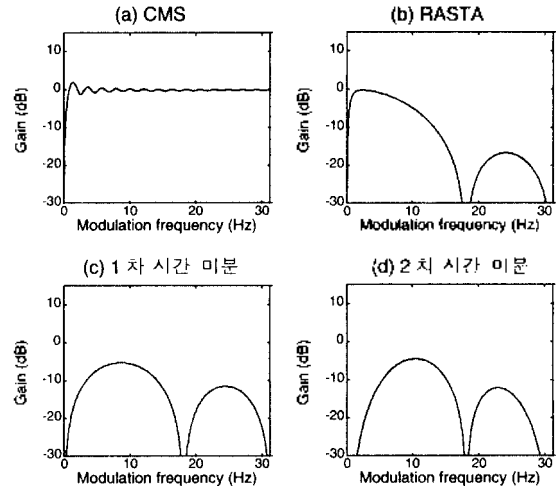


그림 2. CMS, RASTA, 1차 및 2차 시간 미분의 변조 주파수 응답

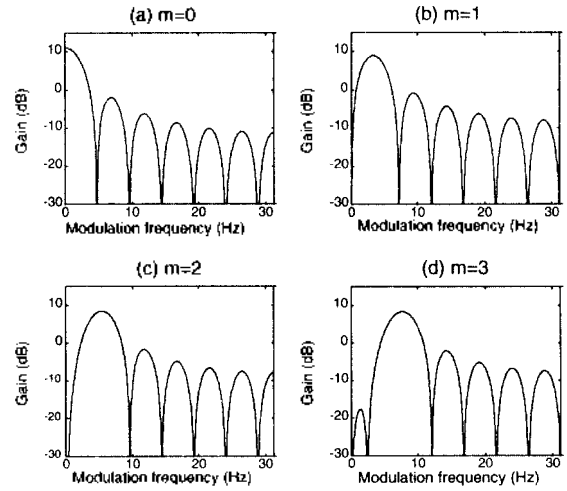


그림 3. 켑스트럼-시간 행렬의 각 열(m)에 대한 변조 주파수 응답

그림 2와 3은 프레임 이동 간격을 16ms로 하였을 경우의 변조 주파수 응답 곡선들이다. 그림 2에서 (a)는 켑스트럼 평균 차감법에 대한 변조 주파수 응답으로 저주파만이 제거된 고역 통과 필터와 같은 역할을 한다. (b)는 RASTA 처리에서 사용하는 식 (1)의 IIR 필터에 대한 변조 주파수 응답이다. 저주파와 고주파를 차단하는 대역 통과 필터의 특성을 갖는다. 이 두 가지 방법의 가장 큰 역할은 약 2Hz 이하의 저주파를 제거하여 채널 왜곡의 영향을 제거하는 것이다.

그림 2(c)는 시간 미분의 변조 주파수 응답을 그린 것

이다. 시간 미분을 구하는 프레임 수를 5로 할 때 시간 미분을 구하는 방법은 식 (1)의 분자항과 동일하게 된다. 그러므로, 그림 2(c)는 그림 2(b)의 RASTA의 응답 곡선과 비슷한데, RASTA의 경우 식 (1)의 분모항에 의하여 극(pole)을 갖고, 이것은 그림 2(c)에 비하여 강조되는 주파수를 낮추면서 저역 차단 주파수를 결정한다. 그림 2(d)는 2차 시간 미분의 응답 곡선으로 그림 2(c)의 경우보다 강조되는 주파수가 조금 높아지고 전체적인 모양은 비슷하다. 여기서, 그림 2(c)와 그림 2(d)는 모두 저주파를 차단시키므로 채널에 강인한 특성을 갖는다.

그림 3은 켑스트럼-시간 행렬을 구하기 위하여 식 (2)의 m을 0에서 3까지 변화시킬 때의 DCT에 대한 변조 주파수 응답을 그린 것이다. m의 값이 커짐에 따라 강조되는 주파수가 점점 높아지게 된다. 여기서 $m=0$인 경우는 채널 성분을 포함하고 있으므로 사용하지 않고, 중요한 정보를 포함하는 대역($m=1,2,3$)만을 이용한다. 켑스트럼-시간 행렬에서 각 열(column)은 서로 다른 변조 주파수 성분을 강조하고 있고, 이는 시간 미분의 역할과 비슷하다.

IV. 특징 변환 방법

채널과 잡음에 의하여 왜곡된 음성 신호의 켑스트럼은 순수한 음성의 켑스트럼에 대하여 회전(rotation), 스케일링(scaling), 이동(translation)에 의하여 변환된다. 이 때, 잡음은 회전 및 스케일링의 작용을 하고 채널 왜곡은 이동의 작용을 하게 되므로[2], 채널과 잡음의 영향을 보상하려면 반대의 과정을 거쳐야 한다. 이러한 과정을 벡터 변환으로 표현하면,

$$y = Ax + b \tag{3}$$

이 된다. 여기서, x는 입력 음성의 켑스트럼 벡터, y는 변환된 후의 켑스트럼 벡터이다. A는 선형 변환 행렬로 회전과 스케일링의 역할을 하고 b는 바이어스 벡터로 이동의 역할을 한다.

만약, 학습 환경 및 인식 환경에 대한 정확한 통계적 정보를 알고 있다면, 인식 데이터를 학습 환경으로 변환시키는 행렬 A 및 벡터 b를 구하여 환경 불일치를 보상할 수 있다[2]. 그러나, 인식 환경에 대한 정확한 정보를 추정할 수 없는 경우에는 학습 데이터와 인식 데이터를 모두 잡음에 강인한 영역으로 변환시킴으로써 인식 성능을 향상시킬 수 있다. 이 때, 선형 변환 행렬 A에 의하여 분별력 높은 영역으로 변환시키고, 바이어스 벡터 b에 의하여 채널 왜곡을 보상해 주는데 이는 2.1절에서 설명한 켑스트럼 영역에서의 채널 바이어스 제거 방법과 일치한다.

본 장에서는 전화 음성으로부터 얻은 켑스트럼을 잡음에 강인한 특징 벡터로 변환시키기 위하여 선형 변환 행렬 A를 구하는 몇 가지 방법에 대하여 설명한다.

4.1 켑스트럼 가중 함수

스펙트럼간의 거리 측정으로 잡음에 강인한 특성을 갖는 가중 켑스트랄 거리 측정 방법(weighted cepstral distance measure)이 많이 이용된다. 이는 다음과 같이 정의되는데,

$$d(c, c') = \sum_{k=1}^{P} w_k^2 (c_k - c_k')^2 = \sum_{k=1}^{P} (w_k (c_k - c_k'))^2 \tag{4}$$

여기서, P는 켑스트럼 계수의 차원, $w \equiv (w_1, \cdots, w_P)$는 가중 함수이고 $c \equiv (c_1, \cdots, c_P)$와 $c' \equiv (c_1', \cdots, c_P')$는 켑스트럼 계수이다. 그런데, 거리 측정에 사용되는 가중 함수들은 켑스트럼에 직접 곱하는 켑스트랄 리프터(cepstral lifter)로 생각할 수 있으므로, 식 (3)에서 변환 행렬 A는 가중 함수들의 값으로 이루어진 대각 행렬(diagonal matrix)로 놓을 수 있다[2].

음성 인식에 사용되어 좋은 성능을 나타낸 가중 함수들로는, 스펙트럼 기울기에 기초를 둔 RPS(Root Power Sum), 가우시안 형태로 스무딩된 선형 리프터(smoothed linear lifter), 지수 함수 리프터(general exponential lifter), 켑스트럼 계수의 높은 차수와 낮은 차수의 바람직하지 못한 변화를 제거하기 위한 밴드 패스 리프터(band pass lifter), 켑스트럼 계수의 통계적인 분포에 따라 가중 함수를 결정한 것 등이 있다[12,13].

본 논문에서는 이러한 여러 가지 켑스트럼 가중 함수들 중에서 RPS에 대하여 실험하였다.

4.2 주성분 분석(PCA: Principal Component Analysis)

주성분 분석은 학습 데이터의 통계적인 특성으로부터 얻어진 변환 행렬에 의한 회전 및 스케일링을 통하여 분별력이 높은 벡터 공간으로 선형 변환시키는 방법이다[14]. 학습 데이터의 공분산 행렬 Σ로부터 고유벡터(eigenvector)와 고유값(eigenvalue)을 계산한 후, 고유벡터들에 의하여 선형 변환 행렬을 구할 수 있다. 고유벡터로 이루어진 행렬을 Φ, 고유값으로 이루어진 대각 행렬을 Λ라고 하면,

$$\Sigma = \Phi \Lambda \Phi' \tag{5}$$

의 관계가 성립되고, 각 차원간의 상관관계가 없어지도록 수직화(orthogonalize)시키는 변환 행렬은 Φ'가 된다.

일반적으로 PCA는 특징 벡터 차원 감소를 주된 목적으로 하며, 고유값이 큰 값들을 갖는 고유벡터들을 선택함으로써 분별력을 높이면서 차원을 감소시켜 준다. 이전의 연구에 따르면, 정적 파라미터와 동적 파라미터를 결합시킨 특징 벡터에 대하여 PCA를 적용하여 벡터 차원을 크게 감소시키면서 인식 성능을 유지하거나 향상시켜 주었다[14].

본 논문에서는 차원 감소의 목적보다는 분별력이 높은 벡터 영역으로의 변환을 주된 목적으로 하며, 정적 파라미터, 즉, 켑스트럼에 대해서만 적용하였다. 켑스트럼의

경우 각 차원에 대한 분산을 같게 해 주는 방법이 잡음에 강인하다고 알려져 있다[13]. 따라서, 변환된 벡터 공간에서도 공분산을 조정해 주어야 한다. 고유벡터로 이루어진 변환 행렬에 의하여 변환된 후의 학습 데이터에서는 차원간의 상관성이 제거되고, 각 차원에 대한 공분산은 Λ에 의하여 결정된다. 만약, 변환 후의 공분산이 단위행렬(identity matrix)이 되도록 백색 잡음화(whitening) 시키려면, 변환 행렬을 $\Lambda^{-1/2}\Phi^t$로 하면 된다. 이렇게 백색 잡음화시킨 영역에서의 유클리디안(euclidean) 거리 측정은 백색 잡음화 이전 영역에서의 Mahalanobis 거리 측정과 같게 된다.

4.3 선형 판별 분석(LDA: Linear Discriminant Analysis)

LDA는 PCA와 같이 전처리 과정에서 분별력이 높은 특성벡터를 얻어내는 것으로서 클래스간의 분별을 최대화하도록 하는 선형 변환을 찾는 방법으로 잡음환경에 효과적으로 적용되었다[15]. 변환은 클래스 내부의 공분산 행렬 W와 클래스들 간의 공분산 행렬 B에 대하여 $tr(W^{-1}B)$을 최대화하도록 정의된다. 여기서 $tr(M)$은 행렬 M의 자취(trace)를 나타낸다.

앞에서의 정의에 의하여, LDA에 의한 변환행렬은 행렬 $W^{-1}B$의 고유벡터들에 의하여 얻을 수 있다. 그런데, $W^{-1}B$가 대칭행렬이 아니므로 고유값 및 고유벡터들의 계산을 쉽게 구할 수 없기 때문에 다음과 같은 변형된 방법을 사용한다. C가 행렬 W를 행렬 L로 대각화하는 유니터리(unitary)행렬이라고 하면,

$$W = CLC^t \tag{6}$$

이다. 먼저 $L^{-1/2}C^t$에 의하여 변환하면 백색 잡음화가 되어, 클래스 내부 공분산은 1이 된다. 이 때, 변환된 공간에서의 클래스간 공분산 S는

$$S = (L^{-1/2}C^t)B(CL^{-1/2}) \tag{7}$$

가 되고 이는 대칭행렬이다. V를 S에 대한 고유벡터들을 열 벡터로 갖는 유니터리 행렬이라고 하면, 변환 행렬 A는 결국

$$A = V^t(L^{-1/2}C^t) \tag{8}$$

와 같이 구해진다. 이 때, 변환된 벡터의 차원을 감소시키려면 V^t에서 고유값이 큰 고유벡터들을 선택한다.

V. 실험 환경 및 인식 시스템

본 장에서는 본 논문에서 사용된 전화 음성 데이터베이스의 구성과 음성 분석 방법 및 인식 시스템에 대하여 설명한다.

5.1 전화 음성 데이터베이스

서울, 경기 지역에 거주하는 20대 남, 녀에게 직접 전화를 걸어, 자연스럽게 발음된 음성을 전화선을 통하여 DAT(Digital Audio Tape)에 48kHz 샘플링으로 녹음하였다. DAT에서 재생시킨 신호를 PC의 사운드 카드를 통하여 16bits, 8kHz 샘플링으로 A/D 변환하였다. 데이터베이스에 사용된 어휘는 음성 다이얼링을 위한 인식 어휘로서 50개의 단어를 임의로 선택하였다. 자세한 구성은 표 1과 같다.

표 1. 데이터베이스의 구성

인식 어휘	음성 다이얼링 50 단어
샘플링 방법	16-bit linear PCM, 8 kHz sampling
학습 데이터	20대 남,녀 각 25명(총 50명), 각 단어 3회씩 발음
인식 데이터	20대 남,녀 각 7명(총 14명), 각 단어 3회씩 발음

5.2 음성 분석 및 특징 추출

샘플링 주파수를 8kHz로 하여 수집된 전화 음성 데이터베이스에 대하여 $1-0.95z^{-1}$의 전달 함수를 갖는 프리엠퍼시스(pre-emphasis) 필터를 사용하여 고역을 강조한다. 신호의 분석은 32ms에 해당하는 256 샘플의 길이를 갖는 해밍윈도우(Hamming window)를 사용하여 16ms(128 샘플)의 간격으로 이동하면서 각 음성 프레임(frame)마다 FFT에 의한 스펙트럼을 얻어 멜 밴드(Mel band)로 변환한 후, 로그 필터 뱅크 에너지에 대하여 DCT를 하여 멜 켑스트럼을 얻는다. 멜켑스트럼으로부터 동적 특징을 얻기 위해서는 차분(difference)에 의하여 구하였다. 현재 프레임을 중심으로 2 프레임 간격에 대한 차분에 의하여 델타(delta) 켑스트럼 계수를 구하고, 델타 켑스트럼 계수들에 대해 다시 전, 후 1 프레임 간격으로 차분하여 델타-델타 켑스트럼 계수를 얻는다. 에너지에 대해서도 위와 같은 방법으로 델타 에너지 및 델타-델타 에너지를 구하여, 이들 결합한 것을 에너지 기반의 특징 벡터로 사용한다. 음성 분석 및 특징 벡터에 대하여 정리하면 표 2와 같다.

표 2. 음성 분석 방법 및 특징 벡터

프리엠퍼시스	$1-0.95z^{-1}$
윈도우 크기	256 샘플 (32 msec)
윈도우 이동 간격	128 샘플 (16 msec)
윈도우 종류	해밍 윈도우
FFT 크기	1024
멜 필터 뱅크 수	18
특징 벡터	켑스트럼 (12차)
	델타 켑스트럼 (12차)
	델타-델타 켑스트럼 (12차)
	델타 에너지 & 델타-델타 에너지 (2차)

5.3 모델 및 인식 시스템

반연속 HMM(semi-continuous Hidden Markov Model)을 기반으로 하여 음성 다이얼링 50개 단어에 대한 고립 단어(isolated-word) 인식 시스템을 구성하였다. 앞에서 구한 4가지 특징 벡터에 대하여 LBG(Linde-Buzo-Gray) 알고리즘을 이용하여 각각의 특징 벡터에 대하여 256의 크기를 갖는 코드북을 생성하였고, 켑스트럼 기반 특징 벡터는 4개, 에너지 기반 특징 벡터는 2개의 혼합 확률(mixture probability)을 사용하여 학습하였다.

학습 모델에 대해서는 다음과 같은 두 가지 형태로 구현하였다. 첫째, 각 단어에 대한 모델을 음절수에 따라 (음절수×5+5)개의 상태 수를 갖는 단어 모델에 각 단어의 전, 후에 2개의 상태 수를 갖는 두 가지 묵음 모델을 결합하여 사용하였다. 둘째, 실험에서 사용된 전화 음성 데이터베이스의 인식 어휘에 나타나는 37개의 문맥 독립 음소 모델을 기반으로 각 3개의 상태 수를 갖는 부분단어(sub-word) 모델을 사용하고, 단어의 전, 후에 앞에서와 같은 2가지의 묵음 모델을 결합하여 모델링하였다.

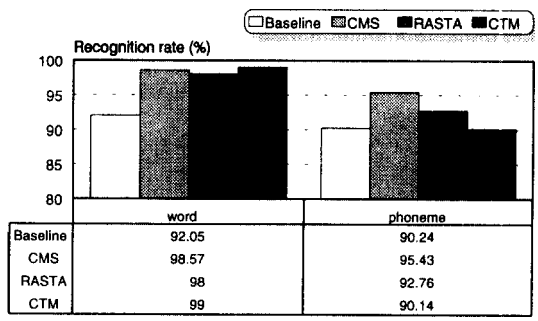
학습 과정에서는 바움-웰치(Baum-Welch) 알고리즘을 이용하여 반복적으로 모델을 추정하였고, 인식 과정에서는 비터비 디코딩(Viterbi decoding) 방법을 이용하여 전체 50단어에 대한 확률을 계산하고, 그 값이 최대가 되는 단어를 선택하도록 하였다.

VI. 실험 및 결과

본 장에서는 단어 모델 및 문맥 독립 음소 모델을 이용하여 채널 보상 방법과 특징 벡터 변환 방법을 실험하고, 그 결과를 분석한다.

6.1 채널 보상 방법에 대한 인식 실험 및 결과

동일한 데이터베이스에 대하여 단어 모델 및 문맥 독립 음소 모델의 두 가지 인식 시스템을 이용하여 채널 보상 방법들에 대한 인식 실험을 하였다. 실험 결과를 그림 4에 제시하였다. 그림에서 "Baseline"은 어떠한 채널 보상 방법도 사용하지 않은 멜켑스트럼 기반의 기준 시스템을 나타낸다.



	word	phoneme
Baseline	92.05	90.24
CMS	98.57	95.43
RASTA	98	92.76
CTM	99	90.14

그림 4. 모델에 따른 채널 보상 알고리즘의 인식 성능 비교

대표적인 바이어스 제거 방법인 켑스트럼 평균 차감법은 간단한 연산량으로 채널 왜곡을 효과적으로 보상해 주어 단어 모델 및 음소 모델에서 모두 성능을 크게 향상시켜 주었다.

RASTA 처리 방법을 적용하였을 때도 인식 성능이 향상되는 데, 그 향상 정도는 인식 모델에 따라 다르게 나타나고 있다. RASTA 처리는 IIR 필터를 사용하므로 현재 시간 이전의 데이터들에 영향받는다. 따라서, 이전의 문맥에 의존적인 특징 벡터를 생성시키게 되는데, 단어 모델의 경우에는 각 단어 내에서의 문맥적인 변화 특성을 완벽하게 표현해 줄 수 있으므로 단어간의 분별력을 높이는 데 유리하게 작용한다. 이에 반하여, 문맥 독립 음소 모델은 특징 벡터에 내재된 문맥을 잘 모델링할 수 없으므로 인식 성능에 악영향을 미치게 된다. 그러므로, 음소 모델의 경우보다 단어 모델의 경우에 인식률 상승폭이 더 크게 됨을 알 수 있다.

켑스트럼-시간 행렬을 구하기 위해서 현재 시간을 기준으로 주위 13개 프레임 동안의 멜켑스트럼의 각 차수에 대하여 시간축으로 1차원 DCT를 하여 최저차를 제외한 3개의 저차항을 취하여 (12×3) 행렬을 추출하였다. 즉, 식 (2)에서 $M=13$이고, $m=1,2,3$에 대하여 얻은 계수로써 켑스트럼-시간 행렬을 얻는다. 인식 실험 결과, 전체단어 모델의 경우 인식률이 크게 상승하여 99.00%로 가장 뛰어난 성능을 얻을 수 있었다. 이에 반해, 부분단어 모델의 경우에는 기준 시스템보다 오히려 인식률이 오히려 낮다. 이러한 이유는 RASTA에서와 같이 문맥에 의한 영향으로 해석할 수 있다. 켑스트럼-시간 행렬에서 시간축 DCT를 계산하는 주위 프레임들의 데이터에 영향을 받게 되므로 RASTA에서와 같이 문맥 의존적인 특징 벡터가 된다. 그림 4에서 RASTA의 인식 성능과 비교해 볼 때, 단어 모델에서는 더 우수하지만 음소 모델에서는 성능이 저조하다. 그러므로, 켑스트럼-시간 행렬의 경우가 RASTA의 경우보다 문맥의 영향에 더 민감함을 알 수 있다.

6.2 특징 벡터 변환 방법에 대한 인식 실험 및 결과

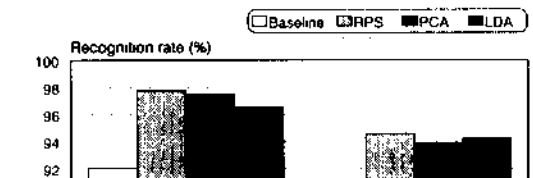
앞에서의 실험에서 살펴보았듯이, 시간적인 정보가 강조된 특징 벡터들은 문맥 독립 음소에 의하여 모델링하기가 어렵다. 그러므로, 문맥 독립 음소 모델을 기반으로 하는 인식 시스템의 성능을 향상시키기 위해서 정적 특징 벡터에 대한 특징 벡터 변환 방법을 적용하여 실험하였다.

먼저 정적 벡터에 대하여 잡음에 강인한 선형 변환 방법들의 성능을 비교하였다. 또한, 선형 변환과 채널 바이어스 제거 방법을 결합한 특징 변환 방법에 의하여 잡음 및 채널 왜곡에 강인한 특징 벡터를 추출하고 이에 대한 성능을 평가하였다.

인식 시스템에 사용되는 4 가지의 특징벡터 중에서 시간적인 정보를 포함하고 있지 않은 켑스트럼에 대하여 가중 함수 및 PCA, LDA를 실험하였는데, 다른 3 가지의 특징벡터에 대해서는 기준 시스템과 동일하게 하였다. 켑스트럼 가중 함수의 경우는 RPS를 사용하였고, PCA와 LDA의 경우에는 벡터 차원 감소의 목적보다는 분별력 있

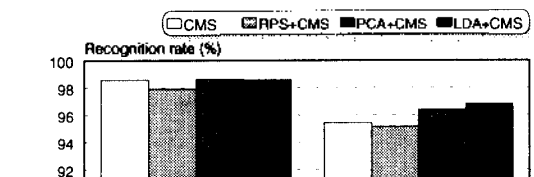
는 벡터 영역으로의 변환을 목적으로 하였기 때문에 차원은 기준 시스템과 같이 12차를 그대로 유지하도록 하였다. PCA의 적용 시 전체 학습 데이터에 대하여 공분산을 계산하고 고유값 및 고유벡터를 구한 후, 4.2절에서 제시한 방법에 의하여 백색 잡음화 과정을 포함하는 신형 변환 행렬을 구하였다. LDA의 경우에는 분별력을 높이고자 하는 클래스를 모델 단위로 정하였다. 즉, 단어 모델에서는 각 인식 단어로 음소 모델에서는 문맥독립 음소로 정의하였다. 먼저, 기준 시스템으로 훈련된 모델을 이용하여 학습 데이터에 대한 비터비 디코딩을 하고, 그 결과에 의해 각 모델에 해당하는 구간을 찾아 클래스를 분류한다. 분류된 클래스를 기준으로 4.3절에서 제시한 과정을 통하여 선형 변환 행렬을 구하였다. 선형 변환과 채널 바이어스 제거 방법을 결합한 특징 변환 방법을 위해서는 채널 바이어스 제거 방법으로 켑스트럼 평균 차감법을 이용하였다.

각각의 특징 벡터 변환 방법에 대한 인식 결과 및 켑스트럼 평균 차감법과 결합하였을 때의 인식 결과를 그림 5와 그림 6에 나타내었다.



	word	phoneme
Baseline	92.05	90.24
RPS	97.76	94.57
PCA	97.48	93.86
LDA	96.52	94.24

그림 5. 선형 변환 방법의 인식 성능 비교



	word	phoneme
CMS	98.57	95.43
RPS+CMS	97.9	95.14
PCA+CMS	98.62	96.38
LDA+CMS	98.57	96.81

그림 6. 선형 변환 및 켑스트럼 평균 차감법의 결합에 대한 인식 성능 비교

켑스트럼 가중 함수인 RPS를 적용하였을 때 인식률이 크게 상승한다. 켑스트럼 평균 차감법(CMS)을 동시에 적용하였을 때는 RPS만 적용한 경우보다는 성능이 향상되지만, 그 향상폭이 아주 작고, 기준 시스템에 CMS만을 적용한 경우보다도 오히려 인식률이 낮다. 즉, 가중 켑스트럴 거리 측정과 CMS의 결합은 그리 효과적이지 못하다.

PCA와 LDA는 그림 5의 결과에서 RPS의 인식률에 미치지 못하였지만, CMS와 결합했을 때는 그림 6의 결과에서처럼 RPS에서의 경우보다 인식률이 높다. 이는 PCA와 LDA가 RPS와는 달리 켑스트럼 계수의 차원간의 상관 관계에 의하여 분별력이 높은 새로운 영역으로 변환시키기 때문이다. PCA와 LDA를 평균 차감법과 결합한 경우, 단어 모델에서는 CMS만을 적용한 경우와 인식 성능이 비슷하지만, 음소 모델에서는 인식 성능이 크게 향상된다.

결과적으로 PCA와 LDA를 채널 바이어스 제거 방법과 결합한 특징 변환 방법에 의하여 전화망 환경에서 우수한 성능을 보여 주며, 특히 문맥 독립 음소 모델에 매우 효과적이다.

VII. 결 론

본 논문에서는 전화망 환경에서 사용되는 음성 인식 시스템의 성능을 개선하기 위하여 특징 벡터 추출 단계에서의 처리 방법들을 연구하였다. 음성 다이얼링을 위한 50 단어의 고립 단어 인식에 대하여 단어 모델과 문맥 독립 음소 모델의 두 가지 인식 시스템을 구성하였다.

먼저 각 시스템에 대하여 채널 바이어스 제거 방법과 시간 정보를 이용한 채널 왜곡 보상 방법에 의한 특징 추출 방법의 성능을 비교하였다. 그 결과, 인식 모델에 따라 다른 성능을 보여 주었다. 대표적인 바이어스 제거 방법인 켑스트럼 평균 차감법은 간단한 연산량으로 채널 왜곡을 효과적으로 보상해 주며, 단어 모델 및 음소 모델에서 모두 성능을 크게 향상시켜 주었다. 그러나, RASTA 처리와 켑스트럼-시간 행렬과 같이 시간적인 정보를 이용하는 특징 벡터 추출 방법들은 문맥에 의존적인 특징 벡터를 생성시키게 되며, 단어 모델에서 성능을 크게 향상시켜 주었다.

두 번째로 정적 특징 벡터에 대한 변환 방법에 의하여 잡음 및 채널 왜곡에 강인한 특징 벡터를 구하는 방법을 제시하고 이를 실험하였다. 정적 특징 벡터인 켑스트럼에 대하여 선형 변환으로 켑스트럼 가중 함수, PCA, LDA의 방법들을 적용할 때 성능이 크게 향상되며, PCA와 LDA는 채널 바이어스 제거 방법인 켑스트럼 평균 차감법과 효과적으로 결합하여 문맥 독립 음소 모델에서 뛰어난 성능을 보여주었다.

본 논문에서의 실험 결과로 볼 때, 전화 환경에서의 음성 인식에서는 시스템이 사용하는 모델에 따라 다른 특징 추출 방법이 적용되어야 한다. 단어 모델을 기반으로 하는 고립 단어 또는 연결음 인식에서는 시간적인 정보를 이용하는 RASTA 처리와 켑스트럼-시간 행렬 등이 효과적이고, 음소 모델을 이용하는 대용량 어휘 인식이나 연속음 인식에서는 정적 특징에 대하여 선형 변환과 채널 바이어스 제거 방법을 결합한 특징 변환 방법이 유효하리라 생각된다.

참 고 문 헌

1. P. J. Moreno and R. M. Stern, "Sources of Degradation of Speech Recognition in the Telephone Network," *Proc. ICASSP*, vol I, pp. 109-112, 1994.
2. R. J. Mammone, X. Zhang, R. P. Ramachandran, "Robust Speaker Recognition - A Feature-based Approach," *IEEE Signal Processing Mag.*, pp. 58-71, September 1996.
3. 전원석, 신원호, 양태영, 김원구, 윤대희, "전화망에서의 음성 인식을 위한 전처리 연구," *한국음향학회지*, 16권 4호, pp. 57-63, 1997.
4. A. Acero and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition," *Proc. ICASSP*, pp. 849-852, April 1990.
5. M. G. Rahim and B. H. 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.
6. A. Sankar and C. H. Lee, "Robust Speech Recognition Based on Stochastic Matching," *Proc. ICASSP*, pp. 121-124, 1995.
7. J. T. Chien, H. C. Wang and L. M. Lee, "Estimation of Channel Bias for Telephone Speech Recognition," *Proc. ICSLP*, pp. 1840-1843, 1996.
8. B. P. Milner, "Inclusion of Temporal Information into Features for Speech Recognition," *Proc. ICSLP*, pp. 256-259, 1996.
9. B. A. Hanson and T. H. Applebaum, "Subband or Cepstral Domain Filtering for Recognition of Lombard and Channel-Distorted Speech," *Proc. ICASSP*, Vol. II, pp. 79-82, 1993.
10. H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. Speech & Audio Processing*, vol. 2, No. 4, pp. 578-589, 1994.
11. B. P. Milner, S. V. Vaseghi, "An Analysis of Cepstral-Time Matrices for Noise and Channel Robust Speech Recognition," *Proc. EUROSPEECH*, pp. 519-522, 1995.
12. J. Junqua and H. Wakita, "A Comparative Study of Cepstral Lifters and Distance Measures for All Pole Models of Speech in Noise," *Proc. ICASSP*, pp. 476-479, 1989.
13. 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.
14. M. Trompf, R. Richter, H. Eckhardt and H. Hackbarth, "Combination of Distortion-robust Feature Extraction and Neural Noise Reduction for ASR," *Proc. EUROSPEECH*, pp. 1039-1042, 1993.
15. O. Siohan, "On the Robustness of Linear Discriminant Analysis As a Preprocessing Step for Noisy Speech Recognition," *Proc. ICASSP*, pp. 125-128, 1995.

▲전 원 석(Won Suk Jun) 1971년 6월 16일생

1996년 8월 : 연세대학교 전자공학과졸업(공학사)

1996년 9월 ~ 1998년 6월 : 연세대학교 대학원 전자공학과 석사과정

1998년 7월 ~ 현재 : LG 종합기술원, 정보기술연구소

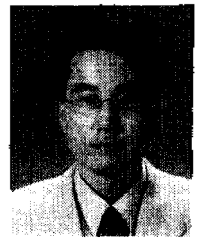
※주관심분야 : 음성인식, 잡음처리

▲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

▲이 충 용(Chungyong Lee)



1983년 3월 ~ 1987년 2월 : 연세대학교 전자공학과(공학사)

1987년 3월 ~ 1989년 2월 : 연세대학교 전자공학과(공학석사)

1990년 3월 ~ 1991년 8월 : 연세대학교 부설 산업기술 연구소 연구원

1991년 9월 ~ 1995년 12월 : Georgia Inst. of Tech., USA, 공학박사

1996년 2월 ~ 1997년 7월 : 삼성전자 System LSI 본부 선임연구원

1997년 9월 ~ 현재 : 연세대학교 기계전자공학부 조교수

※주관심분야 : 통신 신호처리, 비선형 신호처리, 음성인식, 어레이 신호처리

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