방송뉴스 인식에서의 잡음 처리 기법에 대한 고찰*

정용주(계명대)

<차 례>

<Abstract>

A Study on Noise-Robust Methods for Broadcast News Speech Recognition

Yong-joo Chung

Recently, broadcast news speech recognition has become one of the most attractive research areas. If we can transcribe automatically the broadcast news and store their contents in the text form instead of the video or audio signal itself, it will be much easier for us to search for the multimedia databases to obtain what we need. However, the desirable speech signal in the broadcast news are usually affected by the interfering signals such as the background noise and/or the music. Also, the speech of the reporter who is speaking over the telephone or with the ill-conditioned microphone is severely distorted by the channel effect. The interfered or distorted speech may be the main reason for the poor performance in the broadcast news speech recognition. In this paper, we investigated some methods to cope with the problems and we could see some performance improvements in the noisy broadcast news speech recognition.

* Keyword: noise-robust speech recognition, broadcast news recognition

* 본 연구는 한국과학재단 목적기초연구(R05-2001-000-01469-0)지원으로 수행되었음.

1. 서 론

음성인식 기술은 초창기에는 소규모 단어 인식을 위주로 시작되었고, 근래에는 연속어 인식까지 그 기술의 범위가 확대되어 왔다. 특히, 최근에 들어서는 방송뉴스나 비디오 매체물의 오디오에 있는 음성 및 음향을 인식하여 이를 정보 검색에 이용하려는 연구가 진행되고 있다[1,2,3]. CMU의 Informedia 연구는 디지털 라이브러리(digital library) 연구의 일환으로서 기존의 비디오나 오디오 형태의 자료에 들어있는 음성을 인식하여 텍스트 형태로 자료를 변환시킨 후 이를 사용자들이 탐색 기능을 이용하여 자신이 원하는 정보를 쉽고 빠르게 찾아 볼 수 있게 하고자 하는 것이다. 이러한 기능을 수행하기 위해서는 음성인식시스템이 수용할 수 있는 단어가 수만 단어 이상이 되고 또한 다양한 문법구조와 발음현상 등을 고려해야 하므로 많은 연구 과제를 던져준다. 이러한 연구는 결국 무제한 인식시스템을 이루려는 연구의 시작이며 또한 멀티미디어시대의 정보 검색에 매우 필요할 것으로 생각된다. 방송뉴스에 대한 음성인식 연구는 아직 완전하지는 않지만 시제품 제작의 단계까지 발전하였다.

유럽지역의 몇몇 연구소가 공동으로 개발한 방송뉴스 인식시스템인 SPRACH (SPeech Recognition Algorithms for Connectionist Hybrids) 시스템에 관한 연구결과는 다양한 종류의 음질로 구성된 방송뉴스에 대한 인식결과를 제공한다[4]. 예를 들면, 뉴스 앵커가 스튜디오 안에서 원고를 읽는 수준의 깨끗한 음질의 방송으로부터 레포터가 헬기를 타고 가면서 긴장된 목소리로 내보내는 방송까지 여러 가지 형태의 음질에 대한 연구결과를 보여준다. 깨끗한 환경의 방송뉴스 음성에 대해서는 현재의 최신의 음성인식시스템은 약 15% 정도의 단어 오 인식율을 가진다. 이는 10개의 인식단어 중 1.5 개의 단어에서 오류가 발생하는 것으로 대략 해석 될 수 있으며 상당히 좋은 성능을 나타낸다. 이러한 인식결과는 방송뉴스 자동 전사방식의 상업화가 매우 긍정적임을 나타내어 준다. 그러나, 발성환경이 열악해짐에 따라서 인식 성능은 매우 떨어짐을 알 수 있다. 따라서 열악한 환경 하에서의 인식성능을 향상시키기 위해서는 보다 강인한 인식방식을 개발할 필요가 있다.

본 논문에서는 방송뉴스에서 잡음 등의 영향으로 음질이 떨어진 경우에 대해서 인식율의 향상을 위한 방법을 고찰하고자 한다. 특히, 기존의 잡음음성인식에서 가장 널리 사용되는 방법들을 방송뉴스에 적용하여, 최적의 인식율을 얻을 수 있는 방안을 제시하고자 한다. 본 논문의 구성은 2장에서 방송뉴스 인식시스템의 구현에 대한 설명이 있으며, 3장에서는 본 논문에서 이용된 잡음음성인식 기법에 대해서 소개하며, 4장에서 인식결과를 제시한다.

2. 방송뉴스 인식시스템의 구현

본 연구에서는 방송뉴스 인식시스템을 구축하기 위해서 HTK 3.2를 사용하였다[5]. 방송뉴스인식과 같은 연속어 인식시스템을 구축하기 위해서는 음성데이터베이스를 구축하고 이를 이용하여 기본 음향모델을 만들고, 또한 주어진 문장 단위의 텍스트(text)로부터 연속어 인식을 수행하기 위한 문법을 생성하는 것이 필요하다. 다음에는 이에 대해서 간략히 소개한다.

2.1. 음성데이터베이스의 구축

본 연구를 수행하기 위해서는 실제 방송되는 뉴스를 수집할 필요가 있었다. 방송뉴스 데이터베이스를 얻기 위해서 인터넷상에서 방송되는 뉴스를 취합하여 사용하였다. 문장들은 정치관련 뉴스로부터 추출하였으며, 전체적으로 약 1000개의 문장으로 구성되었는데, 시간과 비용 등의 문제 등으로 많은 음성 데이터를 수집하지 못한 점이 아쉽다. 수집된 방송뉴스는 크게 4가지 종류로 나누어지는데, 스튜디오 안에서 아나운서가 발성한 깨끗한 음성과 실외에서 잡음이 조금 있는 경우, 실외의 잡음이 상당히 심한 경우 그리고 음악이 있는 경우로 나누어진다. 이러한 방송뉴스를 이용하여 다양한 조건에서 인식실험이 이루어지게 된다. 한편, 방송뉴스를 이용한 음성인식기의 구현에 있어서, 수집된 방송뉴스 데이터의 양이 너무 작은 관계로 좀 더 신뢰성 있는 연구결과를 얻기 위해서 KAIST에서 제공된 3000 단어의 무역상담 연속어 음성데이터 베이스를 이용하여 인식실험을 병행하였다[1].

2.2 기반 연속어 음성 인식기의 구축

2.2.1 음성데이터의 파라미터화

음성인식을 위해서는 음성파형을 특징 벡터의 열로서 바꾸어 주는 것이 필요하다. 본 연구에서는 Mel-Frequency Cepstral Coefficients (MFCCs)를 특징벡터로서 사용하였다. 여기서는 mel-scale의 주파수 특성을 갖는 필터뱅크 값들이 Discrete cosine transform (DCT)를 이용하여 MFCC로 변환된다. 이외에도 고차의 MFCC 값들이 너무 작은 값을 가지는 것을 보완하기 위해서 켑스트럼 liftering 과정을 거치게 된다. 특히, 음성인식시스템의 성능향상을 위해서는 단순한 MFCC 외에도 시간변화값을 특징벡터로서 사용하는 것이 필요하다. 이를 위해서는 시간변이를 고려하여 MFCC 계수간의 regression 계수 값을 이용한 델타(delta) 와 델타-델타의 계수들을 사용하였다[7]. 따라서, 13차의 MFCC 계수와 더불어 전체적으로는 39차의 특

징벡터들이 사용된다.

2.2.2 음소 HMM의 구현

각각의 음소 HMM은 5개의 연속적인 state 로 구성되어 있으며, 이중에서 처음과 마지막 state는 단지 전후에 오는 다른 음소 HMM 과의 연결 역할만을 하며, 가운데 3개의 state는 각 음소의 음성특징을 결정하도록 한다. 각 음소 state는 여러 개의 mixure의 결합으로 구성된 Gaussian mixture 확률분포로서 결정된다. 따라서, HMM의 모델을 구현하는 것은 Gaussian 확률분포의 평균벡터와 분산 행렬를 추정하는 것이 된다. 이러한 추정방법으로는 가장 많이 사용되는 기법은 Baum-Welch 알고리듬을 이용하였으며[8], 여기서는 ML (maximum likelihood) 추정 기법에 의해서 각 mixture 별로의 평균값과 분산 값을 추정하게 된다.

2.2.3 Tied_State 트라이폰(Triphone)의 구현

음소 HMM의 훈련이 끝난 후, 이를 바탕으로 문맥종속(context-dependent)의 트라이폰(triphone) 모델을 구현하였다[9]. Tied-state 트라이폰 모델링에서는 비슷한 음향정보를 가진 트라이폰의 state들을 동일한 상태(tied-state)로 간주된다. Tied-state 트라이폰 모델의 구현 과정에서 가장 중요한 부분은 tied-state 를 결정하는 과정이다. 본 연구에서는 결정트리(decision tree) 방식을 사용하였다[10]. 결정트리 방식은 각각의 트라이폰에 대해서 좌 우측의 문맥에 관해서 질의하는 방식으로, 음성데이터의 log-likelihood 값에 근거하여, 유사도가 멀리 떨어져있는 최적의 클러스터들을 형성하고자 한다. 예를 들면, 처음에는 어느 음소에 대한 모든 트라이폰의 특정 state들이 하나의 클러스트를 형성한다. 그리고 각각의 질의를 전체클러스터에 적용한다. 이때, 훈련데이터의 log-likelihood를 가장 많이 증가시키는 질의를 이용하여 클러스터는 분리되며, 이러한 작업은 log-likelihood값이 어느 정도 이하로 증가될 때까지 계속되어 클러스터의 수를 증가시키게 된다. HTK 3.2에서는 클러스터의 생성을 제한하는 임계치(TB)를 실험적으로 정하도록 하고 있다. 결정트리 방식은 훈련 중에 나타나지 않은 트라이폰이 인식 중에 나타날 경우, 이 트라이폰을 가장 유사한 tied-state 트라이폰으로 매핑하는 유용한 기능이 있다.

2.2.4 문법의 구성

연속어 음성인식을 위한 문법은 그 형태에 따라서 다양하게 존재한다. 현재 많이 사용되고 있는 문법들로는 word-pair, bigram, trigram 등이 있다[11]. 일반적으로는 bigram 형태의 문법이 많이 사용되는데, 본 연구에서도 주어진 텍스트 문장 수

의 한계 등을 고려하여 bigram 문법을 사용하였다. HTK에서는 HLM이라는 라이브러리를 제공하고 있는데, 본 연구에서는 이를 활용하였다. bigram은 단순히 두개의 단어가 연 이어서 일어날 확률을 추정하는데. 이렇게 할 경우, 자주 발생하지 않는 단어의 시퀀스 값에 대한 확률이 0에 가까워지는 문제가 발생한다. 이를 방지하기 위해서 bigram 모델에서는 큰 값을 갖는 bigram count로부터 일정부분을 빼주고, 이 빼준 값만큼 잘 발생하지 않는 bigram count 값에 더해 주는 과정을 수행하는데, 이를 back-off bigram 이라 한다. Back-off bigram은 일반적인 연속어 음성인식에서 많이 쓰이고 있는데, 본 연구에서도 이를 이용하여, 문법을 구성하였다.

또한 문법의 구성에서 중요한 문제는 인식단위의 선정이다. 즉, 인식단위는 문법을 구성하기 위한 기본 단위인 셈이다. 영어권에서는 인식단위를 단어(word)로 사용한다. 그러나, 한국어에서는 연속어 인식인 경우에 단어를 인식단위로 사용할 경우, 단어의 수가 너무 많아지므로, 훈련 시에 발생하지 않는 단어가 인식 시에 발생하여 out-of-vocabulary 문제를 발생하는 경우가 자주 일어나게 된다. 따라서, 한국어 연속어 음성인식에서는 단어대신 형태소(morpheme)를 인식단위로 사용하는 경우가 있다. 최근에는 형태소를 좀 더 발전시킨 형태의 pseudo-morpheme 형태를 이용한 연구결과도 발표되고 있다 [12]. 본 연구에서는 잡음신호에 강인한 음성인식 방법에 연구의 초점을 맞추었으므로, 인식단위의 최적화에 대한 연구는 차후의 연구과제로 생각을 하고, 일단, 어절 단위를 이용한 음성인식시스템을 구축하였다.

3. 잡음음성인식 기법에 대한 고찰

잡음음성인식을 위해서는 많은 연구 방법들이 제시되어왔다. 본 연구에서는 가장 대표적이라 할 수 있는 잡음 음성인식 기법을 방송뉴스 인식을 위해서 적용하였다. 다음에는 그 기법들에 대해서 간략히 설명한다.

3.1 주파수 차감법(Spectral Subtraction)

주파수 차감법은 잡음 환경에서의 음성인식에서 효과적인 것으로 잘 알려져 있다. 일반적인 주파수 차감법에서는, 잡음신호에 대한 short-term의 주파수 성분 크기를 다음과 같이 잡음음성신호의 스펙트럼에서 빼주도록 한다[13].

$$Y_i(\omega) = H_i(\omega) X_i(\omega)$$

$$H_i(\omega) = (|\widehat{X}_i(\omega)| - |\widehat{N}_i(\omega)|)/|\widehat{X}_i(\omega)|$$

여기서,

$$|\widehat{X}_i(\omega)| = \lambda_x |\widehat{X}_{i-1}(\omega)| + (1-\lambda_x)|X_i(\omega)|$$

$$|\widehat{N}_i(\omega)| = \lambda_n |\widehat{N}_{i-1}(\omega)| + (1-\lambda_n)|N_i(\omega)|$$

이며, $|X_i(\omega)|$와 $|N_i(\omega)|$는 각각 i 번째 프레임의 음성신호와 잡음신호의 short-time 주파수 성분의 크기이다. 일반적으로 λ_x와 λ_n의 값은 0과 1 사이에서 결정된다.

3.2 CMN (Cepstral Mean Normalization)

CMN 방식은 구현하기가 매우 쉬운 점이 큰 장점이며, 그 효과 면에서 아주 우수한 것으로 많은 연구에서 발표되고 있다. 한편 CMN 방식은 대부분의 음성인식시스템에서 활용되고 있다. 여기서는 켑스트럼 특징벡터의 평균값을 구한 다음, 각각의 특징벡터에서 이 평균값을 빼 주게 된다. 이 때의 평균값은 매 문장마다 계산 될 수도 있으며, 또는 누적 평균된 값을 사용할 수도 있다. 만일 문장단위로 이루어질 경우 다음과 같이 주어진다.

$$\hat{O}^c(t) = O^c(t) - \frac{1}{T}\sum_{t=1}^{T} O^c(t)$$

만약, 순수하게 시불변의 콘벌루션 잡음만이 존재할 경우에는 위와 같은 작용을 통해서 원래의 음성신호를 완전하게 복원할 수 있을 것이다. 이 방법을 통해서 음성데이터의 변이를 줄일 수 있으며, 효과적으로 채널잡음을 줄일 수 있을 것이다. 이 과정은 훈련 및 인식용 데이터에 동시에 적용이 되어야 한다. 이 방법은 채널이 단순한 선형적 모델링에 맞지 않을 경우에는 잘 적용이 되지 않을 것이다.

3.3 PMC (Parallel Model Combination)

최근에 많은 관심을 받고 있는 PMC 방식은 기존의 방식과는 다르게 HMM의 모델 파라미터를 잡음음성에 맞도록 변환하는 방식이다[15]. 아래에서는 이 방식에 대해서 간략히 소개하고자 한다. PMC 방식의 구현에서는 켑스트럴 영역의 HMM 파라미터들을 선형주파수 영역으로 변환하는 과정이 먼저 이루어진다. 그런 다음 잡음이 섞인 음성(noisy speech) HMM모델을 만들기 위해서 원래의 깨끗한 음성(clean speech) HMM 파라미터 값과 잡음(noise)에 대한 HMM 파라미터 값을 선형주파수 영역에서 서로 결합하여 준다. 이에 대한 자세한 과정은 다음과 같이

요약된다.

단계 1) 먼저, 잡음과 음성신호의 로그스펙트럼 영역의 HMM 파라미터 값을 역DCT (Discrete cosine transformation) 변환을[4] 이용하여 주어진 각각의 켑스트럼 영역의 HMM 파라미터 값으로부터 구한다.

$$\mu^{l} = C^{-1}\,\mu^{c} \quad , \quad \Sigma^{l} = C^{-1}\,\Sigma^{c}\,(C^{-1})^{T} \qquad (1)$$

여기서 μ^{l} 과 Σ^{l} 는 로그 스펙트럼 영역에서의 평균벡터와 공분산 행렬이며, μ^{c} 와 Σ^{c}는 켑스트럼 영역에서의 값이다.

단계 2) 로그스펙트럼 영역의 HMM 파라미터 값들을 선형 스펙트럼 영역으로 변환한다.

$$\mu_i = \exp(\mu_i^{l} + \frac{\Sigma_{ii}^{l}}{2}), \quad \Sigma_{ij} = \mu_i\,\mu_j[\exp(\Sigma_{ij}^{l}) - 1] \qquad (2)$$

여기서 μ_i 와 Σ_{ij} 는 선형영역의 파라미터 값인 평균벡터 μ 와 공분산 행렬 Σ 의 구성 원소이다.

단계 3) 단계1과 2에서 구한 음성과 잡음의 HMM 파라미터 값을 선형영역에서 결합한다.

$$\hat{\mu} = \mu + \bar{\mu} \quad , \quad \hat{\Sigma} = \Sigma + \bar{\Sigma} \qquad (3)$$

여기서 $\hat{\mu}$ 와 $\hat{\Sigma}$ 는 선형영역에서의 잡음음성의 평균벡터와 공분산 행렬이고, $\bar{\mu}$ 와 $\bar{\Sigma}$ 는 잡음에 관한 것이다.

단계 4) 결합된 HMM 파라미터값에 대해서 다음과 같이 로그변환과 DCT 변환을 취함으로서 최종적으로 켑스트럼 영역에서의 잡음음성의 평균벡터 $\hat{\mu}^{c}$ 와 공분산 행렬 $\hat{\Sigma}^{c}$ 이 얻어진다.

$$\hat{\mu}_i^{l} = \log(\hat{\mu}_i) - \frac{1}{2}\log(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^{2}} + 1), \quad \hat{\Sigma}_{ij}^{l} = \log(\frac{\hat{\Sigma}_{ij}}{\hat{\mu}_i\,\hat{\mu}_j} + 1) \qquad (4)$$

$$\hat{\mu}^{c} = C\;\hat{\mu}^{l}, \quad \hat{\Sigma}^{c} = C\;\hat{\Sigma}^{l}\;C^{T} \qquad (5)$$

Log-add PMC 방식은 위의 식 (2), (4)에서 잡음음성 HMM 평균벡터 값을 구하는 과정에서 잡음과 음성신호HMM의 공분산 값이 충분히 작다는 가정을 함으로서 변환공식이 다음과 같이 단순화된다.

$$\hat{\mu}_i^l = \log(\exp(\mu_i^l) + \exp(\overline{\mu}_i^l)) \qquad (6)$$

이와 같이 로그영역의 평균벡터 값을 구한 다음 위의 식 (5)을 이용하여 켑스트럼 영역의 평균벡터를 구함으로서 log-normal 방식에 비해서 훨씬 간단하게 계산과정을 마칠 수 있다.

4. 연구결과

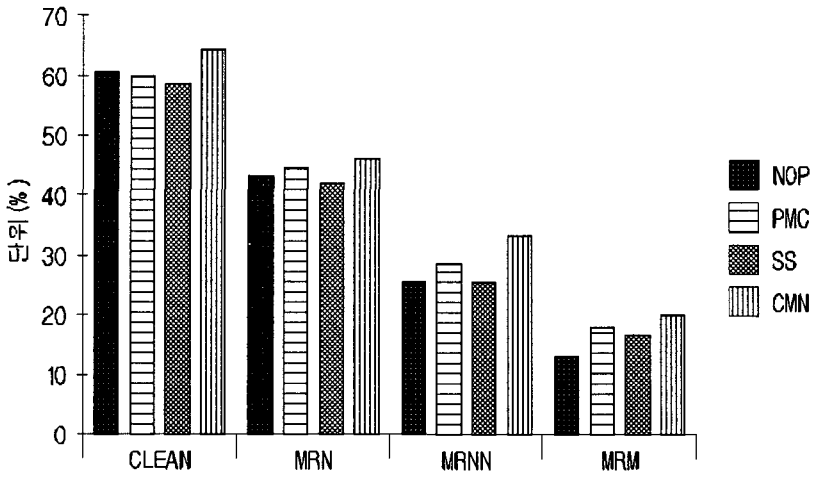
4.1. 방송뉴스 인식에서의 강인성을 위한 연구

방송뉴스에 대한 인식실험을 위해서 기반인식기의 구축이 이루어졌다. 이를 위해서는 tied-state 트라이폰에 기반한 모델링이 이루어졌다. 인식단위로서는 어절을 이용하였는데, 여기서는 주로 잡음음성에 대한 인식률의 향상을 목적에 두었으므로 인식단위에 대한 최적화에 대한 연구는 하지 않았다. 하지만, 형태소를 인식단위로 사용한 경우는 다소 인식성능이 저하되는 것을 확인할 수 있었다. 방송뉴스는 크게 4가지 종류로 나누어진다. 먼저, 기반인식기를 구축하기 위해서 사용된 음성데이터(전체 380 문장, CLEAN)는 스튜디오 안에서 아나운서가 발성하는 깨끗한 음질의 음성신호이며, 그밖에 음악이 존재하는 경우(MRM)의 13문장, 잡음이 어느 정도 있는 경우(MRN)의 70문장 그리고 잡음이 매우 심한 경우 (MRNN)의 70 문장 등으로 구성되어 있다. CLEAN 문장은 그 중에서 310개가 기반인식기의 훈련에 사용되었으며, 나머지 70개는 인식실험에 사용되었다. MRM, MRN 그리고 MRNN 는 모두 인식실험에 사용되었다. <표 1>에는 기반인식기에 대한 인식결과를 나타낸다. 여기서 문법은 back-off의 bigram을 사용하였고, bigram 구성을 위해서 훈련 및 인식용 텍스트를 함께 사용하였는데, 이는 훈련용 텍스트의 부족함에 기인한다.

<표 1> 방송뉴스에 대한 기반인식기의 인식률(Correct Percentage)(%)

인식환경 / mixture 수	CLEAN	MRN	MRNN	MRM
1	54.7	39.8	23.3	11.7
4	60.7	42.9	25.5	8.9
6	60.6	43.2	25.5	13.1

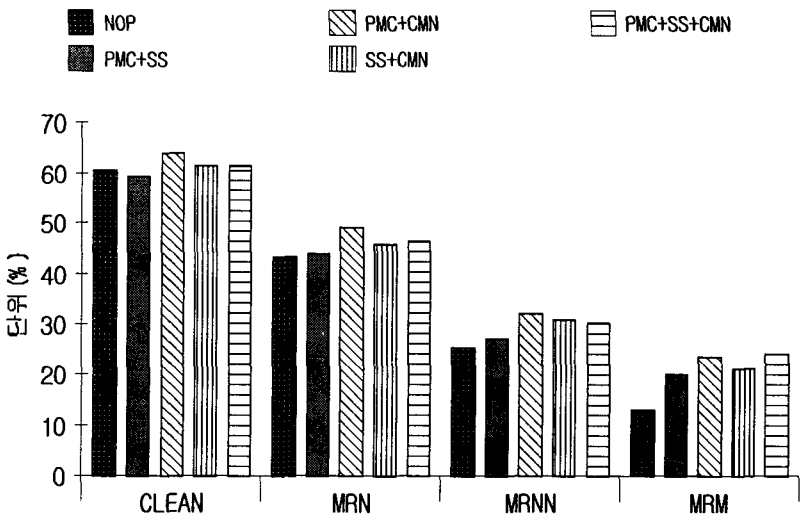
위의 결과에서 알 수 있듯이 기반인식기는 전반적으로 mixture의 개수가 4와 6인 경우에 최고의 인식률을 나타내었다. 인식환경 중에서 CLEAN인 경우에 가장 높은 인식율을 보이고 있으며, 잡음이 존재할수록 인식율은 점점 낮아지는 것을 알 수 있었다. 특히, 음악이 존재하는 MRM의 경우는 인식율의 저하가 몹시 심하게 나타나는 것을 확인할 수 있었다. 예상한대로, 방송뉴스의 종류에 따라서 인식성능의 변화가 몹시 크게 나타난다는 것을 알 수 있었는데, 잡음이 있는 경우에 대처하는 방식을 적용하여 이들의 인식성능을 향상시킬 필요가 있다는 것을 알 수 있었다.



<그림 1> 잡음적응 방식을 방송뉴스에 적용한 경우의 인식률(Correct Percentage)비교
(HMM의 각 state당 mixture 개수는 6일경우)

<그림 1>에는 잡음에 대한 적응 방식중 PMC 방식을 적용한 경우(PMC), 주파수 차감법을 적용한 경우(SS), CMN(cepstral mean normalization)을 각각 적용한 경우의 인식율이 나타나 있다. 인식환경은 CLEAN인 경우, MRN, MRNN 그리고

MRM 등에 대해서 실험하였다. 한편 비교를 위해서 아무런 적응 작업도 하지 않은 경우(NOP)도 함께 나타내었다. 이때, mixture의 개수는 6으로 하였다. 다소의 정도의 차이는 있었지만, 대부분의 잡음 적응방식은 인식율을 향상시키는 것을 볼 수 있었다. 특히, CMN 방식은 다른 방식에 비해서 인식율은 월등히 향상시키는 것으로 나타났다. 이것은 방송뉴스의 음성데이터 자체에 콘벌루션(convolution) 잡음이 상당히 존재하였던 것이 주요 원인이었을 것으로 생각되며, 아마도 방송뉴스를 인터넷상에서 수집한 관계로 채널잡음이 많이 존재한 것으로 보인다. 주파수차감법은 다소 불안정한 모습을 보였는데, PMC 방식이나 CMN 방식에 비해서도 인식율이 떨어지는 것을 확인할 수 있었다. 특히, CLEAN인 환경에 적용했을 경우, 오히려 인식율의 저하를 가져왔다. 이것은 주파수 차감법이 원래의 깨끗한 음성을 복원하고자 하는 과정에서 잡음의 주파수 스펙트럼을 추정하는데, 이때의 추정오차가 미치는 영향이 잡음이 적은 경우에는 오히려 인식성능을 저하시키는 것으로 판단된다.



<그림 2> 잡음적응방식을 결합하여 방송뉴스에 적용한 경우의 인식률 (Correct Percentage) 비교
(HMM의 각 State당 mixture 개수는 6일 경우)

<그림 2>에는 앞에서 적용한 각각의 적응 알고리듬들을 결합하여 적용한 경우의 결과를 나타내고 있다. 이때도 역시 mixture의 개수는 6으로 하였다. 실험결과에 따르면, PMC 방식과 CMN 방식을 결합하였을 경우가 가장 인식율이 좋은 것으로 나타나있다. 오히려 주파수 차감법을 적용한 경우는 MRM 경우를 제외하고는 인식율이 저하되는 것을 알 수 있었다. PMC와 CMN 방식은 각각 부가잡음과

콘벌루션 잡음을 제거하는데 우수한 성질을 가지고 있으므로, 가장 적당한 결합이라고 생각되며, 제시된 연구결과는 이러한 예상에 일치하는 것으로 보인다.

<표 3> PMC 방식과 PMC+CMN 방식을 방송뉴스인식에 적용한 경우의 인식율(Correct Percentage) 비교

인식환경 / 적응방식	CLEAN	MRN	MRNN	MRM
CMN	64.4	46.2	33.3	20.0
PMC+CMN	63.9	49.1	32.1	23.5

<표 3>에는 CMN이 단독으로 사용되었을 경우와 CMN+PMC 방식이 사용되었을 경우에 대하여 좀 더 구체적인 인식율을 제시하고 있다. 전반적으로 CMN+PMC 방식이 잡음음성에 대해서 더 우수한 성능을 보이는 것으로 나타나지만, MRNN의 경우는 CMN 단독의 경우가 다소 우수한 것으로 나타나고 있다. PMC의 적응을 위해서는 잡음구간의 추출이 필요한데, 그러한 구간 추출에서의 오류가 인식율의 저하를 불렀을 가능성이 크다고 생각한다.

5. 결 론

본 논문에서는 잡음에 의해서 음질에 왜곡이 발생한 방송뉴스에 대한 강인한 음성인식방법에 대해서 논의하였다. 이를 위해서는 트라이폰 단위의 HMM 모델링을 이용한 대용량의 연속어 음성인식기를 구축하였다.

잡음 환경 하에서 방송뉴스에 대한 인식성능의 향상을 위해서 PMC 방식, CMN 방식 그리고 주파수차감법 등을 사용하여 인식성능의 향상을 꾀하였다. 그 중에서도 CMN 방식이 인식성능의 향상에 크게 기여하는 것을 볼 수 있었는데, 이것은 방송뉴스 음성데이터에 포함된 채널잡음을 효과적으로 제거했기 때문으로 생각된다. 또한 여러 가지 방식을 결합하여 적용한 결과 부가잡음에 효과적인 PMC 방식과 콘벌루션 잡음에 효과적인 CMN 방식을 결합한 경우 최적의 인식성능을 보임을 알 수 있었다.

참 고 문 헌

[1] J. L. Gauvain, G. Adda et al., "*Transcribing Broadcast news: The LIMSI Nov96 Hub4 System*", *DARPA Speech Recognition Workshop*, pp.56, 1997.

[2] R. Frederking, A. Rudinsky er al., "*Interactive Speech Translation in the DIPLOMAT Project*", Spoken Language Translation Workshop of the Association for Computational Lingustics, ACL-97, Madrid, Spain, July 7-12, 1997.

[3] A. G. Hauptmann, M. J. Witbrock et al., "*Artificial Intelligence Techniques in the Interface to a Digital Video Library*", ACM CHI'97 Conference on Human Factors in Computing Systems, New Orleans LA, March 1997.

[4] Takeo kanade, *Informedia Digital Library System, Annual Progress Report*, 1997, Computer Science Department, CMU

[5] G. Cook et. al., "*An Overview of the SPRACH System for the Transcription of Broadcast News*", DARPA Broadcast News Transcription and Understanding Workshop, Herndon VA, Feb. 1999. ICASSP '96.

[6] S. B. Davis and P. Mermelstein, "*Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*", *IEEE ASSP*, Vol. 28, No. 4, Aug. 1980.

[7] S. Furui, "*Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum*", *IEEE ASSP*-34, 52-59, Feb. 1986.

[8] L. E. Baum, G. S. T. Petrie et al., "*A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*", *Ann. Math., Statist.*, vol. 41, pp.164-171, Jan. 1970.

[9] L. Deng et al,. "Acoustic Recognition Component of an 86,000 word speech recognizer", Proc. ICASSP 90, Albuquerque, NM, pp.741-744, April, 1990.

[10] W. Reichl and W. Chou. "*Robust decision tree state tying for continuous speech recognition*". *IEEE Transactions on Speech and Audio Processing*, 8(5), sep 2000.

[11] M. Weintraub et. al, "*Linguistic constraints in Hidden Markov Model based Speech Recognition*", *Proc. ICASSP 89*, Glasgow, Scotland, pp.699-702, May 1989.

[12] O. Kwon and A, Waibel, "*Korean Broadcast News Transcription Using Morphoneme-based Recognition Units*", *ASK* Vol. 21, March 2002.

[13] P. Lockwood and J. Boudy, "*Experiments with a nonlinear spectral subtraction, hidden Markov models and the projection for robust speech recognition in cars*", Eurospeech 1991.

[14] Y.J. Chung, "*Adaptation method using expectation-maximization for noisy speech recognition*". *IEE Electronics Letters*, June, 2002.

[15] M. J. F. Gales, *Model Based Techniques for Noise Robust Speech Recognition*, Ph. D. Dissertation, University of Cambridge, 1995.

[16] F. Liu et al, "*Efficient Cepstral Normalization For Robust Speech Recognition*", proceedings ARPA human language technology workshop March 1993.

[17] C. J. Leggetter and P. C. Woodland, "*Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models*", *Computer Speech and Language*, vol. 9, pp.171-185, 1995.

접수일자: 2004년 4월 11일
게재결정: 2004년 6월 7일

▶ 정용주(Chung yongjoo)
주소: 대구광역시 달서구 신당동 1000번지
소속: 계명대학교 전자공학과
전화: 053-58-5925
FAX: 053-580-5165
E-mail: yjjung@kmu.ac.kr