한국음향학회지 제32권 제3호 pp. 252~261 (2013)
The Journal of the Acoustical Society of Korea Vol.32, No.3 (2013)
http://dx.doi.org/10.7776/ASK.2013.32.3.252

pISSN : 1225-4428
eISSN : 2287-3775

주파수 변이를 이용한 Parallel Model Combination 모델 적응에 기반한 잡음에 강한 음성인식

Noise Robust Speech Recognition Based on Parallel Model Combination Adaptation Using Frequency-Variant

최숙남, 정현열†

(Sook-Nam Choi, Hyun-Yeol Chung)

영남대학교 정보통신공학과

(접수일자: 2013년 2월 18일; 수정일자: 2013년 4월 15일; 채택일자: 2013년 4월 30일)

초 록: 일반적인 음성인식 시스템은 조용한 인식 환경에서는 높은 인식성능을 나타내지만 잡음이 존재하는 실제 환경에서는 그 성능이 급격히 저하한다. 본 논문에서는 다양한 잡음환경에서도 강인한 음성인식기를 구현하기 위하여, 주파수의 변이도를 이용하여 음성인식을 위한 환경 정보를 얻고 이를 음성 인식을 위한 모델 개선에 적용하여 성능 향상을 도모하는 환경정보 지식에 기반한 주파수 변이 적응 PMC (Parallel Model Combination adaptation using frequency-variant based on environment－awareness : FV-PMC) 방법을 제안한다. 이 방법은 미리 분류된 각 잡음 군 간의 평균 주파수 변이도를 미리 계산하여 임계치로 설정하고 미지의 잡음이 포함된 음성이 입력되면 각 잡음 군과의 주파수 변이도를 다시 계산하여 해당 잡음군의 임계치 보다 높을 경우 그 잡음 군의 잡음이 포함된 음성으로 간주하여 이 잡음 군이 포함된 음성을 이용하여 생성된 인식모델을 이용하여 음성인식을 수행한다. 제안한 FV-PMC 방법을 이용하여 잡음을 분류 하였을 경우 평균 분류 정확도는 56%를 보였고 이를 이용해 음성인식 실험을 실시한 결과 Set A의 평균인식률은 79.05%, Set B의 평균인식률은 79.43%, Set C의 평균인식률은 83.37%로 나타났다. 전체 평균인식률 80.62%로 기존의 깨끗한 모델을 이용한 PMC 인식률 74.93% 보다 5.69% 향상된 결과를 보여 제안한 방법의 유효성을 확인할 수 있었다.

핵심용어: PMC, GMM, 주파수 변이 , 환경인식, 잡음 모델, FV-PMC

ABSTRACT: The common speech recognition system displays higher recognition performance in a quiet environment, while its performance declines sharply in a real environment where there are noises. To implement a speech recognizer that is robust in different speech settings, this study suggests the method of Parallel Model Combination adaptation using frequency-variant based on environment-awareness (FV-PMC), which uses variants in frequency; acquires the environmental data for speech recognition; applies it to upgrading the speech recognition model; and promotes its performance enhancement. This FV-PMC performs the speech recognition with the recognition model which is generated as followings: i) calculating the average frequency variant in advance among the readily-classified noise groups and setting it as a threshold value; ii) recalculating the frequency variant among noise groups when speech with unknown noises are input; iii) regarding the speech higher than the threshold value of the relevant group as the speech including the noise of its group; and iv) using the speech that includes this noise group. When noises were classified with the proposed FV-PMC, the average accuracy of classification was 56%, and the results from the speech recognition experiments showed the average recognition rate of Set A was 79.05%, the rate of Set B 79.43%m, and the rate of Set C 83.37% respectively. The grand mean of recognition rate was 80.62%, which demonstrates 5.69% more improved effects than the recognition rate of 74.93% of the existing Parallel Model Combination with a clear model, meaning that the proposed method is effective.

Keywords: Parallel model combination, Gaussian mixture model, Frequency-variant , Environment-awareness, Noise model, FV-PMC

PACS numbers: 43.71. Gv

†**Corresponding author:** Hyun-Yeol Chung (hychung@yu.ac.kr)
Department of Information and Communication,Yeongnam University 280, Daehak-ro, Gyeongsan, Republic of Korea.
(Tel: 82- 53-810-2496)

I. 서 론

일반적인 음성인식 시스템은 잡음이 없거나 비교적 조용한 실내 환경에서는 좋은 성능을 나타낸다. 그러나 실제 잡음이 혼재하는 환경에서 이용할 경우에는 다양한 잡음들에 의하여 인식 성능이 현저히 저하된다. 이는 실제 환경에서 존재하는 잡음으로 인해 훈련 조건과 인식 조건 사이의 불일치에 기인하기 때문이다.[1] 따라서 잡음에 강인한 음성 인식시스템의 구현을 위해서는 이러한 다양한 잡음으로 인해 야기된 훈련 조건과 인식 조건 사이의 불일치를 보상할 필요가 있다. 이를 보상하기 위한 방법은 음성강화(speech enhancement), 잡음에 강한 특징추출(robust feature extraction), 잡음에 강한 거리측도(robust distance measure)를 이용한 방법, 모델에 기반을 둔 보상방법(model-based compensation) 등이 있다.[2]

음성강화란 배경잡음으로 오염된 음성에서 부가잡음을 제거하고 음성의 질이나 명료도(intelligibility)를 향상시키는 방법을 말한다. 음성강화 알고리즘에는 스펙트럼 크기의 예측에 의한 방법으로 Spectral Subtraction, MMSE(Minimum Mean Square Error), Wiener filtering 등이 있다.[3,4] 잡음에 강한 특징 추출방법으로 대표적인 것으로는 MFCC(Mel-Frequency Cepstral Coefficient), PLP(Perceptual Linear Prediction), SMC (Short-time Modified Coherence) 등을 들 수 있다.[5] 그리고, 인식환경의 변화를 보상하기 위한 특정 파라미터 영역에서의 처리 기법들이 있다. 켑스트럼과 같은 특징 파라미터를 정규화하기 위한 가장 간단한 방법으로서 켑스트럼 벡터의 차수별로 통계적 평균치를 차감하는 방법인 켑스트럼 평균 정규화(Cepstral Mean Nomalization, CMN)등의 기법이 있다.[6]

잡음에 강인한 거리척도를 이용한 방법으로는 음성인식을 위한 특징벡터로 켑스트럼 벡터를 주로 이용하며 인식률 향상을 위해서는 켑스트럼 계수에 가중치를 가해 거리측정을 하는 weighted cepstral distance measure 방법이 널리 연구되어 왔다.[7,8]

또한 다양한 인식환경에서 발생할 수 있는 훈련환경과 인식 환경 사이의 부정합을 보상하기 위한 모델 파라미터에 대한 수정이 요구되는데 그 중의 한 가지 방법이모델 보상방법인 PMC이다. PMC는 훈련환경과 인식환경 사이에 부정합이 나타나지 않을 때 음성인식시스템이 최적의 성능을 보인다는 점에 중점을 두고 간섭부가 잡음 (interfering additive noise)이 있는 경우를 고려한다. 이 경우, 부가적인 잡음(additive noise) 이 부정합성에 나타나는 영향을 알 수 있다면 새로운 테스트 환경에 정합(matching) 시키기 위해서 훈련 데이터를 수정하거나 재훈련 시킬 수 있을 것이다.[9] 이러한 PMC 방법은 다양한 연구가 현재까지 진행되어 우수한 성능을 보이고 있는 방법 중 하나이다. PMC 방법들중 파라미터를 보상하는 방법으로는 공분산의 수축-확대 방법을 동적 파라미터 보상과정에 적용하는 방법[10]에서부터 최근의 정적 및 동적 파라미터의 통합 보상 방법에 이르기까지 다양한 연구가 진행되고 있다.[11] 또한 음성의 전처리 단계에서 MWF(Mel-warped Wiener Filtering) 기법을 이용하여 개선한 음성의 묵음 구간으로부터 잔류 잡음을 취하여 무잡음 모델을 보상함으로써 잡음 환경하의 음성 인식 성능을 향상시키는 방법과[12] 과 PMC 방법으로 모델보상을 하여 생성된 잡음음성을 MMSE를 통하여 잡음을 추정한후, 필터 가중치에 적용하는 후처리 방법도 연구가 되고 있다.[13]

본 논문에서는 인식환경에서 발생할 수 있는 다양한 잡음들을 몇 가지 잡음 군으로 분류하여 각 군별 잡음을 이용하여 인식모델을 훈련한 후, 분류된 잡음 군에 속하는 잡음 환경 하에서 발성된 음성이 입력될 때 이 신호에 포함되는 잡음의 종류를 추정하고 추정된 잡음 군으로 훈련된 인식 모델을 이용할 경우 보다 개선된 음성인식 성능을 달성할 수 있을 것으로 기대할 수 있다.

한편 유사한 특성을 가진 파라미터를 분류하는 척도로서는 주파수 변이도를 이용한 방법이 많이 이용된다.[14] 주파수 변이도는 음성개선 알고리즘의 평가를 하는 데 주로 쓰이는 방법으로, 잡음이 포함되지 않은 원 음성신호와 잡음이 포함된 음성신호의 잡음이 개선된 음성신호들의 각 프레임 간 가중스펙트럼 기울기(weighted spectral distance)를 계산하여 음성의 개선정도를 평가하는 방법이다. 이 방법을 이용하면 유사성분의 많이 포함되어 있는 각 잡음군을 분류하는 데 유용할 것으로 생각된다.

따라서, 본 논문에서는 음성인식 시 혼입이 예상

되는 잡음들을 몇 가지 군으로 분류한 다음, 입력음성에 포함된 잡음과 비교하여 주파수의 변이도를 이용하여 잡음음성인식을 위한 환경 정보를 얻는다. 이를 음성 인식을 위한 모델 개선에 적용하여 성능 향상을 도모하는 환경정보 지식에 기반한 주파수 변이 적응 PMC (FV-PMC) 방법을 제안한다. 이 방법은 미리 분류된 각 잡음 군 간의 평균 주파수 변이도를 미리 계산하여 임계치로 설정한 다음, 미지의 잡음이 포함된 음성이 입력되면 각 잡음 군과의 주파수 변이도를 다시 계산하여 해당 잡음군의 임계치 보다 높을 경우 그 잡음 군의 잡음이 포함된 음성으로 간주하여 이 잡음 군이 포함된 음성을 이용하여 생성된 인식모델을 이용하여 음성인식을 수행하는 방법이다.

본 논문의 구성은 다음과 같다. II장에서는 모델에 기반을 둔 보상방법에 대해서 설명한다. III 장에서는 본 논문에서 제안하는 환경인식 기반의 FV-PMC에 관해서 기술하고 IV에서는 본 논문에서 제안한 FV-PMC의 음성인식을 수행하고, 그 결과를 고찰한 후 V 장에서 본 논문의 결론을 맺는다.

II. 모델에 기반을 둔 보상방법

모델에 기반을 둔 보상방법은 훈련환경과 인식환경 사이의 차이를 통계적인 모델로 특정화하는 방법으로 대표적으로 다음 두 가지 방법이 있다. 즉, 잡음에 오염된 음성을 특정화하기 위해 음성으로부터 얻은 부가잡음에 대한 지식을 이용해서 순수음성으로 훈련된 음소모델의 평균이나 분산을 변환하는 방법으로 HMM 분해법(Hidden Markov Model decomposition)과 부가잡음 뿐만 아니라 선형필터링의 영향도 함께 제거하기 위해서 위의 방법을 확장한 PMC 방법 등이 있다.[15]

2.1 HMM 분해법(Hidden Markov Model decomposition)

HMM 분해법은[15,16] HMM의 구성이 Fig. 1과 같다는 가정 하에서 수행되며 음성과 잡음의 HMM 분해방법 (Speech and Noise Decomposition; SND)은 단순히 잡음의 평균 정보만을 이용하는 것과 달리 잡음의 가우스 분포를 모델 보상에 이용한다.

이 때 특징 벡터로는 로그 필터 뱅크 에너지를 이용한다. 음성 모델과 잡음 모델을 일반화된 Viterbi 디코딩에 적용하면 음성 모델 M과 잡음 모델 M 각각에서의 최적 상태 순서를 얻을 수 있게 된다. 두 개의 모델을 동시에 적용하는 3차원 Viterbi 디코딩은 식 (1)과 같이 표현할 수 있다.

$$P_\tau(j,k\mid M,\tilde{M}) = \max_{u,v} P_{\tau-1}(u,v\mid M,\tilde{M})\cdot a_{u,j}\cdot \tilde{a}_{u,j}\cdot[b_j\otimes\tilde{b}_k](O^l(\tau)) \quad (1)$$

식(1) 에서 윗 첨자 l은 로그 스펙트럼 영역을 의미하며 $P_\tau(j,k)$는 시간 τ 일때 관측 벡터열 $\{(O^l(1),O^l(1),\cdots,O^l(\tau))\}$ 까지 인식한 후 모델 M의 j번째 상태, 모델 $\tilde{M}$의 k번째 상태에 있을 최대 확률을 의미한다. $a_{u,j}$ 와 $\tilde{a}_{v,k}$ 는 각각 모델 M의 u 번째 상태에서 j 번째 상태로 천이할 확률, 모델 $\tilde{M}$의 v 번째 상태에서 k번째 상태로 천이할 확률을 말한다. 또, $[b_j\otimes\tilde{b}_k](O^l(\tau))$는 보상된 모델에서 벡터 $O^l(\tau)$ 의 관측확률을 뜻한다.

SND에서는 잡음이 포함된 관측 벡터의 확률를 계산하기 위해 음성과 잡음 모델 사이에 식(2)와 같은 max 근사를 이용한다. 식(2)에서 i는 벡터의 i 번째 요소이다.

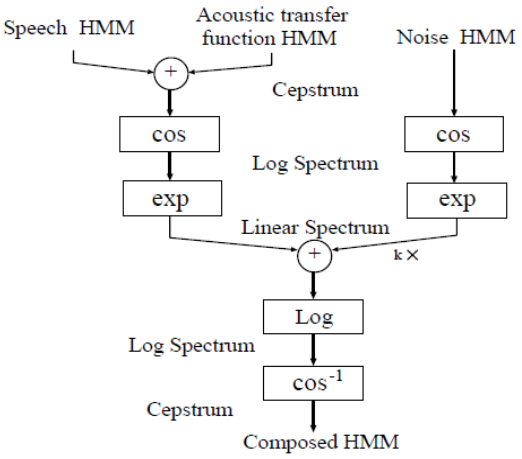

Fig. 1. Block diagram of HMM decomposition.[16]

$$O_i^l(\tau) = \log(S_i(\tau) + N_i(\tau)) \approx \max(S_i(\tau), N_i(\tau)). \quad (2)$$

식(2)의 가정 및 음성과 잡음의 확률 분포는 정규 분포를 이룬다는 가정 하에서 Fig. 1에서 같이 두 종류의 로그 필터 뱅크 에너지를 이용해 훈련된 음성 모델과 잡음 모델의 조합은 식(3)과 같이 근사된다.[17]

$$\begin{aligned}[b_j \otimes \tilde{b}_k](O_i^l(\tau)) &= \int \mathcal{L}(S_i^l(\tau), N_i^l(\tau) | j, k, M, \tilde{M}) \\ &= \int \mathcal{L}(\max(S_i^l(\tau), N_i^l(\tau)) | j, k, M, \tilde{M}) \\ &\approx N(O_i^l(\tau); \mu_i^l, \sum_{ii}^l) C(O_i^l(\tau); \tilde{\mu}_i^l, \tilde{\sum}_{ii}^l) \\ &\quad + C(O_i^l(\tau); \mu_i^l, \sum_{ii}^l) N(O_i^l(\tau); \tilde{\mu}_i^l, \tilde{\sum}_{ii}^l).\end{aligned} \quad (3)$$

식(3)에서 $C()$는 누적 정규 분포를 의미하고 , $N()$은 정규 분포를 나타낸다. 식(3)에서는 주어진 음성모델과 잡음모델에서 관측벡터의 확률값이 가중치의 합으로 표현된다고 볼 수 있다. SND는 디코딩 과정에서 음성 모델을 수정하지 않기 때문에 이 방법을 적용하는데 필요한 시간은 모델의 수보다는 관측 벡터의 수에 의존하므로 대용량의 모델 집합을 가지는 음성 인식 시스템에서 효과적으로 적용할 수 있다. 그러나 이 방법은 로그 스펙트럼 영역에서만 유효하기 때문에 이용할 수 있는 모델에 한계가 있으며 동적 파라미터 보상을 다루기가 쉽지 않다는 단점이 있다.

2.2 PMC

HMM을 기반으로 하는 음성인식시스템의 성능은 훈련환경과 인식환경 사이의 부정합(mismatching)이 증가함에 따라 급속히 저하된다. 따라서 이러한 부정합을 보상(compensation)하기 위한 모델 파라미터에 대한 수정이 요구되는데 그 중의 한 가지 방법이 PMC이다. PMC는 훈련환경과 인식환경 사이에 부정합이 나타나지 않을 때 음성인식시스템이 최적의 성능을 보인다는 점에 중점을 두고 간섭부가 잡음이 있는 경우를 고려한다. 이 경우, 부가적인 잡음이 부정합성에 나타나는 영향을 알 수 있다면 새로운 테스트 환경에 정합시키기 위해서 훈련 데이터를

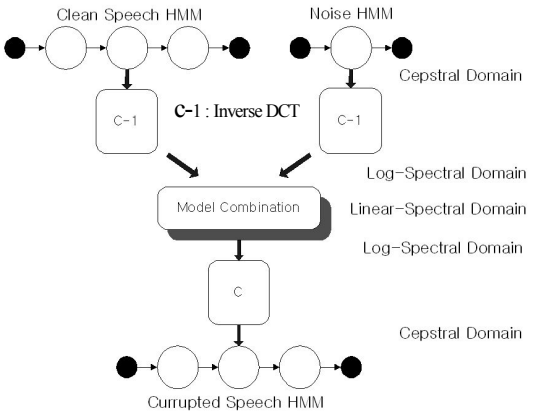

Fig. 2. Basic parallel model combination.[10]

수정하거나 재훈련 시킬 수 있을 것이다.[9]

잡음이 혼합된 음성을 가장 잘 인식할 수 있는 방법은 실 환경과 동일한 잡음 환경에서 자료를 수집하고 인식기를 재학습시키는 것이다. 그러나 이러한 방법은 실용적이지 못하다. 만약 음성 모델이 학습 자료의 통계적 특성을 잘 가지고 있다면 Fig. 2와 같은 모델 파라미터 보상으로 동일한 효과를 얻을 수 있다. 음성과 잡음 신호는 선형 스펙트럼 영역에서 가산적으로 이루어지기 때문에 각 모델의 파라미터를 선형 스펙트럼 영역으로 변환하여 조합한다. 조합할 때 기준이 되는 식을 불일치 함수라 하며, 이는 다음의 가정에 기초한다.[18,19]

1) 음성과 배경 잡음은 상호 독립적이다.
2) 음성과 배경 잡음은 시간 영역에서 가산적이다.
3) 단독 다변량 가우스 모델(single multivariate Gaussian model)로 음성과 배경 잡음 정보를 충분히 알 수 있다.
4) 잡음 첨가 후에도 프레임 및 HMM 모델의 상태 배열은 유지된다.

III. 주파수 변이를 이용한 PMC 모델 적응화에 기반한 잡음에 강인한 음성인식

3.1 잡음군 분류를 위한 거리척도

유사성을 가진 여러 집단의 분류를 위해서는 여러 가지 거리척도가 이용될 수 있으나 본 논문에서는

일반적으로 많이 이용되고 있는 Weighted Spectral Slope와 Cepstral Distance를 이용하기로 한다. 이하 이에 대해 간략한다.

WSS(Weighted Spectral Slope)[14]

이 방법은 필터링, 레벨 변경 포만트 주파수 등 몇 가지 스펙트럼 조작을 받은 음성의 모음사이의 거리를 측정 하려는 요구에서 시작된 것으로 주파수 영역을 인간의 청각 구조에 기초한 임계대역으로 나누고 각 대역에서의 스펙트럼 기울기들 간의 차에 가중치를 준 값을 구하는 방법이다. 이 측정방식은 스펙트럼 간의 기울기, 전반적인 레벨 등의 다른 차이는 무시하고 스펙트럼 피크 위치에 차등을 주어 설계되었으며 이러한 차이는 음성평가에 있어서 두 모음 사이의 거리측정에 효과가 있다고 알려져 있다. 측정 방식은 아래의 과정과 같다.

먼저 각 대역의 스펙트럼 기울기를 찾아 계산한다.

$$\begin{aligned} S_x(k) &= C_x(k+1) - C_x(k) \\ \overline{S_{\hat{x}}}(k) &= \overline{C_{\hat{x}}}(k+1) - \overline{C_{\hat{x}}}(k). \end{aligned} \tag{4}$$

식(4)에서 $C_x(k)$는 원 음성, $C_{\hat{x}}(k)$ 는 개선된 음성의 임계대역 스펙트럼을 데시벨로 표시한다. $S_x(k)$와 $\overline{S_x}(k)$는 원음성성과 개선된 신호의 k번째 대역의 스펙트럼 기울기를 나타낸다. 가중치는 식(4)의 스펙트럼 기울기를 사용하여 계산한다.

두 번째로 스펙트럼 기울기에 가중치를 적용한다. 이때, 가중치는 각 대역의 스펙트럼 피크인지 계곡인지 여부와 그다음 스펙트럼의 가장 큰 피크인지 여부에 따라 차별화 시킨다. 가중치 $W(k)$는 아래와 같이 계산한다.

$$W(k) = \frac{K_{\max}}{[K_{\max} + C_{\max} - C_x(k)} \frac{K_{loc\max}}{[K_{loc\max} + C_{loc\max} - C_x(k)}. \tag{5}$$

식(5) 에서 $C_{\max}$는 전체 대역에서 가장 큰 로그 스펙트럼 크기, $C_{loc\max}$는 k 대역에서 가장 가까운 피크의 값, $K_{\max}$, $K_{loc\max}$는 상수로서 전체적인 성능을 조절하기 위해 변화할 수 있는 파라미터이다.

마지막으로 WSS 는 다음과 같이 음성의 각 프레임에서 식(6)과 같이 계산된다.

$$d_{WSM}(C_x, \overline{C_x}) = \sum_{k=1}^{L} W(k)(S_x(k) - \overline{S_{\hat{x}}}(k))^2. \tag{6}$$

여기서 L은 사용한 임계대역의 수이다.

평균 WSS는 음성의 모든 프레임에서 얻은 WSS 값을 평균하여 얻어진다. 이 방법은 포만트 추출을 필요로 하지 않고 인간의 청각구조에 기초하고 있으므로 다른 척도들에 비해 청자가 느끼는 명료도의 측면을 더욱 잘 반영할 수 있는 유용한 측정방법이다.

Cepstral Distance

Cepstral Distance는 음성신호로부터 추출한 LPC 켑스트럼 계수[20]를 이용하여 구한 인접 프레임들 간의 스펙트럼 거리를 특정 임계치와 비교함으로써 이들 구간들을 구분한다. 이때 n 번째 프레임에서 음성 시작구간의 평균 켑스트럼과의 유클리드 거리 $d[n]$은 식(7)과 같이 구한다. 여기서 p는 켑스트럼 특징 벡터의 차수이다.

$$d[n] = \sum_{i=1}^{p} (c_i[n] - \overline{C_i})^2. \tag{7}$$

두 신호 $c_x, \overline{c_{\hat{x}}}$ 의 켑스트럼 거리 차이는 식(8)과 같이 계산한다.[14]

$$d_{cep}(c_x, \overline{c_{\hat{x}}}) = \frac{10}{\log_e 10} \sqrt{2 \sum_{k=1}^{p} [c_x(k) - c_{\hat{x}}(k)]^2}. \tag{8}$$

3.2 제안된 방법

여기서는 위에서 기술한 여러 방법들을 이용하여 본 논문에서 제안하는 주파수 변이를 이용한 PMC 모델 적응화에 기반한 잡음에 강인한 음성인식시스템에 대해서 단계별로 설명한다. 즉, 음성인식 시 혼입이 예상되는 여러 잡음들을 유사 잡음 별 군으로 분류하여 입력음성에 포함된 잡음과 비교한 다음 발성환경에 대한 정보를 얻어 이를 인식을 위한 모델

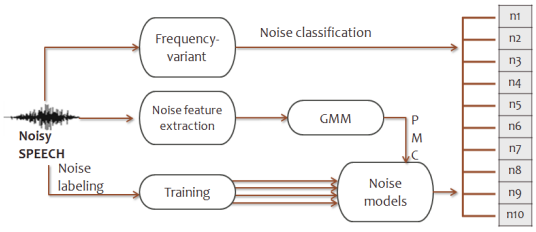

Fig. 3. Block Diagram of FV-PMC

Fig. 3. Block Diagram of FV-PMC.

보상에 적용하여 인식률 향상을 도모하는 방법이다. 제안하는 환경인식 기반 음성인식 시스템의 전체 구성을 Fig. 3에 나타내었다. 이하 이에 대해 각 과정별로 간략한다.

과정 1: GMM을 이용한 개선된 잡음모델 생성

인식시 혼입이 예상되는 여러 종류의 잡음을 GMM을 이용하여 평균과 분산 파라미터를 추출한 후 PMC 알고리즘에 부가한 후 개선된 잡음 모델을 생성한다.

GMM의 추출은 EM (Expectation-Maximization)알고리즘에 의하여, Gaussian 분포를 갖는 각 성분의 평균, 분산 그리고 혼합가중치를 추정할 수 있다. 입력 데이터 집합 $x = x_1, x_2, \cdots, x_N$ 에 대하여 식(9)~(11)은 각각 혼합가중치의 추정치, 평균, 분산 그리고, $\alpha_j, \hat{\mu_j}, \hat{\sigma_j^2}$ 를 나타낸다.

$$\hat{\alpha_j} = \hat{P}(w_j) = \frac{1}{N}\sum_{n=1}^{N} P(w_j|x_n,\theta), \tag{9}$$

$$\hat{\mu_j} = \frac{\sum_{n=1}^{N} P(w_j|x_n,\theta)x_n}{\sum_{n=1}^{N} P(w_j|x_n,\theta)}, \tag{10}$$

$$\hat{\sigma_j^2} = \frac{\sum_{n=1}^{N} P(w_j|x_n,\theta)|x_n - \hat{\mu_j}|^2}{\sum_{n=1}^{N} P(w_j|x_n,\theta)}. \tag{11}$$

과정 2 : 주파수 변이도를 이용한 잡음의 분류

각 잡음 별 주파수 변이 (frequency-variant)의 평균 변이도는 각 주파수 대역별로 가중치를 주어 스펙트럼 기울기(spectral distance)를 측정한 후 주파수의 변이도를 계산한다. 주파수 변이도 $fw\,Var$ 는 식(12)와 같이,[14] 평균 변이도 T는 식(13)과 같이 나타난다.

$$fw\,Var = a_0 + \sum_{j=1}^{k} a_j D_j \tag{12}$$

$$T = \frac{1}{M}\sum_{n=1}^{M} fw\,Var \tag{13}$$

여기서 a_j: regression coefficients (비선형회귀계수)

k: number of bands (대역의 개수) 이다.

식(12)에서 입력 음성의 각 프레임에서 가중 스펙트럼 기울기 D_j는 각 프레임의 왜곡도이며 식(14)와 같다.

$$D_j = \frac{1}{M}\sum_{m=1}^{M} 10\log_{10}[F^2(m,j)/(F(m,j)-\hat{F}(m,j))^2 \tag{14}$$

여기서 $j = 1,2,...,k$이다.

Table 1 에 식(12)에 이용된 대역별 가중치 값을 나

Table 1. Center friquencies (Hz) and weights of Critical Bands.[14]

Band Number	Center Frequency	Weight	Band Number	Center Frequency	Weight
1	50	0.003	14	1148	0.032
2	120	0.003	15	1288	0.034
3	190	0.003	16	1442	0.035
4	260	0.007	17	1610	0.037
5	330	0.010	18	1794	0.036
6	400	0.016	19	1993	0.036
7	470	0.016	20	2221	0.033
8	540	0.017	21	2446	0.030
9	617	0.017	22	2701	0.029
10	703	0.022	23	2978	0.027
11	798	0.027	24	3276	0.026
12	904	0.028	25	3597	0.026
13	1020	0.030			

타낸다. 대역별 가중치는 일정하게 같은 폭을 지니는 대역필터가 아니라 다른 길이를 가지는 임계 대역의 폭에 맞추어 필터를 설계를 하기 위해 만들어졌다. 또한 이 가중치는 본 논문에서는 주파수 변이도의 값을 회귀분석을 이용하여 계산할 때 사용된다. 가중치의 계산은 식(15)와 같이 계산된다.

$$w = \log(bw_{\min}) - \log(bandwidth(i)) \quad (15)$$

여기서 bandwidth 는 대역폭을 의미하며, $bw_{\min}$ 는 최저 대역폭을 말한다.

본 논문에서는 먼저 잡음의 분류를 위해서 사전실험으로 각 잡음군 간의 평균 변이도를 미리 계산하여 정한 임계치를 설정하였다. 이 후 입력음성과 10개의 잡음환경의 음성과 주파수 변이도를 각각 계산하여 특정 임계치 보다 높은 경우 비교한 잡음군의 잡음이 포함된 음성으로 처리한다. 예를 들면 잡음이 확인되지 않은 음성이 입력되어 10개의 잡음음성과의 주파수 변이도를 각각 계산한 후 subway 잡음음성과의 주파수 변이도가 4.61을 넘었다면 subway 잡음으로 분류를 한다.

과정 3: 음성인식 수행

과정 2 에서 각 잡음군 별로 분류된 잡음이 포함된 입력음성들은 과정 1에서 얻어진 각 잡음모델을 이용하여 음성인식을 수행한다. 임계치 이하의 주파수 변이도를 나타내어 미분류된 잡음이 포함된 입력음성들은 기존 PMC 방법을 이용하여 깨끗한 모델과 결합하여 음성인식을 수행한다.

IV. 실험 결과 및 고찰

4.1 실험환경

실험 및 성능평가를 위하여 본 논문에서는 Aurora 2.0 데이터베이스를 이용한다.[21] Aurora 2.0에는 2 종류의 훈련환경 즉, 8440개의 조용한 환경 하여서 발성된 음성 발성으로 구성된 clean-condition과 동일한 발성을 20개의 잡음환경으로 나누어 각 422개의 발성으로 구성된 multi-condition으로 구분되어 있다. 잡음환경은 총 10 종류의 잡음으로 분류되어 있으며, 3개의 Set 즉, Set A(subway, babble, car, exhibition)와 Set B(restaurant, street, airport, station) 그리고 Set A와 Set B에 나타난 2가지 잡음 (subway, street)에 훈련환경과 다른 채널특성을 포함한 Set C로 구성되어 있으며 잡음 레벨을 7가지(Clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB)로 구분되어 있다. 성능 평가에서는 Set A, B, C의 각 잡음의 종류에 대해서 20 dB에서 0 dB까지의 5가지 레벨의 평균 단어 인식률(word accuracy)을 비교한다.

4.2 잡음 군의 분류

임계치를 결정하기 위한 사전실험 결과를 Table 2

Table 2. Thresholds of frequency-variant for noise group classification.

Noise	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train-station	Subway (MIRS)	Street (MIRS)
Subway	**4.61**	2.36	2.74	2.94	3.70	2.77	1.96	2.22	3.30	2.80
Babble	2.34	**4.41**	3.09	2.39	2.76	3.89	2.58	2.75	1.52	2.38
Car	2.62	2.93	**4.58**	2.74	2.76	2.95	3.89	2.89	1.92	2.39
Exhibition	2.90	2.31	2.72	**4.47**	2.37	2.72	1.92	3.42	2.40	2.78
Restaurant	3.68	2.84	2.99	2.51	**4.34**	2.76	2.47	2.68	2.07	2.22
Street	2.60	3.78	2.90	2.62	2.54	**4.21**	2.17	2.46	1.94	3.08
Airport	2.26	2.89	4.30	2.32	2.68	2.63	**4.46**	2.94	1.51	1.95
Train-station	2.40	2.91	3.17	3.63	2.74	2.80	2.80	**4.52**	1.66	2.21
Subway (MIRS)	3.56	1.54	2.00	2.61	2.14	2.19	1.31	1.53	**4.64**	3.01
Street (MIRS)	2.62	2.20	2.21	2.63	1.96	3.12	1.47	1.77	2.61	**4.31**

* 굵은 테두리내의 값은 각 잡음군의 임계치를 나타냄.

에 나타낸다. 이 실험 결과에서 얻은 최적 임계치를 입력음성의 인식을 위한 모델 선택의 기준 값으로 한다.

Table 2에 나타낸 각 잡음군간의 평균 주파수 변이도를 주파수 변이도를 이용한 경우 동일한 잡음환경일 경우 주파수 변이도가 4.21~4.64의 값을 보이고 있으나 다른 종류의 잡음일 경우 이보다 낮은 1~3 사이의 값을 나타내고 있음을 볼 수 있어 타 잡음에 대한 변별력이 있음을 알 수 있다.

Table 3에 본 논문에서 도입한 주파수 변이도를 이용한 잡음분류의 성능을 평가하기 위해 WSS, CEP와의 분류정확도를 비교하기 위하여 실시한 실험결과를 나타내었다.

Table 3 으로부터 알 수 있는 바와 같이 WSS, CEP의 경우는 각각 전체 평균 34.05%, 31.65%, 본 논문에서 도입한 평균 주파수 변이도를 이용한 경우에는 56%로 나타나 주파수 변이 방법은 다른 두 방법에 비해 현저히 높은 분류 정확도를 보임을 알 수 있다.

따라서 본 논문에서는 주파수 변이 방법을 이용하여 잡음을 분류한 후 이를 음성인식에 적용하기로 한다.

4.3 실험결과

Table 4 에 평균변이를 이용한 FV-PMC의 평균 단어인식률을 나타내었다. Table 4로부터 평균 단어인식률은 각각 Set A에서 79.05%, Set B에서 79.43%, Set C에서 83.37%로 나타났다. 전체 평균인식률 80.62%로 기존의 PMC 인식률 74.93% 보다 5.69% 향상된 결과를 보여 제안한 방법의 인식에 대한 유효성을 확인 할 수 있다. 그러나 Fig. 4 에 나타낸 각 잡음별 FV-PMC의 인식률을 살펴보면, babble, exhibition 잡음과 같은 특정 잡음 에서는71.02%, 73.07%로 오히려 평균 인식률이 기존 PMC방법 보다 다소 떨어지는 결과를 보였는데 이는 잡음분류 정확도가 54%, 52%로 다른 잡음에 비해 낮음으로 인해 인식률이 떨어지는 결과를 보이는 것으로 분석된다.

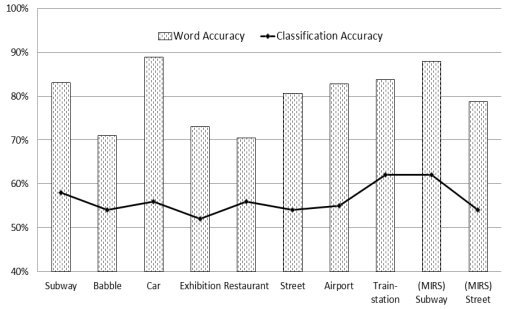

Fig. 4. Word recognition rates according to the noise classification accuracy.

Table 3. Noise classification accuracy of each distance measure (%).

Measure	set A					set B					set C			Avg
	Subway	Babble	Car	Exhibi-tion	Avg	Restau-rant	Street	Airport	Train-station	Avg	Subway (MIRS)	Street (MIRS)	Avg	
wss	32.28	32.87	31.88	33.77	32.70	34.62	35.31	37.56	36.61	36.03	31.05	34.93	32.99	34.05
cep	27.06	28.52	33.43	28.19	29.30	29.57	37.01	32.67	37.10	34.09	25.64	37.33	31.49	31.65
FV	**58.06**	**53.83**	**56.02**	**52.09**	**55.00**	**55.72**	**54.03**	**55.06**	**61.52**	**56.58**	**61.80**	**53.97**	**57.88**	**56.27**

Table 4. Comparison of the word accuracy (%).

set	set A			set B			set C		
method	Baseline	PMC	**FV_PMC**	Baseline	PMC	**FV_PMC**	Baseline	PMC	**FV_PMC**
20dB	95.25	95.23	**95.23**	92.77	94.97	**93.42**	94.30	95.24	**95.67**
15dB	87.33	93.21	**93.26**	81.34	92.57	**89.87**	87.84	92.01	**94.17**
10dB	67.71	86.44	**92.35**	59.01	85.84	**90.48**	74.15	82.39	**91.90**
5dB	39.48	68.83	**70.83**	31.93	67.71	**72.86**	50.24	61.78	**78.91**
0dB	16.95	37.11	**43.59**	13.70	38.5	**50.54**	24.17	32.14	**56.20**
Avg. (20, -0dB)	61.34	76.16	**79.05**	55.75	75.92	**79.43**	66.14	72.71	**83.37**

V. 결 론

본 논문에서는 다양한 잡음환경 하에서 강인한 음성인식 시스템을 구현하기 위하여 FV-PMC 방법을 제안하였다. 이 방법은 혼입이 예상되는 잡음들을 주파수 변이의 평균값을 이용하여 임계치를 정한 후 이를 이용하여 잡음을 수종의 잡음 군으로 분류한 후 잡음 군 별 잡음음성 인식모델을 작성하여 음성인식을 수행하는 방법이다.

실험결과 잡음 군별 분류 정확도는 평균 56%를 보였으며 잡음 군별로 분류된 잡음음성 인식모델을 이용하여 음성인식을 수행한 결과 set A에 대해서는 79.05%, set B에 대해서는 79.43%, set C에 대해서는 83.37%로 나타났다. 그 결과 전체 평균인식률은 80.62%로 기존의 PMC 방법의 74.93%보다 5.69% 향상된 결과를 얻어 제안한 방법의 유효성을 확인 할 수 있었다. 그러나 특정 잡음 예를 들면 babble, exhibition 등에서는 오히려 평균 인식률이 기존 PMC 방법 보다 다소 떨어지는 결과를 보였는데 이는 잡음분류 시스템의 정확도가 54%, 52%로 다른 잡음에 비해 낮음으로 인해 인식률이 떨어지는 결과를 보이는 것으로 분석된다. 향후, 잡음분류의 정확도를 좀 더 향상시킬 수 있는 새로운 방법에 연구가 진행될 예정이다.

References

1. Yao, E. Visser, O. W. Kwon and T. W. Lee, "A seech processing front-end with eigenspace normalization for robust speech recognition in noisy automobile environments," Proc. Eurospeech, 9-12 (2003).
2. Seon-Mi Gang, "Study on speech recognition under noisy environments" (in Korean), J. Inst. Ind. Tech. **3**, 301-318 (1997).
3. J. S. Lim, A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proceedings IEEE, **67,** 1586-1604 (1979).
4. Y. Ephraim and D. Malah, and B. H. Juang, "On the application of hidden markov models for enhancing noisy speech," Proc. ICASSP, 533-536 (1992).
5. J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, (Kluwer Academic Publishers, 1996).
6. Y. H. Suk, S. H. Choi, and H. S. Lee, "Cepstrum PDF normalization method for Speech recognition in noise environment"(in Korean), J. Acoust. Soc. Kr. 4(s) **24**, 224-229 (2005).
7. Hanson, B. A., and Wakita, H., "Spectral slope distance measure with linear prediction analysis for word recognition in noise," IEEE Trans. on ASSP, ASSP-35, **7**, 968-973 (1987).
8. Juang, B. H., Rabiner, L., and Wilpon, J., "On the use og bandpass liftering in speech recognition," ICASSP, 765-768 (1986).
9. A. Nadas, D. Nahamoo and M. Picheny, "Speech recognition using noise adaptive prototypes," Proc. ICASSP, 517-520 (1988).
10. Gue-Jun Jung, Hoon-Young Cho, and Yung-Hwan Oh, "Improved compensation of dynamic parameter in PMC for robust speech recognition"(in Korean), J. Acoust. Soc. Kr. 1(s) **20**, 183-186 (2001).
11. K. C. SIM, M.T. LUONG, "A trajectory-based parallel model combination with a unified static and dynamic parameter compensation for noisy speech recognition," ASRU, 107-112 (2011).
12. G.H. Shen, H.Y. Jung, and H. Y. Chung, "A noise robust speech recognition method using model compensation based on speech enhancement"(in Korean), J. Acoust. Soc. Kr. 4(s) **27**, 191-199 (2008).
13. Hadi Veisi, Hossein Sameti, "Cepstral-domain hmm - based speech enhancement using vector taylor series and parallel model combination," ISSPA, 298-303(2012).
14. Philipos C .Loizou, *Speech Enhancement -Theory and Practice,* (CRC Press, Florida, 2007).
15. Varga A. and Moore R.,"Hidden markov model decomposition of speech and noise," ICASSP, 845-848 (1990).
16. Nakamura, S. Qiang Hou, Shikano, K., "Model adaptation based on hmm decomposition for reverberant speech recognition," ICASSP, 21-24 (1997).
17. G. J. Jung, "Improved on-line model compensation for robust speech recognition"(in Korean), Master's thesis (2002).
18. Gales,M. and Young S.,"HMM recognition in noise using parallel model combination," EUROSPEECH, 837-840 (1993).
19. M. J. F. Gales, S. Young, "Robust continuous speech recognition using parallel model combination," IEEE TSAP, **4**, 352-359 (1996).
20. Rabiner, lr, and Juang, bh, *Fundamentals of Speech Recognition,*(Prentice-Hall, New Jersey,1993).
21. H.-G Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," ISCA ITRW ASR (2000).

저자 약력

▸ 최 숙 남 (Sook-Nam Choi)

1995년 8월: 영남대학교 전자공학과 공학사
2007년 8월: 영남대학교 전기전자통신교육 교육학석사
2011년 8월: 영남대학교 정보통신공학과 박사수료
2008년 8월~ 현재 영남대학교 대학원 정보통신공학과 박사과정
〈관심분야〉 음성인식, 화자인식, 디지털 신호처리

▸ 정 현 열 (Hyun-Yeol Chung)

1989년 일본 동북대학교 정보공학과 공학박사
1989년 3월~현재 영남대학교 전자정보공학부 교수
1992년 7월~1993년 7월 미국 CMU Robotics 연구소 객원연구원
1994년 12월~1995년 2월 일본 토요하시기술과학대학 외국인 연구자
2000년 6월~2000년 8월 미국 Qualcomm Inc. 수석 엔지니어
〈관심분야〉 음성인식, 화자인식, 음성합성 및 DSP 응용 분야