최대우도를 부가한 주파수 변이 PMC 방법의 잡음 음성 인식 성능개선

최숙남[†], 정현열^{††}

요 약

잡음에 강건한 음성 인식을 위한 주파수 변이를 이용한 PMC(Parallel Model Compensation Using Frequency-variant, FV-PMC) 방법은 인식시 입력음성에 혼입이 예상되는 잡음들을 평균 주파수 변이도를 임계치로 하여 몇 가지 잡음 군으로 분류한 후 각 잡음 군 별로 인식을 수행하는 방법이다. 이 방법은 기준 임계치를 이용하여 양호하게 분류된 잡음 음성들에 대해서는 매우 우수한 성능을 보이나, 미 분류된 잡음 음성들에 대해서는 기존의 PMC 방법에서와 같이 무잡음 모델과 결합하여 음성 인식을 수행함으로써 인해 평균 음성 인식률이 낮아지는 문제점이 있다. 이러한 문제점을 해결하기 위하여 본 논문에서는 기존의 방법에서 사용하였던 평균주파수 임계치 방법 대신에 최대 우도를 부가하여 미분류를 방지함으로써 입력 잡음 음성에 포함되는 잡음의 군별 잡음 분류 율을 높여 인식률을 제고하는 개선된 주파수 변이 PMC 인식방법을 제안하였다. Aurora 2.0 데이터베이스를 이용한 인식실험결과, 기존의 FV-PMC 방법에 비해 향상된 결과를 확인할 수 있었다.

Recognition Performance Improvement for Noisy-speech by Parallel Model Compensation Adaptation Using Frequency-variant added with ML

Sook-Nam Choi[†], Hyun-Yeol Chung^{††}

ABSTRACT

The Parallel Model Compensation Using Frequency-variant: FV-PMC for noise-robust speech recognition is a method to classify the noises, which are expected to be intermixed with input speech when recognized, into several groups of noises by setting average frequency variant as a threshold value; and to recognize the noises depending on the classified groups. This demonstrates the excellent performance considering noisy speech categorized as good using the standard threshold value. However, it also holds a problem to decrease the average speech recognition rate with regard to unclassified noisy speech, for it conducts the process of speech recognition, combined with noiseless model as in the existing PMC. To solve this problem, this paper suggests a enhanced method of recognition to prevent the unclassified through improving the extent of rating scales with use of maximum likelihood so that the noise groups, including input noisy speech, can be classified into more specific groups, which leads to improvement of the recognition rate. The findings from recognition experiments using Aurora 2.0 database showed the improved results compared with those from the method of the previous FV-PMC.

Key words: PMC, MLE(최대우도), Frequency-variant(주파수 변이), Speech Recognition(음성인식)

※ 교신저자(Corresponding Author): 정현열, 주소: 경북
경산시 대동 영남대학교 IT관 306호(712-749), 전화: 053)
810-2496, FAX: 053) 810-4742, E-mail: hychung@yu.ac.kr
접수일: 2013 년 4월 16일, 수정일: 2013 년 6월 20일

완료일: 2013 년 7월 17일

[†] 정회원, 영남대학교 정보통신공학과
(E-mail: windy@ynu.ac.kr)

^{††} 정회원, 영남대학교 정보통신공학과

1. 서론

사람과 컴퓨터와의 대화를 통한 기계의 제어를 위해 음성 인식의 중요성을 이해하고 음성 인식 시스템을 구현하고자 하는 시도가 1950년대에 처음 시작된 이래 꾸준한 진전이 이루어져 오고 있으며 최근에는 음성 인식 기반기술에 대한 수요의 증가와 컴퓨터 하드웨어 기술의 발달로 인해 그 응용폭이 점차 확대되고 있다[1]. 현재까지 개발된 음성 인식 시스템은 잡음이 없거나 비교적 조용한 실험실 환경에서는 좋은 성능을 나타낸다. 그러나 실제 여러 종류의 잡음이 존재하는 환경에서는 인식 성능이 현저히 저하된다. 이것은 잡음으로 인해 인식시스템의 훈련 조건과 인식 조건 사이의 불일치가 발생하기 때문이다[2-3].

이러한 환경 불일치를 극복하기 위해 진행된 기존 연구들은 크게 다음의 세 가지로 분류할 수 있다. 첫 번째는 MFCC (Mel Frequency Cepstral Coefficient), RASTA (RelAtive SpecTrAl), PLP (Perceptual Linear Prediction)와 같이 입력되는 음성으로부터 본질적으로 잡음에 강인한 특징을 추출하는 방법이다. 이 방법들은 실내와 같은 조용한 잡음환경에서 발생하는 백색잡음과 같은 잡음에 대해서는 한정한 효과를 나타내고 있으나 거리, 공연장 등과 같은 실환경에서는 한계가 있다[4]. 다음으로는 스펙트럼 차감법(spectral subtraction), cepstrum 평균 차감법(Cepstral Mean Subtraction; CMS)[3], 상태기반 음질 개선(state-based speech enhancement), 신호편향제거(Signal Bias Removal; SBR) 등과 같이 잡음이 포함된 음성 신호에서 잡음을 제거하여 깨끗한 음성 신호를 추출하는 전처리 방법들로서 환경 불일치 요인들이 스펙트럼 영역과 cepstrum 영역에서 음성에 각각 가산된다는 사실에 기초하여 잡음을 추정하고 추정된 잡음을 입력 신호에서 제거하고자 하는 방법으로 원 음성신호의 손상을 피할 수 없다[5]. 마지막으로 잡음 차폐 (noise masking), 음성과 잡음의 모델 분해(speech and noise decomposition; SND), 병렬 모델 조합(Parallel Model Combination; PMC), MLLR (Maximum Likelihood Linear Regression)과 같이 특정 환경에서 학습된 음성 인식 모델을 추정된 사용 환경에 적합하도록 수정하는 방법이 있다 [6-8].

이러한 방법들 중 인식 모델을 수정하는 방법은

음성에 관한 특별한 가설을 필요로 하지 않으며, 또한 입력 신호에 수정을 가하지 않으므로 정보의 손실이 없다는 장점을 가지고 있다[9].

따라서, 본 논문에서는 이들 방법 중 모델 보상 방법의 대표적인 방법 중 하나인 PMC에 관한 연구를 수행하였다. PMC는 cepstrum 영역에서 추출된 MFCC 특징 파라미터를 이용하여 학습된 음성과 잡음의 은닉 마르코프 모델(Hidden Markov Model; HMM) 파라미터들을 잡음이 가산되는 스펙트럼 영역으로 변환하고 두 모델을 조합한다. 조합된 모델은 다시 cepstrum 영역으로 변환하여 환경에 적응된 모델을 얻게 된다[8].

따라서 PMC 방법의 성능을 높일 수 있는 가장 좋은 방법은 실제 환경과 동일한 잡음 환경에서 발생된 음성자료를 수집하고 인식기를 학습시킨 후 동일한 환경 하에서 인식을 수행하는 것이나 잡음환경을 매번 학습하는 것은 실용적이지 못하다.

이와 같은 문제점을 보완하기 위하여 저자들은 다양한 인식환경에서 발생할 수 있는 잡음들을 몇 가지 잡음 군으로 분류하여 각 군별 잡음을 이용하여 인식 모델을 훈련한 후, 분류된 잡음 군에 속하는 잡음 환경 하에서 발생된 음성이 입력될 때 이 신호에 포함되는 잡음의 종류를 추정하고 추정된 잡음 군 환경에서 훈련된 인식 모델을 이용하여 인식하는 FV-PMC (PMC Adaptation using Frequency-variant)를 제안하였다[10].

이 방법은 먼저 주파수 변이도 개념을 도입하여 잡음음성과 각 잡음군과의 평균 변이도를 구하여 각 잡음군 분류를 위한 임계치를 정한다. 입력 잡음음성이 어느 잡음 군에 속하는가를 판정하기 위해서는 기 분류된 잡음군과의 주파수 변이도를 구하여 그 값이 임계치 이상일 경우 해당 잡음군에 속하는 잡음음성으로 분류한다. 인식을 위해서는 분류된 동일 잡음군 내에서 작성한 PMC를 이용하여 인식실험을 실시한다. 실험결과 기존의 PMC방법에 비해 우수한 성능을 얻을 수 있었다. 그러나 주파수 변이도 임계치를 이용해 잡음을 분류할 경우 잡음 분류율이 낮아, 미 분류되는 잡음 음성의 인식을 향상에 한계가 있었다. 이를 해결하기 위하여 본 논문에서는 FV-PMC 방법에서 문제가 되었던 잡음입력음성의 잡음 분류율을 향상시키기 위하여 새로운 분류척도를 도입을 검토하여 FV-PMC 방법에서 미분류되었던 입

력잡음 음성들의 잡음 분류율을 향상시킴으로써 인식성능을 향상시킬 수 있는 방법을 제안하고 그 유효성을 확인하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 잡음의 분류를 위한 척도로써 유용한 거리척도들에 대해 검토하고, 사전 잡음 분류 실험을 수행한다. 3장에서는 본 논문에서 제안한 개선된 주파수 변이를 이용한 인식방법에 대해서 기술한다. 4장에서는 제안한 음성 인식기를 이용하여 인식실험을 수행하고 그 결과를 고찰한 후 5장에서 본 논문의 결론을 맺는다.

2. 잡음 분류율 개선을 위한 거리척도 검토

FV-PMC 방법에서 주파수 변이도 임계치를 이용해 잡음을 분류할 때 잡음 분류율이 낮아 미분류 되었던 잡음들로 인해 야기된 인식률 저하를 개선하기 위해서는 잡음 분류율을 향상시킬 수 있는 거리 척도가 중요하다. 기존 평균 임계치 방법을 이용하여 주파수 분류를 하였을 경우 잡음의 분류 현황을 표 1에 나타내었다. 이하 잡음군 분류에 적용 가능한 거리척도에 대해 검토하기로 한다.

2.1 마할라노비스 거리(Mahalanobis Distance) [11]

마할라노비스 거리는 임의의 입력음성신호와 깨끗한 음성 상호간의 유사도를 측정하여 입력되는 음성의 품질이 깨끗한 음성에 얼마나 가까운지를 측정하는 음성품질의 평가척도로 이용되는 유용한 척도 중의 하나로 입력음성 x 와 깨끗한 음성 $\{I\}$ 간의 마할라노비스 거리는 식(1)과 같이 구할 수 있다.

$$D_i^{Mahalanobis} = (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \quad (1)$$

여기서 μ_i 는 깨끗한 음성 $\{I\}$ 의 평균 벡터를 나타내고, x 는 잡음 분류를 수행할 입력 음성을 나타낸다. Σ_i^{-1} 는 깨끗한 음성 $\{I\}$ 의 공분산 행렬의 역행렬

그림 1. Subway 잡음 음성과 다른 잡음음성군별 간의 마할라노비스 거리

을 나타낸다.

본 논문에서는 잡음의 종류를 모르는 입력 음성 x 와 잡음의 종류를 알고 있는 잡음 음성 $\{I\}$ 간의 유사도를 계산하여, 잡음 분류의 척도로서 효용성이 있는지를 확인하기로 한다. 그림 1에 subway 잡음환경 하에서 발생된 입력 음성과 각 잡음군 간의 마할라노비스 거리를 입력음성의 SNR 별로 나타내었다. 그림에서 알 수 있는 바와 같이 마할라노비스 거리는 입력음성의 SNR 값에 따라 다른 값을 보여주고 있다. 즉 SNR이 낮을수록 마할라노비스 거리는 낮은 값을 나타내는데 이는 공분산의 역행렬을 이용하는 마할라노비스 거리의 특성에 따른 것으로 보인다. 즉, 잡음은 음성에 비해 상대적으로 낮은 분산 값을 가지는 특징이 있고 이 특징은 SNR이 낮을수록 더욱 선명하게 나타나므로, 공분산을 사용하는 마할라노비스 거리는 잡음의 종류의 분류보다는 입력되는 음성의 음성품질 측정에 더 적합한 것으로 보인다. 실험 결과 입력음성의 잡음 분류율이 10% 미만으로 10종의 잡음 종류를 고려하면 잡음 분류의 척도로서는 부적합한 것으로 판단된다.

2.2 Weighted Spectral Slope(WSS)[12]

이 방법은 필터링, 레벨 변경 포맷트 주파수 등

표 1. 평균 임계치 방법을 사용한 잡음 분류율

(단위: %)

Set	Set A				Set B				Set C		Avg
잡음	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train-station	Subway (MIRS)	Street (MIRS)	
정분류율	58.06	53.83	56.02	52.09	55.72	54.03	55.06	61.52	61.8	53.97	56.27
오분류율	24.41	31.30	31.14	12.20	2.26	24.76	35.98	3.91	2.68	9.12	14.86
미분류율	17.53	14.87	12.84	35.71	42.02	21.21	8.96	34.57	35.52	36.91	28.87

몇 가지 스펙트럼 조작을 받은 음성의 모습사이의 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 모습 사이의 거리측정에 효과가 있다고 알려져 있다. 측정 방식은 아래의 과정과 같다.

먼저 각 대역의 스펙트럼 기울기를 찾아 계산한다

$$\begin{aligned} S_x(k) &= C_x(k+1) - C_x(k) \\ \overline{S_x}(k) &= \overline{C_x}(k+1) - \overline{C_x}(k) \end{aligned} \quad (2)$$

식 (2)에서 $C_x(k)$ 는 원 음성, $\overline{C_x}(k)$ 는 개선된 음성의 임계대역 스펙트럼을 데시벨로 표시한다. $S_x(k)$ 와 $\overline{S_x}(k)$ 는 원음성성과 개선된 신호의 k 번째 대역의 스펙트럼 기울기를 나타낸다. 가중치는 식 (2)의 스펙트럼 기울기를 사용하여 계산한다.

다음에는 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.

$$W(k) = \frac{K_{\max}}{[K_{\max} + C_{\max} - C_x(k)]} \frac{K_{\text{locmax}}}{[K_{\text{locmax}} + C_{\text{locmax}} - C_x(k)]} \quad (3)$$

식(3)에서 $C_{\max}$ 는 전체 대역에서 가장 큰 로그 스펙트럼 크기, C_{locmax} 는 k 대역에서 가장 가까운 피크의 값, $K_{\max}$, K_{locmax} 는 상수로서 전체적인 성능을 조절하기 위해 변화할 수 있는 파라미터이다.

마지막으로 WSS는 다음과 같이 음성의 각 프레임에서 식(4)와 같이 계산된다.

$$d_{WSS}(C_x, \overline{C_x}) = \sum_{k=1}^L W(k)(S_x(k) - \overline{S_x}(k))^2 \quad (4)$$

여기서 L 은 사용한 임계대역의 수이다.

평균 WSS는 음성의 모든 프레임에서 얻은 WSS 값을 평균하여 얻는 방법으로 인간의 청각구조에 기초하고 있으므로 다른 척도들에 비해 청취자가 느끼는 명료도의 측면을 더욱 잘 반영할 수 있는 유용한 척도로 알려져 있다.

2.3 PESQ(Perceptual Evaluation of Speech Quality)

PESQ(Perceptual Evaluation of Speech Quality)는 원 신호와 이 신호가 통신 시스템을 통과하여 나온 결과인 감쇄된 신호를 비교하는 평가 방법이다. 원 음성과 감쇄 음성을 입력했을 때 양측의 신호 특성량의 차이에서 음성의 품질을 측정하는 PSQM [13]을 바탕으로 패킷 손실 등을 반영한 PAMS의 장점을 합하여 만든 객관적인 평가 방법이다. 음성 왜곡 및 잡음을 측정할 수 있다. PESQ 값은 0.5부터 4.5 사이이며 일반적으로 1.0에서 4.5 사이의 값을 갖는다. 1미만의 값은 'Bad' 2미만은 'Poor' 3미만은 'Fair' 4미만은 'Good' 그이상은 'Excellent'로 평가한다[14].

2.4 주파수 변이(Frequency-variant)

주파수 변이는 각 주파수 대역별로 음성 가중치 값을 주어 spectral distance 를 측정한 후, 주파수의 변화도를 계산하는 방법으로 주파수 변이 $fwVar$ 는 식(5)와 같이 나타낸다[12].

$$fwVar = a_0 + \sum_{j=1}^k a_j D_j \quad (5)$$

여기서 a_j : regression coefficients (비선형회귀계수) k : number of bands (밴드의 개수) 이다.

식(4)에서 입력 음성의 각 프레임에서 가중 스펙트럼 기울기 D_j 는 각 프레임의 왜곡도를 말하며 식(6)과 같다[12].

$$\begin{aligned} D_j &= \frac{1}{M} \sum_{m=1}^M 10 \log_{10} [F^2(m, j) / (F(m, j) - \hat{F}(m, j))^2] \\ j &= 1, 2, \dots, k \end{aligned} \quad (8)$$

사전 실험 결과, 동일한 잡음환경일 경우 주파수 변이도가 4.21~4.64의 값을 보이고 있으나 다른 종류의 잡음일 경우 이보다 낮은 1~3 사이의 값을 나타내고 있어 타 잡음에 대한 변별력이 있음을 알 수 있었다.

2.5 최대우도를 이용한 잡음군의 선택

기존에는 입력 잡음음성이 속하는 잡음군의 선택에 있어 각 분류척도의 평균치를 임계치로 설정하여 이 임계치를 초과하는 분류값을 가진 잡음군을 선택

하였다. 이 경우 임계치에 미달하는 분류값을 갖는 입력 잡음음성은 미분류로 처리되어 기존의 PMC 인식방법에 의해 인식 처리되어 인식을 저하의 원인이 되었다. 이때 잡음입력음성이 포함되는 잡음군 선택을 위해 최대우도(maximum likelihood estimation, MLE)를 이용할 경우 잡음 분류율이 향상될 것으로 기대된다. 최대우도(maximum likelihood estimation, MLE)는 아래와 같이 표현된다[15].

$$\hat{\theta} = \underset{1 \leq \theta \leq L}{\operatorname{argmax}} \sum_{t=1}^N \log p(\vec{x}|\lambda) \quad (8)$$

2.6 각 거리척도별 잡음군 분류결과

각 거리척도 별 잡음 분류의 결과를 표 2에 나타낸다. 평균값을 임계치로 설정하여 잡음군을 선택하는 기존 방법을 사용하는 경우 PESQ는 평균 23.5%, WSS는 평균 33.9% FV (Frequency-variant)의 경우 56.5%의 잡음 분류율을 보이고 있다. 여기에 최대우도를 부가할 경우, PESQ가 55.5%, WSS가 69.2%, FV가 82.7%의 평균 잡음 분류율을 나타낼 수 있다. 이는 최대우도를 부가하는 방법이 잡음 분류 척도에 있어서 효과적임을 알 수 있다. 최대우도를 부가한 척도중 PESQ의 경우, Car 잡음일 때 97.32%의 잡음 분류율을 보여 매우 우수한 결과를 보였으나 그 외의 잡음군의 경우에는 매우 낮은 잡음 분류율을 보임으로써 평균 잡음 분류율이 55.3%로 나타나 ML을 부가한 거리척도 중 가장 낮은 잡음 분류율을 나타내고 있다. WSS의 경우에도 Car 잡음과 Train-station 잡음 분류율은 각각 99.78%과 85.21%로 나타나 주파수 변이 방법(Frequency-variant)보다 높

은 잡음 분류율을 보이고 있으나, 각 잡음군별 잡음 분류율의 편차가 크고 평균잡음 분류율은 69.19%를 보이고 있다. 주파수 변이의 경우, 각 잡음군별로 비교적 안정적인 잡음 분류율을 보이고 있으며, 평균 잡음 분류율이 82.66%로 나타나 가장 높은 분류결과를 나타낼 수 있다. 따라서 이를 인식실험에 이용하기로 한다.

3. 인식실험 방법(FV-PMCI)

기존의 인식방법(FV-PMC)과 구별하기 위해 본 인식방법을 FV-PMCI로 나타낸다. 그림 2에 제안한 인식시스템의 블록도를 나타낸다. 인식은 3단계 과정으로 나누어 진행되는 데 이하 단계별 인식 과정에 대해 기술한다.

과정 1 : 각 잡음음성군별 PMC(Parallel Mixture Model) 생성

실 환경에서 입력음성에 혼입 가능한 잡음들에 대하여 각 잡음음성군별로 평균과 분산 파라미터를 추

그림 2. 제안하는 인식시스템 블록도

표 2. 각 거리척도 별 잡음 분류율

(단위: %)

Method	setA					setB					setC		
	Subway	Babble	Car	Exhibition	Avg	Restaurant	Street	Airport	Train-station	Avg	Subway (MIRS)	Street (MIRS)	Avg
PESQ	29.6	16.7	16.7	33.2	24.0	21.8	16.7	16.7	35.6	22.7	31.2	16.7	23.9
WSS	32.3	32.9	31.9	33.8	32.7	34.6	35.3	37.6	36.6	36.0	31.1	34.9	33.0
FV	58.1	53.8	56.0	52.1	55.0	55.7	54.0	55.1	61.5	56.6	61.8	54.0	57.9
PESQ +ML	56.8	34.0	97.3	70.2	64.6	29.2	55.2	28.1	82.7	48.8	54.2	52.1	53.2
WSS +ML	77.7	44.1	99.8	85.4	76.8	28.4	61.1	32.3	85.2	51.7	89.1	69.1	79.1
FV +ML	86.9	78.2	87.2	89.2	85.4	78.8	63.2	62.9	84.8	72.4	95.9	84.5	90.2

출한 후 개선된 PMC 잡음 모델을 생성한다.

과정 2: 잡음의 분류

잡음의 분류는 사전 실험에서 가장 높은 잡음 분류율을 보인 주파수 변이와 최대우도를 이용하여 분류한다.

과정 3: 음성인식 수행

단계 2에서 분류된 입력음성이 포함된 잡음음성은 단계 1에서 생성된 해당 잡음음성 PMC를 이용하여 음성 인식을 수행한다. 기존의 FV-PMC 방법에서는 임계치에 도달하지 못하는 입력음성을 무잡음 모델과 결합하여 음성 인식을 수행하였으나, 제안한 방법에서는 최대 우도를 이용하여 분류하므로 미분류되는 경우가 발생하지 않는 장점이 있다.

4. 실험 결과 및 고찰

4.1 실험환경

실험 및 성능평가를 위하여 본 논문에서는 Aurora 2.0 데이터베이스를 이용한다[16]. Aurora 2.0에는 2 종류의 훈련환경 즉, 8,440개의 clean 발성으로 구성된 clean-condition과 동일한 발성을 20개의 잡음환경으로 나누어 각 422개의 발성으로 구성된 multi-condition으로 나누어져 있다. 또, 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)의 훈련환경을 부가하기 전 채널왜곡 특성으로 필터링 된 잡음을 포함한 set C로 나누어져 있다. 성능 평가에서는 set A, B, C 의 각 잡음의 종류에 대해서 잡음을 20dB에서 0dB까지 5dB구간으로 나눈 5 가지 레벨에 대해 평균 단어 인식률(word accuracy)을 구하여 비교한다.

4.2 실험 결과

표 3에 제안한 방법에 따라 실시한 음성 인식 실험 수행 결과를 나타내었다. 표 3으로부터 알 수 있는 바와 같이 set A의 인식률은 84.20%, setB의 인식률은 81.25% , set C의 경우 83.12%로 평균인식률은 82.9%로 나타났다. 분류척도에 따라 잡음음성이 올바른 잡음군으로 분류(정분류)되었을 경우와 그렇지 못할 경우(오분류)에 대해 인식률의 차이를 확인하기 위하여 인식실험을 실시한 결과를 표 4에 나타내었다. 표 4에서 알 수 있는 바와 같이 올바르게 분류되었을 경우의 인식률이 현저히 높은 것으로 나타나 잡음 분류율 향상이 인식률을 향상에 중요한 요소임을 확인할 수 있다.

표 3에는 제안한 FV-PMC II 방법의 음성인식률을 나타내었고 표 4에는 입력되는 음성들이 올바르게

표 3. FV-PMC II 방법의 인식률 (단위:%)

Method	Set A					Set B					Set C		
	Subway	Babble	Car	Exhibition	Avg	Restaurant	Street	Airport	Train-station	Avg	Subway (MIRS)	Street (MIRS)	Avg
20dB	95.29	96.09	96.95	94.63	95.74	93.53	94.75	95.47	93.91	94.41	94.58	94.79	94.69
15dB	92.38	92.91	94.27	91.14	92.67	89.48	92.33	91.31	89.55	90.67	92.69	91.83	91.83
10dB	89.99	89.41	92.07	89.92	90.35	88.83	89.68	89.14	85.63	88.32	89.74	90.06	90.06
5dB	83.91	80.28	83.94	82.30	82.61	76.54	81.03	79.12	76.66	78.34	82.76	80.41	81.59
0dB	63.32	51.59	59.70	63.99	59.65	50.77	58.08	58.41	50.85	54.53	59.14	55.76	57.45
avg	84.98	82.06	85.38	84.40	84.20	79.83	83.18	82.69	79.32	81.25	83.78	82.57	83.12

표 4. 정/오 분류에 따른 평균 음성인식률

(%)

분류	Set A				Set B				Set C	
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train-station	Subway (MIRS)	Street (MIRS)
정분류	86.90	84.58	87.57	86.53	84.72	84.85	86.06	80.73	85.41	84.30
오분류	81.93	77.04	78.88	74.88	63.19	82.94	80.61	76.30	78.26	81.35

게 분류되었을 경우 음성에 포함된 잡음으로 훈련된 잡음모델을 사용하여 음성인식을 수행한 결과와 다른 잡음으로 분류되었을 경우 음성에 포함된 잡음환경이 아닌 다른 잡음모델을 사용하여 음성인식을 수행한 음성인식률을 나타내었다. 표 3의 음성 인식률에서 set B의 Restaurant의 평균 인식률이 79.38% Train-station의 평균인식률이 79.32%로 다른 잡음 음성의 평균 인식률보다 떨어지는 결과를 나타내었다. 같은 set B의 Street, Airport의 경우 잡음 분류율은 Restaurant, Train-station에 비해 낮았으나 음성 인식률의 경우에는 오히려 3% 가량 높게 나타났다. 이는 표 4에 나타나듯이 Restaurant, Train-station 두 잡음 모두 잡음이 정분류된 경우와 오분류된 경우의 음성 인식률이 각각의 평균 음성 인식률에 미치지 못하고 있기 때문인 것으로 분석된다. 특히 Restaurant의 경우 오분류 되었을 경우 평균 음성 인식률이 다른 잡음에 비해 현저히 낮게 나타남을 알 수 있다.

표 5에 Aurora 2 database의 baseline[16]과 본 논문에서 제안한 FV-PMCII 방법의 음성인식 성능을 기존의 PMC 및 FV-PMC 방법과 비교한 결과를 나타내었다. 표 5에서 나타난 바와 같이 본 논문에서 제안한 FV-PMCII 방법은 PMC 방법이나 FV-PMC 방법에 비해 낮은 SNR 특히 5dB와 0dB에서 높은 인식률을 향상을 보이고 있다. 그러나 set B와 set C에서는 높은 SNR에서 기존의 FV-PMC 방법이나, PMC 방법보다 인식률이 떨어지는 문제점이 나타났는데, 이는 비교적 잡음이 적게 포함된 높은 SNR에서는 무잡음 모델과 결합하는 PMC 방법이 오히려, 잡음모델과 결합하여 음성 인식을 수행하는 방법보다 유효한 것으로 분석된다. 그러나 모든 잡음군에 대한 평

그림 3. 잡음 분류율에 따른 음성 인식률

균인식률이 82.9%로 나타나 기존의 FV-PMC보다 2.3% 향상된 인식성능을 나타내어 제안한 방법의 유효성을 확인할 수 있다.

그림 3은 잡음 분류율과 인식률과의 관계를 나타내고 있다. 그림 3에서와 보는 바와 같이 대부분의 잡음환경에서는 잡음 분류율 향상과 더불어 인식률이 향상되는 것으로 나타나고 있으나 set B의 Street, Airport 잡음의 경우, 낮은 잡음 분류율에도 불구하고 다른 잡음음성과 비슷한 음성 인식률을 보이고 있는 데, 이는 표 4에서 알 수 있듯이 오분류된 잡음 음성들의 음성 인식률이 다른 잡음 음성의 인식률에 비해 높은 음성 인식률을 나타내고 있기 때문인 것으로 분석된다.

5. 결 론

본 논문에서는 잡음 환경 하에서의 음성 인식 성능 저하의 주요 원인이 되고 있는 훈련과 인식의 환경적 차이를 줄이기 위한 방법으로 주파수 변이 방법에 최대우도법을 적용하여 잡음을 분류하고, 분류된 잡음 모델로 음성 인식을 수행하는 방법을 제안하였

표 5. 각 알고리즘별 인식률 비교(%)

Method	SET A				SET B				SET C			
	Baseline	PMC	FV_PMC	FV-PMC II	Baseline	PMC	FV_PMC	FV-PMC II	Baseline	PMC	FV_PMC	FV-PMC II
20dB	95.25	95.23	95.23	95.74	92.77	94.97	93.42	94.41	94.30	95.24	95.67	94.69
15dB	87.33	93.21	93.26	92.67	81.34	92.57	89.87	90.67	87.84	92.01	94.17	91.83
10dB	67.71	86.44	92.35	90.35	59.01	85.84	90.48	88.32	74.15	82.39	91.90	90.06
5dB	39.48	68.83	70.83	82.61	31.93	67.71	72.86	78.34	50.24	61.78	78.91	81.59
0dB	16.95	37.11	43.59	59.65	13.70	38.50	50.54	54.53	24.17	32.14	56.20	57.45
avg	67.62	76.16	79.05	84.20	62.96	75.92	79.43	81.25	71.62	72.71	83.37	83.12

다. Aurora 2.0 DB를 이용한 인식 실험 결과, 본 논문에서 제안한 FV-PMCII 방법은 PMC 방법이나 FV-PMC 방법에 비해 낮은 SNR 특히 5dB와 0dB에서 높은 인식률 향상을 보이고 있음을 알 수 있었다.

분류척도에 따라 잡음음성이 올바른 잡음군으로 분류되었을 경우와 그렇지 못할 경우에 대해 인식률의 차이를 확인하기 위하여 인식실험을 실시한 결과 올바르게 분류되었을 경우의 인식률이 현저히 높은 것으로 나타나 잡음 분류율의 향상이 인식률을 향상에 중요한 요소임을 확인할 수 있다. 제안한 FV-PMCII 방법은 기존의 FV-PMC 방법보다 모든 잡음군에 대한 평균인식률이 82.9%로 나타나 기존의 FV-PMC보다 2.3% 향상된 인식성능을 나타내어 제안한 방법의 유효성을 확인할 수 있었다.

참 고 문 헌

- [1] 김상만, 서광석, 김종교, “이산 웨이브렛과 비균일 필터뱅크를 적용한 음성특징 추출,” 정보통신산업진흥원, 2000.
- [2] Gong Y., “Speech Recognition in Noisy Environments: A Survey,” *Speech Communication*, Vol. 16, Issue 3, pp. 261-292, 1995.
- [3] 최숙남, 신광호, 정현열, “캡스트럼 정규화와 캡스트럼 거리기반 목음특징정규화 방법을 이용한 잡음음성 인식” 멀티미디어학회논문지, 제14권, 제10호, pp.1221-1228, 2011
- [4] J.C. Junqua and J.P. Haton, “Robustness in Automatic Speech Recognition: Fundamentals and Applications,” Kluwer Academic Publishers, Netherlands, 1996.
- [5] J.S. Lim, “Speech Enhancement,” Prentice Hall, New Jersey, 1983.
- [6] D.H. Klatt, “A Digital Filterbank for Spectral Matching,” *Proc. ICASSP*, pp. 573-576, 1979.
- [7] A.P. Varga and R.K. Moore, “Hidden Markov Model Decomposition of Speech and Noise,” *Proc. ICASSP*, pp. 845-848, 1990.
- [8] M.J.F. Gales and S. Young, *Model Based Techniques for Noise Robust Speech Recognition*, Dissertation at the University of Cambridge, 1995.
- [9] M.J.F. Gales and S. Young, “An improved Approach to the Hidden Markov Model Decomposition of Speech and Noise,” *Proc. ICASSP-92*, Vol. 1, pp. 233-236, 1992.
- [10] 최숙남, 정현열, “주파수 변이를 이용한 환경인식 기반의 GMM 적용에 관한 연구,” 한국 신호처리·시스템 학회 추계학술대회 논문집, pp. 181-185, 2012.
- [11] Richard O. Duda, Peter E. Hart, and David G. Stock, *Pattern Classification 2nd Edition*, Wiley-Interscience, New York, 2001.
- [12] Philipos C. Loizou, *Speech Enhancement -Theory and Practice*, CRC Press, Florida, 2007.
- [13] ITU-T PSQM, *Objective Quality Measurement of Telephone-band(300-3400Hz) Speech Codecs*, 2001.
- [14] J.G. Beerends, A.P. Heckstra, A.W. Rix, and M.P. Hollier, “Perceptual Evaluation of Speech Quality (PESQ) the New Itu Standard for End-to-end Speech Quality Assessment Part II - Psychoacoustic Model,” *Journal of the Audio Engineering Society*, Vol. 50, No. 10, pp. 765-778, 2002.
- [15] 손영호, 최재훈, 장준혁, “환경인식 기반의 향상된 Minimum Statistics 잡음전력 추정기법,” 한국음향학회지, 제30권, 제3호, pp. 123-128, 2011.
- [16] H.-G Hirsch and D. Pearce, “The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions,” *ISCA ITRW ASR 2000*, 2000.

최 숙 남

1995년 8월 영남대학교 전자공학과 공학사
2007년 8월 영남대학교 전기전자통신교육 교육학석사
2013년 8월 영남대학교 정보통신공학과 박사졸업

관심분야: 음성인식, 화자인식, 디지털 신호처리

정 현 열

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 응용분야