잡음 마스킹 레벨에 따른 복수 모델을 이용한 자동차 소음환경에서의 음성인식

정회인, 송명규**, 권영욱**, 심갑종***, 김형순***

*국방과학 연구소, **부산대학교 전자공학과, ***현대 자동차 승용전자 설계 2 팀

Speech recognition in car noise environments using multiple models according to noise masking levels

Hoi In Hung, Myung Gyu Song**, Young Uk Kwon**, Kab-Jong Shim*** and Hyung Soon Kim***

*Agency of Defense Development, **Dept. of Electronics Eng., Pusan National Univ.,

***Passenger Car E&R Center II, Hyundai Motor Company

E-mail: {mgsong, kimhs}@hyowon.pusan.ac.kr

요약문

음성인식 시스템의 실용화 과정에서 훈련환경과 테스트 환경의 불일치로 인한 인식성능의 저하는 반드시 극복되어야 할 문제이다. 본 논문에서는 잡음 섞인 입력 음성의 비음성 구간에서 잡음레벨을 추정하여 음성 스펙트럼에서 추정된 잡음레벨을 빼는 스펙트럼 차감법(Spectral Subtraction)과 스펙트럼 영역에서 미리 정해진 마스킹 레벨(masking level)보다 낮은 에너지 값을 마스킹 레벨로 올려 주는 잡음 마스킹(noise masking)을 함께 사용함으로써 훈련환경과 테스트환경의 불일치를 줄이는 방법을 제안한다. 그리고 복수의 마스킹 레벨에 대한 모델들을 미리 만들어 두고 추정된 잡음 레벨에 따라 적합한 마스킹 레벨의 모델을 사용하여 인식을 수행하는 다중 모델(multiple model)방법을 적용하였다.

자동차 소음환경에서 두 가지 마스킹 레벨에 대한 모델을 이용한 화자 독립 고립단어 인식 실험을 통하여 본 논문에서 제안한 방식은 정차중 무시동 환경에서 95.8%, 정차중 시동 환경에서 95.6%, 한적한 도로 환경에서 92.8%, 복잡한 시내도로 환경에서 89.6%, 고속도로 환경에서 74.4%의 인식성능을 나타내었으며, 평균 90.7%의 성능을 얻을 수 있었다.

1. 서론

채널 왜곡이 있거나 부가 잡음이 존재하는 상황에서의 음성인식은 훈련환경과 테스트환경 사이의 불일치로 인하여 성능저하가 발생한다. 이 문제의 원인으로 음성을 입력하는 과정에서 발생하는 서로 다른 채널의 특성, 환경에 따른 화자의 발성 방식의 차이, 그리고 주변잡음 등을 들 수 있다. 특히 주행 중인 차량 내에서의 음성인식의 경우 주변잡음의 영향에 의한 불일치가 인식성능 저하의 주 원인이 된다. 이 문제의 해결을 위한 접근 방법은 크게 두 가지 부류로 나눌 수 있는데, 음성 인식을 위한 전처리 단계로 잡음에 의해 손상된 음성신호의 파형 또는 파라미터를 복구하는 음질개선 방법(speech enhancement technique)과 음성 신호의 최적 추정치를 계산해내는 대신에 잡음 환경하에서 인식단계의 모델을 보상하는 모델보상 방법(model compensation technique) 등 이다[1].

본 논문에서는 잡음의 효과를 최소화 하기위해 잡음 섞인 음성의 비 음성구간에서 잡음의 스펙트럼을 추정하여 잡음 섞인 음성의 스펙트럼에서 잡음의 스펙트럼을 크기 성분을 제거하는 스펙트럼 차감법(spectral subtraction)을 수행하였다. 이렇게 추정된 음성 스펙트

럼에는 여전히 잔여 잡음(residual noise)이 존재하는데, 이를 제거하기 위해 스펙트럼 영역에서 미리 정해진 잡음 마스킹 레벨보다 낮은 에너지 값을 잡음 마스킹 레벨로 올려 줌으로써 훈련환경과 테스트 환경 사이의 불일치를 줄이는 방법을 제안하였다. 또한 복수의 잡음 마스킹 레벨에 대한 모델들을 미리 만들어 두고 추정된 잡음 레벨에 따라 적합한 잡음 마스킹 레벨의 모델을 사용하여 인식을 수행하는 다중 모델 인식방법을 적용하였다. 화자독립 고립단어 인식실험을 통해서 제안된 방식이 잡음환경에서 우수한 성능을 나타냄을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 서론에 이어 제 2 장에서는 스펙트럼 차감법 및 잡음 마스킹에 대해 설명하고, 제 3 장에서는 잡음 마스킹 레벨에 따른 다중 모델을 이용한 전처리에 대해서 설명한다. 그리고 제 4 장에서는 실험 방법 및 결과를 기술하고, 제 5 장에서 결론을 맺는다.

2. 스펙트럼 차감법 및 잡음 마스킹

잡음 신호 $n(t)$가 음성신호 $s(t)$에 더해 졌을 때, 잡음 섞인 음성신호 $x(t)$는 다음과 같이 나타낼 수 있다고 가정한다.

$$x(t) = s(t) + n(t) \tag{1}$$

이 신호에 윈도우를 취하여 단구간 푸리에 변환을 하면, 다음과 같다.

$$X(\omega) = S(\omega) + N(\omega) \tag{2}$$

주변 잡음을 제거하는 대표적인 방법인 스펙트럼 차감법[2]은 주변 잡음에 의해 손상된 음성 스펙트럼에서 잡음 스펙트럼의 크기 성분만을 제거하는 방법으로, 음성신호의 크기 추정치를 다음과 같이 구한다[3].

$$|\hat{S}(\omega)| = \begin{cases} |X(\omega)| - \alpha\overline{|N(\omega)|}, \\ \quad if\ |X(\omega)| > (\alpha+\beta)\overline{|N(\omega)|} \\ \beta\overline{|N(\omega)|},\ otherwise \end{cases} \tag{3}$$

여기서 α는 overestimation factor 로 SNR 에 따라서 미리 정의된 상수이며, β는 스펙트럼 차감시 음의 결과를 갖지 않도록 하는 spectral flooring factor 로 잡음 피크(noise peaks)의 진폭을 줄이는 역할도 한다. 실제로 잡음 스펙트럼의 크기 추정치는 음성이 없는 잡음 구간에서 수 프레임의 데이터로부터 구한 샘플 평균을 사용하며, 다음과 같이 구한다.

$$\overline{|N(\omega)|} = \frac{1}{M}\sum_{i=1}^{M}|N_i(\omega)| \tag{4}$$

식 (3)에 의한 스펙트럼 차감법의 적용시 α를 증가시키면 잔여 잡음(residual noise)은 줄어들지만 음성의 왜곡이 심해지고, 반대로 α를 감소 시키면 잔여 잡음은 늘어 나고 음성의 왜곡은 줄어든다. 실제로는 스펙트럼 차감법의 적용 후에도 여전히 잔여 잡음이 존재하게 되는데 이를 제거하기 위해서 본 논문에서는 잡음 마스킹(noise masking)을 도입한다. 잡음 마스킹은 잡음이 존재하는 상황에서 음성 신호의 인지도가 감소하는 심리학적 현상인데, 이러한 효과를 음성인식 시스템에 이용하면 음성 식별에 있어서 에너지가 작은 부분의 기여도를 줄일 수 있다. 즉, 주변 환경의 변화에 영향을 덜 받도록 음성 스펙트럼을 변경할 수 있다[4][5][6]. 본 논문에서는 필터뱅크 영역에서 낮은 에너지 값을 미리 정해진 임계값(masking level)로 올려줌으로써 스펙트럼 차감법 수행후의 잔여 잡음을 제거하고 훈련과 테스트 환경 사이의 비음성 구간의 에너지 불일치를 줄여 줌으로써 인식률 향상을 꾀하였다. 제안된 방식은 스펙트럼 차감후의 필터뱅크의 에너지 값 $|S(\omega_k)|$과 미리 정해진 마스킹 레벨(ML)의 최대값을 취함으로써 구현된다.

$$Y(\omega_k) = \max(|S(\omega_k)|, ML) \tag{5}$$

여기서 $|S(\omega_k)|$는 스펙트럼 차감후의 k 번째 필터뱅크 에너지 값이며, $Y(\omega_k)$는 마스킹된 k 번째 필터뱅크 값이다. 이 방식의 실제적인 문제는 마스킹 레벨을 어떻게 정하느냐 하는 것이다. 마스킹 레벨이 너무 낮으면 잡음에 대한 마스킹 효과가 작아지고, 반대로 너무 높으면 음성신호 자체가 왜곡되게 된다. 본 논문에서는 마스킹 레벨을 여러가지 잡음 환경이 고려된 상황에서 적절한 값을 선정하여 인식 실험을 수행하였다.

3. 잡음 마스킹 레벨에 따른 다중 모델을 이용한 음성인식 시스템

스펙트럼 차감법 및 잡음 마스킹에 의한 hybrid 방식이 잡음환경에서의 음성인식 시스템의 성능 향상에 유용한 것은 사실이나, 단일 마스킹 레벨로는 다양한 잡음환경에 효과적으로 대처하기 곤란하다. 이에 따라, 본 논문에서는 여러가지 마스킹 레벨을 가지는 다중 모델을 미리 구성한 다음, 입력 음성으로부터 추정한 잡음 레벨에 적합한 마스킹 레벨의 모델을 이용하여 인식을 수행하는 방식을 도입하였다.

이 방식에 의한 음성 인식 시스템의 인식 수행 과정은 다음과 같다. 입력 음성신호의 초기 몇 프레임을 비음성 구간이라고 가정하여 초기 몇 프레임의 멜 스케일 필터뱅크 에너지의 평균과 표준편차(σ)를 구한다. 에너지의 평균을 잡음레벨의 추정치로 하여 식 (3)과 같이 스펙트럼 차감법을 수행하고 에너지의 표준편차의 일정 비율에 해당하는 범위를 마스킹하는 정도로 마스킹 레벨을 결정한다. 이는 스펙트럼 차감법으로 현재 입력 음성신호의 잡음레벨의 평균값이 제거되고 난 후의 잔여 잡음의 fluctuation 은 추정된 표준편차와 관련이 있다는 생각에 근거한다. 실제로 마스킹 레벨은 다음과 같이 구한다.

$$ML_{dB} = 20\log(\sigma \times \gamma) \qquad (6)$$

여기서 σ는 잡음의 에너지 표준 편차이고, γ는 표준편차에 대한 weighting factor 이다. 위의 식 (6)에 의해 계산된 ML_{dB}은 무한한 가지 수의 값을 가지게 된다. 그러나 현실적으로는 몇몇 마스킹 레벨에 해당하는 다중 모델을 갖는 인식 시스템을 구성하는 것이 바람직하므로 식 (6)에 의해 구한 값을 시스템이 가진 마스킹 레벨로 사상(mapping)시켜야 한다. 예를 들어 3 가지의 마스킹 레벨 69dB, 74dB, 80dB 에 해당하는 다중 모델이 구성되어 있다면, 식 (6)에 의해 계산된 값이 69dB 미만이면 마스킹 레벨을 69dB 로, 69dB 이상 74dB 미만이면 마스킹 레벨을 74dB 로, 74dB 이상은 마스킹 레벨을 80dB 로 사상시킨다.

결정된 마스킹 레벨에 따라서 스펙트럼 차감된 필터뱅크 출력을 식 (5)과 같이 마스킹하고 그 결과 필터뱅크 출력의 log 값을 DCT 하여 음성 특징 파라메터를 추출한다. 인식 모듈에서는 결정된 마스킹 레벨정보를 받아 그 레벨로 미리 만들어진 모델을 사용하여 인식을 수행한다. 이와 같이 현재 입력데이터의 잡음환경을 자동적으로 분석하여 그에 적합한 모델로 인식을 수행함으로써 인식성능의 향상과 여러가지 다양한 잡음 환경을 고려할 수 있으므로 시스템의 flexibility 를 높일 수 있다.

4. 실험 및 결과

4.1 데이터 베이스 및 인식 시스템

인식 실험에 사용한 음성 데이터 베이스는 자동차 주행 상황에서 각종 기기 조작을 위한 50 개의 고립단어 명령들로 구성되어 있다. 이들 어휘를 59 명의 남성화자가 한 번씩 발성한 것을 8kHz 로 샘플링하여 녹음하였다. 그 중 49 명의 화자의 음성을 고립단어 모델 훈련에 사용하였고, 나머지 10 명의 음성데이터로 인식 실험을 수행하였다. 잡음 섞인 음성데이터는 실차 환경에서 녹음한 여러 가지 상황의 잡음을 실제 상황에 맞게 scaling 하여 잡음 없는 음성 데이터에 더해줌으로써 잡음환경의 음성데이터로 모의 구성하였다. 실차 환경에서 녹음한 잡음 환경은 정차중 무시동(env1), 정차중 시동(env2), 조용한곳 주행(env3), 시내 주행(env4), 고속도로 주행(env5) 등이다. 이들 실차 환경의 SNR 은 각각 21dB, 10dB, 2dB, 0dB, -5dB 였다.

고립단어 모델은 코드북 크기가 64 인 이산(discrete) HMM 으로 구현하였으며, 각 단어를 구성하는 음소수의 3 배가 되도록 state 를 할당하였다. 음성 특징 벡터로는 비교적 잡음에 강인하다고 알려진 mel-frequency cepstral coefficient(MFCC)를 사용하였는데, 26 개의 멜 스케일 필터뱅크를 이용하여 12 차의 MFCC 를 추출하였다.

4.2 인식 실험 및 결과

인식 실험은 잡음 섞인 음성에 대해서 아무런 전처리를 하지 않은 baseline 테스트, 스펙트럼 차감법과 잡음 마스킹을 함께 적용한 테스트를 해보았다. 표 1 은 잡음 없는 음성과 앞에서 언급한 다섯 가지 상황 잡음이 섞

인 음성의 baseline 인식 실험 결과이고, 표 2은 α는 1.0, β는 0.01로 스펙트럼 차감법을 수행하고 마스킹 레벨에 따른 잡음 마스킹을 적용한 인식 실험 결과이다.

표 1. Baseline 인식률(%)

잡음 없는 음성	잡음 섞인 음성					평균
	env1	env2	env3	env4	env5	
96.4	93.0	90.8	76.6	64.6	22.4	74.0

표 2. 스펙트럼 차감법(α=1.0, β=0.01)과 잡음 마스킹을 수행한 경우의 인식률(%)

마스킹 레벨	잡음 없는 음성	잡음 섞인 음성					평균
		env1	env2	env3	env4	env5	
69dB	**96.2**	**95.8**	**95.6**	93.2	**90.2**	61.8	**88.8**
72dB	93.2	93.8	94.0	92.8	87.2	59.6	86.8
74dB	94.4	93.4	94.0	**93.4**	90.0	65.2	88.4
76dB	94.4	93.0	93.4	90.8	88.8	66.6	87.8
77dB	93.4	92.2	92.0	90.2	88.0	69.0	87.5
78dB	93.0	93.6	93.0	91.2	87.4	71.4	88.3
80dB	91.8	90.8	90.6	88.6	86.2	**74.4**	87.1
84dB	89.6	89.6	89.6	87.6	85.0	72.8	85.7
86dB	86.0	84.2	84.4	82.6	79.6	69.8	81.1

위의 표에서 기존의 스펙트럼 차감법을 수행한 후 잔여 잡음을 제거하기 위해 잡음 마스킹을 도입함으로써 평균적인 에러율이 baseline에 비해서는 57%가 줄었음을 알 수 있다. 또한 잡음이 섞이지 않은 음성에 대한 인식실험의 경우 마스킹 레벨을 높일수록 인식성능이 저하되는데 이는 마스킹 레벨이 음성정보의 일부를 마스킹 하여 신호의 왜곡이 발생하기 때문이다.

다중 모델을 두고 입력음성의 잡음레벨을 추정해서 그에 적합한 마스킹 레벨의 모델을 이용하여 인식을 수행하는 실험은 초기 8프레임의 필터뱅크 출력의 평균과 표준편차(σ)를 구한 다음, 평균을 이용해 스펙트럼 차감법을 수행하고 weighting된 표준편차를 마스킹하는 정도의 값으로 마스킹 레벨을 선택하였다. 표준편차에 대한 weighting factor, γ는 0.2에서 3.0까지의 다양한 값을 고려하였다. 다중 모델은 마스킹 레벨에 따라 여러 가지로 구현해 보았으나 실험을 통해 2가지의 마스킹 레벨을 가지는 다중 모델로도 충분한 성능을 얻을 수 있었다. 표 3는 두 가지 마스킹 레벨을 가지는 다중모델의 γ값에 따른 인식률 및 γ와 마스킹 레벨에 따른 테스트 음성 분포를 나타낸다. 예를 들어 γ가 0.3인 경우, 잡음이 섞이지 않은 음성과 잡음이 섞인 음성의 평균 인식률이 90.7%이다. 총 500개의 테스트 데이터에 대해 조용한 곳 주행(env3)환경의 인식률이 92.8%이고, 500개 중에서 478개가 69dB, 나머지 22개가 80dB의 마스킹 레벨로 결정되었다.

표 3. 다중모델을 이용한 인식실험

(a) 가중치 γ에 따른 인식률(%)

γ	잡음 없는 음성	잡음 섞인 음성					평균
		env1	env2	env3	env4	env5	
0.3	96.2	95.8	95.6	92.8	89.6	74.4	90.7
0.7	96.2	95.6	95.6	91.4	87.0	74.4	90.0
1.5	96.2	93.8	95.6	88.6	87.2	74.4	89.3

(b) 가중치 γ와 ML에 따른 테스트 음성의 분포

γ	마스킹 레벨	잡음 없는 음성	잡음 섞인 음성				
			env1	env2	env3	env4	env5
0.3	69dB	500	500	500	478	458	12
	80dB	0	0	0	22	42	488
0.7	69dB	500	488	500	252	95	0
	80dB	0	12	0	248	405	500
1.5	69dB	500	336	489	1	29	0
	80dB	0	164	11	499	471	500

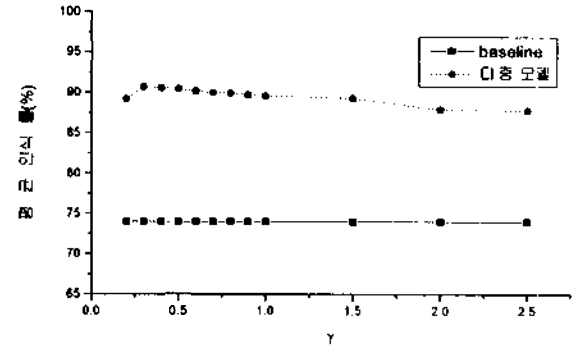

그림 1. 두 가지의 다중모델을 이용한 경우의 γ에 따른 인식률

위의 그림은 두 가지의 다중 모델을 사용한 경우의 γ에 따른 인식률을 보여 준다. 여기서 γ=0.3 근처에서 최적값을 가지나, γ 값의 변화에 따른 인식성능의 변화가 크지 않음을 알 수 있다. 또한 표 3으로부터 잡음 없는 음성 및 잡음 섞인 음성의 각 환경에서의 가장 높은 인

식률의 평균으로 구하 는 인식률의 이론적인 상한이 90.9%임을 감안하면, 두 가지의 다중 모델을 사용하여 얻은 평균 인식률 90.7% (γ = 0.3)는 비음성 구간 에너지의 표준편차에 근거한 마스킹 레벨의 선정이 적절함을 확인 시켜 주었다.

5. 결론

본 논문에서는 자동차 잡음환경에서 음성인식 시스템을 구성할 때의 부가잡음의 영향을 효과적으로 제거하기 위한 음성인식의 전처리 알고리즘을 제안하고 인식실험을 수행하였다. 본 논문에서 도입한 방식은 멜 스케일 필터뱅크 출력에 기존의 스펙트럼 차감법을 적용한 후 잔여 잡음을 제거하기 위해 잡음 마스킹을 수행하는 것이다. 특히, 본 논문에서는 여러 개의 마스킹 레벨에 대한 모델을 미리 구성해 놓고 입력 음성신호의 초기 부분에서 잡음레벨을 추정하여 스펙트럼 차감법을 수행한 다음 적절한 마스킹 레벨을 추정하고 인식과정에서 추정된 마스킹 레벨로 훈련된 모델을 사용하는 다중 모델방법을 적용하였다. 인식 실험 결과 단지 두개의 모델을 사용함으로써 다양한 잡음환경에 걸쳐 우수한 인식성능을 얻을 수 있었다. 또한 본 연구에서는 모델 훈련에 잡음 없는 음성(clean speech)만을 사용했음에도 불구하고 우수한 성능을 얻을 수 있었는데, 실제 잡음 환경의 음성(noisy speech)을 사용하여 모델을 훈련한다면 추가적인 성능향상이 기대된다.

앞으로 주파수 밴드별 마스킹 레벨의 적용을 검토 중이며, 이 경우의 다중 모델의 구현에 대해 연구할 계획이다. 그리고, 인위적으로 잡음을 부가하여 만든 잡음 음성이 아닌 실차 환경에서 얻은 잡음 음성에 대한 인식실험이 진행되고 있다.

참 고 문 헌

[1] J.C.Junqua and J.P.Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, Kluwer Academic Publishers, 1996

[2] S. F. Boll and D.C. Pulsipher, "Suppression of acoustic noise in speech using two microphone adaptive noise cancellation," IEEE Trans. Acoust., Speech Signal Processing, vol. 28, pp.752-755, 1980

[3] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," In Proc. ICASSP, pp.208-211, 1979

[4] D.H. Klatt, " A digital filter bank for spectral matching," In Proc. ICASSP, pp.573-576, 1976

[5] J.N. Holmes and N.C. Sedwick, "Noise compensation for speech recognition using probabilistic models," In Proc. ICASSP, pp.741-744, 1986

[6] A. Varga and K. Ponting, "Control experements on noise compensation in hieedn Markov model based continuous word recognition," In Proc. EUROSPEECH, pp167-170, 1989