Regular paper http://dx.doi.org/10.13067/JKIECS.2014.9.5.543

SNR 매핑을 이용한 환경적응 기반 음성인식

정용주*

Speech Recognition based on Environment Adaptation using SNR Mapping

Yong-Joo Chung*

요 약

다 모델 기반의 음성인식기는 음성인식에서 매우 성공적임이 알려져 있다. 그것은 다양한 신호-대-잡음비(SNR)와 잡음종류에 해당하는 다수의 HMM을 사용함으로서 선택된 음향모델이 인식잡음음성에 매우 근접한 일치성을 가질 수 있기 때문이다. 그러나 실제 사용시에 HMM의 개수가 제한됨에 따라서 음향모델의 불일치는 여전히 문제로 남아 있다. 본 논문에서는 인식잡음음성과 HMM 간의 SNR 불일치를 줄이고자 이들 간의 최적의 SNR 매핑 (mapping)을 실험적으로 결정하였다. 인식잡음음성으로 부터 추정된 SNR 값을 사용하는 대신 제안된 SNR 매핑을 사용함으로서 향상된 인식결과를 얻을 수 있었다. 다 모델 기반인식기에 제안된 방법을 적용하여 Aurora 2 데이터베이스에 대해서 인식 실험한 결과 기존의 MTR 이나 다 모델 기반 음성인식기에 비해서 6.3% 와 9.4%의 상대적 단어 오인식율 감소를 이룰 수 있었다.

ABSTRACT

Multiple-model based speech recognition framework (MMSR) has been known to be very successful in speech recognition. Since it uses multiple hidden Markov modes (HMMs) that corresponds to various noise types and signal-to-noise ratio (SNR) values, the selected acoustic model can have a close match with the test noisy speech. However, since the number of HMM sets is limited in practical use, the acoustic mismatch still remains as a problem. In this study, we experimentally determined the optimal SNR mapping between the test noisy speech and the HMM set to mitigate the mismatch between them. Improved performance was obtained by employing the SNR mapping instead of using the estimated SNR from the test noisy speech. When we applied the proposed method to the MMSR, the experimental results on the Aurora 2 database show that the relative word error rate reduction of 6.3% and 9.4% was achieved compared to a conventional MMSR and multi-condition training (MTR), respectively.

키워드

Speech Recognition, Noise Robustness, Hidden Markov Model, Signal to Noise Ratio
음성 인식, 잡음 강인성, 은닉 마르코프 모델, 신호대 잡음비

Ⅰ. 서 론

잡음환경에서 음성인식이 이루어지는 경우 상당한 인식율의 저하가 발생한다고 알려져 있다. 다양한 연구들이 잡음에 강인한 음성인식을 위해서 진행되어 왔는데 대표적인 방법으로는 잡음에 강인한 특징추출, 음질개선 그리고 음성특징 및 모델 파라미터 보상 등을 들 수 있다 [1-7].

* 교신저자(corresponding author) : 계명대학교 전자공학과 (yjjung@kmu.ac.kr)
접수일자 : 2014. 03. 05 심사(수정)일자 : 2014. 04. 21 게재확정일자 : 2014. 05. 15

한편, HMM을 잡음음성으로부터 직접적으로 훈련하는 방식은 또 다른 잡음음성인식 기법으로 자리 잡아 왔다. 이러한 접근방식은 잡음의 통계적 값이 훈련과 인식시에 큰 차이가 나지 않은 경우 가장 효과적이다. MTR 방식에서는 다양한 잡음환경의 잡음음성 신호들을 모아서 하나의 HMM set을 훈련시킨다[8]. 상당한 인식율의 향상이 MTR방식을 통해서 이루어졌지만 다수의 잡음 조건들을 결합하여 훈련잡음음성을 생성함으로서 음향 모델의 날카로움이 줄어든다는 단점이 있다.

MTR 방식의 단점을 보완하기 위해서 최근에는 MMSR 방식이 제안되어 성공적인 인식율 향상을 이루었다[9-10]. 여기서는 다양한 잡음 종류와 SNR 값별로 별도의 HMM set들이 훈련과정 중에 생성되며 이들 중에서 인식잡음음성과 가장 가까운 하나의 HMM set이 선택되어 인식에 최종적으로 사용된다.

MMSR 방식에서는 실제 음성인식이 실행되기 전에, 인식잡음음성과 가장 일치하는 기준 HMM set을 선택하기 위해서 인식잡음음성에 포함된 잡음의 종류와 SNR 값을 추정하는 과정이 필요하다. 이 과정에서의 오류는 인식율의 저하와 바로 연결되기 때문에 MMSR 방식의 성능은 이 과정에서의 에러를 최대한 줄임으로서 더욱 더 향상 될 수 있다.

기존의 MMSR에 관한 연구에서는 잡음의 종류가 정해지면 추정된 잡음인식음성의 SNR 값과 가장 가까운 기준 HMM이 선택되었다. 그러나 최근의 다른 연구결과들을 종합해 보면 추정된 SNR 값과 약간 상이한 SNR 에 해당하는 기준 HMM을 선택하는 것이 인식성능의 향상을 가져다주는 것으로 판단된다. 본 연구에서는 이러한 관점에서 MMSR 방식의 기준 HMM을 선택하는 과정에서 추정된 SNR값과 선택된 기준 HMM의 SNR 값 사이의 매핑을 통하여 보다 향상된 인식성능을 얻고자 한다. 이를 위해서 다양한 실험을 통해서 최적의 인식성능을 나타내는 SNR 매핑을 결정하고자 한다.

본 논문의 구성은 다음과 같다. MMSR 방식에 대한 개요를 2장에 소개하며 MMSR 방식에서 인식잡음음성과 선택된 HMM set 사이의 SNR 불일치에 대한 실험적 고찰을 3장에서 기술한다. 4장에서는 실험절차와 결과를 제시하며 5장에서 결론을 맺는다.

II. Multiple-model based speech recognition

MMSR 방식에서는 잡음의 종류와 SNR별로 각각의 HMM set이 훈련과정에서 만들어지며 인식시에는 인식잡음음성과 가장 유사한 기준 HMM set이 선택된다. 기준 HMM을 선택하기 위해서는 인식잡음음성의 SNR이 추정되어야 하며 잡음종류에 대한 분류가 필요하다. MMSR 방식의 구조는 그림 1에 나타나 있다.

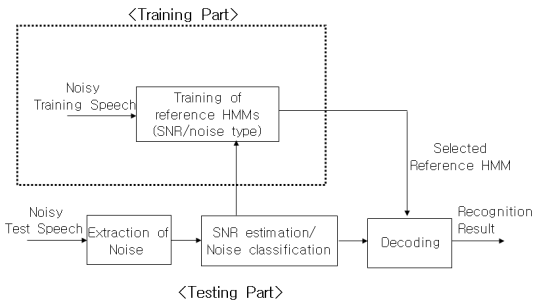

그림 1. 훈련부분과 인식부분으로 나누어진 MMSR 방식의 구조

Fig. 1 Architecture of MMSR method which is divided into training and testing parts

본 논문에서는 기준 HMM이 4가지 종류의 잡음(babble, car, subway, exhibition)에 대해서 2-dB 간격으로 구성되도록 하였다. 일반적으로 추정된 SNR 값에 가장 가까운 기준 HMM을 선택하는 것이 최고의 인식성능을 나타낸다고 알려져 있으나 본 연구에서는 주어진 인식잡음음성에 대해서 최적의 기준 HMM의 SNR을 실험적으로 결정하였다.

MMSR 방식에서 인식잡음음성의 SNR 추정을 위해서는 비교적 간단한 VAD (voice activity detector) 기반의 방식을 사용하였다[11]. VAD는 끝점 추출기와 비슷한 방식으로 동작하며 잡음음성으로 부터의 에너지 임계치를 사용하여 주어진 문장의 음성부분을 추정하게 된다. 추정된 음성부분으로부터 음성신호의 전력 σ_x^2을 구하고 비음성 구간으로 부터 잡음신호의 전력 σ_n^2을 추정하게 된다. 이를 통해서 주어진 문장의 SNR 값은 식(1)로 정의된다.

$$SNR = 10\log\frac{\sigma_x^2 - \sigma_n^2}{\sigma_n^2} \quad (1)$$

III. MMSR 방식에서 SNR 영향

이번 장에서는 MMSR 방식에서 인식잡음음성과 훈련잡음음성간의 SNR 불일치가 인식율에 미치는 영향에 대해서 논의하고자 한다. 이와 같은 논의를 통해서 최고의 인식성능을 나타내는 인식잡음음성과 훈련잡음음성간의 SNR 매핑을 찾을 수 있을 것이다.

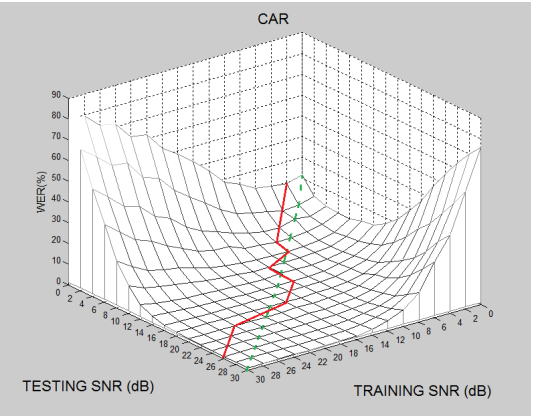

그림 2. 훈련잡음음성과 인식잡음음성간의 SNR 불일치에 따른 단어오인식율의 변화를 보여주는 단어오인식율 평면
(푸른 점선은 동일한 SNR 지점을 나타내며, 붉은 실선은 최저의 단어오인식율을 나타내는 지점을 표시한다)

Fig. 2 The word error rate (WER) surface showing the variation of WERs with the SNR mismatch between the training and test noisy speech.
(The dotted blue line represents the points of equal SNRs and the solid red line marks the points of minimum WERs)

MMSR 방식에서의 SNR 불일치에 대한 영향을 설명하기 위해서 그림 2에서는 Aurora 2 데이터베이스의 자동차(car) 잡음신호를 이용한 경우의 단어오인식율(WER) 평면을 보여준다. 그림 2로 부터 최소의 단어오인식율은 반드시 인식잡음음성과 훈련잡음음성의 SNR 값이 일치하는 경우에 발생하지는 않는다는 것을 보여주며 단순히 SNR 값을 일치시키는 것으로는 충분하지 않음을 알 수 있다. 그러나 이러한 문제는 SNR 값이 높은 영역에서는 단어오인식율 평면이 매우 평편하기 때문에 그리 큰 문제를 야기 하지 않는다. 반면에 SNR 값이 낮은 영역에서는 단어오인식율 평면에는 가파른 경사가 존재하고 이 영역부근에서는 약간의 SNR 값 움직임만으로도 단어오인식율의 변화가 매우 크게 되므로 최적의 인식성능을 위해서는 인식잡음음성과 훈련잡음음성간의 효과적인 SNR 매핑이 중요함을 알 수 있다.

그림 3에는 인식잡음음성의 SNR 값이 각각 0, 2, 4 dB 인 경우 선택된 기준 HMM의 SNR 값에 따른 단어 오인식율의 변화를 보여준다. 예를 들어, 인식잡음음성의 SNR 값이 0 dB 인 경우, 선택된 기준 HMM이 0 dB 의 잡음음성으로 훈련되었다면 단어오인식율은 43.65% 가 된다. 하지만, 기준 HMM이 6 dB의 잡음음성으로 훈련된 경우에는 단어 오인식율은 32.07%로 감소된다. 이 예는 인식잡음음성과 기준 HMM의 SNR 값이 동일할 경우 최고의 인식율을 나타낼 것이라고 판단한 기존의 생각과 대조되며 인식잡음음성과 기준 HMM의 SNR 간의 매핑을 통하여 인식성능을 향상시킬 필요가 있음을 확인시켜 준다.

그림 2에 나타난 단어오인식율 평면을 이용하면 인식잡음음성의 SNR 값이 알려진 경우, 최고의 성능을 나타내는 기준 HMM의 SNR 값을 결정할 수 있다. 이 결과가 표 1에 나타나 있다. 기대한 대로, 인식잡음음성과 기준 HMM의 SNR 간에는 약간이 차이가 있음을 알 수 있는데, 특히 낮은 SNR 영역에서는 실제 추정된 SNR 값 보다 약간 더 높은 SNR의 기준 HMM을 선택하는 것이 인식율 향상을 가져다주는 것을 알 수 있다. 이러한 현상이 발생하는 이유는 여러 가지가 있을 수 있지만, 낮은 SNR의 훈련음성은 음성들 간의 분별력을 낮추기 때문에 비록 인식잡음음성의 SNR값이 낮다고 하더라도 다소 높은 SNR의 기준HMM을 선택하는 것이 유리 할 것이라는 연구결과가 발표되기도 하였다[10].

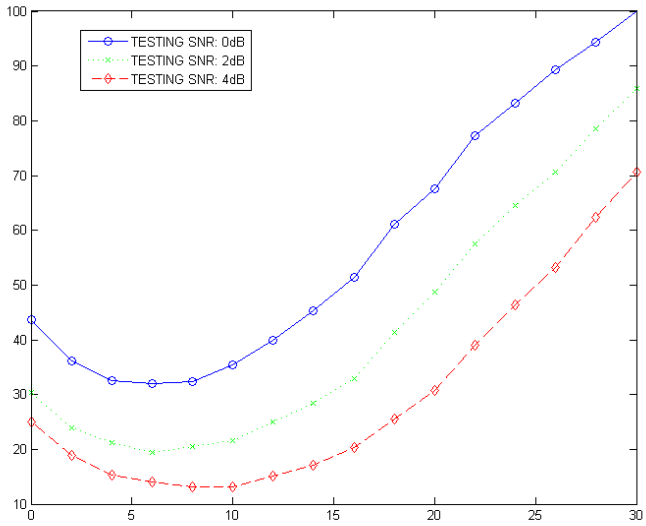

그림 3. 기준 HMM의 SNR 값에 따른 인식율 변화(인식잡음음성의 SNR 값이 각각 0, 2, 4 dB 인 경우)
Fig. 3 The variation of WERs(%) as the SNR of the reference HMM changes(The SNR of test noisy speech is set at 0, 2, 4, dB, respectively)

표 1. 최소의 단어 오인식율을 보이는 기준 HMM의 SNR 값들
Table 1. SNR values of the reference HMM showing the lowest WER

SNR of test speech	SNR of reference HMM showing lowest WER			
	Babble	Subway	Car	Exhibition
0	6	4	2	2
2	6	6	4	4
4	8	6	6	6
6	10	8	8	8
8	10	8	8	10
10	12	12	12	10
12	16	10	14	12
14	18	14	14	16
16	18	16	20	18
18	18	18	18	18
20	20	22	20	20
22	26	22	26	26
24	28	22	28	28
26	28	22	30	28
28	28	22	30	30
30	30	24	30	30

Ⅳ. 실험결과

본 연구를 위해서는 Aurora 2 데이터베이스를 활용하였으며 음향모델 훈련을 위해서는 CLEAN 과 MTR의 두 가지 음성데이터를 이용하였다. CLEAN 데이터는 잡음이 전혀 포함되지 않은 깨끗한 음성을 의미하며 MTR 데이터는 깨끗한 음성과 0~20 dB 사이의 다양한 종류(subway, car, exhibition, babble)의 잡음음성으로 구성되어 있다.

인식실험은 Set A, Set B 그리고 Set C의 3가지 서로 다른 테스트 set에 대해서 이루어졌다. Set A와 Set B는 부가 잡음만으로 오염되었고 Set C는 채널 잡음과 부가잡음 모두에 의해서 오염되었다.

음성특징 방법으로는 2 가지의 많이 알려진 방식을 사용하였다. 첫 번째 방식은 FE 특징으로서[12] 이는 0번째 계수를 제외한 12차의 멜켑스트럼 계수(Mel-frequency cepstral coefficient)와 로그-에너지(log-energy)를 함께 사용한 13차의 기본 특징벡터를 가지고 있다. 기본 특징벡터의 delta와 acceleration 계수가 추가된 39차의 특징벡터가 최종적으로 인식실험

에 사용되었다. 두 번째 방식은 AFE 특징으로서 FE 방식에 비해서 잡음에 강인한 구조로 이루어져 있으며 잡음환경에서 매우 좋은 인식성능을 나타내는 것으로 알려져 있다[13].

표 2. MMSR 방식과 기존의 방법들 간의 성능비교(FE 특징사용)
Table 2. Performance comparison between MMSR method and the conventional methods(using FE feature)

Method	WER(%)			
	Set A	Set B	Set C	Ave.
CLEAN	38.66	44.25	33.86	39.94
VTS	28.23	29.31	24.95	28.00
PMC	20.70	18.82	21.98	20.20
MTR	12.23	13.75	16.42	13.68
MMSR	8.92	16.64	15.09	13.24

표 2에는 MMSR 방식의 단어 오인식율이 기존의 다른 방식과 함께 나타나 있다. 표 2로부터 MMSR 방식은 PMC, CLEAN 그리고 VTS 방식에 비해서 월등히 성능이 뛰어남을 알 수 있으나 MTR 방식에 비해서는 약간의 인식율 상승만을 보임을 알 수 있다. MTR 방식은 MMSR 방식에 비해서 Set B에 대해서 잡음에 대한 강인성이 뛰어난 것을 볼 수 있으며 Set A 와 Set B에 대해서는 MMSR 방식이 MTR 방식에 비해서 오히려 성능이 더 뛰어난 것을 알 수 있다. 그러나 평균적으로는 두 방식의 차이는 거의 미미한 것으로 나타난다 (13.24% vs. 13.68%).

그림 4에는 MMSR 방식에 표 1에서 얻어진 SNR 매핑을 적용한 경우의 인식성능을 나타내었다. 비교를 위하여 매핑을 적용하지 않은 MMSR 방식과 MTR 방식의 인식성능도 함께 표시하였다. 그림에서 MMSR 방식은 MTR 방식에 비해서 평균 3.2% 상대적 단어오인식율 향상을 보임을 알 수 있으며 SNR 매핑을 추가적으로 적용함(SNR-MMSR)으로서 더욱 향상된 성능을 보임을 알 수 있다. SNR-MMSR 방식은 평균 12.40%의 단어 오인식율을 보임으로서 MMSR과 MTR 각각에 대해서 6.3% 와 9.4%의 상대적 단어오인식율 감소를 가져오는 것을 알 수 있다. 그림 4에서 보듯이 SNR-MMSR는 Set A, B, C 모두에 대해서 기존의 MMSR에 비해서 향상된 인식성능을 보이는데 이는 표 1에서 실험적으로 결정된 SNR 매핑이 잡음의 종류에 관계없이 적합함을 말해주는 것이다. 비록 본 논문에서 사용된 SNR 매핑이 알려진 종류의 잡음신호에 대해서 결정되었지만 Set B의 새로운 잡음이나 Set C의 채널잡음에 대해서도 유효함이 확인되었다.

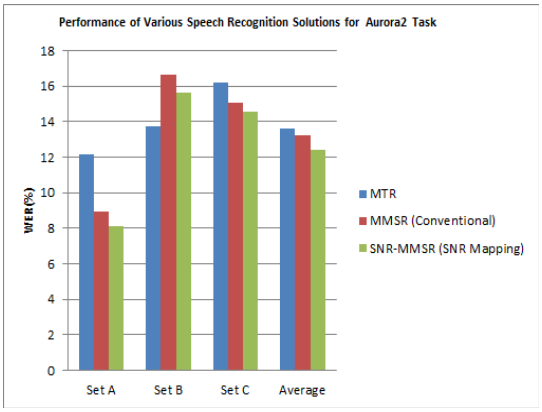

그림 4. SNR-MMSR 방식과 MTR, 기존의 MMSR 방식의 성능 비교(FE 특징사용)
Fig. 4 Performance comparison of SNR-MMSR with MTR and conventional MMSR(using FE feature)

표 3. MMSR 방식과 기존의 방법들 간의 성능비교(AFE 특징사용)
Table 3. Performance comparison between MMSR method and the conventional methods(using AFE feature)

	Set A	Set B	Set C	Ave.
MTR	7.70	8.23	9.26	8.22
MMSR	7.59	10.66	8.57	9.01
SNR-MMSR	6.78	9.56	8.17	8.17

표 3에는 제안된 SNR-MMSR 방식의 성능이 AFE 특징을 사용한 경우에, MTR 방식과 기존의 MMSR 방식과의 성능비교가 나타나 있다. 기존의 MMSR 방식에 비해서 SNR-MMSR 방식의 성능이 월등함이 뛰어난 것이 나타나 있으며 이러한 향상은 Set A, B, C 모두에서 나타남을 볼 수 있다. 이는 제안된 SNR 매핑이 특징의 종류와 상관없이 효과적임

을 나타낸다. AFE 특징을 사용한 경우, 기존의 MMSR 방식은 MTR 방식에 비해서 저조한 성능을 보였지만, 제안된 SNR-MMSR 방식은 향상된 결과를 보여줌으로서 제안된 방식이 잡음음성인식에 있어서 매우 효과적임이 확인되었다.

V. 결 론

잡음환경에 최적화된 MMSR 방식은 기존의 MTR 방식에 비해서 향상된 인식성능을 보임이 알려져 있다. 그러나 훈련잡음음성과 인식잡음음성간의 SNR 불일치는 MMSR 방식이 MTR 방식에 비해서 월등히 나은 성능을 보이기가 어렵게 만들었다. 본 연구에서는 이러한 불일치 문제를 인식잡음음성과 훈련잡음음성간의 SNR 매핑을 통해서 상당히 극복할 수 있었으며 매우 향상된 인식성능을 보일 수 있었다. FE 특징을 사용한 경우 제안된 방식을 통하여 기존의 MMSR 방식과 MTR 방식에 비해서 각각 6.3% 와 9.4%의 상대적 단어오인식율 감소를 이룰 수 있었으며 AFE 특징을 사용한 경우에도 기존의 MMSR에 비해서 9.3%의 상대적 단어오인식율 감소를 이룰 수 있었다.

References

[1] S. Ball, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.* vol. 27, no. 2, 1979, pp. 113-120.

[2] M. J. F. Gales, "Model based techniques for noise-robust speech recognition," Ph.D. Dissertation, *University of Cambridge*, 1996.

[3] P. J. Moreno, "Speech Recognition in noisy environments," Ph.D. Dissertation, *Carnegie Mellon University*, 1996.

[4] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. on Signal Processing*, vol. 39, no. 4, 1991, pp. 795-805.

[5] J. Choi, "Speech and noise recognition system by neural network," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 5, no. 4, 2010, pp. 357-362.

[6] C. Lee and D. kim, "Adaptive noise reduction of speech using wavelet transform," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 4, no. 3, 2009, pp. 190-196.

[7] J.-S. Choi, "Noise reduction algorithm in speech by Wiener filter," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 8, no. 9, 2013, pp. 1293-1298.

[8] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," In P*roc. the Int. Conf. on Spoken Language Processing,* Bejing, China, 2000, pp. 18-20.

[9] H. Xu, Z. H. Tan, P. Dalsgaard, and B. Lindberg, "Robust speech recognition on noise and SNR classification – a multiple-model framework," In *Proc. INTERSPEECH,* Lisboa, Portugal, 2005, pp. 977-980.

[10] H. Xu, Z. H. Tan, P. Dalsgaard, and B Lindberg, "Noise condition dependent training based on noise classification and SNR estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 8, 2007, pp. 2431-2443.

[11] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Process.* vol. 29, no. 4, 1981, pp. 777-785.

저자 소개

정용주(Yong-Joo Chung)

1988년 서울대학교 전자공학과 졸업(공학사)

1990년 한국과학기술원 전기및전자공학과 졸업(공학석사)

1995년 한국과학기술원 전기및전자공학과 졸업(공학박사)

1999년~계명대학교 전자공학과 교수

※ 관심분야 : 음성인식, 멀티미디어신호처리