잡음 첨가된 화자 모델 구성에 의한 잡음 환경의 효과적인 화자확인

안성주*, 강선미**, 고한석*
*고려대학교 전자공학과, **서경대학교 컴퓨터과학과
Email : {sjahn, hsko}@ispl.korea.ac.kr, smkang@bukak.seokyeong.ac.kr

Efficient Speaker Verification in Noise Environment with Noise-added Speaker Model Composition

Sungjoo Ahn*, Sunmee Kang**, Hanseok Ko*
*Department of Electronics Engineering, Korea University
**Department of Computer Science, Seokyeong University

요 약

본 논문에서는 다수의 화자 모델을 구성함으로써 잡음에 강인한 화자확인 방법을 제안한다. Non-stationary한 잡음을 가진 입력음성의 SNR을 측정하는 것은 어렵기 때문에, 각 화자에 대해 잡음이 없을 때의 화자모델에 여러 SNR에 대한 잡음 모델을 결합시킴으로써 여러 개의 잡음 첨가된 화자 모델을 구성한다. 그리고, 화자확인에서는 이렇게 구한 각 모델에 대한 입력 음성의 likelihood를 각각 구해 그중 가장 큰 likelihood만을 선택한다. 이 값을 이용하여 화자확인을 수행한다.

실험 결과, 제안한 방법은 입력음성의 SNR을 모르는 잡음환경에서 일반적으로 하나의 모델을 사용하는 것보다 훨씬 좋은 성능을 보였다.

1. 서론

실생활에 화자확인 시스템을 적용할 때, 많은 시스템이 제한된 학습환경에서 실험한 결과와 비교하여 그 성능이 급격하게 떨어진다[1]. 이러한 성능 감소는 주로 학습환경과 실험환경의 불일치 때문에 일어난다. 그 중에서도 잡음이 가장 큰 영향을 미친다. 실제 잡음 환경에 강인한 화자확인을 위해서, 다양한 연구가 수행되었다. 이것을 분류해 보면, 첫째, 음성으로부터 원천적으로 잡음에 강한 특징을 추출하는 방법, 둘째, 음성 강화 방법, 셋째, 모델에 기초한 잡음 보상 방법을 들 수 있다[5]. 그 중 모델에 기초한 잡음 보상 방법에 대해서도 많은 연구와 알고리즘이 제안되었다[2][3][5].

모델에 기초한 잡음 보상 방법 중, PMC(Parallel Model Combination) 방법은 첨가적인 잡음을 처리하는데 매우 효과적인, 가장 잘 알려진 접근 방법이다[3]. 잡음이 있는 상황의 화자모델을 구성하기 위해서는 다양한 잡음이 섞인 음성 데이터를 수집해야하는 어려움이 있는데, PMC 방법을 이용하면 적은 양의 잡음 데이터로 잡음 HMM을 학습시켜 잡음이 없는 음성 모델을 수정하여 실제 적용하려는 곳의 잡음에 맞게 모델을 바꿔준다. 그러나 PMC 방법은 잡음신호의 SNR을 알아야 하며, 만약 잡음이 non-stationary하면 적용할 수 없는 문제점이 있다.

따라서 본 논문에서는 이러한 알려지지 않은 잡음 신호의 SNR을 해결하기 위해 PMC에 기초한 화자 모델 구성 방법을 제안한다.

본 논문의 구성은 2장에서 PMC 방법에 대해 간단히 소개하고, 3장에서 새로운 모델 구성 방법을 제안한다. 그리고 4장에서 실험결과에 대해 설명하고 마지막으로 5장에서 결론을 맺는다.

2. Parallel Model Combination(PMC)

PMC 방법은 잡음이 없을 때의 음성 모델을 잡음 모델과 결합함으로써 잡음이 있는 음성 모델을 만드는 방법이다[3]. 즉 PMC 방법은 추정된 잡음 모델 파라미터를 이용하여 잡음이 없는 음성 모델의 파라미터인 평균과 공분산을 보정하는 방법이다. 음성과 잡음 신호는 linear-spectrum 영역에서 additive하다고 가정한다. 그러나 음성과 잡음 모델은 cepstral 영역에서 정의되어 있기 때문에 linear spectrum 영역으로 바꿔서 SNR에 따라 더해 주어야 한다.

PMC 방법에 대한 전체적인 과정은 다음과 같다.

과정 1 : cepstral 영역의 모델 파라미터(평균,공분산)를 inverse DCT transformation을 이용하여 log-spectral 영역으로 바꿔준다. 그 관계식은 다음과 같다.

$$\mu^l = C^{-1}\mu^c \quad (1)$$

$$\Sigma^l = C^{-1}\Sigma^c(C^{-1})^T \quad (2)$$

여기서 C는 DCT를 나타내는 matrix이다.

과정 2 : log-spectral 영역의 모델 파라미터를 exponential transformation을 통해 linear-spectral 영역으로 변환시킨다.

$$\mu_i = \exp(\mu_i^l + \Sigma_{ii}^l/2) \quad (3)$$

$$\Sigma_{ij} = \mu_i\mu_j[\exp(\Sigma_{ij}^l) - 1] \quad (4)$$

과정 3 : linear-spectral 영역에서 구하려고 하는 잡음 상황에 맞는 음성의 모델 파라미터인 평균과 공분산을 계산한다.

$$\hat{\mu} = \mu + g\mu_n \quad (5)$$

$$\hat{\Sigma} = \Sigma + g^2\Sigma_n \quad (6)$$

과정 4 : 과정 3에서 구한 linear-spectral 영역의 모델 파라미터를 logarithm transformation을 통해 log-spectral 영역으로 변환시킨다.

$$\hat{\mu}_i^l = \log(\hat{\mu}_i) - \frac{1}{2}\log(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1) \quad (7)$$

$$\hat{\Sigma}_{ij}^l = \log(\frac{\hat{\Sigma}_{ij}}{\hat{\mu}_i\,\hat{\mu}_j} + 1) \quad (8)$$

과정 5 : 마지막으로 log-spectral 영역의 모델 파라미터를 DCT transformation을 이용하여 cepstral 영역으로 변환한다.

$$\hat{\mu}^c = C\,\hat{\mu}^l \quad (9)$$

$$\hat{\Sigma}^c = C\,\hat{\Sigma}^l\,C^T \quad (10)$$

3. 제안한 방법

PMC를 이용하여 모델을 구성하기 위해서는 입력 음성의 SNR을 알 필요가 있다. 만약 잡음이 computer room에서 녹음된 것처럼 power 영역에서 stationary 하다면, 입력 신호의 SNR을 입력 신호의 power와 잡음의 power를 이용하여 구할 수 있다. 그러나 잡음이 전화 부스나 움직이는 차에서 녹음된 것처럼 non-stationary하다면 잡음의 power는 시간적으로 배경 소리나 환경에 의해 바뀌게 된다. 이 경우 SNR을 구하는 것은 어렵다.

따라서 이러한 알려지지 않은 SNR의 문제를 해결하기 위해, 본 논문에서는 PMC에 기초한 다수의 화자 모델을 구성하는 방법을 제안한다.

제안한 방법의 학습 과정은 그림 1과 같다. 잡음이 없는 환경에서 녹음한 음성을 이용하여 각 화자에 대한 모델을 만든다. 그리고 실제 적용할 환경에서 녹음한 잡음 신호를 이용하여 잡음 모델을 만든다. 이렇게 만든 모델을 앞에서 설명한 PMC 방법에 기초하여 각 화자당 여러 가지 SNR에 해당하는 잡음 첨가된 화자 모델을 만든다. 이 모델을 만들 때, 2장에서 설명한 PMC 방법에 기초하여 다음 식을 이용하여 여러 SNR에 대한 화자 모델을 만든다.

$$\hat{\mu} = \mu + g(SNR)\mu_n \quad (11)$$

$$\hat{\Sigma} = \Sigma + g(SNR)^2\,\Sigma_n \quad (12)$$

where $g(SNR) = \lambda\sqrt{\frac{P_s}{P_n}10^{-\frac{SNR}{10}}}$, $0 < \lambda < 1$

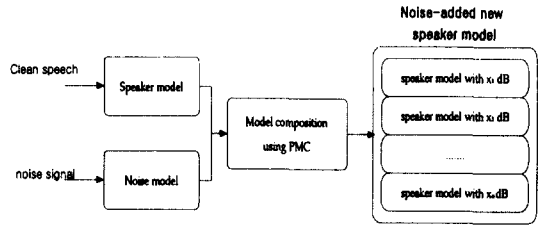

그림 1. 화자 모델 학습 과정

화자확인 과정에서는, 입력 음성이 주어지면 여러 SNR에 대한 잡음 첨가된 화자 모델에 대해, 각각의 likelihood 값이 계산된다. 그리고 이렇게 구한 여러 SNR에 대한 likelihood 값 중에서 가장 큰 likelihood 값을 선택한다. 이 가장 큰 likelihood 값을 threshold와 비교하여 화자확인을 한다. 그 전체적인 과정은 그림 2와 같다.

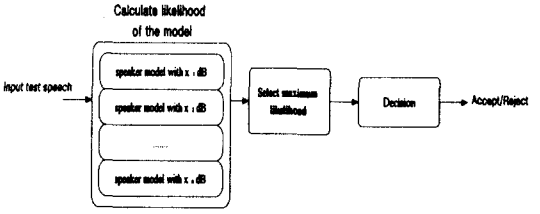

그림 2. 화자확인 과정

4. 실험 및 결과

4.1 실험 환경

본 연구에서는 문맥 종속 화자확인 시스템을 고려한 한국어 단어 DB를 구축하였다. 실험에 사용된 음성데이터는 9명이 발음한 한국어 단어로 구성되었다. 음성은 전화상에서 녹음되었고, 8kHz로 샘플링 되었다. 음성 특징으로는 12차의 MFCC를 사용했다. 그리고 화자모델은 5 state, 8mixture의 CDHMM(Continuous Density Hidden Markov Model)을 사용했고, 잡음 모델은 1 state, 1 mixture의 CDHMM을 사용했다. 실험에 사용된 잡음은 AWGN(Additive White Gaussian Noise)와 전화상의 잡음이다.

4.2 기초 실험

PMC 방법의 성능을 측정하기 위해 제안된 방법을 실험하기 전에 기초 실험이 이루어졌다. 이때 사용된 test data는 깨끗한 음성신호와 잡음이 첨가된 음성신호다. 잡음이 첨가된 음성 신호는 잡음을 깨끗한 음성에 다음 식 (13)과 같이 더함으로써 얻었다. 5, 10, 20, 30 dB의 SNR을 가지는 잡음이 첨가된 음성신호가 실험에 사용되었다. 각 SNR에 대한 test data는 2430(30*9*9)개의 음성

데이터로 구성되었다.

$$x^{test}(t_i) = x^s(t_i) + k(SNR)x^n(t_i), \quad 1 \le i \le N \qquad (13)$$

where $$k(SNR) = \sqrt{\frac{P_s}{P_n} 10^{-\frac{SNR}{10}}}$$

실험 결과는 다음 표 1과 표 2와 같다.

표 1. AWGN의 화자확인 error률(%)

Test data / Trained Model	Clean speech	30dB noisy speech	20dB noisy speech	10dB noisy speech	5dB noisy speech
Clean speech HMM	1.56	1.93	2.31	7.38	9.51
PMC with each dB noise	-	1.52	2.35	6.00	8.21

표 2. 전화잡음의 화자확인 error률(%)

Test data / Trained Model	Clean speech	30dB noisy speech	20dB noisy speech	10dB noisy speech	5dB noisy speech
Clean speech HMM	1.56	8.19	8.31	8.60	9.14
PMC with each dB noise	-	6.09	6.13	6.30	7.08

위의 결과를 보면 clean speech HMM의 화자확인 error률은 AWGN과 전화 잡음 두 경우 모두 잡음의 SNR이 작아질수록 증가함을 볼 수 있다. 그러나 PMC 방법을 사용함으로써 그 error률이 감소하는 것을 볼수 있다. 따라서 이러한 결과는 PMC 방법이 첨가적인 잡음 환경에 효과적이라는 것을 보여준다.

4.3 제안된 방법의 실험

본 실험에서, 제안한 잡음 첨가된 화자 모델을 만들기 위해 3장에서 설명한 것처럼 PMC 방법에 기초하여 식 (11)과 (12)를 사용했다. 또 실험에서 사용된 각 SNR에 대한 모델은 4.2절에서 사용된 PMC 모델과 같다. 4.2절에서처럼 실험에 사용한 test data로는 5, 10, 15, 20, 25, 30 dB의 잡음 첨가된 신호와 깨끗한 음성신호를 사용했다. 전체 test data는 9792(120*9*9)개의 음성 데이터로 구성되었다. 실험 결과는 표 3과 같다.

표 3. 제안한 방법과의 화자확인 error률(%) 비교

Trained Model / Test data	Proposed method	Clean speech	30dB noisy speech	20dB noisy speech	10dB noisy speech	5dB noisy speech
AWGN	6.00	6.58	7.12	8.63	17.95	23.62
Telephone noise	5.45	6.52	8.52	9.41	10.88	12.57

표 3의 결과로부터 제안한 각 화자당 여러 모델을 사용하는 방법이 각 화자당 하나의 모델을 사용하는 것보다 더 좋은 성능을 나타낸다는 것을 알 수 있었다. 그 이유는 test data가 어떤 SNR을 가지고 있는지 알 수 없는 상황에서, 어느 한 SNR에만 적당한 하나의 모델을 사용하는 것보다 여러 SNR에 대해 각각의 모델을 사용하면 어떠한 잡음 신호가 들어와도 그 잡음 신호에 적당한 모델이 선택되어 성능에는 큰 영향을 미치지 않기 때문이다. 그러나 제안한 방법이 하나의 모델을 사용하는 방법보다 매우 큰 성능 향상을 보여주지는 못하고 있다. 그 이유는 여러 SNR에 대한 각각의 화자 모델의 likelihood 값 중에서 최대값을 선택하여 화자확인을 수행하는데, 그 test data에 적당한 모델의 likelihood 값 대신 다른 모델의 likelihood 값이 선택 될 수 있기 때문이다. 여기에서 error가 발생하여 성능이 감소했을 것이다.

그러나 실험 결과로부터, 제안한 방법이 입력 음성의 SNR을 모르는 상황에도 효과적이라는 것을 알 수 있었다.

5. 결론

본 논문에서는 모델 구성 방법을 이용한 잡음에 강한 화자확인 방법을 제안하고 있다. 실험 결과, 제안한 방법이 각 화자당 하나의 모델을 사용하는 것보다 좋은 성능을 나타낸다는 것을 보여준다. 또한 제안한 방법은 잡음에 대해 모르는 상황에서 화자확인에 효과적이다.

앞으로 좀 더 성능이 향상된 화자확인을 위해 최대 likelihood 값을 선택하는 방법 외에 likelihood 값의 선택과 threshold의 결정에 대한 연구가 필요하다. 또한 각 화자당 모델의 수가 증가함에 따라 생기는 memory와 계산량을 줄이는 방법에 대한 연구가 필요하다.

감사의 글

본 논문은 한국과학기술 평가원 '98핵심 S/W 기술개발사업'(과제번호 98-NS-01-08-A-25)의 지원으로 연구되었음.

참고문헌

[1] G. R. Doddington, "Speaker Recognition - Identifying People by their Voice," Proc. IEEE, Vol. 73, pp. 1651-1664, Nov. 1985.

[2] F. Martin, K. Shikano, Y. Minami, "Recognition of noisy speech by composition of hidden Markov models, Proc. Eurospeech-93, pp. 1031-1034, Berlin, 1993.

[3] Gales, M. J. F. & S.J. "HMM recognition in noise using parallel model combination," Proc. Eurospeech93, pp. 837-840, 1993.

[4] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE trans. Speech and Audio Processing, Vol. 2, pp.639-643, 1994.

[5] R. C. Rose, E.M. Hofstetter and D. A. Reynolds, "Iterated models of signal and background with application to speaker identification in noise," IEEE Trans. Speech and Audio Processing, pp. 245-257, 1994.