MFCC 특징 파라미터를 이용한 인식 알고리즘

최재승*
*신라대학교

Recognition Algorithm using MFCC Feature Parameter

Jae-seung Choi*
*Silla University
E-mail : jschoi@silla.ac.kr

요 약

배경잡음은 음성신호의 특징을 왜곡하기 때문에 음성인식 시스템의 인식율 향상의 방해요소가 된다. 따라서 본 논문에서는 배경잡음이 존재하는 환경에서의 음성인식을 실시하기 위해서, 신경회로망과 Mel 주파수 켑스트럼 계수를 사용하여 연속음성 식별 알고리즘을 제안한다. 본 논문의 실험에서는 본 알고리즘을 사용하여 배경잡음이 섞인 음성신호에 대하여 음성인식의 식별율 개선을 실현할 수 있도록 연구를 진행하며, 본 알고리즘이 유효하다는 것을 실험을 통하여 명백히 한다.

키워드

음성인식, 특징파리미터, 멜 주파수 켑스트럼, 배경잡음, 퍼셉트론 네트워크

Ⅰ. 서 론

최근에는 음성인식이 인간과 기계 간의 일반적인 인터페이스로 실용화되고 있다. 이러한 음성인식의 인터페이스를 실용화하기 위해서는 음성이 발성되는 환경이 음성신호뿐만 아니라 배경잡음도 포함된 환경도 고려해야하기 때문에, 잡음에 강건한 음성인식 시스템의 연구개발을 진행할 필요가 있다[1-4].

신경회로망(Neural Network, NN)[5]의 네트워크에 관한 연구가 음성인식의 연구 분야에서도 활발하게 진행되고 있다. 최근에는 음성인식의 연구에 대 어휘에 의한 연속음성의 대응이 필요하기 때문에 연속음성 인식의 연구도 진행되고 있다. 이러한 연구로는 신경회로망을 이용한 음성의 음소검색에 관한 연구, 시간지연신경회로망(Time Delay Neural Network, TDNN) 등을 이용한 음성인식의 연구가 보고되고 있다[6].

잡음이 존재하는 환경에서의 음성인식을 실시하기 위해서 본 논문에서는 신경회로망과 Mel 주파수 켑스트럼 계수(Mel Frequency Cepstral Coefficient, MFCC)[7, 8]를 사용하여 연속음성 식별을 실시하는 인식 알고리즘을 제안한다. 본 논문의 실험에서는 이러한 알고리즘을 음성신호에 중첩된 배경잡음을 억제하기 위하여 음성식별율의 개선이 가능하도록 실험을 진행한다.

Ⅱ. 제안한 인식 알고리즘 및 실험

본 논문에서는 MFCC의 특징벡터를 사용한다. 본 논문에서는 14차의 MFCC 특징벡터를 사용한

다. 또한 본 논문에서는 음성인식을 위하여 3층으로 구성된 다층 퍼셉트론 네트워크를 사용한다. 본 논문의 음성인식에 사용하는 기본적인 알고리즘은 다층 퍼셉트론에 의한 신경회로망의 네트워크이며, 본 네트워크는 각 특징량 벡터에 따라 화자가 다른 입력신호를 다른 계층으로 분류한다.

본 실험에 사용한 음성신호는 8 kHz로 표본화된 일본인 성인 남성화자에 의한 음성 데이터베이스를 사용하였다. 잡음 데이터로는 백색잡음을 사용하여 평가하였다. 본 실험에서는 각 연속음성의 문장에 백색잡음을 부가하며 한 프레임을 256 샘플로 한다.

본 실험에서는 본 알고리즘을 사용하여 14차의 MFCC 켑스트럼 계수터를 사용하여 식별률 실험을 실시하였다. 원 음성을 사용한 경우에 평균 식별율은 99.60%로 나타났으며, 백색잡음이 중첩된 음성의 경우에는 평균 식별률이 99.06%의 결과를 나타냈다. 따라서 본 논문에서 제안한 화자식별 알고리즘의 성능이 배경잡음 하에서 본 논문에서 제안한 알고리즘이 유효하다는 것을 실험 결과로부터 확인할 수 있었다.

Ⅲ. 결 론

본 논문에서는 다층 퍼셉트론 신경회로망에 MFCC 켑스트럼 계수를 입력함으로써 연속어 음성식별을 실시하는 화자식별 모델를 구축하는 알고리즘을 제안하였다. 본 실험에서는 음성신호에 중첩되는 잡음을 억제하기 위해서 배경잡음 하에서 음성식별율의 개선이 가능하도록 하는 연구를 실시하였다.

본 논문의 실험결과로부터, 백색잡음이 부가된 음성신호의 경우에 대해서도 평균 식별률이 양호한 결과를 나타냈다. 따라서 본 논문의 식별 실험에서는 제안한 알고리즘이 배경잡음 하에서 화자식별률의 개선에 유효하다는 것을 알 수 있었다.

참고문헌

[1] H. Wang and F. Itakura, “An implementation of multi-microphone dereverberation approach as a preprocessor to the word recognition system,” J. Acoust. Soc. Jpn. (E) 13, pp. 285-293, 1992.

[2] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 27, No. 2, pp. 113-120, 1979.

[3] G. Huang and M. J. Er, “A novel neural-based pronunciation modeling method for robust speech recognition,” *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 517-522, Dec. 2011.

[4] J. S. Choi, “Text-dependent Speaker Recognition using Characteristic Vectors in Speech Signal and Normalized Recognition Method,” The Journal of Korean Institute of Information Technology, Vol. 10, No. 5, pp. 61-66, May 2012.

[5] D. E. Rumelhart, G. E. Hinton and R. J. Williams, “Learning representations by back-propagation errors”, *Nature*, Vol. 323, No. 9, pp. 533-536, October 1986.

[6] D. Snyder, D. G. Romero and D. Povey, “Time delay deep neural network-based universal background models for speaker recognition,” *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 92-97, Dec. 2015.

[7] B. L. Sturm, M. Morvidone and L. Daudet, “Musical Instrument Identification Using Multiscale Mel-frequency Cepstral Coefficients,” European Signal Processing Conference, pp. 477-481, Aug. 2010.

[8] S. Stevens, J. Volkman and E. B. Newman, “A Scale for the Measurement of the Psychological Magnitude Pitch,” *Journal of the Acoustical Society of America*, Vol. 8, No. 3, pp. 185-190, Jan. 1937.