Journal of Convergence for Information Technology
Vol. 11. No. 6, pp. 1-6, 2021
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2021.11.06.001

음성 인식을 위한 개선된 평균 예측 LMS 필터를 이용한 DNN 기반의 강인한 음성 특징 추출 및 신호 잡음 제거 기법

오상엽
가천대학교 컴퓨터공학과 교수

DNN based Robust Speech Feature Extraction and Signal Noise Removal Method Using Improved Average Prediction LMS Filter for Speech Recognition

SangYeob Oh
Professor, Division of Computer Engineering, Gachon University

요 약 음성 인식 분야에서 DNN이 적용됨에 따라 음성 인식의 이용이 증대되고 있으나 기존의 GMM 보다 병렬 훈련에 대한 계산의 양이 많아야 되며, 데이터의 양이 적으면 오버피팅이 발생한다. 이를 해결하기 위해 데이터의 양이 작은 경우에도 강인한 음성 특징 추출과 음성 신호 잡음 제거에 효율적인 방안을 제시한다. 음성 특징 추출은 음성에 대한 프레임 에너지의 차이와 음성 신호에 영향을 받는 영 교차율과 레벨 교차율을 적용하여 음성 에너지의 효율적 추출을 한다. 또한, 잡음 제거를 위해 음성 신호에 대한 검출에서 음성의 고유 특성을 유지하면서 음성 정보 손상이 적은 평균 예측 LMS 필터를 개선하여 음성 신호의 잡음을 제거하여 데이터양이 적은 경우의 문제를 해결한다. 개선된 LMS 필터는 입력 신호에 대한 활성 파라미터 임계치를 조정하여 입력된 음성 신호에 대한 잡음을 처리하는 방법을 사용한다. 본 논문에서 제안한 방법을 사용하여 기존의 프레임 에너지를 이용한 방법과 비교한 결과 음성의 시작점의 오차율은 7%, 끝나는 점 오차율에서 11% 향상된 성능을 확인하였다.

주제어 : DNN, GMM, 음성 특징 추출, LMS, 잡음 제거

Abstract In the field of speech recognition, as the DNN is applied, the use of speech recognition is increasing, but the amount of calculation for parallel training needs to be larger than that of the conventional GMM , and if the amount of data is small, overfitting occurs. To solve this problem, we propose an efficient method for robust voice feature extraction and voice signal noise removal even when the amount of data is small. Speech feature extraction efficiently extracts speech energy by applying the difference in frame energy for speech and the zero-crossing ratio and level-crossing ratio that are affected by the speech signal. In addition, in order to remove noise, the noise of the speech signal is removed by removing the noise of the speech signal with an average predictive improved LMS filter with little loss of speech information while maintaining the intrinsic characteristics of speech in detection of the speech signal . The improved LMS filter uses a method of processing noise on the input speech signal by adjusting the active parameter threshold for the input signal. As a result of comparing the method proposed in this paper with the conventional frame energy method, it was confirmed that the error rate at the start point of speech is 7% and the error rate at the end point is improved by 11%.

Key Words : DNN, GMM, Speech Feature Extraction, LMS, Noise Elimination

*This research was supported by Global Infrastructure Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT(NRF-2018K1A3A1A20026485)

*Corresponding Author : Sang Yeob Oh (syoh1234@gmail.com)

Received April 20, 2021 Revised May 4, 2021
Accepted June 20, 2021 Published June 28, 2021

1. 서론

음성 인식률의 저하는 음성에 대한 신호처리에서 잡음이 추가되거나 디지털 신호에 양자화 잡음이 부가되어 원래의 음성 신호가 수정되거나 변경되어 발생한다. 그러므로 수정되거나 손상이 발생한 음성 신호에서 잡음을 삭제하기 위한 다양한 연구가 진행되고 있으며, 이를 기반으로 원래 음성 신호로 복원하기 위한 방법이 연구되고 있다. 음성 처리에 대한 방법은 HMM(Hidden Markov Model), GMM(Gaussian mixture model), CHMM(Continuous Hidden Markov Model) 등의 방법과 최근에 인공 지능 분야의 DNN(Deep Neural Network) 방법의 적용으로 음성 인식 분야가 널리 점진적으로 사용되고 있다[1-4]. 잡음환경에 강인한 음성 검출은 잡음에 영향을 받지 않는 신호를 추정하는 것이 필요하며, 신뢰성 있는 음성 추정을 위해 Wiener-filter를 사용한다. 일정한 잡음에서 통계 특성을 가지는 분산 값을 처리하는 기법은 특정 프레임내의 에너지에 대해 기대 값의 변화량을 이용하는 방법을 사용하고, 프레임 내의 전체 대역 에너지가 일정한 에너지 패턴을 보이는 잡음 추정치에 비해 일정 대역에 몰려 있는 음성은 그 분산 값이 서로 상이하게 다른 점을 이용한다. 잡음 제거와 음성 향상을 위한 연구가 활발히 이루어지고 있으며 LMS(Least Mean Square) 적응 필터(adaptive filter)기법, 독립성분분석(Independent Component Analysis)기법, 스페셜필터(special filter)기법 등이 이용되고 있다.

DNN은 HMM, GMM 등의 음성 인식 방법에 비해 음성 인식에 대한 오류가 10에서 30% 정도 낮은 특성을 가지나 음성 인식에 대한 병렬 훈련이 기존의 방법보다 복잡하고, 특히 음성에 대한 학습 데이터의 양을 많이 필요로 하며, 음성 인식 처리에서 데이터의 양이 작으면 오버피팅(overfitting)으로 많은 데이터 처리를 동반해야 되는 단점이 있다. 더욱이 음성 인식 상용에 최소 1000 시간의 학습 데이터를 필요로 한다. 이를 위해 본 논문에서는 DNN 에서 데이터의 양이 작은 경우에 효율적인 처리를 위해 강인한 음성 특징 추출 방법을 적용하고, 음성 신호 잡음 제거 방법을 융합한 방법을 제안하여 오버피팅과 잡음제거 문제를 해결하기 위한 방법을 제안한다. 이 방법을 적용하여 기존의 음성 인식 방법을 개선 및 향상하고, DNN의 데이터양에 대한 문제를 융합하여 해결한다. 본 연구의 음성 특징 추출은 음성에 대한 프레임 에너지의 차이와 음성 신호에 영향을 받는 영 교차율과 레벨 교차율을 적용하여 음성 에너지의 효율적 추출을 한다. 잡음 제거를 위해서는 음성 신호 검출이 중요하며[5-8], 검출은 음성에 대한 고유 특성을 기반으로 음성 신호 변경이 적은 평균 예측 LMS(Least Mean Square) 필터를 개선하여 음성 신호에 대한 잡음을 처리한다. 개선된 LMS 필터는 입력 신호에 대한 활성 파라미터 임계치를 조정하여 입력된 음성 신호에 대한 잡음을 처리하는 방법을 제공하여 잡음을 제거하여 음성 신호에 대한 원 신호에 대한 성능을 향상하도록 하였다. 이 방법을 적용하여 음성신호 특성 추출 처리에서 음성에 대한 고유한 정보를 유지하여 음성 정보에 대한 손실을 줄이도록 하였다. 성능 평가를 위해 음성을 검출한 결과 기존의 프레임 에너지를 이용한 방법과 제안한 방법의 비교 결과 음성의 시작점의 오차율은 7%, 끝나는 점 오차율에서 11% 향상된 성능을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 언급하고 3장에서는 평균 예측 필터를 이용한 잡음 환경에 강인한 음성 검출 방법에 대해 설명하며, 4장에서는 시스템 평가를 수행하고, 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

2.1 LMS 적응 필터

잡음은 다양한 환경적인 특성으로 인해 특정 공간에서 발생하는 잡음은 다양하며, 이와 같은 잡음에 대응할 수 있어야 하며, 실제 사용되는 환경의 잡음을 디지털 영역에서의 계산은 시간의 변화에 대한 임펄스 응답을 사용하는 선형 시변 필터를 가지고 모델링 추출을 수행한다[3,9]. 적응 필터에 대한 입력 신호 $s(n)$, 음성 신호 $v(n)$, 출력을 나타내는 하는 $y(n)$을 사용하며, $\widehat{v}(n)$은 음성 신호에 대한 예측 값을 의미한다. 출력 $y(n)$과 예측 값 $\widehat{v}(n)$의 차이를 오류 신호 $e(n)$로 나타내어 차이 값에 대한 오류 신호가 발생하면 far-end로 나타내고, 차이 값이 없거나 동일한 값을 가지면 near-end 신호를 가지고 처리한다[3].

LMS 적응 필터에서 임펄스 응답 $f(n,l)$은 음성 신호 $v(n)$에 대해 다음의 수식으로 처리한다.

$$y(n)=\sum_{l=0}^{+\infty} f(n,l)\,s(n-1)+noise(n) \quad (1)$$

임펄스 응답 $f(l)$은 지수 함수로 처리되어 l에 대해 비해 매우 작은 값을 가지므로 $f(n,l)=0,\ (l \geq n)$로 나타낸다. LMS 적응 필터는 잡음 제거 필터로 사용되어 정확한 조건에 대해 결과를 처리하므로 예측치 신호 $\hat{v}(n)$과 출력 신호 $y(n)$에 대해 최소의 차이 값을 가지는 조건에 적용된다.

2.2 음성에너지 분포와 음성 특징

음성 검출과 음성 처리를 위한 처리는 발생되는 잡음을 제거하기 위해 발생된 잡음을 계산하고 잡음에 대한 변경 차이를 제거하는 방법을 사용한다[11]. 이와 같은 처리를 위해 잡음 변경에 대해 음성에 대한 현재 프레임과 이전 프레임의 차이에 대한 비교를 수행하거나 음성에 대한 100ms 이상의 시간을 분석하는 방법을 적용하지만 실제 사용하여 응용하기에 적합하지 않은 점이 있다. 이를 보다 효율적으로 처리하기 위해 음성에 대한 분석 구간을 상대적으로 줄이기 위해 구간 퓨리에 분석은 음성 신호 x(t)에 대한 주파수 영역을 위해 DFT(Discrete fourier Transform)을 사용하여 다음 식과 같이 주파수 성분을 Xi로 표현한다.

$$x_i = \sum_{k=0}^{N-1} x_k e^{-j(\frac{2\pi m k}{N})} \quad (2)$$

구간 퓨리에는 음성에너지에 대한 최대화를 적용하기 위해 선형 주파수는 인간의 청각모델에서 사용하는 비선형 주파수 크기로 변환한다. 음성의 기본 주파수는 피치 주파수(pitch frequency)를 가지며, 기본 주파수는 음성 영역의 모든 대역에서 최대 에너지로 표현되며, 최소 에너지는 음성 신호와 무관한 잡음 신호로 표현된다. 각 음성 프레임에서의 표준 편차는 음성 에너지의 분포를 나타내고, 음성 에너지의 주요 데이터는 100Hz ~ 600Hz 대역에서 대부분 사용되는 특성을 가지므로 인간의 가청 영역에서 음성 에너지 크기의 편차가 큰 특징을 가진다. 일반적으로 잡음은 음성 에너지의 가청 영역 전반에 분포되어 편차가 작은 특징을 가진다.

3. 시스템 모델

본 연구에서 제안된 방법의 프로세스는 다음과 같다.

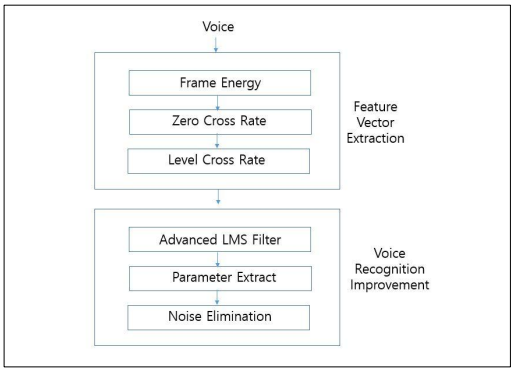

Fig. 1. Process of proposed methode

음성 특징 추출은 음성이 가진 음성 신호의 동적인 특징을 반영하며, 음성 프레임 에너지는 음성 구간 에너지 값보다 무음 구간 에너지 값이 적으므로 이의 차이를 이용하여 음성의 끝점을 추출하기 위한 레임의 에너지 E_1에 대한 식은 다음과 같이 정의한다.

$$E_1 = 10\log[\sum_{n=0}^{n-1} x^2(n)] \quad (3)$$

x(n)은 양자화된 신호, N은 한 프레임내의 샘플 수를 의미한다. 음성은 에너지가 적은 주파수를 가지고, 무음은 에너지가 높은 주파수 대역을 사용하므로 무음 구간에서 사용되는 영 교차율은 무음에 비해 작고 음성보다 크게 된다. 영 교차율 Z는 다음과 같다.

$$z=\frac{1}{2}\sum_{m=0}^{n-1} sgn[x(n-m)]-sgn[x(n-m-1)],$$

$$sgn[x(n)] = \begin{cases} 1, x(n)>0 \\ -1, otherwise \end{cases} \quad (4)$$

음성 신호의 레벨교차율에 대한 식은 다음 식과 같다.

$$z=\frac{1}{2}\sum_{m=0}^{n-1} sgn[x(n-m)]-sgn[x(n-m-1)-L_{th}]$$

,

$$sgn[x(n)] = \begin{cases} 1, x(n) > 0 \\ -1, otherwise \end{cases} \quad (5)$$

L은 레벨 교차율을 의미하고, L_{th}은 오프세트 레벨이며, (5)에서 미분은 시간 축 방향에 대한 필터링을 사용하여 시간 축 방향에 대한 특징 벡터를 사용한다. 이를 기반으로 음성 특징 추출을 위해 MFCC를 이용하여 사용한다.

음성 특징 추출의 인식률 향상[10,12,13]을 위한 잡음 제거를 위해서 LMS(Least Mean Square) 적응 필터를 개선한 방법을 적용한다[14]. 개선된 LMS 적응 필터는 상관 계수가 높은 패턴 모델의 수렴 속도가 저하되어 음성 검출 인식 성능의 저하된다. 이를 개선하여 평균 예측량을 적용한 LMS 검출 방법을 사용하여 잡음을 제거하고 강인한 음성 검출을 수행한다. 평균 예측량을 이용한 필터는 다음 수식으로 정의한다.

$$AMerr[n] = \frac{1}{L}\sum_{k=0}^{L-1} e[n-k] \quad (6)$$

평균 예측 필터는 n 시간에 대한 $e[n]$의 $L-1$개의 입력 값으로부터 평균값을 계산한다. 파라미터는 검출을 위해 다음 수식적으로 나타낼 수 있다.

$$J_{SCLS}(N) = J_{LS}(N) + m\sigma_v^2 \log N \quad (7)$$

m은 활동 파라미터의 수를 나타내고 σ_v^2는 $v(k)$의 편차를 의미한다. 활성 파라미터 임계치를 사용하여 활성 파라미터 검출을 위한 활성 파라미터 측정치를 구하고, k시간에 파라미터 벡터인 $\hat{\theta}_j(k)$를 평균값에 의해 변경하여 다음 수식적으로 나타낼 수 있다.

$$\hat{\theta}_j(K+1) = \alpha^{1-g_j(k)}\hat{\theta}_j(k) + \mu * AMerr * g_j(k)u(k-j) \quad (8)$$

$g_j(k)$는 j번째 $g(k)$의 파라미터 요소이며, 이를 반복 사용하여 파라미터 요소를 추출하여 입력 신호에 대한 잡음을 제거한다.

4. 실험 결과

본 논문에서는 개선된 평균 예측 LMS 필터를 사용하여 잡음을 제거하고, 음성에 대한 프레임 에너지의 차이와 음성 신호에 영향을 받는 영 교차율과 레벨 교차율을 사용하고, 개선된 LMS 필터를 사용하여 잡음에 강인한 음성 검출 방법을 제안하였으며, 실험을 위해 음성정보기술산업지원센터의 음성 데이터베이스를 적용하며, 이 음성 데이터베이스는 8kHz의 다운 샘플링된 데이터로 각각 한국어 음성 50개, 총 100개의 단어에 대해 실험하였다. 환경에 대한 잡음의 강도를 나타내기 위해 각 환경에 대한 신호 대 잡음비(Signal to Noise Ratio)를 구하였으며, 제안한 방법에 의해 음성 데이터의 에너지에 추정한 잡음의 에너지를 감하여 잡음을 줄인 신호 에너지를 구하고, 이 신호와 추정 잡음의 에너지에 대한 신호 대 잡음비를 얻는다.

그림 2는 한국어 "대학교"라는 음성과 엔진 잡음을 합성한 파형의 에너지 분포를 나타낸다.

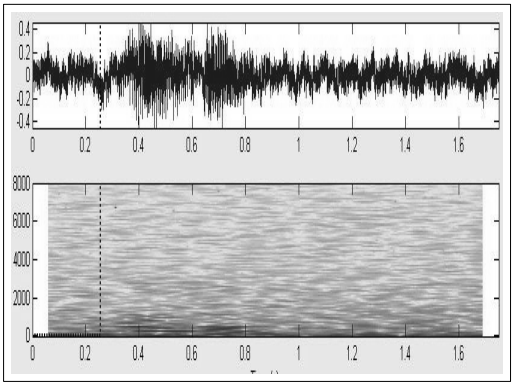

Fig. 2. Energy distribution of the composite waveform

3장에서 설명한 활성 파라미터에 대한 측정값을 사용하여 잡음이 부가된 음성 신호를 필터링하여 원래의 음성 신호에 근접한 신호를 구하였으며, 이를 성능평가에 적용하기 위해 음성 프레임 에너지에 대해 에지 검출 필터를 적용한 기법과 제안한 기법을 비교하여 발생한 오차를 가지고 분석하였으며[15], 오차는 실제 음성 구간에 해당되나 비음성 구간과 음성이 훼손된 구간을 나타내며, 검출의 시작 부분과 끝 부분 기준 신호가 다르면 오차로 처리한다. 정상 검출은 오차가 아니면서 음성 구간이 훼손되지 않은 상태이며, 시작 부분과 끝 부분 기준 신호가 같은 상태이다[13].

Table 1은 100개의 잡음 환경 음성 데이터에를 사용하여 음성 프레임 에너지에 대한 영교차율을 에지 검출 필터를 적용한 결과를 나타낸다[14]. 음성의 시작점 오차율이 27%와 끝나는 점 오차율 31%를 나타내는 것을 확인할 수 있다.

Table 1. Result of existing frame energy

Separation	Starting point (ea)	Ending point (ea)
Error	27	31
Detection	73	69

Table 2는 100개의 잡음 환경 음성 데이터에 대한 제안한 기법의 결과를 나타낸다. 음성의 시작점 오차율이 14%, 끝나는 점 오차율 21%를 확인할 수 있다.

Table 2. Result of suggested methode

Separation	Starting point(ea)	Ending point(ea)
Error	14	21
Detection	86	89

제안한 방법과 기존 프레임 에너지를 이용한 방법과 비교한 결과 시작점의 오차율이 7%와 끝나는 점 오차율 11%향상된 것을 확인할 수 있다.

5. 결론

본 논문에서는 잡음 제거 방법으로 음성신호 의 고유 특성을 유지하고 음성 정보 손상을 제거하기 위한 개선된 평균 예측 LMS 필터를 이용하여 오염된 음성 신호의 잡음을 제거하였으며, 음성 신호에 대한 프레임 에너지의 차이와 음성 신호에 영향을 받는 영 교차율과 레벨 교차율 사용하고, 잡음에 강인한 음성 검출 방법을 제안하였으며, 이 방법을 적용하여 기존의 음성인식 방법을 개선 및 향상하고, DNN의 데이터양에 대한 문제를 융합하여 해결한다. 성능 분석을 위해 잡음환경에서 채집된 데이터를 사용하여 100개의 단어에 대해 실험하여 잡음 환경에서 개선된 평균 예측 LMS 필터를 이용한 잡음 제거 실험을 수행한 결과 신호 대 잡음 비율이 이 향상된 수렴 성능을 확인하였다. 성능 평가한 결과 기존 프레임 에너지를 이용한 방법과 비교하여 음성의 시작점의 오차율과 끝나는 점 오차율이 향상된 것을 확인하였다.

REFERENCES

[1] S. Y. Oh. (2020). Speech Recognition Performance Improvement using a convergence of GMM Phoneme Unit parameter and Vocabulary Clustering. *Journal of Convergence for Information Technology, 10(8),* 35-39.
DOI : 10.22156/CS4SMB.2020.10.08.035

[2] C. S. Ahn & S. Y. Oh. (2012). Gaussian Model Optimization using Configuration Thread Control In CHMM Vocabulary Recognition. *The Journal of Digital Policy and Management. 10(7),* 167-172.
DOI : 10.14400/JDPM.2012.10.7.167

[3] J. Homer & I. Mareels. (2004). LS detection guided NLMS estimation of sparse system. Proceedings of the IEEE 2004 International Conference on Acoustic. Speech, and Signal Processing(ICASSP). Montreal, Quebec, Canada.
DOI : 10.1109/ICASSP.2004.1326394

[4] B. Sisman, J. Yamagishi, S. King & H. Li. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*

[5] B. F. Wu & K. C. Wang. (2005). Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. *IEEE Transactions on Speech and Audio Processing, 13(5),* 762-775.
DOI : 10.1109/TSA.2005.851909

[6] Q. Li, J. Zheng, A.Tsai & Q. Zhou. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Transactions on Speech and Audio Processing, 10(3),* 146-157.
DOI : 10.1109/TSA.2002.1001979

[7] A. Arango, J. P´erez & B. Poblete. (2019). Hate Speech Detection is Not as Easy as You May Think, A Closer Look at Model Validation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 45-54. Paris, France: Association for Computing Machinery.
DOI : 10.1145/3331184.3331262

[8] S. S. Aluru, B. Mathew, P. Saha & A. Mukherjee. (2020). Deep Learning Models for Multilingual Hate Speech Detection, *arXiv preprint arXiv:2004.06465*

[9] E. T. S. I. Standard. (2003). Speech Processing, Transmission and Quality aspects(STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. *ETSI ES 202 050 v.1.1.3.*

[10] P. Scart & J. Filho, (2002). Speech enhancement based on a priori signal to noise estimation. *In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (Vol. 2, pp. 629-632). IEEE.

[11] K. Chung & S. Y. Oh. (2015). Improvement of speech signal extraction method using detection filter of energy spectrum entropy. *Cluster Computing, 18(2),* 629-635.
DOI : 10.1007/s10586-015-0429-9

[12] S. Kamarth & P.Loizou. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. *In ICASSP* (Vol. 4, pp. 44164-44164).

[13] Yi Hu & P. C. Loizou. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing, 16(1),* 229-238.

[14] S. Y. Oh & K. Chung. (2018). Performance evaluation of silence-feature normalization model using cepstrum features of noise signals. *Wireless Personal Communications, 98(4),* 3287-3297.
DOI : 10.1109/TASL.2007.911054

[15] K. C. Wang & Y. H. Tsai. (2008). Voice activity detection algorithm with low signal-to-noise ratios based on spectrum entropy. *In 2008 Second International Symposium on Universal Communication* (pp. 423-428).
DOI : 10.1109/ISUC.2008.55

오 상 엽(Sang-Yeob Oh) **[정회원]**

· 1991년 2월 : 광운대학교 대학원 전자계산학과 (이학석사)
· 1999년 2월 : 광운대학교 대학원 전자계산학과 (이학박사)
· 2007년 2월 ~ 현재 : 가천대학교 IT대학 컴퓨터공학과 교수

· 관심분야 : 인공지능, HCI, 차량 통신, 형상관리, 음성 및 음향 신호처리, 정보검색, 추천 시스템, 기계학습
· E-Mail : syoh1234@gmail.com