Journal of KIIT. Vol. 24, No. 3, pp. 183-188, Mar. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.3.183

AI 음성 학습 효율화를 위한 개선된 SFN을 이용한 음성 인식 성능 향상

오 상 엽*

Voice Recognition Performance Improvement using Supplemented SFN for AI Voice Learning Efficiency

Sang Yeob Oh*

요 약

인공 지능 기반 기술의 발전으로 chatGPT, Gemini 등의 대화형 인공 지능 음성 처리 기술 등이 널리 이용되고 있다. 그러나 지속적으로 발생하는 사용 환경의 잡음, 불규칙한 발음과 연음에 대한 인식 문제로 음성 인식에 대한 성능 저하의 문제가 계속적으로 발생된다. 음성 인식을 위한 인공지능 학습 모델 훈련 방법에서 발생하는 음성 인식률 향상 방법으로 본 논문에서는 개선된 SFN(Silence Feature Normalization)을 이용한 음성 인식 성능 향상 방법을 제안한다. 본 연구에서는 높은 신호 대 SNR에서 음성의 특징에 대한 잡음이 적게 영향을 받고, 낮은 SNR에서는 음성과 비음성의 특징 분포 특성을 이용하여 인식률을 향상하였다. 잡음에 대한 사용자의 음성 인식 정확도를 평가하기 위해 SNR에 대한 Noise Reduction Rate와 PHR(Pause Hit Ratio)을 사용한 정확도를 분석하여 향상된 인식률을 확인하였다.

Abstract

With the advancement of artificial intelligence-based technology, conversational AI speech processing technologies such as chatGPT and Gemini are widely used. However, performance degradation in speech recognition continues to occur due to persistent noise in the user environment, irregular pronunciation, and recognition problems for consonants. In this study, we suggest a way to improve speech recognition performance using improved Silence Feature Normalization (SFN) in the method of training an artificial intelligence learning model for speech recognition. In this paper, the recognition rate is supplemented by taking advantage of the feature distribution characteristics of speech and non-speech at low SNR, while the noise has less effect on speech features at high signal-to-signal ratios. To evaluate the accuracy of user speech recognition in noise, the accuracy was analyzed using the Noise Reduction Rate and Pause Hit Ratio (PHR) for SNR, and an improved recognition rate was confirmed.

Keywords
vocabulary recognition, silence feature normalization, feature extraction, SNR, PHR

* 가천대학교 컴퓨터공학과 교수
- ORCID: https://orcid.org/0000-0002-8002-9588

· Received: Dec. 09, 2025, Revised: Dec. 31, 2025, Accepted: Jan. 03, 2026
· Corresponding Author: Sang Yeob Oh
Dept. of Computer Engineering, Gachon University, Korea
Tel.: + 82-31-750-5798, Email: syoh@gachon.ac.kr

Ⅰ. 서 론

인공 지능 알고리즘 자연어 처리에서 NLP (Natural Language Processing)를 이용하여 컴퓨터가 인간의 음성을 이해하고, 처리하는 AI 챗봇이 널리 사용되고 있다[1]-[4]. 이 기술의 핵심은 학습으로 많은 데이터를 학습하고 적절한 출력에 대한 내용이 알고리즘으로 처리되어 음성의 의미와 문법적 역할을 분석하고 문장의 의도를 인지한다. 그러나 학습된 데이터의 음성 인식 처리에서 사용 환경의 잡음과 잡음 발생에 따른 음성 삭제, 불규칙한 발음과 연음에 대한 인식 문제, 유사 음성의 다른 음성과의 혼동, 음성의 연음으로 인한 음성 에러 등의 문제가 발생된다.

인공지능 음성 처리 알고리즘은 특정 잡음과 낮은 SNR(Signal to Noise Ratio) 환경에서 열화되어 성능이 떨어지는 경우가 있다. 이는 비슷한 주파수 특성을 가진 잡음에 대해서 성능이 떨어지고 위상 검출이 완벽하지 못하기 때문이다. 또한 잡음 제거를 위해 잡음 특성이 반영된 잡음 신호를 확보하기 위한 신호 입력이 별도로 존재하여야 하며, 잡음 환경에서는 낮은 SNR의 성능 저하가 발생된다[5]-[8]. 음성 신호에 부가적 잡음이 포함되어 음성 신호 스펙트럼에 변경이 발생되거나 음성 신호프레임을 탐색하지 못해 음성 인식 성능이 낮아진다. 음성 인식 성능 저하는 음성 인식 모델 훈련 환경에서 많은 차이가 나타난다.

기존의 SFN(Silence Feature Normalization)은 낮은 신호 대 SNR에서 무음 구간의 검출에서 에너지 레벨이 증가하여 음성과 비음성에 대한 불분명한 경계 분류로 인하여 인식 성능에 영향을 미치게 되는 문제점이 있다. 또한, 정규화를 위해 작은 Log-energy을 갖는 비음성 특징을 찾고, 이를 IIR(Infinite Impulse Response) filter를 통과시켜서 사용한 방법에서 기존의 HMM과 CHMM을 이용한 분석에도 1~2% 정도의 인식률 향상을 가진다[9].

이러한 문제점을 개선하기 위해 본 논문에서는 개선된 SFN을 사용한다. 본 논문에서는 음성과 비음성에 대한 분류 처리를 위한 음성 구간 검출에서 개선된 SFN을 이용한 잡음 환경에 강인한 음성 특징 검출 방법으로 캡스트럼 특징을 정규화하고, 잡음이 부가된 음성의 로그 에너지에 대한 가중 함수 처리를 수행하여 음성에 대한 인식률을 향상하기 위한 방법을 제안한다. SNR의 분석을 위해 음성 비서, S-Voice, AI 스피커 등에서 사용자의 음성 인식 정확도를 평가하기 위해 SNR에 대한 PHR(Pause Hit Ratio)을 사용하여 분석하였으며, 자동차와 거리 환경에서 잡음 환경의 잡음 감소율에 대한 음성 신호의 신뢰성 분석을 수행하여 향상된 인식률을 확인하였다.

본 논문은 2장 관련 연구, 3장 개선된 SFN을 이용한 잡음 환경의 차이를 개선한 음성 인식 방법, 4장 시스템 평가, 그리고 5장 결론으로 구성한다.

Ⅱ. 관련 연구

2.1 Zero crossing rate & spectral energy

음성 구간 검출을 위해 시간에 따른 주파수 변화량과 주파수에 따른 시간의 변화량에서 음성 구간을 검출한다. 음성 구간 검출은 계산이 쉽고 비교적 성능이 좋은 ZCR(Zero Crossing Rate)을 사용한다[10]. ZCR은 시간 영역(Time domain)에서 구간의 샘플은 +1 또는 -1로 나타나며, 두 샘플을 곱하여 계산한다. 계산된 결과 음수의 개수에 따라 음성 구간과 비음성 구간을 구분한다. 고주파가 저주파에 비해 음수의 개수가 많이 발생하는 특성을 가진다.

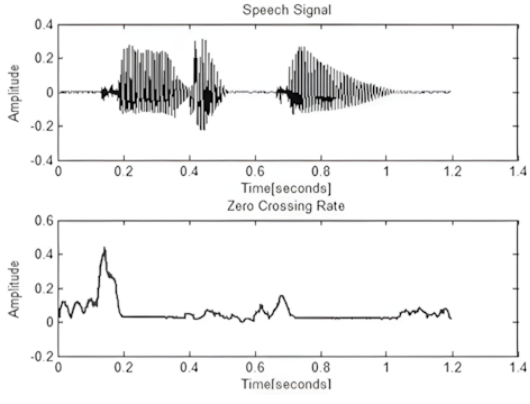

그림 1. 음성신호와 ZCR
Fig. 1. Speech signal and ZCR

그림 1의 x축은 time, y축은 amplitude를 나타내며, ZCR값이 낮은 결과값을 보인 SNR 에너지 스펙트럼은 음성 구간에서 비음성 구간은 높은 에너지 스펙트럼을 가진다. 따라서 음성 에너지 스펙트럼은 비음성 에너지 스펙트럼에 비해 상대적으로 높은 에너지 스펙트럼을 가진다. Shannon은 음성 에너지 스펙트럼이 비음성 에너지 스펙트럼에 비해 상대적으로 높은 에너지 스펙트럼을 가지며, Shannon이 표현한 엔트로피(Entropy)는 DTF(Discrete Fourier Transform)로 계산하였으며 이산 spectral power의 결과 값을 도출하였다. 엔트로피에 의해 처리되는 spectral energy에 대한 확률의 수식은 식 (1)과 같다.

$$P(|Y(k,l)|^2) = \frac{|Y(k,l)|^2}{\sum_{k=1}^{N/2}|Y(k,l)|^2} \quad (1)$$

여기서, k는 Frequency bin 인덱스이며, l은 프레임 인덱스를 의미한다. 프레임에서 frequency bin에 대한 spectral energy 확률을 계산한다. 계산된 각 frequency bin의 확률은 엔트로피로 표현된다.

2.2 Silence feature normalization

SFN은 기준값보다 작은 로그 에너지를 갖는 구간을 정규화(Normalization)하며, 작은 로그 에너지를 갖는 비음성 특징을 찾아서 일정값 이하로 감소한다. 로그 에너지를 IIR 필터를 통과시켜서 에너지를 줄여 사용한다. 에너지의 평균 데이터를 계산하여 음성과 비음성을 분류하는 기준 데이터를 T로 설정한다. 계산된 에너지 평균 데이터가 기준값인 T보다 크면 음성 구간으로 분류하여 로그 에너지 데이터를 유지하고, 기준값인 T보다 낮으면 비음성으로 지정하여 기준보다 낮은 비음성 데이터로 정규화한다. 음성 구간으로 분류된 프레임은 로그 에너지 데이터를 유지하고, 묵음 구간으로 분류된 프레임은 작은 상수 데이터로 정규화한다. 출력에 가중치를 곱하여 음성의 큰 값을 갖도록 가중치를 높이고 비음성은 작은 값을 갖도록 가중치를 낮춘다. 잡음에 오염된 음성은 잡음만 나타나는 신호 구간에 비해 주파수 대역폭이 넓게 나타나며 큰 로그 에너지를 갖는 구간은 작은 로그 에너지를 갖는 구간에 비해 잡음 신호의 영향을 적은 특징을 가진다.

III. 시스템 모델

본 논문에서는 음성과 비음성에 대한 분류를 위한 개선된 SFN을 이용한 음성 인식 성능 향상 방법을 제안한다. 제안한 방법은 높은 SNR에서는 음성과 비음성에 대한 특징 분포 특성을 이용하여 음성이 잡음을 최소화하도록 하였다. 잡음 신호의 캡스트럼 특징은 음성 신호와 비교해서 분산값이 낮은 특성을 이용하며, 캡스트럼 특징에 대한 음성과 비음성 구분의 기준 데이터를 가지고 SNR이 낮은 신호에서 SFN으로 성능을 향상한다. 묵음(Silence) 캡스트럼을 정규화하기 위해서는 음성 신호에서 추출한 캡스트럼 특징을 정규화하였고 인식을 위한 잡음 환경을 훈련 잡음 환경과 잡음의 차이를 비슷하게 구성하였다. 캡스트럼 평균 정규화를 위해 식 (2)에서는 캡스트럼 특징에서 각각의 차수에 대한 평균 데이터 m을 계산하여 차수별로 빼서 구한다.

$$\overline{C_n}(m) = C_n(m) - \mu(m) \quad (2)$$

켑스트럼 특징의 각 차수의 대한 평균값은 식 (3)과 같다.

$$\mu(m) = \frac{1}{n}\sum_{n-1}^{N} C_n(m) \quad (3)$$

캡스트럼 특징의 분산값을 감소시키고 전체 프레임에 대한 각 차수별 캡스트럼 특징의 분산값을 계산하는 수식은 다음과 같다.

$$\sigma^2(m) = \frac{1}{n}\sum_{n-1}^{N}(C_n(m)-\mu(m))^2 \quad (4)$$

캡스트럼 특징은 차수별로 나누어 평균값은 0, 분산값은 1이 되도록 계산한다. $\sigma^2(m)$는 캡스트럼 특징에서 n번째 차수의 성분의 분산값을 의미한다. 식 (4)를 입력된 음성 신호에 적용하여 음성 신호의

특징이 잡음에 영향을 적게 받는 신호 모델을 구성하였다. 잡음화된 음성은 순수 잡음 음성과 비교해 더 큰 주파수 대역폭을 가지며, 큰 로그 에너지 구간은 작은 로그 에너지 구간과 비교해 잡음 영향이 적으므로 가중 함수를 본 연구에서 사용하였다. 가중 함수는 식 (5)와 같이 수식적으로 나타낸다.

$$w(n) = \begin{cases} 1/(1+\exp(-(\log\bar{E}(n) - T_0)/\beta\sigma_1 & \text{if } \log\bar{E}(n) > T_0 \\ 1/(1+\exp(-(\log\bar{E}(n) - T_0)/\beta\sigma_2 & \text{if } \log\bar{E}(n) > T_0 \end{cases} \quad (5)$$

잡음 환경의 음성 신호는 큰 에너지를 가지는 음성 데이터에 대해 잡음 영향이 없지만, 작은 에너지 구간에서는 상대적으로 잡음이 더 발생된다. 작은 로그 에너지 특징 구간에서는 로그 에너지를 증대하여 잡음을 가진 음성신호와 로그 에너지를 유사하게 하며, 이를 위해 각 음성 신호에 대한 로그 에너지가 처리를 위한 DR(Dynamic Range) 함수는 식 (6)과 같다.

$$DR(dB) = 10 \times \frac{Max(\log E_n)}{Min(\log E_n)} \quad (6)$$

여기서, $Max(\log E_n)$는 N 개 프레임에 대한 최대 로그 에너지, $Min(\log E_n)$은 최소의 로그 에너지를 의미한다. DR에 대해 각 음성 신호의 로그 에너지 특징의 최솟값은 식 (7)과 같이 처리한다.

$$T\_\min = 10 \times \frac{Max(\log E_n)}{DR(dB)} \quad (7)$$

N개의 프레임 중에서 로그 에너지의 최대값과 최솟값을 찾아서 Target Minimum이 정해지고, 로그 에너지의 최솟값이 T_min보다 작은 경우에 정규화된 로그 에너지를 가지고 음성 특징을 처리한다.

Ⅳ. 성능 분석

본 논문은 SFN에서 음성과 비음성의 특징 분포 특성을 이용하여 잡음 영향을 적게 받을 수 있도록 하였으며, 성능에 대한 평가를 수행하기 위해 데이터베이스는 Aurora 2.0을 이용하여 분석하였다 [11][12]. Aurora 2.0에는 각각의 잡음 환경과 잡음 신호 레벨로 분류되어 있으며 car, street noise 등 음성 향상 알고리즘의 성능 검증용으로 사용된다. 음성 인식 실험에서 car, street noise 신호들로부터 생성된 잡음 음성이 포함되며, 이들 음서의 특징 벡터를 추출하여 처리하였다. 본 연구에서 제안한 방법의 성능 분석을 위해 음성 인식 실험을 수행하였으며, 잡음 처리는 워너 필터를 사용하고, 음성 인식 실험을 위해 경기도 지역명과 학교 이름 각각 20개를 사용하고, 음성 인식 수행 평가를 위해 각 단어에 대해 5회 발음하여 20개의 지역명에 대해 100단어를 사용하였으며, SNR 변화에 따른 음성 인식 성능을 잡음 환경(15dB, 10dB, 5dB, 0dB)으로 구분하여 실험하였다. ZCR 성능 평가를 위해 비음성에 대한 PHR(Pause Hit Ratio)을 사용하였다. 음성 원본은 8kHz 샘플링 레이트(Sampling rate), 16비트를 이용하였으며 FFT는 256 sample 크기, 1/2 오버래핑을 이용하였고 해밍 윈도우를 사용하며, 실험 결과 표 1에서 SNR에 대해 0dB~15dB에서 잡음이 감소된 결과를 확인할 수 있다. 딥러닝 기반 시스템인 CNN(Convolution Neural Network)와 같은 인공지능 알고리즘도 Noise Reduction Rate를 주요 성능 지표로 사용하며, Noise Reduction Rate는 음성 신호의 품질과 사용자 만족도에 직접적인 영향을 주는 지표이다. 표1과 표2는 본 연구에서 제안한 방법에 의한 결과를 나타낸다. 국내 다른 논문과의 성능 분석 비교를 위해, [9]의 논문에서 음성 신호 특징과 SFN을 이용한 음성 성능의 분석 결과는 음성 종속(Speech dependent) 단계의 평균은 85.2%, 음성 독립(Speech independent) 단계의 평균은 84.4%이며, 이 결과와 비교하여 Noise Reduction Rate가 표 1에서 전체적으로 향상된 것을 확인 할 수 있다.

표 1의 자동차 잡음 환경의 잡음 감소율은 낮은 SNR 0dB와 5dB에서는 87.1%, 89.3%의 정확도를 나타내며, 높은 SNR 10dB와 15dB에서는 91.5%, 93.1%의 정확도를 보였다. 거리 잡음 환경의 감소율은 낮은 SNR 0dB와 5dB에서는 81.5%, 82.1%의 정확도를 가지며, 높은 SNR 10dB와 15dB에서는 83.7%, 85.1%의 정확도를 보였다. 표 2는 잡음 환경

에서 음성 검출 성능을 나타내며, PHR은 음성 인식 시스템이 해당 입력을 실제로 인식해 동작을 실행하는 성공률을 나타내며, PHR 비율이 높을수록 음성 신호의 신뢰성이 높은 것을 나타낸다.

표 1. SNR에 대한 자동차와 거리 잡음 감소 비율
Table 1. Car and street noise reduction rate for the SNR

Noise	SNR (dB)	Noise reduction rate (%)
Car	0	87.1
	5	89.3
	10	91.5
	15	93.1
Street	0	81.5
	5	82.1
	10	83.7
	15	85.1

표 2의 자동차 잡음 환경의 PHR은 낮은 SNR 0dB와 5dB에서는 96.1%, 97.1%의 정확도를 나타내며, 높은 SNR 10dB와 15dB에서는 97.5%, 98.2%의 정확도를 보였다. 거리 잡음 환경의 PHR은 낮은 SNR 0dB와 5dB에서는 89.7%, 91.5%의 정확도를 보였으며, 높은 SNR 10dB와 15dB에서는 93.3%, 93.6%의 정확도를 보였다.

표 2. SNR에 대한 PHR
Table 2. PHR for the SNR

Noise	SNR (dB)	PHR Result (%)
Car	0	96.1
	5	97.1
	10	97.5
	15	98.2
Street	0	89.7
	5	91.5
	10	93.3
	15	93.6

V. 결 론

본 논문은 SFN에서 음성과 비음성에 대한 잡음 영향을 적제 받을 수 있는 방법을 제안하고, 이에 대한 인식 성능을 실험하였다. 다양한 잡음 환경이나 SNR가 낮은 음성 신호에 대해서는 여러 환경 잡음 신호에 민감하게 반응하여 신호 왜곡 현상이 나타나기 때문에 음성 신호에 대한 성능의 저하 원인으로 나타난다. 그러므로 SFN에서 음성 특징에 대한 잡음이 낮은 방법으로 처리하였으며, 낮은 SNR에서는 음성과 비음성의 특징 분포 특성을 가지고 음성에 대한 인식률을 높일 수 있도록 구성하여 잡음이 최소화되는 모델을 제안하였다.

SNR에 대한 자동차와 거리에 대한 잡음 감소 비율은 모두 dB의 증가에 따른 정확도를 성능 분석에서 확인할 수 있었으며, 자동차와 거리 잡음 환경의 PHR은 낮은 SNR구간인 0dB와 5dB보다 높은 SNR구간인 10dB과 15dB에서는 정확도가 향상 되었다. 이와 같은 결과는 AI 기반 환경에서의 음성 인식을 보다 향상 시키는데 기여할 수 있을 것으로 기대하지만, 제한된 실험 환경과 AI의 편향된 데이터 문제를 개선할 수 있는 다양한 방법론의 적용리 필요하다.

References

[1] M. H. Yi and J. H. Shin, "Emotion Recognition Model Using Progressive Transfer Learning for Speech Data Adation", Journal of Korea Multimedia Society, Vol. 27, No. 8, pp. 1004-1013, Aug. 2024. http://doi.org/10.9717/kmms.2024.27.8.1004.

[2] Y. J. Kim, H. J. Cha, and A. R. Kang, "A Study on the Impact of Speech Data Quality on Speech Recognition Model", Journal of the Korea Society of Computer and information, Vol. 29, No. 1, pp. 41-49, Jan. 2024. https://doi.org/10.9708/jksci.2024.29.01.041.

[3] S. Y. Oh, "DNN based Robust Speech Feature Extraction and Signal Noise Removal Method Using Improved Average Prediction LMS Filter for Speech Recognition", Journal of Convergence for Information Technology, Vol. 11, No. 6, pp. 1-6, Jun. 2021. https://doi.org/10.22156/CS4SMB.2021.11.06.001.

[4] E. D. Cahyadi, H. N. H. Soesild, and M. Song, "Enhancing Multimodal Emotion Recognition in Speech and Text with Integrated CNN, LSTM, and BERT Models", The Journal of Convergence

on Culture Technology, Vol. 10, No. 1, pp. 617-623, Jan. 2024. https://doi.org/10.17703/JCCT.2024.10.1.617.

[5] S. Y. Oh, "Improvement Entropy Feature Extraction for AI Voice Recognition Improvement", The Journal of Korean Institute of Information Technology, Vol. 23, No. 7, pp. 149-154, Jul. 2025. http://doi.org/10.14801/jkiit.2025.23.7.149.

[6] S. Y. Oh, "Vocabulary Recognition Rate Enhancement using Clustering Model and Non-parametric Correlation Coefficient", The Journal of Korean Institute of Information Technology, Vol. 22, No. 4, pp. 91-97, Apr. 2024. https://doi.org/10.14801/jkiit.2024.22.4.91.

[7] S. Y. Oh, "Noise Elimination Using Improved MFCC and Gaussian Noise Deviation Estimation", Journal of The Korea Society of Computer and Information Vol. 28 No. 1, pp. 87-92, Jan. 2023. https://doi.org/10.9708/jksci.2023.28.01.087.

[8] S. Y. Oh, "Speech Recognition Performance Improvement using a convergence of GMM Phoneme Unit parameter and Vocabulary Clustering", Journal of Convergence for Information Technology, Vol. 10, No. 8, pp. 35-39, Aug. 2020. https://doi.org/10.22156/CS4SMB.2020.10.08.035.

[9] J. C. Hwang, "Voice Recognition Performance Improvement using the Convergence of Voice signal Feature and Silence Feature Normalization in Cepstrum Feature Distribution", Journal of the Korea Convergence Society, Vol. 8. No. 5, pp. 13-17, May 2017. https://doi.org/10.15207/JKCS.2017.8.5.013.

[10] D. H. Johnson, "Signal-to-noise ratio", Scholar pedia, Vol. 1, No. 12, Art No. 2088, 2006. https://doi.org/10.4249/ scholarpedia.2088.

[11] K. Y. Chung and S. Y. Oh, "Vocabulary optimization process using similar phoneme recognition and feature extraction", Cluster Computing, Vol. 19, No. 3, 1683-1690, Aug. 2016. https://doi.org/10.1007/s10586-016-0619-0.

[12] K. Y. Chung and S. Y. Oh, "Voice Activity Detection Using an Improved Unvoiced Feature Normalization Process in Noisy Environments", Wirekess Personal Communications, Vol. 89, No. 3, pp. 747-759, Aug. 2016. https://doi.org/10.1007/s11277-015-3169-5.

저자소개

오 상 엽 (Sang Yeob Oh)

1989년 2월 : 경원대학교 전자계산학과(공학사)
1991년 2월 : 광운대학교 전자계산학과(이학석사)
1998년 2월 : 광운대학교 전자계산학과(이학박사)
1992년 9월 ~ 현재 : 가천대학교 컴퓨터공학과 교수
관심분야 : 음성 인식, 잡음 검출, 음성 특징 추출, 멀티미디어 데이터 통신