STT 효율 증대를 위한 음성 주파수 correlation 기반 노이즈 필터링 방안

임지원, 황용해, *김규헌
경희대학교
qouop_030@khu.ac.kr, hyh717@khu.ac.kr, *kyuheonkim@khu.ac.kr

Noise filtering method based on voice frequency correlation to increase STT efficiency

Jiwon Lim, Yonghae Hwang, *Kyuheon Kim
KyungHee University

요 약

현재 음성인식 기술은 인공지능 비서, 전화자동응답, 네비게이션 등 다양한 분야에서 사용되고 있으며 인간의 음성을 디바이스에 전달하기 위해 음성 신호를 텍스트로 변환하는 Speech-To-Text (STT) 기술을 필요로 한다. 초기의 STT 기술의 대부분은 확률 통계 방식인 Hidden Markov Model (HMM)기반으로 이루졌으며, 딥러닝 기술의 발전으로 HMM과 함께 Recurrent Nural Network (RNN), Deep Nural Network (DNN) 기법을 사용함으로써 과거보다 단어 인식 오류를 개선하며 20%의 성능 향상을 이루어냈다. 그러나 다수의 화자 혹은 생활소음, 노래 등 소음이 있는 주변 환경의 간섭 신호 영향을 받으면 인식 정확도에 차이가 발생한다. 본 논문에서는 이러한 문제를 해결하기 위하여 음성 신호를 추출하여 주파수성분을 분석하고 오디오 신호 사이의 주파수 영역 correlation 연산을 통해 음성 신호와 노이즈 신호를 구분하는 것으로 STT 인식률을 높이고, 목소리 신호를 더욱 효율적으로 STT 기술에 입력하기 위한 방안을 제안한다.

1. 서론

음성인식 기술은 인공지능 비서, 네비게이션 등 이미 우리 생활의 많은 부분에서 사용되고 있다. 이러한 기술들은 인간의 음성을 디바이스에 전달하기 위해 음성 신호를 텍스트로 변환하는 Speech-To-Text(STT) 기술을 필요로 한다.

초기의 STT 기술은 대부분 확률 통계 방식인 Hidden Markov Model (HMM) 기반으로 이루어졌으며, 2010년대 들어서면서 HMM과 함께 Recurrent Nural Network (RNN), Deep Nural Network (DNN) 방식을 사용하는 것으로 단어 인식 오류를 개선하여 20%의 성능 향상을 이루어 냈다.[1] 그럼에도 불구하고, 주변 잡음이 존재하는 경우 음성의 인식률이 떨어져 STT 기술을 적용하기 힘든 경우가 발생한다.

기존 기술의 제한 사항을 극복하기 위하여 음악과 음성이 혼합된 오디오로부터 음성만을 추출하여 이를 주파수 영역에서 분석하고, 주파수 성분의 correlation 연산을 통해 음성에 남은 잔여 잡음을 제거하고자 한다. 본 논문에서는 음성 신호 사이 주파수 영역에서의 correlation 연산을 통해 음성 신호와 노이즈 신호를 구분하는 것으로 STT 인식률을 높이고, 음성 신호를 더욱 효율적으로 STT 기술에 입력하기 위한 방안을 제시한다.

본 논문의 구성은 다음과 같다. 2 장에서는 주파수 분석을 위해 사용되는 연산과 음원추출 및 문자 변환 기술에 대한설명을 하고, 3 장에서는 본 논문에서 제안한 방안에 대하여 설명한다. 4장에서는 제안한 기법의 성능을 실험을 통해서 확인한다. 마지막으로 5장에서는 본 논문에 대한 결론을 맺으며 마무리한다.

2. 배경 기술

2-1. Short-Time-Fourier-Transform (STFT)

오디오 신호는 시간의 변화에 따른 1차원 신호의 세기 변화로 표현되기 때문에 다양한 신호가 섞여 있을 경우 시간 영역에서 신호의 세기 변화만으로 각 신호의 특징을 구분하기에 어려움이 있다. 차원이 매우 크고 여러 주파수의 합으로 이루어진 오디오 신호의 특성 때문에 시간영역에서 특징을 알아보기에는 한계가 있다.

$$F(u) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ux} dx \quad (1)$$

$$e^{2\pi i\theta} = \cos(2\pi i\theta) + jsin(2\pi i\theta) \quad (2)$$

식 (1)은 Fourier Transform으로 시간 영역에 대한 함수를 다양한 주파수를 가지는 주기함수의 합으로 나타냄으로써 주파수 영역으로 변환하는 연산이다. f(x)는 시간영역에서의 신호, F(u)는 주파수 영역에서의 신호를 나타내며 는 오일러 공식(2)를 의미한다. 오디오 신호에 식(1)을 적용할 경우 해당하는 신호의 주파수 영역 성분을 구할 수 있다.

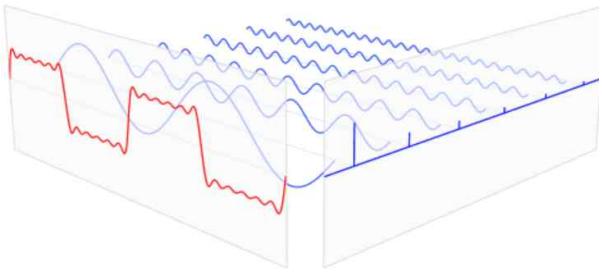

그림 1. 푸리에 변환 예시
Fig 1. Example of Fourier transform

그림 1은 푸리에 변환을 사용한 신호의 변환 예시로 시간 영역에서 여러 주파수를 가진 신호가 섞여 있는 경우 붉은색 신호와 같이 중첩되어 각 신호를 구분하는 것이 힘들지만, 푸리에 변환을 거친 푸른색의 경우 각 주파수에 따른 크기를 구분하는 것이 가능하다.

$$STFT\{x(t)\}(\tau,\omega) \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x(t) w(t-\tau) e^{-i\omega t} dt \quad (3)$$

STFT는 데이터를 시간에 대해 구간을 작게 나누어 Fourier Transform을 하는 방법으로 시간에 따라 주파수 성분이 변화하는 신호의 정보를 효율적으로 분석하기 위하여 STFT를 이용한다. 식 (3)는 STFT 변환 식으로 여기서 w(t)는 윈도우 함수를 나타낸다. 오디오 신호에 STFT를 적용한 결과를 시간과 주파수를 각각 가로, 세로 축으로 설정하고 각 주파수의 세기를 dB 단위의 색 변화로 표현한 3차원 그래프를 spectrogram이라고 한다.

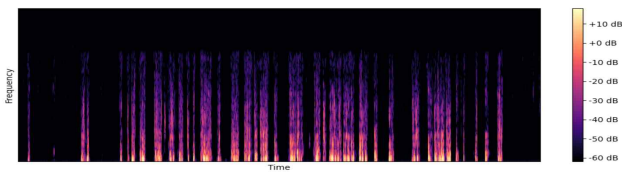

그림 2. 오디오 신호의 스펙트로그램
Fig 2. Spectrogram of audio signal

그림 2는 STFT가 적용된 오디오 신호를 스펙트로그램으로 표현한 예시이다. 스펙트로그램에서 주파수의 세기에 따라 색으로 구분되는데 노란색에 가까울수록 주파수의 세기가 강한 것이고 자색을 띌수록 주파수의 세기가 약한 것을 의미한다.

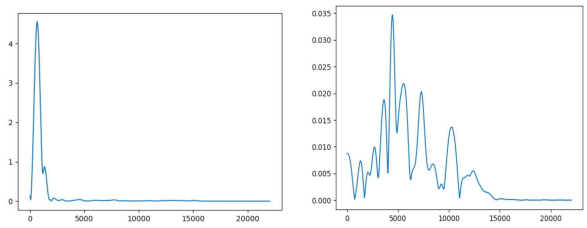

그림 3. 독백의 주파수 성분 (좌), 대화의 주파수 성분 (우)
Fig 3. Frequency of monologue (Left), Frequency of dialogue (Right)

Spectrogram의 세로축 샘플 중 1개를 살펴보면, 주파수의 세기를 그래프를 이용하여 살펴볼 수 있으며, 그림 3과 같이 각각 단일 화자의 음성만을 포함하는 독백과 다수 화자의 음성을 포함하는 대화 구간의 주파수 특징을 확인 할 수 있다. 독백 구간의 경우 대부분 특정한 주파수에 분포하고 있고 대화 구간에서는 다수 화자의 음성이 겹치게 되어 독백의 경우보다 조금 더 다양한 영역의 주파수에서 세기가 강하게 나오는 것을 확인할 수 있다.

2-2. Deep learning-based audio processing

Spectrogram은 1차원의 오디오 신호를 일정한 길이의 구간으로 나눈 다음, 각 구간에 대한 Fourier Transform을 적용하여 가로축에는 시간정보를 세로축에는 주파수의 크기를 dB 단위로 표현한 2차원 이미지로 표현할 수 있으며, 이미지 처리에 사용되는 CNN을 기반으로 하는 딥러닝 프로그램의 학습 데이터로 사용하는 것이 가능하다. [3]

2-2-1. Speech-To-Text (STT)

STT는 음성신호를 문자로 변환하는 기술로 음성인식이라고도 한다. STT 기술은 과거 논리적인 프로그래밍부터 시작하여 통계적 기계학습 방식을 거쳐 현재는 딥러닝을 활용하여 인식하는 방식으로 빠른 속도로 발전하고 있다. 개념적으로 음성 신호를 문자 기호로 해석한다는 차원에서 음성 인식 알고리즘을 디코더(decoder)라고 부르기도 한다.

2-2-2. Singing Voice Separation (SVS)

SVS는 가수의 음성이 함께 녹음된 오디오와 음성을 녹음하기 전 악기의 소리만 녹음된 2가지 종류의 오디오 신호를 사용하여 음성과 악기 소리가 함께 있는 오디오 신호로부터 음성 신호와 악기 신호를 분리하도록 학습된 딥러닝 네트워크이다. 오디오 신호를 Spectrogram으로 변화하고 스펙트럼 변화를 감지함으로써 spectral이급격히 변하는 시점을 기점으로 삼아 음원을 조각 낸

다음에 학습된 정보를 사용하여 음성 분리를 진행하고, 분리된 Spectrogram 데이터를 오디오 신호로 복원하는 과정을 통해 음원 분리 과정이 마무리된다.[4]

3. 음성 주파수 기반 노이즈 필터링

SVS를 사용해 오디오 신호가 음악과 음성으로 분리되는 과정에서 완전히 제거되지 못한 잔여 잡음 남기도 하고 분리된 음원을 재구성하는 과정에서 왜곡되어 새로운 잡음이 생겨나는 한계가 존재하기 때문에 STT 인식률을 높이기 위해서는 이러한 잡음을 제거하는 과정이 필요하다. 따라서 본 논문에서는 비음성 요소에 의해 STT 인식률이 저하되는 문제를 개선하기 위하여 주파수 영역의 correlation을 사용하는 잔여 잡음 제거 필터를 제안한다.

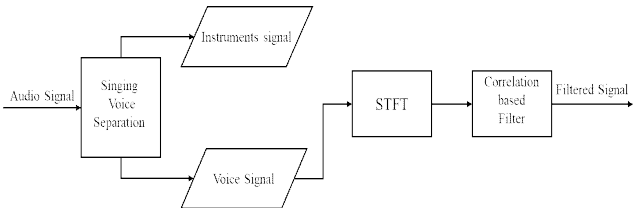

그림 5. 음성 주파수 기반 노이즈 필터링 구조도
Fig 5. Structure of voice frequency-based noise filtering

음성 주파수 기반 노이즈 필터링 구조도는 그림5와 같다. 음악과 음성 신호가 혼합된 오디오 신호로부터 SVS를 사용하여 음성 신호를 분리한다. STFT를 분리된 음성 신호에 적용하여 주파수 영역으로 변환하는 것으로 음성 신호의 각 구간별 주파수 성분을 알아낼 수 있으며, 주파수 영역에서 correlation 연산을 통해 구현한 필터를 이용하여 음원을 필터링 한다.

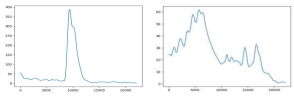

그림 6. 동일 화자 독백의 correlation 결과(좌) 다른 두 화자의 correlation 결과(우)
Fig 6. Correlation result of same speaker's voice signal (left) and different speaker's voice signal (right)

그림6은 단일 화자 독백 신호를 기준으로 동일한 화자의 독백과 다른 두 화자의 대화 구간의 주파수에 correlation 연산을 사용한 결과이다. 동일한 화자의 독백은 특정 주파수 구간에 높은 값을 가지는 결과가 모이고, 여러 화자의 대화는 넓은 범위에 값이 분포되어 전체적으로 낮은 값이 나오는 것을 확인 할 수 있다.

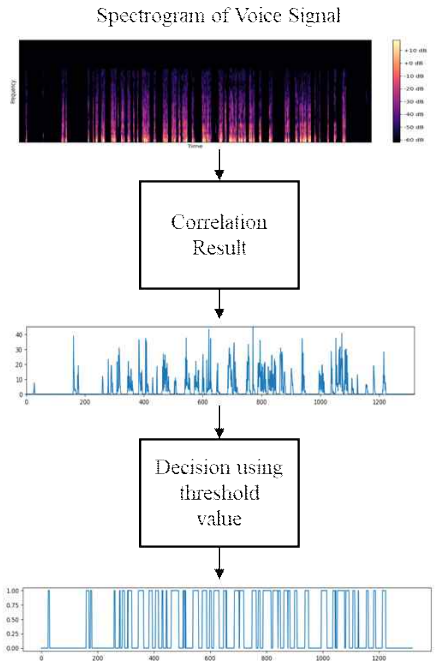

그림 7. Correlation 연산을 이용한 필터 설계
Fig 7. Filter architecture using correlation

주파수 영역의 correlation을 사용하는 잡음 제거 필터는 이러한 결과를 사용하여 그림 7과 같은 과정을 통해 기준으로 사용한 음성 시그널과 각 구간의 correlation 연산 결과의 최대값을 계산하고, 연산 결과의 값이 설정된 한계 값보다 낮은 경우 해당 시간 구간의 음성 신호가 다수 화자의 음성 혹은 잡음을 포함한다고 판단하고 해당 구간의 신호를 제거한다.

4. 실험 결과

본 장에서는 3장에서 제안한 음성 주파수 기반 노이즈 필터링 과정을 실제로 구현하여 검증한 결과를 확인한다. 실험에 사용된 오디오 신호는 음성과 음악 신호를 모두 포함한 콘텐츠를 사용하였으며, MIT 라이센스의 "Vocal Remover" SVS 프로그램을 사용하여 음성 신호를 분리하고, 실험 결과를 평가하기 위해 원본 오디오와 3장에서 제안한 과정으로 처리된 음성 신호의 STT 실험을 진행했다. 본 논문의 실험에서는 전체 신호에서 주파수의 세기가 가장 강한 4개의 신호를 기준 신호로 correlation 연산에 사용하여 그 중 최대값을 가져왔으며, 설정된 한계 값은 correlation연산 결과의 최댓값의 1%에 해당하는 값을 사용했다.

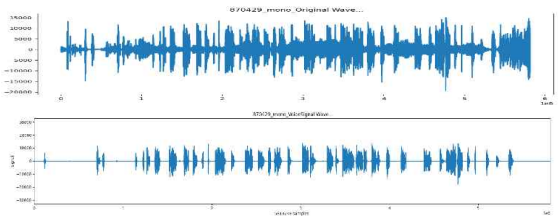

그림 8. 원본신호 파형과 비음성을 제거한 음성파일의 파형
Fig 8. Original signal waveform and voice signal waveform

그림8의 상단의 파형은 음성과 음악이 분리되지 않은 원본의 파형이고 하단의 파형은 SVS를 통해 분리된 음성만의 파형을 나타낸다. 두 파형을 비교해보면 SVS 과정만으로 비음성요소가 제거되면서 깨끗해진 것으로 보인다. 하지만 여전히 두 대사 사이에 제거되지 못한 잡음이 존재하며, 3장에서 제안한 과정을 통해 그림 9와 같이 잡음이 제거되는 것을 확인 할 수 있다.

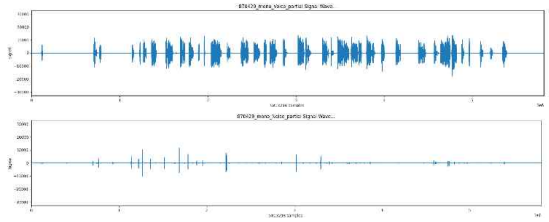

그림 9. 필터링을 거쳐 잡음이 제거된 음성과 제거된 잡음 신호의 파형
Fig 9. Filtered signal waveform and Noise signal waveform

그림9는 필터를 사용하여 추출된 음성에 대한 파형이다. 상단은 필터링 된 음성에 대한 파형이고 하단은 제거한 잡음에 대한 파형이다. 그림8의 잔여 잡음 제거 전의 파형과 비교해보면 대사 사이 남아있던 잔여 잡음이 사라진 것을 확일 할 수 있다.

표 1. 원본 음성과 필터링 음성의 구간별 정확도 비교
Table 1. accuracy of original signal and filtered signal

STT Text ID	Original Confidence	Filtered Confidence
1	0.9586214	0.9989341
2	0.9602871	0.9946679
3	0.9837086	0.9992877
4	0.9516610	0.9894911
5	0.9209005	0.9546975
6	0.9794971	0.9975891
7	0.8585905	0.9604344
8	0.9045419	0.7735642
9	0.9234687	0.9454572
10	0.9962221	0.9992790
11	0.9325716	0.9955888
12	0.9630885	0.9877020
13	0.9542515	0.9934946
14	0.9096265	0.9928818
15	0.9479457	0.9368815
16	0.9364258	0.9480051
17	0.8727486	0.9039986
18	0.9703615	0.9883447
19	x	0.9706504
20	0.9547516	0.9631432
21	0.9345974	0.9972630
22	0.9276739	0.9044928
23	0.8108062	0.8803621
24	0.8496113	0.9982426
25	0.5182038	0.8954286
26	0.9302248	0.9989169

표 1은 원본 오디오 신호와 필터링 과정을 진행한 음성 신호 STT 결과의 각 문장confidence값이다. confidence는 변환된 각 문장의 정확도를 나타내는 지표로 값이 1에 가까울수록 정확도가 높다는 것을 의미한다. 전반적으로 필터링 된 음성이 STT에서 조금 더 높은 정확도를 보이고 있지만, 8번 문장 같은 경우 오히려 정확도가 낮게 나오는 경우가 발생한다. 그러나 오디오 신호를 사람이 직접 듣고 확인한 결과 필터링 과정을 진행한 음성 신호에서 생성된 문장이 더욱 정확한 단어들을 포함한다. 19번 문장의 경우 원본 오디오를 STT에 입력할 경우 목소리로 인식되지 않는 구간이었지만, 제안 기술을 사용한 음성 신호에서는 높은 정확도를 가진 문장을 생성하는 것을 확인했다. STT 결과로 생성된 전체 문장을 비교할 경우 거의 동일한 내용을 포함하기 때문에 표 1의 결과를 통해 STT 결과를 크게 변화시키지 않고 정확도를 상승시키는 결과를 얻었음을 확인 할 수 있다.

5. 결론

본 논문에서는 음성인식 기술이 간섭 신호에 의해 STT 인식률이 저하되는 문제를 해결하기 위해 간섭 신호를 제거하는 방안에 대해서 제안하였다. 주파수영역에서의 correlation 연산을 통해 잔여 잡음의 제거를 위한 필터를 구현하였고 이를 음성에 적용하여 STT 결과를 크게 변화시키지 않고 정확도가 상승하는 결과를 확인할 수 있었다.

그러나 correlation 결과를 판단에 사용할 정확한 한계 값을 실험 과정을 통해 설정하였기 때문에 적절한 한계 값을 찾아내는 방법을 알아내기 위한 추가 연구가 필요하며, 제한된 콘텐츠 데이터의 실험 결과만을 포함하기 때문에 더욱 다양한 오디오 신호를 수집하여 추가적인 검증과 성능 개선을 위한 후속 연구를 통해 STT 성능을 더 높여줄 것으로 기대된다.

참고문헌

[1] 이경님, “음성언어처리기술, 어디까지 왔나”,국립국어원,제27권,제 4호, pages 99-116, 2017
[2] 강경원, 이경민 “스펙트로그램 이미지를 이용한 CNN기반 자동화 기계 고장 진단 기법”, 융합신호처리학회논문지, Vol. 21, No. 3 : 121-126 (accessed Sep. 20, 2020).
[3] Andreas Jansson, Eric Humphrey “SINGING VOICE SEPARATION WITH DEEP U-NET CONVOLUTIONAL” (accessed Oct. 23-27, 2017).