ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 25, No. 4, pp. 223-230, 2019. 4
https://doi.org/10.5626/KTCP.2019.25.4.223

음성-음악 혼재 데이터에서의 음성분리를 위한 확률적 어텐션을 사용한 양방향 LSTM 기반 피치 분류

(Pitch Classification Based on Bidirectional LSTM with Probabilistic Attention for Speech Segregation from Speech-Music Mixtures)

김 한 규 † 장 길 진 †† 박 정 식 ††† 오 영 환 †††† 최 호 진 †††††
(Han-Gyu Kim) (Gil-Jin Jang) (Jeong-Sik Park) (Yung-Hwan Oh) (Ho-Jin Choi)

요 약 Sub-band masking 기반 단일채널 음성분리에서는 음성피치를 추정하여 추정된 피치와 일치하는 주파수 에너지만 통과시키는 필터를 사용하여 배경 잡음으로부터 음성을 분리한다. 음성과 음악은 비슷한 하모닉 구조를 가지고 있어, 음악이 잡음으로 입력될 경우 추정된 피치에 음성 피치와 음악 피치가 공존하게 되며, 이는 음성분리의 성능하락으로 연결된다. 따라서 음성-음악 혼재 데이터에서의 효과적인 음성분리를 위해 음성 피치와 음악 피치를 분류해야 한다. 본 연구에서는 양방향 LSTM을 사용하는 음성/음악 피치 분류 방법을 제안하였으며, 양방향 LSTM의 성능을 향상시키기 위해서 확률적 어텐션 레이어 구조를 제안하였다. 또한 피치 분류 결과로부터 자연스러운 음성분리 결과를 얻기 위해 음악 에너지가 제거된 음성분리 마스크 생성 기법을 제안하였다. 실험결과 확률적 어텐션 기반 양방향 LSTM이 다른 방법에 비해 더 좋은 음성분리 성능을 보여주었다.

키워드: 음성분리, 피치 분류, 양방향 LSTM, 확률적 어텐션

Abstract Speech segregation based on sub-band masking extracts speech signals from audio mixtures via estimation of speech pitch and conservation of signals compatible with the estimated pitch. As speech and music exhibit similar harmonic structures, speech pitch and music pitch coexist in the estimated pitch when speech-music mixture is used as the input, which leads to performance degradation. In order to overcome this limitation, we propose pitch classification using bidirectional

· 본 연구는 한국전력공사의 2018년 착수 에너지 거점대학 클러스터 사업에 의해 지원되었음(과제번호:R18XA05). 또한, 본 연구는 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2017M3C1B6071400)
· 이 논문은 2018 한국컴퓨터종합학술대회에서 '확률적 어텐션을 사용한 양방향 LSTM 기반 음성/음악 피치 분류'의 제목으로 발표된 논문을 확장한 것임

† 비 회 원 : 네이버 Search&Clova 연구원
hangyu.kim@navercorp.com
†† 정 회 원 : 경북대학교 IT대학 전자공학부 교수
gjang@knu.ac.kr
††† 비 회 원 : 한국외국어대학교 ELLT학과 교수
parkjs@hufs.ac.kr
†††† 종신회원 : 한국과학기술원 전산학부 명예교수
yhoh@kaist.ac.kr
††††† 종신회원 : 한국과학기술원 전산학부 교수(KAIST)
hojinc@kaist.ac.kr
(Corresponding author임)

논문접수 : 2018년 8월 28일
(Received 28 August 2018)
논문수정 : 2019년 2월 14일
(Revised 14 February 2019)
심사완료 : 2019년 2월 14일
(Accepted 14 February 2019)

정보과학회 컴퓨팅의 실제 논문지 제25권 제4호(2019. 4)

LSTM. The probabilistic attention layer is also proposed to improve the bidirectional LSTM. Further, musical energy removal for segregation mask generation is also proposed in order to obtain naturally segregated speech with pitch classification. The experiment results show that the proposed pitch classification using bidirectional LSTM based on probabilistic attention outscores other speech segregation methods.

Keywords: speech segregation, pitch classification, bidirectional LSTM, probabilistic attention

1. 서 론

일상생활에서 수집할 수 있는 음성 신호에는 여러 가지 잡음이 섞이게 된다. 잡음이 섞인 음성을 음성인식에 사용할 경우 음성인식의 성능이 저하된다. 따라서 잡음 환경에서도 정확한 음성인식을 진행하기 위해 음성인식의 전처리로 잡음을 제거해야 한다. 음성분리는 음성-잡음 혼재 데이터로부터 깨끗한 음성신호를 추출하는 과정으로써 음성분리의 결과는 음성인식기에 사용되어 음성인식 정확도를 높이게 된다[1].

음성 분리 문제를 해결하기 위해 다양한 방법들이 제안되었었다. Matrix factorization기반 음성 분리는 오디오 스펙트로그램에 반복되어 존재하는 정보를 사용하여 음성을 분리한다[2]. 이 방법은 스펙트로그램에서 반복되어 나타나는 정보를 사용하기 때문에 음성이나 잡음의 특성이 복잡할 경우 원하는 소리를 제대로 분리하지 못하게 된다. Deep Clustering기반 음성분리는 심층 신경망을 사용한 음성분리 방법이다[3]. 이 방법은 신경망을 훈련하기 위해 많은 데이터를 필요로 하며, 신경망이 훈련하지 못한 잡음은 제거하지 못하는 단점이 있다.

단일채널 블라인드 음성분리는 잡음에 대한 사전정보가 없는 상황에서 마이크 하나로부터 입력된 혼재데이터로부터 음성을 추출하는 과정이다. 단일채널 블라인드 음성분리는 특별한 조건 없이 바로 실제 응용에 적용이 가능한 장점이 있다[4]. Sub-band masking은 단일채널 블라인드 음성분리에서 가장 널리 사용되는 기법이다[1]. 이 방법에서는 음성 피치를 추정하여 음성피치와 일치하는 주파수 대역의 신호만 통과시키는 필터를 사용하여 음성분리를 진행한다. 이 방법은 피치가 존재하지 않는 일반 잡음이 섞인 음성데이터에서는 잘 동작하나, 음성-음악 혼재 데이터에서는 음성 피치와 음악 피치가 동시에 존재하기 때문에 피치 추정 알고리즘이 음성 대신 음악 피치를 검출하는 경우가 생겨 음성 분리가 잘 되지 않는다. 따라서 음성-음악 혼재 데이터에서의 효과적인 음성분리를 위해 음성/음악 피치 분류가 꼭 필요하다[2].

본 연구는 [5]의 연구에서 제안하였던 순환신경망 기반 피치 분류 알고리즘을 sub-band masking기반 음성분리에 적용하는 음성-음악 혼재데이터에서의 음성분리 기법을 제안한다. 순환 신경망은 최근에 자연언어 처리, 음성 처리 등 다양한 분야에서 사용되고 있는 시계열 데이터 처리 알고리즘이며, 다른 알고리즘 대비 좋은 성능을 보여주고 있다. [5]에서는 시계열 신호인 연속 피치열을 양방향 LSTM 순환 신경망을 사용하여 혼재데이터로부터 추출한 피치를 분석해서 피치를 음성 혹은 음악으로 분류하는 방법을 제안하였다. 본 연구에서는 순환 신경망이 제공하는 피치 분류 결과를 사용하여 자연스러운 음성을 분리하기 위해 음악에너지가 제거된 음성분리 마스크 생성기법을 제안하였고, 이를 기존의 sub-band masking 기반 음성분리에 적용하였다. 음성과 비슷한 특성을 가지는 음악이 잡음으로 섞이더라도, 음악 에너지를 검출하여 사전에 제거한 이후에 얻은 오디오 스펙트로그램만을 사용하여 음성분리를 진행함으로써, 음성분리 성능 향상을 기대할 수 있다.

본 논문은 다음과 같이 구성되었다. 2장에서는 sub-band masking기반 단일채널 블라인드 음성분리 알고리즘에 대해서 소개하며, 3장에서는 [5]에서 제안하였던 순환신경망 기반 피치분류 방법과 본 연구에서 제안하는 음악 에너지가 제거된 음성분리 마스크 생성 기법에 대해서 설명한다. 4장에서 실험결과를 보여주고 이에 대해 분석을 진행하며 5장에서 결론을 맺는다.

2. Sub-band Masking 기반 음성분리

Sub-band masking기법은 4가지 부분으로 나뉘어진다: 신호분리, 음성피치 추정, 음성분리 마스크 생성, 음성 재합성. 신호분리 과정에서는 입력된 오디오 신호를 여러 주파수 대역의 신호로 분리하게 된다. 본 연구에서는 신호분리를 수행하기 위해 사람귀의 특성을 잘 반영한다고 알려진 Gammatone 필터뱅크를 사용하였다[6]. 음성피치 추정과정에서는 분리된 신호에 대해 분석을 진행하여 음성피치를 추정하게 된다. 본 연구에서는 잡음 환경에서도 강인한 음성피치 추정 알고리즘인 앙상블 칼만필터 기반 음성피치 추정 방법을 사용하였다[7]. 음성분리 마스크는 추정된 피치와 신호분리 결과를 비교하여서 추정된 피치와 일치하는 부분은 1로 마스킹하고 일치하지 않는 부분은 0으로 마스킹 하는방식으로 생성된다. 이렇게 생성된 음성분리 마스크는 혼합 오디오 신호에서 음성의 에너지를 강조하고 잡음의 에너지를 억제할 수 있다. 마지막으로 마스킹된 신호를 Weintraub

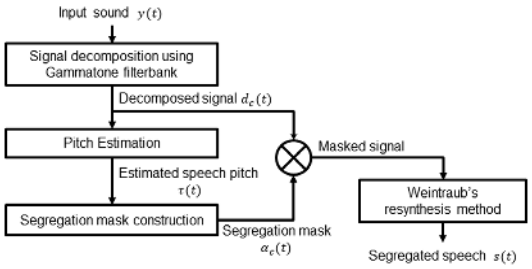

그림 1 Sub-band masking기반 음성분리

Fig. 1 Speech segregation based on sub-band masking

방식을 사용하여 재합성을 진행하게 된다[8]. 이러한 sub-band masking기반 음성분리 과정은 그림 1과 같이 표현된다.

3. 피치 분류를 사용한 음성분리

[7]에서 설명된 sub-band masking기반 음성분리는 일반적인 잡음에 대해서는 잘 동작하나, 음악이 잡음으로 들어오는 경우에는 잘 동작하지 않는다. 이는 음악은 음성과 비슷한 하모닉 구조를 가지고 있어 음성피치 추정 단계에서 음악의 피치와 음성의 피치가 같이 추정되기 때문이다.

음성과 음악 모두 피치를 가지고 있지만, 두 종류의 피치는 다른 특성을 가지고 있다. 음성은 복잡하게 발화되기 때문에 음성의 피치는 안정적이지 않고 계속 움직이는 특성을 보이는 반면, 음악은 특정 음표를 악기가 연주하는 방식으로 만들어지기 때문에 짧은 시간 안에 피치가 변하지 않는다는 특징을 가지고 있다. 이러한 특성은 그림 2에서 확인할 수 있다.

본 연구에서는 음성과 음악 피치의 다른 특성을 딥러닝 알고리즘으로 모델링하는 방법을 제안한다. 여러 딥러닝 알고리즘 중 시계열 처리 분야에서 가장 좋은 성능을 보이고 있는 양방향 LSTM을 사용해서 피치의 특성을 모델링하였다[9,10].

3.1 어텐션을 사용한 양방향 LSTM 기반 피치 분류

양방향 LSTM은 최근 다양한 시계열 데이터 처리 분야에 적용되어 가장 좋은 성능을 보이고 있는 순환 신경망 구조이다. 양방향 LSTM은 정방향 LSTM과 역방향 LSTM으로 구성되어, 정방향 LSTM은 과거 정보가 현재에 미치는 영향을 모델링하고, 역방향 LSTM은 미래 정보가 현재에 주는 영향을 모델링 한다[9].

양방향 LSTM의 state를 효과적으로 취합하기 위해 어텐션 알고리즘이 주로 사용된다[11]. 어텐션 레이어는 모든 state에 대해서 가중치를 계산하고, 이 가중치를 대응되는 state에 곱하여 합을 구하는 방식으로 state정보를 취합한다. 이는 그림 3과 같이 표현된다. 그림 3에서 FC_Att는 어텐션 가중치를 계산하기 위한 fully connected layer이며, 그림 3의 가운데 있는 softmax layer는 가중치의 총합이 1이 되도록 맞춰주기 위해서 사용된다. 최종적으로 네트워크는 입력 피치열이 음성일 확률과 입력 피치열이 음악일 확률로 이루어진 2차원 벡터를 출력하게 된다.

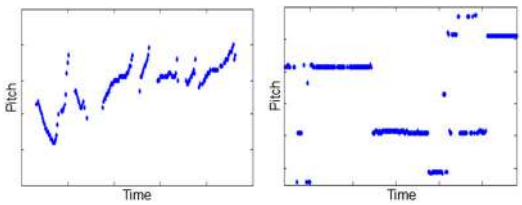

그림 2 음성 피치와 음악 피치 예시

Fig. 2 Examples of speech pitch and music pitch

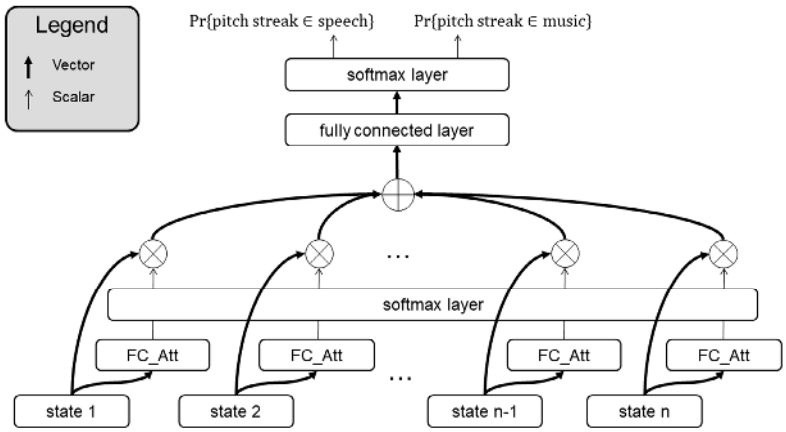

그림 3 어텐션 기반 피치 분류 양방향 LSTM 구조

Fig. 3 Pitch classification using bidirectional LSTM based on attention

3.2 확률적 어텐션을 사용한 양방향 LSTM 기반 피치 분류

어텐션은 대부분의 응용에서 좋은 성능을 보이고 있지만, 기본적으로 softmax함수를 사용하여 state의 가중치를 계산하기 때문에, 모든 state를 유기적으로 종합하지 못하고, 여러 state중 한가지 state에만 집중하는 단점을 가지고 있다[11]. 이러한 집중 현상은 그림 4에서 확인할 수 있다. 하지만 음성-음악 피치 분류는 어느 특정 피치 값만 보고 결정할 수 없고, 모든 피치값의 정보를 종합해서 피치 분류를 진행하여야 한다. 따라서 본 연구에서는 특정 state에 집중하는 기존 어텐션 대신 여러 state를 동시에 고려할 수 있는 확률적 어텐션을 사용한 양방향 LSTM을 제안한다.

확률적 어텐션 레이어의 구조는 그림 5와 같다. 확률적 어텐션에서는 각자 state로부터 미리 네트워크의 출력값을 계산한다. 그 이후 각 출력값에 적절한 가중치를 곱해서 합을 취하는 방식으로 출력을 취합한다. 가중치를 계산하는 과정에서 softmax함수가 쓰이지 않기 때문에, 가중치가 특정 state에 집중되는 현상을 피할 수 있다.

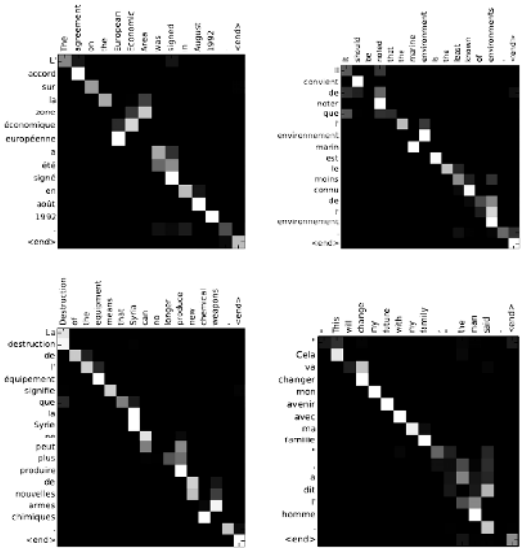

그림 5 어텐션 결과[11]
Fig. 5 Result of attention [11]

3.3 음악 에너지가 제거된 음성분리 마스크 생성

음성-음악 혼재데이터에서 음성 신호만 추출하기 위해서 음성으로 분류된 피치만 사용하여 음성분리 마스크를 생성해야 한다. 가장 단순하게는 음성/음악 피치 분류 결과에서 음성 피치로 분류된 피치만 사용하여 음성분리 마스크 생성하는 방법을 사용할 수 있다. 하지만 위에 서술한 방식으로 음성분리 마스크를 생성하게 된다면, 음성피치로 분류된 구간만 분리가 진행되고, 음악 피치로 분류된 구간은 묵음으로 만들어지게 된다. 하지만 음악 피치로 분류된 구간은 음악 에너지가 음성 에너지보다 강하다는 의미이며, 음성 에너지가 여전히 존재하는 구간이기 때문에, 위에 서술한 방법으로 음성분리 마스크를 생성할 경우, 음성이 존재함에도 불구하고 묵음으로 처리되는 구간이 많아져서 합성된 소리가 부자연스럽게 들리게 된다. 실제로 잡음이 많이 제거되었더라도 음성에 불필요한 왜곡을 많이 만들어내기 때문에 단순히 피치 분류 결과를 직접 사용하여 음성분리 마스크를 생성하는 방법은 오히려 음성분리 성능 하락을 유발할 수 있다.

위에서 서술한 문제점을 해결하기 위해, 본 연구에서는 음악 에너지 제거 후 음성분리 마스크를 생성하는 방법을 제안하였다. 이 방법에서는 음성/음악 피치 분류 결과로부터 음악의 피치를 사용해서 원래 입력된 혼재

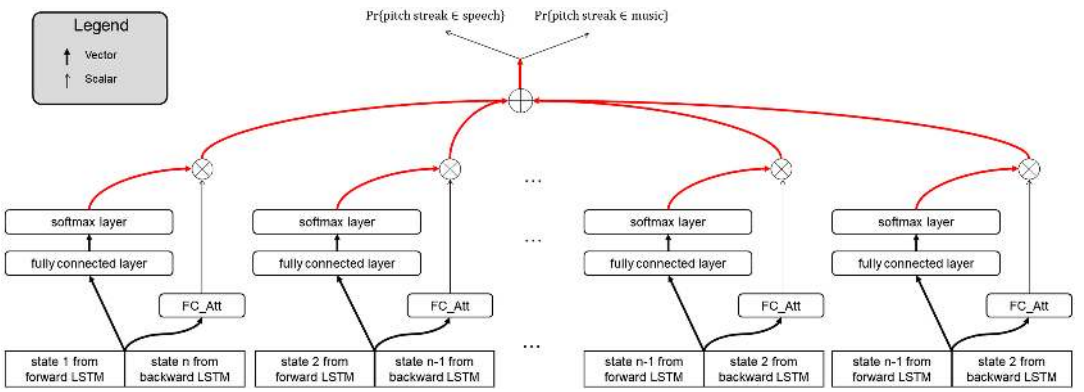

그림 4 확률적 어텐션 기반 피치 분류 양방향 LSTM 구조(빨간색 선은 확률로 이루어진 2차원 벡터)
Fig. 4 Pitch classification using bidirectional LSTM based on probabilistic attention (red lines represents 2-dim vector composed of probabilities)

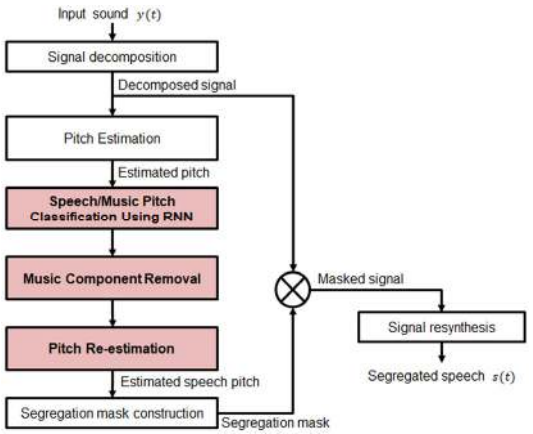

그림 6 피치 분류가 적용된 음성분리

Fig. 6 Speech segregation with pitch classification

데이터로부터 음악 피치와 일치하는 에너지를 분리해서 제거하게 된다. 음악 에너지를 제거한 이후에, 남은 소리에 2장에서 설명한 sub-band masking 기법, 즉 음성 피치 추정, 음성분리 마스크 생성 및 음성 재합성을 수행하여 잡음이 제거된 음성을 추출하게 된다. 음악이 강한 부분에서도 음성에너지를 분리해내기 위해, 음악에너지를 제거한 후 음성피치를 다시 추정하는 방법을 사용하였다. 이러한 방법은 그림 6의 블록 다이어그램으로 표현된다. 그림 6에서의 빨간색으로 표시된 부분이 2장에서 서술한 sub-band masking 기반 음성 분리 알고리즘에서 개선된 부분이다. 이 방법에서는 음악 에너지를 먼저 제거하고 음성 피치를 다시 추정하기 때문에 음악 에너지에 묻혀있는 음성을 효과적으로 분리할 수 있으며, 불필요한 묵음구간이 생겨서 음성이 끊겨서 들리는 왜곡을 방지할 수 있다.

4. 실험결과

4.1 피치 분류

본 연구에서는 음성 피치를 추출하기 위해 TIMIT database의 음성 파일을 사용하였으며, 추가로 다양한 종류의 악기로 연주된 음악 파일을 수집하여 음악 피치를 추출하였다.

음성/음악 피치 분류 실험은 음성-음악 혼재 데이터에서 추출한 피치열에 대해서 진행하였다. 피치 분류 실험에는 비교 실험으로 LSTM과 어텐션 기반 양방향 LSTM (BD-A)가 사용되었으며, 제안한 방법인 확률적 어텐션 기반 양방향 LSTM (BD-PA)도 실험에 사용되었다.

성능 평가 지표로는 정확히 분류된 피치열의 갯수만 통계한 분류 정확도와, 피치열의 길이까지 고려하여 통계한 가중 분류 정확도 두가지가 사용되었다.

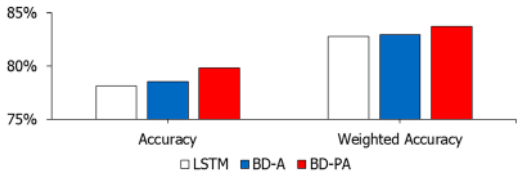

그림 7 피치 분류 실험결과

Fig. 7 Experiment results of pitch classification

실험 결과는 그림 7과 같다. 실제로 음성 분리 성능에 가장 많은 영향을 주는 요소는 피치열의 길이까지 고려한 성능지표인 가중 분류 정확도이다. 실험결과에서 볼 수 있듯이, 해당 지표에 대해서 BD-PA가 LSTM과 BD-A보다 더 좋은 피치 분류 성능을 보임을 확인할 수 있었다. 따라서 음성 분리에서 BD-PA가 다른 알고리즘보다 더 좋은 성능을 보일 것을 기대할 수 있다.

4.2 음성분리

본 음성분리 실험에서는 [7]논문의 앙상블 칼만필터 기반 음성분리 알고리즘(EnKF)을 비교 실험대상으로 사용하였다. 또한 4.1장의 피치분류 알고리즘을 EnKF에 적용하여 EnKF+LSTM, EnKF+BD-A, EnKF+BD-PA에 대한 음성분리 실험을 진행하였다.

음성분리 실험을 위해 TIMIT database에서 선별한 20개의 voiced 음성 데이터가 사용되었으며, 6종류의 일반잡음(N1: 1kHz pure tone, N2: white noise, N3: noise bursts, N4: Cocktail party noise, N5: Siren, N6: Telephone bell)과 6종류의 음악(M1: Piano+drum, M2: Electric guitar+drum, M3: Quiet piano, M4: Acoustic guitar, M5: Muted jazz trumpet, M6: Acoustic guitar+drum)이 사용되었다. 다양한 잡음환경을 시뮬레이션하기 위해서 각 잡음은 -5dB, 0dB, 5dB의 크기로 음성과 혼합한 이후에 음성분리 실험을 진행하였다.

2장 및 3.3장에서 설명된 음성분리 방법에는 피치의 신뢰도 결정을 위한 threshold, 스펙트로그램과 피치의 일치 여부를 판단하기 위한 threshold, 앙상블 칼만 필터의 전이 공분산 행렬 및 관측 공분산 행렬 등 다수의 초매개변수들이 사용된다. 이런 초매개변수들을 임의로 결정할 경우 공정한 성능평가가 이루어지지 않을 수 있다. 본 연구에서는 객관적인 성능평가를 위해 교차검증 방법을 사용하여 실험을 진행하였다. 이 교차검증 방법에서는 전체 데이터를 서로 겹치치 않는 검증셋과 평가셋으로 분리하여, 탐욕 알고리즘을 사용하여 검증셋에서 가장 좋은 성능을 보이는 초매개변수값을 찾은 후, 이 초매개변수값으로 평가셋에 대해 성능평가를 진행한다. 평가셋과 검증셋은 그림 8과 같이 나누어지게 된다. 20개의 음성은 10개씩 묶어서 두 개의 셋으로 나누고, 잡음 혹은 음악 한개와 음성셋 하나가 평가셋이 되고 나머지

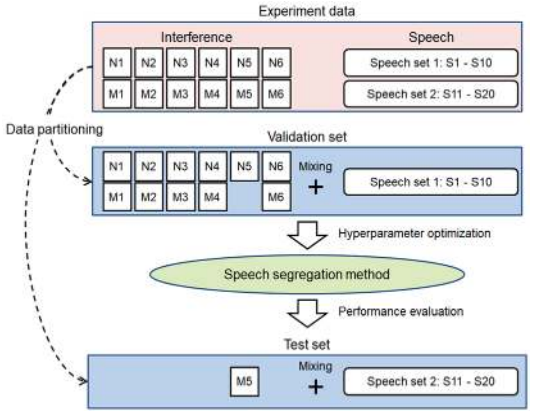

그림 8 교차검증 실험에서 검증셋/평가셋을 구성하는 방법
Fig. 8 Composition of validation and test sets in the cross-validation experiment

데이터가 검증셋이 된다. 모든 데이터가 평가가 한번씩 될 때까지 평가셋/검증셋을 바꿔가면서 실험하게 된다.

음성분리 성능평가에는 SNR(source-to-noise ratio), ELR(Energy Loss Ratio) NRR(Noise Residue Ratio)를 사용하였다. SNR은 분리된 음성에서 음성과 잡음에너지의 비율로, SNR이 높을수록 우수한 음성분리 성능을 뜻하게 된다. ELR은 음성 분리과정 중 손실된 음성에너지의 비율을 뜻하며, NRR은 음성 분리를 진행하였음에도 불구하고 분리된 음성에 여전히 남아있는 잡음에너지의 비율을 뜻한다. ELR과 NRR모두 작을수록 좋은 성능을 의미한다.

음성분리의 SNR 결과는 표 1, 2에서 확인할 수 있으며, ELR, NRR결과는 표 3, 4, 5, 6에서 확인할 수 있다. 표 3, 4에서는 기존의 EnKF방법이 제안한 방법보다 더 좋은 ERL을 보이고 있으며, 표 5, 6에서는 제안한 방법이 기존 EnKF방법보다 월등히 좋은 NRL을 보이고 있다. 본 연구에서 제안하는 피치 분류 알고리즘은 초기에 추정된 피치에 대해 음성/음악 피치 분류를 진행하여, 음악 피치로 생각되는 피치는 제외하고 음성 분리를 진행하게 된다. 하지만 4.1장의 결과에서 보이듯이 피치 분류의 정확도가 100%가 아니기 때문에 음성 피치를 음악 피치로 잘못 분류할 가능성이 있다. 이러한 경우, 실제 존재하는 음성 피치가 최종 음성분리에서 제외되었기 때문에 음성 신호 손실이 발생하여 ERL이 높아질 수 있다. 하지만 일반적인 경우 피치 분류 알고리즘은 음악 피치를 제대로 찾아서 제거하기 때문에 분리된 신호에 음악이 들어갈 확률이 낮아지게 되고 따라서 NRL이 많이 개선되게 된다. 표 3, 4, 5, 6에서 확인할 수 있듯이 NRL에서의 성능 향상이 ELR에서의 성능 손실보다 훨씬 크다. 즉, ELR과 NRL을 종합적으로 보았을 때, 본 연구에서 제안한 피치 분류 알고리즘을 적용할 경우 음성 분리의 성능이 향상된다. 이는 표 1, 2의 SNR결과에서도 확인할 수 있다.

표 1 음성-잡음 혼재데이터의 음성 분리 SNR결과 (dB)
Table 1 SNRs of speech segregation from speech-noise mixtures (dB)

	EnKF	EnKF+LSTM	EnKF+BD-A	EnKF+BD-PA
N1	5.26	5.56	5.51	5.35
N2	2.50	3.68	3.80	3.79
N3	0.55	0.24	0.27	0.28
N4	1.94	1.76	1.92	1.99
N5	6.60	6.66	6.64	6.56
N6	9.58	8.37	8.19	8.40
Avg.	4.41	4.38	4.39	**4.40**

표 2 음성-음악 혼재데이터의 음성 분리 SNR결과 (dB)
Table 2 SNRs of speech segregation from speech-music mixtures (dB)

	EnKF	EnKF+LSTM	EnKF+BD-A	EnKF+BD-PA
M1	-0.17	1.58	1.55	1.63
M2	2.35	1.89	2.26	2.32
M3	0.86	1.34	1.40	1.26
M4	1.44	1.45	1.52	1.52
M5	2.14	1.87	1.94	2.07
M6	1.18	0.96	1.26	1.31
Avg.	1.30	1.51	1.65	**1.69**

표 3 음성-잡음 혼재데이터의 음성 분리 ELR결과
Table 3 ELRs of speech segregation from speech-noise mixtures

	EnKF	EnKF+LSTM	EnKF+BD-A	EnKF+BD-PA
N1	14.5%	23.6%	19.5%	18.5%
N2	15.4%	14.7%	14.1%	13.9%
N3	3.0%	7.1%	9.0%	8.4%
N4	46.3%	44.7%	43.9%	43.7%
N5	14.7%	13.5%	13.5%	14.2%
N6	3.0%	19.8%	6.9%	6.5%
Avg.	**16.2%**	20.6%	17.8%	17.5%

표 4 음성-음악 혼재데이터의 음성 분리 ELR결과
Table 4 ELRs of speech segregation from speech-music mixtures

	EnKF	EnKF+LSTM	EnKF+BD-A	EnKF+BD-PA
M1	41.3%	45.4%	43.8%	43.9%
M2	42.2%	42.2%	40.9%	39.6%
M3	36.5%	35.3%	39.2%	39.7%
M4	47.5%	51.6%	53.1%	52.9%
M5	36.1%	47.3%	39.6%	37.8%
M6	47.8%	55.0%	53.6%	50.9%
Avg.	**41.9%**	46.1%	45.0%	44.1%

표 5 음성-잡음 혼재데이터의 음성 분리 NRR결과
Table 5 NRRs of speech segregation from speech-noise mixtures

	EnKF	EnKF+LSTM	EnKF+BD-A	EnKF+BD-PA
N1	19.8%	17.8%	16.3%	18.0%
N2	30.2%	21.8%	15.6%	16.2%
N3	47.7%	47.1%	47.0%	46.5%
N4	14.0%	15.1%	13.8%	15.0%
N5	3.6%	4.1%	3.6%	4.2%
N6	0.4%	11.4%	1.6%	0.4%
Avg.	19.3%	19.6%	**16.3%**	**16.7%**

표 6 음성-음악 혼재데이터의 음성 분리 NRR결과
Table 6 NRRs of speech segregation from speech-music mixtures

	EnKF	EnKF+LSTM	EnKF+BD-A	EnKF+BD-PA
M1	51.0%	29.5%	27.1%	28.2%
M2	26.1%	25.7%	22.5%	22.6%
M3	40.8%	32.6%	32.2%	34.5%
M4	36.3%	29.3%	30.8%	30.3%
M5	28.3%	30.2%	28.6%	29.3%
M6	40.3%	34.7%	34.8%	33.9%
Avg.	37.1%	30.3%	**29.3%**	**29.8%**

BD-A와 BD-PA의 음성분리 성능비교에서는, BD-PA가 더 좋은 ELR을 보이나, NRL에서는 BD-A의 성능이 더 높았다. BD-PA가 ELR에서 보인 성능 향상이 NRL에서 보인 성능하락보다 약간 더 높았다. 따라서 ELR과 NRL을 종합하여 비교할 때, BD-PA가 BD-A보다 약간 더 좋은 성능을 보이고 있다. 이는 표 1, 2의 SNR결과와도 일치한다. 음성-잡음 혼재데이터, 음성-음악 혼재데이터 두가지 모두의 경우 BD-PA의 SNR결과가 BD-A의 SNR결과보다 약간 더 좋음을 확인할 수 있다.

본 연구는 음성-음악 혼재 데이터에서의 효과적인 음성분리를 목표로 하기 때문에, 음악 종류별로 어떤 음성분리 결과가 나오는지 좀 더 자세히 분석해 보았다. 표 2에서 확인 할 수 있듯이 제안한 피치 분류 알고리즘은 대부분의 음악 종류에 대해서 좋은 성능을 보이고 있지만, M2와 M5에 대해서는 성능 하락을 보이고 있다. 이미 서술한바와 같이 M2는 electric guitar+drum이고 M5는 muted jazz trumpet이다. 이 두 종류의 음악은 다른 음악과 다르게 악기가 특정 음표를 안정적이게 연주하기 보다는 바이브레이션을 넣어서 연주하기 때문에 피치 일정하게 유지되지 않고 계속해서 변화하는 복잡한 패턴을 보이게 된다. 이러한 복잡한 특성으로 인해서 M2와 M5의 피치는 음성피치와 구분하기 더욱 힘들어진다. 따라서 M2와 M5에서는 피치 분류 알고리즘이 제대로 동작하지 못하여서 피치 분류를 적용할 경우 오히려 음성 분류 성능 하락이 발생한다. 하지만 BD-PA가 LSTM과 BD-A보다 복잡한 특성을 모델링 하기 더 좋은 알고리즘이기 때문에, BD-PA를 적용할 경우 성능하락이 거의 발생하지 않는다. 즉 BD-PA는 평균 SNR에서도 BD-A보다 더 좋은 성능을 보이고 있지만, 피치 분류 알고리즘 자체가 잘 동작하지 않는 음악에서도 성능하락을 최소화 하는 효과를 보이고 있기 때문에, 더 안정적인 알고리즘으로 볼 수 있다.

5. 결 론

본 연구에서는 음성-음악 혼재데이터에서 효과적으로 음성을 분리하기 위해 음성-음악 피치 분류 알고리즘을 제안하였다. 피치 분류에서 음성 피치와 음악 피치의 특징을 효과적으로 모델링하기 위해 확률적 어텐션 기반 양방향 LSTM을 제안하였다. 또한, 피치 분류 결과를 자연스럽게 음성분리에 적용하기 위해 음악 에너지가 제거된 음성분리 마스크 생성 기법을 제안하였다. 피치 분류 실험 및 음성분리 실험결과 제안한 확률적 어텐션 기반 양방향 LSTM이 기존의 방법들보다 더 안정적이고 뛰어난 성능을 보여주었다.

References

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge MA, 1990.

[2] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, pp. 1066-1074, 2007.

[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for spgmentation and separation," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 31-35, 2016.

[4] H. Kim, J. Park, G. Jang, and Y. Oh, "Particle filtering by sigmoidal weight update for speech pitch correction," *Proc. of IEEE International Conference on Systems, Man and Cybernatics*, pp. 2574-2579, 2012.

[5] H. Kim, G. Jang, J. Park, Y. Oh, and H. Choi, "Sppech/music pitch classification based on bidirectional LSTM with probabilistic attention," *Proc. of Korea Computer Congress*, pp. 847-848, 2018. (in Korean)

[6] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, Vol. 15, No. 5, pp. 1135-1150, 2004.

[7] H. Kim, G. Jang, J. Park, and Y. Oh, "Monaural speech segregation based on pitch track correction

using an ensemble Kalman filter," *Proc. of Interspeech*, 2013.
[8] M. Weintraub, "A theory and computational model of auditory monaural sounds separation," *Ph.D. thesis, Stanford University*, 1985.
[9] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classication and recognition," *Proc of International Conference on Articial Neural Networks*, pp. 799-804, 2005.
[10] Y. Cheng, S. Shen, Z. He, W. He, H. Wu, M. Sun, and Y. Liu, "Agreement-based joint training for bidirectional attention-based neural machine translation," *arXiv preprint arXiv:1512.04650*, 2015.
[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

김 한 규
2009년 중국 Tsinghua University 전자정보공학과 졸업(학사). 2011년 한국과학기술원 전산학부 졸업(석사). 2018년 한국과학기술원 전산학부 졸업(박사). 2018년~현재 네이버 Speech팀 연구원. 관심분야는 음성신호처리, 기계학습

장 길 진
1997년 한국과학기술원 전산학과 졸업(학사). 1999년 한국과학기술원 전산학과 졸업(석사). 2004년 한국과학기술원 전산학과 졸업(박사). 2004년~2006년 삼성종합기술원 전문연구원. 2006년~2009년 미국 University of California, San Diego 박사후연구원. 2009년~2014년, 울산과학기술대학교(UNIST) 전기전자컴퓨터 공학부 조교수. 2014년~현재 경북대학교 전자공학부 부교수. 최근 5년 연구실적은 국외전문학술지 20여 편, 국제학술회의 발표 18건. 관심분야는 음성신호처리, 기계학습

박 정 식
2001년 아주대학교 정보및컴퓨터공학부 졸업(학사). 2003년 한국과학기술원 전산학과 졸업(석사). 2010년 한국과학기술원 전산학과 졸업(박사). 2012년~2013년 목원대학교 지능로봇공학과 조교수. 2014년~2017년 영남대학교 정보통신공학과 조교수. 2018년~현재 한국외국어대학교 ELLT학과 부교수. 관심분야는 음성신호처리, 기계학습, 인공지능

오 영 환
1972년 서울대학교 공과대학 전자공학과 졸업(학사). 1974년 서울대학교 교육대학원 교육학과 졸업(석사). 1980년 일본 Tokyo Institute of Technology 정보공학과 졸업(박사). 1981년~1985년 충북대학교 공과대학 조교수. 1985년~2012년 한국과학기술원 전산학부 교수. 2006년 대한음성학회 회장. 2009년~2014년 국방소프트웨어설계 특화연구센터 센터장. 2012년~현재 한국과학기술원 전산학부 명예교수. 관심분야는 음성신호처리

최 호 진
1982년 서울대학교 컴퓨터공학 학사 졸업 1985년 영국 Newcastle University 검퓨터과학 석사 졸업. 1995년 영국 Imperial College London 인공지능학 박사 졸업 1982년~1989년 ㈜데이콤 정보통신연구소 선임연구원. 1997년~2002년 한국항공대학교 교수. 2002년~2009년 한국정보통신대학교 교수. 2009년~현재 KAIST 전산학부 교수. 2018년~현재 스마트에너지인공지능연구센터 센터장. 관심분야는 인공 지능, 자연어처리, Biomedical informatics