Journal of The Institute of Electronics and Information Engineers Vol.54, NO.1, January 2017 https://doi.org/10.5573/ieie.2017.54.1.121
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2017-54-1-14

비음수 텐서 분해 및 은닉 마코프 모델을 이용한 다음향 환경에서의 이중 채널 음향 사건 검출

(Dual-Channel Acoustic Event Detection in Multisource Environments Using Nonnegative Tensor Factorization and Hidden Markov Model)

전 광 명*, 김 홍 국**

(Kwang Myung Jeon and Hong Kook Kim©)

요 약

본 논문에서는 다음향(multisource) 환경에서의 음향 사건 검출 정확도를 높이기 위해 비음수 텐서 분해(nonnegative tensor factorization, NTF)와 은닉 마코프 모델(hidden Markov model, HMM)을 이용한 이중 채널 음향 사건 검출 방법을 제안한다. 제안된 방법은 먼저 이중 채널 입력 신호들에 NTF 기법을 적용하여 얻은 각 음향 사건 별 채널 이득을 활용하여 다수의 음향 사건들을 검출한다. 그러고 나서, 채널 이득에 의해 검출된 음향 사건의 발생 여부를 검증하기 위하여 채널 이득을 우도 가중치로 활용하는 HMM 기반의 우도비 검증을 수행한다. 제안된 방법의 검출 정확도를 평가하기 위하여 다양한 잡음과 사건 간 중첩 밀도를 고려하는 다중 사건 발생 환경에 대한 F-measure를 측정하였고, 기존의 혼합 가우시안 모델 및 비음수 행렬 분해 기반의 음향 사건 검출 방법들과 비교하였다. 실험 결과, 제안된 방법이 기존 방법들에 비하여 모든 실험 조건에서 높은 정확도를 보였다.

Abstract

In this paper, we propose a dual-channel acoustic event detection (AED) method using nonnegative tensor factorization (NTF) and hidden Markov model (HMM) in order to improve detection accuracy of AED in multisource environments. The proposed method first detects multiple acoustic events by utilizing channel gains obtained from the NTF technique applied to dual-channel input signals. After that, an HMM-based likelihood ratio test is carried out to verify the detected events by using channel gains. The detection accuracy of the proposed method is measured by F-measures under 9 different multisource conditions. Then, it is also compared with those of conventional AED methods such as Gaussian mixture model and nonnegative matrix factorization. It is shown from the experiments that the proposed method outperforms the convectional methods under all the multisource conditions.

Keywords : Acoustic event detection, dual-microphone signal, NTF, HMM, multisource environment

Ⅰ. 서 론

사고 및 돌발 상황에 대한 검출기술은 그 필요성이 꾸준히 제시되어 왔다.[1~2] 대부분의 사고 감지 기술은 시각 사건 검출(visual event detection, VED) 기술을 근간으로 하여 보안 및 교통 분야에서 활발히 응용되고 있다.[1] 하지만 VED는 야간, 장애물, 그리고 시야각의 한계 등 다양한 제약조건으로 인해 실제의 사건을 인지하지 못하는 경우가 빈번히 발생할 수 있다.[2] 이러한 VED의 한계점에 대한 보완 및 대체 기술로써 음향 사건 검출(acoustic event detection, AED) 기술의 중요성이 부각되고 있다.[2] 음향 사건 검출은 음향 인식 기술의 한 종류로써, 특정 사건에 수반되는 관련 음향을 인식하여 해당 사건의 발생 여부를 검출하는 기술이다.[2] 이러한 음향 사건 검출은 다음과 같은 장점을 지닌다.

* 학생회원, ** 평생회원, 광주과학기술원 전기전자컴퓨터공학부 (School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology)
© Corresponding Author(E-mail : hongkook@gist.ac.kr)
※ 이 논문은 국토부의 재원으로 국토교통과학기술진흥원의 지원을 받아 수행된 연구사업임 (16TBIP-C111209-01)
Received ; September 19, 2016 Revised ; December 13, 2016
Accepted ; December 16, 2016

첫째, 음향의 비교적 낮은 정보량으로 인해 취득과 전송이 용이하다. 둘째, 넓은 지역에서 정보를 쉽게 취득할 수 있다. 그리고 어두운 환경에서도 활용이 가능하다.[2] 이러한 특징들에 힘입어 음향 사건 검출은 감시, 멀티미디어 검색, 보조 기술, 그리고 상황 인지 등 다방면에 활용될 수 있다.[2~6]

음향 사건 검출과 관련된 초기 연구는 특정한 단일 음향 사건들, 예컨대 비명,[6] 총소리,[7] 기계음,[8] 음성/음악 전환,[9] 기침[10] 등을 검출하는데 집중하였다. 이러한 기술들은 감시[6~7]나 멀티미디어 검색[11] 등, 적은 수의 음향 범주를 지니는 분야들에 먼저 활용되었다. 또한 다수의 음향 사건을 지속적으로 기록할 수 있다면 마이크로폰의 인근이 있는 사람들의 행동 양식 등의 정보를 활용하여 다양한 분야에 응용할 수 있다.[12]

음향 사건 검출을 위해 다양한 기술적 접근이 시도되어 왔다.[2~4] 초기의 방법으로는 가우시안 혼합 모델(Gaussian mixture model, GMM) 판별기와 멜-주파수 캡스트럼 계수(mel-frequency cepstral coefficient, MFCC)를 이용하여 다양한 음향 사건을 검출하였다.[2~3] 이러한 방법은 구현 복잡도가 낮다는 장점이 있지만, 동시에 다수의 음향과 잡음이 발생하는 다음향(multisource) 환경에서는 한 시점에 한 가지 검출 결과를 낼 수밖에 없기 때문에, 그 검출 정확성이 떨어지는 한계가 있다.[6]

다음향 환경에서 GMM 기반 방법이 지니는 한계를 극복하기 위해, 비음수 행렬 분해(nonnegative matrix factorization, NMF)를 이용한 음원 분리를 접목하는 방법이 제안되었다.[2, 5~6] 이 방법은 입력 신호로부터 각각의 음향 사건과 배경 잡음을 분리할 수 있기 때문에 다수의 음향 사건을 독립적으로 검출할 수 있다는 장점을 지닌다. 구체적으로, 우선 NMF 음원 분리 기술을 이용하여 음향 사건들로부터 잡음을 분리해 낸다. 다음으로, 분리된 음향 사건들은 은닉 마코프 모델(hidden Markov model, HMM) 기반의 비터비 디코딩(Viterbi decoding)[2, 5] 혹은 우도비 검정(likelihood ratio test)[6]을 통하여 해당 사건으로 판별된다. NMF 음원 분리와 HMM을 이용한 음원 분류의 조합은 다음향 환경에서의 음향 사건 검출 성능을 크게 개선시켰다.[5~6]

하지만 검출하고자 하는 음향 사건의 주파수 분포가 배경 잡음과 유사한 경우, NMF 기법은 정확한 음향 사건 검출에 실패할 수 있다.[13] 따라서 다음향 환경에서의 음향 사건 검출 성능을 높이기 위해서는 보다 정확한 음원 분리 기술이 요구된다.

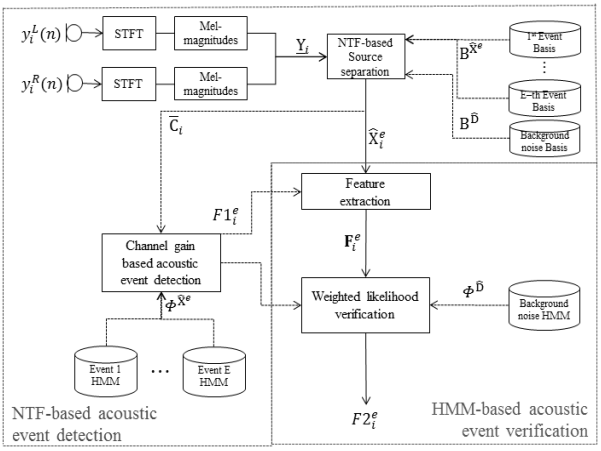

그림 1. 제안된 NTF 및 HMM 기반 이중 채널 음향 사건 검출 방법의 동작 과정

Fig. 1. Procedure of the proposed NTF and HMM based dual-channel acoustic event detection method.

본 논문에서는 비음수 텐서 분해(nonnegative tensor factorization, NTF)와 HMM을 이용하는 이중 채널 음향 사건 검출 방법을 제안한다. 제안된 방법은 두 개의 인접한 마이크로폰으로부터 녹음된 두 채널간의 크기 차이를 활용하여 음원 분리와 검출을 수행한다. 구체적으로, 먼저 NTF를 이용하여 이중 채널 입력 신호로부터 다수의 음향 사건들을 분리한다. NTF 분리 과정에서 각 사건 별 채널 이득이 추정되는데, 이들을 사용하여 1차적으로 음향 사건을 검출한다. 다음으로, HMM 기반의 우도비 검정을 통해 앞서 검출된 음향 사건의 발생 여부를 2차로 검증한다. 최종적으로 단일 판별기의 결과에 의존하는 종래의 방법들과 달리, 제안된 방법은 스테레오 채널 입력에 대한 NTF 분리로부터 획득된 판별 결과들에 대하여 HMM 기반의 2차 검증을 진행하기 때문에 보다 신뢰성 있는 음향 사건 검출 결과를 얻을 수 있다. 제안된 방법의 2차 음향사건 검증은 HMM 기반의 비터비 디코딩 혹은 우도비 검증 방법 모두 적용 가능하다. 본 논문에서는 우도비를 사용하는 기존의 방법인 GMM[4] 및 NMF-HMM[6] 방법과 그 성능을 비교하기 위하여 HMM 기반의 우도비 검정 방법을 사용한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 NTF 및 HMM을 이용한 이중 채널 AED 방법을 제안한다. 이어서 Ⅲ장에서는 다음향 환경에서 제안된 AED 방법의 성능을 평가하고 기존 방법들과 그 성능을 비교한다. 마지막으로 Ⅳ장에서 본 논문의 결론을 맺는다.

Ⅱ. 제안된 이중 채널 음향 사건 검출

1. 개요

제안된 이중 채널 음향 사건 검출 방법은 그림 1에서와 같이 NTF 기반 검출과 HMM 기반 검증 단계로 크게 이루어진다. 그림에서 보는 바와 같이, 제안된 방법의 핵심 동작은 다채널 신호에 대한 NTF 기반 음향 사건 및 잡음 분리, 그리고 이러한 분리 결과의 분석을 통한 1차 음향 사건 검출과 이를 바탕으로 음향 사건의 발생여부를 최종 결정하는 HMM 기반 우도비 검정을 통한 2차 검정으로 구성된다. 상세한 동작 흐름은 다음과 같다.

스테레오 입력의 i번째 프레임, $y_i^L(n)$과 $y_i^R(n)$이 E개의 음향사건, $x_i^{c,e}(n)$과 배경잡음, $d_i^c(n)$의 조합으로 이루어진다고 가정하면, $y_i^L(n)$과 $y_i^R(n)$은 다음과 식과 같이 표현될 수 있다.

$$y_i^c(n) = d_i^c(n) + \sum_{e=1}^{E} x_i^{c,e}(n) \tag{1}$$

여기서 c와 e는 채널과 사건의 인덱스를 각각 나타낸다. 다음으로, $y_i^c(n)$에 단구간 푸리에 변환(short-term Fourier transform, STFT)과 멜-필터뱅크를 적용하여 멜-스펙트럴 진폭(mel-spectral magnitude), $|Y_i^c(m)|$을 얻는다.[16] 이렇게 얻어진 $|Y_i^c(m)|$는 supervised NTF 기법[17]을 통해 E개의 음향 사건의 멜-스펙트럴 진폭, $|S_i^{c,e}(m)|$으로 분리된다. 이후, 분리된 각 음향 사건들의 채널 이득에 대한 평균 대 최대 임계치(mean-to-max threshold)를 계산하여 음향 사건들의 실제 발생 여부를 검출한다. 마지막으로, 검출된 음향 사건들에 대하여 HMM 기반의 우도비 검정을 진행하여 최종적으로 i번째 프레임에서 발생한 음향 사건 결과를 출력한다.

2. NTF 기반 음향 사건 검출

본 절에서는 입력 신호의 멜-스펙트럴 진폭으로부터 어떻게 음향 사건의 발생 여부를 검출하는지를 설명한다. 먼저 주어진 스테레오 입력 신호의 i번째 프레임에 해당하는 멜-스펙트럴 진폭 $|Y_i^L(m)|$과 $|Y_i^R(m)|$을 $(C\times M)$ 차원의 크기를 지니는 채널-주파수 행렬, $\mathbf{Y}_i = \left[|Y_i^L(m)|, |Y_i^R(m)|\right]^T$로 표현한다. 여기서 M과 T는 멜 스펙트럼 차수와 전치행렬 연산자를 각각 의미한다. 본 논문에서는 이중 채널 입력을 다루기 때문에, $C=2$로 정의된다. 다음으로, 과거로부터 연속된 V개의 $\mathbf{Y}_i$들을 연결하는 $(C\times M\times V)$ 차원의 채널-주파수-시간 텐서, $\underline{\mathbf{Y}}_i = \left[\mathbf{Y}_{i-V+1}\cdots\mathbf{Y}_{i-v}\cdots\mathbf{Y}_i\right]$를 정의한다. 여기서 $\underline{\mathbf{Y}}_i$는 NTF 모델로 다음 식과 같이 정의될 수 있다.[17]

$$\hat{\underline{\mathbf{Y}}}_i = \sum_{r=1}^{R'} \overline{\mathbf{C}}_{i;c,r} \otimes \overline{\mathbf{B}}_{m,r} \otimes \overline{\mathbf{A}}_{i;v,r} \tag{2}$$

여기서 $\otimes$는 텐서곱 연산을 의미하며, R'은 모든 음향사건과 잡음의 전체 NTF 기저의 차수를 의미하며 사건 별 기저차수, R과 모든 사건 및 잡음 군의 곱, 즉 $R'=(E+1)R$로 표현된다. 식 (2)에서 $\overline{\mathbf{C}}_i$, $\overline{\mathbf{B}}$, 그리고 $\overline{\mathbf{A}}_i$는 각각 $\hat{\underline{\mathbf{Y}}}_i$를 모델링하는 $(C\times R')$ 차원의 채널 이득 행렬, $(M\times R')$ 차원의 멜-스펙트럴 기저 행렬, 그리고 $(V\times R')$ 차원의 시간 이득 행렬을 각각 의미한다.

각 NTF 행렬들의 기저가 각각 의미하는 사건 및 잡음으로 R개씩 모인 형태로 정렬되어 있다고 한다면 $\overline{\mathbf{C}}_i$, $\overline{\mathbf{B}}$, 그리고 $\overline{\mathbf{A}}_i$는 다음 식과 같이 다시 정의될 수 있다.

$$\overline{\mathbf{C}}_i = [\mathbf{C}_i^{\hat{1}},\cdots,\mathbf{C}_i^{\hat{e}},\cdots,\mathbf{C}_i^{\hat{E}},\mathbf{C}_i^{\hat{\mathbf{D}}}] \tag{3}$$

$$\overline{\mathbf{B}} = [\mathbf{B}^{\hat{1}},\cdots,\mathbf{B}^{\hat{e}},\cdots,\mathbf{B}^{\hat{E}},\mathbf{B}^{\hat{\mathbf{D}}}] \tag{4}$$

$$\overline{\mathbf{A}}_i = [\mathbf{A}_i^{\hat{1}},\cdots,\mathbf{A}_i^{\hat{e}},\cdots,\mathbf{A}_i^{\hat{E}},\mathbf{A}_i^{\hat{\mathbf{D}}}] \tag{5}$$

여기서 $\mathbf{C}_i^{\hat{e}}$, $\mathbf{B}^{\hat{e}}$, 그리고 $\mathbf{A}_i^{\hat{e}}$는 각각 e번째 음향사건의 텐서 추정치, $\hat{\underline{\mathbf{X}}}_i^e = \sum_{r=1}^{R} \mathbf{C}_{i;c,r}^{\hat{e}} \otimes \mathbf{B}_{m,r}^{\hat{e}} \otimes \mathbf{A}_{i;v,r}^{\hat{e}}$를 모델링하는 $(C\times R)$ 차원의 채널 이득 행렬, $(M\times R)$ 차원의 멜-스펙트럴 기저 행렬, 그리고 $(V\times R)$ 차원의 시간 이득 행렬을 각각 의미한다. 마찬가지로 $\mathbf{C}_i^{\hat{\mathbf{D}}}$, $\mathbf{B}^{\hat{\mathbf{D}}}$, 그리고 $\mathbf{A}_i^{\hat{\mathbf{D}}}$는 잡음의 텐서 추정치, $\hat{\underline{\mathbf{D}}}_i = \sum_{r=1}^{R} \mathbf{C}_{i;c,r}^{\hat{\mathbf{D}}} \otimes \mathbf{B}_{m,r}^{\hat{\mathbf{D}}} \otimes \mathbf{A}_{i;v,r}^{\hat{\mathbf{D}}}$를 모델링하는 채널 이득 행렬, 멜-스펙트럴 기저 행렬, 그리고 시간 이득 행렬을 각각 의미한다.

$\hat{\underline{\mathbf{X}}}_i^e$와 $\hat{\underline{\mathbf{D}}}_i$를 추정하기 위하여, 먼저 $\mathbf{B}^{\hat{e}}$와 $\mathbf{B}^{\hat{\mathbf{D}}}$는 표본 기저의 집합으로 사전에 훈련된다. 이때 표본기저는 각 사건 및 배경잡음을 나타내는 훈련 데이터의 멜-스펙트럴 진폭 R개를 무작위로 선별하여 구할 수 있다.[2]

다음으로 $\mathbf{C}_i^{\widehat{\mathbf{X}}^e}$, $\mathbf{C}_i^{\widehat{\mathbf{D}}}$, $\mathbf{A}_{i;c,r}^{\widehat{\mathbf{X}}^e}$, 그리고 $\mathbf{A}_i^{\widehat{\mathbf{D}}}$는 다음과 같은 반복연산을 통해 얻어진다.[17]

$$\overline{\mathbf{C}}_i^{(h)} = \overline{\mathbf{C}}_i^{(h-1)} \circ \frac{\sum_{m=1}^{M}\sum_{v=1}^{V} \mathbf{P}_{i;c,m,v}^{(h-1)} \overline{\mathbf{B}}_{m,r} \overline{\mathbf{A}}_{i;v,r}^{(h-1)}}{\sum_{m=1}^{M}\sum_{v=1}^{V} \overline{\mathbf{B}_{m,r}} \overline{\mathbf{A}}_{i;v,r}^{(h-1)}} \quad (6)$$

$$\overline{\mathbf{A}}_i^{(h)} = \overline{\mathbf{A}}_i^{(h-1)} \circ \frac{\sum_{c=1}^{C}\sum_{m=1}^{M} \mathbf{P}_{i;c,m,v}^{(h-1)} \overline{\mathbf{C}}_{i;c,r}^{(h)} \overline{\mathbf{B}}_{m,r}}{\sum_{c=1}^{C}\sum_{m=1}^{M} \overline{\mathbf{C}}_{i;c,r}^{(h)} \overline{\mathbf{B}}_{m,r}} \quad (7)$$

여기서 $\mathbf{P}_{i;c,m,v}^{(h)} = \underline{\mathbf{Y}}_{i;c,m,v} / \widehat{\underline{\mathbf{Y}}}_{i;c,m,v}^{(h)}$, h는 식 (6)과 식 (7)의 반복 인덱스를 나타내고, $\circ$와 /는 행렬간 요소 곱셈 및 나눗셈 연산을 각각 의미한다. 식 (6)과 식 (7)에서 $\overline{\mathbf{C}}_i^{(h=1)}$과 $\overline{\mathbf{A}}_i^{(h=1)}$은 모두 0과 1 사이의 무작위값으로 초기화된다. 식 (6)과 식 (7)을 반복함에 따라 $\underline{\mathbf{Y}}_i$와 $\widehat{\underline{\mathbf{Y}}}_i$ 간의 Kullback-Leibler divergence가 감소되는데, 반복회수 별 상대적 감소치가 사전 정의된 문턱값보다 작아질 경우 반복이 종료된다.[18]

식 (6)과 식 (7)의 반복을 통해 얻은 $\overline{\mathbf{C}}_i$를 구한 다음, 현재 프레임에서 발생한 음향 사건을 검출한다. 음향 사건 검출은 전체 평균 채널 이득과 각 사건 별 최대 채널 이득간의 평균 대 최대 임계치를 통해 진행된다. 구체적으로, 전체 평균 채널 이득과 사건 별 채널 이득의 최대값은 각각 다음 식과 같이 계산된다.

$$\widetilde{C}_i = \frac{1}{CR'} \sum_{c=1}^{C}\sum_{r=1}^{R'} \overline{\mathbf{C}}_{i;c,r} \quad (8)$$

$$\widehat{C}_i^{\widehat{\mathbf{X}}^e} = \max\left(\frac{1}{R} \sum_{r=1}^{R} \mathbf{C}_{i;c,r}^{\widehat{\mathbf{X}}^e} \right). \quad (9)$$

그림 2는 음향 사건 발생에 따른 각 음향 사건 별 $\widehat{C}_i^{\widehat{\mathbf{X}}^e}$의 변화를 나타낸다. 그림에서 보는 바와 같이, 실제로 발생한 음향 사건에 해당하는 $\widehat{C}_i^{\widehat{\mathbf{X}}^e}$의 값이 상대적으로 크게 발생하는 것을 확인할 수 있다.

다음으로, $\widetilde{C}_i$와 $\widehat{C}_i^{\widehat{\mathbf{X}}^e}$의 비교를 통해, 각 사건 별 음향 사건 발생 여부를 판별한다.

$$F1_i^e = \begin{cases} 1, \text{if} \dfrac{\widehat{C}_i^{\widehat{\mathbf{X}}^e}}{\widetilde{C}_i} > thr1_e \\ 0, \text{otherwise} \end{cases} \quad (10)$$

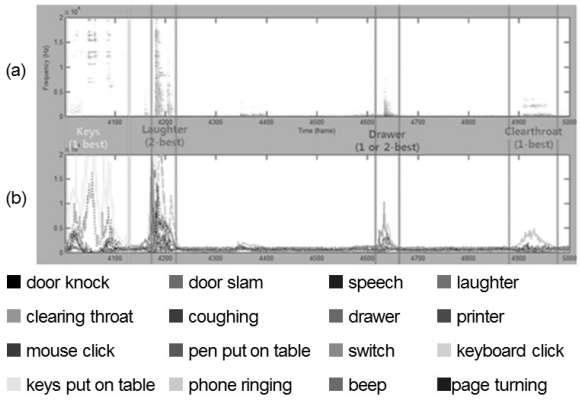

그림 2. 음향 사건 별 최대 채널 이득 값: (a) 음향사건의 스펙트로그램, (b) 사건 별 채널 이득(그림 내 표시된 색깔은 각기 다른 음향 사건에 대한 최대 채널 이득을 나타냄)

Fig. 2. Illustration of maximum channel gains of each event according to occurrence of acoustic events: (a) spectrogram of acoustic events and (b) maximum channel gains of acoustic events, where different colors are used to denote each acoustic event.

여기서 $thr1_e$은 사건 별로 사전 정의된 문턱값으로써 개발용 시험음원들에 대해서 가장 높은 정확도를 보일 때의 값으로 사전에 설정된다. 식 (10)을 통해 $F1_i^e$가 0이면 배경잡음으로 판단하고, 1이면 $\widehat{\underline{\mathbf{X}}}_i^e$는 다음의 HMM 기반 음향 사건 검증 단계로 진입한다.

3. HMM 기반 음향 사건 검증

본 절에서는 앞서 NTF 채널 이득 분석을 통해 판별된 음향 사건들의 실제 발생 여부를 검증하는 방법에 대해 설명한다. 이를 위해 $\mathbf{B}^{\widehat{\mathbf{X}}^e}$와 $\mathbf{B}^{\widehat{\mathbf{D}}}$를 사전 훈련하는데 사용한 훈련 데이터를 이용하여 E개의 음향 사건 및 한 개의 배경잡음의 HMM을 훈련한다. 다음으로, $F1_i^e = 1$에 해당하는 음향 사건들에 대하여 우도비 검증을 수행하여 해당 음향의 발생 여부를 최종적으로 검증 한다.

먼저 HMM 기반 음향사건 검증에서 쓰이는 특징으로는 MFCC[16]을 사용하는데, 특히 앞서 NTF 기반 음향 사건 검출에서 얻은 $\widehat{\underline{\mathbf{X}}}_i^e$로부터, 채널 간 평균을 취한 후, 멜-필터 차수 전환 행렬과 log 연산, inverse discrete cosine transform (IDCT) 연산을 통해 V 프레임에 대한 Q차 MFCC, $\boldsymbol{m}_i^e$을 아래와 같이 얻는다.

$$\mathbf{m}_i^e = \begin{bmatrix} m_{i-V+1}^e(1) \cdots m_{i-V+1}^e(q) \cdots m_{i-V+1}^e(Q) \\ \vdots \quad\quad\quad \vdots \quad\quad\quad \vdots \\ m_{i-v}^e(1) \quad \cdots \quad m_{i-v}^e(q) \quad \cdots \quad m_{i-v}^e(Q) \\ \vdots \quad\quad\quad \vdots \quad\quad\quad \vdots \\ m_i^e(1) \quad \cdots \quad m_i^e(q) \quad \cdots \quad m_i^e(Q) \end{bmatrix} . \tag{11}$$

또한, 본 논문에서는 시간의 변화에 따른 음향 사건 특징의 변화를 모델링하기 위해 MFCC와 이의 delta 및 acceleration으로 $\mathbf{F}_i^e$를 구성한다.[16] 최종적으로, $\mathbf{F}_i^e$는 $(3Q\times V)$ 차원으로 다음과 같이 표현된다.

$$\mathbf{F}_i^e = [\mathbf{m}_i^e, \Delta\mathbf{m}_i^e, \triangle\triangle\mathbf{m}_i^e]. \tag{12}$$

다음으로, 각 음향 사건 및 배경잡음의 HMM이 다음과 같이 훈련된다. e번째 음향사건과 배경잡음의 HMM은 각각 $\Phi^{\hat{\mathbf{X}}^e} = \{\pi^{\hat{\mathbf{X}}^e}, T^{\hat{\mathbf{X}}^e}, \theta^{\hat{\mathbf{X}}^e}\}$와 $\Phi^{\hat{\mathbf{D}}} = \{\pi^{\hat{\mathbf{D}}}, T^{\hat{\mathbf{D}}}, \theta^{\hat{\mathbf{D}}}\}$로 정의되며, 여기서 $\pi^{\hat{\mathbf{X}}^e}$와 $\pi^{\hat{\mathbf{D}}}$, $T^{\hat{\mathbf{X}}^e}$와 $T^{\hat{\mathbf{D}}}$, 그리고 $\theta^{\hat{\mathbf{X}}^e}$와 $\theta^{\hat{\mathbf{D}}}$는 각각 e번째 음향사건과 배경잡음의 초기 상태확률, 상태 천이 확률, 그리고 관측 확률의 행렬을 의미한다.

이들을 학습하기 위하여 해당 음향 데이터로부터 식 (12)의 $\mathbf{F}_i^e$를 추출한다. 여기서, $\Phi^{\hat{\mathbf{X}}^e}$와 $\Phi^{\hat{\mathbf{D}}}$의 구조는 3상태 어고딕(ergodic) HMM을 따르며, 관측 확률 밀도 함수(probability density function, pdf)는 8차 혼합 가우시안으로 모델링된다. 각 HMM 별 파라메터를 훈련하는데 기대-극대화 (estimation-maximization, EM)[19] 알고리즘을 사용한다.

이어서 $F1_i^e = 1$을 만족하는 e번째 음향 사건의 발생여부 검증은 다음과 같이 수행된다. 먼저 주어진 i번째 프레임의 $\mathbf{F}_i^e$에 대한 $\Phi^{\hat{\mathbf{X}}^e}$와 $\Phi^{\hat{\mathbf{D}}}$의 우도를 다음 식과 같이 계산한다.

$$L_i^{\hat{\mathbf{X}}^e} = P(\mathbf{F}_i^e|\Phi^{\hat{\mathbf{X}}^e}) \tag{13}$$

$$L_i^{\hat{\mathbf{D}}} = P(\mathbf{F}_i^e|\Phi^{\hat{\mathbf{D}}}) \tag{14}$$

여기서 $L_i^{\hat{\mathbf{X}}^e}$는 NTF를 통해 얻어진 검출 강도에 따른 가중치를 반영하기 위하여 e번째 채널 이득 정보를 이용하여 다음 식과 같이 변경된다.

$$\hat{L}_i^{\hat{\mathbf{X}}^e} = L_i^{\hat{\mathbf{X}}^e} \frac{\widetilde{C}_i^{\hat{\mathbf{X}}^e}}{\widetilde{C}_i} \tag{15}$$

마지막으로, e번째 음향 사건의 최종 발생 여부의 검증을 위해 다음과 같은 우도비 검정을 수행한다.[20]

$$F2_i^e = \begin{cases} 1, \text{if} \dfrac{\hat{L}_i^{\hat{\mathbf{X}}^e}}{L_i^{\hat{\mathbf{D}}}} > thr2_e \\ 0, \text{otherwise} \end{cases} \tag{16}$$

여기서 $thr2_e$는 우도비 검증을 위해 사건 별로 사전 정의된 문턱값으로써, $thr1_e$과 마찬가지로 개발용 시험 음원들에 대해서 가장 높은 정확도를 보일 때의 값으로 사전에 설정된다. 식 (16)을 통해 $F2_i^e = 1$ 인 경우에만 i번째 프레임에서 e번째 음향 사건이 발생한 것으로 최종 판단한다.

Ⅲ. 다음향 환경에서의 성능 평가

1. 실험환경 설정

제안된 음향 사건 검출 방법의 성능을 평가하기 위하여 다음향 환경에서의 음향 사건 검출 정확도를 측정하였다. 또한, 제안된 방법의 성능을 기존의 GMM 방법[4] 및 NMF-HMM 방법[6]과 비교하였다. 제안된 방법 및 기존 방법들의 음향 사건 검출을 위한 기저행렬과 HMM 모델은 모두 IEEE Audio and Acoustic Signal Processing (AASP) challenge의 훈련용 데이터베이스를 활용하였다.[21] 훈련을 위한 음향 사건들은 16가지 사건으로 정의되었다.* 각 음향 사건들은 각각 평균 1초 길이를 지니는 20종의 각기 다른 음향으로 구성되었다. 배경잡음에 해당하는 기저행렬과 HMM 모델은 60dB sound pressure level(SPL)의 크기를 지니는 사무실 환경의 배경잡음을 스마트폰으로 10분가량 녹음받은 신호를 활용하여 훈련하였다. 평가 음원은 AASP challenge에서 제공된 다양한 Office Synthetic 음원을 활용하여 생성하였다. Office Synthetic 음원은 이중 채널 마이크로폰으로 사전 정의된 총 16가지 음향 사건이 일정하지 않은 간격으로 골고루 발생하는 음원이며, 해당 음원을 바탕으로 세 가지 신호 대 잡음비(signal-to-noise ratio, SNR) 환경, 즉 -6dB, 0dB, 6dB, 그리고 세 가지 중첩 밀도**(0-20%, 20 - 30%, 30 - 40%)의 총 9가지의 다양

* 사전 정의된 음향 사건: door knock, door slam, speech, human laughter, clearing throat, coughing, drawer, printer, keyboard click, mouse click, pen put on table surfaces, switch, keys put on table, phone ringing, short alert(beep), page turning

** 중첩 밀도: 총 음향사건이 발생한 프레임 수 대비 동일한 시점에 두 가지 이상의 음향사건이 발생한 프레임 수의 비율

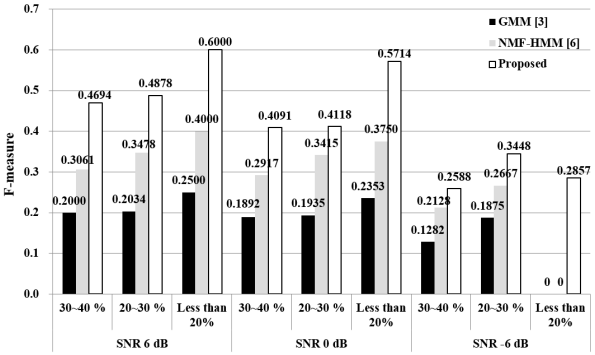

그림 3. 다양한 다음향 조건에서의 음향 사건 검출 정확도 비교

Fig. 3. Comparison of acoustic event detection accuracy under different multisource conditions.

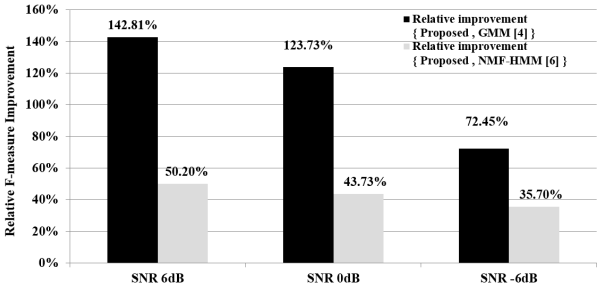

그림 4. GMM 및 NMF-HMM 기반의 기존 방법 대비 제안된 방법의 상대적 F-measure 개선률

Fig. 4. Relative improvement of F-measures of the proposed method against GMM and NMF-HMM based conventional methods.

한 다음향 실험 음원을 생성하였다. 여기서 모든 훈련 및 평가 음원들은 44.1 kHz의 샘플링 주파수와 16 bit의 샘플 해상도를 지녔다.

기존 및 제안된 음향 사건 검출 방법들은 21.53 ms의 프레임 길이를 지녔으며 프레임 간 중첩 비율은 50%로 설정되었다. 또한, 제안된 방법의 NTF 연산과 기존 방법의 NMF 연산에서의 공통 파라메터들은 기존의 NMF-HMM 기반 음향 사건 검출 방법에서의 최적값으로 알려져 있는 값[2, 6] 즉, $M=56$, $V=20$, 그리고 $R=100$으로 설정하였다. 특징 차수와 음향 사건 당 혼합 가우시안 수의 경우, 기존의 GMM 기반 방법은 각각 45차와 16으로, 제안된 방법은 각각 42차와 4로 설정하였으며, 이는 평가 음원에 속하지 않은 개발용 시험음원을 통해 실험적으로 찾은 최적값이다.

2. 정확도 평가

제안된 음향 사건 검출 방법의 정확도를 측정하기 위해 본 논문에서는 F-measure를 사용하였다.[22] F-measure는 정밀도(precision)과 재현율(recall)의 조화평균을 의미한다. 여기서, 음향 사건 검출 측면에서의 정밀도란 검출된 사건들 중 실제로 발생한 사건의 비율로써 다음과 식과 같다.

$$\mathrm{Precision}=\frac{|\{\mathrm{occured}\}\cap\{\mathrm{detected}\}|}{|\{\mathrm{detected}\}|}. \tag{17}$$

또한 재현율이란 실제로 발생한 사건들 중 검출된 사건의 비율로써 다음과 식과 같다.

$$\mathrm{Recall}=\frac{|\{\mathrm{occured}\}\cap\{\mathrm{detected}\}|}{|\{\mathrm{occured}\}|}. \tag{18}$$

F-measure는 식 (17)과 식 (18)로부터 다음과 같이 유도된다.

$$F=\frac{2\bullet\mathrm{Precision}\bullet\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}. \tag{19}$$

그림 3은 다음향 환경에서의 기존 방법들과 제안된 방법의 음향 사건 검출 정확도를 보여준다. 그림에서 보는 바와 같이, 제안된 방법은 기존 방법들보다 모든 조건에서 보다 높은 F-measure를 보였다.

다음으로 기존 방법들 대비 제안된 방법의 성능 개선치를 보이기 위해 상대적 F-measure 개선치를 측정하였다. 상대적 F-measure, F_{Improv}는 다음과 같이 계산하였다.

$$F_{\mathrm{Improv}}=\frac{F_{\mathrm{Prop}}-F_{\mathrm{Conv}}}{F_{\mathrm{Conv}}}. \tag{20}$$

다양한 다음향 환경에서의 성능 개선폭을 확인하기 위하여 SNR 별로 F_{Improv}을 산출하였으며, 이에 대한 결과는 그림 4와 같다. 그림에서 보는 바와 같이, 특히 6 dB 환경에서 제안된 방법은 기존의 GMM 및 NMF-HMM 기반 방법들 대비, 각각 142.81%와 50.20%라는 높은 성능 개선을 얻을 수 있었다.

Ⅳ. 결 론

본 논문에서는 다음향 환경에서의 음향 사건 검출 정확도를 높이기 위하여 NTF 및 HMM 기반의 이중 채널 음향 사건 검출 방법을 제안하였다. 제안된 방법은 이중 채널의 멜-스펙트럴 진폭에 대한 NTF 기반의 다채널 음원 분리에서 얻을 수 있는 채널 이득 정보를 활용하여 1차적인 음향 사건을 검출한 뒤, 검출 결과의

신뢰성을 높이기 위하여 HMM 기반 우도비 검정을 통한 2차 음향 사건 검증을 수행하였다. AASP challenge DB를 이용한 여러 다음향 환경에서의 F-measure를 비교한 결과, 제안된 방법이 기존의 GMM 및 NMF-HMM 기반 방법들보다 크게 우수한 성능을 보였다.

REFERENCES

[1] Y. Hong, S. Yu, C. M. Park, T. Yoon, and J. M. Kim, "Highway incident detection and classification algorithms using multi-channel CCTV", Journal of The Institute of Electronics and Information Engineers, Vol. 39, No. 2, pp. 263-269, Feb. 2014.

[2] J. F. Gemmeke, L. Vuegen, P. Karsmakers, and B. Vanrumste, "An exemplar-based NMF approach to audio event detection", in Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), DOI: 10.1109/WASPAA.2013.6701847, New Paltz, New York, Oct. 2013.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems", in Proc. of IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 21-26, London, UK, Sept. 2007.

[4] L. Vuegen, B. Van Den Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, and H. Van Hamme, "An MFCC-GMM approach for event detection and classification", Tech. Rep., 2013 [Online]. Available: http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/abstracts/OL/VVK.pdf.

[5] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation", in Proc. of Workshop on Machine Listening in Multisource Environments, pp. 36-40, Florence, Italy, Sept. 2011.

[6] K. M. Jeon, D. Y. Lee, H. K. Kim, and M. J. Lee, "Acoustic surveillance of hazardous situations using nonnegative matrix factorization and hidden Markov model", in Proc. of Audio Engineering Society (AES) 137th Convention, Preprint 9203, Los Angeles, CA, Oct. 2014.

[7] C. Clavel, T. Ehrette, and G. Richard, "Event detection for an audio-based surveillance system", in Proc. of IEEE International Conference on. Multimedia and Expo (ICME), pp. 1306-1309, Amsterdam, Netherlands, July, 2005.

[8] L. Atlas, G. Bernard, and S. Narayanan, "Applications of time-frequency analysis to signals from manufacturing and machine monitoring sensors", Proceedings of the IEEE, Vol. 84, No. 9, pp. 1319-1329, Sept. 1996.

[9] J. Pinquier, "Robust speech/music classification in audio documents", Entropy, Vol. 1, No. 2, pp. 2005-2008, Jan. 2002.

[10] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection", EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2013, No. 1, pp. 1-13, Dec. 2013.

[11] J. A. Smith, J. E. Earis, and A. A. Woodcock, "Establishing a gold standard for manual cough counting: video versus digital recordings", Cough, Vol. 2, no. 6, pp. 1-6, Aug. 2006.

[12] Y. Tian, D. Lo, and C. Sun, "Information retrieval based nearest neighbor classification for fine-grained bug severity prediction", in Proc. of Working Conference on Reverse Engineering (WCRE), pp. 215-224, Ontario, Canada, Oct, 2012.

[13] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection", Pattern Recognition Letters, Vol. 31, No. 12, pp. 1543-1551, Sept. 2010.

[14] K. M. Jeon, H. K. Kim, S. J. Lee, and Y. K. Lee, "Nonnegative matrix factorization based adaptive noise sensing over wireless sensor networks", International Journal of Distributed Sensor Networks, Vol. 10, No. 4, Apr. 2014.

[15] Y. Mitsufuji, M. Liuni, A. Baker, and A. Roebel, "Online non-negative tensor deconvolution for source detection in 3DTV audio", in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3082-3086, Florence, Italy, Mar. 2014.

[16] S. Mirsamadi and J. H. L. Hansen, "Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for ASR applications", in Proc. of Interspeech, Singapore, pp. 2828-2832, Sept. 2014.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book Version 3.4. Cambridge University, Cambridge, 2006.

[18] A. Chichocki, R. Zdunek, A. H. Phan, and S. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, John Wiley & Sons: Hoboken, NJ, 2009.

[19] K. M. Jeon, N. I. Park, H. K. Kim, M. K. Choi,

and K. I. Hwang, “Mechanical noise suppression based on non-negative matrix factorization and multi-band spectral subtraction for digital cameras”, IEEE Transactions on Consumer Electronics, Vol. 59, No. 2, pp. 296-302, May 2013.
[20] C. M. Bishop, Pattern Recognition and Machine Learning, Springer: New York, NY, 2006.
[21] K. M. Jeon, D. Y. Lee, N. I. Park, M. K. Choi, and H. K. Kim, “Two-stage impulsive noise detection using inter-frame correlation and hidden Markov model for audio restoration”, in Proc. of Audio Engineering Society (AES) 136th Convention, Preprint 9036, Berlin, Germany, Apr. 2014.
[22] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events”, IEEE Transactions on Multimedia, Vol. 17, no. 10, pp. 1733-1746, Oct. 2015.
[23] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation”, Journal of Machine Learning Technologies, Vol. 2, No. 1, pp. 37-63, Dec. 2011.

저 자 소 개

전 광 명(학생회원)

2010년 2월 세종대학교 정보통신공학과 학사
2012년 2월 광주과학기술원 정보통신공학부 석사
2012년 3월～현재 광주과학기술원 전기전자컴퓨터공학부 박사과정
<주관심분야 : 음성 및 오디오 신호처리, 기계학습, 딥러닝>

김 홍 국(평생회원)

1988년 2월 서울대학교 제어계측공학과 학사
1990년 2월 한국과학기술원 전기 및 전자공학과 석사
1994년 8월 한국과학기술원 전기 및 전자공학과 박사
1990년～1998년 삼성종합기술원 전문연구원
1998년～1998년 MMC Technology 선임연구원
1998년～2003년 AT&T Labs-Research Senior Member Technical Staff
2014년～2015년 City University of New York, Visiting Professor
2003년 8월～현재 광주과학기술원 전기전자컴퓨터공학부 교수
<주관심분야 : 음성인식, 음성 및 오디오 신호처리, 3D 오디오, 딥러닝>