한국정보전자통신기술학회논문지(jkiiect)'23-10, Vol.16 No.5
논문 23-16-05-239 http://dx.doi.org/10.17661/jkiiect.2023.16.5.239

오디오 부호화기를 위한 스펙트럼 변화 및 MFCC 기반 음성/음악 신호 분류

이상길*, 이인성**

Speech/Music Signal Classification Based on Spectrum Flux and MFCC For Audio Coder

Sangkil Lee*, In-Sung Lee**

요 약 본 논문에서는 오디오 부호화기를 위한 스펙트럼 변화 파라미터와 Mel Frequency Cepstral Coefficients(MFCC) 파라미터를 이용하여 음성과 음악 신호를 분류하는 개루프 방식의 알고리즘을 제안한다. 반응성을 높이기 위해 단구간 특징 파라미터로 MFCC를 사용하고 정확도를 높이기 위해 장구간 특징 파라미터로 스펙트럼 변화를 사용하였다. 전체적인 음성/음악 신호 분류 결정은 단구간 분류와 장구간 분류를 결합하여 이루어진다. 패턴 인식을 위해 Gaussian Mixed Model(GMM)을 사용하였고, Expectation Maximization(EM) 알고리즘을 사용하여 최적의 GMM 파라미터를 추출하였다. 제안된 장단구간 결합 음성/음악 신호 분류 방법은 다양한 오디오 음원에서 평균적으로 1.5% 분류 오류율을 보였고 단구간 단독 분류 방법 보다 0.9%, 장구간 단독 분류 방법보다 0.6%의 분류 오류율의 성능 개선을 이룰 수 있었다. 제안된 장단구간 결합 음성/음악 신호 분류 방법은 USAC 오디오 분류 방법보다 타악기 음악 신호에서 9.1% 분류 오류율, 음성신호에서 5.8% 분류 오류율의 성능 개선을 이룰 수 있었다.

Abstract In this paper, we propose an open-loop algorithm to classify speech and music signals using the spectral flux parameters and Mel Frequency Cepstral Coefficients(MFCC) parameters for the audio coder. To increase responsiveness, the MFCC was used as a short-term feature parameter and spectral fluxes were used as a long-term feature parameters to improve accuracy. The overall voice/music signal classification decision is made by combining the short-term classification method and the long-term classification method. The Gaussian Mixed Model (GMM) was used for pattern recognition and the optimal GMM parameters were extracted using the Expectation Maximization (EM) algorithm. The proposed long-term and short-term combined speech/music signal classification method showed an average classification error rate of 1.5% on various audio sound sources, and improved the classification error rate by 0.9% compared to the short-term single classification method and 0.6% compared to the long-term single classification method. The proposed speech/music signal classification method was able to improve the classification error rate performance by 9.1% in percussion music signals with attacks and 5.8% in voice signals compared to the Unified Speech Audio Coding (USAC) audio classification method.

Key words : Classification of Speech/Music, Audio/Speech Coding, Spectrum Flux, MFCC, GMM, Unified Speech and Audio Coding

This work was supported by a funding for the academic research program of Chungbuk National University in 2023.
* School of Information Communication Eng., Chungbuk National University
** Corresponding Author: School of Information Communication Eng., Chungbuk National University (inslee@chungbuk.ac.kr)
Received September 06, 2023 Revised September 19, 2023 Accepted October 07, 2023

1. 서 론

오디오 신호에서 음성 및 음악 신호를 구별하는 기능은 오디오 부호화[1][2], 라디오 방송의 자동 모니터링[3][8], 음성인식 및 일반 오디오 분할[4] 등 여러 멀티미디어 시스템에서 많이 사용되고 있다. 특히 최근의 오디오 부호화 방식은 음성 부호화, 오디오 부호화 방식을 독자적으로 각기 사용하지 않고 음성, 음악 신호를 포함한 모든 오디오 신호를 하나의 통합된 부호화기에서 다양한 전송률로 부호화하고 있다. 입력된 오디오 신호는 음성이나 음악으로 분류되고 음악 신호와 음성 신호의 종류에 따라 하나의 부호화기 내에서 다른 부호화 모델을 사용하여 부호화 한다. MPEG 표준화 USAC(Unified Speech Audio Coding) 부호화기[5]에서 음성 신호는 AMR-WB+ 기반의 선형예측 부호화 모델[6]을 사용하며 음악 신호의 경우 주파수 영역의 부호화 모델 을 사용하고 있다. 효율적인 오디오 부호화를 위해서 입력된 오디오 신호를 음성/음악인지 정확히 구별하고 적절한 부호화 모델을 사용하여야 좋은 음질의 오디오 출력을 만들어낼 수 있다. USAC 오디오 부호화기에서 매 프레임 별로 신호의 종류에 따라 다른 부호화 모델을 사용하고 있고 전송률도 다양하게 사용할 수 있다. 통합 오디오 부호화기에서 정확한 음성/음악 분류는 신호에 따라 최적의 부호화 모델은 결정하여 전체적인 성능에 중요한 요소가 되고 있다. 또한 USAC 부호화기에서 최적의 부호화 모델을 결정하기 위해 여러 다른 모델로 합성하고 신호에 대해 최적의 성능을 갖는 모델을 선택하는 폐루프 방식의 음성/음악 분류 방법을 사용하고 있다.

음성/음악을 분류하기 위해 특징 파라미터로 스펙트럴 중심[7], 스펙트럼 기울기 특성[8], 스펙트럼 변화[7], 스펙트럼 정점[9], 영교차율(Zero Crossing Rate)[8], 크로마 벡터[7], Mel Frequency Cepstral Coefficients(MFCC) 값[10][11] 등을 사용한다. 분류하는 방법으로 GMM 모델[7][12]을 사용한 패턴인식, Support Vector Machines(SVM)[13], 인공 뉴럴모델[14], Covolutional Neural Network(CNN) 딥러닝 모델[15]-[17]을 이용한 방식들을 사용하고 있다.

본 연구에서는 계산량이 적게 요구되며 정확도를 높이기 위해 단구간 분류 방법과 장구간 분류 방법을 결합한 음악/ 음성을 분류하는 방법을 제안한다. 단구간적으로 신호를 분류하기 위해 단구간 특징 파라미터로 MFCC 값을 사용하며, 장구간 신호의 변화를 고려한 스펙트럼 변화(Spectrum Flux) 값을 사용하여 장구간적으로 신호를 분류한다. 전체적인 최종적인 분류는 단구간 분류와 장구간 분류 방법을 결합하여 최종 결정함으로써 정확도를 높이는 방법을 제안한다. 본 논문 2장에서는 사용되는 특징 파라미터에 대해 설명하고 3장에서 GMM 기반 신호 분류 방법과 전체적인 단구간, 장구간 분류 방법에 대해 설명한다. 4장에서 시스템의 성능에 대해 설명하고 5장에서 결론을 맺는다.

2. 음성/음악 분류 특징 파라미터

오디오 신호 분석을 위해 시간 영역 특징 파라미터로 단구간 에너지, 영교차율, 에너지 엔트로피 등이 사용될 수 있으나 정확도를 높이기 위해 주파수 영역에서의 스펙트럼 분석이 필요하다. 오디오 신호의 스펙트럴 특징으로 스펙트럴 중심[7], 스펙트럼 변화, MFCC[10], 스펙트럴 평탄도[17] 등의 특징 파라미터를 사용할 수 있다. 본 연구에서는 오디오 신호를 단구간적으로 음성/음악 분류를 위해 MFCC 값을 사용하며 장구간적으로 음성/음악 분류를 위해 여러 오디오 프레임의 스펙트럼 변화 값을 사용한다.

2.1 단구간 특징 분류를 위한 MFCC 파라미터

음성인식에 쓰이는 특징값으로 음성 발생기관의 모델에 근거한 선형예측계수(Linear Prediction Coefficients) 값이나 음성 청취모델에 근거한 MFCC 값[10][11] 등이 있다. 본 연구에서 단구간 오디오 신호 특징 파라미터로 MFCC 값을 사용하여 음성 신호와 음악 신호를 분류한다.

MFCC 값은 입력 신호를 프레임 단위로 나누어 처리한다. 그림 2는 MFCC 값을 계산하는 과정을 나타낸다. 입력 신호는 프레임 크기로 나눈 후 윈도우 함수를 곱하고 Fast Fourier Transform(FFT)를 취하여 주파수 영역으로 변환하게 된다. 그 후 주파수 대역을 여러 필터 뱅크로 나누고 각 필터 뱅크의 에너지를 구한다. FFT 과정은 식(1)에 나타나 있다.

$$S_i(n,w_k)=\sum_{m=-\infty}^{\infty} s_i[m]w[n-m]e^{-jw_k m} \quad (1)$$
$$,w_k=\frac{2\pi}{N}k$$

멜 스케일 필터 뱅크의 l번째 필터의 주파수 응답을 $R_l(w_k)$라고 하면 i번째 오디오 프레임에 l번째 필터에 대한 멜 에너지는 식(2)로 나타낼 수 있다.

$$E_{mel}(i,l)=\frac{1}{A_l}\sum_{k=L_l}^{H_l}\left|R_l(w_k)S_i(n,w_k)\right|^2 \quad (2)$$

여기서, $A_l=\sum_{k=L_l}^{H_l}\left|R_l(w_k)\right|^2$ 로 주어진다.

멜 에너지를 DCT(Discrete Cosine Transformation)를 적용하여 MFCC 값으로 변환한다. 식 (3)을 이용하여 i번째 음성 프레임의 l번째 MFCC 계수 값을 계산한다. 본 연구에서는 각 프레임 당 13개의 MFCC 값을 신호 분류를 위한 단구간 특징 벡터로 사용한다.

$$C_{mel}[i,l]=\frac{1}{R}\sum_{k=0}^{R-1}\log\{E_{mel}(i,k)\}\cos(\frac{2\pi}{R}kl) \quad (3)$$

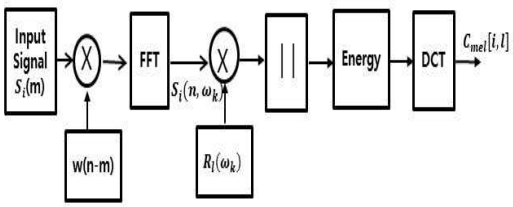

그림 1. MFCC 특징 파라미터 추출 과정
Fig. 1. Block diagram to extract MFCC paameter

2.2 장구간 특징 분류를 위한 스펙트럼 변화

오디오 신호의 프레임 간 스펙트럼의 변화를 감지하기 위해 스펙트럼 변화(Spectrum Flux) 값을 특징 파라미터로 사용한다. i번째 프레임의 샘플 오디오 신호 $s_i[n]\ n=1,...,N$ 은 FFT를 취한 후 바로 전 프레임과의 스펙트럼 변화를 식(4)와 같이 계산한다. 음성 신호와 음악 신호는 스펙트럼 변화 특징값은 다르게 나타나며 음성과 음악 스펙트럼 변화의 분포도는 그림2에 나타나 있다. 음성 신호는 음악 신호보다 시간에 빠르게 변하는 특성이 있어 음성 신호는 음악 신호보다 스펙트럼의 변화 값이 크게 나타난다. 신호 분류를 위한 장구간 특징 파라미터 값으로 12개 프레임의 스펙트럼 변화 값을 사용한다.

$$S_i[m]=FFT[s_i[n]] \quad n=1,...,N \quad m=1,...,N$$
$$\hat{S}_i[m]=\frac{S_i[m]}{\arg\max[S_i[m]]}$$
$$SF[i]=\sum_{m=1}^{N}\left[\left|\hat{S}_i[m]-\hat{S}_{i-1}[m]\right|^2 \quad (4)$$

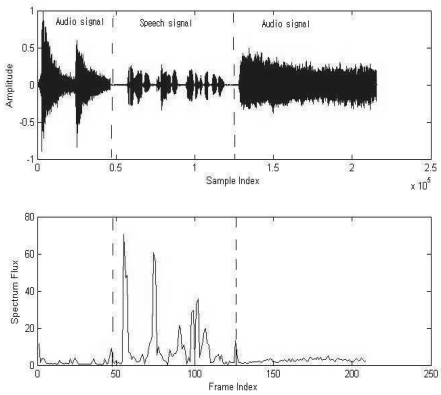

그림 2. 음성과 음악 신호의 스펙트럼 변화 값
Fig. 2. The spectral flux of speech and audio signal

3. GMM 모델을 이용한 음성/음악 신호 분류

음성/음악 신호 분류는 단구간 특징 파라미터 및 장구간 특징 파라미터를 이용하여 가우시안 혼합 모델(GMM)의 로그우도(Log-likelihood) 함수 값을 계산하여 결정한다. 단구간 특징 파라미터 GMM과 장구간 특징 파라미터 GMM은 각각 학습되어지고 최종적인

음성 및 음악 신호의 분류는 두 방법을 결합한 로그 우도 함수로 결정한다.

3.1 GMM 모델 이용한 단구간 오디오 신호 분류

GMM은 L개의 가우시안을 합하여 만들어진 모델로 음향학적인 분포를 표현함에 있어서 매우 뛰어난 것으로 나타난다. GMM은 식(5)로 표현되며, L개의 요소 가우시안 분포에 가중치를 곱하고 합산한다.

$$p_{MF}(\boldsymbol{x}_M|\lambda) = \sum_{k=1}^{L} w_k b_k(x) \qquad (5)$$

특징 벡터 $\boldsymbol{x}_M$, $b_k(x)$는 요소 가우시안 분포, w_k는 k번째 요소 가우시안 분포에 대한 가중치를 나타낸다. 이때 가중치 w_k는 $\sum_{k=1}^{L} w_k = 1$을 만족한다. 각 요소 가우시안 분포 $b_k(x)$는 식(6)에서와 같이 $\boldsymbol{\mu}_k$의 평균 벡터와 $\boldsymbol{\Sigma}_k$의 공분산 행렬을 갖는 가우시안 분포를 갖는다[7][12].

$$b_k(x) = \frac{1}{(2\pi)^{L/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(x-\boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(x-\boldsymbol{\mu}_k)\right\} \qquad (6)$$

L개의 가우시안 확률밀도함수의 선형 결합으로 정의되는 $p_{MF}(\boldsymbol{x}_M|\lambda)$는 각 클래스에 대한 평균, 공분산, 가중치에 관한 함수이며 평균, 공분산, 가중치의 3개의 파라미터는 훈련 과정에서 계산하게 된다. 즉, 학습 벡터 $X = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_t}\}$을 이용하여 각 클래스의 λ_{MF}를 추정한다.

$$\lambda_{MF} = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\} \quad , k = 1, \ldots . L \qquad (7)$$

GMM의 학습은 최대 우도(Maximum Likelihood) 추정 방법을 이용하며 GMM의 우도함수를 최대화하는 파라미터 λ_{MF}를 추정한다. 우도함수를 최대화하는 λ_{MF}를 추정하기 위해 EM(Expectation Maximization) 알고리즘을 사용한다[18].

입력 신호에서 프레임 단위로 MFCC 값을 계산하고 i번째 프레임 단구간 특징 벡터 $\boldsymbol{x}_M(i)$를 구성한다. MFCC 특징 파라미터로 13개를 사용한다. 학습화 과정에서 도출된 음성과 오디오 신호의 GMM에 MFCC 특징 패턴의 로그 우도 함수 값을 비교하여 음성과 음악 신호를 분류한다. 음성 신호는λ_s, 음악 신호는λ_M로 나타내며, 로그 우도 함수가 음성이 클 경우에는 음성 신호로 음악 신호가 큰 경우에는 음악 신호로 판단하게 된다. 식(10)을 만족하면 음성신호 , 식(11)을 만족하면 음악 신호로 결정한다.

$$L_{MF}^{S}(i) = \sum_{n=1}^{L} \log p_{MF}(x_M(i)|\lambda_s) \qquad (8)$$

$$L_{MF}^{M}(i) = \sum_{n=1}^{L} \log p_{MF}(x_M(i)|\lambda_m) \qquad (9)$$

$$If\ L_{MF}^{S}(i) > L_{MF}^{M}(i), then\ C_s(i) \Rightarrow Speech \qquad (10)$$

$$If\ L_{MF}^{S}(i) < L_{MF}^{M}(i), then\ C_s(i) \Rightarrow Music \qquad (11)$$

여기서 $C_s(i)$는 i번째 프레임 단구간 오디오 클래스를 나타낸다.

3.2 GMM 모델 이용한 장구간 오디오 신호 분류

장구간 특징 파라미터로 D개의 스펙트럼 변화 파라미터를 사용한다. 가우시안 확률밀도함수의 선형 결합으로 정의되는 $p_{SF}(x_s|\lambda)$는 각 클래스에 대한 평균, 공분산, 가중치에 관한 함수이며 단구간 특징 파라미터처럼 3개의 파라미터를 학습 과정에서 모델링한다. 학습 샘플을 이용하여 EM 알고리즘을 통해 각 λ_{SF}를 추정한다.

$$\lambda_{SF} = \{w_k, \mu_k, \Sigma_k\} \quad , k = 1, \ldots . D \qquad (12)$$

입력 신호를 프레임 단위로 스펙트럼 변화 값을 구하여 버퍼에 저장된 과거 스펙트럼 변화 값을 포함하여 장구간 특징 벡터 $\boldsymbol{x}_s(i)$ 구성한다. 12 개의 현재 및 과거 스펙트럼 변화 값으로 스펙트럼 변화 특징 벡터를 구성한다. 학습되어 얻어진 음성과 음악 신호의

GMM을 사용하며 스펙트럼 변화 벡터 $\boldsymbol{x}_s(i)$의 로그 우도 함수 값을 비교하여 음성과 음악 신호를 분류한다. 음성 신호는λ_s, 음악 신호는λ_M로 나타내며, 로그 우도 함수 값이 음성이 큰 경우에는 음성 신호로 음악 신호가 큰 경우에는 음악 신호로 판단하게 된다. 식(15)은 음성 신호, 식(16)은 음악 신호를 나타낸다.

$$L_{SF}^{S}(i) = \sum_{n=1}^{D} \log p_{SF}(x_S(i)|\lambda_s) \quad (13)$$

$$L_{SF}^{M}(i) = \sum_{n=1}^{D} \log p_{SF}(x_S(i)|\lambda_M) \quad (14)$$

$$If\, L_{SF}^{S}(i) > L_{SF}^{M}(i), then\ C_l(i) \Rightarrow Speech \quad (15)$$

$$If\, L_{SF}^{S}(i) < L_{SF}^{M}(i), then\, C_l(i) \Rightarrow Music \quad (16)$$

여기서 $C_l(i)$는 i번째 프레임 장구간 오디오 클래스를 나타낸다.

3.3 장단구간 결합 음성/음악 분류기

장구간 및 단구간 특징 파라미터를 이용한 장단구간 결합 음성 및 음악 분류기는 그림3에 나타나 있다. 입력 오디오 신호는 단구간 분류기 및 장구간 분류기에서 각기 특징 벡터를 계산하고 각기 분류한다. 단구간 분류 결과와 장구간 분류 결과가 동일하면 그 결과를 최종 결과로 출력한다. 만약 단구간, 장구간 음성/음악 분류 결과가 상이하다면 장구간 로그 우도 함수값 및 단구간 로그 우도 함수값을 결합한 장단구간 결합 로그 우도 함수를 사용하여 음성 및 음악 신호를 분류한다. 장구간 로그 우도 함수 값과 단구간 로그우도 함수 값을 합한 식을 사용하여 음성 및 음악 신호를 결정한다.

$$If\, L_{SF}^{S}(i) + L_{SM}^{S}(i) > L_{SF}^{M}(i) + L_{SM}^{M}(i), then\ C_j(i) \Rightarrow Speech$$

$$If\, L_{SF}^{S}(i) + L_{SM}^{S}(i) < L_{SF}^{M}(i) + L_{SM}^{M}(i), then\ C_j(i) \Rightarrow Music$$

여기서 $C_j(i)$는 i번째 프레임 장단구간 결합 오디오 클래스를 나타낸다.

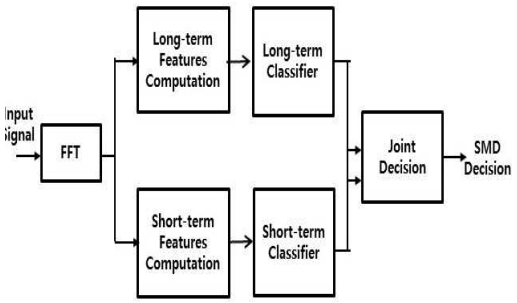

그림 3. 전체적인 장단구간 결합 음성/음악 분류기
Fig. 3. Overall Long-term/Short-term Joint Speech/Music Classifier

4. 장단구간 결합 음성/음악 분류기의 성능 평가

장단구간 혼합 음성 음성 분류기의 가우시안 혼합 모델을 얻기 위해 EM 알고리즘을 사용하여 학습하였다. 실험에 사용한 오디오 신호는 모노(Mono) 채널, 16 bits/sample, 16 kHz로 샘플링 되었으며 프레임의 크기는 1024 샘플을 사용하였다. 35개의 음원을 사용하여 가우시안 혼합 모델을 학습하였고 음성 및 음악 분류기의 성능 시험을 위해 7개의 음원을 사용하였다. 성능 시험에 사용된 7개의 음원에는 Harmonic (백파이프, 오르간), Individual-Line(실로폰, 탬버린), Mixed(노래+연주, 혼합 악기), Generic(Cymbal, Gong), 남성 음성, 여성 음성 신호로 구성되어 있다.

표 1에서는 제안된 음성/음악 분류기의 성능을 나타내었다. MFCC 단구간 특징 파라미터를 이용한 음성/음악 분류기는 평균 2.4%의 오류율을 보였고 스펙트럼 변화 장구간 특징 파라미터를 이용한 음성 음악 분류기는 평균 2.1%의 오류율을 나타내었다. 장단구간 특징을 결합한 분류기는 평균 1.5%의 오류율을 보여 하나의 특징 파라미터만 사용하는 방법보다 0.6% 이상의 성능 개선을 이룰 수 있었다. 심벌(Cymbal) 이나 캐스터네츠(Castanets)와 같은 타악기 오디오 신호에서 USAC 음성/음악 분류기는 많은 성능 저하를 보였으나 제안된 장단구간 혼합 분류기에서는 매우 우수한 성능을 보였다. Individual-line 신호 같은 경우 제안된 알고리즘의 오류율이 2% 미만으로 떨어짐을

볼 수 있었다. 심벌신호인 타악기 오디오 신호에서도 제안된 방법은 1% 미만의 오류율을 보였다. 또한 남성과 여성 음성 신호에서도 제안된 알고리즘의 평균적으로 1.3% 미만의 오류율을 보여 성능 개선이 있음을 볼 수 있었다.

표 1. USAC 방법과 제안한 알고리즘의 신호 분류 오류율(%) 결과
Table 1. Error rate(%) of the signal classification for the USAC and proposed method

Audio Signal	USAC	Short-term (MFCC)	Long-term (Spectrum Flux)	Joint Method
Harmonic (Music)	6.8	1.2	1.0	0.9
Individual-Line (Music)	12.3	2.5	2.0	1.7
Mixed (Music)	7.5	4.7	4.3	3.5
Cymbal (Music)	11.7	2.3	1.7	1.0
Gong (Music)	7.5	0.6	1.2	0.0
Female (Speech)	7.2	3.7	3.1	2.5
Male (Speech)	7.9	1.5	1.2	1.0
Average	8.7	2.4	2.1	1.5

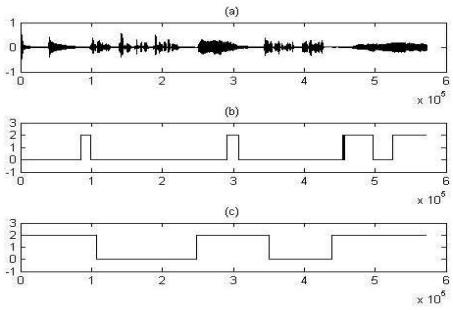

그림 4. 신호 분류 결과 (a) 실험파일 (b) USAC 결과 (c) 제안하는 알고리즘 결과
Fig. 4. The results of the signal classification (a) Test waveform (b) Results of USAC (c) Results of Proposed Method

그림 4에서는 여러 음성/음성 신호에서의 분류 과정 결과를 보여준다. 실험 파일은 음악 신호와 음성 신호를 합하여 구성하였다. 음성으로 분류된 결과는 0(음성)으로 음악으로 분류된 결과는 2(음악)로 나타낸다. 그림 4에서 보듯이 USAC의 신호 분류에서 음악 신호(2)임에도 음성(0)을 나타냄으로 신호를 올바로 판단하지 못하는 것을 볼 수 있었다. 하지만 제안된 알고리즘 방법의 신호 분류에서는 신호 분류의 정확성을 높아진 것을 알 수 있다.

USAC 부호화기에서는 부호화 과정을 수행 한 후 Signal to Noise Ratio(SNR)을 계산하여 성능이 좋은 결과를 선택하는 폐루프 분석 합성 방식이지만 제안한 분류기는 개루프 방식으로 MFCC 특징 파라미터, 스펙트럼 변화 파라미터를 계산한 후 GMM모델의 로그 우도함수를 계산하여 음성 음악을 판단함으로써 적은 연산량이 사용된다.

현재 프레임을 판단하는데 있어 현재 프레임의 특성을 고려할 뿐 아니라 신호의 연속성을 고려한 과거 프레임의 스펙트럼 변화 값을 같이 사용함으로써 현재 프레임 특징만을 사용하는 신호 분류 방법보다 더 나은 성능을 나타내었다. USAC 분류 방법에서는 신호가 갑자기 변하는 어택(Attack) 신호의 경우 음악 신호임에도 음성으로 잘못 판단하여 선형예측 영역 모드로 인코딩하여 음질 저하를 가져왔다. 하지만 제안하는 알고리즘의 신호 분류 방법에서는 어택 신호와 같은 타악기 신호에서도 더 나은 신호 분류의 정확성을 나타내어 개선된 음질을 나타내는 것을 볼 수 있었다.

5. 결론

본 논문에서 음성/오디오 통합 부호화기에서 신호의 종류에 따른 다른 부호화 모델을 사용하기 위해 오디오 신호를 분류 하는 방법을 제안하였다. 신호의 단구간 특성과 장구간 특성을 모두 이용하여 신호를 분류하는 방법을 제안하였다. 단구간 특징 파라미터로 MFCC 파라미터를 사용하였고 장구간 특징 파라미터로 스펙트럼 변화 파라미터 사용하여 분류하였다. 각 특징 파라미터는 GMM을 통한 로그 우도 함수를 사용하여 음성 신호와 음악 신호 분류하였고 최종적으로 두 방법을 결합하여 음성/음악을 결정하였다. MPEG

표준 부호화 방식인 USAC의 신호 분류 방법에서 페루프 분석/합성 방식의 많은 연산량을 요구하는 문제점을 개선하였고, USAC 부호화기에서 신호가 가지는 연속적인 특성을 고려하지 않고 현재 프레임만을 가지고 신호를 분류하였으나, 본 논문에서는 과거 프레임의 스펙트럼 변화 값을 이용한 GMM 분류 방법을 통해 정확성을 높였다. 장구간 및 단구간 오디오 신호 특성을 모두 고려하여 음성 음악 신호를 분류함으로써 단일 방법만 사용하는 방법보다 0.6% 이상의 분류 오류율의 개선을 이룰 수 있었으며 USAC 분류 방법보다는 7% 이상의 분류 오류율 개선을 이룰 수 있었다. 특히 심벌(Cymbal)이나 캐스터네츠(Castanets)와 같은 타악기 오디오 신호에서 USAC 분류 방법의 문제점을 보완할 수 있었다. 단구간, 장구간 변화를 모두 고려하여 음성 및 음악 신호를 분류함으로써 분류 오류율의 성능 개선을 이룰 수 있었고 여러 음성 오디오 통합 코덱에 적용 가능하다.

REFERENCES

[1] M. Neuendorf et el., “Unified Speech and Audio Coding Scheme for High Quality at Low Bit rates,” Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pp.1-4, 2009.

[2] G. Fuchs, "A Robust Speech/music Discriminator for Switched Audio Coding," Proc. of 2015 23rd European Signal Processing Conference (EUSIPCO), pp. 569-573, Nice, France, 2015.

[3] A. Pikrakis, T. Giannakopoulos and S. Theodoridis, “A Computationally Efficient Speech/music Discriminator for Radio Recordings,” Proc. of International Society for Music Information Retrieval Conference, pp.107-110, 2006.

[4] Kos, Marko, Zdravko, and Damjan Vlaj, “Acoustic Classification and Segmentation Using Modified Spectral Roll-off and Variance-based Features,” Digital Signal Processing 23, no.2 pp. 659-674, 2013.

[5] M. Wolters, K. Kjorling, D. Homm, and H. Purnhagen, "A Closer Look into MPEG-4 High Efficiency AAC,"MPEG-4 High Efficiency AAC,” 115th AES Convention, Paper 5871, October 2003.

[6] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami and A. Taleb, "AMR-WB+: a New Audio Coding Standard for 3rd Generation Mobile Audio Services," Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. ii/1109-ii/1112, 2005.

[7] G. Sell and P. Clark, "Music Tonality Features for Speech/music Discrimination," Proc of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2489-2493, 2014.

[8] J. Vavrek, E. Vozáriková, M. Pleva and J. Juhár, "Broadcast News Audio Classification Using SVM Binary Trees," Proc. of 35th International Conference on Telecommunications and Signal Processing (TSP),pp. 469-473, 2012.

[9] M. Bhattacharjee, S. R. M. Prasanna and P. Guha, "Speech/Music Classification Using Features From Spectral Peaks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1549-1559, 2020.

[10] M. Mcknney, J. Breebaart, “Fearures For Audio and Music Classification,” Proc. of International Conference on Music Information Retrieval, (ISMIR-03), 2003.

[11] A. Pikrakis, T. Giannakopoulos and S. Theodoridis, "A Speech/Music Discriminator of Radio Recordings Based on Dynamic Programming and Bayesian Networks," IEEE Transactions on Multimedia, vol. 10, no. 5, pp. 846-857, Aug. 2008,

[12] Hao Zhang, Xu-Kui Yang, W. -Q. Zhang, Wen-Lin Zhang and Jia Liu, "Application of i-vector in Speech and Music Classification," Proc. of 2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 1-5, 2016.

[13] P. Neammalai, S. Phimoltares and C. Lursinsap, "Speech and Music Classification Using Hybrid Form of Spectrogram and Fourier Transformation," Proc. of Signal and

Information Processing Association Annual Summit and Conference, pp. 1-6, 2014 .
[14] M. Srinivas, D. Roy and C. K. Mohan, "Learning Sparse Dictionaries for Music and Speech Classification," Proc. of 19th International Conference on Digital Signal Processing, pp. 673-675, 2014.
[15] David Doukhan and Jean Carrive, “Investigating the Use of Semi-Supervised Convolutional Neural Network Models for Speech/Music Classification and Segmentation, ” Proc. of 9th International Conference of Advances Multimedia, Apr. 2017.
[16] M. Papakostas and T. Giannakopoulos, “Speech-music Discrimination Using Deep Visual Feature Extractors,” Expert Systems with Applications, vol. 114, pp. 334-344, Dec. 2018.
[17] Kruspe, A., Zapf, D. & Lukashevich, . “Automatic Speech/music Discrimination for Broadcast Signals,” INFORMATIK, pp.151-162, 2017.
[18] Y. Wang, X. Yu, W. Wang, L. Liu, “The Research of Audio Clustering with Gaussian Mixture Based on EM Algorithm,” Proc. of International Communication Conference on Wireless Mobile and Computing, p. 389 - 393, 2011.

저자약력

이 상 길 (Sangkil Lee) [정회원]

- 2012년 : 충북대학교 정보통신공학부(학사)
- 2014년: 충북대학교 전파통신공학과(석사)

〈관심분야〉 음성/오디오신호처리, 오디오 음성부호화

이 인 성 (In-Sung Lee) [정회원]

- 1983년 : 연세대학교 전자공학과(학사)
- 1985년 : 연세대학교 전자공학과(석사)
- 1992년 : Texas A&M University Dept. of Electrical Eng. (박사)
- 1986년 ~ 1987년: KT 연구개발단 전임연구원
- 1993년 ~ 1995년 : ETRI 이동통신기술연구단 선임연구원
- 1995년 ~ 현재 : 충북대학교 정보통신공학부 교수

〈관심분야〉 음성/오디오신호처리, 지능신호처리, 통신신호처리