Journal of Rehabilitation Welfare Engineering & Assistive Technology
2017. vol. 11, no. 2, 143-146
http://resko.jams.or.kr
https://doi.org/10.21288/resko.2017.11.2.143

EVS 코덱에서 보청기를 위한 RNN 기반의 음성/음악 분류 성능 향상

Improvement of Speech/Music Classification Based on RNN in EVS Codec for Hearing Aids

강상익*, 이상민
S. I. Kang, S. M. Lee

요 약

본 논문에서는 recurrent neural network (RNN)을 이용하여 보청기 시스템을 위한 기존의 3GPP enhanced voice services (EVS) 코덱의 음성/음악 분류 성능을 향상시키는 방법을 제시한다. 구체적으로, EVS의 음성/음악 분류 알고리즘에서 사용된 특징벡터만을 사용하여 효과적으로 RNN을 구성한 분류기법을 제시한다. 다양한 음악장르 및 잡음 환경에 대해 시스템의 성능을 평가한 결과 RNN을 이용하였을 때 기존의 EVS의 방법보다 우수한 음성/음악 분류 성능을 보였다.

ABSTRACT

In this paper, a novel approach is proposed to improve the performance of speech/music classification using the recurrent neural network (RNN) in the enhanced voice services (EVS) of 3GPP for hearing aids. Feature vectors applied to the RNN are selected from the relevant parameters of the EVS for efficient speech/music classification. The performance of the proposed algorithm is evaluated under various conditions and large speech/music data. The proposed algorithm yields better results compared with the conventional scheme implemented in the EVS.

Keyword : Speech/Music Classification, Recurrent Neural Network (RNN), Enhanced Voice Services (EVS), Hearing Aids

1. 서론

최근 다양한 멀티미디어 기기들의 발달로 인하여 무선통신을 이용한 멀티미디어 서비스가 활발하게 이용되고 있으며 보청기 환경에서도 사운드 및 음성 품질에 대한 요구가 늘어나고 있다. 음성/음악 분류 알고리즘은 가변 비트레이트 음성 코딩에서 필수적인 부분이며 제한된 주파수 밴드를 효율적으로 사용하기 위한 방법이다. 그리고 음성/음악 분류 알고리즘은 보청기와 같이 저전력을 위한 제한된 시스템에서 음성/음악 분류를 통한 전체적인 음향 품질의 향상에도 이용된다. 최근 다양한 적응형 가변 비트레이트 음성 코덱이 연구되어지고 있으며 음성/음악 분류 알고리즘의 성능이 최종적인 음성의 품질에 많은 영향을 미친다 [1]-[5]. 기존의 음성/음악 분류는 CDMA2000 표준 코덱인 selected mode vocoder (SMV) 코덱에서 Gaussian mixture model (GMM), support vector machine (SVM) 을 이용하여 분류 성능향상을 하였다 [6], [7].

본 논문에서는 실시간 음성/음악 분류기법을 기반 가변 전송률 알고리즘을 채택하고 있는 3GPP의 표준코덱인 Enhanced Voice Services (EVS)의 기존 방법을 분석하고 [8] 이를 기반으로 음성/음악 분류성능을 향상시키기 위한 기법을 제안한다. 구체

접 수 일 : 2017.04.20
심사완료일 : 2017.04.26
게재확정일 : 2017.04.27
* 강상익 : 인하대학교 전자공학과 박사과정
rkdtkddlr@gmail.com (주저자)
이상민 : 인하대학교 전자공학과 교수
sanglee@inha.ac.kr (교신저자)
※ 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2016R1A2B4015370)

적으로, 최근 다양한 분야에서 강인한 분류성능을 보이는 neural network를 기반으로 하는 순차적인 데이터에 적합한 recurrent neural network (RNN)을 음성/음악 분류 기법으로 선정하고 기존의 EVS 코덱에서 사용하는 특징벡터를 별도의 계산과정 없이 사용하며 다양한 잡음 환경에서 기존의 EVS 기반의 음성/음악 분류 알고리즘과 성능을 비교하였다.

2. Enhanced Voice Services (EVS)

EVS 코덱은 3GPP에서 표준화한 VoLTE 음성 코덱이다. EVS 코덱은 ACELP와 MDCT 두 가지 모드를 가지고 있는데 각각 중저, 중고 비트레이트에 해당하며 시스템 상황이나 음성 품질 요구에 따라 모드가 선택된다. 효과적인 음성/음악 분류를 위해 먼저 Gaussian mixture model (GMM) 방법을 이용해 음성/음악 확률을 도출하고 스무딩과 샤프닝을 위해 상황 (context) 기반의 음성/음악 분류 기법을 적용하는 2 스테이지의 검출 시스템을 구성한다.

2.1 특징 벡터

EVS 코덱에서 음성/음악 분류 알고리즘의 복잡도는 기존 코덱의 파라미터들을 사용함으로써 감소시킬 수 있다. 음성/음악 분류에 효과적인 특징벡터 선택을 위해 분별 가능성을 Karnebäck의 방법을 이용하여 아래와 같이 구한다 [9].

$$U_{ftr} = \frac{1}{2}\sum_{j=0}^{M}\left| m_{ftr}^{(sp)}(j) - m_{ftr}^{(\mu s)}(j)\right| \qquad (1)$$

여기서 $m_{ftr}^{(sp)}$, $m_{ftr}^{(\mu s)}$ 은 각각 음성과 음악에 트레이닝 데이터베이스에 사용되는 특징벡터의 히스토그램을 나타내며 U_{ftr}은 분별가능성이다.. 분별 가능성 U_{ftr}는 1과 0 사이의 값을 가지며 1일 경우 최대 분별 가능성을 가진다. 그 결과 68개의 초기 파라미터 중에서 12개의 특징벡터를 선별하였다 [3].

2.2 Gaussian mixture model (GMM) 기반의 음성/음악 검출

Gaussian mixture model (GMM) 은 expectation maximization (EM) 알고리즘 [10]에 의해서 학습되며 학습데이터로 음성, 음악 데이터가 사용되며 다음과 같이 K개의 가우시안 밀도함수의 가중합으로 나타낸다.

$$p(\boldsymbol{x}|\theta) = \sum_{k=1}^{K} w_k N(\boldsymbol{x}|\mu_k, \boldsymbol{\Sigma_k}) \qquad (2)$$

여기서 $\boldsymbol{x}$는 현재 프레임의 N 차원 특징벡터 $(N=12)$이며 w_k 는 k 가우시안 밀도함수의 가중치이며 $N(\boldsymbol{x}|\mu_k, \boldsymbol{\Sigma_k})$는 가우시안 밀도함수이다. 최종적으로 도출된 음성, 음악 분류 확률 p_s, p_m을 아래의 결정식을 통하여 음성/음악을 판단한다.

$$f = \log(p_m) - \log(p_s) \qquad (3)$$

여기서 f가 0보다 큰 경우 음악이며 0보다 작을 경우 음성이다.

2.3 상황(context) 기반의 음성/음악 검출

GMM 기반의 음성/음악 분류 기법은 매 프레임마다 음성/음악을 결정할 때 입력신호에 따라 매우 빠르게 반응한다. 따라서 GMM 기반의 음성/음악 검출 알고리즘의 성능향상을 위해 auto regressive (AR) 필터를 이용하여 다음과 같이 스무딩한다.

$$\overline{f} = \gamma_c f + (1-\gamma_c)\overline{f}^{[-1]} \qquad (4)$$

여기서 γ_c는 0과 1사이의 필터 상수이며 $[-1]$은 이전프레임의 값을 의미한다 [8].

상대적인 프레임 에너지는 음성/음악 신호에 가까울수록 1에 가깝고 잡음 환경에 가까울수록 0.01에 가깝다. 따라서 신호 에너지의 크기가 클 때에는 현재 f값에 큰 가중치를 가지는 반면에 SNR 값이 낮은 경우 분류하기 어려운 조건이기 때문에 이전 프레임의 값에 가중치를 줄 필요가 있다. 결과적으로 $f<0, f<f^{[-1]}$ 조건에서 다음과 같은 결정식을 따른다.

$$r = r^{[-1]} + \frac{f^{[-1]} - f}{20}, 0.1 < r < 1 \qquad (5)$$

여기서 r은 GMM 기반의 음성/음악 분류기의 기울기이다.

효율적인 음성/음악 분류 알고리즘을 위해 0~7

프레임의 결과값을 합쳐서 최종 검출 확률을 도출하고 이것을 행오버라고 한다. 행오버 기법을 적용함으로써 과거의 데이터를 가지고 음성이 시작하는 구간에서의 오분류를 줄일 수 있다 [8].

3. 제안된 RNN 기반의 음성/음악 분류 시스템

RNN은 자연어 처리, 음성 인식 등에 많이 사용되는 연속된 신호에 적합한 분류 기법이다. RNN의 입력 값은 시간에 따른 순차적인 값이며 각 프레임에 따라 neural network (NN)이 존재하며 이전의 프레임의 NN과 연결되어 있는 구조이다. 그림 1에서 RNN의 기본구조를 나타낸다.

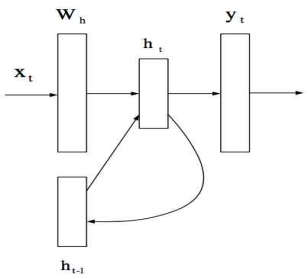

그림 1. RNN 구조
Fig. 1. Recurrent neural network structure

본 논문에서는 음성/음악 신호는 시간의 상관관계가 아주 높은 점을 고려한 RNN 기반의 음성/음악 분류 기법을 제안한다.
길이 T를 가지는 입력 데이터 $\boldsymbol{x}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T]$가 주어질 때 시간 t에서의 RNN의 은닉층 $\boldsymbol{h}_t$을 아래와 같이 정의한다.

$$\boldsymbol{h}_t=\boldsymbol{f}_h(\boldsymbol{x}_t,\boldsymbol{h}_{t-1})=\sigma(\boldsymbol{x}_t\boldsymbol{W}_h+\boldsymbol{h}_{t-1}\boldsymbol{U}_h+\boldsymbol{b}_h) \quad (6)$$

여기서 σ는 비선형 함수이다. 모델의 출력값을 $\boldsymbol{f}_y(\boldsymbol{h}_T)=\boldsymbol{h}_T\boldsymbol{W}_y+\boldsymbol{b}_y$으로 정의하면 RNN을 확률변수 $w=\{\boldsymbol{W}_h,\boldsymbol{U}_h,\boldsymbol{b}_h,\boldsymbol{W}_y,\boldsymbol{b}_y\}$에 따른 확률 모델으로 정의 할 수 있다. 베이시안 NN으로부터 RNN의 목적함수를 유도 할 수 있으며 목적함수 L은 다음과 같다 [11].

$$L\approx-\sum_{i=1}^{N}\log p\left(\boldsymbol{y}_i|\boldsymbol{f}_y^{\hat{w}}(\boldsymbol{f}_h^{\hat{w}}(\boldsymbol{x}_{i,T},\boldsymbol{f}_h^{\hat{w}}(\cdots\boldsymbol{f}_h^{\hat{w}}(\boldsymbol{x}_{i,1},\boldsymbol{h}_0)\cdots)))\right) \quad (7)$$
$$+KL(q(w)\|p(w)).$$

입력 $\boldsymbol{x}$에 대한 분류 확률은 최종적으로 각 레이어의 평균을 다음 레이어로 전달하거나 근사하여 다음과 같이 구한다.

$$p(\boldsymbol{y}^*|\boldsymbol{x}^*,\boldsymbol{X},\boldsymbol{Y})\approx\int p(\boldsymbol{y}^*|\boldsymbol{x}^*,w)q(w)dw \quad (8)$$
$$\approx\frac{1}{K}\sum_{k=1}^{K}p(\boldsymbol{y}^*|\boldsymbol{x}^*,\hat{w}_k)$$

여기서 $\hat{w}\sim q(w)$이다.

4. 실험 결과

본 논문에서는 EVS의 음성/음악 분류 알고리즘에 적용한 RNN 기반의 음성/음악 분류 성능을 알아보기 위해, 기존의 EVS의 알고리즘과 비교하였다. 본 실험을 위해 음성 데이터베이스로 TIMIT 데이터베이스 [12]가 사용되었으며 음악 데이터베이스는 CD로 부터 다양한 장르 (블루스, 클래식, 힙합, 재즈 등)의 음원을 추출하였다.

RNN 기반의 음성/음악 알고리즘의 학습을 위하여 60.1 시간의 음성 데이터베이스와 160.2 시간의 음악 데이터베이스를 사용하였으며 각각의 음성 파일은 6~12 초이며 음악파일의 경우 3~5 분의 길이를 가지고 있다. 실제 환경과 비슷한 성능 평가를 위해 깨끗한 음성 신호에 잡음을 합성하여 잡음 환경에서 음성/음악 분류를 판단하였다. 잡음 환경은 음성 데이터에 car, babble, white, factory 잡음이 20 dB SNR로 부과되었다.

테스트를 위하여 20.4 시간의 음성 데이터베이스 61.5 시간의 음악 데이터베이스를 사용하였으며 학습 데이터와 중복되지 않는 데이터를 사용하였다. 모든 데이터는 16 kHz으로 샘플링 되었으며 프레임 사이즈는 20ms 이고 성능을 평가하기 위해 각각의 프레임에 0 (무음), 1 (음성), 2 (음악)로 작성한 것과 비교하였다.

표 1. EVS와 제안된 음성/음악 분류 성능 비교
Table 1. Comparison of speech/music classification accuracy

		EVS	Proposed
speech	clean	0.9980	0.9981
	car 20dB	0.9927	0.9941
	babble 20dB	0.9905	0.9914
	white 20dB	0.9642	0.9775
	factory 20dB	0.9850	0.9871
music	classic	0.9868	0.9912
	others	0.9311	0.9435

표 1은 EVS와 제안된 RNN 기반의 알고리즘에서 음성/음악 검출 확률을 나타낸다. RNN의 hidden

layer를 3개로 이용하였고 각각 200, 600 그리고 200 개의 유닛을 사용하였다. 표 1을 통해서 기존의 EVS 알고리즘 보다 제안된 RNN 기반의 알고리즘의 전반적으로 우수함을 볼 수 있다.

5. 결론

본 논문에서는 RNN를 이용하여 기존의 3GPP EVS 코덱의 음성/음악 분류 성능을 향상시키는 방법을 제시하였다. EVS의 음성/음악 분류알고리즘에서 사용된 특징벡터만을 사용하여 효과적인 RNN을 구성한 분류기법을 제시하였다. 다양한 잡음환경에서 시스템의 성능을 평가한 결과 기존의 EVS의 방법보다 평균 0.5% 음성/음악 분류 성능 향상을 보였으며 노랫말이 없는 classic 장르의 성능이 다른 장르 (블루스, 힙합, 재즈 등) 보다 성능이 높게 나오는 것을 알 수 있다.

REFERENCES

[1] Y Gao, E Shlomot, and A Benyassine, “The SMV algorithm selected by TIA and 3GPP2 for CDMA applications,” IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 2, pp. 709-712, 2001.

[2] 3GPP2 Spec., Source-Controlled Variable-Rate Multimedia Wideband Speech Codec (VMR-WB), Service Option 62 and 63 for Spread Spectrum Systems, 3GPP2-C.S0052-A, v.1.0, Apr. 2005.

[3] 3GPP Spec., Codec for Enhanced Voice Services (EVS); Detailed Algorithm Description, TS 26.445, v.12.0.0, 2014.

[4] J. Saunders, “Real-time discrimination of broadcast speech/music,” IEEE Int. Conf. Acoustics, Speech, and Processing, vol. 2, pp. 993996, May 1996.

[5] W. Q. Wang, W. Gao, and D. W. Ying, “A fast and robust speech/music discrimination approach,” Int. Conf. Information, Communications, and Signal Processing, vol. 3, pp. 1325-1329, 2003.

[6] J. H. Song, K. H. Lee, J. H. Chang, J. K. Kim, and N. S. Kim, “Analysis and Improvement of Speech/Music Classification for 3GPP2 SMV Based on GMM,” IEEE Signal Process. Lett., vol.15, pp.103-106, 2008.

[7] C. LIM and J.-H. CHANG, “Improvement of SVM-Based Speech/Music Classification Using Adaptive Kernel Technique,” IEICE TRANSACTIONS on Information and Systems, vol.95, no. 3, pp.888-891, 2012.

[8] V .Malenovsky ,T. Vaillancourt, W. Zhe, K. Choo, and V. Atti, “Two-Stage Speech/Music Classifier with Decision Smoothing and Sharpening in the EVS Codec,” IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp.5718-5722, 2015.

[9] S. Karnebäck, “Discrimination between speech and music based on a low frequency modulation feature,” European Conf. on Seech Comm. and Technology, pp. 1891-1984, 2001.

[10] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. Royal Statiscal Soc., vol. 39, no. 1, pp. 1-38, 1977.

[11] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” arXiv preprint arXiv:1506.02142, 2015.

[12] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, “The DARPA speech recognition research database: Specification and status,” DARPA Workshop Speech Recognition, pp. 93-99, 1986.

강 상 익(Sang-Ick Kang)

2007 2월 인하대학교 전자공학과 졸업(학사)
2009 2월 인하대학교 전자공학과 졸업(석사)
2019년 - 현재 인하대학교 전자공학과 박사과정

Interest: Speech Signal Processing, Machine Learning, Hearing Aids

이 상 민(Sang Min Lee)

1987년 2월 인하대학교 전자공학과 졸업(학사)
1989년 2월 인하대학교 전자공학과 졸업(석사)
2000년 인하대학교 전자공학과 졸업(박사)
2006년 - 현재 인하대학교 전자공학과 교수

Interest: Brain-Machine Interface, Bio-signal processing, Hearing Aids, Psyco-Acoustic