symmetry

MDPI

Article

Improvement of Speech/Music Classification for 3GPP EVS Based on LSTM

Sang-Ick Kang and Sangmin Lee *

Department of Electronic Engineering, Inha University, Incheon 22212, Korea; rkdtkddlr@gmail.com
* Correspondence: sanglee@inha.ac.kr; Tel.: +82-70-8251-1549

Received: 18 October 2018; Accepted: 5 November 2018; Published: 7 November 2018

Abstract: The competition of speech recognition technology related to smartphones is now getting into full swing with the widespread internet of thing (IoT) devices. For robust speech recognition, it is necessary to detect speech signals in various acoustic environments. Speech/music classification that facilitates optimized signal processing from classification results has been extensively adapted as an essential part of various electronics applications, such as multi-rate audio codecs, automatic speech recognition, and multimedia document indexing. In this paper, we propose a new technique to improve robustness of a speech/music classifier for an enhanced voice service (EVS) codec adopted as a voice-over-LTE (VoLTE) speech codec using long short-term memory (LSTM). For effective speech/music classification, feature vectors implemented with the LSTM are chosen from the features of the EVS. To overcome the diversity of music data, a large scale of data is used for learning. Experiments show that LSTM-based speech/music classification provides better results than the conventional EVS speech/music classification algorithm in various conditions and types of speech/music data, especially at lower signal-to-noise ratio (SNR) than conventional EVS algorithm.

Keywords: speech/music classification; Enhanced Voice Service; long short-term memory; big data

1. Introduction

Speech/music classification algorithms are an important component of variable rate speech coding and coverage of the communication bandwidth, and provide effective means for enhancing the capacity of the bandwidth. In addition, a major concern in speech coding is optimizing speech input, and many types of input are being investigated, such as music. Speech/music classification algorithms are an essential part for providing high performance sound quality in speech coding. Recently, a number of adaptive multi-rate (AMR) voice codecs have been proposed to efficiently utilize the limited bandwidth resources available [1–3]. Precise determination of speech/music classification is quite necessary, as different bit rate allocations for the correct input/output formats affect the voice characteristic of these adaptive multi-rate voice codecs [4,5].

Recently, further improvements in speech/music classification problems have been achieved by adopting several machine learning techniques, such as the support vector machine (SVM) [6,7], Gaussian mixture model (GMM) [8], and deep belief network (DBN) [9] for the selectable mode vocoder (SMV) codec. The enhanced voice services (EVS) speech/music classifier, which is known as the 3rd-generation partnership project (3GPP) standard speech codec for the voice-over-LTE (VoLTE) network, is also based on GMM, but its features were calculated either at a current frame or as a moving average between those in the current and the previous frames [10]. The speech/music classifier uses a binary classification, but the diversity of music is greater than that of speech, and it can be generally said that it is a multiclass classification method, according to each musical genre. The GMM is not suitable for solving multiclass classification problems due to scalability issues.

In this paper, we propose a robust speech/music classifier based on long-short term memory (LSTM) [11,12], which can solve the vanishing gradient problem [13] better than the RNNs. In the case of an audio signal, a high correlation exists between signal samples in consecutive frames in the sequence. LSTMs, a particular type of RNNs, were basically proposed as a scheme of extending NNs to sequential signals. The extension of recurrent connections allows LSTMs to utilize the prior frame and makes them more robust to manipulating sequential data compared to non-recurrent NNs. The proposed method employs an LSTM using a feature vector derived from EVS codec. To appraise the accuracy of the proposed algorithm, speech/music classification experiments are performed under a variety of simulated conditions.

2. Conventional 3GPP Enhanced Voice Services

The EVS is a 3GPP speech codec designed for the VoLTE network. The EVS codec supports two modes: ACELP (for low and intermediate bitrates) and MDCT (for intermediate and high bitrates). The mode is selected depending on channel capacity and speech quality requirements.

The EVS speech/music classifier operates when the EVS codec is detected as "active" by voice activity detection (VAD) every 20 ms frame.

2.1. Feature Selection

The speech/music classification method applied to the EVS codec reuses the 68 parameters calculated in the early stages of codec preprocessing to minimize complexity. The EVS codec uses the technique proposed by Karnebäck [14], which is a method of analyzing the correlation matrix of all the features, for the initial selection of the features to be used in the Gaussian mixture model. In this way, we analyze the candidate feature sets with minimal cross-correlation and select the feature by calculating the discrimination probability as follows:

$$U_h = \frac{1}{2}\sum_{j=0}^{J}\left|m_h^{(sp)}(j) - m_h^{(mus)}(j)\right| \tag{1}$$

where $m_h^{(mus)}$ and $m_h^{(sp)}$ are the feature histograms that h generated on the music and the speech training database, respectively, and J is the whole number of frames in the database. The following 12 feature vectors are selected from the initial 68 feature vectors through the discriminatory probabilities U_h; five LSF parameters, normalized correlation, open-loop pitch, spectral stationarity, non-stationarity, tonality, spectral difference, and residual LP error energy [3].

2.2. GMM-Based Method

The GMM has been estimated via the expectation maximization algorithm [15] on a speech/music database, and is a weighted sum of L-component Gaussian densities, given by the following equation:

$$p(\boldsymbol{z}|\theta) = \sum_{k=1}^{L}\omega_k N(\boldsymbol{z}|\mu_k, \Sigma_k) \tag{2}$$

where $N(\boldsymbol{z}|\mu_k, \Sigma_k)$ are the component Gaussian densities, ω_k are the component weights, and z is a normalized N-dimensional feature vector. The GMM generates two probabilities, p_m and p_s, for the music probability and the speech probability, respectively.

By analyzing the values of music and speech probabilities in each frame, a discrimination measure between music and speech can be obtained by subtracting the log-probabilities, as:

$$y = \log(p_m) - \log(p_s) \tag{3}$$

2.3. Context-Based Method

The GMM-based speech/music classification responds very rapidly to change-over from speech to music and vice versa. To effectively utilize the discrimination potential of the GMM-based approach, y is sharpened and smoothed by the following adaptive auto-regressive filter:

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

where $[-1]$ denotes the previous frame and γ_c is a filter factor.

If the segment is energetically significant, the scaled relevant frame energy value will be close to 1, and for background noise, will be close to 0.01. Accordingly, if the SNR is high, more weight is given to the current frame, whereas if the SNR is low, the classifier has more dependency on the past data because it is difficult to make accurate short-term decisions. This situation potentially occurs when y is smaller than 0, and is smaller than the value of the previous frame. In this case:

$$g = g^{[-1]} + \frac{y^{[-1]} - y}{20}, 0.1 < g < 1 \tag{5}$$

where g is the gradient of the GMM approach, and $g^{[-1]}$ is initialized to the value of $-y$ each frame. Finally, previous frames of varying sizes (0–7) are combined according to the characteristics of the signal to determine speech/music [10].

3. Proposed LSTM-Based Speech/Music Classification

In this paper, an improved LSTM-based speech/music classification algorithm applicable to the framework of a speech/music classifier is proposed. LSTMs are sequence-based models of key importance for speech processing, natural language understanding, natural language generation, and many other areas. Because speech/music signals are highly correlated in time, LSTM is an appropriate method to classify speech/music. As shown in Figure 1, speech/music classification is performed using LSTM when it makes a decision as active speech in the EVS codec, compared with the conventional EVS codec and RNN-based algorithm. The feature vectors for speech/music classification are limited to 12 feature vectors used in the conventional EVS.

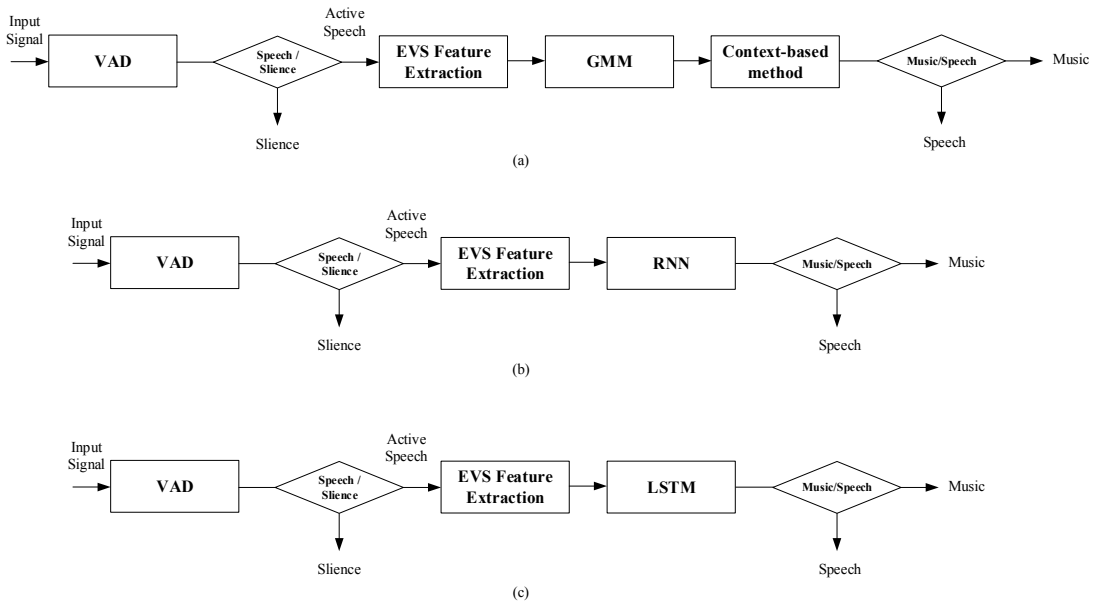

Figure 1. (**a**) Block diagram of the conventional EVS speech/music classification. (**b**) Block diagram of the recurrent neural network based speech/music classification. (**c**) Block diagram of the proposed speech/music classification.

The LSTM unit consists of an input activation function, a single memory cell, and three gates (input i_t, forget f_t, and output o_t), as shown in Figure 2. i_t permits the input signal to change or block the memory cell state. f_t controls what to remember and what to forget in the cell, and avoids vanishing gradients. Finally, o_t allows the memory cell state to have an influence on other neurons, or prevent this influence. With the addition of a memory cell, the LSTM can overcome the gradation problem of capturing and disappearing very complex and long-term dynamics. According to the LSTM unit, for an input x_t, the LSTM calculates a hidden/control state h_t, a block input g_t, and a state of memory cell c_t, which is an encoding of everything the cell has recognized until time t:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{7}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{8}$$

$$g_t = \phi\left(W_{xg}x_t + W_{hg}h_{t-1} + b_g\right) \tag{9}$$

$$c_t = f_t \bigodot c_{t-1} + i_t \bigodot g_t \tag{10}$$

$$h_t = o_t \bigodot \phi(c_t) \tag{11}$$

where W_{ij} are the weight matrices, $\bigodot$ is the point-wise product with the gate value, b_j is the bias, $\phi(x)$ is the activation function, and $\sigma(x)$ is the logistic sigmoid. As shown in Figure 3, LSTM units are gathered together to form layers and are connected at each time step.

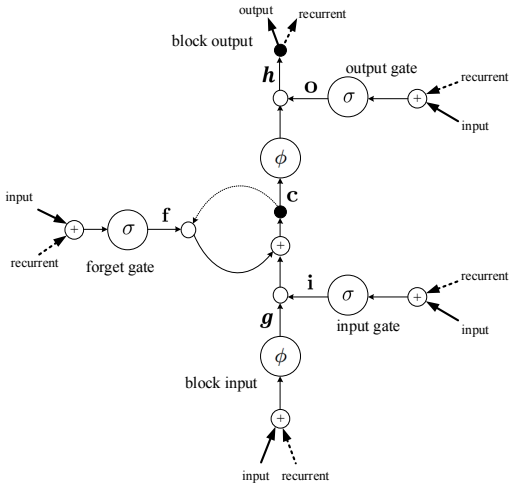

Figure 2. Detailed schematic of the LSTM unit, containing three gates, a block input, and a block output.

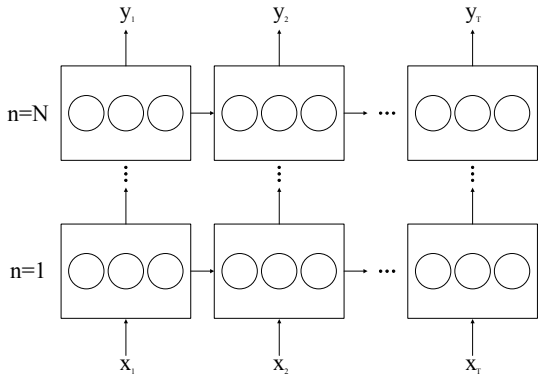

Figure 3. Block diagram of LSTM layers.

4. Experiments and Results

For evaluating the proposed method, we have compared the LSTM-based speech/music classification algorithm with the EVS method. The evaluation was implemented on the TIMIT speech database [16], Billboard year-end CDs from 1980 to 2013, and classical music CDs. From the music CDs, different genres (jazz, hip-hop, classic, blues, etc.) of music were collected. The entire music database was 221 h long and a large amount of data was employed in the experiment.

For training the LSTM-based speech/music classifier, 60 h of speech signal and 160 h of music signal were randomly selected from database. The length of each speech segment ranged from 6 to 12 s, and the length of each music segment ranged from 3 to 12 min. To create noisy environments, we added babble, car, white, pink, factory1, and factory2 noises from the NOISEX-92 database to the clean speech data at 5, 10, 15, and 20 dB SNR. As initial parameters of the proposed LSTM, we have used the parameter setting listed in Table 1.

Table 1. Parameter setting of the proposed LSTM.

Parameter	Value
Number of first hidden layer units	200
Number of second hidden layer units	600
Number of third hidden layer units	200
Learning rate	0.00006
Dropout rate	0.2

For testing, we randomly chose 20 h of speech data and 61 h of music data, which were separated from the training data. The data were sampled at 16 kHz with a frame size of 20 ms. To calculate the accuracy of the algorithm, each frame was manually labelled and compared to the corresponding classification results of the classifier.

For proper understanding of the performance difference, the results of the speech/music classification in conjunction with the test speech/music segment are shown in Figure 4. It can be observed from this figure that the proposed method has effectively classified speech and music, according to manual marking (silence = 0, speech = 1, music = 2), and has yielded better results in white noise (5 dB SNR) conditions when compared with the EVS based algorithm and RNN based algorithm.

To appraise the performance of the proposed algorithm, the speech/music classifier accuracy of the algorithm was investigated. The results are shown in Table 2. The test results verified that the proposed LSTM based method effectively improves the performance of the EVS. In particular, in the case of pink and factory1 noise environments at 5 dB SNR, the accuracy of the proposed method is significantly improved when compared with that of the EVS-based method. In the comparison between RNN and LSTM, the performance of LSTM is shown to have better performance than the RNN in all conditions.

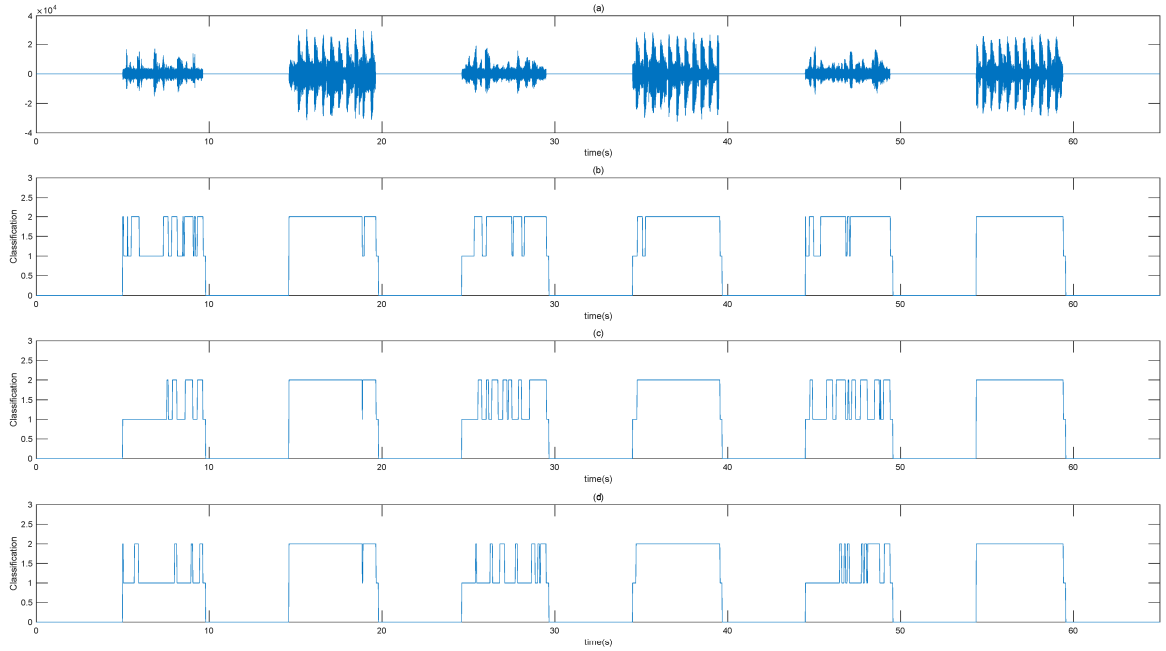

Figure 4. Speech with white noise (5 dB SNR) and music waveform, and the decision result of the speech/music classification algorithms (manual marking:silence = 0, speech = 1, music = 2). (**a**) Test waveform, (**b**) decision of EVS codec, (**c**) decision of RNN based algorithm, and (**d**) decision of LSTM algorithm.

Table 2. Comparison of speech/music classification accuracy.

		dB	EVS	RNN	Proposed
speech	clean	-	0.9980	0.9981	0.9985
	car	5 dB	0.8830	0.8911	0.9135
		10 dB	0.9131	0.9414	0.9517
		15 dB	0.9933	0.9940	0.9951
		20 dB	0.9927	0.9941	0.9943
	babble	5 dB	0.5757	0.6157	0.6524
		10 dB	0.7214	0.7561	0.7752
		15 dB	0.9649	0.9710	0.9760
		20 dB	0.9905	0.9914	0.9920
	white	5 dB	0.6633	0.6917	0.7017
		10 dB	0.8362	0.8581	0.8710
		15 dB	0.8792	0.9012	0.9104
		20 dB	0.9642	0.9775	0.9867
	pink	5 dB	0.3752	0.6016	0.6312
		10 dB	0.6900	0.6925	0.8110
		15 dB	0.9312	0.9450	0.9757
		20 dB	0.9853	0.9937	0.9948
	factory1	5 dB	0.3289	0.5948	0.6391
		10 dB	0.6702	0.7491	0.8471
		15 dB	0.9265	0.9481	0.9513
		20 dB	0.9850	0.9871	0.9906
	factory2	5 dB	0.7251	0.8362	0.8974
		10 dB	0.9119	0.9341	0.9591
		15 dB	0.9711	0.9721	0.9812
		20 dB	0.9855	0.9856	0.9878
music	classical	-	0.9868	0.9912	0.9931
	others	-	0.9311	0.9435	0.9514

5. Conclusions

In this paper, a robust method for enhancing the performance of speech/music classification of the 3GPP EVS codec has been proposed. The proposed method, based on the EVS, uses the feature vectors, which shows statistically superior performance in the encoding process of the EVS. The experimental results have shown that the performance was improved in low-SNR noise environments compared to high-SNR noise environments. The key idea of this paper is to obtain an effective method to integrate the LSTM within the EVS for speech/music classification.

Author Contributions: S.-I.K. wrote this manuscript; S.L. contributed to the writing, direction and content, and also revised the manuscript.

Funding: This research was funded by INHA UNIVERSITY Research Grant.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gao, Y.; Shlomot, E.; Benyassine, A.; Thyssen, J.; Su, H.-Y.; Murgia, C. The SMV algorithm selected by TIA and 3GPP2 for CDMA application. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; Volume 2, pp. 709–712.
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. April 2005. Available online: https://www.3gpp2.org/Public_html/Specs/C.S0052-0_v1.0_040617.pdf (accessed on 6 November 2018).
3. 3GPP Spec. Codec for Enhanced Voice Services (EVS), Detailed Algorithm Description, TS 26.445, v.12.0.0. September 2014. Available online: https://www.etsi.org/deliver/etsi_ts/126400_126499/126445/12.00.00_60/ts_126445v120000p.pdf (accessed on 6 November 2018).
4. Saunders, J. Real-time discrimination of broadcast speech/music. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA, USA, 7–10 May 1996; pp. 993–996.
5. Fuchs, G. A robust speech/music discriminator for switched audio coding. In Proceedings of the Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 569–573.
6. Lim, C.; Chang, J.-H. Improvement of SVM-Based Speech/Music Classification Using Adaptive Kernel Technique. *IEICE Trans. Inf. Syst.* **2012**, *9*, 888–891. [CrossRef]
7. Lim, C.; Chang, J.-H. Efficient implementation techniques of an svm-based speech/music classifier in smv. *Multimed. Tools Appl.* **2015**, *74*, 5375–5400. [CrossRef]
8. Song, J.H.; Lee, K.H.; Chang, J.H.; Kim, J.K.; Kim, N.S. Analysis and Improvement of Speech/Music Classification for 3GPP2 SMV Based on GMM. *IEEE Signal Process. Lett.* **2008**, *15*, 103–106. [CrossRef]
9. Song, J.-H.; An, H.-S.; Lee, S.M. Speech/music classification enhancement for 3gpp2 smv codec based on deep belief networks. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2014**, *97*, 661–664. [CrossRef]
10. Malenovsky, V.; Vaillancourt, T.; Zhe, W.; Choo, K.; Atti, V. Two-stage speech/music classifier with decision smoothing and sharpening in the EVS codec. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Brisbane, Australia, 19–24 April 2015; pp. 5718–5722.
11. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
12. Jozefowicz, R.; Zaremba, W.; Sutskever, I. An empirical exploration of recurrent network architectures. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2342–2350.
13. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef] [PubMed]
14. Karnebäck, S. Discrimination between speech and music based on a low frequency modulation feature. In Proceedings of the Eurospeech, Aalborg, Denmark, 3–7 September 2001; pp. 1891–1984.

15. Dempster, A.P.; Lairdand, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–38.
16. Fisher, W.M.; Doddington, G.R.; Goudie-Marshall, K.M. The DARPA speech recognition research database: Specification and status. In Proceedings of the DARPA Workshop Speech Recognition, Palo Alto, CA, USA, 19–20 February 1986; pp. 93–99.