INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Storyteller's tale: Literature audiobooks Genre classification using CNN and RNN architectures

Nehory Carmi[1], *Azaria Cohen*[1], *Mireille Avigal*[1], *Anat Lerner*[1]

[1]Department of Mathematics and Computer Sciences, the Open University of Israel
{nehorayc,azaria.cohen}@gmail.com, miray@openu.ac.il, anat@cs.openu.ac.il

Abstract

Identifying acoustic properties that characterize reading literary genres can assist in giving a more personal and human tone to the speech of bots and automatic readings.

In this paper we consider the following question: given speech segments of audiobooks, how well can we classify them according to their literary genres? In this study we consider three different literary genres: children, horror and suspense, and humorous audio books, taken from two free audio books sites: Librivox and YouTube.

We ran four classification experiments: three for each pair of genres, and one for all three genres together. We repeated each experiment twice, with two different network architectures: Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN).

Note that, throughout the reading, there are sections that are more typical to the book's genre than others. As the samples were taken sequentially throughout the reading of the books and were short in duration, we did not expect high classification rates. Nevertheless, the accuracy of all the experiments were at least 72% for all the pair's classifications; and at least 57% for both architectures for the three classes classifications.

Index Terms: speech emotion recognition, deep learning, prosody, acoustic features, literary genres.

1. Introduction

Speech processing using deep learning tools is receiving increasing attention during the last few years. New directions are supported by advances in pattern recognition abilities of Deep Neural Networks (DNNs) and the emergence of new more efficient architectures of DNN, such as Gated Recurrent Unit (GRU) [1] and ResNet [2]. These architectures were originally developed for different applications such as Speech Emotion Recognition (SER), Automatic Speech Recognition (ASR) and music classification, but were quickly adapted for speech analysis [3].

The field of speech analysis encompasses topics like SER and various speech classifications. Many papers referred to discrete emotions classifications from human speech for example, in [4] Trigeorgis et al. used a combination of CNN and RNN to classify 2 emotions with 78% accuracy, while in [5], Ghosh et al. used RNN with accuracies of 70.28% for 2 classes and 48.88% for 4 classes.

ASR identifies and processes the human voice for many different tasks, it is primarily used to convert spoken language into text. Additional uses include speaker diarization and biometric identification of the speakers voice fingerprint. Both CNN [6], GRU and Long Short Term Memory (LSTM) [7] has shown very high results in the ASR field.In The field of musical genre classification, Costa et al. [8] classified the GTZAN database, containing 1000 songs of 10 different musical genres. They based the classification on the spectrograms and several acoustic features of each audio segment. They repeated the experiments with both CNN and RNN. The resulted accuracy for the CNN was 89.82, and for the RNN was 97%.

As far as we know, there is no paper with literary genre classification. In the art of storytelling, the narrators tone and voice are highly correlated to the genre of the story. In [9], Crawford offers an audiobook narration course and recommend each future narrator to specialize in a specific genre that is best suited to that person's voice. Audible, Amazons audiobook service, published a list of the most successful narrators for each literary genre and explained the reason why the narrator's voice suits the specific genre.

Storytelling is an art preserved trough myths and history to our days. Storytellers, also known as narrators, convey stories and feelings through their voices, regardless of whether the stories are their own or not. As research shows, the human voice contains vast information of emotions and nuances which we perceive in both conscious and sub conscious levels [10]. The voice is the main artistic tool of professional narrators, encompassing a lot of emotional information.

Since a human listener can usually distinguish between literary genres based only on hearing, we assume a DNN can be trained to classify literary genres as well. Such automatic classification can be useful, for example, for Text to Speech (TTS) missions, by making the sound more ”human-like” while reading a specific genre. In this paper we consider the following question: given speech segments of audiobooks, how adequate can we classify them according to their literary genres? We suggest a supervised deep learning method based on annotated audiobooks. To the best of our knowledge, the use of audiobooks in other papers is mainly for speech synthesis, for text to speech tasks [11] and for automatic speech recognition [12]. For the various analyses, the audiobooks are usually decomposed to its linguistic building blocks, segmented into words and even syllables. Yet by doing that, the literary context of the audiobook and the genre specifications are lost.

Paper's structure. The paper is organized as follows: in section 2, we discuss the methods including the database creation, data pre-processing, feature extraction, and network's architectures. In section 3, we present the results. Finally, in section 4 we conclude and suggest future work.

2. Methods

2.1. Database Description

The audio books were taken from LibriVox [13], a free domain library for downloading audiobooks. This library also provides meta-data for each book, like the literary genre, information regarding the narrator and more. For the current research, we downloaded only three literary genres. We labeled the records

http://dx.doi.org/10.21437/Interspeech.2019-1154

with three classes according to the genre: short children stories (class 0), horror, detective and mystery (class 1) and humor and satire (class 2).

To prevent classification dependency on the narrators, we chose books read by several dozens narrators for each class, and for the test set we used narrators that the network had not been trained on.

Table 1 shows the database specification. The total number of books in the database is 1077, with a total duration of about 180 hours, the metadata was used to balance the different classes of the database based on gender and total reading time (rather than the number of books).

Table 1: *Audiobook database structure*

Attribute	**Class 0**	**Class 1**	**Class 2**
No. Books	556	173	348
No. Narrators	120	125	187
Total time (h)	60:28:38	60:22:58	61:56:38

2.2. Database Pre-processing

Each file in Librivox starts with a 30 seconds introduction that is not a part of the book. We removed these prefixes from the audio files.

The audiobooks length ranged from 5 minutes to 12 hours. We segmented each book into equal-length segments of 8 seconds each, regardless of the contents. We chose a relatively short segment size as we believe that the atmosphere of each genre is reflected throughout the reading. For example, humor in horror books is sometime more like a parody on humor rather than real humor and so is horror in humor books. The segmentation was done using MATLAB.

We then divided each segment into 77 frames using a 250 ms window and a hop-length of 100 ms for the next step of feature extraction.

2.3. Feature extraction

We extracted the following 59 audio features for each frame, using LibROSA [14], a Python package for music and audio analysis: 40 Mel-Frequency Cepstral Coefficients (MFCCs); 12 Chroma and 7 Spectral Contrast (SC). Some of these features were suggested in [15], for temporal analysis and emotion detection. The number of the MFCCs extracted was chosen upon trial and error. The feature extraction resulted in a 77×59 features matrix for each segment.

2.3.1. Mel-frequency cepstral coefficients

MFC is a representation of linear cosine transform on a log of the energy spectrum of an audio signal. MFCCs are the Discrete Cosine Transform (DCT) coefficients of the MFC. They are derived from a type of cepstral representation of the audio file (a nonlinear ”spectrum-of-a-spectrum”). The MFCC frequency bands are equally spaced throughout the Mel-scale, closely imitating the human ear hearing process [16].

As the MFCC closely represent a natural human hearing process, its coefficients are useful in speech classification of either speech recognition [17] or emotion recognition [18].

2.3.2. chroma

Chroma or Chromogram is a spectrogram of 12 frequency bins representing the 12 semitones. It is mostly used in music classification [8]. We assumed that each literary genre consists of a distinct musical element, for example: the long drawn pauses of the horror genre or the high pitch ending of children’s stories. Therefore, the Chroma features might carry the information needed for literary genre classification, although on a much smaller scale than in classifying different musical genres.

2.3.3. Spectral contrast

Spectral contrast is a feature representing the spectral characteristics of a audio signal. The spectral contrast divides each signal into n frequency bands as follows:

$$\left[0, \frac{w_0}{2^n}\right), \left[\frac{w_0}{2^n}, \frac{w_0}{2^{n-1}}\right), \ldots, \left[\frac{w_0}{2^2}, \frac{w_0}{2^1}\right] \qquad (1)$$

in order to get more information from each spectral band. We extracted the 7 spectral contrast features, for the 7 sub-bands (bands 0-6 in Table 2 at a sample rate of 22050Hz.

Table 2: *The frequencies of the seven sub-bands*

Band Number	**Min Frequency**	**Max Frequency**
0	0	200
1	200	400
2	400	800
3	800	1600
4	1600	3200
5	3200	6400
6	6400	12800

$$Peak(b) = \log\left(\frac{1}{\alpha N_b}\Sigma_{i=1}^{\alpha N_b} P_{(b,i)}\right) \qquad (2)$$

$$Valley(b) = \log\left(\frac{1}{\alpha N_b}\Sigma_{i=1}^{\alpha N_b} P_{(b,N_{b-1+1})}\right) \qquad (3)$$

$$SC(b) = Peak(b) - Valley(b) \qquad (4)$$

The spectral contrast parameters can capture changes in the rhythm and in the harmonics. Therefore, they are often used for music genres classifications [19]. As with the Chroma parameters, we believe that narrators apply different rhythms for the various literary genres. We thus hypothesized that the same features used for music genre classification can be useful for literary genre classification as well.

2.4. Networks’ architectures

Analysis of the data was performed using the Keras API [20] with the Tensorflow environment [20].

2.4.1. Database filtering

Using the Pandas package for Python, we filtered data to test the effects of specific parameters on the classification accuracy.

We divided the database into test and train sets, 80% for the train and 20% for the test, based on the total reading time. Table 3 shows the database splitting for the learning stage.

Table 3: *Database splitting*

		Class 0	**Class 1**	**Class 2**
Train-set	No. books	435	138	285
	No. Narrators	100	98	156
	Total time (h)	50:10:32	50:03:47	51:39:28
Test-set	No. books	121	35	63
	No.Narrators	20	27	31
	Total time (h)	10:18:06	10:19:11	10:17:10

2.4.2. Network's structure

For the classification of the audio files we used two different network architectures: RNN and CNN [22].

The first architecture we examined was CNN. Since CNN is widely used for emotion detection, we assumed it could yield satisfactory results in literary genre classification as well.

The suggested model uses a model similar to ResNet [2], a residual CNN with increasing number of filters as we progress with the layers until achieving (or nearly achieving) a one-dimensional array at the output of the network. The network is compiled of five convolution layers (16 neurons, 32 neurons, 64 neurons, 128 neurons and 256 neurons). The first convolution layer is followed by a batch normalization layer, a max pooling layer and a dropout layer. Each one of the following four convolution layers is followed by a max pooling layer and a dropout layer. Following the five convolution layers, we flatten the results to a single dimensional vector and use fully connected layers. We then output the classification using a softmax activation.

The second architecture we examined was RNN for its temporal analyses abilities. RNN enables weights sharing across time, that is, we can share weights of a single layer with a future input vector of the network [22].

We used the RNN version based on Gated Recurrent Units (GRU). GRU is a variation of the Long Short Term Memory (LSTM) in the sense that both architectures apply gates (innate structures). The gates control the flow of information and determine whether the data will be forwarded to the next layer or will be discarded. GRU uses only 2 gates unlike the 4 gates in LSTM, which improves the performance of GRU. Since the order of the audio samples of a book is meaningful, we assumed that a network consisting of memory is due. Therefore, we used RNN trying to capture the unique rhythm of temporal properties each genre possesses.

The network consists of three layers of GRU units (32 neurons, 16 neurons and 8 neurons), followed by a dropout, intended to avoid overfitting. The size of the initial filter is approximately 40% (32 neurons in the input layer versus 77 frames that represents the time steps) of the total number of samples in each segment and it is decreased by a half in each successive layer of the network.

Both RNN and CNN architectures have a tendency to overfitting. To avoid overfitting, we implemented an early stop mechanism in both of them and stop the training session upon three consecutive epochs with no decrease in the loss function value.

3. Results

We conducted two kinds of classification for the 8 seconds segments audio files. The first one is a 2-class classification (children versus horror, children versus humor and horror versus humor), and the second is 3-class classification (children, horror and humor). Table 4 summarizes the performance measures values of the 2-class classification for each of the two architectures, CNN and RNN.

Table 4: *2-class classification performance rates*
Children -0 Horror - 1 Humor - 2

		0 vs 1	**0 vs 2**	**1 vs 2**
CNN	Precision	0.75	0.64	0.74
	Recall	0.81	0.73	0.80
	Accuracy	0.82	0.72	0.80
GRU	Precision	0.71	0.63	0.72
	Recall	0.73	0.73	0.71
	Accuracy	0.74	0.73	0.73

Table 5 summarizes three peformance mesures values of the 3-class classification for each of the two networks architectures used.

Table 5: *3-class classification performance rates*

		children	**Horror**	**Humor**
CNN	Precision	0.40	0.73	0.57
	Recall	0.62	0.66	0.43
	Accuracy	0.57		
GRU	Precision	0.80	0.95	0.29
	Recall	0.90	0.70	0.6
	Accuracy	0.6		

The results presented in table 4 show classification accuracy rates higher than 72% in the three 2-classes cases for both architectures. The CNN architecture performs better than the GRU architecture in all the 2-classes tests. On the contrary, as seen in table 5, for the 3-classes test GRU performs better than CNN. It is worth noting that the classification performance rates of horror versus each of the other two genres in CNN is notably higher than the classification of children versus humor genres, while in GRU it is approximately the same. In the 3-classes test, humor genre seems to have the lowest rates in the two architectures.

4. Discussion and future work

In this study we classified literary genres, studied some key features for distinguishing between the genres and compared the performance of two network's architectures.

We have managed to classify three different genres using short speech segments of eight seconds.

The CNN architecture is usually used for visual analysis. For audio analysis, it is used in emotion detection. We assumed that the CNN results would be better when classifying between genres that greatly differ in terms of the narrator expressed emotions, such as horror and children stories.

RNN is usually used to detect temporal data such as rhythms. Therefore, we pre-assumed that GRU would detect more subtle temporal and rhythmic properties of each genre, even if the genres do not defer on the more prominent emotional level. This might explain the fact that the accuracy of the three 2-classes tests are similar regardless of the emotional possible differences.

In this study, we focused on the narrator's prosody, which is apparently influenced by the targeted audience. We ignored the wording that can be added in future research. In both humor and satire the goal is to make people laugh trough different means, while in children's literature the text itself matters but not as much as the atmosphere the narrator wishes to create.

We conclude that both CNN and GRU models can be used to classify different literary genres. We assume that GRU implies a classification solution unrelated to the genre emotional emphasis while CNN can easily detect genres, which vary greatly by the emotional expression of the narrator. Furthermore, the architectures and features which were used in music genre classification were proved useful in literary genre classification which indicates a certain similarity between the tasks. This work can be extended for synthesizing and creating better TTS systems that will read each literary genre with the nuances a human narrator can achieve. It can also be applied in bot-human communication adjusting the bot prosody to the targeted audience.

5. References

[1] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] Z. Wu and S. King, "Investigating gated recurrent neural networks for speech synthesis," *arXiv preprint arXiv:1601.02539*, 2016.

[4] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[5] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition." in *INTERSPEECH*, 2016, pp. 3603–3607.

[6] D. Palaz, R. Collobert *et al.*, "Analysis of cnn-based speech recognition system using raw speech as input," Idiap, Tech. Rep., 2015.

[7] A. N. Shewalkar, "Comparison of rnn, lstm and gru on speech recognition data," 2018.

[8] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied soft computing*, vol. 52, pp. 28–38, 2017.

[9] C. Crawford, *Chris Crawford on interactive storytelling*. New Riders, 2012.

[10] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural computing & applications*, vol. 9, no. 4, pp. 290–296, 2000.

[11] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[13] https://librivox.org/.

[14] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[15] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech communication*, vol. 52, no. 7-8, pp. 613–625, 2010.

[16] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[17] J. Jo, H. Yoo, and I.-C. Park, "Energy-efficient floating-point mfcc extraction architecture for speech recognition systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 2, pp. 754–758, 2016.

[18] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using mfcc and support vector machine," in *Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on*. IEEE, 2016, pp. 1080–1084.

[19] C.-H. Lee, J.-L. Shih, K.-M. Yu, and J.-M. Su, "Automatic music genre classification using modulation spectral contrast feature," in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 204–207.

[20] F. Chollet *et al.*, "Keras," 2015.

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[22] I. Goodfellow, Y. Bengio, and A. Courvil, *Deep Learning*. MIT Press, 2016.