Sentiment Analysis Using Image-based Deep Spectrum Features

Shahin Amiriparian*†‡, Nicholas Cummins*† Sandra Ottl† Maurice Gerczuk† and Björn Schuller*§
*Chair of Embedded Intelligence for Health Care & Wellbeing, Augsburg University, Augsburg, Germany
†Chair of Complex & Intelligent Systems, Universität Passau, Germany
‡Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
§GLAM – Group on Language, Audio & Music, Imperial College London, London, UK

Email: shahin.amiriparian@tum.de

***Abstract*—We test the suitability of our novel deep spectrum feature representation for performing speech-based sentiment analysis. Deep spectrum features are formed by passing spectrograms through a pre-trained image convolutional neural network (CNN) and have been shown to capture useful emotion information in speech; however, their usefulness for sentiment analysis is yet to be investigated. Using a data set of movie reviews collected from YouTube, we compare deep spectrum features combined with the bag-of-audio-words (BoAW) paradigm with a state-of-the-art Mel Frequency Cepstral Coefficients (MFCC) based BoAW system when performing a binary sentiment classification task. Key results presented indicate the suitability of both features for the proposed task. The deep spectrum features achieve an unweighted average recall of 74.5 %. The results provide further evidence for the effectiveness of deep spectrum features as a robust feature representation for speech analysis.**

1. Introduction

Expressing emotions and sentiments is a central part of human communication [1]. We do this by laughing and grinning when being happy, yawning when being bored or panicking and crying when being upset. Some of those actions are performed quietly, so one can only recognise them via visual information, but other actions like laughing or crying are also noticed by listening [2]. Nowadays, such communications not only take place in person, but also on social networks such as Facebook, Twitter and Youtube where users express their thoughts and emotions using audio, visual, and textual communication channels [3].

The YouTube social media platform has over a billion users who communicate by uploading, watching, and commenting on videos. These videos are from a large variety of different genres such as music, comedy, movie, entertainment, and gaming. Hundreds of hours of video content is added every minute [4], creating a rich source of videos that can be targeted for accumulating data for large data sets. YouTube videos often showing their creators' emotions, these videos convey a broad range of sentiments.

In this study, we perform audio-based sentiment analysis on movie reviews posted to YouTube [5]. This is essentially a polarity classification task [6], in which videos are assigned to a positive or negative class based on whether or not the presenter liked or disliked the movie they are reviewing. This task is generally achieved in a multimodal framework combining linguistic, speech, and visual cues. However, given promising results recently published for speech-based emotion recognition using either *bag-of-audio-words* (BoAW) [7] or novel deep spectrum features [8], this paper discusses the suitability of these feature representations for sentiment recognition.

BoAW features are a sparse audio feature representation in the form of a fixed length histogram. The histogram represents the frequency identified for 'audio words' in a given audio instance as determined through a quantisation procedure [9]. BoAW representations are generally considered to be more robust than using standard acoustic features; the quantisation steps help minimise effects relating to small amounts of noise present in the original feature space. BoAW features have given state-of-the-art performance for the challenging task of continuous emotion recognition on the *REmote COLlaborative and Affective interactions* (RECOLA) data set [7].

Deep spectrum features on the other hand, are computed from audio data by extracting image CNN descriptors from spectrogram plots and have been shown to produce strong results in a range of audio tasks including snore sound detection and speech-based emotion recognition [8], [10]. Recently, deep spectrum features have been shown to outperform expert-designed audio feature sets such as the Geneva Minimalistic Acoustic Parameter Set for affective computing [11], when performing speech based emotion recognition on the FAU-AIBO data set in both clean and noisy recording conditions [8].

The rest of this paper is laid out as follows: Section 2 introduces the Movie Review Data Set, Section 3 outlines our proposed deep spectrum sentiment classification system, our key experimental settings are outlined in Section 4, with the subsequent results given in Section 5. Finally, we summarise our findings and provide concluding remarks in Section 6.

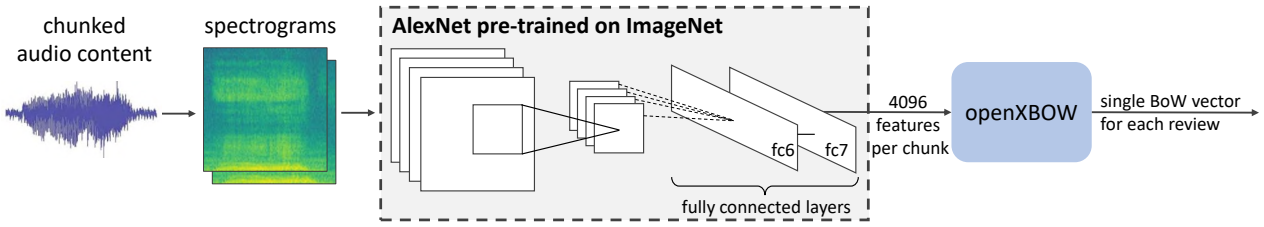

Figure 1: Spectrograms are generated from the chunked audio files they are then used as input for an AlexNet CNN pre-trained on ImageNet and the activations of the last fully connected layer are extracted as large deep spectrum feature vectors. This results in 4096 features for each chunk which are bagged into a single feature vector for each movie review.

2. Movie Review Data Set

The *Movie Review Data Set*, collected by Wöllmer et al. [5], consists of 359 YouTube clips in which people express their opinion on a selection of movies they have previously watched. The video data is in *avi* format, the audio data is in *wav* format, and the included transcriptions of the spoken content is provided in a *trs* format.

The sentiment of the speaker in each video is expressed as integer annotations on a *1–6* Likert scale, *1* the negative and *6* the positive end of the sentiment spectrum. Based on this, the clips are separated into positive and negative, with average scores above *3.5* assigned to *positive* sentiment. This is implemented on each YouTube clip, across each format (avi, wav and trs), dividing the data into three folders video, audio and transcriptions, with each containing two folders positive and negative. Furthermore, for each clip, meta data for each *speaker ID*, including gender, age and ethnicity are listed.

For sentiment classification, to test a trained model on unseen data, it is necessary to divide the data into train, development and test sets. For the creation of the test set an associative array with all speaker IDs is produced, containing the respective gender, age, ethnicity, with clips chained to their respective sentiment value. Random test sets including 55 IDs are then created. The clips for the test sets are chosen via the speaker meta data in this way making the test set speaker independent. To achieve balanced data for successful testing and training of the model, both positive and negative clips and speakers of each age, gender and ethnicity are distributed across all partitions (train, development, and test). The final test set includes precisely 20 % of all negative and almost exactly 20 % of all positive clips from the Movie Review Data Set. Moreover, the average age of speakers in the test set is 19.2 years, with 19.5 % of those being male and 20.7 % female. The same procedure is applied when selecting development set data. The development set ultimately consists of almost 20 % of all positive and negative clips from the data set. Average age for the speakers is 20.1 years, and the percentage of male and female speakers from the entire data set is 20.7 % and 18.8 % respectively. The distribution across the three sets, for the Movie Review Data Set is shown in Table 1.

TABLE 1: Distribution, average length and standard deviation of all videos from the Movie Review Data Set between train, development and test sets.

statistics	total	train	devel	test
# videos	359	215	72	72
# positive videos	209	125	42	42
# negative videos	150	90	30	30
average length (m:s)	2:31	2:32	2:27	2:30
standard deviation (m:s)	0:41	0:40	0:41	0:44

3. Proposed System

Deep spectrum feature extraction consists of three components: a) spectrograms generation from the chunked audio files, b) a pre-trained image-based deep CNN used for the extraction of *deep spectrum* features from spectrogram, and c) openXBOW [9] to create BoAW representation from the deep spectrum features.

3.1. Feature Extraction

Before being able to create BoAWs, audio features are extracted with openSMILE [12], [13] and AlexNet [14], an image CNN descriptor. We use openSMILE to extract 39 Mel Frequency Cepstral Coefficient (MFCC) features, and AlexNet to extract deep spectrum features.

3.1.1. MFCC Features. The 39-dimensional MFCC feature representation (12-MFCCs, 12-ΔMFCCs, 12-ΔΔMFCCs, E, ΔE, and ΔΔE, where E stands for logarithmic energy of the input speech signal) are extracted using the following steps: the frame size is set to 25 ms at a rate of 10 ms. A Hamming function is used to window the frames and a pre-emphasis with $\alpha = 0.97$ is applied. The MFCCs 0/1-12 are computed from 26 Mel-bands computed from the FFT power spectrum. The frequency range of the Mel-spectrum is set from 0 to 8 kHz. This process is done for all files of Movie Review Data Sets' train, development and test set. Afterwards, the respective data for train and development are combined to form the train-devel set. In addition to MFCCs, deep spectrum features, as used in [8], [10]

3.1.2. Deep Spectrum Features. The deep spectrum features used by our system are extracted from spectrograms of

TABLE 2: Comparison of UARs for BoAWs created from either MFCCs or Deep Spectrum features. Different combinations of codebook size and multi-assignment degree a are evaluated. The SVM's cost parameter C is optimised on the development partition. The chance-level is 50 % UAR.

		a=1			a=10			a=20			a=50		
Feature Space	size	C	dev	test	C	dev	test	C	dev	test	C	dev	test
MFCC	2000	1.0	71.7	68.1	10^{-1}	68.6	70.3	10^{-1}	67.7	70.3	10^{-4}	69.3	62.9
MFCC	4000	10^{-2}	**71.9**	72.2	10^{-2}	68.9	72.2	10^{-2}	68.9	69.8	10^{-2}	67.2	69.8
MFCC	8000	10^{-2}	71.2	68.9	10^{-2}	71.7	**73.1**	1.0	69.3	66.5	10^{-5}	66.2	67.7
Deep Spectrum	500	10^{-4}	69.5	**74.5**	10^{-6}	74.3	71.0	10^{-5}	74.3	70.5	10^{-5}	69.5	71.0
Deep Spectrum	1000	10^{-2}	74.5	55.8	10^{-4}	**74.8**	68.6	10^{-5}	74.3	70.5	10^{-6}	71.9	69.8
Deep Spectrum	2000	10^{-4}	75.3	64.1	10^{-4}	70.7	68.6	10^{-5}	73.1	70.5	10^{-5}	71.9	71.0
Deep Spectrum	4000	10^{-3}	69.5	64.5	10^{-4}	70.7	72.2	10^{-5}	71.9	70.5	10^{-5}	71.9	71.0

equally sized chunks of the audio content of a movie review using the pre-trained image classification CNN AlexNet [14]. We obtain a version of AlexNet that has been trained on the large ImageNet corpus from the the Caffe [15] model-zoo[1]. Power spectrograms with Hanning windows of width 16 ms and overlap 8 ms are plotted using Python's matplotlib [16] for 5 s windows with a hop size of 4 s. Forwarding these spectrogram plots through AlexNet and extracting the activations of the second to last fully connected layer yields feature vectors of size 4 096 for each audio chunk.

3.2. Creation of Bag-of-Audio-Words

To obtain a review level feature representation from the deep spectrum features extracted from audio chunks, we use openXBOW to construct bag-of-audio-words. First, all input vectors are min-max normalised to $[0, 1]$, standardisation has been found to lead to inferior results. We then create codebooks by random sampling vectors and quantising all input vectors according to their respective nearest vectors found in the codebook. For each review, this produces a single feature vector from the chunk level deep spectrum features.

4. Key Experimental Settings

The constructed BoAWs are used for training a linear support vector machine (SVM). Unless otherwise specified, input normalisation is applied. The imbalance of the data sets is counteracted by adjusting the weights for the two classes accordingly, i. e. 1.4 for negative and 1.0 for positive instances. Training and evaluating that system is performed using the WEKA's [17], LibLINEAR wrapper with the *L2-regularised L2-loss* solver [18]. Performance of the systems is measured by the *unweighted average recall* (UAR).

5. Results

When training a SVM with BoAWs, two different types of BoAW are observed; BoAWs constructed from MFCCs (cf. Section 5.1) and BoAWs constructed from deep spectrum features (cf. Section 3.1.2). These BoAWs are used for the experiments conducted in Section 5.2.

1. https://github.com/BVLC/caffe/wiki/Model-Zoo

5.1. Bag of MFCCs

BoAWs constructed from a codebook with $a = 10$ are providing the best UAR results (cf. Table 2). This indicates that, when creating a BoAW from a codebook with the size between 2 000 and 8 000, for each feature vector ten suitable vectors are found from the quantised vectors of the codebook. However, there can be made no assumption about the codebook size having influence on the outcome of UAR as at times, the values are best when using a small codebook size and at other times with a bigger size. A trend that can be connected to this result, is observed as results of using BoAWs created from MFCCs are improving with a bigger codebook size for arousal but are getting worse with larger codebooks for valence [7].

5.2. Bag of Deep Spectrum Features

When comparing the results of using BoAWs constructed from MFCCs and BoAWs created from deep spectrum features, it can be observed that these unconventional speech features perform slightly better than MFCCs that are standard and robust speech features. This improvement could be attributed to the substantially higher quantity of extracted MFCC vectors than deep spectrum features. From the Table 2 it can also be discerned that BoAWs coming from codebooks of size 500 and 1 000 are producing the best results. These work better than bigger sizes because the smaller ones are more capable of generalising though not necessarily more discriminative [19].

To improve on the previous results, BoAWs are again created from deep spectrum features with the help of openXBOW. In contrast to previously constructed BoAWs, openXBOW's parameters *supervised* and *norm 1* are now used additionally to *normalizeInput*. Utilising *supervised* generates a codebook for each class separately and then merges all codebooks. The BoAWs are normalised with respect to the length of the input using parameter *norm 1*. Specifically, this parameter states that the term frequencies are divided by the number of input frames. After having created those new BoAWs from codebooks of size 1 000 and $a = 10$ for the Movie Review data, the experiment structure is executed with the help of those BoAWs. The resulting UAR values as depicted in Table 3 show a slightly worse outcome. However, the imbalance of positive and negative

TABLE 3: Comparison of UAR's for different deep spectrum BoAWs systems where the codebooks were formed either in an unsupervised (random) manner or in a supervised (k-mean clustering) manner. C is optimised on the development partition. The chance level is 50 %.

	unsupervised			supervised		
class	C	dev	test	C	dev	test
positive	10^{-4}	92.9	73.8	10^{-8}	88.1	78.6
negative	10^{-4}	56.7	63.3	10^{-8}	56.7	60.0
average	10^{-4}	74.8	68.6	10^{-8}	72.4	69.3

recall values is evened out a bit as the difference of positive and negative recalls is now smaller, see Table 3.

6. Summary and Conclusions

In this study we explored a combination of deep spectrum features and bag-of-audio-words for sentiment analysis. Using an existing corpus of movie reviews collected from YouTube we compared our proposed system with a state-of-the art MFCC BoAW system. Presented results indicate that the deep spectrum based systems consistently outperform the equivalent MFCC system. Given the strong performance of deep spectrum features in the related task of emotion classification [8] this result is not unexpected. This result adds to the growing evidence in the literature that feeding spectrogram representations through pre-trained image CNNs produces salient features suitable for audio classification tasks.

Future work will include fusing information from the different modalities present in the Movie Review corpus to further improve on the deep spectrum results. We also plan to test deep spectrum features extracted from different image CNN's such as VGG19 and GoogLeNet, for the task of sentiment analysis. We also plan to collect further review data from YouTube using our purpose built software [20]; we are interested in performing cross-corpus multimodal sentiment analysis using reviews of different genres and from different cultures.

7. Acknowledgements

This work was supported by the European Unions's seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA).

References

[1] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[2] K. L. Burns and E. G. Beier, "Significance of Vocal and Visual Channels in the Decoding of Emotional Meaning," *Journal of Communication*, vol. 23, no. 1, p. 118, 1973.

[3] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social mediasentiment of microblogs and sharing behavior," *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217–248, 2013.

[4] YouTube, "YouTube - Statistics & Facts," website, available at https://www.statista.com/topics/2019/youtube/; last reviewed on June 15th 2017.

[5] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "Youtube Movie Reviews: Sentiment Analysis in an Audio-visual Context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[6] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

[7] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech." San Francsico, CA, USA: ISCA, 2016.

[8] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *Proc. of the 25th ACM International Conference on Multimedia, MM 2017*. Mountain View, CA: ACM, October 2017, 7 pages.

[9] M. Schmitt and B. W. Schuller, "openXBOW — Introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, 2017.

[10] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proceedings of INTERSPEECH: 18th Annual Conference of the International Speech Communication Association*. Stockholm, SE: ISCA, 2017, 5 pages.

[11] F. Eyben, K. R. Scherer, B. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*. Florence, IT: ACM, 2010, pp. 1459–1462.

[13] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM Multimedia*. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. Orlando, US: ACM, 2014, pp. 675–678.

[16] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for Large Linear Classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[19] S. Pancoast and M. Akbacak, "Bag-of-Audio-Words Approach for Multimedia Event Classification." in *Proc. of INTERSPEECH*. Portland, OR, USA: ISCA, 2012, pp. 2105–2108.

[20] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *Proc. ACII 2017*. San Antionio, TX: IEEE, October 2017, 6 pages.