Improved Multimodal Sentiment Detection Using Stressed Regions of Audio

Harika Abburi, Manish Shrivastava and Suryakanth V Gangashetty
Language Technologies Research Center
IIIT Hyderabad, India
harika.abburi@research.iiit.ac.in, m.shrivastava@iiit.ac.in, svg@iiit.ac.in

***Abstract*—Recent advancement of social media has led people to share the product reviews through various modalities such as audio, text and video. In this paper, an improved approach to detect the sentiment of an online spoken reviews based on its multi-modality natures (audio and text) is presented. To extract the sentiment from audio, Mel Frequency Cepstral Coefficients (MFCC) features are extracted at stressed significant regions which are detected based on the strength of excitation. Gaussian Mixture Models (GMM) classifier is employed to develop a sentiment model using these features. From results, it is observed that MFCC features extracted at stressed significance regions perform better than the features extracted from the whole audio input. Further from the transcript of the audio input, textual features are computed by Doc2vec vectors. Support Vector Machine (SVM) classifier is used to develop a sentiment model using these textual features. From experimental results it is observed that combining both the audio and text features results in improvement in the performance for detecting the sentiment of a review.**

Keywords*—*Stressed significant regions, Sentiment analysis, Gaussian mixture model, Support vector machine, Audio features, Lyric features.

I. Introduction

The task of Sentiment analysis is to classify data into positive, negative and neutral categories based on the opinion. As of now most of the work on sentiment analysis is done on textual data. With increase in social media, people started sharing the information in the form of video, audio along with text. So multimodal data have been requirement with change in conventional source of communication.

To detect the sentiment from natural audio streams, a sentiment detection system is developed based on Maximum Entropy modeling and Part Of Speech tagging. Transcripts from audio streams are obtained using Automatic Speech Recognition (ASR) [1]. Prosodic features can be used to build the sentiment classifier [2]. Speech data is generally extracted from the characteristics of the vocal tract, excitation and prosody. In the literature MFCC, Linear Predictive Cepstral Coefficients (LPCC) are the major spectral features used for emotion recognition [3]. Various acoustic cues such as Energy of the excitation (EoE), loudness, strength of excitation (SoE), instantaneous F0, and their combinations are explored to study the emotion discriminating capabilities of the excitation signal [4]. Strength of excitation of audio input is found using Zero Frequency Filter (ZFF) method. Regions whose strength of excitation fluctuates above and below 30% from the mean strength of excitation referred as emotionally significant regions [5]. From the emotionally significant regions MFCC features are extracted and tested on GMM classifier. A significant improvement in the performance of a system is observed. In this paper, stressed significant regions are used to extract the sentiment of an audio input.

Text based sentiment classification is developed on movie reviews using SVM, Naive Bayes and Maximum Entropy classifiers [6]. Eight different types of features are extracted such as unigrams, bigrams, combination of both and so on. These features are tested on the three classifiers, among them SVM with binary-unigram features gives a high rate of detecting the sentiment. In [7] each sentence is labeled as either objective or subjective and neglected the objective sentences by finding minimum cuts in graphs. This prevents the classifier from considering misleading text. On this data, SVM and Naive Bayes classifiers are implemented to develop a model. From the result, it is observed that a small improvement is obtained. A work on sentiment analysis of online news articles is presented in [8]. By using Machine Learning for Language Toolkit (MALLET), six text-classification algorithms are compared such as the Naive Bayes, the Maximum Entropy, a decision tree rule base, a decision tree with the C4.5 algorithm, the Winnow algorithm and the Balanced Winnow algorithm. Experimental results have shown that the Naive Bayes classifier performs the best. Approach to analyze the sentiment of short Chinese texts is presented in [9]. By using word2vec tool, sentiment dictionaries from NTU and HowNet are extended. Then the feature weight of the words is enhanced to include the words that appear in the sentiment dictionary and the words next to the sentiment words. The model is implemented using the SVM classifier.

Instead of using only text or only audio, research is also done with combinations of both the domains. A survey on multimodal sentimental analysis and methods are discussed in [10] [11]. The joint use of multiple modalities such as video, audio and text features is explored for the purpose of classifying the polarity of opinions in online videos [12] [13]. Both feature level and decision level fusion methods are used to merge affective information extracted from multiple modalities. They have reported an improvement in classification by grouping different modalities rather than single modality. In [14], the authors introduce the Institute for Creative Technologies Multimodal Movie Opinion (ICT-MMMO) database of personal movie reviews collected from YouTube and ExpoTV. It consists of English clips with sentiment annotation of one to two coders. The feature basis is formed by using audio, video and textual features. Based on the textual movie review corpus, different levels of domain-dependence are considered:

978-1-5090-2597-8/16/$31.00

in-domain analysis and cross-domain analysis. This shows that cross-corpus training works sufficiently well. Authors of [15] introduce MOUD database consists of Spanish videos. They have explored the effect of using different combinations of text, speech and video features on classification. They have also carried out the correlation between visual and acoustic features. This is further confirmed on another set of English videos. From the results it is observed that the joint use of three modalities bring significant improvement. Our work differs from [15] as we consider only two modalities such as audio and text.

In this work, a method to combine both the text and audio features is explored for the sentimental analysis of online videos. As of now, less research is done on multimodal classification of online reviews in Spanish language. Our proposed system is implemented in Spanish database. It is observed that in the literature a lot of work has been done by extracting several features using OPENEAR tool and build a system using the SVM classifier. Instead of taking all the features, only MFCC features (13 dimension) are extracted at stressed significant regions of an audio input and build sentiment analysis system using the GMM classifier. Stressed significant regions are the regions where the strength of excitation fluctuates above and below 30% from the average strength of excitation. Textual features which are computed by Doc2Vec vectors are used to build the SVM classifier in order to detect the sentiment of an audio input.

The rest of the paper is organized as follows: Spanish database used in this paper is discussed in section 2. Sentiment analysis using speech features is discussed in section 3. sentiment analysis using text features is discussed in section 4. Multimodal sentiment analysis and experimental results of proposed method for detecting the sentiment of a Spanish videos is discussed in section 5. Finally, section 6 concludes the paper with a mention on the future scope of the present work.

II. Spanish Database

The database used in this paper is obtained from [15]. Dataset consists of videos collected from YouTube, which include reviews for perfumes, movies and books named as MOUD: Multimodal Opinion Utterances Dataset. As the variety of product reviews are used, the database has some degree of generality within the broad domain of product reviews. Total dataset has 100 videos, among them 36 are negative, 42 are positive and 22 are neutral. Only 30 seconds of opinion segments are taken for each video after removing titles and advertisements. Among them 80% is used for training and 20% is used for testing. For text based sentiment classification, transcription and sentiment annotations were manually performed. Annotators were provided with all the modalities i.e. audio and transcribed text to correctly figure out the opinion of the segment.

III. Sentiment Analysis using speech features

Whole audio input may not contain positive or negative sentiment. Most of the part may be neutral. So, in this work stressed regions are considered by omitting neutral regions. Spectral features (MFCC) which are used in this paper are extracted from stressed significant regions of an audio input. These features are then used to build a classifier of positive or negative sentiment. Each audio input is in .wav format, 16 bit, 16000 Hz sampling rate and a mono channel. The process of finding stressed significant regions from an audio input is described in following subsection.

A. Detecting Stressed Effected Regions

Stressed significant regions are detected based on the strength of excitation of an audio signal using zero frequency filter (ZFF) method [16]. Strength of the excitation is defined as the slope of ZFF signal at epoch. Computing the strength of excitation using ZFF method is discussed in the following subsection.

1) Computation of strength of excitation using ZFF method:

- Time-varying low frequency bias is removed from the signal by pre-emphasizing the audio signal using a difference operation.

$$x(k) = s(k) - s(k-1). \tag{1}$$

- Then, the pre-emphasized audio signal is passed through a cascade of two ideal digital resonators at 0 Hz i.e.,

$$y(k) = \sum_{p=1}^{4} a_p y(k-p) + x(k), \tag{2}$$

where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, and $a_4 = -1$. The above operation can be realized by passing the signal $x(k)$ through a digital filter given by

$$H(z) = \frac{1}{(1-z^{-1})^4}. \tag{3}$$

- From the above step an exponential trend is introduced in the $y(k)$, which will be removed in a following manner.

$$zffsignal = y(k) - \bar{y}(k), \tag{4}$$

Where

$$\bar{y}(k) = \frac{1}{2N+1} \sum_{k=-N}^{N} y(k). \tag{5}$$

Here a window size of $2N+1$ is used to compute the local mean, and typically this is the average pitch period when computed over a long segment of speech.

- The trend removed signal $zffsignal$ is referred as the *zero frequency filtered signal* (ZFF signal). The positive zero crossings of the ZFF signal correspond to epochs or the instants of significant excitation.

- Strength of the excitation is given by

$$SoE = |zffsignal(eh_p + 1) - zffsignal(eh_p - 1)|. \tag{6}$$

Where p=1,2,3...M. Here M is the total number of epochs, eh_p is the location of p^{th} epoch and SoE is the strength of the excitation.

2) Detecting stressed significant regions of an audio signal: Stressed significant regions are computed based on the strength of excitation. Strength of excitation is computed at an epoch and keep that same value till the next epoch. From this, step signal is obtained which is subjected to a 20 ms mean smoothing to generate a smooth contour representing the strength of the excitation curve. From this, the average value of the strength of excitation is computed and refereed as $SoE_{avgvalue}$. The regions where the strength of excitation fluctuates above and below 30% of the $SoE_{avgvalue}$ is considered as stressed significant regions.

An approach for detecting stressed significant regions of an utterance is shown in Fig. 1. Spanish speech utterance "pero igual con las lavadas" is shown in Fig. 1(a). Fig. 1(b) shows the zero frequency filtered signal ($zffsignal$). At every epoch strength of excitation is computed using the algorithm described in Section III-A1 is shown in Fig. 1(c). Strength of the excitation which is computed in the above step is mean smoothed with a frame size of 20 ms is shown in Fig. 1(d). Stressed significant regions of an utterance which are detected are marked with red color is shown in Fig. 1(e).

The stressed significant regions computed is used for developing an Sentiment Analysis system. From Table I it is observed that taking whole input as an input and extracting the MFCC features has a low rate of detecting the sentiment of an audio input compared to stressed significant regions. By taking only stressed significant regions the rate of detecting the sentiment of an audio input is improved by 14 %. In comparison with [15] our system outperforms in detecting the sentiment of a review using audio features.

TABLE I. SENTIMENT CLASSIFICATION PERFORMANCE USING AUDIO FEATURES.

Features	GMM accuracy (%)
Whole input	53.6
Stressed significant regions	67.8

IV. SENTIMENT ANALYSIS USING TEXT FEATURES

This section describes the process of extracting the text features of an audio input. These features are then used to build a classifier of positive or negative sentiment. In a preprocessing step, each audio input is transcribed manually and sentiment annotations are assigned. For each of the audio input 300 dimension feature vector is generated for better results.

Many machine learning algorithms require the input to be represented as a fixed-length feature vector. Doc2Vec model is used for associating documents with labels which is an extension of the existing word2vec model. Doc2vec modifies word2vec to an unsupervised learning of continuous representations for larger blocks of text such as sentences, paragraphs or whole documents means Doc2vec learns to correlate labels and words rather than words with other words. In the word2vec architecture, the two algorithms used are skip-gram and continuous bag of words and in the doc2vec architecture, the corresponding algorithms are distributed bag of words and distributed memory . All transcribed inputs are given to the doc2vec which generates a single vector that represents the meaning of a document. This will be used as input to a supervised machine learning algorithm to associate documents with labels. Sentiment analysis based on text can be viewed as a text classification task which can be handled by SVM due to its better classification. SVM classifier is trained with vectors generated from the doc2vec and by using corresponding sentiment tags (positive/negative). Given a test input, the trained models classify it as either positive or negative.

From Table II it is observed that the rate of detecting the sentiment using text features is 65.5 %. When compared these results with [15] our system outperforms in detecting the sentiment using text features.

TABLE II. SENTIMENT CLASSIFICATION PERFORMANCE USING TEXT FEATURES.

Classifier	Accuracy (%)
SVM	65.5

V. MULTIMODAL SENTIMENT ANALYSIS

The advantage that comes to the analysis of audio is voice modularity when compared to their textual data. Textual data will only have the information regarding the words and their dependencies which may be insufficient to convey the exact sentiment of the input. Instead, audio data contain multiple modalities like linguistic and acoustic streams. Both the modalities are hypothesized based on the highest average probability of the classifiers. From our experiments, it is observed that text data gives less accuracy than the audio, so the simultaneous use of these two modalities might help to create a better sentiment analysis model to detect whether the given test input is positive or negative sentiment.

Table III presents the accuracy of sentiment by combining text and audio features using the proposed method. In audio model, the performance of detecting the sentiment from reviews is improved by 21% and for text model the performance is improved by 1% . By combining both audio and text models, rate of detecting the sentiment of an audio input is 74.7% which is improved by 7% compared to [15] model.

TABLE III. SENTIMENT CLASSIFICATION PERFORMANCE FOR DIFFERENT MODELS ON SPANISH MULTIMODAL OPINION DATASET.

Modality in [15]	Accuracy (%)	**Proposed Modality**	**Accuracy (%)**
Audio (Pitch and intensity)	46.75	Audio (Mfcc)	67.8
Text (Unigrams)	64.94	Text (Doc2vec)	65.5
Audio+Text	67.42	Audio+Text	**74.7**

VI. SUMMARY AND CONCLUSIONS

In this paper, we proposed an approach to extract the sentiment of an audio input using both audio and text information. For audio, MFCC features are extracted at stressed significant regions and for text, features generated using Doc2Vec are used to build the classifiers. Stressed significant regions of a speech signal are detected based on the strength of excitation using ZFF based method. From our experiments, it is observed that extracting MFCC features at stressed significant regions are giving better results compared to the features extracted from the entire audio input. It is also observed that by combining both the modalities rate of detecting the sentiment is further improved.

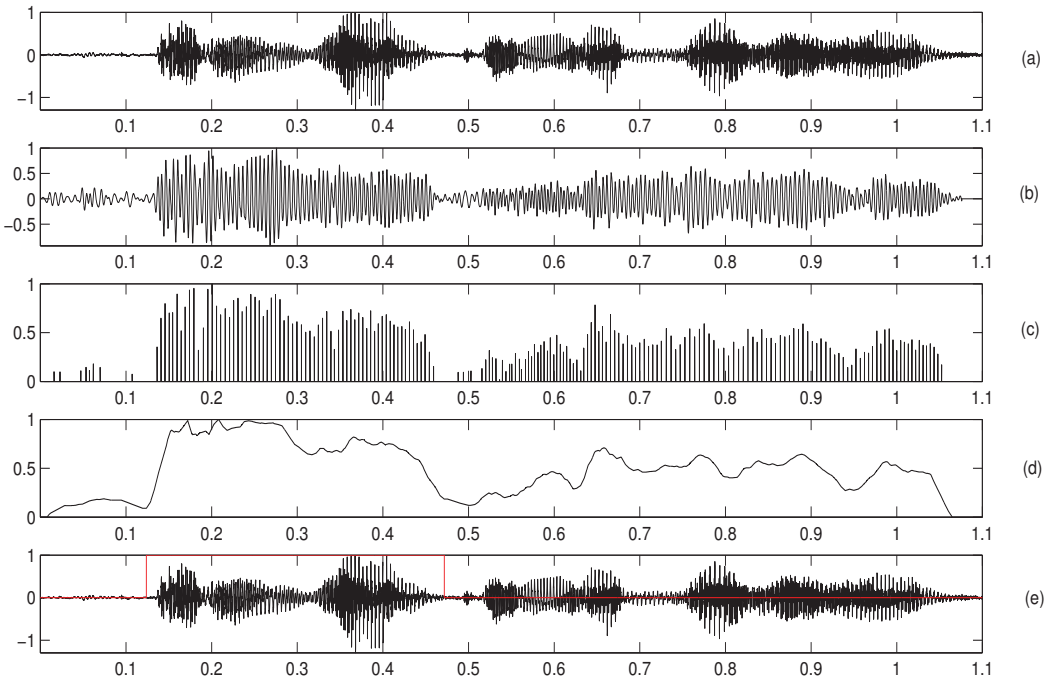

Fig. 1. An approach to detect stressed significant regions of spanish speech utterance "pero igual con las lavadas". (a) Input speech signal, (b) ZFF signal for the input, (c) Strength of the excitation at each epoch, (d) Strength of the excitation which is mean smoothed using a frame size of 20 ms and (e) Stressed significant regions of utterance are detected.

REFERENCES

[1] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Sentiment extraction from natural audio streams," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 8485–8489, 2013.

[2] F. Mairesse, J. Polifroni, and G. D. Fabbrizio, "Can prosody inform sentiment analysis? experiments on short spoken reviews," in *Proc. IEEE International Confernce on Acoustics, Speech and Signal processing (ICASSP)*, pp. 5093–5096, 2012.

[3] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP04)*, vol. 1, pp. 593–596, 2004.

[4] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level," in *Proc. INTERSPEECH*, pp. 1916–1920, 2013.

[5] H. K. Vydana, P. Vikash, T. Vamsi, K. P. Kumar, and A. K. Vuppala, "Detection of emotionally significant regions of speech for emotion recognition," in *Proc. 2015 Annual IEEE India Conference (INDICON)*, pp. 1–6, 2015.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proc. ACL-02 conference on Empirical methods in natural language processing*, vol. 10, pp. 79–86, 2002.

[7] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. Association for Computational Linguistics*, pp. 271–278, 2004.

[8] S. Fong, Y. Zhuang, and J. L. R. Khoury, "Sentiment analysis of online news using mallet," in *Proc. IEEE International Symposium on Computational and Business Intelligence(ISCBI)*, pp. 301–304, 2013.

[9] L. Xing, L. Yuan, W. Qinglin, and L. Yu, "An approach to sentiment analysis of short chinese texts based on svms," in *Proc. 34th Chinese Control Conference*, pp. 28–30, IEEE, july 2015.

[10] S. J. Fulse, R. Sugandhi, and A. Mahajan, "A survey on multimodal sentiment analysis," *International Journal of Engineering Research and Technology (IJERT) ISSN: 2278-0181*, vol. 3, pp. 1233–1238, Nov 2014.

[11] M. Sikandar, "A survey for multimodal sentiment analysis methods," *International Journal of Computer Technology and Applications (IJCTA) ISSN:2229-6093*, vol. 5, pp. 1470–1476, July 2014.

[12] L. P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis:harvesting opinions from the web," in *Proc. 13th International Conference on Multimodal Interfaces (ICMI2011)*, pp. 169–176, November 2011.

[13] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing 174*, pp. 50–59, 2015.

[14] T. K. Martin Wollmer, Felix Weninger and L.-P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intlligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[15] V. Perez-Rosas, R. Mihalcea, and L.-P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intlligent Systems*, vol. 28, no. 3, pp. 38–45, 2013.

[16] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Speech, Audio and language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.