IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Emotion Detection Analysis through Tone of User: A Survey

Ms. Ankita Dutta Chowdhuri[1], Prof. Sachin Bojewar[2]

PG Scholar, Dept of Computer Engg, Alamuri Ratnamala Institute of Engineering and Technology, Mumbai, India[1]

Associate. Prof., Dept of Information Technology, Vidyalankar Institute of Technology, Mumbai University, India[2]

Abstract: Emotion detection through voice is a method that provides information regarding the actual emotion of a person irrespective of the words being used in conversation. Here the emotion is detected through the voice as an input instead of text in conversation. In this paper we focus on various techniques proposed by authors to detect the emotion through user's voice. Estimating the degree of natural emotion through voice corpora by different author's experimental measures is the main aim of this paper.

Keywords: Acoustic Features, Speech Recognition, Support Vector Machines, Anger.

I. INTRODUCTION

Detecting Emotions across different corpora including the combination of corpora having mixed emotions is a major task to be accomplished. As every corpus used in emotion detection is different from one another; many characteristics and parameters are checked to match the training set. The main aim is to detect the emotion of acted corpus instead of acted corpus. Humans can express their emotions in number of ways as there are variations in persons tone and also their culture. If the emotions are expressed in a very unique way than others then the emotion can be detected easily. The variation in tone when recording time is more, brings more complexity for the user to exactly identify the emotion.

II. RELATED WORK

In [1] the author Marie Tahon, Laurence Deville's basically focused at studying differences between acoustic manifestations of anger across corpora which were collected from databases having combined emotions in artificial, modified or realistic context. The authors have further focused on finding measures of realistic emotional corpora. The authors have briefly estimated how a corpus is realistic/naturalistic/induced by utilizing acoustic measures of anger as a source of reference. Firstly they carried out a mixture of measures to estimate the corpus location on the naturalistic/induced scale. First the authors defined three corpora on which they were working. In a second part, they define the acoustic features studied in the paper. The third part consists in the feature analysis and finally they have concluded the possible improvements and outlooks. In [4], the authors describe experiments on the detection of three emotional states (Anger, Positive and Neutral) for two French corpora collected in call centers in different contexts (service complaints and medical emergency). These corpora have a high level of privacy. In order to be comparable with results obtained in the community they used the open EAR acoustic features extraction platform instead of their own library. One of their aims being the comparison of anger and positive emotions across corpora, they basically train models on one corpus and test it on the other to compare their similarities, then conversely. They have further discussed the possible gain in generalization power. In [4] they have performed classification of four basic emotion classes (neutral, sad, happy, and angry) and estimation of emotion primitives using acoustic features. The importance of acoustic features in estimating the emotion primitive values and in classifying emotions into categories was also investigated. An unweighted average recall of 45.5% was obtained for the classification. For emotion dimension estimation, they obtained promising results for activation and dominance dimensions.

III. DATABASES

This section discusses about the data sources used for emotion extraction through voice.

- **Utterances**: A set of recorded voice which can have utterances of many actors both male & female with the combinations of emotional states in it.
- **Corpora of different languages**: Emotion extraction through voice should not be limited by the languages spoken. The languages also have variations in the speech when it's spelled in different languages.
- **Call Center Corpora**: These days many good reviews and complaints are recorded in call centers which have good combination of emotional states. This kind of database having fluctuations of emotions gives a good boost in analyzing the emotions.

IV. ACOUSTIC FEATURES & COMPUTING

Most of the features used by modern automatic speech recognition systems, such as mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive (PLP) coefficients, represent spectral envelope of the

speech signal only. Nevertheless, phase or frequency modulation as represented in recent perceptual models of the peripheral auditory system might also contribute to speech decoding. Furthermore, such features can be complementary to the envelope features. In [1] the authors have proposed a variety of features based on a linear auditory filter bank, the Gammatone filterbank. Envelope features are derived from the envelope of the subband filter outputs. Phase/frequency modulation is represented by the subband instantaneous frequency (IF) and is used explicitly by concatenating envelope-based and IF-based features or is used implicitly by IF-based frequency reassignment. Speech recognition experiments using a standard HMM-based recognizer under both clean training and multi-condition training are conducted on a Chinese mandarin digits corpus. The experimental results show that the proposed envelope and phase based features can improve recognition rates in clean and noisy conditions compared to the reference MFCC-based recognizer. Time-window on which is computed the feature, normalization to the speaker and sound quality require a great attention..

The larger time-window is the voiced segment. According to [4] vowels contain more information (linguistic and paralinguistic) than the rest of the speech signal (consonant, noise, etc.). For this first experiment, we only compute acoustic features on voiced segments. PRAAT gives us fundamental frequency, formants and micro-prosodic features (HNR, VoicedRatio) on voiced segments. With MATLAB, we compute also ZCR, and energy features micro-prosodic features, we compute mean values and variances of each time variables (F0, formants, and energy) on the whole voiced segment. Then a Fast Fourier Transform on the whole voiced signal gives us all spectral and cesptral features. Of course, if there is no voiced part in the segment, it is impossible to compute any feature. We will also not use small segments (duration lower than 50ms).

V. ACOUSTIC CUES, MACHINE LEARNING AND EXPERIMENTAL PROTOCOLS

Authors dealing with the automatic processing of emotions are rooted in specific traditions, e.g. from a general engineering background, from automatic speech recognition, or from basic research (phonetics, psychology, etc.); thus their tools differ as well as the types of features they use. In order to be comparable with results obtained in the community we used the opensource open EAR acoustic features extraction platform instead of their own library. Two different features sets were used: the openEAR Interspeech configuration and the baseline configuration. These features include classical descriptors (F0, ZCR, MFCC, energy...) with various functions applied to them (first and higher order derivatives, min/max, mean and higher order statistics...). We also used classical approach for emotion classification with SVMs using libSVM. We regrouped agent and client segments for these experiments to form balanced classes in every used set. The SVM parameters were optimized using a grid search and cross-validation on the train sets only. The authors wanted to compare the generalization power of each corpus, so they trained models on one corpus and tested on the other conversely using three macro-classes (Anger, Positive and Neutral).The comparisons were made on the basis of the accuracy (unweighted, the classes being already balanced) and class-wise F-score measures, computed from the prediction performed by the classifier on the test sets.

VI. EXISTING TECHNIQUES

In the speech processing community, researchers have worked extensively on identifying emotional characteristics by acoustic parameterization. In [3], the author proposes a framework for recognition of affect in speech through parameters that reflect four main components of speech: intonation, loudness, rhythm and voice quality. The identification of the most appropriate acoustic features for emotional speech classification is presented in [1]. This work suggests a selection method to discover a set of 10 acoustic features that provide best classification. Two new tone-related features are presented in [4]. It uses the K-nearest-neighbor classification method for automatic identification of four basic emotions in human speech. Advanced work exists in the text-analytics community for text categorization. The popular techniques use support vector machines for learning text classifiers from examples [4]. Traditionally, each vector component is assigned a value related to the estimated importance of the word in the document. This is done using the TF-IDF (Term Frequency - Inverse Document Frequency) measure [5]. A comparative study of the feature selection methods in text categorization is provided by [4] and it suggests that information gain is a better method than the commonly used document frequency techniques. Existing work that is most relevant to this paper is present in [5] and [6]. In [5], the authors present a study that explores how people and machines recognize emotions in speech – with application to call-centers. The presented work is able to distinguish between two states agitation and calm. The categorization was used to prioritize voice messages. In [8], the authors focus on identifying emotions from the short utterances that are typical of Interactive Voice Response (IVR) applications. The emphasis was to distinguish anger from speech. However both the techniques focus only on the acoustic parameterization of the speech signal to extract the emotion from call-center type of data.

VII. TWO-STREAM PROCESSING

Spoken language is much more expressive than the written information. So it should be possible to extract certain features that encapsulate the expressiveness of spoken language. Energy and pitch have been believed to be co-related to the emotional status of the speaker. In order to calculate the pitch value from the utterance, we use the sub harmonic-to-harmonic amplitude ratio (SHR). First we locate the position of global maxima (log f1) and then starting from this point, the location of the next local maxima is selected (log f2). The SHR is calculated as:

SHR = 0.5
DA(log f1) − DA(log f2)
DA(log f1) + DA(log f2) (5)

where DA is the difference function that represents the difference of odd and even samples in the log frequency scale. If SHR is less than a certain threshold value, f2 is assigned as the final pitch, else f1 is chosen. Pitch values are calculated for the entire utterance and the following parameters are used as acoustic features:

(a) average pitch over the utterance,
(b) maximum pitch,
(c) minimum pitch and
(d) pitch standard deviation.

These four features are calculated over the first derivative of the pitch contour.
These form the 8 pitch-features. The energy for each frame is also calculated and the following features are extracted for the acoustic features:

(a) average energy,
(b) maximum energy,
(c) minimum energy and
(d) energy standard deviation.

Similar to the pitch values, the first derivative of energy is used to generate 4 more features for energy. Thus the acoustic parameterization consists of the 16 features derived from the pitch and energy of the signal. Since the emotional characteristics of a signal are captured in the contour rather than the signal itself, the first derivatives are able to capture this information. Gaussian mixture models are used to train the 16- dimensional feature space. The output vector va is generated by calculating the likelihood of the input feature vector over the aussians of the different emotion categories.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we presented techniques for emotion recognition that can be used in call-center monitoring and many other databases. The two-stream emotion recognition technique uses the acoustic parameters and the utterance semantics to recognize the emotion category. In our analysis, we showed that the two studied corpora present several differences, in terms of repartitions of fine classes in the Anger and Positive macro-classes, even if we put aside the differences between the annotation schemes. It is possible to have a kind of classification of corpora based on a mix of acoustic features. The results are encouraging and validate the hypothesis that such a joint processing of speech signal is useful. The orthogonal information present in the semantics and the acoustics has been positively exploited by this approach. While we have used the joint processing to recognize emotional characteristic of a call, several interesting applications can be developed with this approach. Complex emotional expressions such as sarcasm can be recognized modeling the mismatch in the acoustic and semantic streams.

REFERENCES

1. Tahon M. and Devillers L., Acoustic measures characterizing anger across corpora collected in artificial or natural context, in Speech Prosody. 2010: Chicago.
2. Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature Appeared in Proceedings of ICECIT-2012 Published by Elsevier.
3. Devillers L., Vidrascu L. and Lamel L., Emotion detection in real-life in spoken dialogs recorded in call center. Journal of Neural Networks, numéro spécial 2005. volume 18.
4. Devillers L., Vaudable C. and Chastagnol C., Real-life emotion-related states detection in call centers, in InterSpeech 2010. 2010: Makhuari, Japan.
5. Devillers L. and Vidrascu L., Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs, in Interspeech 2006. 2006, ISCA: Pittsburg, USA.
6. Polzehl T., Schmitt A., Metze F. and Wagner M., Anger recognition in speech using acoustic and linguistic cues. Speech Communication, 2011. 53(9-10).
7. Gupta P. and Rajput N., Two-Stream Emotion Recognition For Call Center Monitoring, in Interspeech 2007. 2007: Antwerp, Belgium.
8. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification by Stefan Scherer , John Kane, Christer Gobl , Friedhelm Schwenker ELSEVIER June 2005

DOI 10.17148/IJARCCE.2016.55210