Two-Stream Emotion Recognition For Call Center Monitoring

Purnima Gupta

Indian Institute of Technology
Hauz Khas, New Delhi, India
ee1030339@ccsun50.iitd.ernet.in

Nitendra Rajput

IBM India Research Lab
4, Block C, Vasant Kunj, New Delhi, India
rnitendra@in.ibm.com

Abstract

We present a technique for two-stream processing of speech signals for emotion detection. The first stream recognises emotion from acoustic features while the second stream recognises emotion from the semantics of the conversation. A probabilistic measure is derived for each of the individual streams and the emotion category from the two streams is recognised. The output of the two streams is combined to generate a score for a particular emotion category. The confidence level of each stream is used to weigh the scores from the two streams while generating the final score. This technique is extremely significant for call-center data that have some semantics associated with the speech.

The proposed technique is evaluated on the LDC corpus and on the real-word call-center data. Experiments suggest that use of a two-stream process provides better results than the existing techniques of extracting emotion only from acoustic features.

Index Terms: emotion detection, text analytics, speech recognition

1. Introduction

Over the last decade, enterprises have started to use human operated call-centers to provide improved services to their customers. The call-center agents answer customer calls and provide information on different aspects of the services provided by the enterprise that they represent. The evaluation criteria of such call-centers depends on the ability of the agents to satisfy their customer needs in the telephone conversation. Therefore, all call-centers have supervisors who monitor the calls and identify if any agent was not able to satisfy a customer. Since the number of calls received in a typical call-center is very high[1], it is not cost-effective to monitor all calls. So the supervisors monitor a subset of calls and identify if any of them had extreme emotional characteristics (such as happy or angry moods). However the cost involved in human-monitoring of these calls is extremely high. Therefore, automatic monitoring of these calls for recognising the emotional features is a *very important problem from a business perspective.*

There are different ways for humans to express emotions in their speech. The variation is different across people, across geographies and across cultures. If the emotion is expressed very explicitly, then it is easy to interpret and recognise the emotion category for a speech segment. However there is no clear definition or characteristic of expressing a particular emotional style in speech. At certain times, emotion is not completely encapsulated within the speech, but is expressed through other modalities such as gestures and content of speech. Complex emotional expressions such as a sarcasm make the recognition even more difficult. All the above mentioned issues make emotion recognition a *very challenging problem from a research perspective.*

The call-center speech data is a conversation between the agent and the customer about a specific problem. The duration of these calls ranges from 2 minutes to about 30 minutes, with the average being around 5 minutes. However emotion is present at very few (about one or two) utterances in the entire conversation. Identification of the location and the type of emotion expressed is the central problem for this kind of data. In order to recognise the emotional characteristic of the call, we define the following problem to be the focus of this paper:

Given the speech utterance of a call and its corresponding ASR transcribed text, label the emotional characteristics of the call to one of the predefined categories.

In this paper, we present a two-stream technique for emotion recognition of call-center data. The first stream focuses on the speech utterance and extracts the acoustic features from the utterance segments. These features are then used to recognise the the emotion category for the segment. This stream recognises the emotion present in the *rendering* of the. The second stream uses speech to extract the spoken text from the utterance. A speech recognition system is used to generate the ASR (Automatic Speech Recognition) transcribed text for the utterance. Text analysis is then applied to extract the semantics being represented by the speech. Presence of harsh or thankful words in the transcribed text in the *content* provides a clue to the emotion with which this can be rendered in speech. Thus the two streams recognise the emotion in the *rendering* and in the *content* of the conversation. A weighted measure of these two streams is used to generate a combined score. Since the two sources (acoustic-parameters and corresponding text) of emotion detection are orthogonal, we get more information by processing these signals separately.

In Section 2, we briefly describe the previous efforts that have been proposed to solve the problem of emotion recognition. We present an architectural view of the two-stream solution in Section 3. The details of the acoustic stream processing and the semantic stream processing are presented in Section 4. We present the implementation details of the approach, the specifications of the data and the results in Section 5. Finally, the conclusions and discussions are presented in Section 6.

2. Existing Techniques

In the speech processing community, researchers have worked extensively on identifying emotional characteristics by acoustic parameterisation. In [1], the author proposes a framework

[1] a typical medium-sized call-center receives about 100,000 calls per day

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 [2]. 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 [3]. 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 categorisation. 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 categorisation is provided by [6] 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 [7] and [8]. In [7], the authors present a study that explores how people and machines recognise emotions in speech – with application to call-centers. The presented work is able to distinguish between two states *agitation* and *calm.* The categorisation 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 parameterisation of the speech signal to extract the emotion from call-center type of data.

3. Solution Overview

The architecture diagram of the two-stream emotion recognition technique is shown in Figure 1. The input voice sample v_i is passed through two processing streams: (a) Acoustic stream, and (b) Semantic stream. Both the streams get the same input sample and process it to generate two different likelihood vectors. Details of the processing within the two streams are presented in Section 4.

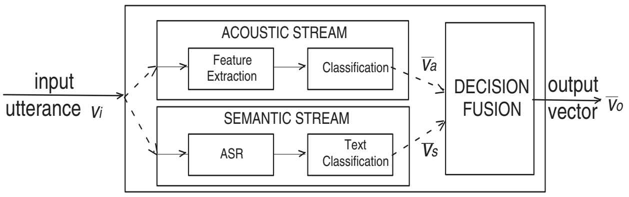

Figure 1: Architecture diagram of the two-stream emotion recognition technique.

At a high level, the *acoustic stream* extracts certain acoustic features based on pitch and energy of the signal. During training, this feature space is labeled with the different emotional categories. Using this space, a likelihood probability map of the extracted features is generated as a output vector $\vec{v_a}$. The size of this vector is defined by the number of emotional categories that were trained in the acoustic feature space. The *semantic stream* performs a speech-to-text conversion of the input speech signal and then the text classification algorithms are applied to classify the text corresponding to the utterance v_i in one of the specified emotional categories. The output of this classification is also probabilistic and is therefore a likelihood probability vector $\vec{v_s}$ that represents the semantic likelihood of the utterance containing the content in a particular emotion.

The output of the two streams is then combined by weighing the two streams based on the confidence of their likelihood. A late-integration decision fusion approach is used for this purpose. For N emotional categories, the output likelihood vector $\vec{v_o}$ is calculated as:

$$v_o^i = w_a * v_a^i + w_s * v_s^i \tag{1}$$

where,
$i = 1, ...N$
$\vec{v_o} = [v_o^1, v_o^2, ..., v_o^N]$
w_a denotes the confidence of the acoustic stream for the input utterance v_i
w_s denotes the confidence of the semantic stream for the input utterance v_i

The values of w_a and w_s are calculated from the distribution of the scores generated in $\vec{v_a}$ and $\vec{v_s}$ respectively. For the N emotion categories, if the distribution of likelihood of the utterance in these categories is well distributed, then this implies that the confidence of the stream in recognising the emotional category is not high. Thus the entropy of $\vec{v_a}$ and $\vec{v_s}$ form good measures to calculate the weights w_a and w_s respectively. Therefore we define the entropy associated with the two streams as:

$$H_a = -\frac{1}{N}\left[\log \prod_{i=1}^{N} v_a^i\right] \tag{2}$$

and

$$H_s = -\frac{1}{N}\left[\log \prod_{i=1}^{N} v_s^i\right] \tag{3}$$

Given the entropy, the weights are calculated as follows:

$$w_a = \frac{H_s}{H_a + H_s} \text{ and } w_s = \frac{H_a}{H_a + H_s} \tag{4}$$

Thus the stream that has a lower entropy, i.e. less confusion, will get a higher weight in decision fusion. This weighing mechanism ensures that if there are any errors in the processing of one particular stream, they are not propagated in the decision fusion. Therefore the robustness of the emotion recognition system is improved with processing of orthogonal information through multiple streams. The emotion category that has the highest probability in the output probability vector $\vec{v_o}$ is selected as the emotion for the input utterance v_i.

4. Two-stream Processing

In this section, we present the techniques that are used for calculating the acoustic and semantic likelihood for the given utterance.

4.1. Acoustic analysis

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 subharmonic-to-harmonic amplitude ratio (SHR). First we locate the position of global maxima $(\log f_1)$ and then starting from this point, the location of the next local maxima is selected ($\log f_2$). The SHR is calculated as:

$$SHR = 0.5\frac{DA(\log f_1) - DA(\log f_2)}{DA(\log f_1) + DA(\log f_2)} \quad (5)$$

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, f_2 is assigned as the final pitch, else f_1 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 parameterisation 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 $\vec{v_a}$ is generated by calculating the likelihood of the input feature vector over the gaussians of the different emotion categories.

4.2. Text analysis

A speech-to-text system is used to convert the entire utterance in text. Additional techniques of speaker-change-detection can be used to label the utterances with the agent or the customer. The distinction between the agent and customer utterances helps in segregating the emotional characteristics of the two speakers. Once the utterance has been converted to text, text analysis is performed to recognise the emotion contained in the text. There could be speech recognition errors while generating the text, however they are not too expensive since the text analysis technique is based on certain keywords and is therefore more robust to non-keywords errors.

Since the utterances are typically very short (one or two sentences), we use a simple technique of TF-IDF to classify the utterance in one of the emotion categories. We create a dictionary of words and phrases that are frequently used in each emotion category. For each emotion category e_i, there is a set of words $w_{ei}^1, w_{ei}^2, w_{ei}^M$ that form the dictionary for that category. The number of words M can be different for each category. This dictionary exists for all the N emotion categories. All words in the text corresponding to the input utterance $w_i^1, w_i^2, ..., w_i^L$ (L is the number of words in the ASR transcribed text) are compared with the dictionary and counts are generated to identify the number of words that exist from each emotion category. The counts are generated as follows:

$$c_{ei} = \sum_{j=1}^{L}\sum_{k=1}^{M} d(w_i^j, w_e^k) \quad (6)$$

where,
$d(w_i^j, w_e^k) = 1$ if $w_i^j = w_e^k$
$= 0$ otherwise

The category that generates the maximum number of counts is the one whose dictionary contains maximum words that appeared in the input text. Thus the probability likelihood $\vec{v_s}$ is generated from these counts as follows:

$$v_s^i = \frac{c_{ei}}{\sum_{j=1}^{N} c_{ej}} \quad (7)$$

4.3. Decision fusion

The likelihood of the two streams are combined to find the joint likelihood for the input utterance. The acoustic stream generates the likelihood every frame whereas the semantic stream generates the likelihood for the entire utterance. Thus the acoustic stream per-frame scores are used to generate a score for the entire utterance. Then the weights are calculated as shown in Equations 2 and 3. These weights are used to generate the final likelihood for the utterance as given by Equation 1.

We have used some basic acoustic features a simple text analytics scheme to demonstrate the usefulness of the two-stream processing technique. Processing with both these streams can be significantly improved by using an enhanced feature set. However this is not considered within the scope of this paper.

5. Implementation and Results

In this section, we describe the implementation details and the two data sets that were used for the experiments. The results obtained from the two data sources are also presented and discussed.

5.1. Implementation

The implementation for extracting the pitch and energy features from the speech signal is done in Java. A 25 msec window with shifts by 10 msec is used to generate the acoustic features from the utterance. We had selected the pitch range to be 50Hz-400Hz. The 16 acoustic features were extracted for each frame. For the semantic stream, we used the IBM ViaVoice based speech recognition system to convert speech to text. The acoustic model was trained on 900 calls constituting about 60 hours of speech. A trigram language model was used to train the system on the call-center utterances. The vocabulary size of the system was more than 10000 words. Text analysis was performed using a simple counts based TF-IDF scheme explained in Section 4.

5.2. Data and experiments

We used two different datasets for the experiments. The first data set is the LDC emotional speech database. This consists of emotion audio and transcripts in 15 different emotional categories that are recorded by professional actors. We used the data from 3 categories (neutral, hot-anger and happy) in our experiments. These are 22 KHz recordings that consist of date and number utterances. Since the LDC corpus does not contain any semantics in the text, it is not possible to verify the two-stream processing technique with this data. So we used real-world data from a call-center in the second experiment. This data-set contains 20 calls, which have one of the three emotions: *anger*, *happy* or *neutral*. These calls are 8 KHz recordings that contain the entire conversation between the customer and the call-center agent. The context of the conversation has enough semantics to perform text analysis on the data.

In the first experiment, 80% of the LDC emotional speech database was used to train the system in three emotional categories (neutral, hot-anger and happy). The rest of the data was used as test data to evaluate the goodness of the *acoustic stream*.

The second experiment consisted of analysing the call-center data in the two streams. Utterances within the call that had some emotional characteristics were manually extracted and 80% of these were used as the training data. The remaining 20% of the utterances were used to test the two-stream processing technique.

5.3. Results and discussions

Table 1 shows the results of the *acoustic stream* processing for the LDC database. As seen, the emotion recognition technique has a high accuracy for the anger and happy emotions. The accuracy is lower for the neutral emotion since it gets classified as anger or happiness in some cases. The accuracies are significantly high since this database has been recorded by professional actors and has explicit emotions. The pitch and energy contours of a couple of sample utterances (Figure 2) clearly suggest that these are distinguishing features for emotion extraction. *These results suggest that the selected acoustic features are able to capture the acoustic part of the emotion.*

Table 1: Acoustic stream processing of LDC database

Emotion	Recognition Accuracy
Happy	93.3%
Anger	80%
Neutral	66.7%

Experiments with the call-center data give more insights into the two-stream processing technique (Table 2). A simple *acoustic stream* processing gives only about 50% to 57% accuracy. However when the scores from *semantic stream* are combined with the acoustic stream, the accuracy improves to about 80%. *This clearly illustrates that the two-stream processing of signals is extremely useful for emotion detection in call-center data.*

Table 2: Two-stream processing of call-center data

Emotion	Acoustics	Semantics	Two-Stream
Happy	57.1%	71.4%	71.4%
Anger	60%	60%	80%
Neutral	50%	66.7%	83.3%

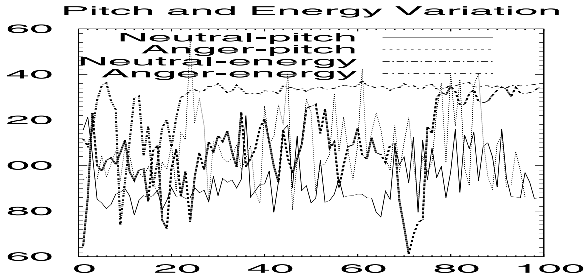

Figure 2: Pitch and Energy variations in two sample utterances of *anger* and *neutral*.

The reason for low *acoustic stream* accuracies for the call-center data is that the stress on different emotions is not very significant when normal people speak long sentences. Thus it becomes critical to use additional orthogonal information derived from the semantics of the data. The semantics contains phrases such as *It was a PLEASURE talking to you*, *THANKS for the information*, *This is DISGUSTING* or *This information is totally USELESS for me*. The keywords such as PLEASURE, THANKS, DISGUSTING, USELESS provide a clue to the emotion with which they will be rendered in speech. This helps in improving the accuracy of the emotion recognition technique.

6. Conclusion and Discussion

In this paper, we presented a technique for emotion recognition that can be used in call-center monitoring. The two-stream emotion recognition technique uses the *acoustic parameters* and the *utterance semantics* to recognise the emotion category. Basic features (pitch and energy) are used to derive the acoustic parameters while TF-IDF is used to perform the text analysis. We validated the acoustic processing by a standard LDC corpus while we used the real-world data from a call-center to validate the two-stream processing technique. 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 recognise emotional characteristis of a call, several interesting applications can be developed with this approach. The categorisation of calls into positive and negative sentiments [9] can also be used to gain several business insights from an enterprise perspective. Complex emotional expressions such as a sarcasm can be recognised by modeling the mismatch in the acoustic and semantic streams.

7. References

[1] Fernandez, R., “A Computational Model for the Automatic Recognition of Affect in Speech,” PhD Thesis, MIT Media Arts and Sciences, 2004.

[2] Ververidis, D., Kotropoulos, C. and Pitas, I., “Automatic emotional speech classification,” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2004.

[3] Yindong Yu, Eric Chang and Cong Li, “Computer Recognition of Emotion in Speech,” The 2002 Intel International Science and Engineering Fair, 2002.

[4] Joachims, T., “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proceedings of European Conference on Machine Learning, 1997.

[5] Yang, Y. and Liu, X., “A re-examination of text categorization methods,” SIGIR, 1999.

[6] Yang, Y. and Pedersen, J. O., “A Comparative Study on Feature Selection in Text Categorization,” Proceedings of International Conference on Machine Learning, 1997.

[7] Petrushin, V. A., “Emotion in Speech: Recognition and Application to Call Centers,” Proceedings of ANNIE, 1999.

[8] Yacoub, S., Simske, S., Lin, X., and Burns, J., “Recognition of Emotions in Interactive Voice Response Systems,” Interspeech, 2003.

[9] Nasukawa, T. and Yi, J., “Sentiment analysis: capturing favorability using natural language processing,” International Conference On Knowledge Capture, 2003.