Speech-based emotion detection in a resource-scarce environment

Olga Martirosian, Etienne Barnard

Human Language Technologies Research Group, Meraka Institute, Pretoria, South Africa

omartirosian@csir.co.za, ebarnard@csir.co.za

Abstract

We explore the construction of a system to classify the dominant emotion in spoken utterances, in a environment where resources such as labelled utterances are scarce. The research addresses two issues relevant to detecting emotion in speech: (a) compensating for the lack of resources and (b) finding features of speech which best characterise emotional expression in the cultural environment being studied (South African telephone speech). Emotional speech was divided into three classes: active, neutral and passive emotion. An emotional speech corpus was created by naive annotators using recordings of telephone speech from a customer service call centre. Features were extracted from the emotional speech samples and the most suitable features selected by sequential forward selection (SFS). A consistency check was performed to compensate for the lack of experienced annotators and emotional speech samples. The classification accuracy achieved is 76.9%, with a 95% classification accuracy for active emotion.

Index Terms: emotion recognition, resource creation, cultural factors

1. Introduction

One of the ways in which companies provide their customers with services is through call centres. To perform quality control over these call centres, people are employed to listen to samples of telephone calls from the call centre and isolate problematic ones. Finding a way to isolate problematic telephone conversations automatically can increase the efficiency of the call centre and allow a company to dedicate more resources to addressing the problem areas found through these telephone conversations. The design of systems for this purpose has therefore been a topic of active research [1, 2].

The expression of emotion in speech will, to a greater or lesser degree, depend on the cultural environment in which the speech occurs. However, this topic has received little experimental attention – in large part because of the lack of publicly available speech corpora containing speech labelled according to emotional content. We have therefore embarked on a study to find whether such a labelled corpus can be created efficiently in a environment where limited resources are available. In the process, we researched the features of speech which can be used to characterise the emotional content of spoken utterances in a particular resource-scarce environment, namely South African telephone-recorded speech.

In line with the call-centre application mentioned above, the goal was to find features which can be used to classify emotions into groups which would be helpful when problematic telephone conversations need to be isolated. This study investigates the use of several global statistical features of speech utterances before selecting an optimal set which achieves the highest classification accuracy between the emotional classes. Emotions were sorted into three classes: active emotion, passive emotion and neutral. Active emotion encompasses emotions which are expressed with energy such as anger, happiness and frustration; passive emotion encompasses sadness and disappointment, and neutral encompasses speech with a negligible amount of emotional content.

Because a study on the expression of emotion in speech has not been done in the South African culture, emotional speech data was not available. A corpus was collected from recordings of telephone calls to a customer service centre, and these recordings were labelled by two naive annotators. An algorithm was designed to act as a consistency measure, which acts as partial compensation for the lack of more sophisticated approaches to annotation which are typically employed in more resource-rich environments [1]. This algorithm acts as a filter, selecting the emotional speech samples that are consistent with most other samples in their emotional class and discarding those samples that are not.

In Section 2 we summarise some of the existing research on emotion recognition, including annotation standards. Section 3 describes the speech corpus employed, and Section 4 describes the methodology employed. Our results are summarised in Section 5.

2. Background

The design of a system that is capable of detecting emotion in speech can be divided into four sub-tasks: how emotional speech is collected from people, the annotation of an emotional speech corpus, the features used to discriminate between emotions and the classifier used for the classification of emotional speech.

This study was centred on performing detection of natural emotion in speech. Natural emotion in speech occurs when a person expresses an emotion through speech without the emotion being elicited or acted. Acted and elicited emotions can be recorded in isolation and with the emotions selected in advance. Acted emotional speech is collected by recording actors speaking with emotions. Acted speech has high emotive content and arousal [2]. However, systems which are trained on acted speech do generally not perform well in real world situations [1, 2]. Elicited emotional speech is collected by using a mood induction procedure called the Velten method [3]. People are asked to read a group of sentences which grow in emotional content, thereby inducing the correct mood in the reader. Natural emotions can only be found in spontaneous speech and thus cannot be recorded in a targeted fashion – rather, sound samples have to be extracted from existing conversations and the emotional label of that utterance deduced from the content [1].

Given the lack of resources available in our developing-world context, we used naive annotators for the collection and annotation of an emotional speech corpus. If following emo-

tion annotation standards [4] in order to keep labelling consistent between different annotators and the same annotator at different times, much time and effort is required from annotators. This study explores the possibility of compensating for lack of experienced and trained annotators and annotation techniques through the use of a consistency checking algorithm. Two naive annotators were used in this study, with no manual consistency checking. However, once the corpus was populated, an automatic consistency check was executed on the data to discard inconsistent samples.

The features used in order to detect emotion in speech are a topic of active research. One must select the features that discriminate the most between different emotional classes. Several studies have found that features derived from the pitch contour are useful in this regard. The pitch contour is induced by the glottal waveform, which depends on the tension of the vocal folds and the sub-glottal air pressure [2]. The glottal volume velocity changes as different emotions are expressed in speech. It has been found to be one of the most reliable features to be used for emotion recognition [5]. The formants of a speech signal are a way of quantitatively describing the vocal tract. They can be used to differentiate well articulated speech from loosely articulated speech. A person under stress or depression does not articulate voiced sounds as well as they do when they are not experiencing emotion [2]. It has been shown that the first and second formants are affected by emotional states more than all the other formants. Other features that have been found useful in discriminating between emotions include features derived from the energy contour of an utterance, since the energy (or amplitude) pattern also conveys the arousal level of emotions in speech [2, 5].

The correct features for emotion detection will play a crucial role in obtaining satisfactory classification accuracy, and we suspect that the appropriate features depend on the environment (linguistic and cultural) in which the speech is collected. Some features may be detrimental to the emotion detection process. Sequential forward selection (SFS) can be used to determine the most suitable features to be used in classification [5]. This method is nevertheless not always needed. Bhatti, Wang and Guan [5] found that, with 17 features, SFS gives an improvement of only 3% in their recognition rate. However, Ververidis and Kotropoulos [6] used SFS to select ten out of eighty seven features to classify the emotional speech best.

The classifier used to classify emotions in this study was a neural network. Studies [7, 5] have found neural networks to be practical for detecting emotion in speech. This is to be expected as neural networks are known to have a good ability to classify data which is not linearly separable.

3. The speech corpus

The speech corpus developed in this research was extracted from recordings of telephone conversations from a call centre that provides customer service for a telecommunications company. All the recordings analysed were transmitted via the GSM network. The corpus contains both male and female voices, in approximately equal quantities. In order to limit the variability of the corpus, only English utterances were employed, but these were spoken by speakers from a range of linguistic backgrounds. (The first languages of the speakers included isiZulu, isiXhosa, Sesotho, English, Afrikaans and Hindi.) The recordings were segmented into samples between 1 and 20 seconds long and the segments labelled with the dominant emotion of the speech contained in them.

The fine emotional labels used were angry, frustrated, happy, friendly, neutral, sad and depressed. These fine labels were combined into three broad classes: active emotion, neutral and passive emotion. Angry, frustrated, happy and friendly were categorised into the active emotion class; neutral was put into the neutral class and sad and depressed were categorised into the passive emotion class.

The final corpus consisted of 2745 samples, 620 of which are active emotion samples, 2022 are neutral samples and 103 are passive emotion samples.

4. Experimental methodology

4.1. Feature set

Based on the prior research summarised in Section 2, global statistical features were selected to be used as the features for classification. All features were extracted using the Praat program for speech analysis [8]. In total, 45 features were examined and are listed below.

- Pitch and intensity mean, maximum, minimum and standard deviation;
- First, second and third formants mean, maximum, minimum and standard deviation;
- Highest and second highest pitch and intensity standard deviation, measured across 300 ms intervals;
- Mean pitch and intensity standard deviation, measured for 300 ms intervals;
- Highest and second highest pitch and intensity gradient and acceleration, measured across 5 ms intervals;
- Mean pitch and intensity gradient, measured across 5 ms intervals;
- Number of peaks of the intensity curve, normalised for number of 5 ms intervals;
- Highest and second highest peak to peak value of the pitch and intensity curve; And
- overall range of the pitch, intensity and formant contours.

Once the feature extraction was completed, the features extracted directly from the pitch, intensity and formant contours of the sample were normalised with regard to their mean and standard deviation using z-score normalisation [9].

Histograms were plotted to estimate the discriminative power of the features between the tiers. One example of these histograms is shown in Figure 1, for a feature derived from the instantaneous intensity contour. The discrimination that the feature provides between the classes can be seen by the separation of the different colours in the histogram – as would be expected, the passive emotions tend to have the lowest relative intensity peaks, whereas the active emotions tend to have the highest values for this feature.

4.2. Classifier

Since the focus of this study was resource collection and feature selection, we did not attempt to find the very best classifier for our data. Instead, a competent non-parametric classifier (a feed forward multilayer perceptron (MLP), using the Levenberg-Marquardt algorithm for training) was used [10]. Initial experiments were run to ascertain the optimal number of hidden nodes. Numbers of hidden nodes between 1 and 10

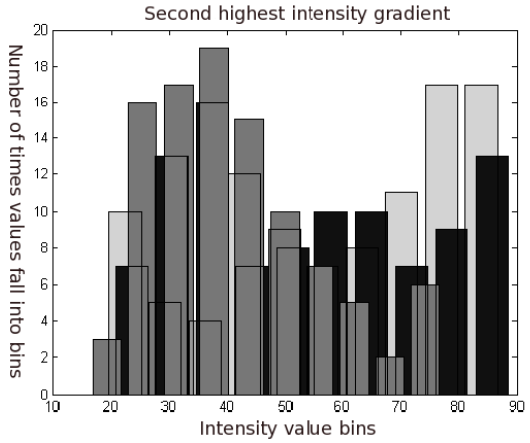

Figure 1: *Histogram of second highest intensity for three tiers (normalised to 100 levels). The darkest colour represents active emotion, the lighter colour represents passive emotion and the lightest colour represents neutral*

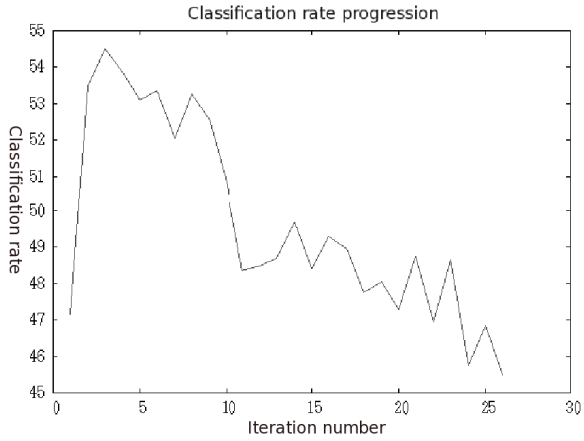

Figure 2: *Relationship between classification accuracy and number of features used for classifying, with features selected according to sequential forward selection*

were used, and the optimal number of 5 was selected taking both time constraints and classification accuracy into account. The neural network was also implemented with 3 output nodes, one for every emotion class. The classification is checked by polling the outputs and checking which output has the highest value. The classification is then checked against the class which the sample falls into to calculate a classification accuracy.

4.3. Feature selection

Sequential forward selection (SFS) was used to select optimal features out of the feature set. SFS starts off with a pool of features and an empty (selected) feature set. A feature is selected to be added to the feature set from the pool by determining which feature will cause the highest reduction in classification error (which, when starting off, will be the error obtained when speech utterances are classified arbitrarily into emotional tiers). Features are added to the set using the same selection technique until none of the features allow for a decrease in classification error. Features are not allowed to be removed from the selected feature set. Figure 2 illustrates the classification accuracy achieved if the SFS algorithm is allowed to continue after the optimal feature set has been found. One can see from the figure that the classification accuracy starts to decrease after 3 features are selected. This is not only because the features become less discriminatory together, it also indicates that 3 is the optimal amount of features for the selected classifier. The SFS algorithm implemented in the emotion detection system would stop searching for features once it gets to the peak classification accuracy at 3 features.

4.4. Data refinement

Once feature selection was completed, the data was checked for consistency. Consistency checking was implemented to compensate for the lack of annotator expertise as well as annotator disagreement, and thus the mislabelling and inconsistent labelling of samples. The consistency checking begins with the passive emotion class because this class contains the smallest number of samples and will thus dictate the number of samples allowed from the other two classes (The same number of samples are used from every class to ensure that the classifier stays unbiased).

Consistency checking makes use of a neural network trained with all samples of the training set. The neural network is then tested with the same samples. The samples which are classified incorrectly by more than a specified margin are discarded and all others (including those classified correctly) are retained. Consistency checking is first executed on the class with the smallest amount of samples, once the number of consistent samples from that class is known, the same amount of consistent samples are extracted from the other two classes to make the new consistent data set. The retained samples are used as class representatives for the training of a new neural network.

5. Experimental results

Two experiments were performed on the emotional speech corpus collected. The first experiment was aimed at getting a base classification accuracy when consistency checking is not implemented. Firstly, because only 100 samples of passive emotional speech were present in the emotional speech corpus, 100 samples of active emotional speech and 100 samples of neutral speech were used in order to not bias the classifier. The samples were selected such that the first 50%, which were used for training data, were selected from a different recording set than the second 50%, which were used as testing and validation data. This would ensure that the classification accuracy achieved would not be over optimistic. Then all 45 features were extracted from these samples. The sequential forward selection algorithm was performed to find the optimal features for discriminating between emotional classes. Three features were selected:

- highest time difference between the peaks of the intensity and pitch contours,
- highest amplitude difference between successive peaks of the intensity contour.

These features, extracted from the training set, were used to train the baseline classifier. The classification accuracy achieved for this experiment was 57.6%. The confusion matrix can be seen in Table 1 and illustrates the classifications of the three classes individually. From the confusion matrix one can

see that active emotion has the highest classification accuracy at 72%.

Because the passive emotion has the least amount of samples, another experiment was run using only active and neutral emotion samples. This allowed the classifier to be trained on 600 samples of each tier. The features selected during this experiment were:

- second highest acceleration measured for the pitch contour,
- mean of the first formant, and
- overall range for the third formant.

The overall classification accuracy achieved for the system was 79.28%.

In the second experiment, all 45 features were extracted from every sample in the corpus. Feature selection was then performed using the sequential forward selection to find the optimal features to use for classification. Once the optimal features were found the samples were filtered using these features to maximise consistency within the classes. The selection margin was set such that (on average) 70% of the samples in each class are retained. These samples were then reduced to ensure equal representation of all classes, and split into a training set ($\frac{1}{2}$), test set ($\frac{1}{3}$) and validation set ($\frac{1}{6}$) and input to the classifier. The average classification accuracy achieved using this method was 76.9%. The confusion matrix, which can be seen in Table 2 illustrates the classifications of the three classes individually. As can be seen in Table 2, Active and neutral emotions are classified with a much better accuracy than the passive emotions. In fact, both the active and neutral classes have a recall rate above 90%, and for active emotions the precision is also better than 90%. Passive emotions are poorly recognised because they are difficult for an annotator to recognise; thus, many of the samples in the passive emotion class are quite similar to those with neutral emotions. Even with consistency checking, there are simply not enough samples with pronounced passive emotions for the classifier to learn to distinguish them from other emotions with high accuracy. In order to ensure that the increase in classification accuracy was indeed due to the improved training of the classifier, the samples that were rejected during filtering were re-classified using the classifier trained on the filtered samples. The average classification accuracy achieved was 61.4%, which is 3.8% higher than the original classification accuracy for non-filtered samples.

Table 1: Confusion matrix of results achieved without consistency checking

True Class	Classified		
	Active emotion	Neutral	Passive emotion
Active emotion	**0.7291**	0.1521	0.1188
Neutral	0.2667	**0.4727**	0.2606
Passive emotion	0.3006	0.2685	**0.4612**

6. Discussion

This study has been centred on the development of a system to detect the expression of emotion in speech in an environment with limited resources. Two matters have received most attention: compensating for lack of resources and finding the features which characterise the expression of emotion.

Table 2: Confusion matrix of results achieved with consistency checking

True Class	Classified		
	Active emotion	Neutral	Passive emotion
Active emotion	**0.9565**	0.04348	0
Neutral	0.087	**0.913**	0
Passive emotion	0	0.3043	**0.6956**

Compensating for lack of resources is very important where speech resources are not freely available. In the context of this study, only a certain number of recordings of emotional speech were available, and these were labelled by naive annotators. The consistency checking has allowed the improvement of both the training and testing process by ensuring samples that are labelled with an emotional class truly contain the emotion that they are labelled with. This means that the classifier is trained to classify emotions with samples that contain less emotional 'noise'. And it is tested on samples that contain the emotions that they are labelled with, thus improving the preciseness of the measurement of classification accuracy. By using the consistency checking algorithm to automatically check the emotional speech samples, a data set was selected which achieves a classification accuracy of 76.9% - this is a 19.3% improvement on the original data set. This improvement came at the cost of discarding the samples which were deemed less distinct by the original classifier; however, our informal listening experiments indicated that these were indeed samples of which the true classes were most debatable. Also, the filtering of samples improved the classifier training to the extent that the samples rejected during consistency checking were classified with a rate of 61.4%, thus showing that the improved system is more capable of classifying even borderline, unseen utterances.

It remains to be seen how well this classifier will perform on an unfiltered set of utterances – based on the experimental results obtained here, it seems feasible that a threshold on the outputs of the neural network can be used to eliminate the samples which do not clearly fall into a particular category. However, even in the absence of such a mechanism, the current system should be able to distinguish between active and neutral utterances with a high precision, which would be useful in practical applications.

It will also be interesting to see how well the methodology followed here will perform for emotion classification in other environments, and for other paralinguistic classification tasks (e.g. dialect classification). Finally, it would be most interesting to compare the optimal features for emotion detection in different languages, and an approach such as ours makes such a comparison feasible.

7. References

[1] L. Devilliers and L. Vidrascu, "Annotation and detection of blended emotions in real human-human dialogs reocrded in a call center," in *Proceedings of IEEE ICME*, Orsay, France, July 2005.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[3] E. Krahmer J. Wilting and M. Swerts, "Real vs. acted emotional speech," in *Proceedings of Interspeech*, Pittsburgh, PA, USA, September 2006.

[4] L. Devilliers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cies on human-human call center dialogs," in *Proceedings of Interspeech*, Pittsburgh, PA, USA, September 2006.

[5] Y. Wang M.W. Bhatti and L. Guan, "A neural network approach for human emotion recognition in speech," in *Proceedings of IEEE International Symposium on Circuits and Systems*, Vancouver, British Columbia, Canada, May 2004, vol. 2, pp. 181–184.

[6] D. Ververidis and C. Kotropoulos, "Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm," in *Proceedings of IEEE ICME*, July 2005, pp. 1500–1503.

[7] V.A. Petrushin, "Emotion recognition in the speech signal: Experimental study, development and application," in *Proceedings of Interspeech*, Beijing, China, 2000.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by commputer 4.5.01 [computer program]," Retrieved October 2, 2006, from http://www.praat.org/.

[9] F.N. Julia and K.M. Iftekharuddin, "Detection of emotional expressions in speech," in *Proceedings of IEEE SoutheastCon*, March 2006, pp. 307–312.

[10] M. Schmid, "Octave's neural network toolbox 0.0.2-23," Retrieved July 2006 from http://octnnettb.sourceforge.net/.