Affect Recognition from Speech

Li ZHANG[1] and Virginia FRANCISCO
School of Computing
University of Teesside, UK
TS1 3BA

Abstract. We aim to provide an automatic anti-bullying component in online text and speech based interaction for young people age 18 – 25. Affect expression in speech generally differs from culture to culture, from female to male. In this study, we focus on affect sensing from speech for different gender user groups for young people. So far our work mainly concentrates on the sensing of five basic emotions (including 'happiness', 'sadness', 'fear', 'surprise', and 'anger') and 'neutral' from speech. Detailed acoustic features have been extracted after analysis of speech data from one chosen male and female speaker. Our affect sensing component has been implemented under the theory of naïve Bayes classifier. We have also evaluated it using new test data. Our work contributes to the conference themes on intelligent technologies – machine learning and affective speech processing.

Keywords. Affect sensing, Speech processing, Naïve Bayes classifier

Introduction

Automatic affect interpretation from speech could be challenging. Especially, affect expression in speech generally differs from one to another. In this study, we mainly focus on affect sensing from speech using acoustic features for different gender groups during online interaction. In our previous work, we have made an in-depth development on affect sensing from open-ended text and metaphorical language. Our final research aim is to equip our system with affect sensing abilities from both speech and text in order to provide an automatic anti-bullying function for online interaction for young people. The work also shows great potential for 'virtual tutors' development in training/learning environments via the automatic interpretation of users' emotion.

1. Related work

There is much work in the area of emotion recognition in speech. Nogueiras et al. [1] have used Hidden Markov Models to recognise emotion from speech. Their study proved that the structure of HMM was useful to capture the temporal behaviour of speech. Grimm et al. [2] have used acoustic features from speech signal and mapped them to an emotion state in a multi-dimensional, continuous-valued emotion space to recognize driver's emotional state while driving. Amir and Cohen [3] have also attempted to characterise emotion in the soundtrack of an animated film. Cichosz and

[1] Corresponding Author.

Ślot [4] reported a symbol-based learning approach to classify emotion in speech. They used a binary decision tree based classifier and evaluated their approach using two databases of emotional speech on German and Polish. Oudeyer [5] made attempts to detect emotion from speech using a genetic algorithm with a set of optimal features.

2. Emotion recognition in speech

In our work, we create our own speech database and adopt naïve Bayes classifier to conduct the recognition task. First of all, we have created 10 neutral and 10 emotional informal conversational sentences for each emotional category. In order to justify the acoustic features discovered for different emotional speech, we have made neutral sentences not only recorded in a neutral way, but also recorded in the other five emotional ways. Similarly, we have also recorded all affective example sentences in a neutral way so that such examples could assist us to remove some of the recovered features from emotional speech data mainly caused by particular speech context.

Thus we have recorded 160 utterances (100 emotional utterances – 20 for each category and 60 neutral utterances – 10 for each category) for each speaker using a standard sound studio. Our emotional speech database has been constructed with the speech samples contributed by 10 speakers mainly age 18 – 27 (4 female and 6 male), most of which have northeast British accent. All these data are used as training data. The test data are recorded in the same fashion, but with totally different emotional and non-emotional informal textual sentences. For each speaker, we have recorded 60 speech samples for testing (10 for each category including neutral).

Praat [6] has been used to analyze the speech data. For each sample, Praat provides an automatic summarized voice report and other files containing detailed information on acoustic features such as pitch, pulses, voicing, jitter, shimmer, intensity etc. After a careful study of the voice reports and other acoustic features of the emotional and neutral speech samples from all the speakers, we have chosen 9 acoustic features (mean and median pitch, standard pitch deviation, minimum and maximum pitch, pulses-per-second, median intensity, minimum and maximum intensity) for further analysis. From the analysis of the chosen speech samples from a male and a female speaker, we extended the 9 features to 45 for the male group and 52 for the female group. We now report our approach to emotional speech recognition.

In our application, naïve Bayes classifier has been used to recognize emotion from speech. Equation 1 has been used to calculate the probabilities of different emotional states for any given test speech sample. The emotional state with the highest probability is regarded as the most probable affective state implied in that instance.

$$V_{max} = \text{argmax}_{vj \text{ in } V} P(v_j) * P(a_1|v_j) * P(a_2|v_j) * ... * P(a_n|v_j) \quad (1)$$

In equation 1, a_1, a_2, … and a_n represent the acoustic features recovered for each speech sample, such as features for mean pitch, median pitch, median intensity etc. We assume that these 9 general features are all independent. Each training speech sample is represented by the set of 9 acoustic features with different values. M-estimate has been adopted to produce the probability of an attribute value given any emotional or neutral classification. A Java application has been implemented based on the above discussion to recognize emotion from speech. In detail, we use 110 training sample data (20 for

each emotional category and 10 neutral speeches) with their corresponding attribute values and the emotion labels indicating which emotional states they imply, as the input to the classifier for each gender group. The 120 test speech samples recorded by the same two speakers have also been represented in such a format, but with totally different sets of values of the acoustic features. Table 1 presents the recognition results.

Table 1. Affect recognition results for the male and female speakers

Male	Neutral	Sad	Happy	Angry	Fear	Surprise
Neutral	**90%**	-	-	10%	-	-
Sad	30%	**10%**	-	-	50%	10%
Happy	-	-	**20%**	50%	-	30%
Angry	-	-	-	**60%**	10%	30%
Fear	-	-	-	20%	**50%**	30%
Surprise	-	-	-	-	-	**100%**

Female	Neutral	Sad	Happy	Angry	Fear	Surprise
Neutral	**90%**	10%	-	-	-	-
Sad	20%	**60%**	20%	-	-	-
Happy	-	-	**60%**	20%	20%	-
Angry	-	-	30%	**20%**	40%	10%
Fear	-	-	30%	60%	**10%**	-
Surprise	-	10%	10%	-	40%	**40%**

A further analysis of the results indicated that for both male and female speakers, an emotional utterance labeled with one negative affective state tends to be recognized to contain another negative affect implication because of the resemblance of the acoustic features in these two emotional categories. E.g. speech samples with 'anger' implication have been interpreted to contain the affective state, 'fear'. Similarly, speech data with 'happiness' implication have been regarded to contain '(good) surprise' taste. These results indicate that our affect sensing component may have extracted some underlying generalization in the recognition of the general positive and negative affective states from the training data, but further improvement is needed in order to effectively distinguish one positive/negative affective state from another. We also aim to extend the speech sample size for training and evaluation as another future direction.

3. Conclusions

We implemented a prototype for affect sensing from speech using Bayes classifier. Although there is room for improvements, the current performance is acceptable. We intend to integrate this component with an intelligent conversational agent who interacts with human users during online speech based interaction so that it would be capable of detecting bullying situation automatically. We are also interested in how it intervenes in such aggressive situations based on the interpretation of the prototype.

References

[1] A. Nogueiras, A. Moreno, A. Bonafante, and J. Maririo. Speech Emotion Recognition using Hidden Markov Models, *Eurospeech 2001*, 2679-2682, 2001.

[2] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, and T. Moosmayr. On the Necessity and Feasibility of Detecting a Driver's Emotional State While Driving. In *Proceedings of ACII 2007*, Lisbon, ACM, Springer, 126-138, 2007.

[3] N. Amir and R. Cohen. Characterizing Emotion in the Soundtrack of an Animated Film: Credible or Incredible? In *Proceedings of ACII 2007*, Lisbon, ACM, Springer, 148-158, 2007.

[4] J. Cichosz and K. Slot. Emotion recognition in speech signal using emotion-extracting binary decision trees. Doctoral Consortium. *ACII 2007*, Lisbon, ACM, Springer, 2007.

[5] P.Y. Oudeyer. The production and recognition of emotions in speech: features and algorithms, *International Journal in Human-Computer Studies*, vol. 59/1-2, 157-183, Special Issue on Affective Computing. 2003.

[6] Praat, a speech processing tool. http://www.personal.rdg.ac.uk/~llsroach/phon2/freespeech.htm, 2008.