A Novel Feature for Emotion Recognition in Voice Based Applications

Hari Krishna Maganti, Stefan Scherer, and Günther Palm

Institute of Neural Information Processing, University of Ulm, Germany

In the context of affective computing, a significant trend in multi modal human-computer interaction is focused to determine emotional status of the users. For a constructive and natural human-computer interaction, the computers should be able to adapt to the user's emotional state and respond appropriately. This work proposes few simple and robust features in the framework of determining emotions from speech. Our approach is suitable for voice based applications, such as call centers or interactive voice systems, which are dependent on telephone conversations. For a typical call center application, it is crucial to recognize and classify agitation (anger, happiness, fear, and disgust) and calm (neutral, sadness, and boredom) callers, for the systems to respond appropriately. For instance, in a typical voice based application, the system should be able to either apologize or appreciate the problem of the caller suitably, if necessary by directing the call to the supervisor concerned.

In [1], the basic emotions were grouped into two categories and features based on fundamental frequency, energy, speaking rate, first three formants, and their bandwidths, along with their vital statistics were used. However, commonly used features such as pitch, energy, and statistics of these may not suffice to extract the relevant information needed to classify emotions accurately [2]. Therefore in this work, a simple feature extraction approach is proposed, which is based on the long term modulation spectrum of speech. The performance of the proposed approach is comparatively accurate, robust and close to real-time, potentially irrespective of the speaker, gender, and speech acquisition channel.

In the first step, the Fast Fourier Transform (FFT) of the input speech signal is computed. Then, the Mel-scale transformation, which imitates the human auditory system, is applied to these vectors. In the second step, the modulations of the signal for each band are computed by taking the FFT, resulting in a sequence of modulation vectors. It is observed that most of the prominent energies are within the frequencies between 2 to 16 Hz. After the computation of the modulation spectrum energy for each band, the median values of these energies are used as features. In our work, the Euclidean distance between the prototypes and the presented feature vectors are considered for classification in a KNN classifier.

All the emotion recognition experiments were performed on a subset of the Berlin Database of Emotional Speech[1], comprising utterances in seven different

[1] Optainable at: `http://pascal.kgw.tu-berlin.de/emodb/`

A. Paiva, R. Prada, and R.W. Picard (Eds.): ACII 2007, LNCS 4738, pp. 710–711, 2007.

emotions recorded from professional actors. The complete details, specification, and structure of the corpus are fully described in [3].

In an earlier work by Petrushin, the basic emotions were grouped as agitation (anger, happiness, and fear) and calm (neutral and sadness) [1]. In the context of a call center application, the best classification, resulting in an accuracy of around 77% was achieved with ensembles of neural networks. However, the simple features based on long term modulation spectrum of speech proposed in this work were efficient in classifying calm and agitated emotions with more than 88% accuracy. Note that, we included seven emotions in "calm" and "agitation", whereas Petrushin [1] considered only five different emotions.

Additionally to the classification results our experiments reveal similarities to the studies by Murray and Arnott, where the relationship between emotion and speech parameters were investigated. This further indicates, that the new features proposed in this work are suitable for emotion recognition [4].

The earlier studies are based on large number of features, and more so prosodic type such as pitch, energy, duration, etc, their variants at frame level and different combinations of the same. The feature extraction process is time-consuming and computational intensive. Apart from requiring huge quantities of data, time, and computational resources, the problem of robustness against unforeseen conditions is an issue of training based approaches. The proposed approach, based on simple features, with the performance close to real-time is independent of gender, speaker, and speech acquisition channel.

In the present work, a novel feature, based on modulation spectrum of speech for emotion recognition, intended for voice based applications is proposed. These simple features as motivated by human auditory system formed the input for classifiers, which performed close to real-time. The performance of automatic emotion recognizers was compared to earlier studies, in which a large set of features, based on pitch, energy, etc were used. The results outperformed earlier work and offer the possibility of extending the experiments further to recognize emotions such as happiness, fear, disgust, anger, boredom, neutral, and sadness separately. Additionally, the proposed features can also be used along with the standard approaches and more sophisticated classifiers to improve the recognition performance, which would be the subject of future research [5].

References

1. Petrushin, V.: Emotion in speech: recognition and application to call centers. In: Proceedings of Artificial Neural Networks in Engineering (1999)
2. Scherer, K.R., Johnstone, T., Klasmeyer, G.: Vocal expression of emotion. In: Handbook of Affective Sciences, pp. 433–456. Oxford University Press, Oxford (2003)
3. Burkhardt, F., et al.: A database of German emotional speech. In: Proceedings of Interspeech (2005)
4. Murray, I.R., Arnott, J.L.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America 93(2), 1097–1108 (1993)
5. Scherer, S., Schwenker, F., Palm, G.: Classifier fusion for emotion recognition from speech. to be published in Proc. of IE07 (2007)