A Study of Emotional Cues in Speech

Yi-Chun Ko, Carol Espy-Wilson

THIRTY YEARS OF SYSTEMS RESEARCH EXCELLENCE

INSTITUTE FOR SYSTEMS RESEARCH
A. JAMES CLARK SCHOOL OF ENGINEERING

Motivation

Speech analytics is an active research area as companies want automatic methods to help understand a customer's experience. Part of this process can involve emotion detection, The importance of automatic emotion recognition has grown with the increasing role of human computer interface applications. Having a device that automatically detects and appropriately responds to its user's emotions could be beneficial to call center managers who want to monitor the quality of the services provided by their agents, and handle very angry customers by specially trained agents.

In e-learning situations, the computer could detect when the user is having difficulty and offer expanded explanations or additional information. What's more, user's interest, stress, and cognitive load can be employed to adapt the teaching pace in an online tutoring system. As a result, the challenging problem of automatic emotion detection has become a research field involving more and more scientists.

Future Work

From our experiment result, we can see that adding these extra features results in a slightly better performance. According to the confusion matrix, most errors come from misclassification between neutral and sad speech. In the future, we plan to look into more different features that are capable of separating neutral and sad speeches. Meanwhile, we will try to acquire more emotional speech database especially those that contain spontaneous speech. We also want to build our own system to automate and simplify our feature extraction process.

References

1. Eyben et al. "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor", *In Proc. ACM Multimedia (MM), Barcelona, Spain*, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013.
2. Lee et al. "An articulatory study of emotional speech production", in Proceedings of InterSpeech, pages 497-500, 2005
3. Deshmukh et al. "Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech." *IEEE Transactions on Speech and Audio Processing*, 13(5):776–786, 2005.
4. Yuan et al. "Speaker identification on the SCOTUS corpus." Journal of the Acoustical Society of America, 123(5), 2008.

Introduction

In previous work, researchers investigated the use of various acoustic parameters for emotion recognition. These parameters include pitch, loudness, mel-frequency cepstral coefficients, and speaking rate among others. Eyben et al. [1] developed the open-Source Media Interpretation by Large feature-space Extraction (openSMILE) toolkit, which is capable of extracting many helpful acoustic features and their statistics for emotion recognition. In this work, we explore additional features related to the vocal source to see if we can significantly improve the performance of the openSMILE toolkit in the detection of the emotions angry, happy, sad and neutral.

Features

For analysis, the Electromagnetic Articulography (EMA) database was used for our study [2]. Three native speakers of American English produced 10 sentences multiple times each with the four emotions. We use only the portion of samples that are evaluated as perfect.

1. Jitter: Jitter is the is cycle to cycle variability of the duration of the pitch period during voiced speech. To measure jitter, we use the aperiodicity, periodicity, and pitch (APP) detector developed by Deshmukh et al. [3]. The APP divides the speech signal into 60 channels, and for each channel the Average Magnitude Difference Function (AMDF) is computed where the AMDF is given by

$$\gamma_n(k) = \sum_{m=-\infty}^{\infty} |x(n+m)w(m) - x(n+m-k)w(m-k)|$$

When speech is quasiperiodic(Fig. 1), the summed output of the channels show clusters of dips at the fundamental period (T_0) and its multiples. More jitter increases the spread of the dip clusters.

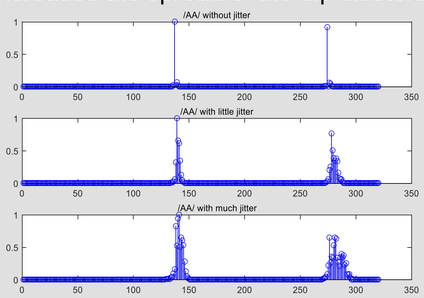

Fig. 1

Fig. 2 shows that speech produced with neutral and sad emotions tends to have higher jitter values relative to speech produced with angry and happy emotions.

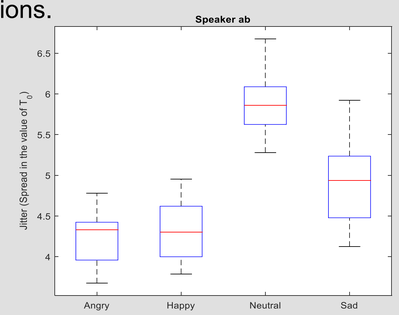

Fig. 2

2. Shimmer: Shimmer is the cycle to cycle variability of the pitch period amplitude. To measure shimmer, we average the height of the first cluster in the dip profiles across each frame. Fig. 3 shows that shimmer is higher for angry, happy and sad speech, relative to neutral speech.

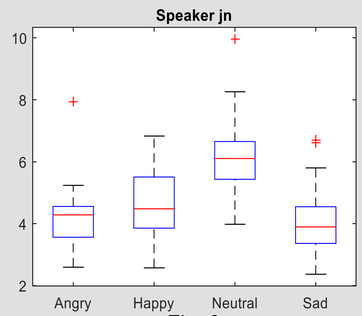

Fig. 3

3. Syllable Rate: Previous studies indicate higher speech rate in anger and happiness, while slower speaking rate in sadness. In this work, speaking rate was calculated as the number of syllables per second. Utterance durations were measured from the corresponding label files produced by the Penn Phonetics Lab Forced Aligner[4]. Our study shows highest rate (5.49 syllable per second) in neutral speech and lowest rate (4.21 syllable per second) produced with sad emotion.

4. Aperiodic Energy: Aperiodic energy is another feature that may contain emotion-related information. For example, when people are sad, their voices tend to be more breathy. Fig. 3 shows speech produced with neutral and sad emotions tends to have more aperiodic energy relative to speech produced with angry and happy emotions.

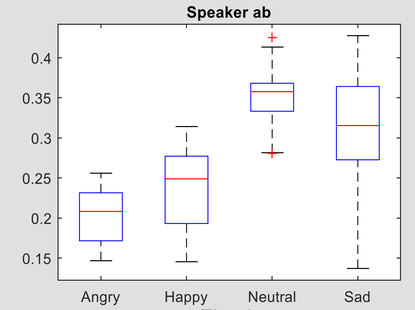

Fig. 4

Experiments

Materials and Methods

The whole process of our feature extraction is depicted in Fig. 5. 10-fold cross validation was done and a 3-nearest neighbors classifier with Euclidean distance is used for training and testing.

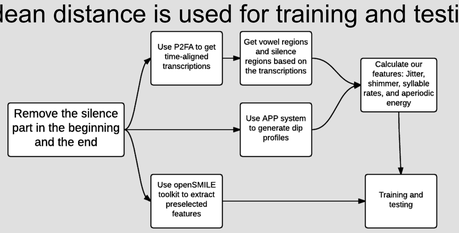

Fig. 5

Results

	Accuracy %
openSMILE features only	79.32
openSMILE features + our additional features	80.34

This work was supported by NSF grant # BCS-0519256

We thank the Signal Analysis and Interpretation Laboratory at University of Southern for sharing the database

Speech Communication Lab