Audio-Based Sentiment Analysis

Bingyue Wang, Mitchell Goff and Justice Amoh

October 1, 2014

1 Introduction

We humans have the innate ability to infer a person's sentiment as we hear them speak. Our brains pick up certain cues in the person's intonation and glean from those the emotions of the speaker; whether they are angry, excited, sad, etc. This skill is referred to as sentiment analysis and we rely on it daily to maintain good dialogue with all people around us. However, sentiment analysis can also be very resourceful in scenarios where it is not practical to employ human capital. For instance, if a company could analyze the sentiments from all phone calls received in customer service, it could, say, gauge the general interest in a product: whether customers are excited or disappointed in that product. In such cases, it is either expensive or not feasible to rely on man-hours since there could be tens or even hundreds of thousands of phone calls to analyze each day. Hence, there is the need for systems that can learn to detect the underlying sentiments or emotions from a person's speech. To that end, sentiment analysis is of great research interest and there is a plethora of published works that explore natural language processing and acoustic signal processing methods for infering speech sentiments [6, 10, 4]. For our CS174 project, we propose a sentiment analysis system based exclusively on audio speech data. Our goal is to train a system on a database of sentiment-labelled speech recordings, such that it can infer the sentiment of new speech data with reasonable accuracy. We are primarily interested in four major sentiment groups:

- Happiness - may include joy, excitement, energetic, etc.
- Neutrality - such as boredom, disinterest, apathy, etc.
- Sadness - including depression, fear, indignation, etc.
- Anger - including outrage, temper, violence, etc.

2 Feature Extraction

In tackling the sentiment learning challenge, we first need to identify useful features that capture essential acoustic properties of speech. The human voice

contains a set of well defined properties such as loudness, pitch, timbre and rate. Intuitively, these properties correlate with the emotions of the speaker and will therefore be reasonable options as features for our learning problem. Besides these, we are also considering more advanced features that have proven resourceful in sentiment analysis studies. One such feature is the fundamental frequency (F0).

According to *Buso et al*, "The fundamental frequency of F0 contour(pitch), which is a prosodic feature, provides the tonal and rhythmic properties of the speech"[3]. Since the fundamental frequency is independent of vocal tract and lexical content, it is likely a good indicator for speaker emotion. Fundamental frequency itself contains a set of characteristics with varying correlation with emotions. Some global aspects of fundamental frequency include minimum, maximum, mean, standard deviation, range. Other aspects such as slope, curvature or inflexion can be used to describe the shape of fundamental frequency.

Besides the fundamental, another feature we are considering is the Mel-Frequency Cepstral Coefficients (MFCCs). The MFCCs describe the spectral content of an audio signal on a frequency scale designed to mimic the acoustic properties the human cochlea[14]. Due to how well they approximate how humans hear sound, the MFCCs have been widely and successfully used in speech recognition research [21, 8]. Some sentiment analysis studies have also employed the MFCCs as features [10, 13, 19].

Having all these features to explore, it will eventually be necessary to determine the relative importance of each feature so we can scale down our feature sets to the most essential ones for our application. A number of feature selection algorithm have been discussed in the literature and one that we are considering for this application is the "KL distance" approach as expressed by *Busso et al* [3]. For example, if we are trying to focus on the pitch of human voice and try to find the relative importance that different characteristics of the pitch such as minimum, mean, standard deviation have on emotion detection, we would follow the following steps. First we will obtain pitch distribution of the audio samples using the software Praat Speech Processing Software [2]. Then we will calculate the distribution of each pitch feature for each emotional category. For each test dataset, the best pitch feature will ideally generate a large KL distance between the feature distribution of the test dataset and the feature distribution of the dataset's incorrect emotional category while the feature will generate a small KL distance between the feature distribution of the test dataset and the feature distribution of the dataset's correct emotional category.

3 Learning Algorithms

Sentiment analysis studies in literature, have employed a number of learning and classification methods such as Gaussian Mixture Models, k-nearest Neighbors, Bayesian Network and Maximum Entropy Model [10, 19, 9]. Two of the learning methods we are considering using are Hidden Markov Models and Binary Emotion Detection.

3.1 Hidden Markov Models (HMM)

The Hidden Markov Model is a statistical model that characterizes time varying patterns as a parametric random process [15, 12]. An HMM can be considered as a finite state machine, that can move from one state to another at any time based on some transition probabilities [11]. HMM is particularly effective in modeling time varying patterns and as a result, has been widely used in speech and pattern recognition applications [7]. In fact, most commonly used speech recognition engines such as Julius, Nuance Dragon Dictate, Sphinx by Carnegie Mellon and HTK by Cambridge University all rely on hidden markov model classifiers. HMM has also been used in emotion recognition studies with promising results [17]. As a learning problem, the HMM approach involves uncovering the "hidden" or unknown finite state machines, that will best yield the different sequence of observations. In our application specifically, we will be attempting to learn hidden markov models for each sentiment from our observations, i.e, our labelled audio speech data. Once we learn these models, we can then evaluate the probability that audio events fit a particular hmm, and hence, classify their sentiments. A three segment procedure for implementing such a hidden markov model based classifier is elaborately discussed by *Rabiner et al* [15]. The 3 segments, each with an associated algorithm, addresses the core challenges in developing an HMM recognition system and has been echoed in several HMM research and applications [20]. One segment provides a framework for computing the probability that a sequence of observations is generated from a particular markov model. The Forward-Backward Algorithm is used here handle this evaluation problem. Another segment involves the optimization of the markov model parameters. This is usually the training phase and relies on the Baum-Welch Algorithm. The third segment then addresses how to select the sequence of states that best explains given observations. This refers to the uncovering-of-hidden-states aspects of the system, which serves as the tool for actually recognizing/classifying new observations. And for this, the Viterbi Algorithm is used. All these processes and algorithms are explained in *Rabiner et al*'s and *Young et al*'s publications [15, 20].

3.2 Binary Emotion Detection

Binary Emotion Detection essentially means tagging the speech as “neutral” or “emotional” without specifically naming the emotion as “angry”, “happy” or sad. Binary emotion detection, unlike explicit emotion detection, suffers less from tagging bias of data collectors and it thus applies better to general situation. Furthermore, binary emotion detection is sometimes the first step in achieving more sophisticated emotion analysis. To achieve binary emotion detection based on the previously selected features, the paper by *Busso et al* has proposed using logistic regression to find the discriminative power of each feature previously selected using KL distance. After selecting the most emotionally salient feature, a Gaussian Mixture Model for each of these features is going to be trained with neutral speech corpus[3]. After all the models are successfully established, the

input speech will be contrasted with the neutral models. The more similar the input is to the speech, the more likely the input speech will be neutral.

4 Dataset

Besides the features and algorithms, the next significant toolkit we need to tackle our defined learning problem is a labelled dataset. In our application specifically, this refers to a corpus of audio speech data that is sentiment labelled. In searching for such a corpus, we found that the AAAC Emotion Research Website contains a multitude of sentiment-tagged data in a wide variety of formats such as video, text, image and video[1]. For the purpose of our project, we will focus mainly on audio data as well as the audio stream extracted from the video data. The language of the audio data is not restricted to English to facilitate the development of a generalized model for sentiment analysis across different languages. Some of sample databases available on the AAAC Emotion Research Website is listed below.

4.1 The SmartKom Database

The SmartKom Database consists of 448 interactive discourse in German by 224 different speakers with each session lasts for about four to five minutes. Each discourse is conducted between a human being and a web machine. Emotion descriptors used to tag the data include "joy, gratification, anger, irritation, helplessness, pondering, reflecting surprise and neutral" [18].

4.2 RECOLA Database

RECOLA stands for REmote COLaborative and Affective interactions which is a French audio database that contains a wide range of audio recordings for socio-affective behaviors conducted in a natural settings. 14 male and 20 female are the subjects of recordings. The tags used in the database - "agreement, dominance, engagement, performance or rapport" - depicts emotions about interaction between two subjects instead of emotion of individual subject [16].

4.3 HUMAINE

Database The HUMAINE Database is a audiovisual dataset that contains 50 clips which last from 5 seconds to 3 minutes. The clips are mainly interactive discourse in English that is conducted in natural settings. Two kinds of tags are used to label the data. The global label applies to the entire clip while time-aligned emotion tracks the emotions of the speakers continuously as the clip proceeds [5].

5 Milestones

By the milestone deadline of October 28th, we plan to achieve the following:

- *Preprocessing & Feature Extraction*: We will implement or obtain the necessary software or code for processing the audio data and extracting the features of interest. The processing will involve actually parameterizing the audio data such correctly sampled matlab vectors for each .wav audio file. For the feature extraction, we will implement the necessary code for extracting all the basic voice features such as pitch, loudness and speaking rate. We obtain external code for computing more complex features like the MFCCs, pitch distribution and the fundamental frequency and it's characteristics. Finally, we will employ the KL distance approach to identify the best features for emotion detection.
- *Learning Algorithms*: We also hope to have implemented two of the three HMM algorithms: the Forward-Backward algorithm, the markov model probability evaluating tool, and the Baum-Welch Algorithm for optimizing model parameters (training phase). With these two, we can develop a somewhat primitive HMM for some sentiments, although we can't actually use the HMM for recognition until the third algorithm.

References

[1] The AAAC Portal.

[2] Paul Boersma and David Weenink. Praat: Doing Phonetics by Computer.

[3] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech and Language Processing*, 17:582–596, 2009.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[5] Ellen Douglas-Cowie, Cate Cox, Jean-Claude Martin, Laurence Devillers, Roddy Cowie, Ian Sneddon, Margaret McRorie, Catherine Pelachaud, Christopher Peters, Orla Lowry, and Others. The HUMAINE database. In *Emotion-Oriented Systems*, pages 243–284. Springer, 2011.

[6] Mohsen Farhadloo and Erik Rolland. Multi-Class Sentiment Analysis with Clustering and Score Representation. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 904–912. IEEE, December 2013.

[7] Mark Gales and Steve Young. The Application of Hidden Markov Models in Speech Recognition. *Foundations and TrendsÂ® in Signal Processing*, 1(3):195–304, 2007.

[8] Chadawan Ittichaichareon, Siwat Suksri, and Thaweesak Yingthawornsuk. Speech Recognition using MFCC. *International Conference on Computer Graphics, Simulation and Modeling*, pages 135–138, 2012.

[9] Lakshmish Kaushik, Abhijeet Sangwan, and John H. L. Hansen. Automatic sentiment extraction from YouTube videos. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 239–244, December 2013.

[10] Loic Kessous, Ginevra Castellano, and George Caridakis. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3(1-2):33–48, December 2009.

[11] Sergio Matos, Surinder S Birring, Ian D Pavord, and David H Evans. An automated system for 24-h monitoring of cough frequency: the leicester cough monitor. *IEEE transactions on bio-medical engineering*, 54(8):1472–9, August 2007.

[12] Sergio Matos, Student Member, Surinder S Birring, Ian D Pavord, David H Evans, and Senior Member. Detection of Cough Sounds in Continuous Audio Recordings Using Hidden Markov Models. 53(6):1078–1083, 2006.

[13] M Murugappan, Nurul Qasturi Idayu Baharuddin, and S Jerritta. DWT and MFCC based human emotional speech classification using LDA. In *2012 International Conference on Biomedical Engineering (ICoBE)*, pages 203–206. IEEE, February 2012.

[14] Dominique Pastor and Andre Goalic. On the Recognition of Cochlear Implant-Like Spectrally Reduced Speech With MFCC and HMM-Based ASR. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):1065–1068, July 2010.

[15] Lawrence R. Rabiner. A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[16] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[17] B. Schuller, G. Rigoll, and M. Lang. Hidden Markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 2, pages II–1–4. IEEE.

[18] Silke Steininger, Florian Schiel, and Angelika Glesner. Labeling procedures for the multi-modal data collection of SmartKom. 2002.

[19] Yongjin Wang and Ling Guan. Recognizing Human Emotional State From Audiovisual Signals. *IEEE Transactions on Multimedia*, 10(4):659–668, June 2008.

[20] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, and Others. *The HTK book*, volume 2. Entropic Cambridge Research Laboratory Cambridge, 1997.

[21] Bin Zhen, Xihong Wu, Zhimin Liu, and Huisheng Chi. On the Importance of Components of the MFCC in Speech and Speaker Recognition. *ACTA SCIENTIARUM NATURALIUM-UNIVERSITATIS PEKINENSIS*, 37(3):371—-378, 2001.