Acoustic Feature-Based Sentiment Analysis of Call Center Data

Master’s Thesis Defense

By: Zeshan Peng

Advisor: Dr. Yi Shang
Roadmap

- Introductions
- Related works
- **Proposed methods**
  - Acoustic feature-based sentiment recognition using classic machine learning algorithms
  - Acoustic feature-matrix-based sentiment recognition using deep convolutional neural network
- Experiment results
- Conclusion and future works
Roadmap

• Introductions
  – Problem
  – Motivation
  – Solution

• Related works

• Proposed methods

• Experiment results

• Conclusion and future works
Introduction

- Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information (Wikipedia).

- With the advancement of machine learning, sentiment analysis on audio signal has attracted attention.

![Sentiment Analysis](image-url)

*Discovering people opinions, emotions and feelings about a product or service*
Problem

- Identify whether a meeting has been successfully scheduled by the customer in a phone call conversation with a representative.

Audio Records

Meeting Scheduled?

Yes  No
Motivation

• Models can be used to identify customer’s attitude and help monitor representative behaviors in real time during the process of a phone call conversation

• Hard to train sale representatives or build good relationships with customers without monitors

• Manpower is limited compared to the huge amount of data

• Sentiment analysis can help making better business decisions
Solution

• Two methods proposed
  – Acoustic feature-based sentiment recognition using classic machine learning algorithms
  – Feature-matrix-based sentiment recognition using deep convolutional neural network

• Machine learning models are developed by different learning algorithms for classification and performances are compared
  – SVM with cubic kernel
  – K-nearest neighbor
  – Neural network
  – Deep convolutional neural network
Roadmap

• Introductions

• Related works
  – Text-based approach
  – Acoustic feature-based approach

• Proposed methods

• Experiment results

• Conclusion and future works
Text-based Approach

• Text corpora is huge and data are generated daily with an increasing speed

• Many models are proposed to capture text characteristics
  – Lexical based (A Naïve-Bayes Strategy for Sentiment Analysis on English Tweets, Gamallo, 2014)
  – Semantic based (Sentiment Analysis on Twitter, Kumar, 2012)

• Transcribe audio data into text and then do sentiment analysis on text data (Sentiment Analysis of Call Centre Audio Conversation using Text Classification, Ezzat, 2012)
Acoustic feature-based Approach

• Mel frequency cepstral coefficients
  – Mel Frequency Cepstral Coefficients For Music Modeling. (Logan, 2000)
  – Musical Genre Classification of Audio Signals. (Tzanetakis, 2002)
    • MFCCs are used for music genre classification
  – Automatic emotional speech classification. (Ververidis, 2004)
    • MFCCs are used for emotion recognition
Acoustic feature-based Approach

- Timbre and Chroma
  - Generating Music from Literature. (Davis, 2014)
  - Classify Music Audio With Timbre and Chroma Features. (Ellis, 2007)
    - “Chroma features are less informative for classes such as artist, but contain information that is almost entirely independence of the spectral features.”
    - Unsupervised Approach to Hindi Music Mood Classification. (Patra, 2013)

- Pitch and speech rate
  - The Organizational Voice: The Importance of Voice Pitch and Speech Rate in Organizational Crisis Communication. (Waele, 2017)
    - Critically important when forming impressions
Roadmap

• Introductions

• Related works

• Proposed methods
  – Acoustic feature-based sentiment recognition using classic machine learning algorithms
  – Acoustic feature-matrix-based sentiment recognition using deep convolutional neural network

• Experiment results

• Conclusion and future works
Method 1 – Classic Machine Learning
Method 1 – Classic Machine Learning
Pre-processing and Cleaning

• Remove ringtone signal and other unrelated parts
  – Several methods have been tried, but none of them worked well
  – Manually
Pre-processing and Cleaning

• Remove ringtone signal and other unrelated parts

• Speaker Diarization

  – Split each audio record into segments by speaker turns using transcripts
Method 1 – Classic Machine Learning

1. Clean/Segmentation
2. Feature Extraction
3. 

- Customer
  - \([x_1, x_2, ..., x_n]\)
- Customer
  - \([x_1, x_2, ..., x_n]\)
- Customer
  - \([x_1, x_2, ..., x_n]\)

- Representative
  - \([x_1, x_2, ..., x_n]\)
- Representative
  - \([x_1, x_2, ..., x_n]\)
- Representative
  - \([x_1, x_2, ..., x_n]\)
Feature Extraction

- Prosody features
- Short-term features
Feature Extraction

- Prosody features
  - Pitch (or frequency)
  - Number of pauses (silence that lasts more than 0.3 seconds)
  - Speech rate (number of words spoken per second)
  - Intensity (or loudness)
  - Duration (total length of a segment)
  - Jitter (variation of frequency)
  - Shimmer (variation of amplitude)
Feature Extraction

- Short-term features
  - Mel-frequency cepstrum coefficients (MFCCs)
  - Chroma
  - Timbre

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCCs</td>
<td>Mel Frequency Cepstral Coefficients from a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale</td>
</tr>
<tr>
<td>Chroma</td>
<td>A representation of the spectral energy where the bins represent equal-tempered pitch classes</td>
</tr>
<tr>
<td>Timbre</td>
<td>The character or quality of a sound or voice as distinct from its pitch and intensity</td>
</tr>
</tbody>
</table>
Method 1 – Classic Machine Learning

1. Clean/Segmentation
2. Feature Extraction
3. Model Training

- Customer
  - [x₁, x₂, ..., xₙ]
- Representative
  - [x₁, x₂, ..., xₙ]
Classification Model

- Build customer model & representative model
  - Based on the according segments group

- Classic machine learning algorithms
  - SVM with cubic kernel
  - K-nearest neighbor
  - Shallow feed forward neural network
Method 1 – Classic Machine Learning

Test Audio Record
Method 1 – Classic Machine Learning

Test Audio Record

Segmentation

Customers

......

Representatives
Method 1 – Classic Machine Learning

Test Audio Record

Segmentation

Customers

Representatives

Yes
Yes
No

......

Yes
No
Yes
Method 1 – Classic Machine Learning

Test Audio Record

Segmentation

Customers

Customer Model

Yes Yes No

Representatives

Representative Model

Yes No Yes

Majority Vote

YES
Method 2 - Deep Learning

• Deep convolutional neural network has shown unprecedented performance on images, but it can be also used for non-image datasets, such as features in the text, for different classification tasks.

• Feeding character-level feature matrix into deep convolutional neural network has achieved competitive results with current state-of-art methods (Zhang, 2015).
Input – Feature Matrix
Input – Feature Matrix

50 ms  50 ms  ...
Input – Feature Matrix

50 ms 50 ms 50 ms 50 ms

34 dimensional

Short-term Feature Vector
Input – Feature Matrix

50 ms  50 ms  50 ms  50 ms

34 dimensional

Short-term Feature Matrix
Input – Feature Matrix

50 ms 50 ms 50 ms 50 ms

34 dimensional

Short-term Feature Matrix
Input – Feature Matrix

50 ms 50 ms 50 ms 50 ms

34 dimensional Short-term Feature Matrix

34 300 300

Deep CNN
Input – Feature Matrix

50 ms 50 ms 50 ms 50 ms

34 dimensional Short-term Feature Matrix

300 300 300

Deep CNN

Label
Deep Learning - 4 Conv. Layer Architecture

1-D convolutional kernel & 1-D max pooling kernel
Roadmap

• Introductions
• Related works
• Data processing pipeline
• Experiment results
  – Dataset
  – Experiment design
  – Experiment results
• Conclusion and future works
Datasets

- 86 audio records (from 30 seconds to 8 minutes)
  - 2588 segments after speaker diarization
  - Around 30 segments per audio record

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Records</td>
<td>31</td>
<td>55</td>
<td>86</td>
</tr>
<tr>
<td>Segments</td>
<td>1284</td>
<td>1304</td>
<td>2588</td>
</tr>
</tbody>
</table>
Experiment Design

• Dataset split
  – Training (85%) Testing (15%)

• Learning algorithm
  – Support vector machine with cubic kernel
  – K-nearest neighbor
  – Shallow feed forward neural network
  – Deep convolutional neural network

• Validation
  – Five fold cross validation
Experiment Results – Per Segment

Customer Model vs. Representative Model

- **Prosody**: Customer Model 0.646, Representative Model 0.57
- **Short-term**: Customer Model 0.82, Representative Model 0.684
- **Together**: Customer Model 0.822, Representative Model 0.694
Experiment Results – Per Record

Model Comparison Based On Majority Votes

Prosody
- Customer: 0.583
- Representative: 0.536
- Together: 0.571

Short-term
- Customer: 0.773
- Representative: 0.685
- Together: 0.739

Together
- Customer: 0.786
- Representative: 0.691
- Together: 0.75

Legend:
- Customer
- Representative
- Together
## Experiment Results

### Prediction Accuracy vs. Number of Convolutional Layers

<table>
<thead>
<tr>
<th>Layers</th>
<th>Feature Vector</th>
<th>Feature Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.623</td>
<td>0.673</td>
</tr>
<tr>
<td>4</td>
<td>0.636</td>
<td>0.691</td>
</tr>
<tr>
<td>5</td>
<td>0.647</td>
<td>0.658</td>
</tr>
<tr>
<td>6</td>
<td>0.632</td>
<td>0.644</td>
</tr>
</tbody>
</table>

Legend: 
- **feature vector**
- **feature matrix**
Experiment Results

1-D CNN vs. 2-D CNN On Different Pooling Kernel Size

- 1-D CNN
- 2-D CNN
Experiment Results

Comparison On Different Window Size

Accuracy
Experiment Results

Comparisons on Different Machine Learning Algorithms

<table>
<thead>
<tr>
<th></th>
<th>Prosody</th>
<th>Short-term</th>
<th>Together</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.583</td>
<td>0.773</td>
<td>0.786</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>0.592</td>
<td>0.735</td>
<td>0.749</td>
</tr>
<tr>
<td>Feed forward NN</td>
<td>0.614</td>
<td>0.792</td>
<td>0.796</td>
</tr>
<tr>
<td>Deep CNN</td>
<td>0.647</td>
<td>0.826</td>
<td>0.813</td>
</tr>
</tbody>
</table>
Roadmap

- Introductions
- Related works
- Data processing pipeline
- Experiment design and results
- Conclusion and future works
Conclusions

• Short-term features, such as MFCCs, Chroma and timbre are good indicators for sentiment in our dataset

• Temporal information can be captured by feeding feature matrixes into deep convolutional neural networks to improve prediction accuracy
Contributions

• Two methods have been proposed and implemented in this work

• Different machine learning methods are compared based on experiment results

• Different parameter settings of training deep convolutional neural network on feature matrixes are experimented and results are compared
Future Works

• Automatic speaker diarization process
  – Transcripts are costly and time consuming

• Create better feature representation to feed into deep neural networks
  – Feature matrix is not the end

• Find better features

• Multimodalities

• Collect more data!!!
Thank You All For Attending!