University of Missouri

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

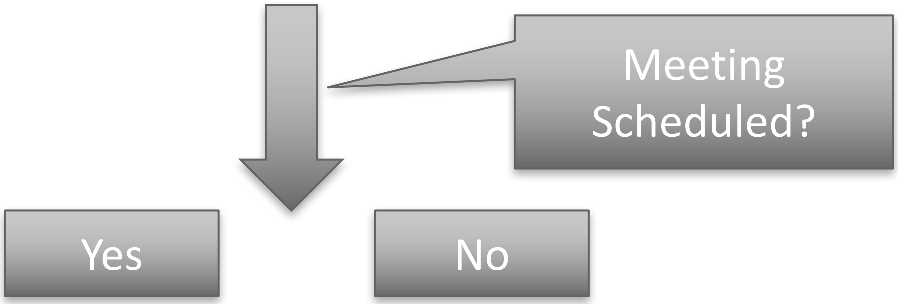
Introduction

- 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

Problem

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

Audio Records



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

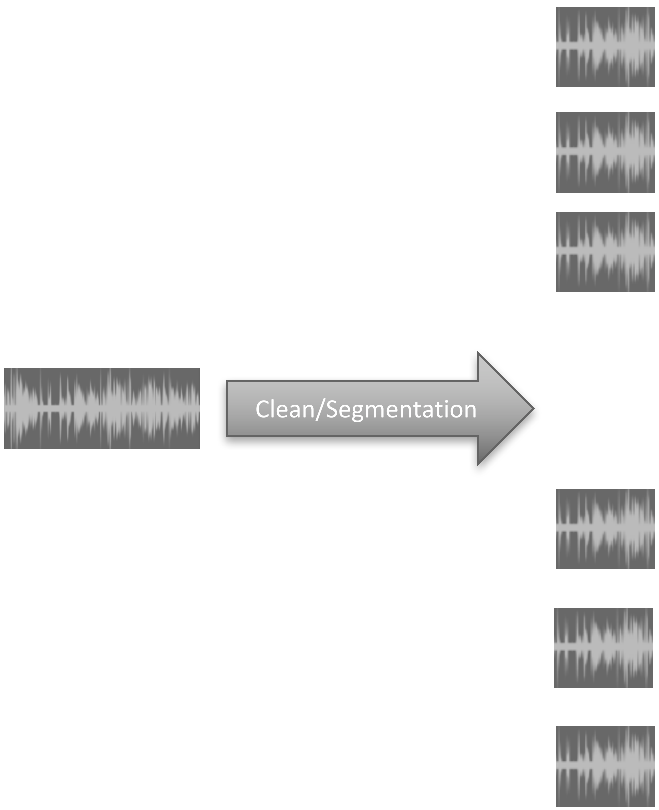
- Timbre and Chroma
  - Generating Music from Literature. (Davis, 2014)
  - Classify Music Audio With Timbre and Chroma Features. (Ellis, 2007)
    - “Chroma features are less informative for classes such as artist, but contain information that is almost entirely independence of the spectral features.”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

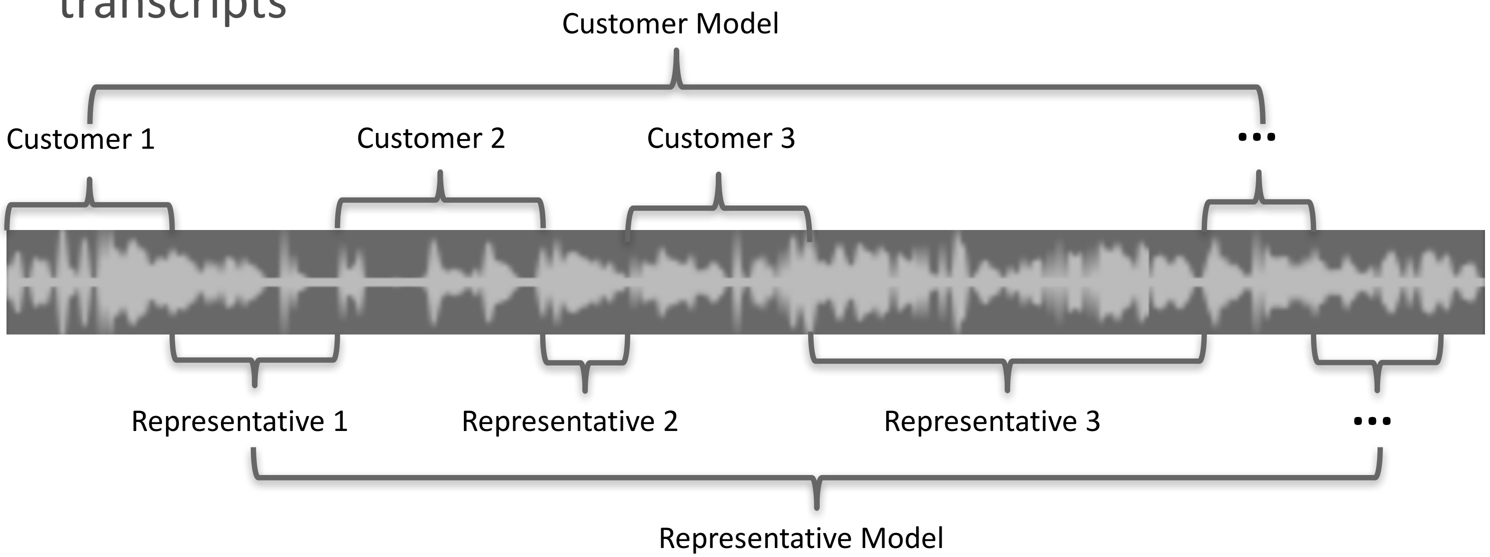


Pre-processing and Cleaning

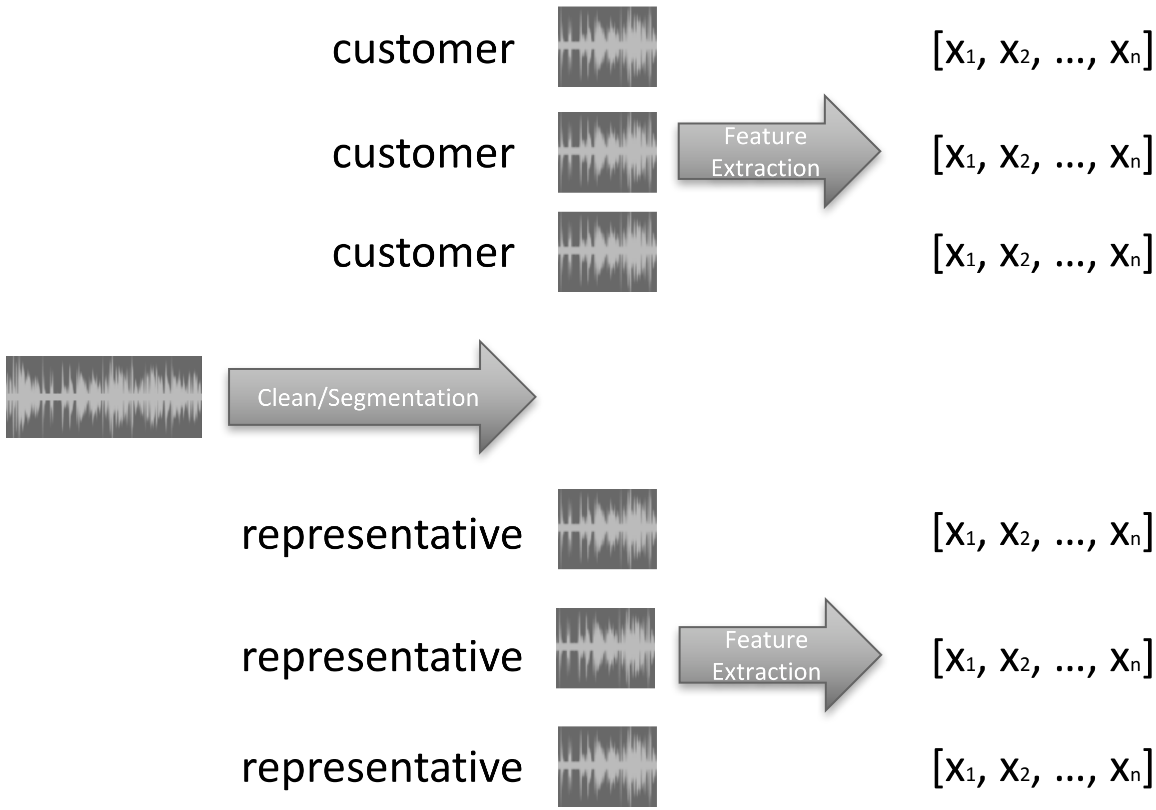
- Remove ringtone signal and other unrelated parts
  - Several methods have been tried, but none of them worked well
  - Manually

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

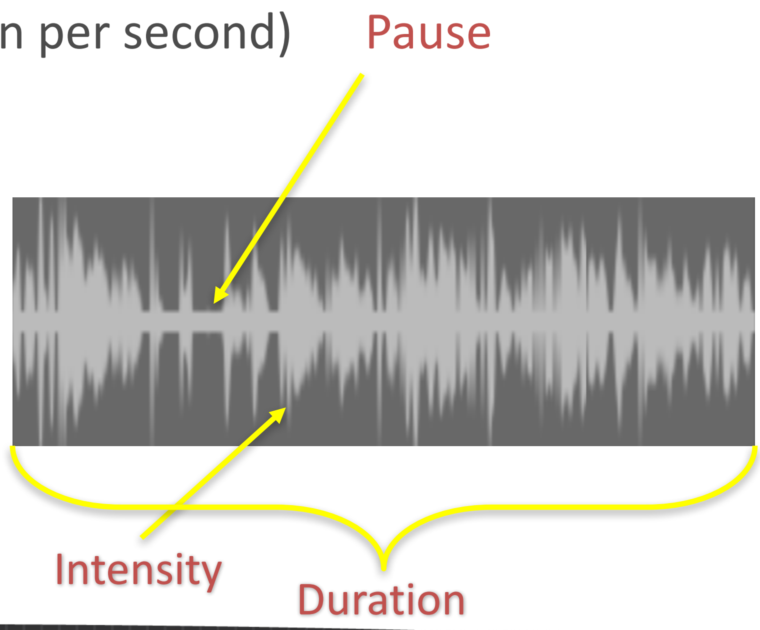


Feature Extraction

- 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)

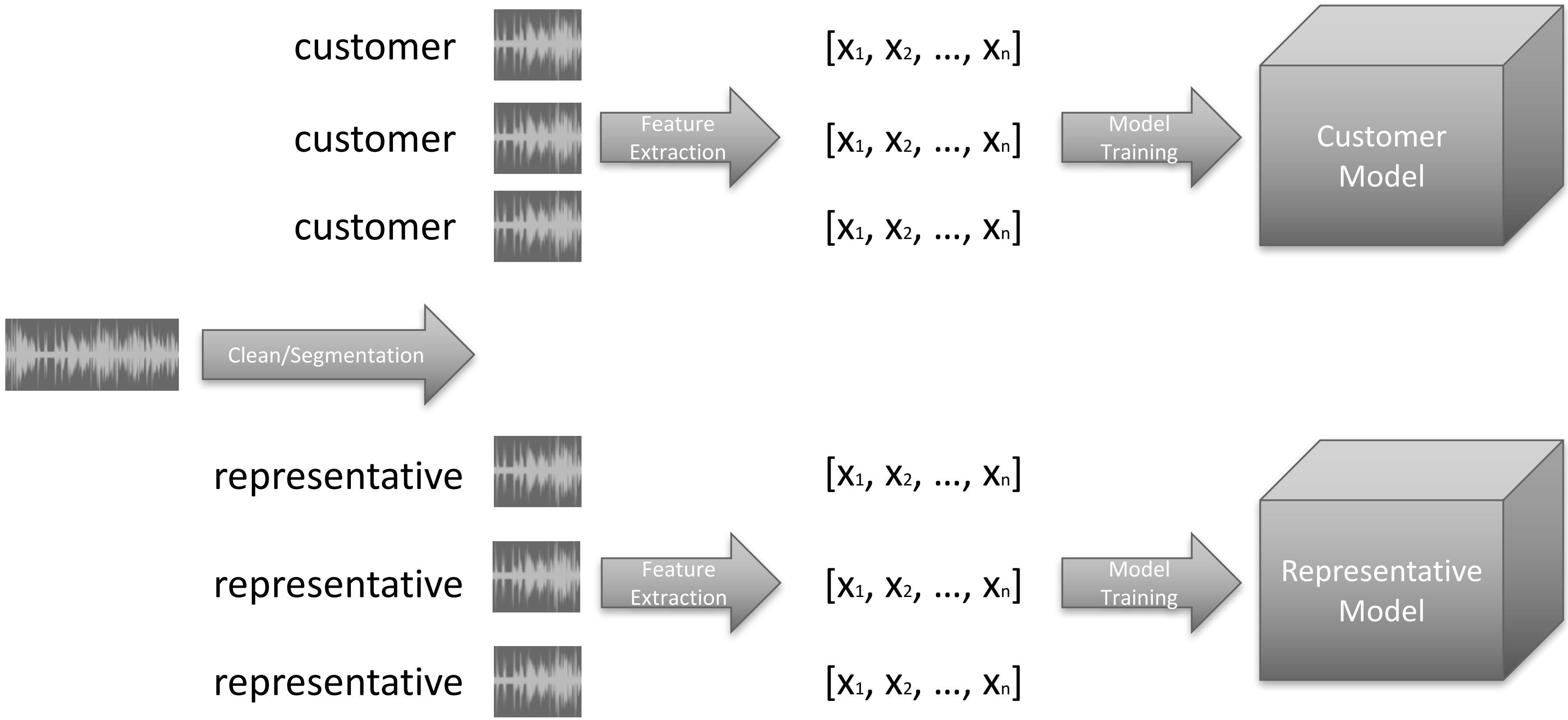


Feature Extraction

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

Feature	Description
MFCCs	Mel Frequency Cepstral Coefficients from a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale
Chroma	A representation of the spectral energy where the bins represent equal-tempered pitch classes
Timbre	The character or quality of a sound or voice as distinct from its pitch and intensity

Method 1 – Classic Machine Learning



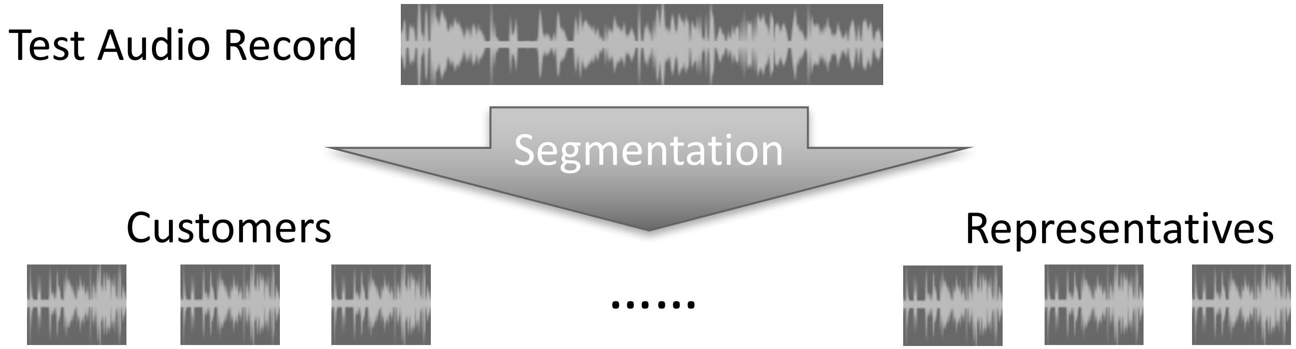
Classification Model

- 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

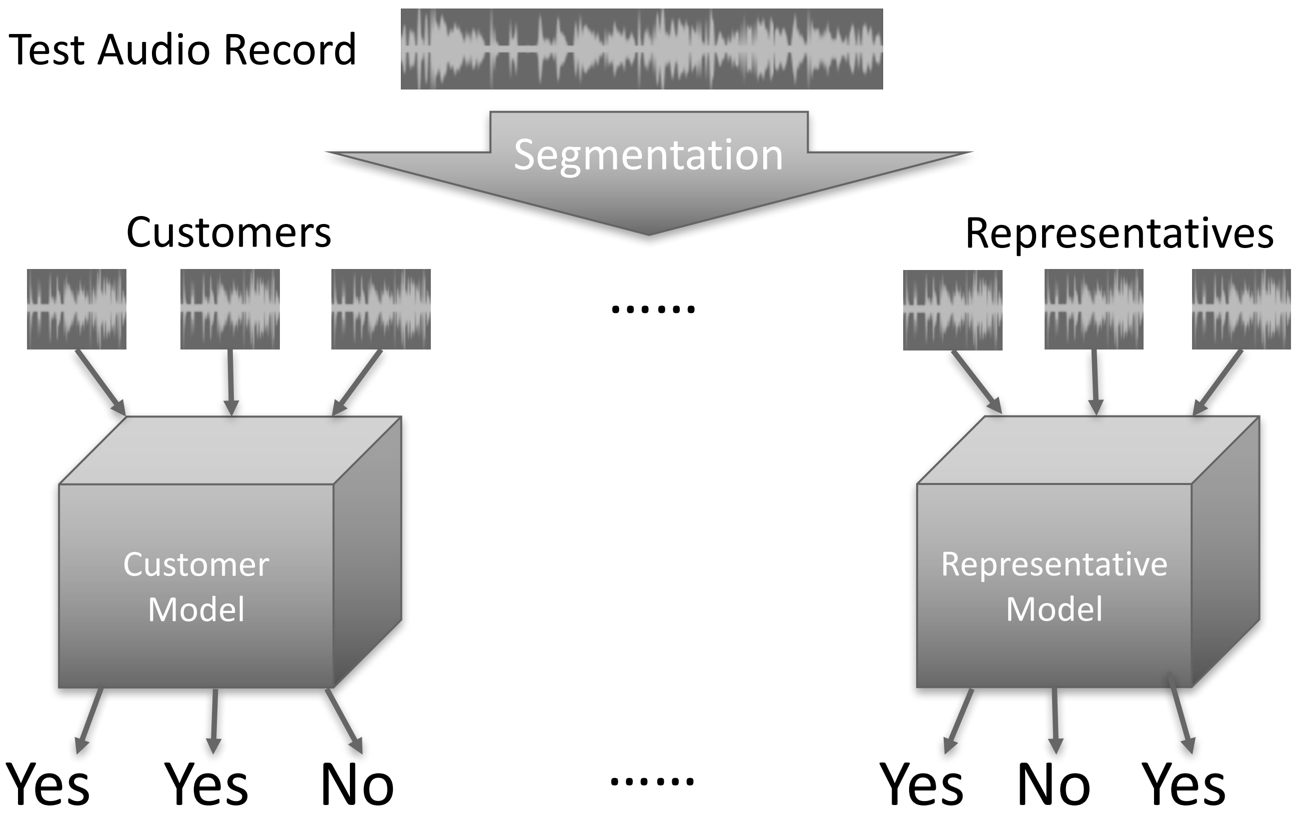
Method 1 – Classic Machine Learning

Test Audio Record

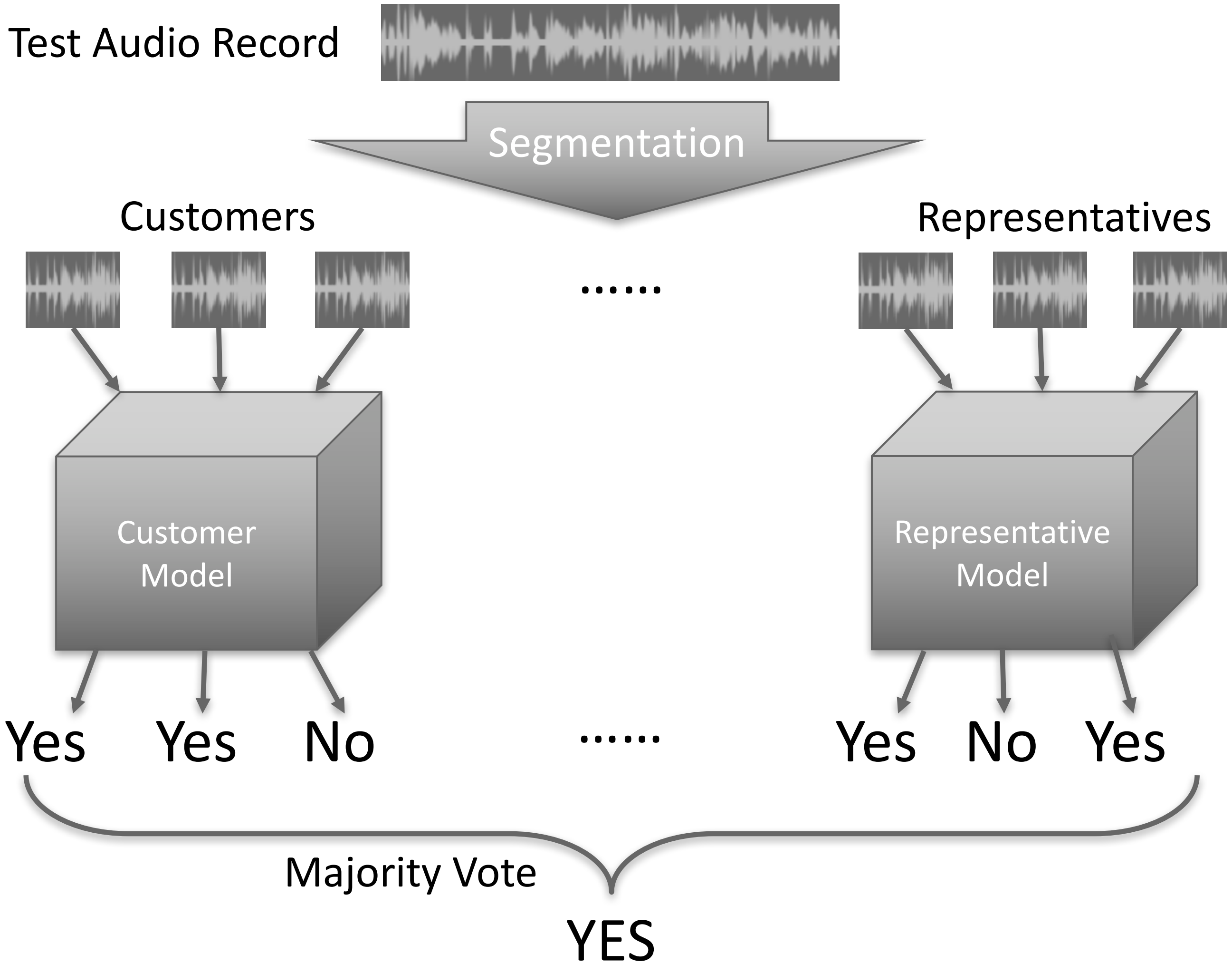
Method 1 – Classic Machine Learning



Method 1 – Classic Machine Learning



Method 1 – Classic Machine Learning



Method 2 - Deep Learning

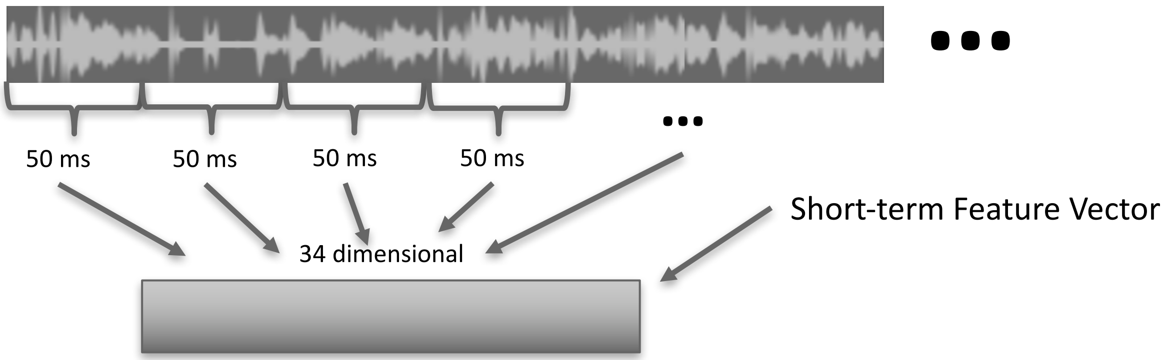
- Deep convolutional neural network has shown unprecedented performance on images, but it can be also used for non-image datasets, such as features in the text, for different classification tasks
- Feeding character-level feature matrix into deep convolutional neural network has achieved competitive results with current state-of-art methods (Zhang, 2015)

Input – Feature Matrix

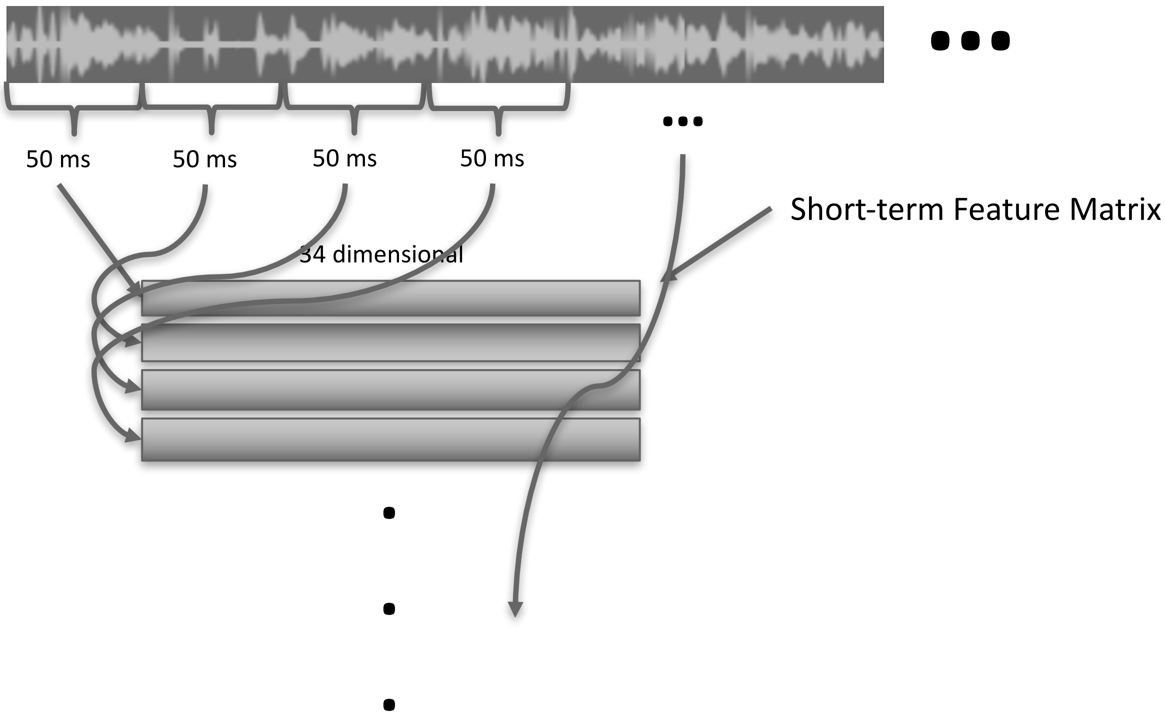
…

Input – Feature Matrix

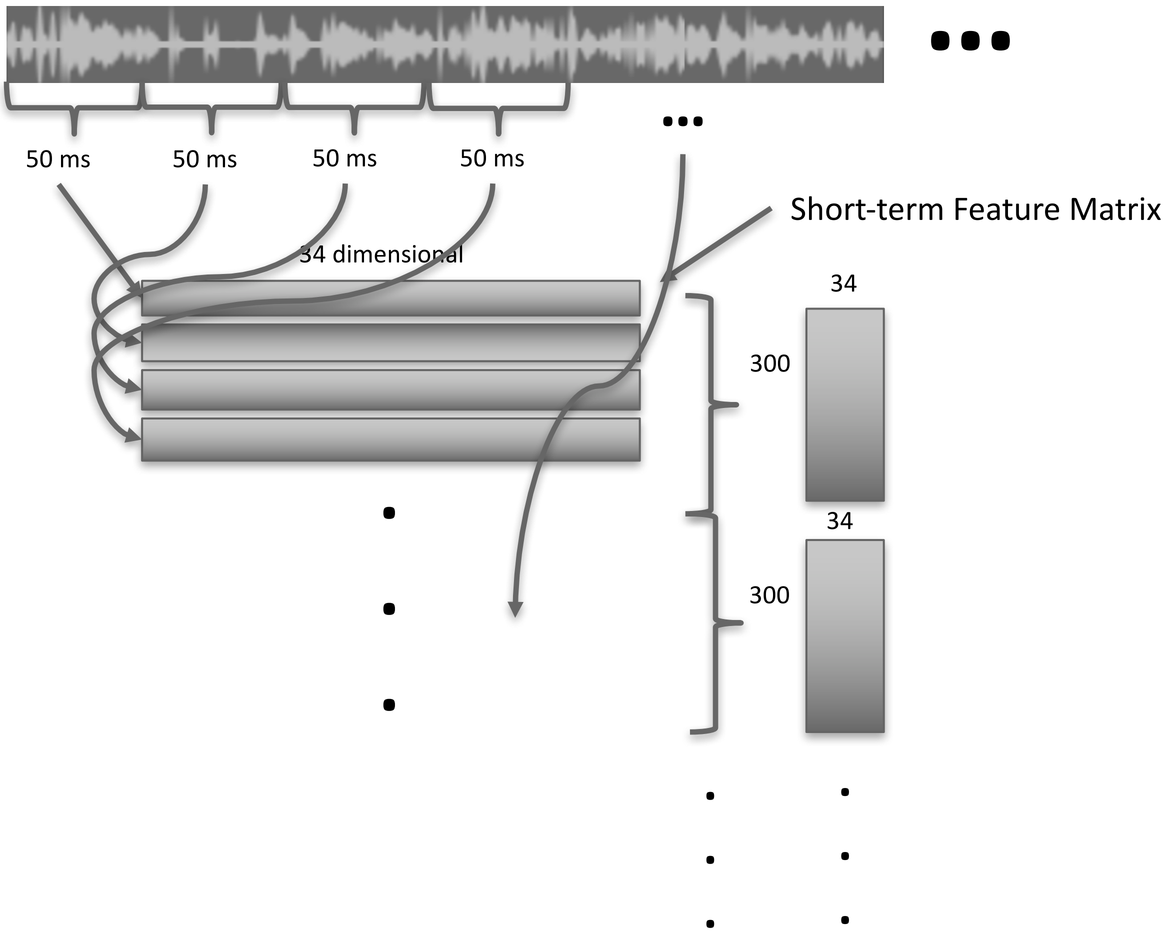
Input – Feature Matrix



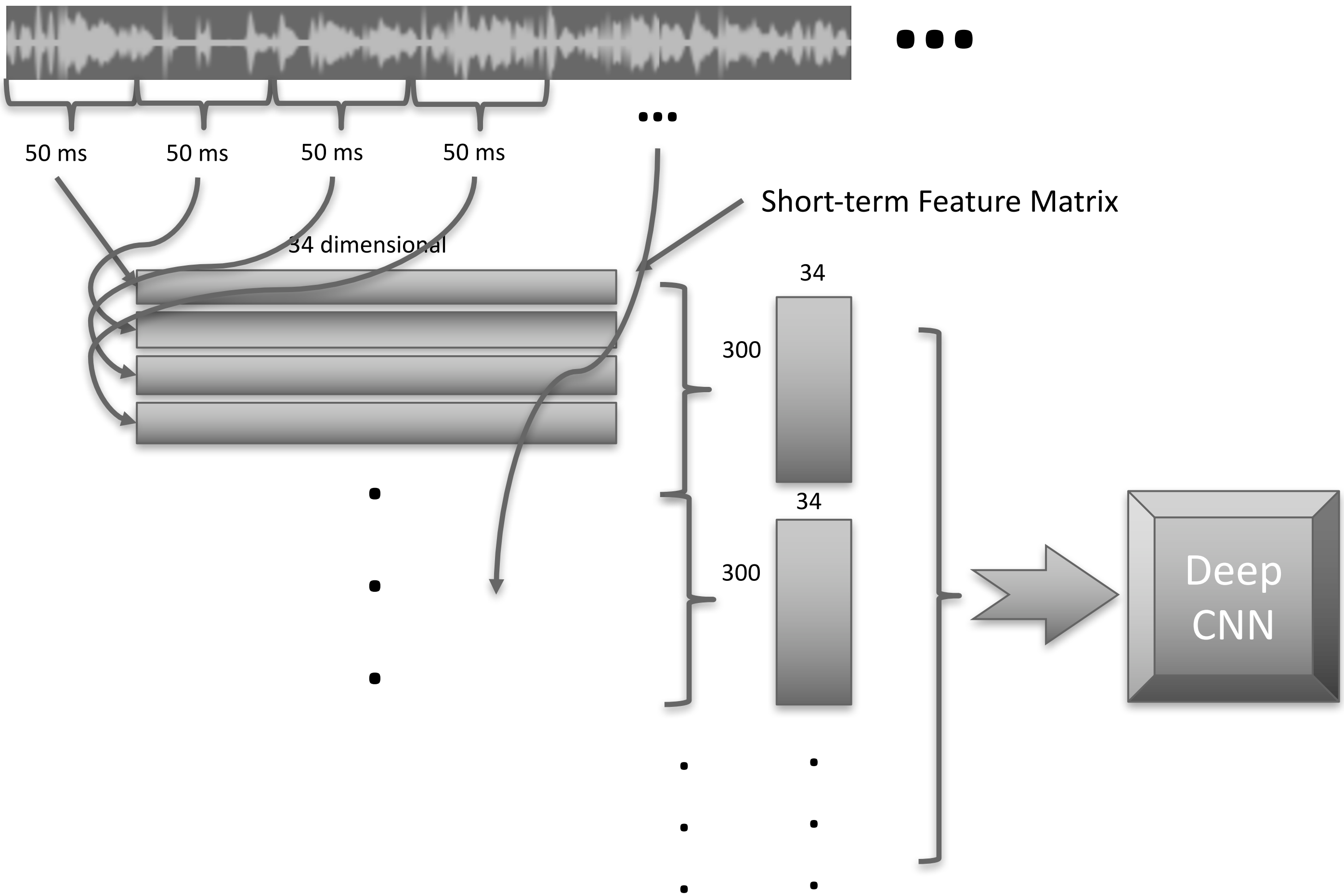
Input – Feature Matrix



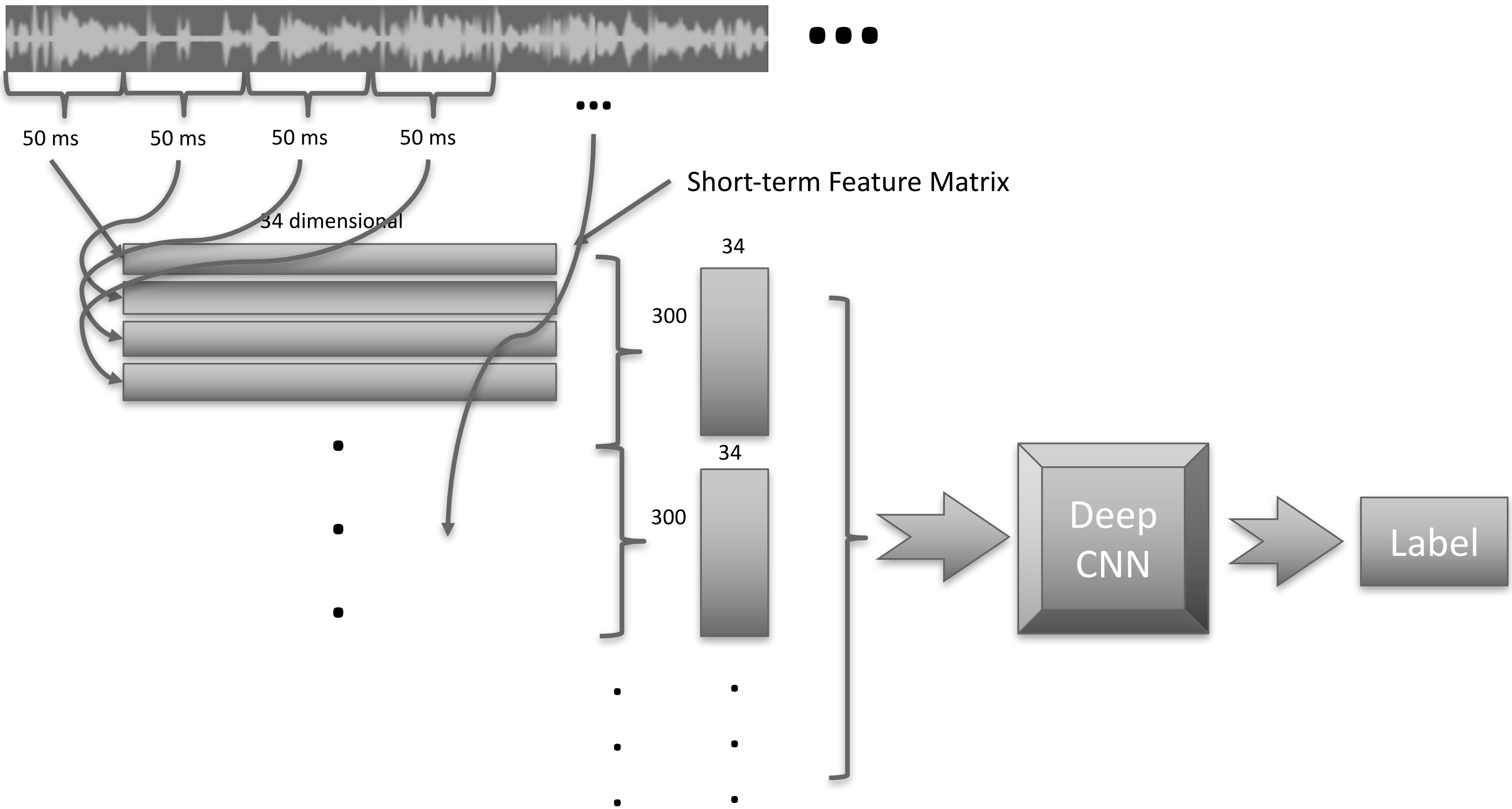
Input – Feature Matrix



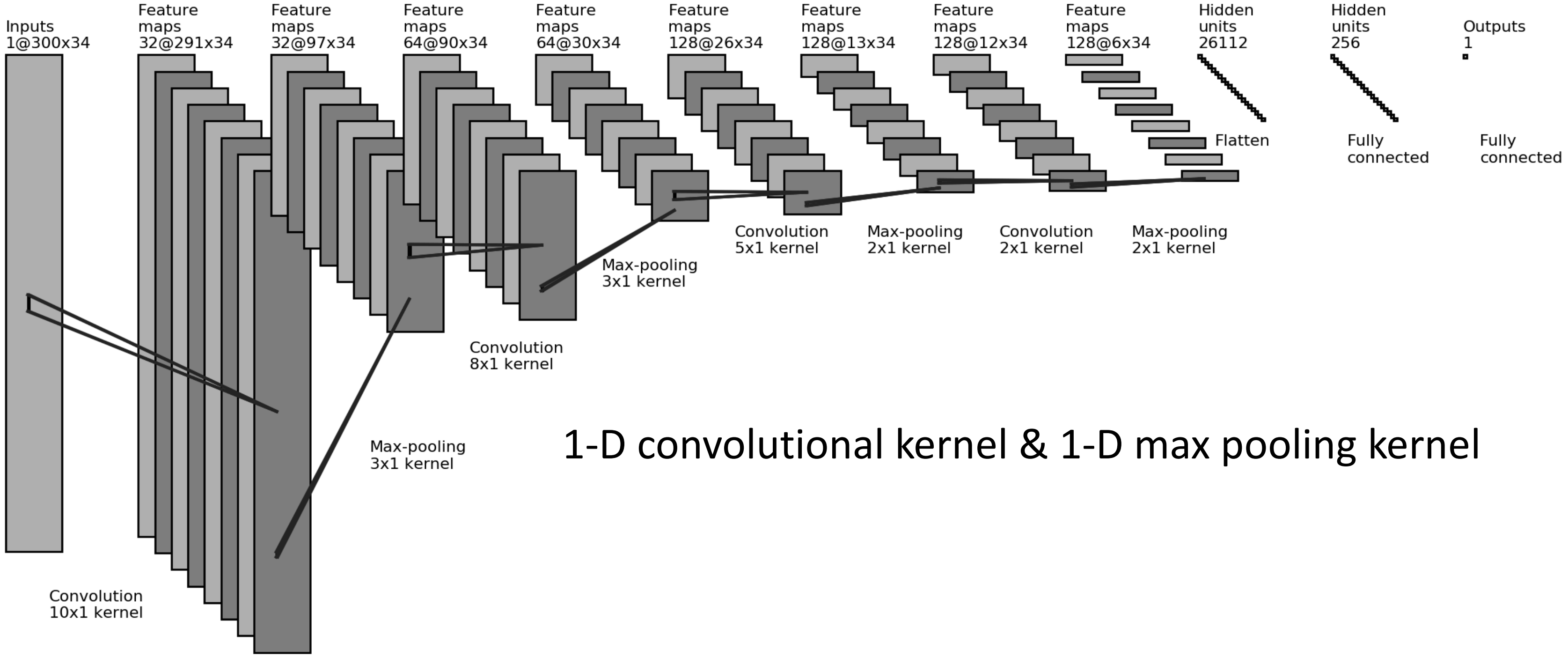
Input – Feature Matrix



Input – Feature Matrix



Deep Learning - 4 Conv. Layer Architecture



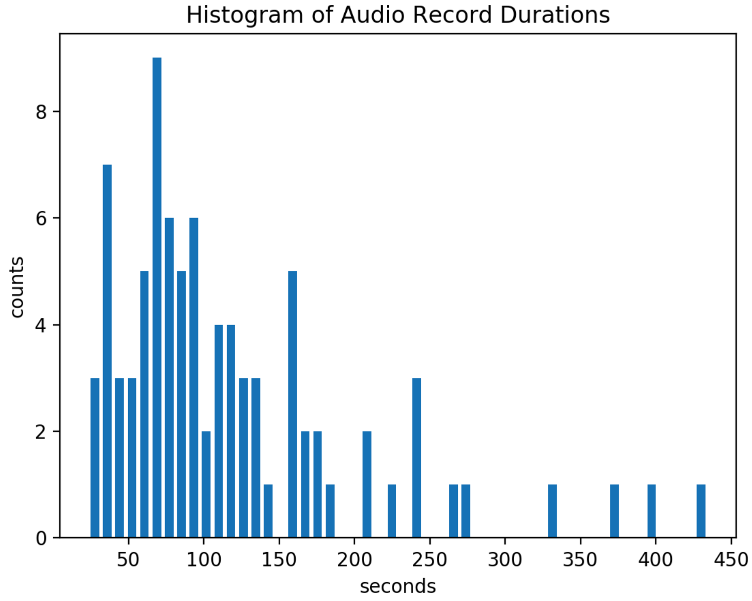
1-D convolutional kernel & 1-D max pooling kernel

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

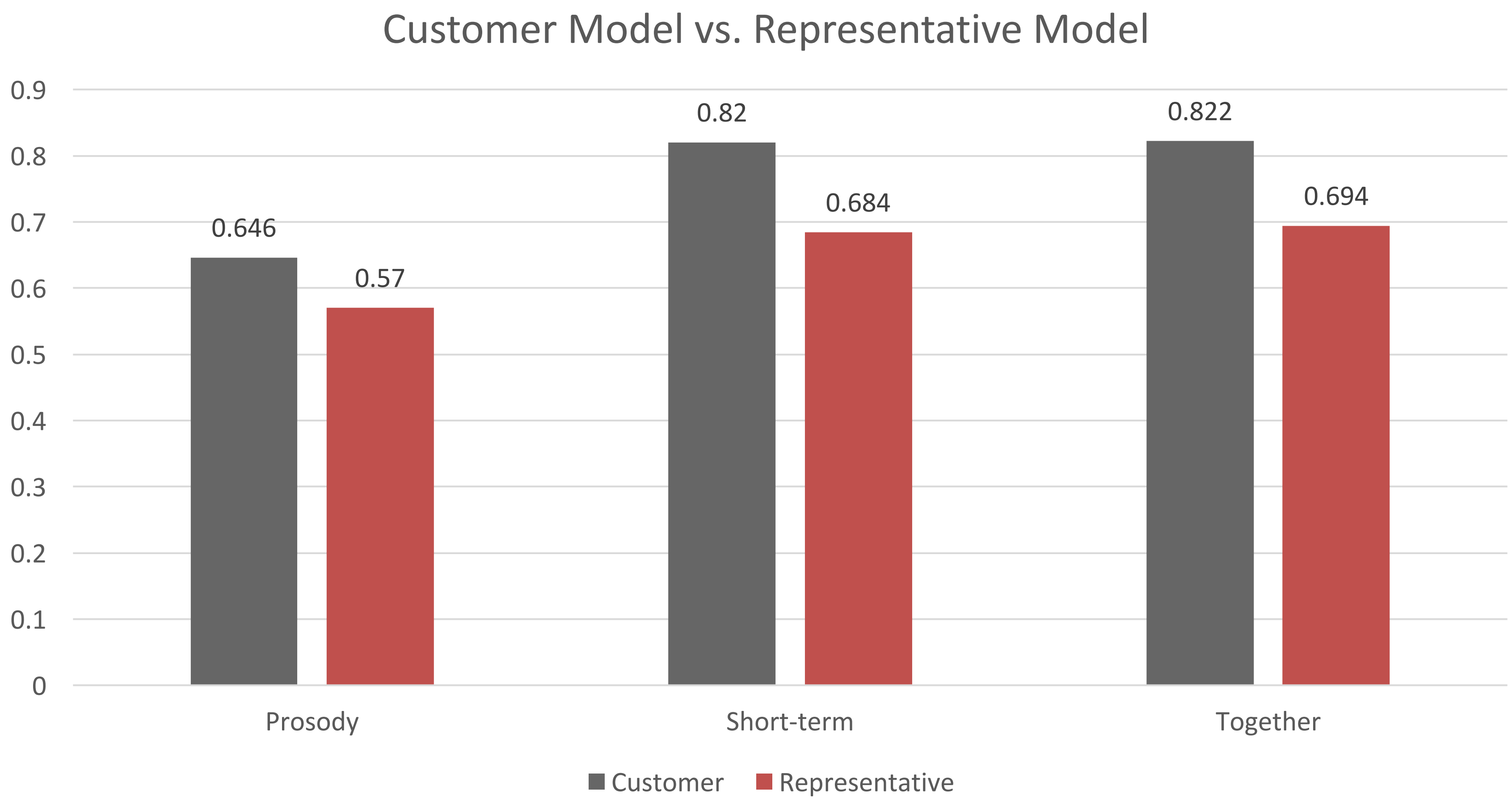


	Positive	Negative	Total
Audio Records	31	55	86
Segments	1284	1304	2588

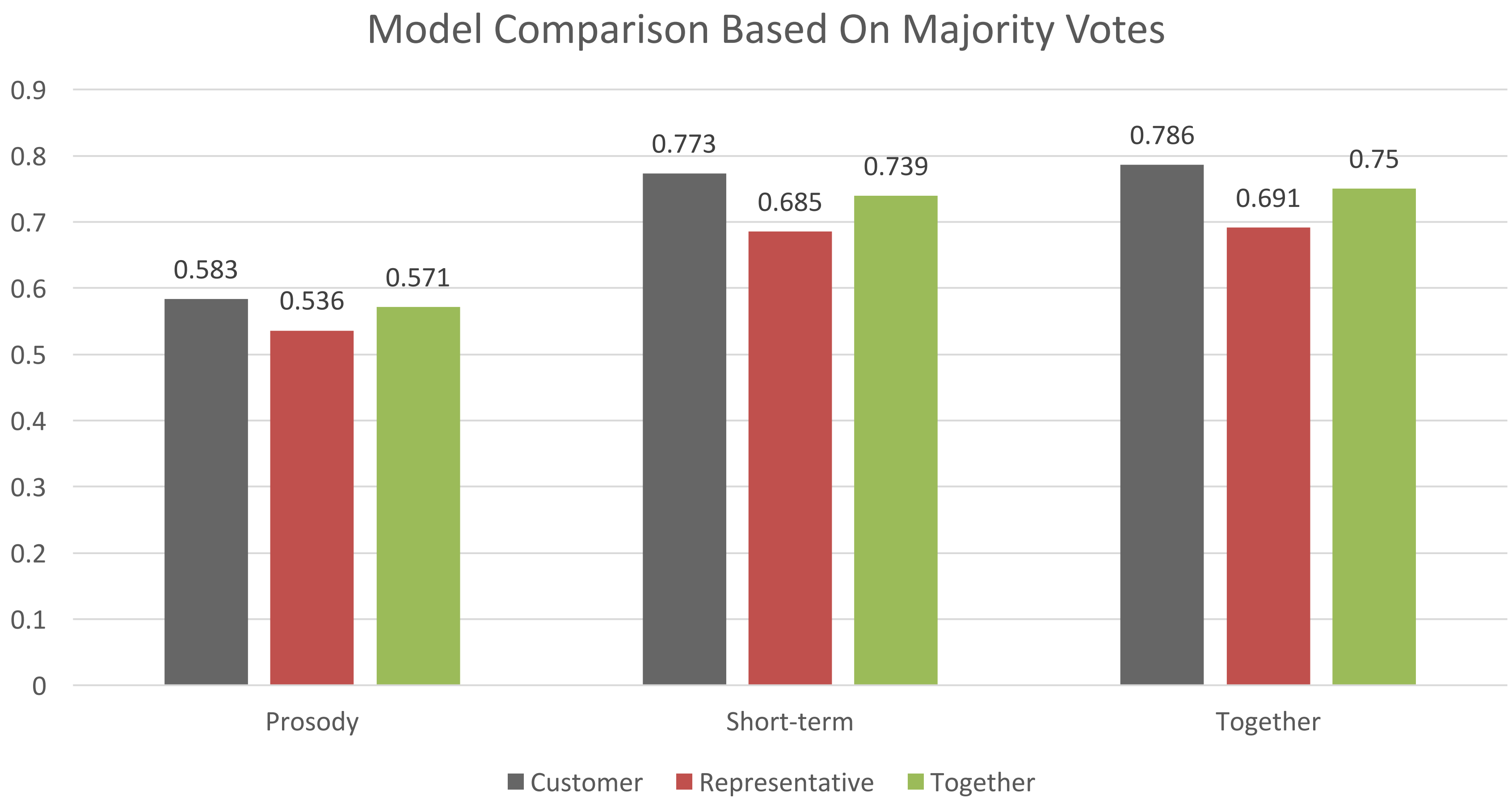
Experiment Design

- 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

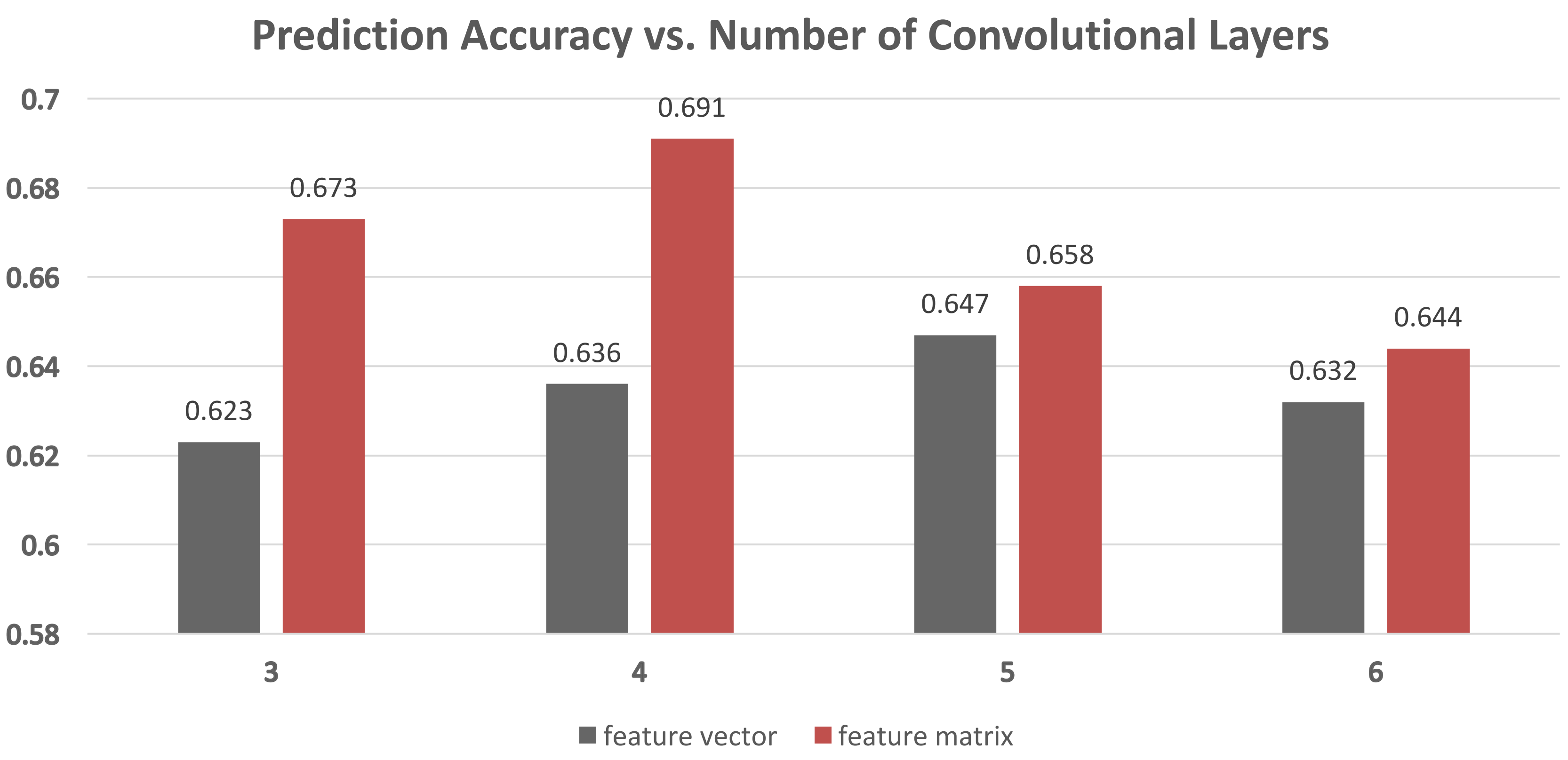
Experiment Results – Per Segment



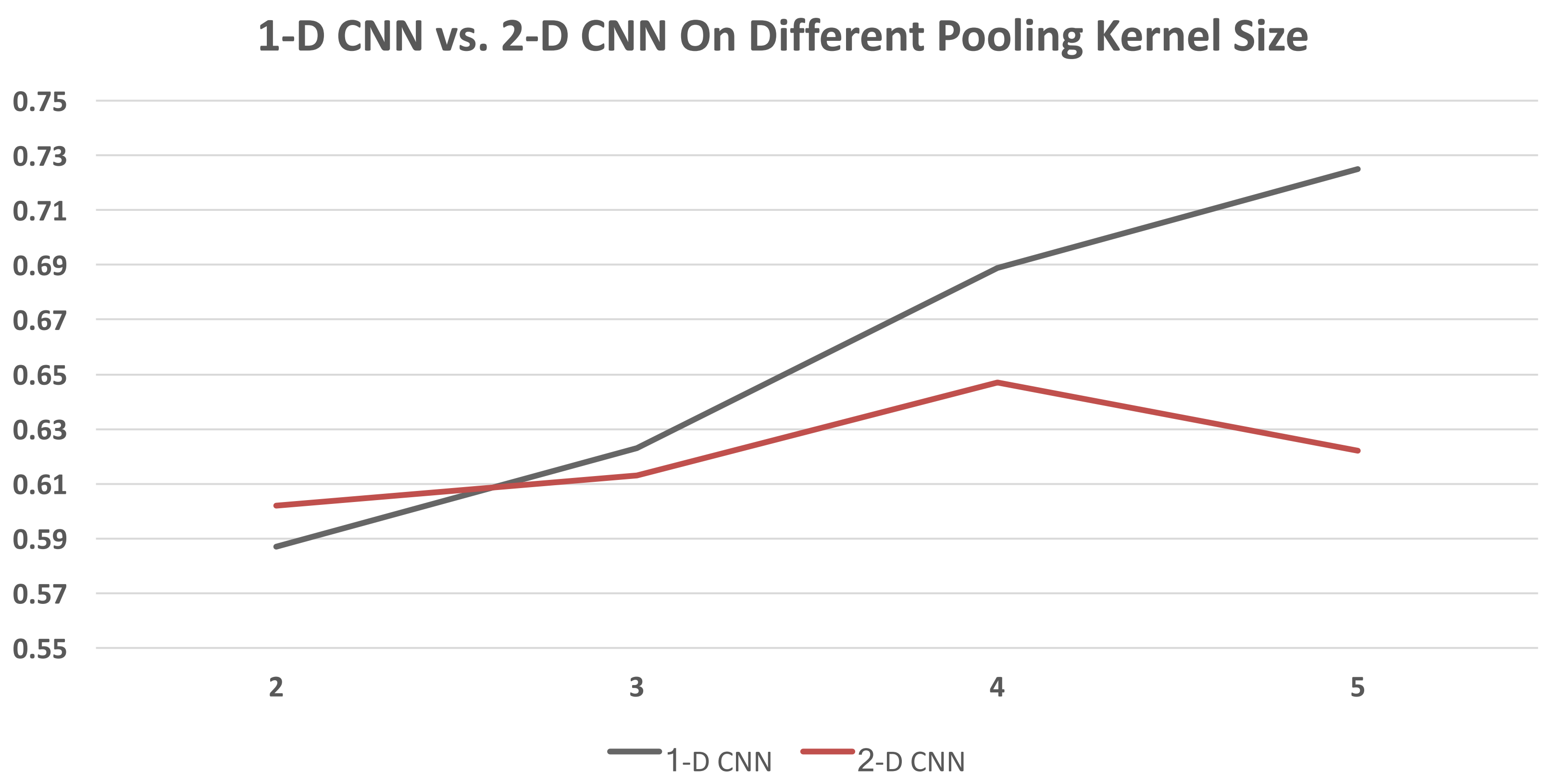
Experiment Results – Per Record



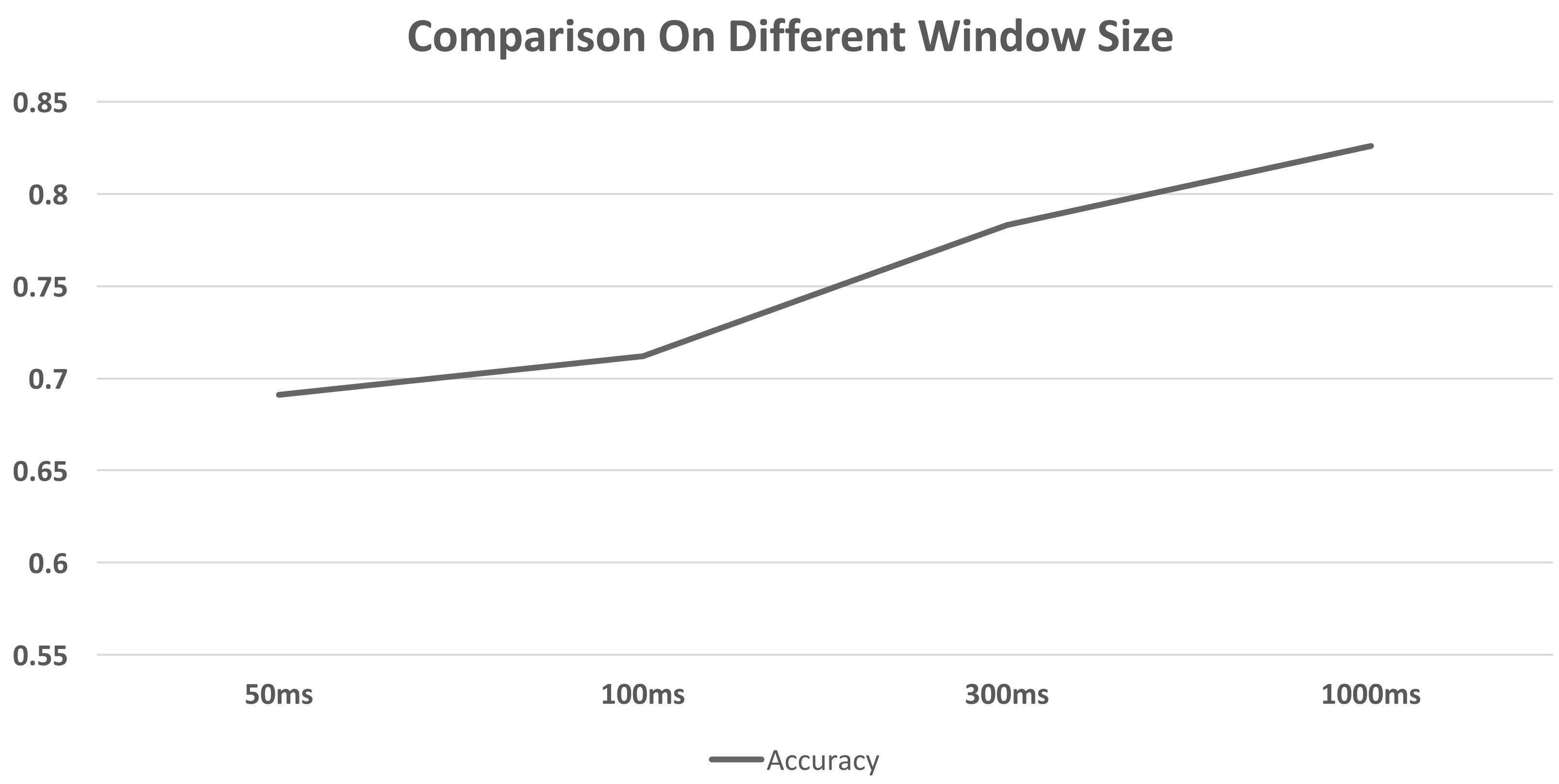
Experiment Results



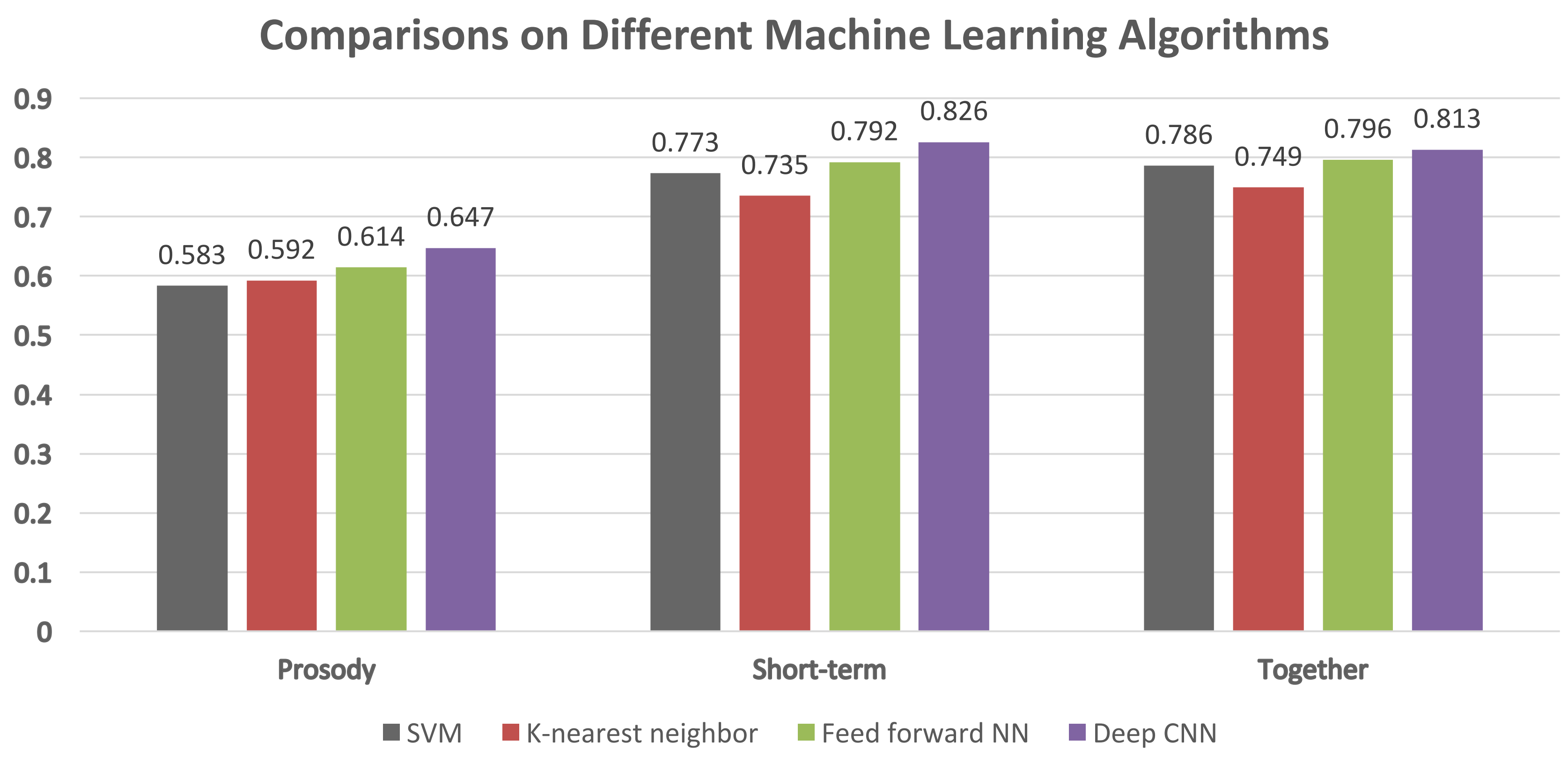
Experiment Results



Experiment Results



Experiment Results



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!