School of Informatics, Computing and Engineering

CSCI-B 659: Project Presentation

Automatic Question Detection in Speech

Taslima Akter, Hasika Mahtta, Khandokar Md. Nayem

INDIANA UNIVERSITY BLOOMINGTON

Overview
Problem Statement

Overview

- **Identifying questions** in human dialogs is an important first step to automatically processing and understanding natural speech.
- Question detection can be viewed as **a subtask of speech act** or dialogue act tagging, which aims to label functions of utterances in conversations, with categories as question/statement/backchannel.
- Question detection is useful for meeting indexing and summarization.
- The main goal of this project is to build a audio classifier capable of recognising questions and declarative sentences for male and female speech corpus.

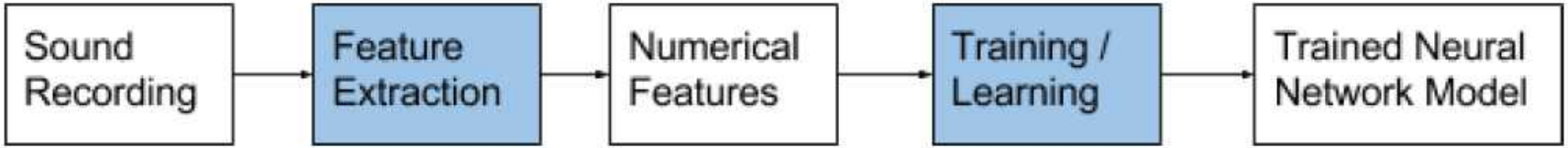
Types of Question

yes-no	did did you do that?
declarative	you're not going to be around this afternoon?
wh	what do you mean um reference frames?
tag	you know?
rhetorical	why why don't we do that?
open-ended	do we have anything else to say about transcription?
or	and @frag@ did they use sigmoid or a softmax type thing?
or-after-YN	or should i collect it all?

Table 1: Examples for each MRDA question category as defined in this paper, based on Dhillon et al. (2004).

Problem Statement

So, the essence of the problem is that : given an audio recording, can a system be trained to classify different types of **Question** (considering intonation and stress).



Motivation & Related Works

Motivation

- The current speech recogniser examples such as Google Home mini and Alexa can understand speech and respond accordingly.
- But these technologies are not smart enough to detect echo questions, rhetorical questions etc.
- Our main motivation is to develop a model that will consider intonation, stress and other speech attributes to classify different groups of question.

Related Works

- **Question Detection in Spoken Conversations Using Texual Conversations-** Comparison between the text-trained model with models trained on manually-labelled, domain-matched spoken utterances with and without prosodic features has been done. ({amargoli,mo}@ee.Washington.edu, Seatttle, WA, USA).

- **Any Questions? Automatics Question Detection in Meetings-** This paper focuses on describing the efforts towards the automatic detection of English questions in meetings using ICSI MRDA Corpus.(Kofi Boaakye, Bnoit Favre, Dilek Hakkini-Tur – Berkeley, CA, USA)

- Integration of prosodic tree model with language model based on words yields best performance accuracy in **detecting questions/question *form*** (Shriberg et al.'98: English)

- Studies of **different types (*functions*) of clarification questions** (Rodríguez & Schlangen'94: German; Edlund et al.'95: Swedish)

Methodology

Corpus

Speech corpus

- Large collections of audio recordings of spoken language.
- Most speech corpora have additional text files containing transcriptions of the words spoken and the time each word occurred in the recording.

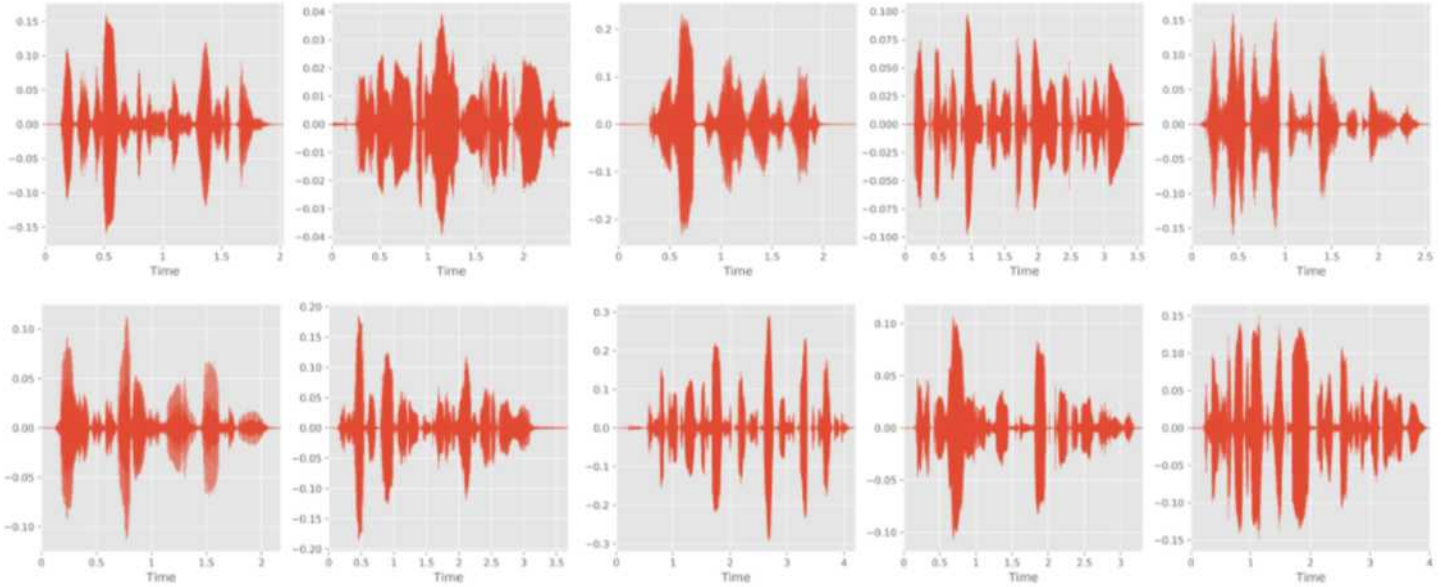
- We use miniature Timit corpus (Kaggle) and some manually recorded audio files (courtesy to professor) in the .wav format.
- We labelled the dataset manually.
- Our corpus contains around only 300 audio input files.

Feature Extraction

Python **LibROSA** feature extraction methods are used.

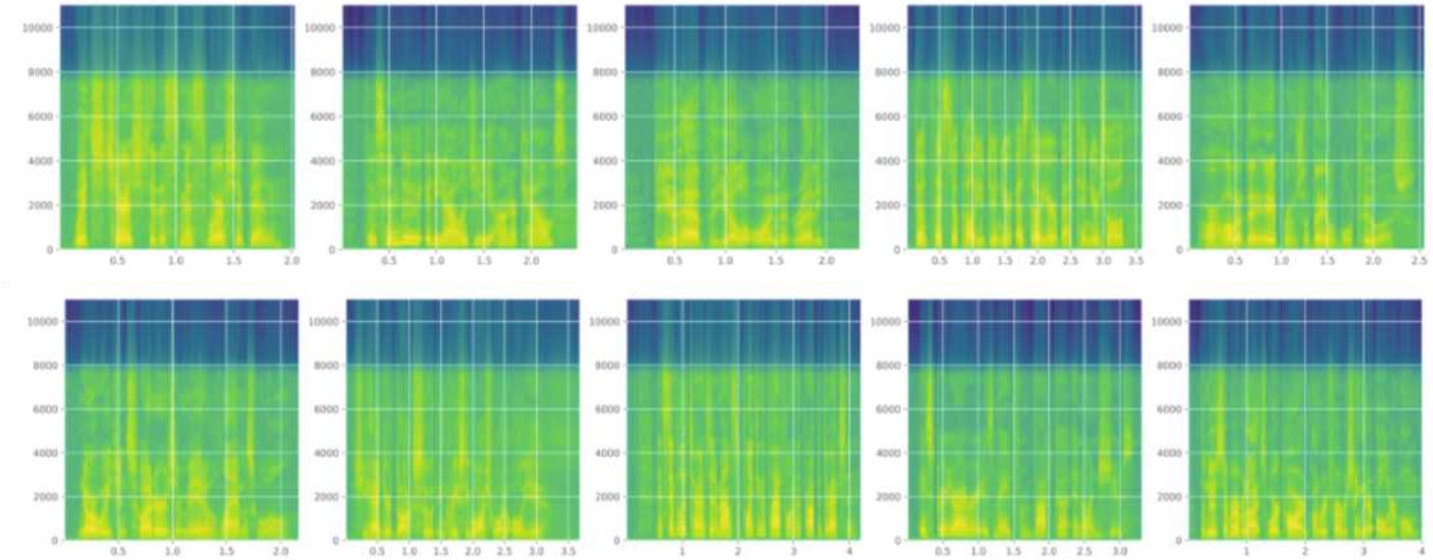
- **Spectrogram:** Time-Frequency representation of the speech signal
- **Mel-frequency cepstral coefficients (MFCC) :** The coefficient that collectively make up the short term power spectrum of a sound.
- **RMSE:** Compute root-mean-square (RMS) energy for each frame, either from the audio samples or a spectrogram.
- **Pitch:** An attribute of auditory sensation in terms of which sounds may be ordered from low to high.

Feature Extraction (Wave plots)



Wave plots for Questions (Top) and Declarative (Bottom) sentences

Feature Extraction (Spectrogram)



Spectrograms for Questions (Top) and Declarative (Bottom) sentences

Models

Model 1

Recurrent Neural Network (RNN)

- **Single .wav single sample:** At each batch, a .wav file is a single sample, fed to RNN cells.
- **Single .wav multiple sample:** .wav files are mini-batched into overlapping samples and fed to RNN.
- **Word chunk:** Each .wav files are broken down into chunks of word .wav and these are fed into RNN.

Model 2

Convolutional Recurrent Neural Network (CRNN)

- **Spectrogram** as feature (inexpensive), no handcrafted feature
- CNN model can be parallelized and faster than RNN (sequential)

Challenges

- No appropriate speech dataset is open/ available for question classification, specially for echo question.
- Available dataset is too small for applying Deep Nets.
- Ground truth can vary depending on context and perspective.
- Word chunking doesn’t 100% accurate which might effect the final result.

Future Works

- Stress and intonation in speech will be considered for better classification.
- Using text annotation with speech .wavs can be interesting to investigate.
- More features can be considered (Chromagram of a STFT, Tonnetz, etc.).

Thank You!

INDIANA UNIVERSITY BLOOMINGTON

FULFILLING *the* PROMISE