Modeling Acoustic-Prosodic Cues for Word Importance Prediction in Spoken Dialogues

RIT | Sushant Kafle, Matt Huenerfauth
Golisano College of Computing and Information Sciences (GCCIS)
Rochester Institute of Technology

Motivation

Not all words are equally important to the meaning of a spoken message; identifying word-importance is useful for speech recognition evaluation or text summarization. Prior models based only on text features are insufficient for conversational speech: Unlike formal text, conversation transcripts may lack capitalization or punctuation, use less formal grammar, contain filler words, or include more out-of-vocabulary words. Moreover, spoken conversations require transcription by a human or an automatic speech recognition (ASR) system, yet ASR is not always feasible or reliable, especially for low-resource languages.

Prosodic Cues in Speech

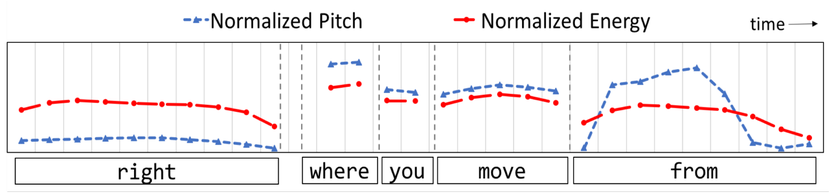

Fig 1: Example of conversational-style of text ("right where you move from") that is difficult to disambiguate without prosodic cues; the correct sentence structure being: "Right! Where you move from?"

Objectives

In this work, we investigate modeling acoustic-prosodic cues for predicting the importance of words to the meaning of a spoken dialogue. We frame this as a sequence labeling task and explore neural architecture for context modeling on speech. Our goal is to incorporate speech features extracted from a sequence of voiced-regions in speech into a prediction model, operating at an utterance-level, to make the prediction of the importance of individual words in the utterance.

Word Importance

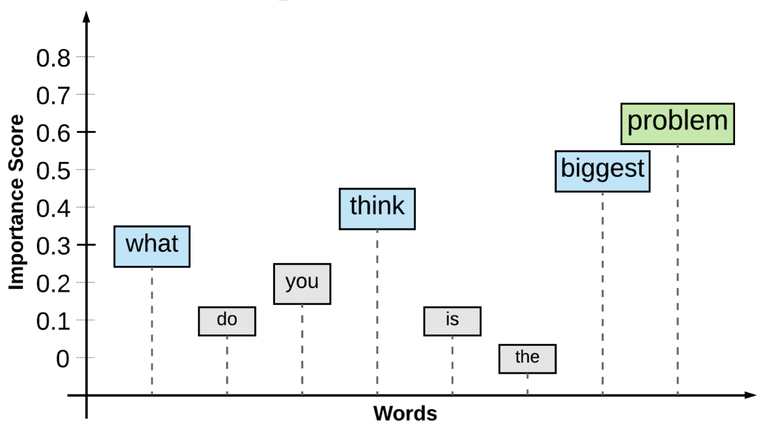

Fig 2: Importance scores assigned to words in an example sentence by human annotators on our project.

Feature Representation

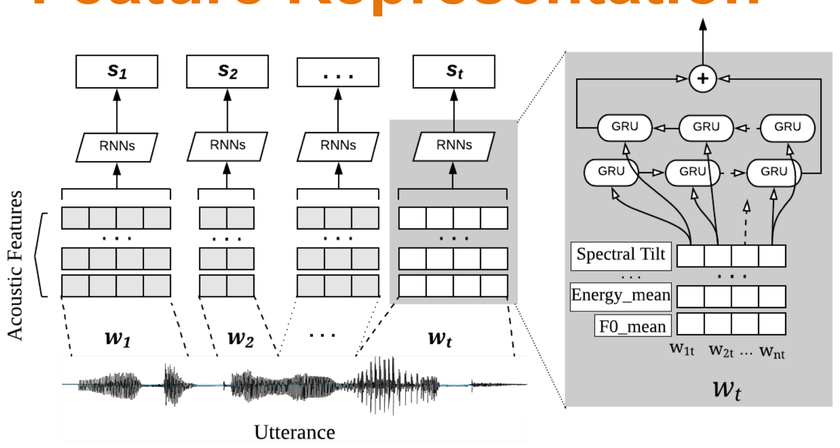

Fig 3: Architecture for feature extraction from words in a time-series speech data.

Approach

For each word-segment in an utterance, a fixed-length (10ms) interval window sliding through the word is used to obtain smaller interval-segments. We examine four categories of features, including: **pitch-related features** (10), **energy features** (11), **voicing features** (3) and **lexical features** (6). Using the neural architecture in Fig 3, a word-level feature representation is learned based on the prosodic features extracted from the interval sequences. Finally, another bi-directional RNN network based on Long-Short Term Memory (LSTM) units, used for context modeling, makes the word-importance prediction utilizing the learned word-level features from the previous network.

This material is based upon work supported by a Google Faculty Research Award and by the National Technical Institute for the Deaf (NTID).

Results

Model	RMS	F1
LSTM-SOFTMAX	0.705	0.60
LSTM-ORD	**0.672**	**0.601**
LSTM-CRF	0.706	0.598

Performance of the speech-based models on the test data under different projection layers: softmax layer (LSTM-SOFTMAX), oridinal softmax layer (LSTM-ORD) and, conditional random field based layer (LSTM-CRF).

Model	RMS	F1
speech-based	0.672	0.601
text-based	0.598	0.73
+ WER: 0.18	0.621	0.658
+ WER: 0.28	0.677	0.59

Comparison of the speech-based model with the text-based models operating on transcripts from two ASR systems, with different levels of Word Error Rate (WER): Watson Speech-to-Text(with WER=0.18 on our test data set) and Google Cloud Speech (with WER=0.28).

Conclusions

Using acoustic features in spoken dialogues, we evaluated neural architectures for the task of identifying the importance of individual words to the overall meaning of a dialogue. Specifically, we evaluated an importance labeler for this classification task. Our model performed competitively against state-of-the-art text-based word-importance prediction models, especially against models operating on imperfect Automatic Speech Recognition output.