INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

R&D SHOWCASE 2020

Exploring the role of pitch in predicting clause and sentence boundaries

Abstract

Pitch and other prosodic features are lost if NLP tools only use raw text as input data. Conversational speech is filled with prosody and it exists in language for a reason, so it is only natural that we make use of it to improve tools in language technology, such as Speech to Speech Machine Translation. Getting information from pitch is a complex process as pitch information is incredibly noisy. So, we use the prosogram[1] to simplify pitch contours into nuclei of salient pitch areas. Then we give these nuclei Low, Mid or High labels and write rules to predict clause and sentence boundaries based on the changing trend of these pitch labels. Rules were written based on our knowledge of pitch patterns in speech and aren't tailored to any particular speaker or language.

Objective

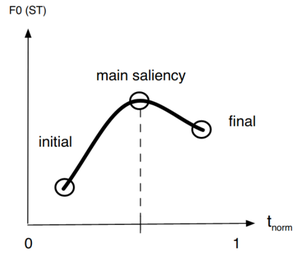
Almost all translation systems today accept text as input data. While this gives more than satisfactory results for several domains and types of documents, once we start to translate speech, we lose valuable prosodic information if we just translate raw text. Current Automatic Speech Recognition (ASR) systems output text without any punctuation or boundary information and Machine Translation gives subpar results on this text. Adding clause and sentence boundary information before translating this text can help improve it significantly. One way to look at it is that this is an effort to enrich raw text with useful information by translating speech features into graphical features, which in turn help Machine Translation systems to disambiguate the input adequately.

ASR Output: you should wait until the movie starts to eat your popcorn
Google Translate (Eng-Hin): आपको तब तक इंतजार करना चाहिए जब तक कि फिल्म आपके पॉपकॉर्न को खाना शुरू न कर दे

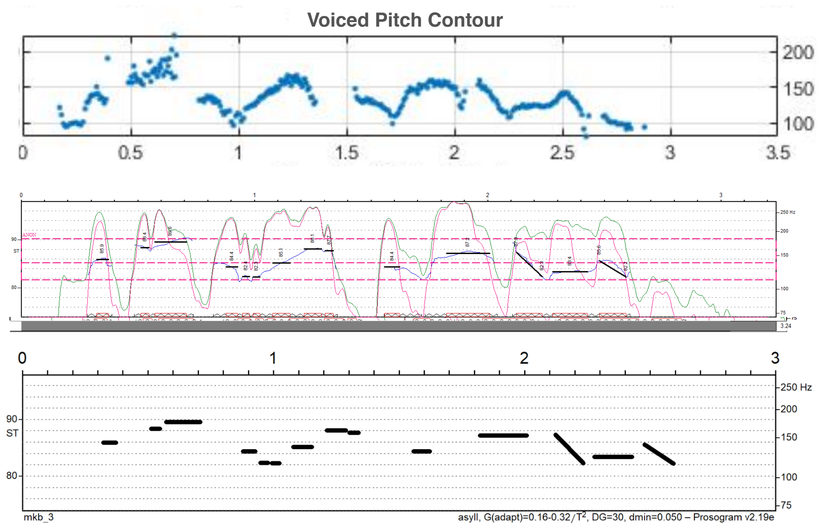
ASR Output(With clause boundaries): you should wait until the movie starts, to eat your popcorn
Google Translate (Eng-Hin): फिल्म शुरू होने तक इंतजार करना चाहिए, अपने पॉपकॉर्न खाने के लिए

It's clear that clause and sentence boundary identification has several benefits, however this project is more about exploring the role of pitch in this task, as opposed to a more complete tool which would almost certainly include linguistic features and even other prosodic features such as pauses, speech rate, etc.

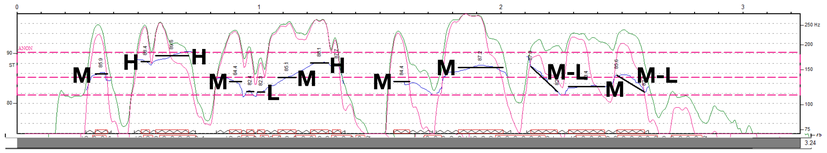
Method



With a wav file as input, the prosogram does **Speech Preprocessing,** i.e. the estimation of the fundamental frequency of speech (F0), and the segmentation of speech into units that are desired for the description of speech prosody. It also does **Acoustic Stylization** of the speech signal by representing the F0 variations that are considered as relevant. The F0 contour is represented by a set of 5 acoustic values - Initial, Final, Main Saliency, Main Saliency Position, and Local Register.



We then created a labelling scheme based on these F0 contours, which enabled us to write rules to predict clause and sentence boundaries. Each unit gets a Low, Mid, or High label based on its proximity to the minimum, maximum, and the mean frequency of the speaker.



The following rules have been written with the knowledge of pitch patterns in speech - a rising tone for commas and a falling tone for sentence boundaries. The change in trend is analysed by the system and the following rules are executed.

Trend Rules and Example Output

Comma
L -(+1)-> M -(-1)-> _L
L -(+1)-> H -(-1)-> _M/L
M -(+1)-> H -(-1)-> _M/L

Sentence Boundary
M -(-1)-> L -(+1)-> _M/H
H -(-1)-> M -(+1)-> _H
H -(-1)-> L -(+1)-> _M/H

mkb_3.wav

Time	Label	Trend	Prediction
0.318	M	0	
0.518	H	1	
0.573	H	1	
0.878	M	-1	Comma/Phrase Boundary
0.953	L	-1	
0.993	L	-1	
1.078	M	1	Sentence Boundary
1.213	H	1	
1.308	M	-1	Comma/Phrase Boundary
1.563	M	-1	
1.828	M	-1	
2.123	M-L	-1	
2.283	L	-1	
2.478	M-L	-1	

Preliminary Experiments

Preliminary Experiments on Hindi and English speech have shown promising results - far better than chance. However, due to the noisy nature of speech, there are several false positives. The next step for this project is to refine these rules.

Future Work

- **Emphasis Detection** using amplitude - reduce false positives.
- **Incorporation with ASR** using timestamps in ASR output.
- **Systematic Evaluation** and declaring precision and recall.
- **Adding more complex rules** combining linguistic features, such as Part-Of-Speech and prosodic features such as pauses.

References

1. Mertens, Piet. (2004). The prosogram: Semi-automatic transcription of prosody based on a tonal perception model.

Authors: Tanmai Khanna, Ganesh Mirishkar, Dipti M. Sharma, Anil K. Vuppala Research Center Name: LTRC