Science.
Applied to Life.™

M*Modal

Anindita(Dita) Nath

NLU(Speech) Intern

PhD.(CS) Candidate, University of Texas at El Paso

Utilizing Prosody to Predict Turns in Medical Conversations

Goal

- Build models using prosody to improve turn-taking in automatic medical transcriptions, eventually to build a better version of the existing transcriber(s).

Prosody In General

- Non-verbal features of speech, not what but *how* words are spoken
- Pitch, volume, speech duration, etc.
- Used to detect speech activities, turns, start or end of speech, emotion(s).

Project Background

- Develop system that automatically transcribes medical conversations to:
  - ✓ relieve physicians/caregivers from the documentation burden increasing their efficiency.
  - ✓ extend functionality to extract and summarize information.
- Existing speech recognizer:
  - ✓ Lower Frame Rate LSTM Acoustic Model:
    - ✓ LSTM based Voice Activity Detector
    - ✓ Speaker ID : LSTM based D-vector embedding for speaker voice

Motivation

- Human-Human conversations have overlapping speech, untimed pauses, etc., yet turn-taking happens spontaneously.
- Not all turns in the existing auto-transcripts are properly aligned or correctly detected.
- Difficult to predict/assign Speaker IDs to short turns.
- Time to think beyond just the *words*.
- Related literature shows that prosody :
  - ✓ is useful in predicting turns in dialogs at speech-onset or at pauses.
  - ✓ has universal utility across languages and speech genres like task-oriented face-to-face dialogs, general telephonic conversations, etc.

Related Works

- Ward et al.[2] used only prosody (*pitch, volume,* etc.) without any lexical annotation to build a general continuous TensorFlow LSTM RNN model to successfully predict:
  - ✓ whether a speaker will yield/hold turn after a pause and at the speech onset.
  - ✓ whether a speaker will continue speaking over a certain future window (0-250ms to 0-3s).

	after a 250ms pause			after a 500ms pause		
Instances	3,405	7,546		2,079	4,608	
% Hold	59.8%	58.8%		57.6%	57.6%	
Model	Skantze	Replica	Ours	Skantze	Replica	Ours
Shift: Precision	0.726	0.776	**0.784**	0.711	0.780	**0.800**
Shift: Recall	**0.703**	0.528	0.601	**0.738**	0.549	0.660
Shift: F-measure	**0.714**	0.628	0.680	**0.724**	0.644	0.720
Hold: Precision	**0.805**	0.730	0.759	**0.802**	0.727	0.778
Hold: Recall	0.822	**0.893**	0.884	0.780	**0.886**	0.879
Hold: F -measure	0.813	0.803	**0.817**	0.791	0.799	**0.825**

Hold/Shift Prediction Results-English Maptask

Prediction Window	English MAE	English % reduction	Japanese MAE	Japanese % reduction
0 to 250ms	0.11	67%	0.21	38%
0 to 500ms	0.19	42%	0.30	12%
0 to 1s	0.27	18%	0.36	-6%
0 to 2s	0.34	-3%	0.40	-18%
0 to 3s	0.36	-9%	0.41	-21%
baseline	0.33		0.34	

Mean Absolute Error- English and Japanese MapTask

	% reduction				
Prediction Window	American English	Japanese	Mandarin	Spanish	Canadian French
0 to 250ms	46%	44%	53%	51%	51%
0 to 500ms	28%	23%	36%	34%	35%
0 to 1s	14%	3%	22%	20%	19%
0 to 2s	4%	-5%	9%	7%	7%
0 to 3s	0%	-8%	7%	2%	5%
baseline	0.39	0.34	0.45	0.41	0.43

Mean Absolute Error-Telephone Corpora

Data

- Audio Set:
  - ✓ Recorded face-to-face medical conversations
  - ✓ Not professional quality
  - ✓ Approx. 36k dialogs (Doctor-Patient)
  - ✓ Remaining data is either doctor dictations (single speaker) or multi-party conversations (Doctor-Patient-Caregiver).
  - ✓ Typical duration is 8-10 minutes, sometimes as long as 45-60mins.
  - ✓ Mono (single channel), .wav audios
- Transcripts:
  - ✓ .stm files
  - ✓ 1 for each audio
  - ✓ Long gaps: mostly *not silence*

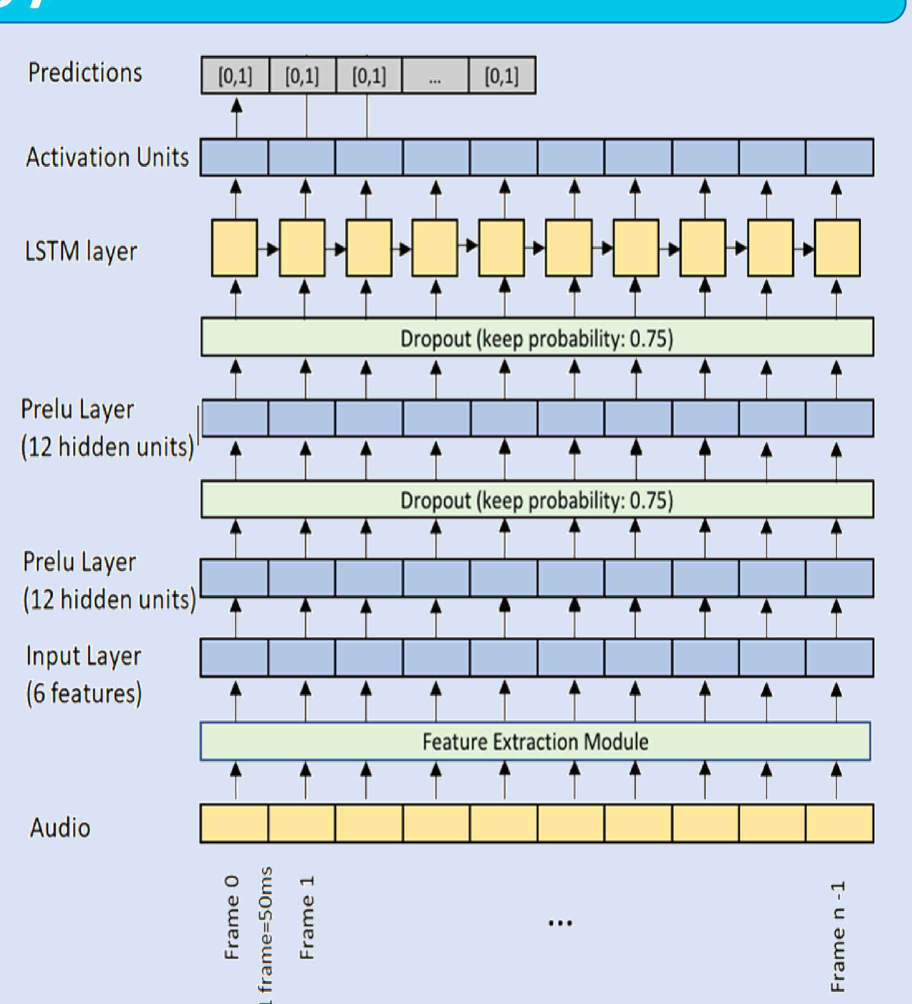
```
WAV.all/100002.wav DR 100002 0.130612 18.5959 all right and then
we close it it is working and we ignore it all right so we were
saying that your Alc is a little higher UH UH this UH this month
UH specifically it's in
WAV.all/100002.wav DR 100002 29.5306 30.8694 days ago from Quest
WAV.all/100002.wav PT 100002 30.902 30.9673 yeah
WAV.all/100002.wav DR 100002 32.0776 35.2939 the Alc was eight
point three usually it's in the 7s
WAV.all/100002.wav PT 100002 35.4898 35.7184 right
```

Feature Set

- 6 low-level extracted prosodic features:
  - ✓ Absolute pitch (log Hz)
  - ✓ Relative pitch (z-normalized pitch)
  - ✓ Volume (energy in dB)
  - ✓ Cepstral Flux
  - ✓ Speaking frame (from pitch)
  - ✓ Speaking frame (from energy)
- Features are extracted for every 50ms window across the entire audio.
- Pitch Tracker: YAAPT algorithm[4]
  Cepstral Flux : James Lyons's Mel Frequency Cepstral Coefficient implementation[3]
- Future set will be extended to include mid-level prosodic features from UTEP's Mid-Level Toolkit[3].

Key Methodology

- Adapted model[2] :
  - ✓ Labels: Turn-Ends as 1, rest 0
  - ✓ Feature Vector: 1min=1200 frames, each 50ms
  - ✓ Prediction Window: 1 (i.e. 50ms, next time step)
  - ✓ Predict turn-ends
  - ✓ Model: Many-to-Many LSTM RNN
  - ✓ Implementation: TensorFlow
  - ✓ Programming Language: Python

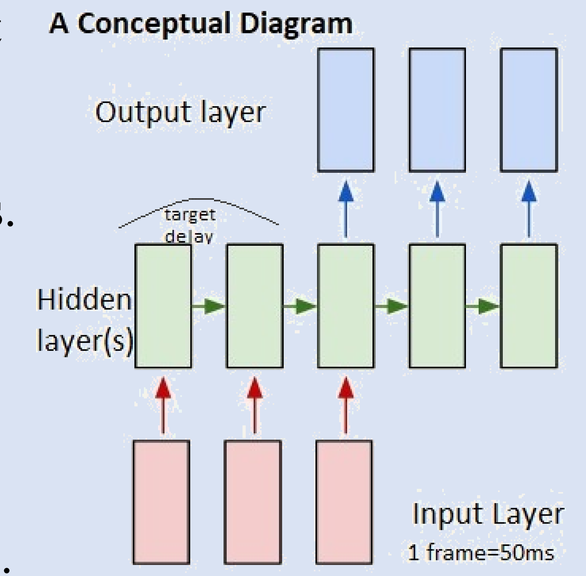


Network Architechture

Variations

- Ongoing Work (Baseline Prosody Models):
  - I. Introduce target delay, future context
  - II. Extend feature vector to include features from the present frame, as well as the past and future frames.
  - III. Include mid-level features

Many-to-Many LSTM with Target Delay: A Conceptual Diagram



- Future Plans:
  Compare and combine performance of :
  - I. Prosody models
  - II. Speaker ID models
  - III. Language Model based turn predictions

References

[1] Chung-Cheng Chiu et al. "Speech recognition for medical conversations," Interspeech,2018.
[2] Ward et al., "Turn-taking Predictions across Languages and Genres using an LSTM Recurrent Neural Network," IEEE Spoken Language Technology Workshop (SLT), 2018
[3] Mid-level Toolkit(Python Version), https://github.com/anath2110/prosodyMonsterPython.git
[4] Stephen A. Zahorian and Hongbing Hu, "A spectral/temporal method for robust fundamental frequency tracking," J. Acoust. Soc. Am. 123(6), June 2008

3M Confidential