Utilizing Prosody To Improve Turn Detection In Medical Conversations

Anindita Nath

NLU(Speech) Intern, 3M, Pittsburgh

PhD. Student (CS), UTEP

Friday, August 30, 2019

Company Background

Science. Applied to Life.™

Minnesota Mining and Manufacturing Company

- multinational conglomerate corporation
- Maplewood, Minnesota, United States, 1902

Science is at the heart of everything we do.

At 3M, science is applied in collaborative ways to improve lives daily.

- Business Groups
  - Safety & Industrial
  - Transportation & Electronics
  - Health Care
  - Consumer

Company Background [..contd.]

- Health Care Divisions
  - Bio Processing
  - Drug Delivery Systems
  - Food Safety
  - Health Information Systems
    - Pittsburgh Office, merger with MModal
  - Medical
  - Medical Device Components
  - Oral Care
  - Personal Health Care

Project Goal

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

3M

mModal

Project Background

- Big picture –

  Develop system(s) that automatically maintains Electronic Health Records so as to:

  - relieve physicians/caregivers from the documentation burden increasing their efficiency.

- Current EHR product, FLUENCY Direct -
  - A single, cloud-hosted voice profile allows clinicians to dictate into their EHR from anywhere, any device, and any care setting.
  - Front-end speech recognition solution.
  - Even dictation takes additional doctors' time, though much reduced than data entry.

3M

mModal

Project Background [..contd.]

- Next Step-

  Develop system(s) that:

  - eliminates any sort of documentation overhead.
  - automatically transcribes real-life medical conversations in real-time.
  - extracts and summarizes useful Patient Health Information.

- Current State-

  Transcribes recorded medical conversations using following in-house speech recognizers:

  - Lower Frame Rate LSTM Acoustic model that uses LSTM based Voice Activity Detector.
  - Speaker ID model that uses LSTM based D-vector embedding for speaker voice recognition and hybrid re-alignment.

Chung-Cheng Chiu et al. "Speech recognition for medical conversations," Interspeech,2018.

3M mModal

Shortcomings

```
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 A1c 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 A1c was eight
point three usually it's in the 7s
WAV.all/100002.wav PT 100002 35.4898 35.7184 right
```

Issues/Flaws:

- Not all turns in the existing auto-transcripts are properly aligned or correctly detected.
- Difficult to predict/assign Speaker IDs to short turns.
- Long gaps: mostly not silence.

3M

mModal

Motivation

Why Prosody?

- 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.

3M

mModal

Related Work

Ward et al. 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

Prediction Window	% reduction: 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

Ward et al., "Turn-taking Predictions across Languages and Genres using an LSTM Recurrent Neural Network," IEEE Spoken Language Technology Workshop (SLT), 2018

Data

Audio Set:

- Recorded face-to-face medical conversations
- Not professional quality
- Approx. 36k dialogs (Doctor-Patient)
  [Dev: 1k, Test:1k, rest-Train but each train batch was of 7k only]
- 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
- generated by speech recognizers, not humans

Features

- 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)
- 64 dimensions Log Filter Bank Energy[1].
- Features are extracted for every 10ms window across the entire audio.
- Pitch Tracker: YAAPT algorithm.[2]
- Cepstral Flux/LFBE : James Lyons's implementation.

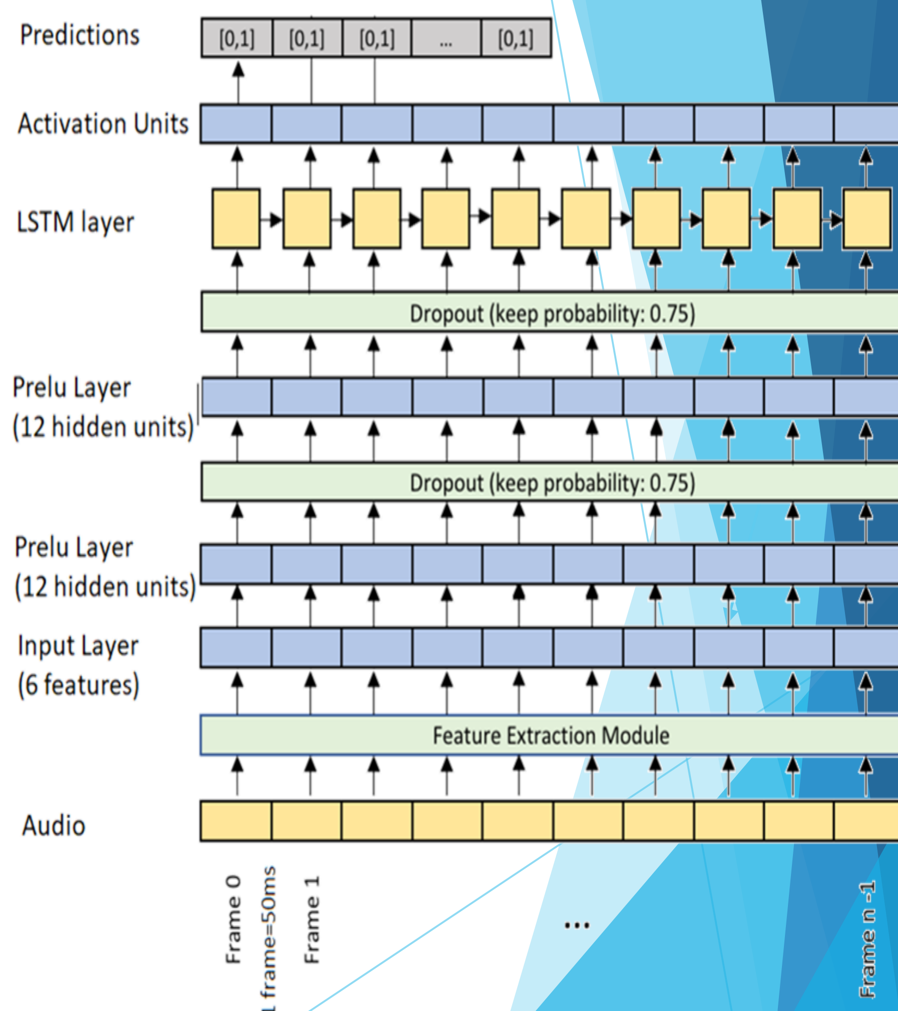
1. Zeynab Raeesy et. al. LSTM-based Whisper Detection, Amazon Alexa
2. Stephen A. Zahorian and Hongbing Hu, "A spectral/temporal method for robust fundamental frequency tracking," J. Acoust. Soc. Am. 123(6), June 2008
3. https://github.com/anath2110/prosodyMonsterPython.git

3M

mModal

Key Methodology

- Labels: Turn-Ends as 1, Speech Zones as 0, rest (unknown zones) as -1
- Feature Vector: 1200 frames, each 10ms
- Prediction Window: 1 frame (i.e. next 10ms)
- Model: Many-to-Many LSTM RNN
- Activation Function: Sigmoid
- Loss Function: Cross-Entropy
- Mass the unknowns while training
- Implementation: TensorFlow with Python



Network Architechture

Variations

- Uni-directional LSTM RNN , 2 hidden PRelu layers of 30 hidden units and a single LSTM layer of 30 units, trained with 6 low-level prosody features.
- Uni-directional LSTM RNN , 2 hidden PRelu layers of 96 hidden units and a single LSTM layer of 96 units, trained with 64 LFBE features.
- Uni-directional LSTM RNN , 2 hidden PRelu layers of 30 hidden units and a single LSTM layer of 30 units, trained with extended feature vector that includes features from the present frame(10ms) concatenated with corresponding 10 past and 10 future frames.

3M

mModal

Variations

- Bi-directional LSTM RNN , 2 hidden PRelu layers of 30 hidden units and a single LSTM layer of 30 units, trained with 6 low-level prosody features.
- Bi-directional LSTM RNN , 2 hidden PRelu layers of 96 hidden units and a single LSTM layer of 96 units, trained with 64 LFBE features.
- Bi-directional LSTM RNN , 2 hidden PRelu layers of 30 hidden units and a single LSTM layer of 30 units, trained with extended feature vector that includes features from the present frame(10ms) concatenated with corresponding 10 past and 10 future frames.

Results

- Models trained with LFBE features performed worse than low-level prosody features.
- Bi-directional LSTM RNN models performed better than LSTM RNN models.
- Differences in loss non-conclusive.

Future Plans

- Train models on the entire train set of dialogs and for more epochs or lower learning rate.
- Use mid-level features as well.
- Evaluate against human transcriptions.
- Compare and combine performance of :
  - Prosody models
  - Speaker ID models
  - Language Model based turn predictions

Challenges -> Solutions

Confidentiality Issues:

Working with confidential Patient Health Information data-

- prevents transfer of data from its confidential remote Linux storage.
- makes it difficult to listen to audios for spot-checking, failure analysis, etc.
- required me to undertake HIPAA training and obtain additional authorization that delayed start of project for nearly a month.

Work-arounds:

- Transfer few audios to local Windows storage, listen and delete immediately.
- Wait until properly authorized and trained.

Challenges -> Solutions

Software Installation Issues:

Working in remote LINUX server as a non-root user-

- makes it difficult to install various packages, e.g. installing Python IDEs, debuggers, media-players, building TensorFlow from source using bazel.

Work-arounds:

- Use Python command-line interpreter.
- Use Vi-editors to debug.
- Open Jupyter notebook installed in your remote server from your local host browser.

License Issues:

- MATLAB's GNU GPL conflicted with corporate license laws.

Work-arounds:

- Translate and modify to Python.

Challenges -> Solutions

Data issues:

Working with (extracting prosody features from and training deep-learning models on) excessively large amount (few TBs) of speech data resulted in:

- Hard-disk storage overheads.
- Memory (RAM) error.
- Slow computation.

Work Arounds:

- Use additional disk storage provided(3 TB).
- Delete the unknown labels and corresponding features from the respective arrays instead of masking.
- Use parallelisation
- Use GPUs

Perks

- Professionally enriching experience.
- Hands-on knowledge of using most of deep-learning techniques with real-life data.
- Co-operative supervisor and team.
- Meet with and showcase my work to the international team from UK and Germany.
- Expand my LinkedIn network.
- Boost in confidence.
- Possible new ideas for my research.

- Escape Room activities.
- One-day trip passes.
- Free Lunch every Wednesdays and Fridays.
- Visit to Minnesota (3M HQs).
- Tickets to my first ever Baseball game.

Thank you for your attention!

Questions?

Comments?

Suggestions?