# Diarized Speech Recognition System for Patient-doctor Communication

Feng Wang, Zachary Bachrach

# Introduction:

**Focus:** tele-conferencing for patient-doctor appointments

Virtual appointments increasingly common (COVID-19).

- Lost ability to detect body language.
- New ability to process speech.

**Need:** efficient and engaged communication, optimize virtual medical appointments
**Goal:** an interface for patients and doctors

# MVP

**An interface**

- where the user can input an audio recording of a doctor’s appointment

**Output**

- the diarized transcript
- highlighted important information
  - physician statements and patient responses

# User stories

**For doctors:**

I want to **extract maximum information** from virtual patient appointments to **diagnose my patient accurately**.

**For Patients:**

I want a **clear written set of instructions** from my doctor.

As a patient, I want **to specify a list of topics** that I'd like to cover in the appointment. I want to **be alerted if we forget to discuss those topics**.

# User stories (cont.)

**For both entities:**

I want **a transcript of the appointment**

- for **research purposes** in the medical community.
- with important highlights for other medical professionals to **reference** before seeing the patient.

# Long-term Product Goals

Split input audio into patient and doctor transcript

Perform NLP on doctor transcript to identify instructions for the patient

Perform NLP on patient transcript to detect/diagnose (Alzheimer’s, deception, etc.)

Perform NLP on patient transcript to highlight important information (patient responses)

Perform NLP on total transcript to make sure topics were covered

# NLP to detect Alzheimer's

## Lexical changes

| Marker | Dementia | Healthy aging |
|---|---|---|
| Vocabulary size | Sharp decrease | Gradual increase, then possible slight decrease |
| Lexical repetition | Pronounced increase | Possible small change |
| Word specificity | Pronounced decrease | Possible small change |
| Word class distribution | Fewer nouns, compensation in verbs | No change |
| Fillers | Pronounced increase | Possible slight increase |

## Syntactic changes

| Marker | Dementia | Healthy aging |
|---|---|---|
| Syntactic complexity | Sharp decline | Little or no change, then possible rapid decline in mid-70s |
| Use of passive voice | Pronounced decrease | Possible small decrease |
| Auxiliary verb in passive voice | *Get* dominates *be* | *Be* dominates *get* |
| Passives without agent | Greater decrease | Moderate decrease |

Lexical and syntactic changes are present in Alzheimer's affected individuals[1].

# NLP to detect deception[2]

Linguistic Inquiry and Word Count (LIWC)- creates a profile of text

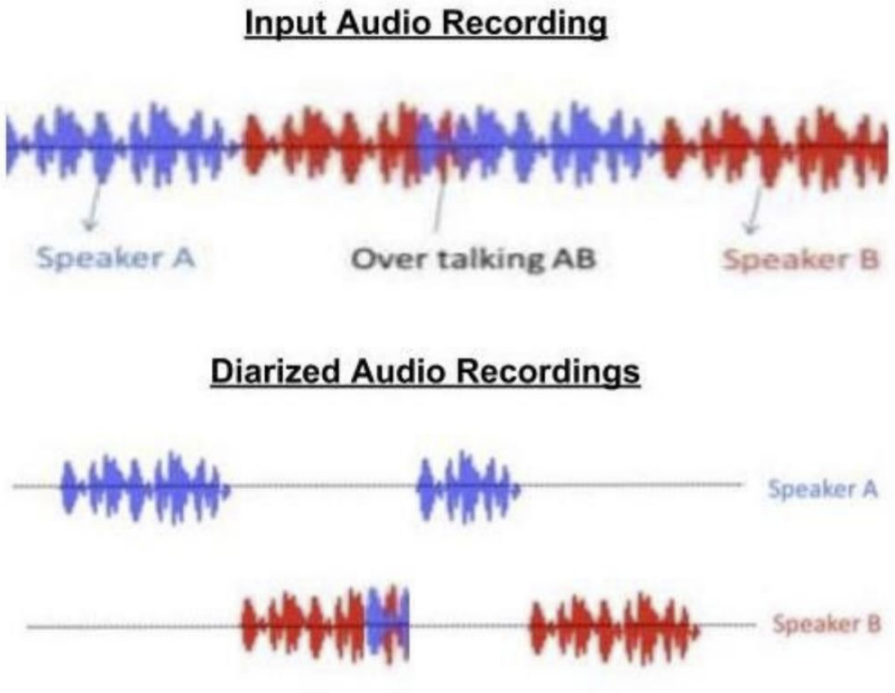
Input: I really try to exercise every day. It depends, I mean I try to run and stuff but I am so busy usually.

Output:

| TRADITIONAL LIWC DIMENSION | YOUR DATA | AVERAGE FOR PERSONAL WRITING |
|---|---|---|
| I-WORDS (I, ME, MY) | 17.4 | 8.70 |
| SOCIAL WORDS | 0.0 | 8.69 |
| POSITIVE EMOTIONS | 0.0 | 2.57 |
| NEGATIVE EMOTIONS | 0.0 | 2.12 |
| COGNITIVE PROCESSES | 34.8 | 12.52 |
| SUMMARY VARIABLES | | |
| ANALYTIC | 2.6 | 44.88 |
| CLOUT | 1.0 | 37.02 |
| AUTHENTICITY | 99.0 | 76.01 |
| EMOTIONAL TONE | 25.8 | 38.60 |

# Speech Diarization

1. Detection
2. Segmentation
3. Embedding
4. Clustering
5. Transcription (ASR)




https://hackernoon.com/speaker-diarization-the-squad-way-2205e0accbda

# Speech Recognition systems (ASR)

ASR receives audio inputs and output text accordingly

1. Signal acquisition
2. Feature extraction
3. Acoustic modelling
4. Language & lexical modelling
5. Recognition of words

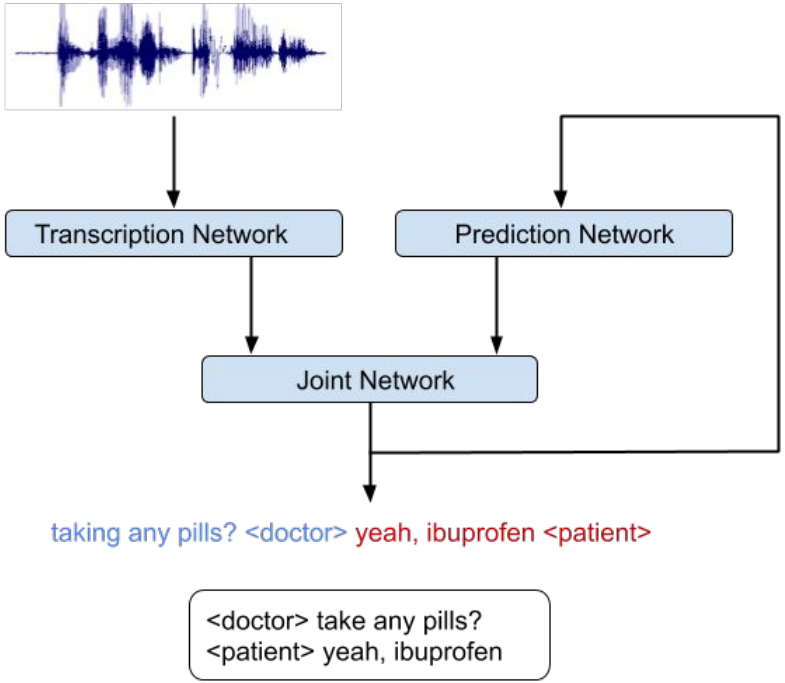
# Diarization Framework Option 1

Resemblyzer to perform segmentation and embedding extraction on input audio data file (diarization)

Google Text-to-speech API to perform audio-to-text conversion (good medical option, but PAID)

Python “speechrecognition” package to perform audio-to-text conversion

# Diarization Framework Option 2: Recurrent Neural Network Transducer (RNN-T))[3]




This RNN-T integrates 3 networks

- A transcription network
- Prediction network for next target label given prior labels
- Joint network to combine results and output probability distribution

# Total Framework

Combine speech diarization and speech recognition *(1)*

Combine *(1)* with tools outlined in user stories to achieve better patient doctor engagement during remote appointment.

# Citations

[1] Graeme, H., Li, X., Lancashire, L., & Jokel, R. (n.d.). Natural language processing methods for the detection of symptoms of Alzheimer's disease in writing. Retrieved from http://www.cs.toronto.edu/pub/gh/Google-talk.pdf

[2] Poesio, M., & Fornaciari, T. (n.d.). Detecting deception in text using NLP methods. Retrieved October 4, 2020, from https://research.signal-ai.com/assets/Deception_Detection_with_NLP.pdf

[3] Shafey, Laurent El, Soltau, Hagen, and Shafran, Izhak. "Joint Speech Recognition and Speaker Diarization via Sequence Transduction." (2019). Web.

Thank you

# Web app

Microsoft Azure platform (cloud hosting)

# Julius

Perform a large vocabulary continuous speech recognition (LVCSR)

Real-time processing

Versatile

Scalable

Free-open source in C.

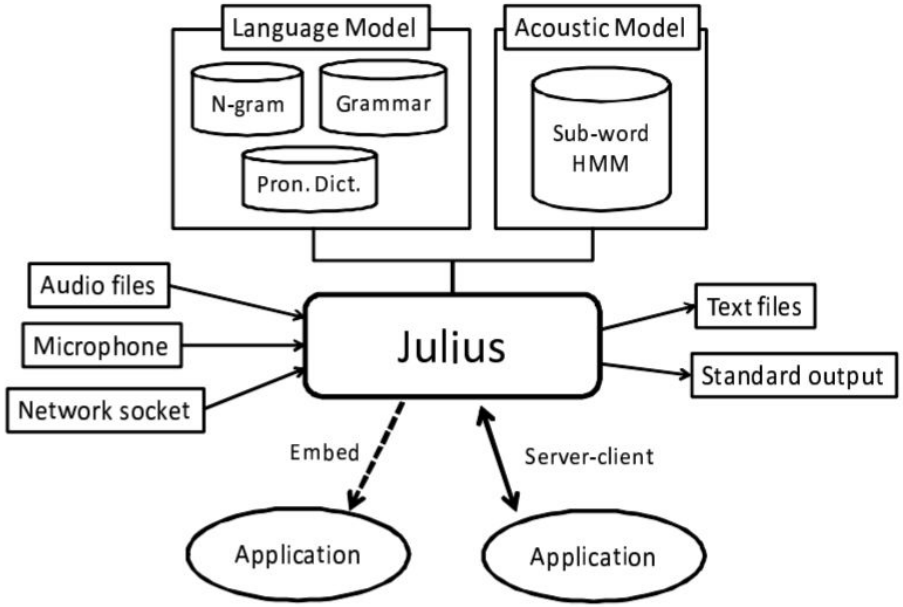



Fig. 1. Overview of Julius.