# Optimized Diarized Speech Recognition System for Patient Doctor Communication

**Feng Wang, Zachary Bachrach**

## ABSTRACT

**More than ever before, doctor's appointments are held virtually through phone calls and video conferencing. The current virtual system has much room for improvement to the patient-doctor relationship. This paper discusses the ways to approach improving the patient-doctor relationship, and improving mental illness diagnosis given the new paradigm.**

## 1. Introduction

Due to COVID-19, almost all doctor appointments start from a phone call examination. This system can be improved to facilitate better patient and doctor communication.

For doctors, an improvement could be clearly conveying the doctor's order which they want the patient to follow strictly. For patients, this could be a trace of nervousness indicating hiding information from the doctor, or early/accurate detection of mental or physical illness.

## 2. Linguistic Cues in Medical Field

Meeting in-person allows one to accurately assess a person's body language. For instance, in-person, it can be easy to tell if someone is hiding information based on their body language. While in-person meetings are much less frequent today, it is still important to detect when someone is hiding information. In Natural Language Processing (NLP), a stylometric approach to feature extraction has been successful in the past [29]. Stylometry studies text based on stylistic features only, and the analyses belong to two main families: surface and lexical features. Surface features include word frequency without regard to meaning. Lexical features try to capture meaning. The Linguistic Inquiry and Word Count (LIWC) is a well known lexical resource used for deception detection. A text can be analyzed and compared to the LIWC dictionary, producing a profile of the text to classify statements made to be deceptive or truthful.

Additionally, because of the virtual platform, it is now possible to perform analysis on speech to detect mental (and possibly physical) illness. More specifically, there has been work showing that early detection of Alzheimer's disease. Alzheimer's disease is a progressive disease that destroys memory and other important mental functions. A University of Toronto presentation [28] discussed the possibilities of using NLP for detecting Alzheimer's symptoms using writing samples. Lexical changes that someone affected by Alzheimer's include a sharp decrease in vocabulary size, lexical repetition increase, decrease in word specificity, fewer nouns and more verbs, and an increased use of filler words. Healthy aging generally exhibits far lower magnitude of change across most categories. These ideas are shown in Table 1. Table 2 shows syntactic changes in a person affected vs a person who is experiencing healthy aging.

## 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 |

Table 1&2 (left and right). Lexical changes and syntactic changes in Alzheimer's vs healthy aging people

### 3. Main Approaches for speech diarization

Speech diarization aims to segment the inputted speech while eliminating non-speech fragments; extract speaker identification factors from speech segments; identify each segments' corresponding speaker identities; and group segmentation according to correlated speakers [1].

Each speaker has their own unique set of characteristics in speaking, such as accent, rhythm, intonation [2]. As a result, speech signals have become a popular biometric technique [2]. There are several features, which are extracted from audio utterance utilized for identifying individual speakers, such as Mel Frequency Cepstral Coefficients (MFCC), Speaker factor, i-vector, Perceptual Linear Predictive (PLP), Neural Predictive Coding (NPC), and so on [3,4].

While these features provide potential future direction in the field. Other characteristics in a daily speech presented challenges for speech diarization. One of the key challenges in the field is speech overlapping [5,6,7,8]. In daily conversations and meetings, ranging from 8% to 17% of the speech are overlapped [6]. These overlapping speech could be deleted during segmentation or misidentified, leading to significant diarization error and lowers the accuracy of the system by 17% [5]. Another challenge is background non-speech segments. Non-speech sounds; including music, loud background chattering, animal sound; need to be removed during diarization and can introduce errors in the system. Further, for a large population, it is unrealistic to train the system with all speakers' voices for identification as a close-set. The algorithm needs to be able to identify speakers who are not in the training set/system library (open-set).

#### a. Signaling processing

Signaling processing has been applied in speech diarization for more than two decades. Conventional speech diarization is simplified into two stages. The first stage detects changes in the acoustic spectrum and marks those changes. The second stage identifies speakers for each acoustic spectrum [9]. One of the earliest approaches uses symmetric Kullback-Liebler (KL2) which is calculated over the 12th order PLP coefficient for the acoustic similarity between audio segments with lengths of 0.5 sec, 1.0 sec, and 4 sec [9]. This creates feature vectors that are then used for identifying non-speech and

speech, and speaker clustering using the generalized likelihood ratio (BIC) [9]. This method is accurate and uses a relatively small amount of computational resources compared with other methods [9]. However, it deals with a simple broadcast scenario which is quite limiting when applied to real-life meetings [9]. In addition, since the audio segment length is preselected, if two persons talk consecutively to each other in one speech segment, this method would fail. Further, it does not take account of overlapping speech.

The probabilistic framework of expectation propagation (EP) is developed for finding accurate stopping points for conversations [10]. Individual speakers have unique identities modeled using Mel-frequency cepstral coefficient (MFCC) [10], which is a bassline acoustic feature set for speech in the 'Mel' scale [4] (1).

$$f_{mel} = 2595 log_{10}(1 + \frac{f}{700}) \quad (1)$$

These unique identities allow EP to choose appropriated joint densities for margin calculation and achieve speaker diarization [10]. Another alternative approach is an adapted Gaussian mixture model, a universal background model (GMM-UBM) system on the frame to frame basis for likelihood ratio computing [11]. The system first extracts input speech features and computes against the hypothesized speaker along with speaker-independent alternatives to calculate the likelihood ratio [11]. GMM-UBM then segments the audio through either internal segmentation or external segmentation [11]. While improved upon KL2, both systems still have several constraints. The most significant constraint is that for both systems each speaker is given unique identities in close-set. As a result, both approaches present a challenge for its application on a large population, when extracting individual acoustic features is too cumbersome. Further, the issue of overlapping speech is also not addressed in this approach. It appears that machine learning needs to be adopted to achieve open-set and overlapping speech diarization.

**b. Machine Learning technique**

I-vector is an audio feature that is widely used in speaker verification [12]. In detail, i-vector is the representation of a speech utterance in total variability space which is a low-dimensional and channel-dependent space [12]. Methods, such as cosine scoring, Variational Bayesian GMMs, mean shift, are explored to utilize i-vector for speaker diarization [13]. However, these approaches are costly and cumbersome [13]. D-vector is explored as a more efficient and accurate replacement for i-vector. D-vector is an average of speaker features which is extracted when deep neural networks are trained to group speakers [14]. D-vector is widely applied in speaker verification but is often dependent on key words detection. By adding a scoring function, the d-vector can directly perform diarization. An alternative approach is named LSTM which achieves speaker diarization by using a text-independent d-vector in combination with nonparametric spectral clustering algorithm [3]. However, the clustering algorithm in majority of these approaches constrains the algorism to close set training data [1]. This restrains the algorism in open-set classification. To solve this restriction, the LSTM system is further improved to unbounded interleaved-state recurrent neural network (UIS-RNN) [1] through an improved clustering algorithm. In UIS-RNN, RNN is an online generative process, which learns speaker information automatically [1]. This approach achieved a diarization error of 7.6% in testing [1] and is introduced by google AI as one of the most accurate open-source voice diarization systems [15].

An issue that is not specified in the approaches above is overlapping speech, which leads to significant diarization errors [5]. To address this issue, a system named overlapping automatic speaker

identification (OSID) is investigated [5]. Unlike previous studies, including the Gaussian Mixture Model (GMM) based system [8], multi-class Vector Taylor series (MC-VTC) system, OSID can identify more than two simultaneous speakers in the overlapping audio [5]. In the two-stage OSID (T-OSID) approach, the system first identifies the number of speakers in the speech segments, then matches the speech segments with all possible combinations with the specific number of speakers presented [5]. Direct speaker identification is achieved through a single neural network classifier in one stage OSID (S-OSID) [5]. An alternative approach is serialized output training with attention-based encoder-decoder (SOT-AED) [7]. Even though both systems can achieve significant success in identifying overlapping speakers [5,7]. It has several pitfalls for future improvements. Mainly, the system is less accurate with a high overlapping energy ratio (OER) and is solved in close-set classification [5,7].

## 4. Main Approaches for speech recognition

Speech Recognition systems (ASR) receives audio inputs and output text accordingly [23]. The general process of ASR follows such workflow: acquisition of inputted speech signals, extracting features from input signals, acoustic modelling based on selected approach, language & lexical modelling and recognition of words [24]. There are several approaches to tackle the problem, namely acoustic-phonetic approach which uses acoustic properties to characteristic distinctive phonetic units in languages; pattern recognition approach which utilizes pattern recognition in mathematical models and trains the algorithm to establish correlation with pre-labelled data [24]. Additional approaches include using Artificial Intelligence(AI) to simulate how humans identify speech (Knowledge based approach); optimizing a network of computing units for storing knowledge and constraints (Connectionist Approach); adapting linear and non-linear classifiers to classify speech (Support Vector Machines)[24].

There are many factors which provided challenges for ASR systems performance. The first one is speech sounds production quality [25]. Psychological traits, speaking habits and signal collection channel, can introduce variabilities and confusion in the system [25]. Further, environmental conditions can result in distortion of sound signals and introduce error[25]. Acoustic manifestations from preceding and following sound in nature utterance requires special attention to spontaneous speech recognition in training [25]. Additional factors include but are not limited to variation in language, accent, ambiguity in language [26].

## 5. CURRENT STATE OF ART OPEN SOURCE SYSTEM

### a. Speech Diarization

In 2018, an unbounded interleaved-state recurrent neural network (UIS-RNN) was introduced through google AI blog and is the most accurate speech diarization system in open source with a 7.6% diarization error rate [15]. By using the UIS-RNN approach, the system can identify new users and associate them with corresponding speech segments [15]. While traditionally, the open set system has lower accuracy, UIS-RNN can utilize the speaker labels and propagate the labeled data through time [15]. This is the major advantage of the system since it can identify new speakers without compensated accuracy rating. This allows the system to have major applications in patient-doctor conversations, due to it's accurate diarization in open-set [15].

### b. Speech recognition

The three most well known systems currently in the ASR field are IBM watson, google cloud speech-to-text, and Wit [23]. Among three, Google cloud speech-to-text runs deep learning neural

network algorithms with superior accuracy compared with other systems [23]. It is able to transcribe domain specific terms in real-time with domain-specific models [24]. By training it with context audios, it can have major applications in different scenarios, such as patient-doctor conversations. IBM Waston combines grammar and language structure into machine learning, and allows personalization of the program through adding out-of-vocabulary words [23]. This feature allows the user to input academic terms into the system to achieve personalization, such as inputting medical terms for patient doctor conversations. However, these two systems both require payments, which could be expensive, especially for individual developers when large amounts of audio data need to be analyzed. Wit on the other hand, is a free open-source application with copartible accuracy with IBM Waston. Another option with the ability to convert audio to text is the "speechrecognition" package for python, which is free and open-source.

## 6. Path Forward

We aim to build an interface where the user can input an audio recording of a doctor's appointment with 1 doctor and 1 patient in the appointment. This interface will then output the diarized transcript with highlighted important information, which will be different for doctors and patients and consist of doctor statements and patient responses. This application is based on the following user stories. Specifically, for patients, they want to receive a clear written set of instructions from the doctor based on the appointment; for doctors, they would want to extract maximum information from virtual patient appointments to diagnose accurately. For anyone in the medical community, it's useful to have a transcript of the appointment for research purposes and for other medical professionals to reference before seeing the same patient. Additionally, it is common for patients to forget what they wanted to ask the doctor. For this, we aim to have a feature for patients to enter a list of topics at the beginning of the meeting that they would like to cover. They system would then alert the meeting towards the end if they haven't discussed those topics.

By utilizing linguistic and acoustic cues, and monitoring voice characteristics such as pitch and loudness, the system would be able to highlight important information for participants. Through a recurrent neural network transducer (RNN-T), automatic speech recognition and speaker diarization can be incorporated together [19]. Traditionally, automatic speech recognition and speaker diarization are run parallel in the system [18]. Speech diarization breaks the segments using voice characteristics, and the deep learning model identifies the words in the speech [18]. The outputs from the two systems are then combined to produce a speech transcript with speaker labels [18]. However, this approach does not utilize the linguistic and acoustic cues in conversation [18]. By incorporating automatic speech recognition and speaker diarization, the system can take advantage of both linguistic and acoustic cues, which decreased word-level diarization error by almost 10 fold to 2.2% [19]. An RNN-T model is adopted in this approach with three networks incorporated, consisting of a transcription network that maps acoustic frames; a prediction network that uses the output from a joint network to predict preceding target labels; a joint network that combined the transcription and prediction network [18,19]. The prediction label allows the system to incorporate linguistic and audio cues [19]. While this approach has no recorded open-source algorism, several existing open-source algorithms, such as ALIZE, can achieve both automatic speech recognition and speaker diarization. Further research and testing are still needed in this topic, but I believe it has huge potential in future patient-doctor appointments.

## 7. Goals and Framework

This section serves to clearly define our product goals and framework direction. The project should perform the following functions: (1) Split input audio into patient and doctor transcript. (2) Perform NLP on doctor transcript to identify instructions for the patient. (3) Perform NLP on patient transcript to detect/diagnose (Alzheimer's, deception, etc.). (4) Perform NLP on patient transcript to highlight important information (patient responses). (5) Perform NLP on the total transcript to make sure topics were covered. As of now there are two potential frameworks we would use to approach diarized transcript creation. The first diarization framework is to use an open source tool such as Resemblyzer to perform the segmentation and embedding extraction steps of the diariation process, and then use an open source tool such as Google speech-to-text api for ACR. The second option is to use an RNN-T model to combine three networks, transcription, prediction, and joint networks.

From there, the total framework involves using the diarization framework to implement the tools described in the user stories.

## 8. Competition

There are several products serving aims to assist efficient communication between remote patient and doctor. However, none of these products adopt linguistic cues for uncovering hidden information. The first app is named CareCloud by CONTINUUM. This app allows voice communication between patients and doctors. In more detail, the app is able to record and transcribe conversation between doctors and patients, which allows the doctors to later send clips or transcriptions of important instructions to the patients [20]. However, the doctors have to identify which audio clips/transcription to send themselves, which requires extra time and attention. Another similar app is from a company named visualDx, which enables telemedicine visits between patients and doctors [21]. Different from CareCloud, after the telemedicine visit, the app will send the patients a handout from the VisualDX system to ensure accurate information and understanding [21]. While it is not clearly stated, the handouts appear to be pre-generated by the company, meaning they are not patient specific according to the telemedicine visit. A more patient specific solution is produced by a company named Nuance. The product has a highly accurate diarized speech recognition system, which enables real time recording of physician statements and patient responses [22]. These recorded information are then added to clinical notes for future references [22]. This product is very close to the aim of this project, as it's able to produce comprehensive clinical notes from conversations. However, it fails to use linguistic cues in the conversations for uncovering hidden information in the conversation. Further, these systems are not interactive with the patients and do not remind patients if they forgot to ask any questions which they might be interested in knowing answers to. As a result, we believe our product has its special niche in the market.

## 9. Github repository

This is the central repository for the project, which will contain important files for the project moving forward:
**https://github.com/zb8787/Medical-Meeting-Interface**

## 10. Reference

[1] Zhang, Aonan, Wang, Quan, Zhu, Zhenyao, Paisley, John, and Wang, Chong. "Fully Supervised Speaker Diarization." (2019): 6301-305.

[2] Sharma, Varun, Bansal, P K, A Review On Speaker Recognition Approaches And Challenges, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 5, May – 2013

[3] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno, "Speaker diarization with lstm," in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5239–5243.

[4] Hossan, M A, Memon, S, and Gregory, M A. "A Novel Approach for MFCC Feature Extraction." (2010): 1-5.

[5] Tran, Van-Thuan, and Tsai, Wei-Ho. "Speaker Identification in Multi-Talker Overlapping Speech Using Neural Networks." IEEE Access 8 (2020): 1. Web.

[6] E. Shriberg, A. Stolcke, and D. Baron, ‘‘Observations on overlap: Findings and implications for automatic processing of multi-party conversation,’’ in Proc. EUROSPEECH, Aalborg, Denmark, 2001, pp. 1359–1362.

[7] Kanda, Naoyuki, Gaur, Yashesh, Wang, Xiaofei, Meng, Zhong, Chen, Zhuo, Zhou, Tianyan, and Yoshioka, Takuya. "Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers." (2020). Web.

[8] Tsai, W. H., & Liao, S. J. (2010). Speaker Identification in Overlapping Speech. Journal of Information Science & Engineering, 26(5).

[9] Meinedo, H, and Neto, J. "Audio Segmentation, Classification and Clustering in a Broadcast News Task." 2 (2003): II-5. Web.

[10] Walsh, John MacLaren, Kim, Youngmoo E, and Doll, Travis M. "Joint Iterative Multi-Speaker Identification and Source Separation Using Expectation Propagation." (2007): 283-86. Web.

[11] Dunn, Robert B, Reynolds, Douglas A, and Quatieri, Thomas F. "Approaches to Speaker Detection and Tracking in Conversational Speech." Digital Signal Processing 10.1-3 (2000): 93-112. Web.

[12] Dehak, Najim, Kenny, Patrick J, Dehak, Réda, Dumouchel, Pierre, and Ouellet, Pierre. "Front-End Factor Analysis for Speaker Verification." IEEE Transactions on Audio, Speech, and Language Processing 19.4 (2011): 788-98.

[13] Garcia-Romero, Daniel, Snyder, David, Sell, Gregory, Povey, Daniel, and McCree, Alan. "Speaker Diarization Using Deep Neural Network Embeddings." (2017): 4930-934.Web.

[14] Variani, Ehsan, Xin Lei, McDermott, Erik, Lopez Moreno, Ignacio, and Gonzalez-Dominguez, Javier. "Deep Neural Networks for Small Footprint Text-dependent Speaker Verification." (2014): 4052-056. Web.

[15] Wang, C. (2018, November 12). Accurate Online Speaker Diarization with Supervised Learning. Retrieved September 10, 2020, from https://ai.googleblog.com/2018/11/accurate-online-speaker-diarization.html

[16] Kiktova, Eva, and Juhar, Jozef. "Comparison of Diarization Tools for Building Speaker Database." Advances in Electrical and Electronic Engineering 13.4 (2015): 314-19. Web.

[17] J. Watada and Hanayuki, "Speech Recognition in a Multi-speaker Environment by Using Hidden Markov Model and Mel-frequency Approach," 2016 Third International Conference on Computing Measurement Control and Sensor Network (CMCSN), Matsue, 2016, pp. 80-83.

[18] Shafey, L., El. (2019, August 16). Joint Speech Recognition and Speaker Diarization via Sequence Transduction. Retrieved September 11, 2020, from https://ai.googleblog.com/2019/08/joint-speech-recognition-and-speaker.html.

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

[20] Casarez, C. (2017, July 25). How Mobile Apps Connect Physicians and Patients -. Retrieved October 04, 2020, from https://www.carecloud.com/continuum/how-mobile-apps-connect-physicians-and-patients/

[21] Telehealth. (n.d.). Retrieved October 04, 2020, from https://www.visualdx.com/clinical-solutions/hospitals-health-systems/telehealth/

[22] Dragon Ambient eXperience Amplification. (n.d.). Retrieved October 04, 2020, from https://www.nuance.com/healthcare/campaign/ppc/dax-group-demos.html?cid=7010W000002T9kWQAS

[23] Filippidou, F., & Moussiades, L. (2020). A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I, 583, 73–82. https://doi.org/10.1007/978-3-030-49161-1_7

[24] Ghai, W.,; Singh, N. (2012). Literature Review on Automatic Speech Recognition. International Journal of Computer Applications, 41(8), 42-50. doi:10.5120/5565-7646

[25] Speech-to-Text: Automatic Speech Recognition  |  Google Cloud. (n.d.). Retrieved October 04, 2020, from https://cloud.google.com/speech-to-text

[26] Samudravijaya, K. (n.d.). Automatic Speech Recognition. Retrieved from https://pdfs.semanticscholar.org/eec8/eeb3686961bfcce83d9971cd5c2c6a8c5019.pdf

[27] Shneiderman, Ben. (2001). The Limits of Speech Recognition. Communications of the ACM. 43. 10.1145/348941.348990.

[29] 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

[30] 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