# Extracting and Using Speaker Role Information in Speech Processing Applications

by

Nikolaos Flemotomos

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

August 2022

*To my parents, Sofi and Antonis.*

# Acknowledgements

This dissertation marks the end of a journey; a journey that I could definitely not have done alone. First and foremost, I would like to thank my advisor, Professor Shrikanth Narayanan, for giving me the opportunity to join SAIL and embark on this journey. Also, I need to extend a big thank you to all my professors at USC who, through their classes, provided me the foundational tools needed for my research, as well as to all the staff, and especially Tanya Avecedo-Lam and Diane Demetras, who supported me and made sure the trip was as smooth as possible.

During this PhD journey, I had the chance to work with some great collaborators. A special thank you to Professors Panayiotis Georgiou, David Atkins, Zac Imel, and Torrey Creed who had a huge impact on my research agenda and on the way I learned to approach real-world problems and research questions. I also need to express my gratitude to all the professors who served as members in my qualification and/or dissertation committees, providing valuable advice and guidance: Keith Jenkins, Maja Matarić, C.-C. Jay Kuo, and Antonio Ortega.

To all my labmates and colleagues I met along the way, many of whom are now my friends. From study groups to music nights, and from writing papers to having fun at international conferences around the world, the SAIL family was always there, to make the journey both more productive and more enjoyable. But also to all the friends outside the lab, both the ones I made in this part of the world and the ones back in Greece. A special thank you to my good friends and roommates who made the journey so much more fun, especially during the weird times of the pandemic, and a big thank you to my girlfriend for her support, patience, and understanding.

Last, and definitely not least, a huge thank you to my family, who supported this journey in every possible way long before it even started. Mom, dad, Evelina, this was only possible because of you.

# Contents

# List of Tables

# List of Figures

# Abbreviations

**ADOS** Autism Diagnostic Observation Schedule.

**AM** Acoustic Model.

**ASR** Automatic Speech Recognition.

**BERT** Bidirectional Encoder Representations from Tranformers.

**BIC** Bayesian Information Criterion.

**CL** Cannot-Link.

**CMLLR** Constrained Maximum Likelihood Linear Regression.

**CNN** Convolutional Neural Network.

**CPTS** Counseling and Psychotherapy Transcripts Series.

**CRF** Conditional Random Field.

**DER** Diarization Error Rate.

**DNN** Deep Neural Network.

**E$^2$CP** Exhaustive and Efficient Constraint Propagation.

**FA** Facilitate.

**GI** Giving Information.

**GMM** Gaussian Mixture Model.

**HAC** Hierarchical Agglomerative Clustering.

**HMM** Hidden Markov Model.

**IRR** Inter-Rater Reliability.

**LDA** Linear Discriminant Analysis.

**LDC** Linguistic Data Consortium.

**LM** Language Model.

**LSTM** Long-Short Term Memory.

**MFCC** Mel Frequency Ceptral Coefficient.

**MI** Motivational Interviewing.

**MIA** Motivational Interviewing - Adherent.

**MISC** Motivational Interviewing Skill Code.

**ML** Must-Link.

**MR** Misclassification Rate.

**NER** Named Entity Recognition.

**NLP** Natural Language Processing.

**PDD** Pervasive Developmental Disorder.

**PLDA** Probabilistic Linear Discriminant Analysis.

**QUC** Closed Question.

**QUO** Open Question.

**RATS** Robust Automatic Transcription of Speech.

**REC** Complex Reflection.

**RES** Simple Reflection.

**SC** Speaker Clustering.

**SER** Speaker Error Rate.

**SRR** Speaker Role Recognition.

**ST** Structure.

**SU** Subword Unit.

**SVM** Support Vector Machine.

**TAL** This American Life.

**TDNN** Time-Delay Neural Network.

**UCC** University Counseling Center.

**VAD** Voice Activity Detection.

**WER** Word Error Rate.

**WFSA** Weighted Finite State Acceptor.

**WFST** Weighted Finite State Transducer.

# Abstract

Individuals assume distinct roles in different situations throughout their lives and people who consistently adopt particular roles develop specific commonalities in behavior. As a result, roles can be defined in terms of observable tendencies and behavioral patterns that can be manifest through a wide range of modalities during a conversational interaction. For instance, an interviewer is expected to use more interrogative words than the interviewee and a teacher is likely to speak in a more didactic style than the student.

Speaker role recognition is the task of assigning a role label in a speech segment where a single speaker is active, through computational models that capture such behavioral characteristics. The approaches that tackle this problem depend on successful pre-processing steps applied on the recorded conversation, such as speaker segmentation and clustering or automatic speech recognition, something that inevitably leads to error propagation. At the same time, accurate role information can provide valuable cues for the aforementioned speech processing tasks.

In this dissertation I propose techniques that combine role recognition with other speech processing modules to alleviate the problem of error propagation. Additionally, focusing on the task of speaker diarization (that answers the question who spoke when), I demonstrate that role-aware systems can achieve improved performance when compared to traditional, state-of-the-art approaches. Finally, I showcase how some of the proposed techniques can be applied in a real-world system, by presenting and analyzing an automated tool for psychotherapy quality assessment, where robust diarization and role identification (i.e., therapist vs. patient) are of critical importance.

# Introduction

## Roles and Human Interactions

Roles are one of the most important concepts in understanding and modeling human behavior. According to social psychology, individuals assume distinct roles in different situations throughout their lives that both guide their own behavioral patterns and create expectations about the behaviors of other people they interact with (Biddle, 1986). Systems of human interaction can be viewed as "*microscopic social systems*" (Bales, 1950) and roles can be defined as stated functions "*associated with a position in a group (a status) with rights and duties toward one or more other group members*" (Hare, 1994).

The underlying social structure, the context, and the end goal of an interaction both enable and constrain the participants' actions and behaviors (Gleave, Welser, Lento, & Smith, 2009), and this is reflected on the participants' roles. For instance, the role of the parent is associated with protecting and caring for offspring, the role of the chief executive officer is linked to taking managerial decisions that will lead to the economic growth of a company, and the role of the lecturer is related to conveying a clear message to their audience. Those roles can be adopted by the same person during different interactions and can occasionally collide and conflict with each other. However, clear role expectations can assist towards better task distribution within a group, promote individual responsibility and accountability, improve group cohesion, and eventually lead to more effective task performance (Mudrack & Farrell, 1995).

Roles can be distinguished into two broad categories. *Formal* roles (e.g., interviewer vs. interviewee) are typically associated with pre-defined objectives of an agent within a group, while *informal* roles (e.g., protagonist vs. supporter in a group discussion) can develop naturally as a result of interpersonal interactions and social dynamics and are sometimes referred to as *emergent* roles (Hare, 1994). Additionally, roles can be assigned to people either *implicitly*, because of the organizational or social structure of the environment where the interaction takes place, or *explicitly*, by requiring participants to perform specific tasks and providing detailed guidelines. Such scripted roles are of particular interest in collaborative learning scenarios, where role playing can foster engagement, discussion, and knowledge sharing (Strijbos & De Laat, 2010), leading to improved learning outcomes when compared to unstructured interactions (Weinberger, Stegmann, & Fischer, 2010). They are also a key aspect in group psychotherapy where patients learn to adhere to specific social and moral values through psychodrama (Kipper, 1992).

In any case, people who consistently adopt particular roles develop specific commonalities in behavior (Gleave et al., 2009). As a result, roles can be defined in terms of observable tendencies and behavioral regularities that can be manifest through a wide range of modalities during a conversational interaction. Thus, different roles may be associated with distinguishable patterns observed in acoustic, prosodic, linguistic, and structural characteristics (Bales, 1950; Knapp, Hall, & Horgan, 2013; Sacks, Schegloff, & Jefferson, 1978). For instance, a teacher is likely to speak in a more didactic style while a student be more inquisitive, an interviewer is expected to use more interrogative words than the interviewee, a doctor is likely to inquire on symptoms and prescribe while a patient describe their symptoms, and so on. All those patterns can be viewed as the structural signatures of the various roles and can be studied through statistical analysis and appropriate computational modeling.

## Computational Analysis of Speaker Roles

The phenomenal growth of multimedia data, including audio recordings, during the last few years, has been connected to heavy demands for efficient data manipulation applications. Speaker role information can be used to facilitate such applications, including audio indexing (Bigot, Ferrané, Pinquier, & André-Obrecht, 2010), topic-based segmentation (Vinciarelli & Favre, 2007), information retrieval (Barzilay, Collins, Hirschberg, & Whittaker, 2000), media browser enhancement (Ordelman, De Jong, & Larson, 2009), and multimedia summarization (Vinciarelli, 2006). At the same time, speaker roles offer valuable cues when studying various aspects of human communication such as entrainment and dominance (Beňuš et al., 2014; Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012). They are, additionally, of critical importance in computer-supported collaborative learning (Strijbos & De Laat, 2010), as well as in social computing and robotics (Beňuš, 2014).

Roles can also be of great value for the development of specialized dialogue models. For instance, in the psychotherapy domain, chatbots can play both the role of a therapist to provide mental health care services (Inkster, Sarda, & Subramanian, 2018), and the role of a patient to assist in training new counselors (Demasi, Li, & Yu, 2020; Tanana, Soma, Srikumar, Atkins, & Imel, 2019). A closely related notion to roles is that of personae. Personae, also known as character archetypes, are classes of characters grouped by similar behavioral traits (Jung, 2014), which means they are affected by the roles they potentially assume. Being able to adopt consistent personae is an essential element of engaging, naturalistic interactions (Roller et al., 2021) and, thus, persona modeling has been a key area of research in developing artificial conversational agents (Demasi et al., 2020; Song, Zhang, Cui, Wang, & Liu, 2019).

Given the importance of speaker roles in multimedia analysis, it is not surprising that there has been an increasing interest in applying computational methods to automatically recognize roles in speech documents. Formal speaker role recognition has been explored in a variety of domains, such as broadcast news programs (Bigot, Fredouille, & Charlet, 2013; Salamin & Vinciarelli, 2012), call centers (Garnier-Rizet et al., 2008), business meetings (Favre, Dielmann, & Vinciarelli, 2009; Sapru & Valente, 2012), psychotherapy sessions (Xiao, Huang, et al., 2016), press conferences (Li et al., 2017), interviews (Rasipuram & Jayagopi, 2018), and medical discussions (Luz, 2009). Other studies have investigated recognition of informal, emergent roles in multi-party interactions occurring in meetings (Sapru & Bourlard, 2015; Zancanaro, Lepri, & Pianesi, 2006) or in computer-assisted learning platforms (Dowell, Nixon, & Graesser, 2019). Usually, role recognition assigns a label from a pre-defined, finite set to a speech segment and is, therefore, viewed as a supervised classification task. However, unsupervised approaches that exploit the structure of the interaction and discover roles through clustering have been also proposed (Dowell et al., 2019; Li et al., 2017).

In order to address the problem of role recognition, appropriate features that capture the distinguishable patterns between different roles have to be extracted. Those features need to exploit characteristics that may be shared between different individuals, since the same role can be played by various speakers. Role-specific regularities can be found in the acoustic (Bigot, Ferrané, et al., 2010), lexical (Garg, Favre, Salamin, Hakkani Tür, & Vinciarelli, 2008), prosodic (Sapru & Valente, 2012), or structural (Salamin & Vinciarelli, 2012) characteristics of the speech signal, with the importance of each modality being task-specific. The extracted features are coupled with machine learning algorithms towards the final task of role classification or clustering. Early works in the field used boosting algorithms, maximum entropy classifiers, and support vector machines (Barzilay et al., 2000; Y. Liu, 2006; Zancanaro et al., 2006). More recently, social network analysis (Garg et al., 2008; Marcos-García, Martínez-Monés, & Dimitriadis, 2015), conditional random fields (Salamin & Vinciarelli, 2012), and deep learning approaches (Li et al., 2017) have also been explored.

Automated role recognition methods that rely on the computational analysis of recorded speech signals typically depend on successful pre-processing steps applied on the conversation, such as speaker diarization (answering the question *who spoke when*) or automatic speech recognition (ASR). At the same time, accurate role information can improve the performance of the aforementioned speech processing tasks (Sapru, Yella, & Bourlard, 2014; Valente, Vijayasenan, & Motlicek, 2011). This interplay between core speech processing and speaker roles is the focus of the current dissertation.

## Research Directions

In this dissertation I build and apply computational models to i) recognize speaker roles using speech and language processing techniques, and ii) use speaker role information to facilitate speech applications. In more detail, I study formal roles within both dyadic and multi-party recorded conversational interactions (e.g., *therapist* during a psychotherapy session, *host* during a podcast, *project manager* during a business meeting) and:

1. I propose a framework for speaker role recognition that alleviates error propagation from pre-processing steps, and

2. I leverage speaker role information to improve the performance of core modules in a speech processing pipeline, with a focus on speaker diarization.

My work can be summarized in the following *research statement*:

*The behavioral patterns found within conversational interactions can help us study speaker roles towards improved performance in speech processing tasks.*

## Outline

The current dissertation is structured as follows:

Introduction defines roles within the context of human interactions and reviews the computational methods that have been proposed in the literature for recognition and analysis of speaker roles, as well as applications for which speaker role information is a useful or even essential sub-task.

In Part I the focus is on how we can effectively use specific speech processing techniques in order to robustly infer speaker roles. To that end, Chapter 1 introduces a framework for the task of speaker role recognition that combines speaker-specific and role-specific information within a conversation from both the acoustic and linguistic modalities. The linguistic information here is acquired through manually-derived transcripts. Chapter 2 describes an effective way to infer speaker roles from transcribed audio data in real-world situations where transcriptions are obtained by an automatic speech recognition system.

In Part II we switch the focus on how speaker role information can help improve the performance of core speech analysis tasks within certain domains. The main area of interest is *speaker diarization*, the problem of answering the quesion "who spoke when" within a conversation. Even though this is typically addressed as an audio-only clustering-based problem, herein I explore ways to provide supplemental information in the form of linguistically extracted speaker roles. Chapter 3 presents a way to reduce the clustering diarization problem into a classification one, answering the question "which role spoke when". A limitation of the proposed approach is that it assumes a one-to-one correspondence between speakers and roles, i.e., each speaker needs to be associated with a unique role within the conversation (e.g., single interviewer vs. single interviewee). To address this limitation, Chapter 4 introduces an alternative, two-step framework where the language-based roles are only used to impose constraints on the subsequent audio-based clustering step.

Part III presents how some of the role-based computational techniques proposed in this work can be successfully applied in a real-world application. To that end, in Chapter 5 a fully automated psychotherapy quality assessment tool, deployed in clinical settings, is described and analyzed. We see why speaker role recognition is an essential element of the system and how the techniques introduced in Chapter 3 can be used to improve the overall performance, with respect to the downstream task of therapy evaluation.

The last chapter, named Conclusions and Future Directions, presents an overview of the dissertation and gives potential directions for future work. While my research has focused on formal roles, computational analysis of informal, emergent speaker roles and their usage within speech processing is an exciting area for future research. The relationship of speaker roles, either formal or informal, with other aspects of a person's identity and with social phenomena is another interesting, and quite unexplored from a computational perspective, research area.

# Part I

# Extracting Speaker Roles

# Chapter 1

# Combined Speaker Clustering and Role Recognition in Conversational Speech

Speaker role recognition (SRR) is usually addressed either as an independent classification task, or as a subsequent step after a speaker clustering module. However, the first approach does not take speaker-specific variabilities into account, while the second one results in error propagation. In this chapter we propose the integration of an audio-based speaker clustering algorithm with a language-aided role recognizer into a meta-classifier which takes both modalities into account. That way, we can treat separately any speaker-specific and role-specific characteristics before combining the relevant information together. The method is evaluated on two corpora of different conditions with interactions between a clinician and a patient and it is shown that it yields superior results for the SRR task.

The work presented in this chapter has been published in (Flemotomos, Papadopoulos, Gibson, & Narayanan, 2018).

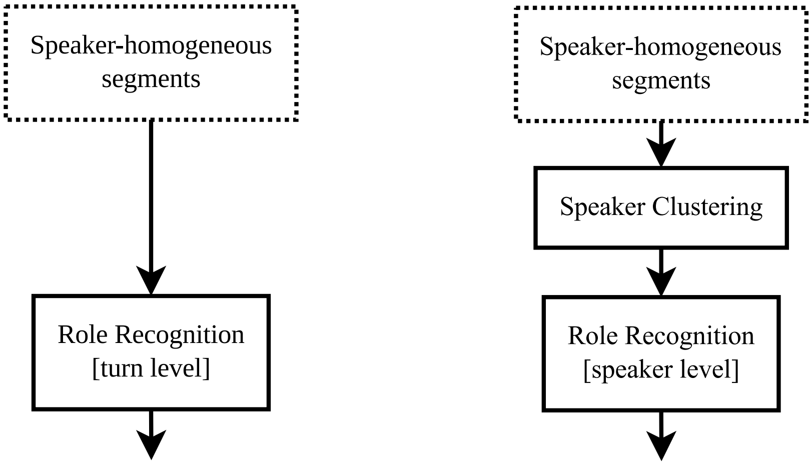
## 1.1 Introduction

Speaker role recognition (SRR) is the task of assigning a specific role to each speaker turn (speaker-homogeneous segment) in a speech signal. This task plays a significant role in numerous areas, such as information retrieval (Barzilay et al., 2000), audio indexing (Bigot, Ferrané, et al., 2010), or social interaction analysis (Biddle, 1986). Most of the research efforts have been focused on identifying roles in broadcast news programs or talk shows (Bazillon, Maza, Rouvier, Bechet, & Nasr, 2011; Damnati & Charlet, 2011a; Laurent, Camelin, & Raymond, 2014; Salamin & Vinciarelli, 2012), while there have been also works dealing with meeting scenarios (Sapru & Valente, 2012), conferences (Li et al., 2017), medical discussions between domain experts (Luz, 2009), and psychotherapy sessions (Xiao, Huang, et al., 2016). There have been presented both supervised (Barzilay et al., 2000; Bigot et al., 2013; Laurent et al., 2014; Rouvier, Delecraz, Favre, Bendris, & Bechet, 2015), and unsupervised (Hutchinson, Zhang, & Ostendorf, 2010; Li et al., 2017) methods.

The approaches towards dealing with the problem of SRR can be distinguished on the basis of whether the final decision is made at the turn level or the speaker level. In the former case (Figure 1.1a), a classifier is built where the input space is the space of speaker turns with no speaker information available. In a real-world application, those turns are obtained through a speaker change detection algorithm. The first works in the field use boosting algorithms (Barzilay et al., 2000) and statistical methods (Barzilay et al., 2000; Y. Liu, 2006) towards this classification task. Sapru and Valente (2012) combine lexical, prosodic, structural, and dialog act information also through boosting algorithms. Damnati and Charlet (2011a) combine audio-based and language-based classifiers with early or late fusion through a logistic regression model. Finally, Rouvier et al. (2015) have more recently applied deep learning techniques to learn turn-level role embeddings.

In the case of speaker-level SRR (Figure 1.1b), the classifier is built in two steps, the first being a speaker clustering (SC) algorithm, or a diarization system in the more general case[1], where turns are grouped into same-speaker clusters in an unsupervised way and then each cluster is assigned a specific role. In this line of work, Vinciarelli (2007) uses a social network analysis approach taking into consideration relational data across different speakers, while Bigot, Ferrané, et al. (2010) and

[1] More details on speaker clustering and speaker diarization are provided in Chapters 3 and 4.


Speaker-homogeneous segments
Role Recognition [turn level]
Speaker-homogeneous segments
Speaker Clustering
Role Recognition [speaker level]

(a) Turn-level SRR.

(b) Speaker-level SRR.

Figure 1.1: Two approaches for speaker role recognition.

Bigot et al. (2013) propose a hierarchical classification system. W. Wang, Yaman, Precoda, and Richey (2011) investigate the effect of various modalities on the final performance of SRR when using boosting algorithms. Dufour, Esteve, and Deléglise (2011) study the relationship between speech spontaneity levels and speaker roles, using a classifier based on boosting methods with decision stumps, which are replaced by small decision trees by Laurent et al. (2014). Bazillon et al. (2011) use question types as features, with results reported both at the speaker and the turn level.

In contrast to tasks such as speaker identification, the features to be extracted for SRR have to exploit characteristics that may be shared between different individuals, since the same role can be shared between various speakers. However, knowledge of speaker-specific information can lead to better classification results (e.g., Bazillon et al., 2011), which is the reason why many SRR-related works operate at the speaker level, employing a SC step. A major drawback of this *piped* approach, presented in Figure 1.1b, is that no matter how good the subsequent classifier is, any potential error in the SC algorithm is propagated and the overall performance is upper-bounded by the performance of the SC module. Thus, it is desirable to effectively combine speaker-specific and role-specific information without such problems.

To that end, Salamin and Vinciarelli (2012) propose an appproach where the final role recognition decision is taken at the turn level, but speaker information, available after a diarization step, is taken into account during feature extraction. However, that information is only used for the extraction of structural features (such as the average time between two turns of the current speaker).

Those are combined with turn-level prosodic features and the final classification is made using conditional random fields (CRFs). It is reported that, when using oracle speaker segmentation, this combination does not lead to improved results over the independent usage of the two different feature sets. Damnati and Charlet (2011b) present a hybrid hierarchical approach, where the SC output is used to distinguish at the speaker level a specific role from all the others, which are then classified at the turn level. However, this approach has been proposed specifically for application in broadcast news shows, taking into consideration different variabilities between the *anchors* and the *reporters* on the one hand and between the *reporters* and *others* on the other.

In this chapter, we present an alternative generic framework to combine a SC algorithm with a turn-level supervised role classifier, in such a way that both speaker-specific and role-specific information is taken into account for the final decision. We evaluate our method on the binary problem of patient-clinician interactions using manually extracted speaker turns. However, the framework presented is generalizable to an arbitrary number of speakers, under the assumption of one-to-one correspondence between speakers and roles in a single speech document, in the sense that each speaker is uniquely linked to a single role within the conversation[2].

## 1.2 Proposed Method

### 1.2.1 General framework

We propose the *combined* architecture presented in Figure 1.2, where the SC and role recognition modules work in parallel and their output is fed as input to a meta-classifier.

We assume that we know a priori the number of speakers in the speech document, say $N$, and that there is a one-to-one correspondence between the set of speakers $\{S_i\}_{i=1}^N$ and the set of roles $\{R_i\}_{i=1}^N$. We treat the outputs of the two modules as continuous-valued scores assigned to each speaker/role label. Thus, the output of the SC algorithm is the sequence of tuples $(p_{1i})_{i=1}^N, (p_{2i})_{i=1}^N, \cdots, (p_{Ti})_{i=1}^N$, such that the $k$-th turn would be assigned the speaker label $S_m$ if and only if $p_{km} = \max_i p_{ki}$. Similarly, the output of the role recognition module is the sequence

[2] This is an assumption we will follow thoughout most of the dissertation. In Chapter 4 we will extensively discuss how such an assumption can be limiting in some domains and we will see an application-specific approach that can be used even in cases where such an assumption does not hold.

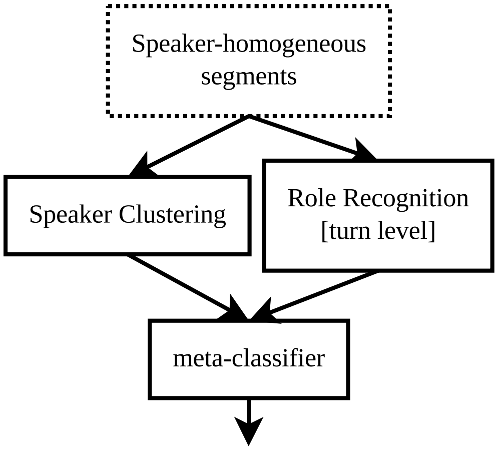


Figure 1.2: Proposed approach for speaker role recognition.

of tuples $(q_{1i})_{i=1}^{N}, (q_{2i})_{i=1}^{N}, \cdots, (q_{Ti})_{i=1}^{N}$, such that the $k$-th turn would be assigned the role label $R_m$ if and only if $q_{km} = \max_i q_{ki}$. In that way, for each turn we have $2N$ scores corresponding to the $N$ speakers/roles. Those are treated as input features for the classifier of the last step of the architecture.

Since there is not a natural correspondence between the two systems' outputs, it is necessary to find the optimal matching between the two sets of labels $\{S_i\}_{i=1}^{N}$ and $\{R_i\}_{i=1}^{N}$. This is a standard step taking place in the more general case of diarization systems output combination (Bozonnet et al., 2010; Tranter, 2005) or for the evaluation of speaker clustering performance (D. Liu & Kubala, 2004). For a small $N$ (which is a realistic assumption for conversational settings), it is easy to find this matching in an exhaustive way. Formally, if we denote such a matching as the mapping $M : \{S_i\}_{i=1}^{N} \to \{R_i\}_{i=1}^{N}$, the optimal matching is defined as

$$\hat{M} = \underset{M}{\operatorname{argmin}} \sum_{k=1}^{T} \mathbb{I}(M(S'_k) \neq R'_k) d_k \tag{1.1}$$

where $S'_k \in \{S_i\}_{i=1}^{N}$ and $R'_k \in \{R_i\}_{i=1}^{N}$ are the labels assigned by the two modules to the $k$-th turn, $\mathbb{I}(\cdot)$ is the indicator function, $d_k$ is the duration of the turn, and $T$ is the total number of turns in the speech document.

### 1.2.2 Speaker clustering module

For the speaker clustering module we use a simple bayesian information criterion (BIC) based hierarchical agglomerative clustering (HAC) algorithm (S. Chen & Gopalakrishnan, 1998; Cheng & Wang, 2003). At each step of the HAC procedure we use one Gaussian to model each cluster, so that the distance metric, known as ΔBIC, between two clusters $x$ and $y$, with $n_x$, $n_y$ members (frames) and with covariance matrices $\Sigma_x$, $\Sigma_y$, respectively, is

$$\Delta\text{BIC}(x,y) = \frac{1}{2}\left(n\log|\Sigma| - n_x\log|\Sigma_x| - n_y\log|\Sigma_y|\right) - \lambda\frac{d(d+3)}{4}\log n \tag{1.2}$$

where $n = n_x + n_y$, $\Sigma$ is the covariance matrix if we merge the clusters $x$ and $y$, $d$ is the dimensionality of the feature vector representing each frame, and $\lambda$ is a penalty factor ($\lambda = 1$ for our experiments). At each step, the pair of clusters with the minimum ΔBIC is merged.

Speaker clustering in this work is purely based on the acoustic information and as features we use the 13 first MFCCs for each frame. At the last step, we have one Gaussian modeling each of the $N$ speakers and the required scores for the turn are the per-frame log-likelihoods with respect to each Gaussian averaged over the voiced frames of the turn. The voiced frames are identified with a voice activity detection (VAD) algorithm, which is also applied at the initial step of the HAC procedure, so that the constructed Gaussians model only the voiced information for each speaker.

### 1.2.3 Role recognition module

We explore two different approaches for the role recognition module; one language-based and one audio-based.

In order to build a language-based role recognizer to exploit the linguistic patterns that are potentially shared between speakers with the same roles, we use similar ideas as in the role matching module presented by Xiao, Huang, et al. (2016). Since we treat role recognition as a supervised classification task, we need a role-labeled training set of speaker turns. On that set we train $N$ n-gram language models (LMs), one for each role. During the test phase, we evaluate the perplexity of the turn to be classified with respect to all the constructed LMs. The required scores to be used as input to the meta-classifier are the $N$ negative log-perplexities.

Even though we use the acoustic information in the SC module, we are interested in exploring the hypothesis that the exact same information has a predictive power over roles, apart from speakers. Following a similar idea as in (Damnati & Charlet, 2011a), we build an acoustic model (AM) for each one of the $N$ roles. The AM for a role is a gaussian mixture model (GMM) fit on the voiced frames of all the turns available in the training set which are labeled with that role. The scores for the turn to be used during the test phase are again, as in the case of the SC algorithm, the $N$ per-frame log-likelihoods with respect to each GMM averaged over the voiced frames of the turn.

## 1.3 Datasets

For this work, we evaluate our proposed method on two different corpora from the psychology domain, featuring interactions between a clinician and a patient. The first corpus is composed of motivational interviewing (MI)—a specific type of psychotherapy—sessions between a therapist (T) and a client (Cl) collected from six independent clinical trials (ARC, ESPSB, ESB21, CTT, iCHAMP, HMCBI; Atkins, Steyvers, Imel, & Smyth, 2014; Baer et al., 2009)[3]. We collectively refer to those sessions as the MI corpus. In this study, we use 343 manually transcribed sessions. The second corpus comprises autism diagnostic observation schedule (ADOS) assessments between a psychologist (P) and a child (Ch) being evaluated for a pervasive developmental disorder (PDD) (Lord et al., 2000). In this study, we use 273 manually transcribed sessions, with a minimum duration of 2 min.

There is a limited number of sessions where there are more than two speakers involved. In such cases, we do not take into account any turns not belonging to the clinician/patient for our analysis. Additionally, there is a limited number of non-pure speaker turns, in the sense that the manually annotated boundaries are not optimal and occasionally overlap. We chose to include such turns in the analysis without any preprocessing, since in a real-world setting (i.e., after automatic segmentation) such problems are impossible to completely avoid.

Some descriptive analysis for the two datasets is presented in Table 1.1. Unfortunately, the exact total number of different clients is not available for the MI dataset. However, under the

[3] Motivational interviewing is studied extensively in Chapter 5.

assumption that it is highly improbable for the same client to visit different therapists in the same study, and having partial information available about the client identities, we made the train/test split in a way that we are highly confident there is no overlap between speakers. Similarly, the exact total number of psychologists is unknown for the ADOS corpus, but the data are collected from two different clinics (in different cities) and we assume that the same clinician does not work for both. So, the data from one clinic is used for training and from the other for testing.

Table 1.1: Descriptive analysis of the corpora used.

| | **MI-train** | **MI-test** | **ADOS-train** | **ADOS-test** |
|---|---|---|---|---|
| #sessions | 242 | 101 | 141 | 132 |
| duration (mean) | 27.24 min | 33.14 min | 3.67 min | 3.67 min |
| duration (std) | 14.40 min | 17.42 min | 1.34 min | 1.65 min |
| duration-T/P | 47.30 h | 26.35 h | 2.63 h | 2.52 h |
| duration-Cl/Ch | 52.96 h | 25.87 h | 2.97 h | 2.98 h |
| #T/P | 123 | 53 | – | – |
| #Cl/Ch | – | – | 89 | 81 |

By *duration-T/P* and *duration-Cl/Ch* we denote the total duration of all the speaker turns labeled as therapist/psychologist and client/child, respectively.
By *#T/P* and *#Cl/Ch* we denote the total number of different therapists/psychologists and clients/children.

## 1.4 Experiments and Results

The two available datasets are split into train and test sets, as explained in Section 1.3, in a way that, with high confidence, there are not overlapping speakers between the sets, in order to ensure that the trained models indeed capture role-specific and not speaker-specific information. The train set is only used to build the LMs and AMs described in Section 1.2.3 corresponding to the different roles.

The LMs are 3-gram models trained (and later evaluated) using the SRILM toolkit (Stolcke, 2002) with manually derived transcriptions of the recordings. In order to ensure a large enough vocabulary that minimizes the unseen words during the test phase, we interpolate those models with a large background model—namely with the pruned version of the 3-gram model of cantab-TEDLIUM (Williams, Prasad, Mrva, Ash, & Robinson, 2015)—giving a weight of 0.9 to the domain-specific LM and 0.1 to the background one.

The AMs are diagonal GMMs, modeling the frames of turns assigned to each role, where frames are represented by 13-dimensional MFCCs. During training, we take into consideration only the voiced frames, by applying to the initial speaker turns a simple, energy-based VAD algorithm, as implemented in the Kaldi speech recognition toolkit (Povey et al., 2011). The same VAD algorithm is applied during evaluation, as well as during the SC step, as explained in Section 1.2.2.

As a meta-classifier we are use a binary linear support vector machine (SVM), since we evaluate on binary problems. All the results are based on a 5-fold cross-validation scheme on the data allocated for testing in each dataset, where, as is the case for the initial train/test split, we use all the available meta-data information to minimize any possible overlapping of speakers between different folds. The reason we are adopting this approach and do not use the training part of the datasets is that we do not want to pipe data already seen by the AMs and/or LMs to the SVM training.

As the evaluation metric of SRR we use the misclassification rate (MR), defined as (D. Liu & Kubala, 2004)

$$\text{MR} = \frac{\#\text{misclassified frames}}{\text{total }\#\text{frames}} = \frac{\sum_k \mathbb{I}(R_k \neq \hat{R}_k) d_k}{\sum_k d_k} \tag{1.3}$$

where the summation is over all the speaker turns, $R_k$ is the role assigned by the algorithm, $\hat{R}_k$ is the reference role, $d_k$ is the duration of the $k$-th turn, and $\mathbb{I}(\cdot)$ is the indicator function.

In Figure 1.3 we can see how MR is affected by the number of Gaussians in the GMM-based AM, when only the audio-based role recognizer is used. Based on that, we use 512 Gaussians for the subsequent experiments both for the MI and the ADOS datasets.

In this work we do not report results for the *piped* architecture presented in Figure 1.1b using an actual classification algorithm as the second step of the pipeline. Instead, in Table 1.2 we give the best possible result with this architecture when using the SC algorithm that we have described. Using a perfect classification algorithm for the SRR task at the speaker level, which we denote as $\mathcal{R}^\dagger$, the overall error of the system is always lower-bounded by the error of the SC algorithm itself. So, the results reported in the *SC+$\mathcal{R}^\dagger$-piped* column of the Table are in fact the MRs of the SC algorithm.

The language-based and audio-based recognizers are evaluated when used independently (*LM-only* and *AM-only*) and when used in the *combined* architecture presented in Figure 1.2 (*SC+LM-*

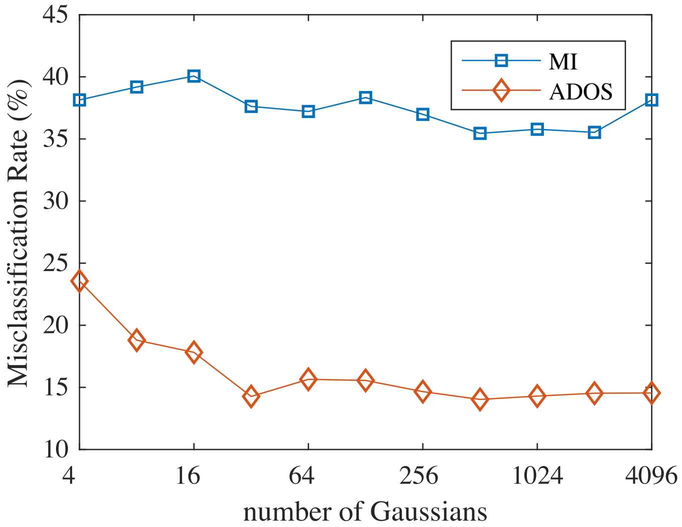


Figure 1.3: SRR misclassification rate when using only the AM-based decision, as a function of the number of Gaussians in the GMM.

Table 1.2: Misclassification rates (%) of the SC algorithm, the language-based recognizer (LM), and the audio-based recognizer (AM), when used independently (*only*) or in a piped (*piped*) or combined (*comb*) architecture for the task of SRR.

| | **SC+$\mathcal{R}^\dagger$ piped** | **LM only** | **SC+LM comb** | **AM only** | **SC+AM comb** | **AM+LM comb** | **SC+AM+LM comb** |
|---|---|---|---|---|---|---|---|
| MI | 3.59 | 9.49 | 2.76 | 35.45 | 3.66 | 9.17 | **2.71** |
| ADOS | 12.67 | 12.37 | 7.70 | 14.03 | 10.58 | 8.02 | **5.98** |

By $\mathcal{R}^\dagger$ an optimal, 0-error classification algorithm is denoted.

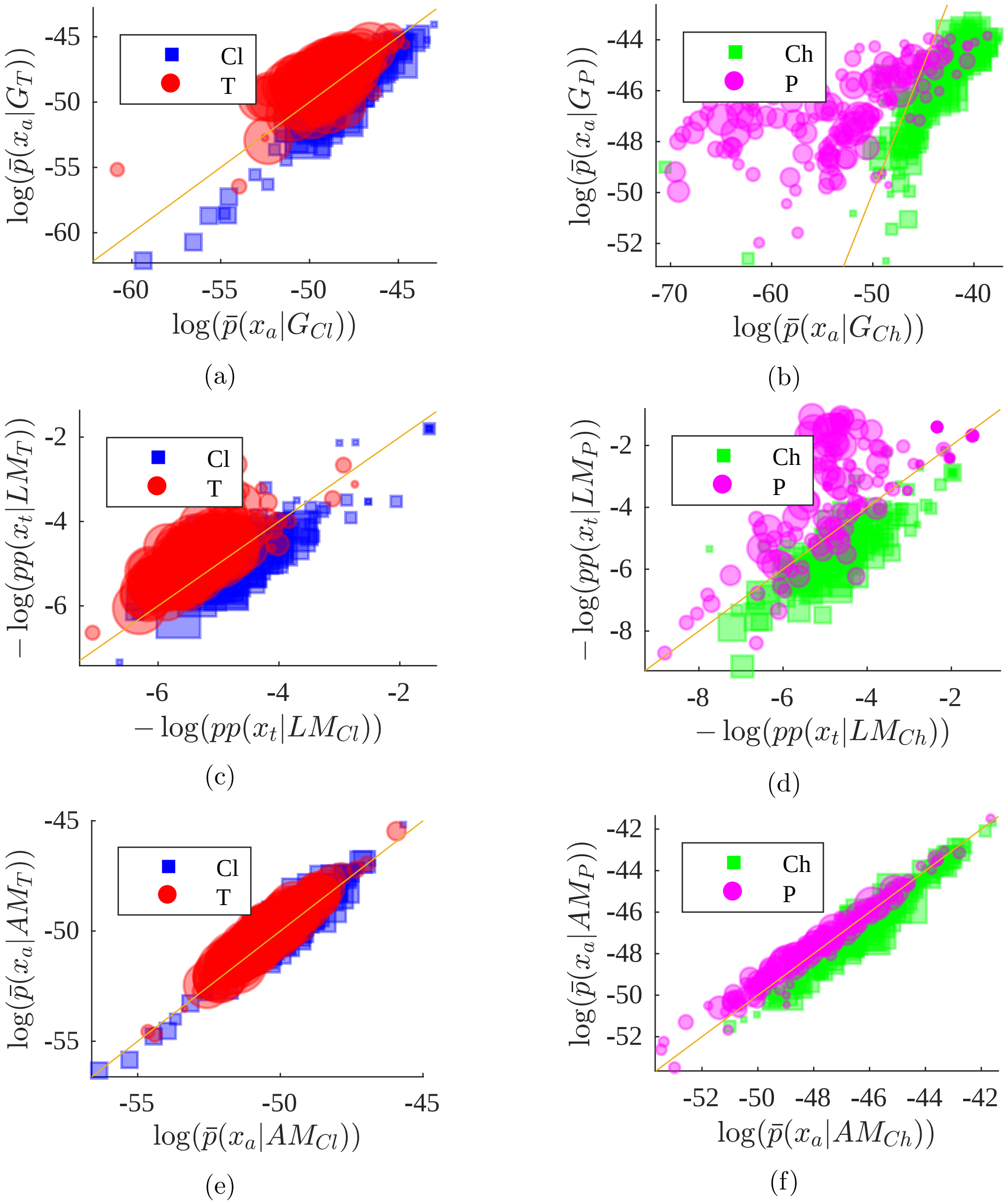
*comb* and *SC+AM-comb*). The results are reported in Table 1.2. As we can see, the LM-based approach has a strong predictive power for both datasets, revealing differences in the linguistic patterns between a clinical provider and a client or a child evaluated for PDD. When this is combined with the SC algorithm which captures the speaker-specific differences within a single session, the results are considerably better, compared not only to the independent classifiers, but also to the *piped* architecture.

On the other hand, the AM approach does not behave in the same manner for the two datasets. As expected, the acoustic characteristics of the children as a whole are different than those of the adult clinicians. This is reflected in the *AM-only* results for the ADOS data, even though they are still worse than the *LM-only* ones. This age distinction between the two different groups of speakers does not exist in the MI dataset. So, although it seems from the results that there is some non-negligible acoustic variability between the clinicians and the clients, the performance gap between the *LM-only* and the *AM-only* approaches is much larger for those data. When combined with the SC algorithm the results are substantially better, because the meta-classifier is affected by the more separated scores which are the output of the SC module. This notion of "separability" is visually depicted in Figure 1.4 where we show how the outputs of the SC, LM, and AM modules are distributed on the plane. It is of high interest that in the case of the ADOS dataset, because of its very special nature, the exact same information (at the feature level) can be used to capture both role-specific and speaker-specific variabilities in a way that if the two modules are combined by our proposed architecture (*SC+AM-comb*), they can improve the overall performance as if they carried complementary information.

As a final experiment, we combine the outputs of the LM- and the AM-based recognizers, again using the linear SVM as the meta-classifier (*AM+LM-comb*) and we also combine all the three constructed modules in an extended *combined* architecture (*SC+AM+LM-comb*). In this latter case the meta-classifier gets $3 \cdot 2$ (in the general case $3N$) inputs for each turn to be classified. We note that the result of the optimal matching between SC and LM was the same as in between SC and AM, so we did not encounter any conflict. When compared to the *LM-only* and the *SC+LM-comb* results, the addition of the acoustic-based recognizer in the architecture does not lead to any substantial improvements, as expected, for the MI data, but does improve the performance of the system for the case of the ADOS sessions. Overall, the relative error improvement with our final system which follows the *combined* architecture is 24.5% for the MI data and 52.8% for the ADOS data, when compared to the *piped* architecture with an optimal recognizer.

## 1.5 Conclusion

In this chapter we proposed a framework to incorporate speaker-specific and role-specific information for the SRR task, by independently implementing an unsupervised SC algorithm and a supervised turn-level role classifier, the output scores of which are fed to a meta-classifier which gives a turn-level final decision. By evaluating our method on dyadic interactions we showed that it yields superior results, compared both to the independent use of turn-level classifiers which do not take speaker-specific variabilities into account, and to systems that use speaker-specific information by applying SC as a first step and predicting the output at the speaker level.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1.4: Distribution of the scores which are the output of the SC ((a),(b)), the LM-based recognizer ((c),(d)), and the AM-based recognizer ((e),(f)) for the MI ((a),(c),(e)) and the ADOS ((b),(d),(f)) datasets. Each data point is a speaker turn with size proportional to the turn length. 300 turns of the test set are randomly shown for each dataset. $x_a$ and $x_t$ are the acoustic and textual representation of a turn $x$. $LM_R$ and $AM_R$ are the LM and AM corresponding to the role $R$. $G_R$ is the Gaussian corresponding to the role $R$ at the end of the SC and after an optimal matching between speakers and roles.

One drawback of our methodology is that it requires additional data for the training of the meta-classifier. Moreover, in a real-world scenario, the speaker boundaries, as well as the language-based features, would be extracted, at least at the evaluation phase, from diarization and automatic speech recognition (ASR) outputs, which can lead to error propagation. In the following chapter, we will explore a technique to mitigate such potential error propagation due to ASR.

# Chapter 2

# Role Specific Lattice Rescoring for Speaker Role Recognition from Speech Recognition Outputs

As shown in the previous chapter, the language patterns followed by different speakers who play specific roles in conversational interactions provide valuable cues for the task of speaker role recognition (SRR). Given the speech signal, existing algorithms typically try to find such patterns either in manually derived transcripts or in the best path of an automatic speech recognition (ASR) system. In this chapter we propose an alternative way of revealing role-specific linguistic characteristics, by making use of role-specific ASR outputs, which are built by suitably rescoring the lattice produced after a first pass of ASR decoding. That way, we avoid pruning the lattice too early, eliminating the potential risk of information loss.

The work presented in this chapter has been published in (Flemotomos, Georgiou, & Narayanan, 2019).

## 2.1 Introduction

In Chapter 1 we introduced the problem of SRR, defined as the classification task of mapping a speaker-homogeneous segment (speaker turn) to an element of a predefined set of roles, where a role is characterized by the task a speaker performs and the objectives related to it. Typical examples of conversational interactions between individuals with specific roles are business meetings (Sapru & Valente, 2012), broadcast news programs (Bigot et al., 2013; Damnati & Charlet, 2011a), psychotherapy sessions (Xiao, Huang, et al., 2016), or press conferences (Li et al., 2017).

In order to address the problem of SRR, appropriate features which capture distinguishable patterns between the different roles have to be extracted. Such patterns can be found in the acoustic (Bigot, Pinquier, Ferrané, & André-Obrecht, 2010), lexical (Garg et al., 2008), prosodic (Sapru & Valente, 2012), or structural (Li et al., 2017; Salamin & Vinciarelli, 2012) characteristics of the speech signal, with the importance of each modality being task-specific. For instance, it is desired that a psychotherapist speaks less than the client, an interviewer is expected to use more interrogative words than the interviewee, etc. However, as validated by our experiments in Chapter 1 where we explored linguistic and acoustic characteristics, it seems that language often carries the most important information for the problem in hand (Damnati & Charlet, 2011a; Sapru & Valente, 2012; W. Wang et al., 2011) and is more robust to unseen conditions (e.g., different speakers) (Rouvier et al., 2015), which is the reason why a great portion of the research efforts has been focused on studying and exploiting the lexical variability between the speaker roles.

The first efforts in the field extract bags of n-grams to represent the lexical information and use them as input features to boosting algorithms or maximum entropy classifiers (Barzilay et al., 2000; Y. Liu, 2006). Boosting approaches have been also followed by W. Wang et al. (2011) and Sapru and Valente (2012) to combine n-gram features with other modalities, with the final classification decision taken either at the speaker (W. Wang et al., 2011) or at the turn level (Sapru & Valente, 2012). Bazillon et al. (2011) first classify the types of questions posed by the different speakers and use that information for the role assignment. Rouvier et al. (2015) explore deep learning approaches by using word embeddings as inputs to convolutional neural networks. Xiao, Huang, et al. (2016) build role-specific n-gram LMs and reduce SRR to the problem of finding the LM which minimizes the perplexity of a speaker turn (or of all the turns assigned to a specific speaker after a speaker clustering step)[1].

Although a bulk of the aforementioned studies use manually transcribed speech data to perform SRR, in a real-world application the lexical information would become available after an ASR step (Rouvier et al., 2015; Xiao, Huang, et al., 2016). Moreover, Damnati and Charlet (2011b) suggest that the quality of ASR transcripts can be used to extract additional features carrying complementary information in specific scenarios. In any case, the ASR output is considered to be the best path of a system that uses generic acoustic and language models.

In this chapter, we propose using role-specific ASR systems, each one of which gives a potentially different output together with a corresponding cost. Then, after passing any given turn through all the systems, we can assign to that turn the role which corresponds to the system producing the minimum cost. In particular, for this study, we create the role-specific systems by rescoring the lattices generated by a generic ASR with role-specific LMs, as explained in Section 2.3. That way, we can exploit any information carried by the decoding lattice before pruning it to find the best path. Based on similar intuitions, Georgiou, Black, Lammert, Baucom, and Narayanan (2011a) and Xiao, Huang, et al. (2016) have previously explored lattice rescoring techniques for binary classification problems in the field of behavioral code prediction. Our method is evaluated on dyadic interactions from the clinical domain, as well as on multi-participant business meeting scenarios, yielding improved results for the task of SRR.

## 2.2 Background

In this section we give an overview of speech lattices and lattice rescoring. In order to better understand lattices, we first explain what a decoding graph is, within the framework of weighted finite state transducers (WFSTs).

[1] This is the baseline SRR approach we followed for our experimentation in the previous chapter, as detailed in Section 1.2.3

### 2.2.1 Weighted Finite State Transducers

Job of a WFST is to transform (or transduce) an input sequence into another output sequence, where the input and output sets of labels (alphabets) may differ[2] (Hori & Nakamura, 2013; Mohri, Pereira, & Riley, 2002). Every WFST is associated with some *semiring*, that enables us to perform various algebraic operations on it. The formal definition of a semiring is given below (Kuich & Salomaa, 1986):

**Definition 1.** *A* semiring*, denoted as* $< \mathbb{A}, \oplus, \otimes, \bar{0}, \bar{1} >$*, consists of a set* $\mathbb{A}$ *with two binary operations* $\oplus$ *and* $\otimes$ *and two constants* $\bar{0}$ *and* $\bar{1}$*, such that the following axioms are satisfied:*

*(i)* $a \oplus \bar{0} = \bar{0} \oplus a = a \; \forall a \in \mathbb{A}$,

*(ii)* $a \otimes \bar{1} = \bar{1} \otimes a = a \; \forall a \in \mathbb{A}$,

*(iii)* *commutativity for* $\oplus$*:* $a \oplus b = b \oplus a \; \forall a \in \mathbb{A}, \forall b \in \mathbb{A}$,

*(iv)* *distributivity:*

$a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$ *and* $(a \oplus b) \otimes c = (a \otimes c) \oplus (b \otimes c) \; \forall a \in \mathbb{A}, \forall b \in \mathbb{A}, \forall c \in \mathbb{A}$,

*(v)* $\bar{0} \otimes a = a \otimes \bar{0} = \bar{0} \; \forall a \in \mathbb{A}$

WFSTs can be viewed as directed graphs where each edge represents a *transition* between two states. During a transition $t$, an input symbol (or the empty symbol) is converted to an output symbol (ot the empty symbol) with some weight $w(t) \in \mathbb{W}$, where $\mathbb{W}$ is the set of a semiring. A valid path $\pi$ is a sequence of finite successive transitions $t_1, t_2, \cdots, t_n$ from an initial state[3] to a final state $f$, associated with some weight $w(f)$. The total cost of the path is

$$w(\pi) = w(t_1) \otimes w(t_2) \otimes \cdots \otimes w(t_n) \otimes w(f) \tag{2.1}$$

For the task of speech recognition, weights are typically negative log probabilities and the *tropical semiring* $< \mathbb{R}_+ \cup \{\infty\}, \min, +, \infty, 0 >$ is the most widely used one.

[2] If there is no new output (or input and output are always the same), we have a weighted finite state acceptor (WFSA).

[3] We can safely assume that every WFST has a single initial state.

One of the most common operations defined on WFSTs is the binary operation of *composition*. Given two WFSTs $T_1$ and $T_2$, their composition $T$ is a WFST that transforms an input sequence $x$ into an output sequence $y$ according to the formula

$$T(x,y) = (T_1 \circ T_2)(x,y) \triangleq \bigoplus_{z} T_1(x,z) \otimes T_2(z,y) \tag{2.2}$$

### 2.2.2 WFST framework for speech recognition

Given a sequence of acoustic features $O$, the job of an ASR system from a traditional point of view[4] is to find, out of the set $\mathcal{W}$ of possible word sequences, the most probable sequence

$$\hat{W} = \operatorname*{argmax}_{W \in \mathcal{W}} P(W|O) = \operatorname*{argmax}_{W \in \mathcal{W}} P(O|W)P(W) \tag{2.3}$$

where $P(O|W)$ is the acoustic likelihood of $O$ for $W$, estimated through the acoustic model (AM), and $P(W)$ is the prior probability of $W$, estimated through a language model (LM). If the pronunciation lexicon mapping words to subword units (SUs) contains the additional information of how probable an SU sequence $V$ is, given the word $W$, then we get

$$\hat{W} = \operatorname*{argmax}_{W \in \mathcal{W}} \sum_{V \in K(W)} P(O|V,W)P(V|W)P(W) \approx \operatorname*{argmax}_{W \in \mathcal{W}} \sum_{V \in K(W)} P(O|V)P(V|W)P(W) \tag{2.4}$$

where $K(W)$ is the set of the possible SU-level representations of $W$. Since decoding is based on the Viterbi algorithm, the summation is replaced by a max function and finally we get in the log domain

$$\hat{W} \approx \operatorname*{argmax}_{W \in \mathcal{W}} \max_{V \in K(W)} \{\log P(O|V) + \log P(V|W) + \log P(W)\} \tag{2.5}$$

In the WFST framework, we have the transducer $\tilde{H}$ that transforms a sequence of acoustic features $O$ into an SU sequence $V$ with a weight $-\log P(O|V)$, the WFST $L$ that transforms an SU sequence $V$ into a word sequence $W$ with a weight $-\log P(V|W)$, and the WFSA $G$ that accepts a word sequence $W$ with a weight $-\log P(W)$ (Hori & Nakamura, 2013; Mohri et al., 2002). $\tilde{H}$ is actually split into a WFST $H$ that transforms a sequence of hidden markov model (HMM) states

[4] as opposed to the end-to-end neural approaches

into an SU sequence and a model $S$ that maps the acoustic observations to HMM states and is trained following either the GMM or the DNN paradigm. Since typically the elementary SUs in ASR are triphones and the pronunciation lexicons give the phoneme-level representation of each word, it is necessary to have one more WFST $C$ that transforms a triphone sequence into a phoneme sequence, where each phoneme is context-independent and is identical to the central phoneme of the corresponding triphone. Those automata are composed into a final WFST[5]

$$N = H \circ C \circ L \circ G \tag{2.6}$$

Given any speech utterance $x$ of $t$ frames, we construct the WFSA $T_x$ that represents $x$ (with $t+1$ nodes and one arc between consecutive nodes for each HMM state) and ASR is now a shortest path problem on $T_x \circ N$, called the *decoding search graph* for the specific utterance.

### 2.2.3 Speech lattices

Conventional ASR systems try to find the shortest path on the decoding graph, such that a sequence of HMM states is transduced to a word sequence with the minimum possible cost. In many cases, however, it is desirable to keep multiple sufficiently probable transcription hypotheses, and not only the best one. This can be done either by keeping a list of $n$-best sequences, or by generating a speech recognition *lattice*. A lattice is a weighted directed acyclic graph (and thus, can be represented as a WFSA) with word labels (Ljolje, Pereira, & Riley, 1999). Each valid path represents an alternative word sequence, weighted by its recognition cost, and an exponential number of such word sequences can be encoded with respect to the number of nodes in the lattice. An example of a word lattice is given in Figure 2.1.

Time and alignment information is also usually included in the lattice. According to Povey et al. (2012), given some cost tolerance $\delta$, any lattice should satisfy the following conditions:

(i) There should be one path for every word sequence within $\delta$ from the one with the minimum total cost,

[5] In practice, optimization operations need to be applied before the final composition (Mohri et al., 2002).

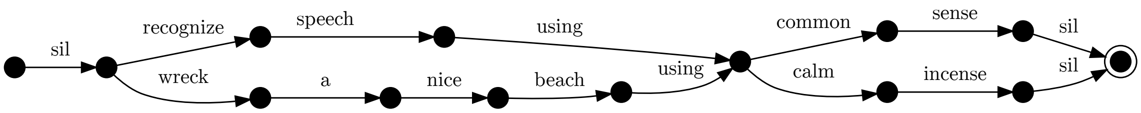


Figure 2.1: Example of speech recognition lattice encoding four alternative transcription hypotheses. For simplicity, no scores or alignments are shown.

(ii) there should only be one path for any distinct word sequence (no duplicate paths allowed),

(iii) the scores and alignments in the lattice should be accurate.

### 2.2.4 Lattice rescoring

One of the main reasons it is desirable to generate lattices during decoding is so that we can later process them and rescore them with more complex, or domain-specific, models. For example, a lattice can be rescored to infuse knowledge-based information (Siniscalchi, Li, & Lee, 2006). Another common scenario is when a relatively simple language model is used during first-pass decoding due to lower computational complexity and the generated lattice is later rescored with a better language model to improve accuracy (Sak, Saraçlar, & Güngör, 2010; Xu et al., 2018).

One of the simplest ways to rescore a lattice is through composition (Povey et al., 2012). Essentially, for LM-rescoring, which is the focus of this study, we want to subtract the old LM cost and add the new LM cost to the weighted automaton representing the lattice. According to the analysis in Section 2.2.2, the lattice should include two scores; the graph cost corresponding to the weight of the WFST $N$ (which, based on equation (2.6), incorporates the LM cost from $G$, the pronunciation cost from $L$, and the HMM-transitions-related cost from $H$) and the acoustic cost corresponding to the model $S$. Storing each weight on the lattice as a pair of graph and acoustic weights $(w_{gr}, w_{ac})$, the lattice $\mathcal{L}_{G_{new}}(x)$ for an utterance $x$, after rescoring with an LM $G_{new}$, can be expressed as

$$\mathcal{L}_{G_{new}}(x) = \left(\mathcal{L}^{\dagger}_{G_{old}}(x) \circ G_{old}\right)^{\dagger} \circ G_{new} \tag{2.7}$$

where $\mathcal{L}_{G_{old}}(x)$ is the lattice generated using the old LM $G_{old}$ and $\dagger$ represents the operation of scaling both the graph and acoustic lattice costs multiplying by $-1$. For $G_{old}$ and $G_{new}$ the weights are of the form $(w, 0)$. A semiring similar to the tropical semiring on $w_{gr} + w_{ac}$ can be used for lattices, but keeping track of the graph and acoustic weights separately. More precisely, the semiring used is equipped with the following operations:

- $(w_{gr_1}, w_{ac_1}) \otimes (w_{gr_2}, w_{ac_2}) = (w_{gr_1} + w_{gr_2}, w_{ac_1} + w_{ac_2})$
- $(w_{gr_1}, w_{ac_1}) \oplus (w_{gr_2}, w_{ac_2}) = \begin{cases} (w_{gr_1}, w_{ac_1}), & \text{if } w_{gr_1} + w_{ac_1} < w_{gr_2} + w_{ac_2} \\ (w_{gr_2}, w_{ac_2}), & \text{if } w_{gr_1} + w_{ac_1} > w_{gr_2} + w_{ac_2} \end{cases}$

Ties in the latter case are broken comparing $w_{gr_1} - w_{ac_1}$ vs. $w_{gr_2} - w_{ac_2}$[6].

## 2.3 Proposed Method

Given a generic ASR system, the goal is to convert the generated decoding lattice for an input turn to multiple, role-specific versions, in such a way that there is one version that reflects the speaker role corresponding to the particular turn. We do this by rescoring the lattice $N$ times, where $N$ is the number of roles, with role-specific LMs. Let's assume we have a background, out-of-domain n-gram LM $\mathcal{G}$ and $N$ role-specific LMs $\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_N$ corresponding to the roles $R_1, R_2, \cdots, R_N$, which are trained using in-domain data. First, we ensure that all the models which are going to be used recognize the same vocabulary. We can efficiently do so by interpolating the individual LMs to get the mixed models $\mathcal{G}^+, \mathcal{R}_1^+, \mathcal{R}_2^+, \cdots, \mathcal{R}_N^+$. To obtain an interpolated model, we assign to each n-gram the weighted average of the probabilities from the input models, and we then we re-normalize the produced model (Stolcke, 2002). Using the symbol $\odot$ to denote LM interpolation, the final models are expressed as

$$\mathcal{G}^+ = w_g \mathcal{G} \odot (1 - w_g)\tilde{\mathcal{R}} \tag{2.8}$$

$$\mathcal{R}_i^+ = w_{g_i} \mathcal{G} \odot w_{r_i} \mathcal{R}_i \odot (1 - w_{g_i} - w_{r_i})\tilde{\mathcal{R}}_i \tag{2.9}$$

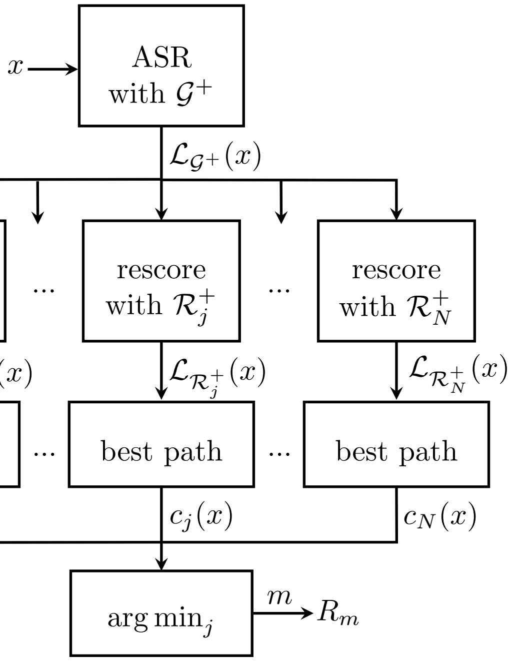
[6] For more details, please refer to (Povey et al., 2012) and to the Kaldi documentation at https://kaldi-asr.org/doc/lattices.html.

where

$$\tilde{\mathcal{R}} = \frac{1}{N}\bigodot_{i=1}^{N}\mathcal{R}_i, \quad \tilde{\mathcal{R}}_i = \frac{1}{N-1}\bigodot_{\substack{j=1\\ j\neq i}}^{N}\mathcal{R}_j$$

and all the weights $w_g, w_{g_i}, w_{r_i}$ are chosen to minimize the perplexity of appropriate role-specific development corpora.

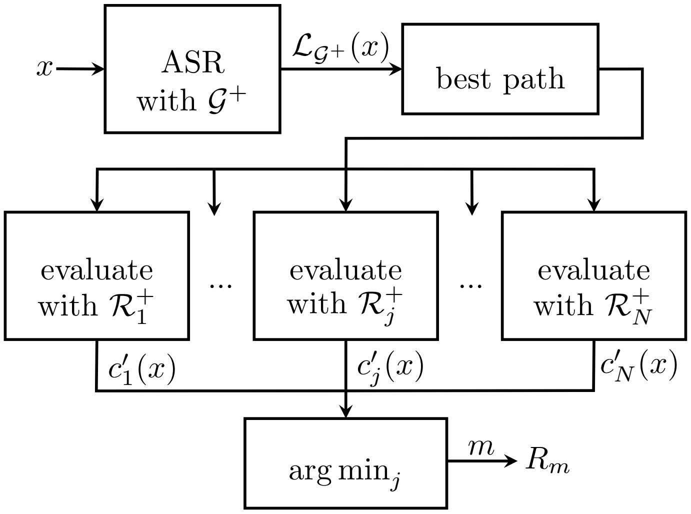
Given an input turn $x$, we first pass it through an ASR system, trained with the LM $\mathcal{G}^+$, producing a decoding lattice $\mathcal{L}_{\mathcal{G}^+}(x)$. The lattice is then rescored with all the LMs $\mathcal{R}_j^+$, $j = 1, 2, \cdots, N$ to produce the lattices $\mathcal{L}_{\mathcal{R}_j^+}(x)$. Denoting as $c_j(x)$ the LM cost of the best path in $\mathcal{L}_{\mathcal{R}_j^+}(x)$, the role assigned to $x$ is $R_m$ where $m = \operatorname{argmin}_j c_j(x)$. The process is visually depicted in Figure 2.2. The difference between this approach and the language-based approach followed in Chapter 1 is that in the second case the evaluation with respect to a role-specific LM would be done using the final output of the ASR, as presented in Figure 2.3. That way, the lattice $\mathcal{L}_{\mathcal{G}^+}(x)$ is pruned using a generic LM, which can potentially lead to loss of valuable information for the task of SRR. This is exactly the problem our approach tries to avoid.



$x$
ASR with $\mathcal{G}^+$
$\mathcal{L}_{\mathcal{G}^+}(x)$
rescore with $\mathcal{R}_1^+$
...
rescore with $\mathcal{R}_j^+$
...
rescore with $\mathcal{R}_N^+$
$\mathcal{L}_{\mathcal{R}_1^+}(x)$
$\mathcal{L}_{\mathcal{R}_j^+}(x)$
$\mathcal{L}_{\mathcal{R}_N^+}(x)$
best path
...
best path
...
best path
$c_1(x)$
$c_j(x)$
$c_N(x)$
$\arg\min_j$
$m$
$R_m$

Figure 2.2: Turn-level SRR by role-specific lattice rescoring.

If the extra information of the speaker who uttered the turn is available, after a speaker clustering step, then the role assignment can be done more robustly at the speaker level instead of the



$x$
ASR with $\mathcal{G}^+$
$\mathcal{L}_{\mathcal{G}^+}(x)$
best path
evaluate with $\mathcal{R}_1^+$
...
evaluate with $\mathcal{R}_j^+$
...
evaluate with $\mathcal{R}_N^+$
$c_1'(x)$
$c_j'(x)$
$c_N'(x)$
$\arg\min_j$
$m$
$R_m$

Figure 2.3: Turn-level SRR by evaluating the text with role-specific LMs.

turn level, as we already saw in Chapter 1. If we denote by $T_i$ the set of turns corresponding to speaker $S_i$, we can define the cost of the speaker-role pair $(S_i, R_j)$ as

$$c(S_i|R_j) \triangleq \sum_{x \in T_i} c_j(x) \tag{2.10}$$

Ideally, we would again like to assign to any speaker $S_i$ the role $R_m$ such that the cost $c(S_i|R_m)$ is the minimum among all $c(S_i|R_j),\ j = 1, 2, \cdots, N$. However, assuming that there is one-to-one correspondence between speakers and roles in a speech document, which is the case for many practical applications, this criterion would fail, since there is no guarantee that for $n \neq m$ we have $\text{argmin}_j\, c(S_n|R_j) \neq \text{argmin}_j\, c(S_m|R_j)$.

Thus, in order to take such a constraint into account, we use Algorithm 1, which is a generalization of the role matching criterion we proposed in (Flemotomos, Martinez, et al., 2018) for the 2-speaker scenario, where the costs were perplexities. The algorithm begins with the entire sets $\tilde{S}$ and $\tilde{R}$ of the available speakers and roles and at every iteration it chooses the speaker $S_k$ such that a confidence metric $C_k$ is the maximum among all $C_i, i = 1, 2, \cdots, |\tilde{S}|$. Then, it assigns to $S_k$ the role $R_{l_k}$ that minimizes the cost $c(S_k|R_j), j = 1, 2, \cdots, |\tilde{R}|$ and removes $S_k$ and $R_{l_k}$ from the available speakers and roles. The confidence metric $C_i$ is designed in such a way that the larger the difference between the minimum cost and the rest of the costs for $S_i$ is, the more confident we are about the role assignment of the particular speaker.

**Algorithm 1** Speaker-level SRR given costs for each (speaker,role) pair.

```
Inputs: speakers $S_1, S_2, \cdots, S_N$
        roles $R_1, R_2, \cdots, R_N$
        costs $c(S_i|R_j) \forall i, j$

$\tilde{S} \leftarrow \{S_i\}_{i=1}^N$;  $\tilde{R} \leftarrow \{R_i\}_{i=1}^N$
while $\tilde{S} \neq \phi$ do
    for $S_i \in \tilde{S}$ do
        $l_i \leftarrow \text{argmin}_m \, c(S_i|R_m),\ R_m \in \tilde{R}$
        $C_i \leftarrow \min_n |c(S_i|R_{l_i}) - c(S_i|R_n)|,\ R_n \in \tilde{R} \setminus \{R_{l_i}\}$
    end for
    $k \leftarrow \text{argmax}_i \, C_i$
    assign $R_{l_k}$ to $S_k$
    $\tilde{S} \leftarrow \tilde{S} \setminus \{S_k\}$;  $\tilde{R} \leftarrow \tilde{R} \setminus \{R_{l_k}\}$
end while
```

## 2.4 Datasets

We evaluate our method on two datasets featuring interactions between individuals under different conditions. The first dataset, to which we will refer as the PSYCH corpus, is composed of motivational interviewing sessions between a therapist (T) and a client (C) and is collected from five independent clinical trials (ARC, ESPSB, ESP21, iCHAMP, HMCBI; Atkins et al., 2014)[7]. The second one is the AMI meeting corpus (Carletta et al., 2005) from which we use the independent headset microphone (IHM) setup of the scenario-only part. This is composed of meetings where each participant plays the role of an employee in a company; the project manager (PM), the marketing expert (ME), the user interface designer (UI), and the industrial designer (ID).

The two datasets are split into training, development and test sets in such a way that there is no speaker overlap between them. For the AMI corpus we follow the scenario-only partition which is officially recommended[8]. For the PSYCH corpus, since the client identities are not available for the HMCBI sessions, the partitioning is done under the assumption that it is highly improbable for the same client to visit different therapists in the same study, as explained in Chapter 1. In both cases, we use the manually derived segmentation. The datasets are presented in Tables 2.1 and 2.2.

[7] Note that this is a subset of the MI dataset used in Chapter 1, described in Table 1.1. In particular here we do not use the 200 CTT sessions (Baer et al., 2009). Those feature scripted interactions between actors playing the roles of therapist vs. patient; here we only consider real-world clinical interactions.

[8] https://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml

Table 2.1: Size of the PSYCH dataset.

| | **PSYCH-train** | **PSYCH-dev** | **PSYCH-test** |
|---|---|---|---|
| #sessions | 74 | 44 | 25 |
| duration-T | 26.43 h | 15.23 h | 7.34 h |
| duration-C | 23.29 h | 12.17 h | 7.54 h |

Durations are calculated based on manual turn boundaries.

Table 2.2: Size of the AMI dataset.

| | **AMI-train** | **AMI-dev** | **AMI-test** |
|---|---|---|---|
| #meetings | 98 | 20 | 20 |
| duration-PM | 16.00 h | 2.95 h | 3.93 h |
| duration-ME | 10.22 h | 2.61 h | 2.51 h |
| duration-UI | 9.71 h | 2.26 h | 1.79 h |
| duration-ID | 11.03 h | 2.02 h | 2.15 h |

Durations are calculated based on manual turn boundaries.

In order to train the required LMs we use the training parts of the PSYCH and AMI corpora, as well as the Fisher English corpus (Cieri, Miller, & Walker, 2004) and the transcribed therapy sessions provided by the counseling and psychotherapy transcripts series[9] (CPTS), as described in Section 2.5. The size of the corresponding vocabularies and the total number of tokens are given in Table 2.3.

Table 2.3: Size of the vocabulary and total number of tokens in the corpora used for LM training.

| | **PSYCH-train** | **AMI-train** | **Fisher** | **CPTS** |
|---|---|---|---|---|
| vocabulary size | 8.17K | 8.54K | 58.6K | 35.6K |
| #tokens | 530K | 479K | 21.0M | 6.52M |

## 2.5 Experiments and Results

First, we train all the necessary LMs, which are 3-gram models with Kneser-Ney smoothing. The generic LM $\mathcal{G}$ is trained using the Fisher English corpus. For the AMI corpus, the 4 role-specific LMs

[9] https://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series

$\mathcal{R}_{PM}, \mathcal{R}_{ME}, \mathcal{R}_{UI}, \mathcal{R}_{ID}$ are trained using only the turns belonging to the corresponding roles in the training set. For the PSYCH corpus, we additionally use the CPTS sessions and get the role-specific LMs $\mathcal{R}_T = w_{o_T}\mathcal{R}_{T,CPTS} \oplus (1 - w_{o_T})\mathcal{R}_{T,PSYCH}$ and similarly for $\mathcal{R}_C$. The mixing weights $w_{o_T}$ and $w_{o_C}$ are optimized so that the perplexity of the turns of the corresponding roles in the development set is minimized. Once we have those LMs, we create the mixed versions according to equations (2.8) and (2.9), where all the appearing mixing weights are again optimized to minimize the perplexity of the development corpora. For the optimization of $w_g$, the corresponding development corpus is the union of all the role-specific development corpora for the dataset we work with. The LM training and weight optimization is done with the SRILM toolkit (Stolcke, 2002). The size of the final mixed vocabulary is 69.5K for the experiments with the PSYCH corpus and 59.6K for the experiments with the AMI corpus, while the phonetic representation of those words is given by the CMU dictionary[10].

The ASR decoding is done with the Kaldi speech recognition toolkit (Povey et al., 2011) using Kaldi's pre-trained ASpIRE acoustic model[11]. The word insertion penalty and the LM weighting factor used during decoding are chosen to minimize the word error rate (WER) on the development set. The evaluation metric used for the final role assignment is the misclassification rate (MR), as defined in equation (1.3).

### 2.5.1 Turn-level SRR

In Table 2.4 we present the results using our method (*lm-resc*) for turn-level (*tl*) SRR, as shown in Figure 2.2, as well as using the approach shown in Figure 2.3 (*lm-asr*) where the cost $c'_j(x)$ is the log-likelihood of the turn $x$ given the LM $\mathcal{R}_j^+$.

As we can see, both *lm-resc-tl* and *lm-asr-tl* fail to beat the baseline classifier which always chooses the majority class (from the training set) for the case of AMI corpus. For the 2-role problem in PSYCH corpus this is not the case, but still *lm-asr-tl* outperforms *lm-resc-tl*. This is because the corpora feature conversational interactions and thus, prior to speaker clustering, utterances are broken into very short speech segments. Each individual segment contains insufficient observations

[10] https://github.com/cmusphinx/cmudict

[11] https://kaldi-asr.org/models/m1

Table 2.4: MR (%) for turn-level SRR.

| | **lm-resc-tl** | **lm-asr-tl** | **maj. class** |
|---|---|---|---|
| PSYCH | 23.58 | **10.75** | 50.67 |
| AMI | 64.70 | 63.40 | **62.22** |

*lm-resc-tl* refers to the system of Figure 2.2.
*lm-asr-tl* refers to the system of Figure 2.3.

to infer speaker role, and since all decisions are independent, that increases error. Such inaccuracies cancel out when we exploit the aggregate score for all the turns of a speaker as we will see in the following section.

### 2.5.2 Speaker-level SRR

Here, the final decision of the role assignment is taken at the speaker level, according to Algorithm 1, which means that a speaker clustering step is required. To that end, a BIC-based HAC is employed on top of an energy-based voice activity detector at the frame level, like in Chapter 1. In order for the clustering to make sense in the case of the AMI corpus, we downmix the 4 headset microphones into one audiofile per meeting. As observed in Table 2.5, our method (*lm-resc-sl*) yields improved results, outperforming both *lm-asr-sl* and the turn-level approaches (Table 2.4). Of course, the final performance depends on the performance of the clustering algorithm used.

Table 2.5: MR (%) for speaker-level (*sl*) SRR and for speaker clustering (BIC-HAC).

| | **lm-resc-sl** | **lm-asr-sl** | **BIC-HAC** |
|---|---|---|---|
| PSYCH† | **0.00** | 7.46 | – |
| PSYCH | **4.41** | 5.83 | 4.08 |
| AMI† | **29.46** | 55.52 | – |
| AMI | **46.16** | 60.94 | 15.63 |

† denotes the use of ground truth speaker clustering information.

### 2.5.3 Effect on speech recognition accuracy

Finally, we want to explore whether the role-specific lattice rescoring can lead to improved results for the task of ASR apart from SRR. To that end, for every turn we assume that the lexical information is given by the best path of the rescored lattice corresponding to the role that was assigned by our algorithm to that turn. The results in Table 2.6 show that this approach, following our per-speaker role assignment, can indeed slightly improve the ASR performance. The slight difference between the WER of the generic ASR model and the combination of the rescored ones, together with the substantial improvements in SRR performance (Table 2.5) suggest that even small role-specific improvements in the text produced by the ASR can be of high value for a reliable role identification.

Table 2.6: WER (%) using the best path of a generic lattice or role-specific rescored lattices.

| | **lm-resc-tl** | **lm-resc-sl** | **generic** |
|---|---|---|---|
| PSYCH | 37.84 | **37.54** | 37.99 |
| AMI | 29.35 | **29.27** | 29.29 |

## 2.6 Conclusion

Here we presented an algorithm that rescores the lattices produced by an ASR system with role-specific LMs in order to exploit the linguistic information in a more robust way for the task of SRR. We experimented with approaches taking the final decision both at the turn and at the speaker level and we identified that the second case leads to more reliable results. This chapter concludes our analysis on how to robustly extract speaker roles from the speech signal. In Chapters 3 and 4 we will focus on how to use the role information in order to improve the performance of a fundamental speech processing task, that of *speaker diarization*. There, we are going to use weaker approaches to extract speaker roles (since we will do so at the turn level for only a few turns) and for that reason we will employ specific confidence criteria.

# Part II

# Using Speaker Roles

# Chapter 3

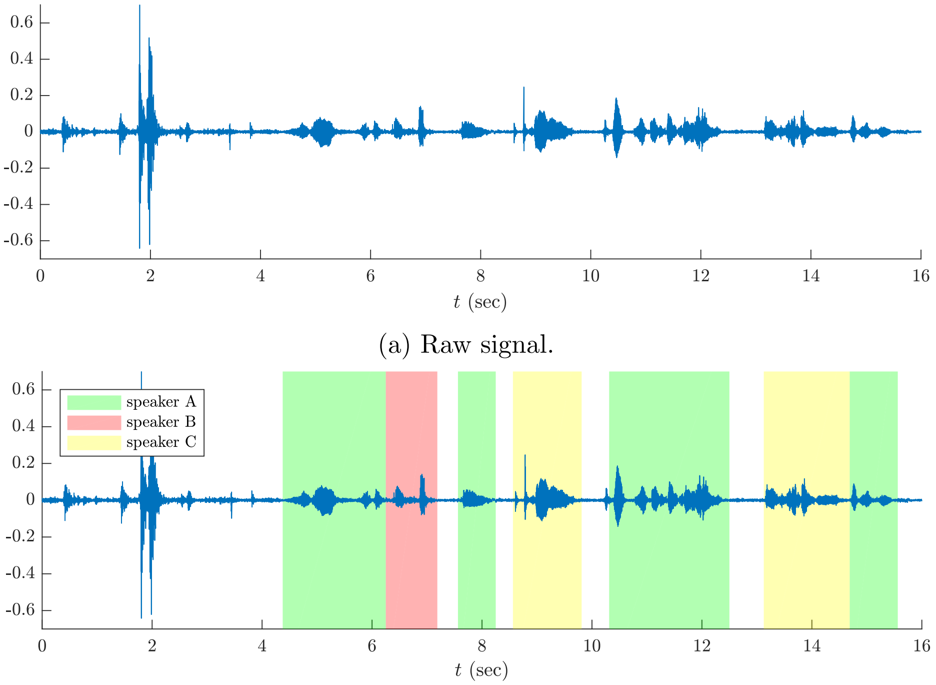
# Linguistically Aided Speaker Diarization Using Speaker Role Information

In the previous chapter we demonstrated how to infer speaker roles from speech recognition outputs and additionally showed that speaker role information can improve the performance of an ASR system. In this chapter, we utilize speaker roles to facilitate another speech processing task, namely speaker diarization. This task relies on the assumption that speech segments corresponding to a particular speaker are concentrated in a specific region of the speaker space; a region which represents that speaker's identity. These identities are not known a priori, so a clustering algorithm is typically employed, traditionally based solely on audio. Under noisy conditions, however, such an approach poses the risk of generating unreliable speaker clusters. Here we aim to utilize linguistic information as a supplemental modality to identify the various speakers in a more robust way. In particular, we show that the different linguistic patterns that speakers are expected to follow in role-based conversational scenarios can help us construct the speaker identities. That way, we are able to boost diarization performance by converting the clustering task to a classification one.

The work presented in this chapter has been published in (Flemotomos, Georgiou, & Narayanan, 2020).

## 3.1 Introduction

Given a speech signal with multiple speakers, diarization answers the question “who spoke when” (Anguera et al., 2012). To address the problem, the main underlying idea is that speech segments corresponding to some speaker share common characteristics which are ideally unique to the particular person. So, the problem is usually reduced to finding a suitable representation of the signal and a reliable distance metric. Under this viewpoint, when the distance between two speech segments is beyond a certain threshold, they are considered to belong to different speakers. The job of a speaker diarization system is visually depicted in Figure 3.1.



(a) Raw signal.

(b) Diarization output.

Figure 3.1: Finding “who spoke when” in a speech signal. In (b), the white regions indicate silence or noise. The 5 detected (colored) speech regions are further segmented into 7 speaker-homogeneous segments which are clustered into 3 same-speaker groups.

In the conventional diarization approach, the input signal is first segmented either uniformly (e.g., Sell et al., 2018) or according to a speaker change detection algorithm (e.g., Zajıc, Kunešová, Zelinka, & Hrúz, 2018). In either case, it is assumed that a single speaker is present in each one of the resulting segments. Since diarization is typically viewed as an unsupervised task, it heavily depends on the successful application of a clustering algorithm in order to group same-speaker segments together. Such a method, however, poses the risk of creating noisy, non-representative speaker clusters. In particular, if the speakers to be clustered reside closely in the speaker space, some speakers may be merged. Additionally, if there is enough noise and/or silence within a recording (possibly not sufficiently captured by a voice activity detection algorithm), it may be the case that one of the constructed clusters only contains the non-speech or distorted-speech segments. This behavior can lead to poor performance even if the exact number of speakers is known in advance.

Even though speaker diarization has traditionally been an audio-only task which relies on the acoustic variability between different speakers, the linguistic content captured in the speech signal can offer valuable supplementary cues. Apart from practical observations such as the fact that it is highly improbable for a speaker change point to be located within a word (Dimitriadis & Fousek, 2017; Silovsky, Zdansky, Nouza, Cerva, & Prazak, 2012), it is widely accepted that each individual has their very own way of using language (Johnstone, 1996). Thus, language patterns followed by individual speakers have been explored in the literature for the tasks of speaker segmentation and clustering, both when used unimodally (Meng, Mou, & Jin, 2017), and in combination with the speech audio (India Massana, Rodríguez Fonollosa, & Hernando Pericás, 2017; Park & Georgiou, 2018; Park, Han, Huang, et al., 2019; Zajíc, Soutner, Hrúz, Müller, & Radová, 2018).

Despite the beneficial effects of using language as an additional stream of information, there is an important practical consideration: how to get access to the transcripts. In a real-world scenario, a high-performing ASR system needs to be applied before any textual data is available. However, speaker diarization is widely viewed as a pre-processing step of multi-talker ASR systems and is often a module that precedes ASR in conversational speech processing pipelines (Huang, Marcheret, Visweswariah, Libal, & Potamianos, 2007; Xiao, Huang, et al., 2016). This is because single-speaker speech segments allow for speaker normalization techniques, including speaker adaptive training through constrained maximum likelihood linear regression (CMLLR; Gales, 1998) and i-vector based neural network adaptation (Saon, Soltau, Nahamoo, & Picheny, 2013). Nevertheless, taking into consideration the error propagating from a non-ideal diarization system to the ASR output, it is nowadays questionable whether diarization can in practice improve recognition accuracy, which is why several modern pipelines start by applying ASR first, achieving state-of-the-art results (Park, Han, Huang, et al., 2019; Yoshioka et al., 2019). In any case, if there are not major computational and/or time constraints, running a second pass of ASR after diarization could be a reasonable approach[1].

Following the aforementioned line of work, we propose an alternative way of using the linguistic information for the task of speaker diarization in recordings where participants play specific roles which are known in advance. In particular, we process the text stream independently in order to segment it in speaker-homogeneous chunks (where only one speaker is active), each one of which can be assigned to one of the available speaker roles. Aggregating this information for all the segments, and aligning text with audio, we can construct the acoustic identities of the speakers found in the recording. That way, each audio segment can be assigned to a speaker through a simple classifier, overcoming the potential risks of clustering. We apply this approach in psychotherapy recordings featuring dyadic interactions between two speakers with well-defined roles; namely those of a *therapist* and a *patient*.

## 3.2 Background: Audio-Only Speaker Diarization

Speaker diarization is the process of partitioning a speech signal into speaker-homogeneous segments and then grouping same-speaker segments together, without having prior information about the speaker identities. Therefore, research effort has been focused on finding i) a representation that can capture speaker-specific characteristics, and ii) a suitable distance metric that can separate different speakers based on those characteristics. The traditional approach has been to model speech segments under some probability distribution (e.g., GMMs), and measure the distance between them using a metric such as the one based on the bayesian information criterion (BIC) (S. Chen & Gopalakrishnan, 1998).

Speaker modeling by GMMs was later replaced by i-vectors (Shum, Dehak, Chuangsuwanich, Reynolds, & Glass, 2011), fixed-dimensional embeddings inspired by the total variability model. In this framework, the cosine distance metric was initially proposed as the divergence criterion to be used, but probabilistic linear discriminant analysis (PLDA) based scoring (Ioffe, 2006; Prince & Elder, 2007) was proved to yield improved results (Sell & Garcia-Romero, 2014). Given two

[1] In Chapter 5 we will see how diarization and ASR can be connected within a larger speech processing pipeline.

embeddings $v$, $r$, PLDA provides a framework to estimate their similarity $s(v,r)$ as the log-likelihood ratio

$$s(v,r) = \log \frac{p(v,r|\text{same speaker})}{p(v|\text{dif. speakers})p(r|\text{dif. speakers})} \tag{3.1}$$

In recent years, with the advent of deep neural networks (DNNs), the embeddings used are usually bottleneck features extracted from neural architectures. Such architectures are trained under the objective of speaker classification, employing a cross-entropy loss function (Snyder, Garcia-Romero, Sell, Povey, & Khudanpur, 2018), or under the objective of speaker discrimination employing contrastive (Garcia-Romero, Snyder, Sell, Povey, & McCree, 2017) and triplet (Bredin, 2017b) loss functions. Typical examples of embeddings that have shown state-of-the-art performance for speaker diarization are the long-short term memory (LSTM) based d-vectors (Q. Wang, Downey, Wan, Mansfield, & Moreno, 2018) and the time-delay neural net (TDNN) based x-vectors (Sell et al., 2018), which are also the embeddings used for the work presented here. The first layers of the architecture used to extract x-vectors operate at the frame level, with deeper layers seeing longer temporal contexts. Then, a statistics pooling layer is used to collect the outputs of the last layer of the TDNN and compute the mean and standard deviation vectors. The next few dense layers operate at the segment level before a softmax inference layer maps segments to speaker labels. The activations of the first dense layer are selected as speaker embeddings.

Speaker diarization usually comprises two steps: first, the speech signal is segmented into single-speaker chunks, and second, the resulting segments are clustered into same-speaker groups (Anguera et al., 2012). Even though speaker change detection is by itself an active research field (Hrúz & Zajíc, 2017; Jati & Georgiou, 2017), it has been shown that it doesn't necessarily lead to improved results within the framework of diarization when compared to a uniform, sliding-window based segmentation (Zajíc, Kunešová, & Radová, 2016; Zajıc et al., 2018), so the latter method is widely used. As far as the clustering is concerned, common approaches include hierarchical agglomerative clustering (HAC; Sell et al., 2018) and spectral clustering (Park, Kumar, et al., 2019; Q. Wang et al., 2018), while methods based on affinity propagation (Yin, Bredin, & Barras, 2018) and generative adversarial networks (Pal et al., 2020) have also been proposed. In order to overcome some of the problems connected with clustering, supervised systems that directly output a sequence of speaker labels have been recently introduced (Fujita et al., 2019; Zhang, Wang, Zhu, Paisley, & Wang, 2019).

## 3.3 Proposed Method: Linguistically-Aided Speaker Diarization

Our proposed approach for speaker diarization in conversational interactions where speakers assume specific roles is illustrated in Figure 3.2. We describe the various modules in detail in Sections 3.3.1–3.3.4.

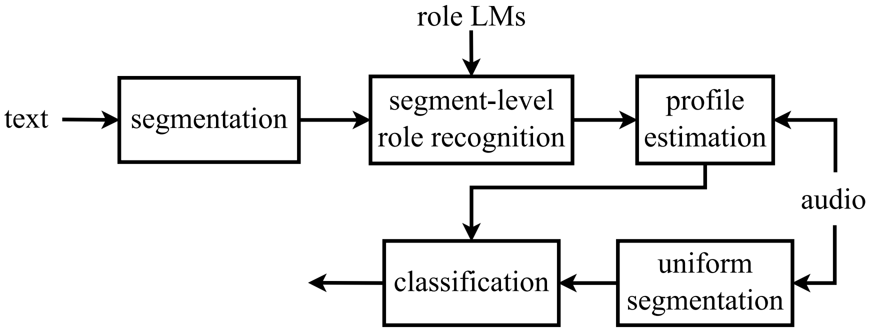


Figure 3.2: Linguistically-aided speaker diarization using role information.

### 3.3.1 Text-based segmentation

Given the textual information of the conversation, our goal is to obtain speaker-homogeneous text segments; that is segments where all the words have been uttered by a single speaker. Those will later help us construct the desired acoustic speaker identities. Even though text-based speaker change detectors have been proposed (Meng et al., 2017), for our final goal we can safely over segment the available document, provided this leads to a smaller number of segments containing more than one speakers (Zajíc et al., 2018). So, we assume that each sentence is with high probability speaker-homogeneous and we instead segment at the sentence level.

To that end, the problem can be viewed as a sequence labeling one, where each word is tagged as either being at the beginning of a sentence, or anywhere else. In particular, we address the problem building a Bidirectional LSTM (BiLSTM) network with a conditional random field (CRF) inference layer (Ma & Hovy, 2016), as shown in Figure 3.3. The input to the recurrent layers is a sequence of words. Each word is given as a concatenation of a character-level representation predicted by a CNN and a word embedding. For our experiments, we initialize the word embeddings with the extended dependency skip-gram embeddings (Komninos & Manandhar, 2016), pre-trained on 2B words of Wikipedia. Those extend the semantic vector space representation of the word2vec model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) considering not only spacial co-occurrences of words within text, but also co-occurrences in a dependency parse graph. That way, they can capture both functional and topic-related semantic properties of words.

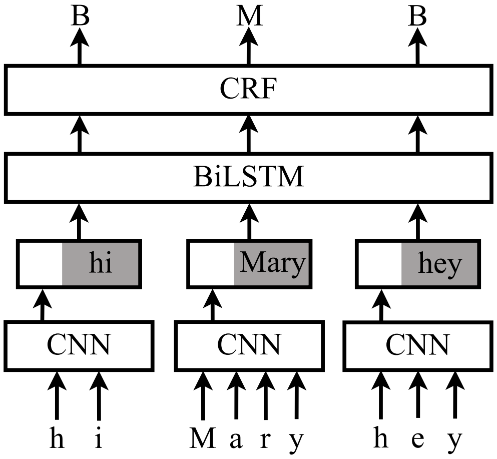


Figure 3.3: Neural network for sentence-level text segmentation: A character representation is constructed for each word through a CNN and is concatenated with a word embedding (here shown in grey). This is the input to a BiLSTM-CRF architecture which predicts a sequence of labels. Here $B$ denotes a word at the beginning of a sentence and $M$ in the middle (any word which is not the first one of a sentence).

### 3.3.2 Role recognition

The next step in our system is the application of a text-based role recognition module. In more detail, assuming we have $N$ speakers in the session ($N = 2$ for our experiments) and there is one-to-one correspondence between speakers and roles (e.g., there is one therapist and one patient), we want to assign one of the role labels $\{R_i\}_{i=1}^N$ to each segment. To do so, we build $N$ LMs $\{\mathcal{R}_i^+\}_{i=1}^N$, one for each role, and we estimate the perplexity of a segment given the LM $\mathcal{R}_i^+$, for $i = 1, 2, \cdots, N$. The role assigned to the segment is the one yielding the minimum perplexity like in Chapters 1 and 2. We note that in our experiments all the perplexities are normalized for segment length.

The required role-specific LMs are n-gram models built as described in Chapter 2, Section 2.3. For this process, we assume that in-domain text data is available for training. We first construct a background, out-of-domain LM $\mathcal{G}$ and $N$ role-specific LMs $\{\mathcal{R}_i\}_{i=1}^N$. $\mathcal{G}$ is used to ensure a large enough vocabulary that minimizes the unseen words during the test phase. Those individual LMs are interpolated to get the mixed models $\{\mathcal{R}_i^+\}_{i=1}^N$[2].

### 3.3.3 Profile estimation

After applying the text-based segmenter and role recognizer, we have several text segments corresponding to each role $R_i$. If we have the alignment information at the word level[3], we can directly get the time-boundaries of those segments. We extract one embedding (x-vector) for each and we estimate a *role* identity $r_i$ as the mean of all the embeddings corresponding to the specific role. Under the assumption of one-to-one correspondence between speakers and roles that we already introduced in Section 3.3.2, those role identities are at the same time the acoustic identities (also known as profiles) of the *speakers* appearing in the initial recording.

We note that role recognition at the segment level does not always provide robust results as explained in the previous chapters, something which could lead to unreliable generated profiles. However, we expect that there will be a fraction of the segments for the results of which we are confident enough and we can take only those into consideration for the final averaging. The proxy used as our confidence for the segment-level role assignment is the difference between the best and the second best perplexity of a segment given the various LMs, similarly to the confidence introduced in Algorithm 1 (Chapter 2). Formally, if segment $x$ is assigned the role $R_i$, and if $pp(x|\mathcal{R}_i^+)$ is the perplexity of $x$ given the LM $\mathcal{R}_i^+$, then the confidence metric used for this assignment is

$$c_x = \min_{j \neq i} |pp(x|\mathcal{R}_j^+) - pp(x|\mathcal{R}_i^+)| \tag{3.2}$$

Then, the corresponding profile is

$$r_i = \frac{\sum_{x \in R_i} \tilde{c}_x u_x}{\sum_{x \in R_i} \tilde{c}_x} \triangleq \frac{\sum_{x \in R_i} \mathbb{I}\{c_x > \theta\} u_x}{\sum_{x \in R_i} \mathbb{I}\{c_x > \theta\}} \tag{3.3}$$

where $u_x$ is the x-vector for segment $x$, $\mathbb{I}\{\cdot\}$ is the indicator function, $\theta$ is a tunable parameter.

[2] For more details, please refer to equations (2.8)–(2.9).

[3] If we have access to the transcripts and the audio, we can force-align. If we generate the text through ASR, we get the desired alignments from the decoding lattices.

### 3.3.4 Audio segmentation and classification

After having computed all the needed profiles $\{r_i\}_{i=1}^N$, in order to perform speaker diarization, we first segment the audio stream of the speech signal uniformly with a short sliding window, a typical approach in audio-only diarization systems. In other words, the language information is used by our framework only to construct the speaker profiles, with the final diarization result relying on audio-based segmentation, as illustrated in Figure 3.2. For each one of the resulting segments an x-vector is extracted. However, instead of clustering the x-vectors, we now classify them within the correct speaker/role. In order to have a fair comparison between common diarization baselines and our proposed system, our classifier is based on PLDA, but we note that any other classifier could be employed instead. In this framework, a segment $x$ with embedding $u_x$ is assigned the label

$$\hat{R}_x = \underset{1\leq i\leq N}{\operatorname{argmax}}\{s(u_x, r_i)\} \tag{3.4}$$

where $s(\cdot,\cdot)$ is the PLDA similarity score estimated in equation (3.1).

## 3.4 Datasets

### 3.4.1 Evaluation data

We evaluate our proposed method on datasets from the clinical psychology domain. In particular, we apply the system to the motivational interviewing sessions introduced in the previous chapters, and specifically the PSYCH corpus described in Table 2.1. As explained there, the train/dev/eval split has been done in such a way that there is no speaker overlap between the subsets. All the results reported are on PSYCH-test.

### 3.4.2 Segmenter and role LM training data

The segmenter presented in Section 3.3.1 is trained on a subset of the Fisher English corpus (Cieri et al., 2004) comprising a total of 10,195 telephone conversations for which the original transcriptions (including punctuation symbols which are essential for the training of our network) are available. This set is enhanced by 1,199 in-domain therapy sessions provided by the counseling and psychotherapy transcripts series (CPTS). The combined dataset is randomly split (80-20 split at the session level) into training and validation sets. Here, we use the same role-specific LMs as the ones trained and used in Chapter 2, employing CPTS and the entire Fisher English corpus for training. Please refer to Table 2.3 for details on the size of the corresponding vocabularies.

## 3.5 Experiments and Results

### 3.5.1 Baseline systems

#### Audio-based diarization with speaker clustering

As an audio-only baseline, we use a diarization system following the widely applied x-vector/PLDA paradigm (Sell et al., 2018). As shown in Figure 3.4, the speech signal is first segmented uniformly and an x-vector is extracted for each segment. The pairwise similarities $s(\cdot, \cdot)$ between all those embeddings are then calculated based on PLDA scoring (equation (3.1)).

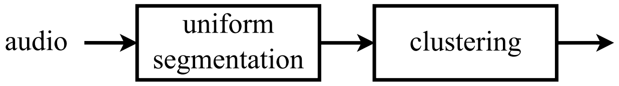


Figure 3.4: Baseline audio-based speaker diarization.

The segments are clustered into same-speaker groups following a HAC approach with average linking. Since our experiments are conducted on dyadic interactions, we force the HAC algorithm to run until two clusters are constructed.

#### Language-based diarization

As a language-only baseline, we use the system of Figure 3.5, which essentially consists of the first steps of the framework in Figure 3.2. After estimating the segment-level role labels as described in Sections 3.3.1 and 3.3.2, we can simply use those as our diarization output labels to evaluate the performance of a system that only depends on linguistic information. In that case, we only utilize audio to get the timestamps of the text segments. If an ASR system is used, this information is already available through the decoding lattices.

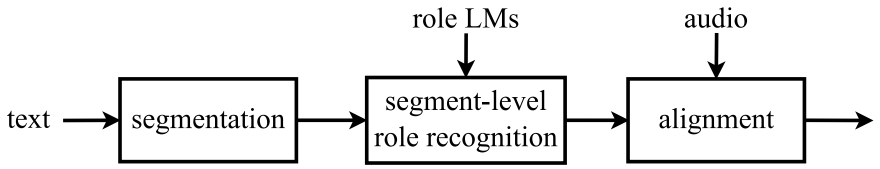


Figure 3.5: Baseline language-based speaker diarization.

### 3.5.2 Experimental setup

As a pre-processing step, the text which is available from the manual transcriptions of the PSYCH corpus is normalized to remove punctuation symbols and capital letters, and force-aligned with the corresponding audio sessions. Based on the word alignments, we segment the audio according to whether there is a silence gap between two words larger than a threshold equal to 1 sec. We should highlight that this initial segmentation is applied before running either one of the baseline systems or our proposed architecture. Thus, the initial segments to be diarized are always the same and those are also the segments that we pass to the ASR system. The diarization ground truth is also constructed through the word alignments, by allowing a maximum of 0.2 sec-long in-turn silence.

The resulting text segments are further subsegmented at the sentence level based on the output of the tagger in Figure 3.3. During training we define as "sentence" any text segment between two punctuation symbols denoting pause, apart from commas. We exclude commas first because they normally do not indicate speaker change points but also because they are too frequent in our training set and they would lead to very short segments, not containing sufficient information for the task of role recognition. The tagger is built using the NCRF++ toolkit (Yang & Zhang, 2018). Following the general recommendations by Reimers and Gurevych (2017) and after our own hyperparameter tuning, the network comprises 4 CNN layers and 2 stacked BiLSTM layers with dropout ($p = 0.5$) and $l_2$ regularization ($\lambda = 10^{-8}$). The length of each word representation is 330 (character embedding dimension = 30, word embedding dimension = 300). The network is trained using the Adam optimizer with a fixed learning rate equal to $10^{-3}$ and a batch size equal to 256 word sequences. The tagger achieves an $F_1$ score of 0.805 on the validation set after 14 training epochs.

All the LMs required for role recognition are 3-gram models with Kneser-Ney smoothing built with the SRILM toolkit, as described in Section 2.5. The audio-based diarization framework is built using the Kaldi toolkit (Povey et al., 2011). We use the VoxCeleb pre-trained x-vector extractor[4] and the PLDA model which comes with it, after we adapt it on the development set of the PSYCH corpus, both for the audio-only baseline and for our linguistically-aided system. The x-vectors are extracted after uniformly segmenting the audio into 1.5 sec-long windows with a window shift equal to 0.25 sec. Those are normalized and decorrelated through a linear discriminant analysis (LDA) projection and dimensionality reduction (final embedding length = 200), mean, and length normalization. The evaluation is always based on the diarization error rate (DER), as estimated by the NIST `md-eval.pl` tool, with a 0.25 sec-long collar, ignoring overlapping speech. DER incorporates three sources of error: false alarms (speech in the output but not in the ground truth), missed speech (speech in the ground truth but not in the output), and speaker confusion (speech assigned to the wrong speaker cluster).

To get the ASR outputs, we use Kaldi's pre-trained ASpIRE acoustic model[5], coupled with the 3-gram LM given in equation (2.8). This ASR system gives a word error rate (WER) equal to 38.02% for the PSYCH-dev and 39.78% for the PSYCH-test set. It is noted that WERs in this range are typical in spontaneous medical conversations (Kodish-Wachs, Agassi, Kenny III, & Overhage, 2018).

### 3.5.3 Results with reference transcripts

Before applying ASR, we employ our system using the manually derived transcripts. That way, we can inspect the usability and effectiveness of our approach, eliminating potential propagation errors because of ASR. Table 3.1 gives the results of our linguistically-aided diarization system in comparison with the audio-only and language-only baseline approaches. First, we notice that, between the two baselines, the one using the acoustic modality yields better results. This came at no surprise since we expected that audio carries the most important speaker-specific characteristics. Hence, we propose using language only as a supplementary stream of information.

When we apply our linguistically-aided system using our sequence tagger to segment at the sentence level (still using all the segments, without applying any confidence criterion) we get a

[4] https://kaldi-asr.org/models/m7

[5] https://kaldi-asr.org/models/m1

Table 3.1: DER (%) following our linguistically-aided approach and the two baselines.

| **transcript source** | **text segmentation** | **audio only** | **language only** | **linguistically aided** | **linguistically aided**† |
|---|---|---|---|---|---|
| reference | oracle | 11.05 | 12.99 | 7.28 | 6.99 |
| | tagger | | 20.09 | 7.71 | 7.30 |
| ASR | tagger | 11.05 | 27.07 | 8.37 | 7.84 |

The text segmentation (when needed) is either performed by our sequence *tagger* or based on the manually annotated speaker changes (*oracle*).

† denotes results when only $a\%$ of the segments we are most confident about are taken into account in each session for the profile estimation, where $a$ is a parameter optimized on the development set.

30.23% DER relative improvement compared to the audio-only approach. In the first row of Table 3.1 we additionally report results when using the oracle speaker segmentation provided by the manual annotations instead of applying the sequence tagger. That way, we can eliminate any negative effects caused by a suboptimal speaker change detector. As expected, the results are indeed better, but it is worth noting the difference in the performance gap between the language-only and the linguistically-aided approaches when we compare the oracle vs. the tagger-based segmentation. Since the sequence tagger operates at the sentence level, its output is over-segmented with respect to speaker changes. As a result, utterances are broken into very short segments, with several segments containing insufficient information to infer speaker role in a robust way. However, when we aggregate all those speaker turns to only estimate an average speaker profile, such inaccuracies cancel out.

Further improvements are observed if for profile estimation we only keep the segments we are most confident about, applying the confidence metric introduced in Section 3.3.3. Instead of directly optimizing for the parameter $\theta$ appearing in equation (3.3), we find the parameter $a$ that minimizes the overall DER on the development set when only the $a\%$ segments we are most confident about are taken into consideration per session. The results on the test set are reported in the last column of Table 3.1 ($a = 70$ for the tagger segmentation and $a = 55$ for the oracle segmentation, after optimizing on the development set). An additional 5.32% relative error reduction is achieved when our tagger is used and similar improvements are noticed in the case of oracle text segmentation.

In Figure 3.6 we plot DER as a function of the percentage of the segments we use to estimate the speaker profiles within a session. Even though the oracle text segmentation consistently yields marginally better results, it seems that if we carefully choose which segments to use to get an estimate of the speakers' identities, our tagger-based segmentation approaches the oracle performance. In fact, the best result we got on the test set (optimizing for $a$ on the same set) using our segmenter was 7.13% DER, while the corresponding number using the oracle segmentation was 6.99%. We should highlight here that the analysis presented in this work is based on using $a$% of the segments within a session, after choosing some $a$ which remains constant across sessions. It is probable that this is a session-specific parameter which ideally should be chosen based on an alternative, session-level strategy.

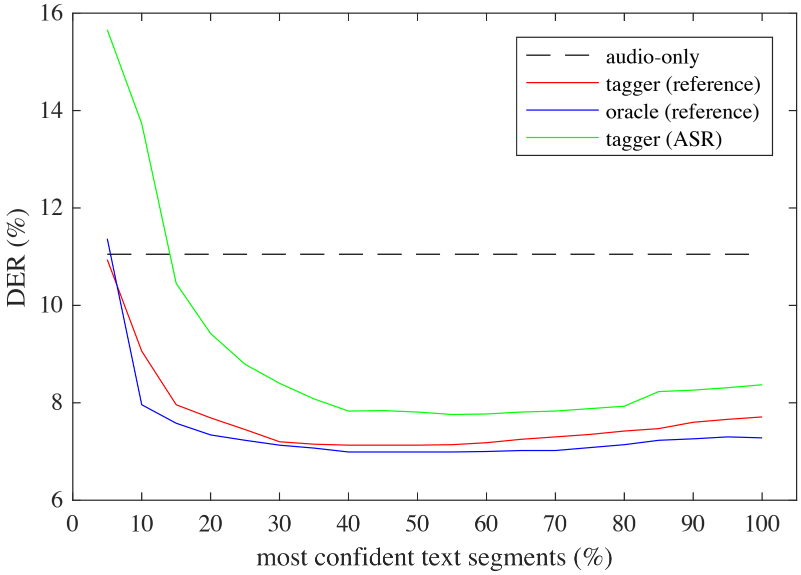


Figure 3.6: DER (%) as a function of the number of text segments we take into account per session for the profile estimation, based on our confidence metric. Text segmentation is either performed by our sequence *tagger* or based on the manually annotated speaker changes (*oracle*). Results are presented both with *reference* and with *ASR* transcripts.

### 3.5.4 Results with ASR transcripts

For the experiments in this Section we apply the same pre-processing steps, but we replace the reference transcripts with the textual outputs of the ASR system and the corresponding time alignments. The results are given in the last row of Table 3.1. Here, we report results only when using the sequence tagger (and not with oracle segmentation), simply because we now assume we have no access to the reference transcripts, so we cannot know the oracle speaker change points.

As we can see, the diarization performance is substantially improved compared to the audio-only system (relative DER reduction equal to 24.25%) even if the WER of the ASR module is relatively high, as reported in Section 3.5.2. It seems that when using the transcripts only for the task of profile estimation, the overall performance is not severely degraded by a somehow inaccurate ASR system. This is not the case for our language-only baseline. Since in that case the final output only depends on linguistic information, the performance gap between using manual and ASR-derived transcripts (*language-only* column in Table 3.1) is large. We should note that this performance gap is not only due to higher speaker confusion in the case of ASR transcripts, but also because of increased missed speech. In particular, the missed speech when using ASR is 2.7% because of word deletions (as opposed to 0.6% when the reference transcrpits and the tagger are used).

As was the case with the experiments in Section 3.5.3, further improvements are observed when only using a subset of the total number of segments per session to estimate the speaker profiles. In particular, if $a = 45\%$ of the segments for which we are most confident about (after optimizing for $a$ on the development set) are used, DER is reduced to 7.84%. The beneficial effects of using our confidence metric to estimate a speaker representation only by a subset of their assigned speech segments is also demonstrated in Figure 3.6.

## 3.6 Conclusion

We proposed a system for speaker diarization suitable to use in conversations where participants assume specific roles, associated with distinct linguistic patterns. While this task typically relies on clustering methods which can lead to noisy speaker partitions, we demonstrated how we can exploit the lexical information captured within the speech signal in order to estimate the speaker profiles and follow a classification approach instead. A text-based speaker change detector is an essential component of our system. For this subtask, assuming each sentence is speaker-homogeneous, we proposed using a sequence tagger which segments at the sentence level, by detecting the beginning of a new sentence and we showed that this segmentation strategy approaches the oracle performance. The resulting segments are assigned a speaker role label which is later used to construct the desired speaker identities and we introduced a confidence metric to be associated with this assignment. Our results showed that such a metric can be used in order to take into consideration only the segments we are most confident about, leading to further performance improvements. When applied to dyadic interactions between a therapist and a patient, our proposed method achieved an overall relative DER reduction equal to 29.05%, compared to the baseline audio-only approach with speaker clustering. When reference transcripts were used instead of ASR outputs, the corresponding overall reduction was equal to 33.94%.

Since role recognition is a supervised task, one drawback of our system when compared to traditional diarization approaches is that it requires in-domain text data in order to build the role-specific LMs. It should be additionally highlighted that the diarization results can be further improved if, for example, a re-segmentation module is employed as a final step, or a more precise audio segmentation strategy is followed instead of relying on uniform segmentation. For instance, an audio-based speaker change detector could be applied both for the audio-only baseline and the linguistically-aided system and in the latter case this could be used in combination with the language-based segmenter. However, our goal in this chapter was mainly to demonstrate the effectiveness of constructing the speaker profiles within a session to be diarized in order to convert the clustering task into a classification one and not to achieve the best possible diarization performance. Additionally, since the initial segmentation was the same both for our system and our audio-only baseline, we expect that any improvements with respect to that part (i.e. more sophisticated segmentation and/or application of re-segmentation techniques) would lead to similar relative improvements to both systems.

Here we essentially modelled each speaker by a single embedding, since for the final profile estimation we averaged over all the speech segments assigned to the corresponding speaker. A potential extension of the current work would be an exploration of alternative speaker identity construction strategies, e.g., representing a speaker by a distribution of embeddings. This is particularly promising in scenarios where recordings are long enough so that they may incorporate various acoustic conditions or different speaking styles corresponding to the same speaker. In any case, to construct the speaker profiles based on roles, we had to assume there is one-to-one correspondence between speakers and roles within a conversation (for our experiments, one speaker assuming the role of *patient* and one speaker assuming the role of *provider*). However, there are domains where such an assumption does not hold. In the following chapter, we are going to study in depth such scenarios and we will provide an alternative role-based approach towards more robust speaker diarization.

# Chapter 4

# Multimodal Speaker Clustering with Role Induced Constraints

In the previous chapter, we introduced a methodology that utilizes speaker roles to reduce diarization from a clustering problem to a classification one, following a multimodal approach where both audio and text were taken into consideration. As we saw, the language used by the participants in a conversation carries information that can supplement the audio modality. However, we assumed that each speaker is linked to a unique speaker role, an assumption that we also followed in Chapters 1 and 2. In this chapter we propose an alternative approach where we employ a supervised text-based model to extract speaker roles and then use this information to guide an audio-based spectral clustering step by imposing must-link and cannot-link constraints between segments. The proposed method, which does not need the aforementioned assumption, is applied on two different domains, namely on medical interactions and on podcast episodes, and is shown to yield improved results when compared to the audio-only approach.

The work presented in this chapter has been submitted for publication. The pre-print is available (Flemotomos & Narayanan, 2022).

## 4.1 Introduction

Speaker diarization, as explained in Sections 3.1 and 3.2, is the task of segmenting a multi-party speech signal into speaker-homogeneous regions (and tagging them with speaker-specific labels) and is a critical component of several applications, including speaker-attributed speech recognition, audio indexing, and speaker tracking (Anguera et al., 2012; Park et al., 2022). Even though recently introduced end-to-end neural diarization offers simplicity and achieves remarkable results in some scenarios (Fujita et al., 2019; Horiguchi, Fujita, Watanabe, Xue, & Nagamatsu, 2020), modular, clustering-based diarization is still widely used and has been an indispensable part of award-winning systems in recent challenges (Medennikov et al., 2020; Y. Wang et al., 2021).

In the conventional diarization approach, the speech signal is first segmented into chunks which are assumed to be speaker-homogeneous, in the sense that a single speaker is active therein. Speaker representations, typically bottleneck feature vectors obtained from a speaker classification neural network (Dawalatabad et al., 2021; Koluguri, Park, & Ginsburg, 2022; Snyder et al., 2018), are then estimated for all the segments and their pairwise similarities are computed. A clustering algorithm that gives the desired labeled speech segments is finally employed. Even though it is generally assumed that no information is known a priori about the speakers, in practice we often need to deploy diarization systems in specific applications, and domain-dependent processing can be used to further improve the final performance. To that end, both the acoustic (e.g., Y. Wang et al., 2021) and the linguistic (e.g., Chapter 3) streams of information can be exploited to either adapt the models or modify the diarization pipeline. The language-based approach, where the transcripts of a recording are taken into consideration during diarization, is especially promising for interactions where speakers play dissimilar roles. It should be noted that several role-playing conversations, such as interviews, clinical interactions, and court hearings, have been included in the evaluation data of recent diarization challenges (Ryant et al., 2021).

In Chapter 3 we used language to identify the roles associated with different speech segments, estimate the acoustic profiles of the participants in the conversation, and eventually reduce the clustering problem into a classification one. Along similar lines, in Chapter 1 we ran an audio-based speaker clustering and a language-based role recognition module in parallel and then combined their outputs through a meta-classifier. However, those systems assume a one-to-one correspondence between speakers and roles, i.e., every speaker is linked to a unique role during a conversation. Even though this is a reasonable assumption in multiple domains (e.g., medical domain with dialogues between a *clinician* and a *patient*), the systems cannot be easily generalized when a single speaker assumes multiple roles or when multiple speakers play the same role (e.g., trials with a single *judge*, a group of *co-defendants*, and multiple *prosecution witnesses*).

To overcome this limitation, here we propose to exploit the linguistically extracted role information only to impose constraints during audio-based clustering. Depending on the domain, we can impose must-link and/or cannot-link constraints, without the need for one-to-one correspondence between speakers and roles. In particular, we use a BERT-based classifier to extract speaker role information from text and we then impose a list of pairwise constraints between segments linked to the same roles or different ones. Using manually-derived speaker-homogeneous segments with oracle transcriptions, we evaluate the effectiveness of the approach on the clustering performance by running experiments on two different domains: i) dyadic clinical interactions, where the roles of interest are the ones of the *therapist* and the *patient*, and ii) multi-party interactions from a weekly radio show with only partial role information available, where the role of interest is the *host*.

## 4.2 Background and Prior Work

### 4.2.1 Spectral clustering for speaker diarization

Clustering is one of the main components in modular speaker diarization. During that step, speech segments are grouped into same-speaker classes, usually following either a HAC (Sell et al., 2018), or a spectral clustering (Q. Wang et al., 2018) approach. This grouping is based on the pairwise similarities between the $N$ segments to be clustered, which are stored in an affinity matrix $\hat{\mathbf{W}} \in \mathbb{R}^{N\times N}$. In Chapter 3 we used PLDA to estimate the elements of the affinity matrix. Another common choice for estimating the affinities uses the cosine distance: given two speaker embeddings $\mathbf{v}_i$, $\mathbf{v}_j$, we have

$$\hat{\mathbf{W}}_{ij} = \frac{1}{2}\left(1 + \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{||\mathbf{v}_i|| \cdot ||\mathbf{v}_j||}\right) \tag{4.1}$$

which ensures that the affinities are in the range $[0, 1]$.

Having constructed the refined affinity matrix $\mathbf{W}$ (where refinements are explained later), spectral clustering is a technique that exploits the eigen-decomposition of $\mathbf{W}$ to project the $N$ elements onto a suitable lower-dimensional space (Ng, Jordan, & Weiss, 2001). To do so, we define the degrees $d_i \triangleq \sum_j \mathbf{W}_{ij}$ and we construct the normalized Laplacian matrix

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} \tag{4.2}$$

where $\mathbf{D} = \mathrm{diag}\{d_1, d_2, \cdots, d_N\}$. Assuming we know the number of speakers $k$, we find the $k$ eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$ corresponding to the $k$ smallest eigenvalues of $\mathbf{L}$ and form the matrix $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_k]$. After normalizing the rows of $\mathbf{X}$ to unit norm, so that $\tilde{\mathbf{X}}_{ij} = \mathbf{X}_{ij}/\sqrt{\sum_j \mathbf{X}_{ij}^2}$, we cluster the $N$ rows of $\tilde{\mathbf{X}}$ through a $k$-means algorithm and assign the original $l$-th segment to speaker $s$ if and only if the $l$-th row of $\tilde{\mathbf{X}}$ is assigned to speaker $s$.

In order to effectively use spectral clustering in diarization settings, several refinement operations have been proposed to be applied on the original affinity matrix (Park, Han, Kumar, & Narayanan, 2019; Q. Wang et al., 2018), the most notable being $p$-thresholding. Given the original affinity matrix $\hat{\mathbf{W}}$, the $(100 - p)\%$ largest values in each row are set to 1 and the rest are either binarized to 0 or multiplied by a small constant $\tau$ (soft thresholding), giving the modified matrix $\hat{\mathbf{W}}_p$. Since this operation may break the symmetry property of the affinity matrix, we re-symmetrize it to get

$$\mathbf{W} = \frac{1}{2}\left(\hat{\mathbf{W}}_p + \hat{\mathbf{W}}_p^T\right) \tag{4.3}$$

Instead of fixing a specific value $p$, an auto-tuning approach which uses the maximum eigengap of the Laplacian matrix can be followed (Park, Han, Kumar, & Narayanan, 2019). The eigengap criterion has its roots in graph theory and is also used to estimate the number of clusters (speakers) $\hat{k}$, when this is not known a priori. $\mathbf{L}$ is a positive semi-definite matrix with $N$ non-negative real eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. If $\mathbf{W}$ is viewed as an adjacency matrix of a graph with $\hat{k}$ perfectly connected components, then $\hat{k}$ equals the multiplicity of the eigenvalue $\lambda_1 = 0$. In practical applications, where we do not expect perfect components, $\hat{k}$ is estimated by the maximum eigengap:

$$\hat{k} = \underset{k}{\mathrm{argmax}} \frac{\lambda_{k+1}}{\lambda_k} \tag{4.4}$$

### 4.2.2 Constrained clustering for speaker diarization

Constrained clustering extends the traditional unsupervised learning paradigm of clustering by integrating supplemental information in the form of constraints (Gançarski, Dao, Crémilleux, Forestier, & Lampert, 2020). Even though several types of constraints have been explored, the most common ones are the instance-level relations, and in particular the must-link (ML) and cannot-link (CL) constraints. Under that viewpoint, if an ML (CL) constraint is imposed between two segments, then those segments must (must not) be in the same cluster.

In speaker diarization, constrained clustering has been applied with constraints imposed either by human input, or by acquired knowledge within a particular framework. C. Yu and Hansen (2017) propose a system where a sufficient number of segments corresponding to all the speakers are first identified by a human expert and the rest of the segments are clustered in a constrained fashion. Bost, Xavier and Linares, Georges (2014) apply a two-step clustering for audio-based speaker diarization in videos, where speakers are first clustered locally in scenes detected to contain dialogues, before a global clustering with CL constraints between segments locally assigned to different clusters. Similarly, in an effort to integrate end-to-end and clustering-based diarization, Kinoshita, Delcroix, and Tawara (2021a, 2021b) first estimate distinct local neural speaker embeddings from short speech chunks, which they then CL-constrain in the subsequent global speaker clustering step. Finally, Tripathi et al. (2022) employ a speaker change detector and impose CL constraints between consecutive segments separated by a speaker change and ML constraints between segments where a speaker change was not detected. To the best of our knowledge, constraints grounded on language, that can provide crucial information in role-based conversational settings, have not been explored.

### 4.2.3 Constrained spectral clustering

Constraints can be combined with several clustering algorithms, such as k-means (Kinoshita et al., 2021b) or HAC (Prokopalo, Shamsi, Barrault, Meignier, & Larcher, 2021). In this work we use a constrained spectral clustering approach, where constraints are integrated via the exhaustive and efficient constraint propagation (E$^2$CP) algorithm (Lu & Peng, 2013), which was recently applied in diarization settings (Tripathi et al., 2022). Applying E$^2$CP, we can propagate an initial set of pairwise constraints to the entire session. In order to do so, we define a constraint matrix $\mathbf{Z} \in \mathbb{R}^{N\times N}$, such that

$$\mathbf{Z}_{ij} = \begin{cases} +1, & \text{if } \exists \text{ ML constraint between } i \text{ and } j \\ -1, & \text{if } \exists \text{ CL constraint between } i \text{ and } j \\ 0, & \text{if } \nexists \text{ any constraint between } i \text{ and } j \end{cases} \tag{4.5}$$

Soft constrains can also be applied within this framework by setting $|\mathbf{Z}_{ij}| < 1$, with $|\mathbf{Z}_{ij}|$ denoting the confidence score that a constraint should be imposed between the $i$-th and $j$-th segments.

The elements of the affinity matrix $\hat{\mathbf{W}}$ are then updated as

$$\hat{\mathbf{W}}_{ij} \leftarrow \begin{cases} 1-(1-\mathbf{F}^*_{ij})(1-\hat{\mathbf{W}}_{ij}), & \text{if } \mathbf{F}^*_{ij} \geq 0 \\ (1+\mathbf{F}^*_{ij})\hat{\mathbf{W}}_{ij}, & \text{if } \mathbf{F}^*_{ij} < 0 \end{cases} \tag{4.6}$$

where $\mathbf{F}^*$ contains the constraints propagated to the entire session based on the initial set of constraints and is estimated as

$$\mathbf{F}^* = (1-\alpha)^2(\mathbf{I}-\alpha\bar{\mathbf{L}})^{-1}\mathbf{Z}(\mathbf{I}-\alpha\bar{\mathbf{L}})^{-1} \tag{4.7}$$

$\bar{\mathbf{L}}$ equals $\bar{\mathbf{D}}^{-1/2}\hat{\mathbf{W}}\bar{\mathbf{D}}^{-1/2}$, where $\bar{\mathbf{D}}$ is a diagonal matrix defined like $\mathbf{D}$ in Section 4.2.1, but using the degrees of $\hat{\mathbf{W}}$. The constant $\alpha \in [0,1]$ is a tunable hyperparameter: a small value penalizes large changes between the initial pairwise constraints in $\mathbf{Z}$ and the new constraints created during propagation, while a large value penalizes large changes between the neighboring segments in the graph described by $\hat{\mathbf{W}}$. Note that for $\alpha = 0$ we get $\mathbf{F}^* = \mathbf{Z}$ which means we only rely on the initial constraints, and for $\alpha = 1$ we get $\mathbf{F}^* = \mathbf{0}$, which means we completely ignore any constraint information. The constraint propagation and integration described here takes place before the refinement and spectral operations described in Section 4.2.1.

## 4.3 Proposed Method

We propose to use a two-step clustering for conversations where speakers assume distinct roles, as depicted in the example of Figure 4.1.

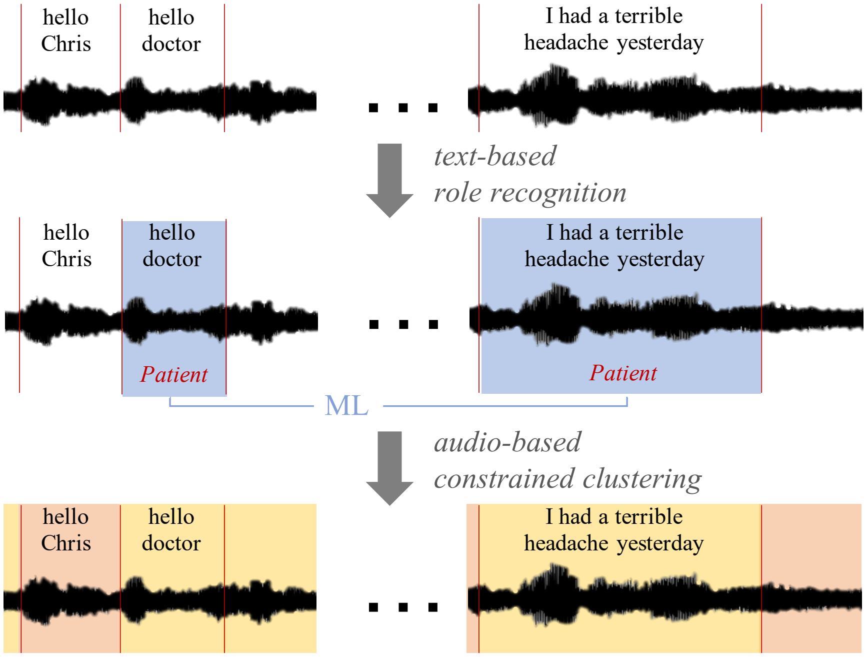

hello Chris
hello doctor
I had a terrible headache yesterday
text-based role recognition
hello Chris
hello doctor
Patient
I had a terrible headache yesterday
Patient
ML
audio-based constrained clustering
hello Chris
hello doctor
I had a terrible headache yesterday

Figure 4.1: Two-step speaker clustering for role-playing interactions. Here, an ML constraint is imposed for two segments both associated with the role *patient*. Those segments have to be in the same cluster after the clustering step.

First, speaker roles are identified from text for each speech segment. To that end, we employ a BERT-based classifier (Devlin, Chang, Lee, & Toutanova, 2019), where we add dropout and a softmax inference layer on top of a pre-trained BERT model and we fine-tune it for the task with in-domain data. If, after classification, we have complete role information available (i.e., each segment is associated with a distinct speaker role), we can directly get a purely text-based diarization result (see also Chapter 3). However, there are multiple scenarios where only partial role information is available (e.g., we have sufficient data to only train a binary classifier to identify *news anchor* vs. *guest* in a broadcast news program with multiple potential *guests* within a show). Additionally, we expect that there will be several segments where the linguistic content is not sufficient to robustly infer the associated speaker role.

So, we only use role information to impose suitable constraints for the following step of audio-based clustering and we take into account only segments where roles are identified with sufficient confidence. Even though it is well known that neural classifiers tend to be over-confident about their decisions and that softmax values are usually not a robust proxy of confidence scores, in practice we saw that we can use a softmax threshold as a threshold of confidence, as discussed in Section 4.5. For those segments where the confidence of their associated role is beyond some specified threshold, we impose ML and CL constraints, according to the domain we are working on. For instance, we can distinguish between the following general scenarios:

1. *different roles are always played by different speakers within a session* (e.g., teacher vs. students during a lecture): apply a CL constraint between any segments associated with different speaker roles,

2. *different speakers always play different roles within a session* (e.g., anchor vs. interviewer vs. guest during a broadcast news program, where anchor and interviewer might be the same person): apply an ML constraint between any segments associated with the same speaker role,

3. *one-to-one correspondence between speakers and roles within a session* (e.g., doctor vs. patient during a doctor's visit): apply both CL and ML constraints as in cases (1) and (2).

Different types of domain-specific strategies can also be followed. The constraints are then integrated within a spectral clustering algorithm, and we proceed as described in Sections 4.2.1 and 4.2.3.

## 4.4 Datasets

We evaluate the proposed speaker clustering approach on two different domains with role-playing interactions. As detailed below, we use a medical dataset drawn from the psychotherapy field and another dataset from the entertainment industry with podcast episodes.

### 4.4.1 Psychotherapy sessions

We use a collection of psychotherapy sessions recorded at a university counseling center (UCC)[1], and specifically the sessions in the sets denoted as $\text{UCC}_{train}$, $\text{UCC}_{dev}$ and $\text{UCC}_{test_1}$ in (Flemotomos et al., 2021)[2]. All the recordings have been normalized to 16 kHz sampling rate, 16 bit precision, and

[1] Note that this is a different dataset than the MI sessions used for the experiments in the previous chapters.

[2] See also Chapter 5, Section 5.4.2.

the two recording microphones suspended from the ceiling of the clinic offices have been combined through acoustic beamforming. Each session is a dyadic conversation between a *therapist* and a *patient*, thus falls under case (3) according to the categorization given in Section 4.3. The dataset comprises 97 participants (23 therapists and 74 patients), with no speaker overlap between the train/dev/eval sets. The sessions have been professionally transcribed, the transcribed segments have been forced-aligned with the beamformed audio, and any utterances consisting of only non-speech vocal sounds (e.g., laughs) have been discarded. More details on the dataset are provided in Table 4.1.

Table 4.1: Size of the UCC dataset.

| | **train** | **dev** | **eval** |
|---|---|---|---|
| #sessions | 50 | 26 | 20 |
| #segments - therapist | 8,766 | 3,959 | 4,146 |
| #segments - patient | 9,052 | 4,246 | 4,245 |
| segment duration (mean) | 7.8 sec | 8.7 sec | 6.4 sec |
| #words per segment (mean) | 21.4 | 22.3 | 18.8 |

### 4.4.2 Podcast episodes

*This American Life*[3] (TAL) is a weekly podcast and public radio show where each episode revolves around a specific theme and is structured as a story-telling act with multiple characters. Mao, Li, McAuley, and Cottrell (2020) have curated a dataset of 663 TAL episodes aired between 1995 and 2020. We use the clean, audio-aligned utterances provided, with the recommended train/dev/eval split, and with the archived audio standardized to 16 kHz, 16 bit precision, mono-channel, wav format (as described by Mao et al., 2020). In each episode there are on average 17.7 speakers (std=8.7) with variable speaking times, while the existing background music poses extra challenges for robust clustering and diarization. The dataset, described in Table 4.2, has been annotated with speaker identities and with three speaker roles, those of *host*, *interviewer*, and *subject*. However, the provided role information was not helpful for our purposes, since, according to the annotations, multiple speakers may play the same role within an episode and, at the same time, a single speaker

[3] https://www.thisamericanlife.org/

may play multiple roles (with some episodes having the same speaker occasionally playing all 3 roles). So, we instead chose to annotate as *host* utterances only the ones spoken by Ira Glass and assign all the other utterances to a *non-host* speaker role. Ira Glass is the host and executive producer of the show and speaks for 18.6% of the time during the entire dataset[4]. Since in this case different roles always denote different speakers (but the inverse does not hold since there are multiple *non-host* speakers), this dataset falls under case (1) according to the categorization given in Section 4.3. Of course, we should note that since this annotation strategy is speaker-dependent, the role recognition algorithm applied is also expected to capture speaker-specific, and not purely role-specific, information.

Table 4.2: Size of the TAL dataset.

| | **train** | **dev** | **eval** |
|---|---|---|---|
| #episodes | 593 | 34 | 36 |
| #segments - host | 26,523 | 1,765 | 1,317 |
| #segments - non-host | 119,295 | 6,869 | 8,039 |
| segment duration (mean) | 14.1 sec | 13.7 sec | 13.4 sec |
| #words per segment (mean) | 37.7 | 36.6 | 36.6 |

## 4.5 Experiments and Results

### 4.5.1 Experimental setup

For both datasets we run experiments using the manually derived speaker segments and the corresponding transcriptions, in order to evaluate the effectiveness of the proposed method without propagating potential errors from automated segmentation and speech recognition modules.

We standardize the text by stripping punctuation, removing non-verbal vocalizations and converting all letters to lower case. We build the binary role classifiers (*therapist* vs. *patient* and *host* vs. *non-host*) using TensorFlow (Abadi et al., 2016) with the pre-trained uncased English BERT-base model provided in TensorFlow model garden (H. Yu et al., 2020), adding a dropout

[4] For reference, the second single most-talking speaker of the dataset is Nancy Updike, speaking for 1.6% of the time.

layer with dropout ratio equal to 0.2. Since very short segments are not expected to have sufficient role-related information, during fine-tuning we only take into account segments containing at least 5 words (65.58% of the available training segments for UCC and 88.65% of the available training segments for TAL). We fine-tune the models for 2 epochs on the training subsets of the datasets, using the development subsets for validation. We use the Adam optimizer with decoupled weight decay (Loshchilov & Hutter, 2019) with initial learning rate equal to $2 \cdot 10^{-5}$ and with a warm-up stage lasting for the first 10% of the training time. The mini-batch size is set to 16 segments and the maximum allowed segment length is set to 128 tokens[5], which means that 2.91% of the initial training UCC segments and 2.06% of the initial training TAL segments are cropped.

The speaker representation of the segments is based on the widely used x-vectors (Snyder et al., 2018) and, to that end, the pre-trained VoxCeleb x-vector extractor from Kaldi (Povey et al., 2011) is used[6]. A single x-vector is extracted per segment, taking into consideration only the voiced frames, as identified by an energy-based voice activity detector. X-vectors are projected through linear discriminant analysis (LDA) on a 200-dimensional space and are further mean- and length-normalized. The segments are then clustered following the described constrained spectral clustering approach with ML and/or CL constraints imposed according to the predicted associated roles[7]. For the UCC dataset, which features dyadic interactions, we group all the segments into two clusters, while for TAL we estimate the number of speakers using the eigengap criterion described in Section 4.2.1, searching in the range 2–50. The value of $p$ for the $p$-thresholding step is found through auto-tuning (Park, Han, Kumar, & Narayanan, 2019), searching in the range 40–95, and we use soft thresholding with $\tau = 0.01$ (Q. Wang et al., 2018).

All the results are reported on the eval subsets of the data. Diarization is evaluated with respect to the diarization error rate (DER), estimated with the `pyannote.metrics` library (Bredin, 2017a) without allowing any tolerance collar around segment boundaries. As explained in Chapter 3 (Section 3.5.2), DER incorporates three sources of error; false alarms, missed speech, and speaker confusion. However, segmentation is always the oracle one provided by human annotators and,

[5] according to the default WordPiece-based BERT tokenizer

[6] https://kaldi-asr.org/models/m7

[7] https://github.com/wq2012/SpectralCluster

since there is almost no speaker overlap in our datasets, by DER we essentially estimate speaker confusion (false alarm is always 0 and missed speech is 0.02% for UCC and 0.13% for TAL).

### 4.5.2 Results and discussion

If we have perfect role information for all the segments and if there is one-to-one correspondence between roles and speakers (e.g., one *therapist* vs. one *patient* in UCC), we can get a perfect diarization result in terms of speaker confusion. In the framework of constrained spectral clustering, this can be done by filling $\mathbf{Z}$ in equation (4.5) with all the corresponding constrains and setting $\alpha = 0$ in equation (4.7) so that $\mathbf{F}^* = \mathbf{Z}$. This is reflected in Figure 4.2 where we see how DER changes as we provide more oracle constraints to the algorithm. This is similar to the expected behavior of the algorithm when constraints are added in the form of human supervision.

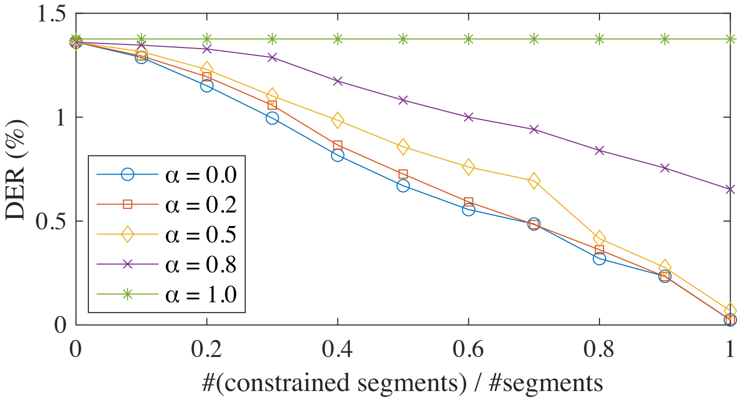


Figure 4.2: DER for the UCC dataset as a function of the (normalized) number of constraints, always providing oracle role information to build the constraints, for different values of $\alpha$ in equation (4.7).

Without having access to the oracle role information, we have to rely on a segment-level role classifier. The classification accuracy of our BERT-based classifiers after fine-tuning is given in Table 4.3 and is compared to a naive majority-class baseline. Even though the classifiers provide reasonable results, we need to ensure that constraints are imposed only on segments which are confidently linked to some role. For this work, apart from only using segments longer than a specified duration (here, containing at least 5 words) to ensure some minimal linguistic content, we use the softmax values associated with the predicted roles as a proxy of the confidence level.

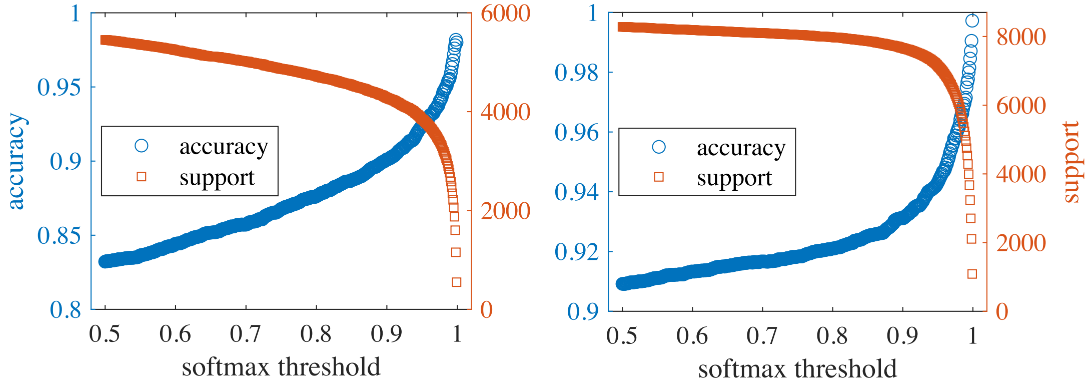
As shown in Figure 4.3, the softmax value can indeed act as a reasonable proxy of confidence for

Table 4.3: Classification accuracy (%) of the BERT-based model and a majority-class baseline.

| | all segments | | segments w. $\geq$ 5 words | |
|---|---|---|---|---|
| | **maj. class** | **BERT** | **maj. class** | **BERT** |
| UCC | 50.59 | 73.63 | 53.50 | 83.22 |
| TAL | 85.92 | 90.87 | 85.41 | 90.92 |

for UCC data: binary problem of identifying *therapist* vs. *patient*
for TAL data: binary problem of identifying *host* vs. *non-host*

our purposes. In particular, if we only consider segments where the corresponding softmax value is above some threshold, accuracy increases monotonically as a function of the threshold. However, the choice of the threshold value is a trade-off decision between accuracy and adequate support so that we have a sufficient number of constraints. With that in mind, we choose a threshold equal to 0.980 for the UCC data (accuracy = 94.66%, support = 3,222) and equal to 0.995 for the TAL data (accuracy = 98.15%, support = 3,674), which leads to imposing constraints on around 40% of the segments in both cases.



(a) UCC

(b) TAL

Figure 4.3: Classification accuracy and support for the BERT-based classifiers when only segments with associated softmax value above some threshold are considered.

After constructing the constraint matrix based on the described role classification only for the segments with sufficient role classification confidence, we perform the constrained spectral clustering algorithm. Our experiments fall under case (3) for the UCC data and under case (1) for the TAL data, according to the categorization given in Section 4.3. The results[8] are reported in the second column of Table 4.4. For comparison, we also provide results for the two extreme cases: i) following a conventional, unconstrained spectral clustering, which ignores any language-based information and ii) following a language-only classification using the results of the BERT-based classifier for all the segments, without setting any softmax threshold, which ignores any audio-based information.

Table 4.4: DER (%) using unconstrained audio-only clustering, constrained clustering with role-induced constraints, and language-only role-based classification.

| | **unconstrained clustering**<br>(audio-only) | **constrained clustering**<br>(multimodal) | **role-based classification**<br>(language-only) |
|---|---|---|---|
| UCC | 1.38 | 1.31 | 10.34 |
| TAL | 42.22 | 23.86 | 63.01* |

*results contain only 2 speakers, since we rely on binary classification

In the case of the UCC data, our approach yields a small improvement (5.1% relative) compared to the unconstrained baseline. We additionally found that adding more constraints (selecting a smaller softmax threshold as our confidence criterion) leads to worse performance. Comparing this finding to the results displayed in Figure 4.2 with oracle constraints, where error approaches 0 given a large number of constraints, we realize that our method is sensitive to the performance of the role classifier. This is because any classification errors can be easily propagated to the clustering step (Figure 4.1). This error propagation becomes, as expected, more evident in the case we constrain all the segments, relying only on the linguistic stream of information (last column of Table 4.4).

Looking at the results with the TAL data, we observe a substantial improvement when going from unconstrained to constrained clustering. We can see that in scenarios with a large number of speakers, even partial role-based information (like the *host* vs. *non-host* classification here) can provide useful cues that robustly guide the subsequent clustering. In more detail, we observed that the imposed constraints changed the final Laplacian matrix in a way that the eigengap criterion led to the detection of more clusters (speakers) per episode. The severe performance degradation with the language-only approach is expected, since the results in that case only contain two speakers (since we only have two role classes), even though each TAL episode features multiple participants.

[8] Those results are for $\alpha = 0.75$ for UCC and $\alpha = 0.50$ for TAL.

## 4.6 Conclusion

In this chapter we proposed to integrate text-based constraints within audio-based clustering to improve the performance of speaker diarization in conversational interactions where speakers assume specific roles. We implemented a BERT-based role classifier solely relying on text data and used its output to construct a constraint matrix for use within constrained spectral clustering. Experimental results in two different domains showed that, after applying a softmax-based confidence criterion, performance can be improved both in cases of one-to-one correspondence between speakers and roles and in cases with only partial available role information, thus overcoming limitations of assumptions we needed to follow for the approaches proposed in the previous chapters.

We performed all our experiments using oracle textual information and oracle speaker segmentation. We should note that, in a real-world scenario, errors would be introduced and potentially propagated to the clustering step both because of a speech recognizer and because of non-ideal segmentation. Speaker segmentation could be included as a separate pre-processing module (e.g., like in Chapter 3), or incorporated with the role recognizer in a named entity recognition (NER)-like approach (e.g., Zuluaga-Gomez et al., 2021). Future work can also investigate a combination of hard and soft constraints for the task, as well as different types of role-induced constraints. Even though here we focused on linguistic characteristics, role-specific behaviors can also be manifest through acoustic, structural, or visual cues, all of which can be potentially used within the framework of role-dependent constrained speaker clustering.

With this chapter, we close our discussion on how linguistically-extracted speaker role information can be used to facilitate the task of speaker diarization. Here we studied how to use this information to impose constraints during audio-based clustering. In Chapter 3 we proposed a technique, suitable in scenarios with one-to-one correspondence between speakers and roles (e.g., *patient-doctor* interactions), to construct the acoustic speaker identities and reduce diarization to a classification problem. In the following chapter we are going to see how the latter technique can be incorporated within a larger speech and language processing pipeline deployed in clinical settings to solve a real-world problem; the one of psychotherapy quality assessment.

# Part III

# Real World Impact

# Chapter 5

# Why Do We Need Roles? Automated Psychotherapy Evaluation as an Example Downstream Application

With the growing prevalence of psychological interventions, it is vital to have measures that rate the effectiveness of psychological care to assist in training, supervision, and quality assurance of services. Traditionally, quality assessment is addressed by human raters who evaluate recorded sessions along specific dimensions, often codified through constructs relevant to the approach and domain. This is, however, a cost-prohibitive and time-consuming method that leads to limited use in real-world settings. To facilitate this process, we have developed an automated competency rating tool able to process the raw recorded audio of a session, analyzing who spoke when, what they said, and how the health professional used language to provide therapy. Since the system focuses on therapist-attributed language, it is essential to robustly differentiate between utterances spoken by the therapist vs. the patient. We present and analyze our platform using a dataset drawn from its deployment in a real-world clinical setting and we show how applying the techniques we introduced in Chapter 3 can have a substantial beneficial effect to the overall performance.

The work presented in this chapter is based on work that has been published in (Flemotomos et al., 2021).

## 5.1 Need for Psychotherapy Quality Assessment Tools

Recent epidemiological research suggests that developing a mental disorder is the norm, rather than the exception, estimating that the lifetime prevalence of diagnosable mental disorders (i.e., the proportion of the population that, at some point in their life, have experienced or will experience a mental disorder) is around 50% (Kessler et al., 2005) or even more (Schaefer et al., 2017). According to data from 2018, an estimated 47.6 million adults in the United States had some mental illness, and 1 in 7 adults received professional mental health services (Substance Abuse and Mental Health Services Administration, 2019).

Psychotherapy is a commonly used process in which mental health disorders are treated through communication between an individual and a trained mental health professional. Even though its positive effects have been well documented (Lambert & Bergin, 2002; Perry, Banon, & Ianni, 1999; Weisz, Weiss, Han, Granger, & Morton, 1995), there is room for improvement in terms of the quality of services provided. In particular, a substantial number of patients report negative outcomes, with signs of mental health deterioration after the end of therapy (Curran et al., 2019; Klatte, Strauss, Flückiger, & Rosendahl, 2018). Apart from patient characteristics (Lambert & Bergin, 2002), therapist factors play a significant and clinically important role in contributing to negative outcomes (Saxon, Barkham, Foster, & Parry, 2017). This has direct implications for more rigorous training and supervision (Lambert & Ogles, 1997), quality improvement, and skill development. A critical factor that can lead to increased performance and thus ensure high quality of services is the provision of accurate feedback to the practitioner (Hattie & Timperley, 2007). This can take various forms; both client progress monitoring (Lambert, Whipple, & Kleinstäuber, 2018) and performance-based feedback (Schwalbe, Oh, & Zweben, 2014) have been reported to reduce therapeutic skill erosion and to contribute to improved clinical outcomes. The timing of the feedback is of utmost importance as well, since it has been shown that immediate feedback is more effective than delayed (Kulik & Kulik, 1988).

In psychotherapy practice, however, providing regular and immediate performance evaluation is almost impossible. Behavioral coding—the process of listening to audio recordings and/or reading session transcripts in order to observe therapists' behaviors and skills (Bakeman & Quera, 2012)—is both time-consuming and cost-prohibitive when applied in real-world settings. It has been reported (Moyers, Martin, Manuel, Hendrickson, & Miller, 2005) that, after intensive training and supervision that lasts on average 3 months, a proficient coder would need up to two hours to code just a 20 min-long session of motivational interviewing (MI), a specific type of psychotherapy which is the focus of the current chapter. The labor-intensive nature of coding means that the vast majority of psychotherapy sessions are not evaluated. As a result, many providers get inadequate feedback on their therapy skills after their initial training (Miller, Sorensen, Selzer, & Brigham, 2006) and behavioral coding is mainly applied for research purposes with limited outreach to community settings (Proctor et al., 2011). At the same time, the barriers imposed by manual coding usually lead to research studies with relatively small sample sizes (Magill et al., 2014), limiting progress in the field. It is, thus, made apparent that being able to evaluate a therapy session and provide feedback to the practitioner at a low cost and in a timely manner would both boost psychotherapy research and scale up quality assessment to real-world use.

In this chapter we investigate whether it is feasible to analyze a therapy session recording in a fully automatic way and provide feedback to the therapist within short time. The focus is on the importance of speaker role modeling within the overall computational approach and on how some of the techniques presented earlier (especially in Chapter 3) can improve the final performance of automated behavioral coding.

## 5.2 Behavioral Coding for Motivational Interviewing

Motivational interviewing (MI; Miller & Rollnick, 2012), often used for treating addiction and other conditions, is a client-centered intervention that aims to help clients make behavioral changes through resolution of ambivalence. It is a psychotherapy treatment with evidence supporting that specific skills are correlated with the clinical outcome (Gaume, Gmel, Faouzi, & Daeppen, 2009; Magill et al., 2014) and also that those skills cannot be maintained without ongoing feedback (Schwalbe et al., 2014). Thus, great effort from MI researchers has been devoted to developing instruments to evaluate fidelity to MI techniques.

The gold standard for monitoring clinician fidelity to treatment is behavioral observation and coding (Bakeman & Quera, 2012). During that process, trained coders assign specific labels or numeric values to the psychotherapy session, which are expected to provide important therapy-related details (e.g., “how many open questions were posed by the therapist?” or “did the counselor accept and respect the client’s ideas?”) and essentially reflect particular therapeutic skills. While there is a variety of coding schemes (Madson & Campbell, 2006), in this study we focus on a widely used research tool, the motivational interviewing skill code (MISC 2.5; Houck, Moyers, Miller, Glynn, & Hallgren, 2010), which was specifically developed for use with recorded MI sessions (Madson & Campbell, 2006). MISC defines behavior codes both for the counselor and the patient, but for the automated system reported here we focus on counselor behaviors.

The MISC manual (Houck et al., 2010) defines both session-level and utterance-level codes. Session-level codes characterize the entire interaction and are scored on a 5-point Likert scale. When coding at the utterance-level, instead of assigning numerical values, the coder decides in which behavior category each utterance belongs. An utterance is a “thought unit” (Houck et al., 2010), which means that multiple consecutive phrases might be parsed into a single utterance and, likewise, multiple utterances might compose a single sentence or talk turn. After the session is parsed into utterances, each one is assigned one of the codes summarized in Table 5.1 (or gets the label NC if it can not be coded). For the work presented in this chapter we focus on utterance-level codes, but we also use session-level summary indicators. In particular, we estimate i) the ratio of *reflections* (simple and complex) to *questions* (open and closed), ii) the percentage of *open questions* (over the total number of questions), iii) the percentage of *complex reflections* (over the total number of reflections), and iv) MI adherence, defined as the percentage of utterances coded with any code other than *advice* (with or without permission), *raise concern* (with or without permission), *confront*, *direct*, *warn*.

## 5.3 Psychotherapy Evaluation in the Digital Era

Psychotherapy sessions are interventions primarily based on spoken language, which means that the information capturing the session quality is encoded in the speech signal and the language patterns of the interaction. Thus, with the rapid technological advancements in the fields of speech and natural language processing (NLP) over the last few years (e.g., Devlin et al., 2019; Xiong et al., 2017), and despite many open challenges specific to the healthcare domain (Quiroz et al., 2019), it is not surprising to see trends in applying computational techniques to automatically analyze and

Table 5.1: Therapist-related utterance-level codes, as defined by MISC 2.5.

| **abbreviation** | **name** | **example** |
|---|---|---|
| ADP | Advise with Permission | Would it be all right if I suggested something? |
| ADW | Advise w/o Permission | I recommend that you attend 90 meetings in 90 days. |
| AF | Affirm | Thank you for coming today. |
| CO | Confront | (C: I don't feel like I can do this.) Sure you can. |
| DI | Direct | Get out there and find a job. |
| EC | Emphasize Control | It is totally up to you whether you quit or cut down. |
| FA | Facilitate | Uh huh. (*keep-going acknowledgment*) |
| FI | Filler | Nice weather today! |
| GI | Giving Information | Your blood pressure was elevated [...] this morning. |
| QUO | Open Question | Tell me about your family. |
| QUC | Closed Question | How often did you go to that bar? |
| RCP | Raise Concern with Permission | Could I tell you what concerns me about your plan? |
| RCW | Raise Concern w/o Permission | That doesn't seem like the safest plan. |
| RES | Simple Reflection | (C: The court sent me here.) That's why you're here. |
| REC | Complex Reflection | (C: The court sent me here.) This wasn't your choice to be here. |
| RF | Reframe | (C: [...] something else comes up [...]) You have clear priorities. |
| SU | Support | I'm sorry you feel this way. |
| ST | Structure | Now I'd like to switch gears and talk about exercise. |
| WA | Warn | Not showing up for court will send you back to jail. |
| NC | No Code | You know, I... (*meaning is not clear*) |

Most of the examples are drawn from the MISC manual (Houck et al., 2010).
Many of the code assignments depend on the client's previous utterance (C).

evaluate psychotherapy sessions.

Such efforts span a wide range of psychotherapeutic approaches including couples therapy (Black et al., 2013), MI (Xiao, Can, et al., 2016) and cognitive behavioral therapy (Flemotomos, Martinez, et al., 2018), used to treat a variety of conditions such as addiction (Xiao, Can, et al., 2016) and post-traumatic stress disorder (Shiner et al., 2012). Both text-based (Imel, Steyvers, & Atkins, 2015; Xiao, Can, Georgiou, Atkins, & Narayanan, 2012) and audio-based (Black et al., 2013; Xiao et al., 2014) behavioral descriptors have been explored in the literature and have been used either unimodally or in combination with each other (Singla et al., 2018).

In this study we focus on behavior code prediction from textual data. Most research studies focused on text-based behavioral coding have relied on written text excerpts (Barahona et al., 2018) or used manually-derived transcriptions of the therapy session (Can, Atkins, & Narayanan, 2015; Gibson et al., 2022; Lee, Hull, Levine, Ray, & McKeown, 2019). However, a fully automated evaluation system for deployment in real-world settings requires a speech processing pipeline that can analyze the audio recording and provide a reliable speaker-segmented transcript of what was spoken by whom. This is a necessary condition before such an approach is introduced into clinical settings since, otherwise, it may eliminate the burden of manual behavioral coding, but it introduces the burden of manual transcription.

An end-to-end system is presented by Xiao, Imel, Georgiou, Atkins, and Narayanan (2015) and Xiao, Huang, et al. (2016), where the authors report a case study of automatically predicting the empathy expressed by the provider. A similar platform, focused on couples therapy, is presented by Georgiou, Black, Lammert, Baucom, and Narayanan (2011b). Even employing an ASR module with relatively high error rate, those systems were reported to provide competitive prediction performance. The scope of the particular studies, though, was limited only to session-level codes, while the evaluation sessions were selected from the two extremes of the coding scale. Thus, for each code the problem was formulated as a binary classification task trying to identify therapy sessions where a particular code (or its absence) is represented more prominently (e.g., identify 'low' vs. 'high' empathy).

## 5.4 Current Study

### 5.4.1 System overview

We analyze a platform able to process the raw recording of a psychotherapy session and provide, within short time, performance-based feedback according to therapeutic skills and behaviors. We focus on dyadic psychotherapy interactions (i.e., one *therapist* and one *client*) and the quality assessment is based on the counselor-related codes of the MISC protocol (Houck et al., 2010). The behavioral codes are predicted by NLP algorithms that analyze the linguistic information captured in the automatically derived transcriptions. The behavioral analysis of the counselor is summarized into a comprehensive feedback report that can be used directly by the provider as a self-assessment method or by a supervisor as a supportive tool that helps them deliver more effective and engaging training.

After both parties have formally consented, the therapist begins recording the session. The digital recording is sent to the processing pipeline and appropriate acoustic features are extracted from the raw speech signal. The baseline system explored here (Figure 5.1) consists of six main steps: (a) voice activity detection (VAD), where speech segments are detected over silence or background noise, (b) speaker diarization, where the speech segments are clustered into same-speaker groups (e.g., speaker A, speaker B of a dyad), (c) automatic speech recognition (ASR), where the audio speech signal of each speaker-homogeneous segment is transcribed to words, (d) speaker role recognition (SRR), where each speaker group is assigned their role (i.e., *therapist* vs. *client*), (e) utterance segmentation, where the speaker turns are parsed into utterances which are the basic units of behavioral coding, and (f) automated behavioral coding where a MISC-based code is assigned to each therapist-attributed utterance. Speaker role recognition is essential in this system, since the goal is to robustly identify and then automatically code the *therapist* utterances.

The architecture design described can inevitably lead to error propagation. Here, we study how errors due to diarization can affect the overall performance of the downstream task of psychotherapy quality assessment and how an alternative framework can help alleviate such error propagation problems. In particular, we compare the system of Figure 5.1 with the architecture of Figure 5.2, which is based on the linguistically aided diarization approach introduced in Chapter 3. Using a collection of real-world psychotherapy recordings acquired after the deployment of our system in

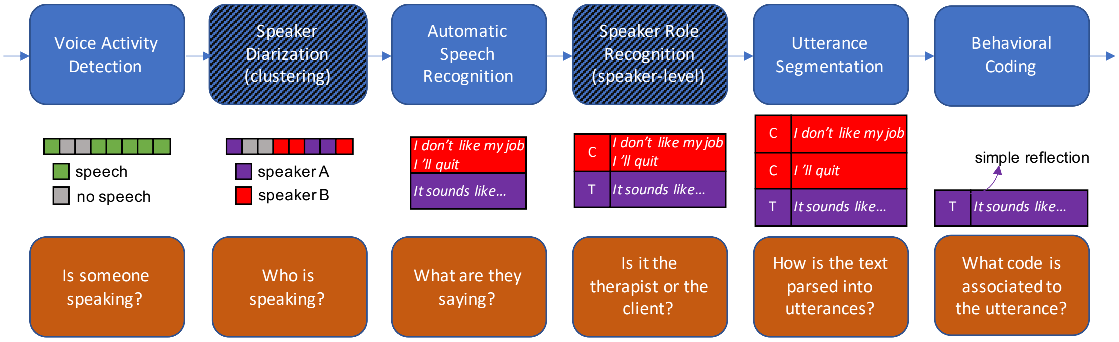


Figure 5.1: Baseline transcription and coding pipeline developed to assess the quality of a psychotherapy session. The focus of the particular study is on the effect of the speaker diarization and role recognition modules on the overall performance.

clinical settings, we show that traditional clustering-based diarization can fail for certain sessions, leading to inaccurate behavior coding results. Employing simple quality and confidence thresholds based on the expected speaking times of the two interlocutors, we can instead use the linguistically-aided approach for those sessions and get significant performance gains.

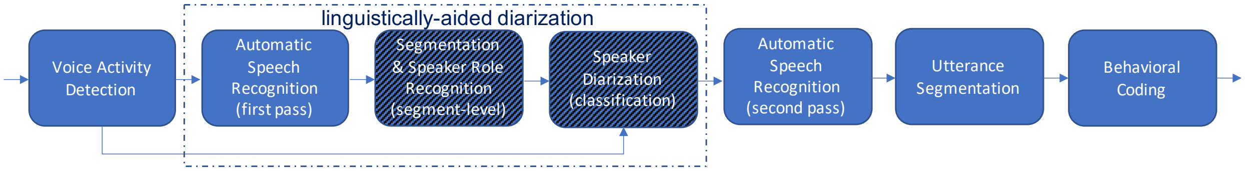

linguistically-aided diarization
Voice Activity Detection
Automatic Speech Recognition (first pass)
Segmentation & Speaker Role Recognition (segment-level)
Speaker Diarization (classification)
Automatic Speech Recognition (second pass)
Utterance Segmentation
Behavioral Coding

Figure 5.2: Transcription and coding pipeline employing linguistically-aided, role-based speaker diarization.

### 5.4.2 Deployment: data collection and pre-processing

Through a collaboration with the counseling center of a large US-based university, we gathered a corpus of real-world psychotherapy sessions to evaluate the system. Therapy treatment was provided by a combination of licensed staff as well as trainees pursuing clinical degrees. Topics discussed span a wide range of concerns common among students, including depression, anxiety, substance use, and relationship concerns. All the participants (both patients and therapists) had formally consented to their sessions being recorded. Study procedures were approved by the institutional review board of the University of Utah. Each session was recorded by two microphones suspended from the ceiling of the clinic offices, one omni-directional and one directed to where the therapist generally sits.

Data were collected between September, 2017 and March, 2020, for a total of 5,097 recordings. Out of those, 188 sessions were selected to be manually transcribed and coded. Coding took place in two independent trials (one in mid 2018 and one in late 2019), with some differences in the procedure between the two. For the first coding trial (96 sessions), the transcriptions were stripped of punctuation and coders were asked to parse the session into utterances. During the second trial (92 sessions), the human transcriber was asked to insert punctuation, which was used to assist parsing. Additionally, for the second batch of transcriptions, stacked behavioral codes (more than one code per utterance) were allowed in case one of the codes is open or closed question (QUO or QUC). We have split the first trial into train (UCC$_{train}$; 50 sessions), development (UCC$_{dev}$; 26 sessions), and test (UCC$_{test_1}$; 20 sessions) sets, while we refer to the second trial as the UCC$_{test_2}$ set. The split for the first trial was done in a way so that there is no speaker overlap between the different sets. For this chapter, we report results on the 112 sessions of the combined UCC$_{test_1}$ and UCC$_{test_2}$ sets.

The manually transcribed UCC sessions do not contain any timing information, which means that we needed to align the provided audio with text. That way, we were able to get estimates of the "ground truth" information required for evaluation. We did so by using the Gentle forced aligner[1], an open-source, Kaldi-based (Povey et al., 2011) tool, in order to align at the word level. However, we should note that this inevitably introduces some error to the evaluation process, since 9.4% of the words per session on average (std=3.4%) remain unaligned.

Another pre-processing step we needed to take in order to have a meaningful evaluation of the system on the UCC data is related to the behavioral labels assigned by the humans and by the platform. In particular, some of the utterance-level MISC codes are assigned very few times within a session by the human raters and the corresponding inter-rater reliability (IRR) is very low (Table A.1); additionally, there are pairs or groups of codes with very close semantic interpretation as reflected by the examples in Table 5.1 (e.g., complex reflections (REC) and reframes (RF)). Thus, we clustered the codes into composite groups resulting in 9 target labels. The mapping between the codes defined in the MISC manual and the target labels, as well as the occurrences of

[1] https://github.com/lowerquality/gentle

those labels in the UCC data, is given in Table 5.2. The facilitate code (FA) seems to dominate the data, because most of the verbal fillers (e.g., *uh-huh*, *mm-hmm*, etc.)—which are very frequent constructs in conversational speech—and single-word utterances (e.g., *yeah*, *right*, etc.) are labeled as FA.

Table 5.2: Mapping between MISC-defined behavior codes and grouped target labels, together with the occurrences of each group in the evaluation UCC sets.

| **group** | **MISC codes** | **count** |
|---|---|---|
| FA | FA | 13,618 |
| GI | GI, FI | 7,661 |
| QUC | QUC | 4,387 |
| QUO | QUO | 2,658 |
| REC | REC, RF | 6,342 |
| RES | RES | 829 |
| MIN | ADP, ADW, CO, DI, RCW, RCP, WA | 987 |
| MIA | AF, EC, SU | 1,839 |
| ST | ST | 2,081 |

MISC abbreviations are defined in Table 5.1.
MIA stands for MI-Adherent codes.
MIN stands for MI-NonAdherent codes.

## 5.5 Experiments

We apply and compare the two systems introduced in Section 5.4.1 and we focus on the performance of the speaker diarization and role recognition modules with respect to the end task of automated behavioral coding. In both cases, all the other modules (VAD, ASR, utterance segmentation, MISC labeling) remain fixed—details on those modules are provided in Appendix B.

### 5.5.1 System with clustering-based diarization

Following a traditional speaker diarization approach, the speech signal is first partitioned into segments where a single speaker is present and then, those speaker-homogeneous segments are clustered into same-speaker groups. In the baseline pipeline of Figure 5.1 we follow the x-vector/PLDA paradigm (Sell et al., 2018), the same baseline audio-only diarization approach we followed in Chapter 3. Each voiced segment, as predicted by VAD, is partitioned uniformly into subsegments of length equal to 1.5 sec with a shift of 0.25 sec. For each subsegment an x-vector (Snyder et al., 2018) is extracted using the pre-trained CallHome x-vector extractor[2] provided by Kaldi (Povey et al., 2011). The subsegments are finally clustered according to hierarchical agglomerative clustering (HAC) with average linking, using probabilistic linear discriminant analysis (PLDA) as the similarity metric. Since each session is expected to have exactly two speakers, we continue the HAC procedure until two clusters are constructed. As a post-processing step, adjacent speech segments assigned to the same speaker are concatenated into a single speaker turn, allowing a maximum of 1 sec in-turn silence.

After diarization, we have the entire set of utterances clustered into two groups; however, there is not a natural correspondence between the cluster labels and the actual speaker roles (i.e., *therapist* and *client*). For our purposes, speaker role recognition (SRR) is exactly the task of finding the mapping between the two. We employ the speaker-level SRR approach used as baseline in Chapter 2, with the text provided by the ASR subsystem (without lattice rescoring), since early experiments showed that this approach gives perfect recognition results for all the sessions where diarization accurately distinguishes the two interlocutors. Let's denote the two clusters identified by diarization as $S_1$ and $S_2$, each one containing the utterances assigned to the two different speakers. We know a priori that one of those speakers is the therapist (T) and one is the client (C). In order to do the role matching, two trained LMs, one for the therapist ($\mathrm{LM}_T$) and one for the client ($\mathrm{LM}_C$), are used. We then estimate the perplexities of $S_1$ and $S_2$ with respect to the two LMs and assign to $S_i$ the role that yields the minimum perplexity. In case one role minimizes the perplexity for both speakers, we first assign the speaker for whom we are most confident. The confidence metric is based on the absolute distance between the two estimated perplexities[3]. The required LMs are 3-gram models trained with the SRILM toolkit (Stolcke, 2002), using the MI-train and CPTS[4] corpora with mixing parameters 0.8 and 0.2, respectively.

[2] https://kaldi-asr.org/models/m6

[3] For more details, please also refer to Algorithm 1 (Chapter 2).

[4] Those datasets have been introduced in Chapter 1 and are also described in Appendix B.

### 5.5.2 System with classification-based diarization

As an alternative, we explore the system of Figure 5.2, where the clustering-based diarization is replaced by a classification-based one. In order to do so, we follow the approach developed in Chapter 3 (Figure 3.2). The voiced segments derived from the VAD module are transcribed with a first pass of ASR and are then sub-segmented based on the textual information. For text segmentation we use the DeepSegment tool[5], which uses a BiLSTM-CRF architecture, similar to the one we built in Chapter 3. SRR is now performed at the turn level using the same LMs—$\text{LM}_T$ and $\text{LM}_C$—as in the baseline system of Section 5.5.1. In order to estimate the acoustic profiles (see Section 3.3) we use the 50% of the role-annotated segments per session about which we are most confident according to the perplexity-based criterion of equation (3.2). Those acoustic profiles are then used during a PLDA-based classification: like in the baseline system, each voiced segment, as predicted by VAD, is partitioned uniformly into subsegments of length equal to 1.5 sec with a shift of 0.25 sec, and each subsegment is labeled as belonging to the interlocutor who maximized the PLDA similarity. We use the same speaker representation as in the baseline system (employing the CallHome x-vector extractor) for both the profile estimation and the sub-segmentation step.

As shown in Figure 5.2, after the linguistically-aided diarization (for which ASR outputs are required), we have a second pass of ASR. The reason is that diarization defines different speech segments than the VAD-based ones used during the first pass and we wanted to have a fair comparison with the baseline system, keeping all the other modules (apart from diarization/role recognition) fixed. However, we should note that, as explained in Appendix B, the second ASR pass does not yield improved recognition results (with respect to the estimated word error rates) compared to the first pass.

[5] https://github.com/notAI-tech/deepsegment This is the same tool used for the utterance segmentation module of both systems (Figures 5.1 and 5.2) before behavioral coding—see also Appendix B.

## 5.6 Analysis and Results

### 5.6.1 Speaker diarization

We first evaluate the two different diarization systems described in Section 5.5. The standard evaluation metric, that we have also used in Chapters 3 and 4, is the diarization error rate (DER; Anguera et al., 2012) and it incorporates three sources of error: false alarms (the percentage of speech in the output but not in the ground truth), missed speech (the percentage of speech in the ground truth but not in the output), and speaker error (the percentage of speech assigned to the wrong speaker cluster after an optimal mapping between speaker clusters and true speaker labels). However, false alarm here is not representative of the algorithms' performance because of the specific implementation followed. In particular, we chose to concatenate adjacent speech segments assigned to the same speaker, if there is not a silence gap between them greater that 1 sec. This step increases DER, since it labels short non-voiced segments as belonging to some speaker, thus introducing false alarms. However, it creates longer speaker-homogeneous segments, which is beneficial to ASR, and, hence, to the overall system. What is important for the downstream task is to identify the therapist speech, and for that reason, we want to minimize missed speech and speaker error. Results with respect to those metrics are reported in Table 5.3. Those are estimated using the NIST `md-eval.pl` tool, with a forgiveness collar of 0.25 sec around each speaker boundary.

Table 5.3: Diarization results (%) for the UCC data.

| **diarization method** | **missed speech** | **speaker error** |
|---|---|---|
| clustering-based | 0.5 | 7.6 |
| classification-based | 0.5 | 4.9 |

*clustering-based* refers to the system of Figure 5.1 and *classification-based* refers to the system of Figure 5.2.

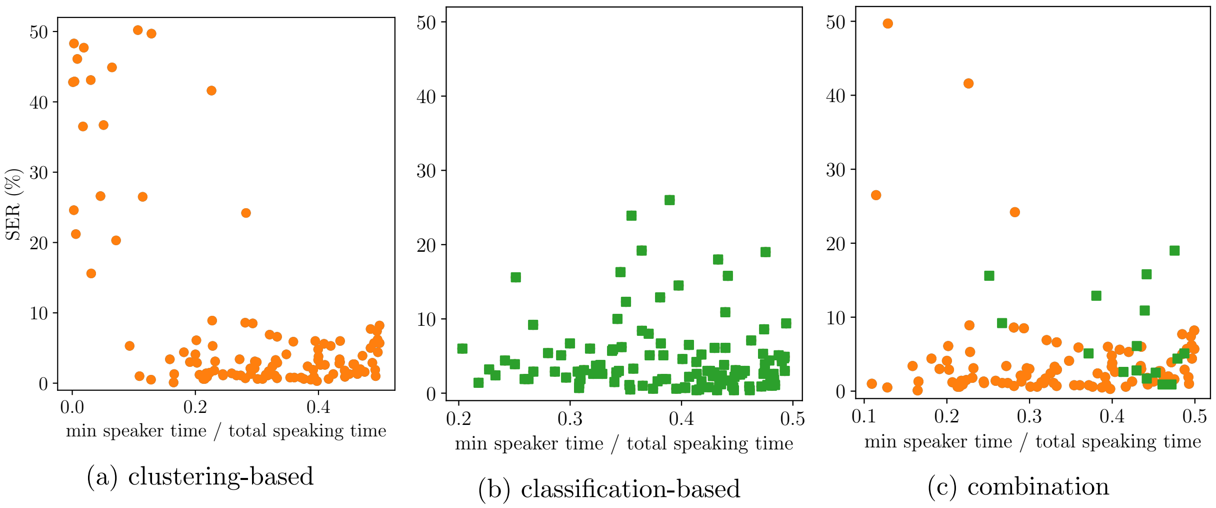
We can see that the overall diarization performance is improved by the classification-based system, thus validating the results of Chapter 3. However, a per-session analysis revealed that most of this performance gap is due to a handful of sessions for which the traditional, clustering-based diarization essentially failed, with a reported speaker error rate as high as 50%. At the same time, the clustering-based system occasionally performs even better than the classification-based one for sessions under clean acoustic conditions and featuring speakers with very dissimilar acoustic characteristics (e.g., male vs. female). In order to get the best of both words, and to avoid the increased computational complexity of the classification-based system whenever this is not needed[6], we propose to start with the clustering-based system for all the sessions and apply a simple proxy of diarization performance. If, according to this proxy, the clustering-based diarization fails, we halt processing and re-run diarization, using the classification-based, linguistically-aided system this time.

According to our proxy, the percentage of speech assigned to each one of the two speakers should be at least $m\%$ of the total speaking time, with $m = 10$ for the results reported here[7]. Since we deal with dyadic conversational scenarios, it is expected that each of the two speakers talks for a substantial amount of time. Even though therapy is not a normal dialogue and the provider often plays more the role of the listener (Hill, 2009), if one of the two interlocutors seems to not be participating in the conversation, then we are highly confident there is some problem. This may be an issue associated either with the audio quality, or with high speaker error introduced by the diarization module because the two speakers have similar acoustic characteristics.

Per-session results in terms of speaker error rate (SER) when using either the clustering-based or the classification-based system for all the sessions, or a combination of those based on the described threshold, are given in Figure 5.3. As we can see (Figure 5.3a), our quality safeguard is a reasonable proxy of diarization performance: most of the sessions with high estimated SER (more than 15%) are sessions where the speaking time of one of the interlocutors is very low (less than 10%), suggesting that the two speaker clusters were collapsed into one. When we choose to continue processing those sessions using the classification-based system (Figure 5.3c), the problem is alleviated.

[6] Note that the specific implementation of the classification-based system requires applying ASR and extracting x-vectors twice.

[7] This was one of the quality safeguards incorporated within the original system, presented in (Flemotomos et al., 2021). Since this system was designed with real-world deployment in mind, it was important to incorporate specific quality safeguards that help us both identify potential computational errors, including ones due to diarization, and determine whether the input was an actual therapy session or not (e.g., whether the therapist pushed the recording button by mistake). Based on those safeguards, if certain quality thresholds were not met, then the final report was not generated and feedback was not provided for the specific session. Instead, an error message was displayed to the counselor.



(a) clustering-based

(b) classification-based

(c) combination

Figure 5.3: Speaker error rate (SER) per UCC session for the different system designs illustrated in Figures 5.1 and 5.2. In (c), we use the *classification-based* system for the sessions where the speaking time of each speaker is not at least 10% of the overall speaking time according to the *clustering-based* diarization output.

### 5.6.2 Psychotherapy evaluation

When diarization fails, the error is propagated throughout the entire pipeline and the system cannot accurately code the therapist utterances. In fact, for seven of the sessions where the clustering-based diarization algorithm failed to sufficiently distinguish between the two speakers, the subsequent speaker-level SRR module (Figure 5.1) failed to find the right mapping between roles and speakers. This is not the case when for the "problematic" sessions we use the linguistically-aided, classification-based diarization where role assignment is done at the turn level (Figure 5.2). When we compare the total number of utterances per session that have been assigned to the therapist by the human annotators and by the automated systems, the Spearman correlation is increased from 0.478 ($p < 10^{-7}$) in the clustering-based system to 0.561 ($p < 10^{-9}$) in the system that uses either the clustering or the classification-based diarization according to the minimum speaking time criterion[8].

This behavior is reflected in the final evaluation of the overall system performance, as well. Evaluation with respect to utterance-level behavioral coding is not straightforward, since the utterances

[8] Those numbers correspond to the utterances predicted by the system after the *utterance segmentation* module (Figures 5.1 and 5.2).

given to the MISC predictor after automatic transcription are not the same as the ones defined by human transcribers. In that case, we use as a simple evaluation metric the correlation between the tallies, i.e., the counts of each MISC label in the manual coding trial and in the automatically generated report. The results are given in Table 5.4.

Table 5.4: Spearman correlation coefficients for the per-session counts of the utterance-level MISC labels between the manually-derived codes and the machine-generated ones for different diarization approaches.

| **MISC** | **clustering-based** | **classification-based** | **combination** |
|---|---|---|---|
| FA | 0.194$^*$ | **0.309**$^\dagger$ | 0.305$^\diamond$ |
| GI | 0.639$^\dagger$ | 0.507$^\dagger$ | **0.627**$^\dagger$ |
| RES | **0.303**$^*$ | 0.187$^*$ | 0.235$^*$ |
| REC | 0.388$^\dagger$ | **0.502**$^\dagger$ | 0.447$^\dagger$ |
| QUC | 0.634$^\dagger$ | 0.475$^\dagger$ | **0.639**$^\dagger$ |
| QUO | 0.524$^\dagger$ | 0.741$^\dagger$ | **0.753**$^\dagger$ |
| MIA | 0.451$^\dagger$ | 0.576$^\dagger$ | **0.596**$^\dagger$ |
| MIA | 0.455$^\dagger$ | 0.390$^\dagger$ | **0.474**$^\dagger$ |
| ST | 0.428$^\dagger$ | 0.549$^\dagger$ | **0.581**$^\dagger$ |
| mean | 0.446 | 0.471 | **0.517** |

*clustering-based* refers to the system of Figure 5.1 and *classification-based* refers to the system of Figure 5.2.

*combination*: use the *classification-based* system for the sessions where the speaking time of each speaker is not at least 10% of the overall speaking time according to the *clustering-based* system.

$^\dagger p < 0.001$, $^\diamond p < 0.01$, $^* p < 0.05$

We additionally report results with respect to session-level functionals commonly used in MI research. Those are the the ratio of reflections to questions (Re2Qu), the percentage of open questions out of all the questions (QUO2Qu), the percentage of complex reflections out of all the reflections (REC2Re), the MI adherence (MI-Adh), defined as the percentage of utterances not assigned the MIN code, as well as the ratio of therapist-attributed over client-attributed speaking time (Ther2Cl). As shown in Table 5.5, Re2Qu and QUO2Qu are reflected more accurately—on average, taking all sessions into account—with the *combination* approach. With respect to MI-Adh and Ther2Cl, results are better when only the classification-based diarization is used for all the sessions. The only metric for which the clustering-based approach yields the best results is REC2Re; however, Spearman correlations are not statistically significant for the particular metric. The reason behind the low overall performance with respect to the REC2Re metric is that our text-based MISC prediction algorithm had a high confusion rate between complex and simple reflections[9] (see also Appendix B).

Table 5.5: Spearman correlation coefficients for the session-level MISC aggregate metrics between the manually-derived codes and the machine-generated ones for different diarization approaches.

| metric | clustering-based | classification-based | combination |
|---|---|---|---|
| Re2Qu | 0.324 | 0.428 | **0.452** |
| QUO2Qu | 0.527 | 0.575 | **0.698** |
| REC2Re | 0.172 | 0.087 | 0.154 |
| MI-Adh | 0.354 | **0.509** | 0.418 |
| Ther2Cl | 0.720 | **0.823** | 0.815 |

*clustering-based* refers to the system of Figure 5.1 and *classification-based* refers to the system of Figure 5.2.
*combination*: use the *classification-based* system for the sessions where the speaking time of each speaker is not at least 10% of the overall speaking time according to the *clustering-based* system.
All correlations are significant ($p < 0.001$), apart from REC2Re ($p > 0.05$).

## 5.7 Conclusion

In this chapter we explored speaker role information and the impact it has on a real-world application. In particular, we presented a processing pipeline used to automatically evaluate recorded dyadic psychotherapy sessions, where accurate estimation of when a therapist talks is critical, since therapist-attributed speech needs to be coded at the utterance level. We applied the linguistically-aided, role-based diarization approach that we presented in Chapter 3 and we compared it with a traditional clustering-based diarization algorithm to study how diarization output can affect the downstream task of psychotherapy quality assessment. Experimental results showed that the linguistically-aided method can significantly outperform the baseline, especially for sessions where the latter fails to identify the two different speakers within the session, and thus the subsequent behavioral coding algorithm is not provided with accurate input.

Here we proposed first trying to process all the sessions with the simpler, and computationally less expensive, clustering-based system and only employing the linguistically-aided diarization for

[9] In the original system deployed in clinical settings, we have grouped complex and simple reflections into a single composite “reflections” label when generating the feedback report.

sessions where the first system failed. Since we cannot directly evaluate the diarization performance on unseen sessions, we used a quality proxy based on the minimum expected speaking time of the two interlocutors. More sophisticated combination methods (e.g., Stolcke & Yoshioka, 2019) and/or confidence metrics (e.g., Vaquero, Ortega, Miguel, & Lleida, 2013) for diarization systems can potentially further improve the final results.

The application of a competency rating tool, like the one we presented, in clinical settings could guarantee the provision of fast and low-cost feedback. Performance-based feedback is an essential aspect both for training new therapists and for maintaining acquired skills, and can eventually lead to improved quality of services and more positive clinical outcomes. Additionally, being able to accurately record, transcribe, and code interventions at large scale opens up ample opportunities for psychotherapy research studies with increased statistical power.

# Conclusions and Future Directions

## Summary and Main Contributions

In the previous chapters I proposed various methods to recognize speaker roles and use the inferred information to facilitate speech processing tasks. A main motivation behind the research conducted has been the reduction of error propagation in pipelined architectures for speech-based applications.

In Chapter 1 I showed that combining audio-based speaker clustering with language-based role recognition at the turn level can lead to substantial performance gains for the task of speaker role recognition (SRR). The linguistic information used for this work, however, was extracted from manual transcriptions. In Chapter 2 I extended the ideas behind language-based SRR for a more realistic scenario where the textual information is drawn from automatically derived transcripts. I did so by producing role-specific ASR outputs, suitably rescoring the decoding lattices produced by a generic ASR system. The proposed approach also led to slight improvements in ASR performance.

Moving to a different speech processing task, in Chapters 3 and 4 I utilized role-related information to improve the performance of speaker diarization, when applied in conversational interactions where speakers assume dissimilar roles. In particular, in Chapter 3 speaker roles were used to construct the acoustic profiles of the interlocutors, thus enabling us to convert speaker diarization from a clustering problem to a classification one. This method, however, assumed that every speaker in the conversation is mapped to a single role and vice-versa. In Chapter 4, I presented a more generic framework, where linguistic, role-based information is used to impose segment-wise constraints during the subsequent step of audio-based clustering.

Finally, in Chapter 5 I presented an end-to-end speech and language processing pipeline developed to transcribe and evaluate psychotherapy sessions to provide performance-based feedback to therapists. Speaker diarization and role recognition are crucial components within this competency rating tool, and I showed how employing a role-aided diarization approach can reduce error propagation and lead to improved overall results.

## Directions for Future Work

This dissertation has focused on the computational analysis of formal speaker roles within conversational interactions and on ways that role information can be used to facilitate core speech processing tasks, such as speaker diarization. With the evolution and success of end-to-end neural architectures, an exciting area of future research is towards unified frameworks where role recognition and other speech processing modules, such as speech recognition and speaker diarization, are combined together. Early works towards that direction have shown promising results, but leave ample room for improvements and further research (El Shafey, Soltau, & Shafran, 2019; Flemotomos, Chen, Atkins, & Narayanan, 2018).

An assumption made throughout this work is that the role concepts we study remain static during a single interaction. Even though this is in general true for formal roles (e.g., *patient* vs. *doctor*), informal roles emerge as a result of interpersonal dynamics and can change over the course of a conversation (Dowell et al., 2019). An interesting direction for future work would be an extension of the tools presented here to the analysis of informal, emergent roles, incorporating this additional element of temporal variability.

Both formal and informal roles can be manifest through specific behavioral patterns; this has been the main overarching idea behind the various models proposed in the previous chapters. However, how an individual behaves within a specific group and under specific circumstances is a function of various aspects, including personality traits and other dimensions of identity. A role that an individual assumes can be viewed as just one such dimension (Hare, 1994). Future research efforts could focus on the analysis and modeling of the relationship between speaker roles and identity characteristics, such as gender, age, and personality.

Finally, an exciting prospect would be the incorporation of role-specific information in voice assistants. While smart conversational agents are becoming part of our everyday lives, generic responses and lack of emotional intelligence remain a shortfall of dialogue generation models, posing obstacles to carrying long and naturalistic conversations (Roller et al., 2021). Allowing intelligent agents to assume specific roles and adopt role-aware behaviors would potentially give them more human-like conversational characteristics and would make them more adaptive to different environments.

# References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation* (pp. 265–283).

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 356-370. doi: 10.1109/TASL.2011.2125954

Anguera, X., Wooters, C., & Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2011–2022. doi: 10.1109/TASL.2007.902460

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*(1), 49. doi: 10.1186/1748-5908-9-49

Baer, J. S., Wells, E. A., Rosengren, D. B., Hartzler, B., Beadnell, B., & Dunn, C. (2009). Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. *Journal of Substance Abuse Treatment*, *37*(2), 191–202. doi: 10.1016/j.jsat.2009.01.003

Bakeman, R., & Quera, V. (2012). Behavioral observation. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 207–225). American Psychological Association. doi: 10.1037/13619-013

Bales, R. F. (1950). A set of categories for the analysis of small group interaction. *American Sociological Review*, *15*(2), 257–263. doi: 10.2307/2086790

Barahona, L. M. R., Tseng, B.-H., Dai, Y., Mansfield, C., Ramadan, O., Ultes, S., … Gasic, M. (2018). Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis* (pp. 44–54). doi: 10.18653/v1/W18-5606

Barzilay, R., Collins, M., Hirschberg, J., & Whittaker, S. (2000). The rules behind roles: Identifying speaker role in radio broadcasts. In *Proceedings of the 7th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence* (pp. 679–684). doi: 10.7916/D8PC39Q9

Bazillon, T., Maza, B., Rouvier, M., Bechet, F., & Nasr, A. (2011). Speaker role recognition using question detection and characterization. In *Proceedings of Interspeech 2011* (pp. 917–920). doi: 10.21437/Interspeech.2011-442

Beňuš, Š. (2014). Social aspects of entrainment in spoken interaction. *Cognitive Computation*, *6*(4), 802–813. doi: 10.1007/s12559-014-9261-4

Beňuš, Š., Gravano, A., Levitan, R., Levitan, S. I., Willson, L., & Hirschberg, J. (2014). Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, *71*(1), 3–14.

Biddle, B. J. (1986). Recent developments in role theory. *Annual Review of Sociology*, *12*(1), 67–92. doi: 10.1146/annurev.so.12.080186.000435

Bigot, B., Ferrané, I., Pinquier, J., & André-Obrecht, R. (2010). Speaker role recognition to help spontaneous conversational speech detection. In *Proceedings of the 2010 International Workshop on Searching Spontaneous Conversational Speech* (pp. 5–10). doi: 10.1145/1878101.1878104

Bigot, B., Fredouille, C., & Charlet, D. (2013). Speaker role recognition on TV broadcast documents. In *Proceedings of the 1st Workshop on Speech, Language and Audio in Multimedia* (pp. 66–71).

Bigot, B., Pinquier, J., Ferrané, I., & André-Obrecht, R. (2010). Looking for relevant features for speaker role recognition. In *Proceedings of Interspeech 2010* (pp. 1057–1060). doi: 10.21437/Interspeech.2010-137

Black, M. P., Katsamanis, A., Baucom, B. R., Lee, C.-C., Lammert, A. C., Christensen, A., … Narayanan, S. S. (2013). Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication*, *55*(1), 1–21. doi: 10.1016/j.specom.2011.12.003

Bost, Xavier and Linares, Georges. (2014). Constrained speaker diarization of TV series based on visual patterns. In *Proceedings of the 2014 IEEE Spoken Language Technology Workshop* (pp. 390–395). doi: 10.1109/SLT.2014.7078606

Bozonnet, S., Evans, N., Anguera, X., Vinyals, O., Friedland, G., & Fredouille, C. (2010). System output combination for improved speaker diarization. In *Proceedings of Interspeech 2010* (pp. 2642–2645). doi: 10.21437/Interspeech.2010-701

Bredin, H. (2017a). pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Proceedings of Interspeech 2017* (pp. 3587–3591). doi: 10.21437/Interspeech.2017-411

Bredin, H. (2017b). Tristounet: Triplet loss for speaker turn embedding. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5430–5434). doi: 10.1109/ICASSP.2017.7953194

Can, D., Atkins, D. C., & Narayanan, S. S. (2015). A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. In *Proceedings of Interspeech 2015* (pp. 339–343). doi: 10.21437/Interspeech.2015-151

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., ... Wellner, P. (2005). The AMI meeting corpus: A pre-announcement. In *Proceedings of the 2nd International Conference on Machine Learning for Multimodal Interaction* (pp. 28–39). doi: 10.1007/11677482_3

Chen, S., & Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop* (pp. 127–132).

Chen, Z., Flemotomos, N., Ardulov, V., Creed, T. A., Imel, Z. E., Atkins, D. C., & Narayanan, S. (2021). Feature fusion strategies for end-to-end evaluation of cognitive behavior therapy sessions. In *Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (p. 1836-1839). doi: 10.1109/EMBC46164.2021.9629694

Cheng, S.-S., & Wang, H.-M. (2003). A sequential metric-based audio segmentation method via the bayesian information criterion. In *Proceedings of the 8th European Conference on Speech*

*Communication and Technology* (pp. 945–948).

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 69–71).

Curran, J., Parry, G. D., Hardy, G. E., Darling, J., Mason, A.-M., & Chambers, E. (2019). How does therapy harm? A model of adverse process using task analysis in the meta-synthesis of service users' experience. *Frontiers in Psychology*, *10*, 347. doi: 10.3389/fpsyg.2019.00347

Damnati, G., & Charlet, D. (2011a). Multi-view approach for speaker turn role labeling in TV broadcast news shows. In *Proceedings of Interspeech 2011* (pp. 1285–1288). doi: 10.21437/Interspeech.2011-430

Damnati, G., & Charlet, D. (2011b). Robust speaker turn role labeling of tv broadcast news shows. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5684–5687). doi: 10.1109/ICASSP.2011.5947650

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 699–708). doi: 10.1145/2187836.2187931

Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., & Na, H. (2021). ECAPA-TDNN embeddings for speaker diarization. In *Proceedings of interspeech 2021* (pp. 3560–3564). doi: 10.21437/Interspeech.2021-941

Demasi, O., Li, Y., & Yu, Z. (2020). A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3623–3636). doi: 10.18653/v1/2020.findings-emnlp.324

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). doi: 10.18653/v1/N19-1423

Dimitriadis, D., & Fousek, P. (2017). Developing on-line speaker diarization system. In *Proceedings of Interspeech 2017* (pp. 2739–2743). doi: 10.21437/Interspeech.2017-166

Dowell, N. M., Nixon, T. M., & Graesser, A. C. (2019). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior Research Methods*, *51*(3), 1007–1041. doi: 10.3758/s13428-018-1102-z

Dufour, R., Esteve, Y., & Deléglise, P. (2011). Investigation of spontaneous speech characterization applied to speaker role recognition. In *Proceedings of Interspeech 2011* (pp. 917–920). doi: 10.21437/Interspeech.2011-370

El Shafey, L., Soltau, H., & Shafran, I. (2019). Joint speech recognition and speaker diarization via sequence transduction. *Proceedings of Interspeech 2019*, 396–400. doi: 10.21437/Interspeech.2019-1943

Favre, S., Dielmann, A., & Vinciarelli, A. (2009). Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models. In *Proceedings of the 17th ACM International Conference on Multimedia* (pp. 585–588). doi: 10.1145/1631272.1631362

Flemotomos, N., Chen, Z., Atkins, D. C., & Narayanan, S. (2018). Role annotated speech recognition for conversational interactions. In *Proceedings of the 2018 IEEE Spoken Language Technology Workshop* (pp. 1036–1043). doi: 10.1109/SLT.2018.8639611

Flemotomos, N., Georgiou, P., & Narayanan, S. (2019). Role specific lattice rescoring for speaker role recognition from speech recognition outputs. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7330–7334). doi: 10.1109/ICASSP.2019.8683900

Flemotomos, N., Georgiou, P., & Narayanan, S. (2020). Linguistically aided speaker diarization using speaker role information. In *Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2020)* (pp. 117–124). doi: 10.21437/Odyssey.2020-17

Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., ... Narayanan, S. (2021). Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*. doi: 10.3758/s13428-021-01623-4

Flemotomos, N., Martinez, V. R., Gibson, J., Atkins, D., Creed, T., & Narayanan, S. (2018). Language features for automated evaluation of cognitive behavior psychotherapy sessions. *Proceedings of Interspeech 2018*, 1908–1912. doi: 10.21437/Interspeech.2018-1518

Flemotomos, N., & Narayanan, S. (2022). Multimodal clustering with role induced constraints for speaker diarization. *arXiv preprint arXiv:2204.00657*.

Flemotomos, N., Papadopoulos, P., Gibson, J., & Narayanan, S. (2018). Combined speaker clustering and role recognition in conversational speech. In *Proceedings of Interspeech 2018* (pp. 1378–1382). doi: 10.21437/Interspeech.2018-1654

Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., & Watanabe, S. (2019). End-to-end neural speaker diarization with self-attention. In *Proceedings of the 2020 IEEE Automatic Speech Recognition and Understanding Workshop* (pp. 296–303). doi: 10.1109/ASRU46091.2019.9003959

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, *12*(2), 75–98. doi: 10.1006/csla.1998.0043

Gançarski, P., Dao, T.-B.-H., Crémilleux, B., Forestier, G., & Lampert, T. (2020). Constrained clustering: Current and new trends. In P. Marquis, O. Papini, & H. Prade (Eds.), *A Guided Tour of Artificial Intelligence Research: Volume II: AI Algorithms* (pp. 447–484). Springer. doi: 10.1007/978-3-030-06167-8_14

Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., & McCree, A. (2017). Speaker diarization using deep neural network embeddings. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4930–4934). doi: 10.1109/ICASSP.2017.7953094

Garg, N. P., Favre, S., Salamin, H., Hakkani Tür, D., & Vinciarelli, A. (2008). Role recognition for meeting participants: An approach based on lexical information and social network analysis. In *Proceedings of the 16th ACM International Conference on Multimedia* (pp. 693–696). doi: 10.1145/1459359.1459462

Garnier-Rizet, M., Adda, G., Cailliau, F., Guillemin-Lanne, S., Waast-Richard, C., Lamel, L., … Waast-Richard, C. (2008). CallSurf: Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 2623–2628).

Gaume, J., Gmel, G., Faouzi, M., & Daeppen, J.-B. (2009). Counselor skill influences outcomes of brief motivational interventions. *Journal of Substance Abuse Treatment*, *37*(2), 151–159. doi: 10.1016/j.jsat.2008.12.001

Georgiou, P. G., Black, M. P., Lammert, A., Baucom, B., & Narayanan, S. S. (2011a). “That’s aggravating, very aggravating”: Is it possible to classify behaviors in couple interactions using automatically derived lexical features? In *Proceedings of Affective Computing and Intelligent Interaction (ACII 2011), Lecture Notes in Computer Science* (Vol. 6974). doi: 10.1007/978-3-642-24600-5_12

Georgiou, P. G., Black, M. P., Lammert, A. C., Baucom, B. R., & Narayanan, S. S. (2011b). "That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features? In *Proceedings of the 2011 International Conference on Affective Computing and Intelligent Interaction* (pp. 87–96). doi: 10.1007/978-3-642-24600-5_12

Gibson, J., Atkins, D., Creed, T., Imel, Z., Georgiou, P., & Narayanan, S. (2022). Multi-label multi-task deep learning for behavioral coding. *IEEE Transactions on Affective Computing*, *13*(1), 508–518. doi: 10.1109/TAFFC.2019.2952113

Gleave, E., Welser, H. T., Lento, T. M., & Smith, M. A. (2009). A conceptual and operational definition of social role in online community. In *Proceedings of the 42nd Hawaii International Conference on System Sciences* (pp. 1–11).

Graff, D., Wu, Z., MacIntyre, R., & Liberman, M. (1997). The 1996 broadcast news speech and language-model corpus. In *Proceedings of the 1997 DARPA Speech Recognition Workshop* (pp. 11–14).

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. doi: 10.20982/tqmp.08.1.p023

Hare, A. P. (1994). Types of roles in small groups: A bit of history and a current perspective. *Small Group Research*, *25*(3), 433–448. doi: 10.1177/1046496494253005

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi: 10.3102/003465430298487

Hill, C. E. (2009). *Helping skills: Facilitating, exploration, insight, and action.* American Psychological Association.

Hori, T., & Nakamura, A. (2013). *Speech recognition algorithms using weighted finite-state transducers.* Morgan & Claypool.

Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., & Nagamatsu, K. (2020). End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. In

*Proceedings of Interspeech 2020* (pp. 269–273). doi: 10.21437/Interspeech.2020-1022

Houck, J. M., Moyers, T. B., Miller, W. R., Glynn, L. H., & Hallgren, K. A. (2010). *Motivational interviewing skill code (MISC) version 2.5.* (Available from http://casaa.unm.edu/download/misc25.pdf)

Hrúz, M., & Zajíc, Z. (2017). Convolutional neural network for speaker change detection in telephone speaker diarization system. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4945–4949). doi: 10.1109/ICASSP.2017.7953097

Huang, J., Marcheret, E., Visweswariah, K., Libal, V., & Potamianos, G. (2007). The IBM Rich Transcription 2007 speech-to-text systems for lecture meetings. In R. Stiefelhagen, R. Bowers, & J. Fiscus (Eds.), *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007* (pp. 429–441). Springer. doi: 10.1007/978-3-540-68585-2_40

Hutchinson, B., Zhang, B., & Ostendorf, M. (2010). Unsupervised broadcast conversation speaker role labeling. In *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing* (pp. 5322–5325). doi: 10.1109/ICASSP.2010.5494958

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*, *52*(1), 19–30. doi: 10.1037/a0036841

India Massana, M. À., Rodríguez Fonollosa, J. A., & Hernando Pericás, F. J. (2017). LSTM neural network-based speaker segmentation using acoustic and language modelling. In *Proceedings of Interspeech 2017* (pp. 2834–2838). doi: 10.21437/Interspeech.2017-407

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, *6*(11), e12106. doi: 10.2196/12106

Ioffe, S. (2006). Probabilistic linear discriminant analysis. In *Proceedings of the 9th European Conference in Computer Vision, Part IV* (pp. 531–542). doi: 10.1007/11744085_41

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., ... Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. I/364–I/367). doi: 10.1109/ICASSP.2003.1198793

Jati, A., & Georgiou, P. G. (2017). Speaker2Vec: Unsupervised learning and adaptation of a speaker manifold using deep neural networks with an evaluation on speaker segmentation. In *Proceedings of Interspeech 2017* (pp. 3567–3571). doi: 10.21437/Interspeech.2017-1650

Johnstone, B. (1996). *The linguistic individual: Self-expression in language and linguistics.* Oxford University Press.

Jung, C. G. (2014). *Two essays on analytical psychology.* Routledge.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, *62*(6), 593–602. doi: 10.1001/archpsyc.62.6.593

Kinoshita, K., Delcroix, M., & Tawara, N. (2021a). Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech. In *Proceedings of Interspeech 2021* (pp. 3565–3569). doi: 10.21437/Interspeech.2021-1004

Kinoshita, K., Delcroix, M., & Tawara, N. (2021b). Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7198–7202). doi: 10.1109/ICASSP39728.2021.9414333

Kipper, D. A. (1992). Psychodrama: Group psychotherapy through role playing. *International Journal of Group Psychotherapy*, *42*(4), 495–521. doi: 10.1080/00207284.1992.11490720

Klatte, R., Strauss, B., Flückiger, C., & Rosendahl, J. (2018). Adverse effects of psychotherapy: protocol for a systematic review and meta-analysis. *Systematic Reviews*, *7*, 135. doi: 10.1186/s13643-018-0802-x

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). *Nonverbal communication in human interaction.* Cengage Learning.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proceedings of Interspeech 2015* (pp. 3586–3589). doi: 10.21437/Interspeech.2015-711

Kodish-Wachs, J., Agassi, E., Kenny III, P., & Overhage, J. M. (2018). A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. In *AMIA Annual Symposium Proceedings* (pp. 683–689).

Koluguri, N. R., Park, T., & Ginsburg, B. (2022). TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing.*

Komninos, A., & Manandhar, S. (2016). Dependency based embeddings for sentence classification tasks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1490–1500). doi: 10.18653/v1/N16-1175

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology.* Sage publications.

Kuich, W., & Salomaa, A. (1986). *Semirings, automata, languages.* Springer Verlag.

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of educational research*, *58*(1), 79–97. doi: 10.3102/00346543058001079

Lambert, M. J., & Bergin, A. E. (2002). The effectiveness of psychotherapy. In M. Hersen & W. Sledge (Eds.), *Encyclopedia of Psychotherapy* (Vol. 1, pp. 709–714). USA: Elsevier Science. doi: 10.1016/B0-12-343010-0/00084-2

Lambert, M. J., & Ogles, B. M. (1997). The effectiveness of psychotherapy supervision. In C. E. Watkins (Ed.), *Handbook of Psychotherapy Supervision* (pp. 421–446). John Wiley & Sons, Inc.

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy*, *55*(4), 520–537. doi: 10.1037/pst0000167

Laurent, A., Camelin, N., & Raymond, C. (2014). Boosting bonsai trees for efficient features combination: application to speaker role identification. In *Proceedings of Interspeech 2014* (pp. 76–80).

Lee, F.-T., Hull, D., Levine, J., Ray, B., & McKeown, K. (2019). Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology* (pp. 12–23). doi: 10.18653/v1/W19-3002

Li, Y., Wang, Q., Zhang, X., Li, W., Li, X., Yang, J., … He, Q. (2017). Unsupervised classification of speaker roles in multi-participant conversational speech. *Computer Speech & Language*, *42*, 81–99. doi: 10.1016/j.csl.2016.09.002

Liu, D., & Kubala, F. (2004). Online speaker clustering. In *Proceedings of the 2004 IEEE In-*

*ternational Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. I–333). doi: 10.1109/ICASSP.2004.1325990

Liu, Y. (2006). Initial study on automatic identification of speaker role in broadcast news speech. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers* (pp. 81–84).

Ljolje, A., Pereira, F., & Riley, M. (1999). Efficient general lattice generation and rescoring. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech 1999)* (pp. 1251–1254).

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223. doi: 10.1023/A:1005592401947

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of the 7th international conference on learning representations.*

Lu, Z., & Peng, Y. (2013). Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications. *International Journal of Computer Vision*, *103*(3), 306–325. doi: 10.1007/s11263-012-0602-z

Luz, S. (2009). Locating case discussion segments in recorded medical team meetings. In *Proceedings of the 3rd Workshop on Searching Spontaneous Conversational Speech* (pp. 21–30). doi: 10.1145/1631127.1631131

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1064–1074). doi: 10.18653/v1/P16-1101

Madson, M. B., & Campbell, T. C. (2006). Measures of fidelity in motivational enhancement: a systematic review. *Journal of Substance Abuse Treatment*, *31*(1), 67–73. doi: 10.1016/j.jsat.2006.03.010

Magill, M., Gaume, J., Apodaca, T. R., Walthers, J., Mastroleo, N. R., Borsari, B., & Longabaugh, R. (2014). The technical hypothesis of motivational interviewing: A meta-analysis of MI's key causal model. *Journal of Consulting and Clinical Psychology*, *82*(6), 973–983. doi: 10.1037/a0036833

Mao, H. H., Li, S., McAuley, J., & Cottrell, G. W. (2020). Speech recognition and multi-speaker diarization of long conversations. In *Proceedings of Interspeech 2020* (pp. 691–695). doi: 10.21437/Interspeech.2020-3039

Marcos-García, J.-A., Martínez-Monés, A., & Dimitriadis, Y. (2015). DESPRO: A method based on roles to provide collaboration analysis support adapted to the participants in CSCL situations. *Computers & Education*, *82*, 335–353. doi: 10.1016/j.compedu.2014.10.027

Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., ... others (2020). The STC system for the CHiME-6 challenge. In *Proceedings of the 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)* (pp. 36–41). doi: 10.21437/CHiME.2020-9

Meng, Z., Mou, L., & Jin, Z. (2017). Hierarchical RNN with static sentence-level attention for text-based speaker change detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2203–2206). doi: 10.1145/3132847.3133110

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (pp. 3111–3119).

Miller, W. R., & Rollnick, S. (2012). *Motivational interviewing: Helping people change.* Guilford press.

Miller, W. R., Sorensen, J. L., Selzer, J. A., & Brigham, G. S. (2006). Disseminating evidence-based practices in substance abuse treatment: A review with suggestions. *Journal of Substance Abuse Treatment*, *31*(1), 25–39. doi: 10.1016/j.jsat.2006.03.005

Mohri, M., Pereira, F., & Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, *16*(1), 69–88. doi: 10.1006/csla.2001.0184

Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M., & Miller, W. R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, *28*(1), 19–26. doi: 10.1016/j.jsat.2004.11.001

Mudrack, P. E., & Farrell, G. M. (1995). An examination of functional role behavior and its consequences for individuals in group settings. *Small Group Research*, *26*(4), 542–571. doi: 10.1177/1046496495264005

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In

*Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (pp. 849–856).

Ordelman, R., De Jong, F., & Larson, M. (2009). Enhanced multimedia content access and exploitation using semantic speech retrieval. In *Proceedings of the 2009 IEEE International Conference on Semantic Computing* (pp. 521–528). doi: 10.1109/ICSC.2009.80

Pal, M., Kumar, M., Peri, R., Park, T. J., Kim, S. H., Lord, C., … Narayanan, S. (2020). Speaker diarization using latent space clustering in generative adversarial network. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6504–6508). doi: 10.1109/ICASSP40776.2020.9053952

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5206–5210). doi: 10.1109/ICASSP.2015.7178964

Park, T. J., & Georgiou, P. (2018). Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks. In (pp. 1373–1377). doi: 10.21437/Interspeech.2018-1364

Park, T. J., Han, K. J., Huang, J., He, X., Zhou, B., Georgiou, P., & Narayanan, S. (2019). Speaker diarization with lexical information. In *Proceedings of Interspeech 2019* (pp. 391–395). doi: 10.21437/Interspeech.2019-1947

Park, T. J., Han, K. J., Kumar, M., & Narayanan, S. (2019). Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. *IEEE Signal Processing Letters*, *27*, 381–385. doi: 10.1109/LSP.2019.2961071

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, *72*, 101317. doi: 10.1016/j.csl.2021.101317

Park, T. J., Kumar, M., Flemotomos, N., Pal, M., Peri, R., Lahiri, R., … Narayanan, S. (2019). The second DIHARD challenge: System description for USC-SAIL team. In *Proceedings of Interspeech 2019* (pp. 998–1002). doi: 10.21437/Interspeech.2019-1903

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the Workshop on Speech and Natural Language* (pp. 357–362). doi: 10.3115/

1075527.1075614

Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proceedings of Interspeech 2015* (pp. 3214–3218). doi: 10.21437/Interspeech.2015-647

Perry, J. C., Banon, E., & Ianni, F. (1999). Effectiveness of psychotherapy for personality disorders. *American Journal of Psychiatry*, *156*(9), 1312–1321. doi: 10.1176/ajp.156.9.1312

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., . . . Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding.* (IEEE Catalog No.: CFP11SRW-USB)

Povey, D., Hannemann, M., Boulianne, G., Burget, L., Ghoshal, A., Janda, M., . . . Vu, N. T. (2012). Generating exact lattices in the WFST framework. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing* (p. 4213-4216). doi: 10.1109/ICASSP.2012.6288848

Prince, S. J., & Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Proceedings of the 2007 IEEE 11th International Conference on Computer Vision* (pp. 1–8). doi: 10.1109/ICCV.2007.4409052

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., . . . Hensley, M. (2011). Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(2), 65–76. doi: 10.1007/s10488-010-0319-7

Prokopalo, Y., Shamsi, M., Barrault, L., Meignier, S., & Larcher, A. (2021). Active correction for speaker diarization with human in the loop. In *Proceedings of IberSPEECH* (pp. 260–264). doi: 10.21437/IberSPEECH.2021-55

Quiroz, J. C., Laranjo, L., Kocaballi, A. B., Berkovsky, S., Rezazadegan, D., & Coiera, E. (2019). Challenges of developing a digital scribe to reduce clinical documentation burden. *npj Digital Medicine*, *2*, 114. doi: 10.1038/s41746-019-0190-1

Rasipuram, S., & Jayagopi, D. B. (2018). Automatic assessment of communication skill in interview-based interactions. *Multimedia Tools and Applications*, *77*, 18709–18739. doi: 10.1007/s11042-018-5654-9

Reimers, N., & Gurevych, I. (2017). Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 338–348). doi: 10.18653/v1/D17-1035

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., ... Weston, J. (2021). Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 300–325). doi: 10.18653/v1/2021.eacl-main.24

Rousseau, A., Deléglise, P., & Esteve, Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp. 3935–3939).

Rouvier, M., Delecraz, S., Favre, B., Bendris, M., & Bechet, F. (2015). Multimodal embedding fusion for robust speaker role recognition in video broadcast. In *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 383–389). doi: 10.1109/ASRU.2015.7404820

Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., ... Liberman, M. (2021). The third DIHARD diarization challenge. In *Proceedings of Interspeech 2021* (pp. 3570–3574). doi: 10.21437/Interspeech.2021-1208

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In J. Schenkein (Ed.), *Studies in the Organization of Conversational Interaction* (pp. 7–55). Elsevier. doi: 10.1016/B978-0-12-623550-0.50008-2

Sak, H., Saraçlar, M., & Güngör, T. (2010). On-the-fly lattice rescoring for real-time automatic speech recognition. In *Proceedings of Interspeech 2010* (pp. 2450–2453). doi: 10.21437/Interspeech.2010-532

Salamin, H., & Vinciarelli, A. (2012). Automatic role recognition in multiparty conversations: An approach based on turn organization, prosody, and conditional random fields. *IEEE Transactions on Multimedia*, *14*(2), 338–345. doi: 10.1109/TMM.2011.2173927

Saon, G., Soltau, H., Nahamoo, D., & Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 55–59). doi: 10.1109/ASRU.2013.6707705

Sapru, A., & Bourlard, H. (2015). Automatic recognition of emergent social roles in small group interactions. *IEEE Transactions on Multimedia*, *17*(5), 746–760. doi: 10.1109/TMM.2015.2408437

Sapru, A., & Valente, F. (2012). Automatic speaker role labeling in AMI meetings: recognition of formal and social roles. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5057–5060). doi: 10.1109/ICASSP.2012.6289057

Sapru, A., Yella, S. H., & Bourlard, H. (2014). Improving speaker diarization using social role information. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 101–105). doi: 10.1109/ICASSP.2014.6853566

Saxon, D., Barkham, M., Foster, A., & Parry, G. (2017). The contribution of therapist effects to patient dropout and deterioration in the psychological therapies. *Clinical Psychology & Psychotherapy*, *24*(3), 575–588. doi: 10.1002/cpp.2028

Schaefer, J. D., Caspi, A., Belsky, D. W., Harrington, H., Houts, R., Horwood, L. J., ... Moffitt, T. E. (2017). Enduring mental health: prevalence and prediction. *Journal of Abnormal Psychology*, *126*(2), 212–224. doi: 10.1037/abn0000232

Schwalbe, C. S., Oh, H. Y., & Zweben, A. (2014). Sustaining motivational interviewing: A meta-analysis of training studies. *Addiction*, *109*(8), 1287–1294. doi: 10.1111/add.12558

Sell, G., & Garcia-Romero, D. (2014). Speaker diarization with PLDA i-vector scoring and unsupervised calibration. In *Proceedings of the 2014 IEEE Spoken Language Technology Workshop* (pp. 413–417). doi: 10.1109/SLT.2014.7078610

Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., ... Khudanpur, S. (2018). Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In *Proceedings of Interspeech 2018* (pp. 2808–2812). doi: 10.21437/Interspeech.2018-1893

Shiner, B., D'Avolio, L. W., Nguyen, T. M., Zayed, M. H., Watts, B. V., & Fiore, L. (2012). Automated classification of psychotherapy note text: implications for quality assessment in PTSD care. *Journal of Evaluation in Clinical Practice*, *18*(3), 698–701. doi: 10.1111/j.1365-2753.2011.01634.x

Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D., & Glass, J. (2011). Exploiting intra-conversation variability for speaker diarization. In *Proceedings of interspeech 2011* (pp. 945–948). doi: 10.21437/Interspeech.2011-383

Silovsky, J., Zdansky, J., Nouza, J., Cerva, P., & Prazak, J. (2012). Incorporation of the ASR output in speaker segmentation and clustering within the task of speaker diarization of broadcast streams. In *Proceedings of the 2012 IEEE 14th International Workshop on Multimedia Signal Processing* (pp. 118–123). doi: 10.1109/MMSP.2012.6343426

Singla, K., Chen, Z., Flemotomos, N., Gibson, J., Can, D., Atkins, D. C., & Narayanan, S. (2018). Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. In *Proceedings of Interspeech 2018* (pp. 3413–3417). doi: 10.21437/Interspeech.2018-2551

Siniscalchi, S. M., Li, J., & Lee, C.-H. (2006). A study on lattice rescoring with knowledge scores for automatic speech recognition. In *Proceedings of Interspeech 2006* (p. paper 1319-Mon3A2O.1). doi: 10.21437/Interspeech.2006-198

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5329–5333). doi: 10.1109/ICASSP.2018.8461375

Song, H., Zhang, W.-N., Cui, Y., Wang, D., & Liu, T. (2019). Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (pp. 5190–5196). doi: 10.24963/ijcai.2019/721

Stolcke, A. (2002). SRILM–an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing* (pp. 901–904).

Stolcke, A., & Yoshioka, T. (2019). DOVER: A method for combining diarization outputs. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop* (p. 757-763). doi: 10.1109/ASRU46091.2019.9004031

Strijbos, J.-W., & De Laat, M. F. (2010). Developing the role concept for computer-supported collaborative learning: An explorative synthesis. *Computers in Human Behavior*, *26*(4), 495–505. doi: 10.1016/j.chb.2009.08.014

Substance Abuse and Mental Health Services Administration. (2019). *Key substance use and mental health indicators in the United States: Results from the 2018 national survey on drug use and health.* Center for Behavioral Health Statistics and Quality.

Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2019). Development and evaluation of ClientBot: Patient-like conversational agent to train basic counseling skills.

*Journal of Medical Internet Research*, *21*(7), e12529. doi: 10.2196/12529

Thomas, S., Saon, G., Van Segbroeck, M., & Narayanan, S. S. (2015). Improvements to the IBM speech activity detection system for the DARPA RATS program. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4500–4504). doi: 1109/ICASSP.2015.7178822

Tranter, S. (2005). Two-way cluster voting to improve speaker diarisation performance. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. I/753–I/756). doi: 10.1109/ICASSP.2005.1415223

Tripathi, A., Lu, H., Sak, H., Moreno, I. L., Wang, Q., & Xia, W. (2022). Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing.*

Valente, F., Vijayasenan, D., & Motlicek, P. (2011). Speaker diarization of meetings based on speaker role n-gram models. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4416–4419). doi: 10.1109/ICASSP.2011.5947333

Vaquero, C., Ortega, A., Miguel, A., & Lleida, E. (2013). Quality assessment for speaker diarization and its application in speaker characterization. *Transactions on Audio, Speech, and Language Processing*, *21*(4), 816–827. doi: 10.1109/TASL.2012.2236317

Vinciarelli, A. (2006). Sociometry based multiparty audio recordings summarization. In *Proceedings of the 18th International Conference on Pattern Recognition* (pp. 1154–1157). doi: 10.1109/ICPR.2006.1063

Vinciarelli, A. (2007). Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, *9*(6), 1215–1226. doi: 10.1109/TMM.2007.902882

Vinciarelli, A., & Favre, S. (2007). Broadcast news story segmentation using social network analysis and hidden markov models. In *Proceedings of the 15th ACM International Conference on Multimedia* (pp. 261–264). doi: 10.1145/1291233.1291287

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L. (2018). Speaker diarization with LSTM. In *Proceedings of the 2018 ieee international conference on acoustics, speech and* *signal processing* (pp. 5239–5243). doi: 10.1109/ICASSP.2018.8462628

Wang, W., Yaman, S., Precoda, K., & Richey, C. (2011). Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5556–5559). doi: 10.1109/ICASSP.2011.5947618

Wang, Y., He, M., Niu, S., Sun, L., Gao, T., Fang, X., … Lee, C.-H. (2021). USTC-NELSLIP system description for DIHARD-III challenge. *arXiv preprint arXiv:2103.10661*.

Weinberger, A., Stegmann, K., & Fischer, F. (2010). Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). *Computers in Human behavior*, *26*(4), 506–515. doi: 10.1016/j.chb.2009.08.007

Weisz, J. R., Weiss, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: a meta-analysis of treatment outcome studies. *Psychological Bulletin*, *117*(3), 450–468. doi: 10.1037/0033-2909.117.3.450

Williams, W., Prasad, N., Mrva, D., Ash, T., & Robinson, T. (2015). Scaling recurrent neural network language models. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 5391–5395). doi: 10.1109/ICASSP.2015.7179001

Xiao, B., Bone, D., Segbroeck, M. V., Imel, Z. E., Atkins, D. C., Georgiou, P. G., & Narayanan, S. S. (2014). Modeling therapist empathy through prosody in drug addiction counseling. In *Proceedings of Interspeech 2014* (pp. 213–217). doi: 10.21437/Interspeech.2014-55

Xiao, B., Can, D., Georgiou, P. G., Atkins, D., & Narayanan, S. S. (2012). Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference.*

Xiao, B., Can, D., Gibson, J., Imel, Z. E., Atkins, D. C., Georgiou, P. G., & Narayanan, S. S. (2016). Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In *Proceedings of Interspeech 2016* (pp. 908–912). doi: 10.21437/Interspeech.2016-1560

Xiao, B., Huang, C., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. S. (2016). A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science*, *2*, e59. doi: 10.7717/peerj-cs.59

Xiao, B., Imel, Z. E., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2015). “Rate my therapist”: Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLOS ONE*, *10*(12), e0143055. doi: 10.1371/journal.pone.0143055

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., … Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2410–2423. doi: 10.1109/TASLP.2017.2756440

Xu, H., Chen, T., Gao, D., Wang, Y., Li, K., Goel, N., … Khudanpur, S. (2018). A pruned RNNLM lattice-rescoring algorithm for automatic speech recognition. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processin* (p. 5929-5933). doi: 10.1109/ICASSP.2018.8461974

Yang, J., & Zhang, Y. (2018). NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of ACL 2018, System Demonstrations* (pp. 74–79). doi: 10.18653/v1/P18-4013

Yin, R., Bredin, H., & Barras, C. (2018). Neural speech turn segmentation and affinity propagation for speaker diarization. *Proceedings of Interspeech 2018*, 1393–1397. doi: 10.21437/Interspeech.2018-1750

Yoshioka, T., Dimitriadis, D., Stolcke, A., Hinthorn, W., Chen, Z., Zeng, M., & Huang, X. (2019). Meeting transcription using asynchronous distant microphones. In *Proceedings of Interspeech 2019* (pp. 2968–2972). doi: 10.21437/Interspeech.2019-3088

Yu, C., & Hansen, J. H. (2017). Active learning based constrained clustering for speaker diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(11), 2188–2198. doi: 10.1109/TASLP.2017.2747097

Yu, H., Chen, C., Du, X., Li, Y., Rashwan, A., Hou, L., … Li, J. (2020). *TensorFlow Model Garden.* https://github.com/tensorflow/models.

Zajíc, Z., Kunešová, M., & Radová, V. (2016). Investigation of segmentation in i-vector based speaker diarization of telephone speech. In *Proceedings of the 18th International Conference on Speech and Computer* (pp. 411–418). doi: 10.1007/978-3-319-43958-7_49

Zajıc, Z., Kunešová, M., Zelinka, J., & Hrúz, M. (2018). ZCU-NTIS speaker diarization system for the DIHARD 2018 challenge. In *Proceedings of Interspeech 2018* (pp. 2788–2792). doi: 10.21437/Interspeech.2018-1252

Zajíc, Z., Soutner, D., Hrúz, M., Müller, L., & Radová, V. (2018). Recurrent neural network based speaker change detection from text transcription applied in telephone speaker diarization system. In *Proceedings of the 21st International Conference on Text, Speech and Dialogue* (pp. 342–350). doi: 10.1007/978-3-030-00794-2_37

Zancanaro, M., Lepri, B., & Pianesi, F. (2006). Automatic detection of group functional roles in face to face interactions. In *Proceedings of the 8th International Conference on Multimodal Interfaces* (pp. 28–34). doi: 10.1145/1180995.1181003

Zhang, A., Wang, Q., Zhu, Z., Paisley, J., & Wang, C. (2019). Fully supervised speaker diarization. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6301–6305). doi: 10.1109/ICASSP.2019.8683892

Zuluaga-Gomez, J., Sarfjoo, S. S., Prasad, A., Nigmatulina, I., Motlicek, P., Ohneiser, O., & Helmke, H. (2021). BERTraffic: A robust BERT-based approach for speaker change detection and role identification of air-traffic communications. *arXiv preprint arXiv:2110.05781*.

# Appendices

# Appendix A

# UCC dataset: Inter-Rater Reliability

In Chapter 5 we presented and used the MISC-annotated UCC data (which we also used in Chapter 4 but without the MISC labels). Here, we present an inter-rater reliability (IRR) analysis of the utterance-level codes assigned by human raters based on a small subset of the available sessions.

Each of the 188 sessions that were selected for professional transcription and coding was coded by at least one of three trained raters. Among those, 14 sessions (from the first trial described in Section 5.4.2) were coded by two or three coders. We estimated Krippendorff's alpha (Krippendorff, 2018) for each code, a statistic which is generalizable to different types of variables and flexible with missing observations (Hallgren, 2012). Since sessions were parsed into utterances from human raters, the unit of coding is not fixed, so we estimated Krippendorff's alpha at the session level by using the per-session occurrences (treated as ratio variables) of each label. The results for all the codes are given in Table A.1.

Table A.1: Krippendorff's alpha ($\alpha$) to estimate inter-rater reliability for the utterance-level codes in the UCC data.

| **code** | **IRR** ($\alpha$) | **code** | **IRR** ($\alpha$) | **code** | **IRR** ($\alpha$) | **code** | **IRR** ($\alpha$) |
|---|---|---|---|---|---|---|---|
| ADP | 0.542* | EC | 0.558 | QUC | 0.897 | RF | 0.093* |
| ADW | 0.422 | FA | 0.868 | RCP | –* | SU | 0.345 |
| AF | 0.123 | FI | 0.784 | RCW | 0.000* | ST | 0.434 |
| CO | 0.497* | GI | 0.861 | RES | 0.268 | WA | -0.054* |
| DI | 0.590 | QUO | 0.945 | REC | 0.478 | | |

MISC abbreviations are defined in Table 5.1.
*the particular code was not used (count=0) by at least 2 coders for at least half of the analyzed sessions.
RCP was never used by any coder.

As described in Section 5.4.2, those MISC labels are grouped into 9 target classes (Table 5.2). Table A.2 gives the results of the IRR analysis for this labeling scheme.

Table A.2: Krippendorff's alpha ($\alpha$) to estimate inter-rater reliability for the utterance-level target labels in the UCC data.

| **group** | **IRR** ($\alpha$) | **group** | **IRR** ($\alpha$) | **group** | **IRR** ($\alpha$) |
|---|---|---|---|---|---|
| FA | 0.868 | QUO | 0.946 | MIN | 0.606 |
| GI | 0.898 | REC | 0.479 | MIA | 0.363 |
| QUC | 0.897 | RES | 0.268 | ST | 0.434 |

The mapping bettween MISC-defined behavior codes and grouped target labels is given in Table 5.2.

# Appendix B

# Psychotherapy Transcription and Coding Pipeline

The following sections provide details related to the several modules of the transcription and coding pipeline introduced in Chapter 5, including training data, hyperparameter values, and evaluation results.

## B.1 Datasets

The design of the system is based on datasets drawn from a variety of sources. We have combined large speech and language corpora both from the psychotherapy domain and from other fields (meetings, telephone conversations, etc.). That way, we wanted to ensure high in-domain accuracy when analyzing psychotherapy data, but also robustness across various recording conditions.

### Out-of-domain corpora

The acoustic modeling was mainly based on a large collection of speech corpora, widely used by the research community for a variety of speech processing tasks. Specifically, we used the Fisher English (Cieri et al., 2004), ICSI Meeting Speech (Janin et al., 2003), WSJ (Paul & Baker, 1992), and 1997 HUB4 (Graff, Wu, MacIntyre, & Liberman, 1997) corpora, available through the linguistic data consortium (LDC), as well as Librispeech (Panayotov, Chen, Povey, & Khudanpur, 2015), TED-LIUM (Rousseau, Deléglise, & Esteve, 2014), and AMI (Carletta et al., 2005). This combined speech dataset consists of more than 2,000 hours of audio and contains recordings from a variety of scenarios, including business meetings, broadcast news, telephone conversations, and audiobooks/articles.

The aforementioned datasets are accompanied by manually-derived transcriptions which can be used for language modeling tasks. In our case, since we need to capture linguistic patterns specific to the psychotherapy domain, the main reason we need some out-of-domain text corpus is to build a background model that guarantees a large enough vocabulary and minimizes the unseen words during evaluation. To that end, we use the transcriptions of the Fisher English corpus, featuring a vocabulary of 58.6K words and totaling more than 21M tokens.

### Psychotherapy-related corpora

In order to train and adapt our machine learning models on in-domain data, in addition to the UCC data collection described in Section 5.4.2, we also used available psychotherapy-focused corpora. In particular, we used a collection of MI sessions (for which audio, transcription and manual coding information were available) from six independent clinical trials (ARC, ESPSB, ESP21, iCHAMP, HMCBI, CTT; Atkins et al., 2014; Baer et al., 2009), as introduced in Chapter 1 (with the *MI-train* subset defined in Table 1.1). The transcripts of those MI sessions were enhanced by data provided by the counseling and psychotherapy transcripts series[1] (CPTS). This included transcripts from a variety of therapy interventions totaling about 300K utterances and 6.5M words. For this corpus, no audio or behavioral coding are available, and the data were hence used only for language-based modeling tasks.

## B.2 System Details

### Audio feature extraction

For all the modules of the speech pipeline (VAD, diarization, ASR), the acoustic representation is based on the widely used mel-frequency cepstrum coefficients (MFCCs), extracted every 10 msec

[1] https://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series

using 25 msec-long windows with the Kaldi toolkit[2]. For the UCC data, the channels from the two recording microphones are combined through acoustic beamforming (Anguera, Wooters, & Hernando, 2007), using the open-source BeamformIt tool[3].

### Voice activity detection

The first step of the transcription pipeline is to extract the voiced segments of the input audio session. The rest of the session is considered to be silence, music, background noise, etc., and is not taken into account for the subsequent steps. To that end, we use a feed-forward neural network with two layers of 512 neurons each and sigmoid activation functions, before a final inference layer giving a frame-level probability. The input is a 13-dimensional MFCC vector characterizing a frame, spliced with a context of 30 neighboring frames (15+15).

This is a pre-trained model, initially developed as part of the robust automatic transcription of speech (RATS) program (Thomas, Saon, Van Segbroeck, & Narayanan, 2015). The model was trained to reliably detect speech activity in highly noisy acoustic scenarios, with most of the noise types included during training being military noises like machine gun, helicopter, etc. Hence, in order to make the model better suited to our task, the original model was adapted using the UCC$_{dev}$ data. Optimization of the various parameters was done with respect to the unweighted average recall (UAR). The frame-level outputs are smoothed via a median filter of 31 taps and converted to longer speech segments which are passed to the diarization sub-system. During this process, if silence between any two contiguous voiced segments is less than 0.5 sec, the corresponding segments are merged together.

### Automatic speech recognition

The linguistic content captured within speech segments is the information supplied to the subsequent text-based algorithms used for speaker role recognition, lignuistically-aided diarization, and behavioral coding. Automatic speech recognition (ASR) depends on two components; the acoustic model (AM), which calculates the likelihood of acoustic observations given a sequence of words,

[2] https://github.com/kaldi-asr/kaldi

[3] https://github.com/xanguera/BeamformIt

and the language model (LM), which calculates the likelihood of a word sequence by describing the distribution of typical language usage. We note that, for the system depicted in Figure 5.2, the same ASR module is used for both the first and the second passes.

In order to train the AM, we build a time-delay neural network (TDNN) with subsampling (Peddinti, Povey, & Khudanpur, 2015). First, word alignments are derived based on the GMM/HMM paradigm. The input feature vectors to the TDNN architecture are 40-dimensional MFCCs which are augmented by 100-dimensional i-vectors, extracted online through a sliding window. The network is trained on a large combined speech dataset composed of the Fisher English, ICSI Meeting Speech, WSJ, 1997 HUB4, Librispeech, TED-LIUM, AMI, and MI corpora. We use the officially recommended training subsets for Librispeech and TED-LIUM and the recommended training and development sets for AMI. We randomly choose 95% of the available Fisher utterances and 80% of the available ICSI, WSJ, and HUB4 utterances. We also use the 242 MI-train sessions (Table 1.1). We have kept the rest of the combined dataset for internal validation and evaluation of the ASR system. Among the aforementioned corpora, TED-LIUM and the clean portion of Librispeech are augmented with speed perturbation, noise, and reverberation (Ko, Peddinti, Povey, & Khudanpur, 2015). The final combined, augmented corpus contains more than 4,000 hours of phonetically rich speech data, recorded under different conditions and reflecting a variety of acoustic environments. The ASR AM is built and trained using the Kaldi speech recognition toolkit (Povey et al., 2011).

In order to build the LM, we independently train two 3-gram models using the SRILM toolkit (Stolcke, 2002). One is trained with in-domain psychotherapy data from the CPTS transcribed sessions. This is interpolated with a large background model, in order to minimize the unseen words during inference. The background LM is trained with the Fisher English corpus, which features conversational telephone data. The two 3-gram LMs are interpolated with mixing weights equal to 0.8 for the in-domain model and 0.2 for the background model.

The evaluation of an ASR system is usually performed through the word error rate (WER) metric which is the normalized Levenshtein distance between the ASR output and the manually-derived transcript and includes errors because of word substitutions, word deletions, and word insertions. Those errors are typically estimated for each utterance given to the ASR module and then summed up for all the evaluation data, in order to get an overall WER. However, when we analyze an entire therapy session which has been processed by the VAD and diarization subsystems, the "utterances" are different than the ones identified by the human transcriber. In that case, we evaluate at the session level, concatenating all the session utterances. The results are reported in Table B.1 using either the oracle segmentation (from the manual transcriptions) or the one generated by the automated systems. For the latter case, we explore VAD-only segmentation (after which we run the first pass of ASR needed for the linguistically-aided diarization as shown in Figure 5.2), as well as the two diarization-based segmentation approaches we explored in Chapter 5: the audio-only, clustering-based one and the linguistically-aided, classification-based one.

Table B.1: ASR results (%) for the UCC data.

| **diarization method** | **substitutions** | **deletions** | **insertions** | **WER** |
|---|---|---|---|---|
| oracle | 15.1 | 14.1 | 2.6 | 31.7 |
| VAD | 16.4 | 12.5 | 3.2 | 32.0 |
| clustering-based | 16.8 | 12.9 | 3.2 | 32.9 |
| classification-based | 17.2 | 12.8 | 3.4 | 33.4 |

WER is estimated as the sum of the substitution, insertion, and deletion rates.
Results are reported when using either the segments derived by the manual transcriptions (*oracle*) or the machine-generated ones, based on only VAD, or based on the two different diarization methods we have explored (Section 5.5).

As we can see, ASR performance is not severely degraded by error propagation due to the preprocessing steps of VAD/diarization (up to about 5% relative WER increase). However, we do note that the degradation observed between the VAD-based and the diarization-based segmentations suggests that ASR can completely precede diarization and an alternative overall architecture than the ones presented in Chapter 5 might provide improved overall performance. This is a direction we did not explore within this study.

Interestingly, comparing the oracle and the machine-generated segmentations, we can see that even though insertion rate is increased, deletion rate is decreased when machine-generated segments are provided. This is explained by the long segments constructed after concatenating consecutive segments given by the VAD and diarization algorithms. On the one hand, labeling silence or noise as "speech" associated with some speaker occasionally leads ASR to predict words where in reality there is no speech activity—thus increasing insertion rate. On the other hand, this minimizes the probability of missing some words because of missed speech. Such deleted words may occur when providing the oracle segments because of inaccuracies during the construction of the "ground truth" through forced alignment.

We note that, even though the estimated error is high, WERs in the range reported and even higher are typical in spontaneous medical conversations (Kodish-Wachs et al., 2018). Error analysis revealed that those numbers are inflated because of fillers (e.g. *uh-huh*, *hmm*) and other idiosyncrasies of conversational speech. It should be additionally highlighted that WER is a generic metric that gives equal importance to all the words, while for our end goal of behavior coding there are specific linguistic constructs which potentially carry more valuable information than others.

### Utterance segmentation

The ASR output is at the segment level, with segments defined by the VAD and diarization algorithms. However, silence and speaker changes are not always the right cues to help us distinguish between utterances, which are the basic units of behavioral coding. The presence of multiple utterances per speaker turn is a challenge we often face when dealing with conversational interactions. Especially in the psychotherapy domain, it has been shown that utterance-level segmentation can significantly improve the performance of automatic behavior code prediction (Z. Chen et al., 2021).

Thus, we have included an utterance segmentation module at the end of the automatic transcription, before employing the subsequent NLP algorithms. In particular, we merge together all the adjacent segments belonging to the same speaker in order to form speaker-homogeneous talk-turns, and we then segment each turn using the DeepSegment tool[4]. DeepSegment has been designed to perform text-based sentence boundary detection having specifically ASR outputs in mind, where punctuation is not readily available. In this framework, sentence segmentation is viewed as a sequence labeling problem, where each word is tagged as being either at the beginning of a sentence (utterance), or anywhere else. DeepSegment addresses the problem employing a bidirectional long-short term memory (BiLSTM) network with a conditional random field (CRF) inference layer (Ma & Hovy, 2016), similarly to the tagger architecture we used in Chapter 3 (Figure 3.3).

[4] https://github.com/notAI-tech/deepsegment

### Utterance-level code prediction

Once the entire session is transcribed at the utterance level, we employ text-based algorithms for the task of behavior code prediction. We focus on counselor behaviors, so we only take into account the utterances assigned to the therapist according to the speaker role recognition module. Each one of those needs to be assigned a single code from the 9 target labels summarized in Table 5.2. This is achieved through a BiLSTM network with attention mechanism (Singla et al., 2018) which only processes textual features. The input to the system is a sequence of word-level embeddings for each utterance. The recurrent layer exploits the sequential nature of language and produces hidden vectors which take into account the entire context of each word within the utterance. The attention layer can then learn to focus on salient words carrying valuable information for the task of code prediction, thus enhancing robustness and interpretability.

The network is first trained on the MI data using the Adam optimizer with learning rate equal to 0.001 and exponential decay equal to 0.9. The batch size is set equal to 256 utterances and we use class weights inversely proportional to the class frequencies. The system is trained on that dataset for 30 epochs with an early stopping strategy, keeping the model with the lowest validation loss. The system is further fine-tuned to the University Counseling Center conditions by continuing training on the UCC$_{train}$ data.

When we use the manually transcribed data to perform utterance-level MISC code prediction, the overall averaged $F_1$ score is 0.517 for the UCC evaluation sets. The $F_1$ scores for each individual code are reported in Table B.2. As expected, the results are better for the highly frequent codes (Table 5.2), such as the one expressing facilitation (FA), since the machine learning models have more training examples to learn from. On the other hand, the models do not perform as well for less frequent codes, such as MI-NonAdherent behaviors (MIN) and simple reflections (RES). Comparing Table B.2 and Table A.1, we can also see that for several of the codes that our system performs relatively poorly (e.g., simple reflections (RES), MI-Adherent (MIA), structure (ST)), the inter-annotator agreement is also considerably low. A notable example which does not follow this pattern is the non-adherent behavior (MIN) where the performance of our system is relatively poor ($F_1 = 0.261$), while there is a substantial inter-annotator agreement ($\alpha = 0.606$). This is partly because of the underrepresentation of the particular code (or cluster of codes) in the training and development sets. It may be also the case that pure linguistic information found in textual patterns may not be enough for the operationalization of the particular code. This example suggests that a hybrid approach where machine learning methods are combined with knowledge-based rules from the coding manuals may be an interesting direction for future research. Finally, by examining the confusion matrices (not reported here), we realized that the system often gets confused between the codes representing questions (QUC vs. QUO) and reflections (RES vs. REC), since those pairs of codes get usually assigned to utterances with several structural and semantic similarities.

Table B.2: $F_1$ scores for the predicted utterance-level codes using the manually transcribed UCC data.

| **FA** | **GI** | **QUC** | **QUO** | **REC** | **RES** | **MIN** | **MIA** | **ST** |
|---|---|---|---|---|---|---|---|---|
| 0.951 | 0.473 | 0.604 | 0.792 | 0.476 | 0.198 | 0.261 | 0.423 | 0.472 |

It is interesting to compare the remarkably good performance of the system with respect to FA with the relatively low correlation reported in Table 5.4, where the MISC predictor is given the automatically generated utterances. The reason behind this is that FA is assigned to a lot of one-word utterances and talk turns. Our speech pipeline, however, often fails to capture turns of such short duration (or concatenates them with neighboring utterances to construct longer segments) which results in a smaller than expected frequency for the specific code.