Linguistically Aided Speaker Diarization Using Speaker Role Information

Nikolaos Flemotomos, Panayiotis Georgiou, Shrikanth Narayanan

Department of Electrical and Computer Engineering
University of Southern California, Los Angeles, CA, USA
flemotom@usc.edu, georgiou@ee.usc.edu, shri@ee.usc.edu

Abstract

Speaker diarization 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, which is traditionally based solely on audio. Under noisy conditions, however, such an approach poses the risk of generating unreliable speaker clusters. In this work we aim to utilize linguistic information as a supplemental modality to identify the various speakers in a more robust way. We are focused on conversational scenarios where the speakers assume distinct roles and are expected to follow different linguistic patterns. This distinct linguistic variability can be exploited to help us construct the speaker identities. That way, we are able to boost the diarization performance by converting the clustering task to a classification one. The proposed method is applied in real-world dyadic psychotherapy interactions between a provider and a patient and demonstrated to show improved results.

1. Introduction

Given a speech signal with multiple speakers, diarization answers the question "who spoke when" [1]. 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.

First, the input signal is segmented either uniformly (e.g. [2]) or according to a speaker change detection algorithm (e.g. [3]). 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 prac-

tical observations such as the fact that it is highly improbable for a speaker change point to be located within a word [4], it is widely accepted that each individual has their very own way of using language [5]. 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 [6], and in combination with the speech audio [7–9].

Using language for diarization can be especially promising in structured scenarios where the speakers assume dissimilar roles with distinguishable linguistic patterns. For example, a teacher is likely to speak in a more didactic style while a student be more inquisitive, a doctor is likely to inquire on symptoms and prescribe while a patient describe symptoms, etc. In our own past work [10, 11] we exploited the differences of language between a therapist and a patient to facilitate the automatic processing of medical conversations. The strong interconnection between speech and language in such scenarios has even led to end-to-end role-annotated Automatic Speech Recognition (ASR) architectures [12, 13] where diarization becomes a byproduct of a rich transcription system.

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 [14, 15]. This is because single-speaker speech segments allow for speaker normalization techniques, including speaker adaptive training through Constrained Maximum Likelihood Linear Regression (CMLLR) [16] and i-vector based neural network adaptation [17]. 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 [9, 18]. 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.

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 present), 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 that of a therapist and a patient.

The rest of the paper is structured as follows: Section 2 provides a background on audio-based speaker diarization systems and on speaker role recognition in structured conversations. In Section 3 we present our approach on how to use the linguistic information in order to estimate the speaker identities instead of relying on clustering. In Sections 4 and 5 we give an overview of the datasets used and we analyze our experimental results. Finally, in Section 6 we discuss limitations of our work, as well as potential directions for future research.

2. Background

2.1. 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., Gaussian Mixture Models - GMMs), and measure the distance between them using a metric such as the one based on the Bayesian Information Criterion (BIC) [19].

Speaker modelling by GMMs was later replaced by i-vectors [20], 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 [21,22] was proved to yield improved results [23]. Given two embeddings v, r, PLDA provides a framework to estimate their similarity $s(v, r)$ as the log-likelihood ratio

$$s(v, r) = \log \frac{p(v, r | \text{same speaker})}{p(v | \text{dif. speakers}) p(r | \text{dif. speakers})} \quad (1)$$

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 [24], or speaker discrimination employing contrastive [25] and triplet [26] loss functions. Typical examples of embeddings that have shown state-of-the-art performance for speaker diarization are the Time-Delay Neural Net (TDNN) based x-vectors [2] and the Long-Short Term Memory (LSTM) based d-vectors [27].

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 [1]. Even though speaker change detection is by itself an active research field [28,29], 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 [3,30], so the latter method is widely used. As far as the clustering is concerned, common approaches include Hierarchical Agglomerative Clustering (HAC) [2] and spectral clustering [27,31], while methods based on affinity propagation [32] and generative adversarial networks [33] 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 [34,35].

2.2. Speaker Role Recognition

In many real-world speech applications speakers play specific roles, in the sense that they perform well-defined tasks within the conversation domain. For example, we can think of the different speakers who may appear during a broadcast news program, a clinical visit, or a lecture. Typically, the domain is given, so the set of available roles is known in advance and Speaker Role Recognition (SRR) is viewed as a supervised classification task [36,37]. Even though various modalities have been explored in the literature [38–40], the linguistic variability between the speaker roles usually provides the most important cues for the task in hand [11].

The lexical information was initially captured by n-gram features which were passed to boosting algorithms or maximum entropy classifiers [36]. Additional language-based features that have been proposed include the types of questions asked by different speakers [41], as well as the errors identified in ASR transcripts [42]. Deep learning approaches have been explored in [39] where SRR relies on word embeddings and Convolutional Neural Networks (CNNs). In [15] SRR is viewed as a problem of finding the role-specific n-gram Language Model (LM) which minimizes the perplexity of a text segment, while in [11] the role assignment is based on the LM scores of role-adapted ASR systems.

3. Linguistically-Aided Speaker Diarization

Our proposed approach for speaker diarization in conversational interactions where speakers assume specific roles is illustrated in Figure 1. We describe the various modules in detail in Sections 3.1-3.4.
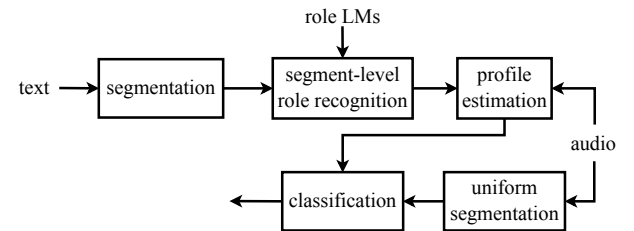

Figure 1: Linguistically-aided speaker diarization using role information.

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 [6], 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 [8]. 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 [43], as shown in Figure 2. 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 [44], pre-trained on 2B words of Wikipedia. Those extend the semantic vector space representation of the word2vec model [45] 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.
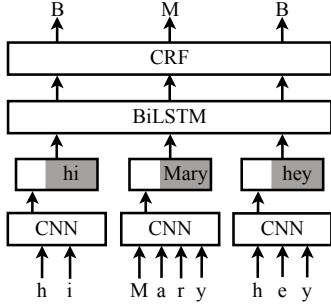

Figure 2: Neural network for sentence-level text segmentation: A character representation is constructed for each word through a CNN and is concatenated with a word embedding (here shown in grey). This is the input to a BiLSTM-CRF architecture which predicts a sequence of labels. Here 'B' denotes a word at the beginning of a sentence and 'M' in the middle (any word which is not the first one of a sentence).

3.2. 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 [10]. 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 [11]. 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 $\{\mathcal{R}_i^+\}_{i=1}^N$. To obtain an interpolated model, the weighted average of the probabilities from the input models is assigned to each n-gram, and the produced model is then re-normalized [46]. Denoting LM interpolation by the symbol $\oplus$, the final models are

$$\mathcal{G}^+ = w_g \mathcal{G} \oplus (1 - w_g)\tilde{\mathcal{R}} \qquad (2)$$
$$\mathcal{R}_i^+ = w_{g_i}\mathcal{G} \oplus w_{r_i}\mathcal{R}_i \oplus (1 - w_{g_i} - w_{r_i})\tilde{\mathcal{R}}_i \qquad (3)$$

where

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

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

3.3. Profile Estimation

After the application of 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[1], 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 3.2, 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 the role recognition at the segment level does not always provide robust results [11], 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. Formally, if the 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 $c_x = \min_{j\neq i} |pp(x|\mathcal{R}_j^+) - pp(x|\mathcal{R}_i^+)|$. Then, the corresponding profile is

$$r_i = \frac{\sum_{x\in R_i} \tilde{c}_x u_x}{\sum_{x\in R_i} \tilde{c}_x} \triangleq \frac{\sum_{x\in R_i} \mathbb{I}\{c_x > \theta\}u_x}{\sum_{x\in R_i} \mathbb{I}\{c_x > \theta\}} \qquad (4)$$

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

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 1. 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 $\hat{R}_x = \arg\max_{1\leq i\leq N}\{s(u_x, r_i)\}$, where $s(\cdot, \cdot)$ is the PLDA similarity score estimated in equation (1).

---

[1]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.

4. Datasets

4.1. Evaluation Data

We evaluate our proposed method on datasets from the clinical psychology domain. In particular, we apply the system to motivational interviewing sessions – a specific type of psychotherapy – between a therapist and a patient, collected from five independent clinical trials (ARC, ESPSB, ESP21, iCHAMP, HMCBI) [47]. We collectively refer to those sessions as the PSYCH corpus. This is split into training, development, and test sets, as shown in Table 1, in such a way that there is no speaker overlap between them. Even though the necessary metadata is available for the rest of the corpus, the patient IDs are not available for the HMCBI sessions. Thus, the partitioning is done under the assumption that it is highly improbable for the same patient to visit different therapists in the same study. All the results reported are on PSYCH-test.

	PSYCH-train	PSYCH-dev	PSYCH-test
#sessions	74	44	25
Therapist	26.43 h	15.23 h	7.34 h
Patient	23.29 h	12.17 h	7.54 h

Table 1: Size of the PSYCH dataset. The total durations of the speech segments corresponding to a therapist and to a patient are calculated based on the manual turn boundaries.

4.2. Segmenter and Role LM Training Data

The segmenter presented in Section 3.1 is trained on a subset of the Fisher English corpus [48] 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[2] (CPTS). The combined dataset is randomly split (80-20 split at the session level) into training and validation sets. CPTS and the entire Fisher English corpus, together with the training part of the PSYCH corpus, are also used to train the required role-specific LMs. The size of the corresponding vocabularies is given in Table 2.

	PSYCH-train	Fisher	CPTS
|voc|	8.17K	58.6K	35.6K
#tokens	530K	21.0M	6.52M

Table 2: Size of the vocabulary and total number of tokens in the corpora used for LM training.

5. Experiments and Results

5.1. Baseline Systems

5.1.1. Audio-based Diarization with Speaker Clustering

As an audio-only baseline, we use a state-of-the-art speaker diarization system following the x-vector/PLDA paradigm [2]. As

shown in Figure 3, 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 (1)).
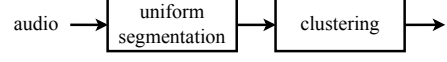

Figure 3: Baseline audio-based speaker diarization.

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.

5.1.2. Language-based Diarization

As a language-only baseline, we use the system of Figure 4, which essentially consists of the first steps of the framework in Figure 1. After estimating the segment-level role labels as described in Sections 3.1 and 3.2, we can simply use those as our diarization output labels to compare the performance of a system that only depends on the 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.
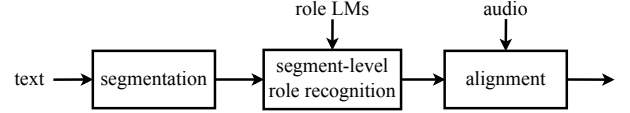

Figure 4: Baseline language-based speaker diarization.

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 2. 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 [49]. Following the general recommendations in [50] 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 [46]. LMs are trained using a combination of in-domain data from the PSYCH-train and CPTS corpora and out-of-domain data from the Fisher English corpus, employing the LM interpolation procedure described in Section 3.2, with mixing weights optimized on PSYCH-dev. SRILM is also used to estimate the needed perplexities.

The audio-based diarization framework is built using the Kaldi toolkit [51]. Both for the audio-only baseline and for our system, we use the VoxCeleb pre-trained x-vector extractor[3] and the PLDA model which comes with it, after we adapt it on the development set of the PSYCH corpus. 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 an 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.

To get the ASR outputs, we use Kaldi's pre-trained ASpIRE acoustic model[4], coupled with the 3-gram LM given in equation (2). 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 [52].

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 idea, eliminating potential propagation errors because of ASR. Table 3 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.

text segmentation	audio only	language only	linguistically aided
oracle	11.05	12.99	7.28
tagger		20.09	7.71

Table 3: DER (%) following our linguistically-aided approach and the two baselines. The text segmentation (when needed) is either performed by our sequence *tagger* or based on the manually annotated speaker changes (*oracle*).

When we apply our linguistically-aided system using our sequence tagger to segment at the sentence level we get a 30.23% DER relative improvement compared to the audio-only approach. In the first row of Table 3 we additionally report results when using the oracle speaker segmentation provided by

[3] https://kaldi-asr.org/models/m7
[4] http://kaldi-asr.org/models/m1

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. 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 we only keep the segments we are most confident about for the profile estimation, applying the confidence metric introduced in Section 3.3. Instead of directly optimizing for the parameter θ appearing in equation (4), 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 Table 4. 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.

text segmentation	DER (%)	DER† (%)
oracle	7.28	6.99
tagger	7.71	7.30

Table 4: Diarization results following our linguistically-aided approach. The text segmentation 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 ($a = 70$ for the tagger segmentation and $a = 55$ for the oracle segmentation).

In Figure 5 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.

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 Table 5. Here, we report results only when using the sequence tagger (and not with oracle segmentation). The reason is simply because we now assume we have no access to the reference transcripts, so we cannot know the oracle speaker change points.
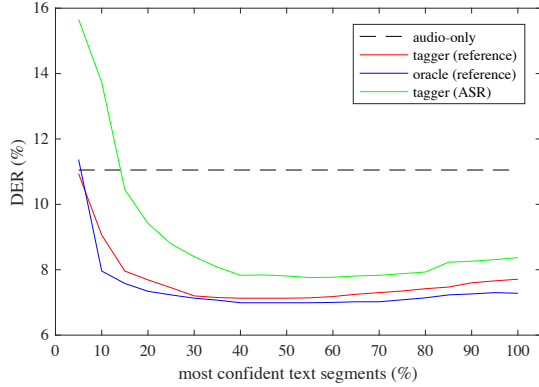
Figure 5: DER (%) as a function of the percentage of the text segments we take into account per session for the profile estimation, based on our confidence metric. The text segmentation is either performed by our sequence *tagger* or based on the manually annotated speaker changes (*oracle*). Results are presented both with *reference* and with *ASR* transcripts.

method	audio only	language only	linguistically aided
DER (%)	11.05	27.07	8.37

Table 5: Diarization results following our linguistically-aided approach and the two baselines. The text segmentation (when needed) is performed by our sequence tagger.

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 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 somewhat inaccurate ASR system. As expected, this is not the case for our language-only baseline. Since in that case the final output only depends on the linguistic information, the performance gap between using manually- and ASR-derived transcripts (2nd column in Tables 3 and 5) 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 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 5.

6. Conclusions and Future Work

We proposed a system for speaker diarization suitable to use in conversations where the 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 can segment 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 work 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 the recordings are long enough so that they may incorporate various acoustic conditions or different speaking styles corresponding to the same speaker. Finally, another direction of future work could be towards an in-depth investigation of the confidence metric used for the text-based role assignments and, accordingly, the number of those segments that we should take into consideration when estimating the speaker profiles. Even though in this work we employed a metric purely based on perplexity differences, it is possible that other attributes (e.g., segment length) provide additional information.

7. Acknowledgements

8. References

[1] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker Diarization: A Review of Recent Research," *Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and Sanjeev Khudanpur, "Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge," in *Proc. Interspeech*, 2018, pp. 2808–2812.

[3] Zbynek Zajíc, Marie Kunešová, Jan Zelinka, and Marek Hrúz, "ZCU-NTIS Speaker Diarization System for the Dihard 2018 Challenge," in *Proc. Interspeech*, 2018, pp. 2788–2792.

[4] Jan Silovsky, Jindrich Zdansky, Jan Nouza, Petr Cerva, and Jan Prazak, "Incorporation of the ASR Output in Speaker Segmentation and Clustering within the Task of Speaker Diarization of Broadcast Streams," in *Proc. International Workshop on Multimedia Signal Processing*, 2012, pp. 118–123.

[5] Barbara Johnstone, *The linguistic individual: Self-expression in language and linguistics*, Oxford University Press, 1996.

[6] Zhao Meng, Lili Mou, and Zhi Jin, "Hierarchical RNN with Static Sentence-Level Attention for Text-Based Speaker Change Detection," in *Proc. Conference on Information and Knowledge Management*, 2017, pp. 2203–2206.

[7] Miquel Àngel India Massana, José Adrián Rodríguez Fonollosa, and Francisco Javier Hernando Pericás, "LSTM Neural Network-Based Speaker Segmentation Using Acoustic and Language Modelling," in *Proc. Interspeech*, 2017, pp. 2834–2838.

[8] Zbyněk Zajíc, Daniel Soutner, Marek Hrúz, Luděk Müller, and Vlasta Radová, "Recurrent Neural Network Based Speaker Change Detection from Text Transcription Applied in Telephone Speaker Diarization System," in *Proc. International Workshop on Temporal, Spatial, and Spatio-Temporal Data Mining*, 2018, pp. 342–350.

[9] Tae Jin Park, Kyu J. Han, Jing Huang, Xiaodong He, Bowen Zhou, Panayiotis Georgiou, and Shrikanth Narayanan, "Speaker Diarization with Lexical Information," in *Proc. Interspeech*, 2019, pp. 391–395.

[10] Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson, and Shrikanth Narayanan, "Combined Speaker Clustering and Role Recognition in Conversational Speech," in *Proc. Interspeech*, 2018, pp. 1378–1382.

[11] Nikolaos Flemotomos, Panayiotis Georgiou, and Shrikanth Narayanan, "Role Specific Lattice Rescoring for Speaker Role Recognition from Speech Recognition Outputs," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 7330–7334.

[12] Nikolaos Flemotomos, Zhuohao Chen, David C Atkins, and Shrikanth Narayanan, "Role annotated speech recognition for conversational interactions," in *Proc. Spoken Language Technology Workshop*, 2018, pp. 1036–1043.

[13] Laurent El Shafey, Hagen Soltau, and Izhak Shafran, "Joint Speech Recognition and Speaker Diarization via Sequence Transduction," *Proc. Interspeech*, pp. 396–400, 2019.

[14] Jing Huang, Etienne Marcheret, Karthik Visweswariah, Vit Libal, and Gerasimos Potamianos, "The IBM Rich Transcription 2007 Speech-to-Text Systems for Lecture Meetings," in *Multimodal Technologies for Perception of Humans*, pp. 429–441. Springer, 2007.

[15] Bo Xiao, Chewei Huang, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan, "A Technology Prototype System for Rating Therapist Empathy from Audio Recordings in Addiction Counseling," *PeerJ Computer Science*, vol. 2, pp. e59, 2016.

[16] Mark JF Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[17] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker Adaptation of Neural Network Acoustic Models Using i-vectors," in *Proc. Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 55–59.

[18] Takuya Yoshioka, Dimitrios Dimitriadis, Andreas Stolcke, William Hinthorn, Zhuo Chen, Michael Zeng, and Xuedong Huang, "Meeting Transcription Using Asynchronous Distant Microphones," in *Proc. Interspeech*, 2019, pp. 2968–2972.

[19] Scott Chen and Ponani Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, vol. 8, pp. 127–132.

[20] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas Reynolds, and James Glass, "Exploiting Intra-conversation Variability for Speaker Diarization," in *Proc. Interspeech*, 2011.

[21] Sergey Ioffe, "Probabilistic Linear Discriminant Analysis," in *Proc. European Conference on Computer Vision*, 2006, pp. 531–542.

[22] Simon JD Prince and James H Elder, "Probabilistic Linear Discriminant Analysis for Inferences about Identity," in *Proc. International Conference on Computer Vision*, 2007, pp. 1–8.

[23] Gregory Sell and Daniel Garcia-Romero, "Speaker Diarization with PLDA i-vector Scoring and Unsupervised Calibration," in *Proc. Spoken Language Technology Workshop*, 2014, pp. 413–417.

[24] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN Embeddings for Speaker Recognition," in *International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[25] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, "Speaker Diarization Using Deep Neural Network Embeddings," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 4930–4934.

[26] Hervé Bredin, "Tristounet: Triplet Loss for Speaker Turn Embedding," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5430–5434.

[27] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno, "Speaker Diarization with LSTM," in *International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5239–5243.

[28] Marek Hrúz and Zbyněk Zajíc, "Convolutional Neural Network for Speaker Change Detection in Telephone Speaker Diarization System," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 4945–4949.

[29] Arindam Jati and Panayiotis G Georgiou, "Speaker2Vec: Unsupervised Learning and Adaptation of a Speaker Manifold Using Deep Neural Networks with an Evaluation on Speaker Segmentation," in *Proc. Interspeech*, 2017, pp. 3567–3571.

[30] Zbyněk Zajíc, Marie Kunešová, and Vlasta Radová, "Investigation of Segmentation in i-vector Based Speaker Diarization of Telephone Speech," in *Proc. International Conference on Speech and Computer*, 2016, pp. 411–418.

[31] Tae Jin Park, Manoj Kumar, Nikolaos Flemotomos, Monisankha Pal, Raghuveer Peri, Rimita Lahiri, Panayiotis Georgiou, and Shrikanth Narayanan, "The Second DI-HARD Challenge: System Description for USC-SAIL Team," in *Proc. Interspeech*, 2019, pp. 998–1002.

[32] Ruiqing Yin, Hervé Bredin, and Claude Barras, "Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization," *Proc. Interspeech 2018*, pp. 1393–1397, 2018.

[33] Monisankha Pal, Manoj Kumar, Raghuveer Peri, Tae Jin Park, So Hyun Kim, Catherine Lord, Somer Bishop, and Shrikanth Narayanan, "Speaker Diarization Using Latent Space Clustering in Generative Adversarial Network," *arXiv preprint arXiv:1910.11398*, 2019.

[34] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang, "Fully Supervised Speaker Diarization," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6301–6305.

[35] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "End-to-End Neural Speaker Diarization with Self-Attention," *arXiv preprint arXiv:1909.06247*, 2019.

[36] Regina Barzilay, Michael Collins, Julia Hirschberg, and Steve Whittaker, "The Rules Behind Roles: Identifying Speaker Role in Radio Broadcasts," in *Proc. National Conference on Artificial Intelligence and Conference on Innovative Applications of Artificial Intelligence*, 2000, pp. 679–684.

[37] Benjamin Bigot, Corinne Fredouille, and Delphine Charlet, "Speaker Role Recognition on TV Broadcast Documents," in *Proc. Workshop on Speech, Language and Audio in Multimedia*, 2013.

[38] Hugues Salamin and Alessandro Vinciarelli, "Automatic Role Recognition in Multiparty Conversations: An Approach Based on Turn Organization, Prosody, and Conditional Random Fields," *IEEE Transactions on Multimedia*, vol. 14, no. 2, pp. 338–345, 2011.

[39] Michael Rouvier, Sebastien Delecraz, Benoit Favre, Meriem Bendris, and Frederic Bechet, "Multimodal Embedding Fusion for Robust Speaker Role Recognition in Video Broadcast," in *Proc. Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 383–389.

[40] Vered Silber-Varod, Sarit Malayev, and Anat Lerner, "Positioning Oneself in Different Roles: Structural and Lexical Measures of Power Relations Between Speakers in Map Task Corpus," *Speech Communication*, pp. 1–12, 2020.

[41] Thierry Bazillon, Benjamin Maza, Michael Rouvier, Frederic Bechet, and Alexis Nasr, "Speaker Role Recognition Using Question Detection and Characterization," in *Proc. Interspeech*, 2011.

[42] Géraldine Damnati and Delphine Charlet, "Robust Speaker Turn Role Labeling of TV Broadcast News Shows," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5684–5687.

[43] Xuezhe Ma and Eduard Hovy, "End-to-End Sequence Labeling via Bi-directional LSTM-CNNs-CRF," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2016, vol. 1, pp. 1064–1074.

[44] Alexandros Komninos and Suresh Manandhar, "Dependency Based Embeddings for Sentence Classification Tasks," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1490–1500.

[45] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in *Proc. Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[46] Andreas Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proc. International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[47] David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth, "Scaling Up the Evaluation of Psychotherapy: Evaluating Motivational Interviewing Fidelity via Statistical Text Classification," *Implementation Science*, vol. 9, no. 1, pp. 49, 2014.

[48] Christopher Cieri, David Miller, and Kevin Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Proc. Conference on Language Resources and Evaluation*, 2004, vol. 4, pp. 69–71.

[49] Jie Yang and Yue Zhang, "NCRF++: An Open-Source Neural Sequence Labeling Toolkit," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 74–79.

[50] Nils Reimers and Iryna Gurevych, "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 338–348.

[51] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. Workshop on Automatic Speech Recognition and Understanding*, 2011.

[52] Jodi Kodish-Wachs, Emin Agassi, Patrick Kenny III, and J Marc Overhage, "A Systematic Comparison of Contemporary Automatic Speech Recognition Engines for Conversational Clinical Speech," in *Proc. American Medical Informatics Association Annual Symposium*, 2018, p. 683.