ROLE ANNOTATED SPEECH RECOGNITION FOR CONVERSATIONAL INTERACTIONS

Nikolaos Flemotomos[1], *Zhuohao Chen*[1], *David C. Atkins*[2], *Shrikanth Narayanan*[1]

[1] Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA
[2] Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

ABSTRACT

Speaker Role Recognition (SRR) assigns a specific speaker role to each speaker-homogeneous speech segment in a conversation. Typically, those segments have to be identified first through a diarization step. Additionally, since SRR is usually based on the different linguistic patterns observed between the roles to be recognized, an Automatic Speech Recognition (ASR) system is also indispensable for the task in hand to convert speech to text. In this work we introduce a Role Annotated Speech Recognition (RASR) system which, given a speech signal, outputs a sequence of words annotated with the corresponding speaker roles. Thus, the need of different component modules which are connected in a way that may lead to error propagation is eliminated. We present, analyze, and test our system for the case of two speaker roles to showcase an end-to-end approach for automatic rich transcription with application to clinical dyadic interactions.

Index Terms— speaker role, automatic speech recognition, weighted finite state transducers, rich transcription, conversational speech

1. INTRODUCTION

Automatic rich transcription of speech is a useful task for various domains featuring interactions between individuals with specific roles. Examples of such interactions include those between a therapist and a patient during a psychotherapy session [1], between an agent in a call center and a customer [2], between a husband and a wife [3], or between interlocutors during an interview [4].

In such cases, the transcription is desired to answer both the questions "what has been said?", thus calling for an Automatic Speech Recognition (ASR) module, and "which role spoke when?", thus calling for a speaker diarization and a Speaker Role Recognition (SRR) module. A diagram presenting the high-level architecture of the overall approach is shown in Fig. 1. After the appropriate feature extraction from the speech signal, the acoustic information is fed as input to the diarization and ASR modules. SRR then maps each detected speaker turn (speaker-homogeneous segment) to some role which belongs to a usually predefined set [5, 6]. In order to make the classification decision, SRR exploits the acoustic [6, 7] or linguistic [1, 8] information (or both [9, 10]) encoded in the speech signal.

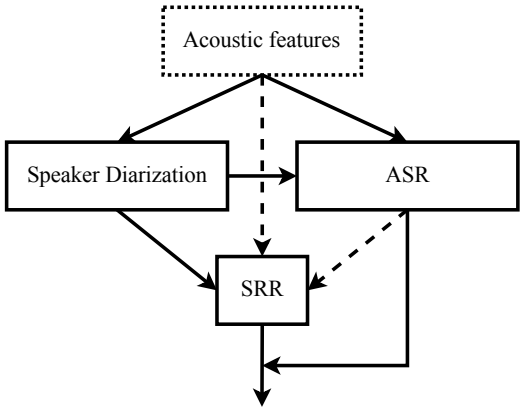

Fig. 1: Traditional approach for automatic transcription of speech documents with role interactions.

Such architectures, where multiple specialized modules are implemented independently and then combined together, may lead to error propagation. For instance, if ASR works with the assumption of speaker-homogeneous segments, then its performance is bounded by the performance of the diarization module [1]. Similarly, SRR performance is greatly affected by both diarization [6] and ASR outputs, in case SRR needs access to textual information. Moreover, those architectures do not allow for information sharing, although modalities that have traditionally been used only for some step can be proved useful for other steps, as well. For example, information extracted by ASR can improve diarization [11, 12], while speaker-specific variabilities taken into account during diarization can be combined with role-specific ones towards better SRR performance [10].

To alleviate the aforementioned problems, we propose a Role Annotated Speech Recognition (RASR) system, an extension of a traditional ASR system that outputs both textual and speaker role information. RASR has a broader goal than that of ASR, by predicting not only a sequence of words corresponding to the input signal, but also the speaker role associated with each individual word, thus doing the job both of an ASR and an SRR module. Since the role prediction is at the word level, this can also be regarded as a diarization

result, by using the alignment of the output text with the audio signal. A comparison of ideal ASR and RASR systems applied to the same input signal is given in Fig. 2.

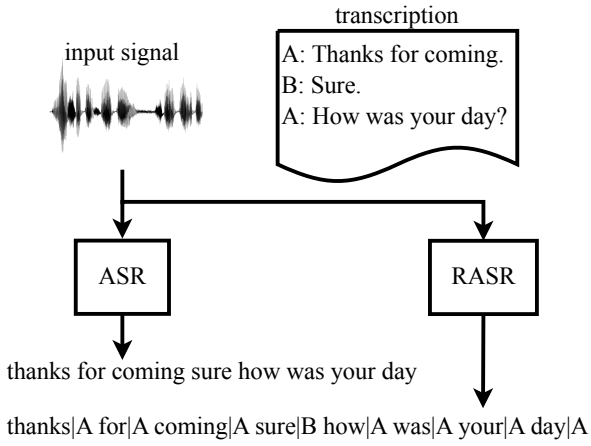

Fig. 2: Example of an ASR and an RASR system applied to the same input signal with two speaker roles, A and B.

In the following sections we present and analyze RASR in the Weighted Finite State Transducers (WFST) framework and we evaluate our method on a dataset of dyadic interactions between a psychotherapist and a patient.

2. METHOD

2.1. WFST framework for ASR

Given a sequence of acoustic features O, the job of an ASR system is to find, out of the set $\mathcal{W}$ of possible sequences of words, the most probable sequence

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmax}} P(W|O) = \underset{W \in \mathcal{W}}{\operatorname{argmax}} P(O|W)P(W)\,,$$

where $P(O|W)$ is called 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 the appearance of an SU sequence V is, given the word W, then we get

$$\begin{aligned}\hat{W} &= \underset{W \in \mathcal{W}}{\operatorname{argmax}} \sum_{V \in K(W)} P(O|V,W)P(V|W)P(W) \\ &\approx \underset{W \in \mathcal{W}}{\operatorname{argmax}} \sum_{V \in K(W)} P(O|V)P(V|W)P(W)\,,\end{aligned}$$

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

$$\begin{aligned}\hat{W} \approx \underset{W \in \mathcal{W}}{\operatorname{argmax}} \max_{V \in K(W)} \{ &\log P(O|V) + \log P(V|W) \\ &+ \log P(W)\}\,.\end{aligned}$$

In the WFST framework [13, 14], we have the transducer $\tilde{H}$ which transforms a sequence of acoustic features O into a sequence of SUs V with a weight $-\log P(O|V)$, the WFST L which transforms a sequence of SUs V into a sequence of words W with a weight $-\log P(V|W)$ and the WFSAcceptor (WFSA) G which accepts a word sequence W with a weight $-\log P(W)$. $\tilde{H}$ is actually split into a WFST H which transforms a sequence of Hidden Markov Model (HMM) states into a sequence of SUs and a model S which maps the acoustic observations to HMM states. 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, which transforms a sequence of triphones into a sequence of phonemes, 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 $N = H \circ C \circ L \circ G$ and ASR is now a shortest path problem on N. S is trained following either the Gaussian Mixture Models (GMM) or the Deep Neural Nets (DNN) paradigm and is directly used during decoding.

2.2. Modifying the component elements for RASR

Working in the same framework described for ASR, we are extending the component elements, that is the AM, the LM, and the pronunciation lexicon, to enable the prediction of annotated words, with the annotations revealing the roles of the corresponding speakers.

First, we extend the phoneme set $\mathcal{V}$ into a set with $R \cdot |\mathcal{V}|$ elements, where R is the number of roles in the dataset, by annotating each phoneme with all the possible speaker roles. An example is given in Fig. 3. That way, we expect to capture micro-variations at the phoneme level which could reveal differences in the acoustic patterns between different speaker roles. The fact that the acoustic characteristics differ substantially between certain roles at the utterance level [10] indicates that there could be distinct patterns identifiable at the phoneme level as well. Moreover, there are studies supporting a correlation between the social role of a speaker and the use of phonological structures which cannot be reflected in the usual written phonetic representation of words [15, 16]. Of course, we do know that the differently annotated versions of the same phoneme are related to each other. One way to take advantage of this fact is to make those phonemes share the same root in the phonetic decision trees constructed during training [17], so that they can share the same HMM states. In any case, if we are working with a dataset where such role-dependent phonetic variations are non-existent, we expect that, given enough data, the transition probabilities of the HMMs as well as the acoustic likelihoods mapping features to HMM states corresponding to the same base phoneme will be equal for the different roles, thus not affecting the RASR performance.

(a) Original phoneme set. (b) Role-annotated phoneme set.

Fig. 3: Extending the phoneme set to include role annotations, assuming two roles, A and B.

Having this extended set of phonemes, we can create the role-annotated pronunciation lexicon L^+ as the union of R WFSTs $\{\tilde{L}_i\}_{i=1}^{R}$, each one corresponding to one of the available speaker roles, so that

$$L^+ = \left(\bigcup_{i=1}^{R} \tilde{L}_i\right)^* = \left(\bigcup_{i=1}^{R}\bigcup_{j} \tilde{W}_{ij}\right)^* ,$$

where $\tilde{W}_{ij}$ is the WFST modeling the phonetic representation of the jth word corresponding to the ith role and * is the Kleene star [13]. An example is given in Fig. 4. This step, apart from providing the necessary information to distinguish same words with different annotations, is also a tool indicating the possible variabilities between the vocabularies used by the different roles, thus facilitating SRR. This is because if a certain word is never uttered by a specific role in the training corpus, then this combination of word and role annotation can be removed from the lexicon.

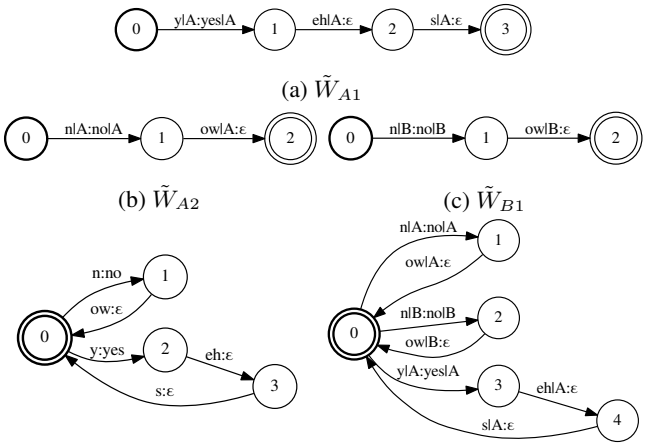

Fig. 4: (a),(b),(c): WFST representation of the words "yes" and "no" corresponding to the roles A and B. (d): Lexicon containing the non-annotated words "yes" and "no". (e): Annotated lexicon $\left(\tilde{W}_{A1} \bigcup \tilde{W}_{A2} \bigcup \tilde{W}_{B1}\right)^*$. For simplicity, no weights are denoted.

Finally, in order to train the LM, after converting all the words in the training corpus from non-annotated to annotated ones, we are exploring two alternatives of how to format the training transcripts. On the one hand, we can list all the speaker turns with start- and end-of-sentence symbols, as shown in Fig. 5b. This would be equivalent to training R LMs, one for each role, and composing them into a final LM. However, this approach would result in an LM which does not contain information about the transitions from one speaker role to the other. Alternatively, we can concatenate all the speaker turns into one "sentence" per training session, as shown in Fig. 5c, thus capturing some information about the interaction between the roles by calculating the occurrence probability of n-grams involving words annotated with different roles.

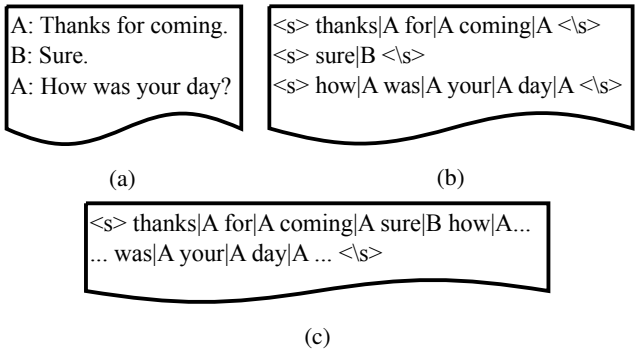

Fig. 5: (a): Segment of a hypothetical original transcription of a conversation between the roles A and B. (b),(c): Two approaches of formatting it as part of the LM training corpus.

2.3. Normalization

An important aspect of any modern ASR system is speaker normalization and adaptation. One of the most common and simplest feature transformations is Cepstral Mean Normalization (CMN) [18], where the statistics are usually collected per speaker, in order to eliminate the convolutive effects of the specific structure of each speaker's vocal tract. Speaker-Adaptive Training (SAT) through Constrained Maximum Likelihood Linear Regression (CMLLR) [19] is also a standard step taking place during ASR training in the GMM paradigm. For acoustic modeling through DNNs, a common technique is to supply i-vectors as input features to the network concatenating them with the regular acoustic feature vectors [20].

However, since in RASR part of the final goal is to segment the input signal based on the deduced speaker role information, we cannot assume speaker-homogeneous segments at evaluation time and, therefore, we cannot use the common speaker normalization techniques, as traditionally implemented. Based on the short-term speaker stationarity hypothesis [21], according to which it is unlikely for speaker

change points to occur very frequently, we can instead normalize per utterance and calculate the CMN statistics and extract i-vectors in an online manner, taking into account only a small history window. In that case, in order to train and test on same features, the same approach for CMN is followed during training and the online i-vector extractor is trained on similarly short segments. The AM is trained with the DNN paradigm, where the DNNs are initialized based on the alignments computed by a GMM-trained model, but without SAT.

But even with that approach, which allows for a reasonable speaker normalization, there is a deeper problem inherent in the very goal of RASR. As explained, feature normalization in ASR aims at discarding any speaker-specific variabilities. However, in the RASR context, this is not necessarily the desired behavior, since speaker-specific characteristics are known to be useful for speaker role prediction [10]. So, in this study we are evaluating RASR both with and without CMN to explore its effects. Either case, the i-vectors are used as already described, since they are not explicitly normalizing the acoustic features, but the DNNs are expected to decide how to better exploit the information they carry. Additionally, their use when extracted online has been shown to lead to improved results for the task of diarization [22].

2.4. Evaluation metrics

In RASR we are interested in the performance of the system with respect both to the output text and to the roles predicted. For evaluating the accuracy in terms of the textual information we are using the Word Error Rate (WER) after discarding role annotations as it is traditionally used for ASR.

For evaluating the role predictions we are using the alignments of the output text together with the role annotations to extract the turn boundaries. We then estimate the error rate the same way Diarization Error Rate (DER) is traditionally calculated, but by enforcing the speaker matching between the reference and the hypothesis to be between the same speaker roles. We call this metric the Role Error Rate (RER). In particular, in diarization, since the output is the result of an unsupervised clustering, there is no natural correspondence between the reference and hypothesis speaker labels; so it is essential to find an optimal matching between them by minimizing a mismatch criterion [23]. In RASR this is no longer the case, since the output is annotated with the same labels as the input roles.

Finally, in order to have an estimate of the performance jointly for the text and role predictions, we can use the Role-Annotated Word Error Rate (RAWER), which is calculated the same way as the WER, but using the annotated words.

3. DATASET

In this work, we evaluate our proposed method on datasets from the clinical psychology domain. Specifically, we apply the RASR system to Motivational Interviewing (MI) sessions between a therapist (T) and a client (i.e., patient) (C) collected from five clinical trials (ARC, ESPSB, ESP21, iCHAMP, HMCBI) [24]. We collectively refer to those sessions as the MI corpus.

Some descriptive analysis for the datasets is presented in Table 1. Unfortunately, the client IDs are not available for the HMCBI sessions, so the exact total number of different clients is not known. However, under the assumption that it is highly improbable for the same client to visit different therapists in the same study, and having the necessary metadata available for the rest of the corpus, we make the train/test split in a way that we are highly confident there is no overlap between speakers. Out of the 143 available sessions, 74 are included in the training set and 69 are held out for testing.

	#sessions	dur-T	dur-C	#T	#C
ARC	9	3.02h	1.47h	3	9
ESPSB	38	17.88h	10.63h	15	38
ESP21	19	8.32h	5.43h	8	19
iCHAMP	7	2.98h	2.53h	5	7
HMCBI	70	7.98h	13.60h	15	–
total (MI)	143	40.16h	33.66h	43	–
MI-train	74	22.40h	18.96h	16	–
MI-test	69	17.76h	14.70h	27	–

Table 1: Descriptive analysis for the MI dataset. dur-T and dur-C are the total speech duration assigned to therapists and clients, respectively, after force-aligning the manual transcriptions with the audio sessions at the word level and allowing a maximum of $0.1\,\mathrm{sec}$ in-turn silence. By #T and #C we denote the total number of different therapists and clients.

In order to train the required LMs, both the training part of the MI corpus and the transcribed sessions provided by the Counseling and Psychotherapy Transcripts Series[1] (CPTS) are used, as described in Section 4. The sizes of the corpora are given in Table 2.

	#words-T	#words-C	|voc-T|	|voc-C|	|voc|
MI-train	289K	243K	5.5K	6.3K	8.1K
CPTS	1.96M	4.56M	20.7K	31.2K	35.6K

Table 2: Size of the corpora used for LM training. #words-T and #words-C are the number of words uttered by therapists and clients, respectively. |voc-T|, |voc-C|, and |voc| are the vocabulary sizes of text assigned to therapists, clients, or both, with voc≡voc-T∪voc-C.

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

4. EXPERIMENTS AND RESULTS

First, we force-align both the training and test sets at the word level using the ASR system developed in [1] which is focused on psychotherapy dialogues based on Motivational Interviewing. For this task, as well as for the rest of the ASR-related tasks, the Kaldi speech recognition toolkit [25] is used. Based on the alignments, we segment the MI sessions in two different ways; according to the manually annotated speaker turns, provided by the available transcripts, and according to whether the silence between two consecutive words is longer than a certain threshold (equal to 1 sec). We call the first segmentation *non-mixed* and the second one *mixed*, since the resulted utterances are not speaker-homogeneous. We use *non-mixed* segments to train the RASR system and *mixed* segments to evaluate it. In a real-world scenario, the *mixed* segmentation could be created by a Voice Activity Detection (VAD) algorithm.

In order to have a rough understanding of the difficulty of the database we are working with in terms of both the problems (ASR and diarization) RASR is called to confront, we are first using the *non-mixed* testing sessions to evaluate with publicly available tools for those tasks, with the results been reported in Table 3. In particular, we estimate the DER using the LIUM SpkDiarization toolkit [26], employing VAD, Generalized Likelihood Ratio (GLR)-based segmentation, Bayesian Information Criterion (BIC)-based clustering, resegmentation based on Viterbi decoding, and reclustering based on cross entropy. The diarization ground truth is obtained through the word alignments, by allowing a 0.25 sec-long collar around the reference boundaries. Additionally, we estimate the WER using Kaldi's pre-trained ASpIRE model[2], a state-of-the-art ASR model for conversational English. The result is based on the best path found on the decoding lattice using the LM-weight which minimizes the WER. For comparison, in the same Table (3rd column) we report the WER using an ASR system trained the exact same way as the RASR system we are describing below, but without considering any role annotations.

LIUM (DER)	ASpIRE (WER)	MI (WER)
39.61	41.27	54.21

Table 3: DER (%) when using the LIUM SpkDiarization toolkit and WER (%) when using the Kaldi's ASpIRE pre-trained model and the RASR system trained on non-annotated MI training data, evaluated on the non-annotated, *non-mixed* testing sessions.

To train and evaluate the RASR system, all the words in the available transcripts have to be converted to their role-annotated equivalent ones. Their phonetic representation is given by the CMU dictionary[3] which is extended with the role annotations T and C like in Fig. 3.

The AM is constructed by the consecutive training steps followed in the standard Kaldi recipes without SAT; namely monophone training, triphone training with deltas and ddeltas, training with Linear Discriminant Analysis (LDA) on spliced frames followed by a Maximum Likelihood Linear Transform (MLLT), and DNN training. For the last step Time Delay Neural Nets (TDNNs) with sub-sampling and p-norm nonlinearities [27, 28], as implemented in the `nnet2` Kaldi setup, are used. For the GMM training 13-dimensional MFCCs are used, while for the DNN training 40-dimensional MFCCs are concatenated with 100-dimensional i-vectors. We report results (Table 4) both when the differently annotated phonemes share the same tree root (*share*) and when they do not (*no-share*), as explained in Section 2.2.

Results are reported for the case when we apply online CMN, as well as when we do not apply CMN at all. For the first case, we are taking into consideration only a 2 sec-long history window of the utterance which is decoded and the same window is used during training as well. Also, the same window is used for the online i-vector extraction. The actual distribution of the duration of the intervals between speaker change points in the entire dataset is shown in Fig. 6, where a speaker change point is here defined as the endpoint of a word when the next one is uttered by a different speaker. As observed, there are actually many short speaker turns (less than 2 sec-long). However, the choice of the window length is a trade-off decision, since very short windows would result in non-robust i-vectors. Additionally, lots of such segments are expected not to greatly affect the final output quality, being e.g. fillers like "mm-hmm", etc.

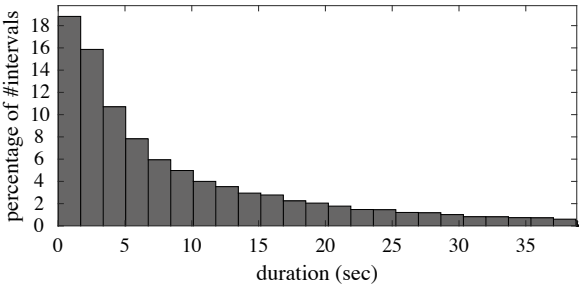

Fig. 6: Distribution of the duration of the intervals between speaker change points in the MI dataset.

The LMs are 3-gram models with Kneser-Ney smoothing, trained with the SRILM toolkit [29]. The training corpus can be created either by concatenating consecutive turns as in Fig. 5c (*conc*), or by considering them independently as in Fig. 5b (*no-conc*). In any case, the created LMs are interpolated with "background" ones which are trained in the same manner, using the transcribed sessions provided by the

[2] `http://kaldi-asr.org/models/m1`

[3] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

Counseling and Psychotherapy Transcripts Series. The mixing weight used is 0.2 for the background LM. It is noted that the pronunciation dictionary is reduced to cover only the vocabulary found in the final LM.

Results, both in terms of WER and RER, are reported in Table 4 for all the aforementioned combinations. All the results are based on the the best path found on the decoding lattice using the LM-weight which minimizes the RAWER. Similarly to the diarization ground truth used for the result presented in Table 3, the reference for RER is obtained through the annotated word alignments, by setting a 0.25 sec-long tolerance collar around the obtained boundaries.

		conc share	conc no-share	no-conc share	no-conc no-share
CMN	WER	58.82	61.47	**58.78**	61.37
	RER	**39.74**	41.32	39.86	41.27
no-CMN	WER	63.64	65.07	63.45	65.13
	RER	41.84	42.63	41.47	43.82

Table 4: WER (%) and RER (%) using RASR on MI with or without CMN for the GMM training. The annotated versions of the same phoneme may or may not share the same root of the phonetic decision trees (*share* vs. *no-share*) and the LM may be trained on a corpus which contains all the speaker turns independently (*no-conc*) or concatenated per session (*conc*).

5. DISCUSSION AND FUTURE WORK

As observed in Table 4, the *share* setup yields improved results compared to the *no-share*, where all the annotated phonemes are treated in a completely independent manner. This comes at no surprise, especially since it has been shown [10] that for the particular dataset the acoustic variabilities between the two roles are not intense (it is noted that in [10] a superset of the dataset that we use in this study is explored since CTT sessions are not used here). Between the *conc* and *no-conc* approaches, no substantial differences are observed, suggesting that alternative ways to capture the dyadic interactions through the LM should be explored in the future.

The relative differences are similar between the various combinations when we apply online CMN and when we do not, but using CMN consistently improves the overall performance in terms of both WER and RER. However, the tasks of ASR and diarization are in general conflicting in that aspect, since the former tries to discard any speaker-specific characteristics in order to capture only the linguistic information carried by the speech signal, thus calling for speaker normalization techniques, while it is the very goal of the latter to find exactly those speaker-specific variabilities. SRR stands somewhere in-between, since it needs to identify patterns that are shared between various speakers but can differentiate the speaker roles. We believe that finding better normalization techniques appropriate for the hybrid task of RASR is an issue that requires further investigation.

By comparing the WER of the RASR system (Table 3) with the baseline results of the 3rd column in Table 2, we observe a non-negligible performance degradation when using the annotated dataset. However, we should keep in mind that the experiment in Table 2 uses speaker-homogeneous segments, thus assuming an ideal diarization step. A point of greater concern is the performance gap between the ASpIRE ASR and the MI ASR models (columns 2 and 3 in Table 2). We believe that the most important factors leading to such a difference in the estimated WER are a) the much better speaker adaptation with SAT and better (for the task of ASR) CMN and i-vector extraction and b) the usage of a much bigger dataset for the training of the ASpIRE model. We note that this model uses the Fisher English corpus [30] with additional data augmentation. Being able to adapt ASR models trained on out-of-domain data in order to be used for the task of RASR is a topic of current research since it is difficult, if not impossible, to obtain large datasets with the desired speaker roles in order to train an RASR system comparable to a modern large-vocabulary ASR system.

On the other hand, the RER results reported in Table 3 are comparable to the baseline DER results of Table 2. We should note, here, that RER is in fact a somehow stricter version of DER since it incorporates both the diarization and the speaker role prediction errors. Importantly, however, diarization by definition does not assume an a priori known number of speakers per session, while RASR assumes predefined speaker roles.

6. CONCLUSION

We proposed a system suitable for automatic rich transcription of conversational data, able to output at the same time textual information and speaker role predictions at the word level. We tested our approach, which we call Role Annotated Speech Recognition, on clinical dyadic interactions and we experimented with different system designs. Based on those experiments, we observed promising results, but also identified downsides and points requiring further research and investigation. The two main directions of our future research efforts will be a) exploring feature normalization methods (or even novel feature sets) suitable for the goal of RASR and b) finding a mechanism for adapting and extending pre-trained ASR models, in order to be used for the task of RASR.

7. ACKNOWLEDGEMENTS

This work was supported by NIH and DARPA. NF is partially supported by the USC Annenberg Fellowship.

8. REFERENCES

[1] 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.

[2] Martine Garnier-Rizet, Gilles Adda, Frederik Cailliau, Jean-Luc Gauvain, Sylvie Guillemin-Lanne, Lori Lamel, Stephan Vanni, Claire Waast-Richard, et al., "Callsurf: Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content.," in *LREC*, 2008.

[3] Matthew P Black, Athanasios Katsamanis, Brian R Baucom, Chi-Chun Lee, Adam C Lammert, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech communication*, vol. 55, no. 1, pp. 1–21, 2013.

[4] Sowmya Rasipuram and Dinesh Babu Jayagopi, "Automatic assessment of communication skill in interview-based interactions," *Multimedia Tools and Applications*, pp. 1–31, 2018.

[5] Benjamin Bigot, Corinne Fredouille, and Delphine Charlet, "Speaker role recognition on tv broadcast documents," in *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

[6] 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, 2012.

[7] Benjamin Bigot, Julien Pinquier, Isabelle Ferrané, and Régine André-Obrecht, "Looking for relevant features for speaker role recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[8] Neha P Garg, Sarah Favre, Hugues Salamin, Dilek Hakkani Tür, and Alessandro Vinciarelli, "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*. ACM, 2008, pp. 693–696.

[9] Géraldine Damnati and Delphine Charlet, "Multi-view approach for speaker turn role labeling in tv broadcast news shows," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[10] Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson, and Shrikanth Narayanan, "Combined speaker clustering and role recognition in conversational speech," in *Interspeech*, 2018.

[11] Tae Jin Park and Panayiotis Georgiou, "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks," in *Interspeech*, 2018.

[12] 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 *Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on*. IEEE, 2012, pp. 118–123.

[13] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 16, pp. 69–88, 2002.

[14] Takaaki Hori and Atsushi Nakamura, "Speech recognition algorithms using weighted finite-state transducers," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–162, 2013.

[15] William Labov, *The social stratification of English in New York city*, Cambridge University Press, 2006.

[16] Jan-Petter Blom, John J Gumperz, et al., "Social meaning in linguistic structure: Code-switching in norway," *The bilingualism reader*, pp. 111–136, 2000.

[17] Steve J Young, Julian J Odell, and Philip C Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 307–312.

[18] Richard Schwartz, Tasos Anastasakos, Francis Kubala, John Makhoul, Long Nguyen, and George Zavaliagkos, "Comparative experiments on large vocabulary speech recognition," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1993, pp. 75–80.

[19] Mark JF Gales et al., "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[20] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors.," in *ASRU*, 2013, pp. 55–59.

[21] Arindam Jati and Panayiotis Georgiou, "Speaker2vec: Unsupervised learning and adaptation of a speaker manifold using deep neural networks with an evaluation on speaker segmentation," *Proc. Interspeech 2017*, pp. 3567–3571, 2017.

[22] Srikanth Madikeri, Ivan Himawan, Petr Motlicek, and Marc Ferras, "Integrating online i-vector extractor with information bottleneck based speaker diarization system," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[23] D Lilt and Francis Kubala, "Online speaker clustering," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–333.

[24] 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.

[25] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[26] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Interspeech*, 2013.

[27] Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 215–219.

[28] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[29] Andreas Stolcke, "SRILM–an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[30] Christopher Cieri, David Miller, and Kevin Walker, "The fisher corpus: a resource for the next generations of speech-to-text.," in *LREC*, 2004, vol. 4, pp. 69–71.