IMPROVING SPEAKER DIARIZATION USING SOCIAL ROLE INFORMATION

Ashtosh Sapru[1,2], Sree Harsha Yella[1,2] and Hervé Bourlard[1,2]

[1] Idiap Research Institute, 1920, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
ashtosh.sapru@idiap.ch,sree.yella@idiap.ch,herve.bourlard@idiap.ch

ABSTRACT

Speaker diarization systems for meetings commonly model acoustic and spatial information, ignoring that meetings are instances of human interactions. Recent studies have shown that social roles influence the interaction patterns of speakers. This paper proposes a novel method to integrate social roles information in the speaker diarization framework. First, we modify the minimum duration constraint in baseline diarization system by using role information to model the expected duration of speaker's turn. Furthermore, we also propose a social role n-gram model as prior information on speaker interaction patterns. The proposed method is integrated in the state-of-the-art diarization system to reduce the speaker error. Experiments are performed on AMI corpus which is annotated in terms of social roles. The proposed method reduces the speaker error by 16% relative to baseline HMM-GMM system. Furthermore, the paper also investigates the performance of the proposed method on other meeting scenarios like those from NIST Rich Transcription campaigns. Experiments on Rich Transcription meetings reveal that speaker error can be reduced by 13% relative to the baseline system, thus demonstrating the potential of the proposed method.

Index Terms— Speaker diarization, Social Roles, HMM-GMM

1. INTRODUCTION

Speaker Diarization aims at inferring who spoke when in an audio stream. Most of the advances in this domain have been due to signal processing techniques for enhancing speech signals [1], and statistical modeling of multistream information [2]. However, speech data used in diarization are instances of multiparty conversations which follow predictable patterns of interaction between participants. Recently, lot of studies have focused on statistical modeling of small group social interactions. In particular, study of roles and automatic role recognition has been investigated on meeting recordings like CMU corpus [3], AMI corpus [4], Mission survival corpus [5] and ICSI corpus [6]. Speaker roles are stable behavioral patterns in multiparty interactions and automatic role recognition assigns a label, i.e., a role to each of the speakers. In literature, the type of roles can be formal roles, which are imposed by scenario of the conversation, for instance chairperson in professional meetings, and socio-emotional (social) roles [5], which describe relation between conversation participants and their roles "*oriented towards functioning of group as a group*". A common format in most studies on automatic role recognition, is to train statistical classifiers on a set of features that represent participant's conversational and prosodic behavior.

This paper investigates whether speaker diarization can be improved by modeling the influence of social roles on the interaction patterns of speakers. In particular, we examine two limitations of current state-of-the-art systems, i.e., a speaker independent minimum duration constraint and a uniform prior on speaker interaction patterns.

Previous studies [7, 8] have shown that speaker sequence modeling can be used to improve diarization. In [7], interaction patterns between speakers where estimated on a per meeting basis and did not consider any role information. In [8], formal roles of speakers were used to estimate the probability of interaction patterns between speakers. However, a limitation of this approach is that formal roles are imposed by specific scenario of meetings and, as such, do not generalize across multiple data sets. In comparison, this study investigates the influence of social roles on turn taking patterns of speakers. Social roles [5, 9] represents a role coding scheme that can be generalized across any type of multiparty conversation and reflects the dynamics of the participant interaction. In this study, roles are used to estimate the expected duration of speaker's turn and social role n-grams are used to model speaker sequence distribution.

Our work extends the commonly used speaker diarization system based on Hidden Markov Model (HMM)-Gaussian mixture (GMM) modeling [10], by including social role information in the speaker segmentation step. Our investigation is carried on the Augmented Multimodal Interaction (AMI) [11] database which is annotated in terms of social roles. Since social roles represent a generalized coding scheme, we also investigate whether the social role statistics estimated on AMI meetings, work on meetings from the Rich Transcription (RT) dataset. The paper is organized as follows: Section 2 describes the dataset and the social role annotation, Section 3 describes the baseline diarization system and its extension to incorporate social role information, Section 4 describes experiments on the AMI meetings and RT meetings. The paper is finally concluded in Section 5.

2. DATA DESCRIPTION

The dataset used in this study consists of meetings selected from AMI corpus. The AMI Meeting Corpus is a collection of meetings captured in specially instrumented meeting rooms, which record the audio and video for each meeting participant. The corpus contains both scenario and non-scenario meetings. In the scenario meetings, four participants play the role of a design team tasked with designing a new remote control. A set of 59 meetings from the scenario portion of AMI Meeting Corpus containing 128 different speakers was annotated in terms of social roles. We partitioned these meetings in to disjoint set of 29 meetings for training and 30 meetings for testing. The annotation process as detailed in [9], is based on annotating each speaker with a social role which can dynamically change within a meeting. The annotation process associates the following set of social roles that a speaker can assume: *Protagonist* - a speaker that takes the floor, drives the conversation, asserts its authority and

assumes a personal perspective; *Supporter* - a speaker that shows a cooperative attitude demonstrating attention and acceptance and provides technical and relational support; *Neutral* - a speaker that passively accepts other speaker's ideas; *Gatekeeper* - a speaker that acts like group moderator, mediates and encourages the communication.

More formally, the annotated data can be described using the following set: $\{(t_1, d_1, s_1), (t_2, d_2, s_2), ..., (t_N, d_N, s_N)\}$, where t_n represents the beginning time of n-th speech turn, d_n represents the duration of the turn, s_n represents the speaker label associated with the turn and N is total number of turns in the recording. Furthermore, each speaker label s_n also has a social role label r_n associated with it as result of annotation process. Several previous studies [5, 12, 13], describe automatic role recognition algorithms that perform non trivial classification of roles based on participants non verbal behavior. A set of structural and prosodic features F_s are extracted for each speaker s. After extracting the feature vector F_s, a supervised classifier can be trained to build a speaker to role mapping $\phi(s) -> r$. The function $\phi(s) = \arg\max_{r \in \mathcal{R}} P(r|F_s, \lambda)$, where λ are the parameters of the role recognition model and $\mathcal{R}$ is the set of social role labels. In the following discussion we assume knowledge of ϕ which is estimated using training data as explained in [13]. The focus of the study is to explore how social role information can be used to improve segmentation of meeting recordings into sequences of speaker turns.

3. SOCIAL ROLES BASED SPEAKER DIARIZATION

Conventional diarization systems are based on agglomerative clustering framework where each speaker is modeled as a HMM state and each state distribution is modeled using a GMM. The baseline system [10] used in this work achieved state-of-the-art performance in several NIST evaluations. It is initialized by uniformly segmenting a given audio recording in to segments representing initial speaker clusters. The number of initial clusters is much higher than the actual number of speakers in the recording. The algorithm iteratively merges the closest clusters until a stopping criterion is met. After each merge, speaker boundaries are realigned based on the estimated speaker models using a Viterbi decoder. For the merging and clustering stopping criteria, a modified version of Bayesian Information Criterion (BIC) is used [14].

In the baseline system, each speaker cluster is associated with a HMM model. We denote by $\mathcal{S} = \{\mathcal{S}^1, ..., \mathcal{S}^K\}$, the set of K models, where $\mathcal{S}^k$ is the model associated with speaker s^k. The individual speaker models can be concatenated to represent a hypothesized speaker turn taking sequence $\{(t_1, d_1, s_1), ..., (t_N, d_N, s_N)\}$. Here N is number of speaker turns and t_n and d_n represent the start time and duration of the n^{th} turn. $s_n \in \{s^1, ..., s^k\}$ is the speaker associated with turn n. The speaker sequence representing the turn taking is denoted by $S = \{s_1, ..., s_N\}$.

Consider a data sequence $X = \{x_1, .., x_t, ..., x_T\}$ made of T speech frames. The re-estimation step in speaker diarization system aims at finding the optimal speaker sequence $\hat{S}$ that maximizes the $P(S|X)$.

$$\hat{S} = \arg\max_{S} P(S|X) \quad (1)$$

$$\hat{S} = \arg\max_{S} P(X|S)P(S) \quad (2)$$

Here $P(S)$ represents the prior probability of a specific speaker sequence and $P(X|S)$ represents the likelihood of the acoustic data. A minimum duration constraint is imposed by associating D^k states with each speaker model $\mathcal{S}^k$ and all the states have the same emission probability modeled with a GMM.

One of the usual assumptions of diarization systems is that minimum duration D^k (2.5 seconds in the baseline system), is independent of conversation dynamics, and is fixed for all speakers across all the recordings. Furthermore, a uniform prior on speaker sequence distribution $P(S)$ is imposed such that transitions between speakers are equally likely. In this study we further investigate the validity of these assumptions.

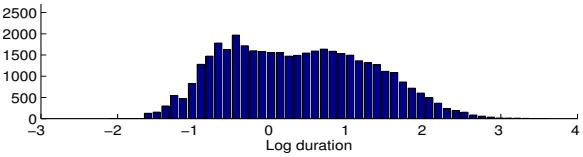

Fig. 1. Histogram of log duration of speaker turns on AMI corpus meetings.

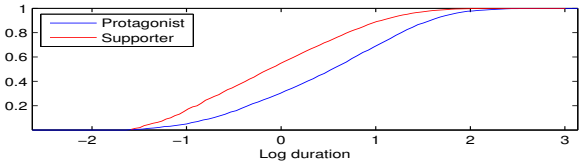

Fig. 2. Cumulative histogram of log duration of speaker turns for social roles of protagonist and supporters.

Figure 1, shows the histogram of log duration of speaker turns in the training set. We can see that distribution of speaker turns is not unimodal and speakers are likely to take shorter turns (less than 1 second) as well as longer turns (greater than 2.5 seconds). This observation is not entirely unexpected, as speaker turns provide information about various conversation phenomena,, such as, dominance and social roles. In particular, previous studies [9] on social roles have shown that, turn duration of speakers is influenced by social roles they assume. To observe this behavior, we plot the cumulative histograms of speaker turn data for protagonists and supporter roles as shown is Figure 2. The figure reveals that protagonists are more likely to produce longer turns compared to supporters. Further analysis reveals that for protagonists, 71% of speech time is associated with longer turns, while for supporters this percentage is only 53%. Furthermore, for protagonists, less than 7% of speech time is due to short turns while for supporters it is 21%.

This analysis leads us to propose a minimum duration constraint based on social role of speakers, i.e, minimum duration D^k for speaker model $\mathcal{S}^k$, that is fixed in traditional diarization systems, is now made a function of speaker's social role $D^k(r)$. The values of $D^k(r)$ were selected by fitting a probability distribution over the log turn duration for each social role in training data. While earlier studies [15] have assumed a log normal distribution for turn duration, we modeled the log duration of turns using a GMM. Note that log normal distribution is equivalent to a GMM with a single mixture component. In this study the number of mixture components was selected based on BIC. Our analysis revealed that the model with two mixture components best explains the variability of turn duration. The mean parameters of the GMM were estimated using maximum likelihood criterion. The minimum duration $D^k(r, c)$ for each speaker is represented as a function of role and mixture

component $c \in \{1, 2\}$. $D^k(r, 1)$ is the expected duration of short turns while $D^k(r, 2)$ is the expected duration of longer turns. The speaker model $\mathcal{S}^k(r, 1)$ is formed by concatenation of $D^k(r, 1)$ states and speaker model $\mathcal{S}^k(r, 2)$ is formed by concatenation of $D^k(r, 2)$ states. The models, $\mathcal{S}^k(r, 1)$ and $\mathcal{S}^k(r, 2)$ have the same emission probability, however, they represent short and long turns produced during participant interaction.

Traditional diarization systems also assume a uniform prior on speaker sequence distribution $P(S)$, considering all turn taking transitions between speakers equally likely. However, the knowledge of social roles can be used to impose more meaningful information on speaker sequence distribution. As an analogy with ASR systems, we propose a role n-gram model, that represents the interaction between speakers in terms of their social roles. Under a Markov assumption,

Table 1. Perplexity of social role sequences for AMI meetings.

	Unigram	Bigram	Trigram
Perplexity	4.4	3.5	2.9

we can factor the speaker sequence distribution in terms of

$$P(S) = P(s_1, .., s_N) = P(s_1, .., s_p) \prod_{p+1}^{N} P(s_n | s_{n-1}, ..., s_{n-p}) \tag{3}$$

Using the mapping $\phi(s) -> r$, the above can be expressed as,

$$P(S) = P(R) = P(r_1, .., r_p) \prod_{p+1}^{N} P(r_n | r_{n-1}, ..., r_{n-p}) \tag{4}$$

The quality of language models in ASR systems is evaluated by computing the perplexity on separate data set. We use the same criterion to select an appropriate language model for conversational turn taking. We calculated the unigram, bigram and trigram estimates on the training data. The observed perplexity on an independent development set is reported in Table 1.

Table 1 reveals a drop in perplexity when moving from unigram to bigram models and their is further decrease in perplexity for trigram model. This reveals that social role of current speaker conditioned on the social roles of previous speakers produces a large reduction in speaker sequence perplexity. The most probable n-grams correspond to protagonists and gatekeepers and their interaction with supporters. For the rest of this study, we have used trigram models as they showed the best performance. By estimating the social roles of speakers using the algorithm in [13], we propose a modified speaker re-estimation step,

$$\hat{S} = \arg\max_{S} P(X|S, R) P(R) \tag{5}$$

Schematically the proposed diarization system can be summarized as follows:

1. Extract acoustic features from the audio file.
2. Speech/non-speech segmentation and reject non-speech frames.
3. Initialize the model for the initial clusters using linear initialization.
4. Perform iterative merging using the following steps:
   (a) Find speaker to role mapping $\phi(s) = r$, from the current speaker segmentation.

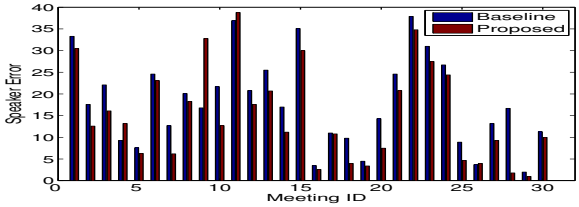

Fig. 3. Per-meeting speaker error for the 30 meetings of the AMI corpus obtained using the baseline system and the proposed system.

   (b) Re-segment the data using role based duration constraints and role trigram model.
   (c) Retrain the speaker models using the Expectation-Maximization (EM) algorithm.
   (d) Select the cluster pair with the largest merge score (based on ΔBIC) that is > 0.
   (e) If no such pair is found, stop and output the current clustering.
   (f) Merge the pair of clusters found in step (d). The models for the individual clusters in the pair are replaced by a single, combined model.
   (g) Go to step (a).

Step 1 to 3 are the same as in baseline system. First multiple distant microphones are beamformed to produce a single enhanced signal using Beamformit toolkit [16]. Acoustic features representing 19 MFCC coefficients are extracted using a 30ms window shifted by 10ms. After that we run an automatic speech/non speech segmentation and eliminate nonspeech regions to extract acoustic features $X = \{x_1, ..., x_T\}$. Step 4 introduces new stages which are different from the baseline diarization system. These include a speaker to role mapping and a novel Viterbi segmentation using social role information. During re-estimation of speaker models, we do not consider frames corresponding to short turns, as previous studies [17] have shown that diarization errors are very high for these segments. The merging and clustering stopping criteria are the same for both the proposed system and baseline system.

4. EXPERIMENTS

In this section, we describe the experiments that were performed to compare performance of the proposed method against the baseline speaker diarization system. First, we detail the experiments on AMI testset and later we also evaluate the generalized performance of the proposed method on NIST Rich Transcription meetings.

For evaluation of diarization performance, most commonly used metric is the Diarization Error Rate (DER). DER is the composed of false alarm time, miss time and speaker error time. Since both, proposed and baseline systems use the same speech/nonspeech segmentation, the difference in performance is evaluated following only the speaker error. Similar to NIST evaluations we have used a collar of 0.25 seconds around reference segment boundaries.

The social role recognition system used in this study is based on the work described in [13]. In brief, this system operates in two stages, first each speaker's acoustic and structural behavior patterns are represented using a high dimensional feature vector, then in the next stage a linear support vector machine (SVM) classifier is used to predict the speaker's role. While [13] also includes linguistic information of speakers, for this study we have only extracted non verbal

features. The social role based trigram language model is estimated using SRI toolkit [18].

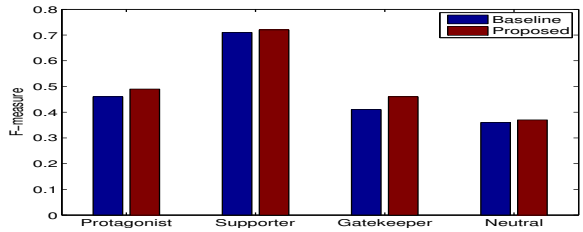

Fig. 4. Social role recognition performance for each of the four roles using the baseline system and the proposed system.

For our first experiment, we include automatic speech/nonspeech segmentation and assume an unknown number of reference speakers. The hypothesized number of final speakers is determined by the BIC based stopping criterion. The performance comparison of the proposed system and the baseline system is shown in Table 2. Under these conditions, experiments reveal that proposed system, which integrates social role information in the diarization system, results in 16% relative improvement over baseline system. The per meeting performance comparison of the two systems is shown in Figure 3. It can be seen that the proposed method outperforms the conventional approach in most of the cases.

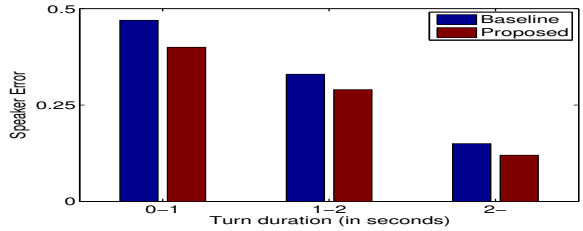

Fig. 5. Variation in speaker error for various sizes of turn duration for the baseline system and the proposed system.

We also compared the role recognition performance from the final output of the two systems, shown in Figure 4. It can be seen that the social roles of speakers are recognized more accurately for the proposed system. The improvement in performance is spread across all the four social roles. These results reveal that iteratively performing role recognition and speaker sequence decoding, improves both final diarization performance and social role recognition. Finally, we also performed an analysis of diarization systems performance as a function of speaker turn duration. For this turns in the reference transcripts are partitioned into three bins, short turns with duration less than 1 second, intermediate turns with duration between 1 second and 2 second and long turns with duration greater than 2 seconds. The results for this analysis are presented in Figure 5. For both systems similar trends are observed, speaker error decreases as turn duration increases. However, the proposed diarization system shows improved performance for all bins.

In order to investigate the performance of the proposed system on other meeting scenarios, we also compared baseline and proposed systems on NIST Rich Transcription (RT) dataset. RT dataset contains meeting domain recordings, representative of spontaneous conversation. Contrary to AMI meetings, these recordings are not necessarily elicited using a specific scenario. However, this does not represent a significant drawback , since social roles represent a generalized role coding scheme, and can conceivably be adapted to multiple conversation scenarios. For our analysis, we selected 15 meeting recordings comprising the evaluation sets RT-07 and RT-09. The social role information represented using parameters of duration model and n-gram role model were obtained on the AMI training set. Table 2 shows the performance of both proposed and baseline systems for this dataset. Table numbers reveal that speaker error drops from 10.2% using the baseline system to 8.9% using the proposed system, which represents around 13% relative improvement in performance. This shows that the proposed system is effective in reducing the speaker error even on unseen scenarios. A meeting wise comparison of the two systems, shown in Figure 6, reveals that proposed system performs better in most of the cases. Further analysis revealed that the few meetings where the proposed system fails to improve over the baseline results correspond to cases that have higher than average missed speech rate.

Table 2. Speaker error obtained from the baseline system and the proposed system on AMI testset and RT dataset.

Dataset	Baseline system Speaker Error	Proposed system Speaker Error
AMI	17.6	14.8(16%)
RT 07,09	10.2	8.9(13%)

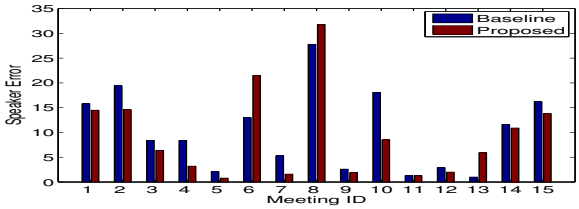

Fig. 6. Per-meeting speaker error for the meetings of the RT07 corpus and the RT09 corpus obtained using the baseline system and the proposed system.

5. CONCLUSION

In this study, we extended the state-of-the-art speaker diarization framework by using social roles to model the expected duration of speaker's turn. The interaction patterns between speakers were modeled using role trigrams. Our investigation carried on AMI corpus meetings, revealed that the proposed method reduces the speaker error by 16% relative to the baseline diarization system. We also investigated how the proposed system generalizes on a completely different corpus. Meetings from the Rich Transcription campaign, multiparty conversations collected in different sites, were used for this purpose. Results revealed a 13% relative improvement for the proposed system compared to the baseline system. In summary, the proposed method leads us to conclude that diarization error can be reduced by incorporating social role information. In future, the proposed method will be investigated in diarization system using both TDOA and MFCC features.

6. ACKNOWLEDGEMENT

This work was funded by the Hasler Stiftung under SESAME grant, the EU NoE SSPNet, and the Swiss National Science Foundation NCCR IM2.

7. REFERENCES

[1] Tranter S.E. and Reynolds D.A., "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14(5), 2006.

[2] Xavier Anguera Jose M. Pardo and Check Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transactions on Computers*, vol. 56, no. 9, September 2007.

[3] Banerjee S. and Rudnick A., "Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants," *Proceedings of ICSLP*, 2004.

[4] Salamin H. et al., "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," *IEEE Transactions on Multimedia*, vol. 11, November 2009.

[5] Zancaro M. et al., "Automatic detection of group functional roles in face to face interactions," *Proceedings of ICMI*, 2006.

[6] Laskowski K. et al., "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," *Proceedings of the 9th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, 2008.

[7] Han K.J. and Narayanan S.S., "Improved speaker diarization of meeting speech with recurrent selection of representative speech segments and participant interaction pattern modeling," in *Proceedings of Interspeech*, 2009.

[8] Fabio Valente, Deepu Vijayasenan, and Petr Motlicek, "Speaker diarization of meetings based on speaker role n-gram models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[9] A. Sapru and Hervé Bourlard, "Automatic social role recognition in professional meetings using conditional random fields," in *Proceedings of Interspeech*, 2013.

[10] Chuck Wooters and Marijn Huijbregts, "The icsi rt07s speaker diarization system," *Multimodal Technologies for Perception of Humans*, pp. 509–519, 2008.

[11] Carletta J., "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, pp. 181–190, 2007.

[12] Valente F. and Vinciarelli A., "Language-independent socio-emotional role recognition in the ami meetings corpus," in *Proceedings of Interspeech*, 2011.

[13] A. Sapru and Hervé Bourlard, "Investigating the impact of language style and vocal expression on social roles of participants in professional meetings," in *Affective Computing and Intelligent Interaction*, 2013.

[14] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *In Proc. IEEE Automatic Speech Recognition Understanding Workshop*, 2003.

[15] John Grothendieck, Allen L. Gorin, and Nash Borges, "Social correlates of turn-taking style," *Comput. Speech Lang.*, vol. 25, no. 4, pp. 789–801, Oct. 2011.

[16] Anguera X., Wooters C., and Hernando H., "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, September 2007.

[17] Mary Tai Knox, Nikki Mirghafori, and Gerald Friedland, "Where did i go wrong?: Identifying troublesome segments for speaker diarization systems," in *Proceedings of Interspeech*, 2012.

[18] A. Stolcke., "Srilm an extensible language modeling toolkit.," *Proc. of ICSLP*, 2002.