SPEAKER DIARIZATION OF MEETINGS BASED ON SPEAKER ROLE N-GRAM MODELS

Fabio Valente, Deepu Vijayasenan and Petr Motlicek

Idiap Research Institute, CH-1920 Martigny, Switzerland
{*fabio.valente,deepu.vijayasenan,petr.motlicek*}*@idiap.ch*

ABSTRACT

Speaker diarization of meeting recordings is generally based on acoustic information ignoring that meetings are instances of conversations. Several recent works have shown that the sequence of speakers in a conversation and their roles are related and statistically predictable. This paper proposes the use of speaker roles n-gram model to capture the conversation patterns probability and investigates its use as prior information into a state-of-the-art diarization system. Experiments are run on the AMI corpus annotated in terms of roles. The proposed technique reduces the diarization speaker error by 19% when the roles are known and by 17% when they are estimated. Furthermore the paper investigates how the n-gram models generalize to different settings like those from the Rich Transcription campaigns. Experiments on 17 meetings reveal that the speaker error can be reduced by 12% also in this case thus the n-gram can generalize across corpora.

Index Terms— Speaker diarization, meeting recordings, multi-party conversations, Speaker Roles, Viterbi decoding.

1. INTRODUCTION

Speaker Diarization aims at inferring *who spoke when* in an audio stream. Most of the recent efforts in the domain of meeting diarization have addressed the problem using acoustic or directional information, e.g., MFCC or TDOA features, ignoring the fact that meetings are instances of conversations.

Conversation analysis has been an active research field for long time [1] but only recently several works have focused on statistical modeling of phenomena in conversations. In between those studies, a lot of attention has been devoted to the recognition of roles. Roles are behavioural patterns [2] that speakers exhibit during the conversations. In literature, the term 'speaker role' is used to refer both to formal roles, for instance the chairperson in a meeting or the moderator in a public debate, as well as to functional roles [3] , i.e., the function that each speaker has in a spontaneous conversation. Automatic role recognition based on statistical classifiers has been applied in meetings recordings like the CMU corpus [4], the AMI corpus [5] and the ICSI corpus [6] as well as Broadcast [7] and telephone [8] conversations. Those works make use of non-verbal features like the speaker turn as well as the speaker sequence statistics. The underlying assumption is that the sequence of speakers in a conversation (thus a meeting), and the roles they have can be jointly modeled and statistically predicted. The sequence of speakers, i.e., the way speakers take turns in a conversation is supposed to be regulated by the role they have in the discussion.

This paper investigates whether the statistical information on the speaker sequence derived from their roles can be used in speaker diarization of meeting recordings. Previous works (see [9]), have successfully included statistics on the speaker sequence (referred as interaction patterns) in speaker diarization. However the information was considered recording dependent and not induced by, or put in relation with, any conversation phenomena. This work proposes to model the speaker sequence using n-gram of speaker roles. N-gram models can be then combined with the acoustic information coming from MFCC features. The approach is largely inspired by the current Automatic Speech Recognition (ASR) framework where the acoustic information from the signal, i.e., the acoustic score, is combined with the prior knowledge from the language, i.e., the language model. The most common form of language model is represented by words n-gram. In a similar way, given a mapping speakers to roles, n-gram models can encode the statistical information on how the participants take turns in the conversation.

The investigation is carried on the Augmented Multimodal Interaction (AMI) database annotated in terms of formal roles. Furthermore, all multi-party conversations share similar characteristics like the presence of a moderator (referred as gate-keeper in the literature on functional roles [3],[10]), thus the n-gram models should be able to generalize across different data sets. This hypothesis is then investigated on meetings from the Rich Transcription (RT) data.

The paper is organized as follows: section 2 describes the data set, the notation and preliminary experiments in terms of perplexity, section 3 describes the baseline diarization system and its extension to the use of role n-gram, section 4 describes experiments on the AMI corpus and section 5 describes experiments on the RT data set. The paper is finally concluded in section 6.

2. SPEAKER ROLES AND N-GRAM

The first investigation is based on the AMI meeting database [11], a collection of 138 meetings recorded with distant microphones for approximatively 100 hours of speech, manually annotated at different levels (roles, speaking time, words, dialog act). Each meeting consists of a scenario discussion in between four participants where each participant has a given role: project manager PM, user interface expert UI, marketing expert ME and industrial designer ID. The scenario consists in four employes of an electronic company that develop a new type of television remote controller. The meeting is supervised by the project manager. The dataset is divided into a training set (98 meetings), an development set (20 meetings) and a test set (20 meetings).

Let us consider the meeting as a sequence of speaker turns; although several definition of speaker turns have been given in literature, we consider here the simplified definition provided by [12] and [13], i.e., speech regions uninterrupted by pauses longer then 300 ms. More formally, for each meeting the following triplets are

This work was funded by the SNF IM2 project, the EU FP7 SSPNet project and the Hasler fundati on SESAME grant.

available:

$$T = \{(t_1, \Delta t_1, s_1), ..., (t_N, \Delta t_N, s_N)\} \quad (1)$$

where t_n is the beginning time of the n-th turn, Δt_n is its duration, s_n is the speaker associated with the turn and N is the total number of speaker turns in the recording. To simplify the problem, the time in overlapping regions between speakers (including back-channels) is assigned to the speaker that currently holds the floor of the conversation. Let us designate with $\varphi(S) \to R$ the one-to-one mapping between the four speakers and the four roles $R = \{PM, UI, ME, ID\}$. Given the speakers sequence $S = \{s_1, ..., s_n\}$, the corresponding sequence of roles will be $\varphi(S) = \{\varphi(s_1), ..., \varphi(s_n)\}$. An example of sequence $\varphi(S)$ extracted from a meeting is reported in the following:

...PM, ME, PM, ME, ID, PM, UI, ME, UI, PM, ME, PM, ME, PM...

where it is possible to notice that most of the turns are regulated by the speaker labeled as PM and regular patterns in the sequence appear in the discussion. The sequence S can be modeled using n-grams of roles $p(\varphi(s_n)|\varphi(s_{n-1}), ..., \varphi(s_{n-p}))$, i.e., the probability of the speaker n depends on the roles of the previous p speakers. Thus the probability of S can be written as:

$$p(S) = p(s_1, ..., s_n) = p(\varphi(s_1), ..., \varphi(s_n)) =$$
$$= p(\varphi(s_1), ..., \varphi(s_p)) \prod_{n=p}^{N} p(\varphi(s_n)|\varphi(s_{n-1}), ..., \varphi(s_{n-p})) \quad (2)$$

As done in language modeling, the quality of the n-gram models can be measured computing the perplexity of a separate data set. The investigation here is limited to unigrams, bigrams and trigrams estimated on the training set composed of 98 meetings. The perplexity of the independent test data set (20 meetings) is then reported in Table 1. The experiment shows a large drop in perplexity when moving from unigrams to bigrams. Trigrams marginally improve over bigrams. This reveals that conditioning the role of a given speaker to the role of the previous speaker, produces a large reduction in the speaker sequence perplexity. The most probable n-gram models are those that contain the role labeled as Project Manager (PM), i.e., the speaker that coordinates and moderates the discussion. Those n-gram models will be referred as *speaker role n-gram* and the paper will investigate how this information can be included as prior knowledge in a speaker diarization system.

Table 1. Perplexity of the role sequences on the test data set

	Unigrams	Bigrams	Trigram
Perplexity	4.0	2.9	2.7

3. SPEAKER DIARIZATION WITH ROLES N-GRAM

Speaker Diarization is the task that aims at inferring *who spoke when* in an audio stream; a common approach is based on agglomerative clustering of speech segments based on acoustic similarity. Often the clustering is followed by a Viterbi re-aglinment step that improves and smooths the speaker sequence [14].

This study is based on the state-of-the-art system described in [15] and briefly summarized in the following. At first multiple distant microphones are beam-formed to produce a single enhanced signal using the Beamformit toolkit [16]. Acoustic features consist of 19 MFCC coefficients extracted using a 30ms window shifted by 10ms. After speech/non-speech segmentation and rejection of non-speech regions, the acoustic features $X = \{x_1, \ldots, x_T\}$ are uniformly segmented into short chunks. Speech segments are then clustered together until a stopping criterion based on Information Theory is met (see [15]). This produces an estimate of the number of speakers in the meeting and a partition of the data in clusters, i. e., it associates each acoustic vector x_t to a speaker s. The initial segmentation into speakers is referred as T^*:

$$T^* = \{(t^*_1, \Delta t^*_1, s^*_1), ..., (t^*_N, \Delta t^*_N, s^*_N)\} \quad (3)$$

It can be notice that T^* is an estimate of the actual speaker turns T (see Eq. 1). After clustering, the speaker sequence is re-estimated using an ergodic Hidden Markov Model/Gaussian Mixture Model where each state represents a speaker. The emission probabilities are modeled as GMMs trained using acoustic vectors x_t assigned to speaker s. This step aims at refining the data partition obtained by the agglomerative clustering and improves the speaker segment boundaries [14]. The decoding is performed using a conventional Viterbi algorithm which implements a minimum duration constraint, i. e. the optimal speaker sequence $\mathbf{S}^{opt}$ (and the associated speaking time) is obtained maximizing the following likelihood:

$$\mathbf{S}^{opt} = \arg\max_{\mathbf{S}} \log\ p(X|S) \quad (4)$$

The emission probability $p(x_t|s_t)$ of the acoustic vector x_t conditioned to speaker s_t is a GMM, i.e., $\sum_r w^r_{st} \mathcal{N}(x_t, \mu^r_{s_t}, \Sigma^r_{s_t})$ where $\mathcal{N}(.)$ is the Gaussian pdf; $w^r_{s_t}, \mu^r_{s_t}, \Sigma^r_{s_t}$ are weights, means and covariance matrix corresponding to speaker model s_t. The output of the decoding step is a sequence of speakers with their associated speaking time.

The decoding only depends on the acoustic scores $p(X|S)$ and completely neglects the fact that not all speaker sequences S have the same probability. This new type of information can be included extending the maximization in Eq. 4 as :

$$\mathbf{S}^{opt} = \arg\max_{\mathbf{S}} \log\ p(X|S)p(S) = \arg\max_{\mathbf{S}} \log\ p(X|S)p(\varphi(S)) \quad (5)$$

In other words, the optimal speaker sequence (and the associated speaker time) can be obtained combining the acoustic score $p(X|S)$ together with the probability of a given speaker sequence $p(S)$. The probability $p(S)$ can be estimated from Eq. 2 if the mapping speakers-roles is known.

This is somehow similar to what is done in Automatic Speech Recognition (ASR) where sentences (i. e. word sequences) are recognized combining acoustic information together with linguistic information captured in the language model (n-gram of words). The acoustic score $p(X|S)$ is a pdf (a GMM), while $p(S)$ is a probability. As in ASR, a scale factor and an insertion penalty are introduced to scale those values to comparable ranges. Equation 5 can be solved using a standard Viterbi decoder. Nevertheless the role of each speaker (thus the mapping $\varphi(.)$) must be known before the decoding. In the experiment section, we will consider the case in which this mapping is given by reference or estimated from data.

4. AMI CORPUS EXPERIMENTS

This section compares the proposed method with a conventional diarization system that does not include any information on the speaker sequence. The experiments are run on the 20 meetings that compose the evaluation set. As the AMI meetings contain four participants, the agglomerative clustering stops whenever the number of

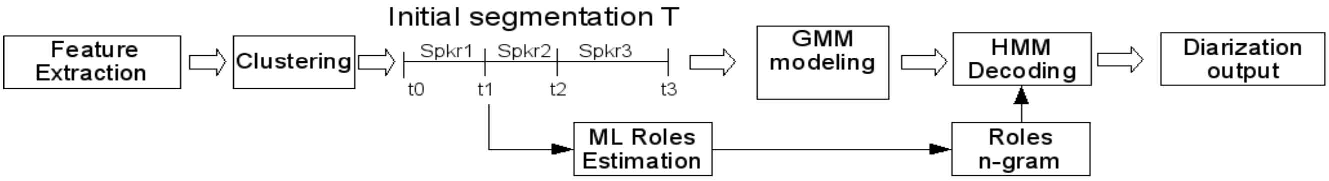

Fig. 1. Schematic representation of the system in case 2 when the speaker roles estimated from data the segmentation T^*

Table 2. Speaker error obtained from the baseline system and the proposed system using unigrams, bigrams and trigrams models on the AMI data. In case 1, the speaker roles are known while in case 2 they are estimated from the data.

Decoding	Case 1	Case 2
no prior	14.4	14.4
unigram	13.8 (+4%)	14.0 (+3%)
bigram	11.8 (+18%)	12.0 (+16%)
trigram	11.5 (+19%)	11.9 (+17%)

Table 3. Speaker error obtained from the baseline system and the proposed system using unigrams, bigrams and trigrams models on Rich Transcription data. Speaker roles are estimated from data.

Decoding	Speaker error
no prior	15.5
unigram	15.0 (+3%)
bigram	13.7 (+11%)
trigram	13.6 (+12%)

actual clusters is equal to four. The most common metric for assessing the diarization performances is the Diarization Error Rate [17] which is composed by speech/non-speech and speaker errors. As the same speech/non-speech segmentation is used across experiments, in the following only the speaker error is reported. Table 2, first row, reports the performance of the baseline system which achieves a speaker error of 14.4%.

The n-gram models that encode the conversational patterns are estimated on the training data composed of 98 meetings. The development set is then used to tune the scale factor and the insertion penalty. The obtained values are evaluated on the independent test set (20 meetings). Let us consider two different cases of increasing difficulty: in the first one, the mapping from speakers to roles $\varphi(.)$ is obtained from an oracle, i.e., from the ground-truth reference, while in the second one it is estimated from the data.

Case 1 Let us assume the mapping $\varphi(.)$ is available from the ground-truth annotation. $P(\varphi(S))$ can be directly estimated using Eq. (2) and included during the Viterbi decoding. Unigram, bigram and trigram models are used in this experiment; for each of those, the language model scale and the insertion penalty are tuned on the separate development data set. Results are reported in Table 2. The use of role n-gram reduces the speaker error by $+4\%$, $+18\%$ and $+19\%$ respectively. The fact that bigram models largely outperform unigrams shows that the improvements are not simply provided by giving more probability to the most common role. The use of trigrams further improves over the bigrams. Those results are consistent with the perplexity measurements presented in Table 1.

Case 2 Let us now consider the case in which the mapping from speakers to roles $\varphi()$ is unknown. An estimate $\varphi^*()$ of this mapping can be obtained from the segmentation T^* (the output of the system before Viterbi realignment) using a simple maximum likelihood estimator:

$$\varphi^* = \arg\max_{\varphi} p(\varphi(s_1^*), ..., \varphi(s_n^*)) = \arg\max_{\varphi} p(\varphi(s_1^*), ..., \varphi(s_p^*)) \prod_{n=p}^{N} p(\varphi(s_n^*)|\varphi(s_{n-1}^*), ..., \varphi(s_{n-p}^*)) \quad (6)$$

The maximization in Eq. 6 is performed exhaustively searching the space of possible mappings speakers-roles, i. e., $\varphi(\{s_h\}) \rightarrow \{PM, UI, ME, ID\}$ and selecting the one that maximize the probability of the speaker sequence S^*. The search space is quite small in this case, making the exhaustive search possible and computationally inexpensive. Approximatively 70% of the speaker time is correctly labeled in terms of roles. The method can be summarized as (see also Figure 1):

1 Perform agglomerative speaker clustering obtaining the initial segmentation in speaker T^*.

2 Estimate the mapping speakers-roles $\varphi^*()$ based on maximization 6.

3 Perform Viterbi decoding combining $p(X|S)$ and $p(\varphi^*(S))$.

Unigram, bigram and trigram models are investigated as before. Results are reported in Table 2. The use of the n-gram models reduce the speaker error of $+3\%$, $+16\%$ and $+17\%$ w.r.t. the conventional diarization system. The degradation with respect to Case 1 (known speaker's roles) is approximatively 2% relative. The per-meeting performance of the two systems is plotted in Figure 3: the proposed technique reduces the speaker error in 18 of the 20 meetings; in two recordings, where the baseline has very high speaker errror, a small degradation in performance is verified. Let us now investigate the differences between the two systems outputs. Figure 2 plots the relative amount of total speaker time correctly attributed to each of the four roles by the baseline diarization and the proposed technique. Those statistics are averaged over the entire test set and normalized dividing by the total speaker time. The largest improvement in performance comes from the time correctly attributed to the speaker labeled as PM. Further analysis shows that the proposed method outperforms the baseline especially on short turns where the acoustic score may not provide enough information to assign the segment to a given speaker.

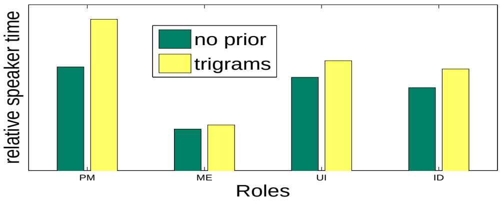

Fig. 2. Relative amount of speaker time correctly attributed to each of the four speakers labeled according to their roles by the baseline diarization and the proposed technique in case 2. Statistics are averaged over the entire test set.

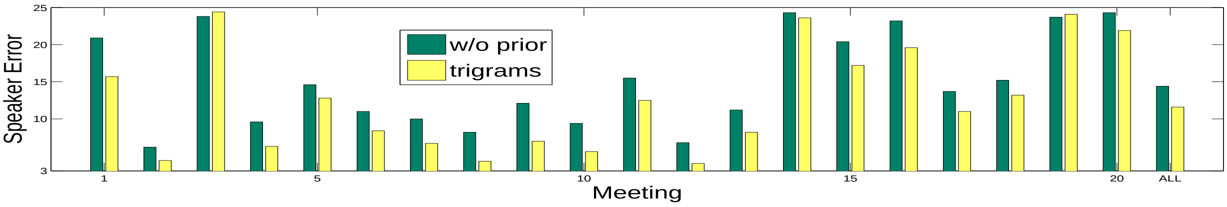

Fig. 3. Per-meeting speaker error for the 20 meetings of the AMI corpus obtained using the baseline system and the proposed system (trigram models - Case 2). Improvements are verified on 18 of the 20 recordings.

5. RICH TRANSCRIPTION EXPERIMENTS

In the previous experiments, the role n-gram models have been estimated and tested on disjoint subsets of the AMI corpus. All the meetings have been elicited using the same scenario, i.e., four participants covering four different roles. In order to investigate how the n-gram models generalize to other types of corpora, the experiments are repeated on a collection of 17 meetings from the Rich Transcription evaluation campaigns 2006 and 2007. In fact all multi-party conversations share common characteristic like the presence of a speaker that moderates the discussion (referred as gate-keeper in the functional role scheme [10],[3]). Those recordings represent spontaneous multi-party conversations collected in five sites. In contrary to the AMI corpus, they are not elicited using a particular scenario.

The number of participants per meeting ranges from 4 to 9 and it is estimated according to a stopping criterion (see [15]). The role of each speaker is obtained using the maximum likelihood estimation in Equation 6. The speakers are thus mapped to one of the four roles PM,ME,UI,ID; the only additional constraint added to the optimization is that only a speaker can be labeled as PM. It is important to notice that the n-gram models are those estimated on AMI corpus, completely different from the evaluation data.

Results are reported in Table 3. The use of the speaker role n-gram reduces the speaker error by 3%, 11% and 12% respectively in case of unigram, bigram and trigram. The improvements are verified on 15 of the 17 recordings thus the n-gram are able to generalize across datasets. However the relative reduction is smaller compared to the AMI corpus.

6. DISCUSSION AND CONCLUSION

Speaker diarization of meetings is typically based on acoustic or directional features and does not consider that meetings are multi-party conversations. This paper investigates whether the information coming from the conversation characteristics can be integrated in a state-of-the-art diarization system.

A number of recent works on meetings data have shown that the way speakers take turn and their roles are closely related and can be statistically modeled [7], [5], [6]. This work studies the use of speaker role n-gram to encode the probability of conversation patterns between speakers. The information is then combined with the acoustic score in the same way the language model is combined with the acoustic score in ASR.

In the first part, the investigation is carried on the AMI corpus annotated in terms of formal roles. Experiments reveal that the speaker error is reduced by $+19\%$ and $+17\%$ respectively when the roles are known or estimated from data. The diarization results are consistent with perplexity measurement. In the second part, the paper investigates how those statistics generalize to a completely different corpus. In fact all multi-party conversations share the same characteristics as for instance the presence of a moderator, i.e, a speaker that mediates the discussion. Meetings from the Rich Transcription campaigns, spontaneous conversations collected in different sites, are used for this purpose. Results reveal that n-gram models estimated on the AMI corpus reduce the speaker error by approximatively 12%. In other words, the role n-gram models generalize to other types of data. It can be noticed that the improvements on RT data are smaller compared to those obtained on the AMI data.

In summary, the speakers sequence in a discussion can be modeled with roles n-grams and this information can be used to reduce the diarization error. In future, this study will be extended considering speaker roles that could potentially generalize better across different conversations like functional roles [3]. Furthermore the use of n-grams will be also be investigated in more complex diarization system which make use of multiple feature streams like MFCC and TDOA.

7. REFERENCES

[1] Sacks H., Schegloff D., and Jefferson G., “A simple systematic for the organization of turn-taking for conversation,” *Language*, , no. 5, 1974.

[2] Hare A.P., “Types of roles in small groups: a bit of history and a current perspective,” *Small Group Research*, vol. 25, 1994.

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

[4] 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.

[5] 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.

[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] Yaman S., Hakkani-Tur D., and Tur G., “Social Role Discovery from Spoken Language using Dynamic Bayesian Networks,” *Proceedings of Interspeech*, 2010.

[8] Grothendieck J et al., “Social correlates of turn-taking behavior.,” *Proceeings of ICASSP 2010*.

[9] 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.

[10] Bales R.F., *Personality and interpersonal behavior*, New York: Holt, Rinehart and Winston, 1970.

[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] Laskowski K., “Modeling norms of turn-taking in multi-party conversation,” in *In proceedings of ACL (Association for Computational Linguistics)*, 2010.

[13] Shriberg E. et al., “Observations on overlap: Findings and implications for automatic processing of multi-party conversation,” in *in Proceedings of Eurospeech 2001*, 2001, pp. 1359–1362.

[14] 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.

[15] Vijayasenan D. at al., “An information theoretic approach to speaker diarization of meeting data,” *IEEE Transactions on Audio Speech and Language Processing*, vol. 17, no. 7, pp. 1382–1393, September 2009.

[16] X. Anguera, “Beamformit, the fast and robust acoustic beamformer,” in *http://www.icsi.berkeley.edu/x̃anguera/BeamformIt*, 2006.

[17] “http://www.itl.nist.gov/iad/mig/tests/rt/,” .