Speaker Diarization of Meetings based on Speaker Role N-gram Models

Fabio Valente, Deepu Vijayasenan, Petr Motlicek

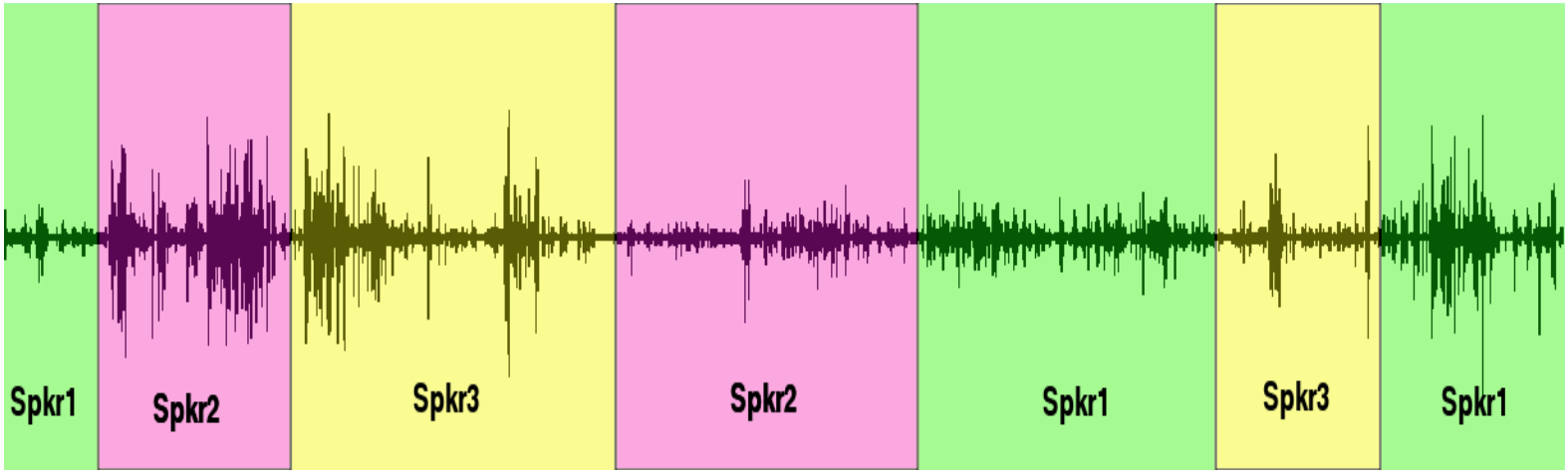
fabio.valente,deepu.vijayasenan,petr.motlicek}@idiap.ch.

Idiap Research Institute

ICASSP 2011

Introduction and Motivation

- Speaker Diarization determines $who\ spoke\ when$ in an audio stream.
- Recent application includes meetings data, spontaneous conversations recorded with Multiple Distant Microphones (MDM).
- Most of the recent efforts have focused on signal processing or statistical methods, e.g., TDOA features, multi-stream modeling, etc.
- Typical speaker diarization methods for meetings ignores the fact that data are instances of human conversations.



Conversation Analysis

- Conversation analysis has been an active research field for long time [Sacks74] but only recently several works have focused on statistical modeling of phenomena in conversations.
- "while appearing unconstrained and spontaneous, [Conversations] are governed by principles and laws and give rise to ordered and predictable behavioral patterns" [Bengt83]
- Roles are $stable\ behavioural\ patterns$ that speakers exhibit during the conversations and influence the way people take-turns in the conversation.
- Terminology $roles$ can be referred to:

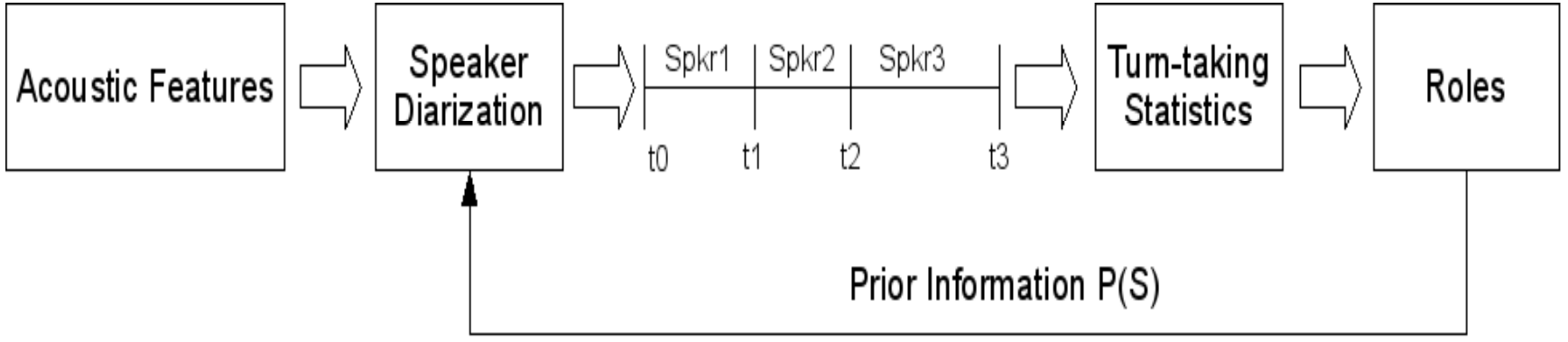
1 Formal roles: the chairperson in a meeting or the moderator in a debate.

2 Functional roles: the function that each speaker has in a spontaneous conversation, e.g., Information provider, Information seeker, Orienter, etc.

3 Social roles, e.g., the way each speaker relates to others in the discussion, e.g., Progatonist, Supporter, Gatekeeper, etc.

Conversation Analysis

- Automatic conversations analysis and especially *role recognition* are often performed using statistics from conversations like turn-taking patterns, turn duration, total speaking time.
- Several meeting corpora have been used for this: CMU meetings [Banerjee04], AMI meetings [Vinciarelli08], ICSI meetings [Laskowski08] and also Broadcast conversations [Yeman10].
- Those statistics are often automatically extracted with a *speaker diarization* system; *here we aim at using those statistics back as prior information into the speaker diarization.*



Data Set

- AMI meeting database, 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).
- 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 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).

Data Labeling

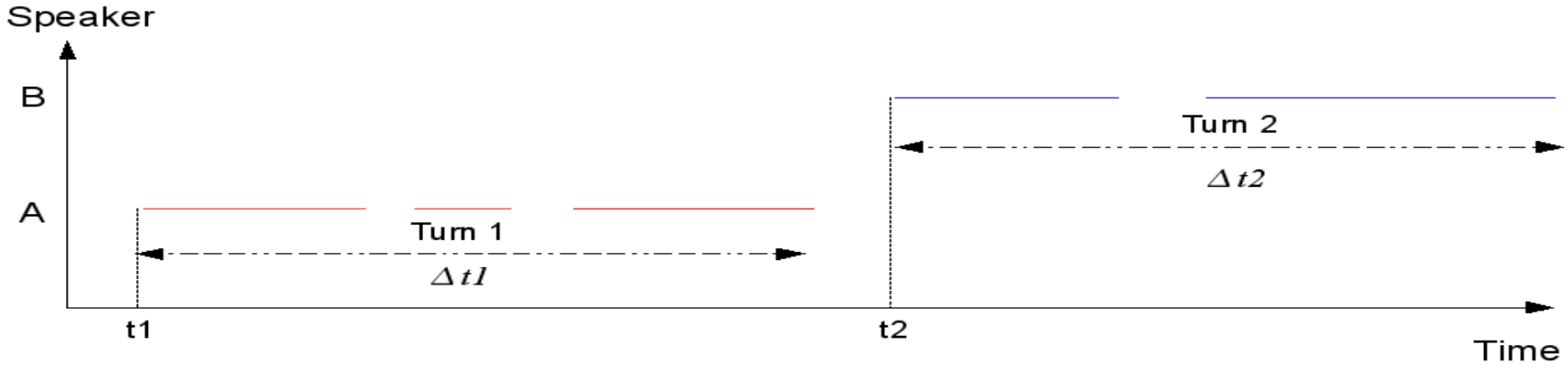
- A meeting is a sequence of speaker turns. Simplified turn definition: speech regions uninterrupted by pauses longer than 300 ms [Shriberg01].
- To further simplify the problem, the time in overlapping regions is given to the floor holder.
- Meeting $T = \{(t_1, \Delta t_1, s_1, r_1), ..., (t_N, \Delta t_N, s_N, r_n)\}$ 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.

  r_n is the role associated with the turn.



Role N-Grams

- Let $S = \{s_1, ..., s_n\}$ be the speaker sequence associated with speaker turns.
- Let $\varphi(S) \rightarrow R$ be the one-to-one mapping between the four speakers and the four roles $R = \{PM, UI, ME, ID\}$.
- The corresponding sequence of roles will be $\varphi(S) = \{\varphi(s_1), ..., \varphi(s_n)\}$.
- 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

$$p(S) = p(s_1, ..., s_n) = p(\varphi(s_1), ..., \varphi(s_n)) =$$

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

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

	Unigrams	Bigrams	Trigram
Perplexity	4.0	2.9	2.7

Diarization with roles N-gram

- Similarly to ASR, the proposed method combines acoustic score (speaker models) with language model score (roles n-gram).
- Diarization method is based on the Information Bottleneck principle [Vijayasenan09] here summarized:

  [1] Speech/non-speech and initial segmentation of the audio $X = \{x_1, \ldots, x_T\}$.

  [2] Information theoretic clustering of those segments which produce an initial clustering/segmentation into speakers is referred as T^*.

  [3] Realignment of the speaker boundaries using an HMM/GMM system and Viterbi decoding which gives the final speaker sequence $\mathbf{S}^{opt}$.

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

- At the moment, no prior information on the speaker sequence S is used.
- Only attempts from [Han2009] made use of $meeting\ dependent$ patterns between speakers.

Diarization with roles N-gram

- $\mathbf{Case\ 1}$: the speaker roles are known, i.e., the mapping speaker-to-roles $\varphi(.)$ is available.
- The prior probability of a speaker sequence $p(S)$ can be computed as $p(\varphi(S))$ and directly integrated in the decoding:

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

- $p(X|S)$ is a pdf.
- $p(S)$ is a probability.
- Similarly to ASR a scaling factor and an insertion penalty are introduced
- They are tuned on the independent development data set in order to minimize the speaker error.

Diarization with roles N-gram

- **Case 2** the mapping speaker-to-roles $\varphi(.)$ is unknown and must be estimated from the diarization output (before realignment).

- A simple Maximum Likelihood estimation $\varphi^*(.)$ can be used:

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

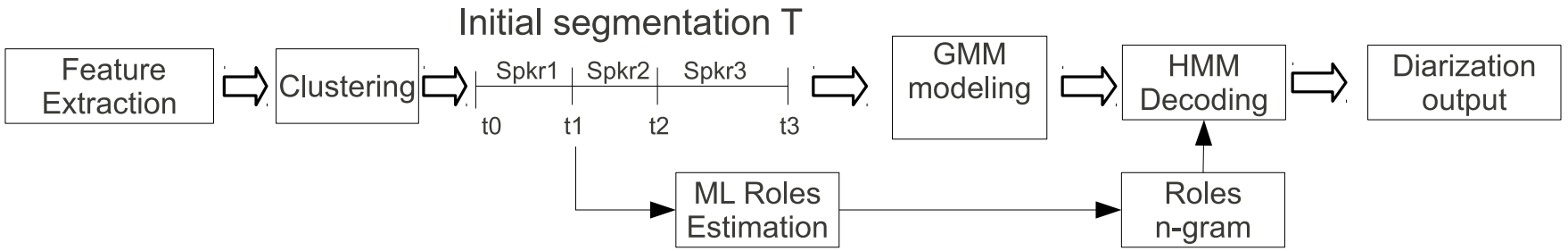
- The decoding becomes:

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

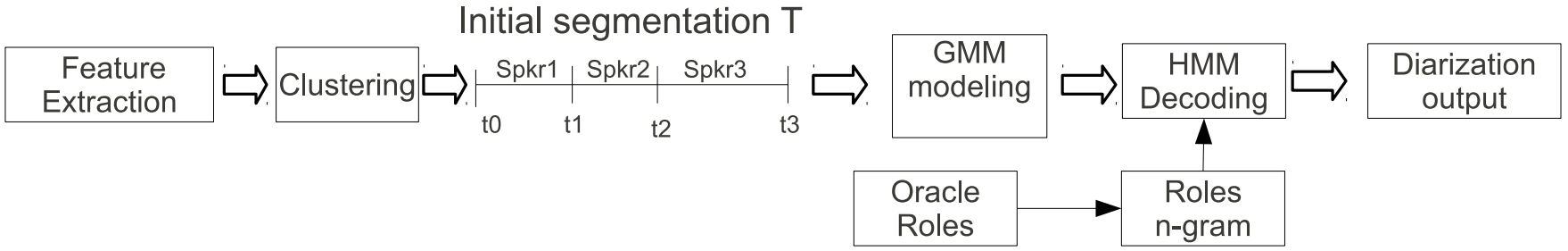
- As before scaling factor and an insertion penalty are introduced and tuned on a separate data set.

Diarization with roles N-gram

- Diarization with known (oracle) roles.



- Diarization with estimated roles.



Experiments AMI

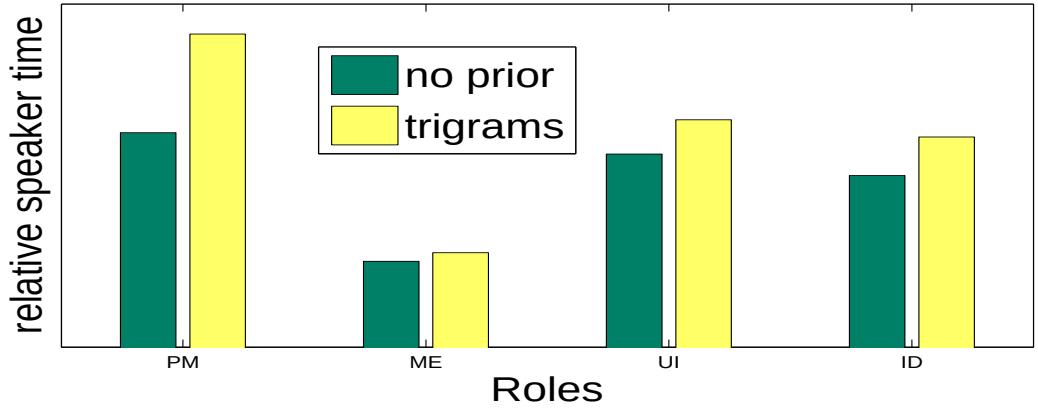
- Evaluation on 20 AMI meetings: 4 speakers and 4 roles: the clustering is forced to converge to 4 speakers.
- Results based on the NIST Diarization Error Rate : speech/non-speech + speaker errors.
- As the same speech/non-speech segmentation is used across experiments, in the following only the speaker error is reported.

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%)

- The largest improvement is obtained with known roles and trigrams consistently with the perplexity measurements.

Experiments AMI

- Relative amount of speaker time correctly attributed to each of the four speakers labeled according to their roles by the baseline diarization.



- The largest improvement in performance comes from the time correctly attributed to the speaker labeled as PM.
- 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
- Does those statistics generalize to other corpora ?

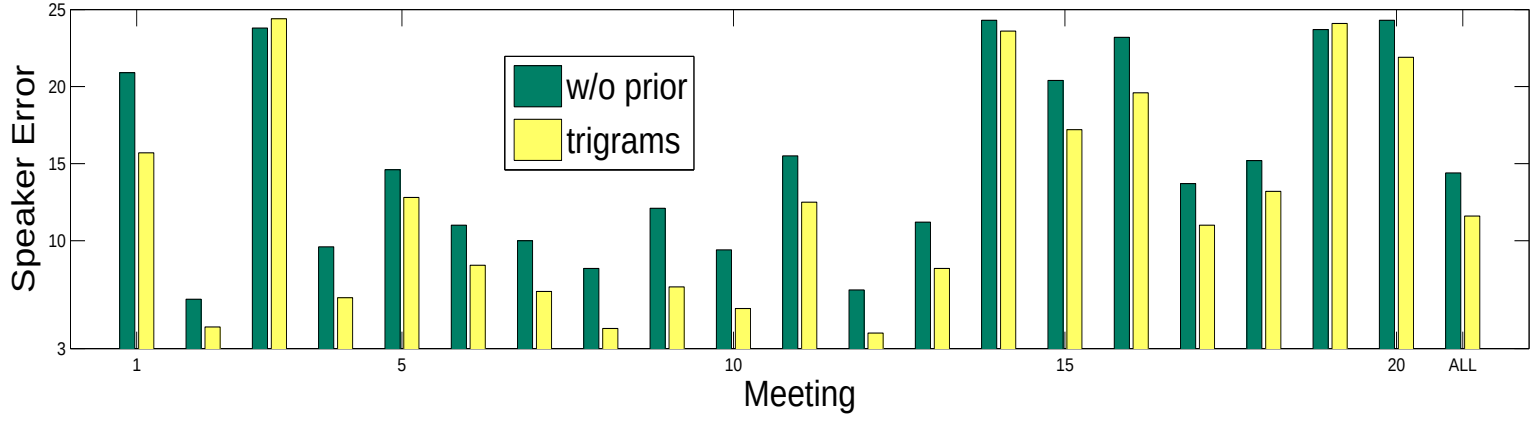
Experiments RT

- The experiments are repeated on a collection of 17 meetings from the Rich Transcription (RT) evaluation campaigns.
- Multiple Distant Microphone conditions (MDM), beam-formed to produce a single enhanced speech signal.
- The rationale is that all multi-party conversations share common characteristic like the presence of a speaker that moderates the discussion (referred as Gate-keeper in the social role scheme).
- A mapping between generic speakers and AMI roles is obtained as before enforcing the constraint that only a speaker can be labeled as Program Manager.

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

Experiments RT

- The number of participants per meeting ranges from 4 to 9 and it is estimated according to a stopping criterion
- Statistics seem to generalize well even if smaller improvements are obtained.
- Improvements are verified on 15 of the 17 the recordings.
- In two recordings a small degradation is verified.



Conclusions

- 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.
- Role n-gram are proposed to encode the probability of conversation patterns between speakers.
- Experiments reveal that the speaker error is reduced by $+19\%$ and $+17\%$ respectively when the roles are known or estimated from data.
- N-gram models estimated on the AMI corpus reduce the speaker error by approximatively $+12\%$ thus generalize to other types of data.
- In future, this study will be extended considering speaker roles that could potentially generalize better across different conversations like functional roles.

Thank You