Automatic Role Recognition Based on Conversational and Prosodic Behaviour

Hugues Salamin[1]
Alessandro Vinciarelli[1,3]
[1]University of Glasgow
G12 8QQ Glasgow (UK)
hsalamin@dcs.gla.ac.uk
vincia@dcs.gla.ac.uk

Khiet Truong[2]
[2]EEMCS, University of Twente
Drienerlolaan 5 - 7522 NB
Enschede (NL)
k.p.truong@ewi.utwente.nl

Gelareh Mohammadi[3,4]
[3]Idiap Research Institute
1920 Martigny (CH)
[4]EPFL - 1015 Lausanne (CH)
gmohamma@idiap.ch

ABSTRACT

This paper proposes an approach for the automatic recognition of roles in settings like news and talk-shows, where roles correspond to specific functions like Anchorman, Guest or Interview Participant. The approach is based on purely nonverbal vocal behavioral cues, including who talks when and how much (turn-taking behavior), and statistical properties of pitch, formants, energy and speaking rate (prosodic behavior). The experiments have been performed over a corpus of around 50 hours of broadcast material and the accuracy, percentage of time correctly labeled in terms of role, is up to 89%. Both turn-taking and prosodic behavior lead to satisfactory results. Furthermore, on one database, their combination leads to a statistically significant improvement.

Categories and Subject Descriptors: H.3.1 [Content Analysis and Indexing].**General Terms:** Experimentation. **Keywords:** Role Recognition, Conditional Random Fields, Multiparty Recordings, Broadcast Data.

1. INTRODUCTION

One of the most common phenomena psychologists observe in social interactions is that people play *roles*, i.e. they display predictable behavioral patterns perceived by others as addressing needs or fulfilling functions in a given interaction setting [10]. Thus, it is not surprising that the computing community has paid significant attention to the automatic recognition of roles, in particular with approaches based on analysis and understanding of nonverbal behavior [14].

This paper proposes an approach for the recognition of roles in formal settings (news and talk-shows) based on turn-taking and prosodic behavior. Turn-taking accounts for who talks when and how much and provides a description of how each person participates in a conversation. Prosodic behavior accounts for the way people talk, i.e. their pitch, loudness and speaking rate. The approach includes three main steps (see Figure 1). The first is the segmentation of the data into turns, time intervals during which only one person is talking. The second is the extraction of turn-taking and prosodic features from each turn. The third is the mapping of the feature vectors extracted from each turn into a sequence of roles with Conditional Random Fields.

The main novelty of this work with respect to the state-of-the-art is that, to the best of our knowledge, this is the first attempt where prosodic features are applied to role recognition. The performances achieved seem to confirm that people playing different roles display different prosodic behaviors, that is they exhibit specific ways of speaking. Furthermore, this is the first work, to the best of our knowledge, where prosodic and turn-taking behavior are combined to provide a full description of nonverbal vocal behavior in conversations. With respect to previous work of the authors in the same domain [12], the main novelty is not only the use of prosodic behavior, but also that the role assignment is performed for each turn rather than for each person. This is a major improvement because it ensures that the same person can play different roles in the same interaction and that role assignment can be performed even if only part of the interaction is actually available.

The results show that both prosodic and turn-taking behavior, when used individually, achieve satisfactory performances (up to 89% accuracy). Moreover, the combination of the two leads to a statistically significant improvement with respect to the best individual performance on one of the databases.

Role recognition is interesting not only from a social interaction analysis point of view [14], but also in an application perspective. Roles can enrich the description of multiparty recordings for indexing and retrieval purposes, can be used in summarization systems to detect interventions more likely to contain important information, or can support browsing systems by allowing a user to quickly identify turns associated to a role of interest.

The rest of the paper is organized as follows: Section 2 proposes a survey of related works, Section 3 describes the proposed approach, Section 4 describes experiments and results, and Section 5 draws some conclusions.

2. RELATED WORK

Two main approaches have been used for the recognition of roles, the analysis of turn-taking, and the modeling of lexical choices. In a few cases, the two approaches have been

MM'10, October 25–29, 2010, Firenze, Italy.

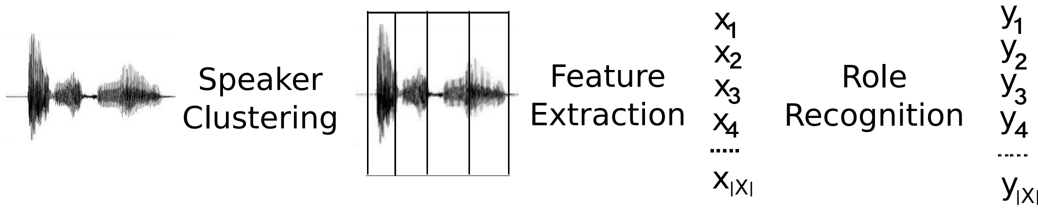

Figure 1: The figure depicts the role recognition approach presented in this work: The audio data is first segmented into turns (single speaker intervals), then converted into a sequence of feature vectors and mapped into a sequence of roles.

combined and some works propose movement based features (fidgeting) as well, resulting into multimodal approaches based on both audio and video analysis. Turn-taking has been used in [12, 13], where temporal proximity of speakers is used to build social networks and extract features fed to Bayesian classifiers based on discrete distributions. Temporal proximity, and duration of interventions, are used in [3, 6, 7, 11] as well, where they are combined with the distribution of words in speech transcriptions. Role recognition is based on BoosTexter (a text categorization approach) in [3], on the combination of Bayesian classifiers (working on turn-taking) and Support Vector Machines (working on term distributions) in [7], and on probabilistic sequential approaches (Hidden Markov Models and Maximum Entropy Classifiers) in [6, 11]. An approach based on C4.5 decision trees and empirical features (number of speaker changes, number of speakers talking in a given time interval, number of overlapping speech intervals, etc.) is proposed in [2]. A similar approach is proposed in [9], where the features are the probability of starting speaking when everybody is silent or when someone else is speaking, and role recognition is performed with a Bayesian classifier based on Gaussian distributions. The only multimodal approaches are proposed in [5, 15], where features accounting for speaking activity and fidgeting are recognized using Support Vector Machines first [15], replaced then with influence models to exploit dependencies across roles [5]. Even if they use fidgeting features, these two works still suggest that audio-based features are the most effective for the recognition of roles.

3. THE APPROACH

The overall approach is depicted in Figure 1. The input data is the audio recording of a multiparty conversation and the first step is the segmentation into turns via a speaker clustering approach (the technique applied in the experiments is fully described in [1] and does not represent the main element of novelty of this paper). The rest of the process includes the feature extraction applied to each turn and the mapping of the resulting observations into roles.

3.1 Feature extraction

From each turn, two types of features are extracted, turn-taking and prosody related, respectively. The former are expected to account for who talks when and how much, the latter for how people talk during their interventions.

Turn-taking related features include duration of current turn (in seconds), total number of turns of the current speaker, time from the beginning of the recording to first turn of the current speaker (in seconds), time after last turn of the current speaker (in seconds), average time between turns of the current speaker (in seconds), time from previous to current turn of the current speaker (in seconds), and number of unique speakers in the N-neighboring turns. All of these features have already been applied in the role recognition literature and they have been shown to be effective. The features are clearly non-independent, but this is not a problem because Conditional Random Fields (see below) do not make any assumption about the independence of the observations.

The extraction of prosody related features includes two steps. The first is the extraction of the low-level features, and the second is the extraction of the turn-level features. Low-level features include pitch, formants, energy and segmentation into voiced and unvoiced intervals, i.e. segments during which there is emission of voice or not, respectively. The extraction of the low-level features is performed with Praat [4], one of the most commonly applied tools in speech analysis. Low-level features are extracted from 30 ms long segments at regular time steps of 10 ms. Thus, low-level features account only for short-term phenomena and are not suitable in their raw form to represent turns that can last from several seconds up to minutes.

The approach applied to address the above problem is to extract turn-level features, i.e. statistics accounting for the distribution of the low-level features on the scale of a turn. In this work, the statistics correspond to the entropy of the low-level features. If f is a low-level feature, the entropy is estimated as follows:

$$H(f) = \frac{\sum_{i=1}^{|F|} p(f_i) \log p(f_i)}{\log |F|} \qquad (1)$$

where $F = \{f_1, \ldots, f_{|F|}\}$ is the set of f values observed in a turn, $|F|$ is the cardinality of F, and f corresponds to one of the low-level features mentioned above. The turn-level features are expected to capture the variability of each low-level feature, the higher the entropy, the higher the number of f values represented a large number of times during the turn and vice-versa.

The turn-level features are not extracted from the whole turn, but from a fraction of the turn centered in its middle and with length corresponding to 90% of the total turn length. The reason is that the speaker clustering process is affected by errors and the turn boundaries are not detected correctly. Thus initial and final part of the turn might include noise.

Corpus	AM	SA	GT	IP	HR	WM
C1	41.2%	5.5%	34.8%	4.0%	7.1%	6.3%
C2	17.3%	10.3%	64.9%	0.0%	4.0%	1.7%

Table 1: Percentage of time each role accounts for in C1 and C2.

3.2 Role Recognition

The role recognition step is performed by labeling the sequence of observations $X = \{x_1, \ldots, x_N\}$ (x_t is the observation vector extracted from turn t and N is the number of turns) with a Conditional Random Field (CRF) [8]. This corresponds to finding the sequence of roles Y^* satisfying the following expression:

$$Y^* = \arg\max_{Y \in \mathcal{Y}} \mathrm{P}(Y \mid X, \alpha) \tag{2}$$

$$= \arg\max_{Y \in \mathcal{Y}} \frac{\exp\left\{\sum_i \alpha_i g_i(X, Y)\right\}}{Z(X)} \tag{3}$$

where the $g_i(X, Y)$ are called feature functions, $Z(X)$ is a normalization constant depending on X, the α_i are the model parameters , $\mathcal{Y}$ is the set of all possible sequences Y, and $Y = \{y_1, \ldots, y_N\}$ is the sequence of roles (y_t is the role assigned to the person talking at turn t). The experiments of this work use a linear chain CRF. This model corresponding to the assumption that two labels are conditionally independent given the observations and one label between them.

Training a CRF boils down to finding the vector α satisfying the following equation:

$$\hat{\alpha} = \arg\max_{\alpha} \sum_j \log \mathrm{P}(Y_j \mid X_j, \alpha) - \frac{\|\alpha\|_2}{\sigma^2} \tag{4}$$

where X_j and Y_j are training sequences, and the second element of the difference is a regularization term (σ is an hyperparameter to be set via cross-validation) aimed at avoiding overfitting (its expression is based on the assumption that the α_i follow a normal distribution). The maximization of the right hand side of the above equation is performed using gradient ascent.

The functions $g_i(X, Y)$ are of two types:

$$g_{r,i}(y_t, x_t) = \begin{cases} x_{t,i} & \text{if } y_t = r \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$g_{r_1,r_2}(y_t, y_{t-1}) = \begin{cases} 1 & \text{if } y_t = r_1 \text{ and } y_{t-1} = r_2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The functions of the first type capture the association between roles and feature values, the functions of the second type capture the dependency of the roles between adjacent turns.

4. EXPERIMENTS AND RESULTS

The experiments have been performed over two corpora, referred to as C1 and C2, containing 96 news bulletins (19 hours in total) and 27 talk-shows (27 hours in total), respectively. The set of roles is the same for both corpora and it includes the Anchorman (AM), the Second Anchorman (SA), the guest (GT), the Interview Participant (IP), the Weather Man (WM), and the Headline Reader (HR).

Corpus	P	T	PT
C1 (A)	83.0%	89.7%	89.3%
C2 (A)	69.5%	84.2%	**87.0%**
C1+C2 (A)	68.1%	86.4%	86.7%
C1 (M)	87.1%	99.1%	99.1%
C2 (M)	76.2%	96.9%	96.2%
C1+C2 (M)	75.8%	96.6%	96.5%

Table 2: Results. This table reports the recognition results, A stands for "*automatic*" (results obtained over the output of the speaker clustering, M for "*manual*" (results obtained over the groundtruth speaker segmentation), P for prosody, T for turn-taking, $P+T$ for the combination of prosody and turn-taking. The value typed in bold corresponds to a statistically significant improvement of P+T with respect to P and T.

However, the distribution of the roles is different in the two corpora (see Table 1) and, even if the roles have the same name, they do not correspond exactly to the same function (e.g., the anchorman is expected to inform in the news and to entertain in the talk shows). The experiments are performed using a k-fold approach ($k = 5$), each corpus has been split into k subsets of equal size and $k-1$ of them have been used for training while the k^{th} one has been left out for test. The experiment has been repeated leaving out for test each of the k partitions. In this way, it is possible to test the approach over the whole corpus while keeping a rigorous separation between training and test set.

The experiments have been performed not only on C1 and C2 separately, but also on their union. In this last case, the role IP has been converted into GT because C2 does not include people playing the IP role (see Table 1).

The accuracy, percentage of data time in the test set correctly labeled in terms of role, is reported in Table 2 for the different experiments. The results are shown for both automatic and manual speaker segmentation. In the first case, the system works over the output of the speaker clustering system described in Section 3, in the second case, the system works over the groundtruth speaker segmentation. This allows one to assess the effect of the speaker clustering errors that corresponds, on average, to roughly 10% decrease of the performance. The reason is that, each time there is a speaker change, the speaker clustering approach takes $1-2$ seconds to switch speaker. The accumulation of this error over all turns amounts to roughly 10% of the data time in the different corpora.

The two types of features work to a satisfactory extent when they are applied separately. On the manual segmentation, their combination does not lead to statistically significant changes. The main reason is probably that the performance of the turn-taking features (close to 100%) is too high to leave an actual margin for improvement. On the automatic segmentation, the combination leads to a statistically significant (p-value < 0.05 measured using the Kolmogorov-Smirnov test) improvement of the performance on C2. On C1, the performance of the turn taking features is already very high and the remaining error is mainly due to the small delays between actual and detected speaker changes. This source of error can be eliminated only by improving the speaker clustering approach and not by working on the features or the role modeling.

In several cases, it has not been possible to extract all the features for a turn. This applies, e.g., to turns that are too short ($2 - 3$ seconds) to extract a meaningful distribution of prosodic features, or to turns that are too close to the boundaries to count the number of speakers in the N-neighboring turns (see Section 3). The missing values have been set to mean of the corresponding feature over the training set. This seems not to affect the performance of the model and represents a good approach to deal with missing data, at least in the case of these experiments.

The approach has been tested on the union of C1 and C2 to assess its robustness with respect to the presence of multiple settings in the data. The results show that the performance is comparable to that obtained over the two corpora separately. Thus, the approach seems actually to deal effectively with different settings at the same time.

5. CONCLUSIONS

This paper has proposed an approach for automatic role recognition based on turn-taking and prosodic behavior. To the best of our knowledge, this is the first work showing that roles, at least in the settings considered, are associated to specific ways of speaking corresponding to different regions of the prosodic features space. Furthermore, the experiments show that, in some cases, the combination of turn-taking and prosodic features improves to a statistically significant extent the performance. The recognition step is performed with linear chain CRFs where the feature functions allow one to capture relationships between roles and observation values or between roles following one another in the turn sequences.

The main source of error in the automatic case is the speaker clustering. The delay between the actual and detected speaker changes results into an accuracy loss of more than 10% that can be eliminated only by obtaining a better speaker segmentation. This means that further progress on role modeling can be obtained only working on other, possibly more spontaneous data, and roles that are less constrained than those considered in this work and possibly relevant to more general human-human interaction scenarii, like, e.g., those described in general theories of social interaction [10]. This might help to identify better directions for the improvement of the models such as the use of kernels exploiting the correlations between features.

Acknowledgments.

This work is supported by the Swiss National Science Foundation (under the NCCR on Interactive Multimodal Information Management), by the European Community's Seventh Framework Programme (FP77 2007-2013), under grant agreement no. 231287 (SSPNet), and by the Scottish Information and Computer Science Alliance (SICSA).

6. REFERENCES

[1] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 411–416, 2003.

[2] S. Banerjee and A. Rudnicky. Using simple speech based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of International Conference on Spoken Language Processing*, pages 221–231, 2004.

[3] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind the roles: identifying speaker roles in radio broadcasts. In *Proceedings of the 17^{th} National Conference on Artificial Intelligence*, pages 679–684, 2000.

[4] P. Boersma and D. Weenink. Praat: doing phonetics by computer, 2005.

[5] W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, pages 271–278, November 2007.

[6] S. Favre, A. Dielmann, and A. Vinciarelli. Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models. In *Proceedings of ACM International Conference on Multimedia*, 2009.

[7] N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür, and A. Vinciarelli. Role recognition for meeting participants: an approach based on lexical information and Social Network Analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 693–696, 2008.

[8] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data, 2001.

[9] K. Laskowski, M. Ostendorf, and T. Schultz. Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In *In proceedings of the 9^{th} ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, pages 148–155, June 2008.

[10] J. Levine and R. Moreland. Progress in small group research. *Annual review of psychology*, 41(1):585–634, 1990.

[11] Y. Liu. Initial study on automatic identification of speaker role in broadcast news speech. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 81–84, June 2006.

[12] H. Salamin, S. Favre, and A. Vinciarelli. Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia, to appear*, 11(7):1373–1380, 2009.

[13] A. Vinciarelli. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(9):1215–1226, 2007.

[14] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.

[15] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proceedings of International Conference on Mutlimodal Interfaces*, pages 47–54, 2006.