Interspeech 2018
2-6 September 2018, Hyderabad, India

Combined Speaker Clustering and Role Recognition in Conversational Speech

Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson, Shrikanth Narayanan

Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

flemotom@usc.edu, ppapadop@usc.edu, jjgibson@usc.edu, shri@sipi.usc.edu

Abstract

Speaker Role Recognition (SRR) is usually addressed either as an independent classification task, or as a subsequent step after a speaker clustering module. However, the first approach does not take speaker-specific variabilities into account, while the second one results in error propagation. In this work we propose the integration of an audio-based speaker clustering algorithm with a language-aided role recognizer into a meta-classifier which takes both modalities into account. That way, we can treat separately any speaker-specific and role-specific characteristics before combining the relevant information together. The method is evaluated on two corpora of different conditions with interactions between a clinician and a patient and it is shown that it yields superior results for the SRR task.

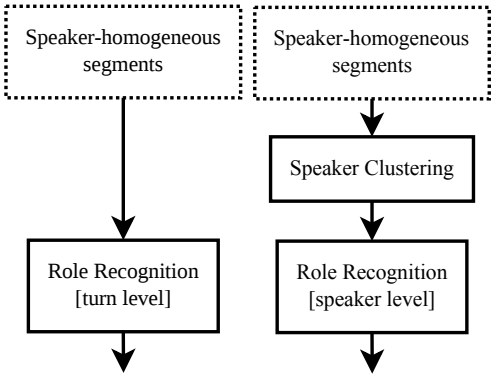
Index Terms: speaker role recognition, speaker clustering, multimodal classification, meta-classifier

1. Introduction

Speaker Role Recognition (SRR) is the task of assigning a specific role to each speaker turn (speaker-homogeneous segment) in a speech signal. This task plays a significant role in numerous areas, such as information retrieval [1], audio indexing [2], or social interaction analysis [3]. Most of the research efforts have been focused on identifying roles in broadcast news programs or talk shows [4–7] , while there have been also works dealing with meeting scenarios [8], conferences [9], medical discussions between domain experts [10], and psychotherapy sessions [11]. There have been presented both supervised [1, 7, 12, 13] and unsupervised [9, 14] methods.

The approaches towards dealing with the problem of SRR can be distinguished on the basis of whether the final decision is made at the turn level or the speaker level. In the former case (Figure 1a), a classifier is built where the input space is the space of speaker turns with no speaker information available. In a real-world application, those turns are obtained through a speaker change detection algorithm. The first works in the field use boosting algorithms [1] and statistical methods [1, 15] towards this classification task. In [8] lexical, prosodic, structural, and dialog act information is combined also through boosting algorithms. Audio-based and language-based classifiers are combined in [5] with early or late fusion through a logistic regression model. Finally, deep learning techniques have been more recently applied [13] in order to learn turn embeddings.

In the case of speaker-level SRR (Figure 1b), the classifier is built in two steps, the first being a Speaker Clustering (SC) algorithm, or a diarization system in the more general case, where the turns are grouped into same-speaker clusters in an unsupervised way and then each cluster is assigned a specific role. In this line of work, [16] uses a social network analysis approach taking into consideration relational data across different speakers, while a hierarchical classification system is proposed in [2] and [12]. The effect of various modalities on the final performance of SRR when using boosting algorithms is investigated in [17]. In [18] the authors study the relation between speech spontaneity levels and speaker roles, using a classifier based on boosting methods with decision stumps. Question types are used as features in [4], with results reported both at the speaker and the turn level.



(a) *Turn-level SRR.* (b) *Speaker-level SRR.*

Figure 1: *Two approaches for Speaker Role Recognition.*

In contrast to Speaker Identification (SID), the features to be extracted for SRR have to exploit characteristics that may be shared between different individuals, since the same role can be shared between various speakers. However, knowledge of speaker-specific information can lead to better classification results (e.g. [4]), which is the reason why many SRR related works operate at the speaker level, employing a SC step. A major drawback of this *piped* approach, presented in Figure 1b, is that no matter how good the subsequent classifier is, any potential error in the SC algorithm is propagated and the overall performance is upper-bounded by the performance of the SC module. Thus, it is desirable to effectively combine speaker-specific and role-specific information without such problems.

To that end, the final role recognition decision in [6] is taken at the turn level, but the speaker information, available after a diarization step, is taken into account during the feature extraction. However, that information is only used for the extraction of structural features (such as average time between two turns of the current speaker). Those are combined with turn-level prosodic features and the final classification is made using Conditional Random Fields (CRFs). It is reported that when using oracle speaker segmentation, this combination does not lead to improved results over the independent usage of the two different feature sets. A hybrid hierarchical approach is presented in [19], where the SC output is used to distinguish at the speaker level a specific role from all the others, which are then classified at the turn level. However, this approach has been proposed specifically for application in broadcast news shows, taking into consideration different variabilities between the *anchors* and the *reporters* on the one hand and between the *reporters* and *others* on the other.

10.21437/Interspeech.2018-1654

In this work, we present an alternative generic framework to combine a SC algorithm with a turn-level supervised role classifier, in such a way that both speaker-specific and role-specific information is taken into account for the final decision. We evaluate our method on the binary problem of patient-clinician interactions using manually extracted speaker turns. However, the framework presented is generalizable to an arbitrary number of speakers, under the assumption of the existence of a one-to-one correspondence between speakers and roles in a single speech document.

2. Method

2.1. General Framework

We propose the *combined* architecture presented in Figure 2, where the SC and role recognition modules work in parallel and their output is fed as input to a meta-classifier.

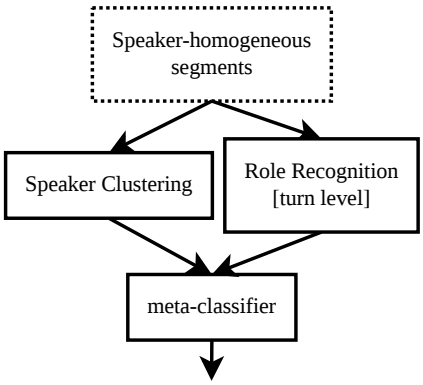


Figure 2: *Proposed approach for Speaker Role Recognition.*

We assume that we know a priori the number of speakers in the speech document, say N, and that there is a one-to-one correspondence between the set of speakers $\{S_i\}_{i=1}^N$ and the set of roles $\{R_i\}_{i=1}^N$. We treat the outputs of the two modules as continuous-valued scores assigned to each speaker/role label. Thus, the output of the SC algorithm is the sequence of tuples $(p_{1i})_{i=1}^N, (p_{2i})_{i=1}^N, \cdots, (p_{Ti})_{i=1}^N$, such that the k-th turn would be assigned the speaker label S_m iff $p_{km} = \max_i p_{ki}$. Similarly, the output of the role recognition module is the sequence of tuples $(q_{1i})_{i=1}^N, (q_{2i})_{i=1}^N, \cdots, (q_{Ti})_{i=1}^N$, such that the k-th turn would be assigned the role label R_m iff $q_{km} = \max_i q_{ki}$. In that way, for each turn we have $2N$ scores corresponding to the N speakers/roles. Those are treated as input features for the classifier of the last step of the architecture.

Since there is not a natural correspondence between the two systems outputs, it is necessary to find the optimal matching between the two sets of labels $\{S_i\}_{i=1}^N$ and $\{R_i\}_{i=1}^N$. This is a standard step taking place in the more general case of diarization systems output combination [20, 21] or for the evaluation of speaker clustering performance [22]. For a small N (which is a realistic assumption for conversational settings), it is easy to find this matching in an exhaustive way. Formally, if we denote such a matching as the mapping $M : \{S_i\}_{i=1}^N \rightarrow \{R_i\}_{i=1}^N$, the optimal matching is defined as

$$\hat{M} = \arg\min_M \sum_{k=1}^{T} \mathbb{I}(M(S'_k) \neq R'_k) d_k$$

where $S'_k \in \{S_i\}_{i=1}^N$ and $R'_k \in \{R_i\}_{i=1}^N$ are the labels assigned by the two modules to the k-th turn, $\mathbb{I}(\cdot)$ is the indicator function, d_k is the duration of the turn, and T is the total number of turns in the speech document.

2.2. Speaker Clustering Module

For the speaker clustering module we are using a simple Bayesian Information Criterion (BIC)-based Hierarchical Agglomerative Clustering (HAC) algorithm [23, 24]. At each step of the HAC procedure we are using one Gaussian to model each cluster, so that the distance metric, known as ΔBIC, between two clusters x and y, with n_x, n_y members (frames) and with covariance matrices Σ_x, Σ_y, respectively, is

$$\Delta\text{BIC}(x,y) = \frac{1}{2}\left(n \log|\Sigma| - n_x \log|\Sigma_x| - n_y \log|\Sigma_y|\right) - \lambda \frac{d(d+3)}{4} \log n$$

where $n = n_x + n_y$, Σ is the covariance matrix if we merge the clusters x and y, d is the dimensionality of the feature vector representing each frame, and λ is a penalty factor ($\lambda = 1$ for our experiments). At each step, the pair of clusters with the minimum ΔBIC is being merged.

The speaker clustering in this work is purely based on the acoustic information and as features we are using the 13 first MFCCs for each frame. At the last step, we have one Gaussian modeling each of the N speakers and the needed scores for the turn are the per-frame log-likelihoods with respect to each Gaussian averaged over the voiced frames of the turn. The voiced frames are identified with a Voice Activity Detection (VAD) algorithm, which is also applied at the initial step of the HAC procedure, so that the constructed Gaussians model only the voiced information for each speaker.

2.3. Role Recognition Module

We explore two different approaches for the role recognition module, one language-based and one audio-based.

In order to build a language-based role recognizer to exploit the linguistic patterns that are potentially shared between speakers with the same roles, we are using similar ideas as in the role matching module presented in [11]. Since we treat role recognition as a supervised classification task, we need a role-labeled training set of speaker turns. On that set we train N n-gram Language Models (LMs), one for each role. During the test phase, we evaluate the perplexity of the turn to be classified with respect to all the constructed LMs. The required scores to be used as input to the meta-classifier are the N negative log-perplexities.

Even though we are using the acoustic information in the SC module, we are interested in exploring the hypothesis that the exact same information has a predictive power over roles, apart from speakers. Following a similar idea as in [5], we build an Acoustic Model (AM) for each one of the N roles. The AM for a role is a Gaussian Mixture Model (GMM) fit on the voiced frames of all the turns available in the training set which are labeled with that role. The scores for the turn to be used during the test phase are again, as in the case of the SC algorithm, the N per-frame log-likelihoods with respect to each GMM averaged over the voiced frames of the turn.

3. Datasets

For this work, we are evaluating our proposed method on two different corpora from the psychology domain, featuring inter-

actions between a clinician and a patient. The first corpus is composed of Motivational Interviewing (MI) sessions between a therapist (T) and a client (Cl) collected from six independent clinical trials (ARC, ESPSB, ESB21, CTT, iCHAMP, HMCBI) [25, 26]. We collectively refer to those sessions as the MI corpus. In this study, we use 343 manually transcribed sessions.

The second corpus is comprised of Autism Diagnostic Observation Schedule (ADOS) assessments between a psychologist (P) and a child (Ch) being evaluated for a Pervasive Developmental Disorder (PDD) [27]. In this study, we use 273 manually transcribed sessions, of minimum 2min duration.

There is a limited number of sessions where there are more than two speakers involved. In such cases, we do not take into account any turns not belonging to the clinician/patient for our analysis. Additionally, there is a limited number of non-pure speaker turns, in the sense that the manually annotated boundaries are not optimal and occasionally overlap. We chose to include such turns in the analysis without any preprocessing, since in a real-world setting (e.g. with automatic segmentation) such problems are impossible to completely avoid.

Some descriptive analysis for the two datasets is presented in Table 1. Unfortunately, the exact total number of different clients is not available for the MI dataset. However, under the assumption that it is highly improbable for the same client to visit different therapists in the same study, and having partial information available about the client IDs, we made the train/test split in a way that we are highly confident there is no overlap between speakers. Similarly, the exact total number of psychologists is unknown for the ADOS corpus, but the data are collected from two different clinics (in different cities) and we assume that the same clinician does not work for both. So, the data from one clinic is used for training and from the other for testing.

Table 1: *Descriptive analysis of the corpora used.* mean_dur *and* std_dur *are the mean and standard deviation values of the session duration. By* dur-T/P *and* dur-Cl/Ch *we denote the total duration of all the speaker turns labeled as therapist/psychologist and client/child, respectively. By* #T/P *and* #Cl/Ch *we denote the total number of different therapists/psychologists and clients/children.*

	MI-train	MI-test	ADOS-train	ADOS-test
#sessions	242	101	141	132
mean_dur	27.24min	33.14min	3.67min	3.67min
std_dur	14.40min	17.42min	1.34min	1.65min
dur-T/P	47.30h	26.35h	2.63h	2.52h
dur-Cl/Ch	52.96h	25.87h	2.97h	2.98h
#T/P	123	53	–	–
#Cl/Ch	–	–	89	81

4. Experiments and Results

The two available datasets are split into train and test sets, as explained in Section 3, in a way that, with high confidence, there are not overlapping speakers between the sets, in order to be certain that the trained models indeed capture role-specific and not speaker-specific information. The train set is only used to build the LMs and AMs described in Section 2.3 corresponding to the different roles.

The LMs are 3-gram models trained (and later evaluated) using the SRILM toolkit [28] with manually derived transcriptions of the recordings. In order to ensure a large enough vocabulary that minimizes the unseen words during the test phase, we are interpolating those models with a large background model–namely with the pruned version of the 3-gram model of cantab-TEDLIUM [29]–giving a weight of 0.9 to the domain-specific LM and 0.1 to the background one.

The AMs are diagonal GMMs with 512 components, modeling the frames of turns assigned to each role, where the frames are represented by 13-dimensional MFCCs. During training, we take into consideration only the voiced frames, by applying to the initial speaker turns a simple, energy-based VAD algorithm, as implemented in the Kaldi speech recognition toolkit [30]. The same VAD algorithm is applied during the evaluation, as well as during the SC, as explained in Section 2.2.

As a meta-classifier we are using a binary linear Support Vector Machine (SVM), since we are evaluating on binary problems. All the results are based on a 5-fold cross-validation scheme on the data allocated for testing in each dataset, where, as is the case for the initial train/test split, we are using all the available meta-data information to minimize any possible overlapping of speakers between different folds. The reason we are adopting this approach and do not use the training part of the datasets is that we do not want to pipe data already seen by the AMs and/or LMs to the SVM training.

As the evaluation metric of the SRR we are using the Misclassification Rate (MR), defined as [22]

$$\text{MR} = \frac{\text{\#misclassified frames}}{\text{total \#frames}} = \frac{\sum_k \mathbb{I}(R_k \neq \hat{R}_k) d_k}{\sum_k d_k}$$

where the summation is over all the speaker turns, R_k is the role assigned by the algorithm, $\hat{R}_k$ is the groundtruth role and d_k is the duration of the k-th turn.

In this work we do not report results for the *piped* architecture presented in Figure 1b using an actual classification algorithm as the second step of the pipeline. Instead, in Table 2 we give the best possible result with this architecture when using the SC algorithm that we have described. Using a perfect classification algorithm for the SRR task at the speaker level, which we denote as $\mathcal{R}^\dagger$, the overall error of the system is always lower-bounded by the error of the SC algorithm itself. So, the results reported in the SC+$\mathcal{R}^\dagger$-piped column of the Table are in fact the MRs of the SC algorithm.

Table 2: *Misclassification Rates (%) of the SC algorithm, the language-based recognizer (LM), and the audio-based recognizer (AM), when used independently (*only*) or in a piped (*piped*) or combined (*comb*) architecture for the task of SRR. By $\mathcal{R}^\dagger$ an optimal, 0-error classification algorithm is denoted.*

	SC+$\mathcal{R}^\dagger$ piped	LM only	SC+LM comb	AM only	SC+AM comb	AM+LM comb	SC+AM+LM comb
MI	3.59	9.49	2.76	35.45	3.66	9.17	**2.71**
ADOS	12.67	12.37	7.70	14.03	10.58	8.02	**5.98**

The language-based and audio-based recognizers are evaluated when used independently (LM-only and AM-only) and when used in the *combined* architecture presented in Figure 2

(SC+LM-comb and SC+AM-comb). The results are reported in Table 2. As we can see, the LM-based approach has a strong predictive power for both datasets, revealing differences in the linguistic patterns between a clinical provider and a client or a child with PDD. When this is combined with the SC algorithm which captures the speaker-specific differences in a single session, the results are considerably better, compared not only to the independent classifiers, but also to the *piped* architecture.

On the other hand, the AM approach does not behave in the same manner for the two datasets. As expected, the acoustic characteristics of the children as a whole are different than those of the adult clinicians. This is reflected in the AM-only results for the ADOS data, even though they are still worse than the LM-only ones. This age distinction between the two different groups of speakers does not exist in the MI dataset. So, although it seems from the results that there is some non-negligible acoustic variability between the clinicians and the clients, the performance gap between the LM-only and the AM-only approaches is much bigger for those data. When combined with the SC algorithm the results are substantially better, because the meta-classifier is affected by the more separated scores which are the output of the SC module. This notion of "separability" is visually depicted in Figure 3 where we show how the outputs of the SC, LM, and AM modules are distributed on the plane. It is of high interest that in the case of the ADOS dataset, because of its very special nature, the exact same information (at the feature level) can be used to capture both role-specific and speaker-specific variabilities in a way that if the two modules are combined by our proposed architecture (SC+AM-comb), they can improve the overall performance as if they carried complementary information.

As a final experiment, we combine the outputs of the LM- and the AM-based recognizers, again using the linear SVM as the meta-classifier (AM+LM-comb) and we also combine all the three constructed modules in an extended *combined* architecture (SC+AM+LM-comb). In this latter case the meta-classifier gets $3 \cdot 2$ (in the general case $3N$) inputs for each turn to be classified. We note that the result of the optimal matching between SC and LM was the same as in between SC and AM, so we did not encounter any conflict. When compared to the LM-only and the SC+LM-comb results, the addition of the acoustic-based recognizer in the architecture does not lead to any substantial improvements, as expected, for the MI data, but does improve the performance of the system for the case of the ADOS sessions. Overall, the relative error improvement with our final system which follows the *combined* architecture is 24.5% for the MI data and 52.8% for the ADOS data, when compared to the *piped* architecture with an optimal recognizer.

5. Conclusions and Future Work

In this work we proposed a framework to incorporate speaker-specific and role-specific information for the SRR task, by independently implementing an unsupervised SC algorithm and a supervised turn-level role classifier, the output scores of which are fed to a meta-classifier which gives a turn-level final decision. By evaluating our method using speech signals from dyadic interactions we showed that it yields superior results, compared both to the independent use of turn-level classifiers which do not take speaker-specific variabilities into account and to systems that use speaker-specific information by applying SC as a first step and predicting the output at the speaker level.

One drawback of our methodology is that it requires additional data for the training of the meta-classifier. Moreover, in a real-world scenario, the speaker boundaries, as well as the language-based features, would be extracted, at least at the evaluation phase, from diarization and Automatic Speech Recognition (ASR) outputs.

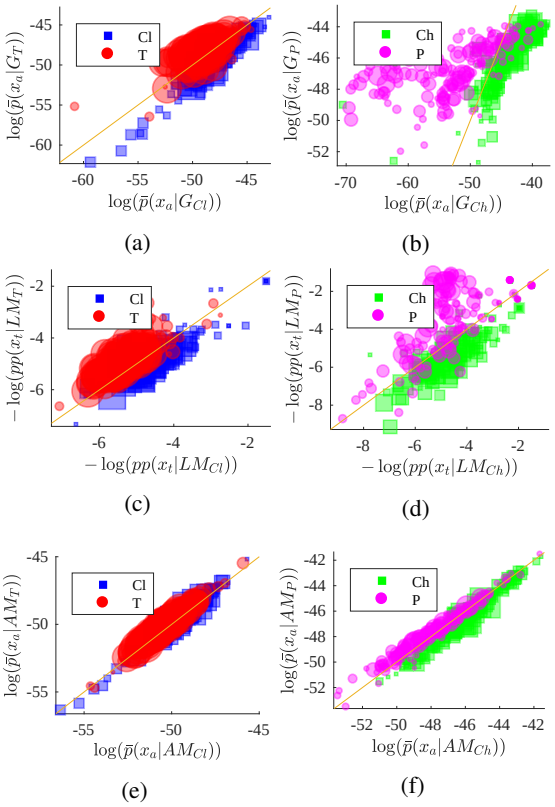


Figure 3: *Distribution of the scores which are the output of the SC ((a),(b)), the LM-based recognizer ((c),(d)) and the AM-based recognizer ((e),(f)) for the MI ((a),(c),(e)) and the ADOS ((b),(d),(f)) datasets. Each data point is a speaker turn with size proportional to the turn length. 300 turns of the test set are randomly shown for each dataset. x_a and x_t are the acoustic and textual representation of a turn x. LM_R and AM_R are the LM and AM corresponding to the role R. G_R is the Gaussian corresponding to the role R at the end of the SC and after an optimal matching between speakers and roles.*

We are planning to apply this framework to multiple-role databases by using multi-class meta-classifiers and to try more sophisticated AMs, LMs, and SC techniques. Additionally, we want to provide a rigid formulation of the framework that can accommodate more than one SC and role recognition modules.

The final goal is to extend the framework in order to combine speaker-specific and role-specific information for speaker segmentation as well so that we can build a fully automatic "role diarization" system.

6. Acknowledgements

NF is supported by the USC Annenberg Fellowship.

7. References

[1] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," in *Proceedings of the 7th National Conference on Articial Intelli-*

gence and 12th Conference on Innovative Applications of Articial Intelligence, 2000, pp. 679–684.

[2] B. Bigot, I. Ferrané, J. Pinquier, and R. André-Obrecht, "Speaker role recognition to help spontaneous conversational speech detection," in *Proceedings of the 2010 international workshop on Searching spontaneous conversational speech*. ACM, 2010, pp. 5–10.

[3] B. J. Biddle, "Recent developments in role theory," *Annual review of sociology*, vol. 12, no. 1, pp. 67–92, 1986.

[4] T. Bazillon, B. Maza, M. Rouvier, F. Bechet, and A. Nasr, "Speaker role recognition using question detection and characterization," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[5] G. Damnati and D. 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.

[6] H. Salamin and A. 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] A. Laurent, N. Camelin, and C. Raymond, "Boosting bonsai trees for efficient features combination: application to speaker role identification," in *Interspeech*, 2014.

[8] A. Sapru and F. Valente, "Automatic speaker role labeling in ami meetings: recognition of formal and social roles," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5057–5060.

[9] Y. Li, Q. Wang, X. Zhang, W. Li, X. Li, J. Yang, X. Feng, Q. Huang, and Q. He, "Unsupervised classification of speaker roles in multi-participant conversational speech," *Computer Speech & Language*, vol. 42, pp. 81–99, 2017.

[10] S. Luz, "Locating case discussion segments in recorded medical team meetings," in *Proceedings of the third workshop on searching spontaneous conversational speech*. ACM, 2009, pp. 21–30.

[11] B. Xiao, C. Huang, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. S. Narayanan, "A technology prototype system for rating therapist empathy from audio recordings in addiction counseling," *PeerJ Computer Science*, vol. 2, p. e59, 2016.

[12] B. Bigot, C. Fredouille, and D. Charlet, "Speaker role recognition on tv broadcast documents," in *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

[13] M. Rouvier, S. Delecraz, B. Favre, M. Bendris, and F. Bechet, "Multimodal embedding fusion for robust speaker role recognition in video broadcast," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 383–389.

[14] B. Hutchinson, B. Zhang, and M. Ostendorf, "Unsupervised broadcast conversation speaker role labeling," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 5322–5325.

[15] 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*. Association for Computational Linguistics, 2006, pp. 81–84.

[16] A. Vinciarelli, "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," *IEEE Transactions on Multimedia*, vol. 9, no. 6, pp. 1215–1226, 2007.

[17] W. Wang, S. Yaman, K. Precoda, and C. Richey, "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5556–5559.

[18] R. Dufour, Y. Esteve, and P. Deléglise, "Investigation of spontaneous speech characterization applied to speaker role recognition," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[19] G. Damnati and D. Charlet, "Robust speaker turn role labeling of tv broadcast news shows," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5684–5687.

[20] S. Tranter, "Two-way cluster voting to improve speaker diarisation performance," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, vol. 1. IEEE, 2005, pp. I–753.

[21] S. Bozonnet, N. Evans, X. Anguera, O. Vinyals, G. Friedland, and C. Fredouille, "System output combination for improved speaker diarization," in *Interspeech*, 2010.

[22] D. Liu and F. Kubala, "Online speaker clustering," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol. 1. IEEE, 2003, pp. I–I.

[23] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA broadcast news transcription and understanding workshop*, vol. 8. Virginia, USA, 1998, pp. 127–132.

[24] S.-S. Cheng and H.-M. Wang, "A sequential metric-based audio segmentation method via the bayesian information criterion," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[25] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification," *Implementation Science*, vol. 9, no. 1, p. 49, 2014.

[26] J. S. Baer, E. A. Wells, D. B. Rosengren, B. Hartzler, B. Beadnell, and C. Dunn, "Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors," *Journal of substance abuse treatment*, vol. 37, no. 2, pp. 191–202, 2009.

[27] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

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

[29] W. Williams, N. Prasad, D. Mrva, T. Ash, and T. Robinson, "Scaling recurrent neural network language models," *CoRR*, vol. abs/1502.00512, 2015. [Online]. Available: http://arxiv.org/abs/1502.00512

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.