DOI: dx.doi.org/10.1109/ICASSP.2019.8683900

ROLE SPECIFIC LATTICE RESCORING FOR SPEAKER ROLE RECOGNITION FROM SPEECH RECOGNITION OUTPUTS

Nikolaos Flemotomos[1], *Panayiotis Georgiou*[1], *David C. Atkins*[2], *Shrikanth Narayanan*[1]

[1] Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA
[2] Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

ABSTRACT

The language patterns followed by different speakers who play specific roles in conversational interactions provide valuable cues for the task of Speaker Role Recognition (SRR). Given the speech signal, existing algorithms typically try to find such patterns in the output of an Automatic Speech Recognition (ASR) system. In this work we propose an alternative way of revealing role-specific linguistic characteristics, by making use of role-specific ASR outputs, which are built by suitably rescoring the lattice produced after a first pass of ASR decoding. That way, we avoid pruning the lattice too early, eliminating the potential risk of information loss.

Index Terms— speaker role recognition, speech recognition, language model, lattice rescoring

1. INTRODUCTION

Speaker Role Recognition (SRR) is defined as the classification task of mapping a speaker-homogeneous segment (speaker turn) to an element of a predefined set of roles, where a role is characterized by the task a speaker performs and the objectives related to it. Typical examples of conversational interactions between individuals with specific roles are in business meetings [1], broadcast news programs [2, 3], psychotherapy sessions [4, 5], or press conferences [6].

In order to address the problem of SRR, appropriate features which capture distinguishable patterns between the different roles have to be extracted. Such patterns can be found in the acoustic [7], lexical [8], prosodic [1], or structural [9, 6] characteristics of the speech signal, with the importance of each modality being task-specific. For instance, it is desired that a psychotherapist speaks less than the client, an interviewer is expected to use more interrogative words than the interviewee, etc. However, it seems that the language often carries the most important information for the problem in hand [10, 2, 1, 5] and is more robust to unseen conditions (e.g. different speakers) [11], which is the reason why a great portion of the research efforts has been focused on studying and exploiting the lexical variability between the speaker roles.

The first efforts in the field extract bags of n-grams to represent the lexical information and use them as input features to boosting algorithms or maximum entropy classifiers [12, 13]. Boosting approaches have been also followed in [10] and [1] to combine n-gram features with other modalities, with the final classification decision taken either at the speaker [10] or at the turn level [1]. In [14] the authors first classify the types of questions posed by the different speakers and use that information for the role assignment. Deep learning approaches have been explored in [11] where word embeddings are used as inputs to convolutional neural networks. In [4] and [5] role-specific n-gram Language Models (LMs) are built and SRR is converted into the problem of finding the LM which minimizes the perplexity of a speaker turn or of all the turns assigned to a specific speaker.

Although a bulk of the aforementioned studies use manually transcribed speech data to perform SRR, in a real-world application the lexical information would become available after an Automatic Speech Recognition (ASR) step [11, 4]. Moreover, in [15] the authors suggest that the quality of ASR transcripts can be used to extract additional features carrying complementary information in specific scenarios. In any case, the ASR output is considered to be the best path of a system that uses generic acoustic and language models.

In this work, we propose using role-specific ASR systems each one of which gives a potentially different output together with a corresponding cost. Then, after passing any given turn through all the systems, we can assign to that turn the role which corresponds to the system producing the minimum cost. In particular, for this study, we create the role-specific systems by rescoring the lattices generated by a generic ASR with role-specific LMs, as explained in Section 2. That way, we can exploit any information carried by the decoding lattice before pruning it to find the best path. Based on similar intuitions, lattice rescoring techniques have been previously explored in [16] and [4] for binary classification problems in the field of behavioral code prediction. Our method is evaluated on dyadic interactions from the clinical domain, as well as on multi-participant business meeting scenarios.

2. METHOD

Given a generic ASR system, the goal is to convert the generated decoding lattice for an input turn to multiple, role-

specific versions, in such a way that there is one version which reflects the speaker role corresponding to the particular turn. We are doing this by rescoring the lattice N times, where N is the number of roles, with role-specific LMs. Let's assume we have a background, out-of-domain n-gram LM $\mathcal{G}$ and N role-specific LMs $\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_N$ corresponding to the roles $R_1, R_2, \cdots, R_N$, which are trained using in-domain data. First, we ensure that all the models which are going to be used recognize the same vocabulary. We can efficiently do so by interpolating the individual LMs to get the mixed models $\mathcal{G}^+, \mathcal{R}_1^+, \mathcal{R}_2^+, \cdots, \mathcal{R}_N^+$ [17]. By using the symbol $\oplus$ to denote LM interpolation, the final models are expressed as

$$\mathcal{G}^+ = w_g \mathcal{G} \oplus (1 - w_g)\tilde{\mathcal{R}} \quad (1)$$

$$\mathcal{R}_i^+ = w_{g_i} \mathcal{G} \oplus w_{r_i} \mathcal{R}_i \oplus (1 - w_{g_i} - w_{r_i})\tilde{\mathcal{R}}_i \quad (2)$$

where

$$\tilde{\mathcal{R}} = \frac{1}{N} \bigoplus_{i=1}^{N} \mathcal{R}_i, \quad \tilde{\mathcal{R}}_i = \frac{1}{N-1} \bigoplus_{\substack{j=1 \\ j \neq i}}^{N} \mathcal{R}_j$$

and all the weights w_g, w_{g_i}, w_{r_i} are chosen to minimize the perplexity of appropriate role-specific development corpora.

Given an input turn x, we first pass it through an ASR system, trained with the LM $\mathcal{G}^+$, producing a decoding lattice $\mathcal{L}_{\mathcal{G}^+}(x)$. The lattice is then rescored with all the LMs $\mathcal{R}_j^+, j = 1, 2, \cdots, N$ to produce the lattices $\mathcal{L}_{\mathcal{R}_j^+}(x)$. Denoting as $c_j(x)$ the LM-cost of the best path in $\mathcal{L}_{\mathcal{R}_j^+}(x)$, the role assigned to x is R_m where $m = \arg\min_j c_j(x)$. The process is visually depicted in Fig. 1.

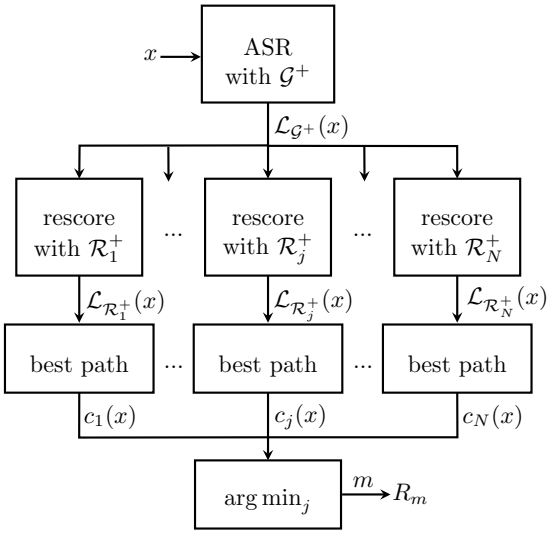

Fig. 1. Turn-level SRR by role-specific lattice rescoring.

The difference between this approach and the language-based approach followed in [5] is that in the second case the evaluation with respect to a role-specific LM would be done using the final output of the ASR, as presented in Fig. 2. That way, the lattice $\mathcal{L}_{\mathcal{G}^+}(x)$ is pruned using a generic LM, which can potentially lead to loss of valuable information for the task of SRR. This is exactly the problem our approach tries to avoid.

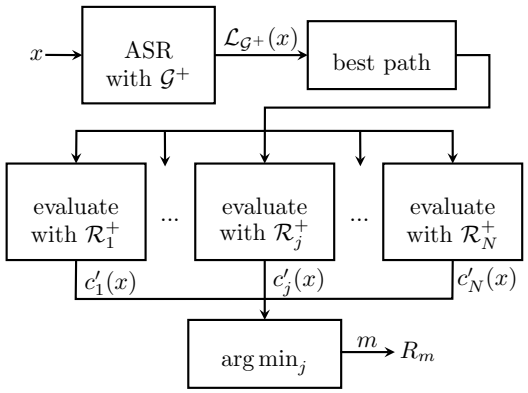

Fig. 2. Turn-level SRR by evaluating the text with role-specific LMs.

If the extra information of the speaker who uttered the turn is available, after a speaker clustering step, then the role assignment can be done more robustly at the speaker level instead of the turn level [14, 5]. If we denote by T_i the set of turns corresponding to the speaker S_i, we can define the cost of the speaker-role pair (S_i, R_j) as $c(S_i|R_j) \triangleq \sum_{x \in T_i} c_j(x)$. Ideally, we would again like to assign to any speaker S_i the role R_m such that the cost $c(S_i|R_m)$ is the minimum among all $c(S_i|R_j), j = 1, 2, \cdots, N$. However, assuming that there is one-to-one correspondence between speakers and roles in a speech document, which is the case for most practical applications, this criterion would fail, since there is no guarantee that $\arg\min_j c(S_n|R_j) \neq \arg\min_j c(S_m|R_j)$ for $n \neq m$.

Thus, in order to take such a constraint into account, we are instead using Algorithm 1, which is a generalization of the role matching criterion proposed in [18] for the 2-speaker scenario, where the costs were perplexities. The algorithm begins with the entire sets $\tilde{S}$ and $\tilde{R}$ of the available speakers and roles and at every iteration it chooses the speaker S_k such that a confidence metric C_k is the maximum among all $C_i, i = 1, 2, \cdots, |\tilde{S}|$. Then, it assigns to S_k the role R_{l_k} that minimizes the cost $c(S_k|R_j), j = 1, 2, \cdots, |\tilde{R}|$ and removes S_k and R_{l_k} from the available speakers and roles. The confidence metric C_i is designed in such a way that the larger the difference between the minimum cost and the rest of the costs for S_i is, the more confident we are about the role assignment of the particular speaker.

3. DATASETS

We evaluate our method on two datasets featuring interactions between individuals under different conditions. The first

Algorithm 1 Speaker-level SRR given costs for each (speaker,role) pair.

Inputs: speakers $S_1, S_2, \cdots, S_N$
roles $R_1, R_2, \cdots, R_N$
costs $c(S_i|R_j) \forall i, j$

$\tilde{S} \leftarrow \{S_i\}_{i=1}^N$; $\tilde{R} \leftarrow \{R_i\}_{i=1}^N$
while $\tilde{S} \neq \phi$ **do**
  for $S_i \in \tilde{S}$ **do**
    $l_i \leftarrow \arg\min_m c(S_i|R_m),\ R_m \in \tilde{R}$
    $C_i \leftarrow \min_n |c(S_i|R_{l_i}) - c(S_i|R_n)|,\ R_n \in \tilde{R} \setminus \{R_{l_i}\}$
  end for
  $k \leftarrow \arg\max_i C_i$
  assign R_{l_k} to S_k
  $\tilde{S} \leftarrow \tilde{S} \setminus \{S_k\}$; $\tilde{R} \leftarrow \tilde{R} \setminus \{R_{l_k}\}$
end while

dataset, to which we will refer as the PSYCH corpus, is composed of motivational interviewing sessions – a specific type of psychotherapy – between a therapist (T) and a client (C) and is collected from five independent clinical trials (ARC, ESPSB, ESP21, iCHAMP, HMCBI) [19]. The second one is the AMI meeting corpus [20] from which we are using the independent headset microphone (IHM) setup of the scenario-only part. This is composed of meetings where each participant plays the role of an employee in a company; the project manager (PM), the marketing expert (ME), the user interface designer (UI), and the industrial designer (ID).

The two datasets are split into training, development and test sets in such a way that there is no speaker overlap between them. For the AMI corpus we follow the scenario-only partition which is officially recommended[1]. For the PSYCH corpus, since the client IDs are not available for the HMCBI sessions, the partitioning is done under the assumption that it is highly improbable for the same client to visit different therapists in the same study [5]. In both cases, we use the manually derived segmentation. The two datasets are presented in Tables 1 and 2.

	PSYCH-train	PSYCH-dev	PSYCH-test
#sessions	74	44	25
dur-T	26.43 h	15.23 h	7.34 h
dur-C	23.29 h	12.17 h	7.54 h

Table 1. Size of the PSYCH dataset. The durations are calculated based on the manual turn boundaries.

In order to train the required LMs we use the training parts of the PSYCH and AMI corpora, as well as the Fisher English corpus [21] and the transcribed therapy sessions provided by the Counseling and Psychotherapy Transcripts Series[2] (CPTS), as described in Section 4. The size of the corresponding vocabularies and the total number of tokens are given in Table 3.

	AMI-train	AMI-dev	AMI-test
#meetings	98	20	20
dur-PM	16.00 h	2.95 h	3.93 h
dur-ME	10.22 h	2.61 h	2.51 h
dur-UI	9.71 h	2.26 h	1.79 h
dur-ID	11.03 h	2.02 h	2.15 h

Table 2. Size of the AMI dataset. The durations are calculated based on the manual turn boundaries.

	PSYCH-train	AMI-train	Fisher	CPTS
|voc|	8.17K	8.54K	58.6K	35.6K
#tokens	530K	479K	21.0M	6.52M

Table 3. Size of the vocabulary and total number of tokens in the corpora used for LM training.

4. EXPERIMENTS AND RESULTS

First, we train all the necessary LMs, which are 3-gram models with Kneser-Ney smoothing. The generic LM $\mathcal{G}$ is trained using the Fisher English corpus. For the AMI corpus, the 4 role-specific LMs $\mathcal{R}_{PM}, \mathcal{R}_{ME}, \mathcal{R}_{UI}, \mathcal{R}_{ID}$ are trained using only the turns belonging to the corresponding roles in the training set. For the PSYCH corpus, we additionally use the CPTS sessions and get the role-specific LMs $\mathcal{R}_T = w_{o_T}\mathcal{R}_{T,CPTS} \oplus (1 - w_{o_T})\mathcal{R}_{T,PSYCH}$ and similarly for $\mathcal{R}_C$. The mixing weights w_{o_T} and w_{o_C} are optimized so that the perplexity of the turns of the corresponding roles in the development set is minimized. Once we have those LMs, we create the mixed versions according to equations (1) and (2), where all the appearing mixing weights are again optimized to minimize the perplexity of the development corpora. For the optimization of w_g, the corresponding development corpus is the union of all the role-specific development corpora for the dataset we are working with. The LM training and weight optimization is done with the SRILM toolkit [17]. The size of the final mixed vocabulary is 69.5K for the experiments with the PSYCH corpus and 59.6K for the experiments with the AMI corpus, while the phonetic representation of those words is given by the CMU dictionary[3].

The ASR decoding is done with the Kaldi speech recognition toolkit [22] using Kaldi's pre-trained ASpIRE acoustic

[1] http://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml

[2] https://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

model[4]. The word insertion penalty and the LM weighting factor used during decoding are chosen to minimize the Word Error Rate (WER) on the development set.

The evaluation metric used for the final role assignment is the weighted Misclassification Rate (MR), defined as

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

where the summation is over all the speaker turns, R_k is the role assigned by the algorithm, $\hat{R}_k$ is the reference role, d_k is the duration of the k-th turn and $\mathbb{I}(\cdot)$ is the indicator function.

4.1. Turn-level SRR

In Table 4 we present the results using our method (`lm-resc`) for turn-level (`tl`) SRR, as shown in Fig. 1, as well as using the approach shown in Fig. 2 (`lm-asr`) where the cost $c'_j(x)$ is the log-likelihood of the turn x given the LM $\mathcal{R}_j^+$.

	lm-resc-tl	lm-asr-tl	baseline
PSYCH	23.58	**10.75**	50.67
AMI	64.70	63.40	**62.22**

Table 4. MR (%) for turn-level SRR.

As we can see, both `lm-resc-tl` and `lm-asr-tl` fail to beat the baseline classifier which always chooses the majority class (from the training set) for the case of AMI corpus. For the 2-role problem in PSYCH corpus this is not the case, but still `lm-asr-tl` outperforms `lm-resc-tl`. This is because the corpora feature conversational interactions and thus, prior to speaker clustering, utterances are broken into very short speech segments. Each individual segment contains insufficient observations to infer speaker role, and since all decisions are independent, that increases error. Such inaccuracies cancel out when we exploit the aggregate score for all the turns of a speaker as we will see in the next Subsection.

4.2. Speaker-level SRR

Here, the final decision of the role assignment is taken at the speaker level, according to Algorithm 1, which means that a speaker clustering step is necessary. To that end, a Bayesian Information Criterion (BIC) - based Hierarchical Agglomerative Clustering (HAC) is employed on top of an energy-based voice activity detector at the frame level, as explained in [5]. In order for the clustering to make sense in the case of the AMI corpus, we downmix the 4 headset microphones into one audiofile per meeting. As observed in Table 5, our method yields improved results, outperforming both `lm-asr-sl` and the turn-level approaches (Table 4). Of course, the final performance depends on the performance of the clustering algorithm used.

[4] `http://kaldi-asr.org/models/m1`

	lm-resc-sl	lm-asr-sl	BIC-HAC
PSYCH†	**0.00**	7.46	–
PSYCH	**4.41**	5.83	4.08
AMI†	**29.46**	55.52	–
AMI	**46.16**	60.94	15.63

Table 5. MR (%) for speaker-level SRR and for speaker clustering (BIC-HAC). † denotes the use of ground truth speaker clustering information.

4.3. Effect on speech recognition accuracy

Finally, we want to explore whether the role-specific lattice rescoring can lead to improved results for the task of ASR apart from SRR. To that end, for every turn we assume that the lexical information is given by the best path of the rescored lattice corresponding to the role that was assigned by our algorithm to that turn. The results in Table 6 show that this approach, following our per-speaker role assignment, can indeed slightly improve the ASR performance. The slight difference between the WER of the generic ASR model and the combination of the rescored ones, together with the substantial improvements in SRR performance (Table 5) suggest that even small role-specific improvements in the text produced by the ASR can be of high value for a reliable role identification.

	lm-resc-tl	lm-resc-sl	generic
PSYCH	37.84	**37.54**	37.99
AMI	29.35	**29.27**	29.29

Table 6. WER (%) using the best path of the generic ASR or the best paths after the role-specific lattice rescoring and the SRR at the turn and at the speaker level.

5. CONCLUSIONS AND FUTURE WORK

We presented an algorithm which rescores the lattices produced by an ASR system with role-specific LMs in order to exploit the linguistic information in a more robust way for the task of SRR. We experimented with approaches taking the final decision both at the turn and at the speaker level and we identified that the second case leads to more reliable results. Our future efforts will focus on extending the proposed ideas to accommodate the scenario where the role of the same speaker changes during the recording, or multiple speakers play the same role, where the assumption of one-to-one correspondence between roles and speakers would not hold.

6. ACKNOWLEDGEMENTS

This work was supported by NIH. Nikolaos Flemotomos is partially supported by the USC Annenberg Fellowship.

7. REFERENCES

[1] Ashtosh Sapru and Fabio Valente, "Automatic Speaker Role Labeling in AMI Meetings: Recognition of Formal and Social Roles," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5057–5060.

[2] Géraldine Damnati and Delphine 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.

[3] Benjamin Bigot, Corinne Fredouille, and Delphine Charlet, "Speaker Role Recognition on TV Broadcast Documents," in *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

[4] Bo Xiao, Chewei Huang, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan, "A Technology Prototype System for Rating Therapist Empathy from Audio Recordings in Addiction Counseling," *PeerJ Computer Science*, vol. 2, pp. e59, 2016.

[5] Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson, and Shrikanth Narayanan, "Combined Speaker Clustering and Role Recognition in Conversational Speech," in *Interspeech*, 2018.

[6] Yanxiong Li, Qin Wang, Xue Zhang, Wei Li, Xinchao Li, Jichen Yang, Xiaohui Feng, Qian Huang, and Qianhua He, "Unsupervised Classification of Speaker Roles in Multi-Participant Conversational Speech," *Computer Speech & Language*, vol. 42, pp. 81–99, 2017.

[7] Benjamin Bigot, Julien Pinquier, Isabelle Ferrané, and Régine André-Obrecht, "Looking for Relevant Features for Speaker Role Recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[8] Neha P Garg, Sarah Favre, Hugues Salamin, Dilek Hakkani Tür, and Alessandro Vinciarelli, "Role Recognition for Meeting Participants: An Approach Based on Lexical Information and Social Network Analysis," in *Proceedings of the 16th ACM International Conference on Multimedia*. ACM, 2008, pp. 693–696.

[9] Hugues Salamin and Alessandro 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.

[10] Wen Wang, Sibel Yaman, Kristin Precoda, and Colleen Richey, "Automatic Identification of Speaker Role and Agreement/Disagreement in Broadcast Conversation," in *ICASSp*. Citeseer, 2011, pp. 5556–5559.

[11] Mickael Rouvier, Sebastien Delecraz, Benoit Favre, Meriem Bendris, and Frederic Bechet, "Multimodal Embedding Fusion for Robust Speaker Role Recognition in Video Broadcast," in *Automatic Speech Recognition and Understanding*, 2015, pp. 383–389.

[12] Yang 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.

[13] Regina Barzilay, Michael Collins, Julia Hirschberg, and Steve Whittaker, "The Rules Behind Roles: Identifying Speaker Role in Radio Broadcasts," in *AAAI/IAAI*, 2000, pp. 679–684.

[14] Thierry Bazillon, Benjamin Maza, Michael Rouvier, Frederic Bechet, and Alexis Nasr, "Speaker Role Recognition Using Question Detection and Characterization," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[15] Géraldine Damnati and Delphine Charlet, "Robust Speaker Turn Role Labeling of TV Broadcast News Shows," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5684–5687.

[16] Panayiotis G. Georgiou, Matthew P. Black, Adam Lammert, Brian Baucom, and Shrikanth S. Narayanan, ""That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?," in *Proceedings of Affective Computing and Intelligent Interaction (ACII), Lecture Notes in Computer Science*, Oct. 2011.

[17] Andreas Stolcke, "SRILM–An Extensible Language Modeling Toolkit," in *Seventh International Conference on Spoken Language Processing*, 2002.

[18] Nikolaos Flemotomos, Victor Martinez, James Gibson, David Atkins, Torrey Creed, and Shrikanth Narayanan, "Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions," *Proc. Interspeech 2018*, pp. 1908–1912, 2018.

[19] David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth, "Scaling Up the Evaluation of Psychotherapy: Evaluating Motivational Interviewing Fidelity via Statistical Text Classification," *Implementation Science*, vol. 9, no. 1, pp. 49, 2014.

[20] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The AMI meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[21] Christopher Cieri, David Miller, and Kevin Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *LREC*, 2004, vol. 4, pp. 69–71.

[22] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2011, number EPFL-CONF-192584.