ABSTRACT

Speaker diarization relies on the assumption that acoustic embeddings from speech segments corresponding to a particular speaker share common characteristics. Thus, they are concentrated in a specific region of the speaker space; a region which represents that speaker’s identity. Those identities however are not known a priori, so a clustering algorithm is employed, which is typically based solely on audio. In this work we explore conversational scenarios where the speakers play distinct roles and are expected to follow different linguistic patterns. We aim to exploit this distinct linguistic variability and build a language-based segmenter and a role recognizer which can be used to construct the speaker identities. That way, we are able to boost the diarization performance by converting the clustering task to a classification one. The proposed method is applied in real-world dyadic psychotherapy interactions between a provider and a patient and demonstrated to show improved results.

Index Terms— diarization, speaker role, x-vector, speaker change detection, sequence labelling

1. INTRODUCTION

Given a speech signal with multiple speakers, diarization answers the question “who spoke when” [1]. The traditional approach has been to model speech segments under some probability distribution (e.g., Gaussian Mixture Models - GMMs), and measure the distance between them using a metric such as the one based on the Bayesian Information Criterion (BIC) [2]. When the distance between two speech segments is beyond a threshold, they are considered to belong to different speakers. Diarization is typically viewed as an unsupervised task, and hence heavily depends on the successful application of a clustering algorithm in order to group same-speaker segments together.

In recent years, with the advent of Deep Neural Networks (DNNs), speaker modelling is performed using fixed-dimensional neural embeddings which capture speaker-specific characteristics encoded in the speech signal. Such embeddings are bottleneck features extracted from architectures usually trained under the objective of speaker classification employing a cross-entropy loss function [3], or speaker discrimination employing contrastive [4] and triplet [5] loss functions. Typical examples of embeddings that have shown state-of-the-art performance for speaker diarization are the Time-Delay Neural Net (TDNN) based x-vectors [6] and the Long-Short Term Memory (LSTM) based d-vectors [7]. The divergence criterion typically used to compare between such embeddings is based on Probabilistic Linear Discriminant Analysis (PLDA) [8].

Even though speaker diarization has traditionally been an audio-only task which relies on the acoustic variability between different speakers, the linguistic content captured in the speech signal can offer valuable supplemental cues. Apart from practical observations such as the fact that it is highly improbable for a speaker change point to be located within a word [9], it is widely accepted that each individual has their very own way of using language [10]. Thus, language patterns followed by individual speakers have been explored in the literature for the tasks of speaker segmentation and clustering, both when used unimodally [11], and in combination with the speech audio [12][13].

Using language for diarization can be especially promising in structured scenarios where the speakers assume dissimilar roles with distinguishable linguistic patterns. For example, a teacher is likely to speak in a more didactic style while a student be more inquisitive, a doctor is likely to inquire on symptoms and prescribe while a patient describe symptoms, etc. In our own past work we exploited the differences of language between a therapist and a patient. We employed a language-based role recognizer and combined it with an audio-based speaker clustering algorithm in [15], while in [16] we used role-specific Language Models (LMs) to recore the output of an Automatic Speech Recognition (ASR) system, showing improved performance for the final task of role recognition. The strong interconnection between speech and language in such scenarios has even led to end-to-end role-annotated ASR architectures [17][18] where diarization becomes a byproduct of a rich transcription system.

 Influenced by this line of work, we propose an alternative way of using the linguistic information for the task of speaker diarization in recordings where participants play specific roles which are known in advance. In particular, we process independently the text stream in order to segment it in speaker-homogeneous chunks, each one of which can be assigned to one of the available speaker roles. Aggregating this information for all the segments, and aligning text with audio, we can construct the identities of the speakers found in the recording. That way, each audio segment can be assigned to a speaker through a simple classifier, overcoming the burden of clustering. We apply this approach in psychotherapy recordings featuring dyadic interactions between two speakers with well-defined roles; namely that of a therapist and a patient.

2. METHOD

As baseline to our system we will employ an audio-only speaker diarization system as depicted in Fig. 1, while in Fig. 2 we illustrate our proposed approach. We describe the various modules in detail in Sections 2.1-2.4.

Fig. 1. Baseline audio-based speaker diarization.
2.1. Audio-based Diarization with Speaker Clustering

As a baseline, we use a state-of-the-art speaker diarization system following the x-vector/PLDA paradigm [9]. As shown in Fig. 1 the speech signal is first segmented uniformly with a short sliding window. For each resulting segment an x-vector is extracted, which is expected to capture speaker-specific characteristics [3]. The pairwise distances between all those embeddings are then calculated based on PLDA scoring [19][20]. Given two x-vectors \( v, r \), PLDA provides a framework to estimate the log-likelihood ratio

\[
s(v, r) = \log \frac{p(v, r | \text{same speaker})}{p(v | \text{dif. speakers})p(r | \text{dif. speakers})}
\]  

(1)

The segments are clustered into same-speaker groups following a Hierarchical Agglomerative Clustering (HAC) approach with average linking. Since our experiments are conducted on dyadic interactions, we force the HAC algorithm to run until two clusters are constructed. Such a method, however, even though being the typical approach, poses the risk of creating noisy, non-representative speaker clusters. In particular, if the speakers to be clustered reside closely in the speaker space and if there is enough noise and/or silence within a recording (possibly not sufficiently captured by a voice activity detection algorithm), it may be the case that one of the constructed clusters only contains the non-speech or distorted-speech segments. This is the problem our proposed method tries to avoid by omitting the clustering step.

2.2. Language-based Segmentation

Given the textual information of the conversation, our goal is to obtain speaker-homogeneous text segments; that is segments where all the words have been uttered by a single speaker. Ideally, for this subtask we need a text-based speaker change detector. Even though such systems have been proposed [11], for our final goal we can safely over segment the available document, provided this leads to a smaller degree of inter-segment speaker contamination [13]. So, we assume that each sentence is with high probability speaker-homogeneous and we instead segment at the sentence level, which is an easier task.

To that end, the problem can be viewed as a sequence labeling one, where each word is tagged as either being in the middle of a sentence, or at the beginning. Following state-of-the-art architectures for the problem of sequence tagging [21], we build a Bidirectional LSTM (BiLSTM) network with a Conditional Random Field (CRF) inference layer, as shown in Fig. 3. The input to the recurrent layers is a sequence of words. Each word is given as a concatenation of a character-level representation predicted by a Convolutional Neural Network (CNN) and a word embedding. For our experiments, we initialize the word embeddings with the extended dependency skip-gram embeddings [23], pre-trained on 2B words of Wikipedia.

2.3. Speaker Role Recognition and Profile Estimation

The next step in our system is the application of a text-based role recognition module. In more detail, assuming we have \( N \) speakers in the session (\( N = 2 \) for our experiments) and there is one-to-one correspondence between speakers and roles (e.g., there is one therapist and one patient), we want to assign one of the role labels \( \{R_i\}^N_{i=1} \) to each segment. To do so, we build \( N \) role LMs \( \{LM_i\}^N_{i=1} \), one for each role, and we estimate the perplexity of a segment given the LM \( R_i \), for \( i = 1, 2, \cdots, N \). The role assigned to the segment is the one yielding the minimum perplexity [15].

At this point, we have several text segments corresponding to each role \( R_i \). If we have the alignment information at the word level\( ^1 \) we can directly get the time-boundaries of those segments. We extract one embedding (x-vector) for each and we estimate a role/speaker identity (also known as profile) \( r_i \) as the mean of all the embeddings corresponding to the specific role. We note that the role recognition at the segment level does not always provide robust results [10], something which could lead to unreliable generated profiles. However, we expect that there will be a fraction of the segments for the results of which we are confident enough and we can take only those into consideration for the final averaging. The proxy used as our confidence for the segment-level role assignment is the difference between the best and the second best perplexity of a segment given the various LMs. Formally, if the segment \( x \) is assigned the role \( R_i \), and if \( pp(x | R_i) \) is the perplexity of \( x \) given the LM \( R_i \), then the confidence metric used for this assignment is \( \min_{j \neq i} [pp(x | R_i) - pp(x | R_j)] \). We note that all the perplexities are normalized for segment length.

2.4. Audio-based Diarization with Speaker Classification

After having computed all the needed profiles \( \{r_i\}^N_{i=1} \), in order to perform diarization, we apply the same uniform segmentation as in the baseline system (Section 2.1) and we extract one x-vector per segment. However, instead of clustering the resulting x-vectors, we need to classify them within the correct role. Our classifier is also based on PLDA, in order to have a fair comparison between the base-

\[ ^1 \text{If we have access to the transcripts and the audio, we can force-align. If we generate the text through ASR, we get the desired alignments from the decoding lattices.} \]
line and our system. In that framework, a segment with embedding $u$ is assigned the label $R = \arg \max_{1 \leq i \leq N} \{s(u, r_i)\}$, where $s(\cdot, \cdot)$ is the PLDA score estimated in equation (1).

3. DATASETS

3.1. Evaluation Data

We evaluate our proposed method on datasets from the clinical psychology domain. In particular, we adopt the system to motivational interviewing sessions—a specific type of psychotherapy—between a therapist (T) and a patient (P), collected from five independent clinical trials (ARC, ESPSB, ESP21, iCHAMP, HMCBI) [23]. We collectively refer to those sessions as the PSYCH corpus. This is split into training, development, and test sets, as shown in Table 1 in such a way that there is no speaker overlap between them. Even though the necessary metadata are available for the rest of the corpus, the patient IDs are not available for the HMCBI sessions. Thus, the partitioning is done under the assumption that it is highly improbable for the same patient to visit different therapists in the same study. All the results reported are on PSYCH-test.

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

Table 1. Size of the PSYCH dataset. dur-T and dur-P are the total durations (calculated based on the manual turn boundaries) of the speech segments corresponding to a therapist and to a patient, respectively.

3.2. Segmenter and Role LM Training Data

The segmenter presented in Section 2.2 is trained on a subset of the Fisher English corpus [24] comprising a total of 10,195 telephone conversations for which the original transcriptions (including punctuation symbols) are available. This set is enhanced by 1,199 in-domain therapy sessions provided by the Counseling and Psychotherapy Transcripts Series (CPTS). The combined dataset is randomly split (80-20 split at the session level) into training and validation sets. CPTS and the entire Fisher English corpus, together with the training part of the PSYCH corpus, are also used to train the required role-specific LMs. The size of the corresponding vocabularies is given in Table 2.

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

Table 2. Size of the vocabulary and total number of tokens in the corpora used for LM training.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

For this work we use the text which is available from the manual transcriptions of the PSYCH corpus, after normalizing to remove punctuation symbols and capital letters, and force-aligning with the audio sessions. Based on the word alignments, we segment the audio according to whether there is a silence gap between two words larger than a threshold equal to 1 sec. The diarization ground truth is also constructed through the word alignments, by allowing a maximum of 0.2 sec-long in-turn silence.

The resulting text segments are further subsegmented at the sentence level based on the output of the tagger in Fig. 1. During training we define as “sentence” any text segment between two punctuation symbols denoting pause apart from commas. We exclude commas first because they normally do not indicate speaker change points but also because they are too frequent in our training set and they would lead to very small segments, not containing sufficient information for the task of role recognition. The tagger is built using the nCRLF++ toolkit [25]. Following the general recommendations in [26] and after our own hyperparameter tuning, the network comprises 4 CNN layers and 2 BiLSTM layers with dropout ($p = 0.5$) and $l_2$ regularization ($\lambda = 10^{-8}$). The length of each word representation is 330 (character embedding dimension = 30, word embedding dimension = 300). The network is trained using the Adam optimizer with a fixed learning rate equal to $10^{-4}$ and a batch size equal to 256 word sequences. The tagger achieves an F1 score of 0.805 on the validation set after 14 training epochs.

All the LMs required for role recognition are 3-gram models with Kneser-Ney smoothing built with the SRILM toolkit [27]. LMs are trained using a combination of in-domain data from the PSYCH-train and CPTS corpora and out-of-domain data from the Fisher English corpus, employing the LM interpolation procedure described in detail in [16], with mixing weights optimized on PSYCH-dev. SRILM is also used to estimate the needed perplexities.

The audio-based diarization framework is built using the Kaldi toolkit [28]. Both for the baseline and for our system, we use the VoxCeleb pre-trained x-vector extractor and the PLDA model which comes with it, after we adapt it on the development set of the PSYCH corpus. The x-vectors are extracted after uniformly segmenting the audio into 1.5 sec-long windows with 50% overlap. Those are normalized and decorrelated through an LDA projection and dimensionality reduction (final embedding length = 200), mean, and length normalization. The evaluation is always based on the Diarization Error Rate (DER), as estimated by the NIST md-eval.pl tool, with a 0.25 sec-long collar, ignoring overlapping speech.

4.2. Results and Discussion

Table 3 gives the results of our language-aided diarization system in comparison with the audio-only baseline approach. When we use our sequence tagger to segment at the sentence level and keep all the resulting segments to estimate the speaker profiles, we get a 26.70% DER relative improvement. Further improvements are observed if we only keep the segments we are most confident about. In particular, we estimate the confidence for each segment of a session using the metric introduced in Section 2.3 and we find the parameter $\alpha$ that minimizes the overall DER on the development set when only the $\alpha\%$ segments we are most confident about are taken into considera-

---

Footnotes:
1. [https://alexanderstreet.com/products/counselling-and-psychotherapy-transcripts-series](https://alexanderstreet.com/products/counselling-and-psychotherapy-transcripts-series)
2. [https://kaldi-asr.org/models/m7](https://kaldi-asr.org/models/m7)
text segmentation — tagger — oracle

| method          | audio-only | language-aided |
|-----------------|------------|----------------|
| DER (%)         | 11.05      | 8.10           |
| DER† (%)        | 7.51       | 6.99           |

Table 3. Diarization results following the audio-only and our language-aided approach. The text segmentation is either done through our sequence tagger or based on the manually annotated speaker changes (oracle). † denotes results when only a% of the segments we are most confident about are taken into account in each session for the profile estimation, where a is a parameter optimized on the development set (a = 55 for the tagger segmentation and a = 60 for the oracle segmentation).

In the last column of Table 3 we report results when using the oracle speaker segmentation provided by the manual annotations instead of applying the sequence tagger. That way, we can inspect the effect of the implemented idea on the diarization performance eliminating any error propagation caused by a suboptimal speaker change detector. As expected, the results are indeed better, but the performance gap is reduced when we choose not to keep all the segments for the speaker profile estimation.

This behavior is also reflected in Fig. 4, where we plot DER as a function of the percentage of the segments we use to estimate the speaker profiles within a session. Even though the oracle text segmentation consistently yields marginally better results, it seems that if we carefully choose which segments to use to get an estimate of the speakers’ identities, our tagger-based segmentation approaches the oracle performance. In fact, the best result we got on the test set (optimizing for a on the same set) using our segmenter was 7.28% DER, while the corresponding number using the oracle segmentation was 6.99%. We should highlight here that the analysis presented in this work is based on using a% of the segments within a session, after choosing some a which remains constant across sessions. It is probable that this is a session-specific parameter which ideally should be chosen based on an alternative, session-level strategy.

5. CONCLUSIONS AND FUTURE WORK

We proposed a system for speaker diarization suitable to use in conversations where the participants assume specific roles, associated with distinct linguistic patterns. While this task typically relies on clustering methods which can lead to noisy speaker partitions, we demonstrated how we can exploit the lexical information captured within the speech signal in order to estimate the speaker profiles and follow a classification approach instead. A text-based speaker change detector is an essential component of our system. For this subtask, assuming each sentence is speaker-homogeneous, we proposed using a sequence tagger which can segment at the sentence level, by detecting the beginning of a new sentence and we showed that this segmentation strategy approaches the oracle performance. The resulting segments are assigned a speaker role label which is later used to construct the desired speaker identities and we introduced a confidence metric to be used for this assignment. Our results showed that such a metric can be used in order to take into consideration only the segments we are most confident about, leading to further performance improvements. When applied to dyadic interactions between a therapist and a patient, our proposed method achieved an overall relative DER reduction equal to 32.04%, compared to the baseline audio-only approach with speaker clustering.

We should note that for our experiments we used the reference transcriptions to demonstrate the effectiveness of the implemented idea. Of course, in a real-world setting the text stream would be generated as the output of an ASR system. Additionally, it should be highlighted that the diarization results can be further improved if, for example, a resegmentation module is employed as a final step, or a more precise audio segmentation strategy is followed instead of relying on uniform segmentation. For example, an audio-based speaker change detector could be applied both for the baseline and the language-aided system and in the latter case this could be used in combination with the language-based segmenter. However, our goal in this work was mainly to demonstrate the effectiveness of constructing the speaker profiles within a session to be diarized in order to convert the clustering task into a classification one.

Herein we essentially modelled each speaker by a single embedding, since for the final profile estimation we averaged over all the speech segments assigned to the corresponding speaker. A potential extension of the current work would be an exploration of alternative speaker identity construction strategies, e.g., representing a speaker by a distribution of embeddings. This sounds particularly promising in scenarios where the recordings are long enough so that they may incorporate various acoustic conditions or different speaking styles corresponding to the same speaker. Finally, another direction of future work could be towards an in-depth investigation of the confidence metric used for the text-based role assignments and, accordingly, the number of those segments that we should take into consideration when estimating the speaker profiles.

6. ACKNOWLEDGEMENTS

This work was supported by NIH.

7. REFERENCES

[1] Xavier Anguera, Simon Bozonnét, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, “Speaker Diarization: A Review of Recent Research,” Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 356–370, 2012.
[1] Scott Chen and Ponani Gopalakrishnan, “Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion,” in Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998, vol. 8, pp. 127–132.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust DNN Embeddings for Speaker Recognition,” in International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5329–5333.

[3] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, “Speaker Diarization Using Deep Neural Network Embeddings,” in Proc. International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 4930–4934.

[4] Hervé Bredin, “Tristounet: Triplet Loss for Speaker Turn Embedding,” in Proc. International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 5430–5434.

[5] Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, et al., “Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge,” in Proc. Interspeech, 2018, pp. 2808–2812.

[6] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio López Moreno, “Speaker Diarization with LSTM,” in International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5239–5243.

[7] Gregory Sell and Daniel Garcia-Romero, “Speaker Diarization with PLDA i-vector Scoring and Unsupervised Calibration,” in Proc. Spoken Language Technology Workshop, 2014, pp. 413–417.

[8] Jan Silovsky, Jindrich Zdansky, Jan Nouza, Petr Cerva, and Jan Prazák, “Incorporation of the ASR Output in Speaker Segmentation and Clustering within the Task of Speaker Diarization of Broadcast Streams,” in Proc. International Workshop on Multimedia Signal Processing, 2012, pp. 118–123.

[9] Barbara Johnstone, The linguistic individual: Self-expression in language and linguistics, Oxford University Press, 1996.

[10] Zhao Meng, Lili Mou, and Zhi Jin, “Hierarchical RNN with Static Sentence-Level Attention for Text-Based Speaker Change Detection,” in Proc. Conference on Information and Knowledge Management, 2017, pp. 2203–2206.

[11] Miquel Àngel India Massana, José Adrián Rodríguez Fonolosa, and Francisco Javier Hernando Pericás, “LSTM Neural Network-Based Speaker Segmentation Using Acoustic and Language Modelling,” in Proc. Interspeech, 2017, pp. 2834–2838.

[12] Zbyněk Zajíc, Daniel Soutner, Marek Krůz, Laděk Müller, and Vlasta Radová, “Recurrent Neural Network Based Speaker Change Detection from Text Transcription Applied in Telephone Speaker Diarization System,” in Proc. International Workshop on Temporal, Spatial, and Spatio-Temporal Data Mining, 2018, pp. 342–350.

[13] Tae Jin Park, Kyu J. Han, Jing Huang, Xiaodong He, Bowen Zhou, Panayiotis Georgiou, and Shrikranth Narayanan, “Speaker Diarization with Lexical Information,” in Proc. Interspeech, 2019, pp. 391–395.

[14] Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson, and Shrikranth Narayanan, “Combined Speaker Clustering and Role Recognition in Conversational Speech,” in Proc. Interspeech, 2018, pp. 1378–1382.

[15] Nikolaos Flemotomos, Panayiotis Georgiou, and Shrikranth Narayanan, “Role Specific Lattice Rescoring for Speaker Role Recognition from Speech Recognition Outputs,” in Proc. International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 7330–7334.

[16] Nikolaos Flemotomos, Zhihao Chen, David C Atkins, and Shrikranth Narayanan, “Role annotated speech recognition for conversational interactions,” in Proc. Spoken Language Technology Workshop, 2018, pp. 1036–1043.

[17] Laurent El Shafey, Hagen Soltau, and Izhak Shafran, “Joint Speech Recognition and Speaker Diarization via Sequence Transduction,” Proc. Interspeech, pp. 396–400, 2019.

[18] Pierre Oifène, “Probabilistic Linear Discriminant Analysis,” in Proc. European Conference on Computer Vision, 2006, pp. 531–542.

[19] Simon JD Prince and James H Elder, “Probabilistic Linear Discriminant Analysis for Inferences about Identity,” in Proc. International Conference on Computer Vision, 2007, pp. 1–8.

[20] Xuezhe Ma and Eduard Hovy, “End-to-End Sequence Labeling via Bi-directional LSTM-CNNs-CRF,” in Proc. Annual Meeting of the Association for Computational Linguistics, 2016, vol. 1, pp. 1064–1074.

[21] Alexandros Kominos and Suresh Manandhar, “Dependency Based Embeddings for Sentence Classification Tasks,” in Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1490–1500.

[22] 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.

[23] Christopher Cieri, David Miller, and Kevin Walker, “The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text,” in Proc. Conference on Language Resources and Evaluation, 2004, vol. 4, pp. 69–71.

[24] Jie Yang and Yue Zhang, “NCRF++: An Open-source Neural Sequence Labeling Toolkit,” in Proc. Annual Meeting of the Association for Computational Linguistics, 2018, pp. 74–79.

[25] Nils Reimers and Iryna Gurevych, “Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging,” in Proc. Conference on Empirical Methods in Natural Language Processing, 2017, pp. 338–348.

[26] Andreas Stolcke, “SRILM–An Extensible Language Modeling Toolkit,” in Proc. International Conference on Spoken Language Processing, 2002, pp. 901–904.

[27] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlícek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Veselý, “The Kaldi Speech Recognition Toolkit,” in Proc. Workshop on Automatic Speech Recognition and Understanding, 2011.