IDIAP RESEARCH REPORT

A BAYESIAN APPROACH TO INTER-TASK FUSION FOR SPEAKER RECOGNITION

Srikanth Madikeri Subhadeep Dey Petr Motlicek

Idiap-RR-07-2020

MARCH 2020

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

A BAYESIAN APPROACH TO INTER-TASK FUSION FOR SPEAKER RECOGNITION

Srikanth Madikeri[1], *Petr Motlicek*[1] *and Subhadeep Dey*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{srikanth.madikeri, petr.motlicek, subhadeep.dey}@idiap.ch

ABSTRACT

In i-vector based speaker recognition systems, back-end classifiers are trained to factor out nuisance information and retain only the speaker identity. As a result, variabilities arising due to gender, language and accent (among many others) are suppressed. Inter-task fusion, in which such metadata information obtained from automatic systems is used, has been shown to improve speaker recognition performance. In this paper, we explore a Bayesian approach towards inter-task fusion. Speaker similarity score for a test recording is obtained by marginalizing the posterior probability of a speaker. Gender and language probabilities for the test audio are combined with speaker posteriors to obtain a final speaker score. The proposed approach is demonstrated for speaker verification and speaker identification tasks on the NIST SRE 2008 dataset. Relative improvements of up to 10% and 8% are obtained when fusing gender and language information, respectively.

Index Terms— Inter-task fusion, Bayesian fusion, speaker recognition

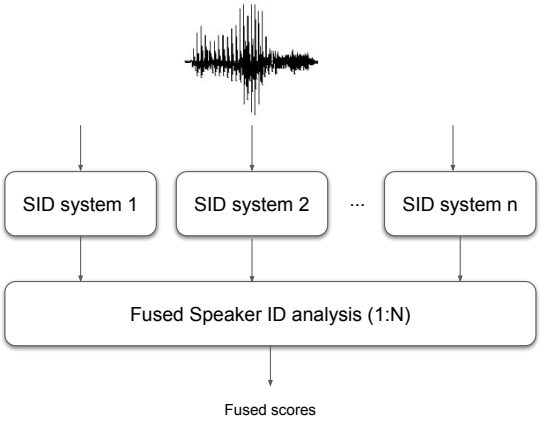

Fig. 1. Block diagram showing the fusion procedure for multiple speaker identification systems, also termed as intra-task fusion.

1. INTRODUCTION

Automatic speaker verification (ASV) and identification (SID) are among two major applications of speaker recognition technologies. This paper focuses on both tasks. Unlike ASV, which aims at authenticating the claimed identity of a speaker based on some speech recording and enrolled speaker model, SID compares the speech recording against the set of N speaker models. SID systems are typically deployed by law enforcement or security agencies in detecting multiple and arbitrary identities used by criminals and terrorists [1].

System submissions in NIST SRE challenges (e.g. UTD-CRSS system submitted in 2016 [2]) have clearly demonstrated that speaker recognition can largely profit from system fusion. In most typical cases, system fusion means *intra-task fusion*, aiming to train a set of independent classifiers with different types of speech features, or acoustic models, and eventually combining the classification scores. The objective of such a fusion is to minimize the errors made by individual systems by exploiting their complementary nature. Numerous techniques exist to combine systems at both model level and score level. In i-vector PLDA (Probabilistic Linear Discriminant Analysis) systems [3, 4], model level fusion may be applied while training the back-end classifier, while output level combination may be achieved by a simple linear combination of the scores. Logistic regression is a commonly employed technique in score-level fusion. In Figure 1, the score combination technique is shown for the case of multiple speaker identification systems. Other approaches such as asynchronous fusion [5] (i.e. fusion of information extracted from different modalities, such as audio and video) can be included in intra-task fusion as well, since different modalities are exploited to train systems for the same task – speaker identification.

This paper proposes an innovative approach towards information fusion - exploiting heterogeneous systems (i,e. systems trained for different tasks), for *inter-task fusion*. In strategic intelligence applications of speaker recognition, it is quite common to have access to metadata of speakers and audios. Speaker metadata includes name, age, gender, nationality, languages spoken, etc. Audio metadata may include channel type (e.g. telephone, social media), language of the recording, phone number, date of the recording, etc. The metadata can be used to validate the speaker recognition system's output. As a trivial example, if a speaker identification system recognizes an unknown speaker's recording as someone who can speak only English while the test recording was in German, the metadata can be used to avoid a false alarm. The goal of inter-task fusion is to exploit such information, either available through external knowledge or through automatic systems (e.g. language identification, accent identification, age identification), for the benefit of speaker recognition. In [6], two methods for inter-task fusion were demonstrated: a score-level fusion based on logistic regression and a model-level fusion that re-used PLDA backends trained for different tasks. Accent identification and language identification systems were trained using the same i-vector PLDA architecture as ASV system. Although the proposed approach does not outperform i-vector based systems, it offers a scalable solution with respect to amount of training data and complementary scores which can be efficiently exploited in system fusion. The improvements observed on NIST SRE 2008 [7] showed that inter-task fusion can be beneficial for speaker recognition ap-

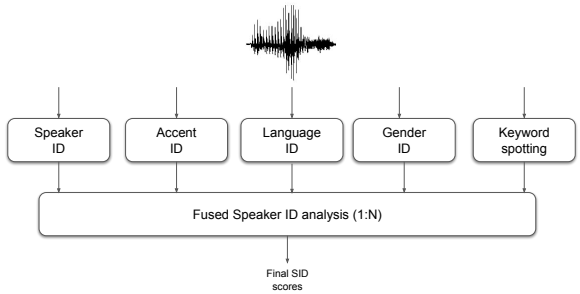

Fig. 2. Block diagram showing the inter-task fusion procedure toward enhanced speaker identification from a speech recording.

plications with access to outputs from different speech processing systems.

The rest of the paper is organized as follows. Section 2 briefly describes the i-vector PLDA framework for speaker recognition. Section 3 provides motivation to the probabilistic approach to inter-task speaker recognition and describes a generic framework to fuse metadata with speaker recognition scores. Section 4 presents the results of experiments conducted on the NIST SRE 2008 dataset. Finally, the paper is summarized in 5.

2. I-VECTOR PLDA FRAMEWORK

Conventional speaker recognition systems are built around the i-vector PLDA framework, where speaker models are extracted by projecting Gaussian mean supervectors on a low-dimensional subspace called *total variability space* (TVS) [3]. The variability model underlying i-vector extraction is given by:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{s}$ is the supervector adapted with respect to UBM-GMM from a speech recording. The vector $\mathbf{m}$ is the mean of the supervectors,$\mathbf{T}$ is the matrix with its columns spanning the total variability subspace and $\mathbf{w}$ is the low-dimensional i-vector representation. I-vectors model the average information content in the audio. In the above model, the i-vector is assumed to have a standard Normal distribution as prior.

LDA, WCCN [8, 3] and PLDA [9, 10] together form the back-end of the i-vector system. The back-end is responsible for projecting the i-vectors in a speaker discriminative space. Two i-vectors are compared as belonging to the same class or as belonging to two different classes, thus generating a likelihood ratio (LR) to score a pair of speech utterances.

Similar to speaker recognition, information such as language and gender can be obtained in an automatic way from language identification and gender identification systems (as shown in Figure 2), respectively.

In this paper, we use a Bayesian approach to use language and gender information as virtual evidence to improve speaker recognition performance. The use of the Bayesian approach provides a fast and simple way, by the nature of the approach, to integrate metadata information. We test our approach on the cross-lingual condition in NIST SRE 2008 dataset [7].

3. BAYESIAN FUSION

In inter-task fusion for speaker recognition, we focus on the scenario where we have access to the output of speech processing systems that analyse the same audio for different characteristics. We use two such systems in this paper: language identification (LID) and gender identification (GID) systems. Using the output of these systems as metadata of the audio, we propose to improve speaker recognition (i.e. identification and verification) performance.

In [6], this was achieved by training an LID and an accent identification (AID) system, and fusing either the i-vectors or the scores with the i-vector system modelling speakers. The i-vector fusion was developed by projecting the speaker i-vectors on PLDA trained for speaker discriminability and language discriminability (i.e. for LID fusion), or accent discriminability (i.e. for AID fusion) and appending another stage of PLDA classification. The score-level fusion was represented by a logistic regression on speaker and language (or accent) scores. In this paper, we generalize the score fusion using a probabilistic framework based on marginal distribution and Bayes' theorem for combining probabilities. The proposed technique addresses the following scenario where we can have an access to GID and LID systems (e.g. as auxiliary information/metadata provided along with audio). One trivial way to gain such access is to retrain the backend of the i-vector system trained for ASV/SID with target classes as languages for LID and gender for GID. The objective is to identify or verify a speaker using additional information from metadata.

Given that we have access to the categories of GID and LID engines, the problem of integrating the outputs of speaker identification, LID and GID identification can be formulated as the computation of the following probability term:

$$P(s|L, G, X), \tag{2}$$

where s is the target speaker, L is the language recognized by LID, G is the gender recognized by GID and X are the features obtained from the test observations. The features can be frame-level features such as traditional Mel Frequency Cepstral Co-efficients (MFCC), or speaker models such as i-vectors. The term is generic and independent of any speaker recognition framework. Interpreting the posterior probability as a belief network, we observe that the model assumes a speaker's identity dependent on the audio and metadata (language and gender in this case).

When using a standalone speaker recognition system, we are interested in (i) comparing two speaker models for speaker verification, or (ii) finding the closest enrolled speaker in speaker identification. The posterior probability of a speaker can be easily used in the latter, while in the former the likelihood ratios can be derived from the posterior. Thus, for a standalone system we are interested in computing (directly or indirectly):

$$P(s|X). \tag{3}$$

When having access to auxiliary information for the same audio, we marginalize this probability as follows:

$$P(s|X) = \sum_{m \in \mathcal{M}} P(s, m|X), \tag{4}$$

where $\mathcal{M}$ represents the sample space of metadata in the discrete case.

If we assume that metadata can be provided independent of the speaker identity, then:

$$P(s|X) = \sum_{m \in \mathcal{M}} P(s|X, m)P(m|X). \tag{5}$$

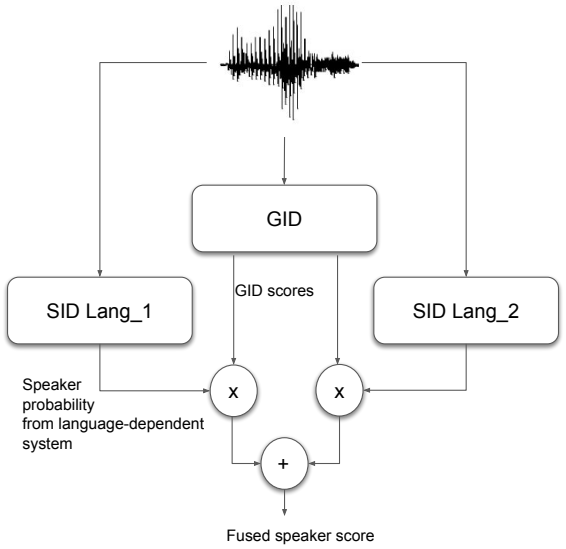

Fig. 3. Block diagram showing the proposed fusion technique to fuse language information in speaker recognition system.

The conditional probability $P(s|X,m)$ can be interpreted as a speaker recognition system developed for the domain represented by the metadata m. A domain refers to an instance in the metadata sample space. For example, if m is one of the languages supported by the language identification engine, then the conditional term represents a speaker recognition system developed for the language m, that is, a language-dependent speaker recognition system.

In this paper, we consider two possibilities for $\mathcal{M}$: one where $\mathcal{M}$ is the set of all languages and another where $\mathcal{M}$ is the gender set. Figure 3 illustrates the fusion of scores from LID with language-dependent speaker recognition system. There are multiple ways to train a system to model the conditional probability $P(s|X,m)$. One strategy is to use PLDA adaptation from a domain-independent model to a domain-dependent model. When sufficient data is available, the domain-dependent PLDA can be directly trained without any adaptation. We employ this strategy for all our systems in this paper.

3.1. Metadata from groundtruth

A trivial example of an implementation of Eq. 5 is the development of gender-dependent speaker recognition systems. According to the model above, $P(m = \text{male}|X)$ is 1.0 for male speakers. Similarly, $P(m = \text{female}|X)$ is 1.0 for female speakers. In general, this is the case when metadata is available as a part of data collection. For instance, we may have access to information that the data belongs to one particular language. Thus, $P(m|X)$ is 1 for the value of m matching the domain (language in this example) and 0 for the rest of the domains. The model can also been seen as a generalization of score fusion where the values of $P(m|X)$ are estimated for optimal performance.

3.2. Virtual evidence as metadata

To integrate the scores obtained from LID and GID with an objective to improve speaker recognition, we employ the concept of virtual evidence from Bayesian networks (Chapter 3 in [11]). Virtual evidence helps modelling the uncertainty in the observation. In cases when we cannot guarantee a speech processing system's output to be completely accurate, we can incorporate the uncertainty in the output in Eq. 5. If the output of a language identification system (or a gender identification system) is $\tilde{m}$, virtual evidence is the probability that the true value is m, that is $P(\mathcal{M}|\tilde{m})$. Therefore, Eq. 5 becomes:

$$P(s|X) = \sum_{m \in \mathcal{M}} P(s|X,m)P(m|\tilde{m},X). \tag{6}$$

The term $P(m|\tilde{m},X)$ is the probability that the true value is m given that the engine corresponding to the metadata has predicted it to be $\tilde{m}$.

3.3. Speaker posteriors

Until now we have discussed two different ways to estimate and use $P(m|X)$. The remaining term $P(s|X,m)$ can be obtained from an i-vector PLDA based speaker recognition system. PLDA systems are usually implemented to produce log likelihood ratios and not posterior probabilities. We can obtain posterior probabilities by applying softmax over likelihoods from enrolled speakers or a cohort set and assuming uniform speaker priors [12]. Given a cohort set $\mathcal{C}$, the softmax for ASV in the i-vector PLDA framework is given by:

$$P(s|X,m) = \frac{exp\left(P(X,\mathbf{w}_s|\Lambda_m)\right)}{\sum_{c \in \mathcal{C} \cup \{s\}} exp\left(P(X,\mathbf{w}_c|\Lambda_m)\right)}, \tag{7}$$

where $\mathbf{w}_s$ is the ivector of the enrolled speaker s, Λ_m are the hyperparameters of the i-vector PLDA system, and $P(X,\mathbf{w}_s|\Lambda_m)$ is the probability that the speaker in the test recording and the enrolled speaker s are the same. The same scoring technique can be used in SID systems by replacing the $\{s\}$ by the set of enrolled speakers.

For the purposes of system evaluation, we circumvent the softmax computation and use logarithmic addition, which is a function to obtain $\log(a+b)$ given $\log(a)$ and $\log(b)$.

4. EXPERIMENTS

Both ASV and SID are conducted on condition 6 of the NIST SRE 2008 data [13]. This condition is best suited to evaluate the effect of cross-lingual trials. The same condition can also be used to see the benefits of incorporating gender information as both male and female trials are available. The evaluation set has 1'788 unique enrollments, 2'569 test files and 35'896 trials across both genders. The Equal Error Rate (EER) of the systems are compared for ASV task. Expected Rank (ER) proposed in [6] is used to report the performance for SID task. ER indicates the average rank of the true speaker. In the results presented here, the minimum possible value of ER is set to 1.0. ASV results are reported on the trials available from the core condition. SID experiments are performed assuming all enrolled speakers are accessible.

4.1. I-vector PLDA configuration

The configuration of the speaker recognition system follows our setup in [14]. Feature vectors consist of 19 dimensional MFCCs extracted every 10 ms over a 25 ms sliding window and post-processed using cepstral mean and variance normalization followed by feature warping over a 3 s long sliding window [15]. A 2'048 component UBM-GMM and 400 dimension i-vector extractor were trained for each system presented using data from Fisher English Parts I and

Table 1. Results on the fusion of speaker and gender information. The results are reported in terms of EER for ASV and ER for SID (Equal Error Rate in % and Expected Rank respectively). (GI-PLDA: Gender-independent PLDA, GD-PLDA: Gender-dependent PLDA.)

System	Male	Female
GI-PLDA	2.8, 1.0	5.6, 2.6
GD-PLDA	3.0, 1.1	5.4, 2.5
Metadata fusion	2.8, 1.0	**5.0, 1.0**
Fusion with virtual evidence	2.8, 1.0	**5.1, 2.3**

II, Switchboard Cellular, Switchboard Phase I, II and III, NIST SRE 2004, 2005 and 2006.

For the fusion experiments with GID, we train both gender-dependent (GD-PLDA) and gender-independent (GI-PLDA) systems. Gender-dependent (LI-PLDA) systems are trained for fusion experiments with LID. All systems are trained using Kaldi [16]. The i-vector systems are trained using the implementation in [17].

4.2. Fusing gender information

In this section, we discuss the results obtained after fusing gender information using techniques presented in Section 3. The results, presented in Table 1, are split by gender. However, the evaluation does not have any cross-gender trials. We noticed that errors introduced by cross-gender trials are insignificant.

In Table 1, the performance of the gender-dependent and gender-independent ASV/SID baselines are presented first. The results for "Metadata fusion", where gender information is assumed to be known, indicate reduced overall error rate over the baseline for individual systems. However, there are no changes to EER or ER for male speakers suggesting that the errors in the ASV/SID baseline are not due to modelling the gender information. The improvements on the female subset is significant. An relative improvement of 10% in EER for ASV task is observed when compared with GI-PLDA. This shows the importance of conditioning the back-end classifier to the domain and that the gender independent system may be imbalanced towards the male subsystem.

The performance achieved using virtual evidence is also promising. There is a graceful degradation of 0.1% in EER with respect to metadata fusion. However, the ASV results on the female subset are still significantly better than the female GI-PLDA baseline.

In terms of ER for SID task, metadata fusion outperforms both GI-PLDA and GD-PLDA baselines reducing from 2.6 to 1.0, which is the optimal rank that can be obtained. For the system using virtual evidence ER reduces from 2.6 to only 2.3 giving a relative improvement of 11%.

4.3. Fusing language information

In this section, we fuse the language information obtained from LID or that available from groundtruth to improve speaker recognition. The results are presented in Table 2 for male and female speakers. Comparing the results of gender-dependent systems in LI-PLDA and gender-specific performance of GI-PLDA or GD-PLDA, it can be seen that domain-independent training does not always achieve optimal performance. LI-PLDA is trained with multilingual data, that is some speakers may have samples from two (or more) languages. The LID problem is simplified by having only two classes: English and non-English. A dot-product based similarity measure is employed to decide if a speech recording is English or not. This reduction to a two-class problem was necessary to simplify the training of domain-dependent back-ends as non-English languages in the dataset do not have sufficient number of examples to estimate the hyperparameters independently. To handle cross-language trials, we use the LI-PLDA system as it has seen both classes.

Table 2. Results on the fusion speaker and language information for male speakers. The results are reported in terms of EER for ASV and ER for SID (Equal Error Rate in % and Expected Rank respectively). (LI-PLDA: Language Independent PLDA).

System	Male	Female
LI-PLDA	2.4, 1.0	4.8, 3.2
Metadata fusion	**2.2**, 1.0	**4.6**, **3.0**
Fusion with virtual evidence	**2.2**, 1.0	**4.6** , **3.0**

For the male system, the ER does not change as the SID baseline performance has already saturated. The EER improves for both fusion approaches by 0.2% absolute and 8% relative. Similar improvements in EER can be observed for female speakers where the ER also improves by 0.2 absolute. An analysis of the scores obtained from all the systems shows that the ER improves as result of the target speaker's rank improving in 4% of the test cases. Importantly, fusion with virtual evidence is effective and a reliable substitute for metadata fusion.

5. SUMMARY

A probabilistic approach to fuse metadata information available either from groundtruth or automatic systems for processing language and gender information was presented. A generic framework to fuse the scores from the gender and language identification systems with speaker posteriors was developed. The framework is capable of either using metadata or virtual evidence when there is uncertainty in the observation about metadata. Results presented on NIST SRE 2008 demonstrate that the ASV/SID systems are effective in utilising such side-information when available.

The proposed inter-task fusion has been successfully tested on real operational data (i.e. a case work) during a field-test under SiiP (Speaker identification integrated project[1]). As criminal activities have increasingly become a cross-border, law enforcement agencies need to deal with cross-lingual data. The analysed case-work has considered this type of data. More specifically, the basic speaker identification engine was first adapted toward target domain (i.e. YouTube recordings). Around 1000+ speakers (from VoxCeleb) were enrolled for the subsequent evaluations. As evaluation data, we had tens of speakers with cross-lingual (English and Arabic) speech recordings. The developed framework of speaker and language identification fusion with virtual evidence was applied. The results reveal that the expected rank has notably improved for most of the target speakers.

6. ACKNOWLEDGEMENT

This work was partially supported by Speaker Identification Integrated Project (SIIP), funded by the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607784.

[1] http://www.siip.eu

7. REFERENCES

[1] Khaled Khelif, Yann Mombrun, Gerhard Backfried, Farhan Sahito, Luca Scarpato, Petr Motlicek, Srikanth Madikeri, Damien Kelly, Gideon Hazzani, and Emmanouil Chatzigavriil, "Towards a breakthrough speaker identification approach for law enforcement agencies: Siip," in *Intelligence and Security Informatics Conference (EISIC), 2017 European*. IEEE, 2017, pp. 32–39.

[2] Chunlei Zhang, Fahimeh Bahmaninezhad, Shivesh Ranjan, Chengzhu Yu, Navid Shokouhi, and John H. L. Hansen, "UTD-CRSS systems for 2016 NIST speaker recognition evaluation," *CoRR*, vol. abs/1610.07651, 2016.

[3] Najim Dehak, Patrick Kenny, Rda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Tran. on Audio, Speech and Language Processing*, pp. 788–798, 2011.

[4] Daniel Garcia-Romero, Xinhui Zhou, and Carol Y Espy-Wilson, "Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4257–4260.

[5] T Wark and S Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Processing*, vol. 11, no. 3, pp. 169 – 186, 2001.

[6] Marc Ferras, Srikanth R Madikeri, Subhadeep Dey, Petr Motlícek, and Hervé Bourlard, "Inter-task system fusion for speaker recognition.," in *INTERSPEECH*, 2016, pp. 1810–1814.

[7] A.F. Martin and C.S. Greenberg, "Nist 2008 speaker recognition evaluation: performance across telephone and room microphone channels," *Annual Conference of the International Speech Communication Association (Interspeech)*.

[8] Richard O Duda, Peter E Hart, and David G Stork, *Pattern classification*, John Wiley & Sons, 2012.

[9] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[10] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV 2006*, pp. 531–542. Springer, 2006.

[11] David Barber, *Bayesian reasoning and machine learning*, Cambridge University Press, 2012.

[12] Douglas A Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[13] Alvin F Martin and Craig S Greenberg, "Nist 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[14] Petr Motlicek, Subhadeep Dey, Srikanth Madikeri, and Lukas Burget, "Employment of subspace gaussian mixture models in speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4445–4449.

[15] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," 2001.

[16] D. Povey, A. Ghoshal, et al., "The kaldi speech recognition toolkit," in *In Proc. of ASRU 2011*, December 2011.

[17] Srikanth Madikeri, Subhadeep Dey, Petr Motlicek, and Marc Ferras, "Implementation of the standard i-vector system for the kaldi speech recognition toolkit," Tech. Rep., No. EPFL-REPORT-223041 Idiap, 2016.