Speech Trax: A Bottom to the Top Approach for Speaker Tracking and Indexing in an Archiving Context

Félicien Vallet[1], Jim Uro[2], Jérémy Andriamakaoly[3], Hakim Nabi[1], Mathieu Derval[1], Jean Carrive[1]

[1]Institut National de l'Audiovisuel, 4 avenue de l'Europe, 94366 Bry-sur-Marne Cedex, France
[2]Université de Technologie de Compiègne, rue du Dr Schweitzer, 60200 Compiègne, France
[3]Télécom ParisTech, 46 rue Barrault, F-75634 Paris Cedex 13, France
firstname.lastname@{ ina.fr[1], etu.utc.fr[2], telecom-paristech.fr[3] }

Abstract

With the increasing amount of audiovisual and digital data deriving from televisual and radiophonic sources, professional archives such as INA, France's national audiovisual institute, acknowledge a growing need for efficient indexing tools. In this paper, we describe the Speech Trax system that aims at analyzing the audio content of TV and radio documents. In particular, we focus on the speaker tracking task that is very valuable for indexing purposes. First, we detail the overall architecture of the system and show the results obtained on a large-scale experiment, the largest to our knowledge for this type of content (about 1,300 speakers). Then, we present the Speech Trax demonstrator that gathers the results of various automatic speech processing techniques on top of our speaker tracking system (speaker diarization, speech transcription, *etc.*). Finally, we provide insight on the obtained performances and suggest hints for future improvements.

Keywords: Demonstration, Audiovisual Indexing, Speech Processing, Speaker Tracking, i-Vectors, TV and Radio shows

1. Introduction

INA[1], France's national audiovisual institute is the world's leading source of digitized audiovisual content. It collects and preserves 80 years of radio archives and 70 years of television programs that form the French collective memory. The missions of archives such as INA being to store but also describe and enhance content, the need for automatic structuring methods is of utmost importance. If storage is not the main issue anymore, indexing methods remain ill-suited for the retrieval of information in multimedia databases. Thus, an important human effort for description is required to enhance the stored data.

In this context, many research works aim at automatically organizing and structuring large quantities of audiovisual documents. Speech processing technologies allow the extraction of valuable information from the audio signal. Discrimination of speech and music zones, automatic segmentation in speaker turns or automatic speech transcription, there are nowadays many tools available. Among these, speaker tracking – the task of finding spoken segments of a particular speaker for which some training material is given – is of great interest. Indeed, famous voices haunt the audiovisual heritage. Politicians, athletes, intellectuals, anchormen, *etc.* our broadcast history resonates with heroic, tender, comical or dramatic accents and the ability to retrieve such persons' interventions in the archive is of great value.

2. Related works

As stated in (Kinnunen and Li, 2010) speaker tracking is derived from the larger field of the speaker recognition technologies that has been a very hot research topic for several decades. However, the challenge of dealing with "big data" is just getting tackled. For instance, in (Jeon and Cheng, 2012), the authors propose a statistical utterance comparison that, coupled with kernelized locality-sensitive hashing (KLSH), they use to retrieve very large population of speakers (about 10,000). Similarly, in (Schmidt et al., 2014), the authors propose a system based on i-vector, the state-of-the-art approach for speaker recognition, and locality-sensitive hashing (LSH) to speed-up the the comparison of a given target with the referenced 1,000 speaker models. Finally, it is worth mentionning the evaluation campaigns Speaker Recognition Evaluation and i-vector Machine Learning Challenge organized and led by the NIST[2].

However, it has to be noted that in all these cases, the speech segments used for the speaker recognition task are not issued from broadcast material. They either come from databases designed for the research on consumer devices as in (Jeon and Cheng, 2012), telephonic and microphone speech as in the NIST challenges or technical talks posted on YouTube as in (Schmidt et al., 2014).

Speaker tracking in audiovisual content is nevertheless

[1]Institut National de l'Audiovisuel: www.ina.fr

[2]NIST SRE and i-Vector Machine Learning Challenge: http://www.itl.nist.gov/iad/mig/tests/spk/ - https://ivectorchallenge.nist.gov/

slowly getting attention. In (Huijbregts and van Leeuwen, 2010), based on speaker diarization techniques, the authors propose to link speakers in an unsupervised fashion for about 1,800 hours of Dutch television broadcasts. Similarly, the prototype developed at BBC[3] gathers 3 years of radio archives amounting to 70,000 programs. In (Raimond and Nixon, 2014), the authors present the speaker recognition feature of the prototype. In this case, 780 speaker models are built, based on Gaussian Mixture Models (GMM), and compared using Kullback-Leibler divergence through a LSH index. Finally, an approach relying on TV material issued form the REPERE challenge (Giraudel et al., 2012) is proposed in (Fredouille and Charlet, 2014). In particular, the authors investigate the i-vector framework and propose a specific protocol to identify candidates using a 533 speakers dictionary.

3. Speaker Recognition

In this section, we focus on the speaker recognition feature of our demonstrator. In particular, we describe how our speaker dictionary is created and detail the implementation choices made for the speaker recognition system itself.

3.1. Speaker Dictionary Constitution

The creation of large-scale multimedia datasets has become a scientific matter in itself. Indeed, the fully-manual annotation of hundreds or thousands of hours of video and/or audio turns out to be practically infeasible. We present here the semi-automatic approach we followed to construct what is, to our knowledge, the largest speaker database issued from broadcast material.

Automatic Collection of Speech Segments

As detailed in our previous work (Salmon and Vallet, 2014) we propose to automatically gather segments for famous speakers. It is based on the simple hypothesis that states that during a TV newscast, when the name of a person appears on screen, that person is presently speaking.

Figure 1: Association between a personality name appearing on screen and a speech turn in a TV newscast.

To this end, an optical character recognition (OCR) software presented in (Poignant et al., 2012) is used to detect names (see Figure 1). A comparison between the transcribed text and a list of referenced people issued from INA's thesaurus is performed using the Levenshtein distance. Then, if a match is found, the corresponding speech turn obtained using the LIUM speaker diarization tool (Rouvier et al., 2013) is associated with the person identity. A total of about 5,000 hours of TV broadcast news have been processed this way. In (Salmon and Vallet, 2014), we present a thorough error analysis of this approach. It highlights that the assumption is correct in most cases – about 72% of the time – but also that great disparities are observed between speakers, few of them showing enough speaking time to produce a reliable speaker model.

In order to counterbalance the produced speaker dictionary, we propose to collect extra speech segments for personalities not appearing often on TV. For this we use an automated video query system on Google Video as well as on Ina.fr website. The returned videos are downloaded and the LIUM speaker diarization tool is run. Then the segments of the person with the longest speaking time are labeled with the name of the personality seeked in the query.

Manual Validation of the Speaker Dictionary

Speaker recognition methods rely very heavily on the quality of the training data to perform correctly. Unfortunately, the automated methods described previously don't allow to reach this necessary quality.

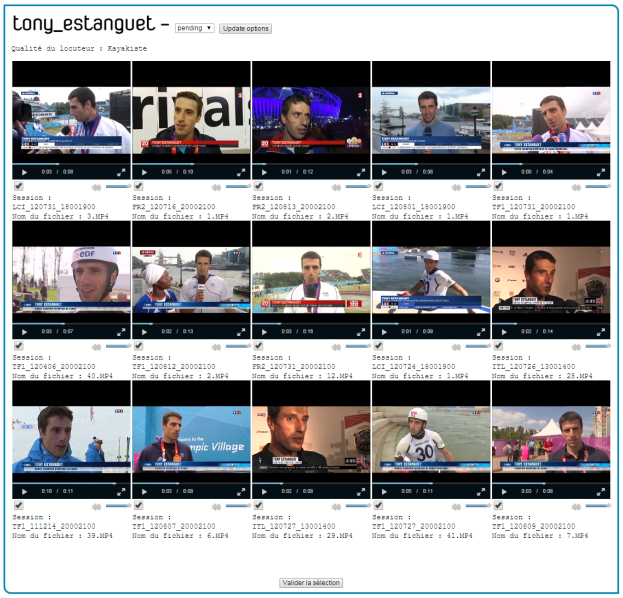

Figure 2: Web interface for the verification of the automatically collected speech segments.

Therefore, a web interface has been designed to manually validate the segments attributed to a given personality (see Figure 2). It allows us to confirm or infirm for each collected segment the identity of the assumed speaker. At the end of the validation process, a total of 2,290 personalities constitute our speaker dictionary (Table 1 shows the detail). However it has to be kept in mind that, despite our efforts to reduce it, such dictionaries are by nature greatly imbalanced.

On top of the number of files and the cumulated time both in total and on average, Table 1 displays the number of sessions. We qualify as belonging to the same acoustic session

[3]The World Service Radio Archive:
www.bbc.co.uk/rd/projects/worldservice-archive-proto

speech segments issued from the same program. This information is of great importance. Indeed for the evaluation of the speaker recognition process, segments belonging to the same session cannot be part of both the training and the testing sets for a given speaker. It would otherwise lead to biaised results since similar acoustic conditions would be both learned and evaluated. However, this claim must be put into perspective. Indeed, it appears that while particularly relevant for news reports, where acoustic conditions vary a lot (background noise, indoor/outdoor differences, room's reverberation, *etc.*), this distinction in acoustic conditions is less notable for studio recordings where settings are quite similar across programs.

	sessions	files	time
total	17,438	32,053	435,254s (∼120h)
average	7.6	14.0	190s (∼3mn10s)

Table 1: Details of the composition of the speaker dictionary in total and on average per speaker.

3.2. Implemented method

State-of-the-art techniques for speaker recognition rely now on the i-vector paradigm. As stated in (Dehak et al., 2011), an i-vector is a compact representation of a speaker's utterance after projection into a low-dimensional, total variability subspace trained using factor analysis and making no distinction between speaker and channel/session information. The speaker and channel dependent GMM supervector, M, issued from the concatenation of speaker GMM means can be defined as :

$$M = m + Tw$$

where m is the mean supervector issued from the Universal Background Model (UBM) representing the speaker and session/channel independent information, the low-rank matrix T defines the total variability space and w represents the speaker and sessions/channel dependent factors in the total variability space, also called i-vectors.
We chose to use a similar system as the one described in (Fredouille and Charlet, 2014) meaning that we rely on the ALIZE v3.0 toolkit (Larcher et al., 2013). However, based on a series of preliminary experiments, we decided to adopt the Probabilistic Linear Discriminant Analysis (PLDA) scoring approach defined in (Prince and Elder, 2007). The Universal Background Model (UBM), the total variability matrix T and the PLDA-related matrices were learnt on data issued from the REPERE challenge (Giraudel et al., 2012).

4. Close-Set Experiment

In order to evaluate the performances of the previously described speaker recognition system, we build a close-set experiment exploiting our speaker dictionary.

4.1. Protocole

In a close-set evaluation all the tested speakers or targets possess a corresponding voice model. For this, the speaker dictionary is split in two parts, one dedicated to the training of the models and the other to the testing. Contrary to the protocole presented in (Schmidt et al., 2014) we take great care not to mix-up, for a given speaker, segments belonging to the same session (*i.e.* issued from the same program) both in the train and in the test set. Besides, to try to ensure a balanced volume of data among the various speakers, we limit the maximum number of sessions considered to 10 in total. We then allocate the sessions and segments in the same fashion as in (Fredouille and Charlet, 2014)), meaning that we set a minimum of 30 seconds and a maximum of 150 seconds for the training phase. Thus, with all this constraints, the number of sessions to be added in the train base is determined as follow for a speaker i:

$$n^i_{\text{train sessions}} = min\left(5; \frac{n_{\text{total sessions}}}{2}\right)$$

$$\text{with 30 sec.} \leq \sum_{j=1}^{n^i_{\text{train sessions}}} \sum_{k=1}^{n^j_{\text{segments}}} duration_{\text{seg } k} \leq 150 \text{ sec.}$$

$$n^i_{\text{test sessions}} = n^i_{\text{total sessions}} - n^i_{\text{train sessions}}$$

It has to be noted that a great variance is observed in the number of segments belonging to one session. Thus, to ensure that the process is statistically significant, we randomly generate 5 training and testing sets and average performances.

4.2. Results

Table 2 displays the averaged results obtained with our speaker recognition system over 1,290 speakers satisfying the previously described conditions. Scores are computed both at the segment and at the session level.

	EER	precision	recall	prec@10	rec@10
segment	7.3%	63.6%	65.9%	89.4%	83.5%
session	6.2%	69.5%	67.4%	94.8%	84.4%

Table 2: Closed-set results.

Besides, the classical metrics precision, recall and equal error rate (EER, the rate at which both acceptance and rejection errors are equal) we prompt prec@10 and rec@10. These metrics allow to hint if the targeted speaker is amongst the first 10 retrieved. It is a slightly different approach than the classical precision-at-n measure that is based on result pages of web search and is defined as $P@n = r/n$ with r the number of relevant documents retrieved at rank n.
The results highlight the fact that the identifiation by session performs better than by segment which is in adequation with the results obtained in (Fredouille and Charlet, 2014) with a referenced segmentation. Also, one can note that compared with classical metrics, the prec@10 et rec@10 measures highlight the fact that targeted speakers are most of the time to be found in the first 10 retrieved candidates.

5. The Speech Trax demonstrator

The idea of the Speech Trax[4] demonstrator is to propose new ways of exploring audiovisual collections based

[4]Speech Trax demonstrator: http://speechtrax.ina.fr

on oral interventions of famous French speakers. On an archiving point of view it also is to issue a raw and imperfect documentation of the contents kept at INA to help their detailed description. To this end, we use a similar approach to the one presented in (Raimond and Nixon, 2014).
As illustrated in the block diagram Figure 3, Speech Trax relies on various automatic speech processing techniques: speech/music discrimination, speaker diarization, speaker recognition and speech transcription.

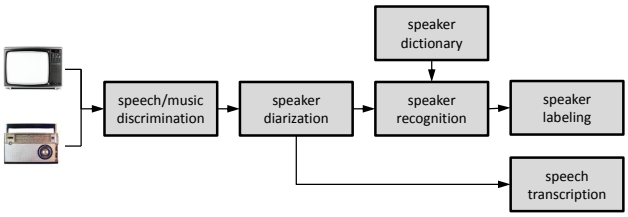

Figure 3: Overall architecture of the Speech Trax system.

5.1. Settings

A corpus of about 250 hours of broadcast news and magazines is selected. It comes from 6 public channels: 3 for TV (France 2, France 5, France 24) and 3 for radio (France Inter, France Info, France Culture). Choice was made to process data from March 2014 due to the important events that occured at that time: French local elections, phone tapping of former president Sarkozy, Ukraine invasion by Russian troops, missing Malaysian Airlines flight 370, *etc.*
Programs are cut in 15 minute slices both to reduce speaker diarization errors and to ease the browsing of the media in the user interface. Besides, adds are manually discarded which is the only manual operation that is performed.

5.2. Implentation

As described in Figure 3, the first step is to identify speech tracks by running a speech/music discrimination methods. To this end, we used the approach proposed in (Pinquier and André-Obrecht, 2006). Once the speech tracks identified, the LIUM speaker diarization tool (Rouvier et al., 2013) is used to provide a speaker segmentation and clustering of the analyzed shows. It then allows to track speakers using the speaker recognition system described earlier. However, in this case, it is worth mentioning that tested speech segments are produced by speaker diarization and not manually validated, inducing potential errors. From the dictionary described in section 3.1., a sub-dataset of 1,783 speakers possessing at least 45 seconds of speech is used. Finally, the commercial system VoxSigma from Vocapia Research is used to provide automatic speech transcriptions.

5.3. Browsing data

The Speech Trax's GUI is a responsive and flat design web-application with all actions accessible on the same single page. The video player used is amalia.js, INA's HTML5 open-source player[5] (Hervé et al., 2015). Figure 4 provides a view of the metadata enriched player amalia.js.
Speech Trax enables a user to navigate inside a corpus of radio and video documents according to the interventions of famous speakers. For this, the user can enter the name of a person he wants to find in the search bar. An autocompletion module will show him if this speaker has been identified in the corpus. Then, the user just needs to select the excerpt of his choice to listen to the speaker's intervention. For TV material, the images of the excerpts are automatically extracted based on the speaker's interventions. Once a program is shown on screen, the user can also browse through the media and access all the interventions of an identified speaker by clicking on the magnifying glass next to his/her name. On the right of the name the user can also take notice of the confidence with which the labeling was made.

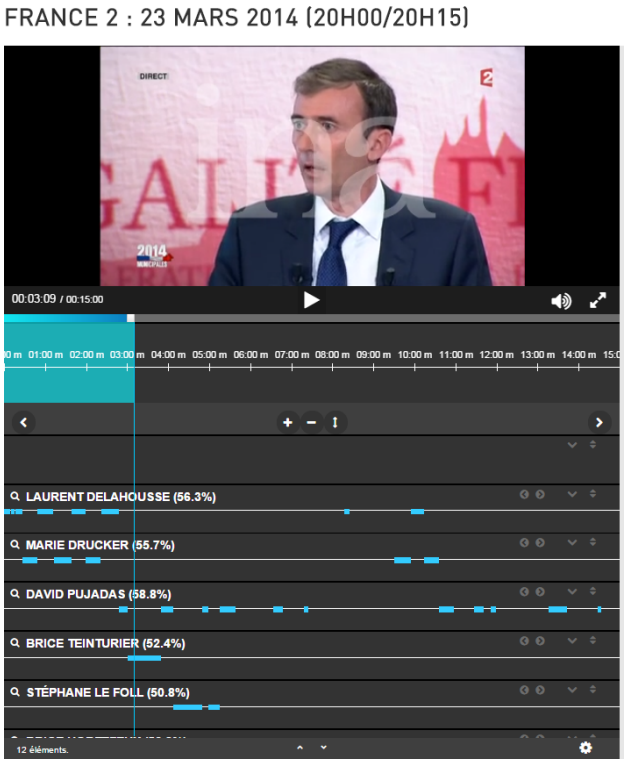

Figure 4: Main view of the Speech Trax demonstrator.

5.4. Performances

The processing of the described corpus enabled the identification of 533 unique speakers, most of them being anchormen, presenters, sportsmen, politicians or experts. It has to be noted that the validation threshold has deliberately been set pretty high in order to increase the user experience. For a professionnal usage, for instance to help collecting interventions for a given speaker, this threshold would be set lower to ensure a greater exhaustivity. However, in this scenario the manual validation of the retrieved speech segments would be necessary.
Better performances appear to be obtained for radio than for TV material. An argument could be made that maybe radio speech is cleaner since no visual cues are available to the listener. Also, it is worth noting that, as highlighted

[5] amalia.js, metadata enriched HTML5 video player: http://ina-foss.github.io/amalia.js/

in Table 3, there are on average less speakers in the radio programs processed than in their TV counterparts.

	FR2	FR5	F24	FIT	FIF	FCR
nb. speakers	19.1	9.0	11.8	11.7	12.5	8.5

Table 3: Number of speakers, identified or not, according to the channel over 15 minutes (the first 3 channels are TV, the last 3 radio).

The repartition of the identified speakers across channels is also of interest. Table 4 reveals that the vast majority of speakers is retrieved on a single channel. A closer look at that population prompts that these persons are for the greater part anchormen, presenters, columnists, *etc.* On the other end of the spectrum, a handful of personalities are retrieved on all or almost all channels. It is interesting to note that all these personalities are politicians: François Hollande, Laurent Fabius, Jean-Claude Gaudin, Marine Le Pen, Jean-François Copé, *etc.* At the same time, it is also worth keeping in mind that politicians are among the speakers with the most training segments in the dictionary on account of their regular appearances on TV.

	1 ch.	2 ch.	3 ch.	4 ch.	5 ch.	6 ch.
nb. speakers	389	85	33	15	8	3

Table 4: Repartition of the identified speakers.

Finally, following the famous zoo introduced in (Doddington et al., 1998), one can notice that several speakers appear to be wolves, meaning that they are particularly successful at imitating other speakers. That is, their speech is very likely to be accepted as that of another speaker. Personalities such as Benjamin Millepied, Manu Payet, Béatrice Idiard-Chamois certainly appear to be wolves in our demonstrator which seems due to noisy speech models. Finally, it is also worth noting that many errors are caused by telephone speech. For instance, the speaker Alain Cayzac, who must have telephonic data in his speech model, is often identified when a speaker is interviewed on the phone. If needed in the future, a simple pass-band identification would easily enable to discard such segments.

6. Conclusion and Perspectives

Relying on various state-of-the-art speech processing techniques, Speech Trax is a first attempt to index and retrieve famous speakers in INA's archiving context. Results are very encouraging. Future uses of Speech Trax could allow users to navigate differently in archives. For instance by generating new queries: "find media where A speaks with B" or "get me contents where C talks about D", *etc.*
However a boost of performances could be obtained by using multimodality to confirm, correct or invalidate the identity of detected speakers. Technologies such as speech transcription, optical character recognition or face recognition could be directly plugged-in to enhance the identification results as it is done in the MediaEval task "Multimodal person discovery in broadcast TV" (Poignant et al., 2015). Besides, if Speech Trax were to be used at INA's industrial scale, it would be necessary to resort to locality-sensitive hashing techniques. Also, as in (Raimond and Nixon, 2014), INA could rely on crowdsourcing to extend the size of the speaker dictionary but also to clean it when necessary, allowing a steady improvement of the system.

ACKNOWLEDGMENT

The authors wish to thank the developers who participated in this project and who can not be named for legal reasons. They also would like to thank Élisabeth Chapalain from Ina for her help validating speech segments in the speaker dictionary as well as academic partners Dr. Sylvain Meignier and Dr. Anthony Larcher from LIUM and Dr. Julien Pinquier from IRIT.

7. References

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788 – 798.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., and Reynolds, D. (1998). Sheep, goats, lambs and woves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *International Conference on Spoken Language Processing*, Sydney, Australia, december.

Fredouille, C. and Charlet, D. (2014). Analysis of i-vector framework for speaker identification in tv-shows. In *Interspeech*, Singapore, september.

Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., and Quintard, L. (2012). The REPERE Corpus : a multimodal corpus for person recognition. In *International Conference on Language Resources and Evaluation*, Istanbul, Turkey, may.

Hervé, N., Letessier, P., Derval, M., and Nabi, H. (2015). Amalia.js : an open-source metadata driven html5 multimedia player. In *ACM Conference on Multimedia*, Brisbane, Australia, october.

Huijbregts, M. and van Leeuwen, D. (2010). Towards automatic speaker retrieval for large multimedia archives. In *International Workshop on Automated Information Extraction in Media Production*, Florence, Italy, october.

Jeon, W. and Cheng, Y.-M. (2012). Efficient speaker search over large populations using kernelized locality-sensitive hashing. In *International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, march.

Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 52(1):12 – 40.

Larcher, A., Bonastre, J.-F., Fauve, B., Lee, K.-A., Lévy, C., Li, H., Mason, J., and Parfait, J.-Y. (2013). Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *Interspeech*, Lyon, France, september.

Pinquier, J. and André-Obrecht, R. (2006). Audio indexing: Primary components retrieval - robust classification in audio documents. *Multimedia Tools and Applications*, 30(3):313 – 330.

Poignant, J., Besacier, L., Quénot, G., and Thollard, F. (2012). From text detection in video to person identification. In *International Conference on Multimedia and Expo*, Melbourne, Australia, july.

Poignant, J., Bredin, H., and Barras, C. (2015). Multimodal person discovery in broadcast tv at mediaeval 2015. In *MediaEval 2015 Workshop*, Wurzen, Germany, september.

Prince, S. J. and Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *International Conference on Computer Vision*, Rio de Janeiro, Brazil, october.

Raimond, Y. and Nixon, T. (2014). Identifying contributors in the BBC World Service Archive. In *Interspeech*, Singapore, september.

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., and Meignier, S. (2013). An open-source state-of-the-art toolbox for broadcast news diarization. In *Interspeech*, Lyon, France, august.

Salmon, F. and Vallet, F. (2014). An effortless way to create large-scale datasets for famous speakers. In *International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, may.

Schmidt, L., Sharifi, M., and Moreno, I. L. (2014). Large-scale speaker identification. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, Florence, Italy, may.