SPEAKER DIARIZATION THROUGH SPEAKER EMBEDDINGS

Mickael Rouvier[1]*, Pierre-Michel Bousquet*[2]*, Benoit Favre*[1]

[1] Aix-Marseille Université, CNRS, LIF UMR 7279, 13000, Marseille, France
[2] Avignon Université, LIA, 84000, Avignon, France

ABSTRACT

This paper proposes to learn a set of high-level feature representations through deep learning, referred to as Speaker Embeddings, for speaker diarization. Speaker Embedding features are taken from the hidden layer neuron activations of Deep Neural Networks (DNN), when learned as classifiers to recognize a thousand speaker identities in a training set. Although learned through identification, speaker embeddings are shown to be effective for speaker verification in particular to recognize speakers unseen in the training set. In particular, this approach is applied to speaker diarization. Experiments, conducted on the corpus of French broadcast news ETAPE, show that this new speaker modeling technique decreases DER by 1.67 points (a relative improvement of about 8% DER).

Index Terms— Speaker Diarization, Deep Neural Network, Speaker Embeddings, Speaker Clustering, i-vector

1. INTRODUCTION

The goal of speaker diarization is to annotate temporal regions of audio recordings with speaker labels, in order to answer the question "who spoke when". A common approach to this task is to perform two steps: namely segmentation of the input speech so that each speech segment belongs to one speaker, and segment clustering in order to regroup all segments of the same speaker. The challenge of speaker clustering is increased by the fact that clustering is performed without prior knowledge of the number of speakers nor their identity.

In speaker diarization, state-of-the-art speaker modeling is based on the i-vectors/PLDA pipeline [1]. Introduced in [2], the i-vector approach provides an elegant way of reducing a large-dimensional input vector (representing the speaker data) to a small-dimensional feature vector, while at the same time retaining most of the relevant information. During speaker clustering the metric used to verify if two i-vectors correspond to the same speaker is based on Probabilistic Linear Discriminant Analysis Scoring (PLDA). Although this approach obtains considerable gains compared to GMM-UBM (Gaussian Mixture Model-Universal Background Model), i-vectors are very sensitive to segment duration.

Indeed, i-vectors are extracted on total variability space, no distinction is made between speaker and channel variation. PLDA scoring is used to disentangle the channel and speaker effects [1]. But it is now well established that the limitation of the i-vector representation of speech segments become apparent when processing short segments, where it's very difficult to disentangle the channel and speaker effects [3, 4, 5, 6]. In order to tackle this problem, we propose to directly estimate a high-level feature representation in the speaker space using DNNs.

We propose to learn high-level speaker identity features with deep models through speaker identification, i.e. classifying speech segments into one of n speaker identities (n = 1.014 in this work). In that context, the hidden layers of the Deep Neural Networks (DNN) are learned to extract information relevant for discriminating between speakers. The main idea is to use one of the hidden layers as the new feature representation. We call this representation a speaker embedding.

The novel contributions in our work are the following:

- We propose an original estimatation a high-level feature representation (called *speaker embedding*) that contains all speaker-specific information, trained with DNNs.
- The new representation is built without Gaussian prior assumptions. We propose an analysis which shows that the PLDA preceded by a conditioning method, as done for i-vectors, can be relevant for speaker embeddings.
- Speaker embeddings are leveraged for Speaker Diarization, leading to a relative improvement of about 8% DER over the baseline diarization system based on i-vector/PLDA.

Our experiments on the ETAPE corpus show that speaker diarization based on i-vectors obtains 20.92% Diarization Error Rate (DER), whereas the proposed approach based on speaker embeddings obtains 19.25% DER (an absolute gain of 1.67 points).

After presenting the related work in Section 2, the speaker embedding paradigm is presented in Section 3. Section 4 describes the speaker verification methods used in speaker diarization. Then, in Section 5, we present a conditioning algorithm to conform speaker embeddings with PLDA's asumptions. The corpus on which experiments are carried and the results of our experiments are presented in Section 6. Finally, we conclude with a discussion of possible directions for future works in Section 7.

2. RELATED WORK

Most methods used for extracting speaker-specific features in speaker diarization come from the speaker verification community.

The i-vector framework [2] is considered as a state-of-the-art method in speaker verification and identification. The idea of i-vectors is to reduce a super-vector to a compact vector. This dimensionality reduction has opened a wide range of perspectives such as the use of Bayesian methods for removing channel variability in the i-vector [1].

Modeling a speaker in the acoustic space has been tackled by many works in the literature. In [7, 8], the authors propose to introduce the technique of anchor modeling. The basic concept of anchor modeling is the representation of a target speech utterance

This work has been carried out thanks to the support of the A*MIDEX project (n° ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French Government program, managed by the French National Research Agency (ANR).

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

with information gained from a set of models pre-trained from a defined set of speakers. Segments of speech are scored against a set of pretrained anchor models. Each of the anchor models yields a likelihood score and the collection of scores is used to form the characterization vector. The characterization vector can be considered a projection of the target utterance into a speaker space defined by the anchor models. Similarly, our approach trains a model to recognize a predetermined set of speakers and extracts features for new speakers through the activations of that model. However, we take advantage of hidden layers in DNNs to leverage a more general representation.

Representation learning has led to interesting improvements in various domains, such as face recognition [9], text modeling [10] or speech recognition [11]. In particular, it relieve researchers of designing features by automatically learning the relevant structure of the input space. As will be shown in the next sections, we strive to take advantage of such advances in order to structure the acoustic space towards relevant features for discrimnating between speakers.

3. SPEAKER EMBEDDINGS

The proposed method differs from the majority of contributions in the field: it learns high-level speaker features with deep models trained to achieve a speaker identification task. When learning a classifier to recognize speaker identities, DNNs compact relevant features in the hidden layers. We propose to create a feature vector from the hidden layer neuron activations, which we call *Speaker embedding*. Although learned through identification, speaker embeddings are shown to be effective for speaker verification, in particular to recognize speakers unseen in the training set. The main idea is to use one of the hidden layers as the new feature representation.

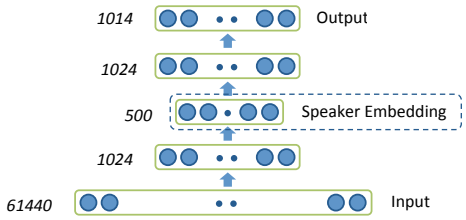

Fig. 1. An illustration of the feature extraction process. Arrows indicate forward propagation direction. The number of neurons in each layer of the deep neural network is labeled besides each layer. Speaker Embedding features are taken from the central hidden layer (this topology obtains the best results on the development corpus).

In our experiments, the DNN is made of five layers from which three are hidden. The output layer is a soft-max layer, and the outputs represent speaker identity classes (there are 1.014 states in our experiments). The number of neurons in the hidden layer is the same for all hidden layers: 1024 neurons (except for the speaker embedding layer). The nonlinearities in the hidden layers are Rectified Linear Unit (ReLu) functions. The objective function is the cross-entropy criterion, i.e. for each frame, the log-probability of the correct class. The weights are updated using mini-batches of size 128 frames. All these parameters are determined on the development corpus in order to obtain the best results.

The input vector is the super-vector obtained from the GMM-UBM of dimension 61440, computed as:

$$s_g = \frac{1}{\sum_t \gamma_g(t)} \sum_t \gamma_g(t)(x_t - \mu_g) \tag{1}$$

where $\gamma_g(t)$ is the posterior probability estimated on the g-th Gaussian component on the frame t, x_t corresponds to the frame t and μ_g is the mean of the GMM-UBM.

Figure 2 shows 500-dimensional Speaker Embeddings extracted from the test corpus for selected speakers. This figure illustrates how speech segments from the same speaker tend to have more activated neurons in common.

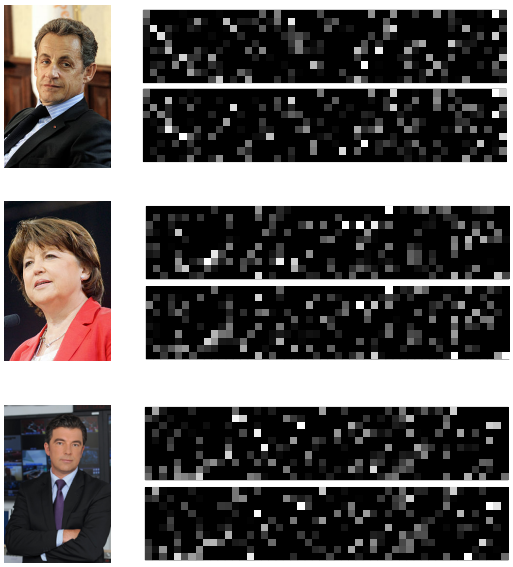

Fig. 2. Examples of the learned 500-dimensional Speaker Embeddings. The figure shows three test pairs from the test corpus. We rearrange them as 10×50 for the convenience of illustration, the ordering of the feature vector is the same for all examples. Feature values are non-negative since they are taken from ReLUs. Approximately 67% of features have non-null values. Brighter squares indicate higher values.

4. SPEAKER VERIFICATION

In speaker diarization, the cluster step requires computing the similarity between pairs of speech segments. This speaker verification step has been successfully performed with PLDA in previous work [1]. PLDA is a probabilistic version of Linear Discriminant Analysis (LDA). This technique projects the input data into a much lower dimensional space with minimal loss of discriminative ability, as the ratio of between-speaker and within-speaker variation is maximized [12].

Speaker verification scoring can be computed as a log-likelihood ratio: Given two speaker embeddings w_i and w_j, the more likely hypothesis is that ($\mathcal{H}_{tar}$) w_i and w_j comes from the same speaker, or that ($\mathcal{H}_{non}$) the speaker embedding come from different speakers. The speaker verification score can be computed as:

$$d(w_i, w_j) = \log \frac{p(w_1, w_2|\mathcal{H}_{tar})}{p(w_1, w_2|\mathcal{H}_{non})} \tag{2}$$

In the PLDA model this equation can be written as:

$$d(w_i, w_j) = w_i^\intercal Q w_i + w_j^\intercal Q w_j + 2 w_i^\intercal P w_j \tag{3}$$

with

$$\begin{aligned} P &= \Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1} \\ Q &= \Sigma_{tot}^{-1}\Sigma_{ac}(\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1} \end{aligned} \tag{4}$$

where $\Sigma_{tot} = VV^{\intercal} + \Sigma_{\text{PLDA}}$ and $\Sigma_{ac} = VV^{\intercal}$. Here, V and Σ_{PLDA} are obtained from the PLDA estimation algorithm which is detailed in [1].

5. SPEAKER CONDITIONING

The main issue still to be addressed is to determine the speaker model and detector to be applied to the new representation. Figure 5 shows the 2D projection of four speakers in the training corpus, using the Principal Component Analysis (PCA) dimensionality reduction.

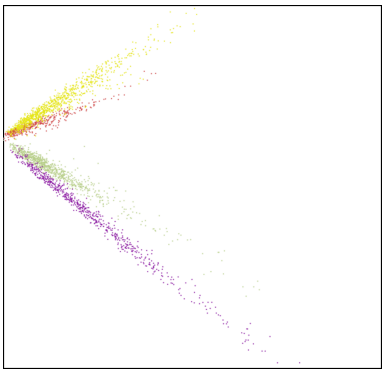

Fig. 3. 2D projection of four speaker embedding in the training corpus, using PCA. Each color represents a speaker and each point represents a segment.

Two comments can be made from this figure. First, the speaker-class distribution has a radial shape. Speaker-class means tend to lie in separate directions. Secondly, there are severe within-class distortions. Figure shows a marked dilatation of the vectors for each speaker from the origin of the space. These observations lead to propose the cosine distance scoring [13] as the final decision score. This metric only focuses on directional proximity, ignoring the length of vectors. But they also suggest that the (standardization / length-normalization / PLDA) solution used for i-vectors may be relevant for the new representation. Length-normalization ignores the vector length and is also known to improve Gaussianity of vectors [14]. By moving data towards a high Gaussian density surface, this technique helps to fit PLDA model to the training set.

We propose to carry out experiments on speaker embeddings based on cosine scoring, then on PLDA modeling preceded by $L\mathbf{W}$-normalization. Introduced in [15][1], this transformation iterates standardization according to the within-class covariance and length-normalization, in order to move data to an isotropic model. Figure 5 shows the spectral graph of our training set before (*0 iteration*) and after two iterations of $L\mathbf{W}$-normalization. Also introduced in the field in [15], this graph displays the variances of the 500 speaker embedding dimensions (*total*), then the proportion of variance due to the between-class (*speaker*, gray line) and within-class (*session*, dashed line) variabilities. As shown in the upper graph, between and within-class variances are initially correlated. After 2 iterations of the transformation, within-class variability is close to isotropy, revealing around 200 dimensions on which the major proportion of variance is due to between-class variability. In view of this graph, applying PLDA with a 200 eigenvoice-rank seems relevant.

[1] and referred to as *Spherical nuisance normalization* in the original paper.

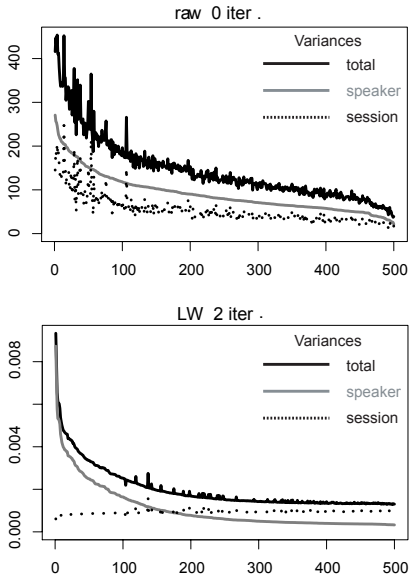

Fig. 4. Spectral graph of training data before and after 2 iterations of $L\mathbf{W}$ normalization. For the 500 dimensions (x axis) of speaker embeddings, the y axis shows the total, speaker and session variances.

6. EXPERIMENTS

6.1. Speaker Diarization

The diarization system used in experiments is the LIUM Speaker Diarization system [16][2]. This system obtained the best results during the ETAPE 2012 and REPERE 2012 French evaluation campaigns.

The speaker diarization system relies on two major steps: segmentation and speaker clustering. The purpose of this segmentation is to produce homogeneous segments that can be exploited in the next steps (i.e., a segment must match a single speaker). Segmentation is performed using a Generalized Likelihood Ratio (GLR) criterion based on GMMs. The next step aims to regroup all segments that belong to the same speaker. Speaker clustering is performed in two steps: BIC Clustering based on GMMs followed by Integer Linear Programming (ILP) Clustering based on i-vectors [17].

In this work, we propose to substitute i-vectors for speaker embeddings during ILP Clustering.

6.2. Data

In the following experiments, we use the ETAPE 2012 data [18]. The data consists in 7 different shows from French TV channels and French Radios, split according to the official train, development and test sets. The development corpus corresponds to 15 recordings (6h53 hours) and is employed to determine the various thresholds of the systems. The evaluation corpus contains 15 recordings (6h57 hours) and is employed to evaluate model performance. For training, we use ESTER1, ESTER2, REPERE and ETAPE training corpora. I-vectors, speaker embeddings and GMM-UBM models are all learned from the training corpus.

We note that on the dev corpus 28.57% speaker are present in the training corpus and on the test corpus 35.25% speakers are present in the training corpus.

[2] Freely distributed at http://www-lium.univ-lemans.fr/diarization/

6.3. Evaluation Metrics

Diarization Error Rate (DER) is the metric used to measure performance in speaker diarization. DER is the fraction of speaking time which is not attributed to the correct speaker, using the best matching between references and hypothesis speaker labels.

$$DER = \frac{\#Spk + \#Miss + \#FA}{\#Total} \quad (5)$$

where $\#Spk$, $\#Miss$ and $\#FA$ are respectively speaker error, missed speech and false alarm speech. The scoring tool we used was developed by the LNE as part of the ETAPE campaign [19].

6.4. I-vectors and Speaker embeddings

Throughout the experiments, speaker embeddings and i-vectors are extracted using 60-dimensional acoustic features, with a 10ms frame rate, composed of 19 MFCCs plus log energy and augmented by the first and second-order deltas. The UBM used for the features is a gender- and channel-independent GMM composed of 1024 Diagonal Gaussians computed with the Kaldi toolkit [20].

The dimension of i-vectors is fixed to 200 (determined on the development corpus from range between 50 and 600). The i-vectors are conditioned with two iterations of $L\mathbf{W}$-normalization.

Concerning speaker embeddings, the DNN used for extracting the speaker embedding is composed of 3 hidden layers of respectively 1024, 500 and 1024 units (the middle one is used for the embedding). This topology obtains the best results on the development corpus. The activation function of the DNN is ReLu. The learning rate was initialized at 0.01 and reduced at the end to 0.001. The weights are updated using mini-batches of size 128 frames. And the model is trained over 6 iterations. The DNN implementation is that of the Kaldi toolkit.

6.5. Results

Table 1 shows the results of a first experiment where we look at which hidden layer has the most potential for being extracted as representation (cosine distance without any normalization). Performance is reported on the test set according to the size of the layer from which embeddings are drawn.

Layer/Dim	300	400	500	600	700
Layer_1	22.11	22.38	20.80	20.10	21.78
Layer_2	21.26	21.08	**20.15**	20.52	20.79
Layer_3	23.97	19.58	21.44	21.73	21.78

Table 1. Results in DER obtained by speaker embedding using cosine metric following vector size and hidden location.

On the dev corpus the best performance is obtained by using the second hidden layer with 500 neurons. We observe that this configuration obtains on the test corpus 20.15% DER. But on the test corpus the best configuration would be to use the first hidden layer with 400 neurons, resulting in 19.58% DER.

In Table 2, we propose to normalize the features with different levels of normalization and substituting the cosine metric by PLDA. The system called *No Norm* is run without any normalization, while systems called *LW-1,2,3* are run with $L\mathbf{W}$-normalization, applied respectively with 1, 2 and 3 iterations.

	No Norm	LW-1	LW-2	LW-3
FranceInter	21.14	**16.48**	16.53	16.81
BFMStory	17.31	**15.55**	15.55	15.55
CaVousRegarde	18.77	**10.28**	10.28	10.28
EntreLesLignes	20.30	**17.61**	17.61	17.61
PileEtFace	35.75	**24.88**	25.73	25.73
TopQuestions	15.81	**13.83**	13.83	13.83
LaPlaceDuVillage	58.19	**45.55**	46.57	46.57
Overall	25.04	**19.25**	19.44	19.54

Table 2. DER results obtained by speaker embeddings using PLDA and $L\mathbf{W}$-normalization with varying iterations.

The *No norm* systems obtains the worse results when using the cosine metric (25.04% DER). We observe that the best results are obtained by the system *LW-1* (ie. which performs $L\mathbf{W}$-normalization with one iteration), ressulting in a DER of 19.25%. The results shows the importance of conditioning the data before applying the PLDA metric.

In Table 3 we compare i-vectors (*i-vector/PLDA*) and speaker embeddings (*speaker-embedding/PDLA*). The two systems use PLDA and are conditioned with $L\mathbf{W}$-normalization (two iterations for i-vectors and one iteration for speaker embeddings.

	i-vector	Speaker Embedding	
	PLDA	Cosine	PLDA
FranceInter	20.88	17.45	**16.48**
BFMStory	16.78	18.42	**15.55**
CaVousRegarde	11.75	11.75	**10.28**
EntreLesLignes	18.08	18.94	**17.61**
PileEtFace	25.73	25.73	**24.88**
TopQuestions	**11.61**	**11.61**	13.83
LaPlaceDuVillage	**43.00**	44.81	45.55
Overall	20.92	20.15	**19.25**

Table 3. Results in DER obtained by i-vectors and speaker embeddings using the cosine and PLDA metrics.

The *i-vector/PLDA* system is considered as the baseline and obtains 20.92% DER. The *Speaker-embedding/Cosine* system gives better results compared to baseline and obtains 20.15% DER (an absolute gain of 0.77 points). Using the PLDA metric on speaker embeddings (*Speaker-embedding/PLDA*), the system gives better results compared to the *Speaker-embedding/cosine* system and obtains 19.25% DER (an new absolute gain of 0.91%).

7. CONCLUSION

This paper proposes to learn effective high-level features revealing speaker identities for speaker diarization. The features are extracted using the activations of the hidden layer of a DNN. By representing a large number of identities with a small number of hidden variables (creating a bottleneck), highly compact and discriminative features are created. Speaker embeddings, in place of i-vectors, obtain a DER decrease of 1.67 absolute DER points on the test corpus of the ETAPE 2012 evaluation campaign. The speaker embeddings give more robust models than i-vectors for this task.

For future work, we will investigate the use of different supervector and the use of Convolutional Neural Networks (CNN) for training the representation. We also plan on extensively testing speaker embeddings on the speaker verification and ASR adaptation tasks.

REFERENCES

[1] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey: The Speaker and Language Recognition Workshop*, 2010.

[2] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[3] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Md Jahangir Alam, and Pierre Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Internation Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[4] Taufiq Hasan, Rahim Saeidi, John HL Hansen, and David A van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Internation Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[5] Giovanni Soldi, Simon Bozonnet, Federico Alegre, Christophe Beaugeant, and Nicholas Evans, "Short-duration speaker modelling with phone adaptive training," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[6] B Vesnicer, J Zganec-Gros, S Dobrisek, and V Struc, "Incorporating duration information into i-vector-based speaker-recognition systems," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[7] Teva Merlin, Jean-François Bonastre, and Corinne Fredouille, "Non directly acoustic process for costless speaker recognition and indexation," in *Intelligent Communication Technologies and Applications*, 1999.

[8] Yassine Mami and Delphine Charlet, "Speaker identification by location in an optimal space of anchor models," in *International Conference on Spoken Language Processing (ICSLP)*, 2002.

[9] Haoqiang Fan, Zhimin Cao, Yuning Jiang, Qi Yin, and Chinchilla Doudou, "Learning deep face representation," *arXiv preprint arXiv:1403.2802*, 2014.

[10] Joseph Turian, Lev Ratinov, and Yoshua Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Association for Computational Linguistics*, 2010.

[11] George Dahl, Abdel-rahman Mohamed, Geoffrey E Hinton, et al., "Phone recognition with the mean-covariance restricted boltzmann machine," in *Advances in neural information processing systems*, 2010.

[12] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *International Conference on Computer Vision (ICCV)*, 2007.

[13] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification.," in *Interspeech*, 2009.

[14] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Interspeech*, 2011.

[15] Pierre-Michel Bousquet, Anthony Larcher, Driss Matrouf, Jean-François Bonastre, and Oldrich Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," in *Odyssey: The Speaker and Language Recognition Workshop*, 2012.

[16] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *InterSpeech*, 2013.

[17] Mickael Rouvier and Sylvain Meignier, "A global optimization framework for speaker diarization," in *Odyssey: The Speaker and Language Recognition Workshop*, 2012.

[18] Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier Galibert, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *Language Resources and Evaluation (LREC)*, 2012.

[19] Olivier Galibert and Juliette Kahn, "The first official repere evaluation.," in *SLAM@ INTERSPEECH*, 2013, pp. 43–48.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *Automatic Speech Recognition and Understanding (ASRU)*, 2011.