INVESTIGATION OF SPEAKER EMBEDDINGS FOR CROSS-SHOW SPEAKER DIARIZATION

Mickael Rouvier, Benoit Favre

Aix-Marseille Université
CNRS, LIF UMR 7279
13000, Marseille, France

ABSTRACT

This paper proposes to investigate speaker embeddings, a representation extracted from hidden layers of deep neural networks trained on a speaker identification task, on cross-show diarization. The new representation brings an improvement over i-vectors, and we show that while shallow hidden layers give best results on the single-show condition, deeper layers yield better performance on cross-show diarization. This confirms that deep representations model higher level features which help generalizing to different acoustic conditions. Experiments, conducted on the French corpus of REPERE, show that the deep speaker embeddings technique decreases DER by 0.82 points.

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

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: 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. In the typical setting, speaker diarization is applied to each recording, without *a priori* knowledge about the speakers or the structure of the show.

Until recently, most speaker diarization systems followed the task proposed by NIST, ie. each show is processed and evaluated independently. The major drawback of this approach is that it does not take into account the fact that some of the speakers attend multiple shows and it would be interesting to predict these events. This situation is very common in broadcast news and TV programs where hosts, anchors and other guests may appear recurrently. The notion of cross-show diarization on a collection has recently been introduced to deal with this kind of situation [1]. Thus, a speaker involved in several shows is always identified by the same anonymous label in each of the recordings. The approaches proposed to tackle this problem are mostly based on variants of speaker clustering [1, 2].

One key challenge in cross-show speaker clustering is how to model speakers, i.e. extract robust speaker-specific features from speech data. The classical approach in speaker modeling is based on the i-vectors/PLDA pipeline [3]. Introduced in [4], 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. Probabilistic Linear Discriminant Analysis Scoring (PLDA) is used as metric to verify if two i-vectors correspond to the same speaker during speaker clustering.

It is now well established that the limitations of the i-vector representation become apparent when processing short segments for which it is very difficult to disentangle useful information from background noise (the channel and speaker effects) [5, 6, 7, 8]. In order to tackle this problem, a novel speaker modeling framework, called speaker embedding, has been proposed in [9]. The main idea is to learn a high-level speaker representation from supervectors with deep neural network (DNN) models trained on a speaker identification task (classifying speech segments into one of n speaker identities). The assumption is that hidden layers of the DNN can extract relevant information for discriminating between speakers regardless of speaker identities. Speaker embeddings are extracted by passing supervectors through the DNN, and extracting the activations at a hidden layer to form a new feature vector representing the speaker. Unlike i-vectors this approach has the advantage to directly estimate a high-level feature representation in the speaker space using DNNs.

This paper extends the work presented in [9] with the following contributions:

- We investigate whether speaker embeddings extracted from different hidden layers of a DNN can be effective for cross-show diarization.
- We also look at the difference in term of optimal hidden layer between the single-show and the cross-show settings, discovering interesting insight on the working of deep neural networks with speech.

Our experiments on the REPERE corpus show that cross-show speaker diarization based on i-vectors obtains 17.10% Diarization Error Rate (DER), whereas the proposed approach based on speaker embeddings obtains a DER of 16.28% (an absolute gain of 0.82 points).

After presenting the speaker embedding paradigm in Section 2. Section 3 describes the speaker verification and conditioning methods used for speaker diarization. The corpus on which experiments are carried and the results of our experiments are presented in Section 4. Results are discussed in Section 5. Section 6 lists related work. Finally, we conclude with a discussion of possible directions for future work in Section 7.

2. SPEAKER EMBEDDINGS

In this section, we summarize the recipe for extracting speaker embeddings. Refer to [9] for details.

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).

Outlined in Figure 3, the proposed method learns high-level speaker features with DNN models trained to achieve a speaker identification task. When learning a classifier to recognize speaker identities, DNNs compact relevant features in the hidden layers. Speaker embeddings are feature vectors extracted from hidden layer neuron activations. 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. We note that the number of neurons in the hidden layer is the same for all hidden layers: 1024 neurons, except for the layer from which we extract embeddings, in order to create a bottleneck and force the network to extract high-level speaker features.

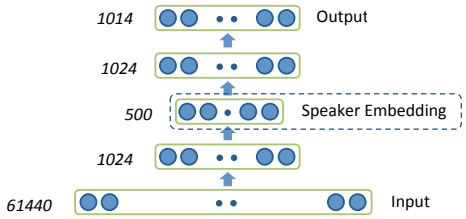

Fig. 1. An illustration of the feature extraction process. Arrows indicate forward propagation direction. The number of neurons in each layer of the deep neural network is labeled besides each layer. Speaker Embedding features are taken from the central hidden layer (this topology obtains the best results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

To construct speaker embeddings, we first extract 60-dimensional acoustic features for each turn (19 MFCC, log energy and first and second-Order deltas). Then, *First-Order* statistics, centred and normalized, are obtained from a Universal Background Model (UBM). For each Gaussian c of the UBM, the statistics are computed as:

$$F^{(c)} = \frac{1}{\sum_t \gamma_c^t} \sum_t \gamma_c^t (o^t - \mu_c) \quad (1)$$

where o^t is the feature vector at frame t and γ_c^t is the occupation probability of the Gaussian c for frame t. The complete *First-Order* statistic is $F_i = (F_i^{(1)}, \ldots, F_i^{(c)})$. F_i is used as input of a DNN.

The aim of the model is to be able to compute features for speakers which are not involved in the identification task, and therefore, at test time, a representation can be computed for any speech, and compared to other speaker models in a verification setting.

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

3. SPEAKER CONDITIONING AND VERIFICATION

In speaker diarization, the clustering step requires to compute the similarity between pairs of speech segments. This speaker verification step has been successfully performed with PLDA in previous work [3].

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 [10].

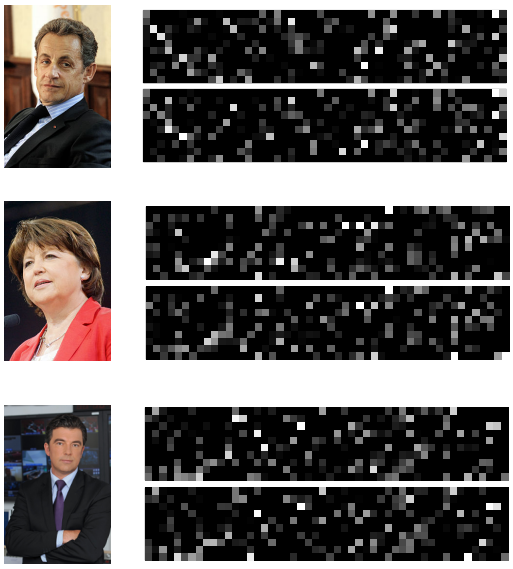

Fig. 2. Examples of the learned 500-dimensional Speaker Embeddings. The figure shows three test pairs from the test corpus. We rearrange them as 10×50 for the convenience of illustration, the ordering of the feature vector is the same for all examples. Feature values are non-negative since they are taken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

But PLDA assumes that the class distribution of the data is Gaussian. Unfortunately, the class distribution of speaker embeddings has a radial shape [9]. In order to tackle this problem we propose to normalize the data by applying the $L\mathbf{W}$-normalization introduced in [11]. This normalization consists in iterating standardization according to the within-class covariance and length-normalization. Thus this normalization moves the data towards a high Gaussian density surface, and helps to fit the PLDA model to the training set.

Finally, given two speaker embedding w_i and w_j, the speaker verification score can be computed in the PLDA model as:

$$d(w_i, w_j) = w_i^\intercal Q w_i + w_j^\intercal Q w_j + 2 w_i^\intercal P w_j \quad (2)$$

with

$$\begin{aligned} P &= \Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1} \\ Q &= \Sigma_{tot}^{-1}\Sigma_{ac}(\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac})^{-1} \end{aligned} \quad (3)$$

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 [3].

4. EXPERIMENTS

4.1. Cross-show diarization architecture

The cross-show diarization system used in these experiments is the LIUM Speaker Diarization system [12][1]. This system obtained the best results during the ETAPE 2012 and REPERE 2012 French evaluation campaigns.

The cross-show diarization system is based on a two-level architecture. We first process the shows separately (individual processing)

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

and then recluster speakers on the whole collection (overall processing).

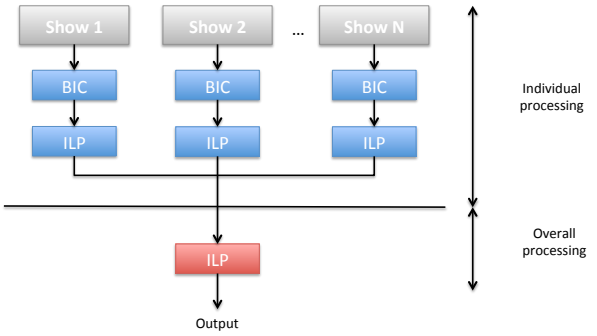

Fig. 3. The cross-show diarization architecture.

The first level 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 at regrouping 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 speaker embeddings [13].

The second level relies on ILP clustering with i-vectors. In this work, we propose to substitute i-vectors for speaker embeddings during the ILP Clustering.

4.2. Data

In the following experiments, we use the REPERE 2013 data [14]. The dataset consists in 9 TV shows recorded on French TV channels BFM and LCP, split according to the official train, development and test sets. The development corpus corresponds to 27 shows (3 hours) and is employed to determine the various hyper-parameters of the systems. The evaluation corpus contains 62 recordings (10 hours) and is employed to evaluate model performance. For training, we use french broadcast news training corpora: ESTER 1 [15], ESTER 2 [16], EPAC [17], REPERE [14] and ETAPE [18]. I-vectors, speaker embeddings and GMM-UBM models are all learned from the training corpus.

4.3. 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.

$$DER = \frac{\#Spk + \#Miss + \#FA}{\#Total} \quad (4)$$

where $\#Spk$, $\#Miss$ and $\#FA$ are respectively speaker error, missed speech and false alarm speech. The scoring tool we used was developed by LNE as part of the REPERE campaign [19]. This tool supports the computation of cross-show DER. The cross-show DER takes into account multiple occurrences of a speaker in several shows, as if all shows were merged into a single show.

4.4. 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 by searching values between 50 and 600). The i-vectors are conditioned with two iterations of $L\mathbf{W}$-normalization.

The DNN used for extracting the speaker embedding is composed of 3 hidden layers (one hidden layer is used for the embedding). The number of neurons in the hidden layer is the same for all hidden layers: 1024 neurons, except for the layer from which we extract embeddings. 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. Finally the speaker embeddings are conditioned with one iteration of $L\mathbf{W}$-normalization.

Note that when we vary the hidden layer from which embeddings are extracted, we always set the size of the embedding layer to 500 and the size of the other layers to 1024.

4.5. Results

In a first experiment we look at which hidden layer has the most potential for being extracted as representation in a single-show diarization system. These results are reported in Figure 4 where we plot DER according to the distance metric threshold (Factor) used in the ILP constraints. The distance metric threshold is a stopping criterion of the clustering process. The point on the curves corresponds to the systems obtained by using the thresholds determined using the development corpus.

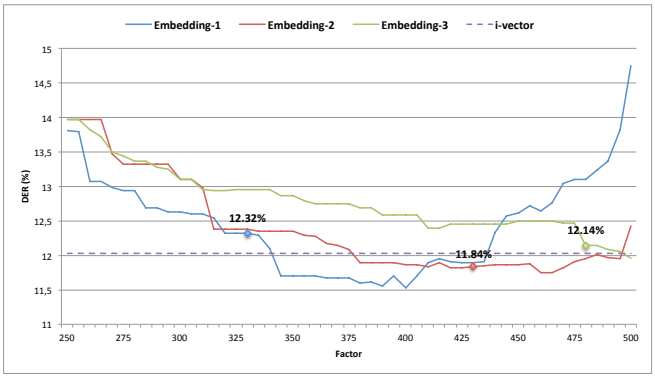

Fig. 4. Results in DER obtained by using the representation extracted from the different hidden layers in single-show diarization.

The dashed purple curve corresponds to the baseline system in which we use the i-vector paradigm. This system, for which thresholds are tuned on the development corpus, obtains a DER of 12.03%. The red, green and blue curves correspond to the systems by using the representation extracted from the first, second and third hidden layer. These systems are called *Embedding-1*, *Embedding-2* and *Embedding-3*. The *Embedding-2* system gives better results compared to the baseline and obtains 11.84% DER (an absolute gain of 0.19 points)

Show name	i-vector	Embedding-1	Embedding-2	Embedding-3
BFMStory	10.46	10.37	**10.22**	10.72
CultureEtVous	35.03	**34.28**	34.83	34.50
RuthElkrief	14.35	12.59	**12.59**	12.59
CaVousRegarde	13.56	14.38	**13.57**	13.57
EntreLesLignes	**11.55**	11.84	11.84	11.84
LCP_Actu	**5.95**	7.93	6.02	7.93
LCP_Info	9.22	8.86	**8.56**	8.56
PileEtFace	**8.80**	11.54	8.89	8.89
TopQuestions	**10.62**	10.36	11.18	11.67
Overall	12.03	12.32	**11.84**	12.14

Table 1. Results in DER obtained by i-vector, embedding-1, embedding-2 and embedding-3 systems.

We observe that on the development corpus the best performance is obtained by using the second hidden layer. But on the test corpus the best configuration would be to use the first hidden layer, resulting in 11.53% DER. Table 1 shows results obtained on each system by TV show.

In a second experiments, we look, as reported in Figure 5, at which hidden layer has the most potential for being extracted as representation for cross-show diarization. As previously, the point on the curves corresponds to the systems obtained by using the thresholds determined using the REPERE development corpus.

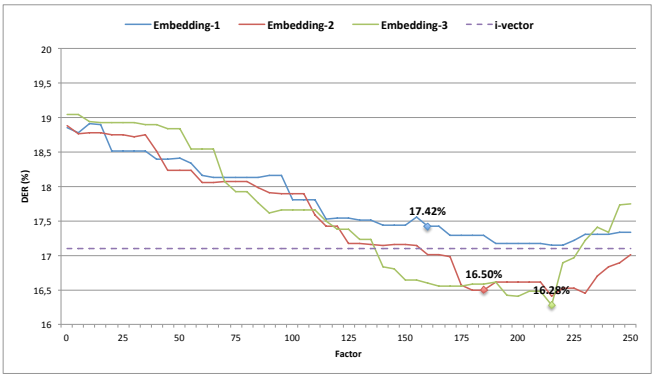

Fig. 5. Results in DER obtained by using the representation extracted from the different hidden layer in cross-show diarization.

We observe that the i-vector system obtains 17.10% DER, while the *Embedding-1*, *Embedding-2* and *Embedding-3* obtain respectively 17.42%, 16.50% and 16.28% DER. The representation extracted from the third hidden layer obtains the bests results.

5. DISCUSSION

In the artificial vision community, it is hypothesized that deeper hidden layers extract higher-level features from input pixels. Shown through activation maximization or sampling [21], deeper units tend to be sensitive to more complex shapes and semantic primitives. This phenomenon remains to be explored when speech is used as input of deep neural networks. The fact that deeper hidden layers yield better performance on cross-show diarization while shallow layers are more effective for single-show diarization tends to confirm the hypothesis that deeper layers convey higher-level features. Indeed, in the cross-show task, the model has to account for variability between recording conditions among the shows, while on the single-show task, it benefits from overtraining on low-level acoustic conditions which might not be speaker-specific but which correlate with speakers.

One of the biases of our experimental setting is that some speakers appear both in the training data for the DNN identification task and in the test data where we apply diarization. In fact, 32.04% of speakers are in both sets[2] and it might very well be that the embedding system allocates more generalization power to model them than unknown speakers. We performed an extra evaluation where we remove all known speakers from scoring. In that setting, the i-vector approach results in a DER of 18.33% and the speaker embedding approach results in 18.30%. On those speakers, it seems that the two approaches are not significantly different. Evidently, future work should investigate how to create representations that clearly outperform i-vectors on all conditions.

6. RELATED WORK

In [22, 23], the authors propose to introduce the technique of anchor modeling. The basic concept of anchor modeling is the representation of a target speech utterance 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 pre-trained anchor models. Each of the anchor models yields a likelihood score and the collection of scores is used to form the representation vector. This vector can be considered as 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 [24], text modeling [25] or speech recognition [26]. In particular, it is supposed to relieve researchers of designing features by automatically learning the relevant structure of the input space.

Concerning cross-show speaker diarization, most efforts have focused on reducing computation time. In [27], the authors propose to consider clustering as a connected graph in order to simplify the problem. In [28], the authors propose to extract speaker features with binary key speakers. The main advantage is that this approach provides very competitive time performance.

7. CONCLUSION

This paper investigates the speaker embedding modeling framework on a cross-show diarization task. In place of i-vectors, speaker embeddings obtain a DER decrease of 0.82 absolute points on the test corpus of the REPERE evaluation campaign. We observe that the best performance is obtained by using the third hidden layer of a DNN trained on a speaker identification task. This could be explained by the fact that relevant information is promoted more and more by each hidden layer so that the last hidden layer contains only speaker-specific information. Speaker embeddings give more robust models than i-vectors for this task.

For future work, we will investigate the use of different supervectors for training the representation. An interesting possibility is the DNN/i-vector paradigm proposed in [29]. This paradigm proposes to estimate the i-vector statistics by using DNN trained for ASR. We also plan on extensively testing speaker embeddings on the speaker identification and ASR adaptation tasks.

[2]It is also the case on other publications with the REPERE corpus.

REFERENCES

[1] Viet-Anh Tran, Viet Bac Le, Claude Barras, and Lori Lamel, "Comparing multi-stage approaches for cross-show speaker diarization.," in *Interspeech*, 2011.

[2] Grégor Dupuy, Mickael Rouvier, Sylvain Meignier, and Yannick Esteve, "I-vectors and ilp clustering adapted to cross-show speaker diarization.," in *Interspeech*, 2012.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Mickael Rouvier, Pierre-Michel Bousquet, and Benoit Favre, "Speaker diarization through speaker embeddings," in *European Signal Processing Conference (EUSIPCO)*, 2015.

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

[11] 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.

[12] 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.

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

[14] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard, "The repere corpus: a multimodal corpus for person recognition.," in *Language Resources and Evaluation (LREC)*, 2012.

[15] Sylvain Galliano, Edouard Geoffrois, Guillaume Gravier, Jean-François Bonastre, Djamel Mostefa, and Khalid Choukri, "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news," in *Language Resources and Evaluation (LREC)*, 2006.

[16] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts.," in *Interspeech*, 2009.

[17] Yannick Esteve, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas, "The epac corpus: Manual and automatic annotations of conversational speech in french broadcast news.," in *Language Resources and Evaluation (LREC)*, 2010.

[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*, 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.

[21] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Visualizing higher-layer features of a deep network," *Dept. IRO, Université de Montréal, Tech. Rep*, vol. 4323, 2009.

[22] 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.

[23] 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.

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

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

[26] 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.

[27] Grégor Dupuy, Sylvain Meignier, Paul Deléglise, and Yannick Esteve, "Recent improvements on ilp-based clustering for broadcast news speaker diarization," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[28] Hector Delgado, Xavier Anguera, Corinne Fredouille, and Javier Serrano, "Fast single-and cross-show speaker diarization using binary key speaker modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, 2015.

[29] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Moray McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Internation Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.