MULTIMODAL EMBEDDING FUSION FOR ROBUST SPEAKER ROLE RECOGNITION IN VIDEO BROADCAST

Mickael Rouvier, Sebastien Delecraz, Benoit Favre, Meriem Bendris, Frederic Bechet

Aix-Marseille Université, CNRS, LIF, Marseille, France
{firstname.lastname}@lif.univ-mrs.fr

ABSTRACT

Person role recognition in video broadcasts consists in classifying people into roles such as anchor, journalist, guest, etc. Existing approaches mostly consider one modality, either audio (speaker role recognition) or image (shot role recognition), firstly because of the non-synchrony between both modalities, and secondly because of the lack of a video corpus annotated in both modalities. Deep Neural Networks (DNN) approaches offer the ability to learn simultaneously feature representations (embeddings) and classification functions. This paper presents a multimodal fusion of audio, text and image embeddings spaces for speaker role recognition in asynchronous data. Monomodal embeddings are trained on exogenous data and fine-tuned using a DNN on 70 hours of French Broadcasts corpus for the target task. Experiments on the REPERE corpus show the benefit of the embeddings level fusion compared to the monomodal embeddings systems and to the standard late fusion method.

***Index Terms*—** Speaker role recognition, multimodal speaker embeddings, broadcast News

1. INTRODUCTION

Person role recognition in video broadcasts consists in classifying a person (speaking and/or visible) among a list of possible roles such as anchor, journalist, guest, etc. In this context, the audio and image modalities are complementary since role characteristics appear in the audio, speech transcription and scene analysis features. Most of the approaches proposed so far for person role recognition only consider one modality for two reasons: first, the presence of a person is not always synchronous between modalities. Indeed, a speaker is not always visible and all visible faces are not talking. In addition, the lack of labelled multimodal data limits the possibility for jointly training multimodal systems which generally assume synchrony between the modalities.

Recently approaches based on Deep Neural Networks (DNN) have achieved state-of-the-art performance on several tasks for audio and image processing. The main advantage of such techniques is the ability to learn simultaneously feature representations and classification functions. The initialization of feature representations can be performed on a large generic corpus not necessarily related to the target task, resulting in embeddings that can be jointly fine-tuned for that task. This approach has been proposed for synchronous tasks such as lip/speech activity detection and recognition [1].

In this paper, we want to classify speakers into four roles using the audio, image and text modalities:

- R1: anchors. These speakers are characterized by their presence throughout the show, without discontinuity.
- R2: journalists. They are TV professionals appearing one time or more during the show.
- R3: reporters. Similar to the R2 role, they are correspondents covering events outside the set of the show.
- R4: guests and others. They are invited to discuss the news, because of their expertize or fame, under the guidance of the anchor. They are neither part of the organization committee, nor the leaders of debates. They can appear in different TV shows, especially during highly publicized events. Others refers to everyone else who could appear, like interviewed people in a report.

We present an alternative to the standard late fusion paradigm based on multimodal embeddings refined for the Speaker Role Recognition (SRR) task. The main novelty of our approach is a fusion at the embedding level that characterizes multimodal information without assuming a synchrony between modalities. Experiments on the French REPERE corpus show the benefit of this approach with respect to monomodal strategies and standard late fusion methods.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 presents a general description of our approach. Sections 4, 5 and 6 present speaker role embeddings over the text, audio and visual modalities.

This work has been carried out thanks to the support of the A*MIDEX project (n^{o} ANR-11-IDEX-0001-02) funded by the “Investissements d’Avenir” French Government program, managed by the French National Research Agency (ANR).

Section 7 focuses on the fusion system. Experiments are presented in Section 8.

2. RELATED WORK

Automatic Speaker Role Recognition (SRR) assumes that roles are characterized by specific acoustic, visual and textual features such as language style or prosody. In the literature, SRR methods have been studied in purpose of chaptering audio-visual documents (talk shows and broadcast news). Existing methods are divided according to the features extracted (audio and/or text), the decision level (for each speaker turn [2, 3] or globally on all the turns of a given speaker [4, 5, 6, 7, 8]) and classification techniques (supervised [4, 5, 2, 3] or unsupervised [6, 7, 8]).

In [4], based on the hypothesis that spontaneous speech classification is a clue for SRR, authors proposed an application of a spontaneous speech detector for the SRR task using prosodic and linguistic features (local) and a contextual model (global). Promising results on radio broadcasts are showed for 10 speaker roles classification. However, confusion analysis shows the difficulty of identifying specific roles. In [6] authors proposed an unsupervised system that clusters speakers according to their role based on structural and lexical features. A partition selection algorithm is used on speaker clusters on Mandarin and English talk shows data. On the same corpus, a sentence pattern extraction method is proposed in [7]. Then, spectral clustering is used for unsupervised SRR allowing to classify hosts, expert guests and soundbites. In [8], authors described an unsupervised SRR system on English, Arabic and Mandarin data based on structural, lexical, social network analysis features and a boosting classifier. A loss of 1.1% in accuracy is shown when automatic features are extracted compared to the use of manually labelled linguistic phenomena. In [5], authors used temporal, acoustic and prosodic features to classify roles at the speaker cluster level. Authors distinguish between punctual and non punctual speaker roles and train Support Vector Machine (SVM) and Gaussian Mixture Model (GMM) classifiers hierarchically. Several feature selection methods are compared (Principal Component Analysis/Canonical Discriminant Analysis/Sequential Backward Feature Selection). Experiments on a French broadcast corpus achieved good results. However, those methods make decisions on speaker clusters.

[3] classifies speaker roles (anchor/reporter/other) using HMM and Maximum Entropy classifiers. In the HMM classifier, speaker roles are states and pronounced sentences are observations. Maxent classifies speaker roles using the first and last pronounced sentences during the speaker turn. Best results on Mandarin broadcast news are achieved by the Maxent-based system enriched with contextual information (previous and next sentences). However, ASR and speaker turns were manually labelled. [2] focuses on speaking style features to classify speaker roles. Dynamic Bayesian Networks (DBN) models have been used to classify speaker turns role depending on the previous role and recent speaker role. [9] present a multimodal system based on lexical and acoustic features for SRR. Authors proposed to classify speaker turns hierarchically: first anchor, then reporter/other. Two classifiers are used: a boosting-based text classifier (icsiboost) and MFCC-based GMMs. Two classifier fusion are compared: one adding the GMMs score to textual features of icsiboost. A second late fusion based on logistic regression of GMMs and icsiboost scores.

Considering the visual modality of broadcast videos, to the best of our knowledge, there are no work based on image features for SRR. Some work have been done for visual shot role recognition. For example in [10] authors proposed a generalized anchor shot detector based on deep neural networks with a sampling strategy. They obtain interesting results on large-scale broadcast news videos (30 different TV channels). However, this method does not characterize speakers.

This study takes advantage of all these previous approaches by the set of features used (acoustic and linguistic) and the fusion paradigm between modalities. The main novelties of our approach are firstly to introduce image features into speaker characterization; and secondly to propose a real multimodal fusion framework that goes beyond late fusion and overcomes most of the problems linked to early fusion methods (lack of data annotated synchronously in all modalities and difficulties to define multimodal features).

We applied our model to the speaker turn characterization task, independently of all the other turns from the same speaker in a given show. While this decision level is suboptimal, it will allow us to demonstrate the effectiveness of our multimodal approach independently of any global decision strategy that could be defined on the top of the local decisions.

3. APPROACH

Our proposed approach consists in creating representations in each modality tailored for the SRR task.These representations are used as input of a multimodal classifier which can take advantage of cross-modal features drawn from the concatenation of the monomodal representations. Figure 1 illustrates that approach.

Each monomodal representation is trained on a large monomodal corpus not necessarily linked to the SRR task. The multimodal annotated corpus is only used when training the fusion. This method allows us to take advantage of both early and late fusion at the same time: we can use large amount of monomodal data for which we do not have synchronous annotations in the other modalities as can be done with late fusion; we can train multimodal classifiers that can build multimodal features directly from each modality like in early fusion.

For the text modality we train Convolutional Neural Networks (CNN) that start from word embeddings trained on a large text corpus. The audio modality uses a representation extracted from a DNN modeled after a speaker recognition system, trained on the SRR task. The image modality relies of a representation extracted from an ImageNet concept recognizing neural network, and repurposed for the target task. Fusion consists in concatenating hidden layers of the monomodal systems and adding fully-connected layers, which create building blocks for merging decisions according to relevant features from multiple modalities. The following sections detail the architecture of the monomodal and fusion components.

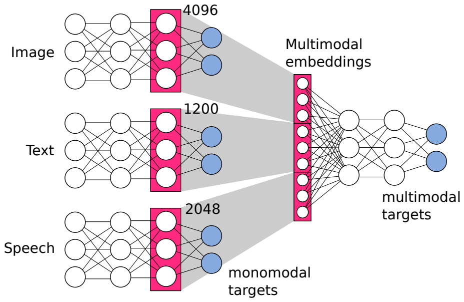

Fig. 1. Illustration of the embedding fusion approach.

4. TEXT EMBEDDINGS

Recent work has shown that CNN are powerful for Natural Language Processing (NLP) classification problems [11]. A CNN is a deep model having multiple convolutional and pooling layers followed by a simple classifier (usually a Multilayer Perceptron). The main advantage of using convolution is the ability to process variable input dimensions (sentences in our case). In addition, multiple convolutional filters extract local N-grams semantics with different granularities while pooling layers extract global semantics of the input. In our work, we use speech transcriptions from the current speaker turn for SRR.

First, each word is represented by a 300 dimension continuous and real-valued vector called word embedding [12]. In our experiment, word embeddings [1] are trained on Wikipedia using the skip-gram model (window size = 7, 5 iterations). This strategy allows to characterize semantic and grammatical associations between words.

Then, word embeddings for the words of the current turn are passed through three convolutional filters which select the best 3-grams, 4-grams and 5-grams. They are combined with a Max-Over-Time pooling layer (400 dimension) and a standard Soft-Max fully connected layer [13]. We used dropout to disable randomly 40% of neurons at each iteration, which acts as regularization.

Finally, *text embeddings* of 1200 dimensions are extracted from the Max-Over-Time pooling layer and used later for the multimodal system.

5. AUDIO EMBEDDINGS

In previous work, it was proposed to learn high-level acoustic features for speaker identification [14] (called Speaker embeddings). In the same way, we propose to learn high-level speaker role features, called "*audio embeddings*", using deep models trained to achieve the SRR task.

The *audio embeddings* are trained as follows: first, a 60-dimensional acoustic feature vector is extracted for each turn[2] i with a 10ms frame rate (19 MFCCs, log energy and first and second-order deltas). Then, first-order statistics Centred-Normalized obtained from a Universal Background Model (UBM) are generated. Thus for each gaussian component c, the first-order statistics are extracted as follows:

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

where F^c is the first-order statistics for gaussian component c, o^t is the feature vector at frame t, γ_c^t is the occupation probability of the gaussian c for frame t and μ_c is the mean of the gaussian c. The complete first-order statistic is $F_i = (F_i^{(1)}, \ldots, F_i^{(c)})$. The UBM used is a gender- and channel-independent GMMs of 1024 diagonal gaussians computed with the Kaldi toolkit [15].

Then, the first-order statistics are used as input of a DNN having two 2048-dimensional hidden layers. The nonlinearities of the hidden layers is corrected by a Rectified Linear Unit (ReLU) function. The output layer is a Soft-Max. Training is performed by optimizing the cross-entropy criterion. In our experiments, we optimized DNN parameters on the development set. Weights were updated using 512 mini-batches over 8 iterations and the learning rate initialized at 0.04 is reduced to 0.004 when converging.

Finally, *audio embeddings* of 2048 dimensions are extracted from the last hidden layer of the DNN and used later for multimodal fusion.

6. IMAGE EMBEDDINGS

The visual grammar in talk-shows and news is a true source of information for speaker role detection. In this section, we describe image features used in the speaker role recognition system. We used *image embeddings* based on DNN which have shown to be extremely good for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

We fine tuned the *AlexNet* CNN proposed in [16] trained on the ILSVRC-2012 [17] corpus for image classification. The architecture consists of five convolutional layers, three

[1] We used the *Word2vec* toolkit

[2] Average segment duration is 7.8 seconds in the training corpus.

fully-connected layers, max-pooling and normalization layers. It uses ReLU as activation function to accelerate learning and dropout after the first two fully-connected layers to prevent over-fitting. This model takes resized (256 × 256) and normalized image inputs with a 512 batch size. Weights are updated following the same rules as described in [16].

We adapted the *AlexNet* architecture by changing the last fully-connected layer to predict only four classes and finedtuned the already learned weights of the AlexNet model to obtain a new CNN for SRR. We increased the learning rate on the last fully-connected layer (10 times the global LR) in order to regularize fine tuning. We trained the network for 270 epochs on 19k images using Caffe [18] on GPUs.

Finally, *image embeddings* are extracted from the second fully-connected layer, providing 4096 dimension vectors for use in the multimodal fusion system.

7. MULTIMODAL FUSION

There are two approaches commonly used for deep multimodal fusion: early and late fusion. Late fusion considers that the modalities are independent by first applying classification separately on each modality and then merging the output using a high-level classifier. Unfortunately, the classifier cannot model the correlations among modalities. The early fusion approach tackles this problem by learning features and class relationships to model the interaction between modalities. In [1], authors proposed to learn multimodal feature representations based on auto-encoders for audio-visual speech classification. In [19], authors proposed to learn a common image and text embedding space based on multimodal auto-encoders for word similarity and categorization tasks. In [20], authors proposed to learn visual and linguistic features jointly for image labelling and retrieval tasks. Two fusions are presented: one forcing word embeddings to take into account fixed visual features by maximizing the similarity during training. In the second fusion, the authors proposed to add a layer in the DNN that merges images and words representations. While late fusion cannot benefit from multimodal feature correlations, early fusion requires lots of synchronous training data.

We propose an early fusion approach based on DNNs where the input is task-specific embeddings in all modalities. First, DNNs are trained independently for each modalities allowing to extract general monomodal representations (text, audio and image embeddings). Then, these embeddings are used as input of a new DNN trained to learn from multimodal features to classify speaker roles. Unlike late fusion, our method can take advantage of relevant feature subspaces (embeddings) from multiple modalities.

In our experiments, the DNN used for early fusion is composed of two 1024-dimensional hidden layers. The nonlinearities of the hidden layers are corrected by a ReLU function. Weights were updated using a mini-batch size of 512, trained over 6 iterations. The learning rate was initialized at 0.01 and reduced till 0.001. Our experiments also show results of a late-fusion based on the SVM classifier. All probabilities given by each modalities are grouped in a vector, and a linear SVM classifier is trained on these probability vectors to predict speaker roles.

8. EXPERIMENTS

We present experiments performed on the multimodal REPERE corpus [21]. We compare the results obtained by several baselines systems in all modalities with our DNN embeddings method. The DNN fusion is also compared to a standard late fusion approach consisting of a combination of decisions output at each modality.

8.1. Experimental setup

Speaker diarization is carried out using the LIUM opensource speaker diarization toolkit [22]. First speaker segmentation is used to detect fine-grained speaker changes using Generalized Likelihood Ratio (GLR). Then hierarchical agglomerative clustering is used to group segments that belong to the same speaker using the Bayesian Information Criterion (BIC) followed by a Clustering based on Integer Linear Programming described in [23]. This system performs a Diarization Error Rate (DER) of 12.03% on the REPERE corpus.

Then, speech transcripts are generated using the Kaldi Automatic Speech Recognition tool [15]. The speech transcription process is carried out in two passes: (1) An automatic transcript is generated with a GMM-HMM model of 7000 states and 150000 Gaussians. (2) Word-graphs output by the first pass are used to compute a fMLLR transform on each speaker cluster. Then, the second pass is performed using DNN acoustic model trained on acoustic features normalized with the fMLLR matrix [24].

The acoustic models are trained on 227 hours of wideband recordings (167 hours from ESTER 1 and 2 campaign and 60 hours from EPAC [25, 26]). The language model is based on trigram LMs on a lexicon of 95k words. Sources for training the LM are the audio corpus transcript, the French gigaword [27] and additional data collected from the Web. To estimate and interpolate these models, the SRILM [28] toolkit is employed using modified Knser-Ney discounting without cut-off.

The system is fully described in [29] and obtains a Word Error Rate (WER) of 19.67% on the REPERE test set.

In TV broadcasts, speakers appear only 60% of the time and observable speakers talk only 30% of the time [30]. When processing the image modality, due to this asynchrony, we choose to select one image per turn as follows: for each speaker turn t, we selected the longest video shot v that matches t ($t \cap v \neq \emptyset$); then, we choose the frame f as the center of the intersection $t \cap v$.

8.2. Results

Experiments are performed using the REPERE corpus [21] which consists of about 70 hours of video broadcast from 9 French speaking channels ranging from news with an anchor and field reports, to talk shows and tabloid shows. Each speaker turn in the corpus is manually annotated with transcripts, speaker identities and speaker roles among the four classes: anchor/host (R1), commentator (R2), reporter (R3), invited speaker/other (R4[3]).

The corpus is split into train (18951 turns), development (1402 turns) and test (4627 turns) sets used respectively for training the systems, validating the structure of the neural networks and the hyper-parameters of the classifiers, and evaluating the results. The test set contains shows that occur in the train and development sets (of course not on the same days), as well as a new show that is completely unknown from models trained on the train and development sets, to check the capacity of generalization of our models. Table 1 describes the distribution of roles on the test set.

Role	% of turns
Host (R1)	23.34
Commentator (R2)	11.28
Reporter (R3)	14.22
Other (R4)	51.16

Table 1. Repartition of roles on the test set.

All results are given using accuracy (number of role correctly identified) and the Diarization Error Rate (DER). The DER consists in computing the SRR errors at the frame level, the same way it is done in the speaker diarization task. The main advantage of this metric is to allow us to compare two different SRR output with a different speaker segmentation, as we consider each frame independently. This is the case when we compare results obtained using reference transcripts and speaker segments (DER-Man for manual annotation) versus ASR and automatic speaker diarization (DER-Auto for automatic annotation).

First, table 2 compares baseline and monomodal deep learning approaches. Among the baselines we have:

- *Majority*: this baseline simply chooses the most frequent role for each speaker frame.
- *Adaboost*: this is a boosting-based classifier [31] applied to word n-grams from textual segments to classify speaker roles.
- *JFA*: this baseline trains Joint-Factor-Analysis models to characterize speaker roles in the audio modality [**?**].
- *SVM-HOG*: this is a SVM-based classifier using full frame histogram of gradient features to find the best role for a given image [32].

Results in Table 2 clearly indicate that the DNN approaches consistently outperform baselines. In addition, the audio modality offers the best monomodal classifier.

System	Modality	Acc-Man	DER-Man	DER-Auto
JFA	A	26.76	37.48	42.54
DNN-Audio	A	**77.52**	**19.79**	**25.43**
Adaboost	T	62.13	28.80	34.33
CNN-Text	T	67.50	29.11	32.66
SVM-HOG	I	62.76	36.97	42.04
CNN-Image	I	70.48	25.69	35.25

Table 2. Monomodal accuracy and DER results on the test set, with baseline and neural network systems. Modalities are T for text, A for audio and I for image. Acc-Man is the accuracy using the reference transcription. DER-Man is the DER using the reference transcription and DER-Auto is the DER using automatic transcription.

In a second set of experiments, we analyse the performance of several multimodal systems according to the type of modality used and the fusion method (late of early with embeddings).

Results, presented in Table 3 show that merging decisions at the embeddings level performs better than late decisions. They also justify the use of multimodal models for the task: the gain of performance in the multimodal setting compared to the monomodal one is very important. The best monomodal DER-Man was 19.79 (respectively 25.43 for DER-Auto) in the monomodal setting and only 13.84 (respectively 19.9) in the multimodal setting. We can also observe that it is the fusion of all modalities which gives the best results.

Fusion	Modality	Acc-Man	DER-Man	DER-Auto
Majority	-	51.16	39.77	44.54
Late	A+T	78.49	18.67	24.11
Late	A+I	80.98	17.26	22.98
Late	I+T	78.02	21.16	27.60
Late	A+I+T	82.36	15.37	20.97
Embedding	A+T	80.16	15.90	21.82
Embedding	A+I	82.16	15.45	20.65
Embedding	I+T	76.01	22.83	28.60
Embedding	A+I+T	**85.28**	**13.84**	**19.79**

Table 3. Multimodal DER results for the posterior-level late fusion and the embeddings level fusion. Modalities are T for text, A for audio and I for image. Acc-Man is the accuracy using the reference transcription. DER-Man is the DER using the reference transcription and DER-Auto is the DER using automatic transcription and speaker segmentation.

[3] R4 and R5 classes from the original corpus are merged as annotator agreement is low on that pair.

In order to study the robustness of our methods, Table 4 shows the accuracy and DER on a subset of the test corpus corresponding to unseen conditions (different shows). The system based on text embeddings is robust to unseen conditions while the audio and image modalities results decrease when processing these new shows. This is particularly true for the image modality which goes from a DER of 25.69 in the whole test set to 43.29 on the unseen show. In this condition it is not suprising that the fusion methods do not provide better results over the single best modality.

System	Modality	Acc-Man	DER-Man	DER-Auto
CNN-Text	T	**70.07**	**26.65**	**28.42**
DNN-Audio	A	65.69	29.88	37.31
CNN-Image	I	51.09	43.29	46.47
Late	A+I+T	70.07	27.77	34.61
Embedding	A+I+T	66.42	32.80	34.06

Table 4. Results on the unseen conditions (different show) which represents 5% of the test set.

These results point out one of the weaknesses of a multimodal approach when all the modalities don't have the same generalization capacity to process unseen events. If the textual modality is really robust, audio and image have difficulties to process unseen events.

9. CONCLUSION

In this paper, we introduced a speaker role recognition system based on multimodal embeddings fusion for asynchronous data. Experiments on the REPERE corpus using manual and automatic speaker diarization showed that merging text, audio and visual features improves greatly speaker role classification performance with respect to monomodal approaches. Our multi-modal embeddings allows to capture speaker role features in multiple views and the use of embeddings level fusion obtained the best results with 19.79% of DER on automatic speaker diarization. Our method allows us to take advantage of both early and late fusion at the same time: we can use large amounts of monomodal data for which we do not have synchronous annotations in the other modalities as would be performed by late fusion; we can train multimodal classifiers that can build multimodal features directly from each modality like in early fusion.

However one of the drawbacks of this method is the lack of generalization of the audio and image models when processing unseen shows. Increasing robustness to unseen events is the line of research we are following now to improve our multimodal SRR system.

10. REFERENCES

[1] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning (ICML)*, Bellevue, USA, June 2011.

[2] Sibel Yaman, Dilek Hakkani-Tür, and Gökhan Tür, "Social role discovery from spoken language using dynamic bayesian networks," in *InterSpeech*, 2010.

[3] Yang Liu, "Initial study on automatic identification of speaker role in broadcast news speech," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, Stroudsburg, PA, USA, 2006, NAACL-Short '06, pp. 81–84, Association for Computational Linguistics.

[4] Richard Dufour, Yannick Esteve, and Paul Deléglise, "Investigation of spontaneous speech characterization applied to speaker role recognition," *Reporter*, vol. 11, no. 18, pp. 0h10, 2011.

[5] Benjamin Bigot, Isabelle Ferran, Julien Pinquier, and Rgine Andr-Obrecht, "Speaker Role Recognition to help Spontaneous Conversational Speech Detection (regular paper)," in *International workshop on Searching Spontaneous Conversational Speech SCSS (SCSS)*. octobre 2010, pp. 5–10, ACM.

[6] Brian Hutchinson, Bin Zhang, and Mari Ostendorf, "Unsupervised broadcast conversation speaker role labeling," in *ICASSP*, 2010, pp. 5322–5325.

[7] Bin Zhang, Brian Hutchinson, Wei Wu, and Mari Ostendorf, "Extracting Phrase Patterns with Minimum Redundancy for Unsupervised Speaker Role Classification," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, June 2010, pp. 717–720, Association for Computational Linguistics.

[8] Wen Wang, Sibel Yaman, Kristin Precoda, and Colleen Richey, "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5556–5559.

[9] Géraldine Damnati and Delphine Charlet, "Multi-view approach for speaker turn role labeling in TV broadcast news shows," in *InterSpeech*, 2011, pp. 1285–1288.

[10] Bailan Feng, Jinfeng Bai, Zhineng Chen, Xiangsheng Huang, and Bo Xu, "Anchor shot detection with deep neural network," in *Proceedings of the 15th Pacific-Rim Conference on Advances in Multimedia Information Processing — PCM 2014 - Volume 8879*, New York, NY, USA, 2014, pp. 304–312, Springer-Verlag New York, Inc.

[11] Ronan Collobert, "Deep learning for efficient discriminative parsing," in *AISTATS*, 2011.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[13] Yoon Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[14] Mickael Rouvier, Pierre-Michel Bousquet, and Benoit Favre, "Speaker diarization through speaker embeddings," in *EUSIPCO*, 2015.

[15] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," Idiap-RR Idiap-RR-04-2012, Idiap, Rue Marconi 19, Martigny, 1 2012.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural network," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, pp. 1–42, April 2015.

[18] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[19] Carina Silberer and Mirella Lapata, "Learning grounded meaning representations with autoencoders," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, 2014, pp. 721–732.

[20] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni, "Combining language and vision with a multimodal skip-gram model," *CoRR*, vol. abs/1501.02598, 2015.

[21] Aude Giraudel, Matthieu Carr, Valrie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard, "The repere corpus : a multimodal corpus for person recognition," in *LREC*, 2012.

[22] Mickael Rouvier, Gregor 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.

[23] Mickael Rouvier and Sylvain Meignier, "A global optimization framework for speaker diarization," in *Speaker Odyssey*, 2012.

[24] Mark JF Gales and PC Woodland, "Mean and variance adaptation within the mllr framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[25] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts.," in *Interspeech*, 2009, vol. 9, pp. 2583–2586.

[26] 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 *LREC*, 2010.

[27] Angelo Mendonça, David Graff, and Denise DiPersio, "French gigaword," 2009.

[28] Andreas Stolcke et al., "Srilm-an extensible language modeling toolkit.," in *InterSpeech*, 2002.

[29] Mickael Rouvier and Benoit Favre, "Speaker adaptation of dnn-based asr with i-vectors: Does it actually adapt models to speakers?," in *InterSpeech*, 2014.

[30] Meriem Bendris, Delphine Charlet, and Gérard Chollet, "Talking faces indexing in tv-content," in *Content-Based Multimedia Indexing (CBMI), 2010 International Workshop on*. IEEE, 2010, pp. 1–6.

[31] Benoit Favre, Dilek Hakkani-Tür, and Sebastien Cuendet, "Icsiboost," http://code.google.come/p/icsiboost, 2007.

[32] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Factor analysis simplified.," in *ICASSP*, 2005.

[33] Yanwei Pang, Yuan Yuan, Xuelong Li, and Jing Pan, "Efficient hog human detection," *Signal Processing*, vol. 91, no. 4, pp. 773–781, 2011.