INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Acoustic word embeddings for ASR error detection

Sahar Ghannay, Yannick Estève, Nathalie Camelin, Paul deléglise

LIUM - University of Le Mans, France
firstname.lastname@univ-lemans.fr

Abstract

This paper focuses on error detection in Automatic Speech Recognition (ASR) outputs. A neural network architecture is proposed, which is well suited to handle continuous word representations, like word embeddings. In a previous study, the authors explored the use of linguistic word embeddings, and more particularly their combination. In this new study, the use of acoustic word embeddings is explored. Acoustic word embeddings offer the opportunity of an *a priori* acoustic representation of words that can be compared, in terms of similarity, to an embedded representation of the audio signal.

First, we propose an approach to evaluate the intrinsic performances of acoustic word embeddings in comparison to orthographic representations in order to capture discriminative phonetic information. Since French language is targeted in experiments, a particular focus is made on homophone words. Then, the use of acoustic word embeddings is evaluated for ASR error detection. The proposed approach gets a classification error rate of 7.94% while the previous state-of-the-art CRF-based approach gets a CER of 8.56% on the outputs of the ASR system which won the ETAPE evaluation campaign on speech recognition of French broadcast news.
Index Terms: ASR error detection, acoustic word embeddings, neural networks, speech recognition.

1. Introduction

Recent advances in the field of speech processing have led to significant improvements in speech recognition performances. However, recognition errors are still unavoidable. This reflects their sensitivity to the variability, *e.g.* to acoustic conditions, speaker, language style, *etc.* These errors can have a considerable impact on applications based on the use of automatic transcriptions, like speech to speech translation, spoken language understanding, information retrieval, *etc.*

Error detection aims to improve the exploitation of ASR outputs by downstream applications. For two decades, many studies have focused on the ASR error detection task. State of the art approaches are based on the use of Conditional Random Fields (CRF) [1]. In [2], authors detect error regions generated by Out Of Vocabulary (OOV) words. They propose an approach based on a CRF tagger which takes into account contextual information from neighboring regions instead of considering only the local region of the OOV words. A similar approach for more general ASR errors is presented in [3]: the authors propose an error detection system based on a CRF tagger using various ASR-derived, lexical and syntactic features.

The most recent approaches have focused on neural network classifiers. In [4], authors propose to use a neural network classifier furnished by a stacked auto-encoders (SAE) structure in order to tackle the problem of imbalanced data, the SAE helping to learn the error word representation. This approach is compared to Support Vector Machines, MaxEnt, and Extremely Randomized Tree classifiers, but not to CRFs. In [5], we proposed a neural approach to detect errors in automatic transcriptions, and to calibrate confidence measures provided by an ASR system. In this previous study, we experimented the use of several information sources: syntactic, lexical, ASR-based feature, however, we did not investigate the use of acoustic information, except to some prosodic features.

Acoustic information can be obtained through acoustic word embeddings. Acoustic word embeddings are obtained by incorporating an arbitrary or fixed dimensional speech segment in a fixed-dimensional space such that speech segments of words that sound similarly will have similar embeddings. These representations were successfully used in a query-by-example search system [6, 7] and in a segmental ASR lattice re-scoring system [8].

In this paper, we propose to build acoustic word embeddings with the purpose to evaluate their performance in the framework of the ASR error detection task. Moreover, we propose to compare them to orthographic representations in order to evaluate whether they capture discriminative phonetic information. Since French language is targeted in experiments, a particular focus is made on homophone words.

2. ASR error detection system

An ASR error detection system attributes a label *Error* or *Correct* for each word in the automatic transcription. This is accomplished by analyzing each word within its context, this analysis being based on a set of features defined below. The window context size used in this study is 2 on either side of the current word (the entire window size is 5: 2 on left and 2 on right, in addition to the current targeted word). The proposed neural architecture is a feed forward neural network, based on a multi-stream strategy to train the network, named MultiLayer Perceptron MultiStream (MLP-MS). The MLP-MS architecture is used in order to better integrate the contextual information from neighboring words. In addition, this architecture is designed to be fed by different types of features, including word embeddings. A detailed description of this architecture is presented in a previous study [9]. The features used in this study, are nearly the same as the ones presented by [3], which are detailed as follows: *ASR features* are the posterior probabilities generated from the ASR system at the word level. *Lexical features* are the length of the current word and three binary features indicating if the three 3-grams containing the current word have been seen in the training corpus of the ASR language model. *Syntactic features* are POS tag, dependency labels and word governors, which are extracted using the MACAON NLP Tool chain[1] [10] to process the ASR transcriptions. The orthographic representation of a word is used in CRF approaches, as in [3]. With our

[1] http://macaon.lif.univ-mrs.fr

http://dx.doi.org/10.21437/Interspeech.2016-784

neural approach, we will use word embeddings which permit us to take advantage of some generalizations extracted during the construction of this continuous representation. Different approaches have been proposed to build word embeddings through neural networks. These approaches can differ in the type of architecture and the data used to train the model. Hence, they can capture different types of information: semantic, syntactic, *etc.* In our previous studies [5, 11], we evaluated different kinds of word embeddings, including *word2vecf* on dependency trees [12], *skip-gram* provided by *word2vec* [13], and *GloVe* [14]. These evaluations were carried on ASR error detection, natural language processing, analogical and similarity tasks. We revealed that the combination of word embeddings through auto-encoder yields the best results compared to the other combination approaches (PCA and simple concatenation). Based on the results of these studies, we propose to use the best word embeddings (the three ones cited above) retained from the evaluation task [11] and to combine them with auto-encoder as in [5]. A detailed description of the word embeddings and the combination approaches is presented in [11, 5].

3. Acoustic word embeddings

3.1. Building acoustic word embeddings

The approach we used to build acoustic word embeddings is inspired from the one proposed in [8]. Word embeddings are trained through a deep neural architecture, depicted in figure 1, which relies on a convolutional neural network (CNN) classifier over words and on a deep neural network (DNN) trained by using a triplet ranking loss [8, 15, 16]. This architecture was proposed in [8] with the purpose to use the scores derived from the word classifier for lattice rescoring. The two architectures are trained using different inputs: speech signal and orthographic representation of the word.

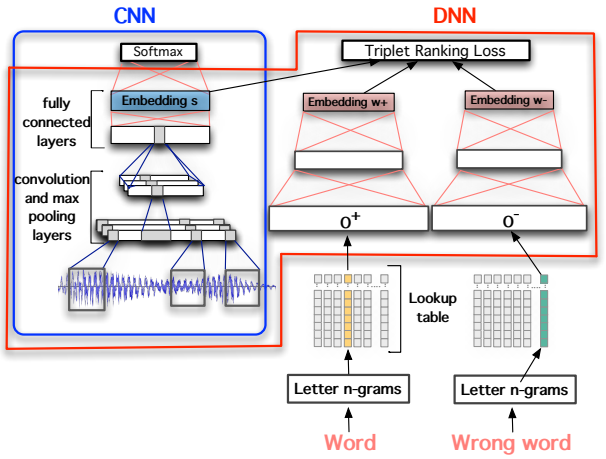

Figure 1: Deep architecture used to train acoustic word embeddings.

The CNN is trained to predict a word given an acoustic sequence of T frames as input. It is composed of a number of convolution and pooling layers, followed by a number of fully connected layers which feeds into the final softmax layer. The final fully connected layer just below the softmax one is called embedding layer **s** (it was called **e** in [8]). It contains a compact representation of the acoustic signal. This representation tends to preserve acoustic similarity between words, such that words are close in this space if they sound alike.

The idea behind using the second architecture is to be able to build an acoustic word embedding from orthographic word representation, especially in order to get an acoustic word embeddings for words not already observed in an audio speech signal. More, a such acoustic word embedding derived from an orthographic representation can be perceived as a canonical acoustic representation for a word, since different prononciations imply different embeddings **s**.

Like in [8], orthographic word representation consists on a bag of n-grams ($n \leq 3$) of letters, composed of 10222 trigrams, bigrams, and unigrams of letters, including special symbols *I* and *J* to specify the start and the end of a word. Then, we use an auto-encoder to reduce the size of this bag of n-grams vector to d-dimension. To check the performance of the resulting orthographic representation, a neural network is trained to predict a word given this orthographic representation. It reaches 99.99% of accuracy on the training set composed of $52k$ words of the vocabulary, showing the richness of this representation.

Similar to [8], a DNN was trained by using the triplet ranking loss [8, 15, 16] in order to project the orthographic word representation to the acoustic embeddings **s** obtained from the CNN architecture, which is trained independently. It takes as input a word orthographic representation and outputs an embedding vector of the same size as **s**. During the training process, this model takes as inputs the acoustic embedding **s** selected randomly from the training set, the orthographic representation of the matching word $\mathbf{o}^+$, and the orthographic representation of a randomly selected word different to the first word $\mathbf{o}^-$. These two orthographic representations supply shared parameters in the DNN.

We call $t = (\mathbf{s}, \mathbf{w}^+, \mathbf{w}^-)$ a triplet, where **s** is the acoustic signal embedding, $\mathbf{w}^+$ is the embedding obtained through the DNN for the matching word, while $\mathbf{w}^-$ is the embedding obtained for the wrong word. The triplet ranking loss is defined as:

$$Loss = \max(0, m - Sim_{dot}(s, w^+) + Sim_{dot}(s, w^-)) \quad (1)$$

where $Sim_{dot}(x, y)$ is the dot product function used to compute the similarity between two vectors x and y, and m is a margin parameter that regularizes the margin between the two pairs of similarity $Sim_{dot}(\mathbf{s}, \mathbf{w}^+)$ and $Sim_{dot}(\mathbf{s}, \mathbf{w}^-)$. This loss is weighted according to the rank in the CNN output of the word matching the audio signal.

The resulting trained model can then be used to build an acoustic embedding ($\mathbf{w}^+$) from any word, as long as one can extract an orthographic representation from it.

3.2. Evaluation

In the literature [6, 7], the evaluation of the acoustic word embeddings was conducted on a word discrimination task developed for this purpose [17]. This task consists on deciding whether two words are similar or not based on their acoustic representation. In [6, 7], the authors use two collections of words (train and test) from the Switchboard English corpus for the evaluation. For each pair of words in the test set the cosine distance is computed between their embeddings. The two words are classified as similar or different by applying a threshold on their distance, and a precision-recall curve is obtained by varying the threshold.

In this study, we propose to build different evaluation sets in French language in order to assess the acoustic word embeddings ($\mathbf{w}^+$) performances on *orthographic* and *phonetic similarity* and *homophones detection* tasks. As a remainder, the

acoustic word embedding $\mathbf{w}^+$ is a projection into the space of acoustic signal embeddings $\mathbf{s}$ of an orthographic word representation $\mathbf{o}^+$. In our evaluation, we would like to measure the loss of orthographic information carried by $\mathbf{w}^+$ and the potential gain of acoustic information due to this projection, in comparison to the information carried by $\mathbf{o}^+$.

The evaluation sets are built along the following lines: given a list L of n frequent words (candidate words) in the vocabulary V composed of m words, a list of $n * m$ word pairs was created. Then, an alignment was performed between each word pair based on whether their orthographic (letters) or phonetic (phonemes) representations, using the sclite[2] tool. From these alignments, two *edition distances* were computed corresponding to the alignment results of orthographic or phonetic representations.

Once we have the two *edition distances* for each candidate word in the list L, we extract its orthographically and phonetically nearest words, which results the following lists: Orthographic: word pairs orthographically similar; Phonetic: word pairs phonetically similar; those lists are used for *orthographic* and *phonetic similarity* tasks, while for *homophone* detection, Homophones list is used, which contains pairs of homophone words. Orthographic and Phonetic lists are sorted according to the *edition distances* of the word pairs.

In the case of the orthographic (*Sim-Ortho*) and phonetic similarity (*Sim-Phon*) tasks, the evaluation of the acoustic embeddings is performed by ranking the pairs according to their cosine similarities and measuring the Spearman's rank correlation coefficient (Spearman's ρ) with the corresponding *edition distance*. However, for homophone detection task (*Homoph-Det*), the evaluation is performed in terms of precision (P), which is defined as the ratio of the number of correct homophones found over the number of homophones in the list.

4. Experiments on acoustic word embeddings

4.1. Experimental data

The training set for the CNN consists of 488 hours of French Broadcast News with manual transcriptions. This dataset is composed of data coming from the ESTER1 [18], ESTER2 [19] and EPAC [20] corpora.

It contains $52k$ unique words that are seen at least twice each in the corpus. All of them corresponds to a total of 5.75 millions occurrences. In French language, many words have the same pronunciation without sharing the same spelling, and they can have different meanings; *e.g.* the sound [so] corresponds to four homophones: *sot* (fool), *saut* (jump), *sceau* (seal) and *seau* (bucket), and twice more by taking into account their plural forms that have the same pronunciation: *sots*, *sauts*, *sceaux*, and *seaux*. When a CNN is trained to predict a word given an acoustic sequence, these frequent homophones can introduce a bias to evaluate the recognition error. To avoid this, we merged all the homophones existing among the $52k$ unique words of the training corpus. As a result, we obtained a new reduced dictionary containing $45k$ words and classes of homophones.

Acoustic features provided to the CNN are log-filterbanks, computed every 10ms over a 25ms window yielding a 23-dimension vector for each frame. A forced alignment between manual transcriptions and speech signal was performed on the training set in order to detect word boundaries. The statistics computed from this alignment reveal that 99% of words are shorter than 1 second. Hence we decided to represent each word by 100 frames, thus, by a vector of 2300 dimensions. When words are shorter they are padded with zero equally on both ends, while longer words are cut equally on both ends.

[2]http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

4.2. Architectures

The CNN and DNN deep architectures are trained on 90% of the training set and the remaining 10% are used for validation.

- CNN: this architecture predicts a word given a sequence of 100 frames. It contains two convolution and max-pooling layers followed by two fully-connected layers, which feed into the final Softmax layer over $45k$ words and classes of homophones. The convolution layers have respectively 15 and 10 filters over 8 frames. While, the max pooling layers performs max pooling over 4 units. More, the two fully connected layers composed with 500 and 100 units respectively. The hyperbolic tangent ($Tanh$) function is used as an activation function for all the layers. This model achieves 61.51% of accuracy (*i.e.* word recognition precision) on the validation set.
- DNN: this architecture has to map the word orthographic representation into the acoustic embeddings space obtained by the CNN model. It is composed of two fully connected $Tanh$ layers of 300 and 100 units each. When replacing in the CNN the signal embedding $\mathbf{s}$ (*Cf.* figure 1) by the acoustic word embedding $\mathbf{w}^+$ computed by the DNN model from the orthographic embedding $\mathbf{o}^+$, the accuracy on the validation set becomes 50.15%. This shows that even if acoustic word embeddings are less precise than signal embeddings, they are able to capture relevant acoustic information in a continuous space close to the signal embeddings space.

4.3. Acoustic word embeddings evaluation

This section reports the evaluation results of the acoustic word embeddings $\mathbf{w}^+$ on orthographic similarity, phonetic similarity, and homophones detection tasks. The embeddings we evaluate are built from two different vocabularies:

1. the one used to train the neural network models, composed of the $52k$ words presented in the manual transcriptions of the 488 hours of audio.
2. the vocabulary of the ASR system used in our experiments to process French broadcast news, composed of $160k$ words. The words present in the $52k$ vocabulary are nearly all present in the $160k$ vocabulary.

The evaluation sets described on section 3.2 are generated from these two vocabularies: in the $52k$ vocabulary, all the acoustic word embeddings $\mathbf{w}^+$ are related to words which have been observed during the training of the CNN. This means that at least two acoustic signal embeddings have been computed from the audio for each one of these words; in the $160k$ vocabulary, about $110k$ acoustic word embeddings $\mathbf{w}^+$ were computed for words never observed in the audio data. Results are summarized in table 1.

They show that the acoustic word embeddings $\mathbf{w}^+$ are more relevant for the phonetic similarity task, while $\mathbf{o}^+$ are obviously the best ones on the orthographic similarity task.

As presented above, $\mathbf{w}^+$ is a projection of $\mathbf{o}^+$ into the acoustic space of $\mathbf{s}$. These results show that the projection of the orthographic embeddings into the acoustic embeddings space

Task	52K Vocab.		160K Vocab.	
	$\mathbf{o}^+$	$\mathbf{w}^+$	$\mathbf{o}^+$	$\mathbf{w}^+$
Sim-Ortho	**54.28**	49.97	**56.95**	51.06
Sim-Phon	40.40	**43.55**	41.41	**46.88**
Homoph-Det	64.65	**72.28**	52.87	**59.33**

Table 1: Evaluation results of similarity ($\rho \times 100$) and homophone detection tasks (*precision*).

changes their properties: they have captured more information about word pronunciation while they have lost information about spelling. So, in addition to making possible a measure of similarity distance between the acoustic signal (represented by **s**) and a word (represented by $\mathbf{w}^+$), acoustic word embeddings are better than orthographic word embeddings to measure the phonetic proximity between two words.

For the homophone detection task, we compute all the homophone pairs from the $160k$ vocabulary: that results to 53869 pairs in total. The $52k$ vocabulary contains 13561 homophone pairs which are all included in the pairs present in the $160k$ vocabulary. As we can see, the $\mathbf{w}^+$ acoustic embeddings outperform the orthographic ones on this task in both cases. This confirms that acoustic word embeddings have captured additional information about word pronunciation than the one carried by orthographic word embeddings. While the precision measure is dependent to the number of events, for this metric we cannot compare the results between the ones got from the $52k$ vocabulary and the ones obtained from $160k$ vocabulary. For the Spearman's correlation, a comparison is roughly possible and results show that the way to compute $\mathbf{w}^+$ is effective to generalize this computation to word not observed in the audio training data.

5. Experiments on ASR error detection

Experimental data are based on the entire official ETAPE corpus [21], composed by audio recordings of French Broadcast News Shows with manual transcriptions. It is enriched with automatic transcriptions generated by the LIUM ASR system, detailed in [22]. This ASR system won the ETAPE evaluation campaign. This corpus is divided into three sets: Train, Dev and Test, which are composed respectively, of $349K$, $54K$, and $58K$ words. Their word error rates are 25.3%, 24.6% and 21.9% respectively.

This section reports the experimental results of the ASR error detection system *MLP-MS*. The performance of the proposed approach is compared with a state-of-the-art system based on the CRF tagger provided by *Wapiti*[3] and applied to the set of features presented in section 2, named *baseline features* (**B-Feat.**) in the remainder of the paper. The performance is evaluated by using recall (R), precision (P) and F-measure (F) for the erroneous word prediction and global Classification Error Rate (CER). CER is defined as the ratio of the number of misclassifications over the number of recognized words.

Our baseline neural system is based on the use of an effective combined word embeddings. The latter results from the combination of word2vecf, skip-gram, and GloVe, using an auto-ecoder. The combination of these embeddings results to a 200-dimensional vector. The 200-dimensional word embeddings were computed from a large textual corpus composed of about 2 billions of words. This corpus was built from articles of the French newspaper "Le Monde", the French Gigaword corpus, articles provided by Google News, and manual transcriptions of about 400 hours of French broadcast news. It contains dependency parses used to train word2vecf embedding, while the unlabeled version is used to train skip-gram and GloVe.

[3]http://wapiti.limsi.fr

Experimental results are summarized in Table 2. We observe that our MLP-MS system, called NN (**B-Feat.**) in Table 2, yields significant improvements compared to the state-of-the-art CRF approach, it achieves 6.02% and 5.72% of CER reduction on Dev and Test.

With the purpose to evaluate the performance of acoustic word embeddings on ASR error detection, they were used as additional features into the MLP-MS system. Three settings are compared corresponding to different types of features used in MLP-MS: baseline features (**B-Feat.**) alone, then adding the signal acoustic embeddings **s**, and finally adding the acoustic embeddings $\mathbf{w}^+$. Experimental results reported in Table 2 show the usefulness of the acoustic embeddings **s**, that is able to characterize some suspicious acoustic segments. It yields an improvement in terms of CER reduction compared to the results of the baseline (NN (**B-Feat.**)). An additional slight improvement is observed by adding the acoustic embedding $\mathbf{w}^+$. Comparing to the CRF system, our entire neural approach using acoustic signal embeddings **s** and acoustic word embeddings $\mathbf{w}^+$ achieves statistically significant improvements, by respectively 8.18% and 7.24% in terms of CER reduction on Dev and Test.

Corp.	App.	Label error			Global CER
		P	R	F	
Dev	CRF	68.11	55.37	61.08	10.38
	NN (**B-Feat.**)	70.50	57.56	63.38	9.79
	+ **s**	**71.98**	57.63	64.01	9.54
	+ **s** + $\mathbf{w}^+$	71.70	**58.25**	**64.28**	**9.53**
Test	CRF	67.69	54.74	60.53	8.56
	NN (**B-Feat.**)	69.66	57.89	63.23	8.07
	+ **s**	69.64	**59.13**	63.95	7.99
	+ **s** + $\mathbf{w}^+$	**70.09**	58.92	**64.02**	**7.94**

Table 2: Performance of acoustic word embeddings.

6. Conclusions

In this paper, we have investigated the intrinsic evaluation of acoustic word embeddings, and their impact on the ASR error detection task. We have proposed two approaches to evaluate the performances of these acoustic word embeddings and compare them to their orthographic embeddings: orthographic and phonetic performance by ranking pairs and measuring the Spearmans rank correlation coefficient (Spearman's ρ), and by measuring the precision in a homophone detection task. Experiments were made on automatic transcriptions generated by LIUM ASR system applied on the ETAPE corpus (French broadcast news). They show that the proposed neural architecture, using the acoustic word embeddings as additional features, outperforms state-of-the-art approach based on the use of Conditional Random Fields (CRF).

7. Acknowledgements

This work was partially funded by the European Commission through the EUMSSI project, under the contract number 611057, in the framework of the FP7-ICT-2013-10 call, by the French National Research Agency (ANR) through the VERA project, under the contract number ANR-12-BS02-006-01, and by the Région Pays de la Loire.

8. References

[1] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[2] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual Information Improves OOV Detection in Speech," in *Human Language Technologies: Proceddings of the North American Chapter of the Association for Computational Linguistics (NAACL'10)*, 2010.

[3] F. Béchet and B. Favre, "ASR error segment localization for spoken recovery strategy," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 6837–6841.

[4] S. Jalalvand and D. Falavigna, "Stacked auto-encoder for ASR error detection and word error rate prediction," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2142–2146.

[5] S. Ghannay, Y. Estève, N. Camelin, C. Dutrey, F. Santiago, and M. Adda-Decker, "Combining continous word representation and prosodic features for ASR error prediction," in *3rd International Conference on Statistical Language and Speech Processing (SLSP 2015)*, Budapest (Hungary), November 24-26 2015.

[6] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *arXiv preprint arXiv:1510.01032*, 2015.

[7] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 410–415.

[8] S. Bengio and G. Heigold, "Word embeddings for speech recognition." in *INTERSPEECH*, 2014, pp. 1053–1057.

[9] S. Ghannay, Y. Estève, and N. Camelin, "Word embeddings combination and neural networks for robustness in ASR error detection," in *European Signal Processing Conference (EUSIPCO 2015)*, Nice (France), 31 aug.-4 sept. 2015.

[10] A. Nasr, F. Béchet, J.-F. Rey, B. Favre, and J. Le Roux, "Macaon: An nlp tool suite for processing word lattices," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*. Association for Computational Linguistics, 2011, pp. 86–91.

[11] S. Ghannay, B. Favre, Y. Estève, and N. Camelin, "Word embedding evaluation and combination," in *10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož (Slovenia), 23-28 May 2016.

[12] O. Levy and Y. Goldberg, "Dependency based word embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, 2014, pp. 302–308.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013.

[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global Vectors for Word Representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[15] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[16] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *IJCAI*, vol. 11, 2011, pp. 2764–2770.

[17] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid Evaluation of Speech Representations for Spoken Term Discovery." in *INTERSPEECH*, 2011, pp. 821–824.

[18] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French Broadcast News." in *Interspeech*, 2005, pp. 1149–1152.

[19] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts." in *Interspeech*, vol. 9, 2009, pp. 2583–2586.

[20] Y. Estève, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas, "The EPAC Corpus: Manual and Automatic Annotations of Conversational Speech in French Broadcast News." in *LREC*. Citeseer, 2010.

[21] G. Gravier, G. Adda, N. Paulsson, M. Carr, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

[22] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?" in *Interspeech*, Brighton, UK, September 2009.