Enriching confusion networks for post-processing

Sahar Ghannay, Yannick Estève, Nathalie Camelin

Enriching confusion networks for post-processing

Sahar Ghannay, Yannick Estève, Nathalie Camelin

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

Abstract. The paper proposes a new approach for *a posteriori* enrichment of automatic speech recognition (ASR) confusion networks (CNs). CNs are usually needed to decrease word error rate and to compute confidence measures, but they are also used in many ways in order to improve post-processing of ASR outputs. For instance, they can be helpfully used to propose alternative word hypotheses when ASR outputs are corrected by a human on post-edition. However, CNs bins do not have a fixed length, and sometimes contain only one or two word hypotheses: in this case the number of alternatives to correct a misrecognized word is very low, reducing the chance of helping the human annotator.

Our approach for CN enrichment is based on a new similarity measure presented in this paper, computed from acoustic and linguistic word embeddings, that allows us to take into consideration both acoustic and linguistic similarities between two words.

Experimental results show that our approach is relevant: enriched CNs (for a bin size equals to 6) increase the potential correction of erroneous words by 23% than initial CNs produced by an ASR system. In our experiments, a spoken language understanding task is also targeted.

Index Terms: Confusion networks, post processing, acoustic and linguistic word embeddings, similarity measure.

1 Introduction

Despite of the recent advances in the field of speech processing, automatic speech recognition errors are still unavoidable. This reflects the sensitivity of this technology to 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 subtitling, computer assisted transcription, speech to speech translation, spoken language understanding, information retrieval, *etc.* Error detection and correction aim to improve the exploitation of ASR outputs by downstream applications.

Many studies have focused on ASR error detection and correction, some of them [1–3] have attempted to improve recognition accuracy for many tasks such as keyword search, spoken language understanding and other tasks by using discriminative post-processing on ASR outputs.

Other studies consider the use of automatic speech recognition confusion networks (CNs) to decrease word error rate and to compute confidence measure. These networks were introduced in [4]. They rely on *posterior* probabilities and were used to represent a set of alternative sentences. Authors in [5] propose an approach to automatically correct erroneous words in the CNs. It depends on the use of the n-grams and the semantic score between words that are located far from each other based on Normalized Relevance Distance. Confusion networks can be used as well in many ways to improve post-processing of ASR outputs. For instance, they can be used to propose alternative word hypotheses when ASR outputs are corrected by a human on post-edition [6]. However, CN bins, sets of competing hypothesis between two nodes in the CN, do not have a fixed length, and sometimes contain only one or two word hypotheses: in this case the number of alternatives to correct a misrecognized word is very low, reducing the chance of helping the human annotator.

In this study, we propose an approach for CN enrichment, which is based on a similarity measure computed from acoustic and linguistic word embeddings. This measure allows us to take into consideration both acoustic and linguistic similarities between two words. Since our assumption is that word hypotheses in a same bin should be close in term of acoustics and/or linguistics, we propose to use this particularity to enrich confusion networks by applying this new similarity measure. This enrichment will be evaluated in the context of a human post-edition of automatic transcripts. Moreover, the proposed similarity measure can be used in a spoken language understanding (SLU) system in order to propose semantically relevant alternative words to ASR outputs. Last, this similarity measure is applied as well for prediction of potential ASR errors for rare words.

2 Word embeddings

Many neural approaches have been proposed to build word embeddings, they can be based on continuous bag of words, syntactic dependency, co-occurrences matrix, and even audio signal. Hence, they can capture different types of information: semantic, syntactic, and acoustic.

2.1 Linguistic word embeddings

Word embeddings were initially introduced through the construction of neural language models [7, 8]. They are defined as projections in a continuous space of words in a manner that preserve semantic and syntactic similarities.

Following the results published in [9] in which different kinds of word embeddings are evaluated and different word embeddings combinations are compared, we use a combination of word embeddings to get better results. It has been shown in [9] that the combination through PCA (Principal Component Analysis) achieves the best performance in the analogical and similarity tasks. Since the approach we propose is based on the cosine similarity too, we suggest to use PCA to combine *word2vecf* on dependency trees [10], *skip-gram* provided by *word2vec* [11], and *GloVe* [12].

We considered word embeddings presented here as linguistic representations of words, since they are built based on lexical, contextual, and syntactic information.

2.2 Acoustic word embeddings

Recent studies have started to reconsider the use of whole words as the basic modeling unit in speech recognition and query applications, instead of phonetic units. These systems are based on the use of a function that embeds an arbitrary or fixed dimensional speech segment into a continuous space, named acoustic embeddings, in a such way that speech segments of words that sound similarly will have similar embeddings. These representations were successfully used in a query-by-example search system [13, 14], in a segmental ASR lattice re-scoring system [15] and recently for ASR error detection [16].

In [15], the authors proposed an approach to build acoustic word embeddings of words observed in an audio corpus, and also of words never observed in this corpus, by exploiting their orthographic representation. Moreover, a such acoustic word embedding derived from an orthographic representation can be perceived as a canonical acoustic representation for a word, since different pronunciations imply different acoustic embeddings. This approach relies on the use of two neural architectures: a convolutional neural network classifier over words trained independently to build acoustic embeddings, and a deep neural network (DNN) trained by using a triplet ranking loss function [15, 17, 18] in order to project the orthographic word representation to the acoustic embeddings space, that results the acoustic word embeddings $\mathbf{w}^+$. The orthographic word representation consists on a bag of n-grams of letters ($n \leq 3$), in which we reduce its dimension using an auto-encoder, that results the orthographic embeddings $\mathbf{o}^+$.

In another previous study [19], we have investigated the evaluation of the intrinsic performances of acoustic word embeddings, and compare them to their orthographic embeddings, on orthographic, phonetic similarities and homophone detection tasks. As a reminder, we report in table 1 some results obtained in that study.

Tasks	160K Vocab.	
	$\mathbf{o}^+$	$\mathbf{w}^+$
Orthographic	**56.95**	51.06
Phonetic	41.41	**46.88**
Homophone	52.87	**59.33**

Table 1: Evaluation results of similarity ($\rho \times 100$) and homophone detection tasks (*precision*). ρ corresponds to the Spearman's rank correlation coefficient.

As shown in this table, the acoustic word embeddings are better than orthographic ones to measure the phonetic proximity between two words. Moreover, they are better too to detect homophone words. These results confirm that acoustic word embeddings have captured additional information about word pronunciation in addition to the information carried by their spelling. In this study, the acoustic word embeddings are used as an acoustic representation of the words.

3 Similarity measure to enrich confusion networks

In this study, we propose to use both linguistic and acoustic word embeddings to *a posteriori* enrich confusion networks, in order to improve post-processing of ASR outputs. Due to the nature of acoustic and langage models involved in an ASR system, our assumption is that word hypotheses in a same bin should be close from acoustic and/or linguistic points of view.

Since we aim to enrich confusion networks by adding nearest neighbors of the recognized word hypotheses, this neighborhood has to be characterized: it is done through the cosine similarities of acoustic and linguistic word embeddings.

With the purpose to take benefit from both linguistic and acoustic similarities, we propose to use a linear interpolation to combine them. This results to a similarity called $LA_{SimInter}$, defined as:

$$LA_{SimInter}(\lambda, x, y) = (1 - \lambda) \times L_{Sim}(x, y) + \lambda \times A_{Sim}(x, y) \tag{1}$$

where x and y are two words, λ is the interpolation coefficient, while L_{Sim} and A_{Sim} are respectively the linguistic and acoustic similarities computed with the cosine similarity applied to respectively the linguistic and acoustic word embeddings of x and y.

Since our goal is to enrich confusion networks and use them to propose alternative word hypotheses to correct ASR outputs, we aim to optimize the λ value for this purpose. To estimate λ, a list of known substitution errors made by an ASR system is used. Let define h an erroneous word hypothesis and $\overline{r}$ the reference word that is substituted with h.

For each word pairs $(h, \overline{r})$ in the list, we compute the probability of using h when the reference word $\overline{r}$ is wrong, *i.e.* the probability of substituting the reference word with the hypothesis one, which is defined as:

$$P(h|\overline{r}) = \frac{\#(h, \overline{r})}{\#\overline{r}} \tag{2}$$

where $\#(h, \overline{r})$ refers to the number of occurrences of the word pair and $\#\overline{r}$ is the number of errors (deletion + substitution) on the reference word.

Based on the similarity score $LA_{SimInter}(\lambda, h, \overline{r})$ and the probability $P(h|\overline{r})$, we choose the interpolation coefficient $\hat{\lambda}$ that minimizes the mean squared error (MSE) as:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\, MSE(\forall(h, \overline{r}) : P(h|\overline{r}), LA_{SimInter}(\lambda, h, \overline{r})) \tag{3}$$

This choice is not optimal since similarities are not normalized in function of the number of errors related to one word in the vocabulary whereas probabilities are, but we assume this approach provides an acceptable approximation in the search of the λ value that aims to combine $L_{Sim}(x, y)$ and $A_{Sim}(x, y)$ in order to predict the confusability betwen x and y.

By using $LA_{SimInter}$ with $\hat{\lambda}$, it is now possible to propose for a given word its linguistically and acoustically nearest neighbors.

Table 2 shows an example of hypothesis word and its nearest neighbors. As expected, the neighbors of any given word seem linguistically similar when induced by linguistic word embeddings, and sound like it when they are induced by the acoustic ones. By combining acoustic and linguistic word similarities ($LA_{SimInter}$), it is also possible to restrict the neighborhood to words close to any given word both linguistically and acoustically.

Nearest neighbors of the French word 'portables', pronounced \pɔʁtabl\		
L_{Sim}	A_{Sim}	$LA_{SimInter}$
téléphones, ordinateurs, portable, portatif *telephones, computers, portable, portable* \telefɔn\\\ɔʁdinatœʁ\\\pɔʁtabl\\\pɔʁtatif\	portable, portant, portants, portait *portable, carrying, racks, carried* \pɔʁtabl\\\pɔʁtã\\\pɔʁtã\\\pɔʁtɛ\	portable, portant, portatif, portait *portable, carrying, portative, carried* \pɔʁtabl\\\pɔʁtã\\\pɔʁtã\\\pɔʁtɛ\

Table 2: Nearest neighbors of the hypothesis word 'portables', with some translations in English and their pronunciation in French. 'portables' is a French word pronounced \pɔʁtabl\that can be translated to the same word 'portables' in English

4 Experimental setup

We present in this section the performance of the similarity measure $LA_{SimInter}(\lambda, h, \overline{r})$ on two tasks: prediction of ASR potential errors for rare words and enrichment of confusion networks.

4.1 Computation of linguistic and acoustic embeddings

The 100-dimensional linguistic word embeddings result from the combination of word2vecf, skip-gram, and GloVe, using PCA. The 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 embeddings, while the unlabeled version is used to train skip-gram and GloVe [20].

The training set for the convolution neural network consists of 488 hours of French Broadcast News with manual transcriptions. This dataset is composed of data coming from the ESTER1 [21], ESTER2 [22] and EPAC [23] 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.

Acoustic features provided to the convolution neural network are log-filterbanks, computed every 10ms over a 25ms window yielding a 23-dimensions vector for each frame. Each word is represented 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. Once the acoustic embeddings are built, we build orthographic embeddings from the vocabulary compose of $52k$ words, and train the DNN architecture to build the acoustic word embeddings.

The resulting model is used to build 100-dimensional acoustic word embeddings from the same vocabulary as the linguistic ones. A detailed description of the training data of the architectures used to build these acoustic word embeddings is presented in [16].

4.2 Experimental data

Experimental data is based on the entire official ETAPE corpus [24], composed by audio recordings of French broadcast news shows, with manual transcriptions. This corpus is enriched with automatic transcriptions generated by the LIUM ASR system, detailed in [25], which won the ETAPE evaluation campaign. Its vocabulary contains 160k words.

The automatic transcriptions have been aligned with reference transcriptions using the *sclite*[1] tool. From this alignment, one can derive the lists of errors produced by our ASR system. The experimental data is divided into two sets: Train and Test, which are composed respectively of $399K$ and $58K$ words. Their word error rates are 25.2% and 21.9% respectively. More, they have respectively 10.3% and 8.3% of substitution errors.

For this task, we will use the list of substitution errors of Train to compute the interpolation coefficient $\hat{\lambda}$, while the list of Test will be used to evaluate the performance of our approach to enrich confusion networks and to correct erroneous word hypotheses. This list is composed of 4678 occurrences of substitution error pairs, named Sub_{Test} further in the paper. For these substitution error pairs we use their corresponding confusion bins.

Figure 1 illustrates the percentage of the confusion bins according to the number of their alternative words (*i.e* words in concurrence with the 1-best hypothesis). The CN bins do not have a fixed length and 55% of them contain none or only one alternative word, that justify our aim about CN enrichment. The CN bins that have a size between 6 and 12 are grouped into a single class [6-12].

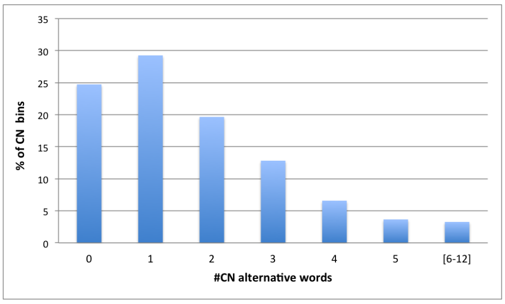

Fig. 1: Percentage of confusion network bins according to their size.

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

4.3 Tasks and evaluation score

We propose in this study two evaluation tasks: the prediction of errors for rare words (task1) and the correction of ASR errors (task2).

Given a word pair (a,b) in a list L of m substitution errors, the evaluation tasks consist on looking for the word b in the list $N(a, \mathit{\Gamma}, n)$ of the n nearest neighbors of a, computed through the similarity $\mathit{\Gamma}$. In our experiments, the similarity can be L_{Sim}, A_{Sim} or $LA_{SimInter}$.

The evaluation score is calculated by varying the size n and computing the precision at n of finding the word b. The precision at n computed for all the word pairs in the list L, taking into account their occurrence frequencies in the evaluation corpus, is called $S(\mathit{\Gamma}, n)$ and computed as follows:

$$S(\mathit{\Gamma}, n) = \frac{\sum_{i=1}^{m} f(i, \mathit{\Gamma}, n) \times \#(a_i, b_i)}{\sum_{i=1}^{m} \#(a_i, b_i)} \tag{4}$$

where f is defined as :

$$f(i, \mathit{\Gamma}, n) = \begin{cases} 1 & \textit{if}\, b_i \subset N(a_i, \mathit{\Gamma}, n) \\ 0 & \textit{otherwise} \end{cases}$$

where i corresponds to the i^{th} word pair (a_i, b_i) of L, a_i and b_i are defined according to the evaluation tasks:

- task1: b_i corresponds to the hypothesis word h and a_i is the reference word $\overline{r}$,
- task2: b_i corresponds to reference word $\overline{r}$ and a_i is the hypothesis word h.

5 Experimental results

5.1 Prediction of potential ASR errors for rare words

To compare the performance of the combined similarity to the linguistic and acoustic ones, we evaluate them on ASR errors prediction task for rare words. These latter are defined as the reference words not seen in the training corpus of the ASR system. This is why the Sub_{Test} list was filtered to keep only the errors (misrecognized reference words) not seen in Train. It is composed of 538 pairs of substitution errors, named $Sub_{TestRarewords}$. For each reference word $\overline{r}$ in the $Sub_{TestRarewords}$ we derive their 30 nearest neighbors from the ASR vocabulary, based on linguistic, acoustic or combined similarities. That results to three similarity lists named respectively $List_{SimL}$, $List_{SimA}$, and $List_{SimInter}$.

Figure 2 illustrates the results of predicting errors for rare words using the lists described above, by varying their sizes from 1 to 30. We observe that the results are in favor of $List_{SimInter}$ followed by $List_{SimA}$: this shows that our proposition to optimize the interpolation weight to combine $List_{SimL}$ and $List_{SimA}$ is relevant. The interesting area in this figure is the left part, which shows the results of the prediction when the list of errors is short. When this list is composed of only one word, the prediction is correct 11% of the time. This must be analyzed in light of the vocabulary size of the ASR system, which contains 160k words: each word of the vocabulary can be selected in a list of error prediction. The prediction is correct 20% of the time when the size of the $List_{SimInter}$ list is 12. It seems that this list reaches then a plateau. The combined similarity will be used for the remaining experiments.

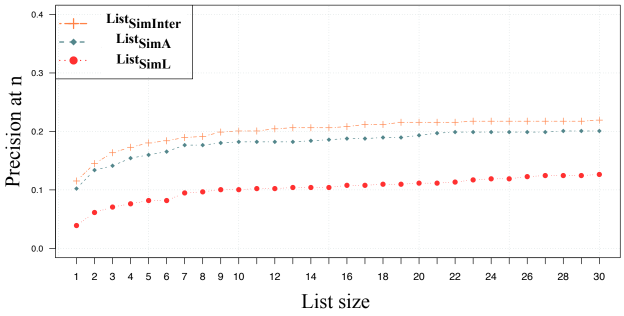

Fig. 2: Performance of predicting ASR errors for rare words by varying the size of the lists.

5.2 Enrichment of confusion networks

The enrichment of confusion networks can be used for post-processing of automatic transcriptions, or to enrich the automatic transcriptions provided for a spoken language understanding system.

Post-processing of automatic transcriptions For each hypothesis word (h) in Sub_{Test} we derive their 6 nearest neighbors from the ASR vocabulary, based on the combined linguistic and acoustic similarity $LA_{SimInter}$. The resulting list is named $List^{h}_{SimInter}$. Then, for each word pair $(h, \overline{r})$ in Sub_{Test} we enrich their corresponding confusion bins with the nearest neighbors of the hypothesis word (h) from $LA_{SimInter}$, to have a bin size equals to 6 (this size seems relevant to visualize alternative words in a graphical user interface in a computer-assisted transcription software [26]). The list of competing words in the confusion bin is named $List_{CN}$, and the one in the enriched confusion bin is named $List_{EnrichCN}$.

We evaluate the performance of the resulting lists on erroneous word hypotheses correction task. In this task, we try to see, when there is a recognition error, whether the correct word $(\overline{r})$ was in the nearest neighbors (or confusion bin) of the recognized word (h). As shown in table 3, experimental results show that our approach is relevant: enriched confusion networks permit to increase the precision at 6 of more than 23% in comparison to CN produced by our ASR system. Notice that the P@6 value for the alternative words proposed by the $List_{\mathbf{SimInter}}$ alone is 0.11.

Spoken language understanding task The approach we propose can be useful for a spoken language understanding task, to correct the semantically relevant erroneous

	List$_{\text{CN}}$	**List**$_{EnrichCN}$
P@6	0.17	**0.21 (+23.5%)**

Table 3: Performance of CN enrichment: evaluation of $List_{CN}$ and $List_{EnrichCN}$ on erroneous word hypotheses correction task in terms of precision at 6.

word hypotheses. However, in the case of only the 1-best ASR hypotheses was provided to the dialogue manager, one can use the proposed similarity metric to expand this 1-best hypotheses and build confusion networks. This scenario in which getting access to only the 1-best ASR hypotheses is frequent in industry, especially when the semantic interpretation module is fed by outputs generated by an ASR system from a third party in the cloud.

For this experiment, we use the automatic transcriptions of the French MEDIA corpus [27,28] generated by a variant of the ASR system developed by LIUM that won the last evaluation campaign on French language [29]. This variant system contains 2.5K words and its language model is adapted to the MEDIA data. The purpose of the MEDIA corpus is to evaluate spoken language understanding systems. Often the SLU task derived from the MEDIA corpus consists on labeling recognized words with semantic concepts [30]. For a such task, a misrecognized word implies usually an error of labeling, that can be prevented by using confusion networks or word-lattices [31], when available. We expect to propose relevant alternative words in the scenario where only the 1-best hypothesis is available.

The automatic transcriptions were aligned with the reference ones in order to extract the list of substitution errors produced by the ASR system. This list is divided into two sets: Train to compute the interpolation coefficient $\hat{\lambda}$, which is enriched with Train Etape used for the previous experiments. Test is used for the evaluation, and has been filtered to keep only 1204 occurrences of semantically relevant erroneous words, based on the semantic labels. Since the size of MEDIA vocabulary is limited to 2.5K words, it is enriched with the vocabulary composed with 160K words.

For each hypothesis word (h) in Test list, we derive their 6 nearest neighbors from the ASR MEDIA vocabulary, based on the combined linguistic and acoustic similarity $LA^{h}_{SimInter}$.

By using the resulting list, one can enrich the one-best hypotheses produced by the ASR system and compute how many times we propose the correct word to recognize as an alternative in this list. Experimental results show that, thanks to our proposition, it is possible to potentially retrieve 20.6% of semantically relevant words that were initially misrecognized.

6 Conclusions

Assuming that word hypotheses in a same confusion network bin should be close in term of acoustics and/or linguistics, we propose to take benefit from linguistic and acoustic word embeddings to enrich confusion networks, in order to improve post-processing of ASR outputs.

We propose an approach to compute a similarity function, called $LA_{SimInter}$, which is optimized to ASR error correction. We show that this function allows us to compute relevant lists of nearest neighbors linguistically and acoustically. This list is used successfully to enrich the confusion networks and to increase the potential correction of erroneous words by 23% in comparison to initial confusion networks provided by the ASR system. Moreover, when used in a spoken language understanding task, this approach permits to propose 6 alternative words to 1-best hypotheses carrying on semantics to be exploited by the SLU module. When the ASR hypothesis is wrong on a word supporting a semantic concept, these alternatives contain the correct word in 20.6% of the cases.

Through our contribution and experimental results, we show that it is possible to relevantly enrich confusion networks by applying a similarity computed from linguistic and acoustic word embeddings.

References

1. S. Stoyanchev, P. Salletmayr, J. Yang, and J. Hirschberg, "Localized detection of speech recognition errors," in *Spoken Language Technology Workshop (SLT), 2012 IEEE.* IEEE, 2012, pp. 25–30.
2. E. Pincus, S. Stoyanchev, and J. Hirschberg, "Exploring features for localized detection of speech recognition errors," in *Proc. of the SIGDIAL Conference. ACL*, 2013, pp. 132–136.
3. V. Soto, E. Cooper, L. Mangu, A. Rosenberg, and J. Hirschberg, "Rescoring confusion networks for keyword search," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on.* IEEE, 2014, pp. 7088–7092.
4. L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.
5. Y. Fusayasu, K. Tanaka, T. Takiguchi, and Y. Ariki, "Word-error correction of continuous speech recognition based on normalized relevance distance." in *IJCAI*, 2015, pp. 1257–1262.
6. A. Laurent, S. Meignier, T. Merlin, and P. Deléglise, "Computer-assisted transcription of speech based on confusion network reordering," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 4884–4887.
7. Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," vol. 3. JMLR.org, mar 2003, pp. 1137–1155.
8. H. Schwenk, "CSLM-a modular open-source continuous space language modeling toolkit." in *INTERSPEECH*, 2013, pp. 1198–1202.
9. 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.
10. 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.
11. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Workshop at ICLR*, 2013.
12. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, vol. 12, 2014.
13. H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *arXiv preprint arXiv:1510.01032*, 2015.

14. 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.
15. S. Bengio and G. Heigold, "Word embeddings for speech recognition." in *INTERSPEECH*, 2014, pp. 1053–1057.
16. S. Ghannay, Y. Estève, N. Camelin, and P. Deleglise, "Acoustic word embeddings for asr error detection," in *Interspeech 2016*, San Francisco (CA, USA), 9-12 September 2016.
17. 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.
18. J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *IJCAI*, vol. 11, 2011, pp. 2764–2770.
19. S. Ghannay, Y. Estève, N. Camelin *et al.*, "Evaluation of acoustic word embeddings," *ACL 2016*, p. 62, 2016.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. P. Cardinal, G. Boulianne, M. Comeau, and M. Boisvert, "Real-time correction of closed-captions," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, 2007, pp. 113–116.
27. H. Bonneau-Maynard, M. Quignard, and A. Denis, "Media: a semantically annotated corpus of task oriented dialogs in french," *Language Resources and Evaluation*, vol. 43, no. 4, p. 329, 2009.
28. L. Devillers, H. Maynard, S. Rosset, P. Paroubek, K. McTait, D. Mostefa, K. Choukri, L. Charnay, C. Bousquet, N. Vigouroux *et al.*, "The french media/evalda project: the evaluation of the understanding capability of spoken language dialogue systems." in *LREC*. Citeseer, 2004.
29. A. Rousseau, G. Boulianne, P. Deléglise, Y. Estève, V. Gupta, and S. Meignier, "LIUM and CRIM ASR system combination for the REPERE evaluation campaign," in *International Conference on Text, Speech, and Dialogue*. Springer, 2014, pp. 441–448.
30. C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding." in *Interspeech*, 2007, pp. 1605–1608.
31. C. Servan, C. Raymond, F. Béchet, and P. Nocéra, "Conceptual decoding from word lattices: application to the spoken dialogue corpus media," in *The Ninth International Conference on Spoken Language Processing (Interspeech 2006-ICSLP)*, 2006.