Evaluation of acoustic word embeddings

Sahar Ghannay, Yannick Estève, Nathalie Camelin, Paul deléglise
LIUM - University of Le Mans, France
firstname.lastname@univ-lemans.fr

Abstract

Recently, researchers in speech recognition have started to reconsider using whole words as the basic modeling unit, instead of phonetic units. These systems rely on a function that embeds an arbitrary or fixed dimensional speech segments to a vector in a fixed-dimensional space, named acoustic word embedding. Thus, speech segments of words that sound similarly will be projected in a close area in a continuous space. This paper focuses on the evaluation of acoustic word embeddings. We propose two approaches to evaluate the intrinsic performances of acoustic word embeddings in comparison 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.

1 Introduction

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 acoustic word embedding, which are projection of arbitrary or fixed dimensional speech segments into a continuous space, in a manner that preserve acoustic similarity between words. Thus, speech segments of words that sound similarly will have similar embeddings. Acoustic word embedding were successfully used in a query-by-example search system (Kamper et al., 2015; Levin et al., 2013) and in a ASR lattice re-scoring system (Bengio and Heigold, 2014).

The authors in (Bengio and Heigold, 2014) proposed an approach to build acoustic word embeddings from an orthographic representation of the word. This paper focuses on the evaluation of these acoustic word embeddings. We propose two approaches to evaluate the intrinsic performances of acoustic word embeddings in comparison to orthographic representations. In particular we want to evaluate whether they capture discriminative information about their pronunciation, approximated by their phonetic representation. In our experiments, we focus on French language whose particularity is to be rich of homophone words. This aspect is also studied in this work.

2 Acoustic word embeddings

2.1 Building acoustic word embeddings

The approach we used to build acoustic word embeddings is inspired from the one proposed in (Bengio and Heigold, 2014). The deep neural architecture depicted in figure 1 is used to train the acoustic word embeddings. It relies on a convolutional neural network (CNN) classifier over words and on a deep neural network (DNN) trained by using a triplet ranking loss (Bengio and Heigold, 2014; Wang et al., 2014; Weston et al., 2011).

The two architectures are trained using different inputs: speech signal and orthographic representation of the word, which are detailed as follows.

The convolutional neural network classifier is trained independently to predict a word given a speech signal as input. It is composed of convolution and pooling layers, followed by fully connected layers which feed the final softmax layer. The embedding layer is the fully connected layer just below the softmax one, named **s** in the figure 1. This representation contains a compact representation of the acoustic signal. It tends to preserve acoustic similarity between words, such that words are close in this space if they sound alike.

The feedforward neural network (DNN) is used

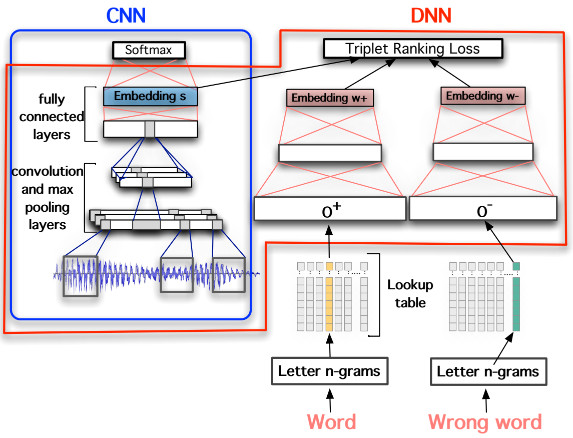

Figure 1: Deep architecture used to train acoustic word embeddings.

with the purpose to build an acoustic word embedding for a word not observed in the audio training corpus, based on its orthographic representation. It is trained using the triplet ranking loss function in order to project orthographic word representations to the same space as the acoustic embeddings **s**.

The orthographic word representation consists in a bag of n-grams ($n \leq 3$) of letters, with additional special symbols *[* and *]* to specify the start and the end of a word. The size of this bag of n-grams vector is reduced using an auto-encoder.

During the training process, this model takes as inputs acoustic embeddings **s** selected randomly from the training set and, for each signal acoustic embedding, 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.

The resulting DNN model can then be used to build an acoustic word embedding ($\mathbf{w}^+$) from any word, as long as one can extract an orthographic representation from it. This acoustic word embedding can be perceived as a canonical acoustic representation for a word, since different pronunciations imply different signal embeddings **s**.

2.2 Evaluation

In the literature (Kamper et al., 2015; Levin et al., 2013; Carlin et al., 2011), a word discrimination task was used to evaluate acoustic embeddings **s**. Given a pair of acoustic segments, this task consists on deciding whether the segments correspond to the same words or not. This evaluation task can be performed on many ways, for example through the use of a dynamic time warping (DTW) to quantify the similarity between two segments when using frame level embeddings (Thiollière et al., 2015), or by using the euclidean distance or the cosine similarity between embeddings representing the segments.

In (Kamper et al., 2015) the evaluation was conducted on two collections of words (train and test) coming from the Switchboard English corpus. After training the model on the training corpus, the cosine similarity is computed between the embeddings of each pair of words in the test set. These pairs 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 two approaches to evaluate acoustic word embeddings $\mathbf{w}^+$. We suggest to build different evaluation sets in order to assess the acoustic word embeddings ($\mathbf{w}^+$) performances on *orthographic* and *phonetic similarity* and *homophones detection* tasks. We remind that the acoustic word embedding $\mathbf{w}^+$ is a projection of an orthographic word representation $\mathbf{o}^+$ into the space of acoustic signal embeddings **s**. 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 as follows: given a list L of n frequent words (candidate words) in the vocabulary composed of m words, a list of $n \times m$ word pairs was created. Then, two alignments were performed between each word pair based on their orthographic (letters) and phonetic (phonemes) representations, using the sclite[1] tool.

From these alignment two *edition distances* are computed with respect to the alignment results of orthographic and phonetic representations. The Edition distance is computed as follows:

$$SER = \frac{\#In + \#Sub + \#Del}{\#symbols\ in\ the\ reference\ word} \times 100 \quad (1)$$

where SER stands for Symbol Error rate, *symbols* correspond to the letters for orthographic representations, and to the phonemes for phonetic ones, and In, Sub and Del correspond respectively to insertion, substitution and deletion.

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

Next, we compute two *similarity scores* that correspond to the orthographic and phonetic similarity scores sim_score attributed for each pair of words, which are defined as:

$$sim\_score = 10 - \min(10, SER/10) \quad (2)$$

where $\min()$ is a function used to have an edition distance between 0 and 10. Then, for each candidate word in the list L we extract its orthographically and phonetically 10 nearest words. This results in two lists for *orthographic* and *phonetic similarity* tasks. For each candidate word in the list L, the Orthographic list contains its ten closest words in terms of orthographic similarity scores and the Phonetic list contains its ten closest words in terms of phonetic similarity scores. Finally, the Homophones list, used for the *homophone* detection task, contains the homophone words (*i.e.* sharing the same phonetic representation).

Table 1 shows an example of the content of the three lists.

List	**Exampls**
Orthographic	très près 7.5 très ors 5
Phononetic	très frais 6.67 très traînent 6.67
Homophone	très traie très traient

Table 1: Example of the content of the three lists.

In the case of the orthographic and phonetic similarity 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). This approach is used in (Gao et al., 2014; Ji et al., 2015; Levy et al., 2015; Ghannay et al., 2016) to evaluate the linguistic word embeddings on similarity tasks, in which the similarity scores are attributed by human annotators.

For the homophone detection task, the evaluation is performed in terms of precision. For each word w in the Homophones list, let $L_H(w)$ be the list of k homophones of the word w, and $L_{H\_neighbour}(w)$ be the list of k nearest neighbours extracted based on the cosine similarity and $L_{H\_found}(w)$ be the intersection between $L_H(w)$ and $L_{H\_neighbour}(w)$, that corresponds to the list of homophones found of the word w.

The precision P_w of the word w is defined as:

$$P_w = \frac{|L_{H\_found}(w)|}{|L_H(w)|} \quad (3)$$

where $|.|$ refers to the size of a list. We define the overall homophone detection precision on the Homophones list as the average of the P_w:

$$P = \frac{\sum_{i=1}^{N} P_{w_i}}{N} \quad (4)$$

where N is the number of candidate words which have a none-empty Homophones list.

3 Experiments on acoustic word embeddings

3.1 Experimental setup

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 (Galliano et al., 2005), ESTER2 (Galliano et al., 2009) and EPAC (Estève et al., 2010) corpora.

It contains $52k$ unique words that have been 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.

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

3.2 Acoustic word embeddings evaluation

The embeddings we evaluate are built from two different vocabularies: the one used to train the neural network models (CNN and DNN), composed of $52k$ words present in the manual transcriptions of the 488 hours of audio; and another one composed of $160k$ words. The words present in the $52k$ vocabulary are nearly all present in the $160k$ vocabulary.

The evaluation sets described in section 2.2 are generated from these two vocabularies: in the $52k$ vocabulary, all the acoustic word embeddings 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 were computed for words never observed in the audio data.

3.2.1 Quantitative Evaluation

The quantitative evaluation of the acoustic word embeddings $\mathbf{w}^+$ is performed on orthographic similarity, phonetic similarity, and homophones detection tasks. Results are summarized in table 2.

Task	52K Vocab.		160K Vocab.	
	$\mathbf{o}^+$	$\mathbf{w}^+$	$\mathbf{o}^+$	$\mathbf{w}^+$
Orthographic	**54.28**	49.97	**56.95**	51.06
Phonetic	40.40	**43.55**	41.41	**46.88**
Homophone	64.65	**72.28**	52.87	**59.33**

Table 2: Evaluation results of similarity ($\rho \times 100$) and homophone detection tasks (*precision*).

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.

These results show that the projection of the orthographic embeddings $\mathbf{o}^+$ into the acoustic embeddings space $\mathbf{s}$ changes their properties, since 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 $\mathbf{s}$) and a word (represented by $\mathbf{w}^+$), acoustic word embeddings are better than orthographic ones to measure the phonetic proximity between two words.

For the homophone detection task, the Homophones list is computed from the $160k$ vocabulary: that results to 53869 homophone pairs in total. The $52k$ vocabulary contains 13561 homophone pairs which are 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 on the two data sets. This confirms that acoustic word embeddings have captured additional information about word pronunciation than the one carried by orthographic word embeddings. For this task we cannot compare the results between the two vocabularies, since the precision measure is dependent to the number of events. 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.

3.2.2 Qualitative Evaluation

To give more insight into the difference of the quality of the orthographic word embeddings $\mathbf{o}^+$ and the acoustic ones $\mathbf{w}^+$, we propose an empirical comparison by showing the nearest neighbours of a given set of words. Table 3 shows examples of such neighbour. It can be seen that, as expected, neighbour of any given word share the same spelling with it when they are induced by the orthographic embeddings and arguably sound like it when they are induced by the acoustic word ones.

Candidate word	$\mathbf{o}^+$	$\mathbf{w}^+$
grecs	i-grec, rec, marec	grec, grecque, grecques
ail	aile, trail, fail	aille, ailles, aile
arts	parts, charts, encarts	arte, art, ars
blocs	bloch, blocher, bloche	bloc, bloque, bloquent

Table 3: Candidate words and their nearest neighbours

4 Conclusion

In this paper, we have investigated the intrinsic evaluation of acoustic word embeddings. These latter 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. 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 Spearman's rank correlation coefficient (Spearman's ρ), and by measuring the precision in a homophone detection task.

Experiments show that the acoustic word embeddings are better than orthographic ones to measure the phonetic proximity between two words. More, they are better too on homophone detection task. This confirms that acoustic word embeddings have captured additional information about word pronunciation.

Acknowledgments

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.

References

Samy Bengio and Georg Heigold. 2014. Word embeddings for speech recognition. In *INTERSPEECH*, pages 1053–1057.

Michael A Carlin, Samuel Thomas, Aren Jansen, and Hynek Hermansky. 2011. Rapid Evaluation of Speech Representations for Spoken Term Discovery. In *INTERSPEECH*, pages 821–824.

Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas. 2010. The EPAC Corpus: Manual and Automatic Annotations of Conversational Speech in French Broadcast News. In *LREC, Malta, 17-23 may 2010.*

Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, and Guillaume Gravier. 2005. The ESTER phase II evaluation campaign for the rich transcription of French Broadcast News. In *Interspeech*, pages 1149–1152.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Interspeech*, volume 9, pages 2583–2586.

Bin Gao, Jiang Bian, and Tie-Yan Liu. 2014. Wordrep: A benchmark for research on learning word representations. *CoRR*, abs/1407.1640.

Sahar Ghannay, Benoit Favre, Yannick Estève, and Nathalie Camelin. 2016. Word embedding evaluation and combination. In *10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož (Slovenia), 23-28 May.

Shihao Ji, Hyokun Yun, Pinar Yanardag, Shin Matsushima, and S. V. N. Vishwanathan. 2015. Wordrank: Learning word embeddings via robust ranking. *CoRR*, abs/1506.02761.

Herman Kamper, Weiran Wang, and Karen Livescu. 2015. Deep convolutional acoustic word embeddings using word-pair side information. In *arXiv preprint arXiv:1510.01032*.

Keith Levin, Katharine Henry, Anton Jansen, and Karen Livescu. 2013. Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 410–415. IEEE.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Roland Thiolliere, Ewan Dunbar, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. 2015. A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In *Proc. Interspeech*.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, pages 2764–2770.