Evaluation of acoustic word embeddings

Sahar Ghannay, Yannick Estève, Nathalie Camelin, and Paul Deléglise

LIUM-University of Le Mans

Introduction

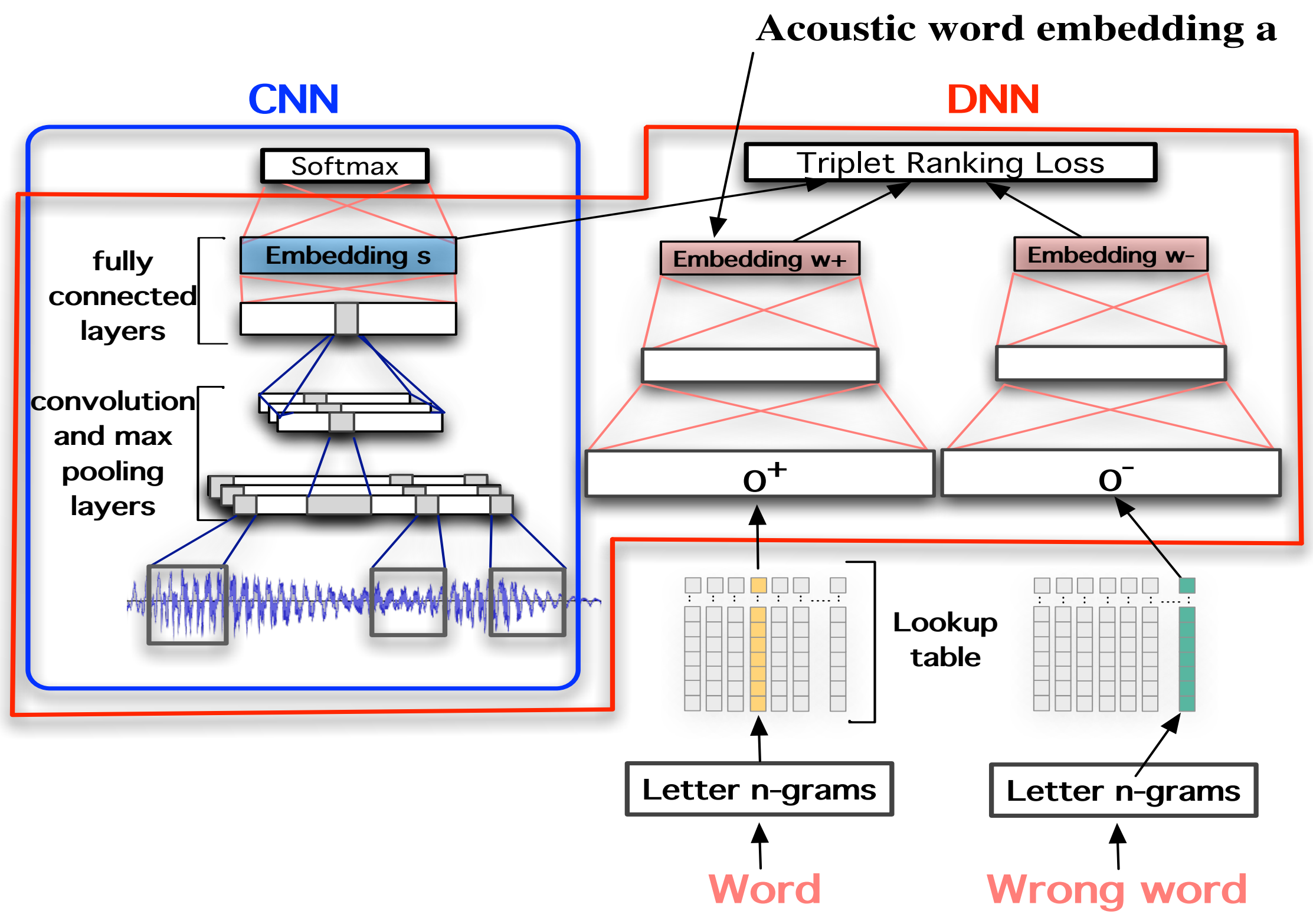
Acoustic embeddings:

f: speech segments → $\mathbb{R}^n$ is a function for mapping speech segments to low-dimensional vectors.

→ words that sound similar = neighbors in the continuous space

Architecture:

Building acoustic word embeddings from an orthographic representation of the word



Goal:

→ Evaluation of acoustic word embeddings (**a**) in comparison to the orthographic embeddings (**o**)

Evaluation of acoustic word embeddings

Objective:

Measure:

- Loss of orthographic information carried by **a**
- Gain of acoustic information in comparison to **o**

Benchmark tasks:

- Orthographic and phonetic similarity tasks
- Homophones detection task

Evaluation sets:

Building three evaluation sets:

- Lists of n x m word pairs
  - n: number of frequent words
  - m: number of words in the vocabulary
- Alignment of word pairs
  - Orthographic representation (letters)
  - Phonetic representation (phonemes)
- Edition distance and similarity score:

$$SER = \frac{\#Ins + \#Sub + \#Del}{\#symbols\ in\ the\ reference\ word} \times 100$$

$$Similarity\_score = 10 - \min(10, SER/10)$$

Example of the three lists content:

List	Examples
Orthographic	très [tʁɛ] près [pʁɛ] 7.5 très [tʁɛ] tris [tʁi] 7.5
Phonetic	très [tʁɛ] frais [fʁɛ] 6.67 très [tʁɛ] traînent [tʁɛn] 6.67
Homophone	très [tʁɛ] traie [tʁɛ] très [tʁɛ] traient [tʁɛ]

Experiments

Setup:

Acoustic word embeddings:

Data: 488 hours of French Broadcast news
Vocabulary size: 52k

Evaluation sets:

Data:
Vocabulary of the audio training corpus: 52k
ASR Vocabulary: 160k
Language: French
Size:
Orthographic: 1000 pairs
Phonetic: 1000 pairs
Homophone: 53869 homophone pairs for 160k vocab.
13651 homophone pairs for 52k vocab.

Evaluation metrics:

Similarity tasks:

- Spearman's rank correlation coefficient

Homophone detection task:

- Precision of the word

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

- Overall precision

$$P = \frac{\sum_{i=1}^{N} P_{w_i}}{N}$$

Results:

Quantitative evaluation:

Performed on orthographic similarity, phonetic similarity and homophones detection tasks:

Tasks	52K Vocab.		160K Vocab.	
	o^+	w^+	o^+	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**

Qualitative evaluation:

Empirical comparison between **a** and **o** by showing the nearest neighbors of a given word :

Candidate word	Orthographic word embedding o	Acoustic word embedding a
grecs [gʁɛk]	i-grec [igʁɛk], rec [ʁɛk], mare [maʁ]	grec [gʁɛk], grecque [gʁɛk], grecques [gʁɛk]
ail [aj]	aile [ɛl], trail [tʁaj], fail [faj]	aille [aj], ailles [aj], aile [ɛl]
arts [aʁ]	parts [paʁ], charts [ʃaʁ], encarts [ɑ̃kaʁ]	arte [aʁte], art [aʁ], ars [aʁ]
blocs [blɔk]	bloch [blɔk], blocher [bloʃɛʁ], bloche [blɔʃ]	bloc [blɔk], bloque [blɔk], bloquent [blɔk]

Conclusion

✚ 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.

✚ Evaluation of acoustic word embeddings (**a**) in comparison to the orthographic embeddings (**o**) on **orthographic** and **phonetic similarity** tasks and **homophone detection** task.

→ Acoustic word embeddings are better than orthographic ones:

- to measure phonetic proximity between words
- on homophone detection task

✓ Acoustic word embeddings have captured additional information about word pronunciation