# *Acoustic word embeddings for ASR error detection*

Sahar Ghannay, Yannick Estève, Nathalie Camelin and Paul Deléglise

LIUM, IICC, Université du Maine Le Mans, France

INTERSPEECH 2016, SAN FRANCISCO

10/09/2016

# INTRODUCTION

- Why error detection is still relevant ?
  - MGB 2015 challenge results for ASR task on BBC data

| | Best Sys | CRIM/ LIUM | Sys1 | Sys2 | Sys3 | LIUM | Sys4 | Sys5 | Sys6 | Sys7 | Sys8 | Sys9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Overall WER(%) | 23.7 | 26.6 | 27.5 | 27.8 | 28.8 | 30.4 | 30.9 | 31.2 | 35.5 | 38.0 | 38.7 | 40.8 |

- The ASR errors may due to the variability:
  - Acoustic conditions, speaker, language style, *etc.*
- Impact of ASR errors:
  - Information retrieval,
  - Speech to speech translation,
  - Spoken language understanding,
  - Named entity recognition,
  - *Etc.*

ASR error detection can help

# RELATED WORK (1/2)

## ASR ERROR DETECTION

- Approaches based on Conditional Random Field (CRF):
  - OOV detection [C. Parada et al. 2010]
    - Contextual information
  - Errors detection [F. Béchet & B. Favre 2013]
    - ASR based, lexical and syntactic features
  - Errors detection at word/utterance level [Stoyanchev et al. 2012]
    - Syntactic and prosodic features

- Approach based on neural network:
  - MLP for errors detection [T. Yik-Cheung et al. 2014]
    - Complementary ASR systems, RNNLM, confusion network
  - MLP furnished by a stacked auto-encoders for errors detection [S. Jalalvand *et al.* 2015]
    - Confusion network, textual features
  - MLP-Multi-stream for errors detection and confidence measure calibration [S. Ghannay *et al.* 2015]
    - **Combined word embeddings**, syntactic, lexical, prosodic and ASR-based features

# Related work (2/2)

## 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

- Successfully used in:
  - Query-by-example search system [kamper *et al*, 2015, levin *et al*, 2013]
  - ASR lattice re-scoring system [Bengio and Heiglod *et al*, 2014]

# Contributions

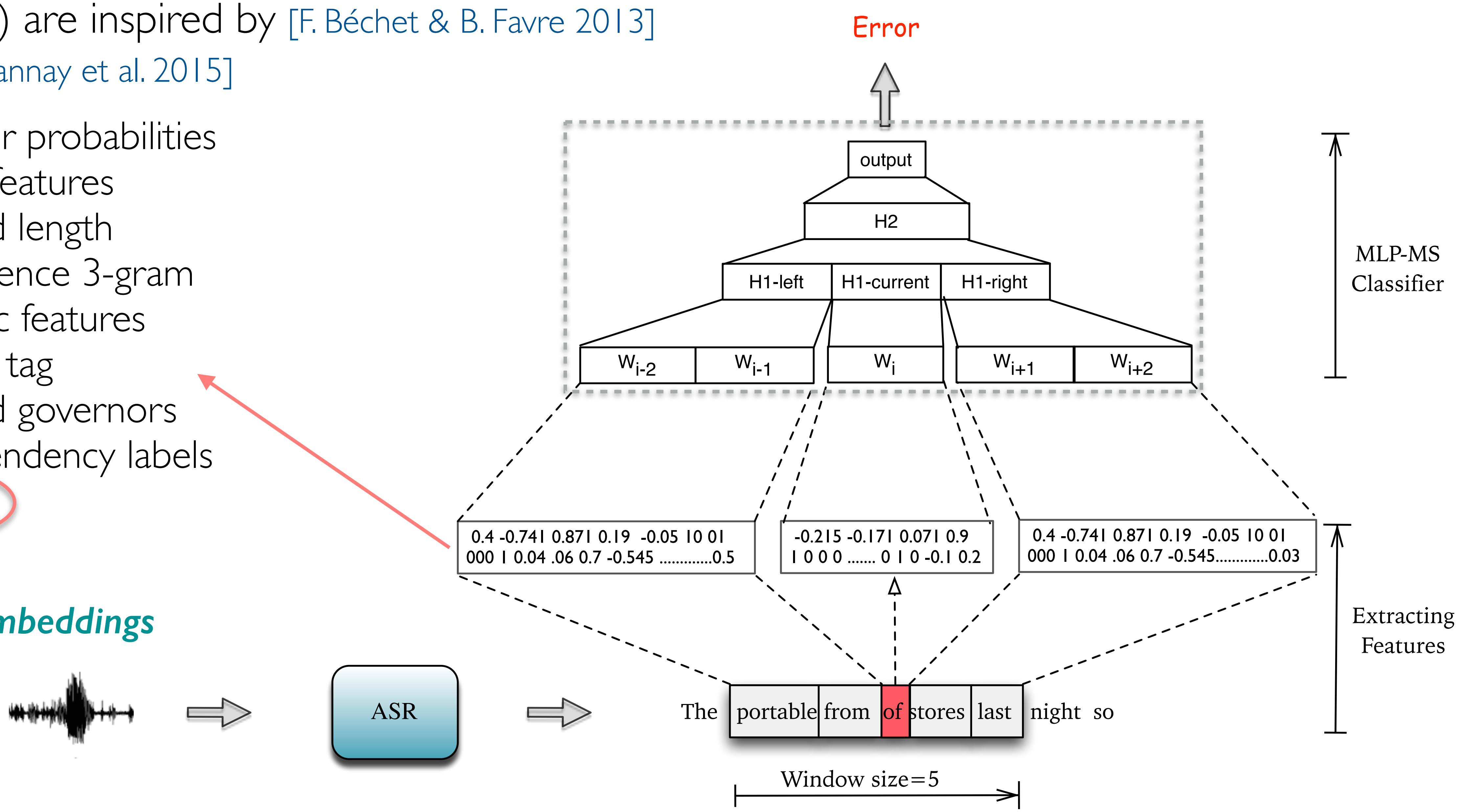
- Building acoustic word embeddings
- Evaluation of their impact on ASR errors detection
- Comparison of their performance to orthographic embeddings
  - Evaluate whether they capture discriminative phonetic information

# ASR ERROR DETECTION SYSTEM

Features (B-Feat.) are inspired by [F. Béchet & B. Favre 2013] and used in [S.Ghannay et al. 2015]

- Posterior probabilities
- Lexical features
  - word length
  - existence 3-gram
- Syntactic features
  - POS tag
  - word governors
  - dependency labels
- Word

***Combined word embeddings***



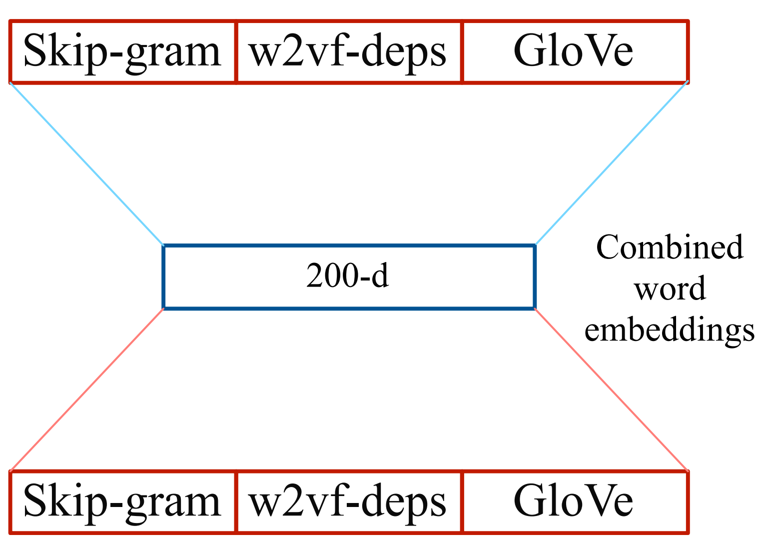
# COMBINED WORD EMBEDDINGS

Evaluation and combination of word embeddings
[S.Ghannay *et al.* SLSP 2015, LREC 2016]

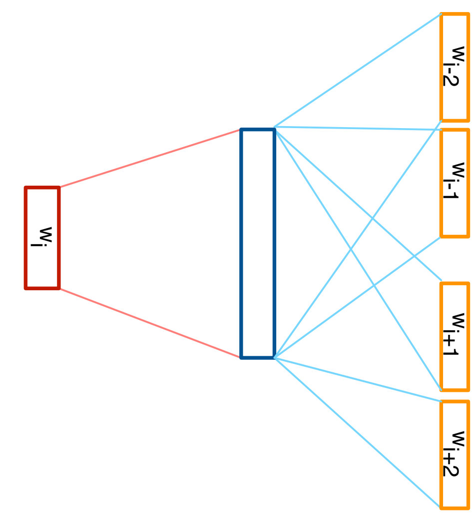
- ASR error detection
- NLP tasks
- Analogical and similarity tasks

➡ Combination of word embeddings through auto-encoder yields the best results

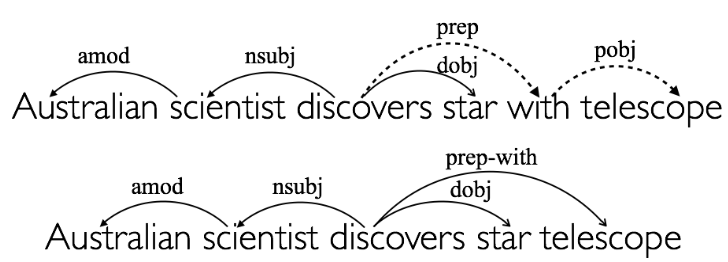
## Auto-encoder




## Skip-gram [T. Mikolov *et al.* 2013]




## w2vf-deps [O. Levy *et al.* 2014]



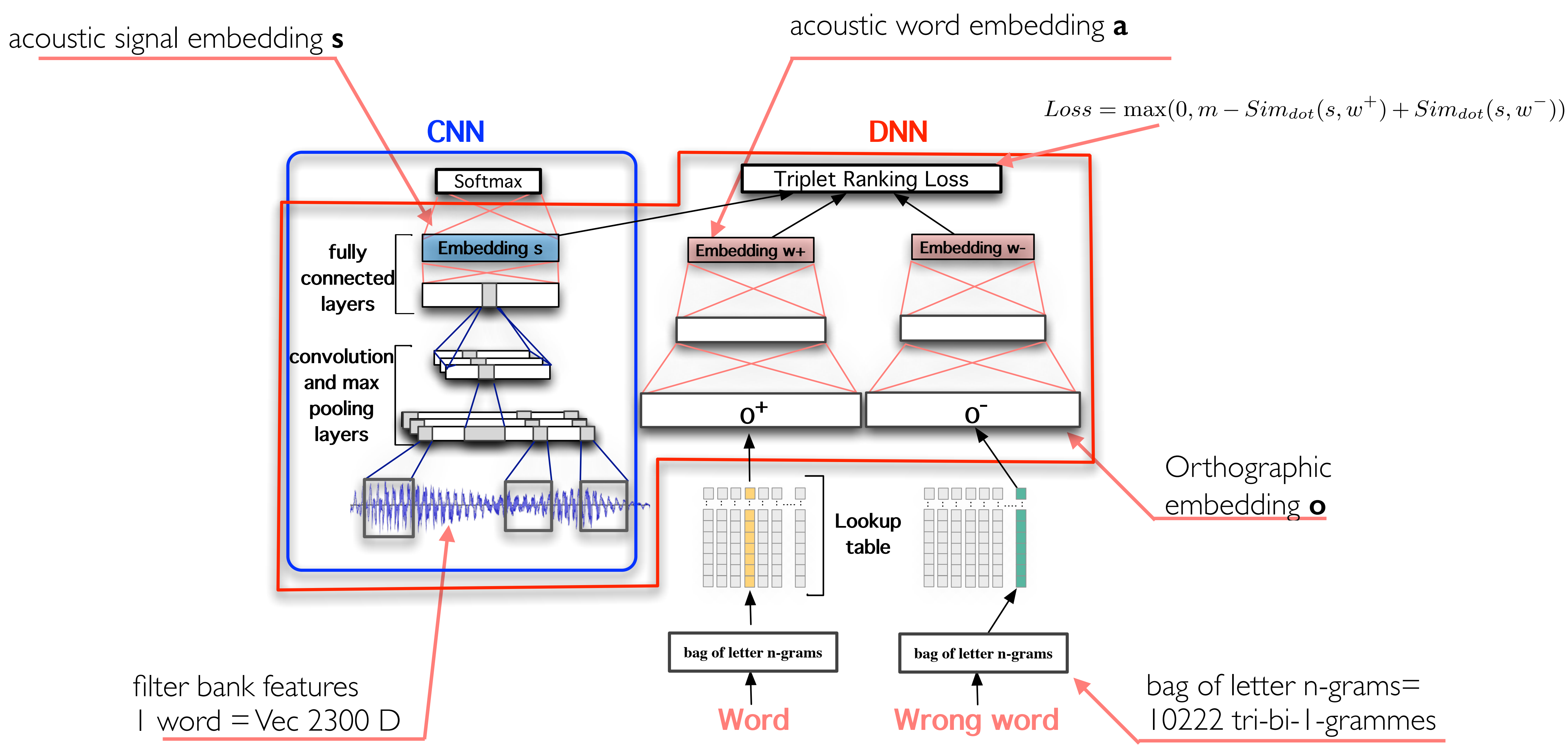

## GloVe [J. Pennington *et al.* 2014]

- building a co-occurrence matrix
- estimating continuous representations of the words

# ACOUSTIC EMBEDDINGS

## ARCHITECTURE

Inspired by [Bengio and Heiglod *et al*, 2014]



# ACOUSTIC EMBEDDINGS

## EVALUATION APPROACHES (1/2)

- Measure:
  - Loss of orthographic information carried by acoustic word embeddings (**a**)
  - Gain of acoustic information in comparison to the orthographic embeddings (**o**)

- Benchmark tasks:
  - Orthographic and phonetic similarity tasks
  - Homophones detection task

# ACOUSTIC EMBEDDINGS
## EVALUATION APPROACHES (2/2)

- 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<br>très [tʁɛ] tris [tʁi] 7.5 |
| Phonetic | très [tʁɛ] frais [fʁɛ] 6.67<br>très [tʁɛ] traînent [tʁɛn] 6.67 |
| Homophone | très [tʁɛ] traie [tʁɛ]<br>très [tʁɛ] traient [tʁɛ] |

# EXPERIMENTAL DATA

## Training data of acoustic word embeddings

- 488 hours of France Broadcast news (ESTER1, ESTER2 et EPAC)
- Vocabulary : 45k words and classes of homophones
- Occurrences : 5.75 millions

## Training of the ASR error detection systems

Automatic transcriptions of the ETAPE Corpus, generated by:

- ASR: CMU Sphinx decoder
  - acoustic models: GMM/HMM

## Training data of the word embeddings

Corpus composed of 2 billions of words:

- Articles of the French newspaper "Le Monde",
- French Gigaword corpus,
- Articles provided by Google News,
- Manual transcriptions: 400 hours of French broadcast news

Description of the experimental corpus

| Name | #words REF | #words HYP | WER |
|---|---|---|---|
| Train | 349K | 316K | 25.3 |
| Dev | 54K | 50K | 24.6 |
| Test | 58K | 53K | 21.9 |

# EVALUATION METRICS

- Similarity task
  - Spearman's Rank correlation coefficient $\rho$
- Homophone detection task
  - Precision $P = \frac{\sum_{i=1}^{N} P_{w_i}}{N}$ , where Pw is the precision of the word $P_w = \frac{|L_{H\_found}(w)|}{|L_H(w)|}$
- Error detection task

➡ Neural architecture vs. CRF [F. Béchet & B. Favre 2013]

  - Error label: Precision (P), Recall (R), and F-measure (F)
  - Overall classification: CER (Classification error rate)

# Acoustic word embeddings evaluation

## Evaluation sets

- Data:
  - Vocabulary of the audio training corpus 52k
  - ASR vocabulary 160k
- Language:
  - French

## Evaluation results

| Tasks | Metrics | 52k Vocab. | | 160K Vocab. | |
|---|---|---|---|---|---|
| | | o | a | o | a |
| Orthographic | $\rho$ | **54.28** | 49.97 | **56.95** | 51.06 |
| Phonetic | $\rho$ | 40.40 | **43.55** | 41.41 | **46.88** |
| Homophone | P | 64.65 | **72.28** | 52.87 | **59.33** |

# ASR ERROR DETECTION TASK

## Performance of acoustic word embeddings

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

# CONCLUSION

- Evaluation of acoustic word embeddings **a** in comparison to the orthographic **o** ones on:
  - Orthographic and phonetic similarity tasks
  - Homophones detection task
    - **a** are better than **o**
      - to measure phonetic proximity between words
      - on homophone detection task
    - **a** have captured additional information about word pronunciation
- Evaluation of their impact on ASR error detection task
  - Neural approach using the acoustic word embeddings
    - significant improvement by 7.24% in terms of CER relative to CRF on Test.

Thank you !