Continuous word representation and prosodic features for ASR error detection

Sahar Ghannay, Yannick Estève, Nathalie Camelin,
Camille Dutrey, Fabián Santiago, and Martine Adda-Decker

LIUM, IICC, University of Le Mans France
LPP - Université Sorbonne Nouvelle, Paris, France

SLSP 2015, Statistical Language and Speech Processing, Budapest, Hungary

24/11/2015

Introduction

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

Introduction

MGB 2015 challenge result

Detailed performance of the best system

Show	CU
Daily Politics	10.4
Magnetic North	11.6
Dragons'Den	11.5
Eggheads	14.1
Athletics London	14.7
Point of View	13.5
Syd Barrett	21.3
Top Gear	21.8
Blue Peter	24.6
Legend of the Dragon	21.7
The North West 200	27.7
Holby City	32.1
The Wall	33.7
One Life Special Mum	35.3
Goodness Gracious ME	37.2
Oliver Twist	**41.4**
Overall WER(%)	***23.7***

Introduction

ASR errors have impact on downstream applications:

- ✤ Information retrieval
- ✤ Speech to speech translation
- ✤ Spoken language understanding
- ✤ etc.

⇨ ASR error detection can help

Introduction

✓ Related work

- Approaches based on Conditional Random Field (CRF)
  - OOV detection [C. Parada *et al.* 2010]
    - Contextual informations
  - Errors detection [F. Béchet & B. Favre 2013]
    - ASR based, lexical and syntactic informations
  - Errors detection at word/utterance level [Stoyanchev *et al.* 2012]
    - Syntactic and prosodic features
- Approach based on neural network
  - Errors detection [T. Yik-Cheung *et al.* 2014]
    - Complementary ASR systems

Introduction

✓ Contributions

✤ Neural approach

- Word embeddings combination
- Prosodic features
- Confidence measures produced by the neural system

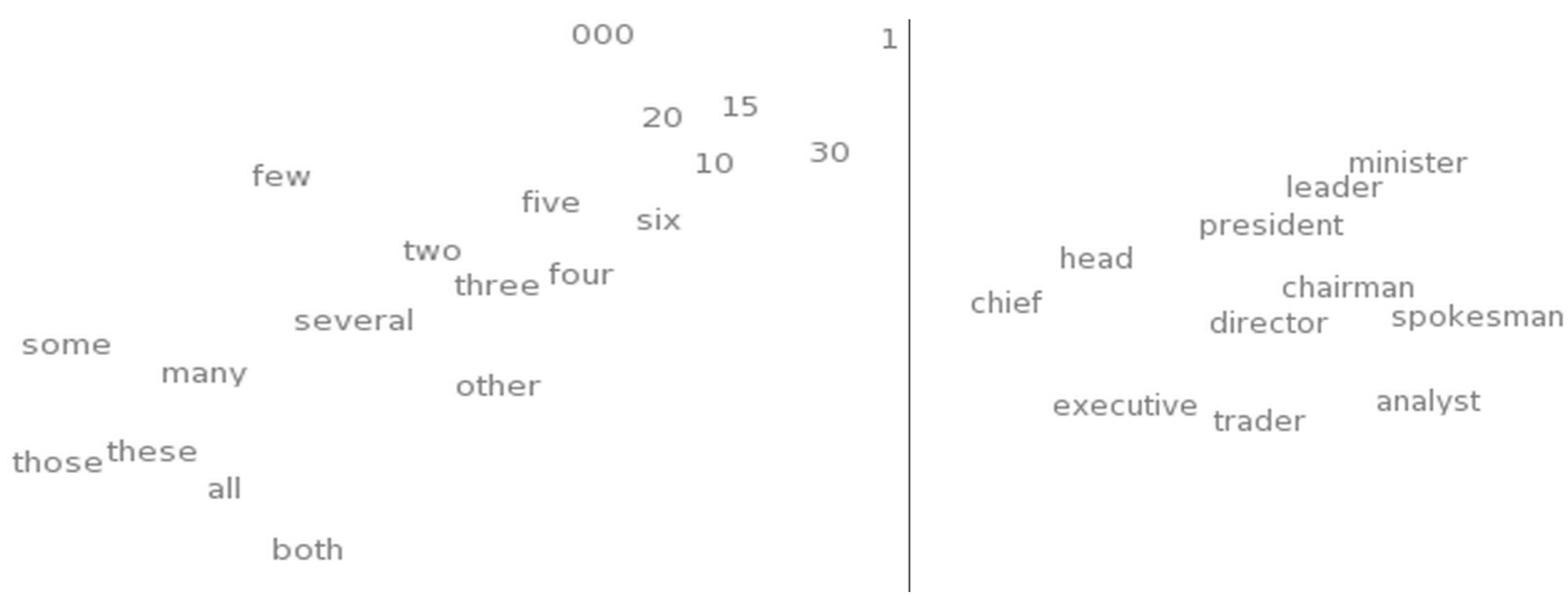
Word embeddings

Mapping words to high-dimensional vectors (e.g. 200 dimensions)

$$R : Words = \{W_1, ..., W_n\} \rightarrow Vectors = \{R(W_1), ..., R(W_n)\} \subset R^d$$

Distance between vectors indicates the relation between words

$$R(W_1) \approx R(W_n) \rightarrow W_1 \approx W_n$$



2D t-SNE visualizations of word embeddings. Left: Number Region; Right: Jobs Region [J.Turian *et al* . 2010]

FRANCE	JESUS	XBOX
AUSTRIA	GOD	AMIGA
BELGIUM	SATI	PLAYSTATION
GERMANY	CHRIST	MSX
ITALY	SATAN	IPOD
GREECE	KALI	SEGA
SWEDEN	INDRA	PSNUMBER
NORWAY	VISHNU	HD
EUROPE	ANANDA	DREAMCAST
HUNGARY	PARVATI	GEFORCE
SWITZERLAND	GRACE	CAPCOM

What words have embeddings closest to a given word? [R.Collobert *et al* . 2011]

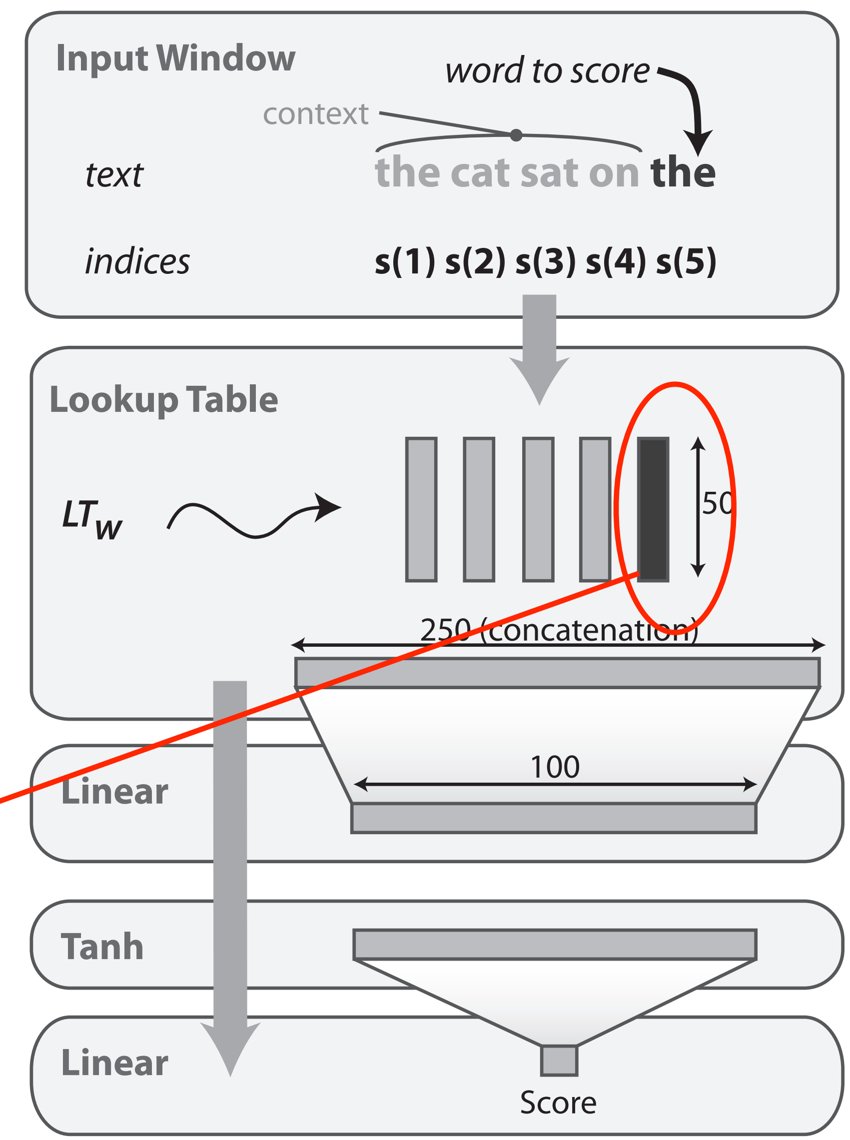
Word embeddings approaches(1/3)

1. Tur: Collobert and Weston embeddings revised by Joseph Turian [J.Turian *et al.* 2010]

- Existence n-gram
- Training criterion: score (n-gram) > score (corrupted n-gram) + some margin

➡ Morpho-syntactic similarities

Word embedding of «**the**»



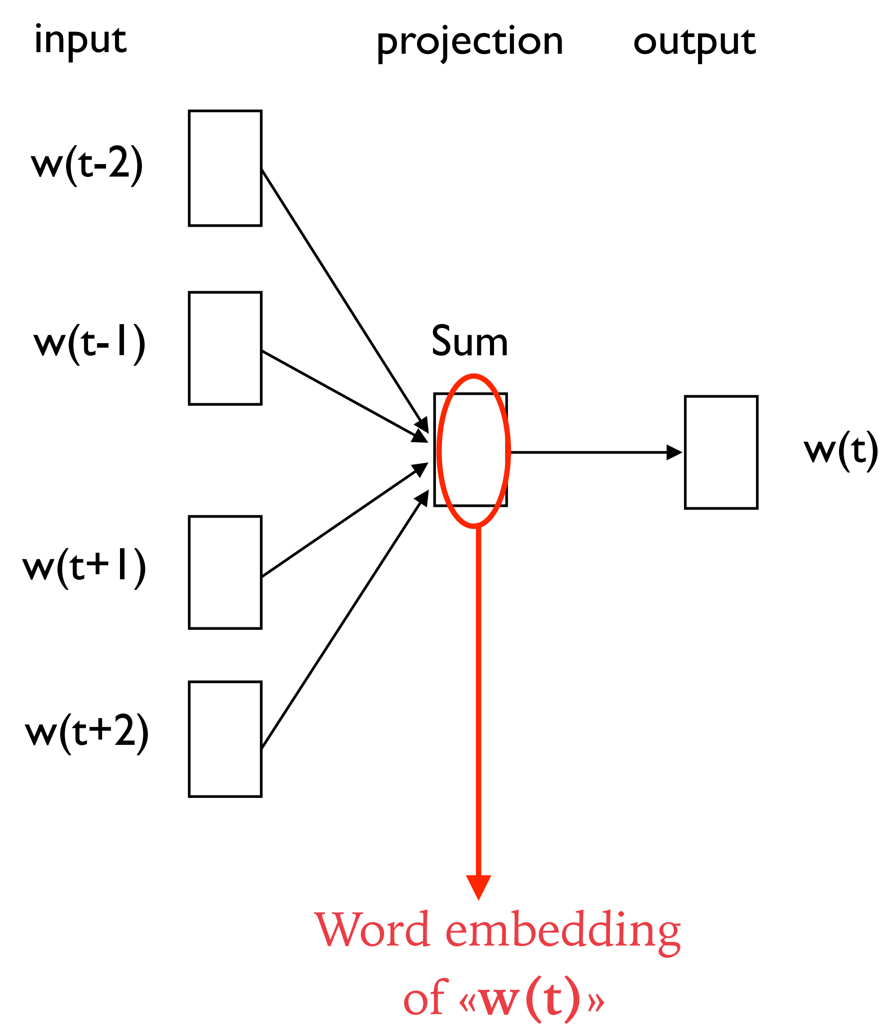
Neural architecture to compute 50 dimensional word embeddings

Word embeddings approaches(2/3)

2. Word2vec [T.Micolov *et al.* 2013]

- Continuous bag of words (CBOW)
  - predicting the current word based on its context

➡ Syntactic modeling



CBOW architecture

Word embeddings approaches(3/3)

3. Glove: global vector for word representation [J.Pennington *et al.* 2014]

- ✤ Analysis of co-occurrences of words in a window
  - building a co-occurrence matrix
  - estimating continuous representations of the words

➡ Semantic similarities

Word embeddings combination (1/3)

1. Simple concatenation (**GTW**)

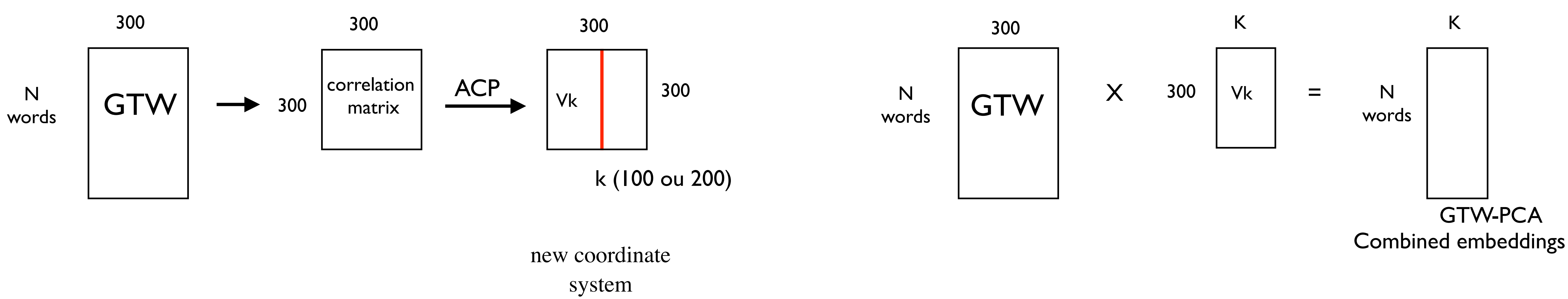
   - concatenation of 100 dimensional word embeddings: glove, tur et w2v
   - word = vector of 300 dimensions

glove	tur	w2v	300 D

Word embeddings combination (2/3)

2. Principal Component Analysis (**PCA**)

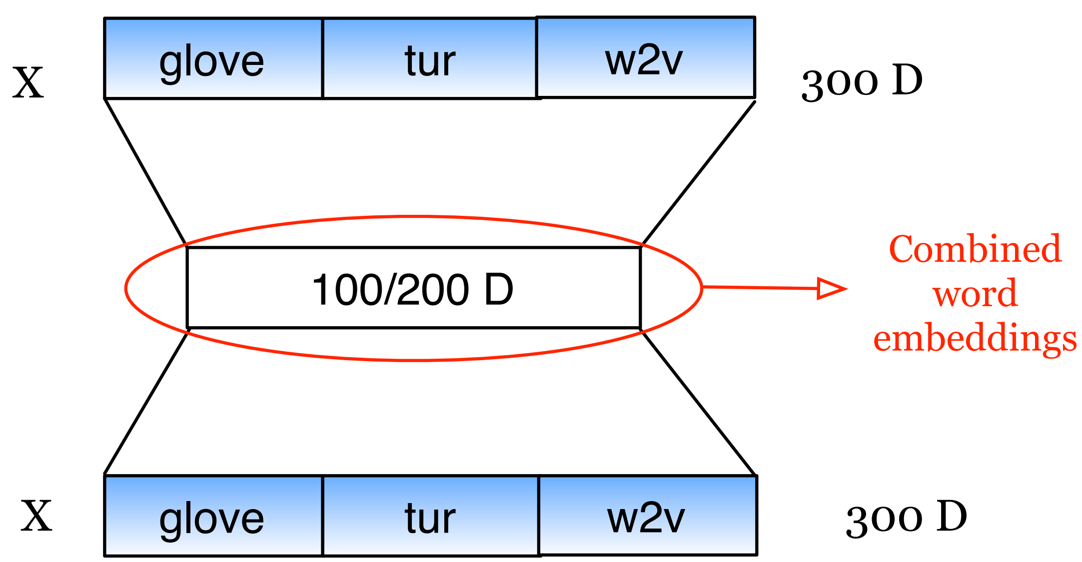
   - Convert correlated variables into uncorrelated variables called principal components.



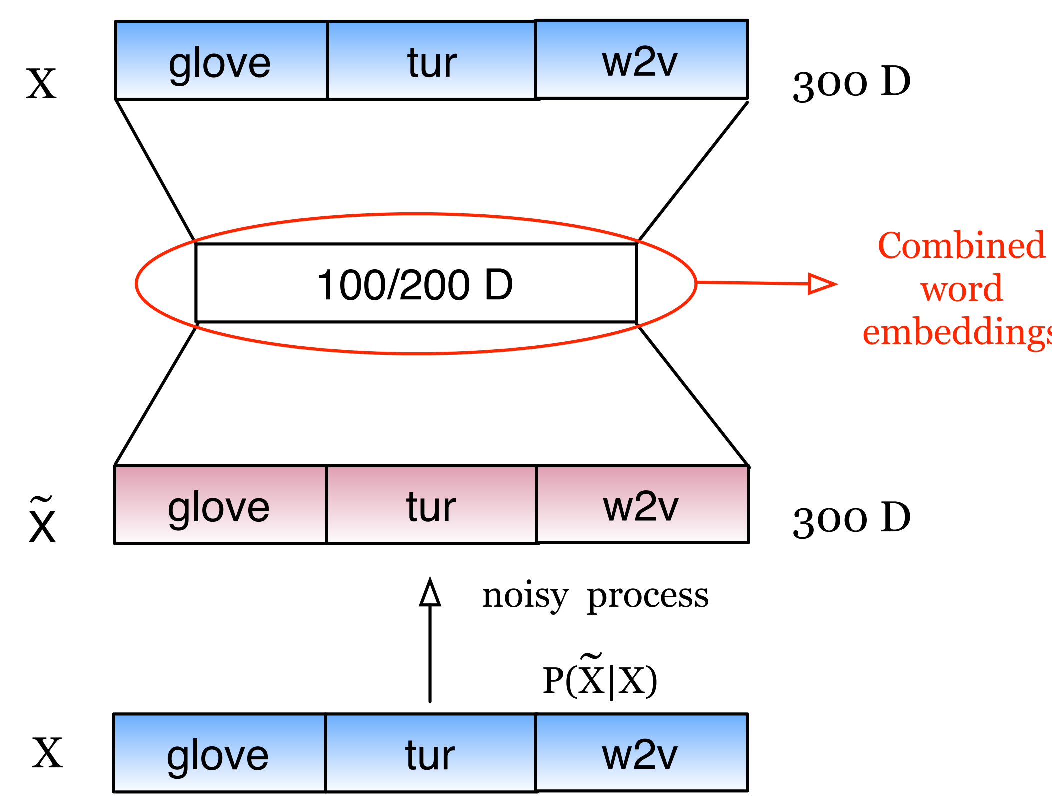
Word embeddings combination (3/3)

3. Auto-encoders

- ✤ Ordinary auto-encoder (**GTW-O**)



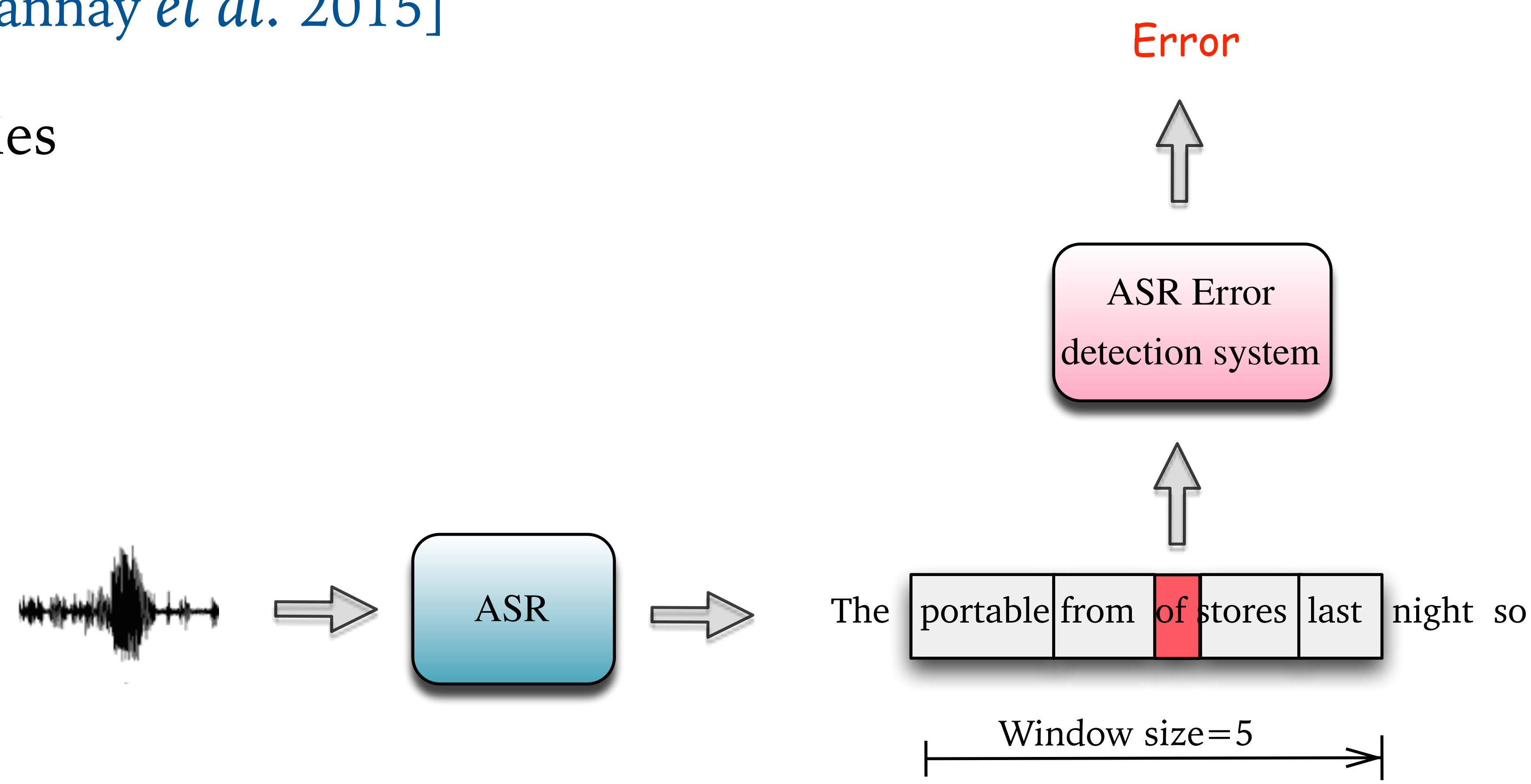
- ✤ Denoising auto-encoder (**GTW-D**)



Set of features

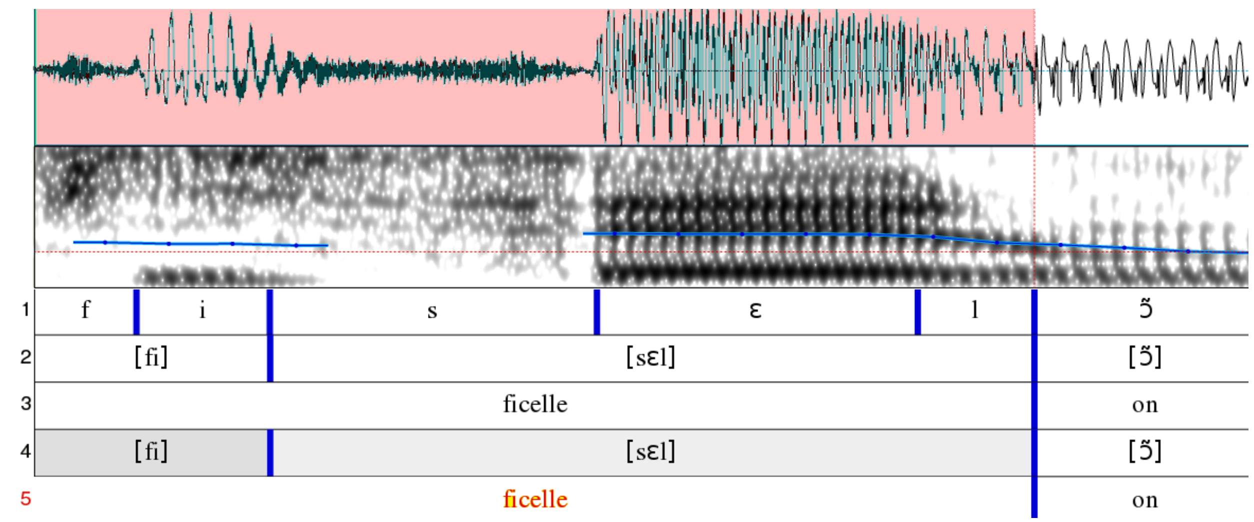
Features used in [S.Ghannay *et al.* 2015]

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

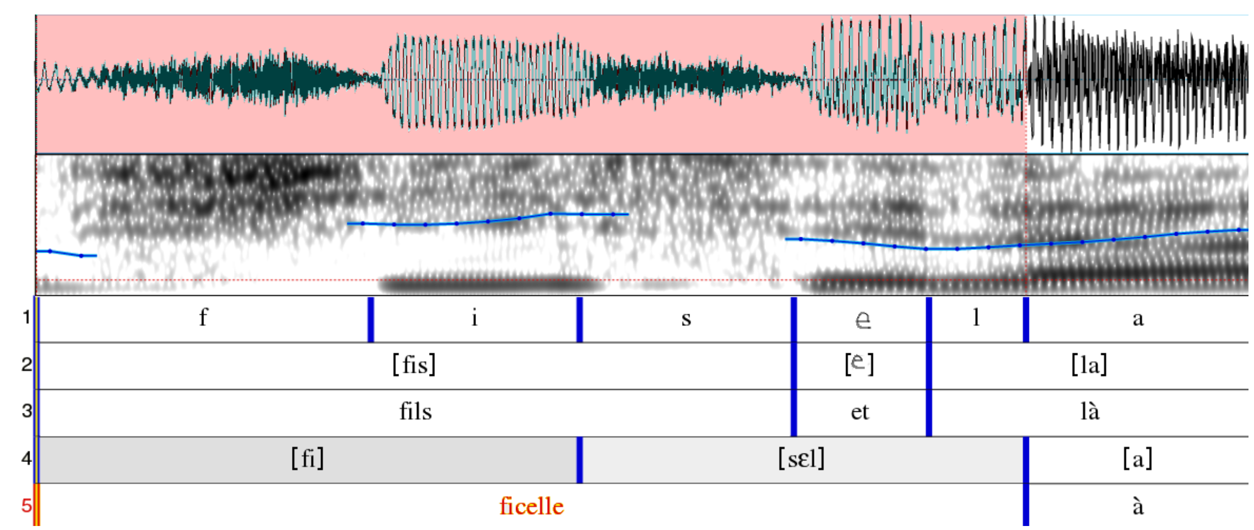


Set of features

- ✤ Prosodic features
  - ✦ number of phonemes
  - ✦ average duration of phonemes
  - ✦ average f0 of the word
  - ✦ f0 delta of the last word
  - ✦ f0 semitone delta last word
  - ✦ duration of the previous pause



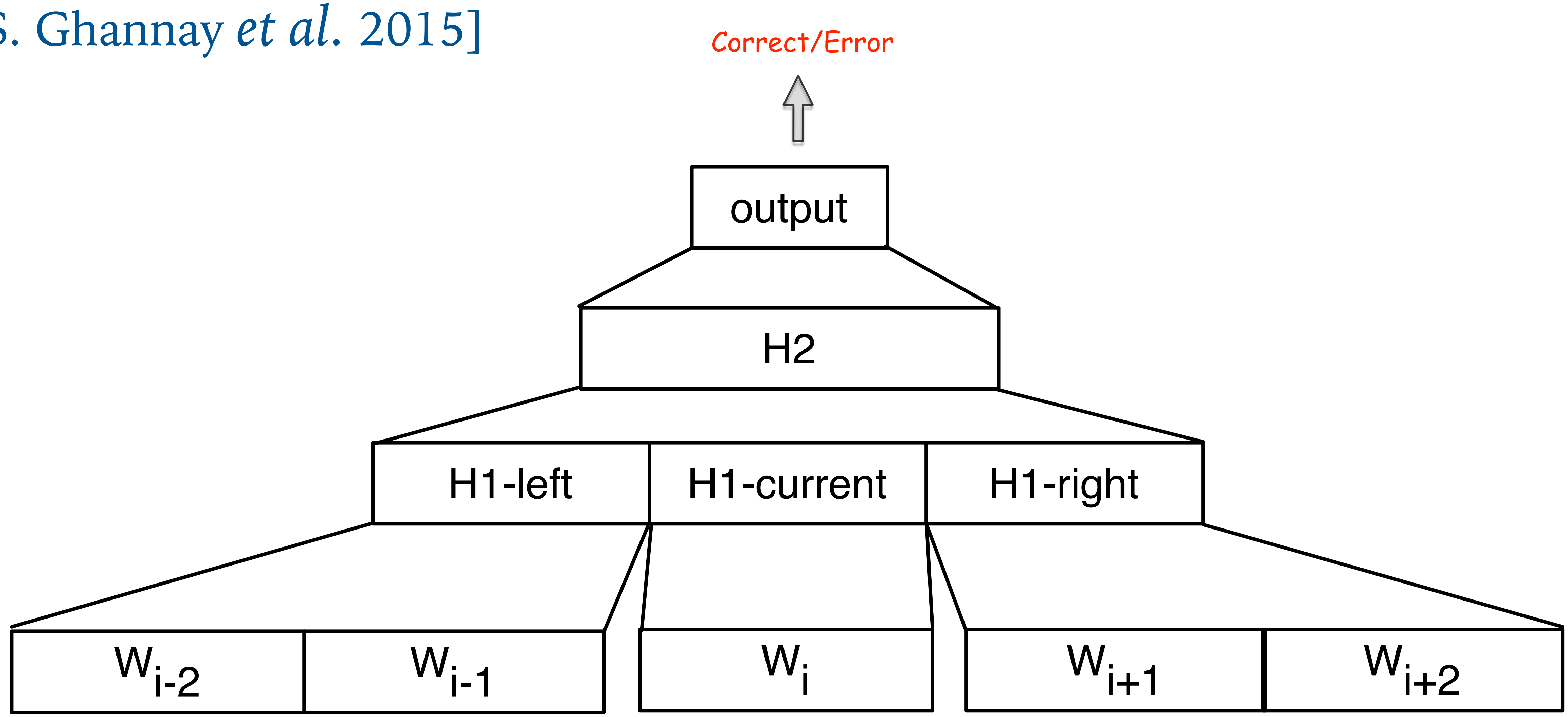
(a)



(b)

Neural architecture: MLP-Multi-Stream

[S. Ghannay *et al.* 2015]



Experimental data

Training of the neural system:

Automatic transcriptions of the ETAPE Corpus [G.Gravier *et al.* 2012], generated by:

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

ASR	Name	#words REF	#words HYP	WER
Sphinx GMM	Train	349K	316K	25.3
	Dev	54K	50K	24.6
	Test	58K	53K	21.9

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.

Evaluation results

- Neural architecture vs. CRF
- Evaluation metrics:
  - Error label: F-measure
  - Overall classification: CER
  - NCE: confidence measures

Experimental results

Comparison of different word embeddings (Dev corpus)

Without prosodic features

		Label error	Global
Neural architecture	Embeddings	F-measure	CER
MLP-MS	Glove	59.64	10.60
	tur	57.58	10.54
	w2v	56.69	10.49
	GTW 300	59.71	10.38
	GTW-PCA100	59.04	10.39
	GTW-PCA200	57.09	10.48
	GTW-O100	56.43	10.28
	GTW-O200	61.86	**9.86**
	GTW-D100	61.63	10.12
	GTW-D200	**63.42**	9.89

Experimental results

Performance of MLP-MS on Test corpus

Without prosodic features

	Label error	Global
Approach	F-measure	CER
CRF(baseline)	57.52	8.79
GTW-O200 GTW-D200	61.83 **62.25**	**8.10** 8.25

Experimental results

Corpus	Approach	Label error	Global
		F-measure	CER
Dev	*CRF(baseline)*	59.48	10.41
	GTW-O200	61.86	9.86
	GTW-D200	63.42	9.89
Test	*CRF(baseline)*	57.52	57.52
	GTW-O200	61.83	8.10
	GTW-D200	62.25	8.25

- prosodic features

Performance of MLP-MS (Dev+Test corpus)

With prosodic features

Corpus	Approach	Label error	Global
		F-measure	CER
Dev	*CRF(baseline)+pros*	60.20	10.29
	GTW-O200+pros	64.80	9.67
	GTW-D200+pros	64.11	9.55
Test	*CRF(baseline)+pros*	**59.17**	**8.57**
	GTW-O200+pros	64.73	7.96
	GTW-D200+pros	64.42	8.03

+ prosodic features

Experimental results

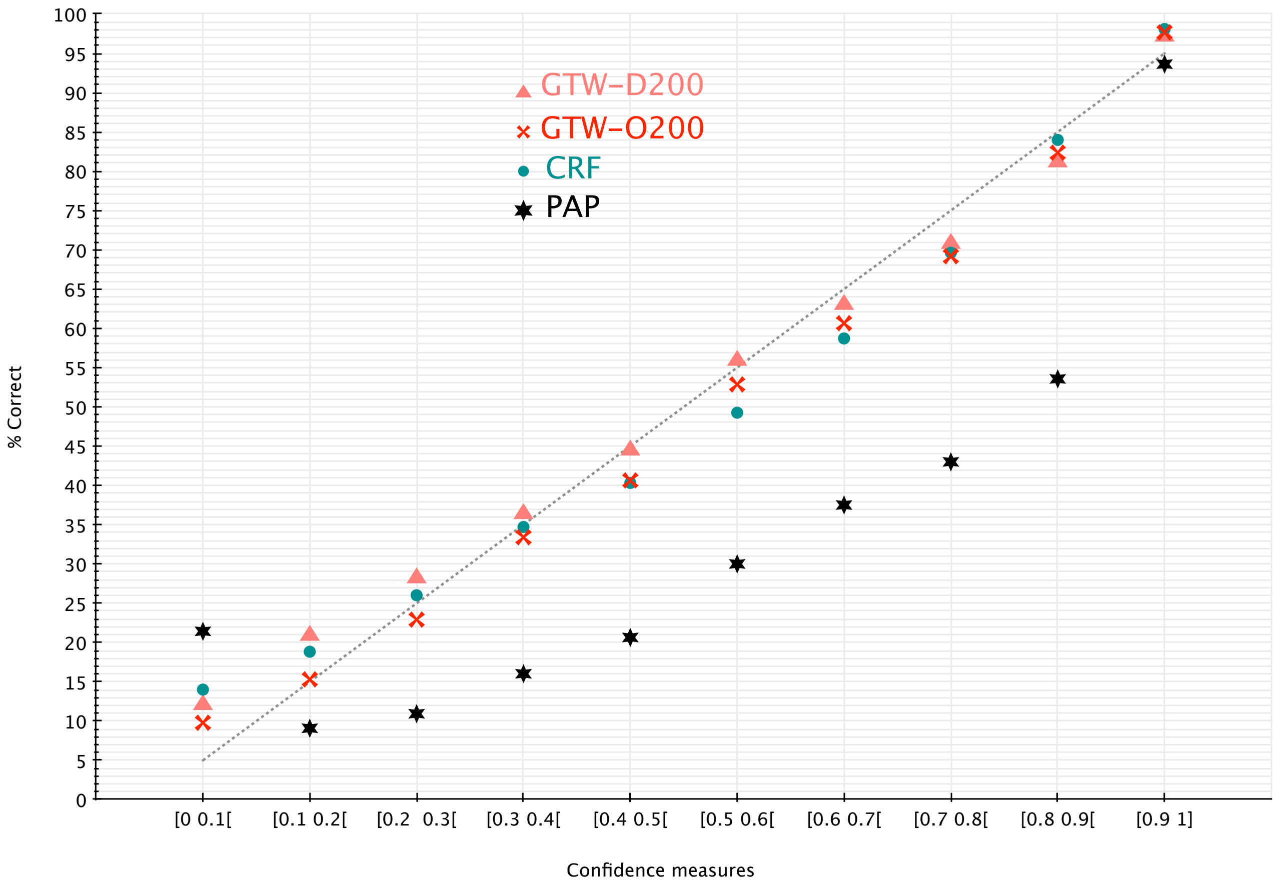
Calibrated confidence measure

Name	PAP	Softmax proba GTW-D200	Softmax proba GTW-O200	CRF
Without prosodic features				
Dev	-0.064	0.0425	0.443	**0.445**
Test	-0.044	0.448	**0.461**	0.457
With prosodic features				
Dev	-0.064	0.461	**0.463**	0.449
Test	-0.044	0.471	**0.477**	0.463

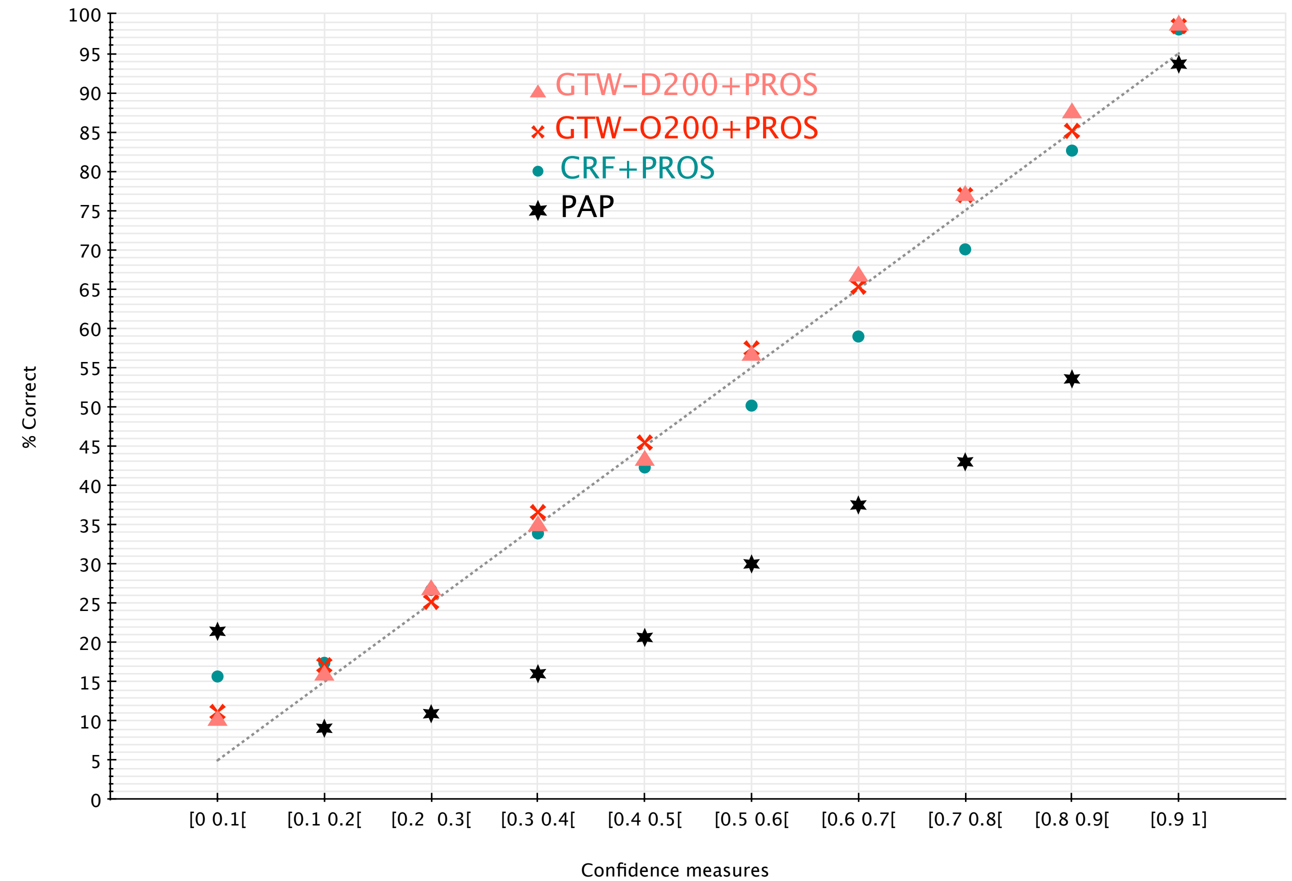
NCE for PAP and the probabilities resulting from MLP-MS and CRF

Experimental results

Calibrated confidence measure



- prosodic features



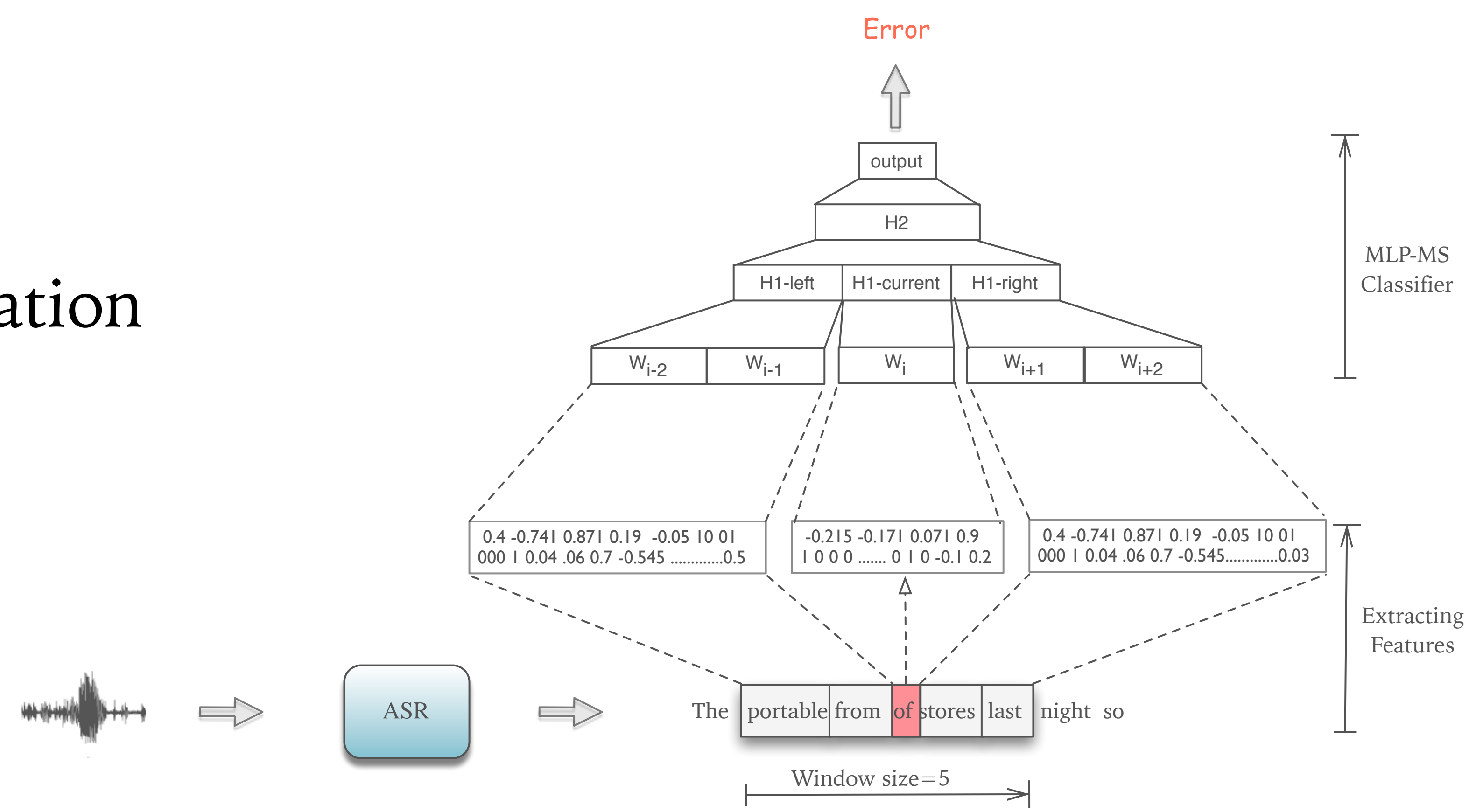
+prosodic features

Pourcentage of correct words based on PAP and confidence measures derived from MLP-MS and CRF

Conclusions

ASR error detection system

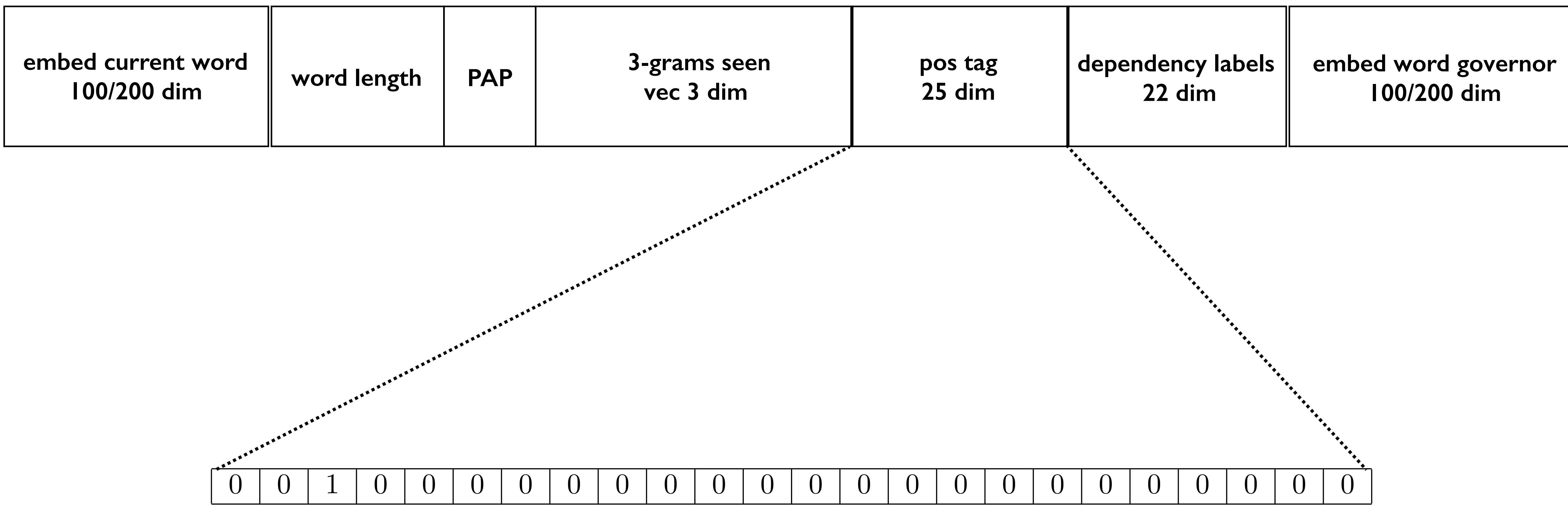
- ✤ Word embeddings combination
- ✤ Prosodic features
- ✤ MLP-MS architecture



- ➡ Outperforms CRF approach
- ➡ Produces calibrated confidence measures

Thank you

Neural network input feature vector format



Example: 25 POS tags, 3rd POS tag

Figure 11 : Neural network input feature vector format (252/452 D)