



**HAL**  
open science

# ASR Performance Prediction on Unseen Broadcast Programs using Convolutional Neural Networks

Zied Elloumi, Laurent Besacier, Olivier Galibert, Benjamin Lecouteux

► **To cite this version:**

Zied Elloumi, Laurent Besacier, Olivier Galibert, Benjamin Lecouteux. ASR Performance Prediction on Unseen Broadcast Programs using Convolutional Neural Networks. Blackbox NLP Workshop and EMLP 2018, 2018, Bruxelles, Belgium. 2018. hal-02102831

**HAL Id: hal-02102831**

**<https://hal.science/hal-02102831v1>**

Submitted on 17 Apr 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

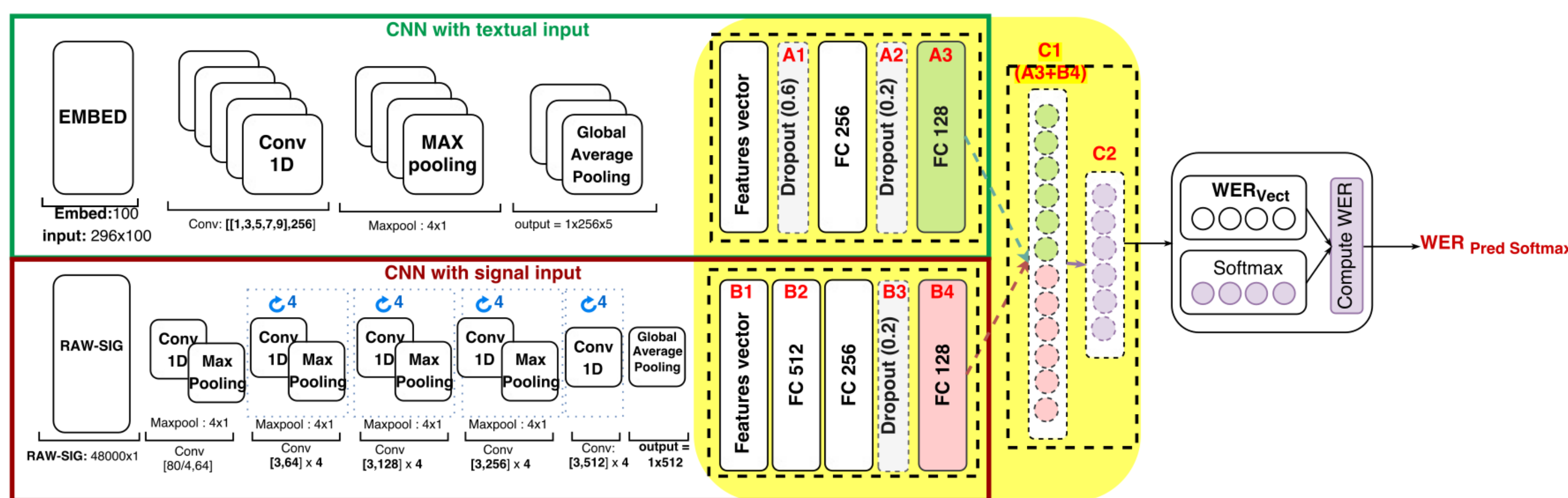


## In short

- ◆ **Task** : prediction of ASR performance on unseen broadcast programs at utterance level
- ◆ **Goal**: understand which information is captured by our deep model (Elloumi et al., 2018) and its relation with different conditioning factors
- ◆ **Main results** : a clear signal is captured about speech style, accent and broadcast type

## Our ASR performance prediction system

- ◆ In (Elloumi et al., 2018), we proposed a new approach using convolution neural networks (CNNs) to predict ASR performance from a collection of heterogeneous broadcast programs (both radio and TV)
- ◆ We particularly focused on the combination of **text** (ASR transcription) and **signal** (raw speech) inputs which **both** proved useful for CNN prediction



- ◆ The network input can be either a pure text input, a pure signal input (raw signal) or a dual (text+speech) input at utterance level
- ◆ Our best approach gave 19.24 % in terms of MAE (Mean absolute error)

## Evaluating learned representations

### Methodology

- ◆ Generate utterance level features (colored in yellow) from our deep model
- ◆ Follow (Belinkov and Glass, 2017) approach to better understand which information is captured by our deep model and its relation with different conditioning factors: Speech style, accent and broadcast program origin

➢ **Classification task**: build three shallow feed-forward neural network classifiers (SHOW, STYLE, ACCENT) with a similar architecture: one hidden layer of 128 units followed by dropout (rate of 0.5), a ReLU non-linearity and a *softmax* layer for mapping onto the label set size

➢ **Visualization task**: t-SNE algorithm to plot hidden representations

### Data

- ◆ Data set from (Elloumi et al., 2018) divided into 3 subsets: TRAIN (67.5K), DEV (7.5K) and TEST (6,7K) → The TEST set contains unseen broadcast programs that are different from those present in TRAIN and DEV

Category	TRAIN	DEV	TEST
Non Spontaneous	54250	6101	<b>3109</b>
Spontaneous	<b>13277</b>	<b>1403</b>	3728
Native	44487	4945	5298
Non Native	<b>23040</b>	<b>2559</b>	<b>1539</b>

Show	TRAIN	DEV	TEST*
FINTER-DEBATE	7632	833	-
FRANCE3-DEBATE	928	77	-
LCP-PileEtFace	<b>4487</b>	525	-
RFI	25565	2831	-
RTM	24198	2745	-
TELSOINNE	4717	<b>493</b>	-
<b>Total</b>	<b>67527</b>	<b>7504</b>	-

- ◆ Extract a balanced version of our TRAIN/DEV/TEST sets by filtering among over-represented labels

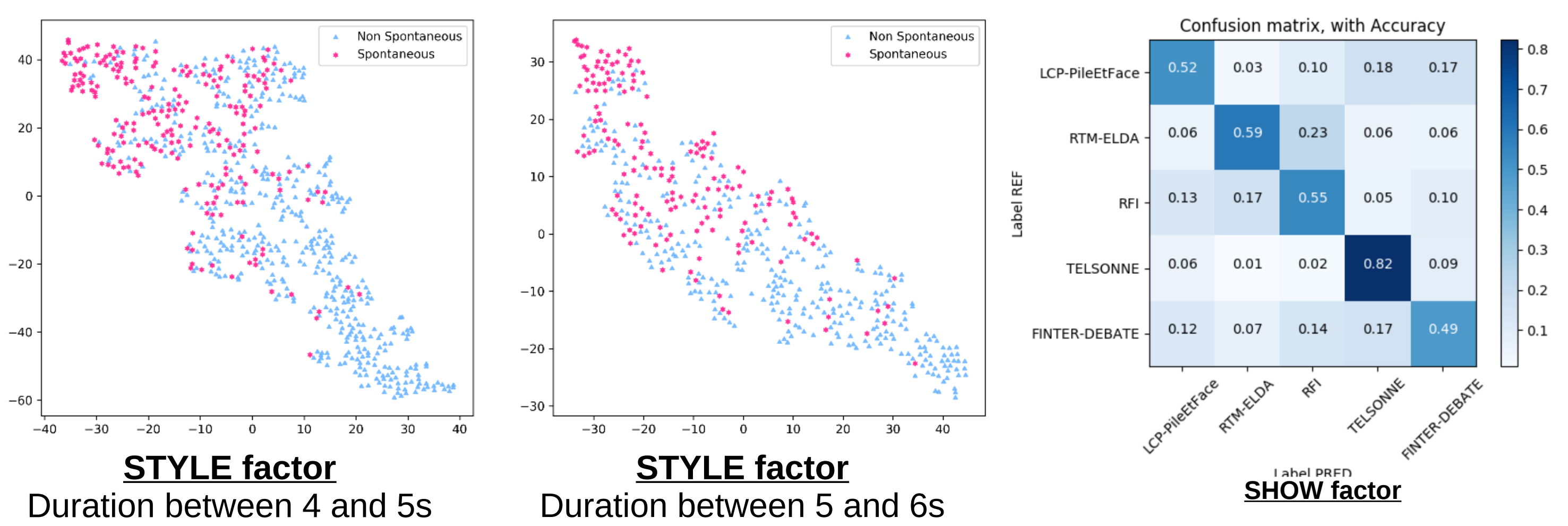
	#Catg	Turns of speech per category		
		TRAIN	DEV	TEST
<b>SHOW</b>	5	4487 <sub>×5</sub>	493 <sub>×5</sub>	-
<b>STYLE</b>	2	13277 <sub>×2</sub>	1403 <sub>×2</sub>	3109 <sub>×2</sub>
<b>ACCENT</b>	2	23040 <sub>×2</sub>	2559 <sub>×2</sub>	1539 <sub>×2</sub>

## Results

### Classification task

Layer	Dim.	SHOW	STYLE	ACCENT
<b>TXT</b>				
A1	1280	<b>57.12</b>	80.72	68.99
A2	256	54.89	80.01	<b>69.56</b>
A3	128	51.04	79.23	68.27
<b>RAW-SIG</b>				
B1	512	<b>42.35</b>	72.92	58.64
B2	512	41.22	72.20	58.41
B3	256	41.22	72.38	58.44
B4	128	40.77	72.38	58.52
<b>TXT + RAW-SIG</b>				
C1 (A3+B4)	256	<b>57.04</b>	81.29	70.36
C2	128	53.06	79.62	<b>70.55</b>
<b>Random</b>	-	<b>20.00</b>	<b>50.00</b>	<b>50.00</b>

### Visualisation task



## Multi-task learning

- ◆ We perform multi-task learning providing the additional information about broadcast type, speech style and speaker's accent during training
- ◆ The architecture of the multi-task model is similar to the single-task WER prediction model but we add additional outputs: a Softmax function is added for each new classification task after the last fully connected layer (C2)

Models	Performance prediction task		Classification tasks		
	MAE	Kendall	SHOW	STYLE	ACCENT
<b>Baseline: Mono-task</b>					
WER (Elloumi et al., 2018)	15.24	19.24	45.00	46.83	-
<b>2-task</b>					
WER SHOW	<b>14.83</b>	19.15	<b>47.25</b>	47.05	<b>99.29</b>
WER STYLE	15.07	19.66	45.92	45.49	-
WER ACCENT	15.05	19.60	46.17	45.60	-
<b>3-task</b>					
WER STYLE ACCENT	15.12	20.23	45.75	44.09	-
WER SHOW ACCENT	14.94	19.76	46.19	43.61	98.38
WER SHOW STYLE	14.90	<b>19.14</b>	45.87	<b>47.28</b>	99.12
<b>4-task</b>					
WER SHOW STYLE ACCENT	15.15	19.64	45.59	45.42	99.04
WER ALL COMBINED OUTPUTS	<b>14.50</b>	<b>18.87</b>	<b>48.16</b>	<b>48.63</b>	-

## Conclusion

- ◆ We proposed an analysis of learned representations of our deep ASR performance prediction system
- ◆ Experiments show that hidden layers convey a clear signal about speech style, accent, and broadcast type
- ◆ We proposed a multi-task learning approach to simultaneously predict WER and classify utterances according to style, accent and broadcast program origin
- ◆ A slight improvements on the test set are observed for MAE and Kendall metrics using multi-task systems