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

Zied Elloumi, Laurent Besacier, Olivier Galibert, Benjamin Lecouteux

# Analyzing Learned Representations of a Deep ASR Performance Prediction Model

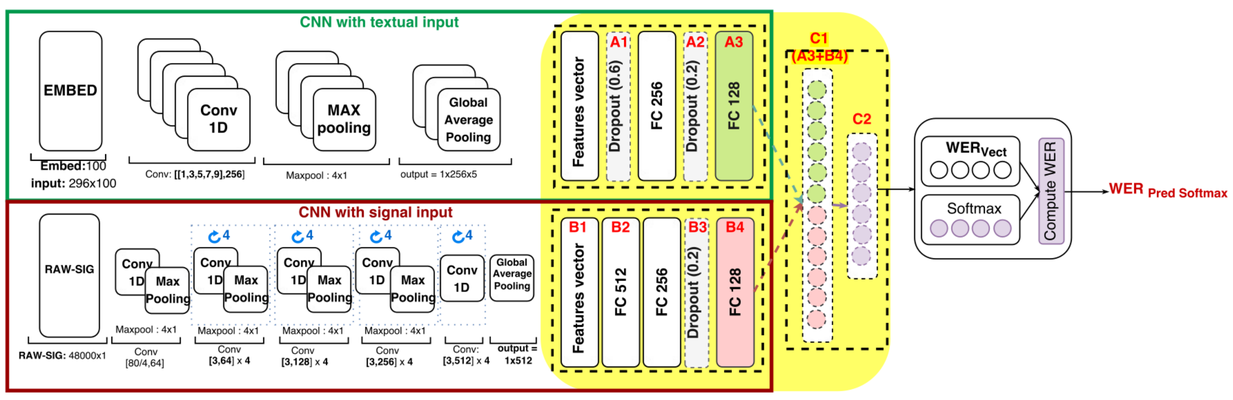
*Zied Elloumi, Laurent Besacier, Olivier Galibert, Benjamin Lecouteux*

## 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 utterence level
- Our best approach gave 19.24 % in terms of *MAE (Mean absolut 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 | **3109** |
| Spontaneous | **13277** | **1403** | 3728 |
| Native | 44487 | 4945 | 5298 |
| Non Native | **23040** | **2559** | **1539** |

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

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

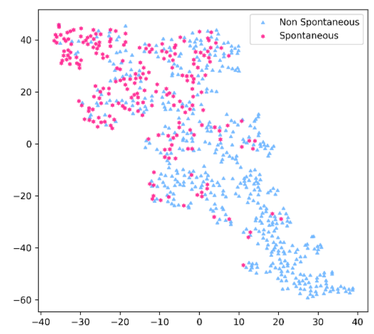
| | #Catg | Turns of speech per category TRAIN | DEV | TEST |
|---|---|---|---|---|
| **SHOW** | 5 | $4487_{\times 5}$ | $493_{\times 5}$ | - |
| **STYLE** | 2 | $13277_{\times 2}$ | $1403_{\times 2}$ | $3109_{\times 2}$ |
| **ACCENT** | 2 | $23040_{\times 2}$ | $2559_{\times 2}$ | $1539_{\times 2}$ |

## Results

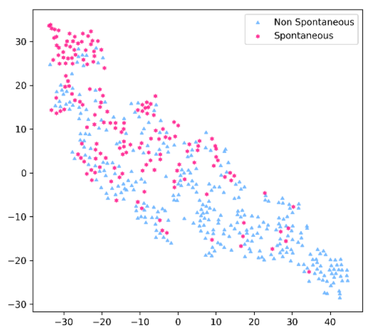
### Classification task

| Layer | Dim. | SHOW | STYLE | ACCENT |
|---|---|---|---|---|
| | | TXT | | |
| A1 | 1280 | **57.12**\|\|- | **80.72**\|\|68.99 | **70.75**\|\|66.54 |
| A2 | 256 | 54.89\|\|- | 80.01\|\|**69.56** | 69.30\|\|69.43 |
| A3 | 128 | 51.04\|\|- | 79.23\|\|68.27 | 68.25\|\|**70.89** |
| | | RAW-SIG | | |
| B1 | 512 | **42.35**\|\|- | **72.92**\|\|**58.64** | 64.60\|\|**55.85** |
| B2 | 512 | 41.22\|\|- | 72.20\|\|58.41 | 64.44\|\|54.84 |
| B3 | 256 | 41.22\|\|- | 72.38\|\|58.44 | 64.50\|\|54.65 |
| B4 | 128 | 40.77\|\|- | 72.38\|\|58.52 | **64.74**\|\|54.87 |
| | | TXT + RAW-SIG | | |
| C1 (A3+B4) | 256 | **57.04**\|\|- | **81.29**\|\|70.36 | **71.41**\|\|**65.98** |
| C2 | 128 | 53.06\|\|- | 79.62\|\|**70.55** | 70.01\|\|65.20 |
| **Random** | - | **20.00** | **50.00** | **50.00** |

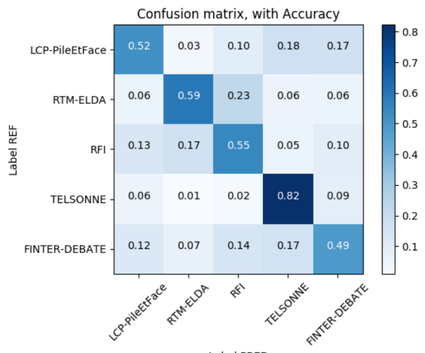
### Visualisation task




**STYLE factor**
Duration between 4 and 5s




**STYLE factor**
Duration between 5 and 6s




**SHOW factor**

## 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 MAE | Kendall | Classification tasks SHOW | STYLE | ACCENT |
|---|---|---|---|---|---|
| | **Baseline: Mono-task** | | | | |
| **WER** (Elloumi et al., 2018) | 15.24\|\|19.24 | 45.00\|\|46.83 | - | - | - |
| | **2-task** | | | | |
| **WER SHOW** | **14.83**\|\|19.15 | **47.25**\|\|47.05 | **99.29**\|\|- | - | - |
| **WER STYLE** | 15.07\|\|19.66 | 45.92\|\|45.49 | - | 99.01\|\|65.24 | - |
| **WER ACCENT** | 15.05\|\|19.60 | 46.17\|\|45.60 | - | - | 91.72\|\| 75.30 |
| | **3-task** | | | | |
| **WER STYLE ACCENT** | 15.12\|\|20.23 | 45.75\|\|44.09 | - | 98.63\|\|69.07 | 88.99\|\| **77.46** |
| **WER SHOW ACCENT** | 14.94\|\|19.76 | 46.19\|\|43.61 | 98.38\|\|- | - | 89.87\|\|71.44 |
| **WER SHOW STYLE** | 14.90\|\|**19.14** | 45.87\|\|**47.28** | 99.12\|\|- | **99.47**\|\|**81.98** | - |
| | **4-task** | | | | |
| **WER SHOW STYLE ACCENT** | 15.15\|\|19.64 | 45.59\|\|45.42 | 99.04\|\|- | 99.29\|\|81.55 | **91.92**\|\|73.60 |
| **WER ALL COMBINED OUTPUTS** | **14.50** \|\|**18.87** | **48.16**\|\|**48.63** | - | - | - |

## 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

**Contact :**
**Zied.elloumi@lne.fr**

**Reference:**

*Elloumi, Z., Besacier, L., Galibert, O., Kahn, J., & Lecouteux, B. (2018). ASR Performance Prediction on Unseen Broadcast Programs using Convolutional Neural Networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary-Canada.*

*Belinkov, Y., & Glass, J. (2017). Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems. In Advances in Neural Information Processing Systems (pp. 2441-2451).*