ASR performance prediction on unseen broadcast programs using convolutional neural networks

Zied Elloumi, Laurent Besacier, Olivier Galibert, Juliette Kahn, Benjamin Lecouteux

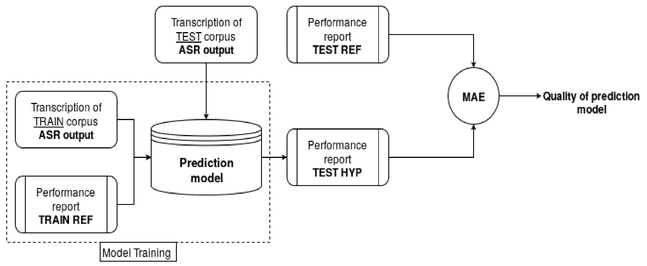
ASR Performance Prediction on Unseen Broadcast Programs using Convolutional Neural Networks

Zied Elloumi, Laurent Besacier, Olivier Galibert, Juliette Kahn, Benjamin Lecouteux

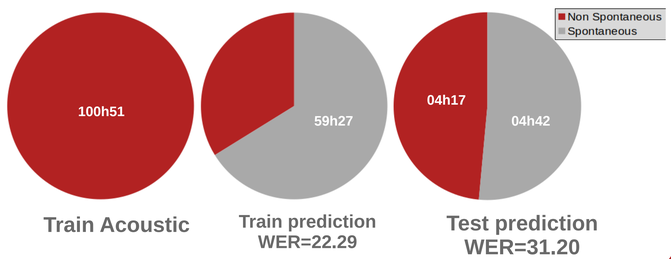
Abstract

- Propose an heterogenous French corpus dedicated to performance prediction task
- Compare two prediction approaches : regression **(engineered features)** based vs a new strategy based on convolutional neural networks **(learnt features)**.
- The joint use of textual and signal features did not work for the regression baseline while the combination of inputs for CNNs leads to the best WER prediction performance.
- CNN prediction predicts the WER distribution on a collection of speech recordings

Evaluation framework

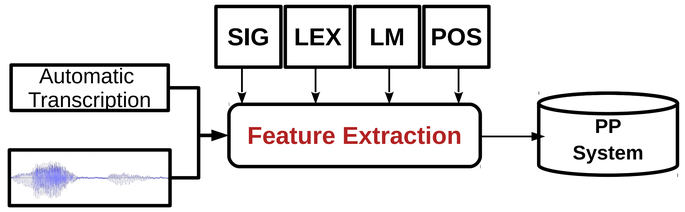


- We built our own French ASR system based on the KALDI toolkit
- We used a French data from different broadcast collections : ESTER, ETAPE, REPERE, Quaero



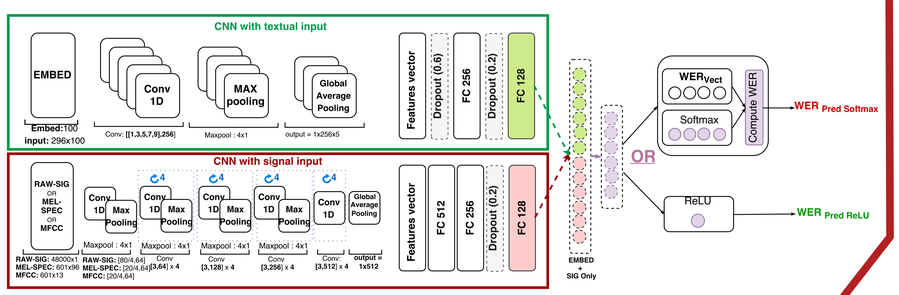
Regression Baseline

- We used an existing tool named TranscRater for baseline regression approach
- We adapted the TranscRater tool from English to French that requires **engineered features** to predict the WER performance



Our proposed approach

- We proposed a new End-to-End performance prediction system based on CNNs.
- Several approaches to encode speech signal
- **Input** : textual, speech signal or the both textual+ speech signal
- **Output** : two proposed methods to predict a continous value ($CNN_{Softmax}$ and CNN_{ReLU})



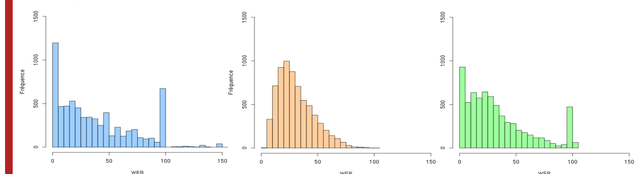
Results & Analysis

Model	Input	MAE	τ
Textual features			
Regression	POS+LEX+LM	22.01	**44.16**
$CNN_{Softmax}$	EMBED	**21.48**	38.91
CNN_{ReLU}	EMBED	22.30	38.13
Signal features			
Regression	SIG	25.86	23.36
$CNN_{Softmax}$	RAW-SIG	25.97	23.61
CNN_{ReLU}	RAW-SIG	26.90	21.26
$CNN_{Softmax}$	MEL-SPEC	29.11	19.76
CNN_{ReLU}	MEL-SPEC	26.07	24.29
$CNN_{Softmax}$	MFCC	**25.52**	**26.63**
CNN_{ReLU}	MFCC	26.17	25.41
Textual and Signal features			
Regression	POS+LEX+LM+SIG	21.99	45.82
$CNN_{Softmax}$	EMBED+RAW-SIG	**19.24**	**46.83**
CNN_{ReLU}	EMBED+RAW-SIG	20.56	45.01
$CNN_{Softmax}$	EMBED+MEL-SPEC	20.93	40.96
CNN_{ReLU}	EMBED+MEL-SPEC	20.93	44.38
$CNN_{Softmax}$	EMBED+MFCC	19.97	44.71
CNN_{ReLU}	EMBED+MFCC	20.32	45.52

- Analysis of predicted WERs

	NS	S	NS + S
WER_{REF}	21.47	38.83	31.20
WER_{Pred} Regression	**22.08**	28.72	25.82
WER_{Pred} $CNN_{Softmax}$	18.93	**33.99**	**27.37**
#Utterances	3,1k	3,7k	6,8k
$\#Words_{REF}$	49.8k	63.3k	113,1k

- Distribution of speech turns according to their WER



(a) real (b) $Regression_{Pred}$ (c) CNN_{Pred}

Conclusion

- We presented an evaluation framework for evaluating ASR performance prediction on unseen broadcast programs
- CNNs were very efficient encoding both textual (ASR transcript) and signal to predict WER.
- **Future work :** Analyze the learnt representations of our performance prediction system

Contact :
Zied.elloumi@lne.fr

2018 IEEE International Conference on Acoustics, Speech and Signal Processing
15–20 April 2018 • Calgary, Alberta, Canada