INVESTIGATING ROBUSTNESS OF A DEEP ASR PERFORMANCE PREDICTION SYSTEM

Zied Elloumi, Olivier Galibert, Benjamin Lecouteux, Laurent Besacier

INVESTIGATING ROBUSTNESS OF A DEEP ASR PERFORMANCE PREDICTION SYSTEM

Zied Elloumi [1] [2], *Olivier Galibert* [1], *Benjamin Lecouteux*[2], *Laurent Besacier*[2]

[1] Laboratoire national de métrologie et d'essais (LNE) , France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France
zied.elloumi@univ-grenoble-alpes.fr, laurent.besacier@univ-grenoble-alpes.fr

ABSTRACT

In this paper, we address a relatively new task: prediction of ASR performance on unseen broadcast programs with ASR system considered like a black-box. In a previous study, we compared two different prediction approaches: a baseline performance prediction based on engineered features and a new strategy based on learnt features using CNNs which combines both textual (ASR-transcription) and signal inputs. In this new contribution, we analyze more deeply the robustness of both ASR prediction approaches (learnt and engineered features) by studying the effect of speech style, training set size and ASR system considered a training or test time. Performance prediction is shown to be more difficult on spontaneous speech. Effect of training size of the predictor is also investigated and it is found that while CNN predictor is better than the baseline predictor, it is also more sensible to training size reduction. Finally, we investigate the robustness of error prediction when the predictor is trained with outputs of a particular ASR system and used to predict performance on unseen broadcast programs and unseen (new) ASR system.

Index Terms— ASR Performance Prediction, Large Vocabulary Continuous Speech Recognition, TV shows, Convolutional Neural Networks

1. INTRODUCTION

Predicting automatic speech recognition (ASR) performance on unseen speech recordings is an important Grail of speech research. From a research point of view, such a task helps understanding automatic (but also human) transcription performance variation and its conditioning factors. From a technical point of view, predicting ASR difficulty is useful in applicative workflows where transcription systems have to be quickly built (or adapted) to new document types (predicting learning curves, estimating amount of adaptation data needed to reach an acceptable performance, etc.).

Related works Other works propose to use more features types than acoustic, [2] exploit ASR, textual, hybrid and acoustic features to predict a WER on different conditions. By exploiting previous works in ASR and machine translation performance prediction tasks [2, 3, 4, 5], [6] proposed an open-source tool named *TranscRater* based on feature extraction (lexical, syntactic, signal and language model features) and regression (WER prediction) or classification (if multiple ASR outputs are provided). Evaluation was performed on CHiME-3 data. For both regression and classification tasks, it was shown that *signal* features did not help WER prediction. Finally, [1] proposed a new ASR performance prediction approach based on CNN. It is based on both textual and raw signal features. Evaluation was performed on a French corpus of TV programs. We give more details on this work and analyze our results more deeply in the next sections.

Contribution Extending our previous work on ASR-performance prediction (PP) task [1], the current work investigates the robustness of PP systems evaluated on unseen broadcast programs. Firstly, we present a large and heterogenous French corpus (containing *non spontaneous* and *spontaneous* speech), an evaluation framework, as well as both *engineered features* and *learnt features* approaches dedicated to performance prediction task. In this study, we focus only on the combination of both textual (ASR transcription) and speech signal, while, ASR system is considered as a black-box. Secondly, we propose a deep analysis in order to evaluate the robustness of ASR-performance prediction systems by studying: i) the effect of speech style on predictor system quality, ii) the influence of training set (for PP) size on ASR performance prediction systems, iii) the robustness of error prediction when the predictor is trained with outputs of a particular ASR system and used to predict performance on shows transcribed with a different ASR system.

Outline The paper is organized as follows. Section 2 details our evaluation framework. Section 3 presents both ASR performance prediction approaches. Section 4 is a deep analysis of the robustness of PP approaches by studying the effect of speech style, training set size and ASR system considered. Finally, section 5 concludes this work.

2. FRAMEWORK FOR ASR-PERFORMANCE PREDICTION

We focus on ASR performance prediction on unseen speech data. Our hypothesis is that performance prediction systems should only use ASR transcripts (and the signal) as input

in order to predict the corresponding transcription quality (WER). Obviously, reference (human) transcriptions are only available at training of the prediction system. A $Train_{pred}$ corpus contains many pairs {*ASR output, Performance*} (more than 75k ASR turns in this work), a $Test_{pred}$ corpus only contains ASR outputs (more than 6.8k turns in this work) and we try to predict the associated transcription performance. Reference (human) transcriptions on $Test_{pred}$ are used to evaluate prediction quality . In order to evaluate *WER* prediction task, we use *Mean Absolute Error* (*MAE*) metric.

Data The data used in our protocol comes from different broadcast collections in French: *Quaero*[1], *ETAPE* [7], *ESTER 1 & ESTER 2* [8] and *REPERE* [9]. As described in Table 1, the full data contains non spontaneous speech (NS) and spontaneous speech (S). The data used to train our ASR system ($Train_{Acoustic}$) is selected from the non-spontaneous speech style that corresponds mainly to broadcast news. The data used for performance prediction ($Train_{pred}$ and $Test_{pred}$) is a mix of both speech styles (S and NS). It is important to mention that shows in $Test_{Pred}$ data set were unseen in the $Train_{Pred}$. Moreover, more challenging (high WERs) shows were selected for $Test_{Pred}$.

	$\text{Train}_{Acoustic}$	Train_{Pred}	Test_{Pred}
NS	100h51	30h27	04h17
S	-	59h25	04h42
Duration	100h51	89h52	08h59

Table 1: Distribution of our data set between non-spontaneous (NS) and spontaneous (S) styles

ASR systems To obtain speech transcripts (ASR outputs) for the prediction model with different qualities, we built our own French ASR systems based on the KALDI toolkit [10]. For the acoustic modelling (AM), we used $\text{Train}_{Acoustic}$ dataset (100 hours of broadcast news from ESTER, REPERE, ETAPE and Quaero) to learn 3 acoustic models (following a standard Kaldi recipe) with 13 dimensions mel-frequency cepstral coefficients (MFCC). These acoustic models are named and trained as following: i) **GMM**: we learnt triphone models with GMM distributions; ii) **SGMM**: we learnt triphone models with SGMM (subspace gaussian mixture models) distributions; iii) **DNN**: we learnt a hybrid HMM/DNN system using DNNs of 4 hidden layers (with 1024 units).

For language modelling (LM), we use both *3-gram* and *5-gram* language models trained on several French corpora[2] using SRILM toolkit [11]. For the pronunciation model, we used lexical resource BDLEX [12] as well as automatic grapheme-to-phoneme (G2P)[3] transcription to find pronunciation variants of our vocabulary (limited to 80k). Finally, the LNE-Tools [13] are used to evaluate the ASR performance in terms of Word Error Rate (WER), knowing that overlapped speech and empty utterances are removed.

ASR systems	AM	LM	Train_{Pred}	Test_{Pred}
ASR1 [1]	**DNN**	**5-gram**	**22.29**	**31.20**
ASR2	DNN	3-gram	23.64	32.80
ASR3	SGMM	3-gram	24.58	34.01
ASR4	GMM	3-gram	27.02	36.79

Table 2: Description of 4 ASR systems produced and their WER performance evaluated on our Train_{Pred} and Test_{Pred} sets

In Table 2, we show 4 different ASR systems learnt to obtain speech transcripts of Train_{Pred} and Test_{Pred} datasets. The results show that ASR systems have different qualities with a higher WER (due to the effect of spontaneous speech) on Test_{Pred}. In addition, we notice that $ASR1$ system generated the best transcription quality while $ASR4$ system performed worse with a difference of +4,73% and +5,59% on Train_{Pred} and Test_{Pred} respectively. In next sections, we use these four ASR systems to obtain all transcripts of Train_{Pred} and/or Test_{Pred}. We note them as Train_i and Test_i sets where $i = 1, 2, 3, 4$ denotes the ASR system used.

3. ASR-PERFORMANCE PREDICTION SYSTEMS

3.1. Engineered features based

An open-source tool for automatic speech recognition quality estimation, *TranscRater* [6], is used for the baseline regression approach (named as TR system in our experiments). It exploits Extremely Randomized Trees algorithm [14] which is a very competitive algorithm in WER prediction and successfully used in [2, 3, 4, 5]. Features selection was performed using Randomized Lasso [15]. *TranscRater* requires *engineered features* to predict the WER performance. These features are extracted for each utterance and are of several types: *Part-of-speech (POS)* features capture the plausibility of the transcription from a syntactic point of view,[4] *Language model (LM)* features capture the plausibility of the transcription according to a N-gram model (fluency),[5] *Lexicon-based (LEX)* features are extracted from the ASR lexicon,[6] *Signal (SIG)* features capture the difficulty of transcribing the input signal (general recording conditions, speaker-specific accents).[7] This approach, based on *engineered features*.One

[1] http://www.quaero.org

[2] 3323M words in total - from EUbookshop, TED2013, Wit3, GlobalVoices, Gigaword, Europarl-v7, MultiUN, OpenSubtitles2016, DGT, News Commentary, News WMT, LeMonde, Trames, Wikipedia and transcriptions of our $Train_{Acoustic}$ dataset

[3] https://goo.gl/NCwpxz

[4] *Treetagger* [16] is used for POS extraction in this study

[5] We train a 5-gram LM on 3323M words text already mentioned

[6] A feature vector containing the frequency of phoneme categories in its prononciation is defined for each input word

[7] For feature extraction, *TranscRater* computes 13 MFCC, their delta, acceleration and log-energy, F0, voicing probability, loudness contours and pitch for each frame. The SIG feature vector for the entire input signal is obtained by averaging the values of each frame

drawback is that its application to new languages requires adequate resources, dictionaries and tools which makes the prediction method less flexible.

3.2. Learnt features based

In [1], 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. We also observed that our system remarkably predicts WER distribution on a collection of speech recordings. The network input can be either a pure text input, a pure signal input (raw signal) or a dual (text+speech) input. To avoid memory issues, signals are downsampled to 8khz and models are trained on six-second speech turns (shorter speech turns are padded with zeros). For text input, the architecture is inspired from [17]: the input is a matrix of dimensions 296x100 (296 is the longest ASR hypothesis length in our corpus ; 100 is the dimension of pre-trained word embeddings on a large held out text corpus of 3.3M words). For speech input, we use the best architecture (*m18*) proposed in [18] of dimensions 48000 x 1 (48000 samples correspond to 6s of speech). For WER prediction, we used ASR1 system (see Table 2) to obtain speech transcripts of our training and evaluation datasets. Our best approach (called CNN$_{Softmax}$) used $softmax$ probabilities and an external fixed WER$_{Vector}$ which corresponds to a discretization of the WER output space (see [1] for more details). The best performance obtained is 19.24% MAE using text+speech input. Our ASR prediction system is built using both *Keras* [19] and *Tensorflow*.[8]

4. DEEP ANALYSIS OF OUR PROPOSED APPROACH

4.1. Effect of speech style on ASR preformance prediction quality

In order to better understand the behavior of the systems for different conditioning factors, we propose in this section to analyze the effect of speech style on PP outputs at broadcast show instance level and at speech style level.

In Figure 1, we compare TR and CNN systems in terms of MAE by calculating the difference between their performances (MAE(TR) - MAE (CNN). If Δ_{MAE} is positive, then CNN is better, else TR is better. The results obtained show that our CNN system is better than the TR system on 80.51% of the shows (95 over 118). In addition, we notice that CNN's prediction is good for both *NS* (green) and *S* (red) speech styles. Notably, for *S* speech, CNN is better than TR on

[8] https://www.tensorflow.org

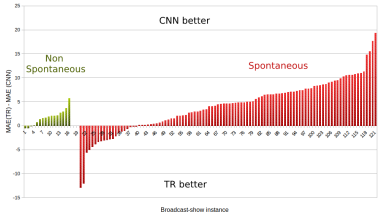

Fig. 1: Evaluation of TR and CNN systems in terms of Δ_{MAE} (CNN is better when $\Delta_{MAE} > 0$) on Test$_1$ (ASR1) dataset at broadcast show instance level and for both *NS* (green) and *S* (red) speech styles

82/102 broadcast show instances by a large margin (50 show instances present a Δ_{MAE} larger than 5%).

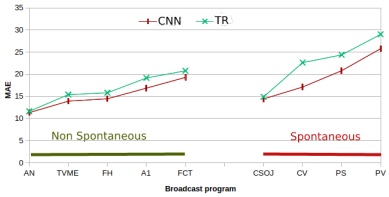

Fig. 2: Evaluation of PP system on Test$_1$ (ASR1) dataset in terms of MAE at broadcast program level

In Figure 2, we compare both CNN and TR systems in terms of MAE on Test$_1$ (ASR1) set at broadcast program level. The performance obtained show that *Spontaneous* (S) is more difficult to predict the performance than *Non Spontaneous* (NS) speech style. In *Spontanous* part, we notice that the gap between CNN and TR curve is wider than for *Non Spontanous* speech. That means that CNN is able to predict a high WER, while TR predicts a performance around the mean WER observed on training data [1]. To confirm this hypothesis, we created an artificial reference by attributing the mean WER observed on training data (22.29%) to all utterances. Evaluating our systems' outputs with this basic reference lead to the following MAE scores: 13.15% and 21.58% on TR and CNN systems respectively, which confirms our intuition.

4.2. Effect of training set size on the quality of ASR performance prediction

Training-set size and its influence on systems' quality remains always an important issue for many tasks (speech recognition, machine translation, image classification, etc). In this section, we attempt to understand what is the effect of training set size on our PP systems (TR and CNN). We build new ASR performance prediction systems with less training data using subsets of Train$_1$ (ASR1). We selected randomly 20% (overall WER of 21.50%) and 50% (overall WER of 22.40%) of the full Train$_1$ (ASR1). PP systems using *engineered features*

(TR) and *learnt features* (CNN) were rebuilt from these training subsets.[9] Finally, we applied the PP systems on all our test sets Test_i (using ASR_i systems to produce ASR outputs).

Evaluation sets	**TR-100%**	**TR-50%**	**TR-%20**
Test_1	**21.99**	**22.50**	**21.81**
Test_2	22.15	22.67	22.01
Test_3	23.23	23.68	22.94
Test_4	23.00	23.43	22.64

Table 3: Evaluation of new **TR** systems on 4 evaluation datasets Test_i (ASR_i) in terms of MAE

Evaluation sets	**CNN -100%**	**CNN-50%**	**CNN-%20**
Test_1	**19.24**	**20.55**	**21.53**
Test_2	19.67	20.79	21.87
Test_3	20.64	21.70	22.90
Test_4	21.34	22.44	23.62

Table 4: Evaluation of new **CNN** systems on 4 evaluation datasets Test_i (ASR_i) in terms of MAE

Tables 3 and 4 summarize experimental results obtained with 6 ASR-performance prediction systems (3 TR and 3 CNN systems) learnt on 100%, 50% and 20% of the whole Train_1 set. These systems are evaluated on 4 $Test_i$ sets in order to measure robustness of PP systems in terms of *MAE*. We emphasize on the fact that all evaluation sets (Test_i) correspond to the same speech collection, the only difference is that texts correspond to different ASR outputs (see table 2). First of all, we notice that CNN systems outperform all TR systems in terms of MAE for 11 train/test conditions over 12 (the exception is Train-20%/Test_4).

If we focus on the difference between evaluation sets (lines), results show that Test_1 obtained the best prediction in terms of MAE on CNN and TR systems, knowing that Test_1 (average WER of 31.20%) has the best ASR output quality in table 2. We also notice that ASR output quality (see Table 2) and PP system quality seem correlated (when ASR quality is lower - eg $i = 4$ - MAE of PP systems increases). This confirms the trend, already noticed for spontaneous speech, that it is harder to predict higher WERs. Anyway, it is interesting to note that a PP system learnt for a particular ASR system (ASR_1 for instance) is not too much degraded when applied on ASR outputs obtained with a different transcription system (ASR_i for $i = 2, 3, 4$ for instance).

Looking at the amount of training data factor (columns), we observe that reducing training set size increases MAE for the CNN system. For example, on Test_1 set, we obtained respectively 19.24% and 21.53% on CNN-100% and CNN-20% systems in terms of MAE. It means that training set size have a strong influence on the performance of the PP system based on CNNs. Unlike CNNs, Table 3 shows that TR approach is not too much degraded when training size decreases (surprisingly TR-20% has better quality than TR-100% !).

[9] results corresponding to the full training data are those reported in [1] and named respectively CNN-100% and TR-100%

4.3. Effect of ASR output quality at training time for performance prediction

In previous sections, we used ASR_1 system to obtain speech transcripts and learn PP systems. In this section, we aim to investigate the effect of ASR output quality at training time for performance prediction. We learn 4 PP systems for each prediction approach named TR_i and CNN_i using speech transcripts of Train_i (ASR systems $i = 1, 2, 3, 4$) and apply them to Test_i sets. We obtain a 4x4 matrix of results for each PP system. Results are given in Table 5 and Table 6.

PP systems	Test_1	Test_2	Test_3	Test_4
$\textbf{TR}_1$	21.99	22.15	23.33	23.00
$\textbf{TR}_2$	21.68	21.72	22.67	22.33
$\textbf{TR}_3$	21.62	21.67	**22.37**	22.13
$\textbf{TR}_4$	**21.58**	**21.60**	22.66	**21.95**

Table 5: Effect of ASR output quality at training time for performance prediction - TR systems evaluated with MAE

PP systems	Test_1	Test_2	Test_3	Test_4
$\textbf{CNN}_1$	**19.24**	19.67	20.64	21.34
$\textbf{CNN}_2$	19.75	19.78	20.54	21.18
$\textbf{CNN}_3$	19.87	19.81	20.62	21.39
$\textbf{CNN}_4$	19.26	**19.28**	**19.94**	**20.22**

Table 6: Effect of ASR output quality at training time for performance prediction - CNN systems evaluated with MAE

The main result of this experiment is that both PP systems (CNN and TR) are rather stable whatever the ASR output quality is at training time. It is remarkable to note that CNN_4 system trained on Train_4 is actually slightly better to predict performance on unseen broadcast programs transcribed with better ASR systems: the last line of Table 6 displays better MAE on Test_2, Test_3 and Test_4. This result (robustness of PP systems to ASR quality at both training and test time) is important for the portability and application of performance prediction systems in practical scenarios.

5. CONCLUSION

The main goal of this research was to analyze more deeply the robustness of two ASR prediction approaches (CNN and TR) by studying the effect of speech style, training set size and ASR system considered. Performance prediction was shown to be more difficult on spontaneous speech. We also investigated the robustness of error prediction when the predictor is trained with outputs of a particular ASR system and used to predict performance on unseen broadcast programs transcribed with unseen (new) ASR systems. It was found that performance prediction is rather robust whatever the ASR output quality is at training time. Finally, effect of training size of the predictor was also investigated and it was found that while CNN predictor is better than TR predictor, it is also more sensible to training size reduction.

6. REFERENCES

[1] Zied Elloumi, Laurent Besacier, Olivier Galibert, Juliette Kahn, and Benjamin Lecouteux, "Asr performance prediction on unseen broadcast programs using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[2] Matteo Negri, Marco Turchi, José GC de Souza, and Daniele Falavigna, "Quality estimation for automatic speech recognition.," in *COLING*, 2014, pp. 1813–1823.

[3] José Guilherme Camargo de Souza, Christian Buck, Marco Turchi, and Matteo Negri, "Fbk-uedin participation to the wmt13 quality estimation shared task," in *Proceedings of the eighth workshop on statistical machine translation*, 2013, pp. 352–358.

[4] Shahab Jalalvand, Matteo Negri, Falavigna Daniele, and Marco Turchi, "Driving rover with segment-based asr quality estimation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, vol. 1, pp. 1095–1105.

[5] José GC de Souza, Hamed Zamani, Matteo Negri, Marco Turchi, and Falavigna Daniele, "Multitask learning for adaptive quality estimation of automatically transcribed utterances," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 714–724.

[6] Shahab Jalalvand, Matteo Negri, Marco Turchi, José GC de Souza, Daniele Falavigna, and Mohammed RH Qwaider, "Transcrater: a tool for automatic speech recognition quality estimation," *Proceedings of ACL-2016 System Demonstrations. Berlin, Germany: Association for Computational Linguistics*, pp. 43–48, 2016.

[7] Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier Galibert, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *LREC-Eighth international conference on Language Resources and Evaluation*, 2012, p. na.

[8] Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, and Guillaume Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news.," in *Interspeech*, 2005, pp. 1149–1152.

[9] Juliette Kahn, Olivier Galibert, Ludovic Quintard, Matthieu Carré, Aude Giraudel, and Philippe Joly, "A presentation of the repere challenge," in *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*. IEEE, 2012, pp. 1–6.

[10] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[11] Andreas Stolcke et al., "Srilm-an extensible language modeling toolkit.," in *Interspeech*, 2002, vol. 2002, p. 2002.

[12] Martine De Calmès and Guy Pérennou, "Bdlex: a lexicon for spoken and written french," in *Proceedings of 1st International Conference on Langage Resources & Evaluation*, 1998, pp. 1129–1136.

[13] Olivier Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech.," in *INTERSPEECH*, Frdric Bimbot, Christophe Cerisara, Ccile Fougeron, Guillaume Gravier, Lori Lamel, Franois Pellegrino, and Pascal Perrier, Eds. 2013, pp. 1131–1134, ISCA.

[14] Pierre Geurts, Damien Ernst, and Louis Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[15] Nicolai Meinshausen and Peter Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[16] Helmut Schmid, "Treetagger— a language independent part-of-speech tagger," *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, vol. 43, pp. 28, 1995.

[17] Yoon Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[18] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das, "Very deep convolutional neural networks for raw waveforms," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 421–425.

[19] François Chollet et al., "Keras," https://github.com/fchollet/keras, 2015.