INVESTIGATING ROBUSTNESS OF A DEEP ASR PERFORMANCE PREDICTION SYSTEM

Zied Elloumi, Olivier Galibert, Benjamin Lecouteux, Laurent Besacier

To cite this version:

Zied Elloumi, Olivier Galibert, Benjamin Lecouteux, Laurent Besacier. INVESTIGATING ROBUSTNESS OF A DEEP ASR PERFORMANCE PREDICTION SYSTEM. [Research Report] LIG lab; LNE. 2019. hal-02131931

HAL Id: hal-02131931
https://hal.archives-ouvertes.fr/hal-02131931
Submitted on 16 May 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.
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 heterogeneous 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\textsubscript{pred} corpus contains many pairs \{ASR output, Performance\} (more than 75k ASR turns in this work), a Test\textsubscript{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\textsubscript{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\textsuperscript{1}, ETAPE \textsuperscript{[7]}, ESTER 1 \& ESTER 2 \textsuperscript{[8]} and REPERE \textsuperscript{[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\textsubscript{Acoustic}) is selected from the non spontaneous speech style that corresponds mainly to broadcast news. The Train\textsubscript{pred} is a mix of both speech styles (S and NS). It is important to mention that shows in Test\textsubscript{pred} data set were unseen in the Train\textsubscript{Pred}. Moreover, more challenging (high WERs) shows were selected for Test\textsubscript{pred}.

<table>
<thead>
<tr>
<th>ASR systems</th>
<th>AM</th>
<th>LM</th>
<th>Train\textsubscript{pred}</th>
<th>Test\textsubscript{pred}</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR1 \textsuperscript{[1]}</td>
<td>DNN</td>
<td>5-gram</td>
<td>22.29</td>
<td>31.20</td>
</tr>
<tr>
<td>ASR2</td>
<td>DNN</td>
<td>3-gram</td>
<td>23.64</td>
<td>32.80</td>
</tr>
<tr>
<td>ASR3</td>
<td>SGMM</td>
<td>3-gram</td>
<td>24.58</td>
<td>34.01</td>
</tr>
<tr>
<td>ASR4</td>
<td>GMM</td>
<td>3-gram</td>
<td>27.02</td>
<td>36.79</td>
</tr>
</tbody>
</table>

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

Table 2: Description of 4 ASR systems produced and their WER performance evaluated on our Train\textsubscript{Pred} and Test\textsubscript{Pred} sets

In Table 2, we show 4 different ASR systems learnt to obtain speech transcripts of Train\textsubscript{Pred} and Test\textsubscript{Pred} datasets. The results show that ASR systems have different qualities with a higher WER (due to the effect of spontaneous speech) on Test\textsubscript{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\textsubscript{Pred} and Test\textsubscript{Pred} respectively. In next sections, we use these four ASR systems to obtain all transcripts of Train\textsubscript{pred} and/or Test\textsubscript{pred}. We note them as Train, and Test, 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 \textsuperscript{[6]}, is used for the baseline regression approach (named as TR system in our experiments). It exploits Extremely Randomized Trees algorithm \textsuperscript{[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 \textsuperscript{[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, Language model (LM) features capture the plausibility of the transcription according to a N-gram model (fluency), Lexicon-based (LEX) features are extracted from the ASR lexicon, Signal (SIG) features captures the difficulty of transcribing the input signal (general recording conditions, speaker-specific accents). This approach, based on engineered features.

\footnote{\textsuperscript{1}http://www.quaero.org \textsuperscript{2}3232M words in total - from EUbookshop, TED2013, Wi3, GlobalVoices, Gigaword, Europarl-v7, MultiUN, OpenSubtitles2016, DGT, News Commentary, News WMT, LeMonde, Trames, Wikipedia and transcriptions of our Train\textsubscript{Acoustic} dataset \textsuperscript{3}https://goo.gl/NCwpxz}

\footnote{\textsuperscript{3}We train a 5-gram LM on 3323M words text already mentioned}

\footnote{\textsuperscript{5}A feature vector containing the frequency of phoneme categories in its pronunciation is defined for each input word}

\footnote{\textsuperscript{6}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 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 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).

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).

3.3. Effect of training data size on 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 $\Delta_{\text{MAE}}$ (CNN is better when $\Delta_{\text{MAE}} > 0$) 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 $\Delta_{\text{MAE}}$ larger than 5%).

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 $\Delta_{\text{MAE}}$ 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

---

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. Finally, we applied the PP systems on all our test sets Test1 (using ASR1 systems to produce ASR outputs).

### Table 3: Evaluation of new TR systems on 4 evaluation datasets Testi (ASRi) in terms of MAE

<table>
<thead>
<tr>
<th>Evaluation sets</th>
<th>TR-100%</th>
<th>TR-50%</th>
<th>TR-%20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test1</td>
<td>21.99</td>
<td>22.50</td>
<td>21.81</td>
</tr>
<tr>
<td>Test2</td>
<td>22.15</td>
<td>22.67</td>
<td>22.01</td>
</tr>
<tr>
<td>Test3</td>
<td>23.23</td>
<td>23.68</td>
<td>22.94</td>
</tr>
<tr>
<td>Test4</td>
<td>23.00</td>
<td>23.43</td>
<td>22.64</td>
</tr>
</tbody>
</table>

### Table 4: Evaluation of new CNN systems on 4 evaluation datasets Testi (ASRi) in terms of MAE

<table>
<thead>
<tr>
<th>Evaluation sets</th>
<th>CNN-100%</th>
<th>CNN-50%</th>
<th>CNN-%20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test1</td>
<td>19.24</td>
<td>20.55</td>
<td>21.53</td>
</tr>
<tr>
<td>Test2</td>
<td>19.67</td>
<td>20.79</td>
<td>21.87</td>
</tr>
<tr>
<td>Test3</td>
<td>20.64</td>
<td>21.70</td>
<td>22.90</td>
</tr>
<tr>
<td>Test4</td>
<td>21.34</td>
<td>22.44</td>
<td>23.62</td>
</tr>
</tbody>
</table>

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

In previous sections, we used ASR1 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, and CNN, using speech transcripts of Train1 (ASR systems i = 1, 2, 3, 4) and apply them to Testi sets. We obtain a 4x4 matrix of results for each PP system. Results are given in Table 5 and Table 6.

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

<table>
<thead>
<tr>
<th>PP systems</th>
<th>Test1</th>
<th>Test2</th>
<th>Test3</th>
<th>Test4</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR1</td>
<td>21.99</td>
<td>22.15</td>
<td>23.33</td>
<td>23.00</td>
</tr>
<tr>
<td>TR2</td>
<td>21.68</td>
<td>21.72</td>
<td>22.67</td>
<td>22.33</td>
</tr>
<tr>
<td>TR4</td>
<td>21.58</td>
<td>21.60</td>
<td>22.66</td>
<td>21.95</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>PP systems</th>
<th>Test1</th>
<th>Test2</th>
<th>Test3</th>
<th>Test4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN1</td>
<td>19.24</td>
<td>19.67</td>
<td>20.64</td>
<td>21.34</td>
</tr>
<tr>
<td>CNN2</td>
<td>19.75</td>
<td>19.78</td>
<td>20.54</td>
<td>21.18</td>
</tr>
<tr>
<td>CNN3</td>
<td>19.87</td>
<td>19.81</td>
<td>20.62</td>
<td>21.39</td>
</tr>
<tr>
<td>CNN4</td>
<td>19.26</td>
<td>19.28</td>
<td>19.94</td>
<td>20.22</td>
</tr>
</tbody>
</table>

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 CNN4 system trained on Train1 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 Test2, Test3 and Test4. 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.

---

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


