Prediction of Speech Recognition Accuracy for Utterance Classification

Maxim L. Korenevsky1,3, Andrey B. Smirnov1,3, Valentin S. Mendelev1,2

1ITMO University, Saint-Petersburg, Russia
2Speech Technology Center Ltd., Saint-Petersburg, Russia
3STC-Innovations Ltd., Saint-Petersburg, Russia

korenevsky@speechpro.com, smirnov-a@speechpro.com, mendelev@speechpro.com

Abstract

The paper deals with the problem of predicting speech recognition quality and filtering poorly recognized utterances in the case when no reference transcriptions are available. In the proposed system, word error rate (WER) predictions for individual utterances are made using conditional random fields (CRF), and classification based on a given threshold is performed afterwards. We propose using a boosting technique, which significantly increases recall for high precision values. We also apply Recurrent Neural Networks (RNN) directly to the utterance classification task and obtain comparable results but with a much simpler system. All experiments were carried out on Russian spontaneous conversational speech.

Index Terms: word error rate (WER) prediction, conditional random field (CRF), precision-recall, boosting, recurrent neural nets.

1. Introduction

Speech recognition systems are becoming more common in various areas of human activity, including the processing of contact center calls and the analysis of dialogs between customer and operator. Such an analysis can be used for assessing service quality as well as in speech analytics systems [1], [2], [3] aimed at identifying problems that most frequently concern the majority of customers. In many cases recognition accuracy is low and is not sufficient for reaching a conclusion about the reason of the call, the customers problem or the operators qualification and service quality. Since the number of calls may reach tens of thousands per day, it is important to reduce the load on the processing and analytics systems. One possible way of doing this is to pre-filter calls according to their speech recognition accuracy. For this reason the task is to assess speech recognition accuracy automatically and reject the calls whose recognition accuracy is below a given threshold.

In this paper we propose a system for classifying utterances by speech recognition accuracy without using reference transcriptions. The goal of the classification is to separate utterances into two non-overlapping classes corresponding to recognition accuracy below and above a given threshold. The task is complicated by the fact that the equipment of the contact center for which our system is designed is unable to record customer and operator channels separately, so the input of our system is monaural recordings in which customer and client speech alternates and often overlaps. That is why, prior to speech recognition, these recordings are split into speaker turns according to the segmentation obtained from a diarization system. Furthermore, the average accuracy of speech recognition of separate utterances is quite low (about 60%), which is due to both recording conditions on the customer side and the large number of spontaneous artefacts in the customers speech, such as hesitations, filled pauses, repetitions, etc. These circumstances substantially reduce the reliability of confidence measures which are the main a posteriori metrics of recognition accuracy of individual words, and make the information about parts of speech and syntactic patterns in recognized text virtually useless.

The task of automatic estimation of speech recognition accuracy has been widely studied from different perspectives. Significant efforts have been made to assess speech intelligibility (see, for example, Speech Transmission Index, STI [4], [5], Speech Intelligibility Index, SII [6], etc.), but in order to compute these indices one should have information about geometrical and physical properties of the room, which are measured by recording specially generated signals, as well as about relative positions of the speech source and the receiver. The authors of [7] estimate the influence of acoustic conditions, such as the spectral structure and intensity of noise, on WER in noisy speech as compared to clean speech. They manage to achieve a significant correlation of estimated and true WER changes.

However, the main trend in recognition accuracy estimation in recent years has been to use a variety of features obtained from the recording (including those obtained during recognition) and to combine them using different classification models. The feature set usually consists of lexical and positional features (words themselves, their lengths, number of words and word positions in the utterance, etc.), syntactic features (part-of-speech tags and their confidences, functional word tags, the portion of functional words in the utterance, etc.) and ASR-based features (log-likelihoods of the words, probabilities of N-grams containing the words, posterior probabilities, etc.). In [10] a large original feature set is optimized by a greedy algorithm to improve the quality of clarifying questions in a decision-tree based dialog system. A conditional random field (CRF) model [11] is used for the same goal in a voice input SMS sys-
tem in [12]. Based on a large feature set, CRF predicts the positions in the utterance which are most likely to contain errors. In [13], [14] CRF models are used to predict types of recognition errors (substitution, insertion or deletion) for individual words, and the obtained information in turn helps to accurately estimate WER from the recognition results. The important contribution of these papers is the proposed word alignment network (WAN) based features which substantially improve prediction accuracy.

In this paper CRFs are also used for the prediction of recognition accuracy and the rejection of low accuracy recordings, but there are several significant differences from previous studies. Importantly, our task involves estimating recognition accuracy separately for each recording. In [13], very high WER prediction accuracy is obtained for a whole lecture recording, but the variation of prediction accuracy for separate utterances is quite large, which is crucial for the task of reliably rejecting low recognition accuracy utterances. Next, due to the low average recognition accuracy on our test set, there are few relatively long utterances in the target class (with WER<Threshold), so a significant part of the target class consists of short utterances. They are more difficult for the system to classify because WER prediction is less reliable for shorter utterances. To improve classification accuracy for short utterances we propose to use an additional classifier which takes into account sentence length as well as the predicted number of various recognition errors.

The main advantage of CRFs is an ability to effectively process sequential information. Another group of classifiers that share such a property are recurrent neural nets (RNNs) [15], which “have memory” and are capable of interpreting sequential data as a whole. To the best of our knowledge, RNNs have not been used previously for the classification of speech recognition results. However, our experiments show that using RNNs is a good alternative to the previously developed CRF-based approach.

Even though the proposed methods are designed primarily for data with relatively low recognition accuracy, additional experiments show that they are equally applicable for better-quality recognition results, so they are basically universal.

The rest of the paper is organized as follows. In Section 2 we briefly describe WAN-based features proposed in [13] and propose their modification that does not require constructing word confusion networks (WCNs) and uses raw word lattices from the recognizer output instead. Section 3 describes the prediction of word statuses based on CRF. In Section 4 we discuss the dependence of prediction accuracy on sentence length and propose a way to improve classification using an additional classifier (MultiBoost). Section 5 describes the classification of speech recognition results using recurrent neural nets. Section 6 contains a description of the experimental setup and results. Section 7 concludes the paper.

## 2. WAN-based features

In [13] it was proposed to use word confusion networks (WCNs) for primary estimates of recognized words statuses (correct, substitute, insert & delete). WCN (also known as “sausage”) [16] is a directed graph (Fig. 1, upper part) with relatively few nodes, which correspond to boundaries of segments in a recognized utterance, and many arcs, which correspond to hypotheses of words recognized within each segment and contain posterior probabilities of every word. The sequence of arcs (words or empty symbols eps) with maximum probability in each segment determines the 1-best recognition result. The posterior of these 1-best arcs may be treated as probabilities of the correct recognition of corresponding words. However, the posteriors of other arcs also contain useful information. The total posterior of all the words other than the 1-best one in a given segment may be treated as the probability of a substitution (if the 1-best arc corresponds to a real word) or deletion (if the 1-best arc contains the eps symbol) error. Next, if the 1-best arc contains a real word, the posterior of the eps symbol may be treated as the probability of an insertion error. In [13], a WCN-like graph (Fig. 1, lower part) whose arcs contain statuses of 1-best recognized words and the probabilities of these statuses estimated as described above is called a word alignment network (WAN), so these probabilities are further referred to as WAN-features.

WCN generation from the recognizer output is widely used. However, WCNs are not a direct outcome of ASR decoding, rather, they are the result of the consensus decoding [16] of basic word lattices and usually contain a different hypotheses set compared to them. Besides, the algorithm of word lattice-to-WCN conversion has several parameters which can significantly influence the result. To overcome this, we propose an alternative way to compute WCN-like probabilities of correct recognition, substitution, insertion and deletion, which uses word lattices directly and produces comparable prediction quality. The algorithm is based on calculating these probabilities for every frame (similarly to separate segments in the case of WCNs) for each arc of the 1-best path of the word lattice and averaging them over the arc duration. A detailed pseudo-code description of the algorithm is given below.

**Input:**
- word lattice, each arc $a$ of which contains the word $w(a)$ (or possibly eps) and its posterior probability $p(a)$;
- sequence $A$ of arcs corresponding to the 1-best recognition hypothesis.

**Output:** set of probabilities $P(\text{cor}|a)$, $P(\text{sub}|a)$, $P(\text{ins}|a)$, $P(\text{del}|a)$ for each $a \in A$.

**Algorithm:**

```python
for arc $a \in A$
    $P(\text{cor}|a) \leftarrow p(a)$;
    $P(\text{sub}|a) \leftarrow P(\text{ins}|a) \leftarrow P(\text{del}|a) \leftarrow 0$;
    Len($a$) $\leftarrow$ $\text{end}(a) - \text{start}(a)$;
    for time $t \in [\text{start}(a); \text{end}(a)]$
        for arc $\tilde{a} \neq a$, such that $t \in [\text{start}(\tilde{a}); \text{end}(\tilde{a})]$
            $P(\text{cor}|a|t) \leftarrow P(\text{cor}|a|t) + P(\text{cor}|\tilde{a}|t)$;
            $P(\text{sub}|a|t) \leftarrow P(\text{sub}|a|t) + P(\text{sub}|\tilde{a}|t)$;
            $P(\text{ins}|a|t) \leftarrow P(\text{ins}|a|t) + P(\text{ins}|\tilde{a}|t)$;
            $P(\text{del}|a|t) \leftarrow P(\text{del}|a|t) + P(\text{del}|\tilde{a}|t)$;
```

**Figure 1:** Word confusion network and its conversion to word alignment network (reproduced from [13])

<table>
<thead>
<tr>
<th>Segment $i - 1$</th>
<th>Segment $i$</th>
<th>Segment $i + 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(\text{cor}</td>
<td>a)$</td>
<td>$p(\text{sub}</td>
</tr>
<tr>
<td>$p(\text{del}</td>
<td>a)$</td>
<td>$p(\text{cor}</td>
</tr>
<tr>
<td>$p(\text{del}</td>
<td>a)$</td>
<td>$p(\text{cor}</td>
</tr>
<tr>
<td>$p(\text{del}</td>
<td>a)$</td>
<td>$p(\text{cor}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word confusion network</th>
<th>Word alignment network</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>$p_0(C) = 0.4$</td>
<td>$p_0(C) = 0.4$</td>
</tr>
<tr>
<td>$p_0(S) = 0.2$</td>
<td>$p_0(S) = 0.2$</td>
</tr>
<tr>
<td>$p_0(D) = 0.6$</td>
<td>$p_0(D) = 0.6$</td>
</tr>
</tbody>
</table>
if \((w(a) = w(\hat{a}))\) then
\[ P(\text{cor}|a) \leftarrow P(\text{cor}|a) + p(\hat{a}); \]
else if \((w(a) \neq \text{eps})\) then
\[ P(\text{sub}|a) \leftarrow P(\text{sub}|a) + p(\hat{a}); \]
else
\[ P(\text{del}|a) \leftarrow P(\text{del}|a) + p(\hat{a}); \]
\]
\[ P(\text{cor}|a) \leftarrow P(\text{cor}|a)/\text{Len}(a); \]
\[ P(\text{sub}|a) \leftarrow P(\text{sub}|a)/\text{Len}(a); \]
\[ P(\text{ins}|a) \leftarrow P(\text{ins}|a)/\text{Len}(a); \]
\[ P(\text{del}|a) \leftarrow P(\text{del}|a)/\text{Len}(a); \]
\]

3. Improving word status prediction using CRFs

WAN-features obtained with the algorithm from Section 2 (or the one in [13]) by themselves have a good correlation to true word statuses and therefore can be used directly to predict WER according to the formula
\[
W = \frac{\frac{1 + D + S}{N_{\text{rec}}} - 1}{\frac{1 + D + S}{N_{\text{rec}}} - S - I},
\]
where \(N_{\text{rec}}\) and \(N_{\text{rec}}\) are the number of words in reference and recognized sentences respectively and \(S, I, D\) are the expected numbers of substitutions, insertions and deletions, calculated by summation probabilities \(P(\text{sub}|a)\), \(P(\text{ins}|a)\) and \(P(\text{del}|a)\) respectively over all arcs \(a\) of 1-best recognition path \(A\).

However, it is reasonable to assume that better WER prediction can be obtained by using these probabilities as inputs to a classifier which takes into account information not only from a single word but from the whole word sequence. Conditional random field (CRF) is a type of classifier that has such properties. As mentioned above, CRFs were already used for speech recognition accuracy prediction. In [13] using CRFs significantly improves the quality of WER prediction for lecture recordings.

In our experiments we used the open-source CRF++ toolkit [17] and trained three different CRF models. The first one is for predicting the statuses of recognized words, \(P(\text{cor}|w), P(\text{sub}|w)\) and \(P(\text{ins}|w)\) and the two others are for predicting the number of deletions immediately before and immediately after every recognized word. Using these values we can estimate WER according to (1), calculating \(S\) and \(I\) as before by summation probabilities over the recognized words, and calculating \(D\) as:
\[
D = D_s(w_1) + D_a(w_{N_{\text{rec}}}) + \sum_{i=1}^{N_{\text{rec}}} \left( D_s(w_i) + D_a(w_i) \right)/2.
\]

We used the following features for every recognized word as input variables for CRF training:
- The word itself;
- Its confidence obtained from posteriors calculated on the word lattice;
- A binary value indicating whether the language model has a trigram of the current word and its two neighbors;
- The average confidence of word phoneme and its difference from the maximum phone confidence in the word;
- Word length in letters and phones;
- WAN-probabilities of correct recognition, substitution, insertion and deletion.

In order to take context dependence into account, the same parameters of neighboring words were used as well.

4. Improvement of target class prediction by using an additional classifier

To assess the quality of 2-class classification we used conventional metrics, namely Precision and Recall, defined as
\[
P = \frac{|D_{\text{rel}} \cap D_{\text{retr}}|}{|D_{\text{retr}}|}, \quad R = \frac{|D_{\text{rel}} \cap D_{\text{retr}}|}{|D_{\text{rel}}|},
\]
where \(D_{\text{rel}}\) is the set of all target class utterances (relevant) and \(D_{\text{retr}}\) is the set of utterances classified as belonging to the target class (retrieved).

After the very first experiments we observed two problems which make utterance classification by recognition accuracy much more difficult. First, on average, the predicted WER values tend to be lower than true WER values, so the classification based on these predicted values makes Precision lower, because higher WER utterance are classified as belonging to the target class. The other and much more serious problem is that the variance of differences between true and predicted WER is quite large, and it increases when utterances get shorter. As a result, many short utterances have predicted WER that is much lower than the true one, and thus are classified as belonging to the target class. So high Precision may be attained only at the cost of very low Recall.

To mitigate this difficulty we propose to use an additional classifier, which takes into account the probabilities of substitution, insertion and deletion predicted by CRFs, as well as the recognized sentence length, in order to make utterance classification more reliable.

4.1. Correcting table

The first classifier we tried was very simple. We divided the whole training set into 6 groups with similar sentence length and of approximately equal size. We used groups of (recognized) sentence lengths 0–1, 2, 3, 4–6, 7–12, >12 words, and for each group we estimated the classification Precision depending on the predicted WER threshold. After that, these dependencies were approximated by piecewise-linear functions for each group. During classification, the group for each utterance is determined, and for every Precision value the WER threshold is inferred from the corresponding function. This approach made it possible to increase Recall significantly for the high Precision values (the range of our main interest), although Recall became slightly lower for moderate Precision.

4.2. Boosting

The success of this very simple table classifier encouraged us to seek an advanced variant, and we chose to use boosting. Boosting is a way of constructing a strong classifier based on a linear combination of many weak classifiers. We used the open-source MultiBoost toolkit [18] as a tool for constructing such a classifier. The variables we used as input for MultiBoost classifier were the recognized sentence length as well as probabilities of
insertion, substitution and deletion from CRF++ and the WER estimate based on them. The AdaBoostMH was used as a strong classifier.

Using the MultiBoost classifier made it possible to obtain additional Recall increase for high Precision values.

5. Using recurrent neural nets

For primary experiments with recurrent neural nets we used the RNNLIB library [19]. For RNN training the objective was to classify input data sequences into two classes, namely WER>Thresh and WER<=Thresh. The input features were the same as for CRF learning. The neural net consisted of one hidden layer with 50 Bidirectional Long Short-Time Memory (BLSTM) cells [15].

6. Experiments and results

For spontaneous speech recognition we used the ASR system developed at Speech Technology Center [20] with a vocabulary of 210K words and a trigram language model of 7 million N-grams. Available recordings with a total duration of about 100 hours were divided into three parts at a 4:4:2 ratio. The first part was used to train the CRF models, the second one to train the MultiBoost classifier, and the last part was used for testing. The target class for classification consists of utterances for which the true recognition error (WER) is not higher than 20%.

6.1. WAN-features: WCNs vs. word lattices

The Table 1 shows the values of correlation coefficients between values of true sentence-level WER and several its estimates. The first two lines compare using WAN-features derived from WCNs and from word lattices. Results are given for both all test sentences and for sentence of length greater than some thresholds. As we can see, the proposed algorithm for extracting WAN features provides better values of correlation for all sentence length. For this reason every subsequent results relate only lattice-based WAN features.

Table 1: Sentence-level correlation of the predicted and true WERs

<table>
<thead>
<tr>
<th>WER estimation</th>
<th>All &gt; 5 words</th>
<th>&gt; 10 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCN-based WAN</td>
<td>0.46</td>
<td>0.50</td>
</tr>
<tr>
<td>lattice-based WAN</td>
<td>0.53</td>
<td>0.59</td>
</tr>
<tr>
<td>CRF</td>
<td>0.47</td>
<td>0.64</td>
</tr>
</tbody>
</table>

6.2. Using CRF++

As can be seen from the 3rd line of Table 1, using CRF models to predict word statuses and WER further improves the correlation for sufficiently long sentences. It is explained by the ability of CRFs to capture long-term dependencies between input symbols which is inapplicable for short utterances.

Figure 2 shows Precision-Recall graphs when using raw WAN features and CRF-predicted probabilities. It’s evident that CRF outperforms raw WAN features significantly.

6.3. Using MultiBoost

The Precision-Recall graphs after using the additional MultiBoost classifier described in Section 3 are also shown in Figure 2. We can see that using additional classifier improves Recall in the high Precision range.

6.4. Using RNN

The RNN was trained by the online gradient descent method with a learning rate of 1e-4 and using a momentum of 0.9. The training data were the same 85% of the whole dataset as for CRF training, but 5% of them were separated to form a cross-validation (CV) set. The stopping criterion was the condition of no classification improvements on the CV set during 30 successive epochs.

As we can see from Figure 2, the Precision-Recall graph corresponding to RNN classification is below the graph for the CRF and MultiBoost combination. However, the difference is minor, and RNN is a single and relatively simple classifier, whereas the best result is obtained by using three CRF models and applying MultiBoost over their results. Moreover, at the moment of writing this paper only preliminary experiments were carried out, and we believe their results can be made significantly better by careful tuning of training options and parameters.

7. Conclusions and future work

We described the results of our research on classification of separate utterances from two phone dialog participants according to speech recognition accuracy. To predict speech recognition error we used WAN-based features presented in [13], however we developed an alternative algorithm to calculate them directly from word lattices and not word confusion networks as in [13]. To improve WER prediction we used CRF models which take into account features of the whole sequence of recognized words. To further improve utterance classification we proposed to use an additional MultiBoost classifier which effectively utilizes information about the length of the recognized sentence.

Another possible choice for the classification of speech recognition results is using recurrent neural nets (RNNs). The first experiments we carried out on RNNs showed promising results, which demonstrates the potential of this approach.

8. Acknowledgements

The work was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.579.21.0008, ID RFMEFI57914X0008, and by the Government of the Russian Federation, Grant 074-U01.
9. References


