Analysing Recognition Errors in Unlimited-Vocabulary Speech Recognition

Teemu Hirsimäki and **Mikko Kurimo**
Adaptive Informatics Research Centre
Helsinki University of Technology
P.O. Box 5400, 02015, TKK, Finland
teemu.hirsimaki@tkk.fi

Abstract

We analyze the recognition errors made by a morph-based continuous speech recognition system, which practically allows an unlimited vocabulary. Examining the role of the acoustic and language models in erroneous regions shows how speaker adaptive training (SAT) and discriminative training with minimum phone frame error (MPFE) criterion decrease errors in different error classes. Analyzing the errors with respect to word frequencies and manually classified error types reveals the most potential areas for improving the system.

1 Introduction

Large vocabulary speech recognizers have become very complex. Understanding how the parts of the system affect the results separately or together is far from trivial. Still, analyzing the recognition errors may suggest how to reduce the errors further.

There exist previous work on analyzing recognition errors. Chase (1997) developed error region analysis (ERA), which reveals whether the errors are due to acoustic or language models. Greenberg et al. (2000) analyzed errors made by eight recognition systems on the Switchboard corpus. The errors correlated with the phone misclassification and speech rate, and conclusion was that the acoustic front ends should be improved further. Duta et al. (2006) analyzed the main errors made by the 2004 BBN speech recognition system. They showed that errors typically occur in clusters and differ between broadcast news (BN) and conversational telephone speech (CTS) domains. Named entities were a common cause for errors in the BN domain, and hesitation, repeats and partially spoken words in the CTS domain.

This paper analyzes the errors made by a Finnish morph-based continuous recognition system (Hirsimäki et al., 2009). In addition to partitioning the errors using ERA, we compare the number of letter errors in different regions and analyze what kind of errors are corrected when speaker adaptive training and discriminative training are taken in use. The most potential error sources are also studied by partitioning the errors according to manual error classes and word frequencies.

2 Data and Recognition System

The language model training data used in the experiments consist of 150 million words from the Finnish Kielipankki corpus. Before training the n-gram models, the words of the training data were split into morphs using the Morfessor algorithm, which has been shown to improve Finnish speech recognition (Hirsimäki et al., 2006). The resulting morph lexicon contains 50 000 distinct morphs. A growing algorithm (Siivola et al., 2007) was used for training a Kneser-Ney smoothed high-order variable-length n-gram model containing 52 million n-grams.

The acoustic phoneme models were trained on the Finnish SpeechDat telephone speech database: 39 hours from 3838 speakers for training, 46 minutes from 79 speakers for development and another similar set for evaluation. Only full sentences were used and sentences with severe noise or mispronunciations were removed.

Proceedings of NAACL HLT 2009: Short Papers, pages 193–196,
Boulder, Colorado, June 2009.

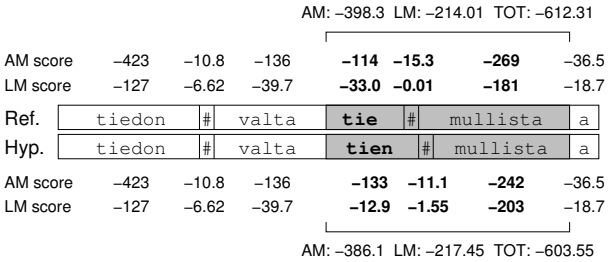

Figure 1: An example of a HYP-AM error region. The scores are log probabilities. Word boundaries are denoted by '#'. The error region only contains one letter error (an inserted 'n').

	Letter errors		
Region	ML	ML+SAT	ML+SAT+MPFE
HYP-BOTH	962	909	783
HYP-AM	1059	709	727
HYP-LM	623	597	425
REF-TOT	82	60	15
Total	2726	2275	1950
LER (%)	6.8	5.6	4.8

Table 1: SpeechDat: Letter errors for different training methods and error regions. The reference transcript contains 40355 letters in total.

The acoustic front-end consist of 39-dimensional feature vectors (Mel-frequency cepstral coefficients with first and second time-derivatives), global maximum likelihood linear transform, decision-tree tied HMM triphones with Gaussian mixture models, and cepstral mean subtraction.

Three models are trained: The first one is a *maximum likelihood* (ML) model without any adaptation. The second model (ML+SAT) enhances the ML model with three iterations of *speaker adaptive training* (SAT) using *constrained maximum likelihood linear regression* (CMLLR) (Gales, 1998). In recognition, unsupervised adaptation is applied in the second pass. The third model (ML+SAT+MPFE) adds four iterations of discriminative training with *minimum phone frame error* (MPFE) criterion (Zheng and Stolcke, 2005) to the ML+SAT model.

3 Analysis

3.1 Error Region Analysis

Error Region Analysis (Chase, 1997) can be used to find out whether the language model (LM), the acoustic model (AM) or both can be blamed for an erroneous region in the recognition output. Figure 1 illustrates the procedure. For each utterance, the final hypothesis is compared to the forced alignment of the reference transcript and segmented into correct and error regions. An *error region* is a contiguous sequence of morphs that differ from the corresponding reference morphs with respect to morph identity, boundary time-stamps, AM score, LM score, or n-gram history[1].

By comparing the AM and LM scores in the hypothesis and reference regions, the regions can be divided in classes. We denote the recognition hypothesis as HYP, and the reference transcript as REF. The relevant classes for the analysis are the following. REF-TOT: the reference would have better total score, but it has been erroneously pruned. HYP-AM: the hypothesis has better score, but only AM favors HYP over REF. HYP-LM: the hypothesis has better score, but only LM favors HYP over REF. HYP-BOTH: both the AM and LM favor HYP.

Since the error regions are independent, the letter error rate[2] (LER) can be computed separately for each region. Table 1 shows the error rates for three different acoustic models: ML training, ML+SAT, and ML+SAT+MPFE. We see that SAT decreases all error types, but the biggest reduction is in the HYP-AM class. This should be expected. In the ML case, the Gaussian mixtures contain much variance due to different unnormalized speakers, and since the test set contains only unseen speakers, many errors are expected for some speakers. Adapting the models to the test set is expected to increase the acoustic score of the reference transcript, and since in the HYP-AM regions the LM already prefers REF, corrections because of SAT are most probable there.

On the other hand, adding MPFE after SAT seems

[1] A region may be defined as an error region even if the transcription is correct (only the segmentation differs). However, since we are going to analyze the number of letter errors in the error regions, the "correct" error regions do not matter.

[2] The words in Finnish are often long and consist of several morphs, so the performance is measured in letter errors instead of word errors to have finer resolution for the results.

	Letter errors					
Class label	Total	HYP-BOTH	HYP-AM	HYP-LM	REF-TOT	Class description
Foreign	156	89	**61**	6		Foreign proper name
Inflect	143	74	26	**43**		Small error in inflection
Poor	131	37	**84**		10	Poor pronunciation or repair
Noise	124	21	**97**	6		Error segment contains some noise
Name	81	29	**29**	23		Finnish proper name
Delete	65	29	9	**27**		Small word missing
Acronym	53	44	**6**	3		Acronym
Compound	42	11	8	**23**		Word boundary missing or inserted
Correct	37	15	**19**	3		Hypothesis can be considered correct
Rare	27	11	3	**13**		Reference contains a very rare word
Insert	9	3	**6**			Small word inserted incorrectly
Other	1082	421	**379**	277	5	Other error

Table 2: Manual error classes and the number of letter errors for the ML+SAT+MPFE system.

to reduce HYP-BOTH and HYP-LM errors, but not HYP-AM errors. The number of search errors (REF-TOT) also decreases.

All in all, for all models, there seems to be more HYP-AM errors than HYP-LM errors. Chase (1997) lists the following possible reasons for the HYP-AM regions: noise, speaker pronounces badly, pronunciation model is poor, some phoneme models not trained to discriminate, or reference is plainly wrong. The next section studies these issues further.

3.2 Manual Error Classification

Next, the letter errors in the error regions were manually classified according to the most probable cause. Table 2 shows the classes, the total number of letter errors for each class, and the errors divided to different error region types.

All errors that did not seem to have an obvious cause are put under the class *Other*. Some of the errors were a bit surprising, since the quality of the audio and language seemed perfectly normal, but still the recognizer got the sentences wrong. On the other hand, the class also contains regions where the speech is very fast or the signal level is quite low.

The largest class with a specific cause is *Foreign*, which contains about 8 % of all letter errors. Currently, the morph based recognizer does not have any foreign pronunciation modeling, so it is natural that words like *Ching, Yem Yung, Villeneuve, Schumacher, Direct TV, Thunderbayssa* are not recognized correctly, since the mapping between the written form and pronunciation does not follow the normal Finnish convention. In Table 2 we see, that the acoustic model prefers the incorrect hypothesis in almost all cases. A better pronunciation model would be essential to improve the recognition. However, integrating exceptions in pronunciation to morph-based recognition is not completely straightforward. Another difficulty with foreign names is that they are often rare words, so they will get low language model probability anyway.

The errors in the *Acronym* class are pretty much similar to foreign names. Since the letter-by-letter pronunciation is not modelled, the acronyms usually cause errors.

The next largest class is *Inflect*, which contains errors where the root of the word is correctly recognized, but the inflectional form is slightly wrong (for example: *autolla/autolle, kirjeeksi/kirjeiksi*). In these errors, it is usually the language model that prefers the erroneous hypothesis.

The most difficult classes to improve are perhaps *Poor* and *Noise*. For bad pronunciations and repairs it is not even clear what the correct answer should be. Should it be the word the speaker tried to say, or the word that was actually said? As expected, the language model would have preferred the correct hypothesis in most cases, but the acoustic model have chosen the wrong hypothesis.

The *Name* and *Rare* are also difficult classes. Contrary to the foreign names and acronyms, the pronunciation model is not a problem.

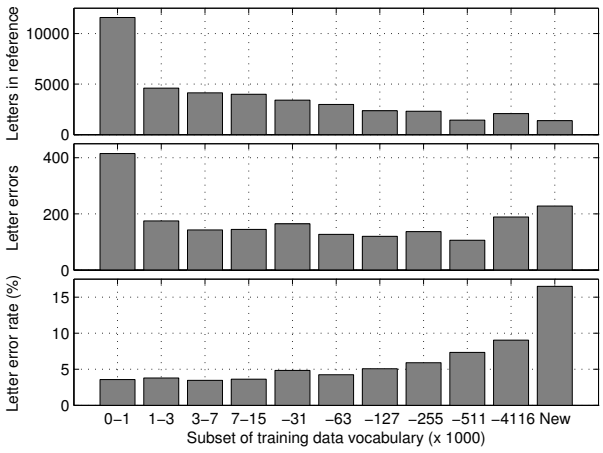

Figure 2: Frequency analysis of the SAT+MPFE system. Number of letters in reference (top), number of letter errors (middle), and letter error rate (bottom) partitioned according to word frequencies. The leftmost bar corresponds to the 1000 most frequent words, the next bar to the 2000 next frequent words, and so on. The rightmost bar corresponds to words not present in the training data.

The *Compound* errors are mainly in HYP-LM regions, which is natural since there is usually little acoustic evidence at the word boundary. Furthermore, it is sometimes difficult even for humans to know if two words are written together or not. Sometimes the recognizer made a compound word error because the compound word was often written incorrectly in the language model training data.

3.3 Frequency Analysis

In order to study the effect of rare words in more detail, the words in the test data were grouped according their frequencies in the LM training data: The first group contained all the words that were among the 1000 most common words, the next group contained the next 2000 words, then 4000, and so on, until the final group contained all words not present in the training data.

Figure 2 shows the number of letters in the reference (top), number of letter errors (middle), and letter error rate (bottom) for each group. Quite expectedly, the error rates (bottom) rise steadily for the infrequent words and is highest for the new words that were not seen in the training data. But looking at the absolute number of letter errors (middle), the majority occur in the 1000 most frequent words.

4 Conclusions

SAT and MPFE training seem to correct different error regions: SAT helps when the acoustic model dominates and MPFE elsewhere. The manual error classification suggests that improving the pronunciation modeling of foreign words and acronyms is a potential area for improvement. The frequency analysis shows that a major part of the recognition errors occur still in the 1000 most common words. One solution might be to develop methods for detecting when the problem is in acoustics and to trust the language model more in these regions.

Acknowledgments

This work was partly funded from the EC's FP7 project EMIME (213845).

References

Lin Chase. 1997. *Error-Responsive Feedback Mechanisms for Speech Recognizers*. Ph.D. thesis, Robotics Institute, Carnegie Mellon University.

Nicolae Duta, Richard Schwartz, and John Makhoul. 2006. Analysis of the errors produced by the 2004 BBN speech recognition system in the DARPA EARS evaluations. *IEEE Trans. Audio, Speech Lang. Process.*, 14(5):1745–1753.

M. J. F. Gales. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12(2):75–98.

Steven Greenberg, Shuangyu Chang, and Joy Hollenback. 2000. An introduction to the diagnostic evaluation of the Switchboard-corpus automatic speech recognition systems. In *Proc. NIST Speech Transcription Workshop*.

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20(4):515–541.

Teemu Hirsimäki, Janne Pylkkönen, and Mikko Kurimo. 2009. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Trans. Audio, Speech Lang. Process.*, 17(4):724–732.

Vesa Siivola, Teemu Hirsimäki, and Sami Virpioja. 2007. On growing and pruning Kneser-Ney smoothed n-gram models. *IEEE Trans. Audio, Speech Lang. Process.*, 15(5):1617–1624.

Jing Zheng and Andreas Stolcke. 2005. Improved discriminative training using phone lattices. In *Proc. Interspeech*, pages 2125–2128.