CONFIDENCE ESTIMATION, OOV DETECTION AND LANGUAGE ID USING PHONE-TO-WORD TRANSDUCTION AND PHONE-LEVEL ALIGNMENTS

[+]*Christopher White,* [++]*Geoffrey Zweig,* [*]*Lukas Burget,* [*]*Petr Schwarz,* [**]*Hynek Hermansky*

[+]HLT Center of Excellence, JHU, Baltimore, MD, USA
[++]Microsoft Research, Redmond, WA, USA
[*]Brno University of Technology, Brno, Czech Republic
[**]IDIAP Research Institute, Martigny, Switzerland

ABSTRACT

Automatic Speech Recognition (ASR) systems continue to make errors during search when handling various phenomena including noise, pronunciation variation, and out of vocabulary (OOV) words. Predicting the probability that a word is incorrect can prevent the error from propagating and perhaps allow the system to recover. This paper addresses the problem of detecting errors and OOVs for read Wall Street Journal speech when the word error rate (WER) is very low. It augments a traditional confidence estimate by introducing two novel methods: phone-level comparison using Multi-String Alignment (MSA) and word-level comparison using phone-to-word transduction. We show that features from phone and word string comparisons can be added to a standard maximum entropy framework thereby substantially improving performance in detecting both errors and OOVs. Additionally we show an extension to detecting English and accented English for the Language Identification (LID) task.

Index Terms— Speech Processing, Speech Recognition, Maximum Entropy Methods

1. INTRODUCTION

Automatic Speech Recognition (ASR) systems make errors during search when handling various phenomena and it is useful to predict whether each hypothesized word is incorrect. A confidence estimate predicts the reliability of the recognition result, in this case the probability of error at the word level. It enables the system to discard a result and can prevent an error from propagating in applications such as Spoken Information Retrieval or Speech Translation. Standard methods use a rich lattice that is re-normalized to provide a posterior probability, which acts as a confidence estimate (CE)[1]. We compare against a baseline estimate described in [1] as 'Cmax', which uses the maximum a posterior probability after re-normalization. Recently, state-of-the-art confidence estimation predicts the probability of error using several observations taken from a recognition lattice to train a statistical model such as a maximum entropy classifier [2].

Recognition errors can result from issues including acoustic artifacts (e.g. noise, poor articulation) or language artifacts (e.g. rare words, out of vocabulary words (OOV)). Regardless of the quality or relevance of the input, an ASR system will emit a sequence of words that is the best match to the acoustics. State-of-the-art systems currently apply all the information sources at their disposal simultaneously in this decoding process. These sources consist of the (context dependent) acoustic models, the pronunciation dictionary, and the language model, which are combined in the Finite State Transducer paradigm [3]. During the decoding process, each source influences the recognition result for better or worse. We investigate, beginning with the Johns Hopkins University Summer Workshop 2007 (JHUWS07), whether influence of linguistic constraints (from the language model and lexicon) compels the ASR system to make errors. In the OOV case this is true by definition as the spoken word is not the lexicon and therefore cannot be hypothesized, but the theory extends to other cases. This thesis is tested by considering the output of two systems in parallel, one with linguistic constraints (strong recognizer, main ASR system) and one without (weak recognizer, phone recognizer). If the two systems produce inconsistent output (e.g. acoustic likelihood, phones) during a segment of speech then perhaps the strong system has made an error (due to an OOV or otherwise).

The bulk of research completed during the workshop dealt with frame-based comparisons, whereas this work considers string-based comparisons at the word and phone-level. We examine how the one-best word and phone strings coming from the ASR system become altered during decoding and differ from the one-best phone string without linguistic constraints and the word string that is produced from a phone-to-word transducer operating on the unconstrained phone stream (described below). While it is known that multiple sources of information help in detecting errors and OOVs [4, 5] (e.g. language model factor tuning), this work is only similar in that regard and in the idea that the language model plays a role in ASR errors. We introduce two novel methods for estimating confidence: phone level comparison based on Multi-String Alignment (MSA), and word level comparison based on phone-to-word transduction. These two methods construct a coarse-to-fine (e.g. word-to-phone) grain space for selecting features for a maximum entropy (MaxEnt) framework. We

show that they provide complementary information, which combines well with a standard confidence measure. Also, our features have the benefit of not requiring a lattice, which was a major concern in [2].

2. STRING COMPARISONS

Figure 1 describes the entire comparison procedure through an example taken from our test set originating in the Wall-street Journal Part 1 test set. We consider two types of phone strings: 1) those output by a linguistically unconstrained recognizer, denoted 'HMM'; 2) those implied by the word level output of a fully constrained recognizer, denoted 'ASR'. Line (1) shows the correct transcript, which includes two OOVs: numerous, art. Line (2) shows the output of the ASR system, which includes errors: new, morris, are, part. Line (3) shows the output after transducing the HMM phone string to words. It can be seen that the ASR and transduced word strings differ exactly in the regions where the ASR output contains errors. There are three lines of phone strings, each letter represents one English phone: (4) corresponds to the phone string that comes from standard 3-gram word decoding, (6) corresponds to the string using the same acoustic models but with a 2-gram phonotatic language model, (5) corresponds to the phone string that results from altering the HMM phone string in order to get the words in line (3). For example, the transducer had to consume the phones 'm', 'R', and 's' from the HMM phone string in order to produce 'new works'. Again, the various strings differ in the regions where the ASR system has erred. The ASR system provides the word boundary in terms of phones, which can be seen as a segmentation in lines (7), (8), and (9).

```
(1)REF:  numerous works of art are based on the story
(2)ASR: new MORRIS works ARE PART are based on the story
(3) TD: new ****** works *** THAT are based on the story
(4)ASR: nUmqriswRksGpartarbYstanDJstqrI
(5) TD: nU-----wRksD-A-tarbYstanDJstqrI
(6)HMM: nUmR--swRksJvQrtQrbYstanDJstqrI
(7)ASR: nU mqris wRks G part..
(8) TD: nU ----- wRks D -A-t..
(9)HMM: nU mR--s wRks J vQrt..
```

Fig. 1. Word and Phone-Level Alignment.

2.1. Phone-Level Comparison

Initially, we have two strings of phones (e.g. lines (4) and (6) in Figure 1) that are variable length and must be aligned before comparison. Although in this case the alignment could be done with a standard Levenshtein distance, we will consider aligning three strings with the possibility of a varied alphabet and employ symbol dependent alignment cost. Our previous work [6] applies techniques in Bioinformatics for aligning multiple streams of phones using MSA, and is similarly used here to provide a feature space. In Figure 1, we can see the result of the MSA using three strings and note the additional symbol '-', which denotes an insertion.

Feature	*Description*
W	Does the ASR word match in the transduced sequence?
P	Normalized number of insertions, deletions, substitutions, and cost in alignment of transduced phones to ASR phones
J	Normalized number of insertions, deletions, substitutions, and cost in alignment of HMM phones to ASR phones
D	Number of repeated phones within word
R	#HMM phn /#ASR phn
C	Cmax

Table 1. Feature List

2.2. Word-Level Comparison

Our word-level comparison features are based on taking a decoded phone sequence (either from a full blown ASR system or from a more lightweight phone recognizer), converting it to words, and then comparing these words with those of the recognizer. The primary difference between this two-step approach and the standard recognition process is that it operates at the phone level rather than the frame level. The input is a 1-best phone string; of course, there may be subsequences within this string which cannot be matched to words, and an error model is used to assign costs to the corrections which are necessary in order to recover words. The process can be understood in terms of a noisy channel model in which we assume that a speaker utters an intended word sequence with an underlying intended phone sequence, and we receive a corrupted version of the phone sequence. We then want to recover the intended words. This can be more precisely stated if we let $\mathbf{w}_i$ denote the intended words, $\mathbf{p}_i$ denote the intended phone sequence, and $\mathbf{p}_c$ denote the corrupted phone sequence. The job of the decoder is then to determine

$$\begin{aligned}
\arg\max_{\mathbf{w}} P(\mathbf{w}|\mathbf{p}_c) &= \arg\max_{\mathbf{w}} P(\mathbf{w})P(\mathbf{p}_c|\mathbf{w}) \\
&= \arg\max_{\mathbf{w}} P(\mathbf{w}) \sum_{\mathbf{p}_i} P(\mathbf{p}_i, \mathbf{p}_c|\mathbf{w}) \\
&= \arg\max_{\mathbf{w}} P(\mathbf{w}) \sum_{\mathbf{p}_i} P(\mathbf{p}_i|\mathbf{w})P(\mathbf{p_c}|\mathbf{p}_i, \mathbf{w}) \\
&\approx \arg\max_{\mathbf{w},\mathbf{p_i}} P(\mathbf{w})P(\mathbf{p}_i|\mathbf{w})P(\mathbf{p_c}|\mathbf{p}_i)
\end{aligned}$$

The components here have straightforward interpretations: $P(\mathbf{w})$ is given by the language model; $P(\mathbf{p}_i|\mathbf{w})$ by the pronunciation model; and $P(\mathbf{p_c}|\mathbf{p}_i)$ by an error model. In this work, the error model contains insertion, deletion and substitution probabilities at the phone level. The decoding process can be implemented in terms of a transduction from phones to words, and further details are available in [7]. The intuition behind doing word comparisons based on a transduction of the phone sequence is that when correct phone sequences are present, the transduction process will tend to produce the same output as the ASR system, but where OOVs or mistakes are present, the transducer will have to guess and is unlikely to produce the same output as the recognizer. In this sense, it is similar to the method of language model jitter investigated in [4].

3. EXPERIMENTAL SETUP

3.1. Corpus

Our setup mirrors the JHUWS07, with the exception of using a 3-gram ASR decoder rather than a 2-gram, and is described in detail in the final report of the workshop. For acoustic training, the ASR system used the Switchboard Corpus obtained from the Linguistic Data Consortium (LDC). For development and testing, we use the Wallstreet Journal (WSJ) Corpus also obtained from LDC. We show results for a combined test set of WSJ0 and WSJ1 (about 19000 words) using a standard 5K vocabulary associated with WSJ0 and a 3-gram ASR decoder and 3-gram transducer. The word error rate was approximately 12% with an OOV rate of 5%.

3.2. Setup

The ASR and HMM phone strings as well as the ASR word string were generated with a system derived from AMI[DA] LVCSR [8]. The baseline score, Cmax, comes from normalizing the lattice as in [1] and represents a conventional confidence estimate. Our maximum entropy approach computes the probability that a word is incorrect or OOV based on a set of word-level features. Because the ASR system provides time-marks, we can segment the phone strings by words (Lines 7-9 in Fig. 1), thus allowing us to define word-level features based on both phone and word strings. The features we use are listed in Table 1. The MaxEnt features associated with the 'Cmax+phn:ASR+HMM' curve in Figure 2 and Figure 3 are extracted by comparing the lines (4) and (6) in Figure 1, whereas features associated with the other curves (besides Cmax) use all three phone strings. These features are shown in Table 1 as 'P, 'J', 'D', and 'R'. Features 'P' and 'J' are simple substitutions, insertions, and deletions to go from one phone string to another normalized by the number of phones in a word. Also, the cost feature comes from taking the negative log probability of confusion from the HMM phone confusion matrix after decoding the development data. 'D' counts the number of times a phone is repeated, and 'R' simply counts the ratio of the number of phones in two strings.

For the results in Figure 2 and Figure 3, the curves with 'TD:HMM' show results transducing the HMM phone string, and those with 'TD:ASR' show results transducing the ASR phone string. The word comparison features are extracted from a simple alignment between the lines (2) and (3) in Figure 1 with the capitalized words in line (2). These features are shown in Table 1 as 'W' for a binary word match. In all of our experiments, we combine features using a MaxEnt classifier similar to the one described in [2]. For example, consider the word 'part' in the ASR hypothesis in 1 line (2). Feature 'W' would be given a value of 1 since there's a mismatch between lines (2) and (3). Features P would include the 0.5 for insertions (2/4), 0.25 substitutions, and a cost of 1.14 (1.27 ins 'p', 1.60 sub 'a' for 'A', 1.53 ins 'r', 0.17 sub 't' for 't'). It can be seen that these features reflect how the various strings differ in regions of ASR error.

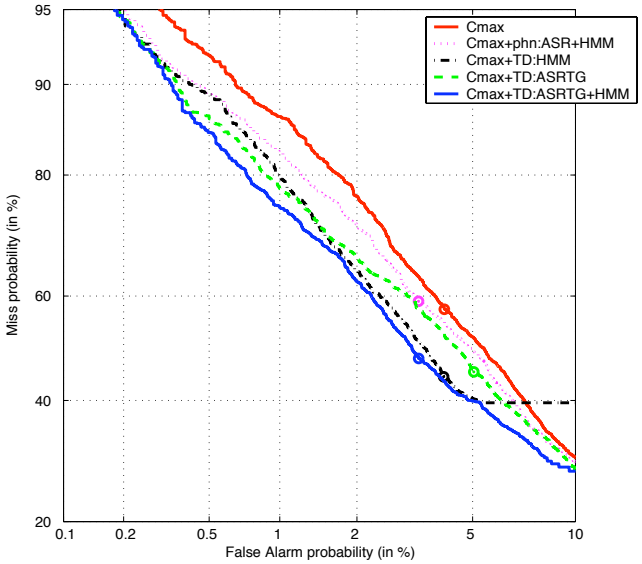

Fig. 2. ERROR Detection DET Curve.

4. RESULTS

4.1. Error and OOV Detection Results

Figures 2 and 3 show Decision Error Tradeoff (DET) curves, a standard for performance analysis used in National Institute of Standards (NIST) evaluations. In both figures, the solid line furthest from the origin shows the results using the baseline, Cmax. Figure 2 describes the performance when trying to detect all errors, where an error is defined by using a Levenshtein alignment from the ASR transcript to the reference. In Figure 3, only errors which overlap with an OOV by 5 or more frames are considered. In both figures, adding information from phone comparison and word comparison substantially improve detection performance. We can see an improvement by adding phone-level comparison information, and a further improvement by adding word-level comparison information. For example, accepting only 2% false alarms, this method detects 40% of the errors and 30% of the OOVs compared to 25% and 15% using the baseline, respectively.

5. EXTENSION TO LID

We apply this comparison technique to detect English in a language recognition evaluation (LRE) setting. In the 2005 NIST LRE several teams did poorly detecting English because many of the English segments were spoken with an Indian accent. The theory above applies here: given a phone string, if it is English then the ASR decoder and the transducer shouldn't have to alter many phones to produce an English word output, and should be similar on the phone and word levels. On the contrary, if the language is not English, then many changes should be made to the phone string in order to produce English words (which we should be able to detect). Using the LID setup in Brno [9], we can see the

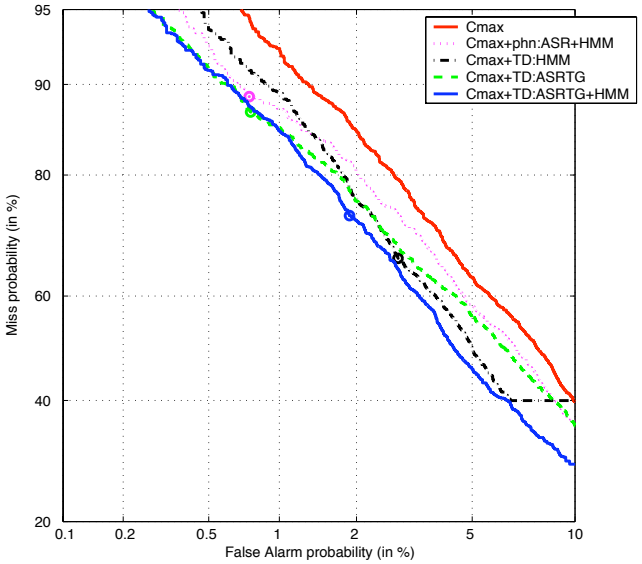

Fig. 3. OOV Detection DET Curve.

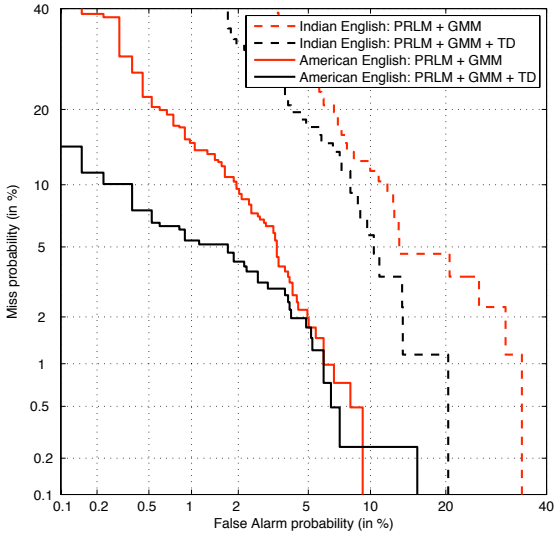

Fig. 4. English Detection (LRE 05) DET Curve.

results of adding the features described above to a state-of-the-art Gaussian Mixture Model (GMM) (trained with Maximum Mutual Information)/Parallel Phone Recognition with Language Modeling (PPRLM) English detection system. In Figure 4, half of the LRE 05 data were used to train a MaxEnt classifier, and results are shown on the other half. The dotted lines are for Indian accented English, the solid for American English. Although this is preliminary work, the results look quite promising.

6. SUMMARY AND DISCUSSION

This paper studies the utility of comparing phone streams of a conventional ASR output with varied linguistic constraint and converting from phones-to-words in order to detect errors and OOVs. We use a phone-to-word transducer for word recovery, which requires only a one-best phone string from the first stage and uses an error model on phones to recover from mistakes in the input. String based comparisons without a lattice can effectively detect errors and OOVs for WallStreet Journal speech. Coarse-to-fine comparison using a phone strings from an ASR system with and without a traditional decoder and a phone-to-word transducer facilitates robust confidence estimation, even in the difficult setting of detecting errors with a low WER. Furthermore, we successfully apply this technique to another task: detecting English in a LID setting.

7. ACKNOWLEDGMENTS

The authors wish to thank our colleagues Milind Mahajan for his help with the conditional MaxEnt training. This research was performed while one of the authors was on appointment as a U.S. Department of Homeland Security (DHS) Fellow under the DHS Scholarship and Fellowship Program, a program administered by the Oak Ridge Institute for Science and Education (ORISE) for DHS through an interagency agreement with the U.S Department of Energy (DOE). ORISE is managed by Oak Ridge Associated Universities under DOE contract number DE-AC05-06OR23100. All opinions expressed in this paper are the authors and do not necessarily reflect the policies and views of DHS, DOE, or ORISE.

8. REFERENCES

[1] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 3, 2001.

[2] C.M. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *Proc. ICASSP*, 2007.

[3] M. Mohri, F. Pereira, and M. Riley, "Weighted finite state transducers in speech recognition," *Computer Speech and Language*, vol. 16, no. 1, 2002.

[4] Lin Lawrence Chase, *Error-Responsive Feedback Mechanisms for Speech Recognizers*, Ph.D. thesis, 1997.

[5] I. Bazzi, *Modeling Out-of-Vocabulary Words for Robust Speech Recognition*, Ph.D. thesis, 2002.

[6] C.M. White, I. Shafran, and J-L. Gauvain, "Discriminative classifiers for language recognition," in *Proc. ICASSP*, 2006.

[7] G. Zweig and J. Nedel, "Empirical properties of multilingual phone-to-word transduction," in *Tech. Report MSR-TR-2007-125*, 2007.

[8] Thomas Hain, Luks Burget, John Dines, Giulia Garau, Martin Karafit, Mike Lincoln, and Vincent Wan, "The AMI Meeting Transcription System," in *Proc. NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation Worskhop*, 2006.

[9] Matejka Pavel, Burget Lukas, Schwarz Petr, and Cernocky Jan, "Brno university of technology system for nist 2005 language recognition evaluation," in *Proc. of Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006.