INTERSPEECH
2009 BRIGHTON

A Comparison of Audio-free Speech Recognition Error Prediction Methods

Preethi Jyothi, Eric Fosler-Lussier

Department of Computer Science and Engineering, Ohio State University, Columbus
jyothi@cse.ohio-state.edu, fosler@cse.ohio-state.edu

10.21437/Interspeech.2009-350

Abstract

Predicting possible speech recognition errors can be invaluable for a number of Automatic Speech Recognition (ASR) applications. In this study, we extend a Weighted Finite State Transducer (WFST) framework for error prediction to facilitate a comparison between two approaches of predicting confusable words: examining recognition errors on the training set to learn phone confusions and utilizing distances between the phonetic acoustic models for the prediction task. We also expand the framework to deal with continuous word recognition and we can accurately predict 60% of the misrecognized sentences (with an average words-per-sentence count of 15) and a little over 70% of the total number of errors from the unseen test data where no acoustic information related to the test data is utilized.

Index Terms: Finite State Transducer, Automatic Speech Recognition, Error prediction

1. Introduction

Speech recognition systems often have the problem of being unable to discern the difference between acoustically similar words. In the literature, we find several proposals of distances between words to detect confusable words and hence restrict them from appearing in the lexicon of an Automatic Speech Recognition system [1][2]. Moving one step ahead from confusion detection, predicting confusable words allows the generation of large corpora of simulated speech recognition errors. These corpora can be invaluable in developing discriminative language models and building lexicons with minimal sets of confusable words. They could also potentially be used as a source of information for the evaluation and optimization of spoken language systems.

Fosler-Lussier et al. [3] detail a weighted finite state transducer (WFST) framework that makes use of a confusion matrix between phones to model acoustic errors made by the recognizer. Information regarding acoustic confusability between words such as proposed in [4] was not incorporated within this system. Furthermore, experiments with the framework were conducted on an isolated word task and did not deal with continuous word recognition. This paper seeks to address both of these concerns.

The paper is organized as follows: the basis of our framework is discussed in the next section. Section 3 describes the details of our experimental setup including specifics regarding the training/test data split and the computation of the acoustic confusability scores between phones. Section 4 presents our design and analysis of two sets of experiments. The first set of experiments examined various confusion matrix FSTs on an isolated word task to study more about the influence of the acoustic model and the pronunciation model on the prediction task. For the second set of experiments, we applied the framework to continuous speech and also studied the influence of acoustic model information on the predictive capabilities of our experimental setup. Finally, Section 5 draws the conclusions and suggestions for future work.

2. Prediction Framework

We use the framework outlined in [3] which integrates possible acoustic confusions, pronunciation modeling and language model information into a single framework to determine confusable words for the given vocabulary. The speech recognition process can be viewed as a composition of WFSTs [5] given by the following equation:

$$W_{recog} = bestpath(F \;\; o \;\; Ac \;\; o \;\; P \;\; o \;\; Lm) \qquad (1)$$

where *F* is a finite state automaton (FSA) representing the acoustic features computed from the input utterance, *Ac* is an FST that maps acoustic features to phones using the acoustic model scores, *P* is the pronunciation model FST mapping phones to words and *Lm* is the language model FSA (with n-gram scores). On account of the invertible nature of transducers, given a word sequence W, we can generate a lattice of all possible confusable word sequences using:

$$W_{conf} = (W \; o \; Lm^{-1} \; o \; P^{-1} \; o \; Ac^{-1} \; o \; Ac \; o \; P \; o \; Lm) \quad (2)$$

Lm is a finite state automaton, and thus $Lm^{-1} = Lm$. The initial composition of W with Lm can be omitted without any trade-offs in accuracy if Lm is an n-gram grammar (as it is in our task) since it only serves to scale scores of W deterministically. Taking into consideration the infeasibility of the task of modeling the continuous space of acoustic features as an FST, [3] introduces a confusion matrix C to replace $Ac^{-1} \;\; o \;\; Ac$ that essentially captures the acoustic errors made by the recognizer. Equation (2) now reduces to

$$W_{conf} = W \; o \; P^{-1} \; o \; C \; o \; P \; o \; Lm \qquad (3)$$

Fosler-Lussier et al. [3] use counts of phonetic confusions obtained from inspecting the recognition errors to determine the confusion matrix C. The weighted lattice W_{conf} can be decoded to generate likely confusable words/sentences ordered by rank that is determined by sorting them according to the sum of the weights along a path in the phone lattice. We experiment with different methods to arrive at an appropriate representation for C that attempts at closely modeling the phonetic confusions prevalent in the recognizer and apply continuous speech to this WFST framework to further test its predictive abilities.

3. Experimental Setup

3.1. Corpus Information

Following [3], our first set of experiments were designed with a focus on developing a fitting model for the phonetic confusion

Set	Num. of Utterances	Num. of Words
Training	60349	6400
Test	13889	1205

Table 1: Training/Test sets for the Phonebook corpus.

matrix and thus we chose to use the NYNEX Phonebook isolated word recognition task [6] that eliminates the influence of the language model in the speech recognition pipeline. Acoustic models were developed on one portion of the corpus and testing was done on another part of the corpus with a completely disjoint vocabulary. We have used a similar partition as in [7]: we used their "small" training set for training our acoustic models and performed ASR transcription of the rest of the corpus using the Hidden Markov Model Toolkit [8] that gave a 15% word error rate (WER). This relatively high WER was a result of using the entire Phonebook vocabulary of 8000 words in the final decoding step and a uniform language model. The recognized corpus was further split into a training set and a test set of completely disjoint words that were used to build the confusion matrix and test its performance, respectively (Table 1). We performed both word recognition (using the CMU dictionary) and phone recognition (with 39 phoneme outputs) to better understand the influence of a lexicon restriction on the prediction task at hand. Section 3 talks more about the motivation behind generating recognized outputs on the word level and the phone level and how we make use of both these outputs in our experiments.

We designed a second set of experiments to evaluate the performance of the prediction framework as described in [3] when applied to a corpus of continuous speech (Wall Street Journal (WSJ0) corpus [9]). We used the WSJ0 training set containing 7236 speaker independent utterances to train our acoustic models and the WSJ0 standard 5K non-verbalized closed bigram language model to run the recognizer. To further explore the functionality of our prediction framework when used with recognizers of varying word error rates, we developed two acoustic models with the state output distributions from each of the phonetic Hidden Markov Models (HMMs) modeled as single Gaussians and 16-component Gaussian mixtures. Using each acoustic model, we then performed ASR transcription of a subset of the training set (1155 sentences with no out of vocabulary words) which was in turn used to build the confusion matrix. We tested its performance using the standard 5K non-verbalized test set (330 sentences) and obtained WERs of 15.0% and 7.1% for the recognizers with the single-Gaussian and 16-Gaussian acoustic models, respectively. The number of word errors per sentence is larger for the single-Gaussian recognizer thus making the prediction task more difficult as compared to the 16-Gaussian mixture model system. Section 4 further elaborates on the rationale for using these two recognizers.

3.2. Confusion Matrix Construction

We compute an alignment between the phonetic transcriptions of the actual word in the training set and the recognized word or phones. The dynamic programming alignment approach that we implemented used substitution costs based on a phonetic distance metric determined by counting the different articulatory features between two phones. As an example alignment, the word "airspeed" recognized as "emptied"' would produce the alignment "ey:eh r:m s:p p:t iy:iy d:d." The cost of each phone-to-phone mapping in the WFST confusion matrix C was set to the negative log-likelihood of observing the confusable phone given the correct phone. Once the confusion matrix has been computed, using equation (3), one can calculate a confusable word lattice for each word/sentence in the test set.

This method of constructing the confusion matrix mainly utilizes information from the pronunciation model and not the underlying HMM topology of the word. A dissimilarity measure between phones can also be computed by calculating the distance between their HMMs [4]. The distance between two HMMs is considered to be a weighted sum of the average distance between the Gaussian Mixture Models (GMMs) of the aligned states for each alignment Q, normalized by the sum of all weights.

$$d_{HMM}(p1, p2) = \begin{cases} \dfrac{\sum_{Q} P(Q) \frac{1}{L} \sum_{i=1}^{L} d_{GMM}(M_{q1i}, M_{q2i})}{\sum_{Q} P(Q)} & \text{if } p1 \neq p2 \\ 0 & \text{if } p1 = p2 \end{cases} \tag{4}$$

where Q is an alignment between the states of the HMMs of the phones $p1$ and $p2$, $P(Q)$ is the probability of the alignment Q, L is the length of the alignment, $q1i$ and $q2i$ are states of HMMs $p1$ and $p2$ that are aligned according to Q and M_{q1i},M_{q2i} are the corresponding GMMs. Specifics regarding which alignments are taken into consideration and the computation of $P(Q)$ is detailed in [4]. $d_{GMM}(.)$ is a distance measure between two GMMs that is computed using a 0.5-weighted sum of inter-dispersions (dispersion between two different GMMs A and B is a weighted double sum over all the distances between the monomodal Gaussian distributions in both the GMMs) normalized by self-dispersions (dispersion of a GMM with itself) [10]. The distance between the monomodal Gaussian distributions could be computed using Mahalanobis, Kullback-Leibler or Bhattacharya distance measures: we found that Bhattacharya distances worked best in our experiments.

4. Experimental Design and Analysis

As mentioned in Section 3, we conducted two sets of experiments to compare the performance of different representations for the confusion matrix and to evaluate the performance of this prediction framework using continuous speech.

4.1. Experiments Using the Phonebook Corpus

We adopted two evaluation methods for the first set of experiments that measured statistics about the predicted rank of observed ASR errors in the test set. The first metric computes the fraction of test set errors that are recalled when a threshold of "top n" predicted errors is applied. Typically, one is interested only in the top hypotheses from the confusable word lattice and thus, we plot the trends up to the top 100 hypotheses. This evaluation method, however, does not take into account the distribution of the ranks of the words falling below the threshold of 100. The Mean Reciprocal Rank (MRR) metric counters this by computing the mean of the reciprocal ranks (inverse of the rank of the highest ranking true positive) over all the words in the test set. From Figure (1), we observe that all our experiments perform better than chance where we define random chance as choosing the top n words from a randomly ordered list of the test set vocabulary.

First, we build a confusion matrix (Exp 1: *FST-PhoneRec*) using training data obtained from the phone recognizer which outputs phone sequences free from the restrictions of dictionary

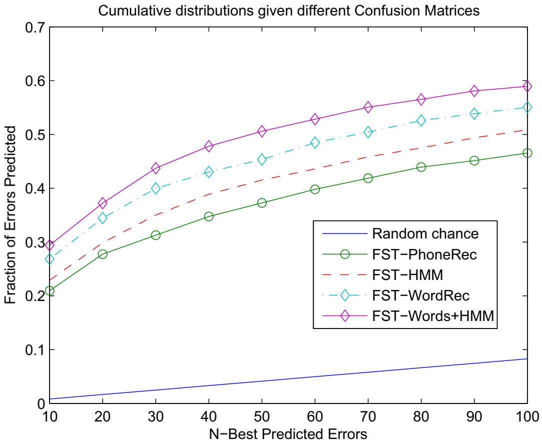

Figure 1: Recall rank of recognized errors in Phonebook corpus

entries. The second experiment uses the isolated word recognizer (described in Section 3) where phone sequences must correspond to one of the unseen words. We then build the confusion matrix based on these realized words from the recognizer (Exp 2: *FST-WordRec*). The basic difference between these two methods is the lexicon restriction.

From Figure 1, we also see that lexical restriction helps the confusion matrix generalize better by almost 10% at the threshold value of 100; this improvement is statistically significant at the $p<0.0001$ level. Thus, we observe that the lexicon introduces restrictions which are beneficial in improving the prediction abilities of the confusion matrix.

To test the utility of building confusion matrices directly from acoustic models, we build the confusion matrix for the FST framework by setting the costs between confusable phones as the distance between the two phone HMMs (from Equation (4) – Exp 3: *FST-HMM*). Costs of the insertion and deletion alignments are taken to be the average of the HMM Bhattacharya distances between all the phones. This model is indicative of the information that is obtained from the acoustic variations in the HMM representations of the phones.

From Figure 1, we observe that *"FST-HMM"* performs significantly better than *"FST-PhoneRec"* at the $p<0.005$ level which further highlights the positive influence of the lexicon restriction on the quality of prediction. We also see that *"FST-WordRec"* performs significantly better than *"FST-HMM"* at the $p<0.0001$ level. This suggests that the model which uses only distance measures between the acoustic models might do better if augmented with the likelihood of observing a confusable phone given the actual phone. We test this hypothesis by building a confusion matrix with its costs set to HMM distances weighted by the negative log-likelihood values from the word-based confusion matrix (Exp 4: *FST-Words+HMM*). We observe in Figure 1 that *"FST-Words+HMM"* does indeed perform better than all the other methods. The increase in fraction of errors predicted over *"FST-WordRec"* is statistically significant at the level of $p<0.005$. From this, we infer that *"FST-Words+HMM"* discounts phones that are close together in terms of HMM distances but not the word-based confusion matrix and retains phones which are further off in terms of HMM distances but not the word-based confusion matrix, hence improving the prediction capability of the confusion matrix.

Method	FST-PhoneRec	FST-HMM	FST-WordRec	FST-Words+HMM
MRR	0.1186	0.1341	0.1542	0.1671

Table 2: Mean Reciprocal Ranks of the Methods

Table 2 shows the MRRs for the different representations of the confusion matrix FST where a higher value indicates a better prediction strategy. We observe that the patterns are similar to the threshold metric: *"FST-Words+HMM"* method has the highest metric value and *"FST-PhoneRec"* has the lowest value. Thus, the MRR values reaffirm the hypotheses that we derived using the rank evaluation method.

4.2. Experiments using the Wall Street Journal Corpus

The second set of experiments aim at examining the functionality of our error prediction framework in the continuous speech domain. As described in Section 2, we utilize two ASR systems of single Gaussian and 16 GMM acoustic models with varying WERs. Using recognizers with different WERs strengthens the conclusions we derive regarding the positive influence of using acoustic model information for our prediction task. Similar to the recall rank metric in Figure 1, Figures 2 and 3 show measures of how often a misrecognized sentence has a rank better than a given threshold (we use a threshold of 500 for these experiments given the longer sentence lengths).

To test the importance of using a language model in the composition step (Lm in Equation 3), we generate confusable sentences from the resulting word lattice (W_{conf}) with and

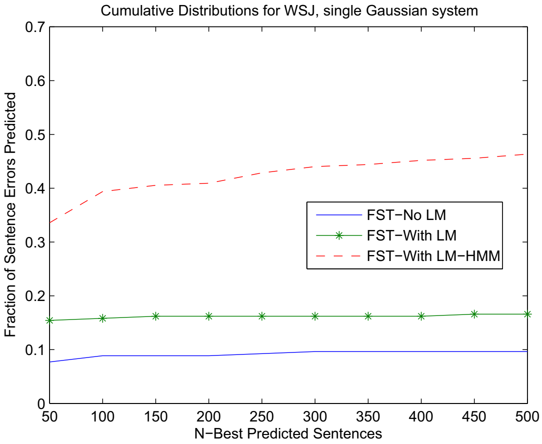

Figure 2: Recall rank of recognized sentence errors in test set (WSJ, single Gaussian system)

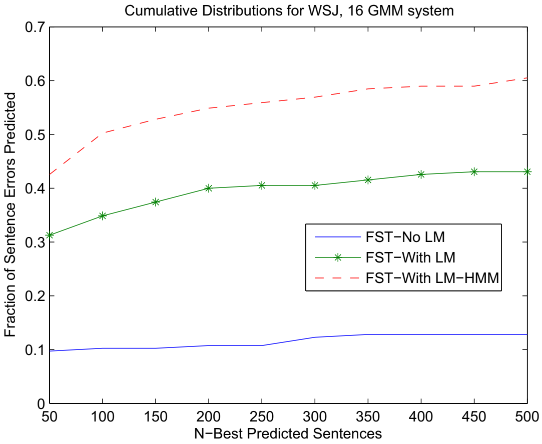

Figure 3: Recall rank of recognized sentence errors in test set (WSJ, 16 GMM system)

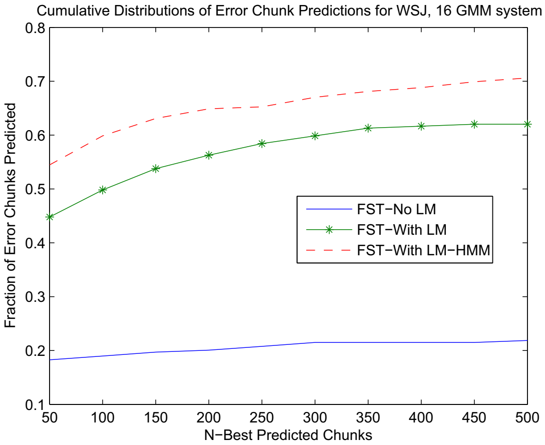

Figure 4: Recall rank of recognized error chunks in test set (WSJ, 16 GMM system)

without the influence of the WSJ standard 5K non-verbalized closed bigram language model encoded as an FST ("*FST-With LM*" and "*FST-No LM*" respectively). Keeping with our intuition, from Figures 2 and 3, it is clear that application of the language model was crucial to improving the prediction performance. The confusion matrices in "*FST-No LM*" and "*FST-With LM*" are computed using phone confusion counts as the costs between confusable phones and do not account for any information from the acoustic model. Experiment "*FST-With LM-HMM*" attempts at resolving this by constructing the confusion matrix with the costs between confusable phones set to the phone HMM distances (Equation 4). From Figures 2 and 3, we observe an increase in the fraction of sentence errors predicted for "*FST-With LM-HMM*" that is statistically significant at the level of p<0.001. This can be explained by the fact that the phone distance information, that is indicative of the confusability of the phones in the acoustic space, discounts for any incomplete information from the phone confusion counts which in turn is completely dependent on the phone distribution of the training set used to create the confusion matrix.

Predicting the complete misrecognized sentence, composed of 15 words on an average, might be too hard an evaluation criterion. We suggest a slightly more tolerant evaluation rule that calculates the number of "error chunks" that are correctly predicted by the framework within a given threshold. "Error chunks" are isolated by removing the longest common subsequence of the correct sentence and the recognized sentence from each of these sentences and grouping together the leftover parts of the sentence in sequence. For example, the sentence "At N. E. C. the need for international managers will keep rising" misrecognized as "At any see the need for national managers will keep rising" has the longest common subsequence "At the need for managers will keep rising" that is removed and the remaining words put together in sequence to generate the error chunks - "N. E. C. : any see"' and "international : national". From figure 4, we see that almost 70% of the total number of "error chunks" are predicted correctly for a threshold of 500 (as opposed to 60% of complete sentences predicted correctly from figure 3). We also found that a little more than 80% of the sentences have at least one error chunk predicted correctly. All these results are very encouraging considering that it is useful in certain ASR applications (for example, spoken dialogue systems) to be able to correctly determine parts of misrecognized sentences.

5. Conclusions

Two interesting results fall out from our experiments on an isolated word task. One is that the linguistic restrictions imposed by the lexicon on the recognizer help model the confusability between phones better as when compared to using a phone recognizer with no lexicon. This could be because the lexicon helps disambiguate the inherent confusability among phones by providing a more accurate reflection of the decoding process and hence providing a more accurate picture of the possible phone confusions. The second result is that using a confusion matrix FST with costs combining HMM phone distance information along with word-based phone confusion log likelihoods performs better than all the other methods. This corroborates our intuition that imposing lexical restrictions on the HMM costs is beneficial to the prediction task at hand.

The results from our experiments on a continuous speech domain look promising. We can accurately predict almost 60% of the erroneous sentences within a threshold of 500 and more importantly, we can predict a little more than 50% of the misrecognized sentences correctly within a modest threshold of 100 sentences implying that the confusion matrix generalizes reasonably well to the test set. The prediction capabilities of the confusion matrix improve when HMM distance information is taken into account thus emphasising the importance of using information from the acoustic models as was ascertained from the isolated word task. We also learn that the flexibility of the WFST framework allows us to integrate language model information and information from the acoustic models of the phones easily. Future work will include extending the continous word prediction task to incorporate contextual information.

6. References

[1] Tan, B. T., Gu, Y. and Thomas, T., "Word confusability measures for vocabulary selection in speech recognition", Proceedings of ASRU, 1999.

[2] Chen, J-Y., Olsen, P. A. and Hershey, J. R., "Word Confusability - Measuring Hidden Markov Model Similarity", Interspeech, Antwerp, August 2007.

[3] Fosler-Lussier, E., Amdalb, I. and Kuo, H-K. J., "A framework for predicting speech recognition errors", Speech Communication ,46(2):153–170, 2005.

[4] Anguita, J., Hernando, J., Peillon, J. and Bramoulle, S., "Detection of confusable words in automatic speech recognition", IEEE Signal Processing Letters, 8(12):585–588, 2005.

[5] Mohri, M., Riley, M., Hindle, D., Ljolje, A. and Pereira, F., "Full expansion of context-dependent networks in large vocabulary speech recognition", Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Seattle, 1998.

[6] Pitrelli, J. and Fong, C., "PHONEBOOK: NYNEX isolated words", Linguistic Data Consortium, 1995.

[7] Dupont, S., Bourlard, H., Deroo, O., Fontaine, V. and Boite, J-M., "Hybrid HMM/ANN systems for training independent tasks: experiments on Phonebook and related improvements", Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Munich, 1997.

[8] Young, S., "The HTK Hidden Markov Model Toolkit: Design and Philosophy", Online: http://htk.eng.cam.ac.uk, 1993.

[9] Garafalo, J., Graff, D., Paul, D. and Pallett, D., "CSR-I (WSJ0) Complete", Linguistic Data Consortium, 2007

[10] Wang X., Xuan, P. and Bingxi, W., "A GMM-based telephone channel classification for Mandarin speech recognition", Proceedings of International Conference on Signal Processing, 2004