IMPROVING PRONUNCIATION INFERENCE USING N-BEST LIST, ACOUSTICS AND ORTHOGRAPHY

Gopala Krishna Anumanchipalli [†‡] *Mosur Ravishankar* [‡] *Raj Reddy* [‡]

[†] Language Technologies Research Center (LTRC)
International Institute of Information Technology - Hyderabad
Gachibowli, Hyderabad, India 500032

[‡] Institute for Software Research (ISR), Carnegie Mellon University
Pittsburgh, PA 15213, USA
{gopalakr, rkm, rr}@cs.cmu.edu

ABSTRACT

In this paper, we tackle the problem of pronunciation inference and Out-of-Vocabulary (OOV) enrollment in Automatic Speech Recognition (ASR) applications. We combine linguistic and acoustic information of the OOV word using its spelling and a single instance of its utterance to derive an appropriate phonetic baseform. The novelty of the approach is in its employment of an orthography-driven n-best hypothesis and rescoring strategy of the pronunciation alternatives. We make use of decision trees and heuristic tree search to construct and score the n-best hypotheses space. We use acoustic alignment likelihood and phone transition cost to leverage the empirical evidence and phonotactic priors to rescore the hypotheses and refine the baseforms.

Index Terms— n-best list, Out-of-Vocabulary, letter-to-sound rules, pronunciation modeling, automatic pronunciation learning

1. INTRODUCTION

This work is motivated from and in part addresses our long term goal of incorporating adaptive learning in ASR applications. In the typical scenario of a dictation system, which can be considered a minimal component of most ASR applications, a user utterance is decoded into textual output. Being no exception from other ASR applications, dictation systems have a limitation of a closed vocabulary. It may be non-trivial for an end-user to manually include a new word. Hence, our attempts focus on building systems that can themselves adapt to the user, learning from the corrections to the system's output. Of the many things that can be 'learnt' from the user feedback, the current work describes our efforts to deal with the following: If the corrected word (after the correction) is an OOV, we plan to enroll it, thereby making the system potentially capable of recognizing it in a future encounter. In other words, we aim to include it into the ASR lexicon with its pronunciation derived from the spelling and an instance of its utterance. It is to be noted that the algorithm for retrieving the part of the whole utterance corresponding to a given OOV word is beyond the scope of this paper.

Several earlier approaches dealt with the problem of pronunciation inference- using either acoustics or spelling (letter-to-sound rules) or both to arrive at an appropriate baseform. Most acoustics driven methods [1] [2] [3] implement a viterbi decoding on the utterance using sub-phone (arc) acoustic units and a phone transition model to derive one or more pronunciations for each word. Orthography based methods widely use Finite State Transducers (FST) or decision trees to determine the pronunciation [9]. However, the quality of orthography based pronunciations is dependent on the grapheme-phoneme correspondence of the language. Hence, they cannot be directly used as baseforms in the ASR lexicon.

Of late, techniques combining both linguistic and acoustic information have gained focus owing to the wide range of application scenarios providing such a setting. To name a few are automatic lexicon generation [6] and systems supporting dynamic vocabularies [3] [4]. [4] and [5] for example, use syntactic and semantic information respectively to incorporate dynamic classes allowing OOV detection and enrollment. Also, [6] applies a letter-to-sound (FST) constrainer within the decoder to take advantage of the spelling of the OOV word. In this work, we exploit the linguistic information further by efficiently constructing the n-best list of pronunciation alternatives and scoring them using decision trees. The hypotheses are further rescored with costs in acoustic alignment and phone transition, achieved here using a smoothed phone bigram model. The remaining sections present our approach in detail followed by a thorough performance evaluation and analysis.

1-4244-0728-1/07/$20.00 ©2007 IEEE

2. N-BEST LIST: GENERATION AND RESCORING

Conventional approaches use the acoustics to generate the n-best list of possible phone/sub-phone strings. The n-best alternatives are re-ranked using additional knowledge sources, like a language model, to improve the intelligibility of the first best alternative, typically the output of the decoder. This is sub-optimal in situations where the orthography is available but only one instance of acoustic evidence exists. The novelty of the method proposed here lies in inverting this relationship. Our method uses the spelling information to generate an n-best list of pronunciation hypotheses, which can be subsequently rescored using available acoustic evidence and phone transition costs. The bias towards using the orthography for generating the n-best list is justified by the fact that, on an average, spelling can give more information about the pronunciation than a single acoustic exemplar, as borne out by our results (Section 4, below).

The following subsections present the decision-tree based approach in the generation of n-best pronunciation hypotheses and their subsequent rescoring using acoustic and phone transition costs.

2.1. Learning grapheme-to-phoneme rules

Decision trees offer flexibility in length of the modeling context and hence are chosen as the statistical paradigm for capturing letter-to-phone rules from a large training lexicon. Separate decision trees are trained for each letter of the alphabet. The leaves of the tree are discrete probability distributions of the phones and the internal nodes are questions about the neighboring context (*e.g.*, next letter=‘*a*’? etc.). Training and testing set features for each letter were extracted from CMUDICT [12] of 130K words. The trees were built using the letter-to-sound module within the FESTVOX [9] framework. Various context lengths of 1, 2, 3 and 4 letters on either side of the target letter were tried and the performance of the resulting trees in predicting the phone produced by a letter in an untrained word was studied. (However, we subsequently discarded 1-letter context as being overly general, and 4-letter context for overtraining the decision trees.) Fig. 2.1 shows the relative performance of the 2-letter and 3-letter context trees on a held-out set, consisting of 10% of the lexicon. As would be expected, three letter context trees outperform two letter context trees. Also, it is interesting to note that irrespective of the context length, relative performance within the letters remains the same in both cases. Furthermore, letters that produce vowel sounds (a, e, i, o, u etc.) perform significantly worse than the other consonant letters, which also agrees with intuition.

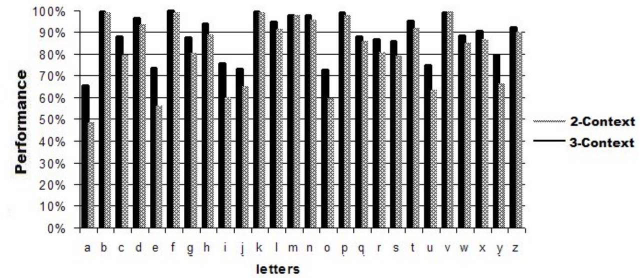

Fig2.1: Performance of the 2-letter and 3-letter context trees on the held-out data.

2.2. Orthography based n-best pronunciation generation

Given the G->P decision trees, we generate multiple (n-best) hypotheses of pronunciations for a given OOV word, as follows. From the spelling of the given OOV word, features are drawn for each letter using the same context length (2 or 3 letters) as that of a chosen set of trees. When queried with these features, the corresponding G->P trees return a list of phones, with their probabilities, for each letter in the OOV word. A variant of best first search algorithm traverses through all of the phones predicted for each letter, thus generating several pronunciation alternatives. Each pronunciation also receives a score which is the product of probabilities of the constituent phones, as determined by the decision trees. This product, which we refer as the *n-best likelihood*, is also used in the n-best list rescoring process described further below.

2.3. Acoustic alignment

Each hypothesis in the n-best list of pronunciations is aligned (using Viterbi alignment) against the single speech sample of the word, producing an acoustic likelihood for the hypothesis. The acoustic likelihood is used in re-ranking the n-best list, as described in Section 4.1. For the alignment, we used acoustic models consisting of three state context independent phone models with left-to-right topology, and 8 gaussian mixture components per state. The models were trained on the officially designated training set of the TIMIT data [7]. The *sphinx3_align* tool from the Sphinx suite [8] was used for the viterbi alignment.

2.4 Phone transition model

The function of the phone transition model is similar to that of a language model in continuous speech recognition. It provides a prior probability to each hypothesis in the n-best list. Word beginning and ending markers are also considered while computing the transitions. For our purpose, we trained a phone bigram model from CMUDICT [12]. The model was then smoothed with a uniform distribution, to avoid over-fitting to the training data. The smoothing was done as follows: If N is the number of phones, and $P(\alpha|\beta)$ the unsmoothed probability of transitioning from phone β to phone α, the smoothed transition probability is given by:

$$P_{Interpolated}(\alpha|\beta) = \omega * P(\alpha|\beta) + (1-\omega)/N$$

The scaling factor $0 < \omega \leq 1$ can be chosen according to the reliability and comprehensiveness of the dictionary. The cleaner and larger the dictionary, the higher ω can be. An optimal value for ω can be determined empirically using deleted interpolation. $\omega = 0.5$ in our experiments reported here.

2.5 N-best rescoring criteria

The n-best list of pronunciations generated according to Section 2.2 is rescored by combining the three scores: n-best likelihood, acoustic likelihood, and phone transition costs. Since the three have widely differing ranges, we propose a combination function as follows. For each alternative Φ in the n-best list, the function $\xi(\Phi)$ is computed, where:

$\xi(\Phi)$ = (Acoustic likelihood) * (n-best likelihood)$^{\eta}$ * (Phone transition penalty)$^{\gamma}$

The exponentiation weights 'η' and 'γ' are determined empirically (similar to the "language weight" in most speech recognition systems). The highest ranking pronunciation, according to ξ, is chosen for the OOV word.

3. EVALUATION

We used phone error rates (PER) of the inferred baseforms as the performance measure in our experiments. The baseline for our comparison is the PER of the top hypothesis in the original n-best list (before rescoring). For the test data[1], we chose them to be exclusively proper names, which are a good representative of OOV words in many applications. Furthermore, the peculiarities of the spoken form of proper names as opposed to their written form, makes them an appropriate tough test for the current problem. We use 100 randomly selected first and last names from the OGI names corpus [10]. This test set was excluded from the training data for acoustic, G->P trees, and phone transition probability models. We chose to use 3-letter context decision trees in the n-best list generation step.

4. RESULTS AND DISCUSSION

In Table 4.1, we present baseline PERs of the top hypothesis of the original n-best list, re-ranked by each of the three scores individually (*i.e.*, not in combination with any of the others). The table shows the average error rates obtained on the test data. The table suggests that orthography determines the pronunciation more reliably than a single instance of the speech. This may change when more than just a single instance is provided. (Furthermore, relying solely on phone transition probability to rank the n-best list is clearly useless, and is only included here for the sake of completeness.)

Baseline	PER
Orthography based n-best	22.9%
Acoustic alignment	37.8%
Phone transition	68.6%

Table 4.1: Baseline phone error rates of the factors contributing to rescoring criterion

We use the orthography-based performance of 22.9% PER as the baseline in the following sections, which deal with combining the three sources of information effectively in re-ranking the n-best list of pronunciations.

4.1 Effect of n-best likelihood + acoustic match

We first examine the effectiveness of combining acoustic likelihood with n-best likelihood in n-best selection, ignoring phone transition costs. To study this combination, we tried a wide range of values for η, measuring the PER from the best re-ranked n-best hypothesis in each case. Fig 4.1 shows the performance with varying η. The dotted line represents the baseline performance of 22.9% PER using n-best likelihood alone. We observe that as η increases, the PER drops rapidly from the acoustic-likelihood baseline of 37.8% (η=0), and reaches a minimum of approximately 19.5%. The combined information from orthography and acoustics is able to provide a 3.4% absolute improvement (14.8% relative improvement) over our n-best likelihood baseline performance of 22.9% PER.

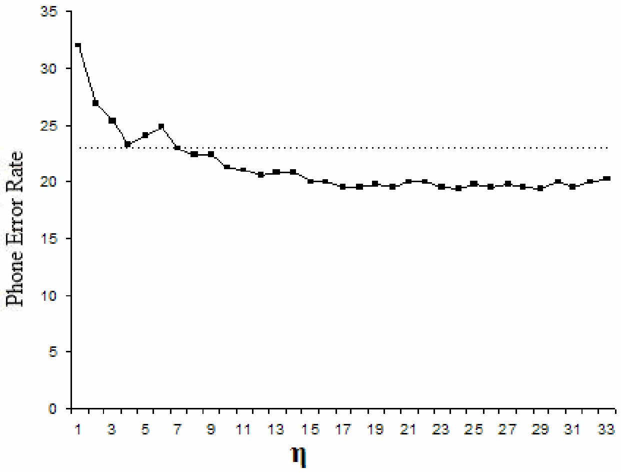

Fig 4.1: PER for varying n-best likelihood weights η

[1] The authors may be contacted to get a copy of the actual test set used.

4.2 Effect of n-best likelihood + Acoustic match + Phone transition penalty

The performance can be further improved by bringing in phonotactic constraints via the phone transition penalty. To study the effect of this factor, the n-best likelihood weight η is kept constant around the middle of the steady-state region in Fig 4.1 (we chose η=28). The phone transition penalty weight γ is varied in computing $\xi(\Phi)$ and the error rates from the re-ranked n-best list are recorded. Fig 4.2 summarizes the behavior. As shown, we are able to achieve a further reduction in PER, reaching a minimum of around 18%, which is a 21.4% relative improvement over the orthography baseline of 22.9% PER.

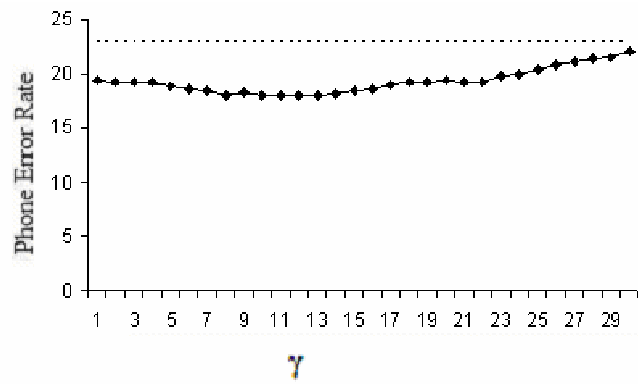

Fig 4.2: PER for varying phone transition penalty weights γ

5. CONCLUSION

We have presented a new technique for pronunciation inference for OOV words, employing an orthography-driven n-best list generation and rescoring using acoustic and other evidence. We have shown that orthographic information is more accurate than a single spoken exemplar. Accordingly, we have based our n-best list generation on the richest information available, and used the other information to re-rank the list. A comprehensive evaluation and analysis of the approach is made. We have shown that the n-best list likelihoods, combined with acoustic match likelihoods and phone transition priors can be used to reduce phone error rates of the inferred pronunciation significantly. On our test set, the PER is reduced from the orthographic baseline of 22.9% to about 18%, a 21.4% relative reduction. It remains to be seen to what extent the improvement in PER translates into improvement in word error rates in various applications. Obviously, this is highly application dependent, and we have not attempted this characterization in this paper. It also remains to be seen how far this approach can be extended when more than a single spoken exemplar is available for pronunciation inference.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation, ITR/IERI Grant No. REC-0326153. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsor(s) or of the United States Government.

The authors would like to thank David Huggins-Daines and Dr. Alan Black for help and useful suggestions.

7. REFERENCES

[1] B.Ramabhadran, L.R.Bahl, P.V. DeSouza and M. Padmanabhan, “Acoustics-Only Based Automatic Phonetic Baseform Generation” in Proc. of *ICASSP, 1998.*

[2] Sabine Deligne, Lidia Mangu, “On the Use of Lattices for Automatic Generation of Pronunciation” in Proc. of *ICASSP, 2003, Hong-Kong, China.*

[3] Sabine Deligne, Benoit Maison and Ramesh Gopinath, “Automatic Generation and Selection of Multiple Pronunciations for Dynamic Vocabularies” in Proc. of *ICASSP 2001, Salt Lake City, USA,.*

[4] Grace Chung, Stephanie Seneff, Chao Wang, and Lee Hetherington, “A Dynamic Vocabulary Spoken Dialogue Interface”, in Proc. of *ICSLP 2004, Jeju Island, Korea, October, 2004.*

[5] I. Bazzi and J. Glass, “A multi-class approach for modelling out-of-vocabulary words”, in Proc. of ICSLP, Denver, CO, September 2002.

[6] Grace Chung, Chao Wang, Stephanie Seneff, Ed Filisko, Min Tang, “Combining Linguistic Knowledge and Acoustic Information in Automatic Pronunciation Lexicon Generation*”,* in Proc. of *ICSLP 2004, Jeju Island, Korea, October, 2004.*

[7] John S. Garofalo, Lori F. Lamel, William M. Fisher, Johnathan G. Fiscus, David S. Pallett, and Nancy L.Dahlgren, “*The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM”*, Linguistic Data Consortium,1993.

[8] CMUsphinx, The Carnegie Mellon Sphinx Project http://cmusphinx.sourceforge.net

[9] Festvox : building of new synthetic voices,“http://festvox.org”

[10] Names v1.3, the CSLU, OGI Names corpus, http://cslu.cse.ogi.edu/corpora/names/

[11] Alan W Black, “Lexicons and Letter-to-sound rules”, http://www.cs.cmu.edu/~awb/papers/ISCA01/flite/node7.html

[12] The CMU Pronunciation Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict