

DISCRIMINATIVE PRONUNCIATION LEARNING USING PHONETIC DECODER AND MINIMUM-CLASSIFICATION-ERROR CRITERION

Oriol Vinyals¹, Li Deng², Dong Yu², and Alex Acero²

²Microsoft Research, Redmond, WA

¹International Computer Science Institute, Berkeley, CA

ABSTRACT

In this paper, we report our recent research aimed at improving the pronunciation-modeling component of a speech recognition system designed for mobile voice search. Our new discriminative learning technique overcomes the limitation of the traditional ways of introducing alternative pronunciations that often enlarge confusability across different lexical items. Instead, we make use of a high-quality phonetic recognizer to generate pronunciation candidates, which are then evaluated and selected using the global minimum-classification-error measure, guaranteeing a reduction of the training-set error rate after introducing alternative pronunciations. A maximum entropy approach is subsequently used to learn the weight parameters associated with the selected pronunciation candidates. Our experimental results demonstrate the effectiveness of the discriminative pronunciation learning technique in a real-world speech recognition task where pronunciation of business names presents special difficulty for high-accuracy speech recognition.

Index Terms— Pronunciation modeling, discriminative learning, MCE objective function, phonetic decoding, greedy search

1. INTRODUCTION

In current speech recognition technology, virtually all components of recognition systems are automatically learned. One notable exception is the “pronunciation” component, which determines how each lexical item (e.g., word) is composed of a sequence of constituent units such as phones. Standard pronunciations are derived from existing dictionaries, and do not cover all diversity of possible lexical items that represent various ways of pronouncing the same word. Such diversity is prevalent, and is caused by many factors including different speakers, accents, speaking conditions/styles, and different word contexts.

A number of earlier studies explored the use of alternative pronunciations or pronunciation networks [2, 3, 4, 5, 7]. On the one hand, alternative pronunciations, obtained typically

by manual addition or by maximum likelihood learning, increase the coverage of pronunciation variability. On the other hand, they may also lead to greater confusability between different lexical items. These two opposing factors usually result in either no recognition performance improvement or minor one compared with the use of standard dictionaries (e.g., [2]).

To overcome the difficulty of increased lexical confusability while introducing alternative pronunciations, one can use discriminative learning to intentionally minimize the confusability. This forms the core idea of discriminative pronunciation modeling presented in this paper. Some earlier work used the discriminative criterion of minimum classification error (MCE) to adjust the weighting parameters in the alternative pronunciations (e.g., [6]). In our work, we directly exploit the MCE criterion for selecting the most “discriminative” pronunciation alternatives, where these alternatives are derived from high-quality N-best lists or lattices in the phonetic recognition results.

The rest of this paper is organized as follows. In Section 3, we introduce our discriminative pronunciation learning framework, and in particular, the MCE objective function and its approximation for selecting the alternative pronunciations. In Section 3, we describe our experiments and present the results demonstrating the effectiveness of our pronunciation learning method. Finally, we draw conclusions and outline the future work that will extend the current work presented in this paper.

2. DISCRIMINATIVE PRONUNCIATION LEARNING

In speech recognition exploiting alternative pronunciations, the decision rule can be approximated by

$$\begin{aligned}\widehat{W} &= \arg \max_W P(X|W)P(W) \\ &= \arg \max_W \sum_q P(S_q, X|W)P(W) \\ &\cong \arg \max_W \sum_q P(S_q|W)P(X|S_q)P(W) \\ &\cong \arg \max_W \max_q P(S_q|W)P(X|S_q)P(W)\end{aligned}\quad (1)$$

The work was carried out during Oriol Vinyals’ internship at Microsoft Research, Redmond, WA. Contact: deng@microsoft.com

where X is the sequence of acoustic observations (generally feature vectors extracted from the speech waveform), W is a sequence of hypothesized words or a sentence, and $q = 1, 2, \dots, N$ is the index to multiple phone sequences S_q that are alternative pronunciations for sentence W . Each S_q is associated with the q -th path in the lattice of word-pronunciations and has an implicit dependency on W (which we drop for ease of reading).

In this work, we develop a discriminative technique that selects pronunciation(s) from the phonetic recognition results for each word such that the sentence error rate in the training data is minimized. In particular, we use the MCE objective function to approximate the empirical error rate in the training set according to

$$l(d(\Lambda)) = \sum_r l_r(d_r(\Lambda)) \quad (2)$$

with

$$l_r(d_r(\Lambda)) = \frac{1}{1 + e^{-d_r(\Lambda)}}$$

and

$$d_r(\Lambda) = -D_i(X_r, \Lambda) + \log \left\{ \frac{1}{M} \sum_{j \neq i} \exp [D_j(X_r, \Lambda)] \right\}$$

$$D_j(X_r, \Lambda) = \max_q P(S_q|W_j)P(X_r|S_q)P(W_j)$$

where X_r is the observation from the r -th sentence or utterance, W_i refers to the correct sequence of words for the r -th sentence, W_j refers to all the incorrect hypotheses (obtained from the N-best lists produced by the word decoder), M is the total number of sentence hypotheses considered for a given sentence, and Λ is the family of parameters to be estimated to optimize the objective function. These parameters include the probability weights of $P(S_q|W_j)$ for each pronunciation and the parameters of HMMs. In the experiments reported in this paper, the HMM parameters are fixed independent of the current pronunciation learning, and $P(S_q|W_j)$ is optimized by the MAXENT method [1].

We now define the ‘‘MCE score’’ of

$$\Delta^{(p,w)} = l(d(\Lambda)) - l^{(p,w)}(d(\Lambda)) \quad (3)$$

where (p, w) is a pronunciation-word pair, $l^{(p,w)}(d(\Lambda))$ is the MCE objective function associated with the added new pronunciation p for word w , and $l(d(\Lambda))$ is the same MCE objective function but associated with the canonical pronunciation. Note that $\Delta^{(p,w)}$ in (3) is an approximation of the number of corrected errors after using the new pronunciation p . Thus, if the value is positive, then there is an improvement of the recognizer performance measured by training-set error reduction with the new pronunciation p for word w .

This, therefore, turns discriminative pronunciation learning to the problem of searching for all possible ways of adding

new pronunciations to incorrectly recognized words so that $\Delta^{(p,w)}$ in (3) is positive (with a margin). That is, our method searches for the subset S (in a greedy manner) such that

$$S = \left\{ (p, w) \in T \mid \Delta^{(p,w)} > \epsilon \right\} \quad (4)$$

where set T denotes all possible (p, w) pairs.

One of the limitations of finding the pairs in a greedy way is the assumption that errors corrected by a particular pair are uncorrelated with errors corrected by any other pair. Therefore, in order to gain the greatest performance improvement, it is essential to schedule the search steps in pronunciation candidate selection over the words and sentences in the training data. In the experiments reported in Section 3 of this paper, we explored only some heuristic scheduling strategies.

Another practical issue encountered in implementing the technique described above is the computational cost since the MCE objective function depends on the entire set of training data X and the number of paths in the phone lattices produced from the phonetic decoder is very large. This makes the ranking of the possible alternative phone sequences based on the MCE objective function very expensive in computation. In our implementation, we explored two ways of computation saving. First, we approximated the MCE objective function so that it no longer depends on training data X . Second, using no approximation, we pre-stored and cached all quantities related to training data X in a huge memory and then accessed them efficiently during the execution of the discriminative pronunciation learning/selection algorithm.

3. EXPERIMENTAL EVALUATION

In the experimental evaluation of the discriminative pronunciation learning technique described in the preceding section, we used the Windows-Live-Search-for-Mobile (WLS4M) database, which consists of very large quantities of short utterances from spoken queries to the WLS4M engine. The usual behavior of a user who called in the system is to ask for a particular business and/or street and city. Often times, the business names are not in any standard lexicon and our current system uses Letter-To-Sound rules to produce the ‘‘canonical pronunciations’’ in the lexicon for these words or phrases.

The data selected for training pronunciations correspond to all the queries collected during a fixed period of the system usage, which contains approximately 100,000 utterances about business names for which we had a ‘‘click’’ by the users. When using the system, the user is prompted to click one of the options from the N-best list that the speech recognizer produces. If the user clicks an option, the corresponding query is recorded. We use these clicks as a noisy ‘‘ground truth’’. We found that in most cases (90% counted from a subset of the data) there was a clear correspondence between the click and the realistic ground truth. Therefore, our training data are not

Word	New pron.	Canonical pron.	MCE score
Whataburger	/w/ao/t/ax/r/b/er/r/g/ax/r/	/w/ao/t/ax/b/er/g/er/	29.9
Donalds	/d/aa/n/ax/l/z/	/d/aa/n/ax/l/d/z/	28.6
Stations	/t/ey/sh/ax/n/	/s/t/ey/sh/ax/n/z/	-31.3
Hall	/ao/l/	/hh/ao/l/	-18.0

Table 1. Examples of (p, w) pairs and their MCE scores. Note the different qualities of the pronunciations as related to the positive or negative MCE scores.

quite “pur” in terms of the transcription quality (which may account for some experimental results shown later in this section). However, the test data used in our experiments consist of 11,000 utterances that were all manually transcribed.

3.1. Baseline System

The baseline system used in our experiments is a standard HMM-GMM speech recognizer provided as part of the Microsoft Speech API (SAPI). The acoustic model is trained on about 3,000 hours of speech, using PLP features and an HLDA transformation. The bigram language model is used and trained with realistic system deployment data. The recognizer has a dictionary with 64,000 distinct entries. The only component that is modified from the above baseline system in our experiments reported below is the entries in the dictionary as well as the weights associated with the entries.

3.2. Training — Candidate pronunciation selection

To select the set T in (4) for the (p, w) candidates, we first performed a phone-to-word alignment on the training data. The alignment was obtained using an independent phone recognizer based on 500 single-phone plus pre-defined multi-phone units (which were the most common sequences of phones in the dictionary). Using the temporal information of the 10-best phone lists and the word recognition results, all the observed pairs (p, w) were initially selected. Among these pairs, we selected only ones (as possible new pronunciation candidates) with more than 30 occurrences and with at most three phone sequence candidates for each word in order to prune out noisy sequences of the recognized phones. This pruning step is necessary because the phone recognition accuracy is approximately 60%. Some examples of (p, w) pairs observed in the training data can be found in Table 1, where we also listed the related MCE scores defined in (3). We note that the positive scores, which correspond to reduction of the training-set error rate after introducing the new pronunciation, are associated with the new phone sequences (produced by the phonetic recognizer) with high quality, while the negative scores tend to be associated with poor quality (e.g., missing /s/ or /h/).

It is worth noting that the sparsity of the data is a limiting factor in our current approach: only 10,000 words out of the 64,000 appeared in the training data. Further, among

these 10,000 words, only 500 of them appeared frequently enough to give robust selection and parameter estimation for the pronunciation candidates against the “noise” added by the imperfection of the phonetic decoder. Thus, the cardinality of the set of candidates T was undesirably low, only 1,100 in our experiments. Applying the MCE score of (3) as described in the previous section, we obtained 400 valid candidates in the set of S . The examples in Table 1 are among the extreme pairs (i.e., with the best and worst MCE scores) from set S .

3.3. Training — Weight optimization

After the pronunciation candidates are selected via the evaluation and ranking of the MCE score, we then need to determine the weight parameters $\Lambda = P(S_q|W_j)$ before applying the decision rule of (1) for speech recognition. Among various possible approaches, we in this work adopted the maximum entropy (MAXENT) one, as this approach does not require full re-decoding of the entire training data (which is computationally expensive). Our MAXENT approach determines the weight for each of the valid pronunciations by re-scoring the N-best list produced by the original decoder. Specifically, for each utterance, we extract the following set of features:

- $AC_o + LM_o$
- $\max(0, AC_{(p,w)} - AC_o) \forall (p, w) \in S$

where AC_o and LM_o are the original acoustic and language model scores, respectively, and $AC_{(p,w)}$ is the acoustic score, obtained using forced alignment after adding a particular pronunciation pair (p, w) . Note that the second set of features will be zero almost always except when word w appears in a particular candidate from the N-best list and the new pronunciation aligns better than the baseline pronunciation. The MAXENT weights are expected to be positive. This is because when a useful pronunciation is found to align better than the baseline pronunciation, it gives a clue that the hypothesis should be ranked higher in the N-best list. In our analysis, it was found that a pronunciation that corrected more errors often had a higher weight than one that did not correct as many errors.

The actual training of the weights using MAXENT consists of two steps:

- Extract features from each item in the N-best list for each sentence — one feature from the original scores, and 400 features from each valid pair of (p, w) ;
- Learn the weights using the MAXENT formulation

3.4. Performance results

In our experiments with the manually transcribed test set, we used the set of pronunciations learned from the training data as described

Test Sets	New pron.	Sent. acc.(Baseline system)	Sent. acc. (New system)
Full		73.86%	74.52%
Whataburger	/w/ao/t/ax/r/b/er/r/g/ax/r/	58.8%	97.05%
Js	/jh/ey/s/	32.11%	44.03%
Donalds	/d/aa/n/ax/l/z/	38.65%	71.42%
Subset with full coverage		64.79%	67.35%

Table 2. Speech recognition accuracies of the baseline system vs. the new system after pronunciation learning. Results are presented for the full test set, as well as for commonly mis-recognized words and for a subset of the test set for which there is full coverage of words in the training set.

in the earlier portion of this paper. For practical reasons, we used a re-scoring method rather than a computationally expensive, full re-decoding one. The comparative performance results between the baseline recognizer and the new one with MCE-based pronunciation learning are summarized in Table 2. The sentence accuracy (Sent. acc.) is used as the performance measure for both systems. For the full test set (where we suffered from data sparsity problem), pronunciation learning helped reduce the error rate by absolute 0.7%. When the sparsity problem is artificially removed by counting the errors for only the words that are fully covered in the training set, then more than three times of performance improvement (2.6% absolute error rate reduction) are observed.

In Table 2, we also listed several common words in the test set whose recognition accuracy is drastically enhanced after adding the new pronunciation learned by the discriminative technique described earlier in this paper.

In our analysis of the experimental results, we found that when the recognizer made errors, about one quarter of times the acoustic score was right (in terms of N-best ranking) and the language model score was wrong, and about the same one quarter of times when both scores were wrong. However, about half of the times, the acoustic score was wrong while the language model score was right. Therefore, continuing the work on getting a better acoustic score (e.g. developing a better pronunciation model) is key to achieving higher accuracy for the full recognizer in our future work.

4. DISCUSSIONS AND FUTURE WORK

Pronunciation modeling in speech recognition has a long history and is known to be a very difficult problem (e.g., [2, 3, 7]). The earlier, maximum-likelihood-based approach to generating multiple pronunciation creates greater flexibility in ways that people pronounce words, but the addition of new entries often vastly enlarges confusability across different lexical items. This is because the maximum likelihood learning does not take into account such lexical confusability. Our new approach presented in this paper directly makes use of the popular MCE concept, not for learning HMM parameters in the past but for selecting pronunciation candidates produced by a phonetic recognizer. The MCE objective function, which we used measures the recognizer error rate and hence the lexical confusability, for evaluating and selecting the pronunciation candidates. If a candidate causes undesirable confusability increase, then it would not be added. (This kind of decision could not be made in the traditional maximum-likelihood-based approach.) As a result, our proposed technique may not produce realistic pronunciation for a word, but the deviations would be consistent so as to reduce the overall recognition errors in the training set as the MCE objective

function dictates.

Due to the several experimental limitations, we reported only moderate performance improvement for the overall test set. However, our experiments showed promising results when the limitations such as the data-sparsity problem were eliminated. Our future work will involve significant increase of the training set to reduce or remove the data-sparsity problem. We will also work on enriching the sources of pronunciation candidates by not only using the phonetic decoding results but also generating noisy versions of canonical pronunciations. Finally, due to the large computational cost, we only carried out the experiments using the N-best re-scoring paradigm. We expect greater performance improvement using re-decoding instead of re-scoring after the computational bottle-neck is overcome.

5. ACKNOWLEDGEMENTS

We are grateful for many helps that we received from our colleagues, Drs. G Zweig, P. Nyugen, and J. Odell in particular, in carrying out the experiments reported in Section 3 of this paper.

6. REFERENCES

- [1] S. Chen and R. Rosenfeld. "A survey of smoothing techniques for ME models" IEEE Trans. on Speech and Audio Processing, Vol. 8, 2000, pp. 37 – 50.
- [2] A. Finke and A. Waibel. "Speaking mode dependent pronunciation modelling in large vocabulary conversational speech recognition." Proc. Eurospeech, 1997, pp. 23792382.
- [3] E. Fosler-Lussier. "Dynamic Pronunciation Models for Automatic Speech Recognition," ICSI Technical Report tr-99-015, 1999.
- [4] T. Hain. "Implicit modelling of pronunciation variation in automatic speech recognition," Speech Communication, Vol. 46, 2005, pp. 171-188
- [5] T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu, Pronunciation modeling using a finite-state transducer representation, Proc. ITRW PMLA, 2002.
- [6] F. Korkmazskiy and B-H. Juang. "Discriminative training of the pronunciation networks." Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 1997, pp. 137144.
- [7] M. Riley. "A statistical model for generating pronunciation networks," Proc. Eurospeech, 1991.