AUTOMATIC GENERATION OF CONTEXT-DEPENDENT PRONUNCIATIONS

Ravishankar, M. and Eskenazi, M.
School of Computer Science
Carnegie Mellon University, Pittsburgh, PA-15213, USA.
Tel. +1 412 268 3344, FAX: +1 412 268 5576, E-mail: rkm@cs.cmu.edu

ABSTRACT

We describe experiments in modelling the dynamics of fluent speech in which word pronunciations are modified by neighbouring context. Based on all-phone decoding of large volumes of training data, we automatically derive new word pronunciation, and context-dependent transformation rules for phone sequences. In contrast to existing techniques, the rules can be applied even to words not in the training set, and across word boundaries, thus modelling context-dependent behavior. We use the technique on the Wall Street Journal (WSJ) training data and apply the new pronunciations and rules to WSJ and broadcast news tests. The changes correct a significant portion of the errors they could potentially correct. But the transformations introduce a comparable number of new errors, indicating that perhaps stronger constraints on the application of such rules are needed.

1. INTRODUCTION

Modern large vocabulary, continuous speech recognition systems have three knowledge sources: acoustic models, language models, and pronunciation lexicons. A lexicon provides pronunciation information for each word in the vocabulary in *phonemic* units, which are modelled in detail by the acoustic models. The language model provides the *a priori* probabilities of word sequences.

Whereas acoustic and language models can be trained automatically from large amounts of data ([1,2]), pronunciation lexicons are still mostly hand-crafted. In a few cases, the lexicon indeed has been either generated or tuned automatically (*e.g*, see [3,4].) However, the state of the art in this technology is restricted to learning word pronunciations in isolation that are *static*, *i.e*., that remain unchanged during recognition.

Real speech, however, is *dynamic*. Between-word co-articulation is a major problem in the recognition of continuous, fluent speech. For example, the phrase “DID YOU” often sounds something like “DIDJA”. In other words, the exact pronunciation of a word is dynamically determined by its context. This has been handled in a limited way by further handcrafting of static pronunciations for common phrases ([5, 6]). Our task is to build a model of the context-dependent dynamics of speech, and evaluate its effect on recognition accuracy.

A second problem with the conventional approach is that we need a good quantity of training data for every word in the vocabulary. Modifications learnt for one word cannot be applied to others.

In this paper we study ways of automatically or semi-automatically tuning pronunciations, in isolation and in context, and their effect on recognition accuracy. The basic principle relies on statistics gathered by processing a large set of training data using an *all-phone* recognizer. It has been tried in the past, for example in [4], to tune word pronunciations. Our approach produces a set of *word-independent phonetic transformation rules* that capture the ways in which sequences of phones in the training set are transformed into other sequences. Moreover, the transformations can be context-dependent. That is, they are qualified by the neighboring phonemes, and can only be applied in selected contexts.

Transformation rules may be applicable entirely within a word, or span across word boundaries. In the first case, they can, of course, be incorporated statically in the lexicon. In the second case, the rules must be invoked dynamically in a speech recognizer at run time, because the contexts are not known beforehand and are too numerous to be enumerated exhaustively.

As an aside, even if improving the pronunciation of a particular word has only a minor effect on recognition accuracy, it is still desirable to incorporate it in the lexicon. For example, a word may be correctly recognized in spite of an inferior pronunciation. However, the acoustic likelihood of the sentence it occurred in would be worsened and increase the chances of an error elsewhere in the utterance. Secondly, since the acoustic models are also trained from a given lexicon, they can benefit from an improvement in the latter. However, the results presented in this paper are without any retraining of the acoustic models.

The rest of this paper is organized as follows. In Section 2 we describe the details of the pronunciation learning mechanism and the extraction of context-dependent pronunciation rules. In Section 3 we provide several results; the specific modifications applied to the lexicon as well as their effect on recognition accuracy on independent data. We conclude the paper in Section 4.

2. PRONUNCIATION LEARNING

In this section we describe our process for tuning the pronunciation of words encountered in the training data, as well as extracting context-dependent transformation rules that can be applied to the entire lexicon.

2.1. Processing of Training Data

Our procedure for the identification of pronunciation errors is straightforward and has been used before in [4], as mentioned. We extend it to generate word-independent pronunciation transformation rules that are context-dependent. This training process is applied to a large volume of pre-transcribed data. It consists of the following steps:

1. Perform a *forced-recognition* of the training speech data using the corresponding transcripts and an initial lexicon. The result is a time-segmentation for each word instance (and its phoneme sequence) in the training data.
2. Decode the training data using an *all-phone recognizer*, producing the best possible phonetic transcription for each utterance.
3. Time-align the all-phone recognition result to the forced recognition result (using a conventional dynamic programming, or DP, algorithm).
4. For each word segment in the forced recognition result, extract the corresponding segment from the all-phone result as indicated by the above alignment. This is the *observed pronunciation* for the word.
5. Identify the *error regions* in the DP alignment. An error region is a maximal contiguous sequence of phonemes in the forced recognition that is different from the corresponding all-phone segment. An error region, together with its left and right phonetic contexts, forms a context-dependent pronunciation *transformation rule*.

We stress that transformation rules are derived without regard to word boundaries, *i.e.*, purely from differences in phone sequences. Hence, they are applicable to any relevant word or phrase derived from the lexicon, not just those that occur in the training data.

2.2. Extracting Pronunciations

The *observed pronunciations* obtained for individual words in Step 4 above can be incorporated directly into the lexicon. However, the observed pronunciation of a word may differ from its lexical definition for two reasons: a genuine difference between the lexical entry and what was actually spoken, or an error in the all-phone recognition. Clearly, the latter kind is spurious and should be separated from the former. This is indeed possible because a genuine difference in pronunciation would show up as a systematic and predictable pattern, while all-phone errors would exhibit a somewhat random behavior. With enough training data, the systematic changes can be isolated based on their higher frequency of occurrence. The details are covered in Section 3.1.1.

Even if the lexicon is well tuned to begin with, and there are few corrections to it, the above process is useful because it serves as a sanity check on the basic principle of producing pronunciations from all-phone results. In other words, given a good quality lexicon, most observed pronunciations should already exist in it if the process is reliable. This aspect is also covered in Section 3.1.1.

Occurrence count	Total words	Existing pron.	New pron.
10	2949	2812 (.95)	777
20	1739	1698 (.98)	308
30	1260	1236 (.98)	188
40	998	985 (.99)	123
50	829	820 (.99)	90

Table 1: No. of words (total, existing pronunciations, new pronunciations) with different occurrence counts.

In the case of the transformation rules, also, one must rely on frequency of occurrence to isolate the genuine cases of pronunciation transformation. Otherwise, errors in all-phone recognition would corrupt the results.

3. EXPERIMENTS AND RESULTS

We applied the processing described in Section 2 to the Wall Street Journal SI-284 training set ([7]). This set consists of a little under 36K sentences, with about 800K word or 2,800K phoneme occurrences. The number of distinct words is a little under 14K. The all-phone recognition was performed using fully continuous, triphone acoustic models trained on the same data. The raw phoneme error rate was about 18% (*i.e.*, the result of the DP alignment between the forced-recognition and all-phone results, step 3, in Section 2.1). It reflects both all-phone recognition errors as well as genuine differences between actual and lexical pronunciations.

3.1. Details of Pronunciation Generation

Table 1 shows the raw performance of the pronunciation extraction procedure. It is best explained by example. Taking the first row, a total of 2949 distinct words occurred *at least* 10 times in the training set. The observed word pronunciations were separated into those already existing in the lexicon, and those that did not. 2812 distinct words that had existing pronunciations occurred at least 10 times, and 777 words with new pronunciations were observed at least 10 times. (The sum of the latter two is greater than the first since the same word can show up in both the categories, with different pronunciations.)

As the minimum occurrence count is increased, the ratio of words with existing pronunciations to total words (shown in parentheses) gets closer to 1. It demonstrates that above a certain minimum count, the procedure picks the correct pronunciation with very good accuracy.

3.1.1. New Word Pronunciations

The raw set of new word pronunciations were pruned to eliminate spurious pronunciations as follows:

Word	New Pronunciation
Thousand	TH AW Z AX N
Hundred	HH AH N D AXR DD
Financial	F AY N AE N SH AX L
Asked	AE S TD
July	JH AX L AY
Actually	AE K SH AX L IY

Table 2: Sample new pronunciations.

1. New pronunciations that occurred fewer than 20 times or less than 5% of the total occurrences of the word were eliminated.
2. If an observed pronunciation was identical to an existing lexical entry for a different word, it was dropped to minimize the risk of acoustic confusion.
3. The remaining list checked by hand and unlikely pronunciations were dropped.

As a result, 144 new pronunciations were selected for addition to the testing lexicon. Table 2 lists a few examples (using the CMU Sphinx phone set, see [8]).

3.1.2. Context-Dependent Transformations

Count	Lexical phone Sequence	All-phone sequence
790	N DD S	N S
703	IX N K	IX NG K
171	IH TD IX	IH DX IX
156	AX S S	AX S

Table 3: Sample phone sequence transformations.

Similarly, we obtained pronunciation transformation rules from the high-count error regions. About 200 of them occurred 100 or more times. Table 3 lists a few rules and the frequency of their occurrence in the training set. Most transformations consist of a single phoneme being either substituted with another or entirely deleted in specific contexts. By manual inspection, we further classified the rules into the following categories:

- *Stop deletion*: Stop phonemes entirely deleted, especially at word ends when preceded and followed by non-vowel phonemes. For example, in the first row in Table 3, the DD phoneme is dropped.
- *Geminates*: Identical or related phonemes merged at word boundaries (*e.g.*, as in LAST TIME).
- *Contractions*: A series of stop phones contracted into a single stop (*e.g.*, ASKED sounds like AST).
- *Substitutions*: *E.g.*, an N at the end of a word is transformed into an M when following by a P or a B (IN PERFECT may sound like IM PERFECT).

We concentrated on geminates and stop deletion in the recognition experiments.

3.2. Recognition Experiments and Results

The new pronunciations and transformation rules were applied in recognition experiments in three ways:

	New Pronunc.	Geminate Merging	Stop Deletion
Baseline err	31/746	16/746	21/746
Corrected	8 (26%)	1 (6%)	6 (29%)
Introduced	3	3	6

(a)

Baseline err	110/1917	?/1917	?/1917
Corrected	21 (19%)	6 (?)	13 (?)
Introduced	23	1	17

(b)

Baseline err	59/1199	16/1199	65/1199
Corrected	23 (39%)	3 (19%)	16 (25%)
Introduced	14	3	18

(c)

Table 4: No. errors corrected and introduced by lexical modifications. (a) 1996 broadcast news devtest F0, (b) F1 conditions, (c) 1994 H1-C0 test set.

1. The observed word pronunciations were added to the test lexicon and used during recognition.
2. The geminate and stop-deletion models were independently incorporated into the recognition algorithm and tested.
3. Hand-selected transformation rules were applied to chosen words of the test lexicon (without reference to context), and tested.

The test sets were chosen from the following:

1. The DARPA 1996 broadcast news development and test set's F0 and F1 conditions [5]. F0 is clean, high quality, prepared speech, and F1 is similar but spontaneous speech. Uses a 51K word vocabulary.
2. The DARPA 1994 H1-C0 test set [7]; read speech from business news; pre-defined 20K vocabulary.

They were decoded using the Sphinx-3 decoder with fully continuous acoustic models ([5]).

Table 4 shows the number of baseline word errors that could have been corrected by each of the techniques., on several test sets. (*E.g.*, the first entry 31/746 means that 31 out of a total of 746 errors could have been corrected by the new pronunciations added. These figures were determined manually, and were not available for all test cases.) The table also shows the number of errors actually corrected in each case. The numbers in parentheses show the fraction of correctable errors that were actually corrected. Clearly, they are quite significant. Unfortunately, in most cases there were a comparable number of new errors introduced, substantially or completely negating the gains.

The context transformation rules were also applied to isolated word pronunciations, as mentioned. In particular, they indicated the occurrence of *displaced stress*; *i.e.*, a word being stressed at the "wrong" place. The 27 most frequent rules were processed by hand and resulted in the addition of about 920 new pronunciations to the 1996 evaluation 51K lexicon. (Most of them turned out to be corrections to existing pronunciations.) For example, the pronunciations with a dropped T:

ENTER EH N AXR
ATLANTA AX T L AE N AX

were created in this manner.

The new lexicon was tested on the 1996 broadcast news evaluation's F0 and F1 conditions. The word error rates for the two conditions changed from 28.9 and 33.6 in the baseline to 28.8 and 34.0, respectively.

3.3. Discussion

Overall, the experimental results are inconclusive. However, from a detailed analysis of the errors, similar to [9], we obtained the following insights.

The generation of new word pronunciations does work. There is a small overall gain on the three test sets. Moreover, even though the same words may be recognized, the new pronunciations are preferred in about 2.3% of the total words. Finally, the acoustic likelihood is improved in about 95% of the utterances in the H1-C0 test. These facts indicate that the techniques do help, but there are confounding factors.

Let us consider the context-dependent pronunciation transformations. Both geminate merging and stop deletion result in effectively new pronunciations that can conflict with existing ones. For example, ATROCITIES SINCE and ATROCITY SINCE became phonetically indistinguishable after the S phones in the former were merged. Hence, both have identical acoustic likelihoods, with only the language model discriminating between them. More generally, the transformations considered, when applied to words that differ only in case, tense, etc. effectively produce several homophones. This is one possible source of errors. A detailed examination of the language model probabilities provides no definite answers at this time.

Secondly, short words often behave as *garbage models*; they readily substitute for unintelligible portions of speech. As both forms of pronunciation transformations shorten the average duration of words, the number of garbage words covering the same portion of speech rises. This also increases the word error rate.

Finally, it is possible that the context constraints employed are too weak and the transformations should be applied more restrictively. Also, the experiments have been conducted with no retraining of the acoustic models after tuning the lexicon. Both these questions are under investigation.

4. CONCLUSION

We have shown the use of all-phone recognition on large volumes of training data to generate word pronunciations as well as context-dependent transformation rules that translate phone sequences into others. Such rules can be applied to arbitrary words or word sequences to model the dynamic patterns of fluent speech, in which word pronunciations are influenced by neighboring words or phonemes. We derived 144 new pronunciations and almost 1000 transformations from the Wall Street Journal SI-284 training data. The latter were eventually condensed into a few broad categories of geminates, and stop deletion in non-vowel context. Tests on broadcast news and WSJ data using these modifications show that the transformation rules have significant positive and negative impact on recognition. We believe the negative impact is effectively due to the creation of a large number of homophones. It is probably necessary to further restrict the transformation rules contextually. Also, retraining the acoustic models with the modified lexicon should give us a clearer view of the benefits of the approach.

ACKNOWLEDGEMENTS: We would like to thank Mei-Yuh Hwang, Kevin Markey and Raj Reddy for their comments and discussions on this topic.

This research was sponsored by the department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES

[1] Rabiner, L.R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Readings in Speech Recognition*, Ed. Waibel&Lee, pp. 267-296. Morgan Kaufmann Publishers.
[2] Katz, S.M., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", *IEEE Trans. on ASSP*, Vol. ASSP-35, Mar. 87, pp. 400-401.
[3] Ljolje, A. *et al*, "The AT&T 60,000 Word Speech-To-Text system", *Proc. DARPA Spoken Lang. Sys. Tech. Workshop*, Jan 1995, pp. 162-165.
[4] Sloboda, T., "Dictionary Learning for Spontaneous Speech Recognition", Proc. ICSLP, Oct. 1996.
[5] Placeway, P. *et al*, "The 1996 Hub-4 Sphinx-3 System", *Proc. DARPA Speech Recognition Workshop*, Feb. 1997.
[6] Gauvain, J-L. et al, "Acoustic Modelling in the LIMSI Nov96 Hub4 System", *Proc. DARPA Speech Recognition Workshop*, Feb. 1997.
[7] Kubala, F. "Design of the 1994 CSR Benchmark Tests", *Proc. DARPA Spoken Language Systems Technology Workshop*, pp. 41-46, Jan. 1995.
[8] Ravishankar, M., "Efficient Algorithms for Speech Recognition", Ph.D. thesis, *TR. CMU-CS-96-143*, May 1996.
[9] Chase, L., "Error-Response Feedback Mechanisms for Speech Recognizers", Ph.D. thesis, *TR CMU-RI-TR-97-18*, Apr. 1997.