# Pronunciation Dependent Language Models

*Andrej Ljolje*

AT&T Labs - Research
Florham Park, NJ 07932-0971
U.S.A.

alj@research.att.com

## Abstract

Speech recognition systems are conventionally broken up into phonemic acoustic models, pronouncing dictionaries in terms of the phonemic units in the acoustic model and language models in terms of lexical units from the pronouncing dictionary. Here we explore a new method for incorporating pronunciation probabilities into recognition systems by moving them from the pronouncing lexicon into the language model. The advantages are that pronunciation dependencies across word boundaries can be modeled including contextual dependencies like geminates or consistency in pronunciation style throughout the utterance. The disadvantage is that the number of lexical items grows proportionaly to the number of pronunciation alternatives per word and that language models which could be trained using text, now need phonetically transcribed speech or equivalent training data. Here this problem is avoided by only considering the most frequent words and word clusters. Those new lexical items are given entries in the dictionary and the language model dependent on the chosen pronunciation. The consequence is that pronunciation probabilities are incorporated into the language model and removed form the dictionary, resulting in an error rate reduction. Also, the introduction of pronunciation dependent word pairs as lexical items changes the behavior of the language model to approximate higher order n-gram language models, also resulting in improved recognition accuracy.

## 1. Introduction

Historically, speech recognition started by attempting to solve the easiest, yet important, recognition tasks. Those tasks invariably had simple language models, small vocabularies and well defined applications. Examples include digit recognition, alphabet recognition or simple lists of commands. Given the scope of the tasks it was easy to collect training data that allowed whole word acoustic models which more recently became either context dependent whole word models or context dependent head-body-tail word fraction models. As the size and scope of the recognition tasks grew, our ability to provide such training data coverage diminished, and context dependent sub-word units became the acoustic units of choice. In all of those cases, the basic unit for building the language models had always been the basic lexical unit, the word. In rare cases this model for the structure of the recognition system was broken, mostly to try to account for major pronunciation changes due to heavy coarticulation that occurs in some short phrases, which would be given a new lexical entry and appropriate dictionary entry accounting for the pronunciation changes from the baseline phonemic baseforms[1]. In the language modeling domain there have been numerous attempts to model short frequent phrases as lexical items which were mostly successful [2] [3], although the only example on the database used here was not successful [1]. In recent years our ability and willingness to collect ever more transcribed speech, albeit that the transcriptions were often noisy due to the need to do the transcriptions quickly and inexpensively, has resulted in several databases that are relatively generally available and suitable for building recognition models that until recently would have been impossible. An example is a successful attempt to build a huge acoustic model using full covariances for tens of thousands of Gaussian components [4]. To build that model all the available speech training data in the EARS program from the Switchboard database was used.

Given such large databases it is possible to start considering approaches that would bring back old approaches to speech modeling like whole world or short phrase models and thus implicit pronunciation modeling, if not for all the words then certainly for the most frequent types that cover most of the tokens in the training database. In the experiments below we describe the consequences of building such a recognition system, where a lexical item can be a phrase rather than a word, which is spoken with a particular pronunciation. The same Switchboard database is used for building the acoustic models as well as for building language models. These language models are entirely built using the transcriptions of the acoustic training data as obtained by forced alignment of the conventional triphonic acoustic model with the manually obtained lexical transcriptions.

## 2. Training Data

The training data in all the experiments is the complete set of transcribed Switchboard recordings as available for training acoustic models for the EARS-04 evaluation. This includes the transcriptions of Call Home data, Switchboard 1 data, Switchboard cellular data and the Fisher Switchboard data. This totals 23.5 million words of text and about 2200 hours of speech. One of the peculiarities of most conversational speech transcriptions is that some words are much more frequent than others, more so than in other types of collections like Broadcast News or Wall Street Journal. This means that a few types (lexical items) make up most of the training data (lexical tokens). This can easily be seen in Fig 1.

The training data is processed with preserving the chosen pronunciation for each word, including the ”silence” word, which can also have multiple pronunciations. The most frequent 100 word/pronunciation pairs are given special lexical labels. In practice, they are the lexical entry with the pronunciation suffix, an integer representing which pronunciation is associated with the lexical entry. All the other words in the lexicon are left unchanged.

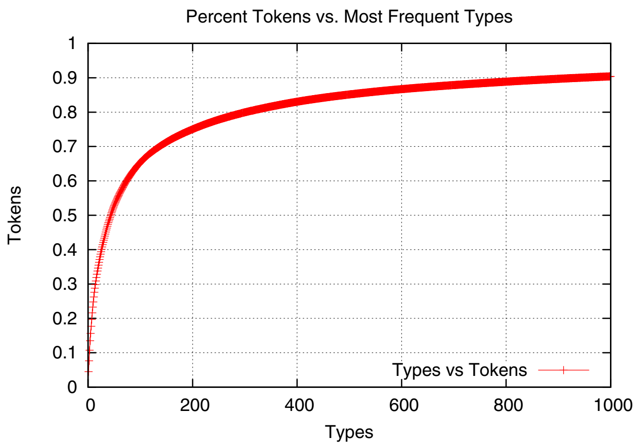

Figure 1: A few words (types) occur frequently so that the most frequent 1,000 types (out of more than 60,000) represent 90% of all tokens

The top 100 words were selected as a reasonable compromise between the amount of available data for each word, the complexity of the acoustic model and the tokens' coverage by the most frequent types. In the 100 most frequent types there were 10 words with multiple pronunciations, one with as many as four pronunciations. In addition to those multiple pronunciation examples there were an additional seven words whose most frequent pronunciation was in the top 100 word pronunciations, while the remaining pronunciation(s) did not get a special treatment but are nevertheless preserved in the dictionary as conventional words with all the pronunciations excluding the one which is contained as one of the 100 most frequent word pronunciations. This makes a total of 17 words whose pronunciation probabilities (when used) could be explicitly provided in the dictionary or alternatively combined with language model probabilities in the language model. In the experimental results described later it will be important to notice where those alternative pronunciations were left in the lexicon and where the language model reflected the word pronunciation probability as part of the n-gram probabilities. In other words, there are two important components to the pronunciation models, which pronunciation alternatives to preserve in the dictionary, and which probability should be associated with those pronunciations. Similar treatment of alternative pronunciations has also been applied on up to 240 most frequent word pairs which did not include the silence word. When a silence is inserted between the words in a frequent word bigram, the silence was treated as a lexical item and word pairs which included the word ”silence” were ignored. In the top 240 bigrams there were 15 word pairs with two alternative pronunciations and one with three pronunciations. Most of the words making up the 240 most frequent bigrams were included in the 100 most frequent words.

## 3. Single Word Pronunciations

All the results in this section were based on an identical acoustic model, using a conventional phoneme set, a typical front end and a triphonic tied state model structure. The model is trained using 60-dimensional features obtained by concatenating 9 frames of 13 dimensional MFCC cepstra reduced to 60 features by LDA and de-correlated using a single semi-tied covariance. The model used 9300 tied states to represent more than 25,000 triphonic 3-state left-to-right HMMs. The triphonic context classes were determined using decision trees. All the differences in experimental results are the result of the differences in the implementation of pronunciation probabilities and the effect of combining lexical entries into longer span entries on the language model.

All the experiments were run on Xeon 2.4 GHz processors on the Switchboard evaluation data from 2002, which included Switchboard 1, Switchboard 2 and Switchboard cellular data in equal amounts.

The baseline results used transcription based statistics to determine the probability of pronunciation alternatives for the 100 most frequent types. There are 10 words whose pronunciation alternatives are included in the top 100 word pronunciations. There are an additional 7 words whose most frequent pronunciation is included in the top 100, while the rest of them are not. Nevertheless, in the rest of the experiments it will be seen that as the 100 most frequent pronunciations are given a separate lexical entry, it will still constitute pronunciation modeling for those words when the pronunciation probabilities are incorporated as part of the language model. The performances with no pronunciation probabilities, pronunciation probabilities for the 10 most frequent words, and 17 most frequent words with alternative probabilities incorporated within the dictionary are shown in Fig 2. The language model used in these experiments was trained on the lexical transcriptions of the acoustic training data, with only a single type for each word which had multiple pronunciations in the dictionary.

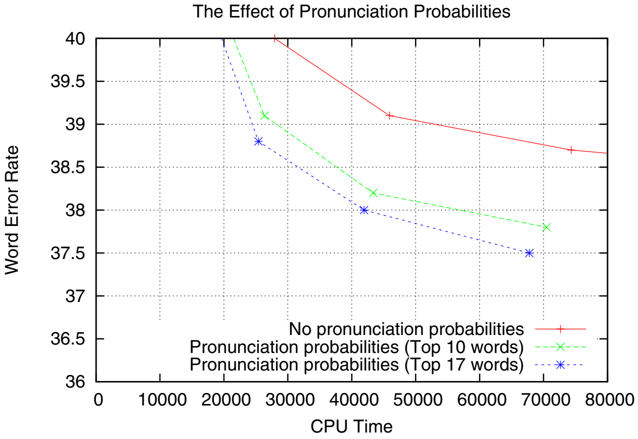

Figure 2: The effect of providing pronunciation probabilities for the most frequent 10 and 17 words with pronunciation alternatives

It is clear that adding pronunciation probabilities to the lexicon is important, when the acoustic model was trained with those alternatives. The impact on the recognition accuracy is significant even when only a handful of words have pronunciation alternatives especially if they are some of the most frequent words.

The next set of results used a different lexical representation than the previous experiments, so direct comparison to the results in Fig 2 is not possible. However it will be possible to compare what happens when it is not dictionary that contains the pronunciation probabilities, but the lexicon contains unique entries for each pronunciation variant thus moving pronunciation probabilities from the dictionary into the language model. This comparison is shown in Fig 3.

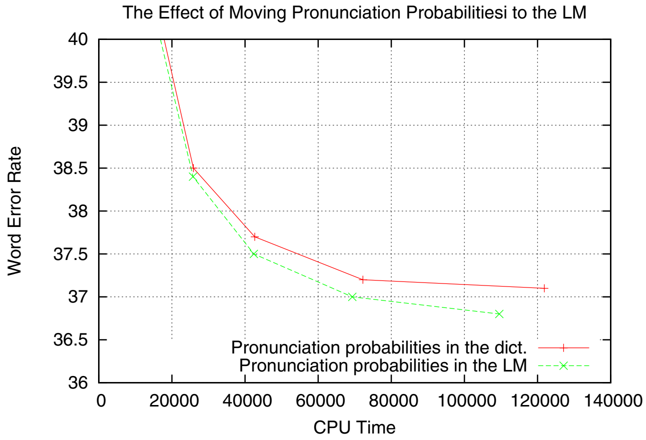

Figure 3: The effect of moving pronunciation probabilities for the most frequent 17 words with pronunciation alternatives from the dictionary to the LM

It is clear that it is advantageous to allow the language model to extract all the available information from the fact that some words commonly have different pronunciations. In addition to exploiting geminates, which is not easy through pronunciation modeling within a dictionary, the distinction between common variants where the reduction in a vowel determines the difference between the verb and the noun type (eg. reset /r ax s eh t/ vs. /r iy s eh t/ respectively) could also significantly contribute to performance improvements. In the past it was discovered that elaborate pronunciation models with many alternatives per word were in fact not performing pronunciation modeling, but were in fact allowing context dependent models to automatically form speaker clusters. Shifting those models into the language model would allow for the constraint of staying within the same cluster throughout the utterance, which was not possible to enforce, explicitly or implicitly as it would be with longer span constraints of n-gram language models.

## 4. Word Cluster Pronunciations

Here we investigate the consequences of combining words into new longer lexical items in the language model. When two words are merged to form a new lexical item (eg. "I know" becomes "I_know", than it implicitly increases the context of the n-gram type language model. This is often done with acronyms, so that context is not entirely lost (eg. "A. T. & T." would become "A._T._&_T."). In this case we do it on the most frequent word pronunciation pairs, starting from none and trying up to 240 most frequent pairs. The training data is otherwise identical, as is the method for building the language model. The only difference is the transformation of word (pronunciation specific) pairs into a single word, which is also pronunciation specific. The largest number of pairs tried was 240 which included 17 pairs with more than one pronunciation. In other words, there were 223 word pairs, with 16 pairs having 2 alternative pronunciations and one having three. It is important to note that the 223 word bigrams whose 240 alternative pronunciations make up the most frequent word pronunciation pairs have in principle a total of 399 possible pronunciations, given the alternatives in the dictionary for individual words that make up the the most frequent word pairs. The benefit is similar to using bigram language models instead of unigram language models.

The result of this change for 29, 60, 120 and 240 pairs is shown in Fig 4.

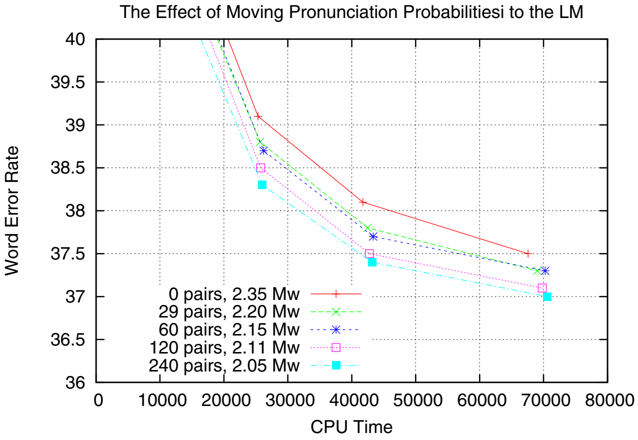

Figure 4: The effect of converting word pronunciation pairs into single lexical items for language modeling

For each of the conditions in Fig 4 the total number of lexical items is also shown in millions of words (Mw), indicating how frequent are the most frequent pronunciation pairs. Another experiment was conducted by adding the most frequent pronunciation triples (effectively the same as adding word trigrams), and despite their low frequency of occurrence, it still resulted in a modest performance improvement, as shown in Fig 5

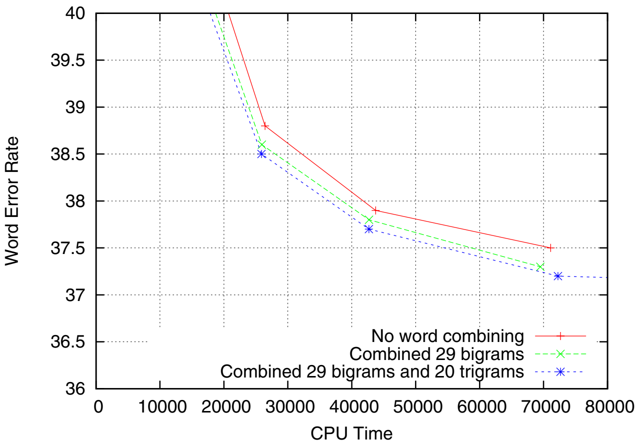

Figure 5: The effect of converting word pronunciation triples into single lexical items for language modeling

The final question is how does the merging of words or pronunciation variants into longer span lexical entries, and consequent expansion of the n-gram span, compare to simply using higher order n-gram models. In all the experiments above a simple trigram with a shrink of zero was used. We compare that with quadrigram and quintagram models with different shrink values. The results are shown in Fig 6.

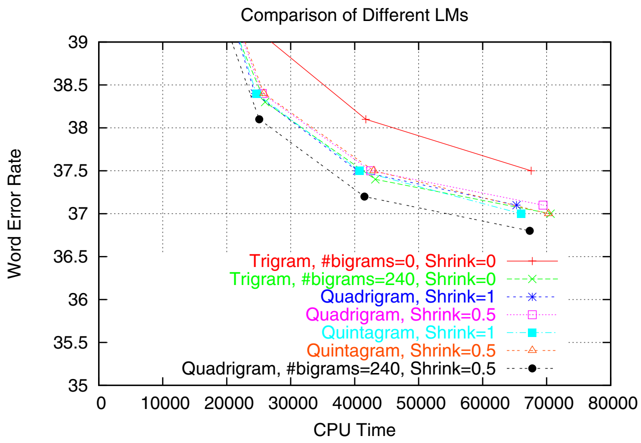

Figure 6: Comparison between higher order n-gram models and merging multiple lexical items into longer span single lexical items for language modeling

Although Fig 6 is too crowded to compare the different language models it is clear that either higher order n-gram models or the trigram language model with 240 pronunciation merges perform equally well and all perform significantly better than the baseline trigram LM, and that the quadrigram model with 240 pronunciation merges performs even better. Although the Switchboard acoustic training database is considered very large, its transcriptions make up but a small language modeling database, and that might be the reason why the differences between the different higher order n-gram language models is so negligible.

## 5. Conclusions

The first consequence is that lexical items become pronunciation dependent and that pronunciation modeling becomes the provenance of the language model and not the dictionary any more. This allows for better modeling of coarticulation and part-of-speech differences in pronunciation, resulting in a modest gain using a modest number of thus modeled pronunciations, on a total of 17 words in a vocabulary of over 60,000 words. Such transfer from the dictionary to the language model can be performed for all of the words in the dictionary, with expectations of similar improvements, given adequate training data.

The issue of sufficient training data could be overcome by creating speaker-specific pronunciation models, and using them to transform lexical data into many speaker dependent phonemic transcription based datasets by using the pronunciation models in the generative mode.

Merging words into longer lexical items (phrases), has the beneficial consequence of changing the behavior of lower order n-gram models to be more like one or two orders higher n-gram models (trigram behaves like quadrigram or quintagram models), which might make a significant difference for some recognition decoder architectures. This improvement appears to hold even when the amount of the language modeling training data is significantly reduced to match the amount of the acoustic training data. How to implement language model-based pronunciation model and extend the training to more text-only data is, however, yet to be explored.

## 6. References

[1] Michael Finke and Alex Waibel, “Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition,” in *Proceedings of Eurospeech 97, Rhodos, Greece*, 1997.

[2] G. Saon and M. Padmanabhan, “Data-driven approach to designing compound words for continuous speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 327–332, 2001.

[3] X. Aubert, P. Beyerlein, R. Haeb-Umbach, M. Ullrich, A. Wendemuth, D. Klakow and P. Wilcox, “Language model investigations related to broadcast news,” in *Proc. DARPA Broadcast News and Transcription Workshop, Lansdowne, VA*, Feb. 1998.

[4] L. Mangu, D. Povey, G. Saon, H. Soltau, B. Kingsbury and G. Zweig, “The ibm 2004 conversational telephony system for rich transcription,” in *Proceedings of ICASSP’05, Philadelphia*, 2005.