WHAT KIND OF PRONUNCIATION VARIATION IS HARD FOR TRIPHONES TO MODEL?

Dan Jurafsky, Wayne Ward, Zhang Jianping, Keith Herold, Yu Xiuyang, and Zhang Sen

Center for Spoken Language Research
University of Colorado, Boulder

ABSTRACT

In order to help understand why gains in pronunciation modeling have proven so elusive, we investigated which kinds of pronunciation variation are well captured by triphone models, and which are not. We do this by examining the change in behavior of a recognizer as it receives further triphone training. We show that many of the kinds of variation which previous pronunciation models attempt to capture, including phone substitution or phone reduction, are in fact already well captured by triphones. Our analysis suggests new areas where future pronunciation models should focus, including syllable deletion.

1. INTRODUCTION

Many studies of human-to-human speech have shown that pronunciation variation is a key factor contributing to the high error rates of current recognizers. For example [1] showed that Switchboard word error decreased from 40% to 8% if the dictionary pronunciation matched the actual pronunciation.

While the need for better pronunciation modeling is widely acknowledged, and many previous researchers have attempted to build models of the lexicon which capture this variation, very few of these previous models have had significant success in reducing error rates. For example one solution that has often been implemented is to build an 'allophone network' [2], as shown in Figure 1.

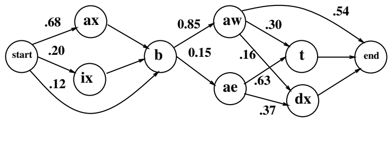

Fig. 1. A multiple-pronunciation network for *about*

But our own research, and that of others, has shown that these allophone networks do not perform well. Both [3] and [1] showed that blindly adding multiple pronunciations to a dictionary, even those shown to improve the performance of a single utterance, substantially increased the word-error of a Switchboard recognizer. We and others have shown similar problems with this method on WSJ [4], and Switchboard [5].

The problem with adding large numbers of pronunciation is that the benefit of finding the correct pronunciation of some words is offset by errors caused by increased ambiguity among other words. The solution to this problem requires dynamically adjusting the lexicon to contextual factors. For example, many factors dynamically effect pronunciation variation, including the surrounding phones, the prosodic/accent context, the identity and probability of neighboring words, and the presence of disfluencies or silence near the target word [6, 7, 8].

While these factors play a clear role in pronunciation variation, many of them may already be well-captured by current lexicons because of the power of triphone models to capture contextual effects. For example, triphones almost certainly already do a good job of modeling phone changes that are caused by neighboring phones. This means that future research should probably not focus on the kinds of phonetic context that have been studied in some previous work [2, 9, 4, 10].

Our goal in this paper is to perform a background analysis of which kinds of pronunciation variation are well captured by triphone models, and which are not. We do this by examining the phonetic factors which differentiate sentences which are successfully modeled by a triphone recognizer from sentences which are not successfully modeled.

2. METHODOLOGY

Our experiments were run using the CMU Sphinx-II speech recognition system, a speaker-independent recognizer with Viterbi decoding, semi-continuous acoustic models and trigram language models. We used a very preliminary prototype Switchboard system, which has a 23,000 word vocabulary, and was built by taking acoustic models bootstrapped by CMU on their Communicator system, and then retrained on 69,607 utterances from the version of the Switchboard training data released by Mississippi State, together with the

Thanks to the National Science Foundation for partial support of this work via IIS-9978025.

Mississippi State lexicon and with a language model generously supplied by Andreas Stolcke of SRI. Our prototype system currently has a word error rate of 64.7%.

We used a special test set drawn from the 3.5-hour portion of Switchboard that was phonetically hand-labeled at ICSI [11]. Berkeley students were given wavefiles together with their word transcription, and a rough automatic phonetic transcription. They then corrected this rough phonetic transcription, using an augmented version of the ARPAbet. The hand-labeled data consists of 5132 utterances. The version of the phonetic transcriptions we used had been force-aligned to the Switchboard word transcriptions by Eric Fosler-Lussier.

We selected 2780 of these utterances as a test set, and then used Sphinx-II in forced alignment mode, to time-align the word transcriptions to the speech files, resulting in an acoustic score for each utterances in the test set. We performed this forced alignment twice; once for the 'canonical' Mississippi State Switchboard lexicon, and once for a special 'cheating', or 'surface' lexicon.

The 'surface' lexicon was actually a collection of 2780 lexicons, consisting of a separate cheating lexicon for each sentence in the test set. In each cheating lexicon, the pronunciation for each word was taken from its ICSI hand-labeled pronunciation, converted to triphones. When performing forced alignment, we switched lexicons dynamically, using the matching lexicon for the test sentence. For example for one test sentence whose transcription was "That is right", we had the following two lexicons:

Word	Canonical Lexicon	Surface Lexicon
that	dh ae t	dh ae
is	ih z	s
right	r ay t	r ay

Each of the 2780 utterances in the test set received two forced-alignment scores; one from the surface (cheating) lexicon, and one from the canonical (Mississippi State) lexicon. We then examined which sentences in the test set received a higher force-alignment score from which lexicon. For example, if a certain class of sentences receive a higher score from a canonical lexicon, this tells us something about what kinds of variation a canonical lexicon already captures.

But we are more interested in knowing what kinds of variation could be further accounted for as a triphone system based around a canonical lexicon received more training data. That is, we assumed that as a triphone system received more training, that the triphones would do a better and better job of capturing variation in the test set. Thus we examined how the forced-alignment scores of sentences changed over time. We looked at two stages of our embedded training. First, we considered at our beginning system, whose triphone acoustic models were trained only on CMU Communicator data, with no training at all on Switchboard. Then we considered our system after it had been trained on the Switchboard dataset. We examined those sentences that had a higher acoustic score with the surface lexicon in the initial Communicator system, but switched to having a higher acoustic score with the canonical lexicon after training on Switchboard.

In other words, we looked at those utterances whose score with the canonical lexicon improved after more exposure to data. We call these 807 utterances SC, because they began with higher scores from the S (surface) lexicon, but ended up with higher scores from the C (canonical) lexicon). We compare these 807 utterances with the 1047 utterances which began with a higher score from the surface lexicon, and remained with a higher score from the surface lexicon. We call these utterances SS. We also examined the 750 CC utterances, those which always had a higher score from the canonical lexicon.

In the rest of this paper, then, we study the kinds of variation which cause certain sentences (in SC) to improve with a canonical lexicon, as their triphones see more data, while other sentences (SS) do not improve their forced alignment score with the canonical lexicon.

We looked at two kinds of factors. First was the exposure to more triphones in training; presumably the canonical triphone lexicon improves on sentences in the test set whose triphones were seen more in the training set. The second class of factors was the kinds of phonetic variation in the sentence; we investigated whether certain kinds of phonetic variation (such as syllable deletion) are difficult for triphones to model. For each factor, we compare their effects on sentences in three subsets of our test set.

3. AMOUNT OF TRIPHONE TRAINING

Our first factor was the amount of triphone training data for the triphones in the test sentences. We hypothesized that sentences which had higher scores with the canonical lexicon might consist of triphones that occurred more often in training. That is, since our embedded training regime used the canonical lexicon, it is possible that sentences in the test set which better matched the triphone characteristics of the training set would perform better with a canonical lexicon. The table below shows the counts for triphones from the three subcategories of our testset:

Set	% of test triphones types in training	
SC	5359/5689	94%
CC	3997/4233	94%
SS	6074/6407	95%

The percentage of triphones which occured in both training and test sets was very high, and did not differ across sets. Thus the percentage of triphones types which had received some training did not play a role in whether a test sentence was better modeled by the canonical or surface lexicons.

We then investigated whether some triphones might have received more training samples. For each triphone which occurred in each of the three training sets, we computed the average number of times that triphone was seen in the test set.

Set	Average # of times test set triphones occurred in training set
SC	142
CC	179
SS	128

We found a significant difference in these averages. SS sentences had the least amount of training data per triphone, SC sentences had somewhat more, and CC had the most. Thus the amount of training data each triphone receives does play an important role in whether a canonical lexicon is able to model pronunciation variation.

4. PHONETIC VARIATION PER SENTENCE

We next investigated the amount of phonetic variation in each sentence. Our hypothesis was that sentences which had a higher acoustic score with the canonical lexicon (CC sentences) would have less phonetic variation than SS or SC sentences. We defined phonetic variation as any difference between the canonical (dictionary) phone sequence and the surface (hand-labelled) phone sequence. For example, the following sentence had 7 changed phones (4 substitutions and 3 deletions), and 3 unchanged phones:

WORD:	you		know		and			one		
CAN:	y	uw	n	ow	ae	n	d	w	ah	n
SURF:	-	ih	-	uh	ah	nx	–	w	ah	n

The following table shows the percentage of phones which changed for each of the 3 categories:

Set	% Phones Changed
CC	27%
SS	34%
SC	29%

Indeed, the number of changed phones per sentence (or, equivalently, the total percentage of phones in each set which change) does distinguish between sentences which are modeled well by the surface lexicon (SS) (34% of the phones change), and those which switched to the canonical lexicon after more training (SC) (29% of the phones change). This suggests that sentences with less phonetic variation, in which pronunciations are closer to a canonical pronunciation, are better modeled by a canonical lexicon.

This result confirms earlier studies that show that the significant phonetic variation in Switchboard does cause problems for canonical lexicons. It also acts as an important test of our methodology. Since our methodology is shown to be sensitive to the effect of phonetic variation, we can now test to see which particular kinds of phonetic variation cannot be easily handled by triphone training.

5. SYLLABLE DELETION

The previous section showed that sentences with fewer phonetic changes are more easily modeled by the triphones as they see more data. This suggests that triphones, while they are able to model some amount of phonetic variability, aren't able to model all of it.

In the next three sections we study 3 kinds of particular phonetic variation to see if any of them are particularly easy or difficult for triphones to capture. The three kinds of variation are syllable deletion, vowel reduction, and other (non-reduction) cases of phone substitution.

Syllable deletions are cases in which the canonical pronunciation has an entire syllable which is missing in the surface pronunciation. For example, the hand-transcription of the word *variety* below shows that it has only 2 syllables ([v r ay] and [dx iy]), while the canonical dictionary entry has 4 syllables ([v ax], [r ay], [ih], and [t iy]):

WORD:	variety						
CAN:	v	ax	r	ay	ih	t	iy
SURF:	v	-	r	ay	-	dx	iy

The syllable deletion rates were quite different for the three sets:

Set	% Syllables Deleted
CC	2.6%
SS	3.3%
SC	1.8%

As the table above shows, the syllable deletion rates for the SC category were not only lower than the SS category, but even lower than the CC category. That is, the sentences which matched the canonical lexicon after received extra triphone training had a particularly low level of syllable deletion. This suggests that syllable deletion is not well modeled by simply having more training data for the triphones.

6. REDUCED VOWELS

The next subcategory of phonetic change we examined was vowel reduction. Vowel reduction is the process in which many vowels in unstressed syllables reduce to a shorter, more neutral vowel like [ax], [axr], or [ix]. Vowel reduction is an important category of phonetic change to investigate because it is strongly linked with prosodic effects; vowels which are *stressed* or *accented* are not reduced. Whether a syllable received lexical stress or not is already modeled in the lexicon. But accent is a more complicated phenomenon, whose location is much more dependent on semantic and syntactic context.

Thus if the triphone training of the canonical lexicon allows it to handle cases of reduced vowels, this implies that stress or accent are modeled sufficiently by current triphone systems. Again, we checked the rates of each subcategory of sentences:

Set	% of Vowels Reduced
CC	6.3%
SS	9.4%
SC	10.0%

As the table above shows, the SC sentences do *not* show less reduction than the SS sentences. That is, the sentences which switched to the canonical lexicon did not have less vowel reduction than the sentences which were still better modeled by the surface lexicon. This suggest that triphones do learn to capture vowel reduction, and that it may not be a kind of reduction that we should focus on.

7. PHONE SUBSTITUTION

The final category of phonetic variation that we investigated was phone substitution. By phone substitution, we mean any case where the surface (hand-labeled) phones are distinct from the canonical (dictionary) phones except for cases of reduction. This kind of variation is usually caused by phone coarticulation, and thus is the kind of local phonetically-induced variation that we expected triphones would do a good job of modeling.

Set	% of Phones Substituted
CC	7.7%
SS	7.0%
SC	7.2%

The table above shows that, as expected, the SC sentences did not, in general, have less phone substitution than the SS sentences (if anything, they had more). This means that the triphones are in fact able to model this kind of phonetic variation

8. CONCLUSIONS

We investigated three factors related to phonetic variation to study which of them might cause problems for a triphone-based recognizer. Our method was to examine how the addition of training data enabled a canonical lexicon to successful model more sentences. As expected, one of the ways this data helped was by increasing the triphone training; but we showed that this training data played a role not by increasing the number of types of triphones, but by increasing the average number of training instances for each triphone.

We also showed that the more phonetic variation a sentence had, the less well it was modeled by a canonical lexicon. But not every kind of phonetic variation was problematic; the canonical lexicon did not have problems in modeling phone substitution, nor in modeling vowel reduction. This last fact suggests that prosodic (pitch accent) factors may not be a large cause of problems in pronunciation models. Rather, of the types of phonetic variability we looked at, the the main factor which caused sentences to be poorly modeled by a canonical lexicon was syllable deletion.

We are currently investigating these factors in more detail and exploring more sophisticated statistical measures of the differences.

9. REFERENCES

[1] Don McAllaster, Larry Gillick, Francesco Scattone, and Mike Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," in *ICSLP-98*, Sydney, 1998, vol. 5, pp. 1847–1850.

[2] Michael H. Cohen, *Phonological Structures for Speech Recognition*, Ph.D. thesis, University of California, Berkeley, 1989.

[3] Murat Saraclar, "Automatic learning of a model for word pronunciations: Status report," in *Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation*, Baltimore, MD, 1997.

[4] Gary Tajchman, Eric Fosler, and Daniel Jurafsky, "Building multiple pronunciation models for novel words using exploratory computational phonology," in *EUROSPEECH-95*, 1995, pp. 2247–2250.

[5] Michael D. Riley, William Byrne, Michael Finke, Sanjeev Khudanpur, Andrei Ljolje, John McDonough, Harriet Nock, Murat Saraclar, Chuck Wooters, and George Zavaliagkos, "Stochastic pronunciation modeling from hand-labelled phonetic corpora," *Speech Communication*, vol. 29, pp. 209–224, 1999.

[6] Daniel Jurafsky, Alan Bell, Eric Fosler-Lussier, Cynthia Girand, and William D. Raymond, "Reduction of English function words in Switchboard," in *ICSLP-98*, Sydney, 1998, vol. 7, pp. 3111–3114.

[7] Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond, "Probabilistic relations between words: Evidence from reduction in lexical production," in *Frequency and the emergence of linguistic structure*, Joan Bybee and Paul Hopper, Eds. Benjamins, Amsterdam, 2000, To appear.

[8] Eric Fosler-Lussier, *Dynamic Pronunciation Models for Automatic Speech Recognition*, Ph.D. thesis, University of California, Berkeley, 1999, Reprinted as ICSI technical report TR-99-015.

[9] Michael D. Riley, "A statistical model for generating pronunciation networks," in *IEEE ICASSP-91*. IEEE, 1991, pp. 737–740.

[10] Michael Finke and Alex Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *EUROSPEECH-97*, 1997.

[11] Steven Greenberg, Dan Ellis, and Joy Hollenback, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *ICSLP-96*, Philadelphia, PA, 1996, pp. S24–27.