To appear in ESCA Tutorial and Research Workshop on ``Modeling pronunciation variation for automatic speech recognition'', Kerkrade, Netherlands, May 1998

EFFECTS OF SPEAKING RATE AND WORD FREQUENCY ON CONVERSATIONAL PRONUNCIATIONS

Eric Fosler-Lussier *Nelson Morgan*

International Computer Science Institute and University of California, Berkeley
1947 Center Street, Suite 600, Berkeley, CA 94704, USA
{fosler,morgan}@icsi.berkeley.edu

ABSTRACT

The possible set of pronunciations in continuous speech corpora change dynamically with many factors. Two variables, speaking rate and word predictability, seemed to be promising candidates for integration into dynamic ASR pronunciation models; however, our initial efforts to incorporate these factors into phone-level decision tree models met with limited success. In this paper, we confirm the intuition that these factors have an effect on ASR systems, and analyze the relationship between these factors and pronunciations in order to shed light on why the decision trees models failed. We present a statistical exploration of the effects of these factors at the word, syllable, and phone level in the Switchboard corpus. We show that both increased speaking rate and word likelihood can induce a significant shift in probabilities of the pronunciations of frequent words. Using these data, we hypothesize reasons for the difficulty in incorporating these dynamic measures into phone-level decision trees.

1. INTRODUCTION

One of the foremost issues in pronunciation modeling for ASR is how to increase the coverage of pronunciations without increasing acoustic confusability. We argue that a model of word pronunciations should be dynamic; the changes in ASR models should be influenced by contextual factors that determine a probability distribution over pronunciations. The need to constrain the number of pronunciation alternatives was shown in a diagnostic study [1] where Switchboard lattices (baseline: 46% word error) were rescored with a dictionary which was augmented on a per-utterance basis with the "correct" pronunciations in the utterance (as determined by phone constraint decoding), reducing word error to 26%. However, when the dictionary was augmented with the "correct" pronunciations for all of the test set, the word error rate degraded to 38%; in other words, the benefit of having the correct pronunciation was often offset by the presence of unnecessary competing pronunciations. This showed the importance of dynamically selecting appropriate pronunciations.

We have appealed to linguistic studies to find conditions under which one should dynamically change the selection of pronunciations available to the recognizer. For instance, linguists have recognized that word frequency affects the perception and production of phones; speech researchers have used the concept of *function words* as an approximation to this factor, although we argue that this binary decision should be smoothed. Unusually slow or fast speaking rate has also been shown to have an adverse effect on recognizers, and linguists have also found that rate of speech variations can affect phone perception and production.

1.1. Integration into pronunciation model

We have been using decision-tree pronunciation models derived from work at the 1996 Johns Hopkins LVCSR Summer Research Workshop [2, 3]. These decision trees determine mappings from baseform phonemes to realized phones using information-theoretic clustering of surrounding phonemic contexts, similar to [4, 5].

Our goal was to place variables corresponding to speaking rate and word predictability directly into the pronunciation model as a feature for determining splits in the decision tree growing process. Therefore, we marked every baseform/realized phone pair with the following candidate features:

Unigram probability
The probability of the word in which the phone occurs, determined from frequency counts of the word in the reference transcription of the corpus.

Trigram probability
The probability of the word dependent on the previous two words.

Transcribed syllable rate
The interpausal rate determined from hand transcriptions. For interactive systems, this particular measure is not determinable at recognition time. As speaking rate estimators can sometimes be unreliable, however, we wanted to see what the best-case scenario might be.

Mrate
A signal-processing speaking rate measure that we have been developing at ICSI [6]. We defer description of this measure and its relationship to our studies until section 4.

Since we were still developing our baseline Switchboard ASR system, we decided to train and test the decision trees on the portion of the Switchboard corpus that was hand transcribed by linguists at ICSI [7]. However, we found that the trees on the whole failed to use our new factors as splitting criteria. This was surprising, particularly as other researchers have found that an earlier measure of ours called *enrate* (which did not correlate as well with transcribed rate as our current version of *mrate*) was a useful feature in determining a "hidden mode," which they then used to determine pronunciation probabilities (see [8] for details).

Thus, we decided to check our assumptions, and conduct an investigation of the relationship between the above factors and pronunciation errors. We set out to answer the following questions:

- What is the effect of non-canonical pronunciations on recognizer performance for the Switchboard corpus?

- Do the factors we have chosen (speaking rate and word predictability) have an effect on recognizer performance for Switchboard?
- Taking these static and dynamic factors into account, can we find systematic trends in pronunciations? If so, why do they not fit into the phone decision tree paradigm?

1.2. Materials

The Switchboard data used are approximately four hours of phonetically hand transcribed utterances with syllable boundary markers, provided by ICSI for the Johns Hopkins Summer Research Workshop series [7]. We generated a mapping from syllabified Pronlex dictionary baseforms to these hand transcriptions using a dynamic programming technique which uses phonetic features to calculate a distance metric between phones, as in [2]. In the cases where multiple pronunciations existed in the dictionary,[1] the closest baseform (in terms of the distance metric) to the realization was used. Pronunciation maps were generated for every baseform word, syllable and phone.

We also used the syllabic boundaries marked in the hand transcriptions to calculate the interpausal syllabic rate, and calculated the *mrate* for the same region. Using the word-level transcriptions, we also computed the unigram and trigram probability of each word, using a back-off grammar trained from Switchboard data.

We also used the 1996 JHU workshop HTK recognizer trained with the same Pronlex dictionary (hereafter referred to as the WS96 recognizer) to provide recognition hypotheses for error analysis. While analyzing only one recognizer can certainly highlight the idiosyncrasies of that system, the speech community has seen previously [9, 10] that speaking rate affected the output of all WSJ systems in a 1993 evaluation, so we have hope that these studies may be applicable to more than just the WS96 system.

1.3. Error metrics

One of the difficulties we encountered when embarking upon this study was finding a good way to characterize the behavior of pronunciations as a function of the factors we would like to study. We have experimented with a number of metrics; each has some advantages and disadvantages.

Probability of a single pronunciation

We track the probability of canonical (dictionary) pronunciations, which is particularly useful when estimating how well the pronunciations given to the baseline recognizer match the transcribed data. In later experiments, we also track the behavior of the single most likely pronunciation, under the assumption that a system which performs automatic baseform learning would also have that pronunciation in its dictionary. Unfortunately, analysis becomes complex when tracking more than just a few pronunciations.

Entropy

This is a traditional measure for pronunciation learning systems [5], and is a good measure of the spread of pronunciations in a training set. It becomes unwieldy, however, if one tries to use it to predict how well models will perform on a particular test set (i.e. relative entropy), as pronunciation models are typically pruned to some cutoff (assigning zero probability to some test events), which causes relative entropy to approach infinity.

Phonetic distance score

We also developed a metric which was smoother than the hard binary decision of whether a pronunciation was canonical or not by using the phonetic feature distance between the two pronunciations as utilized in the dynamic programming technique described above. We interpret this distance as a measure of how far the pronunciation has deviated from expected. This procedure can also be extended to give a smoothed score using a particular pronunciation model; the distance between each baseform pronunciation in the model and the target phone sequence is weighted by the probability of the baseform pronunciation. It is difficult, however, to give a statistical or information-theoretic interpretation to this metric.

2. RECOGNIZER ERROR ANALYSES

2.1. Previous work

As one moves to spontaneous speech corpora, such as Switchboard, the variability in word pronunciations increases. An immediate observation is that typical conversational speech is faster than typical read speech. The differences between these two speaking modes are more complex, however. Bernstein *et al.* [11] show that pronunciations in spontaneous speech are different from fast read speech. Although the number of words per second in spontaneous speech is similar to fast reading for most speakers, the number of phones per second for spontaneous speech is more like that for normal reading. This suggests that speakers tend to delete phones rather than reduce durations during spontaneous speech— many pronunciation variations and a high phone deletion rate can be expected.

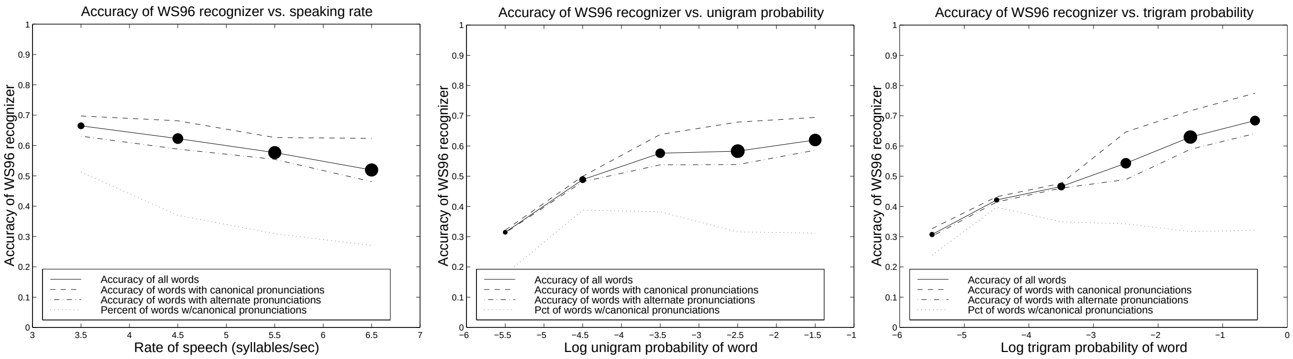
As we have reported previously [3], in an initial study at WS96 we attempted to characterize this variability for the Switchboard corpus. A comparison of the phonetic transcription of the development test set with citation-form pronunciations of the transcribed words revealed that 12.5% of the phonemes in the "standard" pronunciation are deleted; substitutions and insertions of phones also changed pronunciations so that only 67% of the phones in the pronunciations obtained from the dictionary were identified as correct by the hand transcriptions.

2.2. Pronunciation models as a cause of error

In an elaboration of this study, we have tried to characterize the effects of these phone-level statistics on word-level pronunciations. We found that while 67% of phones retained "canonical" form in spontaneous speech, only 33% of word pronunciations found in the Switchboard development test set (using ICSI hand transcriptions) were found in the Pronlex dictionary. Thus, the phone transformations we have observed are not concentrated in a few words, but rather are spread throughout the corpus.

What remains to be shown is that these pronunciation errors have an effect on our recognizers. Intuitively, one would believe that recognizers would fail miserably if 67% of word pronunciations are not in the dictionary. However, recognizers are not necessarily learning the linguistic ideas with which they are being seeded— for instance, the distributions of acoustic models may be less sharp to compensate for the pronunciation deviance. We analyzed the errors made by the WS96 HTK recognizer, comparing recognizer results in conditions where linguists determined that pronunciations were canonical versus conditions where alternative pronunciations were used by the speaker.

[1] The average number of pronunciations per word was 1.07.



a. Speaking rate b. Unigram probability c. Trigram probability

Figure 1. Accuracy of WS96 Switchboard recognizer dependent on several factors. In these graphs, the solid line indicates the overall accuracy trend as each factor changes. The size of the dots indicate the proportion of data found in that particular histogram bin. The dashed and dot-dashed lines indicate recognizer scores when the hand transcription of the word did or didn't match the recognizer pronunciation, respectively. The dotted line indicates the percentage of words that had canonical pronunciations for that histogram bin.

For this study, we examined 439 sentences from the Switchboard development test set which were also phonetically transcribed. Each word in the test set transcriptions was annotated with whether it was correctly recognized, substituted, or deleted, and whether the transcribers observed a canonical or alternative pronunciation, as defined by the Pronlex dictionary (i.e. the recognizer lexicon). Recognizer insertions were disregarded; although pronunciations certainly have an effect on insertions, it is difficult to mark them as canonical or alternative pronunciations compared to the hand transcriptions, as the speakers did not actually utter the inserted words.

	Overall	Canon. Pron.	Alt. Pron.
% correct	57.4	65.0	53.9
% deleted	12.0	8.1	13.9
% substituted	30.5	26.1	32.2
# of words	4085	1337	2748

Table 1. Breakdown of word substitutions and deletions with WS96 Switchboard Recognizer for Canonical and Alternative Pronunciations.

There is a significant improvement in recognizer performance when the linguists' transcription matches the dictionary pronunciation (Table 1). There is a large (70% relative) increase in the recognizer word deletion rate for words with alternative pronunciations, as well as a significant increase in recognizer substitutions.

It is difficult to separate the effects of different factors on word error rates; for instance, a mispronounced word can result in a substitution, causing a language model error for the following word. However, these numbers suggest that there is a real effect of pronunciations on word error. The numbers also show that solving "the pronunciation problem" will not necessarily solve the speech recognition problem, but will contribute towards reducing error rates.

2.3. Relationships between static and dynamic factors and recognizer error

Although we have seen that there is an effect of the pronunciation model on recognizer errors, it is not clear what the relationship is between recognizer errors and speaking rate, unigram probability, and trigram probability. Thus, we labeled every word in the above dev-test set with the transcribed syllable rate, unigram probability, and trigram probability of the word. We then partitioned the words into histogram bins, and determined the recognizer accuracy for each bin, as seen in Figure 1. Included on each graph is the percentage of words which had canonical pronunciations, and scores for words with or without canonical pronunciations, as marked by the transcribers.

In Figure 1a, we see that there is a 14% drop in recognizer accuracy as the speaking rate moves from very slow to very fast speech. This is due mainly to the poorer performance on words pronounced non-canonically, of which there are more in fast-speech conditions. Note that for this test set, the number of fast utterances is non-trivial (35% of the data); thus, there is a real and significant effect from fast speech for this set. One additional note: as in Section 2.2., these graphs do not include insertions. Since rate is calculated over an interpausal region, we can calculate insertion rates for each speaking rate, however. Insertions, as expected, do decrease from 7.7% to 2.3% as the speaking rate increases; when these are taken into account in the word error rate, the difference in errors between slow and fast speech is still roughly 9%.

In the case of language model probabilities (Figures 1b and 1c), we do see that recognizer performance improves as words become more likely. This is not surprising, as language models in the recognizer tend to favor more likely words during recognition. As one would expect in this case, the trigram graph has a larger spread (from 30% to 69%) than the unigram (31% to 61%), as the recognizer (which utilizes a bigram grammar) takes into account more information than unigram probabilities. What is interesting here is that, even though the recognition rate increases as words become more likely, the percentage of words with canonical pronunciations decreases.[2] For higher probability words (e.g. log(trigram)>-3), there is a gap in performance for canonical versus non-canonical pronunciations. On the other hand, for low probability words the language model in the recognizer dominates the error, and it does not matter as much whether the pronunciation was canonical or not.

[2] It is unclear why the probability of canonical pronunciations drops for low probability words. These words do tend to be longer on average, so *a priori* there is an increased chance of a single phone changing in a word. This class makes up 5% of the words in the test set.

Therefore, it seems that there is a relationship between language model probabilities and pronunciations, although one has to be careful to tease the effects apart from the influence of the language model itself in the recognizer.

3. RELATIONSHIPS BETWEEN FACTORS AND PRONUNCIATION ERRORS

We have shown above that there exist correlations between recognition errors and pronunciation errors, as well as recognition errors and the factors of speaking rate and word predictability. In the next few studies, we attempt to understand more directly the correlation between pronunciations and these factors. We begin with some word and phone-level studies, although for the amount of data we had, the most significant findings are at the syllabic level. Data from these studies provide us with some hypotheses about our decision tree problem.

3.1. Word-level experiments

In a pilot experiment, we extracted the word-pronunciation pairs for the 117 most frequent words from a 2-hour subset of the transcriptions. For each word, we divided the pronunciation population in half based on speaking rate and compared the probability of both the most likely transcribed pronunciation and canonical pronunciation (as given in the Pronlex dictionary) between partitions. A sample output of the comparison for "been" is shown in Table 2.

Pronunciation	Low Syllable Rate	High Syllable Rate
Canonical	0.6087 b ih n	0.3636 b ih n
Alternative	0.3913 others	0.6364 others

Table 2. Distribution of the pronunciation probabilities for 45 realizations of the word "been."

We also partitioned the data based on trigram scores. We found that there was a significant ($p<0.05$) shift in the probability of the canonical pronunciation for 30% of the words due to rate differences; for trigram probabilities, 18% of words had a significant shift. Similar results were seen with the most likely pronunciations.

3.2. Phone-level experiments

In the word-level study, we had few words with enough data for statistical analysis. Therefore, we examined how dictionary phonemes are realized as phones in the hand transcriptions. For each dictionary phoneme, we extracted the corresponding hand transcribed phones, along with the applicable speaking rate. We then observed the overall trends for all of the phones.

As seen in Figure 2, we found that from very slow to very fast speech, the phoneme deletion rate rises from 9.3% to 13.6%; the phone substitution rate also changes significantly ($p<0.05$), rising from 16.9% to 24.2%. We also found that as speaking rate increases, the entropy of the distribution of pronunciations also increases.

In the next step of this study, we wanted to examine the effects of rate for each phone. However, we discovered that this was a futile effort if we did not take into account the phonetic context of the phoneme; since we had difficulty building decision trees that incorporated context, stress, *and* rate, we decided to work with larger linguistic units.

3.3. Syllable-level experiments

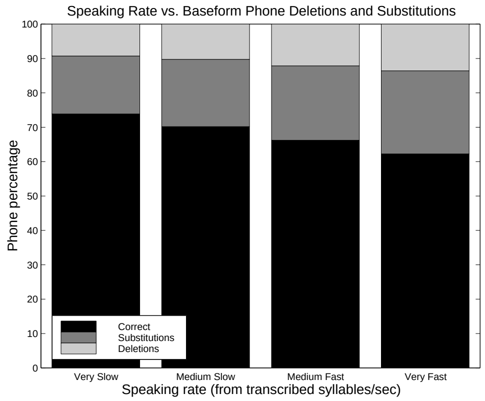


Figure 2. Phone-level statistics for effects of speaking rate on pronunciations.

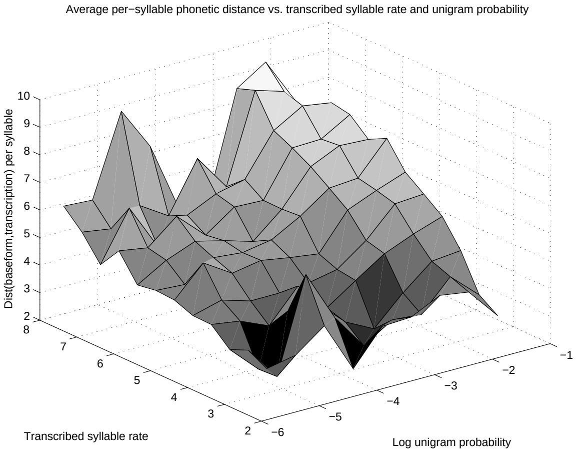


Figure 3. Average syllable distance from baseform as a function of speaking rate and unigram probability. For low unigram probability and very high or low speaking rates, the number of samples is low; the peaks in the graph for these extremes are probably statistical noise.

As a middle ground between the word level and phone level, we decided to examine pronunciations on the syllable level. This allowed us to cluster some of the data from the word-level experiments, but also gave us more context than on the phone level. In addition, it has been suggested that pronunciation phenomena are more often affected by syllabically internal rather than external context [12].

As can be seen from Figure 3, there is a connection between unigram probability, speaking rate, and the average distance for each syllable from the Pronlex baseforms: in less frequent words there is some increase in mean distance as rate increases, but for syllables occurring in more frequent words, the rate effect is more marked. There is a relatively strong interaction between these two variables, which may be a clue for our decision tree problem— as trees partition the training data as they are built, we may have lacked sufficient data to get good estimates of the joint distribution when taking phonemic context, rate, and word predictability into account.

When we looked at the probability of canonical pronunciations for this same data, we did not see this sharp effect, which puzzled us greatly. The first key to solving the puzzle was to notice that the probability of canonical pronunciations *did* change as a function of rate when we took lexi-

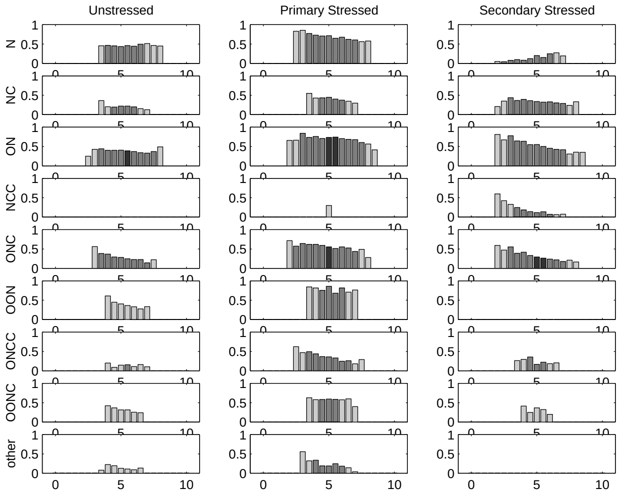


Figure 4. Probability of canonical pronunciations for different speaking rates (in syllables/second), partitioned by lexical stress and syllabic structure. Light grey bars have 20-100 samples, medium grey 100-1000, dark grey >1000. O=onset, N=nucleus, C=coda.

cal stress and syllabic structure into account. In Figure 4, we see that for some syllable types, (e.g. primary stressed nucleus-only), rate has a strong effect, but for others the effect is negligible. For one case (secondary stressed nucleus-only), a surprising reverse effect occurs— the probability of canonical pronunciation increases as rate increases. These data also confirm the commonly held intuition that syllabic stress is an important factor in pronunciation models, as shown by other researchers [13, 2].

We then chose to examine the 200 most frequent syllables in the Switchboard corpus, which provide 77% syllable coverage of the 4-hour transcription set, and 75% of the corpus at large. We clustered the data for each syllable into speaking rate bins, and determined the probability of the canonical and most likely pronunciations[3] for each syllable as a function of the rate bin. We reclustered data in a similar fashion using trigram probability as the clustering criterion.

	# of syls w/ significant differences			
Clustering on:	Canon.	Most Likely	Either	Both
Speaking rate	85	81	95	71
Trigram prob	64	59	70	53

Table 3. Number of syllables (out of 200) with significant ($p<0.05$) differences in pronunciation probabilities for the extremes of speaking rate and trigram probability.

As we see in Table 3, 95 of 200 syllables showed a significant change in the probability of either the canonical or the most-likely pronunciation. In general, the probabilities shift smoothly as a function of rate (as in Figure 5a) or trigram probability (Figure 5c); this may be another reason why it was difficult for the decision tree algorithm to make hard data splits based on speaking rate.

The major characteristic that describes the class of syllables with significant rate shifts is that these syllables are often more frequent. The mean unigram log probability for these syllables is -2.33; for non-affected syllables the mean unigram log probability is -3.03.

For some syllables (Figure 5b), there is a tradeoff between the most-likely and canonical pronunciations as a function of rate. However, for faster examples, the sum of the probabilities is lower than for slower examples; the rest of the probability mass shifts to other pronunciations. This is also shown by the fact that the entropy of the distribution of syllabic pronunciations increases as a function of increased rate for roughly the same set of syllables with significant pronunciation differences.

To this point, we have been treating unigram and trigram scores as roughly equivalent. For the vast majority of cases, it appears that using trigram scores provides little extra modeling power, as the trigram is often correlated with the unigram. However, for a small number of frequent syllables it distinctly helps to have the trigram score. For example, in Figure 5c, the syllable *ih_f*, which corresponds only to the word *if* in our training set (i.e. all examples share the same unigram probability), is significantly reduced in very likely word sequences.

A critical observation is that the effects of these variables, even for similar syllables, can vary widely. For example, the central vowel in the syllables *th_ih_ng, th_ih_ng_k,* and *th_ih_ng_z* all share similar acoustic contexts (at least in a system which only looks one phone to the left and right). However, the vowels of these syllables change very differently under different speaking rates— in *th_ih_ng,* the vowel shows a significant shift towards *ix* or *iy,* but for the other two syllables, it remains relatively constant with respect to rate. Again, this may show a lack of data— in order to capture these effects, we may have needed more phonetic context in our decision trees; however, 4 hours of data may not be sufficient training data to allow more than one phone of context during tree building.

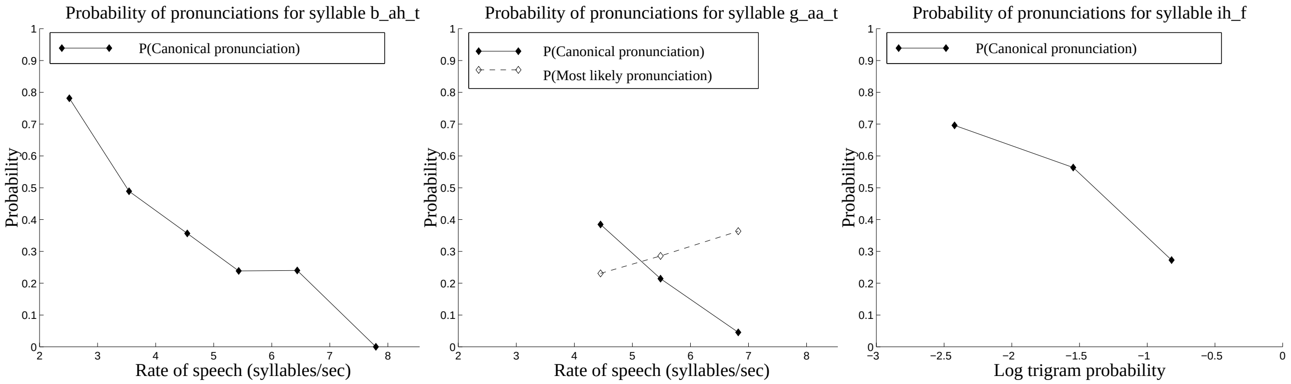
4. ESTIMATION OF SPEAKING RATE

As the reader may have noticed, most of our speaking rate investigations have been conducted using transcribed syllable rate; however, it is not feasible to have linguistic transcribers determine the speaking rate at the runtime of the recognizer. As mentioned before, we are developing a signal processing measure of rate called *mrate.* A full description of the algorithm can be found in [6]. The measure correlates pretty well with transcribed syllable rate ($\rho \sim .67$), although for faster speaking rates, it tends to underestimate the rate somewhat.

The correlation of *mrate* with pronunciation reductions is also reduced somewhat; only 54 of the 200 syllables show significant shifts in the probability of the canonical or most-likely pronunciations when *mrate* is used as the partitioning criterion. We hypothesized that *mrate* underestimates the true rate when pronunciations are non-canonical (as reduced pronunciations might have less sharp acoustic distinctions). Preliminary evidence supports this hypothesis. When *mrate* matches or overestimates the true (i.e. transcribed) rate, the probability of a canonical syllabic pronunciation is roughly 50%. However, as the amount that *mrate* underestimates the true rate increases, the canonical probability drops, reaching 33% when the rate is underestimated by 40% or more.

We are continuing work on improving our rate estimate; in the future, we will be integrating information from syl-

[3] The canonical and most likely pronunciations differed for 55 of the 200 syllables.



a. "b_ah_t" for varying rates b. "g_aa_t" for varying rates c. "ih_f" for varying trigram probs

Figure 5. Pronunciation probabilities of syllables dependent on dynamic factors.

labic onsets. We are also looking into more localized measures of rate (i.e. estimating over a few syllables). Psychologists have suggested that many rate effects are localized [14], and we have found that estimating rate using a small number of syllables locally, as opposed to interpausally, sharpens some of the pronunciation distinctions seen above.

5. CONCLUSIONS

This work originated as an analysis of pronunciation data to determine why we were not able to integrate the dynamic real-valued attributes of speaking rate and word predictability into phone-level decision trees. We have shown that pronunciations are strongly dependent on rate and language model probabilities. Thus, there is no fundamental problem with these variables as predictors of pronunciation variation. The complexity of the relationship between factors, the smoothness of change in pronunciation probabilities as a function of these factors, and the lack of both context in our phone trees and enough data to train the trees probably contributed to the poor performance of the trees.

Furthermore, we have seen that these factors all affect recognizer performance on the Switchboard database. We will soon be integrating the lessons learned here into the pronunciation models for our recognition systems.

ACKNOWLEDGMENTS

We would like to thank Dan Jurafsky, Gethin Williams, Murat Saraclar, Sanjeev Khudanpur, Steve Greenberg, Brian Kingsbury, three anonymous reviewers, and the ICSI Realization Group for helpful discussions of this work. We would also like to thank Bill Byrne for help with the WS96 recognizer and Steve Greenberg and the ICSI Switchboard Transcription Project for providing data. This work was supported by a grant from the Center for Language and Speech Processing at The Johns Hopkins University, NSF SGER grant IRI-9713346, and NSF grant IRI-9712579.

REFERENCES

[1] M. Saraclar. Automatic learning of a model for word pronunciations: Status report. In *Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation*, Baltimore, MD, May 1997.

[2] M. Weintraub, E. Fosler, C. Galles, Y.-H. Kao, S. Khudanpur, M. Saraclar, and S. Wegmann. WS96 project report: Automatic learning of word pronunciation from data. In F. Jelinek, editor, *1996 LVCSR Summer Research Workshop Technical Reports*, chapter 3. Center for Language and Speech Processing, Johns Hopkins University, April 25 1997. Research Notes No. 24.

[3] E. Fosler, M. Weintraub, S. Wegmann, Y-H Kao, S. Khudanpur, C. Galles, and M. Saraclar. Automatic learning of word pronunciation from data. In *ICSLP-96*, October 1996.

[4] F. Chen. Identification of contextual factors for pronounciation networks. In *IEEE ICASSP-90*, pages 753–756, 1990.

[5] M. Riley. A statistical model for generating pronunciation networks. In *IEEE ICASSP-91*, pages 737–740, 1991.

[6] N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *IEEE ICASSP-98*, Seattle, WA, May 1998.

[7] S. Greenberg. WS96 project report: The switchboard transcription project. In F. Jelinek, editor, *1996 LVCSR Summer Research Workshop Technical Reports*, chapter 6. Center for Language and Speech Processing, Johns Hopkins University, April 25 1997. Research Notes No. 24.

[8] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In F. Jelinek, editor, *1996 LVCSR Summer Research Workshop Technical Reports*, chapter 4. Center for Language and Speech Processing, Johns Hopkins University, April 25 1997. Research Notes No. 24.

[9] Nikki Mirghafori, Eric Fosler, and Nelson Morgan. Fast speakers in large vocabulary continuous speech recognition: Analysis & antidotes. In *Eurospeech-95*, 1995.

[10] M. A. Siegler and R. M. Stern. On the effects of speech rate in large vocabulary speech recognition systems. In *IEEE ICASSP-95*, 1995.

[11] J. Bernstein, G. Baldwin, M. Cohen, H. Murveit, and M. Weintraub. Phonological studies for speech recognition. In *DARPA Speech Recognition Workshop*, pages 41—48, February 1992. Palo Alto, California.

[12] S. Greenberg. On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, April 1997.

[13] M. Finke and A. Waibel. Flexible transcription alignment. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 34–40, Santa Barabara, CA, December 1997.

[14] Q. Summerfield. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Performance and Perception*, 7:1074–1095, 1981.