INCORPORATING CONTEXTUAL PHONETICS INTO AUTOMATIC SPEECH RECOGNITION

Eric Fosler-Lussier[⋆†], Steven Greenberg[†], and Nelson Morgan[⋆†]

[⋆]*University of California, Berkeley, USA*
[†]*International Computer Science Institute, USA*

ABSTRACT

This work outlines the problems encountered in modeling pronunciation for automatic speech recognition (ASR) of spontaneous (American) English speech. We detail some of the phonetic phenomena within the Switchboard corpus that make the recognition of this speaking style difficult. Phonetic transcribers found that feature spreading and cue trading made identification of phonetic segmental boundaries problematic. Including different forms of context in pronunciation models, however, may alleviate these problems in the ASR domain. The syllable appears to play an important role, as many of the phonetic phenomena seen are syllable-internal, and the increase in pronunciation variation compared to read speech is concentrated in coda consonants. In addition, we show that other forms of context – speaking rate and word predictability – help indicate increases in variability. We present a dynamic ASR pronunciation model that utilizes longer phonetic contextual windows for capturing the range of detail characteristic of naturally spoken language.

1. INTRODUCTION

ASR systems typically perform more poorly on spontaneous speech than on corpora containing scripted and highly planned material. Although some of this deterioration in performance reflects the wide range of acoustic background conditions typical of natural speech, much of the decline in recognition accuracy can be attributed to a mismatch between the phonetic sequence recognized and the representation of words in the system's lexicon. Finding ways to predict when and how the phonetic realization of an utterance deviates from the norm is likely to improve recognition performance.

In NIST's recent evaluation of speech recognizers [11], it was clear that all current systems perform much worse in spontaneous conditions. In Figure 1 we show the error rates of recognizers running on the Broadcast News corpus, a collection of radio and television news programs, for two different focus conditions: *planned* studio speech, in which announcers read from a script, and *spontaneous* studio speech, in which reporters conducted more natural interviews.[1] All of the recognizers in the evaluation had 60 to 100% more errors in the spontaneous condition. Since the acoustic environment of these two conditions is similar, the most plausible explanation of the variation in ASR performance is the difference in speaking style.

Recognizers' diminished performance on spontaneous speech can be attributed to many factors, such as differences in sentence structure or additional disfluencies that would affect the ASR language model [6, 13]. One of the biggest influences, however, is the variation in pronunciations seen in spontaneous speech. We have observed [2] that an increase in errors made by ASR systems correlates with situations in which phonetic transcriptions of the test speech data do not match the pronunciations found in the recognition dictionary. For example, one system tested on the Switchboard corpus of spontaneous speech produced one-third more errors for words pronounced non-canonically.

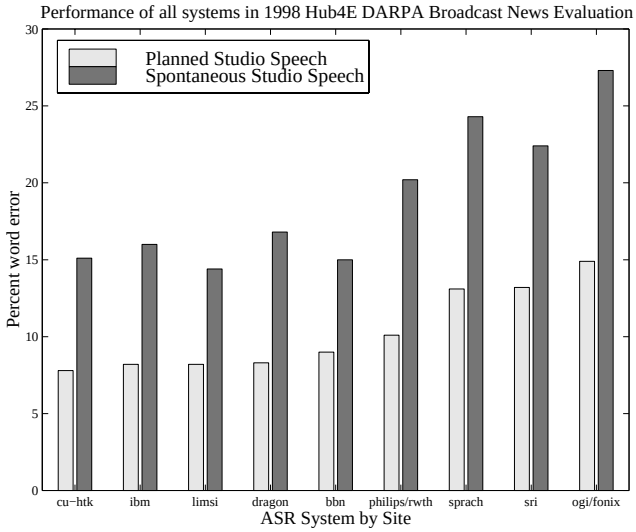

Figure 1. ASR system error for nine recognizers on planned and spontaneous studio speech in the Broadcast News corpus.

McAllaster *et al.* [10] used simulated acoustic data with their Switchboard recognizer to normalize the effects of misclassifications made by the acoustic (phonetic categorization) model; focusing on the differences between the phonetic transcript of the Switchboard test set and pronunciation models in the dictionary, they found that reductions and phonological variations in Switchboard were the single most significant cause of errors in their recognizer. Thus, a critical step for training a casual-speech recognition system is the determination of when and how pronunciations can vary in this speaking style.

2. HOW IS SPONTANEOUS SPEECH DIFFERENT?

Since the above experiments suggest that the pronunciations of spontaneous speech are different enough to cause substantial mismatches with standard recognizer pronunciation models developed primarily for read speech, it is important to characterize how these differences are realized both acoustically and with respect to features other than segmental context. We present here some observations from our transcription of the Switchboard corpus.

Syllable constituent	Switchboard (spontaneous) # instances	% Canonical	TIMIT (read) # instances	% Canonical
Onset	39214	84.4	57868	90.0
Simple [C]	32851	84.7	42992	88.9
Complex [CC(C)]	6363	89.4	14876	93.3
Nucleus	48993	65.3	62118	62.2
with/without onset	35979 / 13104	69.6 / 53.4	50166 / 11952	64.7 / 51.8
with/without coda	26258 / 15101	64.4 / 66.4	32598 / 29520	58.2 / 66.6
Coda	32512	63.4	40095	81.0
Simple [C]	20282	64.7	25732	81.3
Complex [CC(C)]	12230	61.2	14363	80.5

Table 1. Frequency of phone transcription matches against the lexicon's canonical pronunciation for Switchboard and TIMIT

2.1. Transcribing Switchboard

For the 1996 and 1997 Johns Hopkins Large Vocabulary Continuous Speech Recognition Summer Research Workshops, linguists at ICSI transcribed phonetically roughly four hours of the Switchboard corpus [4]. The difficulty of transcribing this data provided valuable insights into how the assumptions made for read-speech transcription did not fit this database.

The original transcription system used was modeled after the guidelines developed for transcribing the TIMIT corpus of prompted speech [3]. Transcribers were asked to segment words into individual phones, as most ASR systems require. However, the transcribers often found phenomena that defied the given segmentation and identification criteria. Irregular phonetic expression of segments was a common occurrence. The linguists cited the following difficulties in transcription:

Feature spreading: Many segments are deleted entirely in production, though their influence is often manifest in the phonetic properties of their segmental neighbors. This makes it difficult to determine hard phonetic boundaries. For example, the character of vowels neighboring `/r/` or following `/j/` are colored almost completely by the consonant; it was impossible to say where the segmental boundary lay. Nasals often spread into adjoining stops (*e.g.*, `/nd/` clusters in syllable codas), eliminating the closure but preserving the stop burst.

Cue trading: Alternative phonetic realizations often occur in place of canonical acoustic patterns. For example, dental and nasal flaps are occasionally demarcated by dips in waveform amplitude, rather than by any noticeable change in the formant trajectories. Often, there was almost no acoustic evidence for very predictable words (*e.g.*, more *of* that); however, a vestigial timing cue would indicate the presence of a word that could be filled in from context.

These observations instigated a slight shift in transcription focus for later phases of the project. Since phonetic boundaries were difficult to determine, and many of the observed phenomena were syllable-internal, the linguists were instructed to give the phonetic identities of segments, but only mark junctions between syllables. While not every boundary was unambiguous, this did ease the decision process for transcribers, speeding transcription greatly. For more examples from the Switchboard transcription project, visit `http://www.icsi.berkeley.edu/real/stp`.

2.2. TIMIT versus Switchboard

Syllabic constraints exert influence on pronunciation variation in both read and spontaneous speech; the differences between the two speaking styles also stand out when examining phones within syllabic contexts. Greenberg [5] has previously demonstrated with the Switchboard corpus that the probability of canonical pronunciation of a phone depends on the position of the phone within the syllable. We compared these results with the TIMIT read-speech corpus in order to determine whether syllabic constraints caused characteristic pronunciation variation effects.

We compared the pronunciations transcribed for each word in Switchboard and TIMIT to the closest pronunciation given for the word in the Pronlex pronunciation dictionary [9], using automatic syllabification methods to determine syllabic positions, as described in [2].[2] This procedure highlighted marked similarities and differences between pronunciations in the two corpora. As we see in Table 1, onset consonants are pronounced canonically more often than other phones in both corpora, particularly in the case of complex consonant clusters. These segments are often acoustically strong, perhaps to demarcate the start of a syllable. Also, vowel nuclei match the *a priori* pronunciation approximately as often in read as in spontaneous speech. This is a surprising fact — it suggests that the acoustics of vowels are influenced by context, but still remain relatively variable. Nuclei without preceding onset consonants are much less likely to be canonical than those with onsets, probably because they are influenced more by the varying preceding syllable than by the (usually canonical) onset.

The biggest difference between spontaneous and read speech is the large increase in variability of the coda consonants — essentially a 20% change. Thus, in spontaneous speech coda segments are about as canonical as nuclei, whereas in read speech their canonicity compares to that of onset consonants. Keating's [8] analysis of a different portion of this corpus concurs with this finding: most of the variation phenomena she discusses involve changes either in vowel qualities or in the final consonant.

The implication of these findings is that words may be identified most strongly by the syllable-initial portion of the word. Less variation is observed in onsets because they are used to discriminate between lexical items. Given the words in the transcribed portion of the Switchboard corpus, we located pairs of words that differed by one phone in the Pronlex dictionary (*e.g.*, *news* and *lose*). These pairs were classified by whether the phone difference was in onset, nucleus, or coda position. Onset discrepancies outnumbered nucleus discrepancies by a factor of 1.5 to 1, and coda discrepancies by 1.8 to 1, indicating that at least for this crude measure, onsets appear to be more important for word discriminability.

2.3. Word Frequency and Speaking Rate

Phonetic context is not the only factor that can affect the acoustic realization of words. We have been investigating other non-segmental factors (word frequency and speaking rate) that can determine how pronunciations can vary [2].

We computed an average syllabic distance measure between the phonetic transcription and the Pronlex dictionary for all of the syllables in the transcribed portion of the Switchboard corpus; an

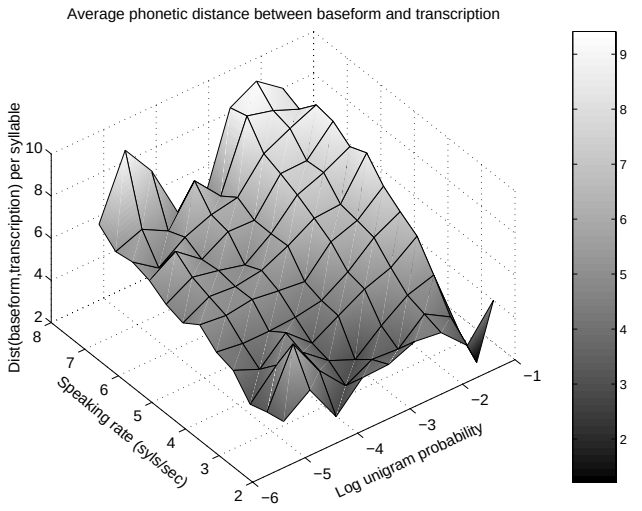

Figure 2. Distance from canonical pronunciation as a function of word frequency and speaking rate (from [2]). Higher frequency words are to the right on this graph; faster speaking rates are to the left/rear.

increase in this measure corresponds to further divergence in pronunciation in terms of a phonetic feature space. In Figure 2, this measure is plotted against the unigram frequency of the word and local interpausal speaking rate, as given by the transcribers.

There is an interaction 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. This complex interdependency between these three variables makes sense from an information-theoretic viewpoint — since high-frequency words are more predictable, more variation is allowed in their production at various speaking rates, as the listener will be able to reconstruct what was said from context and few acoustic cues.

Other factors besides speaking rate and word predictability can affect pronunciations. Jurafsky *et al.* [7] have studied how filled pauses, disfluencies, segmental context, speaking rate, and word predictability relate to the realization of the ten most common function words in the Switchboard corpus. For many of these variables, they found significant independent effects on function word reductions.

3. IMPLICATIONS FOR ASR MODELS

It is clear that the context in which a phone appears has a significant effect on the acoustic (and articulatory) realization of the phone; this effect is very prominent in spontaneous speech. The increased variability in the phonetic realization must be considered in building statistical models for ASR systems. Many of the "problematic" phonetic phenomena described here can be modeled by examining the extended context for each phone: either the neighboring phones or the containing syllable or word.

Many speech recognizers already incorporate triphone models [12] that are dependent on the previous and subsequent phones in context. In essence, one builds finer and finer models of phonetic categories; so that one does not have to build a model of *every* possible phonetic context, clustering techniques [15] that either use phone categories (*e.g.* manner or place of articulation) or a blind statistical criterion of similarity can effectively reduce the number of models needed.

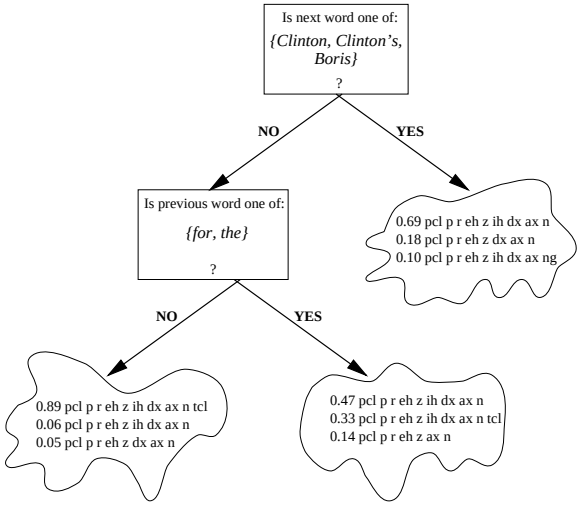

Figure 3. Decision tree model for *president*.

Another option is to determine which pronunciation models match acoustic examples under different contexts [14, *inter alia*]. In this scenario, a recognizer, trained using a baseline pronunciation representation, generates a phonetic transcription of some training data unconstrained by the word sequence. One can then use automatic techniques to find how the unconstrained ASR phone models differ from the dictionary pronunciation, given the surrounding phones as context — a quasi-phonological approach. Instead of concerning ourselves with the interrelation of phonemes and phones, we are determining how phones relate to recognizer models in different contexts.

As we have seen, all phones are not created equal — syllabic position can influence the phonetic realization of segments. Since many of the phenomena we studied are syllable-internal, syllable and word models can be used explicitly to model internal context. Rather than spending modeling power on learning the contexts in which phones change pronunciation, we allow segmental context to determine the set of models we use. We can then learn how other factors (*e.g.*, speaking rate) affect pronunciations within this longer context and *dynamically* choose appropriate pronunciation models during recognition.

We trained decision trees (d-trees) to predict the pronunciation of words based on information about surrounding words. D-trees [1] are statistical classifiers that can select a set of features to improve the prediction of events (in this case the probability of a particular pronunciation). Thus, we can present the d-tree algorithm with a substantial number of features, such as the identities and features of surrounding phones or extra-segmental features like speaking rate and word predictability, and have the algorithm automatically select the best combination of these features to improve pronunciation classification.

Using roughly 74 hours of training data from the Broadcast News corpus, we built models for the 550 most frequent words using surrounding word identities and the identities, manner, place, and syllabic position of neighboring phones as features in the d-tree. We also included information about word length, several estimates of speaking rate, and the trigram probability of the word. Slightly less than half of the trees in each case used a distribution other than the prior (*i.e.*, were grown to more than one leaf).

The automatic analyses provided by the d-tree algorithm located several linguistically plausible pronunciation changes. For example, in the tree for *president* (shown in figure 3), when the

Dictionary	All conditions	Planned studio	Spontaneous studio
Baseline	26.7%	15.4%	27.2%
Word trees	26.5%	15.0%	27.0%
Syllable trees	26.3%	15.3%	25.8%

Table 2. Broadcast News word error rate for dynamic tree models.

following word was *Clinton*, *Clinton's*, or *Boris*, the final /t/ closure was very likely to be deleted. In addition, the velarization of /n/ to [ng] was possible, a likely consequence of the following /k/ in *Clinton('s)*. It is important to note that the velarization requires the deletion of /t/ to be possible; it is easier for the recognizer to learn these co-occurrences when units larger than individual phones are modeled.

We also trained roughly 800 d-trees to model syllables, giving about 70% coverage of the syllables in the corpus. Each word was given a single canonical syllable transcription so that words with similar alternative syllabic-internal pronunciation in the baseline dictionary shared the same syllable model. In addition to the features found in the word trees, we informed the the syllable trees about the lexical stress of the syllable, position within the word, and the word's identity.

We found the 100 best hypotheses for each utterance using our baseline recognizer in a 30-minute subset of the 1997 Broadcast News (Hub 4) English evaluation test set. The word and syllable d-trees were used to expand each hypothesis into a large pronunciation graph that was then rescored; hypotheses were then re-ranked using an average of the old and new acoustic scores.

The word-based d-trees gave a slight improvement over the baseline, though the syllable trees boosted results a bit more. Notably, the word trees provided incremental improvements under each focus condition, whereas the syllable trees contributed primarily to an improvement specific to spontaneous speech. Given the distinct effects of syllabic structure on spontaneous pronunciations demonstrated in Section 2, the improvement on this speaking style is not unexpected; however, the exact relationship between these phenomena is uncertain, and bears further investigation.

4. CONCLUSIONS

Spontaneous speech presents a difficult challenge to speech researchers; engineers and phoneticians should work together to build coherent models of the pronunciation variability inherent in this speaking style. Mostly due to this variability, current recognizer technology for spontaneous speech lags behind that for recognition of planned speech.

The pronunciation variability inherent in Switchboard is accompanied by a number of non-traditional phonetic phenomena, including feature spreading and cue trading. We have found that a syllabic orientation can help explain some of these phenomena, as the onsets of syllables in casual speech tend to be more stable than the rime (nucleus/coda segments).

In order to integrate these phonetic observations into our recognizer, we developed statistical models of syllables and words which took into account an extended context that included word predictability and speaking rate, as well as segmental context. An initial implementation of this model showed improvement particularly for the spontaneous speech portion of the Broadcast News corpus; we are encouraged by these results, and are continuing development of these models.

ACKNOWLEDGMENTS

This work was supported by the European Community basic research grant SPRACH, NSF SGER grant IRI-9713346 and NSF grant IRI-9712579.

NOTES

1. The corpus also comprises several other focus conditions, including degraded acoustics and foreign accents.
2. The results reported here deviate slightly from those listed in [5:Table 6] due to differences in how the canonical dictionary pronunciation was chosen, as well as issues of normalizing phonesets between the Switchboard and TIMIT transcriptions.

REFERENCES

[1] Breiman, L., Friedman, J., Olshen, R., and Stone, C. 1984. *Classification and Regression Trees*. Belmont: Wadsworth.

[2] Fosler-Lussier, E. and Morgan, N. 1998. Effects of speaking rate and word frequency on conversational pronunciations. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 35–40, Kerkrade, Netherlands.

[3] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. 1993. Darpa timit acoustic-phonetic continuous speech corpus. Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD.

[4] Greenberg, S. 1997. WS96 project report: The Switchboard transcription project. In Jelinek, F. (ed.), *1996 LVCSR Summer Research Workshop Technical Reports*, chapter 6. Center for Language and Speech Processing, Johns Hopkins University.

[5] Greenberg, S. 1998. Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 47–56, Kerkrade, Netherlands.

[6] Heeman, P. and Allen, J. 1997. Intonational boundaries, speech repairs, and discourse markers: Modeling spoken dialog. In *Proceedings of the 35th ACL*, Madrid, Spain.

[7] Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., and Raymond, W. 1998. Reduction of English function words in Switchboard. In *ICSLP-98*, Sydney, Australia.

[8] Keating, P. 1997. Word-level phonetic variation in large speech corpora. To appear in an issue of *ZAS Working Papers in Linguistics*, ed. Berndt Pompino-Marschal. Available as http://www.humnet.ucla.edu/humnet/linguistics/people/keating/berlin1.pdf.

[9] Linguistic Data Consortium (LDC). 1996. The PRONLEX pronunciation dictionary. Available from the LDC, ldc@unagi.cis.upenn.edu. Part of the COMLEX distribution.

[10] McAllaster, D., Gillick, L., Scattone, F., and Newman, M. 1998. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. In *ICSLP-98*, pp. 1847–1850, Sydney, Australia.

[11] Pallett, D., Fiscus, J., Garofolo, J., Martin, A., and Przybocki, M. 1999. 1998 Broadcast News benchmark test results: English and non-English word error rate performance measures. In *DARPA Broadcast News Workshop*, Herndon, Virginia.

[12] Schwartz, R., Chow, Y., Roucos, S., Krasner, M., and Makhoul, J. 1984. Improved hidden Markov modeling of phonemes for continuous speech recognition. In *IEEE ICASSP-84*, pp. 35.6.1–4, San Diego, CA.

[13] Stolcke, A. and Shriberg, E. 1996. Statistical language modeling for speech disfluencies. In *IEEE ICASSP-96*, pp. 405–409. Atlanta, GA.

[14] Weintraub, M., Fosler, E., Galles, C., Kao, Y.-H., Khudanpur, S., Saraclar, M., and Wegmann, S. 1997. WS96 project report: Automatic learning of word pronunciation from data. In Jelinek, F. (ed.), *1996 LVCSR Summer Research Workshop Technical Reports*, chapter 3. Center for Language and Speech Processing, Johns Hopkins University.

[15] Young, S. J., Odell, J. J., and Woodland, P. C. 1994. Tree-based state tying for high accuracy acoustic modelling. In *IEEE ICASSP-94*, pp. 307–312.