A Cost-Benefit Analysis of Hybrid Phone-Manner Representations for ASR

Eric Fosler-Lussier
Department of Computer Science and Engineering
Department of Linguistics
Ohio State University
Columbus, OH 43210
fosler@cse.ohio-state.edu

C. Anton Rytting
Department of Linguistics
Ohio State University
Columbus, OH 43210
rytting@ling.ohio-state.edu

Abstract

In the past decade, several researchers have started reinvestigating the use of sub-phonetic models for lexical representations within automatic speech recognition systems. Lest history repeat itself, it may be instructive to mine the further past for models of lexical representations in the lexical access literature. In this work, we re-evaluate the model of Briscoe (1989), in which a hybrid strategy of lexical representation between phones and manner classes is promoted. While many of Briscoe's assumptions do not match up with current ASR processing models, we show that his conclusions are essentially correct, and that reconsidering this structure for ASR lexica is an appropriate avenue for future ASR research.

1 Introduction

Almost every state-of-the-art large vocabulary automatic speech recognition (ASR) system requires the sharing of sub-word units in order to achieve the desired vocabulary coverage. Traditionally, these sub-word units are determined by the phones or phonemes of a language (depending on desired detail of representation). However, phonetic (or phonemic) representation has its pitfalls (*cf.* (Ostendorf, 1999)). Among the problems cited in the literature are that (1) segments are often difficult for machines to recognize from the acoustic cues alone, because the acoustic cues to a particular phoneme are multi-faceted, and (2) the intended words and phrases are not always recoverable even from correctly recognized segments, because speakers themselves will also fail to articulate words with the dictionary-listed phonemes. The first of these problems refers to the *discriminability* of phonemes within an inventory; the second to the *reliability* of (actual) phone sequences mapping to the canonical phonemic representations of words. This is particularly true in conversational speech (such as that found in the Switchboard corpus), where pragmatic context and conversational conventions assist human comprehension (but not current ASR systems).

A common approach for handling pronunciation variation is to introduce alternative entries into the lexicon. However, phones that are perceived as non-canonical (for example, when an /eh/ is heard as an /ih/ by linguistic transcribers) often are closer in acoustic space to the Gaussian means of the canonical phones, rather than the perceived phones (Saraçlar et al., 2000). This insight suggests that acoustic models need to be cognizant of potential pronunciation changes. Thus the lexical and acoustic models should work hand in hand.

Another way to model this type of pronunciation variation is to find the commonalities that the canonical and perceived phone share in terms of a sub-phonetic representation. In the past decade, a significant community in acoustic-phonetic ASR research has been turning to distinctive features (Jakobson et al., 1952) for building ASR lexica. While an exhaustive description of these approaches is beyond the scope of this paper, estimates of phonological feature probabilities have been combined to obtain phone probabilities (Kirchhoff, 1998), or incorporated into "feature bundles" that allow representa-

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 571–578, Vancouver, October 2005.

tion of phonological processes (Finke et al., 1999).

More recent work has integrated phonological features into graphical models (Livescu et al., 2003) and landmark based systems (Juneja and Espy-Wilson, 2004). The common thread among this research is the notion that acoustic models should be sensitive to sub-phonetic information. With this trend in phonological representation research, it is time to re-examine some older hypotheses about lexical access and speech processing in order to gain some insight in this current featural renaissance.

Sub-phonetic ASR research is also driven by the fact that deviations from canonical pronunciation and from correct perception of phones is far from random; indeed, there have been a number of studies demonstrating that both of these variations have defined, modelable trends. Deviations from canonical pronunciation can be described by phonological rules, and errors in perception also tend to conform to phonological patterns. By and large, confusions occur (at least in humans) between phones with phonological features in common (e.g., (Miller and Nicely, 1955)). In particular, three features (voicing, manner, and place) have been postulated as relatively invariant (see e.g., (Stevens, 1981), quoted in (Church, 1987)). It follows from this phonetic detection based on the most reliable features may handle highly variable speech more robustly than systems which demand full identity over all the features for a given phone or phone sequence.

Consequently, a number of researchers have previously suggested using certain broad classes of segments, rather than full phonemic identification, for a first pass on recognition. For instance, Shipman and Zue (1982), working on large-vocabulary isolated word recognition, used both two-way consonant-vowel distinctions and a six-way distinction based on manner in order to divide their 20,000-word dictionary into "cohorts" or groups of words. They found that this partial specification of segments reduced the search space of word candidates significantly. Carlson et al. (1985) found similar results for English and four other languages.

2 A suggested compromise: a hybrid phone-manner representation

Briscoe (1989) extended this broad-class approach to address the problem of lexical access on connected speech. However, Briscoe argues against the use of broad, manner-based classes at all times. He argues that manner cues provide no particular advantage for stressed syllables, but that all cues are sufficiently reliable in stressed syllables to justify a full segmental analysis. Working with a 30,000-word lexicon, Briscoe shows that the manner-based broad classes for weak (reduced) syllables, together with full identification of strong (unreduced) syllables constrained the set of possible candidates satisfactorily. Unfortunately, he only provides results for one sentence from his corpus.

This approach proposes to adjust the granularity of recognition dynamically, depending on the stress level of the current syllable. The details of how this would be managed are left somewhat vague. As it stands, it would seem to depend crucially on first detecting the stress of each frame, so as to determine which alphabet of symbols to apply to incoming input. Alternatively, it could recognize the broad class as a first pass, and then refine this into a full phonemic analysis for stressed syllables in a second pass, at the cost of multiplying passes through the speech data. It is not possible in this system to recover from the miscategorization of stress.

One possible remedy is to bypass a hard decision on stress and run both a manner-based broad-class detector and a traditional phonemic system in parallel. These then may be combined according to the probability of lexical stress, such that those frames judged less likely to be stressed weight the broad-class analysis more heavily, and those judged more likely to be stressed weight the narrow phonemic analysis more heavily. Its advantage is that a full phonemic analysis is recoverable for each frame and phone, but those in weak syllables (and hence less likely to be accurate) weigh in less heavily.

Briscoe's analysis is in terms of lexical access activations: taking a cue from the lexical access community, he assumes that any "partially activated" word (e.g., "boat" and "both" being active after processing "bo") will contribute linearly to the processing time in ASR. However, most large-vocabulary ASR systems today use a tree-based lexicon where common phonetic prefixes of words are processed only once, thus invalidating this conjecture. Briscoe experimented with several triggers for starting a new word — at every phone, at the beginnings of syllables, at the beginnings of syllables with unreduced vowels, and at the beginnings of word boundaries.

Category	Example (Vietnamese)	Cohorts
# of Words		6247
Stress pattern only (lower bound)	0001	84
Identify phones in stressed syllables	0001:m_iy_z	4709
+CV pattern of unstressed syllables	0001:CV:VC:CV:m_iy_z	5609
+manner pattern of unstressed syllables	0001:FV:CS:NV:m_iy_z	6076
Phonetic prons. (upper bound)	v iy . eh t . n aa . m iy z	6152

Table 1: Cohorts for varied lexical representations

However, the latter three require oracle information as to where word or syllable boundaries can occur. A more appropriate measure commensurate with current ASR practice would be to only allow words to start where a previous word hypothesis ends.

In the remainder of the paper, we seek to validate (or invalidate) Briscoe's claim that a hybrid phonetic and feature model is appropriate for ASR processing. In the 15 years since Briscoe's paper, the ASR community has developed large phonetically transcribed corpora and more advanced computational tools (such as the AT&T Finite-State Toolkit (Mohri et al., 2001)) that we can apply to this problem.

3 Experiment 1: Effective Partitioning by Manner-based Broad Classes

Our first experiment explores various types of broad classes to determine the effects of these encodings on cohort size within a sample 6,000 word dictionary.[1] Here we use the lexical stress-marked dictionary provided with the TIMIT database (Garofolo et al., 1993), which was syllabified using the NIST Tsylb2 syllabifier (Fisher, 1996).

Rather than calculate cohort size directly, we calculate the number of cohorts into which our dictionary is partitioned, a measure which Carlson et al. (1985) showed to correlate well with expected cohort size. (Note that this is an inverse correlation.) This describes the static *discriminability* of the lexicon: systems that have words with the same lexical representation will not be able to discriminate between these two words acoustically and must rely on the language model to discriminate between them.

[1] "Cohort size" is used here (as with Shipman and Zue (1982)) to mean the number of distinct vocabulary items that match a particular broad-class encoding. It is not intended to imply a particular theory of lexical access.

Before proceeding, it may useful to set upper and lower bounds for this exercise (Table 1). An obvious upper bound is the full phonemic disambiguation of every word. Of the 6247 words in the dictionary, 6152 unique pronunciations are found (a few cohorts consisting of sets of homophones). A convenient lower bound is the lexical stress pattern of the word, devoid of any segmental information: e.g., "unidirectional" has its stress on the 4th of 6 syllables; hence, 000100 is its lexical-stress profile. 84 unique lexical-stress profiles exist in the dictionary.

Between these two bounds, three variant broad-class partitions were explored for isolated word recognition. All three use the lower-bound stress profile as a starting place, combined with full phonemic information for the syllable with primary stress. The first, with no additional segmental information, produces 4709 distinct cohorts. The second adds a consonant-vowel (CV) profile for the unstressed syllables, which boosts the number of distinct cohorts to 5609. The final partition replaces the CV profile with a six-class manner-based broad-class partition (Nasals, Stops, Fricatives, Glides, Liquids, and Vowels). Including a manner-class representation for unstressed vowels increases the number of cohorts to 6076, which is very close to the upper bound. Thus, there is not much loss of lexical discriminability when using this type of representation.

3.1 Caveats

Now, for this scheme to be maximally useful for recognition, several conditions must obtain. First, we have assumed that we can reliably detect lexically stressed syllables within the speech signal. Waibel (Waibel, 1988) has shown that stress correlates with various acoustic cues such as spectral change. As a side experiment, we have shown that very basic methods provide encouraging results (only sketched here due to space constraints). We re-annotated TIMIT with lexical stress markings, where all frames of each stressed syllable (including onset and coda consonants, not just the nucleus) were marked as stressed. A multi-layer perceptron with 100 hidden units was trained to predict $P(\mathrm{Stress}|\mathrm{Acoustics})$ with a nine-frame context window. No additional phonetic information besides the binary label stressed/unstressed was used in training. Frame-by-frame results on the TIMIT test set were 75% accurate (chance: 52%), and when

MLP output was greater than 0.9, a precision of 89% was obtained (recall: 20%). While far from perfect, this result strongly suggests that even very simple methods can predict lexical stress fairly reasonably.

A second assumption in the above analysis was that words occur in isolation. It is clear that in connected speech, there are a larger number of potential lexical confusions. A third assumption is that those features we are relying upon in our partitions (namely, all features within stressed syllables, and manner of articulation for unstressed syllables) are perfectly reliable and discriminable. In the next two sections, we relax these assumptions by applying extensions of this method to connected speech.

4 Experiment 2: What does a hybrid representation buy you?

As Experiment 1 shows, the hybrid phone/feature representation does not drastically decrease the discriminability of the (albeit small) lexicon. It is also possible that such a representation reduces pronunciation variation, by allowing the canonical representation to more closely match actual pronunciations. For example, we have demonstrated that for common ASR corpora (Switchboard and TIMIT), segments in unstressed syllables were much more likely to deviate from their canonical lexical representation (Fosler-Lussier et al., 1999). If phones that deviate from canonical still keep the same manner class, then a dictionary built with Briscoe-esque representations should more closely match the actual pronunciations of words in running speech (as transcribed by a phonetician).

4.1 Method

In order to test this theory, we used phonetic data from (Fosler-Lussier et al., 1999) in which the ICSI phonetic transcripts of the Switchboard corpus (Greenberg et al., 1996; NIST, 1992) were aligned to a syllabified version of the Pronlex dictionary (Linguistic Data Consortium (LDC), 1996), which has 71014 entries for 66293 words. In this alignment, for every canonical phone given by the lexicon, there were zero or more corresponding realized phones. From these data we extracted the canonical and realized pronunciation of each word token, for a total of 38,527 tokens. Generally, high-frequency function words show the most variation, so they may benefit most from a manner-based representation.

Lexicon type	Strict matching	Matching w/ deletion
1) Phonetic units	37.0%	50.1%
2) Manner-based function words	50.2%	69.6%
3) + Manner for unstressed syls	53.4%	74.6%
4) + Manner for secondary stress	55.7%	77.9%
5) Manner for all syls	60.7%	85.2%

Table 2: Percent of words pronounced canonically for phonetic and hybrid lexical representations

Given these word pronunciation data, we can examine how many word tokens have transcriptions that match their dictionary-listed pronunciations, given the broad-class mappings for various sets of syllables. We built lexica and mapped phonetic transcriptions according to five different criteria:

1. Every segment is phone based (no classes).
2. Function words use manner-based classes.
3. Unstressed syllables and function words use manner only.
4. Secondary stressed syllables also use manner. (Primary stressed syllables are phone based.)
5. Every segment uses manner-based classes.

We noted in the data (as others have done) that a large proportion of the pronunciation variation was due to phone deletion (29% of words) — which would not be handled by the manner-based lexicon. However, it is likely that not every phone deletion leads to an ASR error (as attested by the fact that state-of-the-art Switchboard ASR error rates are typically less than 29%). Often there is enough residual phonetic evidence of the deleted phone, or enough phonetic evidence in other parts of the word, to recognize a word correctly despite the deletion. Thus, we decided to use a two-part strategy in calculating canonical pronunciation (Table 2). The first column, "strict matching", allows no insertions or deletions when comparing the canonical and realized pronunciation. "Matching with deletion" reports the ideal situation where phone deletions were perfectly recoverable in their canonical form. Including and ignoring deletions provides upper and lower bounds on the true lexical access results. (Insertions are relatively rare and not anticipated to affect the results significantly, and hence are not examined.)

4.2 Results and Discussion

In Table 2, we see that a standard ASR lexicon approach (strict matching 1), does not match the tran-

scribed data very well, with only 37% of words pronounced according to the dictionary. The strict matching hybrid scenario on line 3 most closely resembles Briscoe's experiment, and shows a marked improvement in matching the dictionary and realized pronunciations; comparing the two, we see that using manner-based broad classes reduces mismatch by 25% of the total error (from 63% error to 47%), most of which comes from improved modeling of function words (line 2). Whether this gain in representation is worthwhile will depend of course on the cost in terms of the increased hypothesis space.

By allowing for perfect deletion recovery (which will of necessity entail another large expansion of the hypothesis space), a somewhat more optimistic is obtained. Comparing the "matching with deletion" columns of lines 1 and 3, we see that a little over half of the non-deletion pronunciation variation is due to manner changes in unstressed syllables. Again, a good chunk of this is in function words. By moving to manner class for stressed syllables as well would bring the hypothetical error from 25% to 15%, but at the cost of a huge explosion in the hypothesis space (as Briscoe rightly points out and as discussed in the next section).

One interesting implication of this data is that over all types of segments (stressed and unstressed), roughly three-quarters of word pronunciation variants differ from the canonical only in terms of within-manner variation and phonetic deletion.

The moral of this story is that manner-based broad classes may be a useful type of back off from truly reduced and variable syllables (particularly function words), but the full benefit of such a maneuver would only be realized after a reasonable solution for recovering large-scale deletions is found. This may come from predicting with increased specificity where deletions are likely to occur (e.g., complex codas), and what reduced realizations (e.g., of function words) are most common.

5 Experiment 3: What is the cost of a hybrid representation?

Briscoe measured the cost of hybrid representation in terms of the number of lexical activations that a partially-completed word creates (see Section 2). Yet Briscoe's methodology has several shortcomings when applied to today's ASR technology; a summary of the arguments presented above are: (1) Tree-based lexica now share processing for words with identical prefixes. (2) New words are activated only when other word hypotheses end. (3) We now have a large amount of phonetically transcribed, spontaneous speech. (4) Perfect stress detection is not really achievable.

Given criticism 1, a better measure of potential processing requirements is to generate a lattice of hypothesized words and count the number of arcs in the lattice. This lattice can be constructed in such a way that criticism 2 is satisfied. In the next section, we present a finite state machine formalism for generating such a lattice.

We apply this technique to the phonetic transcription of the Switchboard corpus (thus alleviating criticism 3). However, this introduces several problems. As Experiment 2 shows, many words have pronunciations that do not appear in the dictionary. Thus, we must find a way to alleviate the mismatch between the phonetic transcription and the dictionary in a way that is plausible for ASR processing.

We can address criticism 4 by creating phone-based and manner-based transcriptions that will run in parallel; thus, the lattice generator would be free to choose whichever representation allows the matching to a dictionary word.

5.1 Method

In this experiment we consider a finite-state transducer model of the strategy described above. This corresponds not to the ASR system as a whole, but rather to the pronunciation model of a traditional system. We assume that the pronunciation as given by the transcriber is correct, but we model the transformation of realized phones into canonical dictionary pronunciations. Since we are only investigating the combined acoustic-phonetic-lexical representation, we have left out the re-weighting and pruning of hypotheses due to integration of a language model, discourse model, or any other constraints.

Specifically, this model consists of three finite state transducers composed. The first FSM, **R**, encodes the representation of the **r**ealized phonetic transcription of the spoken corpus. In order to match this to dictionary pronunciations, we train a confusion matrix on all realized/canonical phone pairs, to obtain $P(\text{dictionary phone}|\text{transcribed phone})$; these confusion probabilities are encoded as a finite state transducer **C**. Thus, **C** is derived by computing

the strength of all correspondences between the phonetic transcription of what was actually said at the phone level and the canonical pronunciation of the corresponding words. This confusion matrix consists of three parts, corresponding to substitutions, insertions, and deletions.

1. Pairwise substitutions are counted to yield a standard confusion matrix.

2. Where two or more realized phones correspond to a single canonical phone (a rare occurrence, as in e.g., *really* /r iy l iy/ → [r ih ax l iy]), each realized phone is allowed (independently) to be either deleted or substituted with its pairwise confusions from (1).

3. Deleted phones are assumed to be potentially recoverable (as in Experiment 2), so both an epsilon transition and the canonical pronunciation are preserved in the confusion matrix.

In each of these confusion matrices, we have always preserved the pathway from each realized utterance to its canonical representation for the whole corpus. So for this seen corpus, it is always possible in theory to recover the canonical representation, such that the right answer is always one of the possible hypotheses. While this may seem a bit strange, here we can only overestimate the potential hypothesis space (by adding the correct string and by assuming that deletions are recoverable); the point of this exercise is to see the number of total hypotheses (the search space) generated under such a system.

The third transducer, **D**, is the ASR dictionary that we wish to test. Thus, composing $\mathbf{R} \circ \mathbf{C} \circ \mathbf{D}$ will give the graph of all potential complete hypotheses in this space. Figure 1 shows a pruned hypothesis graph for the phrase "it's really sad" (the full hypothesis graph has 12216 arcs).

5.2 Results and Discussion

By choosing different sub-word representations, we can test Briscoe's contention that backing off to manner-based broad classes for certain (e.g., unstressed) syllables will reduce the search space and/or facilitate recovery of the intended word string. When a phone is substituted with a manner class, we construct **C** so that the generated confusions are over manner classes rather than phones.

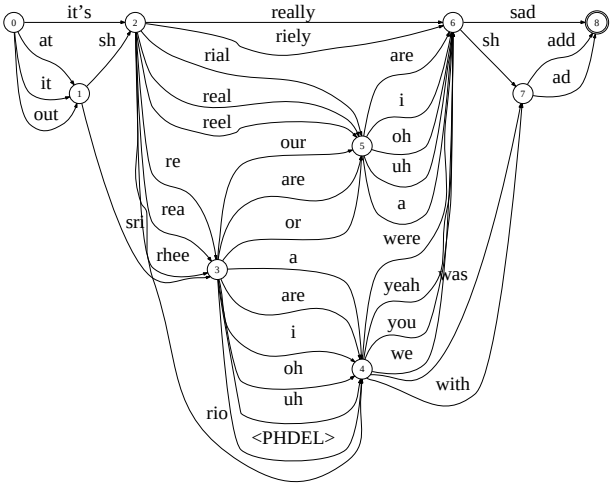

Figure 1: Pruned hypothesis graph for *It's really sad*

Figure 2 shows how the number of hypotheses per word changes as a function of the number of words in the hypothesis. Note that if the relationship were linear, we would expect to see a flat line. The figure demonstrates that that Briscoe's conclusions were correct, given the assumption that one can accurately detect lexical stress (as illustrated by the line with circles on 2). Across all utterances, the average number of hypotheses per word for the hybrid dictionary was 510 (roughly 1/3 of the phone-based average of 1429). However, when one allows for the fact that stress detection is not perfect, one sees an *increase* in the amount of necessary computation: the non-ideal hybrid dictionary has an average of 3322

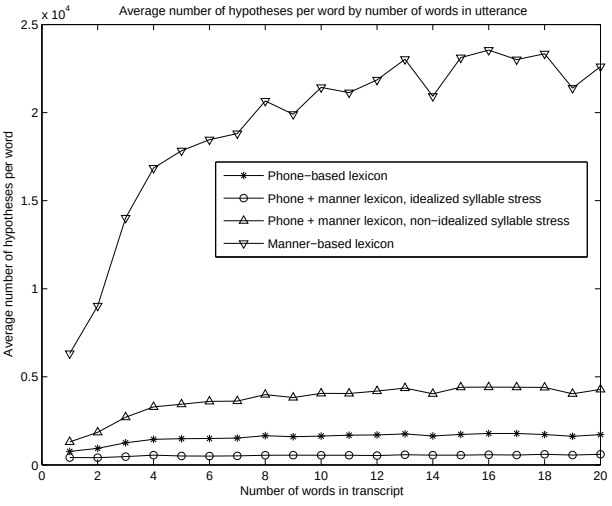

Figure 2: Average number of hypotheses per word as a function of number of words in utterance

hypotheses per word (2.3 times the phone-based average). Yet this is much lower than the potential growth of the hypothesis space given with manner-only dictionaries. This dictionary generated a hypothesis space 12 times as large as a phone based dictionary (17186 hypotheses/word average); moreover, the curve grows significantly as a function of the number of words, so longer utterances will take disproportionately more space. Thus, Briscoe's hypothesis that purely manner-based decoding is too expensive seems to be confirmed.

6 Integration into ASR

This paper has investigated hybrid representations along computational phonology lines, but we have also trained an ASR system with a hybrid lexicon for the Wall Street Journal (WSJ0) corpus. Space does not permit a full explanation of the experiment here (for more details, see (Fosler-Lussier et al., 2005)), but we include the results from this experiment as evidence of the validity of the approach.

In this experiment, we trained phonetic and manner-based acoustic models for all segments using the flat-start recipe of the HTK recognizer (Young et al., 2002). After a number of iterations of EM-training, we constructed a hybrid set of acoustic models and lexicon in which phones in unstressed syllables were replaced with manner classes (Hybrid-all). We also derived a lexicon in which the recognizer could choose whether a manner or phonetic representation was appropriate for unstressed segments (Hybrid-choice). During evaluation, we found that the Hybrid-choice lexicon degraded only slightly over a phone-based lexicon (9.9% word error vs. 9.1%), and in fact improved recognition in mild (10dB SNR) additive car noise (13.0% vs. 15.4%). The Hybrid-all was worse on clean speech (13.1% WER) but statistically the same as phone-based on noisy speech (15.8%). While not conclusive, this suggests that hybrid models may provide an interesting avenue for robustness research.

7 Conclusion

Our studies verify to some degree Briscoe's claim that a hybrid representation for lexical modeling, with stressed syllables receiving full phonetic representation and unstressed syllables represented by manner classes, can improve ASR processing. However, our analysis shows that the argument for this hypothesis plays out along very different lines than in Briscoe's study. A hybrid phone-manner lexicon can theoretically benefit ASR because (a) the discriminative power of the lexicon is not reduced greatly, (b) such a representation is a much better model of the types of pronunciation variation seen in spontaneous speech corpora such as Switchboard, and (c) the theoretical average hypothesis space increases only by a little over a factor of 2. This last fact is contrary to Briscoe's finding that the search space would be reduced because it incorporates more realistic assumptions about the detection of stressed versus unstressed syllables.

These experiments were designed primarily to investigate the validity of Briscoe's claims, and thus we attempted to remain true to his model. However, it is clear that our analysis can be extended in several ways. We have begun experimenting with pruning the hypothesis graph to remove unlikely arcs – this would give a more accurate model of the ASR processing that would occur. However, this only makes sense if language model constraints are integrated into the processing, since some word sequences in the graph would be discarded as unlikely. This analysis could also benefit from a more accurate model of the ASR system's transformation between realized phones and lexical representations. This could be achieved by comparing the Gaussian acoustic model distributions in an HMM system or sampling the acoustic model's space (McAllaster et al., 1998). Both of these extensions will be considered in future work.

The results clearly indicate that further investigation and development of a hybrid lexical strategy in an ASR system is worthwhile, particularly for spontaneous speech corpora where the problem of pronunciation variation is most rampant.

Acknowledgments

The authors would like to thank Keith Johnson, Monica Rajamanohar, and Yu Wang for discussion of this work. This work was funded in part by NSF grant ITR-0427413; the opinions and conclusions expressed in this work are those of the authors and not of any funding agency.

References

E. J. Briscoe. 1989. Lexical access in connected speech recognition. In *Proc. 27th Annual Meeting of the Association for Computational Linguistics*, pages 84–90.

R. Carlson, K. Elenius, B. Granström, and H. Hunnicutt. 1985. Phonetic and orthographic properties of the basic vocabulary of five european languages. In *STL-QPSR 1/1985*, pages 63–94, Stockholm. Speech Transmission Laboratory, Dept. of Speech Communication, Royal Institute of Technology.

K. W. Church. 1987. *Phonological Parsing in Speech Recognition*. Kluwer, Dordrecht.

M. Finke, J. Fritsch, and D. Koll. 1999. Modeling and efficient decoding of large vocabulary conversational speech. In *Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation*.

W. Fisher, 1996. *The tsylb2 Program: Algorithm Description*. NIST. Part of the tsylb2-1.1 package.

E. Fosler-Lussier, S. Greenberg, and N. Morgan. 1999. Incorporating contextual phonetics into automatic speech recognition. In *Int'l Congress of Phonetic Sciences*, San Francisco, California.

E. Fosler-Lussier, C. A. Rytting, and S. Srinivasan. 2005. Phonetic ignorance is bliss: Investigating the effects of phonetic information reduction on asr performance. In *Proc. Interspeech*, Lisbon, Portugal.

J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren. 1993. DARPA TIMIT acoustic-phonetic continuous speech corpus. Technical Report NISTIR 4930, NIST, Gaithersburg, MD.

S. Greenberg, J. Hollenbach, and D. Ellis. 1996. Insights into spoken language gleaned from phonetic transcription of the switchboard corpus. In *Proc. 4th Int'l Conference on Spoken Language Processing*. Philadelphia, PA.

R. Jakobson, G. Fant, and M. Halle. 1952. Preliminaries to speech analysis. Technical Report 13, Acoustics Laboratory, Massachusetts Instutite of Technology.

A. Juneja and C. Espy-Wilson. 2004. Significance of invariant acoustic cues in a probabilistic framework for landmark-based speech recognition. In *From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication*, Cambridge, MA. MIT.

K. Kirchhoff. 1998. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In *Proc. 5th Int'l Conference on Spoken Language Processing*, Sydney.

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

K. Livescu, J. Glass, and J. Bilmes. 2003. Hidden feature models for speech recognition using dynamic bayesian networks. In *Proc. 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland.

D. McAllaster, L. Gillick, F. Scattone, and M. Newman. 1998. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. In *Proc. 5th Int'l Conference on Spoken Language Processing*, pages 1847–1850, Sydney, Australia.

G. Miller and P. Nicely. 1955. Analysis of some perceptual confusions among some english consonants. *Journal of Acoustical Society of America*, 27:338–52.

M. Mohri, F. Pereira, and M. Riley, 2001. *AT&T FSM Library*$^{\text{TM}}$ *– General-Purpose Finite-State Machine Software Tools*. AT&T, Florham Park, New Jersey. Available at http://www.research.att.com/sw/tools/fsm.

NIST. 1992. Switchboard Corpus: Recorded telephone conversations. National Institute of Standards and Technology Speech Disc 9-1 to 9-25.

M. Ostendorf. 1999. Moving beyound the 'beads-on-a-string' model of speech. In *1999 IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado.

M. Saraçlar, H. Nock, and S. Khudanpur. 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. *Computer Speech and Language*, 14:137–160.

D. W. Shipman and V. W. Zue. 1982. Properties of large lexicons: Implications for advanced isolated word recognition systems. In *Proc. Int'l Conference on Acoustics, Speech, and Signal Processing*, volume 82, pages 546–549, Paris, France.

K. Stevens. 1981. Invariant acoustic correlates of phonetic features. *Journal of Acoustical Society of America*, 69 suppl. 1:S31.

A. Waibel. 1988. *Prosody and Speech Recognition*. Morgan Kaufmann, San Mateo, California.

S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, 2002. *The HTK Book*. Cambridge Unveristy Engineering Department. http://htk.eng.cam.ac.uk.