Stochastic pronunciation modelling from hand-labelled phonetic corpora

Michael Riley [a,*], William Byrne [b], Michael Finke [c], Sanjeev Khudanpur [b],
Andrej Ljolje [a], John McDonough [b], Harriet Nock [d], Murat Saraclar [c],
Charles Wooters [e], George Zavaliagkos [f]

[a] *AT&T Labs – Research, Room E141, 180 Park Florham Park, NJ 07932, USA*
[b] *Johns Hopkins University, Baltimore, MD, USA*
[c] *Carnegie-Mellon University, Pittsburgh, PA, USA*
[d] *Cambridge University Engineering Department, Cambridge, UK*
[e] *US Department of Defense, Fort Meade, MD, USA*
[f] *BBN, Cambridge, MA, USA*

Abstract

In the early 1990s, the availability of the TIMIT read-speech phonetically transcribed corpus led to work at AT&T on the automatic inference of pronunciation variation. This work, briefly summarized here, used stochastic decision trees trained on phonetic and linguistic features, and was applied to the DARPA North American Business News read-speech ASR task. More recently, the ICSI spontaneous-speech phonetically transcribed corpus was collected at the behest of the 1996 and 1997 LVCSR Summer Workshops held at Johns Hopkins University. A 1997 workshop (WS97) group focused on pronunciation inference from this corpus for application to the DoD Switchboard spontaneous telephone speech ASR task. We describe several approaches taken there. These include (1) one analogous to the AT&T approach, (2) one, inspired by work at WS96 and CMU, that involved adding pronunciation variants of a sequence of one or more words ('multiwords') in the corpus (with corpus-derived probabilities) into the ASR lexicon, and (1 + 2) a hybrid approach in which a decision-tree model was used to automatically phonetically transcribe a much larger speech corpus than ICSI and then the multiword approach was used to construct an ASR recognition pronunciation lexicon. © 1999 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Anfang der neunziger Jahre hat die Verfügbarkeit des phonetisch umschrifteten TIMIT Korpus für gelesene Sprache bei AT&T zur Forschung über die automatische Inferenz von Aussprachemodellen geführt. In dieser Arbeit wurden Decision Trees auf der Basis von phonetischen wie linguistischen Merkmalen trainiert und auf den DARPA North American Business News Korpus für die automatische Erkennung gelesener Sprache angewendet. In jüngerer Zeit wurde für die an der Johns Hopkins Universität ausgerichteten LVCSR Summer Workshops 1996 und 1997 der phonetisch transkribierte ICSI Korpus für spontane Sprache erstellt. Eine Arbeitsgruppe des Workshops 1997 (WS97) hat sich auf Basis dieses Korpus mit der Inferenz von Aussprachemodellen zur Anwendung auf den DoD Switchboard Korpus für spontane Telefonsprache befaßt. Wir beschreiben im folgenden verschiedene Ansätze in diese Richtung. Sie umfassen im einzelnen: ein zu dem AT&T Ansatz (1) analoges Modell, einen Ansatz (2), der auf der Arbeit am WS96

und an der CMU aufbaut und die automatische Generierung und Aufnahme von Aussprachevarianten für Folgen von einem oder mehreren Worten ('multiwords') in das ASR Aussprachewörterbuch (mit aus dem Trainingskorpus abgeleiteten Wahrscheinlichkeiten) behandelt, und ein hybrides Modell $(1 + 2)$, in dem ein Decision Tree basiertes Modell zur automatischen phonetischen Transkription eines sehr viel größeren als des ICSI Korpus verwendet wird um dann mittels des 'multiword'-Ansatzes ein ASR Aussprachewörterbuch zu generieren. © 1999 Elsevier Science B.V. All rights reserved.

Résumé

Au début des années 90, la disponibilité du corpus TIMIT de transcription phonétique de parole dictée conduisit les chercheurs de AT&T à travailler sur l'inférence automatique des variations de prononciation. Ce travail, brièvement résumé ici, utilisa des arbres de décision construit à partir de traits phonétiques et linguistiques, et fut appliqué à la reconnaissance de la parole dictée dans le domaine du DARPA North American Business News. Plus récemment, le corpus ICSI de transcription phonétique de parole spontanée fut mis au point à l'occasion des ateliers de travails d'été à l'université Johns Hopkins (WS) en 1996 et 1997. Un groupe de travail du WS97 mit l'accent sur l'inférence de la prononciation à partir de ce corpus afin de l'utiliser dans le domaine de la parole spontanée téléphonique DoD Switchboard. Nous décrivons plusieurs approches suivies à cette occasion. Ces approches incluent (1) une approche analogue à celle de AT&T, (2) une autre approche inspirée par le travail accompli à WS96 et à CMU qui nécessite l'addition de variantes de prononciation d'une séquence d'un ou plusieurs mots du corpus ('multimots', avec des probabilités extraites de ce corpus) au dictionnaire de prononciation du système de reconnaissance de la parole, et $(1 + 2)$ une approche hybride où un model d'arbres de décision est utilisé pour déterminer automatiquement la transcription phonétique d'un corpus bien plus long que celui d'ICSI, et où l'approche multimots est utilisée pour construire un dictionnaire de prononciation pour la reconnaissance automatique de la parole. © 1999 Elsevier Science B.V. All rights reserved.

1. Introduction

Most speech recognition systems rely on pronunciation dictionaries that contain few alternate pronunciations for most words. In natural speech, however, words seldom adhere to their citation forms. The failure of ASR systems to capture this important source of variability is potentially a significant source for recognition errors, particularly in spontaneous, conversational speech. We report methods used to address this issue applied to read speech at AT&T (Riley and Ljolje, 1995) and to spontaneous speech at and after WS97, the Fifth LVCSR Summer Workshop, held at Johns Hopkins University, Baltimore, in July–August 1997 (Byrne et al., 1998). From our perspective, the primary goal of modelling pronunciation variation is improving the accuracy of automatic speech recognition systems and we will judge it accordingly.

To date, there have been a variety of approaches to the modelling of pronunciation variation. Strik and Cucchiarini (1998) give a taxonomy of these methods. One dimension they use to distinguish between methods is whether the *information source* for the pronunciation models is data-driven (Byrne et al., 1997, 1998; Chen, 1990; Cremelie and Martins, 1997; Finke and Waibel, 1997; Fukada and Sagisaka, 1997; Holter, 1997; Randolph, 1990; Riley and Ljolje, 1995; Weintraub et al., 1996; Wooters and Stolcke, 1994) or knowledge-driven (Downey and Wiseman, 1997; Kipp et al., 1997; Lamel and Adda, 1996; Tajchman et al., 1995). The work from AT&T, in fact, was among the first to apply the data-driven approach, in particular, in its use of hand-labelled phonetic corpora for building statistical pronunciation models. The AT&T work was based on the TIMIT *read*-speech phonetic corpus (Fisher et al., 1987). Applying analogous data-driven methods to American English *spontaneous* speech required the collection of a similarly-sized phonetically-labelled conversational-speech corpus, since none was currently available. This was done for the DoD

Switchboard telephone-speech corpus by the International Computer Science Institute (ICSI) under contract from the LVCSR Summer Workshop (Greenberg, 1996). As such, the work described here is perhaps the first to use a hand-labelled phonetic corpus for building statistical pronunciation models of spontaneous American English speech.

Another dimension Strik and Cucchiarini used to distinguish between methods is whether the *information representation* is formalized (e.g., rules) or enumerated (e.g., an explicit dictionary). We explore both methods (and, as we shall later see, their hybrid). One viewpoint we explore is that pronunciation variability may be modelled by a statistical mapping from canonical pronunciations (baseforms) to symbolic surface forms, and we use decision trees to capture this mapping. A second way we exploit the hand transcriptions is by enhancing the dictionary using frequently seen pronunciations. While the former has the potential to generalise to unseen words and pronunciations, the latter is more conservative and hence potentially more robust.

As many researchers have observed earlier, simply adding several alternate pronunciations to the dictionary increases the confusability of words to the extent that the gains from having them are often more than nullified. We address this problem in two ways. We assign costs to alternate pronunciations so that, e.g., if a frequent pronunciation of 'cause' and an infrequent pronunciation of 'because' are identical, a penalty is incurred to attribute the pronunciation to 'because' rather than 'cause'. In addition, we account for context effects so that, e.g., 'to' is allowed the pronunciation [ax], which is a frequent pronunciation of 'a', only if 'to' is preceded by 'going', as in [g aa n ax].

Our pronunciation modelling efforts may be divided into two broad categories. In our *tree based dictionary expansion* experiments, we apply decision tree based pronunciation models to entries in our baseform dictionary to obtain alternate pronunciations, which are then used in testing. In our *explicit dictionary expansion* experiments, we apply the decision tree based pronunciation models first to the training corpus, and perform a forced alignment with the acoustic models to

'choose' amongst the alternatives. The dictionary is then explicitly augmented with novel pronunciations which occur sufficiently often. The tree based expansion implicitly adds many more new pronunciations than the explicit expansion. The explicit expansion models cross-word coarticulation in a word-specific manner by allowing as dictionary entries a select set (cf. Finke and Waibel, 1997) of *multiwords* – word pairs and triples.

We demonstrate in Sections 2 and 3 that the tree-based method gives a reduction in the word error rate (WER) for the read-speech North American Business (NAB) News task while both methods give reductions for the conversational telephone speech Switchboard task over baseline systems using only a citation-form dictionary. Further, we show in Sections 4 and 5 that reductions persist when the baseline systems are improved by coarticulation sensitive acoustic modelling and improved language modelling.

We felt that it was important for completeness to present here our full set of results. To assist the reader, however, we have placed auxiliary experiments in Appendix A. The reader may wish to skip these on first reading.

2. Tree based dictionary expansion

Our tree based pronunciation models were inspired by phonological rules in acoustic phonetic studies (cf. e.g., Ladefoged, 1975) which characterize allophonic variations in certain phonemic contexts. Fig. 1 illustrates a typical word sequence along with its phonemic transcription (shown at the top), based on a pronunciation dictionary, together with the phonetic (surface form) transcription of an instance of the utterance (shown at the bottom). The phenomenon of deletion of some phonemes or their realisation as a different phone are treated as a probabilistic event and modelled by decision trees. Similar methods or variations based on replacing decision trees with rewrite rules have been successfully used to model pronunciation variability and constraints by other researchers (e.g., Chen, 1990; Randolph, 1990; Finke and Waibel, 1997; Tajchman et al., 1995; Weintraub et al., 1989, 1996). Throughout this paper, we will

I   HAVE   FORMULATED   A

ay   hh   ae   v   f   ow   r   m   y   ax   l   ey   t   ih   d   ax

ay ϕ ϕ v...vl   f   ao   r   m ϕ ih...n   l   ey   dx..t   ih dx...d ax
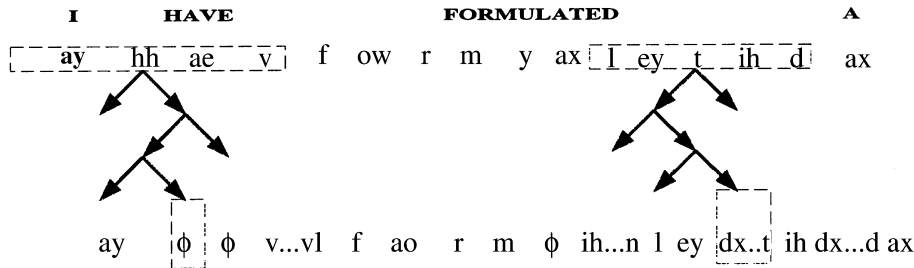
Fig. 1. Decision trees as phone predictors.

use the computer-readable phonetic alphabet ARPABET (Shoup, 1990) to represent phonemes and its TIMIT extensions (Fisher et al., 1987) to represent allophones.

2.1. Decision trees from hand-labelled data

The decision trees built for these tasks drew from the TIMIT and ICSI corpora, which were used as the means to bootstrap our pronunciation models. The TIMIT corpus is read wide-band-width speech that was phonetically annotated under contract from DARPA in the 1980s (Fisher et al., 1987). The ICSI corpus is spontaneous conversational telephone speech that was phonet-ically annotated under contract from WS97 (Greenberg, 1996). Approximately three-quarters (134,000 phones) of the TIMIT corpus and ap-proximately 3.5 hours (96,000 phones) of 4 hour ICSI corpus were used as the training sets in our experiments. The remainder of these corpora were held out for use as test sets needed in some of the experiments reported below.

The labels used by the annotators were slightly richer than the phonetic labels in the pronunciation dictionary (PronLex LDC, 1995) used for these experiments. However, since the acoustic models for the baseline system were trained using the PronLex set, the hand annotations were mapped down to this phone set for reasons of consistency. [1]

Next, based on the orthographic transcriptions and the pronunciation dictionary, a phonemic transcription of the utterances was obtained. Whenever the dictionary permitted more than one pronunciation for a word, a choice was made via a forced alignment [2] of the acoustic signal with the alternatives using the baseline acoustic models.

The phonemic transcriptions were then lined up with the phonetic labels, using as the alignment criterion the minimization of the phonetic feature distance between the two symbol streams (Riley and Ljolje, 1995). Table 1 gives an example alignment from the ICSI corpus. This gave us a corpus of *phoneme-to-phone* transformations to-gether with the phonemic environment or *context* for each instance.

Decision tree models were then built to repre-sent this phoneme to phone mapping. The tree growing criterion was minimization of the empir-ical entropy of the surface phone, the stopping criterion was a minimum sample count at both parent and child nodes, and the trees were pruned via internal five-fold cross-validation (Brieman et al., 1984). A separate tree was grown for each phoneme. The context for decision making in-cluded the identity of the phoneme to be mapped as well as three neighbouring phonemes on either side, the lexical stress on neighbouring vowels as

---

[1] The ICSI labels were first stripped of diacritics. Glottal stops were then deleted and non-PronLex syllabic consonants were expanded into schwa plus consonant (e.g., [em] → [ax m]). Any remaining non-PronLex labels were mapped to their most phonetically similar PronLex labels.

[2] Note that time-alignment was not a goal of the alignment process. Since the PronLex dictionary contains a single pronunciation for an overwhelming majority of the words and a few (2–3) for most others, we feel that the forced alignment procedure was not of much significance other than a sensible means to resolve the few pronunciation ambiguities that did arise.

Table 1
Phoneme-to-phone alignment for an ICSI transcription fragment

Phoneme	Phone	Word
f	f	for
ao	ao	
r	–	
dh	dh	the
ax	iy	
f	f	for
ao	–	
r	–	
dh	dh	the
ax	ax	
d	jh	drug
r	r	
ah	ah	
g	g	

Table 2
Phoneme encoding scheme

●*Consonant-manner*
voiced stop, unvoiced stop, voiced fricative, unvoiced fricative, voiced affricate, unvoiced affricate, nasal, rhotic, lateral, not–applicable.

●*Consonant-place*
bilabial, labiodental, dental, alveolar, palatal, velar, pharyngeal, not–applicable.

●*Vowel-manner*
monophthong, r–colored vowel, w–diphthong, y–diphthong, glide, not–applicable.

●*Vowel-place*
front–low, front–mid–low, front–high, central–mid–low, central–mid–high, back–low, back–mid–low, back–mid–high, back–high, not–applicable.

obtained from the pronunciation dictionary, and the distance of the phoneme from the nearest segment boundary on either side, where the segment could be either a syllable, a word or a phrase. We also investigated the effect of leaving out a few of these features as well as adding a few others, and the results of this investigation are presented in subsequent sections.

The identity of each phoneme is specified by an encoding in terms of its phonetic features, using the four-element feature vector of (`consonant-manner, consonant-place, vowel-manner, vowel-place`) (see Table 2). For example, /s/ is encoded as (`voiceless fricative, palatal, n/a, n/a`) and /iy/ is encoded as (`n/a, n/a, y-diphthong, high-front`). This encoding was chosen to guide the decision trees in asking concise, phonetically-relevant questions about phoneme identity (Riley and Ljolje, 1995). Given that there are forty-odd distinct English phonemes, some knowledge-based encoding as we have used or some data-driven clustering (Chou, 1990) is necessary, since the decision tree cannot entertain all 2^{40} possible binary partitions of the phoneme set per tree node split (Brieman et al., 1984).

2.1.1. Predicting surface forms from baseforms

Each leaf in a tree thus assigned probabilities in some context to more than one surface form

realisation of the phoneme it modelled. A way to judge the goodness of these trees, therefore, was to apply them as predictors on a held out portion of the hand-labelled corpora. Test sets from TIMIT and ICSI corpora were held out for this purpose. The results in Table 3 summarize the predictive ability of the trees on these sets. [3] Relative to the context independent distribution of surface form realisations of a phoneme, decision trees built on the TIMIT portion of the trainig set reduce the entropy by about 50%, when tested on TIMIT. Those built on the Switchboard portion of the training set reduce the entropy by about 30%, when tested on Switchboard. Trees based on TIMIT alone are much less effective on the Switchboard test set (20%), but adding them to the Switchboard training data (ICSI + TIMIT) results in a small additional gain (32%). These results suggest there is more variability in pronunciations in Switchboard, relative to TIMIT, which is not captured by either the phonemic context cues or the modelling paradigm we considered.

Several additional experiments were conducted to compare the relative importance of the features, e.g., extent of phonetic neighbourhood or lexical

---

[3] So that test observations in contexts unseen in training do not make entropy figures infinite, the worst 10% of the test data (i.e., highest $\log_2$-prob) is removed from each entropy measurement in this paper.

Table 3
The entropy (in bits) of predicting the surface-form of a phoneme[a]

Training data	Average $\log_2$-prob (efficiency)	
	ICSI-test	TIMIT-test
TIMIT	$0.76 \rightarrow$ 0.60 (20%)	$0.34 \rightarrow 0.17$ (51%)
ICSI	$0.72 \rightarrow$ 0.50 (30%)	
ICSI+TIMIT	$0.71 \rightarrow$ 0.48 (32%)	

[a] The number preceding the arrow corresponds to prediction based on the relative frequency of the surface form and the one following the arrow corresponds to using a decision tree pronunciation model. Numbers in parentheses indicate relative reduction in entropy.

stress, that we have used as context in out pronunciation models. Some experiments were also conducted to quantify the importance of additional features which have not been used here. Though these results are of significant interest, we relegate them to Appendix A.1 for clarity of presentation.

2.2. Generating automatic phone transcriptions

Using decision trees is a data intensive modelling technique. Large quantities of automatic phonetic transcriptions were generated to augment the hand-labelled corpora using the 37,000 sentence acoustic training corpus from 284 speakers for the NAB task at AT&T and using the 60-hour acoustic training corpus for the Switchboard task at WS97. Unlike (Weintraub et al., 1996), where unconstrained phone recognition was used to generate phone transcriptions, we constrained the words in our training utterances to assume only pronunciations generated by application of the

decision trees to their phonemic baseforms: a forced alignment was performed on the resulting network of alternate pronunciations in an utterance and the most likely sequence of pronunciations was chosen to be the phonetic transcription. Pronunciation probabilities derived from the trees were used as 'language model' weights during alignment, and since the word transcription was provided, word level language model weights are redundant and were not used.

Anecdotal evidence suggests that this method of obtaining automatic transcriptions is reasonable: it agrees more with human annotations than the phonemic baseforms do, though not by much. For the hand-labelled portion of the ICSI corpus, for instance, we aligned the baseforms with the hand labels and found the *phone error rate* of the citation form transcription to be about 30%. The error rate for the automatic transcriptions for the same portion was 25%.

It is also not clear if total agreement with the hand labels is desirable. Occasionally, as in the transcriptions shown in Table 4, a large number of human listeners preferred the automatic transcriptions to those of the annotators (who matched the dictionary in this case)! Readers who would wish to listen to this particular utterance can find it on the 1997 LVCSR workshop pronunciation project web page at www.clsp.jhu.edu.

2.2.1. Building decision trees from automatic transcriptions

These transcriptions were used to build new decision trees. Two options were explored to use this large amount of training data – retain the topology (i.e., the sequence of questions, or the equivalence classification of the contexts) of the original phonetically hand-labelled corpus trees, and only update their leaf distribution by *pouring* this new training data down those trees; or

Table 4
Comparing phonemic (dict), manual (icsi) and automatic (auto) phonetic transcriptions

words	just	because	they're	grandparents $\cdots$
dict	jh ah s t	b ax k ah z	dh ey r	g r ae n p ey r ih n t s
icsi	jh ah s t	b ax k ah z	dh ey r	g r ae n p ey r ih n t s
auto	jh ax s	b ax k ao z	er	g r ae n p eh r s

Table 5
Rebuilding versus retuning the pronunciation trees

Model	# Trn tokens	$\log_2$-prob
ICSI+TIMIT	96,040	0.525
Recount weights	2.36 million	0.585
Rebuild trees	2.36 million	0.542

rebuild the trees altogether. When applied to Switchboard, there is very little difference between the two methods in terms of prediction entropy on a held out set, as illustrated in Table 5.

It is also not surprising that the prediction entropy of these trees is higher than the ICSI+TIMIT trees trained on hand labels alone, because there is an obvious mismatch between the automatic derivation of the training transcriptions, and the hand-labelling of the test set. The fully rebuilt trees were named *Retrained trees*.

Since we now had much more training data, we also built trees which additionally included in the context the previously realised surface form so as to capture some of the dependency in the surface string. Trees built this way were named *Retrained2 trees*.

2.3. Dictionary expansion using pronunciation trees

We applied the ICSI+TIMIT trees of Table 3 to successive phonemes of each baseform in the WS97 baseline dictionary to obtain a weighted pronunciation network as described in (Riley and Ljolje, 1995). Fig. 2 illustrates such a network for the word `pretty`.

Applied statically, this resulted in an expanded dictionary which we call the *ICSI+TIMIT dictionary*.

We also applied the Retrained trees to baseforms in the baseline dictionary as before, to obtain a second enhanced dictionary, which we call the *Retrained dictionary*. Finally, expanding the baseforms in the baseline dictionary using the Retrained2 trees resulted in the *Retrained2 dictionary*.

2.4. Testing with tree based dictionaries

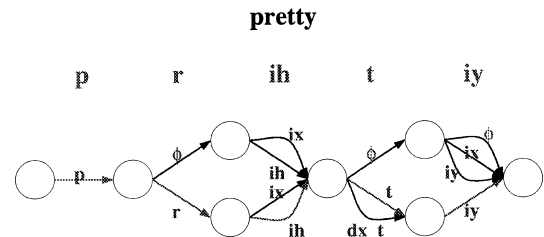At AT&T, both trees built on TIMIT and retrained trees built on the automatically transcribed

Fig. 2. Pronunciation network for `pretty`.

Table 6
NAB recognition results with enhanced dictionaries

Dictionary	WER (%)
TTS	12.7
TIMIT	10.8
Retrained2	10.0

NAB 284-speaker acoustic training corpus were used to constuct recognition dictionaries for the NAB Eval '95 test set. These were compared with a baseline system whose pronunciations came from the AT&T TTS text-to-speech system (Coker, 1985). Table 6 shows these recognition results. We see that the TIMIT-based trees gave a 1.9% WER reduction over the citation-form TTS dictionary, while trees retrained on the NAB 284-speaker acoustic training corpus gave an additional 0.8% reduction. In this earlier work, the full TIMIT phone set (minus the stop closures) was used, which contained 53 phones compared to the TTS inventory of 41 phonemes. Thus, new acoustic models had to be built for the larger phone set. In other words, the acoustic models used for the TIMIT and Retrained2 entries in Table 6 were different than the TTS-based test. At WS97, this was not required, since as mentioned before, the phone realisations had been forceably mapped to the PronLex set.

At WS97, bigram lattices for the WS97 development-test were rescored using the enhanced dictionaries described above using the WS97 baseline acoustic models. [4] Table 7 shows recognition

---

[4] The baseline acoustic models were state clustered cross-word triphones comprising about 7000 states, each with twelve-component Gaussian mixture output densities, trained on about sixty hours of Switchboard data.

Table 7
Switchboard recognition results with enhanced dictionaries

Dictionary	WER (%)	DEL (%)	SUB (%)	INS (%)
PronLex	44.66	10.85	29.47	4.34
ICSI + TIMIT	46.14	11.65	30.39	4.10
Retrained	43.99	10.90	29.08	4.02
Retrained2	43.75	10.87	28.85	4.02

performance using the three static but weighted enhancements to the dictionary based on the ICSI + TIMIT trees, the Retrained trees and the Retrained2 trees.

The degradation in performance from the ICSI + TIMIT dictionary came as a surprise, especially since the AT&T NAB experiments showed an apparently opposite effect. There were, however, many differences between the two tests including (1) a read speech, low error task versus a spontaneous speech, high error rate task, (2) the TTS-dictionary versus the PronLex dictionary, (3) 53 phone set versus a 43 phone set, (4) potentially different phonetic transcription conventions between the TIMIT and ICSI labellers and (5) acoustic model retraining in the AT&T experiments but not in the WS97 experiments.

Regarding (2): the TTS dictionary (using the 41 ARPABET phoneme set) has few alternative pronunciations and favors citation forms (Coker, 1985), while the PronLex dictionary (using the 41 ARPABET phonemes plus syllabic l and n), has more alternative pronunciations, some containing (in principle) predictable variants (LDC, 1995). At AT&T, we used the PronLex dictionary to build an alternate NAB system and we used the TTS dictionary to build an alternate Switchboard system. We found little difference in the baseline performances (within 0.2% WER) of the alternate dictionary systems compared to their respective original dictionary systems. Regarding (5): preliminary attempts at WS97 to retrain acoustic models using tree-based pronunciation lexicons lead to significantly worse results (Byrne et al., 1997).

There were various conjectures made why the ICSI + TIMIT dictionary gave a worse result and we launched a series of experiments to investigate them. Again, while the results of these experiments are of significant interest, we relegate their descriptions to Appendix A.2 in order to maintain clarity of presentation. From Table 7, we also see that the Retrained and Retrained2 trees yielded modest but statistically significant improvements in word error rate over the WS97 baseline system. [5]

2.5. Summary of tree based experiments

- Pronunciation probabilities based on TIMIT trees for NAB helped performance (+1.9%) and reestimated trees helped more (+2.7%).
- Pronunciation probabilities based on ICSI + TIMIT trees for Switchboard hurt performance (−1.5%), but those from reestimated trees help (+0.9%).
- Reestimated pronunciation probabilities still don't agree with empirical frequencies in training (see Appendix A.2). Word level pronunciation probabilities should be examined.
- Words have variable tendencies to be mispronounced. Perhaps all words in the dictionary should not be expanded equally (see Appendix A.2).

3. Explicit dictionary expansion

The degradation in performance due to the ICSI + TIMIT dictionary admits the possibility that the ICSI + TIMIT trees either generalise incorrectly or do a poor job of assigning costs to the alternate pronunciations. Both of these are crucial to the success of dictionary enhancement based methods. An alternate, more conservative

[5] The 95% confidence interval on the word error rate on our test set of 18,198 words is approximately 0.7% (assuming a binomial distribution on the word errors).

Table 8
Lattice-rescoring with explicitly expanded dictionaries

Dictionary	WER (%)	DEL (%)	SUB (%)	INS (%)
PronLex	44.7	10.9	29.5	4.3
ICSI multiword	44.6	10.3	29.7	4.6
Auto multiword	43.8	10.4	29.1	4.3

approach to dictionary enhancement was therefore examined at WS97. As such, all experiments from here on apply to Switchboard.

3.1. ICSI multiword dictionary

The PronLex dictionary is first enhanced with all the pronunciations for words seen in the hand-labelled (ICSI) portion of the corpus. A candidate list of 172 multiwords (cf. Finke and Waibel, 1997) is also appended to the dictionary to capture coarticulation, and pronunciations for these are similarly appended using the hand-labelled corpus. The word transcription of the training corpus is then expanded using these alternate pronunciations and aligned with the acoustics using our baseline models. New pronunciations which are chosen sufficiently often (in at least 20% of the tokens for each type) are deemed *bona fide* entries to the *ICSI Multiword dictionary*; the others are discarded. Pronunciations are assigned weights based on their relative frequency.

3.2. Auto multiword dictionary

Instead of the forced alignment among alternate pronunciations extracted from the hand-labelled portion of the corpus as described above, new pronunciations for words and multi-words may be chosen from the large automatically transcribed corpus described in Section 2.2. This alternative approach yields the *Auto Multiword dictionary*. Qualitatively speaking, this dictionary invokes the decision tree pronunciation models to generate alternatives, but keeps only those which occur frequently enough in the automatic transcription. Again, weights are assigned to each pronunciation based on its relative frequency.

3.3. Recognition results using expanded dictionaries

Bigram lattices for the WS97 dev-test, generated using the PronLex dictionary, are rescored using the enhanced dictionaries described above. The multiwords are incorporated into the lattices as additional lattice arcs that parallel their corresponding single word arcs. These multiword arcs are correctly weighted so that sentence hypotheses receive the same language model score whether or not a multiword pronunciation is chosen. Table 8 shows recognition performance using the two dictionaries. The 0.9% improvement due to the Auto Multiword dictionary is encouraging, particularly in contrast to the lack of improvement obtained from the ICSI Multiword dictionary. This comparison further reinforces the impression that the hand-labelled data is good for bootstrapping, but not reliable enough for directly estimating pronunciation models. At the least, incorporation of human expert knowledge into statistical information processing systems has been shown again to be a difficult problem in which naive approaches do not work as well as the modelling techniques that match the supplied knowledge to the capabilities of the system.

4. Coarticulation sensitive clustering

Context dependent acoustic models such as triphone HMMs are capable of implicitly modelling some allophonic variation. However, the models in our baseline system do not distinguish between word-internal and cross-word triphones, and one may hypothesise that the gains above, especially those from the Multiword experiments, are due to better modelling of common cross-word effects. To investigate this possibility, the triphone clustering procedure in our HTK-based system is enhanced, as described next (Young et al., 1995).

Table 9
Lattice-rescoring with new AMs

Dictionary	WER (%)	DEL (%)	SUB (%)	INS (%)
Baseline acoustic models				
PronLex	43.4	9.8	29.4	4.1
INTWBD acoustic models				
PronLex	41.8	10.1	27.8	3.9
Retrained2	41.3	10.2	27.5	3.7
Auto multiword	41.1	9.7	27.5	4.0

The major deviation from the baseline system is to mark the phones in the the PronLex dictionary to permit acoustic triphone state clustering routines to make explicit use of information about word boundary location. Another important modification is the use of a specific interjection phone set. This is not so much to model interjections better as to prevent the very frequent interjections from overwhelming the clustering and modelling of phones in noninterjections. Acoustic model training is carried out in the same manner as the baseline system, with the difference that the question set for triphone state clustering is augmented with questions regarding the word boundary tags and interjection phone set. A set of acoustic models, named the *INTWBD models*, comparable to the baseline in terms of the number of states and Gaussian components, is thus estimated.

Next, the training data is retranscribed using these models and the pronunciation networks of Section 2.2. The Retrained2 dictionary and the Auto Multiword dictionary of Sections 2.2 and 3.2, respectively, are then regenerated from these transcriptions.

4.1. Recognition results using improved acoustic models

Table 9 shows the results [6] of rescoring the WS97 dev-test set using the INTWBD acoustic

models, and indicates that enabling the state clustering to take advantage of word boundary information and separate phones for interjections result in significant improvement in performance (1.6%). Observe that the two dictionary enhancement techniques continue to provide added improvements (0.7%), though to a slightly smaller extent now.

5. Language model improvements

In the spirit of investigating whether pronunciation modelling via the two expanded dictionaries continues to be of benefit when other components of the system are improved, lattices generated by a bigram language model and the baseline PronLex dictionary are rescored using a trigram language model and the Retrained2 and Auto Multiword dictionaries. The results in Table 10 are therefore directly comparable with those in Table 9, which are based on bigram scores.

Observe that the improvement from the INTWBD models over the baseline models is 1.5%, which matches the 1.6% improvement with the bigram language model. The additional improvement of 0.5% from the Retrained2 dictionary also continues to hold, and the improvement from the Auto Multiword dictionary over the PronLex dictionary actually increases from 0.7% to 0.9%. All these results indicate that our straightforward pronunciation models and the coarticulation sensitive acoustic modelling provide gains which are additive to language model improvements.

---

[6] Though these results are for the same baseline system and test set, the baseline performance here differs slightly from the one shown in Tables 7 and 8. This is mostly due to a change in the acoustic segmentation of the test set between the two experiments, evidently for the better, and to a smaller extent due to a small change in the scoring software.

Table 10
Lattice-rescoring with new AMs and a trigram LM

Dictionary	WER (%)	DEL (%)	SUB (%)	INS (%)
Baseline acoustic models				
PronLex	40.9	8.9	27.8	4.2
INTWBD acoustic models				
PronLex	39.4	9.2	26.2	4.0
Retrained2	38.9	9.2	25.9	3.8
Auto multiword	38.5	8.6	25.8	4.2

5.1. The effect of multiwords

The performance of the Auto Multiword dictionary with the INTWBD acoustic models given in Table 10 is the best obtained in the experiments we report here. To determine the actual contribution of the multiword pronunciation models to this result, we removed the multiwords from the Auto Multiword dictionary and rescored the trigram lattices using the INTWBD models. We found that the 38.5% result did not change. From this we conclude that while it is possible to identify and obtain pronunciations for multiwords using automatic methods, including multiwords in lattice rescoring experiments neither helps nor hurts performance. This is an encouraging result, in that it suggests the benefits we see are due to newly learned pronunciations and not merely from additional model complexity. We note also that this result does not contradict earlier work which found gains from multiwords (Finke and Waibel, 1997). Multiwords do make it possible to selectively incorporate cross-word acoustic context in single pass recognition systems.

6. Acoustic model retraining

The baseline as well as the INTWBD acoustic models are trained on the PronLex dictionary, prompting the concern that these models are not appropriate for use with the new dictionaries. In particular, given the prevalence of reduced variants in the new dictionaries, the acoustic contexts upon which the triphone states are clustered in the baseline system are suspected to be poorly matched to the new dictionaries. This section describes a procedure used to retrain models better matched to the ICSI Multiword dictionary.[7] This work makes use of the training techniques developed by the Hidden Pronunciation Mode group at the 1996 LVCSR Workshop.

First, the state clustered triphone INTWBD models and the regenerated ICSI Multiword dictionary of Section 4 are used to obtain a phonetic transcription of the corpus, which then remains fixed during training. Untied triphones for this transcription are then cloned from the monophone HMMs created during the training of the baseline system. Finally, the training procedure for the INTWBD models is mimicked starting with triphone HMM reestimation, followed by state clustering, etc. The resulting HMMs, comparable in the number of states and Gaussian components to the baseline system, are called *MWINTWBD models*.

6.1. Recognition results using retrained acoustic models

Bigram lattices for the WS97 dev-test, generated using the baseline acoustic models and the PronLex dictionary, are rescored using the MWINTWBD acoustic models and the ICSI

---

[7] Acoustic retraining was not on our best (Auto Multiword) dictionary for historical reasons: the ICSI Multiword dictionary was obtained first, and a retraining effort was started before the superiority of the Auto Multiword dictionary was established.

Table 11
Lattice-rescoring with retrained acoustic models

Model	WER (%)	DEL (%)	SUB (%)	INS (%)
Bigram LM				
INTWBD	41.8	10.1	27.8	3.9
MWINTWBD	41.3	9.6	27.5	4.2
Trigram LM				
INTWBD	39.4	9.2	26.2	4.0
MWINTWBD	39.0	8.7	26.1	4.2

Multiword dictionary. Table 11 shows the results of the rescoring experiment.

Recall from Table 8 that the ICSI Multiword dictionary gives essentially no gain by itself, and thus the gain here (0.4%) may be attributed to the acoustic retraining. It is expected that substantially higher gains will be attained by acoustic retraining with better phonetic transcription such as those obtained using the Auto Multiword dictionary.

7. Conclusion

This research suggests that significant improvement in speech recognition can be made by suitably modelling systematic pronunciation variation. Further, our results indicate that while a hand-labelled corpus is very useful as a bootstrapping device, estimates of pronunciation probabilities, context effects, etc., are best derived from larger amounts of automatic transcriptions, preferably done using the same set of acoustic models which will eventually be used for recognition.

On NAB, using pronunciation modelling with acoustic model retraining, we saw a 2.7% reduction in WER over a TTS baseline system. On Switchboard, without acoustic model retraining, we saw a 0.9% reduction in WER over a PronLex baseline system, which is demonstrably additive to improvements in language (2.5%) and acoustic (1.5%) modelling, and to gains from adaptation (not reported here). Work is underway to develop effective acoustic model retraining methods for Switchboard when these statistical pronunciation lexicons are employed.

While we were heartened by the improvements seen and the knowledge gained, there are nevertheless many issues and details left unexplored. Briefly, these include:

- Effective acoustic model retraining when using pronunciation models for spontaneous speech;
- Improved cross-word and phone-realisation dependency modeling with the decision tree approach;
- Generalisation to unseen contexts with the multiword approach;
- Dynamic pronunciation modelling – e.g., effective use of duration/rate;
- Speaker normalization/clustering in pronunciation modelling;
- Discriminative training of the pronunciation models to reduce confusability.

We conclude by mentioning that our effort at the WS97 Workshop, which involved individuals from various locations with a common interest in pronunciation modelling, not only produced some answers, but raised questions and formed relationships that will hopefully serve to further our understanding of pronunciation modelling in the future.

Acknowledgements

Appendix A

A.1. Useful features for predicting pronunciations

The tree based models of Section 2.1.1 were studied to evaluate the relative merits of the features used for predicting alternate pronunciations, as also to gauge the gain from using additional features not used in the models described in Section 2.1.1.

Table 12
Leaving out features from the context

Features provided as context	$\log_2$-prob
All features	0.485
Third phoneme from the left and right excluded	0.484
Second and third phonemes from the left and right excluded	0.485
Lexical stress excluded	0.487
Segment boundary cue excluded	0.490
Vowels (manner and place) excluded	0.497
Stress and segment boundary cues excluded	0.498
Consonants (manner and place) excluded	0.527
Right phonemic context excluded	0.537
Left phonemic context excluded	0.547
Entire phonemic context excluded	0.606
All context excluded (root trees)	0.714

A.1.1. The effect of leaving out features

In order to investigate the conditional utility of each of our contextual features given the others, trees were built at WS97 by leaving features out from the context one at a time. Table 12 summarizes the results of these experiments. The trees were trained on all of the Switchboard and TIMIT data mentioned above, and the test set was the same as the one used for the Switchboard results of Table 3.

Note that, at least for this corpus size, there was little additional predictive power in neighbouring phonemes more than one position away, when the triphone context, word boundary, and lexical stress related information was specified.

It seems reasonable to conclude that most of the modeling gain for the pronunciation trees comes from the immediate $+/-1$ phonemic context, lexical stress and segment boundary location information. The contribution of the other features is small, and they may be excluded if computational resources are an issue.

A.1.2. The impact of some additional features

We also experimented at WS97 with adding new features to the decision trees.

• Based on the number of distinct pronunciations of a word that were seen in the ICSI-portion of the corpus, words were categorized into ten bins: from words having many pronunciations to words having few pronunciations. The bin number was then provided to the trees for each phoneme of the word. It was hoped that knowing how stable a word's pronunciation was would help predict the surface form better.

• It is varyingly conjectured that frequently used words, function words or low information bearing words often tend to be mispronounced. The frequency of occurrence of a word in the 60 hour acoustic training corpus was provided to the trees for each phoneme of the word.

• In the hope of capturing limited phonotactics, as well as to indirectly model deletion or reduction of units larger than phonemes, the trees were provided the surface form realisation of the previous phoneme.

Table 13 summarizes our results. All the features used by the ICSI+TIMIT trees of Table 3 are already present in the context. Note that while we were unable to successfully exploit the information about empirical pronunciation variability or frequency of a word, knowing the previous surface form seems to be of significant value in this modelling paradigm, perhaps because it compensates for some of our conditional independence assumptions in modelling the phoneme to phone mapping very locally.

A.2. Diagnosis of models based exclusively on hand-labelled data

It was noted in Section 2.4 that unlike our experience on the NAB corpus, where a pronunciation model constructed from the TIMIT corpus was helpful in reducing recognition errors, the model build from the ICSI+TIMIT labelled data resulted in higher error rates on the Switchboard

Table 13
Adding new features to the context

Features added to context	$\log_2$-prob
None	0.485
Word level pronunciation variance (10 bins)	0.481
Word frequency (from 60 hour training)	0.483
Pronunciation variance and word frequency	0.483
Pronunciation variance, word frequency and previous phone	0.451

Table 14
Scaling pronunciation probabilities

Dictionary	Weights	WER (%)	DEL	SUB	INS
PronLex	–	44.66	1987	5398	796
ICSI+TIMIT	$\sum = 1$	46.14	2134	5568	751
ICSI+TIMIT	max $= 1$	46.13	1904	5653	893

Table 15
Word-internal versus cross-word pronunciation modelling

Dictionary	Context	WER (%)	DEL	SUB	INS
PronLex	–	44.66	1987	5398	796
ICSI+TIMIT	None	46.14	2134	5568	751
ICSI+TIMIT	1 Phone	46.09	2112	5590	743

corpus. We investigated several possible causes for this effect and they are reported below.

A.2.1. Are words with many pronunciations being penalized?

It is conceivable that a word such as `and`, which admits many pronunciations, may be unnecessarily penalized relative to a word with few pronunciations such as `an`. E.g., the phones [ae n] are the most likely pronunciation for both `an` and `and` in conversational speech. Since they have a much higher likelihood amongst pronunciations of `an` than amongst those of `and`, other things being equal, it costs less to map these phones to the word `an`.

If Viterbi decoding is employed, many researchers have suggested that this problem is alleviated by scaling the pronunciation probabilities of every word so that the most likely pronunciation has unit weight. [8] We scaled our enhanced ICSI+TIMIT dictionary in this manner, and found an insignificant gain (see Table 14), ruling

this out as the major cause of the degradation in performance.

A.2.2. Do the trees badly need crossword context?

Recall that the enhanced dictionaries were obtained by applying the pronunciation trees to baseforms in isolation, and thus they could not utilise crossword context. We wrote additional software utilities so that the pronunciation model could be applied to the bigram lattices directly. However, looking at three neighbouring phonemes across word boundaries would have resulted in a drastic expansion of the lattice. We therefore implemented crossword pronunciation trees which looked at only one neighbouring phoneme in the context. This, we expected, would not be a severe limitation in light of the fact (from Table 12) that the deleted context is of little additional value in prediction.

The results in Table 15 indicate that this too is not the reason for the poor performance of the ICSI+TIMIT dictionary. We conjecture that crossword pronunciation context is perhaps more important for some words than others (e.g., `and I`, `want to`).

A.2.3. Are the trees generalising too much?

The motivation for using local decision tree based models is to be able to observe phoneme to

---

[8] This would perhaps be unnecessary if the likelihoods were properly summed over all pronunciations of a word, but is a sensible adjustment for Viterbi decoding, as is the additional scaling of the pronunciation probabilities by the language model scale (12) to bring them on par with the acoustic scores.

Table 16
Expanding only the most frequent words using trees

Dictionary (# expanded)	WER (%)	DEL	SUB	INS
PronLex	44.66	1987	5398	796
ICSI + TIMIT (All words)	46.14	2134	5568	751
ICSI + TIMIT (Top 100)	45.50	2213	5456	666

phone transformations which are universally applicable. However, it may be argued that since many words exhibit remarkably stable pronunciations in the hand-labelled data set, the pronunciation model when applied to these words creates confusion without generating useful new pronunciations. We therefore expanded only the hundred most frequent words in the corpus using the ICSI + TIMIT trees, and tested using this instead of the ICSI + TIMIT dictionary.

As the results in Table 16 indicate, this is a significant improvement over expanding all dictionary entries, and should be investigated further. However, the performance continues to be below that of the baseline system. This suggests that expanding only the 100 most frequent words simply brings the system closer to the baseline, and the recognition performance tracks this regression.

A.2.4. Can the weights in dictionary be improved?

Application of the decision tree model one phoneme at a time entails a conditional independence assumption between the surface forms given the baseforms, much as in a hidden Markov model (HMM). Thus the resultant probability of a pronunciation (obtained as a product of the conditional probabilities of the surface phones) is, at best, as reflective of the observed frequency of the pronunciation as the goodness of this independence assumption. To check this, we compared the probabilities of the pronunciations in the ICSI + TIMIT dictionary for a few hand-picked words with their relative frequency in our automatic transcriptions. Table 17 suggests that the tree probabilities, and perhaps the independence assumption as well, are very unsatisfactory. Much room for research and improvement remains here.

Since the Retrained trees were based on much more data (which also happened to be the same data from which the empirical probabilities of the

Table 17
Empirical versus ICSI + TIMIT dictionary probabilities

Pronunciation	Probability	
	ICSI + TIMIT dictionary	Empirical
WANT TO: [w aa n t t ax]		
w aa n ax	0.04	0.34
w aa n t ax	0.20	0.28
w aa t t ax	0.05	–
WANT TO: [w ah n t t ax]		
w ah n ax	0.05	0.37
w ah n t ax	0.26	–
w ah n t ah	0.06	–

Table 18
Empirical versus retrained dictionary probabilities

Pronunciation	Probability	
	Retrained2 dictionary	Empirical
WANT TO: [w aa n t t ax]		
w aa n ax	0.08	0.34
w aa n t ax	0.49	0.28
w aa n t uw	0.08	–
WANT TO: [w ah n t t ax]		
w ah n ax	0.10	0.37
w ah n t ax	0.61	–
w ah n t uw	0.10	–

pronunciations were inferred), we conducted a similar comparison for the Retrained2 dictionary.

The example in Table 18 further reinforces our conclusion that it is the HMM-like independence assumption more than the leaf probability estimation which skews the tree based pronunciation probabilities away from their empirically observed values. Alternative probability assignments at the surface string level should be investigated in the future.

We also conducted an experiment, which clearly brings out the importance of correct pronunciation weight estimation even when the HMM-like independence assumption *is* made. Since we were

Table 19
Impact of reestimating pronunciation tree probabilities

Dictionary	Weights	WER (%)	DEL	SUB	INS
PronLex	–	44.66	1987	5398	796
ICSI+TIMIT	ICSI+TIMIT	46.14	2134	5568	751
ICSI+TIMIT	Retrained	44.05	1982	5351	736

not satisfied with the pronunciation probabilities of the ICSI+TIMIT trees, we poured the 60 hours of automatically transcribed data down the trees and reestimated the leaf distributions, as described in the context of Table 5. These trees continued to assign mismatched pronunciation probabilities to words, much as above, but they had considerably better recognition performance, as indicated in Table 19.

References

Brieman, L. et al., 1984. Classification and Regression Trees. Wadsworth & Brooks, Monterey, CA.

Byrne, W. et al., 1997. Pronunciation modelling for conversational speech recognition: A status report from WS97. Presented at the 1997 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara, CA, December.

Byrne,W. et al., 1998. Pronunciation modelling using a hand-labelled corpus for conversational speech recognition. In: Proc. ICASSP '98, Seattle, WA.

Chen, F., 1990. Identification of contextual factors for pronunciation networks. In: Proc. ICASSP '90, S14.9.

Chou, P., 1988. Applications of information theory to pattern recognition and the design of decision trees and trellises. Ph.D. thesis, Stanford University, Stanford, CA.

Coker, C., 1985. A dictionary-intensive letter-to-sound program. J. Acoust. Soc. Amer. 78 (1).

Cremelie, N., Martins, J.P., 1997. Automatic rule-based generation of word pronunciation networks. In: Proc. Eurospeech '97, pp. 2459–2462.

Downey, S., Wiseman. R., 1997. Dynamic and static improvements to lexical baseforms. In: Proc. Eurospeech '97, pp. 1027–1030.

Finke, M., Waibel, A., 1997. Speaker mode dependent pronunciation modelling in large vocabulary conversational speech recognition, In: Proc. Eurospeech '97.

Fisher, W., Zue, V., Bernstein, J., Pallet, D., 1987. An acoustic-phonetic data base. J. Acoust. Soc. Amer. 81 (1).

Fukada, T., Sagasaki, Y., 1997. Automatic generation of a pronunciation dictionary based on a pronuniciation network. In: Proc. Eurospeech '97, pp. 2471–2474.

Greenberg, S., 1996. The switchboard transcription project. In: 1996 LVCSR Summer Workshop Technical Reports, http://www.icsi.berkeley.edu/real/stp/.

Holter, T., 1997. Maximum likelihood modelling of pronunciation in ASR. Ph.D. thesis, Norwegian University of Science and Technology, December.

Kipp, A. et al., 1997. Automatic detection and segmentation of pronunciation variants in german speech corpora. In: Proc. Eurospeech '97, pp. 1023–1026.

Ladefoged, P., 1975. A Course in Phonetics. Harcourt Brace Jovanovich, New York.

Lamel, L., Adda, G., 1996. On designing pronunciation lexicons for LVCSR. In: Proc. ICSLP '96.

Linguistics Data Consortium. 1995. COMLEX English Pronunciation, http://www.ldc.upenn.edu/ldc/catalog/html/lexical_html/comlexep.html.

Randolph, M., 1990. A data-driven method for discovering and predicting allophonic variation. In: Proc. ICASSP '90, S14.10.

Riley, M., Ljolje, A., 1995. Automatic generation of detailed pronunciation lexicons. In: Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer, Boston.

Shoup, J., 1990. Phonological aspects of speech recognition. In: Lea, W. (Ed.), Trends in Speech Recognition. Prentice-Hall, New York, pp. 125–138.

Strik, H., Cucchiarini C., 1998. Modelling pronunciation variation for ASR: Overview and comparison of methods. In: Proceedings of the ETRW Workshop on Modelling Pronunciation Variation for ASR. Rolduc, The Netherlands, 4–6 May.

Tajchman, G., Fosler, E., D. Jurafsky, 1995. Building multiple pronunciation models for novel words using exploratory computational phonology. In: Proc. Eurospeech '95.

Weintraub, M., Murveit, H., Cohen, M., Price, P., Bernstein, J., Baldwin, C., Bell, D., 1989. Linguistic constraints in hidden markov model based speech recognition. In: Proc. ICASSP '89, S13.2.

Weintraub, M., Fosler, E., Galles, C., Kao, Y., Khudanpur, S., Saraclar, M., Wegmann, S., 1996. Automatic learning of word pronunciation from data. In: 1996 LVCSR Summer Workshop Technical Reports.

Wooters, C., Stolcke, A., 1994. multiple-pronunciation lexical modelling in a speaker independent speech understanding system. In: Proc. ICSLP '94, pp. 1363–1366.

Young, S., Jansen, J., Odell, J., Ollasen, D., Woodland, P., 1995. The HTK Book, Version 2.0, Entropic Cambridge Research Laboratory, Cambridge, England.