Extracting Phoneme Pronunciation Information from Corpora

Ian Thomas, Ingrid Zukerman
Department of Computer Science
Monash University
Clayton, VICTORIA 3168
AUSTRALIA
{iant,ingrid}@cs.monash.edu.au

Bhavani Raskutti
Artificial Intelligence Section
Telstra Research Laboratories
Clayton, VICTORIA 3168
AUSTRALIA
b.raskutti@trl.oz.au

Abstract

We present a procedure that determines a set of phonemes possibly intended by a speaker from a recognized or uttered phone. This information will be used to allow a speech recognizer to take pronunciation into account or to consider input from a noisy source during lexical access. We investigate the hypothesis that different pronunciations of a phone occur within groups of sounds physically produced the same way, and use the Minimum Message Length principle to consider the effect of a phoneme's context on its pronunciation.

1 Introduction

When trying to match spoken words to dictionary entries during speech recognition, it is useful to be able to generate alternative versions of the spoken sequences of phones to account for the manner in which different speakers pronounce a phone. If we know the probabilities that the component sounds in a sequence of phones are pronounced like other sounds, then likely alternative pronunciations of that sequence of phones can be generated to match against a lexicon of known words. Furthermore, if we also have some idea of how the context within which a phone was uttered affects its pronunciation, we have extra information which can be used to produce more realistic alternative pronunciations.

This paper considers the task of automatically extracting statistical information about how various sound sections of words (phonemes) are pronounced by speakers (as phones) by matching intended phonemes and uttered phones from a transcribed speech corpus. The same approach could be used to gather statistics about how phones recognized (or mis-recognized) by a speech recognizer match the phonemes intended by a speaker.

This information extraction process is part of the training phase for the lexical access component of a speech recognition system, where the pronunciation probabilities are generated from a training corpus. The study was done on the TIMIT corpus (Fisher *et al.*, , 1986) — a collection of American-English read sentences with correct time-aligned acoustic-phonetic and orthographic (word-aligned) transcriptions.[1] The corpus contains 3696 sentences spoken by 462 speakers from 8 different dialect divisions across the United States.

Previous work by Riley (1989) and Withgott and Chen (1993) used Classification and Regression Trees (CART) on a large number of different features of the corpus (such as gender, dialect and speaking rate) to obtain pronunciation information of intended phonemes. Our system obtains similar results using positional information and context, and using exact matches from uttered phones to intended phonemes to guide other matches.

Work by Cohen *et al.*, (1987) on pronunciation used a couple of set sentences for multiple speakers, but did not cover a wide range of words (and thus different phone contexts). Our study considers the pronunciation patterns of a wide range of different speakers using a large collection of words.

A tree-based system by Luccassen and Mercer (1984) uses an information theoretic approach for deciding alternative pronunciations based on the classification of a large context feature vector. However, when building their decision tree, they do not evaluate the quality of the resulting tree, i.e., they keep testing attributes until a boundary situation is reached. In contrast, our system initially uses the relative positioning of uttered phones and intended phonemes to determine the phonemes possibly intended by a speaker when uttering a particular phone. The context of a phone is considered only as

[1] Transcriptions were made by a combination of hand transcriptions using multiple parametric representations of sentences as a guide, and automatic alignment (Zue and Seneff, 1988). The use of different representations is claimed as a good way of overcoming dialect biases during transcription (Withgott and Chen, 1993).

Ian Thomas, Ingrid Zukerman and Bhavani Raskutti (1998) **Extracting Phoneme Pronunciation Information from Corpora.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning,* ACL, pp 175-183.

```
The Mayan neoclassic scholar disappeared while surveying ancient ruins.

the     |mayan          |neoclassic                            |
DH   AX |M   AY   ax  N |N   IY   ow   K   L   AE   S   ih   k   |
DH   AX |M   AY   eh  N |N   IY   ix   K   L   AE   S   ix   kcl |
                         C   C

scholar               |disappeared                          |while           |
S   K   AA   L   AXR |D   ih   S   ax   P   iy   r     d  |hh   W   ay    L |
S   K   AA   L   AXR |D   ix   S   ix   P   ih   axr   dx |ax   W   aa    L |


surveying                   |ancient                          |ruins                 |
S   axr   V   EY   ix   NG |-   EY   N   SH   ix   n    t   |R   uw   ih   N    Z |
S   er    V   EY   -    NG |q   EY   N   SH   -    en   tcl |R   ux   ix   N    Z |
```

Figure 1: A typical alignment of lexicon entry phonemes with input phones.

an extra source of information to qualify these predictions. Further, the method we apply for building decision trees evaluates whether context is meaningful in terms of its predictive power.

Riley (1991) implements a similar system using a different method for tree induction, but estimates the probability of an uttered phoneme given a phoneme context and a partial phone context, whereas we are inferring an intended phoneme from an uttered phone context.

Our specific aim is to test two hypotheses: (1) that phonemes are pronounced as phones in the same broad sound category (for example, vowels for vowels and fricatives for fricatives), and (2) that the context of a phone, that is, the attributes of the phones immediately preceding and following it, influence the pronunciation of this phone.

2 Determining Alternative Pronunciations

For each word in each sentence in the corpus, we match the transcribed phones with the phonemes in that word as recorded in a lexicon, and record the frequency of occurrence of each phoneme/phone match over the entire corpus. The major difficulties in this process are (1) transcriptions include extra phones that do not appear in the phoneme sequences corresponding to words in the lexicon, and (2) there are expected phones that were not pronounced. As a result, the phone sequences rarely align exactly with the phonemes corresponding to the words in the lexicon. This makes a simple alignment unreliable, and calls for a more flexible method of matching.

Our method consists of examining the words of the corpus aligned with their lexicon entries, and choosing phoneme to phone pairs that we are confident are a match. This information is then used to make further matches in less certain areas. This process has three main steps: (1) aligning each corpus word with its corresponding lexicon entry, (2) building initial data structures, and (3) iteratively making additional certain matches.

2.1 Aligning Corpus Words with Lexicon Entries

We use the dynamic string alignment algorithm described by Sandoff and Kruskal (1983) to determine the minimum number of substitutions, insertions and deletions needed to turn one string into another. This algorithm can produce several edit sequences with the same cost, so the edit sequence with the highest number of exact matches is selected. Table 1 describes selected symbols from the ARPAbet symbol set (Shoup, 1980) used for representing phonemes and phones in TIMIT.[2] A typical alignment of words in a sentence is given in Figure 1. The first row contains phonemes from the lexicon words, and the second row contains phones from the corresponding words in a corpus sentence. Note the use of `C` to represent the sharing of the **N** phone between *mayan* and *neoclassic* due to co-articulation.

The TIMIT transcriptions of sentences use some symbols that are not present in the lexicon entries. For example, stop closures (**bcl, dcl, gcl, pcl, tcl, kcl**) and releases (**b, d, g, p, t, k**) are provided in the corpus transcriptions, but only releases are used in the lexicon entries. In order to prevent this mismatch from causing the surrounding phones to

[2]ARPAbet is a type-written version of the standard International Phonetic Alphabet.

Table 1: Selected phoneme and phonetic symbols from the ARPAbet set

Symbol	Example Word	Possible Phonetic Transcription
d	day	DCL D ey
p	pea	PCL P iy
t	tea	TCL T iy
k	key	KCL K iy
q	bat	bcl b ae Q
s	sea	S iy
sh	she	SH iy
z	zone	Z ow n
v	van	V ae n
dh	then	DH e n
m	mom	M aa M
n	noon	N uw N
en	button	b ah q EN
l	lay	L ey
r	ray	R ey
w	way	W ey
y	yacht	Y aa tcl t
hh	hay	HH ey
hv	ahead	ax HV eh dcl d
el	bottle	bcl b aa tcl t EL
iy	beet	bcl b IY tcl t
ih	bit	bcl b IH tcl t
eh	bet	bcl b EH tcl t
ey	bait	bcl b EY tcl t
ae	bat	bcl b AE tcl t
aa	bott	bcl b AA tcl t
ay	bite	bcl b AY tcl t
ah	but	bcl b AH tcl t
ow	boat	bcl b OW tcl t
uw	boot	bcl b UW tcl t
ux	toot	tcl t UX tcl t
er	bird	bcl b ER dcl d
ax	about	AX bcl b aw tcl t
ix	debit	dcl d eh bcl b IX tcl t
axr	butter	bcl b ah dx AXR

be misaligned, we remove stop closures when they appear before stop releases. For example, in Figure 1 we have removed the **kcl** closure that preceded the first **K** phone in *neoclassic*, but the final **kcl** of that word was not removed because there was no release found for that closure (we don't want to eliminate the evidence of the **k** sound completely).

2.2 Building Initial Data Structures

We scan through each aligned word pair in every sentence in the corpus and record *certain matches* and *uncertain areas*, generating frequency counts of the certain matches.

A **certain match** is a pairing between a phoneme in a lexicon word and a phone in the same corpus word which we are confident represent the same intended sound. Initially, the certain matches are insertions, deletions and substitutions bounded immediately on

		did you			
d	ih	d	|	y	uw
dx	ix	jh	|	jh	ux
		C		C	

(a)

	did	
d	ih	d
dx	ix	dcl

(b)

Figure 2: Example of (a) uncertain co-articulation, and (b) multiple choice matches

the left and right by an exact match or the beginning or end of a word. Since we know the boundaries of words from the transcriptions, we can reliably match up phonemes on word boundaries provided they are bounded on the other side by an exact match. Exact matches are also recorded as certain matches. In the sentence in Figure 1, **ow/ix** in *neoclassic* is a certain match, as is **ih/ix** in *disappeared*, **hh/ax** and **ay/aa** in *while*, and **ix/-** in *surveying*.

An **uncertain area** is a group of *two* or more operations (insert, delete or substitute) bounded by a certain match or the beginning or end of a word. Examples of uncertain areas in Figure 1 are the last three operations in *ancient* and the second and third operations in *ruins*. Uncertain areas potentially have matches within them, but we have not committed to an alignment at this stage.

TIMIT uses a number of phonological rules for sharing and deleting phones on the boundaries of the words in the transcriptions (such as shown between the second and third word in Figure 1). These rules correspond to cases of co-articulation of phones (Giachin *et al.*, , 1991). Such co-articulated phones remove word boundaries, resulting in the concatenation of the end of a word with the beginning of the next word. For example, Figure 2(a) illustrates how co-articulation of the words *did you* renders the entire phone sequence an uncertain area. In contrast, the co-articulated **N/N** phones between *mayan* and *neoclassic* in Figure 1 constitutes a certain match.

2.3 Making Additional Certain Matches

We use frequency counts of certain matches obtained in Step 2 to generate additional certain matches from the uncertain areas. To this effect, we consider each possible phoneme/phone match in each uncertain area, and select the phoneme/phone match with the highest frequency of certain matches. For instance, if the match **n/en** occurs more often than **n/tcl**, then **n/en** will be chosen in the uncertain area between **SH** and the end of *ancient* in Figure 1.

Whenever there are multiple instances of the same phoneme in an uncertain area, it is matched to the phone that is positionally closest. For example, both

```
                                 D  A  T  A     P  H  O  N  E  S

                         stops      affricates  nasals glides&         vowels
                                    fricatives         semi-
                                                       vowels

                    -bdgptkdqbdgptk jcsszzftvd mnneeen lrwyhhe iieeaaaaaaoouuueaiaa peh
                           x cccccc hh h h h h   gmnnx    hvl yhhyeawyhoywhwxrxxxx ap#
                             llllll                g                          r- ui
                                                                               h

                  - .AAAAAAANAAAABA AAAAA..AAA AA..AAA AAAAAAA AAAA.A..AA.AAAAABAAA AC. -
     stops        b AV..A...AC..... .........A. ....... .A..... .................... ... b
               →  d D.L..A.DA.G..A. AAA.AA..AA .A....A ...AA.. ................AA.. ... d
                  g A..U..A....E... .......... ....... ...A... .................... ... g
                  p AA..V...A...D.. ......A... ....... ....... .................... ... p    A:0-4%
               →  t B.A..M.DB.A.AF. .AAA..AA.. .A..... ....... ................AA.. .A. t
                  k A.....U.A..A.AE .......... ....... ....... .................... ... k    B:4-8%
                 dx ............... .......... ....... ....... .................... ... dx
                  q ............... .......... ....... ....... .................... ... q    C:8-12%
                bcl ............... .......... ....... ....... .................... ... bcl
                dcl ............... .......... ....... ....... .................... ... dcl  D:12-16%
                gcl ............... .......... ....... ....... .................... ... gcl
                pcl ............... .......... ....... ....... .................... ... pcl  E:16-20%
                tcl ............... .......... ....... ....... .................... ... tcl
                kcl ............... .......... ....... ....... .................... ... kcl  F:20-24%

    affricates   jh A..A......A..A. XA..AA.... ....... ...A... .................... ... jh
                 ch A...........A.  .X.A...... ....... ....... ................AA.. ... ch   G:24-28%

    fricatives    s B....A.......A. ..XAA..... ....... ....... .................... ... s
                 sh A.............. .AAY...... ....... ....... .................... ... sh   H:28-32%
                  z A.A..A.A..A.... ..CAVA...A ....... ....... .................... ... z
                 zh ..........A.... CA..AV.... ....... ....... .................... ... zh   I:32-36%
                  f A...A.......... ......YAA. ....... ....... .................... ... f
                 th A.A..A....A..A. .......X.A ....... ....... .................... ... th   J:36-40%
                  v B........A..... ..A...A.XA ....... ....... .................... ... v
                 dh B.A..A.AA....A. ..A.A..B.V .A..... ....... .................... ... dh   K:40-44%

L   nasals        m AA............. .......... XAAA..A ..A.... .................... ... m
E                 n A.............. .......... AUA.B.C ....... ...A................ ... n    L:44-48%
X                ng A..........A... .......... ABW..AA ....... .................... ... ng
I                em D.............. .......... ...V... ....... .................... ... em   M:48-52%
C                en A.............. .......... .I.AP.. ....... ................A... ... en
O               eng ............... .......... ....... ....... .................... ... eng  N:52-56%
N                nx ............... .......... ....... ....... .................... ... nx

    glides &      l A......A....... .......A.. ....... X.A...B .................... ... l    O:56-60%
    semi-         r B.......A...... ..A....... ....... ATA.... .........A......BAAC. ... r
    vowels        w A.............. .......... ....... .AY.... .................... ... w    P:60-64%
                  y D.............. AA........ ....... ...V... AA.................. ... y
                 hh E....A......... .......... ....... .A..LI. ................A... ... hh   Q:64-68%
                 hv ............... .......... ....... ....... .................... ... hv
                 el A.............. ..........  ...... D.....U ...........AA...AA.. ... el   R:68-72%

P   vowels       iy A.............. .......... ....... ...A... WA.A....A....AA.AB.A ... iy
H                ih A.......A...... .......... ....... .A..... AMAAA..AA..AAAAAAIAA .A. ih   S:72-76%
O                eh A.............. .......... ....... ....... AAVAA...AA.AA..AAAAA ... eh
N                ey A.............. .......... ....... ....... AAAYA..AA.......AA.. ... ey   T:76-80%
E                ae A.............. .......... ....... ....... AAEAPAAAA...A...ACAA .A. ae
M                aa B.......A...... .......... ....... ....A.. .AA.AVAAAA.A....AAA. ... aa   U:80-84%
E                aw A.............. .......... ....... ....... ..A.ABW.A..A........ ... aw
S                ay A.............. .......... ....... ....... AAAAAA.XA.A.....AA.. ... ay   V:84-88%
                 ah A.............. .......... ....... ....... .AA..AA.UA.AA...BAAA ... ah
                 ao B.......A...... .......... ....... ....... .A...CA.ATAAA..AAAA. ... ao   W:88-92%
                 oy ............... .......... ....... ....... .........AXA........ ... oy
                 ow A.............. .......... ....... ..A...A .A...A..AA.XAAAAAAAA ... ow   X:92-96%
                 uh G.............. .......... ....... ....... .AA.....AA.ALAAAADAA .A. uh
                 uw A.............. ........... ...... ..A.... AAA.....AA.AAFO.ACAA .A. uw Y:96-100%
                 ux ............... .......... ....... ....... .................... ... ux
                 er ............... .......... ....... .A..... .A.....AA...A.TAAFA ... er   Z:100%
                 ax B....A..A.....A .......... ....... .A....A BBAAAA.ABA.AA.AAJIAA .A. ax
                 ix C.............. .......... ....... ....... BDA.A...A...A.A.BNAA .A. ix
                axr A.............. .......... ....... .B....A .A...A..A...A.AGAAPA ... axr
               ax-h ............... .......... ....... ....... .................... ... ax-h
                pau ............... .......... ....... ....... .................... ... pau
                epi ............... .......... ....... ....... .................... ... epi
                 h# ............... .......... ....... ....... .................... ... h#

                    -bdgptkdqbdgptk jcsszzftvd mnneeen lrwyhhe iieeaaaaaaoouuueaiaa peh
                           x cccccc hh h h h h   gmnnx    hvl yhhyeawyhoywhwxrxxxx ap#
                             llllll                g                          r- ui
                                                                               h
```

Figure 3: Phonemes (rows) pronounced as phones (columns) after stabilization of the matching algorithm.

d phonemes in the uncertain area in Figure 2(b) match the phones **dx** and **dcl**. In this case, we match the first **d** to **dx** and the second **d** to **dcl**. During this process, we consider only potential phone/phoneme matches, and we ignore any match involving a "–" symbol (which indicates an insertion or a deletion). Insertions and deletions that are certain matches are collected for statistical purposes but do not influence the match decision process.

After a phoneme/phone match has been determined, the phonemes and phones of the uncertain area are shifted so that the matched phoneme and phone are lined up. This new match is then removed from the uncertain area and recorded as another certain match. This process can create other bounded matches that are also removed from the uncertain areas and added to the certain matches. For example, in Figure 1, finding the certain match **ih/ix** in *disappeared* in Step 2 suggests that the match to be made in the uncertain areas in *ruins* and *neoclassic* is **ih/ix**. The match in *ruins* in turn creates a **uw/ux** pairing on its left-hand side, which is also recorded as a certain match, eliminating this uncertain area completely. Similarly, the match in *neoclassic* yields the **k/kcl** match on its right-hand side. This step is repeated until the number of matches being made levels off.

2.4 Evaluation

For the TIMIT corpus, the number of certain matches levels off at 123115 from 109898 initial matches (after six iterations of Step 3), and the number of remaining uncertain matches falls from 13913 to 908. Figure 3 shows the percentage of phones (columns) found for every phoneme in the lexicon words (rows) after six iterations of matching attempts in each uncertain area. For example, between 12-16% of **t** and **d** are pronounced as **dx**. It is important to note that exact phoneme/phone matches registered in the 1000-2000 range, while most of the alternative pronunciations had a frequency in the few dozen.

Figure 3 suggests that phonemes are often mispronounced as phones in the same broad sound group, i.e., both the intended phonemes and the uttered phones are generated by the same physical method of production. This result is most evident for vowels. Some of the other types of sounds, such as fricatives and nasals, have not been differentiated so clearly.

3 Considering Context

To investigate the effect of the context of a phone (the attributes of the phones before and after this phone) on its pronunciation, we examined sentences aligned like the sentence in Figure 1 and recorded phoneme/phone pairs, along with acoustic features for the phones that appear before and after each phoneme/phone pair.[3] It includes the following attributes: broad sound category, voiced/voiceless, sibilant/non-sibilant and sonorant/obstruent. Table 2 shows how phones are classified according to these acoustic features, from (Yannakoudakis and Hutton, 1987) and (Rabiner and Juang, 1993).

The contexts for all the phones were fed into an inductive inference program by Wallace and Patrick (1993) in order to find functions of the context attributes (i.e., acoustic features) that are good predictors of the phoneme intended by a speaker when s/he utters a particular phone. The inference program uses the Minimum Message Length (MML) principle (Georgeff and Wallace, 1984) to measure the significance of these functions. These functions are realized by a decision tree from the contexts for each uttered phone, with nodes splitting on the values of particular attributes.[4] Leaf nodes in a decision tree represent a partition of the context sample. Each leaf node contains a collection of contexts in the sample and the phoneme intended by a speaker for each context. An internal node of the tree is split on an attribute only when doing so creates a statistically significant reduction in the number of different types of intended phonemes in the leaf nodes (better than that expected from random effects). Ideally, each leaf node should contain several contexts, all of which have the same intended phoneme. This means that the attributes these contexts have in common are sufficient to predict this intended phoneme from an uttered phone.

The MML principle is based on the following premise: if a sender knows both the attribute values and the class of the objects in a set, and wants to send the class of each object to a receiver (who knows the attribute values but not the classes), the sender aims to send the shortest possible message (in bits). The MML criterion is used to produce the decision tree that can be sent by means of the shortest possible message. A split is made in the decision tree only if it decreases the message length for transmitting the intended information. The decision tree is then sent in a two-part message: (1) instructions for the receiver on how to reconstruct the tree; and

[3] Uncertain areas were treated as a single phoneme/phone pair involving complicated sounds.

[4] Words which are common in the corpus generate contexts that appear with high frequency. We assume that these frequencies are representative of those in English.

Table 2: Classification of TIMIT phones according to acoustic features.

b,d,g	Stop Release, voiced, obstruent, non-sibilant
p,q,t,k,dx	Stop Release, unvoiced, obstruent, non-sibilant
bcl,dcl,gcl	Stop Closure, voiced
pcl,tcl,kcl	Stop Closure, unvoiced
jh	Affricate, voiced, obstruent, sibilant
ch	Affricate, unvoiced, obstruent, sibilant
z,zh	Fricative, voiced, obstruent, sibilant
v,dh	Fricative, voiced, obstruent, non-sibilant
s,sh	Fricative, unvoiced, obstruent, sibilant
f,v,th	Fricative, unvoiced, obstruent, non-sibilant
m,n,nx,ng,em,en,eng	Nasal, voiced, sonorant
l,r,y,w,el	Semivowel/Glide, voiced, sonorant
hh	Semivowel/Glide, unvoiced, obstruent, non-sibilant
hv	Semivowel/Glide, voiced, obstruent, non-sibilant
iy,ih,eh,ae	Vowel, Front Position
aa,er,ah,ax,ao	Vowel, Mid Position
uw,uh,ow	Vowel, Back Position
axr,ax-h	Vowel
pau,epi	Pause
wb	Word boundary
sb	Sentence boundary
–	Missing phoneme

(2) the labels for the classes of the objects in the leaves of the tree. Standard coding techniques show that the encoded set of labels for the classes in a leaf node will be short if most of the objects in that leaf node have the same label, and will be longer if the labels of the objects are equally likely. Therefore, if the objects in each leaf node are predominantly of one class, then the message encoding the decision tree will be shorter than a message which simply encodes the class label for each object in the set (without the tree).

Given an uttered phone, the resulting decision tree shows the significant attributes and values for classifying the intended phoneme, which is the effect that the surrounding sounds have on predicting the intended phoneme.

3.1 Evaluation

As indicated in Figure 3, in the majority of cases the uttered phone and the intended phoneme are the same. Table 3 summarizes the decision trees for situations where uttered phones are different from intended phonemes. This summary shows the effect of contextual phonetic information on the intended phoneme for each possible uttered phone (one line per phone). For example, for the uttered phone **d**, the intended phoneme was **t** when the next sound was neither a consonant nor a vowel, i.e., a word boundary, a sentence boundary or missing; this occurred in 12 of the samples. Also, the intended phoneme was missing (i.e., **d** was uttered when no phoneme was intended) when the next sound was an obstruent; this happened in 2 of the samples. In 220 samples, the uttered phone was **iy** with intended phoneme **ax** when **iy** was the last sound in a word, and the previous sound was a fricative.

Some uttered phones found in Figure 3 are missing from Table 3 because either there were not enough samples to create a tree (the stops **b**, **g** and **p**, the nasal **m**, and the pause), or more often because the trees produced had no discriminatory power. This occurred when each leaf node in a decision tree had an evenly spread mixture of intended phonemes, or when the same intended phoneme appeared throughout the tree.

The decision trees were evaluated using test contexts from 1344 sentences (out of 5040 sentences) spoken by 26% of the speakers. No speaker was in the test and training sets. Each phone and its context was classified into a leaf node using the attributes in the context. A phoneme prediction was considered correct when the intended phoneme was the same as the most common phoneme in the leaf node (determined during training). 75% of the different test samples were predicted correctly.

4 Discussion

We have analyzed a large corpus of sentences read by a large number of speakers with a view to determining possible mis-pronunciation of phonemes and the context in which such mis-pronunciations occur. The results of our analysis support our hypotheses that phonemes are mis-pronounced as phones in the same broad sound categories, and that the context of mis-pronunciation provides valuable information

uttered phone	number of samples	intended phoneme (samples)	prev sound	next sound	intended phoneme (samples)	prev sound	next sound	intended phone (samples)	prev sound	next sound
d	2369	t(12)		non-cons non-vow	— (2)		obst			
t	3905	— (41)		obst	d (8)		non-vow			
k	3788	— (6)	nasal		g (1)	fricat				
dx	1069	t(183)	frontV		t (171)	non-vow non-cons	voiced			
q	1318	— (912)		vowel	t (154)	non-cons vowel	wordB			
jh	987	zh(21)		non-vow	d y (2)		backV			
ch	778	sh(5)	nasal		jh (2)	vowel		t (2)	fricat	
sh	1276	s(24)		obst unvoiced	s ch(3)	vowel	voiced	— (3)	wordB	voiced
f	2204	v (23)	vowel	non-cons	p(2)	vowel	obst			
v	1978	b (7)	vowel		— (2)	wordB				
n	6133	ng (26)	vowel		en (2)	fricat		dh (1)	wordB	
em	109	ax m(41)	stopRe		ah m (19)		semiGl			
en	516	ix n(149)	obst unvoiced	frontV	ax n (54)	stopR	obst	ae n d (42)	non-cons	wordB
eng	16	ng (3)	vowel		ix ng (3)	StopR		ih n (2)	wordB	
r	4635	— (54)	non-cons	vowel	axr (34)	obst				
w	2184	— (17)	non-cons		— (3)	obst	vowel	uw(2)	obst	semiGL
hh	915	— (24)	unvoiced		d (2)	voiced				
el	911	ih l (56)	unvoiced semiGl		ax l (33)	voiced	vowel	uh l (10)	fricat unvoiced	
iy	4481	ax (220)	fricat	wordB	ix (93)	non-vow	nasal	dh ax (21)	non-cons unvoiced	wordB
ih	3838	ax (52)	fricat obst	wordB	ix (52)	non-obst unvoiced	nasal	iy (28)	semiGl	wordB
eh	2920	ae (151)	obst	semiGl	ae (63)	non-obst unvoiced	nasal	ax (24)	non-obst voiced	nasal
ey	2209	ax (92)	unvoiced	wordB	ae (6)		nasal			
ae	2273	ih (3)	nasal		aa (3)		semiGl			
aa	2137	ao (53)	non-obst	nasal	aw (34)	wordB	semiGl	ay(11)	nasal	wordB
ay	1935	aa (4)	nasal		ax (2)	semiGL		ax (2)	wordB	
ah	2045	ax (84)	unvoiced	obst	ax (65)	unvoiced	non-cons voiced			
ao	1840	aa (12)	voiced obst	semiGl	uh (11)	unvoiced	semiGl	ah (2)	unvoiced semiGL	semiGL
oy	296	ow ix(3)		nasal	ao (3)		semiGl			
ow	1643	ax (8)	non-cons		ao r(4)	sonor	ao r(3)		obst	
uh	467	ax (10)	sonor	obst	ih (9)	sonor	sonor	uw (10)	obst	non-cons
uw	522	ow (3)		vowel	er (2)		wordB	uh (2)		semiGl
ax	3091	uw (59)	stopRe unvoiced	wordB	ae (55)	wordB	nasal			
ix	5958	ax (505)	fricat non-sibila	wordB	ih (404)	unvoiced non-obst	nasal	uw (186)	stopRe unvoiced	wordB
axr	2181	ao r (130)	fricat unvoiced	wordB	aa r (63)	wordB	wordB	r (51)	vowel	non-cons
ax-h	303	uw (79)	stopRe	non-cons wordB	ix(29)		obst sibila	ax (26)	fricat voiced	wordB
—	2136	d (332)	nasal	unvoiced non-obst	t (84)	fricat unvoiced sibila	wordB	hh(119)	wordB	semiGl

Table 3: Summary of most significant values in the decision tree for each uttered phone.

about the intended phoneme.

As indicated in Figure 3, mis-pronunciation of affricates and fricatives is rare (over 84% certain matches), though when they are mis-pronounced, the uttered phone may be one of the stop consonants. Vowels are often mis-pronounced, but the uttered phone is almost always from the same broad sound category. The stop consonants, **d** and **t**, the nasal **en**, and the semi-vowel **hh** are often mis-pronounced. Analysis of the decision trees generated for these mis-pronunciations indicates that the attributes of the phones surrounding the mis-pronounced phoneme do indeed provide information about the intended phoneme. These decision trees are particularly informative for vowels owing to the large number of mis-pronunciations as well as the regularity of these mis-pronunciations. The attributes that are most useful vary for the different pronunciations of each phoneme. For instance, **ay** is the pronunciation for **aa** when the previous phone is nasal, while **ao** is pronounced for **aa** when the preceding phone is a voiced obstruent and the next phone is a semivowel or glide.

The frequency of the matches in Figure 3 combined with the decision trees produced using the MML principle may be used to generate alternate pronunciations of phonemes in word models. This will assist in the recognition of mis-pronounced words during automatic speech understanding. The decision tree is weakest for uncommon contexts, because of a lack of training data for constructing the tree (the message length for encoding phonemes in such contexts is no better than an efficient encoding of the context classes using a Huffman code). In this case, the matrix in Figure 3 should be used to predict alternative pronunciations. However for more common contexts, the decision trees are preferred, as they use more information than the matrix to determine the intended phoneme.

The system described in this paper investigates dependencies between an intended phoneme and a pronounced phone, but it may be easily adapted to determine relationships between an intended sound and a recognized sound, i.e., the output of a speech recognizer. Relationships determined in this manner may be used during speech recognition, and thus account for mis-recognition as well as mis-pronunciation.

Acknowledgments

The authors thank Jon Oliver and Chris Wallace for their advice on MML encoding.

REFERENCES

Cohen, M., Baldwin, G., Berhnstein, J., Murveit, H., and Weintraub, M., Studies for an Adaptive Recognition Lexicon. In *Proceedings of the DARPA Speech Recognition Workshop*, Report No. SAIC-87/1644, 1987.

Fisher, W.M., Doddington, M., George, R., and Goudie-Marshell, K.M., The DARPA Speech Recognition Database: Specifications and Status. In *Proceedings of the DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, 1986.

Georgeff, M.P., and Wallace, C.S., A General Criterion for Inductive Inference. In *Proceedings of the Sixth European Conference on Artificial Intelligence*, pp. 473-482, Pisa, Italy, 1984.

Giachin, E.P., Rosenberg, A.E., and Lee, C., Word Juncture Modeling using Phonological Rules for HMM-based Continuous Speech Recognition, *Computer Speech and Language* **5**:155-168, 1991.

Lucassen, J.M., and Mercer, R.L., An Information Theoretic Approach to the Automatic Determination of Phoneme Baseforms. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 42.5.1–42.5.4, 1984.

Rabiner, L.R., and Huang, B., *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs NJ, 1993.

Riley, M.D., Some Applications of Tree-based Modeling to Speech and Language. In *DARPA Speech and Language Workshop*, pp. 339-352, 1989.

Riley, M.D., A Statistical Model for Generating Pronuncation Networks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 737-740, 1991.

Sankoff, D. and Kruskal, J.B., *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison Wesley, London, 1983.

Shoup, J.E., Phonological Aspects of Speech Recognition. In *Trends in Speech Recognition*, W.A. Lea, Ed., Prentice-Hall, Englewood Cliffs, NJ, pp. 125-138, 1980.

Wallace, C.S., and Patrick, J.D., Coding Decision Trees, *Machine Learning* **11**:7-22, 1993.

Withgott, M.M. and Chen, F.R. *Computational models of American Speech*, CSLI Lecture Notes, No. 32, Stanford, CA, 1993.

Yannakoudakis, E.J. and Hutton, P.J., *Speech Synthesis and Recognition Systems*, Ellis Horwood Limited, 1987.

Zue, V.W. and Seneff, S., Transcription and Alignment of the Timit Database. In *Second Symposium on Advanced Man-Machine Interface through Spoken Language*, Oahu, Hawaii, 1988.