DYNAMIC AND STATIC IMPROVEMENTS TO LEXICAL BASEFORMS

Simon Downey and Richard Wiseman

downey@saltfarm.bt.co.uk, richard@saltfarm.bt.co.uk

Speech Technology Unit, BT Laboratories,
Martlesham Heath,
Suffolk, UK.

ABSTRACT

One limitation of many speaker independent recognition systems is their dependence on a single baseform dictionary to model word pronunciations. These dictionaries typically contain only a single (or 'ideal') pronunciation for each word. Previous work on improving dictionary models to include multiple pronunciations has met with mixed success - the alternatives may increase ambiguity in some cases. This paper investigates two approaches to improve lexical baseforms. The first is a 'bottom-up' approach in which 'ideal' transcriptions of utterances looked up in a pronunciation dictionary are compared to phonemic level hand-annotated transcriptions. Analysing the differences between the two transcriptions reveals many common mispronunciations, accent-based alternatives, false-starts and incorrect word substitutions. Each of these problems is illustrated in the paper, where it is also shown that unfamiliar words are prone to large numbers of alternative pronunciations. The second approach is more 'top-down'. Phonologically developed rules and transforms are described which modify the lexical representation of the utterance and a pronunciation network is thus derived. This approach has the advantage of being able to explicitly model cross-word coarticulation effects, whereas the former approach models them implicitly to a certain extent. The relative merits of each technique are investigated using a set of experiments performed on a phonetically rich database.

1. INTRODUCTION

It is widely accepted that word models derived from the concept of 'ideal' pronunciations are often not rich enough for speaker independent, continuous speech recognition tasks. Common words and suffixes are particularly susceptible to large acoustic variations in fluent speech. This can result in significant reductions in recognition accuracy especially for talkers whose accents bear considerable pronunciation differences from the citation dictionary baseforms.

Speaker adaptive systems are able to overcome this shortcoming to some extent by tailoring the system to a talker's characteristics through iterative reestimation of the acoustic models. For many telephony applications the duration of a call is too short to collect sufficient data to successfully adapt the models.

An alternative approach attempts to modify the recognition lexicon, for instance to deal effectively with words having more than one widely accepted pronunciation. Improved modelling of continuous speech coarticulation phenomena which are likely to cause additional variations in the word initial and word final phonemes is also considered. These two word variation mechanisms may be defined as intra-word and inter-word variations respectively.

Intra-word pronunciation variants are often speaker-dependent: the speaker's dialect can have a significant influence. Several techniques have been established for dealing with intra-word variations including more consistent database transcriptions [1], alternative pronunciations [2], dictionary baseform optimization [3] and phonologically developed rules/transforms to modify lexical representations to fit the speakers dialect [4].

The problem of inter-word variations is even more complicated. Word initial and final phonemes can be deleted, substituted or elided depending on the particular context. These variations are to some extent speaker independent. Techniques for improved acoustic-phonetic modelling at word boundaries that have been investigated include emphasis on landmarks (or points of time defining speech 'entities' rather than identification of steady state regions), Linear Discriminant Analysis on phone classes during training to improve phonetic segmentation and explicit modelling of adjacent word co-articulation effects – cross-word triphones can deal with some small effects, but not the more abrupt ones.

The work presented in this paper describes how the above techniques impact upon recognition accuracy for large vocabulary continuous speech tasks. The phonetically hand-annotated Subscriber database [5] is used to analyse the fluent speech effects described above. A set of experiments then investigates the key shortcomings in the current citation-form pronunciation dictionary. Results are presented for both rule-based and phonetically derived techniques. Recognition accuracy for each technique is compared with that of an unadapted baseline system. The results of the experiments were also used to investigate whether certain key continuous speech effects as described in [6] are observable in the recognition output.

2. RULE-BASED NETWORK GENERATION

Much of the phonemic variation present in fluent speech is due to contextual effects and the speaker's dialect. The contextual variation for British English is described in detail in [6]. A set of rules was developed to deal with the effects of assimilation, coalescence, consonant elision, phonemic elision, intrusive 'r' and allophonic variation. These rules were then applied to the baseform transcription of an utterance, resulting in a transform network describing possible pronunciation variants for that utterance. An example network derived from the sentence "Would you give me my hand bag" is shown in Figure 1.

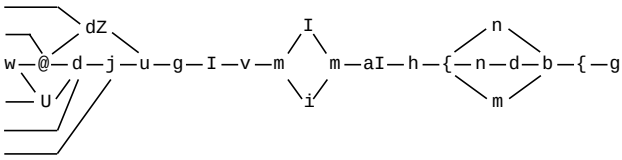

Figure 1: Rules based network for "Would you give me my hand bag"

To ensure that all possible transcriptions of a sentence are generated, the rules were applied to a phonemic transcription

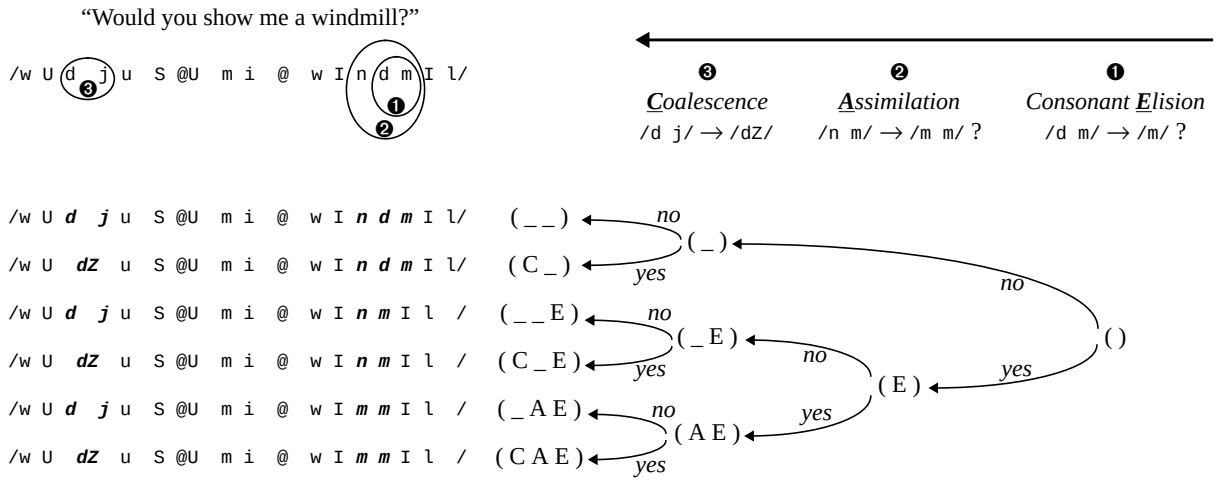

Figure 2: An example of the application of rules to a phrase.

according to the following points:

- rules are applied by considering pairs of phonemes, along with up to three surrounding phonemes; only the central pair may undergo changes by application of the rules
- the changes caused by one rule can allow another rule to be applied to a phoneme pair (generally) earlier on in the phonemic transcription. The transcription is therefore traversed in reverse order to allow these effects to propagate through the phonetic sequence
- every time a rule can be applied, two copies of the phonemic transcription are made, one where the rule is applied and one where it is suppressed; rule analysis of the two copies is then continued separately. This is necessary because changes due to one rule can *prevent* the application of another

The non-application of rules is effectively the application of a 'null' rule, and will now be considered as such.

2.1. Rules Application

An example of how the application of rules can affect phoneme pairs earlier in the transcription is shown in Figure 2. The phrase to be analysed is "Would you show me a windmill?" Three 'pairs' of phonemes to which rules can be applied are circled in the top left of the diagram.

Phonemes are traversed in reverse order, as indicated by the arrow, so that effects propagate through the transcription. Each variant of the transcription is generated through the application of rules and has a unique label, in which code letters indicate rules that *were* applied, and underscores indicate rules that were *not*. These variant labels are parenthesised in the diagram, and the rules relating to the code letters are shown with examples in Table 1.

The right-hand half of Figure 2 shows the application of the rules. Every time a rule can be applied, two new instances of the phonemic transcription are created, one where the rule is applied, and one where it is not. The first pair of phonemes to which a rule can be applied is indicated by ❶: two new copies of the transcription are generated, and the /d/ consonant is elided in one of these. Pair ❷ can undergo assimilation only

Code	*Rule*	*Example*
A	Assimilation	"tin can" /t I **n k** { n/ → /t I **N k** { n/
C	Coalescence	"would you" /w U **d j** u/ → /w U **dZ** u/
E	Consonant Elision	"old man" /Q **l d m** { n/ → /Q **l m** { n/
P	Phonemic Elision	"run along" /r V **n @ l** Q N/ → /r V **n l** Q N/
R	Intrusive 'r'	"far away" /f **A @** w eI/ → /f **A r @** w eI/
r	Intrusive 'r' + /@/ elision	"[is] Asia a [large place?]" /eI **S @ @**/ → /eI **S r @**/
V	Allophonic Variation	"how old" /h **aU Q** l d/ → /h **{ Q** l d/

Table 1: The code letters used in variant labels, the rules they represent and an example of each

where the /d/ in pair ❶ has been elided. (This is why pair ❷ appears to contain three phonemes—the rule can only be applied if pair ❶ is reduced to one phoneme.) In all cases, phoneme pair ❸ can coalesce, and the six final transcriptions (including the original) are shown on the left with the affected phonemes hilighted.

The presence of words which can have strong and weak forms may also result in more than one baseform transcription of a sentence, since the phonetic realisation of the utterance will depend on factors such as emphasis. As this is unknown at recognition time all possible combinations are added to the network. For example, in a sentence such as "The Olympic torch shines as a symbol of hope which has pushed aside barriers of race," words with weak/strong forms are: "the" (2 forms/transcriptions), "as" (2), "a" (3), "of" (4) and "has" (5). Allowing for all possible combinations, 960 different baseform transcriptions may be generated by the recognition network for this sentence.

3. COMPARING TRANSCRIPTIONS

Fluent speech pronunciation variations were investigated using the Subscriber database. The database consists of utterances collected over the UK telephony network from over 1000 talkers throughout the British Isles who were selected as a demographically balanced sample of the adult population. A detailed descrip-

```
CITATION: m A   t I  n  | @ n d  | k r eI g   | g r @U  | d w O    f  | t j u l I p s  |
MANUAL  : m A r t    n  | { n d  | k r eI g Q | g r @U  | d w Q r  f  |  tS u l I p s  |
CITATION: @ n d  | e k z I b I t  | D @ m  | O l | @U v @ | D @ | k aU n t   i |
MANUAL  : A n    | I g z I b @ t  | D e m  | Q l | @U v @ | D @ | k V  n t r I |
```

Figure 3: DP Match of "*Martin and Craig grow dwarf tulips and exhibit them all over the county*"

tion of the database and the accent categories can be found in [5]. Utterances from the database were annotated manually at the phonetic level using a rich phoneme set comprising 74 different speech sounds. A DP-match was used to align the dictionary-based transcription and the hand-annotated version. A slightly modified version of the standard algorithm was used which enabled word boundary markers (based on the citation-form transcriptions) to be inserted in the match. The word boundary markers enable different transcriptions of individual words to be examined—an example of the match is illustrated in Figure 3.

Certain continuous speech effects are immediately obvious—for instance the final /n/ in "Martin" acting as a syllabic consonant and the strong vowel form of the first "and". More accent-specific effects can also be observed such as the pronunciations of "tulips" and "exhibit". These effects are studied in more detail in Section 4.2. Finally, the last word of the sentence has been incorrectly uttered as "country", giving obvious problems if the database were to be used for whole word modelling.

Results from the DP match revealed typically 75–80% phoneme agreement between the hand labelling and citation form transcriptions. An analysis of the transcriptions indicated that approximately 30% of the words were always transcribed identically by both methods. For 67% of the words, the hand labelling gave at least one different transcription. The remaining 3% of cases were identified as dp matching errors and subsequently corrected by hand.

On average an additional 4 transcriptions were generated for each word. Some of the longer and more unfamiliar words generated many more than 4 alternatives, for instance 'trapezoidal' generated 20 pronunciations. Table 2 illustrates the 10 pronunciation variants obtained for the word 'power'[1].

p aU w @ r	p aI r
p @U @	p aU @
p @U @ r	p aU @ r
p @U r	p aU r
p @U w @	p aU w @

Table 2: Alternative Pronunciations of the word 'power'

4. EXPERIMENTAL RESULTS

Experiments were performed using the Subscriber database. A subset of the database was used in this set of experiments, comprising the 5 phonetically rich sentences recorded by each talker. The complete subset consisted of 4874 sentences and contains a total of 1243 different words. A standard cepstral front-end feature parameterisation was used.

4.1. Rule-Based Experiments

Experiments were performed in order to determine the usefulness of the rules. It is possible that recognition performance using alternate transcriptions generated using the rules could be improved or degraded, depending upon which rules are used (individually or in combination).

In the first experiment the recogniser was forced to choose one transcription of a particular sentence from all the transcriptions generated using the rules. This indicates how often the spoken phrase matches the rule-altered transcriptions more closely than the baseform generated transcription.

A second experiment comprised six sub-experiments, one for each rule (see Table 1; 'R' and 'r' were considered together). Rules were investigated individually in much the same way as above.

The results from the first set of experiments were analysed to determine not only how often rules *were* used, but also where they specifically were *not* used. With reference to Figure 2, for example, if the variant labelled as "C _ E" was recognised as the best representation of the spoken phrase (through having the best log probability), then clearly rules 'C' (coalescence) and 'E' (consonant elision) were used in the generation of the transcription. However, the underscore indicates that a rule was specifically not applied to yield that transcription. It can be determined (by looking at other variant labels) which rule was not applied—in this case rule 'A' (assimilation). Thus, it is significant both that rules 'C' and 'E' were used, but also that rule 'A' was not. The results are shown in Table 3:

Rule	*No. times used in chosen variants (a)*	*No. times not used in chosen variants (b)*	*Ratio a/b*
A	392	514	76%
C	144	102	141%
E	938	373	251%
P	305	612	49%
R	104	141	73%
r	35	9	388%
V	147	247	59%

Table 3: A comparison of the number of times each rule was/ was not used in generating the chosen transcription variants

The last column of the table can be considered a "helpfulness" measure, since it is a comparison of where the rule was helpful (where it was chosen) and where it was not (and was specifically not chosen). Where this measure is 100%, both of these occurred with equal frequency, and so the rule can be considered neutral—it neither increased nor decreased recognition accuracy. Values greater than 100% are desired, since for these the rule improves the log probability.

The second set of experiments gave the results in Table 4. These results differ from those of Table 3 in that the results for each rule were generated separately in a total of six experiments (rules 'r' and 'R' being considered together).

1. Phonetic Transcriptions in this paper use the SAM Phonetic Alphabet

Rule	*Number of times used in chosen variants (a)*	*Number of times not used in chosen variants (b)*	*Ratio a/b*
A	476	648	73%
C	145	101	143%
E	947	367	258%
P	311	618	50%
R	100	138	72%
r	40	7	571%
V	148	246	60%

Table 4: A comparison of the number of times each rule was and specifically was not used in generating the chosen transcription variants

Looking at the results where rules were allowed in combination (Table 3) and where each of the six rule types was individually investigated (Table 4), there are few differences between the "helpfulness" measures (the values in the last column).

It is not surprising that the values for the individual experiments are better than the combined, since there is a certain amount of redundancy between the rules. For example, the /d b/ in "old boat" can be altered as a result of both consonant elision and assimilation (giving similar, though not identical results). Thus, because of a rule hierarchy imposed to keep repeated transcriptions to a minimum, some rules which might not normally be applied can take effect. Therefore, the "helpfulness" of a rule does not depend upon whether or not it is the only rule which can be applied. Furthermore, rules for which the "helpfulness" is better than neutral (i.e., greater than 100%), are 'C', 'E' and 'r': coalescence, consonant elision and intrusive 'r' with /@/ elision respectively.

Applying the strong/weak form rules to the database sentences resulted in certain sentences having up to 960 transcriptions, this rose to a maximum of 6912 when all the rules were applied. Although these numbers are fairly large, they can be easily minimised to compact networks for recognition given the large amount of common sections in each transcription variant.

4.2. Alternative Pronunciation Experiment

Several recognition experiments were performed to explore unconstrained phoneme recognition performance of the Subscriber sentences. The results were examined for certain key continuous speech effects described in [6] and outlined below. A comparison was made between the phonetically annotated data and the unconstrained phoneme recognition. This gave some indication as to how well the speech models are behaving in the particular contexts, and hence provides indicators as to how likely increased recognition accuracy may be achieved by improved modelling of these contexts.

Syllabic Consonants – final syllabic /n/ frequently occurs following /t d f v s z S Z/ as in 'cotton, sudden, often etc...), in other sequences an intervening /@/ is common ('open', 'broken'). The recogniser output considerably favours the former sequence, recognising 7% more syllabic consonant sequences than are labelled as such in the phonetically annotated data. This corresponds to a similar drop in the latter sequences by the unconstrained recognition output compared to the annotations—indeed this sequence is very rarely recognised at all by either model set, which may be due to the 'broad' nature of the schwa unstressed vowel model.

Plosive aspiration – only about one third of aspirated plosive sequences labelled as such by hand were recognised correctly. This may be important because aspiration can give clues to phoneme identity (e.g. 'pin' is distinguished from 'bin' very largely by the aspiration and voicing onset time accompanying /p/) and these sequences are relatively common. The errors may be affected by the 16ms data rate, with the restricted telephony bandwidth an additional factor.

Other effects, such as devoicing of liquids/glides, vowel reduction and restricted consonant distribution, whilst still significant, have less of an impact on recognition performance. They are discussed further in [7].

Recognition accuracy of the alternative pronunciations was investigated using an unconstrained word recognition grammar. Results of this experiment show a 5.4% increase in recognition accuracy, over the citation original (from 18.9% to 24.2%). The improvement is consistent when a phoneme bigram language model is used.

5. CONCLUSIONS

This paper has investigated some of the effects present in fluent speech which can influence the phonetic realisation of utterances. Using a phonetically labelled database, various continuous speech traits have been identified which are absent in the citation form of the utterance. These have been used to define a set of alternative pronunciations for words in the Subscriber database. Experiments with the alternative pronunciations show an increase in recognition accuracy of over 5% compared to an otherwise identical system based solely on the baseform transcriptions of the words. The nature of many telephony applications limits the use of automatic methods for optimising pronunciation dictionaries for a particular accent group, however these techniques may be useful for improved modelling of coarticulation effects which are largely speaker independent. The rule-based network generation experiments have identified several rules which have a significant impact on recognition accuracy. These rules will be investigated further in the next stage of the work.

6. REFERENCES

[1] X Aubert, 'Improved Acoustic-Phonetic Modelling in Philips' Dictation system by handling liaisons and multiple pronunciations', Proc. Eurospeech Vol 1 pp.767–770 Madrid 1995.

[2] P Schmid R Cole & M Fanty, 'Automatically Generated Word Pronunciations from Phoneme Classifier Output', Proc. ICASSP 93 II pp.223–226.

[3] T Svendsen et al., 'Optimizing Baseforms for HMM-Based Speech Recognition' Proc. Eurospeech Vol 1 pp.783–786 Madrid 1995.

[4] N Cremelie & J P Martens, 'On the Use of Pronunciation Rules for Improved Word Recognition' Proc Eurospeech Vol 3 p.1747 Madrid 1995.

[5] A Simons & K Edwards, 'Subscriber – A Phonetically annotated Telephony Database', Proc. IoA Vol 14 Pt 6 Windermere 1992 p. 3

[6] A C Gimson, 'An Introduction to the Pronunciation of English', 4th Edition, Edward Arnold publishers 1989.

[7] S N Downey, 'Analysing Alternative Pronunciations to Improve Dictionary Baseforms', Proc IoA Vol 18 Pt 9 Windermere 1996.