Proceedings of the Institute of Acoustics

ANALYSING ALTERNATIVE PRONUNCIATIONS TO IMPROVE DICTIONARY BASEFORMS.

Simon Downey
downey@saltfarm.bt.co.uk
BT Labs, Martlesham Heath, Ipswich, Suffolk, IP5 7RE, UK

1. INTRODUCTION

It is widely accepted that word models derived from the concept of 'ideal' pronunciations are often too simple 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 ideal 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, however, the duration of the call is too short to collect sufficient data to successfully adapt the models. The adaptation approach is also unable to deal with words having more than one widely accepted pronunciation or continuous speech coarticulation phenomena which are likely to cause additional variations in the word initial and word final phonemes. 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 and 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 [5], Linear Discriminant Analysis on phone classes during training [1] and explicit modelling of adjacent word co-articulation effects (cross-word triphones can deal with some small effects, but not the more abrupt ones [6]).

The work presented in this paper investigates some of the above techniques in order to increase recognition accuracy of large vocabulary continuous speech tasks. The phonetically hand-annotated Subscriber database [7] 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. Possible avenues for improving the effectiveness of the dictionary model are also explored, these include the use of syllable-based speech models. The results of the experiments were also used to investigate whether certain key continuous speech effects as described in [10] are observable in the recognition output.

2. COMPARISON OF HAND TRANSCRIPTIONS WITH CITATION BASEFORMS

The Subscriber database contains sentences annotated manually at the phonetic level using a rich phoneme set comprising 74 different speech sounds. A method was required for comparing these phonetic transcriptions with 'idealised' phonemic transcriptions of the same sentence generated from a baseform pronunciation dictionary. To achieve this, a dp-match was performed between the dictionary-based transcription and the hand-annotated version. A slightly modified version of the 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 1[1].

```
CIT: 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  |
ACT: 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  |
CIT: @ n d  | e k z I b I t  | D @ m  | O l  | @U v @ | D @ | k aU n t  i |
ACT: A n    | I g z I b @ t  | D e m  | Q l  | @U v @ | D @ | k V  n t r I |
```

Figure 1: DP Match of *"Martin and Craig grow dwarf tulips and exhibit them all over the county"*

As can be seen, 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 2.1. 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 indicate typical phoneme string equivalences of 75-80%. An analysis of the word transcriptions generated indicated that approximately 30% of the total utterances are transcribed identically by both dictionary and hand methods. 67% of the utterances give different transcriptions, and the remaining 3% were labelled as dp matching errors[2], and subsequently checked by hand.

In total, 4968 validated alternative transcriptions were generated, giving on average an additional 4 transcriptions for each word. Some of the longer and more unfamiliar words generated many more than 4 alternatives, for instance 'trapezoidal' generated 20 pronunciations.

Table 1 illustrates the 10 pronunciation variants obtained for the word 'power'.

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

Table 1: Alternative Pronunciations of the word 'power'

1. Phonetic transcriptions in this paper use the SAM Phonetic Alphabet.
2. This was often caused by incorrect recitation of database sentences.

2.1 Continuous Speech Observations

Analysis of the dp matched transcriptions revealed several common traits in the fluent speech which are not present in the citation form:

- many final consonant deletions
- some schwa /@/ insertions at word endings
- initial/final phoneme deletions
- /t/ realisations as a glottal-stop (certain - 'cer-n'), /t/ is also sometimes deleted
- /dZ/ pronounced as /d Y/ (e.g. 'during')
- vowel substitutions abundant (encircled pronounced 'encircled', 'incircled', 'ancircled' or even just 'ncircled')
- common phrases slurred ('that are' /D { t A/ -> 'tha-uh' /D { @/)
- common suffixes are particularly prone to alternatives, for instance '-iasm' has 6 pronunciation variants in subscriber: /I z @ m/, /i z @ m/, /j z @ m/, /I z I m/, /i z I m/, /j z I m/
- Unusual words prone to large pronunciation variations - 19 different transcriptions were obtained for the words 'bourgeoisie' and 12 for 'ceremony'
- citation forms of word pairs where the end phoneme of the first word is the same as the initial phoneme of the next (i.e. 'it told' /I t t @U l d/) can cause recognition errors as the cross word phoneme is only realised once ('it-old' /I t @U l d/)
- elision/assimilation ('was shopping' -> 'wash-opping')

Not all of the effects may be dealt with simply by improved acoustic modelling, or the addition of alternative pronunciations to the recogniser vocabulary. The nature of the following three bullet points are user errors which may be handled more appropriately at the language modelling stage:

- false starts (ack-acknowledge et-etc...)
- incorrect utterances (anyone - anybody, bank - bang, county - country etc...)
- out-of-vocabulary utterances - words that don't actually exist in the English Language e.g. 'filch' pronounced as 'flench' (Scandinavian ?)

The other effects may be further divided into 'accent specific' and 'fluent speech' effects. Some of the observations may fall into both camps - for instance the co-articulation effects described above are particularly noticeable in certain regional accents. Other effects result entirely from the particular word sequence - specifically elision, assimilation, r-insertion[1] and strong/weak vowel forms ('the apple' /D i { p l/ vs 'the car' /D @ k A/). Finally, some effects are dependent on particular phoneme sequences - for instance syllabic consonants and vowel reductions.

1. Many words ending in /@ 3 1@ e@ A O/ with an 'r' in the spelling will usually cause an /r/insertion when followed by a vowel. For example: 'for' /f O/ , 'Arthur' /A T @/ , but 'for Arthur' -> /f O r A T @/.

2.2 Common Phone Confusions

The dp alignment process typically separates phone mismatches into insertion, deletion and substitution errors. A modification was made in order that substitution errors may be further separated into intra-word, word-initial and word-final substitutions. This allows investigation of 'data-driven' continuous speech articulation rules, which may not necessarily be inferred from a study of phonology alone.

Vowel substitutions dominate in all cases, the most frequent ones are listed in Table 2. The substitutions are categorised as intra-word, word-initial and word-final substitutions. For entries with two categories, the most frequent confusion is listed first. It should be noted that the substitutions are not simply vowel reductions to schwa, in fact the reverse is quite common. Vowel identity is strongly related to place of articulation, and hence will be affected by surrounding contexts given the physical constraints of the articulators. Different regional accents will also influence vowel production.

Citation	Realisation	Position	Example
I	@	intra-word/ word-initial	remind (ri-, ruh-)
I	i	intra-word/word-final	freshly (-lih, -lee)
V	U	intra-word	am*o*ngst (-mong-,-monk-)
A	(	intra-word/word-initial	mous*ta*che (-arsh,-ash)
@	(	word-inital/intra-word	*a*nd (uh-,a-)
@	U	word-final	to go (t'-go, to-go)
eI	aI	intra-word	sh*a*me (-ame,-ime)
(I) @	I@	intra-word (cross syllable)	re*a*lised (re-alised,realised)
@	Q	word-initial/intra-word	*o*f ('uv','ov')
i	I	word-final/intra-word	becoming ('beek-','bik-')
@	r	word-final	heat*e*r ('heat-uh','heat-err')
O	r	word-final/intra-word	*sure* ('shaw','shure')
@	i	word-final	*the* ('th-uh','th-ee')
t	@	word-final	bu*t* now ('but now', 'buh-now')

Table 2: Most Frequent Phoneme Substitutions

Non-vowel substitutions occur most frequently in word-final positions, where elision and assimilation effects may take place. The most common effect observed is the replacement of a word final /t/ sound with schwa /@/, this is possibly occurring because there is no glottal stop speech model. An example is given in the above table.

3. EXPERIMENTS

The variations between citation-based and manually annotated transcriptions were studied, allowing valid alternative transcriptions to be collated and added to the baseform dictionary. The final checking of the alternative pronunciations was performed manually. The performance of the enhanced baseform dictionary was then determined by a series of experiments. Variations in the model set complexity, and model size were also explored.

3.1 Database

The Subscriber 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 description of the database and the accent categories can be found in [7]. A subset of the database was used in this set of experiments, comprising the 5 phonetically rich sentences recorded by each talker. These sentences form a selection of those used in the 'SCRIBE' database collection. The complete subset consisted of 4874 sentences and contains a total of 1243 words.

Recognition experiments were based on speech models built from a combination of the phonetically rich sentences and 2 extra accent diagnostic sentences. Three sets of speech models were used. The first is based on the common UK english speech sounds, comprising 44 phonemes, the second uses a more rich symbol set with a total of 74 phoneme models. The final model set was built at the syllable level from the 1300 syllables present in the training database.

The syllable models used a variable mode, variable number of states topology, details of which may be found in [9]. The number of states allocated to a syllable was proportional to the number of phonemes contained in that syllable.

A standard cepstral frontend feature parameterisation was used.

3.2 Unconstrained Phoneme Recognition Experiments

Several recognition experiments were performed to explore unconstrained phoneme recognition performance of the Subscriber sentences. As well as providing an indicator to the effectiveness of the phoneme models, the resulting transcriptions can be used as a first stage in implementing an automatic baseform generation system (see [2]). Results are presented in Table 3.

The first experiment used the 44 phoneme model set to obtain a baseline performance figure. Increasing the richness of the models to the full 74-phoneme set caused a slight fall off in accuracy (Phon-74 result), possibly because some of the fixed-mode models were undertrained. Virtually identical recognition figures were obtained by mapping the 74 phoneme set model labels onto the nearest 44-set phoneme label and rescoring the experiment (Map-44).

An analysis of the phoneme confusion matrix indicated that the unstressed vowel /@/ and unvoiced plosive /t/ had the broadest set of confusions (these are also the most frequently observed phonemes in the database).

As the overall accuracies were so low, a phoneme bigram language model experiment was conducted. The phoneme bigram penalties were generated from the Subscriber phonetically rich sentence data, and add an extra 8% to recognition accuracy for the 74 model phoneme set.

The phoneme recognition performance was still considered poor. Because the speech models were trained and tested on the same data, much higher accuracy should have been expected. A possible explanation for the problem may be found by considering the phoneme duration statistics for the hand-annotated labels. Durations for some of the phonemes of less than 4ms occur. Given the data rate used in training and recognition is typically 16ms, and hence a 3-state phoneme model must have a duration of at least 16X3 = 48ms, these labels will not be correctly trained and may adversely affect surrounding phonemes during training.

3.3 Unconstrained Syllable Recognition Experiment

Recognition of the test set using the 1300 syllable models gave a lower overall recognition accuracy than the phoneme models (see Table 3). This was found to be due to a large number of the syllable models being undertrained due to the paucity of the training data for those syllables. The addition of a bigram model to the syllables gives a large increase in accuracy, resulting from the tighter grammar constraints inherent in the syllable bigram model.

Phone Set	% accuracy	% correct	%sub	%Ins	%del
Phon-44	29.4	38.9	42.5	9.5	18.6
Map-44	30.2	38.2	41.4	7.9	20.4
Phon-74	27.0	34.5	45.5	7.5	19.9
Phon74-bigram	35.4	40.8	36.2	5.4	23.0
Syllable	15.2	20.8	67.4	5.6	11.8
Syllable-bigram	36.5	51.3	43.1	14.8	5.5

Table 3: Unconstrained Phoneme/Syllable Recognition

3.4 Analysis of Continuous Speech Effects

The unconstrained phoneme recognition results were examined for certain key continuous speech effects described in [10] and outlined below. A comparison was made between the phonetically annotated data and the unconstrained phoneme recognitions. This gives 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.

Consonant Distribution - certain consonants have a restricted distribution, for instance /h r w j/ occur only before a vowel, however both the 44 and 74 model sets results indicated these phonemes were followed by a consonant in 8% of all /h r w j/ recognitions.

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 /t d f

v s z S Z/-/@ n/ 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.

There are three distinct allophones of the phoneme /l/. A 'clear' /l/ before a vowel or /j/ (for example 'leaf', 'million') and a 'dark' /l/ before a consonant and as a syllabic sound (e.g. 'feel', 'help', 'middle'. A third, partially or wholly devoiced /l/ sound follows stressed unvoiced plosives (/p t k/), e.g. 'please, 'clean'. Here the 74 phoneme model set results closely matched the occurrences labelled by the annotated data, whereas the 44 model set slightly favoured /l/-consonant sequences.

The other liquid /r/ and glides /w j/ are similarly devoiced following /p t k/, for these the gap between recognition output and annotation is slightly greater, with approximately half the amount of devoiced sequences being recognised are present in the annotations.

Plosive aspiration - unvoiced plosives /p t k/ are typically accompanied by aspiration, except in certain stressed-syllable initial sequences i.e. /sp- st- sk-/, and where the plosive is followed by another plosive or affricate - here the first plosive has no audible release (e.g. 'September' or 'object'). However, only about one third of these sequences are 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 again be affected by the 16ms data rate, with the restricted telephony bandwidth an additional factor.

Vowel Reduction - the length of long vowels /i: A: O: u: 3:/ and diphthongs is very much reduced when they occur in syllables closed by unvoiced consonants /p t k tS f T s S/ (e.g. 'park', 'cheap'). In these cases, vowel duration provides a significant clue to meaning. Both sets of models match the occurrences of vowel reduction sequences well with the annotated data, indicating that the vowel models may be inherently robust to this durational effect.

3.5 Unconstrained Word Recognition Experiments

The alternative transcriptions generated in Section 2 were added to a dictionary containing the citation forms of the Subscriber phonetically rich sentences vocabulary. Using the 44 phoneme model set and an unconstrained word recognition grammar, the recognition accuracy of the new dictionary was compared with that of the original citation-only form. Results of this experiment show a 5.4% increase in recognition accuracy, from 18.9% to 24.2%. The improvement is consistent when a phoneme bigram language model is used. Using the syllable models, a baseline figure of 16.7% is achieved - this is again greatly improved (to 71.2%) through use of the syllable bigrams.

4. 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 traits 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. Experiments with

syllable-based models indicate that they are able to out-perform phoneme based models if a good syllable language model is available. Poor training of uncommon syllables due to limitations in the available training data reduce their effectiveness in unconstrained recognition experiments.

Common continuous speech effects have been analysed which, when compared to a hand-annotated version of the data, indicate where particular models are performing poorly. It is suggested that improved acoustic and contextual modelling of these effects will benefit recognition accuracy.

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. A preliminary study of applying such methods to telephony databases has given promising results.

5. . 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 pp223-226.

[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 pp 1747 Madrid 1995.

[5] K N Stevens, 'Applying Phonetic Knowledge to Lexical Access', Proc. Eurospeech Vol 1 pp 3-10 Madrid 1995.

[6] J C Simon ed, 'Spoken Language Generation and Understanding', D Reidel publishers 1979. pp311-335.

[7] A Simons & K Edwards, 'Subscriber - A Phonetically annotated Telephony Database', Proc. IoA Vol 14 Pt 6 Windermere 1992 pp3-15.

[8] D G Ollason, "Variable Pool Size Tied Parameter Systems for Context-Dependent Sub-Word Unit Speech Recognition", Proc. IoA Vol 16 Pt 5, Nov. 1994.

[9] R Jones, "Syllable-based Word Recognition", MSc thesis, Univ Wales, Swansea, Oct. 1996.

[10] D Jones & A C Gimson, 'English Pronouncing Dictionary', J M Dent publishers 1982.

[11] M Edgington et al., 'Overview of current text-to-speech techniques Part 1', BTTJ Vol 14 No 1 1996.

[12] J Wells, 'Computer-coded phonetic notation of individual languages of the European Community', J IPA 19, pp32 1989.