

Speech Recognition, Syllabification and Statistical Phonetics

Melvyn J. Hunt

Novauris Laboratories UK

Cheltenham, England

melvyn.hunt@Novauris.com

Abstract

The classical approach in phonetics of careful observation of individual utterances can, this paper contends, be usefully augmented with automatic statistical analyses of large amounts of speech. Such analyses, using methods derived from speech recognition, are shown to quantify several known phonetic phenomena, most of which require syllable structure to be taken into account, and reveal some apparently new phenomena. Practical speech recognition normally ignores syllable structure. This paper presents quantitative evidence that prevocalic and postvocalic consonants behave differently. It points out some ways in which current speech recognition can be improved by taking syllable boundaries into account.

1. Introduction

This paper has three main purposes: (i) to argue that taking account of the location of syllable boundaries can make an important contribution to automatic speech recognition; (ii) to describe an automated, statistical approach to phonetics, which complements the traditional, detailed approach; and (iii) to present phonetic observations on American English in statistical form, some of which confirm expectations, some apparently run counter to expectations and some appear to be new.

In traditional phonetics, careful observations are made of individual utterances and the mechanisms by which particular sounds are produced are studied. What this leaves out is the quantitative study of phonetic phenomena over populations of speakers, for which we are proposing the term *statistical phonetics*. In contrast to other useful large-scale analyses that examined hand-labeled conversational telephone speech [1], the approach in this paper utilizes speech recognition techniques to analyze more carefully spoken speech material of better acoustic quality fully automatically.

The work reported here builds to some extent on the work of Greenberg [2] on the importance of syllable structure.

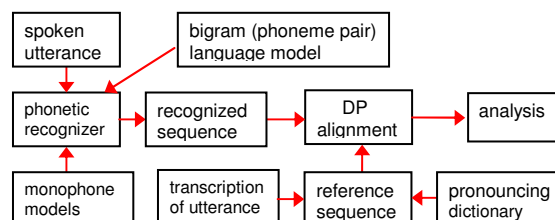
2. The Statistical Analysis Technique

In our statistical analysis technique, a corpus of utterances is submitted to a speech recognizer with a set of phoneme-sized acoustic models. Since the speech recognizer has no grammatical constraints, any phonetic unit can follow any other. However, there is a simple bigram language model, derived from our pronouncing dictionary.

For each utterance, we also use this pronouncing dictionary to generate from the lexical transcription of the utterance an expected phonetic sequence (or a network of phonetic sequences when there are multiple possible pronunciations).

This expected sequence is aligned to the phonetic sequence produced by the recognizer using dynamic programming symbol-sequence alignment with knowledge of the insertion, deletion and substitution probabilities of the phonetic units. To the extent that the recognizer provides an accurate phonetic transcription of each utterance and that the alignment process is reliable, we can observe how the pronunciations differ from what the dictionary predicts.

Fig 1: Block diagram of the analysis process



2.1. The Phonetic Inventory and Acoustic Models

The inventory of phonetic symbols used in our dictionary is similar to that used in the CMU dictionary [3]. As with that dictionary, vowels are given one of three levels of stress, there are no special symbols for syllabic consonants, and voiced /h/ is not distinguished from unvoiced /h/. However, unlike the CMU dictionary, the dictionary in this work does not use a postvocalic /r/ symbol, but rather has a set of rhotic vowels. Also, unlike the CMU dictionary, it marks alveolar flaps. They are assumed to occur when medial /t/, /d/, /l/ and /n/ are immediately preceded by a vowel and immediately followed by an unstressed vowel. Prevocalic /t/ and /d/ are given optional pronunciations [ch] and [jh] before /r/ in our dictionary.

The *monophone* acoustic models used in this work have three states and 64 Gaussian components with no tying. Although context-dependent models, such as triphones, are known to be potentially much more effective for speech recognition in general, they would be unsuitable for the investigations here. By defining the context, triphones exclude the possibility of observing context effects, or at least obscure them. For example, if a phoneme B when preceded by phoneme A and followed by C had a realization acoustically closer to phoneme D, it would not be apparent in the triphone ABC, which would resemble ADC, but with monophones, D would be preferred to B. Moreover, phonetic boundaries in triphones are rather arbitrary, making it difficult to study durations. With monophones, boundaries are more meaningful.

2.2. Methodological Questions

Statistical phonetic analysis is attractive because it allows the study of large amounts of speech from a population of speakers and it provides information on the frequencies of occur-

rence of the phenomena observed. Not depending directly on human judgments, the results are in a sense objective.

However, automatic statistical analysis poses some problems. There is a problem of circularity because the acoustic models are built using a dictionary, which itself presumes how words will be pronounced. The method relies on the assumption that the dictionary accurately represents the pronunciation of most words (using multiple pronunciations in some cases). Moreover, the language model, derived from the dictionary, exacerbates any circularity problem.

A second problem is that the phonetic decisions made by the decoder will often differ from those that humans would make. Partly, this is because of the inadequacy of current speech recognition technology, resulting in effectively random errors. Partly, it is because phonemes often cannot be identified context-independently as monophone models require. However, differences also occur because the decision criteria applied by the recognizer are systematically different from those applied by human listeners. We shall see later that differences of this kind can sometimes be interesting. The danger of being misled by these two properties of the decoder can be alleviated by: (i) seeking systematic behavior rather than isolated phenomena; (ii) by listening to a selected subset of the recordings to check the automatic decisions; and (iii) by working with recordings of good acoustic quality and confining ourselves initially to reasonably careful speech.

A third potential problem may arise from the uneven distribution of the vocabulary. For instance, in our corpus of names and addresses the word “street” is very common. Consequently, almost all examples of the prevocalic consonant sequence /s t r/ occur in this word, and any observed properties of the consonant string may be peculiarities of “street”.

The final problem is that in any reasonably large population of speakers there will be differences in the phonology. Consequently, the phoneme must be regarded as a useful idealization rather than a rigidly defined unit.

3. Syllabification

By syllabification, we mean determination of the location of syllable boundaries in the region between pairs of vowels. With up to three consonants before the vowel and up to four after it, English syllable structure is relatively complex, making syllabification not obvious. We assume here that syllable boundaries are always located between phonemes rather than within a phoneme.

As described by Wells [4], several criteria can be used to decide where to place a boundary. In some words, syllable boundaries are unambiguously defined by the rules of English phonotactics, since only a specific set of consonants strings can occur before a vowel in a syllable and another specific set can occur after the vowel. For example, the word “ringlet” can have its syllable boundary only between /ng/ and /l/, since the sequence /ng/ + /l/ cannot occur within a syllable.

These rules are less clear-cut for non-native words, such as “Renoir”. More seriously, the phonotactic rules are not sufficient to determine many syllable boundaries uniquely.

In these ambiguous cases, we take account of three ranked phonetic tendencies and a potentially competing morphological tendency. In decreasing order of strength, the phonetic tendencies are: (i) associating a consonant with the higher stress vowel (with schwa treated as having lower stress than other unstressed vowels); (ii) attaching a consonant to a pre-

ceding short (or *lax*) vowel (/ih/, /ah/, /uh/ or /eh/) but not to preceding long vowels or diphthongs; (iii) making prevocalic consonant strings as long as possible. The potentially competing tendency is to make syllable boundaries coincide with morpheme boundaries. Thus, the first two syllables in the word “beefeater” may be split using the phonetic tendencies (no. (iii) applies here) after the first vowel; alternatively, the syllables may be split on morphological grounds after the /f/. Which location is chosen depends on whether the speaker sees the word as a compound (someone who eats beef) or an integrated unit (meaning a guard at the Tower of London).

By applying these rules, partly automatically and partly manually, our pronouncing dictionary of almost 300,000 lexically unique items has been syllabified. Inevitably, not all decisions will coincide with all productions, but our analyses suggest that the great majority do.

4. The Speech Material Analyzed

The speech material used in this work is a subset of our US name-and-address corpus, recorded in a quiet office. The speakers were mainly Harvard students, from all parts of the USA. They read aloud US names and addresses presented on a monitor as they might appear typed on an envelope.

The resulting speech is more careful than typical conversational speech. However, since the speakers were paid to record a given number of names and addresses, to minimize the time taken some spoke extremely quickly.

In the analyses, recordings from 43 male and 51 female speakers were used, providing in all 7,977 different name-and-address combinations. Excluding the ZIP codes, these contained 81,626 words, comprising a vocabulary of 20,162 orthographically distinct items, and about 372,000 phonemes.

5. Statistical Analyses

5.1. Detection of Prevocalic and Postvocalic Consonants

Table 1 compares the detection rates of consonants in pre- and post-vocalic position in stressed and unstressed syllables.

Table 1: Proportions of certain consonants correctly recognized (C) according to the dictionary, and detected (D) as the correct consonant or as some other consonant.

	PRE-VOCALIC CONSONANT						POST-VOCALIC CONSONANT					
	In Stressed Syllable			In Unstressed Syllable			In Stressed Syllable			In Unstressed Syllable		
	No.	C %	D %	No.	C %	D %	No.	C %	D %	No.	C %	D %
p	2564	93	99	817	71	96	482	66	83	76	65	93
t	4764	86	94	4718	56	78	5288	41	53	1308	41	55
k	5923	93	99	2322	75	98	4406	87	94	462	81	89
b	3411	92	97	1485	83	92	220	71	98	11	27	100
d	1740	79	90	2241	73	80	3065	74	77	1788	50	54
g	1398	87	98	918	84	94	372	68	93	175	65	93
f	4912	84	97	508	70	92	1116	30	76	99	19	90
s	7732	91	95	4261	92	98	4020	92	97	2379	90	98
v	1711	44	76	2240	71	95	3049	88	91	20	60	70
z	554	85	99	791	70	99	1533	53	89	970	47	91
m	3631	96	99	2173	88	98	1434	68	93	426	64	96
n	3805	82	94	1602	75	92	9046	79	94	7151	89	97
l	5918	76	93	1471	66	90	3856	81	92	2028	80	92
sh	557	94	100	437	96	100	470	96	99	17	94	100
ch	2695	94	100	288	74	100	541	99	100	10	80	100
jh	3805	94	100	745	91	100	327	76	98	105	59	98

In the analyses shown in Table 1 and elsewhere, examples of

digits were excluded. This was because digits were overrepresented in the training material, making the recognition accuracies of phonemes in digits anomalously high.

Cross-word geminates were also excluded, because geminates are almost always interpreted by the recognizer as a single phoneme, and the association by the alignment of such a unit to the prevocalic or postvocalic consonant is arbitrary.

Table 1 shows that prevocalic plosives are more reliably detected and identified than the corresponding postvocalic consonants, at least when the number of examples available for analysis exceeds a few hundred. This is not surprising, since postvocalic plosives are typically not released.

We can also see, again not surprisingly, that with both prevocalic and postvocalic plosives detection and identification is more reliable in stressed than in unstressed syllables.

The same tendency is seen with the other consonants, though less consistently. However, even for consonants such as /l/ that do not show the behavior seen for plosives, the pattern of confusion with other phonemes made by the recognizer is quite different for the prevocalic and postvocalic forms, indicating that they are acoustically different.

5.2. Consonants Preceding Rhotic Vowels and /r/

When /t/ is followed by /r/ in a prevocalic sequence, the /t/ is recognized with a [ch] model in about 97% of the cases, and /d/ in the same context is equally frequently recognized with the [jh] model. Sometimes, the segment sounds clearly affricated, but often it sounds like a normal plosive. When /t/ and /d/ precede a rhotic vowel, the recognizer prefers affricate models about 67% of the time, even though the consonant almost never sounds affricated. Evidently, the metric applied by the recognizer finds the presumably retroflexed plosives to be acoustically closer to the affricate models than to the plosive models. (*Postvocalic /t/ and /d/ show no affrication before rhotic vowels, and 6% and 21% respectively before /r/.*)

Rhotic vowels also have a surprising effect on the way in which /v/ is recognized. When preceding a rhotic vowel, /v/ is recognized as [b] in 31% of the examples, compared with 37% as /v/, while before a non-rhotic vowel, it is recognized as [b] in only 11% of the examples, against 63% as [v].

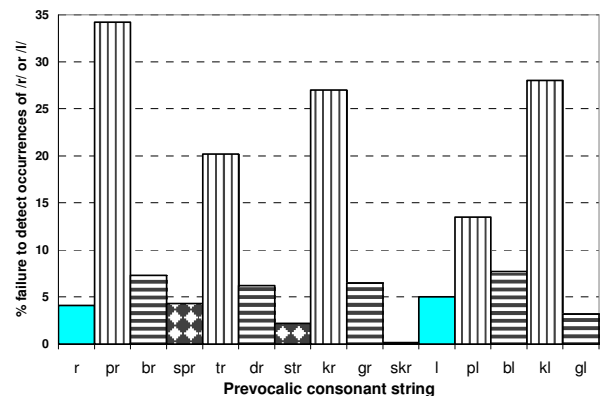
5.3. Aspiration following Unvoiced Plosives

In American English prevocalic unvoiced plosives are normally aspirated, at least when preceding a stressed vowel. Aspiration might be observed directly by the insertion of [h] in the output of the decoder between a prevocalic unvoiced plosive and a vowel immediately following it. This does indeed happen before stressed vowels in 0.6% of cases after /p/, 1.2% after /t/, and 1.7% after /k/. The low incidence is presumably because the acoustic model of the plosive itself models aspiration. By contrast, in the 867 cases when a vowel follows an unvoiced plosive marked postvocalic, only one example of [h] insertion was found, an incidence of around 0.1% (it was in the digit sequence “eight-oh,” and presumably resulted from a syllabification shift of the /t/ to the start of “oh”).

A prevocalic /s/ immediately before the plosive is known to suppress aspiration, and indeed no [h] insertion at all occurs in this case. On the other hand, when there is a postvocalic /s/ in a preceding word with no pause before the plosive, the incidence of [h] insertion appears to be unaffected. This and the other results in this section give us confidence that our syllabification decisions are reasonably reliable.

When there is another consonant between a prevocalic unvoiced plosive and the vowel, aspiration manifests itself as devoicing, or partial devoicing, of the consonant. The speech recognizer tends to miss such devoiced consonants. Fig 2 shows this effect with /t/ and /l/, comparing the detection failure rates when the consonant occurs in a prevocalic sequence: (a) on its own, (b) preceded by an unvoiced plosive, (c) preceded by the corresponding voiced plosive, and (d) preceded by /s/ and an unvoiced plosive. The aspiration following the unvoiced plosive causes a high rate of detection failure. The aspiration-suppressing effect of /s/ reduces the detection failure rate. In the material analyzed, there were only six examples of /s k r/ (with no /r/ detection failures), and no examples of /s k l/ or /s p l/ (/s t l/ is impossible).

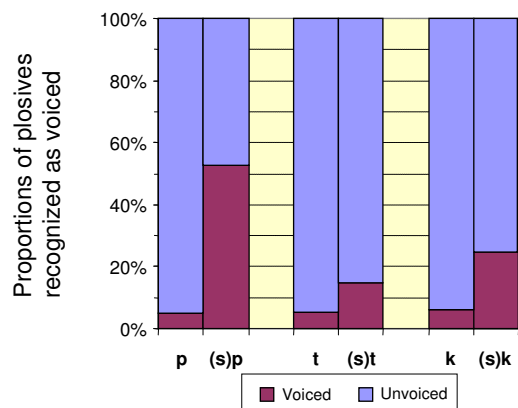
Fig 2. Frequencies with which the recognizer failed to detect /t/ and /l/ in prevocalic consonant contexts.



When a syllable boundary occurs between /s/ and /t r/ (the only sequence for which there are enough examples), the detection failure of /t/ is 31%, *i.e.* as if the /s/ has no effect. Similarly, when a syllable boundary occurs between /p/, /t/ or /k/ and /r/, the detection rate of /r/ is 100%, as if the plosives have no effect.

The aspiration-suppressing effect of /s/ preceding a nominally unvoiced plosive can also be seen in the way that the plosive is recognized. Fig 3 shows that the probability of unvoiced plosives being recognized as their voiced counterparts is much greater when they are preceded by /s/.

Fig 3. Proportions of times in which an unvoiced plosive is recognized as voiced when it is (i) syllable-initial, and (ii) preceded by /s/ in the same syllable.



5.4. Effect of Postvocalic Consonants on Vowels

When a vowel is followed by a postvocalic nasal consonant, we expected that anticipatory nasalization of the vowel would depress the accuracy with which it was identified. In confirming this expectation, we confined ourselves to vowels with primary stress, since they are generally identified most reliably. Since we noticed that postvocalic laterals also depressed identification accuracy, presumably through anticipatory velarization, we systematically measured this effect as well.

Rather to our surprise, we found that, when we had an adequate number of examples (judged to be 100), the reduction in identification accuracy of vowels preceding a postvocalic lateral consonant was almost always greater than that for the same vowel preceding a postvocalic nasal consonant. The single exception was the vowel /ih/, whose identification accuracy was only 29% before nasals, compared with 55% before laterals and 66% in all other contexts. However, this vowel one of only three in our system to have separate stressed and unstressed acoustic models. When the identifications of stressed /ih/ with the unstressed acoustic model are counted as correct, and the identification accuracy before a nasal consonant increases to 69%, while that before a lateral increases to only 59%, and that of the rest increases to 80%. This now fits the general pattern.

5.5. Place Assimilation of Alveolar Consonants

English is known to allow the alveolar consonants /t/, /d/ and /n/ to assimilate their place of articulation to that of a following consonant [5, 6]. Since our recognizer frequently reports only one phoneme when two plosives or two nasals occur sequentially, we can search for alveolar place assimilation only when the manner of production of a consonant pair is also different (*i.e.*, plosive followed by nasal, approximant or fricative, or nasal followed by plosive, fricative or approximant).

We could find little evidence of place assimilation within words, possibly because such assimilation is often already anticipated in the orthography (*e.g.* “import” not **inport*”).

Across words, there is some evidence of place assimilation, but it is more limited than might be expected. The only large effect is word-final /n/ recognized as the velar nasal [ŋ] 29% of the time before word-initial /k/ and 28% of the time before word-initial /g/. For comparison, the overall incidence of word-final /n/ being recognized as [ŋ] is just 3%.

The overall incidence of word-final /n/ recognized as [m] of 2%, does not increase at all before /w/ or /f/. It rises to 4% before word-initial /p/ and 5% before /b/, but with only 11 and 17 instances of /m/, this small effect may not be real.

No place assimilation at all was detected with /t/ or /d/.

Phoneticians also document [5, 6] the palatalization of word-final alveolar fricatives, /s/ and /z/ to [sh] or [zh] when followed by palatal or palato-alveolar consonants (/y/, /sh/, /zh/, /ch/ and /jh/). We could see evidence of this only before /jh/, where /s/ was recognized as [sh] 6% of the time and /z/ as /sh/ 4%. While these incidences are much higher than for word-final contexts not preceding /jh/ (40 times higher for /s/ and 14 times for /z/), the absolute incidence rates are small, and the effect could be illusory, since only 10 utterances in total show it. (Compare this with an undoubtedly real effect, namely the disappearance of word-final /t/ and /d/ after /n/ and before any consonant or silence, where the incidence is 73% and 64% respectively, where over 600 utterances show the effect.)

5.6. Cross-Word Anticipatory Assimilation of Voicing

For words whose dictionary entry ends in /z/, this phoneme is recognized as [z] 66% of the time and as [s] 20% of the time when the next word starts with a vowel or voiced consonant. On the other hand, when the next word starts with an unvoiced consonant, the recognition as [z] drops to 36%, while its recognition as [s] rises to 53%, suggesting that the phoneme is assimilating the voiceless property of the following sound.

A much smaller effect in the reverse direction is seen with word-final /s/: before unvoiced consonants, the recognition rate as [s] is 93% and as [z] it is just 1%, while before vowels and voiced consonants the rates become 90% and 3%, respectively. No voicing assimilation effects have been found with any other word-final consonants.

6. Implications for Speech Recognition

In context-dependent triphone models for speech recognition, the locations of syllable boundaries are not usually taken into account [7]. The results in Section 5 show that such boundaries have a major influence on context effects (the surprisingly weak effects reported in Subsection 5.5 occur *across* syllable boundaries). Using syllable boundaries in context specification and in context-tree building consequently offers an opportunity to improve the models and hence speech recognition accuracy. Insights from statistical phonetic analyses may help design better question sets for context-tree building.

While not a novel proposal, the potential of using automatic alignment methods to improve the phonetic specifications in dictionaries used for speech recognition is underlined by the work reported here. Such methods are particularly attractive for assigning probabilities to alternative pronunciations. We have preliminary evidence that our own dictionary has benefited significantly from our automatic analyses.

7. Acknowledgements

The author thanks his former colleagues at Novauris UK for help with this work, and especially John Bridle, who wrote the alignment analysis system.

8. References

- [1] Byrne W., Finke M., Khudanpur S., McDonough J., Nock H., Riley M., Saraclar M., Wooters C. and Zavalagkos G., “Pronunciation Modelling Using a Hand-Labelled Corpus for Conversational Speech Recognition,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, 1998.
- [2] Greenberg S., “Speaking in Shorthand – A Syllable-Centric Perspective for Understanding Pronunciation Variation,” *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Keldrae, pp. 47-56, 1998.
- [3] <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
- [4] Wells J.C., “The English phonemic system and its notation: Syllabification,” *The Longman Pronunciation Dictionary*, pp. xiv-xvi, Longman Group UK Ltd, 1990.
- [5] O’Connor J.D., *Phonetics*, Pelican, 1973, p. 250.
- [6] Shockey L., *Sound Patterns of Spoken English*, Blackwell, Oxford, 2003, p. 18.
- [7] Young S.J. *et al*, *The HTK Book*, <http://htk.eng.cam.ac.uk>