THE EFFECTS OF WORD BOUNDARY AMBIGUITY IN CONTINUOUS SPEECH RECOGNITION.

JONATHAN HARRINGTON[+] ANNE JOHNSTONE[++]

The Centre for Speech Technology Research, University of Edinburgh, Scotland.
Also Department of Linguistics[+] and Department of Artificial Intelligence[++], University of Edinburgh, Scotland.

ABSTRACT

This study assesses the effect of employing different phonological units on parsing a given string of phonemes into words in a continuous speech recogniser. It is shown that when the input utterance is encoded using a representation intermediate between the broad classes in Huttenlocher & Zue [3] and the 44 phonemes of Received Pronunciation, the number of possible word strings found from the input utterance is usually in excess of 10 million. Even when all 44 phonemes are implemented, an input utterance of 4 - 10 words in length can be parsed into in excess of 10,000 word strings if word boundaries are not identifed prior to lexical access. In the final part of the study, it is shown that the number of such parses can be reduced if stress is represented in the input utterance and lexicon.

INTRODUCTION

In some of the speech understanding systems of the *ARPA* project [1] as well as the feature-based, continuous speech recogniser being developed at Edinburgh University, one of the main tasks of the syntactic and semantic components is to filter out grammatically unacceptable and meaningless word strings which are the output of a lexical access component: they must, then, be able to identify the target word string *please let us know*[1] from a list of other strings such as *please letter snow* and *please let a snow*, both of which are possible word parsings of the input phonemic string /p l ii z l e t @ s n ou/[2]. The total number of word strings which can be derived from a given phonemic string depends on several factors, such as the number of entries in the lexicon, the parsing strategy and the units which are used to represent words phonemically. This paper forms part of a larger study of which the main goal is to devise a set of units which is optimal both from the point of view of acoustic-phonetic processing (i.e. it must ultimately be possible to identify such units with a high degree of accuracy from the acoustic waveform) and from the point of view of syntactic/semantic filtering (i.e. the number of word strings passed to the syntactic/semantic components should be minimal).

An initial aim has been to implement a *mid-class* representation in the continuous speech recogniser [2] being developed at Edinburgh University. The prime motivation for analysing the acoustic waveform into mid-classes is that they should be easier to identify than the 44 phonemes of Received Pronunciation (R.P.): for example, an analysis of the acoustic waveform into mid-classes such as /B/, voiced stop, is (arguably) likely to result in better identification scores than its analysis into the members of /B/, that is /b/, /d/, /g/. At the same time, it has been shown that when all the words of a 20,000 word lexicon are represented in classes that were much 'broader' than our mid-classes (i.e. there are fewer broad-classes than mid-classes and therefore a greater number of phonemes, on average, in a broad-class than a mid-class), around 1/3 of the words are still uniquely identifiable [3]; when the lexicon is represented in mid-classes, the percentage of uniquely identifiable words will presumably be considerably greater. This may, therefore, be a strong argument for analysing the acoustic waveform as far as the mid-class level and allowing the syntactic and semantic components to filter out the impermissible word strings which have resulted from using mid-classes rather than phonemes. However, the statistics on discriminability in the lexicon do not take account of the fact that in continuous speech, word boundaries are more difficult to identify from a given mid-class string compared with a phonemic string. Thus, while at a phonemic level, the sequence /m g l/ can only be parsed into /m # g l/ [4] (e.g. *same glass*), at a mid-class level (i.e. /N B L/), the unambiguous identification of the word-boundary is no longer possible: since the mid-class category /N/ includes /n/ and since /B/ includes /d/, /N B L/ could also be parsed as /N B # L/ (e.g. *sand layer*), or indeed /N B L #/ (e.g. *sandle*). Since phonotactic constraints often no longer successfully apply at the mid-class level, the total number of ways of parsing a given mid-class string into words is likely to increase considerably despite the fact that the lexicon remains highly discriminable when represented in mid-classes. The first experiment was designed to determine the magnitude of this increase and to assess whether this would place an unmanageable burden on syntactic and semantic filtering.

METHOD

The lexicon of the continuous speech recogniser includes the 4000 most frequent words from the American Heritage Dictionary [5]. Each entry consists of an orthographic form, a phonemic citation form (R.P.) and a key for accessing syntactic tag information. Using a phonological rule interpreter written in INTERLISP-D to run on the Xerox 1100 series [6], a set of phonological reduction rules was applied to this lexicon to derive fast speech forms (known as reduced forms) which were stored together with the citation form under the corresponding orthographic entry. Details of the reduction rules are given in [7].

The lexicon containing citation and reduced forms was then compiled into a discrimination tree in which, working from left-to-right, phonemic entries with identical phoneme sequences share the same branch(es). Thus, *tee, tea, teach, teacher* and *tedious* share the same branches as far as the second phoneme /ii/ at which point there is a division to /d/ (the continuation through the tree for *tedious*) and to /ch/ (the continuation for *teach* and *teacher*). A terminal branch is attached to a phoneme node whenever a sequence of phonemes forms a word. A fragment of the tree is shown in Figure 1.

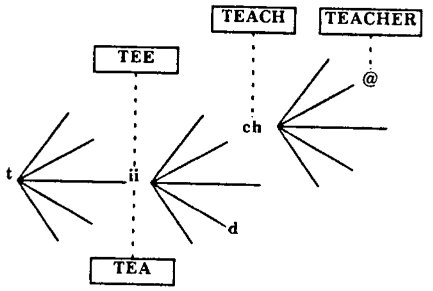

FIGURE 1: The tree-structured phonemic lexicon implemented in the continuous speech recogniser.

The acoustic front end of the continuous speech recogniser analyses the acoustic waveform into a string of phonemes (henceforth *input utterance*) which are matched against the discrimination tree to locate possible word boundaries. The matching process takes place from left to right through the input utterance and when a word is matched, it is stored on a word lattice. Thus, if the first phonemes of a string were /t ii ch i ng w i l/ (*teaching will...*), *tea* and *tee* would be stored on the word lattice. Subsequently, two searches, or paths, are continued: the first is the continuation from /ii/ to /ch/ and /i/; the second is from the initial /ch/ node that begins words such as *chide*, *choke* etc. to the /i/ node, in this case, of words such as *chin* and *chimpanzee*. The second path in this example would be terminated for two reasons: there are no citation, or reduced forms, beginning with /ch i ng/; and also because the fragment /ch i/ is not a citation or reduced form of any word. Only those paths which enable a complete parsing of the input utterance are passed to the syntactic and semantic components *and only such complete paths are considered in the statistics on total number of paths in this paper*. An example of the parsing process is shown in Figure 2.

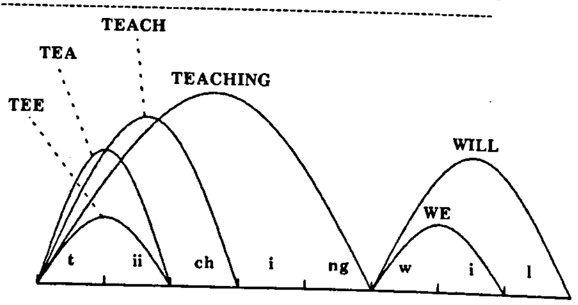

FIGURE 2: the paths show that there is only one possible parsing of the sequence /t i ch i ng w i l/, into *teaching + will*.

When an input utterance is represented in mid-classes, it is first expanded into all possible phonemic representations which are then each matched against the tree as described above.

Phonemic transcriptions were made by a trained phonetician of 50 sentences produced by one R.P. speaker: these sentences included a mixture of 'phonemically dense' sentences (e.g. *I know no minimum* whose consonants are entirely nasal); sentences taken from a 'phonemically balanced' passage; sentences from Section H of the Lancaster-Oslo-Bergen corpus [8]; sentences from a corpus of business dictation collected at C.S.T.R; and sentences that consisted of words which were uniquely identifiable (in isolation) when represented in mid-classes. These transcriptions were then converted automatically to their corresponding mid-class representations[3].

The hand transcriptions, rather than the phonemic strings derived automatically by the acoustic front end of the continuous speech recogniser, were input to the discrimination tree. As such, the hand transcriptions can be considered as a perfect analysis by the continuous speech recogniser of an acoustic waveform into a string of mid-classes excluding any representation for word boundaries, syllable boundaries or stress.

RESULTS I

Parses into words	$< 10^3$	$10^3 - 10^4$	$10^4 - 10^5$	$10^5 - 10^6$	$10^6 - 10^7$	$>10^7$
Number of utterances	1	4	2	4	7	32

TABLE 1: distribution of mid-class input utterances in terms of number of word strings found. The first column on the left denotes, for example, that 1 (out of 50) utterances was parsed into less than 1000 word strings.

From Table 1 it can be seen that 32 out of the 50 input utterances of 4 - 10 words in length were parsed into 10 million, or more, word strings. The smallest number of parses was 82, the largest 3.25 x 10^{18}. The average number of parses was 8.47 x 10^{16}.

DISCUSSION I

When *isolated* words are represented in mid-classes, the number of words that are uniquely identifiable decreases from around 98% (for words represented phonemically) to 70%. Such a statistic would suggest that analysing the acoustic waveform to the mid-class level is a viable alternative to performing a phonemic analysis. However, if word boundaries cannot be identified prior to matching the input string against the lexicon, the syntactic and semantic components could have to find the correct word string from over 10 million competing strings. This may be too much of a burden to place on higher level processing; furthermore, the hand-transcriptions represent the best possible mid-class analysis of the acoustic waveform by a connected speech recogniser. The number might increase if the input utterance were derived from an acoustic front end analysis that contained a large number of errors. Second, the lexicon only contained around 4000 lexical items; if larger lexicons are implemented (e.g. 20,000 words), the number of word strings found will undoubtedly increase. Finally, the speaker produced utterances of 4 - 10 words in length. It is certainly possible to produce longer utterances without pausing; in this case, the total number of ways in which the corresponding mid-class representation could be parsed into words would again increase.

The next section reports statistics on the number of possible word parsings of *phonemic* input utterances. In this case, the original hand-transcriptions were matched against the tree-structured lexicon.

RESULTS II.

Parses into words	0 - 10	11 - 100	$100 - 10^3$	$10^3 - 10^4$	$> 10^4$
Number of utterances	15	16	8	8	3

TABLE 2: distribution of phonemic input utterances in terms of number of word strings found. The first column on the left denotes, for example, that 15 (out of 50) input utterances were parsed into 10, or less, word strings.

Table 2 shows that when the input utterances are phonemically represented, over half of them parse into 100, or less, competing word strings. The average number of word strings was 2491; there were two utterances that could only be parsed into one word string; the maximum number of word strings for any utterance was 66,528.

It was not necessarily the case that longer input uttterances (where length is defined as number of phonemes or number of words intended by the speaker) necessarily gave the greatest number of parsings into word strings: there was no correlation between number of phonemes in the utterance and number of parses into words (r = -0.07, not significant); neither was there a significant correlation between number of words produced by the speaker and number of possible parses of its phonemic representation into word strings (r=0.11), although there is a trend to show that these two variables are positively correlated.

DISCUSSION II

Even when the input utterance is phonemically encoded, the syntactic and semantic components could have to filter out over 10,000 competing word strings from the target word string. While this figure may not be unmanageable for syntactic and semantic filtering, it is clearly desirable to seek to reduce this figure further.

As a means towards reducing the number of word strings, we considered the possibility of increasing the number of 'sound units' by using *allophones* in both the input utterance and the lexicon. The fact that the number of word strings should decrease using an allophonic representation is easily demonstrable. Phonemically, *plea* is represented as /p l ii/ which also embeds the lexical item *Lee*, phonemically /l ii/. On the other hand, *Lee* would not be embedded within *plea* in an allophonic representation since these would be encoded as [li] and [pl̥i] respectively. However, this advantage would be lost if the allophones that were the product of word-internal context-effects were also caused by context-effects across word boundaries: thus if /l/ in *Lee* were realised as a voiceless [l̥] in a moderately fast production of *..type Lee some results*, *Lee* would once more be embedded within *plea* even at an allophonic level of representation. There is some experimental evidence [9] to suggest that such word-boundary coarticulation of /l/ is possible. If the majority of identifiable allophones can occur as a result of coarticulation both across word boundaries and word-internally, the case for introducing this kind of phonetic representation is considerably weakened. Furthermore, the acoustic front end is faced with the additional difficulty of identifying *specific* allophones; given the fact that our recogniser does not obtain a perfect analysis even at the mid-class level, it may be unrealistic to assume that it would be able (in the short-term, at least) to differentiate, for example, between aspiration before stressed vowels, the various allophones of /h/ and [l̥].

An alternative means of increasing the number of units in the input utterance, and thereby decreasing the number of word strings found, would be to include *stress* in the lexicon and input utterance. In order to test this hypothesis, stressed vowels were differentiated from unstressed vowels in the lexicon by inserting a '*' symbol before the former; no distinction was made between different levels of stress; thus, *conversation*, which is normally marked for secondary stress on /k o n/ and primary stress on /s ei/ is represented as /k *o n v @ s *ei sh @ n/ in citation form. With this type of representation, in which each vowel phoneme (except schwa) can be marked for stress, an additional 19 units are introduced into the phonemic inventory. The same set of hand-labelled transcriptions were then re-transcribed including the '*' symbol to identify word-stress. Most of the function words were not marked for stress either in the lexicon or in the hand-labelled data. These modified transcriptions were matched against the modified tree-structured lexicon as described above.

RESULTS III

Parses into words	0 - 10	11 - 100	$100 - 10^3$	$10^3 - 10^4$	$> 10^4$
Number of utterances (U)	15	16	8	8	3
Number of utterances (S)	20	14	10	5	1

TABLE 3: distribution of phonemic input utterances in terms of number of word strings found. (U)/(S) denote the input utterances unmarked/marked for stress as described above.

The results in Table 3 show that when the input utterance and the lexicon are marked for stress, 34 utterances are parsed into less than 100 word strings and 6 utterances are parsed into 1000, or more, word strings; the corresponding results for a phonemic input utterance unmarked for stress are 31 and 11 respectively. For the stress-marked input utterances, the average number of word strings for a given utterance was 624 compared with 2481 word strings for the unstressed, input utterances.

Table 4 shows similar statistics for stressed and unstressed *mid-class* utterances. The stressed mid-class utterances were derived from the stressed phonemic utterances by automatic conversion into mid-classes, but with stressed vowels marked: thus, /k *o n v @ s *ei sh @ n/ is represented as /P *BV N V CV S *D S CV N/ where /P/, /BV/, /CV/, /S/ and /D/ are the mid-classes *voiceless stop*, *back vowel*, *central vowel*, *strong fricative* and *diphthong* respectively.

Parses into words	$< 10^3$	$10^3 - 10^4$	$10^4 - 10^5$	$10^5 - 10^6$	$10^6 - 10^7$	$>10^7$
Number of utterances (U)	1	4	2	4	7	32
Number of utterances (S)	3	6	5	8	8	20

TABLE 4: distribution of mid-class input utterances in terms of number of word strings found. (U)/(S) denote the input utterances unmarked/marked for stress as described above.

As in the phonemic utterances, the inclusion of stress in the mid-class utterances improves performance: there are 14 stressed, mid-class utterances that were parsed into 100,000, or less, word strings and 20 such utterances parsed into 10 million, or more, word strings: the corresponding results for the mid-class, *unstressed* utterances are 7 and 32 respectively. The average number of word strings for mid-class, stressed utterances is nevertheless very high at 5.54×10^{13}.

DISCUSSION III

The results show that including stress in the process of matching an input utterance to the lexicon clearly decreases the average number of word strings found. Implicit in this result is the assumption that the acoustic front end would be able to identify stressed vowels in an utterance. There are some reports [10], [11] of a high level of success in the automatic identification of stressed vowels; work in this area on our own continuous speech recogniser is currently in progress.

This study has also not been able to take account of *sentence* stress which could cause some unstressed vowels in our lexicon to be stressed. Thus, since *can* (auxiliary) is marked as unstressed in the lexicon, our lexical access model would fail to find the appropriate word string in the utterance *I can come* (emphasis on *can*); the effects of sentence stress on the automatic identification of lexically stressed vowels is the subject of a future investigation.

CONCLUSIONS

If the acoustic front end of an automatic speech recognition system is unable to locate word boundaries, the syntactic and semantic components must be implemented to identify the target word string from a potentially large number of competing word strings. The number of ways in which a given input utterance can be parsed into word strings depends on at least two factors: the type of parsing strategy and the units which are used for the phonemic representation of words. This study has been concerned with the latter problem and has shown that an input utterance represented entirely in mid-classes, or even phonemically, may place an unmanageable burden on syntactic and semantic filtering. The possibility of introducing stress into the input utterance and lexicon was explored with preliminary, promising results. The introduction of stress in this way may provide a basis for implementing a mixed mid-class and phonemic representation: if stress enables a substantial reduction in the number of word strings found, it may be possible to represent some of the phonemes which are notoriously difficult to identify from the acoustic waveform (such as weak fricatives /th, dh, h, f, v/) by their mid-classes.

REFERENCES

[1] Smith A. & Sambur M. (1980) Hypothesizing and verifying words for speech recognition. In Lea W. (ed.) *Trends in Speech Recognition* 139-165 Englewood Cliffs: New Jersey.

[2] Dalby J, Laver J. & Hiller S.M. (1986) Mid-class phonetic analysis for a continuous speech recognition system. In Lawrence R. (ed.) *Proceedings of the Institute of Acoustics* 8.7, 347-354. Institute of Acoustics: Edinburgh.

[3] Huttenlocher D.P & Zue V.W. (1983) Phonotactic and lexical constraints in speech recognition. *Proceedings of the American Association for Artificial Intelligence Conference* 172-176.

[4] Lamel L. & Zue V.W. (1984) Properties of consonant sequences within words and across word boundaries. *IEEE Institute of Acoustics, Speech and Signal Processing* 42.3.1 - 42.3.4.

[5] Carroll J., Davies P. & Richman B. (1971) *The American Heritage Word Frequency Book*. Houghton-Mifflin: New York.

[6] Cutting D. & Harrington J.M. (1986) Phongram: an interpreter for phonological rules in automatic speech recognition. In Lawrence R. (ed.) *Proceedings of the Institute of Acoustics* 8.7, 461-470. Institute of Acoustics: Edinburgh.

[7] Harrington J.M., Laver J. & Cutting D. (1986) Word-structure reduction rules in automatic, continuous speech recognition. In Lawrence R. (ed.) *Proceedings of the Institute of Acoustics* 8.7, 451-460. Institute of Acoustics: Edinburgh.

[8] Johannson S., Leech G.N. & Goodluck H. (1978) *The Lancaster-Oslo/Bergen Corpus of British English*. Oslo University: Department of English.

[9] Bladon R.A.W. & Al-Bamerni A. (1976) Coarticulation resistance in English /l/. *Journal of Phonetics*, 4, 137-150.

[10] Marshall C. & Nye P. (1983) Stress and vowel duration effects on syllable recognition. *Journal of the Acoustical Society of America* 74, 433-443.

[11] Lea W. (1980) Prosodic aids to speech recognition. In Lea W. (ed.) *Trends in Speech Recognition* 139-165 Englewood Cliffs: New Jersey.

NOTES

1 Our thanks to Jim Hurford for this example.

2 The machine readable alphabet for the 44 phonemes of R.P. is shown below:

/p/	*pea*	/f/	*fan*	/l/	*lee*
/b/	*bead*	/v/	*van*	/r/	*road*
/t/	*tea*	/th/	*think*	/w/	*win*
/d/	*day*	/dh/	*then*	/y/	*you*
/k/	*key*	/s/	*sing*	/m/	*man*
/g/	*guy*	/z/	*zoo*	/n/	*name*
/ch/	*chew*	/sh/	*shoe*	/ng/	*sing*
/jh/	*judge*	/zh/	*measure*		
		/h/	*hat*		
/ii/	*we*	/o/	*hot*	/ei/	*stay*
/i/	*hit*	/oo/	*saw*	/ai/	*sigh*
/e/	*head*	/u/	*could*	/oi/	*toy*
/a/	*had*	/uu/	*who*	/au/	*now*
/aa/	*hard*	/@/	*the*	/ou/	*go*
/i@/	*here*	/u@/	*sure*	/e@/	*there*
/@@/	*first*				

3 We thank Maggie Cooper for her assistance in converting from phonemes to mid-classes.

4 We are grateful to Julian Kupiec for writing software to count the number of word strings.

This research was supported by SERC grant number GR/D29628 and is part of a collaborative project with Plessey Research & Technology, the Husat Research Group at Loughborough, Imperial College of Science and Technology, and C.S.T.R., Edinburgh University. Our thanks to John Laver and Henry Thompson for many helpful comments on an earlier draft of this paper.