



THE APPLICATION OF PHONEME SEQUENCE CONSTRAINTS TO WORD BOUNDARY IDENTIFICATION IN AUTOMATIC, CONTINUOUS SPEECH RECOGNITION.

Jonathan Harrington*, Ian Johnson^o, Maggie Cooper*.

ABSTRACT

This study examines the set of CV, VC, CVC and some CCVC sequences which are non-occurring in monomorphemic words in a 20,000 word lexicon. A preliminary analysis suggests that many sequences in which the prevocalic and postvocalic consonants are similar, or identical, are excluded. The sequences are discussed in relation to 'reduced forms', characteristic of fast speech, word boundary assimilation and lexical access.

INTRODUCTION

In every language, there are restrictions, often referred to as *phonotactic constraints*, on the serial order of phonemes. In English, for example, it is a phonotactic constraint of CCCV syllables, that the only segment that can occur as the first consonant is /s/; that the second consonant must be a voiceless stop and that the third segment has to be one of the approximants /l, r, y, w/¹. As Lamel & Zue (1984) have suggested, a knowledge of all the legal phonotactic sequences of a language can provide important clues to the identification of word boundaries. Thus, since the sequences /m g/ and /m.g l/ are excluded in English, a word boundary can be unambiguously located after the first segment, i.e. /m # g l/ as in *same glass*.

This study is concerned with the identification of phoneme sequences which could provide clues to the presence of word boundaries, but differs from some earlier investigations both in the linguistics (Cairns & Feinstein, 1982) and speech technology literature (Lamel & Zue, 1984), in analysing constraints that may hold across CV, VC and CVC sequences, rather than across a string of consonants. The reason for the choice of constraints across CV, VC and CVC is that such consonant-vowel sequences are fundamental to the phonological structure of utterances in English; this, in turn, suggests that it may be possible to apply such constraints relatively frequently. In (1), for example,

(1) /sh i sh a l n a u b a i sh e @ z @ t/ (*she shall now buy shares at...*)

there is little scope for the application of the phonotactics of *consonant* sequences in the attempted location of word boundaries; however, since there are no words in the 4000 word lexicon described in Harrington, Laver & Cutting (1986) that contain the sequences /sh i sh/, /au b/ and /ai sh/, at least three word boundaries may be detectable using a knowledge of sequence constraints on VC and CVC sequences. Furthermore, it may be the case that some CV, VC and CVC sequences which are impermissible in citation forms of monomorphemic words, are also excluded in their corresponding reduced forms. Reduced forms, such as the elision of the schwas in *solicitor* to derive reduced form /s l i s t @/ are common in the production of continuous speech at a moderate tempo and are therefore quite likely to occur in the input utterance to a continuous speech recogniser. Clearly, if reduced forms violate the sequence

*The Centre for Speech Technology Research and Department of Linguistics, University of Edinburgh, Scotland.

^oPlessey Research & Technology, Roke Manor, Romsey, England.

constraints that are derived from an analysis of the phonemic composition of citation forms, such sequences may be of limited use in locating word boundaries. However, in the 9300 word lexicon (4000 citation forms, 5300 reduced forms) discussed in Harrington, Laver and Cutting (1986), there were no reductions containing the sequences /ai sh/, /au b/ or /sh i sh/. Such sequences could be applied, therefore, to a phoneme lattice containing either citation or reduced form pronunciations for the detection of possible word boundaries.

METHOD

From Rockey's (1973) *Phonetic Lexicon*, a list was made of all CVC sequences which were non-occurring in monosyllabic words and this list was matched against a modified version of the *Oxford Advanced Learner's Dictionary of Current English* (Hornby, 1974) containing around 20,000 entries in order to derive CVC sequences which are excluded in the phonemic entries of lexical items. The program for the matching task was written by Ian Johnson in Interlisp-D to run on a Xerox 1100 workstation.

The phonemic entries of the lexicon (Received Pronunciation) were modified to accord with the criteria outlined in Harrington, Laver & Cutting (1986) for deriving citation forms. None of the sequences containing the phonemes /i@/ (*fear*), /u@/ (*poor*) or /e@/ (*there*) were examined in this study, but CVC sequences with medial /@/, which are excluded in Rockey's *Phonetic Lexicon*, were included in the match against the lexicon.

RESULTS

CV, VC and CVC sequences which are excluded from the phonemic, citation form entries of the lexicon are given below. When exceptions to the sequence constraints are given, lexical items that are morphologically related to the exceptions are not listed (thus, only *trapezoid*, but not *trapezoidal*, is listed as an exception to the /z oi/ constraint in (2)).

Constraints on CV sequences.

The following CV sequences are excluded:

- (1) /v u/, /v uu/, /th u/, /th uu/, /dh u/, /dh uu/. Exceptions - /v u/: *bivouak*; /th uu/: *thulium*; /v uu/: *voodoo, rendezvous*.
- (2) /th oi/, /dh oi/, /z oi/, /sh oi/, /y oi/. Exception - /z oi/: *trapezoid*.
- (3) /dh aa/, /dh @@/, /dh ai/, /dh au/. The only exceptions to /dh ai/ and /dh au/ are archaic function words such as *thy* and *thou*. /dh au/ can also occur across an internal word boundary, e.g. *without*.
- (4) /ch au/, /jh au/. Exceptions - /ch au/: *chowder*; /jh au/: *jowl, joust*.
- (5) /r @@/.
- (6) /y ai/, /y au/. Exceptions - /y ai/: *yikes*; /y au/: *yowl*.

Constraints on VC sequences.

The following VC sequences are excluded:

- (7) /au b/, /au m/, /au fl/, /au v/, /au sh/, /au jh/, /au k/, /au g/. Exceptions - /au jh/: *gouge*; /au b/: only across an internal word boundary e.g. *cowboy*.

(8) /oi p/, /oi b/, /oi m/, /oi f/, /oi v/, /oi th/, /oi dh/, /oi sh/, /oi zh/, /oi ch/, /oi jh/, /oi k/, /oi g/. (i.e. /oi/ only precedes alveolars). Exceptions - /oi m/: only across a stem-suffix boundary, e.g. *employment*; /oi f/: *coif*; /oi k/: *boycott*.

(9) /uv/, /u th/, /u dh/.

(10) /oo v/ Exception - *wharve*.

(11) /ai sh/.

(12) /uh th/. Exception - *nothing*.

(13) long vowels + /ng/, where long vowels include /ii, aa, oo, uu/ and all diphthongs.

Constraints on CVC sequences.

The following sequences are excluded:

(14) /f V p/, /z V p/, /th V p/. Exceptions - /f V p/: *fop*; /th V p/: *Orthopaedic*, *thorp*, *absorption*.
/z V p/: *marzipan*, *zip*, *zap*, *usurp*.

(15) /th V v/. Exception - *thieve*.

(16) /v V f/, /th V f/. Exceptions - /v V f/: *vivify*; /th V f/: *thief*.

(17) /v V dh/, /th V dh/, /ch V dh/, /jh V dh/. Exception - /v V dh/: *nevertheless*.

(18) /th V th/, /z V th/, /ch V th/, /jh V th/.

(19) Where /V/ ≠ /i/ or /@/, /z V t/. Exceptions - *dessert*, *ersatz*, *exhort*, *resort*, *exotic*, *gazette* and *rosette*.

(20) Where /V/ ≠ /i/ or /@/, /z V d/. Exceptions - *trapezoid*, 'Z' (the letter of the alphabet), *Zodiac*.

(21) Where /V/ ≠ /i/ or /@/, /z V z/. Exception - *disease*.

(22) /sh V s/ and, where /V/ ≠ /i/ or /@/, /z V s/. Exceptions - /sh V s/: *chassis*; /zVs/: *disaster*, *exhaust*, *possess*, *zest*.

(23) /th V sh/, /sh V sh/, /ch V sh/, /jh V sh/. Exceptions - /sh V sh/: *shush*, *hashish*, *shish kebab*.

(24) /z V ch/, /jh V ch/, /sh V ch/.

(25) /th V jh/, /sh V jh/, /ch V jh/. Exceptions - /th V jh/: *lethargy*; /ch V jh/: *charge*. It should be noted that these sequence constraints are upheld at a more 'abstract' level of phonemic representation in which *lethargy* and *charge* contain a medial /r/.

(26) /g V k/. Exceptions - *gecko*, *gherkin*, *gobbledygook* and *oligarchy*.

(27) Where /V/ ≠ /i/ or /@/, /z V g/, /ch V g/. Exception - /ch V g/: *chug*.

(28) (see also (12)). Where /V/ ≠ /i/, /n V ng/.

DISCUSSION

In many of the C_1VC_2 sequences which are non-existent, or at least extremely rare, C_1 and C_2 are similar, where degree of similarity might be defined by comparing their feature matrices. Thus, with the exception of a small number of words, neither /f V p/ nor /g V k/ occur. In the fricative set, /th V th/, /z V z/, /sh V sh/, /ch V sh/ and /jh V sh/ are, with few exceptions, also excluded.

With the exception of /n V ng/, all of the constraints on CVC sequences discussed above involve the interaction between pairs of fricatives and stops. A subsequent investigation of CCVC sequences has shown that certain repetitive sequences which have nasals and liquids as the second consonant in the onset and in the coda are also excluded. Thus, in monomorphemic words, /s N V N/ is excluded, where /N/ is any nasal (the only exception is *smarmy*, which contains a medial, underlying /r/ at a more abstract level of representation); and with the exception of *flail*, *slalom* and *haplology*, /C l V l/ sequences are also excluded, where /C/ is any consonant. There is also a parallel in repetitive stop sequences of a similar structure since /s p V p/ (exception: *dyspepsia*), /s p V b/, /s k V g/, /s k V k/ are also excluded. Finally, the absence of /s m V p/, /s m V b/, /s m V f/ and /s m V v/ (in monomorphemic words) demonstrates again the similarity of the segments in the onset and coda. A detailed investigation is currently in progress in order to determine whether, for all the CVC and CCVC sequences that do not occur in monomorphemic words, the prevocalic and postvocalic segments are more similar than would be predicted by chance.

As discussed above, it is only possible to implement such constraints for detecting word boundaries in continuous speech recognition if they are not violated in the phonological structure of reduced forms (characteristic of a moderate tempo). While a preliminary investigation has suggested that reduced forms also obey many of the constraints, some constraints are not upheld if word boundary assimilation rules apply (e.g. /t/ \Rightarrow /k/ in a fast production of *Scott could go* resulting in the 'impermissible' sequence /s k V k/). Therefore, the identification of word boundaries from such constraints requires the prior 'on-line' application of word-boundary assimilation rules on the incoming phoneme lattice derived from the acoustic waveform. It is also evident that some of the sequence constraints are only upheld in *monomorphemic* words (thus /s N V N/ occurs in *snow#man* and *displace+ment*). The application of the sequence constraints discussed in this paper would be equally valid in a continuous speech recogniser in which lexical suffixes and stems were autonomously represented, and accessed.

REFERENCES

- C.E. Cairns. & M.H. Feinstein. *Linguistic Inquiry* 13, 193-225 (1982).
J. Harrington, J. Laver & D. Cutting. *Proc. Inst. Acoust.* 8.7, 451-460.
A.S. Hornby *The Oxford Advanced Learner's Dictionary of Current English*. (1974).
L. Lamel & V. Zue. *IEEE Inst. Acoustics, Speech and Sig. Proc.* 42.3.1 - 42.3.4 (1984).
D. Rockey. *Phonetic Lexicon*. Heyden : Oxford (1973).

NOTES

1 Details of the Machine Readable Phonetic Alphabet are given in Harrington, Laver & Cutting (1986).

This research was supported by SERC grant number GR/D29628. Our thanks to Richard Shillcock for comments on an earlier version of this paper.