STATISTICAL ANALYSIS OF LARGE-SCALE LEXICAL CORPUSSES IN THE CONTEXT OF CONTINUOUS SPEECH RECOGNITION SYSTEMS (CSR SYSTEMS)

Michael Bundgaard
Institute of General and Applied Linguistics, University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen S, Denmark

Abstract
A number of statistical properties of allophonic and prosodic distributions of the 8000 most frequent words of Danish were investigated. With CSR in mind the distributions of the words were examined with different degrees of partial description of the allophones and considering also to some extent the metric characteristics of the initial syllable at word level was uncovered in the efforts to find a way of detecting syllable boundaries.

1. Introduction
In connection with the development of HEAD, a Continuous Speech Recognition system (CSR), a number of statistical characteristics of the most frequent Danish words were investigated. The aim was to clarify the allophonic and prosodic distributions and constraints in large scale lexica, and to take advantage from these insights in constructing the HEAD system.

The Danish 'HEAD' system [1],[2] is a phonetically based CSR system consisting of 3 components. The first is an input component, which with the aid of a neural net converts the speech signal to phonetic feature vectors, one every 10 msec. These vectors form the input to the phonetic decoding component, which by means of a two step algorithm produces a sequence of allophone segments. The first step is a segmentation into intervals with homogeneous feature vectors. The second step is an allophonic description of the segments in the shape of a set of distinctive features. Corresponding to a partial description of the allophone.

The last component is the expert system performing the symbolic reasoning on the allophonic segments that were detected by the phonetic decoding component. The allophones are used as input to a lexicon, which through a search and a verification phase (henceforth subsumed in the term 'access') produces word hypotheses. The upper level consists of an island driven parser which combines the words into legal phrases with respect to the present grammar.

In CSR-systems based on two-step allophone detection algorithms, consisting of segmentation and a classification-part (like our HEAD system), the broad phonetic specification is realistically based as well on partial segmentation of the allophone sequence, as on partial allophonic specifications. To cover the phonetic entities the prosodic features of the large scale vocabularies are also drawn into account.

 Accordingly the distribution of words in the Danish lexicon is investigated with respect to (a) a partial allophonic specification, (b) a partial segmentation of the allophone sequence and (c) combined with different levels of prosodic description.

In this study several allophonic and prosodic phenomena in large-scale lexica will be investigated and the implications for CSR-systems will be discussed.

The database used in this investigation consists of the 8000 most frequent words in Danish. In addition to an allophonic transcription of each word it includes a frequency count of the word. The frequency calculation is based on the Aarhus School of Business corpus of standard Danish [6]. The transcription was made under the Danish Speech Synthesis Project [7]. In the database the phonetic transcription was based on 37 allophones and 5 prosodic symbols. Subsets of the 8000 words were applied in several tests and the subsets were always taken from the most frequent words.

2. Partial allophonic specification
First the distributions of the 17 vowels and 20 consonants in the 8000-word corpus were computed. The ten most frequent
allophones are \{C, @, s, n, d, l, g, i, e, E\}, and there were 58.9% consonants and 41.1% vowels. In Fig. 1 the average number of consonants and vowels per word are shown for different lexicon sizes. The tendency towards short high-frequency words is evident. One-syllable words constitute 37% in the 1000-word corpus and 17% in the entire 8000 word corpus, whereas the corresponding figures for words with more than 3 syllables show an increase from 2% to 13%.

Fig. 1. The average number of vowels and consonants per word for the 4 different lexicons.

In the following a partial description of the allophones is considered, corresponding to a 4-class description of the vowel and consonant categories. The organization of the allophones into these 8 groups are listed in Table I, where also the distribution of the groups in the 8000 corpus is displayed.

Table I. An 8-class description of the Danish allophones.

| Stop consonants | B, p, t, k, b, d, g | 19% |
| Fricatives | s, s', sh | 12% |
| Nasals | n, m, n | 11% |
| Sonorants | l, r, v, w, j, d | 17% |
| High front vowels | i, e, y, l | 9% |
| Low front vowels | e, æ, e, æ | 9% |
| High back vowels | u, o, å, ø | 13% |
| Low back vowels | A, æ, Æ, O, C | 10% |

If the input to the lexicon from a acoustic-phonetic component is not a full specification of each allophone, but rather a broad class specification like that in Table I, some words receive identical phonetic descriptions. Take the word 'start' - the full specification [sdAd] will, according to Table I, be converted to [SBAB], which coincides with 'stop' - [sdCb] -> [SBAB]. The set of words having the same pattern is by Shipman and Zue [3] termed the 'cohort' of the pattern. In the following different degrees of breadth in the phonetic description is investigated with respect to cohort size.

The broadest type is achieved by simply replacing each consonant with a 'C' and each vowel with a 'V'. With the 1000 word corpus 109 different patterns are found with this 2-class description. Fig. 2 displays the cohort sizes for CV-patterns as a function of lexicon size. The figure shows that on average there are 9 words with identical pattern in the 1000-word vocabulary and 13 out of the 8000. In worst case a lexical access with 2-class patterns yields 146 rising to 606 word matchings. The frequency normalized average cohort is perhaps the most interesting with respect to CSR, as it takes the frequency of the words in the cohort into account.

In Fig. 2 a comparison is made with two phonetic classifiers. The first one works with the 4 vowel groups of Table I and only one consonant group, whereas the second has 4 consonant groups and a single vowel group. From Fig. 3 it is clear that the consonant-based classifier is in every respect superior to the vowel-based one, although both classify the allophones into 5 groups.

In both cases the standard average does not increase significantly as a function of lexicon size, in fact the cohort average is smaller for the 8000 than for the 4000 words for both classifiers. The consonant-based 5-way classifier gives cohort sizes at the same level as found in English [3] with the 6-class description, which also only operated with one vowel group. In the following when referred to a 5-group classifier the consonant-based is meant.

Increasing the number of groups to the 8 from Table I produces the relatively small cohort sizes shown in Fig. 4. The average is constant at 1.4 words per cohort, and the frequency normalized average is not exceeding 7 words. Finally the number of groups has been extended to 12 (8 consonants and 4 vowels) giving less than 4 words in the frequency normalized average.
The experience with manual segmentation of Danish shows two kinds of undecided segment boundaries. The first is the general problem of demarcating two vowels of different quality or two different fricatives in sequence. In the 5 and 8-group partial descriptions these neighbouring sounds correctly end up in the same class. In the 8000-word corpus these identical-category clusters are on average found in more than 10% of the words based on the 8-class phonetic description, while under the 5-class consonant based description the figure increases to more than 20% of the words.

The other problem is attached to the group of sonorant consonants that in consonant clusters like [skr-, pl-] and in postvocalic positions like [-aD] are very hard to segment.

The 8-class and 5-class descriptions were adjusted to meet problem with the identical-category allophone clusters. For example the two words [Emn0] and [en0] were both transformed to [ENV] in the 8-class description and therefore belonged to same cohort. Table II 6 shows a comparison of cohort sizes for the standard 8-class and the revised 8-class descriptions. The difference in averages is very small, and it is constant for the different lexicon corpuses. Similarly small differences were obtained with the 5 class approach.

Table II. Comparison of cohorts for the standard 8-group and the revised 8-group classifier (8000 words corpus).

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Norm. av.</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard 8-class</td>
<td>1.44</td>
<td>6.56</td>
<td>25</td>
</tr>
<tr>
<td>Revised 8-class</td>
<td>1.53</td>
<td>7.71</td>
<td>26</td>
</tr>
</tbody>
</table>

Another method for eliminating the effect of the different types of missed segment boundaries is the so-called superphone strategy [2]. This means that not all allophone classes are taken into account at the lexicon search phase. The superphones alone count and the rest of the allophones are taken into account only in the verification phase of the lexical access. As the group of sonorant consonants are generally troublesome with respect to segmentation, all allophones of this group are ignored, leaving 7 groups to describe the words in the lexicon. The cohort results —excluding an allophone group containing 17% of the allophones—are shown in Fig. 6.

3. Segmentation

In the two-step allophone detection algorithm the segmentation part is critical. It gives rise to many of the inaccuracies or faults in the description. Experience with the segmentation problems has been gained in segmenting and labelling the EUROM speech database. More particularly, focus has been directed to those cases where expert phoneticians either deliberately decided not to commit themselves to the placing of a segment boundary or those points where their segmentation diverged. The EUROM speech database is established within the ESPRIT Speech Assessment Methodology (SAM) project [8].

Applying the basic principle of "not describing the allophones further than you're sure what you say is correct" to the segmentation task implies that false alarms are really serious segmentation errors. The "partial segmentation" reflects the abovementioned situations where the expert phoneticians were hesitant to draw a segment border. Still, these cases are classified as missed boundaries, because it produces segments containing more than one allophone.

The experience with manual segmentation of Danish shows two kinds of undecided segment boundaries. The first is the general problem of demarcating two vowels of different quality or two different fricatives in sequence. In the 5 and 8-group partial descriptions these neighbouring sounds correctly end up in the same class. In the 8000-word corpus these identical-category clusters are on average found in more than 10% of the words based on the 8-class phonetic description, while under the 5-class consonant based description the figure increases to more than 20% of the words.

The other problem is attached to the group of sonorant consonants that in consonant clusters like [skr-, pl-] and in postvocalic positions like [-aD] are very hard to segment.

The 8-class and 5-class descriptions were adjusted to meet problem with the identical-category allophone clusters. For example the two words [Emn0] and [en0] were both transformed to [ENV] in the 8-class description and therefore belonged to same cohort. Table II 6 shows a comparison of cohort sizes for the standard 8-class and the revised 8-class descriptions. The difference in averages is very small, and it is constant for the different lexicon corpuses. Similarly small differences were obtained with the 5 class approach.

Table II. Comparison of cohorts for the standard 8-group and the revised 8-group classifier (8000 words corpus).

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Norm. av.</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard 8-class</td>
<td>1.44</td>
<td>6.56</td>
<td>25</td>
</tr>
<tr>
<td>Revised 8-class</td>
<td>1.53</td>
<td>7.71</td>
<td>26</td>
</tr>
</tbody>
</table>

Another method for eliminating the effect of the different types of missed segment boundaries is the so-called superphone strategy [2]. This means that not all allophone classes are taken into account at the lexicon search phase. The superphones alone count and the rest of the allophones are taken into account only in the verification phase of the lexical access. As the group of sonorant consonants are generally troublesome with respect to segmentation, all allophones of this group are ignored, leaving 7 groups to describe the words in the lexicon. The cohort results —excluding an allophone group containing 17% of the allophones—are shown in Fig. 6.
4. Prosody

It is assumed that the prosodic features may contribute to lexical access in two ways: as a cue for the essential word boundary detection, and together with allophonic information to minimize the cohort sizes.

The first interesting point is the distribution of stress in Danish words. In English, Cutler and Carter [5] found a predominance of strong initial words enough to suggest a working where strong syllables form the onset of lexical words. In continuous speech many words lose the stress, accordingly a database consisting of spontaneous spoken sentences would be the most appropriate for stress experiments with respect to CSR. Unfortunately no such database exists in Danish, so the distribution of word level stress is examined here. Table III shows the metrical status of the initial syllable in the 8000 word corpus.

Table III. The percentage of stressed / unstressed initial syllables in the 8000-word corpus.

<table>
<thead>
<tr>
<th>Monosyllable words</th>
<th>Polysyllables with</th>
<th>Polysyllables with</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>initial stressed</td>
<td>initial unstressed</td>
</tr>
<tr>
<td></td>
<td>syll.</td>
<td>syll.</td>
</tr>
<tr>
<td>16.8 %</td>
<td>63.4 %</td>
<td>19.8 %</td>
</tr>
</tbody>
</table>

In Table III the secondary stressed syllables are regarded as unstressed. This means that 80.2% of the words in the vocabulary have stressed word onsets. If the frequency of occurrence of the words is drawn into account the words with a stressed initial syllable increases to 94.7. These results show the same tendency as found by Cutler and Carter in English at word level. Yet further analysis of spoken sentences in Danish is needed to establish a basis for the design of a word boundary algorithm.

To test the minimization effect on cohort sizes the 5-class consonant-based allophone description was augmented with the prosodic Danish features: length, stressed, primary stress and secondary stress. This reduced the cohort sizes to a level close to the 8-class description mentioned above. The frequency normalized average cohort was as follows for the 1000-word corpus: 5 classes = 9.3, 5 classes + prosody = 3.9 and 8 classes = 3.1.

Which suggests that the prosodic features contribute to minimization of cohort sizes roughly as much as the 4-class division of the vowels.

The last topic of the investigation was based on the 'islands of clarity' principle based on the fact that the allophones in stressed syllables are more distinctly pronounced than in the unstressed syllables. In general this principle plays a role at the expert system level in HEAD [1], but can perhaps also affect the structure of the lexicon. The idea is to operate with a very broad description of the allophones in the stressed syllables, and a fairly distinct classification in the stressed syllables, or as proposed in Cutler and Carter [5], 'that only strong syllables are input to any lookup process'.

In the present experiment the 8-group classifier was used in the stressed syllables and the 2-group (CV) classifier in the other cases. The cohort size for the 8000-word corpus is very close to the ones obtained with the 8-class description of all syllables, as shown in Table IV. The results in Table IV show that the 2-class description of allophones in unstressed syllables yields small cohort sizes close to the standard 8-class description. Whereas ignoring the unstressed syllables tends to increase the cohort sizes considerably.

5. Conclusion

The frequency normalized cohort sizes achieved by the 8 and 8+2 class descriptions are less than 10 words when based on the 8000 word corpus (the 5+prosody and the 8+8-'identical' category) classifiers were also under this limit. The superphone method and the 5-group classifier achieved less than 25 words. This leaves relatively few words for the verification phase and consequently provides an acceptable level of precision for accessing the lexicon. These results were obtained with known word boundaries. The 10-million word strings in small sentences without word boundaries were based on a 4000-word corpus [4]. Assuming that word boundaries are given, a simple computation of how many word strings would match a 5-word sentence (based on the 4000 corpus and on the 5-class description) amounts to 1.2 million word strings and 2600 for the 8-class description. Roughly estimated a good word verifier could distinguish between the words in the cohort and reduce it into half, i.e. reducing the word strings from 1.2 million to 38,450 and from 2600 to 81. This leads to the conclusion, that the 5-group classification is perhaps acceptable as input to the lexicon search, but only the 8-class description combined with at least some word boundaries detected comes close to an acceptable performance, giving the syntactic parser 81 combinations of word strings to work with. It is obvious that the word boundaries, which reduce the word strings from more than 10 millions to 2600 are an object for closer research. In Danish an algorithm based on prosodic features appears to be the key to this problem.

References:


