AUTOMATIC RECOGNITION OF WORDS
DIFFERING IN DISTINCTIVE QUANTITY

G. Kuhn and K. Ojamaa
96 Leigh Avenue Princeton, NJ 08540 USA

ABSTRACT
We report the results of an experiment on talker-dependent, connected recognition of 10 Estonian
CVCV words that differ in distinctive quantity. The words were spoken, and recognized, in sentence pairs of
the form "Did you say (word 1, word 2, word 3)? No, I said (word 4, word 5, word 6)." The test sentences were
spoken either at the same rate as the training sentences, or at a much faster rate. Each word was modelled with
spectral estimates for four variable-duration states.

The best recognition results obtained on the test words spoken at the training (faster) rate, were 88%
(641) without probabilities or likelihoods of durations, 87% (683) with likelihoods of durations, and 85% (771) with likelihoods of duration
ratios.

We conclude that speech rate can be a major problem for automatic recognition of these words, and that in
these experiments the problem was not completely overcome using ratios of successive state durations.

INTRODUCTION

In the field of automatic speech recognition, there is new interest in implicit [1] and explicit [2,3]
modelling of speech state durations. However, unless there is a correction for speech rate, expected state
durations may be inappropriate. In languages which use distinctive quantity, like Estonian or Finnish,
inappropriate state durations could lead to misrecognition of a large number of words.

In this paper, we report the results of an experiment on automatic recognition of 10 Estonian CVCV
words that differ in distinctive quantity. Estonian is described as having three consonant
quantities and three vowel quantities: short, long and
overlong [4,5,6,7]. Within our vocabulary of 10 Estonian
words to be recognized, 4 words participated in 2 two-way
quantity contrasts: 

CORPUS

Speech was recorded while one of the authors (KD)
read a prepared text. The text consisted of a
randomization of 36 occurrences of each of the 10 words,
embedded in 60 repetitions of the sentence pair "Kas sa
ültesid (Did you say) 'word 1, word 2, word 3'? No, I
said 'word 4, word 5, word 6'." The randomization was constrained so that each word occurred
6 times in each position in each sentence of the pair.

The text was recorded 3 times. In the first two
recordings, one sentence pair was spoken every 6 seconds.
In the third recording, one sentence pair was spoken
every 4 seconds. The first recording was used to train
the word models, while the second and third recordings
were used for the recognition tests.

Each recording was digitized at 10000 samples/s.
The digitized recordings were parameterized in
centisecond frames using a 10-channel, filter-bank
spectrum analyzer.

WORD MODELS

We used 15 "word" models, one each for Kas sa,
ültesid, Ei ma, ültesin, (pause), and the 10 CVCV words.
The models for ültesid and ültesin had six states. The
models for all other words had four states. Each state
had an initial segment of fixed duration, a center
segment of variable (possibly 0) duration, and a final
segment, again of fixed duration. The minimum duration of
a state was thus the sum of the durations of its
initial and final segments. The minimum durations of the
four states in the 10 CVCV words were 3+1, 3+3, 3+2, and
2+3 cs.

The word models were trained using two passes
through the training productions. Pass 1 started with DP
alignments [8] to the "miniai". The miniai for each word is
that training production which has minimum average
distance to all training productions of the word. Pass 1
alignments minimized the distance between each training
production and the miniai. Means and a covariance matrix
were computed over the spectra aligned to each segment of
each hand-marked state of the miniai. Pass 2 alignments
maximized the probability of the training productions
given the Pass 1 means and covariances. Duration
estimates (minimum, average, maximum) for each state were
produced from the Pass 2 alignments.

In some experimental conditions, spectral estimates
were tied across word models, i.e., the weighted average
of the means and the weighted average of the
outer-product matrices were computed over corresponding
segments of the states looped together below:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(RECON DITION)

The routines for connected recognition computed a spectral match score for the best path through an entire recording of |0,1,2| |3|
|    |    |    |
|    |    |    |
|    |    |    |
|    |    |    |
|    |    |    |

With independent likelihoods of state durations given the contrast group, the spectral score for each possible duration of each state was multiplied by

\[ L(0,1,2,3; G(w)) \]

and the contrast group, so the routines always chose the first trisyllabic word of each pair. However, the average similarity score increased with the tied models, from \( \frac{1}{2} \) to \( \frac{9}{2} \).

Restricted word order did not significantly affect the recognition or similarity scores.

Conditions 4-9 of Figure 1 used expanded durations, tied models, and independent recognition options 3-5, respectively. Conditions 5-9 used recognition options 6-8, respectively.

DISCUSSION AND CONCLUSION

As Figure 1 shows, the best recognition results obtained in test words spoken at the training (faster) rate, were 486 (6443) without probabilities or likelihoods of durations or duration ratios, 876 (681) with likelihoods of durations, and 876 (782) with likelihoods of duration ratios.

Figure 1 is a plot of likelihoods of each word in the contrast groups (from top to bottom) with \( \frac{|V|}{|V|} \) and \( \frac{|V|}{|V|} \). Figure 3 is the analogous plot for ratio, the solid curves are for the models made from the training productions. The dashed curves are for models with post hoc from the 4s/pair productions. As modellled, the ratio, contrast boundaries lie along radii or along rays. Figure 6 is a scatter plot of the values of the duration of \( V_2 \) and \( C_3 \) observed while modelling the CVVC words in the training recordings. Figure 7 is the analogous plot for the 4s/pair test recording. Prior coordinates were used for these plots, i.e., the radius is ratio, and the angle is \( \frac{V_2}{C_3} \). With the expanded range of durations, the contrast groups (from top to bottom) with \( V_2 = /e/, /u/, /o/ \) or \( /a/ \). Figure 3 is the analogous plot for ratio, the solid curves are for the models made from the training productions. The dashed curves are for models with post hoc from the 4s/pair productions. As modellled, the ratio, contrast boundaries lie along radii or along rays.