AUTOMATIC RECOGNITION OF WORDS
DIFFERING IN DISTINCTIVE QUANTITY

G. Kuhn and K. Ojamaa

96 Leigh Avenue
Princeton, NJ 08540
USA

Library of Congress
Washington, DC 20540
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% (64%) without probabilities or likelihoods of durations or duration ratios, 87% (68%) with likelihoods of durations, and 85% (77%) 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: tee:de-teete and kyde-kuu:de; and 6 words participated in 2 three-way contrasts: toode-toote-too:te and kade-kate-katte.

CORPUS

Speech was recorded while one of the authors (KO) 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 ütlesid (Did you say) 'word 1, word 2, word 3'? Ei ma ütlesin (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.

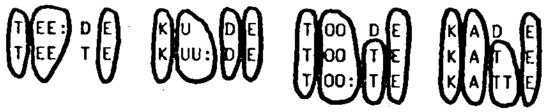
Each recording was digitized at 10000 samples/s. The digitized recordings were parameterized in centisecond frames using a 10-channel, filter-bank spectrum analyzer.

WORD MODELS

We used 15 "word" models, one each for Kas sa, ütlesid, Ei ma, ütlesin, (pause), and the 10 CVCV words. The models for ütlesid and ütlesin 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+2, 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 "miniav". The miniav 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 miniav. Means and a covariance matrix were computed over the spectra aligned to each segment of each hand-marked state of the miniav. 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:



(Here we refer to the states by name. We feel that this is justified because of the good correspondance over alignments to the miniav). The weights were the number of spectra aligned to each segment. When spectral estimates were tied, there was no spectral difference between the models in word pairs kude-kuu:de, toote-too:te, and kate-katte.

RECOGNITION

The routines for connected recognition computed a spectral match score for the best path through an entire recording [9,10]. That score was the maximum product of the likelihoods of the observed spectra, over all segments of all states of all words on the path. The likelihood of a single spectrum O_t under the continuous multivariate-gaussian probability density function (pdf) for spectral shape in segment j of state i of word w, was

$$L(O_t|j,i,w) = \frac{P(O_t|j,i,w)}{\Sigma_j \Sigma_i \Sigma_w P(O_t|j,i,w)} .$$

The recognition routines used the notion of a "contrast group". Let G(w) be the contrast group for word w, i.e., the group of words including word w that we expected to be confusable under a pure spectral match score. Kas sa, Ei ma and (pause) were each assigned to a one-word group. Ütlesid and ütlesin were assigned to a two-word group. The 10 CVCV words were assigned to four contrast groups, one for each V_1: /e/, /u/, /o/ or /a/.

The recognition options were:

1) expanded range of state durations;
2) restricted word order;
3) independent probabilities of state durations;
4) independent likelihoods of state durations given the contrast group;
5) multivariate likelihood of state durations given the contrast group;
6) independent likelihoods of a pair of state duration ratios given the contrast group;
7) independent likelihoods of a second pair of state duration ratios given the contrast group;
8) multivariate likelihood of the second pair of state duration ratios given the contrast group.

With expanded state durations, durations in the range $0.5*min_{i,w}$ through $1.5*max_{i,w}$ were permitted.

With restricted word order, Kas sa could only follow (pause); ütlesid could only follow Kas sa; Ei ma could only follow (pause), ütlesin could only follow Ei ma; while the other 10 words and (pause) could follow one another any number of times.

With independent probabilities of state durations, the spectral score for each possible state duration of each state was multiplied by $P(d_i|i,w)$. $P(d_i|i,w)$ is the probability of duration d_i in state i of word w, under a discrete binomial state duration pdf parameterized by $(min_{i,w}, average_{i,w}, max_{i,w})$.

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

$$L(d_i|i,w,G(w)) = \frac{P(d_i|i,w)}{\Sigma_{m \in G(w)} P(d_i|i,m)} .$$

With the multivariate likelihood of state durations given the contrast group, the spectral score for each word w was multiplied. by the tri-variate gaussian $L(d_{S-2},d_{S-1},d_S|w,G(W))$, where S is the number of states in word w.

With independent likelihouds of a pair of state duration ratios, the spectral score for each word w was multiplied by $\Pi_r L(ratio_r|w,G(w))$, r=1,2. The underlying duration ratio pdf's, $P(ratio_r|w)$, were discrete binomials parameterized by the (min,expected,max) values of $ratio_r$. The first pair of duration ratios tested [11] was

$$ratio_1 = d_{S-2}/(d_{S-2} + d_{S-1})$$
$$ratio_2 = d_{S-2}/(d_{S-2} + d_S) .$$

The second pair of duration ratios tested [12] was

$$ratio_1 = d_{S-2}/(d_{S-2} + d_{S-1})$$
$$ratio_3 = (d_{S-2}+d_{S-1})/(d_{S-2}+d_{S-1}+d_S) .$$

With the multivariate likelihood of the second pair of state duration ratios given the contrast group, the spectral score for each word was multiplied by the bi-variate gaussian $L(ratio_1,ratio_3|w,G(w))$.

RESULTS

Boxes are drawn on the confusion matrix in Table 1. Let the count in the boxes divided by the count in the 10 rows be a "similarity score" (these words were at least recognized as a word in the same contrast group). Then this confusion matrix shows how a recognition score of 88% and a similarity score of 99.4% was obtained when a baseline system was run on the 6s/pair test recording. The baseline system used the observed range of state durations, separate spectral models, unrestricted word order, and a path score based only on the spectral match.

Figure 1 gives recognition results in terms of recognition scores on each test recording, and average similarity score over the two test recordings. The curve of recognition scores for the 6s/pair test recording is labelled "6". The curve of recognition scores for the 4s/pair test recording is labelled "4". The curve of average similarity scores is labelled "SIM".

Under conditions 0-3 in Figure 1, the baseline system was used (condition 0), or the baseline system modified by three cumulative changes: expanded range of durations (condition 1), tied models (condition 2), and restricted word order (condition 3).

Not surprisingly, both the recognition score for the 4s/pair recording, and the average similarity score, improved with the expanded range of durations.

The recognition score for both recordings decreased with the tied models, because there was no difference between the models in word pairs kude-kuu:de, toote-too:te, and kate-katte, so the routines always chose the first listed word of each pair. However, the average similarity score increased with the tied models, from 97.2% to 98.2%.

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

Conditions 4-9 of Figure 1 used expanded durations, tied models, and restricted word order. Conditions 4-6 used recognition options 3-5, respectively. Conditions 7-9 used recognition options 6-8, respectively.

DISCUSSION AND CONCLUSION

As Figure 1 shows, the best recognition results obtained on the test words spoken at the training (faster) rate, were 88% (64%) without probabilities or likelihoods of durations or duration ratios, 87% (68%) with likelihoods of durations, and 85% (77%) with likelihoods of duration ratios.

Figure 2 is a plot of $L(ratio_1|w,G(w))$ for the CVCV contrast groups (from top to bottom) with V_1 = /e/, /u/, /o/ or /a/. Figure 3 is the analogous plot for $ratio_3$. The solid curves are for the models made from the training productions. The dashed curves are for models made post hoc from the 4s/pair productions. As modelled, the $ratio_1$ contrast between toote and too:te was neutralized at the faster rate of speech.

Figure 4 is a scatter plot of the values of $ratio_1$ and $ratio_3$ observed while modelling the CVCV words in the training recording. Figure 5 is the analogous plot for the 4s/pair test recording. Polar coordinates were used for these plots, i.e., the radius is $ratio_3$, and the angle is $ratio_1*\pi/2$. Assuming independence, quantity contrast boundaries lie along radii or along rays.

Figure 6 is a scatter plot of the values of the durations of V_1 and C_2 observed while modelling the CVCV words of the training recording. Figure 7 is the analogous plot for the 4s/pair test recording. The minimum permitted state durations were apparently somewhat long for the 4s/pair recording.

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.

REFERENCES

[1] R.K. Moore, M.J. Russell and M.J. Tomlinson, "Locally constrained dynamic programming in automatic speech recognition", Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing, 1982, pp. 1270-1273.
[2] T.H. Crystal and A.S. House, "Characterization and modelling of speech-segment durations", Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing, 1986, pp. 2791-2794.
[3] M.J. Russell and A.E. Cook, "Experimental evaluation of duration modelling techniques for automatic speech recognition", to appear in Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing, 1987.
[4] P. Ariste, "A quantitative language", Proc. Third Intl. Cong. Phonetic Sciences, 1938, pp. 276-280.
[5] G. Liiv, "On the quantity and quality of Estonian vowels of three phonological degrees of length", Proc. Fourth Intl. Cong. Phonetic Sciences, 1962, pp. 682-687.
[6] I. Lehiste, "Temporal Compensation in a quantity language", Ohio State University Working Papers in Linguistics, 12, 1972, pp. 53-67.
[7] A. Eek, "Estonian quantity: notes on the perception of duration", Estonian Papers in Phonetics, 1979, pp. 5-29.
[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimisation for spoken word recognition", IEEE Trans. ASSP-26, February 1978, pp. 43-49.
[9] T.K. Vintsyuk, "Element-wise recognition of continuous speech consisting of words of a given vocabulary", Kibernetika, 7, 1971, pp. 133-143.
[10] J.S. Bridle, M.D. Brown and R.M. Chamberlain, "Continuous connected word reconition using whole word templates", Radio & Electronic Engineer, 53, 1983, pp. 167-173.
[11] U. Lippus, "Prosody analysis and speech recognition strategies: some implications concerning Estonian", Estonian Papers in Phonetics, 1978, pp. 56-62.
[12] K. Ojamaa, "Temporal aspects of phonological quantity in Estonian", Ph.D. Thesis, Univ. of Connecticut, 1976.

	0	1	2	3	4	5	6	7	8	9
0 TEE:DE	36									
1 TEETE		36								
2 KUDE			36							
3 KUU:DE				36						
4 TOODE.					36					
5 TOOTE					2	17	17			
6 TOO:TE						11	25			
7 KADE						1		35		
8 KATE						1		1	25	9
9 KATTE										36

Table 1. Confusion matrix obtained when a baseline system was run on the 6s/pair test recording.

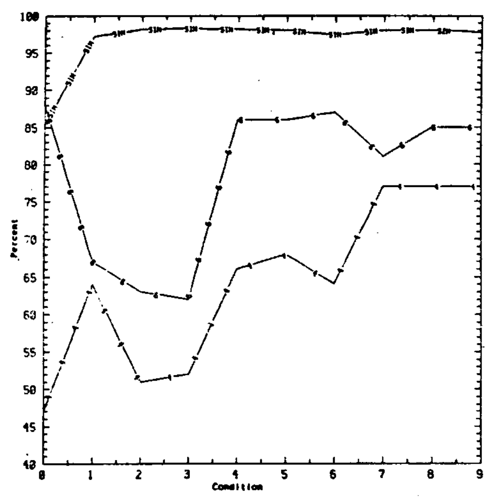


Figure 1. Recognition and similarity scores as a function of experimental condition.

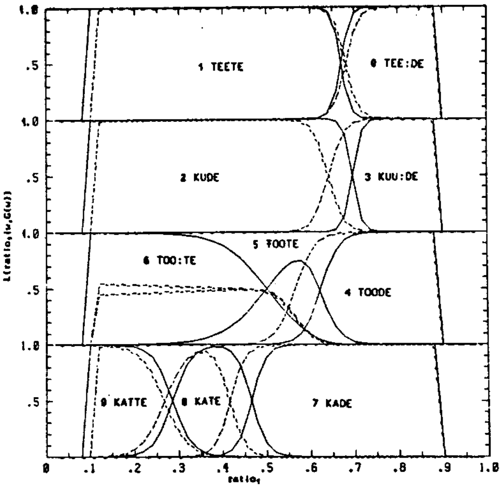


Figure 2. $L(ratio_1|w,G(w))$ for $w=6,9$ and training (solid line) or 4s/pair (dashed) models.

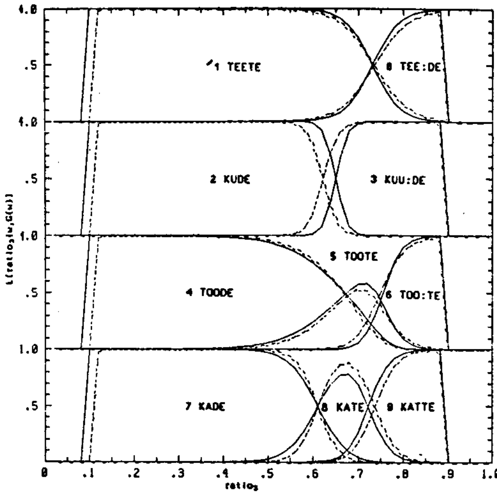


Figure 3. $L(ratio_3|w,G(w))$ for $w=6,9$ and training (solid line) or 4s/pair (dashed) models.

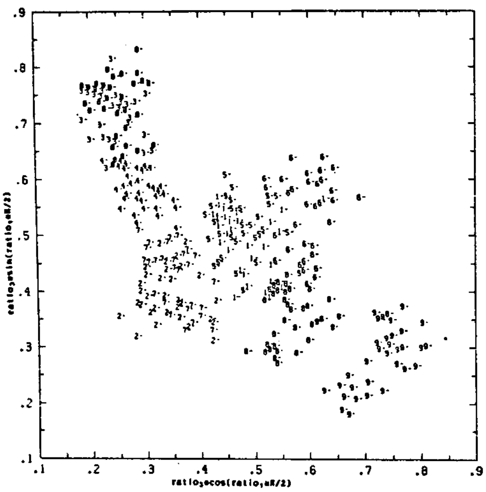


Figure 4. Values of $ratio_1$ and $ratio_3$ observed while modelling the CVCV words in the training recording. Radius=$ratio_3$. Angle=$ratio_1*\pi/2$.

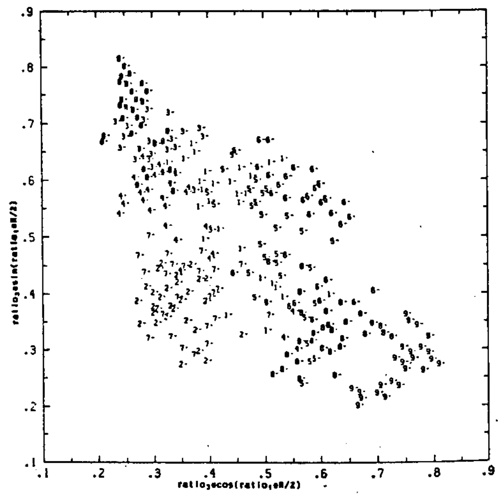


Figure 5. Values of $ratio_1$ and $ratio_3$ observed while modelling the CVCV words in the 4s/pair recording. Radius=$ratio_3$. Angle=$ratio_1*\pi/2$.

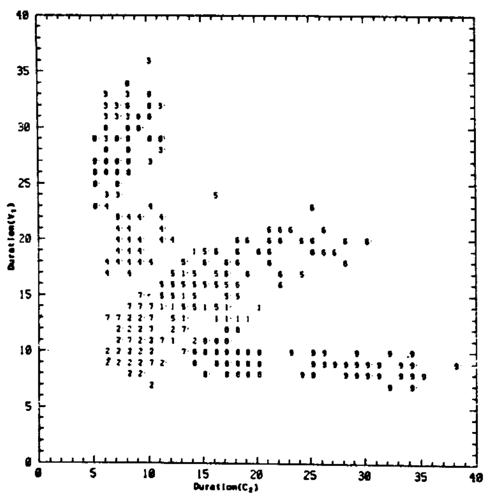


Figure 6. Durations of V_1 and C_2 observed while modelling the productions of the CVCV words in the training recording.

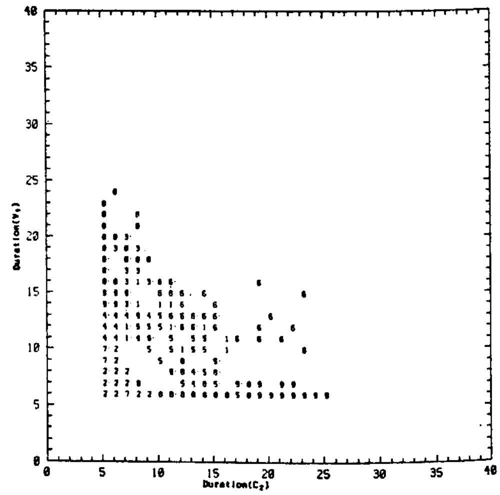


Figure 7. Durations of V_1 and C_2 observed while modelling the productions of the CVCV words in the 4s/pair recording.