ON THE INDEPENDENCE OF DIGITS IN CONNECTED DIGIT STRINGS

J.W. Koolwaaij *L. Boves*

Department of Language and Speech, Nijmegen University
P.O. Box 9103, 6500 HD Nijmegen, the Netherlands
E-mail: koolwaaij,boves@let.kun.nl

ABSTRACT

One of the frequently used assumptions in Speaker Verification is that two speech segments (phonemes, subwords, words) are considered to be independent. And therefore, the log-likelihood of a test utterance is just the sum of the log-likelihoods of the speech segments in that utterance. This paper reports about cases in which this observation-independence assumption seems to be violated, namely for those test utterances which call a certain speech model more than once. For example, a pin code which contains a non-unique digit set performs worse in verification than a pin code which consists of four different digits. Results illustrate that violating the independence assumption too much might result in increasing EERs while more information (in form of digits) is added to the test utterance.

1. INTRODUCTION

In Speaker Verification (SV) systems using passwords in the form of fixed or prompted digit strings it seems usual to compute the log-likelihood of a claimant being the true speaker for individual digits, and the total score is obtained as some simple combination of the digit log-likelihoods. In doing so, it is usually assumed that the individual digits are independent, and that each digit contains essentially the same amount of relevant information. In our experiments with the scope card numbers in the SESP database this basic assumption is called into question, because the 14 digit numbers by necessity repeat some digits. Moreover, because the scope card numbers in SESP adhere to the ISO standard for card numbers, and because the numbers were issued as 'real' company card numbers (i.e., with a fixed prefix even beyond the first six digits that are determined by the ISO standard) the number of non-unique digits is relatively large. This made it natural to investigate the indepenence assumption for our experiments. It is shown that the assumption does not hold, and alternative ways for combining the evidence form non-unique utterances are suggested.

2. THE SESP CORPUS

The SESP corpus is developed to support research into speaker verification over the telephone network. It contains calls made by 24 males and 22 females, each of whom called 25 times with different handsets (including some calls from mobile phones), from a wide variety of places (in restaurants, public phones on airports, etc., in addition to home, office and quiet hotel rooms). In each call subjects produced their own 14 digit scope number and the attendant 4 digit PIN code twice. In addition, they produced the scope number and PIN of a different same sex fellow subject to provide impostor utterances. For the experiments reported in this paper only utterances consisting of sequences of 14 digits (also called *words*) were used. (A word is one of the Dutch digits /nul(0)/, /een(1)/, /twee(2)/, ..., /negen(9)/.) Four sessions from quiet environments were set apart for enrolment. For the experiments in this paper both productions of their own scope number were used, making for a total of eight enrolment utterances for each client.

3. PERFORMANCE FOR DIFFERENT TEST UTTERANCE LENGTHS

Speaker models in this study are HMM's; separate client models are trained for all digits that occur in this client's scope card number. The HMM topology used is left-to-right HMM, with 2 states per phoneme, 3 mixtures per state and diagonal covariance matrix. Acoustic features are 12 liftered zero-mean cepstra (LPC based) together with the log energy and their delta's and delta-delta's. In addition to the client models there is a single set of sex independent world models, one for each of the ten digits. Finally, there is a silence model (or maybe better: non-speech model), that is shared by all clients and the world.
Scoring is based on the sum of the log-likelihoods obtained for the individual words in each test utterance. This enables us to compute client scores based on any subset of the words in a test utterance. Test results are summarised in terms of dynamic Same Sex Equal Error Rate (SS EER): for each experiment the accept/reject threshold is computed that yields the same proportion of false accepts and false rejects.

To begin, we ran an experiment with different test utterance lengths by using only the first n words of the scope numbers. Due to their adherence to the ISO standard and their origin from a set of company card numbers, the first 10 words are always /8931002042/, the 11th word is almost always /4/, only words 12-14 are more or less uniformly distributed over the digit space. Words with a variable digit content are indicated by /x/. Table 1 shows the SS EERs for both the first n and the last n words. It can be seen that the EER does not decrease monotonically with the numbers of words used for verification (cf. the words

Table 1. Same Sex Equal Error Rates (SS EER) as a function of the number of words used for verification. W(n-m) implies that words n-m are used for verification. D(n) is the digit content of word n.

W(1-n)	SS EER	W(n-14)	SS EER	$D(n)$
1-1	10.47	1-14	0.49	/8/
1-2	3.96	2-14	0.83	/9/
1-3	1.81	3-14	1.16	/3/
1-4	1.33	4-14	1.31	/1/
1-5	0.89	5-14	1.45 ⇐	/0/
1-6	1.43 ⇐	6-14	1.34	/0/
1-7	1.27	7-14	1.52 ⇐	/2/
1-8	1.30 ⇐	8-14	1.40	/0/
1-9	1.15	9-14	1.75	/4/
1-10	1.04	10-14	2.04	/2/
1-11	1.10 ⇐	11-14	2.08	/4/
1-12	0.76	12-14	3.28	/x/
1-13	0.65	13-14	6.20	/x/
1-14	0.49	14-14	10.22	/x/

Table 2. SS EER for a single occurrence of the digit /0/, and for /0/ combined with one additional digit.

Test utterance is ...	the EER is
/0/	11.46
When a ... is added	the EER becomes
/8/	5.50
/9/	3.40
/3/	4.32
/1/	5.03
/0/	9.48
/0/	9.35
/2/	5.42
/4/	4.98
/2/	4.98
/4/	5.14
/x/	4.31
/x/	6.25
/x/	4.90

marked with a ⇐ in Table 1). All increases of EER happen when the added word is a digit which was already present in the string.

3.1. EERs for words and word pairs

From the raw data in Table 1 the cause of the discontinuities in the EER as function of the number of words is not evident. To get more insight into the verification power of pairs of words, we performed all possible verifications with a test utterance length of 1 or 2 (so in total 105 experiments). This experiment has two goals: to establish whether some words perform better than others, and to investigate how performance improves if a specific word is added to a given word. Table 2 shows the results for the digit /0/ in combination with other digits: When only a single /0/ (actually, the third occurrence in the scope numbers) is used as test utterance, the SS EER is 11.46%. Depending on the word that is added, SS EER drops to a value between 9.48% and 3.40%. The best score

Table 3. SS EER for individual words, and for the best c.q. worst performing additional word.

$D(n)$	SS EER	Best		Worst	
/8/	10.47	3.23	/3/	6.14	/0/
/9/	7.45	2.64	/3/	4.02	/x/
/3/	8.68	2.64	/9/	5.44	/x/
/1/	9.88	3.82	/9/	6.34	/0/
/0/	11.70	3.39	/9/	10.24	/0/
/0/	14.01	3.37	/9/	10.24	/0/
/2/	9.67	3.45	/9/	6.71	/2/
/0/	11.46	3.40	/9/	9.48	/0/
/4/	12.03	2.78	/9/	8.45	/4/
/2/	9.14	3.31	/9/	6.71	/2/
/1/,/4/	10.02	3.14	/9/	8.45	/4/
/3/-/9/	11.88	4.02	/9/	5.61	/x/
/0/-/9/	12.85	3.11	/9/	6.68	/0/
/0/-/9/	10.22	2.99	/9/	6.54	/4/

is obtained by adding /9/, the worst by adding another /0/. These experiments are done for all digits in the scope number. Table 3 summarises the results; it shows the SS EERs for the individual words, and the highest/lowest scoring when another word is added.

From Table 3 several observations can be made. First it is obvious that there is a large range for the performance of individual words, with the word /9/ ([n e X @ (n)] in Sampa transcription) showing the best performance while the word /0/ ([n Y l]) is on the opposite end of the range. Several explanations can be put forward for the rank of the word /9/: it contains up to two occurrences of the phoneme /n/ (cf. [1]), it contains the phoneme /X/ which is known to vary a lot between dialects, and it is the word with the longest duration (411 ms on average, compared to an overall average of 274 ms for the other nine digits; so it simply contains most information in the form of frames). The word final /n/ in /9/ is optional; it will not occur in the speech of most speakers of standard Dutch, unless perhaps the word /9/ is followed by a word with a word initial vowel (in the digit vocabulary the words /1/ and /8/). If the word final /n/ is realised in other positions, that is certainly characteristic for the speaker in question. However, not too much emphasis should be placed on the speaker specificity of the phoneme /n/, since it also occurs in the word /0/.
Also, and compatible with the finding that /9/ is the best scoring individual word, adding /9/ to any other word is the best one can do. Last but not least, Table 3 shows that repeating the same digits yields the lowest possible improvement.

One explanations for the small improvement with repeated digits is that summing unweighted log-likelihoods over the words in an utterance is just too simple if this utterance contains the same digit more than once, because then the independence assumption underlying the additive combination is violated too much. In order to explore this hypothesis we first of all studied the contribution that an added word makes depending on the correlation between log-likelihoods of individual words. Figure 1 displays the gain factor as

Figure 1. Performance improvement for word pairs as function of the correlation between the performance of the individual members of the pair.

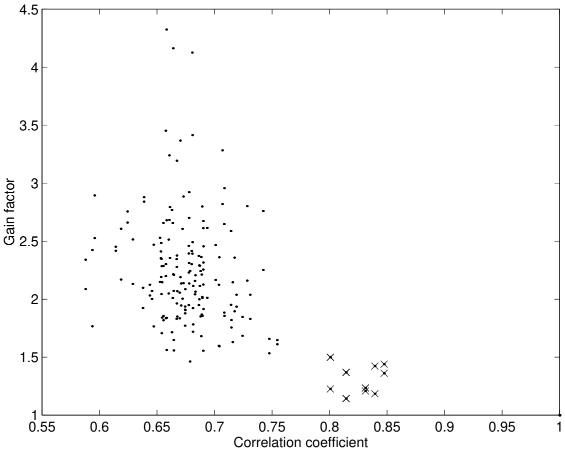
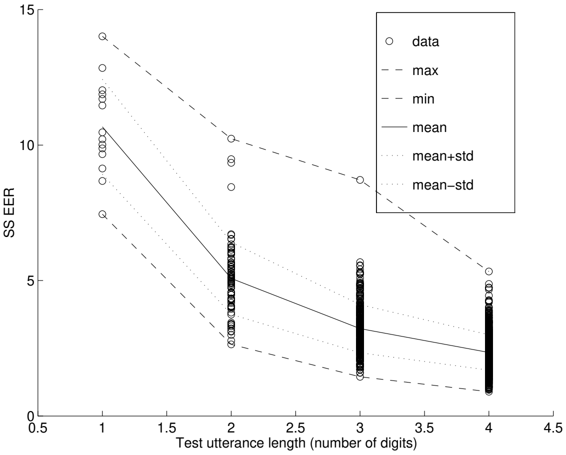


Figure 2. SS EER for test utterances comprising 1, 2, 3, and 4 words.



function of the correlation between the log-likelihoods of those words. The gain factor is defined as the quotient of the EER when only the first word is used as test material and the EER with both words. And it can be seen that the digit pairs with high correlation have the lowest gain factor. In fact, the data points marked with $\times$ in the lower right corner of the plot are the pairs /00/ (6$\times$), /22/ (2$\times$) and /44/ (2$\times$).

3.2. EER for utterances of 1,…,4 words

The next question that arises is how the non-uniqueness of the digits in the test utterance influences the eventual EER. Therefore we used all possible combinations of the digits of the scope numbers to form a test utterance with 1, 2, 3 or 4 digits. For each card number utterance a large number of test utterances were constructed, viz. 14 single digit test utterances, 91 ($= 14 \times 13/2$) test utterances of 2 digits, 361 test utterances of 3 digits and 1001 test utterances of 4 digits. Fig. 2 displays a summary of the results. It can be seen that the difference between the best and worst scoring test utterances is quite large. The worst scoring utterances, connected by the dashed line in Fig. 2 happen to be test utterances with a non unique digit set. For example the data point with test_utterance_length=3 and SS_EER=8.71 is the test utterance /000/. On the other hand, the test utterances with lowest EER always contain a /9/.

Table 4 summarises the data in Fig. 2 in a different way. It shows the best and worst combinations in the test sets of length $1, \ldots, 4$ (x means that this word can be any digit /0, …, 9/).

Table 4. SS EER for utterances of $1, \ldots, 4$ words, separately for male and female speakers.

		EER		
	Test utt.	SS	MM	FF
1	/9/	7.45	11.30	3.60
	/3/	8.68	7.92	9.43
	/2/	9.14	8.83	9.44
	…	…	…	…
	/4/	12.03	10.43	13.62
	/x/	12.85	13.02	12.67
	/0/	14.01	15.84	12.19
2	/93/	2.64	3.26	2.03
	/94/	2.78	3.45	2.11
	/9x/	2.99	4.20	1.79
	…	…	…	…
	/00/	9.35	11.96	6.74
	/00/	9.48	12.34	6.62
	/00/	10.24	13.1‘	7.37
3	/90x/	1.45	2.08	0.82
	/93x/	1.60	2.08	1.11
	/93x/	1.60	2.20	1.01
	…	…	…	…
	/00x/	5.57	6.79	4.34
	/00x/	5.68	7.30	4.06
	/000/	8.71	11.39	6.04
4	/94xx/	0.90	0.88	0.92
	/93xx/	0.96	1.18	0.75
	/904x/	0.99	1.59	0.38
	…	…	…	…
	/1000/	4.74	6.07	3.41
	/0020/	4.87	5.20	4.54
	/000x/	5.34	6.74	3.93

3.3. Improving EER by leaving out information

To further check the hypothesis that simply adding log-likelihoods for individual words is an too easy way, an experiment was done in which SS EERs were computed for the full 14 digit card numbers and for a reduced 13 digit number, i.e., the original sequences with the second occurrence of the digit /2/ left out. This choice was made because the pair /22/ has the highest correlation (even if it is not the digit pair with the lowest combined EER). EER for the full 14 digit sequence is 0.45%; for the reduced 13 digit sequence EER drops to 0.29%! So with one digit less test material we have about 41% improvement in EER!

For the time being the performance improvement obtained by ignoring part of the information in the test utterances is difficult to understand. The most likely explanation invokes the assumption that the position

Figure 3. SS EER as function of the duration of the test utterances. Encircled data points refer to utterances with repeated digits.

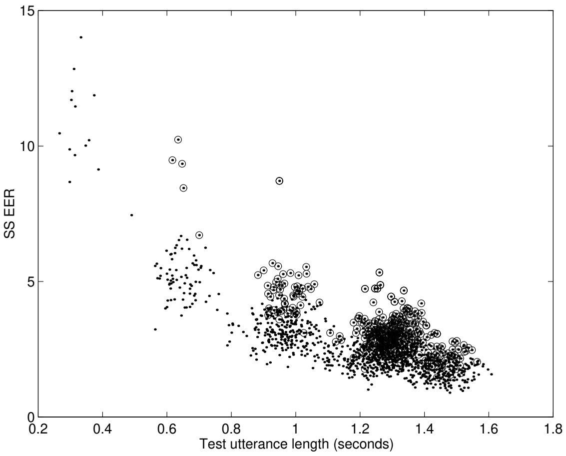
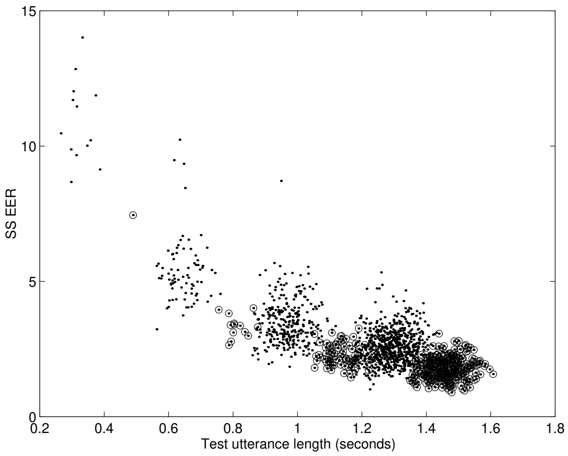


Figure 4. EER as a function of utterance duration. Now utterances containing the digit /9/ are encircled.



of the word in the prosodic structure of the utterances induced so much within speaker variation that it is better left out, especially since what information is contributed by the sounds in the word /2/ is already discounted for by the other occurrence of the word.

3.4. Utterance length

It has been said before that the bi-syllabic word /9/ may outperform the other words simply by the fact that it is longer. In more general terms, there might well be a strong correlation between utterance duration and SS EER. To check this hypothesis SS EER is plotted as a function of the duration of the $1, \ldots, 4$ word test utterances in Fig. 3.

It is obvious that there is a strong correlation between utterance length and SS EER. However, the test utterance with a double digit (the encircled data points in Fig. 3) appear to perform worse compared to test utterances with the same duration but without double digits (the normal data points).

Fig. 4 contains the same data as fig. 3, only now the utterances containing the word /9/ are encircled. It can be seen that the /9/ in a test utterance has two impacts: they tend to have the longest average duration in their class and the lowest EER.

3.5. Predicting SS EER from utterance duration

With the aid of the 1470 data points per speaker obtained from $1, \ldots, 4$ word test utterances we estimated the relation between test utterance length in seconds and EER. Estimation is done using the method of least squares.

$$\text{SS EER} = 3.14 \cdot t^{-1.15} \tag{1}$$

This formula is also quite good when it is used for extrapolations: a 14 digit test utterance is about 4.7 seconds long and the EER is estimated at 0.52%, and in reality this is 0.49%.

4. DISCUSSION

The fact that test utterances with a non unique digit set, for example /1000/, perform worse than test utterances with a unique digit set, seems to give some ground to assume that the indepence assumption underlying the summation of log-likelihoods for individual words is violated. The observation-independence assumption, which is commonly used, states that the acoustic vectors are not correlated or, in other words, that the probability that a particular acoustic vector will be emitted at time t depends only at the transition taken at that time, and is conditionally independent of the past (cf. [2]). In terms of equations the independence assumption can be expressed by

$$\frac{\mathcal{L}\left(s_1 \oplus s_2 \mid X\right)}{\mathcal{L}\left(s_1 \oplus s_2 \mid \Omega\right)} \tag{2}$$

$$= \frac{\mathcal{L}\left(s_1 \,|X_1\right) \mathcal{L}\left(s_2 \,|s_1, X_2\right)}{\mathcal{L}\left(s_1 \,|\Omega_1\right) \mathcal{L}\left(s_2 \,|s_1, \Omega_2\right)} \tag{3}$$

$$\approx \frac{\mathcal{L}\left(s_1 \,|X_1\right) \mathcal{L}\left(s_2 \,|X_2\right)}{\mathcal{L}\left(s_1 \,|\Omega_1\right) \mathcal{L}\left(s_2 \,|\Omega_2\right)}. \tag{4}$$

(s is the speech, X are the client models, and Ω are the world models.) And because s_1 and s_2 are assumed to be independent, it is valid to state that $\mathcal{L}\left(s_2 \,|s_1, X_2\right)$ is very closely approximated by $\mathcal{L}\left(s_2 \,|X_2\right)$. And in practice we are able to work with this assumption. At least, as long as s_1 and s_2 are different digits. Otherwise the only possible statement can be that

$$1 \leq \frac{\mathcal{L}\left(s_2 \,|s_1, X_2\right)}{\mathcal{L}\left(s_2 \,|s_1, \Omega_2\right)} \leq \frac{\mathcal{L}\left(s_2 \,|X_2\right)}{\mathcal{L}\left(s_2 \,|\Omega_2\right)} \tag{5}$$

And since this quotient is indeterminable, its contribution to the likelihood ratio is also indeterminable, so in other words, noisy information is added to the SV system, which eventually results in worse EERs.

REFERENCES

[1] H. van den Heuvel, *Speaker variability in acoustic properties of Dutch phoneme realisations*, Ph.D.Thesis, Nijmegen, 1996

[2] H.A. Bourlard, N. Morgan, *Connectionist speech recognition, a hybrid approach*, Kluwer, Boston, 1994