A NEW PROCEDURE FOR CLASSIFYING SPEAKERS IN SPEAKER VERIFICATION SYSTEMS

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

In this paper we propose a new measure to classify speakers with respect to their behaviour in speaker recognition systems. Taking the proposal made by EAGLES as a point of departure we show that it fails to yield results that are consistent between closely related speaker recognition methods and between different amounts of speech available for the recognition task. We show that measures based on a straightforward confusion matrices, that take only the 1-best classification into account, cannot result in consistent classifications. As an alternative we propose a measure based on n-best scores in a speaker identification paradigm, and show that it yields more consistent performance.

1. CONTEXT

In Speaker Verification (SV) research it has been customary to base comparisons and evaluations on Equal Error Rates (EER) [2]. Recently, Receiver Operator Curves (ROC) seem to gain popularity, as a richer representation of performance results. However, ROC representations are still based on relatively straightforward binary accept/reject decisions.
Despite their great value for system evaluation purposes, EERs and ROCs are less valuable when diagnostic information on the behaviour of a system is needed for improving performance. Also, EER nor ROC give information on the contribution to the overall performance of individual subjects in a test corpus. However, for effective deployment of SV systems insight into the behaviour for individual customers is needed, if only in order to be able to predict whether a prospective customer will turn out to be a so called 'goat'. Since it is quite possible that someone's 'goatiness' may depend on the SV method employed, we have compared several such methods. In this paper we can only present data from two systems.
In this paper we propose new measures to characterise the performance of SV systems. Instead of producing a single number, we develop measures that allow one to see to what extent the performance results are determined by specific subjects in the test corpus.

This research is based on the YOHO Speaker Verification Corpus [1]. The YOHO vocabulary consists of two-digit numbers ("thirty-four", "sixty-one", etc), spoken in sets of three (e.g. "36-45-89"). There are 138 speakers (106 male, 32 female); for each speaker, there are 4 enrolment sessions of 24 utterances each, and 10 verification sessions of 4 utterances each, for a total of 136 utterances in 14 sessions per speaker.

2. METHOD DEFINITION

We have compared two SV systems, both based on left-to-right HMMs [3]. The only difference between the systems is the number of Gaussians in each subword model (1 *vs.* 5). We have used 1 state per phoneme in each *subword*. Diagonal covariance matrices were obtained from 96 training utterances per speaker. In YOHO the set of *subwords* comprises {one, ..., seven, nine, ty, Twen, ..., Nine} [1]. We also will use the subset of *tens* {Twen, ..., Nine} and the subset of *units* {one,...,seven, nine} (cf. table 2). An *utterance* is defined as a sequence of three numbers.
We applied a preemphasis with factor 0.97 and used a Hamming window (length 25.6 ms, step 10.0 ms) to calculate 13 Mel-frequency zero-mean cepstral coefficients $c_0, \ldots, c_{12}$ and their first and second time derivative, yielding 39-dimensional feature vectors.

3. THE ENROLMENT AND IDENTIFICATION SETS

We used all available enrolment speech of 118 speakers (96 male, 22 female); the other 20 speakers in YOHO were used to train a world model. On average we have $4 \times 24 \times 3/8 = 36$ occurrences of each *ten*, 36 occurrences of each *unit* and $96 \times 3 = 288$ *ty*'s per speaker.

YOHO has 10 verification sessions per speaker with 4 utterances each. Our identification set was defined as all the speech of two of those sessions per speaker, randomly chosen out of the ten available sessions. Thus with 8 utterances, the test material consisted of 24 words and 48 subwords per speaker. Speakers $1, \ldots, 22$ are females and speakers $23, \ldots, 118$ are males.

4. SCORING PROCEDURE

For each speech unit the log-likelihood ratio (LLR) that this unit has been uttered by the claimant speaker is computed. Because LLR's for utterances yielded virtually 100% correct identification, we decided to compute LLR's for subwords and words too. LLR's for words and utterances were obtained by summing the LLR's for the subwords making up the larger units (for the definition of the units, cf. section 2.).

Table 1. Identification results per method

Method		$\bar{\gamma}_F$	$\bar{\gamma}_M$	$\bar{\gamma}_{MF}$
5 Mixs	Subw.	13.73	16.25	14.99
per	Word	3.60	2.52	3.06
state	Utt.	1.14	0.13	0.63
1 Mix	Subw.	39.49	33.27	36.38
per	Word	10.80	6.94	8.87
state	Utt.	0.00	0.52	0.26

Method		$\hat{\gamma}_F$	$\hat{\gamma}_M$	$\hat{\gamma}_{MF}$	γ
5 Mixs	Subw.	10.69	16.34	13.52	15.78
per	Word	0.97	2.92	1.95	2.72
state	Utt.	0.00	0.35	0.17	0.32
1 Mix	Subw.	27.92	34.04	30.98	34.43
per	Word	6.04	7.49	6.77	7.66
state	Utt.	0.00	0.44	0.22	0.42

Table 2. Misclassification rates per subword ($\times$100%) for 1 and 5 mixtures

	Mix/State			Mix/State	
Units	1	5	Tens	1	5
one	0.18	0.10			
two	0.61	0.19	Twen	0.44	0.13
three	0.41	0.12	Thir	0.25	0.14
four	0.34	0.15	Four	0.29	0.14
five	0.27	0.15	Fif	0.38	0.17
six	0.55	0.12	Six	0.31	0.15
seven	0.11	0.13	Seven	0.11	0.19
			Eigh	0.72	0.42
nine	0.20	0.07	Nine	0.38	0.16

5. CLOSED SET IDENTIFICATION RESULTS

According to the EAGLES document [2], the most natural measure for the performance of a SI system is the relative number of times the system fails to identify an applicant speaker correctly. This so called *misclassification rate* γ was computed and averaged over the speakers, both gender dependent ($\bar{\gamma}_F, \bar{\gamma}_M$) and gender independent ($\bar{\gamma}_{MF}$); the same was done for the *mistrust rate* ($\hat{\gamma}$), i.e., the relative number of times the system falsely assigns an attempt to a registered speaker. Because the misclassification rate on utterance level is extremely small, we are also interested in the misclassification rates on word and subword level. Table 1 shows the results of identification on those three levels. It can be seen that even at the word level misclassification rate is very small. On utterance level γ roughly corresponds with the SI EERs reported in [3] (0.109% with 5 and 0.666% with 1 mixture/state).

To provide insight in the behaviour of individual subwords, the misclassification rates per subword are listed in table 2. It can be seen that *Eigh* causes most trouble, followed by *two*. These units are both very short: *Eigh* consists of a single phoneme, and *two* of only two.

Males perform better than females with low complexity models, whereas the situation is reversed with more complex models. This corroborates the results in [3].

6. ANIMAL FARM

In the literature on speaker recognition classes of subjects are often given animal names; unfortunately, the SV literature is inhabited with different animals than the literature on Speaker Identification (SI). For instance, in SI one has

Goat	Unreliable applicant speaker (with high misclassification rate)
Sheep	Dependable applicant speaker (with low misclassification rate)
Lamb	Vulnerable registered speaker (with high mistrust rate)
Ram	Resistant registered speaker (with low mistrust rate)

The literature on SV is inhabited by

Goat	high false reject rate
Sheep	low false reject rate
Lamb	high false accept rate on claimed id (easy to impersonate)
Ram	low false accept rate on claimed id (difficult to impersonate)
Wolf	high false accept rate on true id (successful impersonator)
Badger	low false accept rate on true id

A wolf (SV) can be identical with a goat (SI), since persons with a high misclassification rate in SI are probably able to intrude into a SV system under the guise of one of the persons for whom (s)he was mistaken in a SI experiment. (From now on we will use *client* for a registered speaker and *speaker* for an applicant speaker.)

For an operational SV system it is important to be able to predict that a new customer will turn out a goat or a lamb, because such persons run an increased risk of finding themselves in trouble; the goat due to too many false rejects, the lamb due to too many successful break in attempts into her/his account. Therefore, we need effective techniques to classify speakers, preferably such that we can understand why a speaker is put into a given class. In general it is dangerous to rely on classifications derived from a single SV method, because one may be confounding characteristics of the speaker with idiosyncrasies of the method. Therefore, in order to answer the question "Is there a reliable measure to classify speakers as resistant, dependable, unreliable or vulnerable?" we will do the classification with different methods and compare the results to see if they are method independent.

In the experiments with the two SI systems under investigation the conventional confusion matrices on utterance level resulted in only 3 misclassifications. That is obviously too little data to base any classification on. We can increase the number of confusions by reducing the amount of speech the classification is

Table 3. Dividing clients into categories (1-best)

	1 Mix/state	
	Subword	Word
Goat	71 19 14	71 97 19
Sheep	42 100	-
Ram	21 36 4	-
Lamb	56 89	11 40

	5 Mixs/state	
	Subword	Word
Goat	71 35 97 11	71 11
Sheep	60	-
Ram	72 3	-
Lamb	40	66

based on. Table 3 shows the results for the classification on the basis of words and subwords. It is evident that the differences between the classifications obtained from the two SV systems (as well as from the two sets of text material) are very large. But the problem is more fundamental than that it could be solved by taking ever shorter stretches of speech to increase the number of confusions. In a conventional confusion matrix only data on the first best client (i.e., the one with highest log likelihood ratio) is used; all data relating to the distance of other speakers to the one with the highest LLR is discarded. Consequently, there are many interesting questions which cannot be answered by simply looking at the confusion matrix. Examples of such questions are: How many clients were almost as likely to be the speaker as the client who is eventually chosen by the system? What is the distance of the first best and the second best speaker? Which speakers can easily imitate a particular client? (In this case 1-best do not really suffice, because, for example, a client always being second best for each speaker is more a lamb than a client being first best for few speakers.) Which clients are very typical? Which male sounds most like a female and vice versa?

In order to answer these questions we need to perform SI experiments, using a scoring measure which takes into account the 1-best till the T-best identity, where T varies depending on the application. (We will choose T equal to the number of clients.) Therefore we defined a ranking function $R(s,c,t)$:

> The number of times that client c reaches the t^{th} rank in the attempts of speaker s.

The goat-sheep-curve GS is then defined as

$$GS(s,t) = \sum_{i=1}^{t} \frac{R(s,s,i)}{n_u(s)} \qquad (1)$$

for $t = 1, \ldots, n_c (= T)$. $GS(s,t)$ is very similar to the definition of confidence intervals in EAGLES, except that it is normalised such that $GS(n_s, n_c) = 1$. And the ram-lamb-curve RL is defined as

$$RL(s,t) = \frac{1}{(n_c - 1)} \sum_{i=1}^{t} \sum_{c \neq s} \frac{R(s,c,i)}{n_u(s)} \qquad (2)$$

Figure 1. GS and RL curves on subword level for the system with 5 Mixs/state

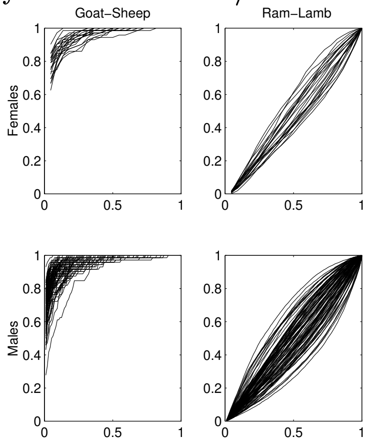
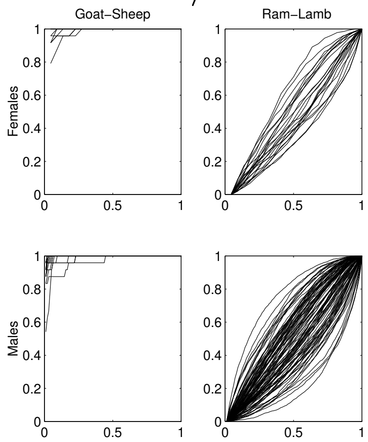


Figure 2. GS and RL curves on word level for the system with 5 Mixs/state



for $t = 1, \ldots, n_c (= T)$.

In equations (1,2) n_c is the number of clients, n_s is the number of speakers and $n_u(s)$ is the number of attempt units of speaker s. To clarify these definitions suppose that we make 8 attempts to identify speaker #1 and all 8 times she is recognised correctly (she is dependable). Moreover, when we make one identification attempt for all other 117 clients in the database, speaker #1 always turns out to be in the lowest rank: 118 (she is resistant too). Then $R(1,1,1) = n_u(1) = 8$, $R(1,c,118) = n_u(c) = 1$ and so $GS(1,t) = 1, t = 1, \ldots, n_c$ and $RL(1,t) = \delta(t, 118)$. To determine which clients are lambs we could have chosen the speakers who obtain scores $RL > \alpha$, like is done with the confidence intervals. However, because we want to give an overall performance measure, we prefer an integration over the ranking t. So the goat-sheep-score and the ram-lamb-score can then be defined as

$$S_{GS}(s) = \frac{1}{n_s} \sum_{t=1}^{n_s} GS(s,t)$$

$$S_{RL}(s) = \frac{1}{n_s} \sum_{t=1}^{n_s} RL(s,t)$$

Table 4. Dividing clients into categories (N-best)

	1 Mix/state	
	Subword	Word
Goats	71 6 65	71
Sheep	28 24 107 42	-
Ram	72 52	52 72 56 89 64
Lamb	80 63	80 66

	5 Mixs/state	
	Subword	Word
Goats	71 35 6 79	35 97 71 11
Sheep	24	-
Ram	72 64 89 28	89 64 72 28
Lamb	66 43 40	66 43 81

These scores are normalised between 0 and 1. We can now review the definition of goat, sheep, ram and lamb (although we are aware of the fact that is sounds more legitimate to base the definition of goat and sheep on the first best results only, but we also want to know if the results of those two selection methods correspond and which one is the most consistent.)

Goat	Unreliable speaker (with low goat-sheep-score)
Sheep	Dependable speaker (with high goat-sheep-score)
Ram	Resistant client (with low ram-lamb-score)
Lamb	Vulnerable client (with high ram-lamb-score)

In table 4 the classifications obtained with the two SV systems are shown. Speakers who are not mentioned are modal by definition. We plotted the RL curves and the GS curves for the SV system with 5 Mixs/state operating on subwords and words in figures 1 and 2. These figures show that a better performing system (the one based on words) has GS curves moving into the upper left corner and RL curves becoming more convex. Taking into account the confidence intervals of these results, which are not depicted in this paper, we can say that speaker 71 is a goat. And it is nice to see that on subword level he is by far the most unreliable speaker. There are no real sheep, because the system performs very well overall, and there is no real ram but there is quite a set of resistant speakers. The most obvious lamb is client 66.

The advantage of the new classification method compared to the confusion matrix method is that, because it uses the overall performance of speakers and clients and not only their champion performance, it gives a more consistent classification, both with respect to the two SV systems and the amount of speech available for testing. The consistency level can be expressed by means of correlation coefficients: the correlation coefficient between the results with subwords and words increase from 0.47 for 1-best GS-classification to 0.55 for n-best GS-classification and from 0.18 for 1-best RL-classification to 0.99 for n-best RL-classification.

Finally it is important to know how this knowledge can be used in real life applications. Suppose a new client is enrolled for such a real life application. Then the person responsible for the maintenance of that application, should want to know the classification of this client, to get a be able to predict his/her future performance. Now one can calculate the GS and the RL score of this new client and compare them with the scores of all registered speakers. If he/she appears to be a sheep or a ram everything is fine and one need not expect severe problems with this person. But when he/she is a lamb, one should be careful with attempts assigned to that particular client and maybe even add additional security measures. If the new client turns out to be a goat, one should consider giving this new client access to the application via a human operator and not via the automatic SV/SI system.

7. CONCLUSION

From this paper it becomes clear how difficult it is to make a method independent classification of speakers. We have made a new proposal for the goat, sheep, ram, lamb classification, which is more consistent than the existing definition when the classification is based on different length speech segments. Also, our newly proposed classifier appears to outperform the classifier proposed by EAGLES, at least in terms of consistency between SV methods and between tests with different amounts of speech in the test samples.
Additional research into the classification or characterisation of speakers is necessary, since it appears to be very difficult to construct a classifier which is totally method independent. Yet, in the YOHO database speaker 71 stands out as especially unreliable, while speaker 66 is the most vulnerable.
Awaiting the advent of a truly method independent classifier, for each different SV/SI method a new classification has to be done, because the results show that within each method the different classifiers have a much higher degree of agreement. Further we showed that especially for the ram-lamb-classification 1-best is not enough. Therefore we propose to use a n-best classifier.

REFERENCES

[1] Higgins A., Bahler L. and Porter J. (1991) *Speaker Verification Using Randomized Phrase Prompting*, Digital Signal Processing, **1**, pp. 89-106.

[2] European Advisory Groups on Language Engineering Standards (1995), Spoken Language Working Group.

[3] Bimbot F., Hutter H.-P, Jaboulet C, Koolwaaij J., Lindberg J., Pierrot J.-B. (1997), *Speaker verification in the telephone network: An overview of the technical development activities in the CAVE project.* Proceedings EUROSPEECH-97.