# S13.7

## NORMALIZATIONS AND SELECTION OF SPEECH SEGMENTS FOR SPEAKER RECOGNITION SCORING

*Kung-Pu Li and Jack E. Porter*

ITT Defense Communications Division
San Diego, California 92131

## ABSTRACT

It has been shown that a continuous speech recognizer with a set of sub-word templates can provide excellent performance as a text-independent speaker identifier. The performance of this kind of speaker identifier becomes poor when the speech samples are very short utterances, due to variations in matching scores arising from the unpredictable phonetic content of the small speech sample. This paper describes new normalization and selection techniques which improve speaker recognition accuracy using very short uncontrolled speech samples. The first normalization depends on the means and variances of scores of a short sample of unknown matched to different models of many speakers. The selection procedure discards portions of a speech sample with poor speaker discrimination ability. A second normalization is based on the range of matching scores of the claimed speaker's model against other speakers' models. It facilitates setting acceptance thresholds for speaker verification against an open population.

## INTRODUCTION

Higgins and Wohlford [1] have shown that a continuous speech recognizer with sub-word templates can be used as a text-independent speaker identifier. The recognizer scores unknown speech by matching it against speaker models in the form of a small number (about 50) of short (< 100 msec) templates, with the identity of the unknown speaker inferred from the best scoring model.

When test data are subdivided into segments of very short durations, the speaker's scores vary widely at some segments, whereat the speaker cannot be identified reliably by ascribing the speech to the best scoring model. The variable reliability of short, uncontrolled speech segments is therefore a major cause of increasing speaker recognition error rate using short unknown speech samples.

The variation of scoring reliability is graphically illustrated in Figure 1. It shows the distribution of scores against successive 0.4 sec. segments of connected speech from which silence has been removed. The matching scores for the true speaker's model is plotted as a heavy line. The dots are scores against 25 other speakers' models. The logatirhmic amplitude of each test segment is shown at the bottom. The distribution of scores is seen to vary widely from segment to

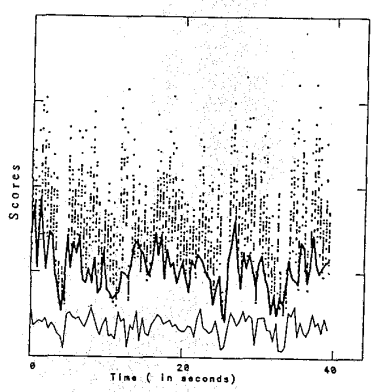


Figure 1. CSR 400 msec Segmental Scores from Speaker and Imposter Models

segment. Although the true speaker's model usually gives the lowest score among all models, that is not always the case. Both the mean and the spread of other speakers' scores are quite variable. The rate of change of the score distributions suggests that it is closely related to the phonetic content of the unknown speech.

Overall, high amplitude segments give better discrimination of the true speaker than do low amplitude segments. Good discrimination of the true speaker also tends to be associated with a large variance of scores across all models.

Speaker verification requires reaching a decision as to whether the speaker is who he claims to be, based on a comparison of the observed speech with what is expected from the claimed speaker *vs.* what is expected from the general population. The decision is reached by comparing some form of score with a threshold, and the threshold is adjusted to obtain a desired balance between false acceptance and false rejection errors. The variability of matching scores, shown in Figure 1 for a single speaker, clearly shows that no absolute threshold on raw scores can be chosen which will give reliable decisions even if the threshold is speaker dependent. Speaker identification, in contrast, uses no threshold, as it compares match scores of the unknown speech against a finite set of talker's models. Identification therefore avoids the threshold setting problem, but is sensitive to model biases.

Verification system performance is often reported using



CH2561-9/88/0000-0595 $1.00 © 1988 IEEE

the "equal error rate", which is based on posterior probabilities derived from test data. It is often difficult to realize, or even approach, these results with an *a priori* threshold on raw scores, due to their inconsistent distributions. This paper presents new normalization and selection techniques which improve performance with short utterances and also improve speaker verification with *a priori* thresholds, as distributions of the resulting scores are more stable than those of raw scores.

## DATABASE AND FRONT-END PROCESSING:

*Database:*

The database used in this study is a set of conversations recorded by 26 male speakers. All recording sessions were at least one week apart for all speakers. Front-end processing includes adaptive pre-emphasis and extraction of ten LPC-derived cepstra and ten log-area-ratio values, comprising the 20 parameters used. The frame rate is 20 msec.

*Baseline Training:*

The training (extraction of speaker models) is the same as for the speaker identification system described earlier [1]. A set of fifty 4-frame templates which are generated for each speaker from 40 sec. of his recorded conversational speech is used as the reference model. The raw score is obtained from dynamic programming matching of each 400 msec. (20 frame) segment to templates in the set comprising a model. A score obtained by matching speech with the true speaker's model is called the true speaker score, and scores against others' models are denoted imposter scores. In this study, 26 reference models for 26 speakers give one true speaker score and 25 imposter scores, for each 400 msec. segment.

*Scoring and Additional Recording Sessions*

True speaker scores are very variable. Scores against the same speech from which the model was built are particularly biased, and not at all indicative of true speaker scores which will be obtained from other speech or other sessions. These factors exacerbate the problem of determining thresholds, and necessitate using speech from different sessions to create models and to set thresholds. We therefore distinguish three sets of speech data; training data used for forming models, development data used for setting thresholds, and test data used for evaluation. In this study all three sets of data were from distinct recording sessions.

## NORMALIZATION AND SELECTION:

Figure 1 shows the true speaker score is fairly stable relative to the distribution of imposter scores, although the imposter score distribution itself varies widely. This observation suggests that normalizing the raw score by making it relative to the mean and standard deviation of the distribution of imposter scores will remove most of the variability across segments and produce a more stable numerical score, and one which is simply related to the conditional probability that the 400 msec. segment was spoken by the true speaker, given the observed scores against all models, irrespective of the speech segment. This concept of normalization is similar to the likelihood scoring decision process [1], which also improves accuracy of speaker verification.

### Segment by segment normalization:

Figure 2 shows a block diagram of the procedure for the normalization and selection. The first segment-by-segment normalization is done on every short utterance segment (every 400 msec.). Scores at this interval result from matching about four or five templates to the unknown speech. One

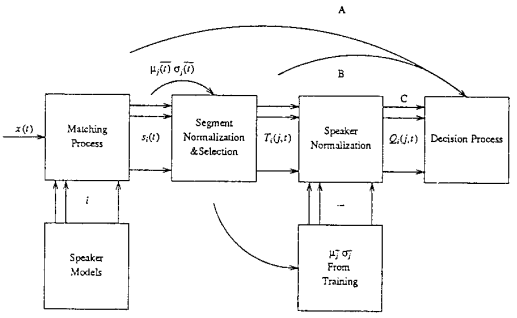


Figure 2. Block Diagram of Segment and Speaker Normalization Process

might expect improvement in performance with this normalization alone, especially since the accumulation becomes optimal if the distribution of scores is a Gaussian. However, we have found that identification results become somewhat poorer, which indicates that the weighting of segments with small original score standard deviations causes more errors.

### Segmental Selection:

Fig 1. shows that portions of speech with low level and small raw score variance frequently cause poor true speaker scoring. From the same set of data, we may obtain a scatter plot of normalized score against the standard deviation of raw scores, as shown in Fig. 3. It clearly shows the high propensity to error when the standard deviation is small for this speaker, and that selection of segments becomes desirable before the decision process. Since the probability of error is higher for both low levels of speech and small value of raw score standard deviation, a selection procedure is introduced to discard those segments. Criteria are set so that about 20 to 25% of the non-silence segments are discarded. The same speech scores as shown in Fig 1. are plotted again after the normalization and selection, in Fig. 4. The improvement afforded by the normalization and selection is apparent in the improved stability and reduced relative spread of the true speaker normalized score.

### Speaker model normalization:

After the segmental normalization and selection just described, each speaker model is seen to have characteristic distributions of true speaker and imposter scores * across

* Hereafter, "score" means the segment - normalized score, as opposed to the un-normalized raw score.

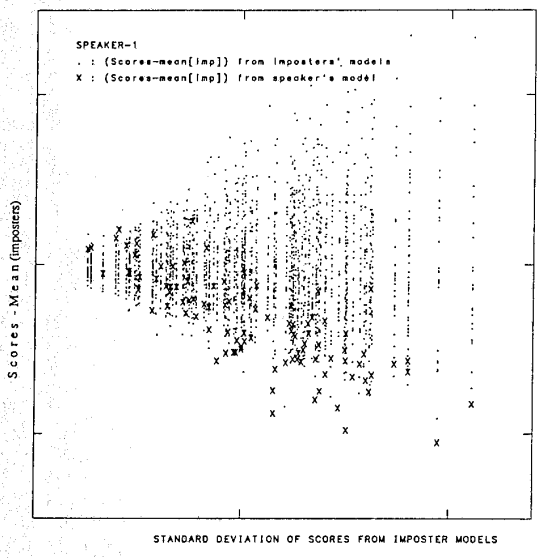


Figure 3. Scatter Plot of CSR Scores vs Standard Deviation of Imposter Scores

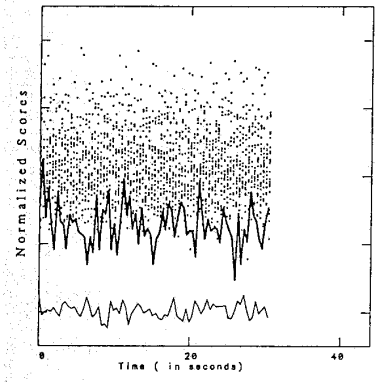


Figure 4. Normalized and Selected Segmental Scores from Speaker and Imposter Models

segments which vary from one model to the next. It is found that the true scores and imposter scores tend to be positively correlated; that is, models tend to produce biased scores. In speaker identification, where scores against two models are compared, this bias can lead to error if it is not removed. In speaker verification, this model bias adds to the difficulty of setting thresholds. It is therefore desirable to further normalize scores with respect to the distribution of scores of imposter models across segments **. The mean and standard deviation of the segment-normalized scores is estimated from scores of each model against the training data. The second normalization is accomplished by subtracting the mean and dividing by the standard deviation, analogous to the segment normalization procedure.

** It usually occurs that the amount of speech training data for each speaker is too limited to estimate the parameters of the true speaker score distribution.

## EXPERIMENTAL RESULTS:

### Speaker identification:

Speaker identification results are shown in Fig 5 as a function of the test sample duration. It shows that the

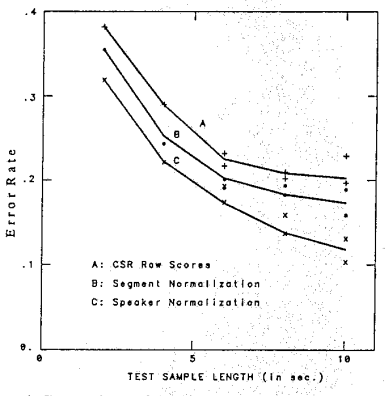


Figure 5. Comparison of Speaker Identification Results

improvement due to the normalizations is significant. There is a larger proportional reduction of error at the longer test sample durations than for short duration samples.

### Speaker verification:

*Criteria for Determining Decision Thresholds:*

There are three different kinds of results to report the performance by using posterior probabilities. Fig. 6 shows the accumulated probability function of false rejection and false acceptance, when the decision is made every 1.2 sec by accumulating three scores of 400 msec. segments. The abscissa of the crossover of these two curves is the threshold to obtain equal error rate (EER), and the ordinate there is the EER. The figure also shows the minimum average false rejection rate and false acceptance rate (MAFRFA). The third scoring technique is to use a constant false acceptance (CFA) rate to establish the threshold to measure the average false rejection rate. This result is usually higher than both previous

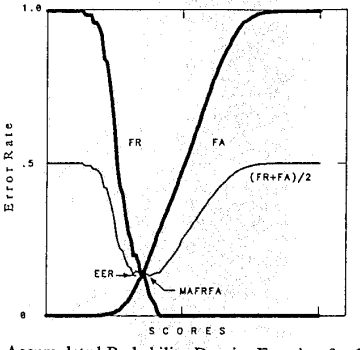


Figure 6. Accumulated Probability Density Function for Speaker and Imposters.

error rates. Fig 7 is a comparison of results from development and testing recording sessions using these three scoring techniques.

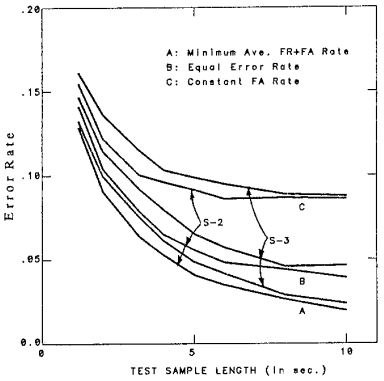


Figure 7. Comparison of Different Scoring Techniques on Two Recording Sessions.

*Verification Results of Normalizations:*

Figure 8 shows speaker verification results using the average MAFRFA rate as a performance criterion, using raw

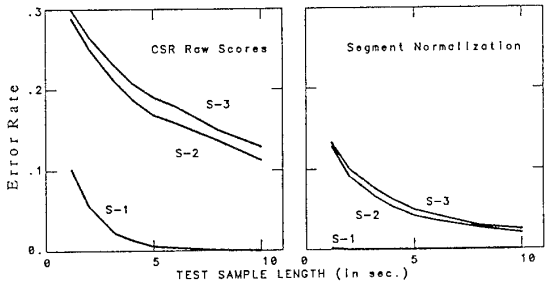


Figure 8. Comparisons of CSR Raw Scores and Different Normalization Scorings

CSR scores and segment normalized scores. It shows that segment normalization does provide a significant improvement over the raw score. We also find that there are almost no errors at any sample length when testing against training data S-1. (Speaker normalization has no effect on speaker verification, as it only rationalizes scores between models.)

*Threshold decision results:*

The development database was used to find the optimal thresholds either by EER, MAFRFA, or CFA, which were then used as the decision criterion on the test database. The thresholds were estimated speaker dependently, and the results are shown in Fig 9. We find that the best results by pre-determined threshold is about 2-3 time higher than the average MAFRFA error rate obtained for the database. The most encouraging result is that the distribution of imposter scores is very stable and each model has the same mean and spread of imposter scores. Thresholds found in this way can therefore be used for an open set of speakers.

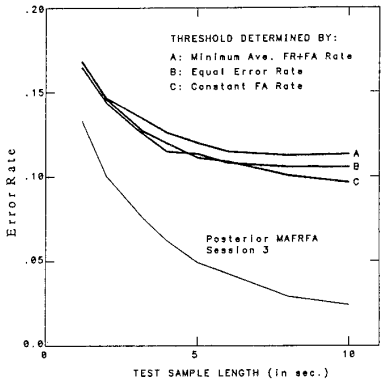


Figure 9. Speaker Verification Results of Speaker Dependant Threshold Determined from Other Sessions

The performance with thresholds determined by CFA is slightly better than when they are based on the other two criteria, but long duration utterance recognition has shown a flatter curve which indicates the less improvement for longer utterances.

The major problem remains session-to-session changes for some speakers. The problem becomes more serious with longer duration decisions based on accumulating scores. At 10 sec. test duration length, more than 90% of the errors are caused by 30% of the total population. Most speakers have a much lower error rate (<1/2%) compared with the poorest speakers (usually > 20%). At this duration one or two speakers have been 100% falsely rejected by their own models.

## References

1. A. L. Higgins and R. E. Wohlford, "A New Method of Text-Independent Speaker Recognition," *ICASSP-86*, vol. 2, no. 1, April 1986. Tokyo, Japan