A New On-Line Model Quality Evaluation Method for Speaker Verification

Javier R. Saeta[1] and Javier Hernando[2]

[1] Biometric Technologies, S.L., 08007 Barcelona, Spain
j.rodriguez@biometco.com
[2] TALP Research Center, Universitat Politecnica de Catalunya, Barcelona, Spain
javier@talp.upc.es

Abstract. The accurate selection of the utterances is very important to obtain right estimated speaker models in speaker verification. In this sense, it is important to determine the quality of the utterances and to establish a mechanism to automatically discard or accept them. In real-time speaker verification applications, it is decisive to obtain on-line measures to ask the speaker for more data if necessary. In this paper, we introduce a new on-line quality method based on a male and a female Universal Background Model (UBM). These two models act as a reference for new incoming utterances in order to decide if they can be used to estimate the speaker model or not. Text-dependent experiments have been carried out by using a telephonic multi-session database in Spanish. The database has been recorded by the authors and has 184 speakers.

1 Introduction

In a Speaker Verification (SV) system, the user enrolls the system by pronouncing some utterances in order to estimate a speaker model. The enrollment procedure is one of the most critical stages of a SV process. At the same time, it becomes essential to carry out a successful training process to obtain a good performance. The importance and sensitiveness of the process force us to pay special attention on it. Consequently, it is necessary to protect the enrollment procedure by giving the user some security mechanisms, like extra passwords or by providing a limited physical access. There are a lot of cases where the training process is vulnerable. One of the most common ones is when the enrollment is done by phone.

In such cases, occasional impostors could seriously damage the speaker model especially if we are training the model in several sessions. This could also be applied to models which are adapted from new utterances coming from the speaker.

But sometimes background noises, distortions or heavy colds can produce similar effects to the ones with real impostors. For this reason, it is convenient to control the quality of the speaker utterances to detect low quality ones and prevent from errors in the estimation of the speaker model.

The quality of a model mainly depends on the reliability and variability of the utterances and on the training and test conditions. It is crucial that the speaker model includes the most discriminative speaker characteristics. In real applications, one can normally afford one or two enrollment sessions only. In this context, it is important to control the content and quality of the recorded voice samples, when the enrollment process is 'open', i.e., when the speaker is talking and the utterances are being recorded.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 572–579, 2005.

Quality model measures evaluate how discriminative a model is by comparing client and/or impostor utterances against the model. Some approaches to the problem of model quality evaluation have traditionally dealt with outliers, i.e., those client scores which are distant with respect to the mean in terms of Log-Likelihood Ratio (LLR). They use the distance between the training model and the utterances used to estimate the model. The 'leave-one-out' method [1] has the problem of an excessive computational cost. The Z method [2] uses impostor data. The method introduced by the authors in [3] overcomes these two problems but, as it happens with the first two methods, it needs the speaker model to evaluate quality.

In this paper, we introduce a new on-line quality method to detect non-profitable or non-representative utterances coming from an impostor or from the own speaker. When an undesired utterance is located, the system asks the user for a new one. The method compares an utterance against a male and a female UBM, previously estimated from a collected corpus. Two scores are obtained. These scores are used to locate the utterance with respect to the UBMs. In principle, utterances from the same speaker are similar enough between them so when a new utterance is compared against the UBMs, the score should be similar to the ones obtained before for the rest of the speaker utterances. This is the basis of the on-line quality model method.

A theoretical view of the state-of-the-art is reported on the next section. New proposers are developed in section 3. The experimental setup and the evaluation with empirical results are described in section 4, followed by conclusions in section 5.

2 Theoretical Approach

Some approaches have been previously shown in literature concerning the evaluation of quality models. In [1], a model quality checking method called 'leave-one-out' is introduced. It uses N-1 utterances from a total of N utterances to train the model. N scores are obtained by testing every utterance against the model. The model that yields the highest score on the test utterance is the most representative model. The lowest scores belong to utterances which can be considered as outliers. The main problem of the 'leave-one-out' is that it estimates the model N times to detect the best representative one. It implies a huge computational cost.

Another different approach [2] to check model quality introduces the distance Z between LLR scores from clients and from impostors for a given model:

$$Z = \frac{\max\{0, \mu_C - \mu_I\}}{\sigma_I} \tag{1}$$

where μ_C is the mean LLR score on client utterances of the given model and μ_I and σ_I are, respectively, the mean and standard deviation of LLR scores on a set of impostor utterances. Z shows how discriminative a model is. If Z is close to zero, a low discrimination is expected. Z method has the problem of using data from impostors and it is common to deal only with client data in some applications.

Another measure which has been introduced by the authors in [3] uses an algorithm to determine the quality level of a speaker model. This algorithm decides if an utterance is a good representation of the model according to an iterative process. We define s_n as a LLR score obtained by testing an utterance against its own model. We assume that an utterance has an acceptable degree of quality when it surpasses the following interval:

$$s_n \geq \mu_C - \alpha\sigma_C \tag{2}$$

where μ_C and σ_C are the mean and standard deviation of LLR scores on the utterances used to train the model. The coefficient α is empirically determined.

The method is applied to the enrollment data in combination with an algorithm to find the less representative utterances for every speaker. Once these outliers are located, they can be suppressed or replaced by new ones coming from the same speaker. It classifies the speaker models according to their quality. The classification will detect reduced quality models. Models will be placed into different groups depending on the degree of similarity of their utterances with their respective models.

The use of this method with client data is especially useful when it is difficult to obtain data from impostors, for instance in phrase-prompted cases. When using words or phrases as passwords – except in connected digits-, this method will be generally more suitable than the one explained before which employed Z to determine the model discrimination, because that method used data from impostors.

On the other hand, in comparison with the 'leave-one-out' method, the last method is more effective in terms of computational cost. If N is the number of client model utterances, the 'leave-one-out' method trains N models per client to evaluate quality while the method showed in (2) trains a maximum of the whole part of N/5 models.

But the problem of the last method – and also of the first two ones – is that it is not possible to ask the user for new data until the model is estimated. And this inconvenience is especially critical when we use only one session for training or when we are in the second session of a two-session enrolment process. If there are some low quality utterances, we lost the opportunity of obtaining more voice samples from the speaker when (s)he is just recording them. It could lead to wrong estimated or undertrained models.

3 New On-Line Method

Until now, the evaluation of the quality of the speaker model took place once the model was estimated but the use of quality measures has several disadvantages then. The main problem is that, in real-applications, we do not have the option of asking for more utterances to the speaker. This is basically because modern systems use to train the speaker in one – maximum two – session(s). Furthermore, the number of utterances tends to be small.

In this case, even when we detect that an utterance has a bad quality or comes from an intentional impostor, it is not possible to ask the speaker for a new one.

With on-line model quality measures we solve this problem because we decide if an utterance has a sufficient degree of quality before estimating the model and, what is more important, before adding this utterance to the speaker model.

The algorithm works as follows:

1. Obtain LLR scores $\{s_{1m}, s_{2m}, s_{3m}...\}$ and $\{s_{1f}, s_{2f}, s_{3f}...\}$ from incoming utterances $\{U_1, U_2, U_3...\}$ against $\{UBM_m, UBM_f\}$
2. Estimate $\{\mu_m , \mu_f\}$ from the previous scores

3. Ask for a new utterance U_n and obtain $\{s_{nm}, s_{nf}\}$ against $\{UBM_m, UBM_f\}$
4. Calculate a distance $d_{mf} = |\mu_m - s_{nm}| + |\mu_f - s_{nf}|$
5. If $d_{mf} > \Theta$, quality is considered as sufficient. If $d_{mf} \leq \Theta$, then go to 3

First of all, we obtain a pair of scores for every utterance $\{U_1, U_2, U_3...\}$, one against a male UBM_m and another one against a female UBM_f. From the moment we obtain some new utterances, we estimate the mean $\{\mu_m , \mu_f\}$ for every pair of scores. Thus, a comparison takes place when new incoming utterances (U_n) are obtained for the speaker. They should not be far – in terms of LLR – from that estimated mean if they really belong to the speaker. The process is shown in the following scheme:

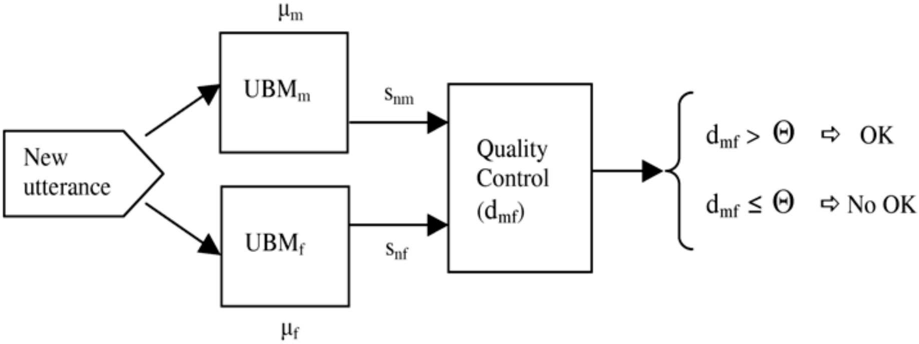

Fig. 1. Block diagram for the on-line quality algorithm

Finally, we set a maximum distance d_{mf} and reject utterances that surpass that distance because they have not reached the minimum degree of quality required, fixed by a threshold Θ. D_{mf} is a conventional distance which has been shown as suitable in our experiments. Of course, more work could be done to find a more optimized one.

The threshold Θ is empirically determined. It is obvious that the quality estimation becomes more robust if using as more utterances as possible to established the maximum distance allowed to considerate an acceptable degree of quality.

The on-line quality method has similarities to the Tnorm [4] normalization technique because the score is obtained on-line by comparing – in the Tnorm case – the test utterance to the client model and to some impostor models.

4 Experiments

4.1 Database

The database used in this work has been recorded by the authors and has been especially designed for speaker recognition. It includes land-line and mobile telephone sessions. 184 speakers were recorded by phone, 106 male and 78 female. It is a multi-session database in Spanish, with 520 calls from the Public Switched Telephone Network (PSTN) and 328 from mobile telephones. One hundred speakers have at least 5 or more sessions. The average number of sessions per speaker is 4.55. The average time between sessions per speaker is 11.48 days.

Each session includes:

a) 4 different sequences of 8-digit numbers, repeated twice.
b) 2 different sequences of 4-digit numbers, repeated twice.
c) 6 different isolated words.
d) 5 different sentences.
e) 1 minute long read paragraph.
f) 1 minute of spontaneous speech.

4.2 Experimental Setup

In our experiments, utterances are processed in 25 ms frames, Hamming windowed and pre-emphasized. The feature set is formed by 12th order Mel-Frequency Cepstral Coefficients (MFCC) and the normalized log energy. Delta and delta-delta parameters are computed to form a 39-dimensional vector for each frame. Cepstral Mean Subtraction (CMS) is also applied.

Left-to-right HMM models with 2 states per phoneme and 1 mixture component per state are obtained for each digit. Client and world models have the same topology. The UBM for each digit is estimated with data from a balanced subset of speakers of the database, concretly those speakers which have recorded one to four enrollment sessions only (over 25 speakers).

The speaker verification is performed in combination with a speech recognizer for connected digits recognition. During enrollment, those utterances catalogued as "no voice" are discarded. This ensures a minimum quality for the threshold setting.

Clients have a minimum of 5 sessions. It yields 100 clients. We use 4 sessions for enrollment and the rest of sessions to perform client tests or for adding more data to the speaker in quality model experiments. Speakers non-included in digit UBMs with more than one session and less than 5 sessions are used as impostors. 4 – and 8-digit utterances are employed for enrollment and 8-digit for testing. We apply verbal information verification [5] as a filter to remove low quality utterances. The total number of training utterances per speaker goes from 8 to 48. The exact number depends on the number of utterances discarded by the speech recognizer. During test, the speech recognizer discards those digits with a low probability and selects utterances which have exactly 8 digits.

It is important to note that fixed-line and mobile telephone sessions are used indistinctly to train or test. This factor increases the error rate.

4.3 Verification Results

Our verification experiments with connected digits show the Equal Error Rate (EER) for the baseline, the 'leave-one-out' method [1], the method introduced in [3] and the on-line quality method introduced in this paper.

The 'leave-one-out' method has been used here without predefined thresholds. Like the other experimental methods presented here, it uses the Speaker Dependent Threshold (SDT) method of the following equation [6, 7]:

$$\Theta_x = \alpha\,\mu_I + (1-\alpha)\,\mu_C \tag{3}$$

where μ_C is the client scores mean, μ_I is the impostor scores mean and α is a constant which has to be optimized from a pool of speakers.

Table 1. Error rates for a set of speakers in connected digit verification experiments with SDT defined in (3)

Quality methods	EER (%)
Baseline	2.23
Leave-one-out	2.02
Without outliers	5.86
On-line method	2.00

As we can see in Table 1, the baseline experiments give an EER over 2%. The 'leave-one-out' method slightly improves the baseline experiments, but its enormous computational cost makes it unaffordable.

In the third method, an average of 2.3 utterances per speaker were removed for the 44 speakers with low quality. The error rates dramatically increased by removing only a few utterances considered as outliers. That reflects the importance of data when estimating a model. In our case, we have found that it is better to keep data even when we have realized that they are not the best representation of the speaker. This is especially important when we do not use too much data to estimate the speaker model or when the handsets for training and testing are different because it can cause errors in the selection of outliers.

On the other hand, in case we replace outliers by new and more representative data from the speaker, we reduce error rates by around 40% and the system performs better than the baseline (EER = 1.39%).

The on-line quality measure consists of a simulation for an enrollment procedure with 4 training sessions per speaker. The algorithm tests the quality of the utterances by means of the on-line quality method and decides if there are non-representative utterances. If the measure reveals bad quality utterances, they are replaced by new ones from the fifth session of the speaker. If the number of non-representative samples exceeds the number of valid utterances of the fifth session, bad quality utterances are removed anyway. In this case, some models are trained with a smaller number of utterances than initially – a reduction of 8% of the data with respect to the baseline. This increases the error rates.

The whole process can be done in real-time because the model is not estimated until the minimum number of utterances is reached. The use of on-line quality measure reduces the error although not very significantly because the threshold is estimated using impostor data. In this case, the influence of non-representative utterances can be better minimized than in cases when only material from clients is available. Furthermore, not every utterance discarded by the on-line method was replaced by a new one from the fifth session. Some of them could not be replaced because of the bad quality of the utterances of the fifth session for some speakers. Anyway, the on-line quality method has the advantage of determining the quality before the creation of the model.

The following table shows a comparison of the EER (%) for threshold estimation methods with client data only, without impostors:

The baseline SDT method for Table 2 is defined as follows [8]:

$$\Theta_x = \mu_C - \alpha\, \sigma_C \tag{4}$$

where μ_C is the client scores mean, σ_C is the standard deviation from clients and α is a constant empirically determined.

Table 2. Comparison of threshold estimation methods in terms of EER (%) with data from clients only

Quality methods	Baseline	On-line method
Baseline	5.89	4.50
Baseline + 2 impostor utterances	6.19	4.72

Two intentional impostor utterances per speaker are added here to the baseline during training to taint the enrollment process. We add two utterances from a male voice for men and two female utterances for women.

The on-line quality method discards the 94% of these utterances. At the same time and despite the presence of intentional impostors and the elimination of some training data, the on-line method reduces the error rate with respect to the baseline.

As we can see from table 2, the on-line measures, with and without 2 impostors, perform better than their respective baselines.

5 Conclusions

A new on-line model quality evaluation algorithm has been introduced here. It outperforms the 'leave-one-out' method in terms of computational cost and it has the advantage of using only data from clients, which is strongly recommended when dealing with words or phrases as passwords and it is difficult to obtain data from impostors.

The new algorithm has the advantage of estimating quality without needing the speaker model. This implies that the quality can be measured on-line. In our experiments, the method was capable of rejecting 94% of intentional impostor utterances while preserving client utterances. The best on-line quality performance was achieved with a threshold that used impostor data.

Although the improvement is not very sensitive when adding two impostor utterances, further work will show that if the number of intentional impostor utterances is increased, the use of the on-line quality evaluation method will result in a substantial improvement.

References

1. Gu, Y., Jongebloed, H., Iskra, D., Os, E., Boves, L.: Speaker Verification in Operational Environments-Monitoring for Improved Service Operation, ICSLP'00, Vol. II, 450-453, Beijing (2000).
2. Koolwaaij, J., Boves, L., Os, E. den, Jongebloed, H.: On Model Quality and evaluation in Speaker Verification, ICASSP'00, 3759-3762, Istanbul (2000).
3. Saeta, J.R., Hernando, J.: Model Quality Evaluation during Enrollment for Speaker Verification, 8th International Conference on Spoken Language Processing (ICSLP), 352-355, Jeju (South Korea), October (2004).
4. Auckentaler, R., Carey, M., Lloyd-Thomas, H.: Score Normalization for Text-Independent Speaker Verification Systems, Digital Signal Processing, Vol. 10, 42-54, 2000.
5. Li, Q., Juang, B.H., Zhou, Q., Lee, C.H.: Verbal Information Verification, Proc. Eurospeech'97, 839-842.

6. Pierrot, J.B., Lindberg, J., Koolwaaij, J., Hutter, H.P., Genoud, D., Blomberg, M., Bimbot, F.: A Comparison of A Priori Threshold Setting Procedures for Speaker Verification in the CAVE Project, Proc. ICASSP'98, 125-128.
7. Lindberg, J., Koolwaaij, J., Hutter, H.P., Genoud, D., Pierrot, J.B., Blomberg, M., Bimbot, F., Techniques for A Priori Decision Threshold Estimation in Speaker Verification, Proc. RLA2C, Avignon (1998) 89-92.
8. Saeta, J.R., Hernando, J.: Automatic Estimation of A Priori Speaker Dependent Thresholds in Speaker Verification, Proc. 4th International Conference in Audio- and Video-based Biometric Person Authentication (AVBPA), ed. Springer-Verlag, 70-77 (2003).