Model Quality Evaluation during Enrollment for Speaker Verification

Javier R. Saeta¹, Javier Hernando²

¹ Biometric Technologies, S.L. Barcelona, Spain
j.rodriguez@biometco.com

² TALP Research Center. Universitat Politècnica de Catalunya, Spain
javier@talp.upc.es

Abstract

The amount of data usually determines the robustness of speaker models in speaker recognition. In this sense, it is convenient to set a model quality measure for every speaker and to classify models into different categories according to their quality level. We propose a new quality measure, which uses only data from clients, based on the number of training utterances that surpass a predefined threshold. If the desired quality is not high enough, the quality measure allows for the detection of non-representative utterances. Once selected, these utterances, considered as outliers, can be removed or better replaced by new ones coming from the same speaker. A database of 184 speakers in Spanish is used to obtain empirical results with connected digits. Our experiments removing outliers and replacing them by new utterances coming from the same speaker outperform the baseline experiments by 40%.

1. Introduction

Training robust models is an essential issue in speaker verification. The speaker model should include the most discriminative speaker characteristics. Moreover, the more training data available, the more robust model can be estimated. However, in real applications, we typically have only a few utterances for enrollment. Furthermore an unknown telephone channel aggravates the problem [1]. In this context, the impact of those utterances affected by adverse conditions becomes more important than in such cases where a great amount of data is available. Background noises, distortions or strange articulatory effects may produce non-representative scores, i.e., outliers. In terms of Log-Likelihood Ratio (LLR), we define the outliers as those client scores which are distant with respect to mean.

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 [2, 3] have traditionally dealt with outliers. They use the distance between the training model and the utterances used to estimate the model.

We introduce in this paper a new model quality measure in order to detect reduced quality models. The most important feature of this new measure is that it uses only data from clients. The measure 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. The selection of suitable data in the training period causes an important improvement in the performance of a speaker verification system in terms of Equal Error Rate (EER).

In this paper, we also introduce a classification of speaker models according to their quality. The classification will provide a method to validate good quality models and to detect reduced quality models. Models are placed into different groups depending on the degree of similarity of their utterances with their respective models. We have defined four levels of quality in our experiments.

Applying these techniques will result in a substantial improvement of the performance by adding new data or by retraining the model without the presence of outliers.

A theoretical view of the state-of-the-art and a new technique for evaluating the model quality are reported on the next section. The experimental setup, a new model quality classification and the empirical results are described in section 3, followed by conclusions in section 4.

2. Theoretical approach

Some approaches to model quality measures have been previously shown in literature. In [2], 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.

Another different approach [3] to check model quality introduces the distance Z between LLR scores from clients and from impostors for a given model:
\[ Z = \frac{\max \left\{ 0, \mu_c - \mu_i \right\}}{\sigma_i} \quad , \]  

(1)

where \( \mu_c \) is the mean LLR score on client utterances of the given model and \( \mu_i \) and \( \sigma_i \) 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.

We propose here a new algorithm to determine the quality level of a speaker model. Once the model is estimated from an initial set of utterances, the next step consists in checking the model quality and deciding if it is high enough. If not, the less representative score/utterance is replaced by another one. The model quality measure is applied again and a new decision is taken until the quality becomes higher than a certain value or the maximum number of iterations is reached.

If \( N \) is the number of client model utterances, the maximum number of iterations for this client will be the whole part of \( N/5 \). This number has been empirically determined from a pool of speakers. The minimum \( N \) from which we decide to use our method will be \( N=5 \).

In order to apply the proposed algorithm, we introduce here a new model quality measure. 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:

\[ s_n \geq \mu_c - \alpha \sigma_c \quad . \]  

(2)

where \( \mu_c \) and \( \sigma_c \) are the mean and standard deviation of LLR scores on the utterances used to train the model. The coefficient \( \alpha \) is empirically determined.

Our method to set the quality uses only data from clients. This is especially important 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–, our 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, ours 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 our method trains a maximum of the whole part of \( N/5 \) models.

The a priori speaker-dependent threshold (SDT) method proposed here to estimate the threshold uses data from clients and impostors [4, 5, 6] according to:

\[ \Theta_x = \beta \hat{M}_x + (1 - \beta) \tilde{M}_x \quad , \]  

(3)

where \( \hat{M}_x \) is the client scores mean, \( \tilde{M}_x \) is the impostor scores mean and \( \beta \) is a constant empirically determined.

3. Experiments

3.1. Database

The database used in this work has been recorded by the authors and especially designed for speaker recognition. It includes fixed-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:
- 4 different sequences of 8-digit numbers, repeated twice.
- 2 different sequences of 4-digit numbers, repeated twice.
- 6 different isolated words.
- 5 different sentences.
- 1 minute long read paragraph.
- 1 minute of spontaneous speech.

3.2. Experimental setup

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 Hidden Markov Models (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 silence model is shared by the client and the Universal Background Model (UBM).

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 selection ensures a minimum quality for the threshold setting.
Our experiments include speakers with a minimum of 5 recorded sessions for enrollment. It yields 98 clients. We use 4 sessions for enrollment, one additional session to add data to the model if necessary and the rest of sessions to perform client tests. Speakers with more than one session and less than 5 sessions have been used as impostors. 8-digit and 4-digit utterances are employed for enrollment whereas 8-digit utterances are used for tests. The number of training utterances varies with the speaker because of the application of the speech recognizer, which discards low quality or silence utterances. This number is between 16 and 48 utterances per speaker.

Tests are carried out with 8-digit utterances. 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.

### 3.3. Model quality classification

Our proposal is to detect and replace an outlier by a new utterance and to define some quality levels where we can place every model according to its characteristics. At this point we define four quality measures depending on the number of LLR scores $n_s$ that accomplishes Equation (2):

- **I**: $n_s \geq 95\%$
- **II**: $95\% > n_s \geq 85\%$
- **III**: $90\% > n_s \geq 95\%$
- **IV**: $n_s < 85\%$

A model belongs to a certain quality level according to these percentages of utterances. For instance, quality I means that the 95% of the LLR scores ($s_n$) used to train the model fulfills the condition defined in Equation (2). If a speaker model is included in quality groups I or II, we consider that the quality is enough for our experiments and do not use the algorithm. Otherwise we iterate and stop when $n_s \geq 90\%$.

### 3.4. Verification results

Our verification experiments with connected digits show the False Acceptance (FA) and False Rejection (FR) rates for the baseline and the ‘leave-one-out’ method. Furthermore, they also show the effect of removing low quality utterances and how the error rates improve if we substitute them by another ones coming from the same speaker.

The ‘leave-one-out’ method has been used here without predefined thresholds. Like the other experimental methods presented here, it uses the SDT method of the Equation (3).

<table>
<thead>
<tr>
<th>Qualities</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>8</td>
<td>46</td>
<td>43</td>
<td>1</td>
</tr>
<tr>
<td>Without outliers</td>
<td>12</td>
<td>83</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>Without outliers + new data</td>
<td>15</td>
<td>81</td>
<td>2</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 1**: Quality groups for a set of speakers

Only 8 models obtain the maximum quality in baseline experiments. The majority are of quality II and III and even one of them achieves the lowest quality, as we can see in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>FAR</th>
<th>FRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>2.25</td>
<td>2.21</td>
</tr>
<tr>
<td>‘Leave-one-out’</td>
<td>2.04</td>
<td>2.00</td>
</tr>
<tr>
<td>Without outliers</td>
<td>5.89</td>
<td>5.84</td>
</tr>
<tr>
<td>Without outliers + new data</td>
<td>1.40</td>
<td>1.37</td>
</tr>
</tbody>
</table>

**Table 2**: Error rates for a set of speakers in connected digit verification experiments
As we can see in Table 2, the baseline experiments give FRR and FAR over 2%. The ‘leave-one-out’ method slightly improves the baseline experiments.

In the whole process, an average of 2 utterances per speaker where removed for the 44 speakers with low quality. The error rates have dramatically increased by removing only a few utterances considered as outliers. That reflects the importance of data when estimating a model. In our case, it is better to keep data even when we have found 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.

3.5. Discussion

These experiments may be influenced by the random use of sessions for training and testing because the speaker was allowed to call from a fixed-line or a mobile telephone. There are cases where every training session comes from a fixed-line phone and its corresponding tests use only utterances recorded from mobile phones for the same speaker. In this context, it could be cases where only a few utterances coming from a mobile phone are used to estimate the model. If some of them are selected as outliers and removed, the model will probably perform worse with new mobile telephone test utterances coming from impostors or clients.

The channel mismatch between training and testing could produce some unexpected errors in the outliers selection. In this context, it would be suitable in our case to analyze model by model the proportion of training utterances of every channel and especially those catalogued as outliers and the relations of unbalanced models with errors in tests. The careful selection of outliers could lead to an improvement in general performance.

4. Conclusions

In this paper, a new model quality measure has been introduced. It lets the classification of models into different categories according to the number of LLR client scores which exceed a certain threshold. 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.

Our empirical results have shown that the elimination of those utterances that reduce quality increases the error rates if these utterances are not replaced by new ones that better reflect speaker features. The impact of removing data becomes very significant and suggests us to be careful when selecting outliers and removing the utterances, especially if we are not able to replace them by more speaker data.

The analysis of results should take into account that the random choice of handset to train and test deteriorates the general performance and probably yields some unexpected errors when deciding if a utterance can be considered as an outlier or not. If we are able to replace those utterances catalogued as outliers by new ones coming from the speaker, the baseline system is outperformed by 40%.

Further work should include how to select those models which are particularly sensitive to a little increase or decrease in the amount of data.

5. References