

AUTOMATICALLY CORRECTING BIAS IN SPEAKER RECOGNITION SYSTEMS

Yosef A. Solewicz and Moshe Koppel

Dept. of Computer Science, Bar-Ilan University, Ramat-Gan, Israel

ABSTRACT

In this paper we present a general machine learning framework for score bias reduction and analysis in Speaker recognition systems. The general principle is to learn a meta-system using recognition systems' errors, given the training and testing conditions in which they occurred. In the context of speaker recognition, the proposed method is able to reduce the bias introduced in scores due to a variety of factors such as channel mismatch, additive noise, gender mismatch, different speaking styles, etc. Moreover, this framework enables a deep understanding of the origins of score bias in any system, which will support an optimized system redesign. Preliminary results obtained with several state-of-the-art systems showed considerable improvement in original performance, in addition to identifying sources of system bias.

1. INTRODUCTION

In this paper, we introduce a new system, ABIE, for automated bias identification and elimination, designed for use in speaker recognition systems.

In general, the goal of the speaker recognition task is to determine whether two utterances come from the same speaker or not. Current speaker recognition benchmarks dictate a series of true and impostor trials in which one has to compute a score expressing the degree of similarity between each pair of training and testing utterances. The performance of a system can be estimated computing the equal error rate (EER) associated with false alarm and false reject errors caused by this system.

One of the main problems that speaker recognition systems must deal with is that of train-test mismatch: speaker models are learned from training utterances recorded under a particular set of conditions but these models are then applied to testing utterances recorded under a different set of conditions. Such mismatch in data

environment typically distorts speech features, thus introducing bias in the computed scores and finally causing light to severe degradations in system performance. Mismatch factors such as transmission channel, recording media and background noise have been the subject of intense research. In recent years, several techniques for channel and additive noise compensation have been proposed. These compensation techniques can be employed at the feature level [1, 2, 3, 4, 5], aiming to compensate for the bias introduced in speech features' distributions, or at the score level [6]. Current systems normally compensate at both levels sequentially. Score normalization, although effective, represents a blind and costly solution. Normalization is performed based on the behavior of scores obtained by a reference population in similar operating conditions. Thus, besides a large computational overload in operating mode, no explicit indications of bias sources can be obtained. Additionally, higher order feature levels have been recently explored for speaker verification [7]. Naturally, proper compensation schemes should be developed for these novel feature sets.

By contrast, ABIE is a general score compensation technique, which uses explicit feature-level information, from both high-level and traditional low-level speech features. The core of ABIE is a meta-system that learns the errors of any recognition system, based on side-information reflecting the environment in which the utterances were recorded. The type of error (false alarm or false reject) of each erroneous trial is associated with side-information extracted from the training and testing utterances comprising this trial. Once trained, ABIE should be able to predict whether a recognition error is expected given the side-information of the training and testing utterances of some trial. ABIE's outcome can then be used either to redefine a system design or to correct its scores.

This approach offers several advantages. Although operating on the score level, ABIE provides an explicit insight into bias sources due to mismatched conditions in training and testing utterances. In addition, ABIE can be overlaid on those methods previously described, yielding further improvement in accuracy, with a very low computational overload.

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

The outline of this paper is as follows. In Sections 2 and 3, we introduce the meta-system and its settings. In Sections 4 and 5, we assess our approach on NIST evaluations. In Section 6, we show how this method can be applied as diagnosis tool for improved recognition systems design.

2. ABIE OVERVIEW

The proposed methodology for score bias analysis and reduction is based on a meta-system that learns a given speaker recognition system’s flaws. The meta-system tries to correlate erroneous trials of the recognition system with corresponding side-information that presumably reflects the causes of the errors. Then, in operating mode, given a particular training/testing trial, the meta-system attempts to correct the score obtained by the speaker recognition system. Alternatively, the reasons for bias can be deduced from the meta-system learning parameters and the system can be redefined as to correct the spotted deficiency. The ABIE operation scheme is summarized in Fig. 1.

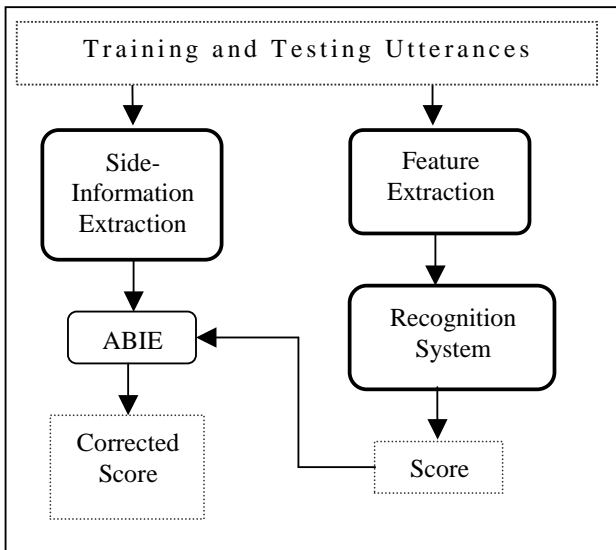


Figure 1 - Schematic representation of ABIE

More precisely, ABIE consists of the following phases:

1. For a given speaker recognition system, divide development data into training sets and validation sets. Use the training examples to learn speaker models and apply these models to examples in the validation set.
2. Detect and label false-negative (+1) and false-positive (-1) ‘errors’ committed by the recognition system. These could be either misclassified or within distances ‘Dt’ and ‘Di’ respectively from the

boundaries of the target and impostor scores’ distributions (where ‘Dt’ and ‘Di’ are parameters (Fig. 2)). In fact, Dt and Di define the score ranges for which target and impostor trials will be considered, respectively, false-negative and false-positive ‘errors’ and thus used to train the meta-system.

3. For each detected error, estimate and record “side-information” consisting of a variety of factors, such as noise and channel mismatch between training examples and poorly classified validation examples that presumably could be a cause of bias in the specific recognition system. (Details will be discussed in Section 3 below.)
4. Train some machine learning system to distinguish between the two types of errors, on the basis of the side-information captured from the detected errors. The system should output a “soft” score that represents the eventual bias expected from a particular side-information assessment. For example, we use an SVM classifier (implemented using the *SVMlight* package [8]) and record as our soft score signed distance of an example from the SVM boundary.
5. Correct the recognition system’s scores as a function of the meta-system output:

$$S' = S + k * T$$

where S' is the corrected score; S is the original recognition score; k , the correction step, is a constant, and T is the (soft) score output by the meta-system for the relevant trial. If learners other than SVM are used, the same method can be used after mapping their output to confidence scores using standard methods.

6. Iterate steps 2 to 5 above, optimizing parameters ‘Dt’, ‘Di’ and ‘k’. Use hold out partitions of the development data, searching for values that maximize performance within the sets.
7. Train the meta-system and correct the scores on an evaluation-set using the optimized parameters found in 6. Optionally, redesign the recognition system after an analysis of bias sources.

In short, we use a development set (split into train and test utterances) to learn the optimal correction to speaker verification scores on the basis of features (“side-information”) representing the dissimilarity of train and test utterances. In the following section we consider what features are appropriate for this purpose.

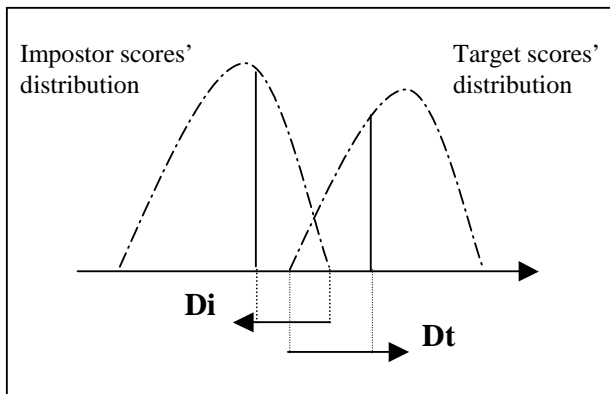


Figure 2 – ABIE’s training region

3. SIDE-INFORMATION

Proper choice of side-information is essential to ABIE’s operation. Side-information differs from the speech features used by the speaker recognition system. While those should maximize recognition performance, side-information is supposed to reflect the environment in which an utterance was recorded and should encompass a variety of factors that presumably could be a cause of bias in the specific recognition system. Additionally, more efficiency is expected the more both systems are tightly coupled at the feature extraction level. Therefore, it would be recommended that ABIE use side-information directly derived from the recognition system’s feature extraction scheme. It should be noted, however, that in these experiments the recognition systems and their corresponding meta-systems don’t use a common feature extraction scheme.

The side-information vector used in the present experiments basically attempts to cover a rough estimate of channel, noise and prosody aspects found both in target and test utterances and is composed of the following 63 attributes:

- Absolute differences between target and test mean cepstral parameters (19 components). This is an indication of channel mismatch between both utterances [3].
- Sum of target and test standard deviation of the cepstral parameters (19 components). This is an indication of the amount of additive noise in the trial [3]. (For simplicity, we denote hereafter cepstrum standard deviation by ‘Quality’.)
- Absolute differences between target and test standard deviation of the cepstral parameters (19

components). This is an indication of quality mismatch between both utterances [3].

- Absolute differences of: mean pitch, pitch standard deviation, “rate of speech” (zero-crossing of 1st cepstrum), between target and test (3 components). This would roughly reflect mismatch in higher-levels: gender, style, excitement.
- Sum of: mean pitch, pitch standard deviation, “rate of speech”, between target and test (3 components).

In fact, for each selected flawed trial of the development set, we estimate the above attributes and concatenate them into a single vector. This vector is labeled as ‘+1’ in case the corresponding trial is a misclassified target trial (i.e. a false-negative), or as ‘-1’, in case of a misclassified impostor trial. The collection of labeled attribute vectors is normalized and finally used to train ABIE.

4. EXPERIMENTS

The proposed technique was applied to several state-of-the-art speaker evaluation systems. Experiments were conducted using the NIST’04 evaluation as a development set and NIST’05 evaluation as a test set. These evaluations typically consist of 10-20 thousand trials, involving 500+ speakers recorded in a variety of landline/cellular lines [9].

Initially, we obtain scores and extract side-information vectors for all trials in a partition of the development set. We construct and label side-information vectors for each trial within the spotted error range as described in Section 3. The side-information vectors are then used to train ABIE, which is implemented as a linear SVM classifier. At this point, we extract side-information vectors from all trials belonging to another partition of the development set and input to the trained SVM. The recognition score of each such trial is corrected using its correspondent SVM soft score as described in Section 2. This procedure is iterated varying ‘Dt’, ‘Di’ and ‘k’. The values which maximize the performance on this partition are kept.

Finally, using the whole development set, we re-train the SVM classifier with all trials having scores placed within the optimized ‘Dt’ and ‘Di’ regions. We extract side-information vectors from the testing set trials and input to the SVM. Each recognition score of this data set is then corrected using the correspondent SVM’s soft output and the optimal ‘k’. In addition, the learned meta-systems parameters can be analyzed in order to reveal bias sources in the recognition system.

In the present experiments, we used scores made available by SRI for seven different systems. Their brief description can be found in [10] and they are listed in Table 1. Basically, the systems can be categorized either into acoustic (Systems 1 to 3) or stylistic (Systems 4 to 7) oriented. The acoustic systems are based on derivations of cepstral features and components of the maximum likelihood linear regression (MLLR) transforms. By contrast, the stylistic systems explore counts and duration of words and other prosodic features extracted over automatically estimated syllables (SNERF). All these systems use either GMM or SVM classifiers for modeling their respective feature sets.

System	Description
1	Cepstral GMM
2	Cepstral SVM
3	MLLR transform SVM
4	State Duration
5	SNERF
6	Word Duration
7	Word N-gram SVM

Table 1. Systems used in the experiments

5. RESULTS

Table 2 compares the original performance in EER (%) of each of the SRI systems for NIST'05, with respective corrected results. The corrected results for each system were obtained after training corresponding meta-systems in NIST'04 as described in Section 4.

As can be seen, consistent improvement was achieved on a variety of speaker recognition systems based on diversified high and low level speech features. The results obtained are very encouraging considering that state-of-the-art systems already fully optimized were analyzed and endorse ABIE as a universal and complementary bias reduction framework.

Optimized 'k', 'Dt' and 'Di' are shown in Table 3, for each of the systems. For simplicity, 'Dt' and 'Di' are expressed in terms of percentages of the total amount of target and impostor trials used for training ABIE (Fig. 2). Thus, increasing D means that we progressively consider 'good' trials as 'errors'. (In these evaluations impostor trials are ten times more frequent than target trials. Therefore, 'Dt' values ten times bigger than 'Di'

correspond to the same number of positive and negative examples.)

System	Original EER (%)	Corrected EER (%)	Relative Improvement (%)
1	7.2	6.5	9.6
2	7.3	6.6	9.2
3	10.3	9.6	7.5
4	15.4	14.1	8.1
5	14.1	13.5	4.0
6	19.3	16.7	13.5
7	24.6	20.2	17.6

Table 2. Original and corrected system performances

We note that 'k', 'Dt' and 'Di' roughly correlates with systems' EER. This is intuitive, since we expect better performing systems to require less tuning. Moreover, we observe that optimal ABIE training relies on a broad range of 'false-negatives', far beyond the recognition systems' threshold vicinity. On the other hand, only 'false-positives' strictly within the correspondent error rates are considered.

System	k	Dt (%)	Di (%)
1	0.10	30.0	6.0
2	0.10	40.0	6.0
3	0.12	25.0	8.0
4	0.26	60.0	8.5
5	0.22	70.0	2.0
6	0.40	60.0	6.0
7	0.45	75.0	9.5

Table 3. Optimum 'Dt', 'Di' and 'k' for the systems

6. BIAS ANALYSIS

As an attempt to understand the origin of bias in the several systems, we identified the meta-systems' prominent weights for each of those systems (high

absolute values of weights in the learned linear SVM). Table 4 illustrates bias analysis for some of the systems.

System	Identified Sources of Bias
1	Quality
5	Channel mismatch Rate of speech Mean pitch Quality
6	Mismatch in mean pitch Mismatch in pitch standard deviation Quality mismatch
7	Mismatch in mean pitch Mismatch in pitch standard deviation Quality mismatch

Table 4. Bias analysis

Recall that we denote by ‘Quality’, the variation strength of cepstral features. High variation indicates good quality or, correspondingly, absence of additive noise in the utterances of a trial. As could be expected, since System 1 is definitely an acoustic system, eventual bias is rooted in spectral features. On the other hand, essentially non-acoustic systems, like 5, 6 and 7 are mostly affected by bias in high-level features. Moreover, it is interesting to note that since there are no cross gender trials, mismatch in pitch related features could be a sign of different speaking styles between training and testing utterances.

7. CONCLUSIONS AND FUTURE WORK

A simple machine learning framework for bias identification and reduction in speaker verification systems was presented. The proposed meta-system represents a valuable tool for system design, diagnosis and correction. Within this framework even the impact of high-level features in any recognition system can be analyzed. It can offer a deep understanding of bias origins in any recognition system. In this paper, several state-of-the-art speaker recognition systems were analyzed and had their scores corrected within this framework. In general, significant performance improvement was obtained with all the systems. Nevertheless, extensive evaluation with diversified data is needed in order to further validate the conclusions.

This work could be tuned to other related applications, such as forensic speaker recognition, speech recognition and even to the development of novel feature

sets for robust speech processing. It would be interesting to enrich the side-information representation in order to improve results and to be more precise in detecting bias sources. Moreover, we would like to elaborate ABIE’s current training methodology and investigate alternative formalizations of the proposed scheme. For instance, we plan to train separate meta-systems, trained either in true and impostor trials and using their output ratio in a Bayesian framework.

ACKNOWLEDGMENTS

The authors are grateful to SRI for kindly making their scores available.

8. REFERENCES

- [1] S. Furui, "Cepstral analysis technique for automatic speaker verification". *IEEE Trans. Acoust. Speech Signal Process.* 29:254-272, 1981.
- [2] H. Hermansky and N. Morgan, "RASTA processing of speech". *IEEE Trans. Speech Audio Process.* vol.2, no.4, pp. 578-589, 1994.
- [3] O. Viikki, and K. Lurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition", *Speech Communication* (25), pp. 133-147, 1998.
- [4] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop on Speaker Recognition - 2001: A Speaker Odyssey*, Crete, Greece, pp. 213-218, June 2001.
- [5] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping". *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Minneapolis, U.S.A., pp. 53-56, 2003.
- [6] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.
- [7] SuperSID webpage: <http://www.clsp.jhu.edu/ws2002/groups/supersid>.
- [8] Joachims T., "Making large-Scale SVM Learning Practical", *Advances in Kernel Methods - Support Vector Learning*. Schölkopf B., Burges C. and Smola A. (ed.), MIT-Press, 1999.
- [9] NIST Speaker Recognition Evaluation plans, <http://www.nist.gov/speech/tests/spk>
- [10] L. Ferrer, E. Shriberg, S. S. Kajarekar, A. Stolcke, K. Sonmez, A. Venkataraman, and H. Bratt. "The Contribution of Cepstral and Stylistic Features to SRI’s 2005 Nist Speaker Recognition Evaluation System", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006.