SPEAKER IDENTIFICATION USING DATA-DRIVEN SCORE CLASSIFICATION

HOCK GAN IOSIF MPORAS SAEID SAFAVI REZA SOTUDEH

School of Engineering and Technology University of Hertfordshire
(h.c.gan, i.mporas, s.safavi, r.sotudeh)@herts.ac.uk

Abstract. We present a comparative evaluation of different classification algorithms for a fusion engine that is used in a speaker identity selection task. The fusion engine combines the scores from a number of classifiers, which uses the GMM-UBM approach to match speaker identity. The performances of the evaluated classification algorithms were examined in both the text-dependent and text-independent operation modes. The experimental results indicated a significant improvement in terms of speaker identification accuracy, which was approximately 7% and 14.5% for the text-dependent and the text-independent scenarios, respectively. We suggest the use of fusion with a discriminative algorithm such as a Support Vector Machine in a real-world speaker identification application where the text-independent scenario predominates based on the findings.

Key words. speaker identification, classification, machine learning, multiple classifier system, fusion, robustness

1 Introduction

An increasing use of biometric technology can be found in areas such as the password checks in the banking sector, security checks at airports, human-computer interfaces that allow computer navigation and login checks. One of the most widely used modalities in this area is speaker recognition in voice-based biometrics. Speaker recognition biometrics offer convenience to the users as well, as they do not rely on special sensors for capturing the biometric input but rely on conventional microphones, which are available in most electronic devices. Speaker recognition is briefly categorized as speaker verification and speaker identification. In speaker verification, the system verifies or rejects a claimed identity, while in speaker identification the user is assigned to an identity from a set of speakers. Speaker identification uses voice as a unique characteristic to identify a person's identity. This task can further be classified as closed and open-set speaker identification. In closed-set speaker identification, an unknown voice input will be assigned to one of the known speaker reference templates with the highest level of similarity, based on the assumption that the unknown input belongs to one of the given set of speakers. In the open set case, the input speaker might not be assigned to any of the closed-set speakers and thus be deemed as an unknown speaker. In addition to this discrimination, the speaker identification task can also be divided into a text-dependent or text-independent task [7, 3]. While in the text-independent case [27, 29, 12, 31] the speech content is not known apriori, in the text-dependent case the users produce a pre-determined pass-phrase [18, 30]. Systems using just text dependency or independency suffer from a "liveness" problem where segments of the text can be reconstructed digitally. A text-prompted modality serves to remedy the situation. This mode constrains the user to repeat text phrases that the system chooses at random at the point of input albeit not being effective for a fixed-passphrase [2]. Text prompting requires cooperation from the claimant and suggests the use of an utterance verification engine to check the prompted phrase [34]. The state of the art technology in speaker identification is based on short-time analysis of the voice signal and post-processing by a pattern recognition algorithm. The dominating features at the speaker recognition task are the Mel frequency cepstral coefficients (MFCCs) [11, 21]. The estimated MFCC parametric representations of the speech signals are used to train speaker models. Modeling of speakers using the Gaussian Mixture Models (GMMs) [26] is widely considered to be a benchmark for modern speaker recognition. GMM uses a common statistical model (Universal Background Model or UBM) derived from a large number of speakers as a template for individual speaker models. The speaker models use a means-only adaptation process (Maximum A-Posteriori or MAP) to provide the differentiation to the UBM. Whilst the GMM represents a generative model approach, popular discriminative model approaches such as support vector machines (SVMs) have also been used [32]. Hybrid approaches using both SVM and GMM have also been trialled. The means for each feature of each Gaussian component in the GMM can be concatenated into a super-vector. SVM is then used to discriminate between the vectors [32]. Recent methods for dimensionality reduction, like i-vectors [9] offer low dimensional fixed length representation of a speech utterance that preserves the speaker-specific information. In this method, a factor analysis (FA) model is used to learn a low-dimensional sub-space from a large collection of data. A speech utterance is then projected into this sub-space and its coordinates vector is denoted as i-vector [9]. In specific experimental setups, the i-vector method has outperformed the classic GMM-UBM approach. However, GMM-UBM based modeling offers more stable results with respect to the availability of significantly large amount of training data or not, thus in this article we relied on Gaussian modeling. The use of fusion as a means of improving system performance has been exploited in many fields [19], especially in biometrics [28, 22]. For a speaker identification task, instead of a single classifier to identify which speaker the voice input belongs to, multiple classifiers are used and the results combined to provide the speaker identity [20]. Early use of fusion in the speaker identification task use simple fusion rules in the decision logic such as the minimum or maximum value of scores that for example, represent distortion measures of the speech [33]. Research in the area became more widespread and the fusion rules became more sophisticated with the building of theoretical frameworks for schemes such as the product, sum and majority voting rules [15, 16, 10, 4]. As machine-learning algorithms such as SVMs and multi-layer perceptrons (MLPs) became popular, they became alternatives for fusion rules [9, 25]. These algorithms have a progression curve of their own and new developments such as the extension of SVMs from a binary classifier to a multi-class SVM [14] make themselves available to the speaker identification task. The required performance of a single classifier suffers because of a mismatch between conditions present during training and testing [24]. Fusion addresses this in a number of ways: (1) complementary fusion where the sources of fusion are independent of each other but are combined to give a more complete view of the verification (2) cooperative fusion where information is provided by sources of fusion that would not be available from a single source alone. (3) A third form of fusion, which is competitive fusion, is based on several independent assessments of the same input using different devices to mitigate errors from inputs of singular systems. In this paper, we exploited a form of redundancy that potentially makes use of complementary and competitive fusion. Classifiers were used to produce a positive outcome for one classifier and a negative one for all other cases under ideal conditions. Under lesser conditions, all results may be negative. However, the fusion of all the results may still provide a correct result as specific patterns of negative outcomes may still lead to a good identifications. In this paper we evaluate the performance of different machine learning algorithms to be used as fusion rules in the context of text dependent and text independent speaker identification. GMM-UBM implementations of single-mode classifiers used speaker specific models to generate scores, which were combined using a classifier to estimate the identity of an unknown input speaker. The choice of fusion algorithms for evaluation were primarily by popularity but also contain both generative and discriminative classifiers [5]. The remainder of the article is organized as follows. In Section 2, we present the methodology for speaker identity selection using a classification model. In Section 3, we describe the experimental setup that was followed and in Section 4, the evaluation results are presented. Finally, in Section 5, we conclude this work.

2 Speaker Identity Selection

The traditional speaker identification decision is based on the selection of the maximum score, i.e. the speaker model with the maximum likelihood to have produced the input speech observation is selected as the detected speaker identity. However, the underlying information between the per speaker scores is not exploited in this case and especially when the difference between the maximum score and the scores of the following top speaker models is not significant. Thus, instead of applying a maximum selection criterion, we investigate the use of a classification model as a speaker identity selector. A voice sample serves as input to the system shown in Fig. 1 and after preprocessing and feature extraction, the input is processed by a set of speaker models, which correspond to a closed-set of speakers. Each model will produce a score indicating the probability or distance of the test utterance from it. These scores will be concatenated in a score vector and used by a classification algorithm to assign a speaker identity to the input test utterance. The method described forms the basis of our proposal for a fusion based speaker identification.

Let us denote the input test utterance after preprocessing and parameterization as X. A number of speaker models is used in order to estimate a score, i.e.

$$S_i = g(X, M^i) \tag{1}$$

where M^i is the model for the i_{th} speaker, with $1 \leq i \leq N$ and S_i is the corresponding score. Instead of selecting the maximum (or minimum) score, we concatenate the estimated scores in a single feature vector, $V \in R^N$, which is used as input to a fusion classifier as

$$d = f(V) \tag{2}$$

where f denotes the fusion classification model and d is the decision, i.e. the detected speaker identity. The fusion classification model will capture underlying information among the scores and in contrast to a simple maximum score selection, will provide a more robust estimation of the user's identity.

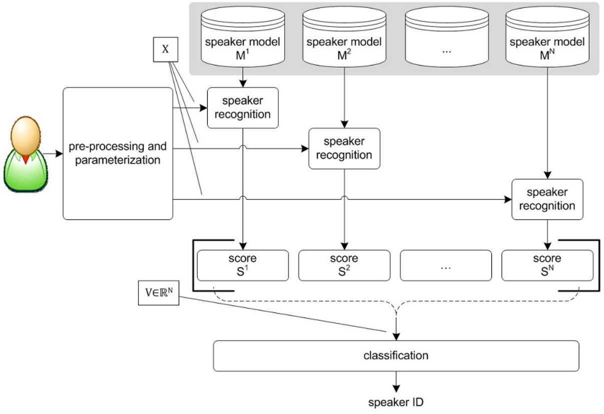

Fig. 1: Block diagram of the classification based selection of unknown speaker's identity

3 Experimental setup

Based on the framework described in Section 2, the implementation details of the experiment are described here. This includes the split of the speech corpus to form data inputs, the speaker- identification engine implementation, and the mechanics of the speaker identification.

3.1 Evaluated Speech Corpus

The speech corpus used to provide data for the experiments in this paper was RSR2015 [18], a research database established to support text-dependent verification of speaker verification systems. RSR2015 contains recordings made up by 157 male and 143 female speakers. Each speaker is put through nine recording sessions. Each recording session generates 73 utterances. The testing protocol established uses three of those sessions for training the speaker verification engines and the other six sessions for testing. The recordings used a sampling frequency of 16 kHz and used linear sampling with a 16-bit quantization resolution. However, another corpus (TIMIT [8] was used for generating the universal background model (UBM). This corpus is bigger, holding recordings from 630 speakers. The recordings had a similar specification, also having a sampling frequency of 16 kHz and a sampling resolution of 16 bits.

3.2 Speaker Identification Engine

For training speaker identification models, we relied on the GMM-UBM approach [27]. The speaker recordings had to be pre-processed first. The pre-processing provided a number of parameters that were used for training the speaker verification engine. The pre-processing compresses the speech data. An initial stage consists of removing the "silence" of the speech using a speech activity detector that discriminates speech to be retained using an energy threshold. A specific time window was used to capture the spectral characteristics of the speech that has a balance between statistical stationarity and information capture. A sliding Hamming window of 20 ms with a 10 ms overlap between successive frames was chosen. The first 19 Mel frequency cepstral coefficients (MFCCs) were captured for each frame. The differential and acceleration coefficients corresponding to the first (delta) and second (double-delta) derivatives were also generated. In total, a feature vector with 57 features were obtained. Finally, as part of the feature extraction post-processing procedures to clean up the feature data from noise and reduce the effect of handset mismatch, RASTA [13] and CMVN [36] processing was applied. The universal background model (UBM) was trained with all recordings of the 630 speakers from the TIMIT corpus. The GMM used a 128-mixture model. Once the UBM was trained, the individual speaker models were trained using recordings from the RSR2015 corpus. The training of the individual speaker models used a means only adaptation of the UBM.

3.3 Speaker identity classification selection

In this evaluation, we present a set of results using the recent RSR2015 corpus with the intention of benchmarking different classification algorithms for the selection of the speaker identity. In particular, training and trial lists (definition of speaker pairs) are designed to simulate system evaluation of two different operational modes concerning speech content, (a) text-prompted phrases and (b) text-independent engines. The first mode refers to a scenario whereby a system prompts a randomly selected phrase out of a closed subset of pass-phrases. The second mode is essentially a text-independent scenario with arbitrary enrolment and test phrases. Speaker identification is evaluated for two different circumstances, (a) and (b).

Tab. 1 shows how test resources were allocated where text dependent and text independent testing was carried out. The Mode column indicates the operational modes for testing. Under the two modes, (a) and (b), different enrolment (train) and trial lists were designed shown by the Train and Trial columns. The experiments were conducted on a subset of the male section of the recently released RSR2015 dataset. For all modes, 43 speakers were used which were identified by the range 101 to 143 as seen in the Speaker rows. (If the number of speakers were to be increased, there is an expectation for the performance to be poorer related to the data representation made for training the class and the machine learning algorithm used. For example, a Deep Neural Network would have a point where performance is optimal given the training data.) In mode (a), speakers were enrolled with 15 different pass-phrases which were identified by specific subranges in the Sentence rows. For each speaker sentences 01 to 05 were taken from session 04, sentences 06 to 10 were taken from session 01 and the rest (sentences 11 to 15) were taken from session 07. All of the 15 sentences used in the enrolment were prompted during testing.

For mode (b), the enrolment is done in similar way as the previous mode. However, the test data is exclusively different from the enrolment data. Here, the remaining 15 sentences (from 16 to 30) of the speaker were used in testing. For the classification selection stage, we relied on a number of well-known and widely used machine learning algorithms based on statistical signal modeling. Specifically, the following algorithms were used:

Tab. 1: Resource allocations during train and trial cycles of text dependent and independent testing

Protocol	Resource	Train	Trial
(a)	Speaker	101-143	101-143
Text	Sesion	1,4,7	2,3,5,6,8,9
dependent	Sentence	6-10,1-5,11-15	16-30
(b)	Speaker	101-143	101-143
Text in-	Sesion	1,4,7	2,3,5,6,8,9
dependent	Sentence	6-10,1-5,11-15	16-30

(i) support vector machines (SVM) [17] using the sequential minimal optimization implementation, (ii) multilayer perceptron neural networks (MLP) [23], (iii) C4.5 decision trees [37], (iv) k-nearest neighbors (Weka-IBk) [1]. For the implementation of these algorithms, we used the WEKA toolkit [35]. For the SVM algorithm we used the radial basis function kernel, with empirically selected parameters of C=30 and gamma=0.01. These classification algorithms were trained with the scores estimated by the speaker recognition engine described in the previous subsection (B).

4 Experimental Results

In Section 2 the framework was described and Section 3 provided the implementation details. In this section, the results obtained by a 10-fold cross validation protocol used by the fusion will be described. The metric used for the speaker identification was the identification accuracy.

The experimental results for the evaluated classification algorithms for both text-dependent and text-independent modes of speaker identification operation are tabulated in Tab. 2. The identification accuracy using the maximum selection criterion is considered as the baseline methodology for speaker identity selection. As can be seen in Tab. 2. the best performing classification algorithm for selecting the speaker identity based on the scores generated by the use of speaker GMM-UBM models was the support vector machines, both for the text-dependent and the text-independent case. Specifically

Tab. 2: Speaker identification rates (in percentages) for text-dependent and text-independent modes of operation

Method	Text-Dependent	Text-Independent
Maximum-Selection	89.08	73.93
Support Vector Machine	96.16	88.40
Multi-Layer Perceptron	94.36	84.47
Decision tree (C4.5)	66.19	49.13
k-nearest neighbour (IBK)	91.19	73.87

SVM achieved identification accuracy equal to 96.16% for the text-dependent operation mode and 88.40% for the text-independent mode. This corresponds to an absolute improvement of approximately 7% and 14.5% for text-dependent and text-independent respectively.

Both discriminative algorithms, i.e. the support vector machines and the multilayer neural network achieved significantly higher scores compared to the rest of the evaluated classification algorithms. The superiority of the SVMs arises because they perform well in higher dimensional spaces - an advantage that can be exploited since they do not suffer from the curse of dimensionality (potential increase of data that causes over-fitting). Moreover, SVMs have the advantage over other approaches, such as neural networks, in that the cost function in their training always reaches a global minimum [6]. For the rest of the evaluated algorithms, the IBk algorithm did not demonstrate any significant improvement compared to the baseline maximum-selection approach, while the C4.5 decision tree achieved worse speaker identification accuracy than the baseline. It is worth mentioning that the use of classification models as speaker identity selectors improved the performance of both the text-dependent and text-independent operation modes. We deem that this methodology can be used in real-world applications where speaker recognition systems are exposed to various types of interferences. Especially in the case of text-independent speaker identification, which is more often met in real-life applications, the presented significant improvement could be necessary.

5 Conclusion

Speaker identification accuracy when using clean speech is in general high, especially for a text-dependent scenario. However, the text-independent scenario is closer to realistic and everyday applications. In this paper, we presented an evaluation of different classification algorithms for selecting the identity of a user (speaker) based on the model scores of a closed-set of speakers. The experimental results indicated that discriminative algorithms and especially support vector machines, can significantly improve the speaker identification rate when comparing with the baseline maximum score selection criterion. Since the classification selection scheme operated equally well in the text-independent scenario, it is appropriate for use in real-world applications where robust identification of the user (speaker) is required. Also, the work done in this paper has provided a range of performances which suggests that fusion could be extended in directions other than a discrimination of identity such as speaker verification, which involves two classes as opposed to the multi-class problem in speaker identification. Within the two parallels of speaker identification and verification there is also scope to consider the robustness of the technique against spoofing attacks.

Acknowledgment

The work was partially supported in part by the European Horizon 2020 (H2020) Objective Control for Talker Verification (Octave) Project, funded by the European Commission (EC) under Grand Agreement number 647850. The authors also extend their thanks to Dr Md Sahidullah, Dr Nicholas Evans and Dr Tomi Kinnunen for their support.

References

[1] Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185

[2] Beigi, H. (2011). Speaker Recognition, Encyclopedia of Cryptography and Security, Springer, pp. 1232–1242

[3] Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Reynolds, D.A. (2004). A tutorial on text-independent speaker verification. *EURASIP journal on applied signal processing*, 2004, 430–451

[4] Bishop, C.M. (2008, June). A new framework for machine learning. In *IEEE World Congress on Computational Intelligence* (pp. 1–24). Springer Berlin Heidelberg

[5] Bouchard, G. (2007). Bias-variance tradeoff in hybrid generative-discriminative models. In *Machine Learning and Applications. ICMLA 2007. Sixth International Conference* on (pp. 124–129). IEEE

[6] Burges, C.J.C., Ben, J.I., Denker, J.S., LeCun, Y., Nohl, C.R. (1993). Off line recognition of handwritten postal words using neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04), 689–704

[7] Campbell, J.P. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9), 1437–1462

[8] Campbell, J.P., Reynolds, D A. (1999, March). Corpora for the evaluation of speaker recognition systems. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference* on (Vol. 2, pp. 829–832). IEEE

[9] Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788–798

[10] Damper, R.I., Higgins, J.E. (2003). Improving speaker identification in noise by subband processing and decision fusion. *Pattern Recognition Letters*, 24(13), 2167–2173

[11] Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2), 254–272

[12] Ganchev, T., Siafarikas, M., Mporas, I., Stoyanova, T. (2014). Wavelet basis selection for enhanced speech parametrization in speaker verification. *International Journal of Speech Technology*, 17(1), 27–36

[13] Hermansky, H., Morgan, N. (1994). RASTA processing of speech. *IEEE transactions on speech and audio processing*, 2(4), 578–589

[14] Hsu, C.W., Lin, C.J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2), 415–425

[15] Kittler, J., Hatef, M., Duin, R.P., Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3), 226–239

[16] Kuncheva, L.I., Alpaydin, E. (2007). Combining Pattern Classifiers: Methods and Algorithms, *IEEE Transactions on Neural Networks*, 18(3), 964–964

[17] Kung, S.Y. (2014). *Kernel methods and machine learning*. Cambridge University Press. pp. 341–342

[18] Larcher, A., Lee, K.A., Ma, B., Li, H. (2014). Text-dependent speaker verification: Classifiers,

databases and RSR2015. *Speech Communication*, 60, 56–77

[19] Mitchell, H. B. (2007). *Multi-sensor data fusion: an introduction*. Springer Science & Business Media

[20] Monte-Moreno, E., Chetouani, M., Faundez-Zanuy, M., Sole-Casals, J. (2009). Maximum likelihood linear programming data fusion for speaker recognition. *Speech Communication*, 51(9), 820–830

[21] Najafian, M., Safavi, S., Weber, P., Russell, M. (2016). Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems. ODYSSEY

[22] Nandakumar, K., Jain, A. K. (2008, September). Multibiometric template security using fuzzy vault. In *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference* on (pp. 1–6). IEEE

[23] Pal, S.K., Mitra, S. (1996). Noisy fingerprint classification using multilayer perceptron with fuzzy geometrical and textural features. *Fuzzy sets and systems*, 80(2), 121–132

[24] Ramachandran, R.P., Farrell, K.R., Ramachandran, R., Mammone, R.J. (2002). Speaker recognition–general classifier approaches and data fusion methods. *Pattern Recognition*, 35(12), 2801–2821

[25] Raudys, Š. (2006). Trainable fusion rules. I. Large sample size case. *Neural Networks*, 19(10), 1506–1516

[26] Reynolds, D.A., Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on speech and audio processing*, 3(1), 72–83

[27] Reynolds, D.A., Quatieri, T.F., Dunn, R.B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, 10(1), 19–41

[28] Safavi, S., Gan, H., Mporas, I., Sotudeh, R. Fraud Detection in Voice-based Identity Authentication Applications and Services. In *The IEEE International Conference on Data Mining series (ICDM)*, 2016

[29] Safavi, S., Hanani, A., Russell, M., Jancovic, P., Carey, M.J. (2012). Contrasting the effects of different frequency bands on speaker and accent identification. *IEEE Signal Processing Letters*, 19(12), 829–832.

[30] Safavi, S., Jancovic, P., Russell, M.J., Carey, M.J. (2013). Identification of gender from children's speech by computers and humans. In *INTERSPEECH* (pp. 2440–2444)

[31] Safavi, S., Najafian, M., Hanani, A., Russell, M.J., Jancovic, P., Carey, M.J. (2012). Speaker Recognition for Children's Speech. In *INTERSPEECH* (pp. 1836–1839)

[32] Safavi, S., Russell, M.J., Jancovic, P. (2014, September). Identification of age-group from children's speech by computers and humans. In *INTERSPEECH* (pp. 243–247)

[33] Soong, F.K., Rosenberg, A.E., Juang, B.H., Rabiner, L.R. (1987). Report: A vector quantization approach to speaker recognition. *AT&T technical journal*, 66(2), 14–26

[34] Sukkar, R.A., Lee, C.H. (1996). Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(6), 420–429

[35] Witten, I.H., Frank, E., Hall, M.A. (20011). Embedded Machine Learning. *Data Mining: Practical ma-*

chine learning tools and techniques. Elsevier BV, pp. 531–538

[36] Viikki, O., Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1), 133–147

[37] Zhang, S., Zhu, L. (2013). A packet classification algorithm based on improved decision tree. *Journal of Networks*, 8(12), 2864–2871