Evaluation of Classification Algorithms for Text Dependent and Text Independent Speaker Identification

Iosif Mporas, Saeid Safavi, Hock Chye Gan and Reza Sotudeh
Division of Electronics and Communications, School of Engineering and Technology
University of Hertfordshire
Hatfield, AL10 9AB, Hertfordshire
{i.mporas, s.safavi, h.c.gan, r.sotudeh}@herts.ac.uk

***Abstract*—We present a comparative evaluation of different classification algorithms for the task of speaker identity selection based on GMM-UBM speaker identification scores. The performance of the evaluated classification algorithms was examined in both text-dependent and text-independent operation modes for speaker identification. 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.**

Keywords— speaker identification; classification; machine learning.

I. INTRODUCTION

Biometric technology is used over the last years to a number of applications, such as security access control to physical places, secure login to computer systems and mobile devices, online banking, personalized human-machine interfaces etc. One of the most widely used modalities in this area is voice-based biometrics and particularly speaker recognition. Speaker recognition biometrics offer convenience to the users as well as they do not rely on special sensors for capturing the biometric input rather than on conventional microphones, which are available in most electronic devices. Speaker recognition is briefly categorized to 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 to 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 be assigned to not belong to any of the closed set speakers and thus is assigned as an unknown one. Except this discrimination, speaker identification task can also be divided to text-dependent and text-independent [1-2]. While in the text-independent case [3-6] the speech content is apriori known, in the text-dependent case the users pronounce a pre-determined pass-phrase [7-9]. The pass-phrases are either unique or prompted by the system, e.g. in a screen.

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) [10-11]. The estimated MFCC parametric representations of the speech signals are used to train speaker models. Modeling of speakers using the Gaussian Mixture Models (GMMs) [12] is widely considered to be a benchmark for modern speaker recognition. GMM technology has proved to perform well using universal background models (UBMs) trained from a large number of background speakers and maximum a-posteriori (MAP) adaptation or means-only adaptation of the UBM to speaker specific data. Except GMMs, other approaches such as support vector machines (SVMs) have also successfully been used in the task of speaker recognition [13]. SVMs have also been used in combination with GMMs by concatenating the means of the Gaussian components of the GMMs to super-vectors and afterwards apply discriminative classification on them [13]. Recent methods for dimensionality reduction, like i-vectors [14] 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 subspace and its coordinates vector is denoted as i-vector [14]. In specific experimental setups, 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.

In this paper we evaluate the performance of different machine learning algorithms for classification on the tasks of text dependent and text independent speaker identification. The scores produced by speaker specific models using the GMM-UBM approach are used as input to a classifier to estimate the identity of the unknown input speaker.

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.

II. 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 user is providing the system with a voice sample and after pre-processing and feature extraction the input is processed by a set of speaker models which correspond to a close-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 in order to assign a speaker identity to the input test utterance. The proposed methodology for score fusion based speaker identification is illustrated in Fig. 1.

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

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

where M^i is the model for the i-th speaker, with $1 \le i \le N$, 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, $V \in \Re^N$, which is used as input to a fusion classifier as

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

where f denotes the fusion classification model and d is the decision, i.e. the detected speaker identity.

We deem the fusion classification model will capture

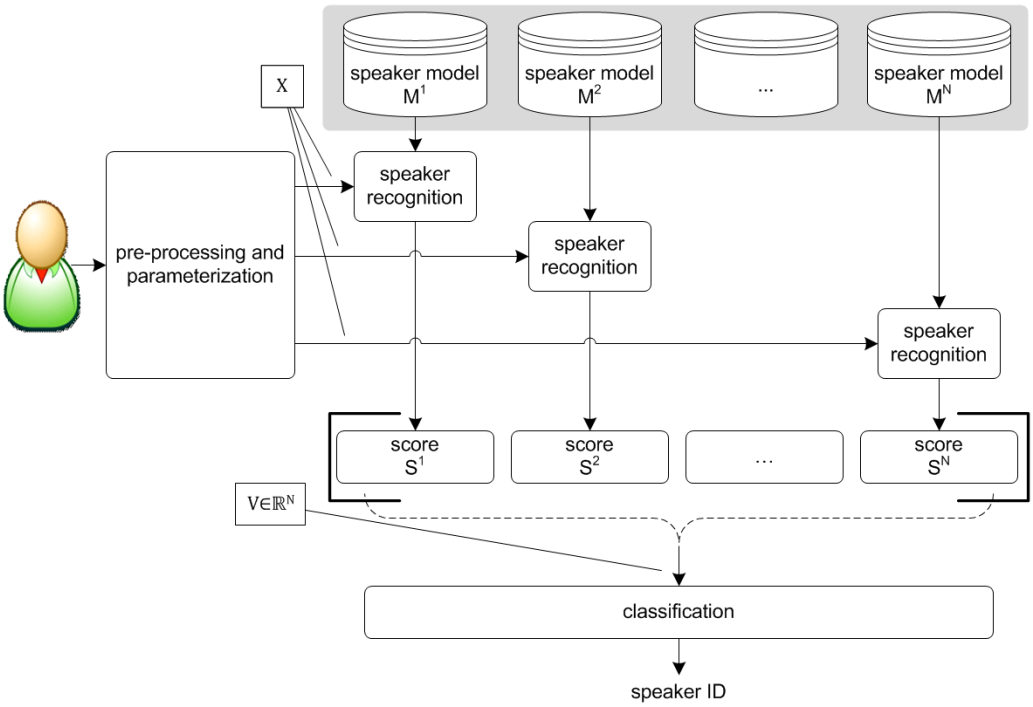

Fig. 1. Block diagram of the classification based selection of unknown speaker's identity.

underlying information between the scores and in contrast to a simple maximum score selection, will provide more robust estimation of the user's identity.

III. Experimental Setup

The experimental setup for the evaluation of the methodology described in Section 2, is presented here. Specifically, we describe the dataset used in the evaluation, the setup of the speaker identification engine and the setup of the classification based stage for the selection of the input user/speaker identity.

A. Evaluated Speech Corpus

In this evaluation RSR2015 speech corpus is used [9]. RSR2015 consists of recordings from 300 speakers (157 males, 143 females). For each speaker, there are 3 enrolment sessions of 73 utterances each and 6 verification sessions of 73 utterances each. In total there are 657 utterances distributed in 9 sessions per each speaker. The sampling frequency of the speech recordings is 16 kHz and the speech samples are stored with analysis equal to 16 bits. In addition to RSR2015, we used TIMIT [15] for training a universal background model. TIMIT consists of recordings of 630 speakers, sampled at 16 kHz with resolution analysis equal to 16 bits per sample.

B. Speaker identification engine

For training speaker identification models we relied on the GMM-UBM approach [13]. Each voice input was initially pre-processed and parameterized. During pre-processing an energy-based speech activity detector was applied to retain the speech only parts. The speech input was frame blocked using a time shifting Hamming window of 20 milliseconds length with 10 milliseconds time shift between successive frames. For each frame the first 19 Mel frequency cepstral coefficients (MFCCs) were estimated, which were further expanded to their first (delta) and second (double-delta) derivatives, thus resulting to a feature vector of length equal to 57. In order to reduce the effect of handset mismatch and make the features more robust RASTA [16] and CMVN processing were applied to the MFCC features.

The universal background model (UBM) was built by a mixture of 128 Gaussian distributions and was trained using all utterances from 630 speakers from TIMIT. For each of the speakers of the RSR2015 database we applied means only adaptation on the UBM model, using the speaker-specific enrollment data.

C. Speaker identity classification selection

In this evaluation we present a set of results on the recent RSR2015 corpus intended for 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 configurations concerning speech content, (a) text-prompted phrases and (b) text-independent engines. The first protocol refers to a scenario whereby a system prompts a randomly selected phrase out of a close subset of pass-phrases. The second scenario is essentially a text-independent scenario with arbitrary enrolment and test phrases. Speaker identification is evaluated for two different circumstances, (a) and (b).

To assess the performance for two protocols, different enrolment and trial lists were designed. The experiments are conducted on a subset of male section of recently released RSR2015 dataset. For all three protocols 43 speakers are used. In the protocol (a), speakers are enrolled with 15 different pass-phrases. For each speaker sentences 01 to 05 are taken from session 04, sentence 06 to 10 are taken from session 01 and rest sentence 11 to 15 are taken from session 07. Out of the 15 sentences used in the enrolment are prompted during testing.

For protocol (b), the enrolment is done in similar way as the previous protocol. But the test data is exclusively different from the enrolment data. Here, the rest 15 sentences (from 16 to 30) are used in testing.

For the classification selection stage we relied on a number of well known and widely used in the area of statistical signal modeling machine learning algorithms. Specifically, the following algorithms were used: (i) support vector machines (SVM) [17] using the sequential minimal optimization implementation, (ii) multilayer perceptron neural networks (MLP) [18], (iii) C4.5 decision trees (C4.5) [19], (iv) k-nearest neighbors (IBk). For the implementation of these machine learning algorithms for classification we used the WEKA toolkit [20]. For the SVM algorithm we used radial basis function kernel, with empirically selected parameters C=30 and gamma=0.01. These classification algorithms were trained with the scores estimated by the speaker recognition engines described in the previous subsection.

IV. Experimental Results

The classification based methodology for speaker identification presented in Section 2 was evaluated based on the experimental setup described in Section 3. For all evaluations a 10-fold cross validation protocol were applied. The performance of the speaker identification task was evaluated in terms of identification accuracy.

The experimental results for the evaluated classification algorithms for both text-dependent and text-independent protocols of speaker identification operation are Tabulated in Table 1. The identification accuracy using the maximum selection criterion is considered as the baseline methodology for speaker identity selection.

As can be seen in Table 1, the best performing classification algorithm for selecting the speaker identity based on the scores of the speaker GMM-UBM models is the support vector machines, both for the text-dependent and the text-independent case. Specifically SVM achieved identification accuracy equal to 96.16% for the text-dependent operation protocol and 88.40% for the text-independent protocol. This corresponds to an absolute improvement of approximately 7% and 14.5% for TD and TI respectively.

Both discriminative algorithms, i.e. the support vector machines and the multilayer neural network achieved

significantly higher scores comparing to the rest evaluated classification algorithms. The superiority of the SVMs is due to the fact that perform well in higher dimensional spaces since they do not suffer from the curse of dimensionality. Moreover, SVMs have the advantage over other approaches, such as neural networks, etc, that their training always reaches a global minimum [21].

TABLE I. 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
SVM	96.16	88.40
MLP	94.36	84.47
C4.5	66.19	49.13
IBk	91.19	73.87

For the rest of the evaluated algorithms, the IBk algorithm did not demonstrate any significant improvement comparing 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 both for the text-dependent and text-independent operation modes. We deem that this methodology can be used to real-world applications were speaker recognition systems are exposed to various types of interferences. Especially for the case of text-independent speaker identification, which is more often met in real-life applications the presented significant improvement could be a must.

V. 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 the 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 to be used to real-world applications were robust identification of the user/speaker is required.

Acknowledgment

This work was partially supported by the H2020 OCTAVE Project entitled “Objec-tive Control for TAlker VErification” funded by the EC with Grand Agreement num-ber 647850. The authors would like to thank Dr Md Sahidullah, Dr Nicholas Evans and Dr Tomi Kinnunen for their support in this work.

References

[1] Joseph P. Campbell, Jr., Senior Member, IEEE, “Speaker Recognition: A Tutorial”, Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, September 1997.

[2] F. Bimbot, et al., “A tutorial on text-independent speaker verification,” EURASIP J. Appl. Signal Process., no. 1, pp. 430–451, 2004.

[3] Douglas A. Reynolds, Thomas F. Quatieri, Robert B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing, Volume 10, Issues 1–3, January 2000, Pages 19-41, ISSN 1051-2004,

[4] S. Safavi, A. Hanani, M. Russell, P. Jancovic and M. J. Carey, "Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identification," in IEEE Signal Processing Letters, vol. 19, no. 12, pp. 829-832, Dec. 2012.

[5] Safavi, Saeid, Maryam Najafian, Abualsoud Hanani, Martin J. Russell, Peter Jancovic, and Michael J. Carey. "Speaker Recognition for Children's Speech." In INTERSPEECH, pp. 1836-1839. 2012.

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

[7] H. Aronowitz, R. Hoory, J. Pelecanos, D. Nahamoo, "New Developments in Voice Biometrics for User Authentication", in Proc. Interspeech, 2011.

[8] M. Hébert, M. Sondhi, Y. Huang, “Text-Dependent Speaker Recognition”, Book Section, Springer Handbook of Speech Processing, pp. 743-762, 2008

[9] Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li, “Text-dependent speaker verification: Classifiers, databases and RSR2015, Speech Communication, Volume 60, May 2014, Pages 56-77, ISSN 0167-6393, http://dx.doi.org/10.1016/j.specom.2014.03.001.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, Aug 1980.

[11] S. Furui, "Cepstral analysis technique for automatic speaker verification," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 2, pp. 254-272, Apr 1981.

[12] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," in IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, Jan 1995.

[13] Campbell, W.M., Sturim, D.E. and Reynolds, D.A., 2006. Support vector machines using GMM supervectors for speaker verification. Signal Processing Letters, IEEE, 13(5), pp.308-311.

[14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 4, pp. 788 - 798, May 2010.

[15] J. P. Campbell and D. A. Reynolds, “Corpora for the evaluation of speaker recognition systems,” in Proc. Of ICASSP’99, 1999, vol. 2, pp. 829-832.

[16] H. Hermansky and N. Morgan, "RASTA processing of speech," in IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, pp. 578-589, Oct 1994.

[17] Schölkopf B, Burges CJ. Advances in kernel methods: support vector learning. MIT press; 1999.

[18] Pal SK, Mitra S. Multilayer perceptron, fuzzy sets, and classification. Neural Networks, IEEE Transactions on. 1992 Sep;3(5):683-697.

[19] J. R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research, 4:77-90, 1996.

[20] Witten, I.H., Frank, E., Hall, M.A., Data Mining: Practical machine learning tools and techniques, 3rd Edition. Morgan Kaufmann, San Francisco, (2011).

[21] Burges C., “A tutorial on Support Vector Machines for Pattern Recognition”, Data Mining and Knowledge Discovery, Vol. 2, Number 2, p.121-167, Kluwer Academic Publishers, 1998.