# A Unified Approach for Audio Characterization and its Application to Speaker Recognition

*Luciana Ferrer*[1], *Lukas Burget*[1], *Oldrich Plchot*[2], *Nicolas Scheffer*[1]

[1]Speech Technology and Research Laboratory, SRI International, California, USA
[2]Brno University of Technology, Speech@FIT, Czech Republic
{lferrer,burget,scheffer}@speech.sri.com
iplchot@fit.vutbr.cz

## Abstract

Systems designed to solve speech processing tasks like speech or speaker recognition, language identification, or emotion detection are known to be affected by the recording conditions of the acoustic signal, like the channel, background noise, reverberation, and so on. Knowledge of the nuisance characteristics present in the signal can be used to improve performance of the system. In some cases, the nature of these nuisance characteristics is known a priori, but in most practical cases it is not. Most approaches used to automatically detect the characteristics of a signal are designed for a specific type of effect: noise, reverberation, language, type of channel, and so on. We propose a method for detecting the audio characteristics of a signal in a unified way, based on iVectors. We show results for the detector itself and for its use as metadata during calibration of a state-of-the-art speaker recognition system based on iVectors extracted from Mel frequency cepstral coefficients. Results show relative gains in equal error rate of up to 15% in a variety of recording conditions.

## 1. Introduction

The performance of systems designed to solve speech processing tasks like automatic speech recognition, speaker recognition, language identification, and emotion detection can be severely affected by the recording conditions of the acoustic signal. These conditions can include the effect of the channel, background noise, reverberation, the mood of the speaker, and other characteristics that are not the one that needs to be detected by the system (see [1, 2, 3, 4, 5] for some examples of the effect of different conditions on speech processing tasks). For example, language variations are a nuisance when detecting speaker identity, while speaker variations are a nuisance when detecting language. Being able to detect the characteristics of the signal that might affect the performance of the classifier of interest can lead to improvements in the system's performance, since this knowledge can be used to predict optimal parameters of the system under the particular conditions detected (see, for example, [4, 6, 7, 8]).

For some applications, the nature of the nuisance characteristics in the signals is known a priori, but in most practical cases it is not. For those cases, different detectors have been designed that can automatically estimate the nuisance characteristics present in the audio, like the signal-to-noise ratio (SNR) (e.g., [9]), reverberation time (RT) (e.g., [10]), language being spoken (e.g., [11]), and so on. These approaches are appropriate when it is known that the signals will contain only certain types of nuisance variability. For example, if the signal is known to contain only babble noise at different SNRs, then only an SNR detector is needed to detect the nuisance characteristics present in the signal. If both noise and reverberation are likely to be present, then two detectors need to be run under this paradigm.

This approach cannot easily handle new types of variability since new detectors have to be designed for those cases. For example, if the data is now likely to also contain channel variability apart from noise and reverberation, a new detector to estimate the channel has to be designed from scratch, since the standard SNR and RT detectors cannot be used to detect channel types.

We propose a unified approach for detecting the characteristics of an acoustic signal based on low-dimensional vectors extracted from the signal. In its basic form, the approach extracts for each signal iVectors containing information about all characteristics present in the signal, as those currently used in many state-of-the-art speaker recognition systems [12]. It then models the distribution of these vectors for different audio classes present in the training data. Given a new signal, a vector of posterior probabilities for the classes found in training is generated. This approach can inform the user when a test signal's characteristics are not well represented within the training data – a situation that would result in unpredictable system performance. Finally, any new type of variability that might be faced by the system can be easily handled by the proposed approach simply by adding training data with that kind of variability and, perhaps, adding a class to be predicted.

When the proposed system is used as input to a speaker recognition system, the iVectors used for both systems can be the same. In such case, the characteristics that affect the performance of the speaker recognition system should be easily detected by the audio characterization system since these nuisance characteristics can only affect the speaker recognition performance if they are encoded in the iVectors. Hence, the posteriors predicted by the audio characterization system should be particularly useful if we want to compensate for the effect that these nuisance characteristics have in speaker recognition performance.

## 2. Audio Characterization System

The detection of the audio characteristics of a speech signal can be a goal of its own, or, as mentioned earlier, it can be just another step of a bigger speech processing system aimed at solving a different problem. In this case, the system adapts to the detected characteristics, with the goal of improving its overall performance. Here, we describe the proposed audio characterization system independently of its intended use. The system

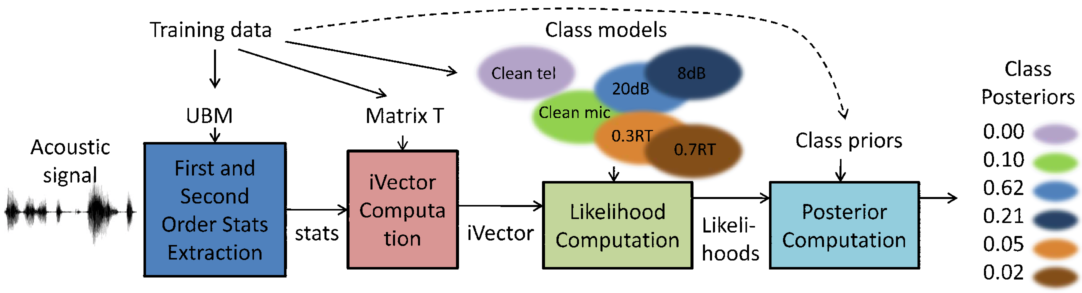


Figure 1: Schematic figure for the proposed audio characterization system. The training stage generates the UBM, T matrix for iVector extraction and the set of class models. The class priors can be obtained from the training data or set arbitrarily based on the expected distribution of classes on the test data. Given a test signal, the first- and second-order statistics are obtained, and, based on those, the corresponding iVector is estimated. Finally, likelihoods for each of the class models are obtained and converted into posteriors using the given priors.

is based on the iVector idea [12], on which most state-of-the-art speaker recognition systems are currently based [13]. We also present results for the audio characterization task itself on a database containing different types of variability. Section 3 shows one application of the proposed system for improving speaker recognition results.

### 2.1. System Description

Figure 1 shows a schematic figure of the proposed audio characterization system. The different steps and inputs in this figure are explained below.

#### *2.1.1. Universal Background Model*

In our proposed framework for audio characterization, features extracted over the acoustic signal are modeled using a Gaussian mixture model (GMM) as commonly done in speaker recognition systems. The features can be different types including, but not restricted to, the commonly used Mel frequency cepstral coefficients (MFCCs). Given a set of held-out data, a universal background model (UBM) represented by a GMM is trained to model the overall distribution of these features. In speaker recognition, only features extracted over speech regions (as detected by a voice activity detection algorithm) are used throughout the process. Our initial implementation of the audio characterization system uses this same restriction, even though, in our case, using features over pauses could improve performance for certain types of variability. This is a topic for research in the future.

#### *2.1.2. iVector Extraction*

We represent each signal using the iVector approach. An iVector is a single fixed-length vector of relatively low dimension that should contain all relevant information in the signal. In our case, the relevant information comprises all types of nuisance variability that the system is trying to detect. This iVector is extracted, in our current implementation, using a total variability subspace model [12] given by

$$M = m + Tw, \tag{1}$$

where $m$ is the UBM supervector, composed by the concatenation of the means from all Gaussians in the UBM, $T$ is a low-rank rectangular matrix estimated using held-out data, and $w$ is a random vector having a standard normal distribution. $M$ is the supervector corresponding to the observed sample. The vector $w$ is a hidden variable. Its posterior distribution is estimated to maximize the likelihood of the feature vectors extracted from the signal given the above model (that is, the UBM with the new means given by $M$). The mean of the estimated posterior distribution of $w$ is then taken to be the desired iVector for the signal.

#### *2.1.3. Training the Classifier*

Given the iVectors for the training data (which can coincide with the held-out data used for training the UBM and the $T$ matrix), the classifier can be trained. For this, each iVector is labeled with one of $N$ nuisance characteristic labels. The choice of $N$ and the labels themselves depends on what type of data is available for training and the purpose for which the system is being designed. In Figure 1 six classes are shown, corresponding to clean telephone signal, clean microphone signal, noisy signal with 8 dB SNR, noisy signal with 20 dB SNR, signal with reverberation of 0.3 RT and signal with reverberation of 0.7 RT.

The labeled iVectors are then used to train a model for each nuisance class. In our implementation, each class is represented by a single Gaussian where the mean is estimated as the mean of all iVectors in the class. The covariance of all Gaussians is forced to be identical and is estimated by subtracting the mean of the corresponding class from each iVector and estimating the covariance of the resulting class-centered iVectors. The resulting Gaussian models are shown schematically in Figure 1 for a two-dimensional iVector.

#### *2.1.4. Classifying an Acoustic Signal*

Given a new waveform for which the audio characteristics need to be estimated, the first step is to obtain the first and second order statistics with respect to the UBM. These statistics are then used to estimate the iVector corresponding to the signal [12]. Next, the likelihood of each of the models corresponding to the different classes given the iVector are obtained. These likelihoods can then be transformed into posterior probabilities using Bayes rule and a set of priors which can be either estimated from the training data, taken as uniform, or arbitrarily defined based on prior belief of what the distribution of classes will be during testing.

The resulting vector of posteriors can be used directly as a representation of the characteristics of the audio found in the

| Class | | Description | #Train | #Test |
|---|---|---|---|---|
| **clean telephone** | | Waveforms collected from telephone conversations recorded over telephone channels | 67822 | 11878 |
| **clean microphone** | | Waveforms collected from telephone conversations and interviews recorded over different types of microphones | 11018 | 11034 |
| **noisy** | 8dB SNR | Clean microphone signals with added noise at 8 dB SNR. | 1830 | 2176 |
| | 15dB SNR | As above but at 15 dB SNR. | 1830 | 2176 |
| | 20dB SNR | As above but at 20 dB SNR. | 1830 | 2176 |
| **reverberated** | 0.3 RT | Clean microphone signals distorted with reverberation at an RT of 0.3. | 1830 | 2176 |
| | 0.5 RT | As above but with an RT of 0.5. | 1830 | 2176 |
| | 0.7 RT | As above but with an RT of 0.7. | 1830 | 2176 |

Table 1: Classes used in the audio characterization system. Both the clean telephone and clean microphone sets contain waveforms that are not necessarily completely clean but might have some background noise and channel distortion. We call them clean to differentiate them from the noisy and reverberated ones.

data. This way, if a sample contains a mix of two or more characteristics only considered as separate classes for training, the posteriors for those classes should all be large. In our example, if a sample contains both noise at around 8 dB and reverberation at around 0.3 RT, then the posteriors corresponding to those two classes should both be large[1]. Alternatively, if a decision about the sample's class has to be made, the class with the largest posterior can be selected. Finally, note that, depending on how the output of the system will be used, the vector of likelihoods can be kept as it is, without converting it to a vector of posteriors.

### 2.2. Results

To test the audio characterization system for its ability to predict the same classes with which it was trained, we use the PRISM evaluation set described in detail in [1]. We use the training data for training the classifier and all sessions used in speaker recognition trials in that database for testing. These two sets are disjoint, in the sense that they do not have any speakers in common.

The training set is composed of data from Fisher 1 and 2, Switchboard phases 2 and 3 and Switchboard cellphone phases 1 and 2, along with data from all National Institute of Standards and Technology (NIST) speaker recognition evaluations (SRE) from 2004 to 2008. Simulated noisy and reverberated signals were also added to the training set, starting from a set of held-out lavalier mic data from SRE08. To create the noisy signals, real waveforms from FreeSound.org [14] containing cocktail noise collected in bars, cafeterias, offices, and airports are added to the signal using the FaNT tool [15]. The reverberation effect is added to the clean waveform with the rir tool [16] using different parameters for the room size, microphone and speaker location, wall, floor and ceiling reflection coefficients, and so on.

The test set comprises data from SRE05, SRE06, SRE08 and SRE10. As mentioned above, no speakers are shared between the training and the test sets. The test set also contains simulated noisy and reverberated signals created from lavalier mic data from SRE08 and SRE10. The noise waveforms added to the signal and the reverberation parameters used in the test set are different from those used in the training set to avoid testing on highly matched cases.

[1] This hypothesis has not yet been confirmed in practice, since the test data used in our experiments contains similar kinds of characteristics as those found in our training data. Confirming this hypothesis is part of our future work.

Both training and test sets are composed of signals with the same type of nuisance characteristics. In our experiments, we divide these characteristics into eight different classes. Table 1 lists the eight classes with their characteristics and the number of signals available for each of them in the training and test sets.

Table 2 shows the confusion matrix obtained with our proposed audio characterization system on the test data described. In this case, MFCCs are used as input to the system. Details on parameters used for extraction of the iVectors are given in Section 3.2. To compute the confusion matrix we assign to each sample the class with the highest posterior as estimated by the system. The rows of the confusion matrix have been scaled to add up to 100, to facilitate comparisons. A confusion matrix for a perfect classifier would have 100 in the diagonal and 0s elsewhere. In this case, we see that both the clean telephone data and clean microphone data are detected very consistently, with microphone data being confused 11% of the time with the cleanest noisy condition. This is very reasonable given that the noisy data was created by adding noise to clean microphone data.

Noisy signals are also detected very consistently as being noisy, even though there is some confusion across SNR levels. This confusion happens mostly for some noise signals. For example, most signals for which 8 dB or 15 dB noise is confused with 20 dB noise correspond to the same two noise signals (added to different clean signals). That is, these two noise signals are such that, when added to clean signals, they do not result in a significant degradation of the iVectors.

In the case of reverberated signals, the detection is very unreliable for the lower RT values. In fact, most reverberated signals with the two lowest RT values are detected either as microphone signals or noisy signals with 15 or 20 dB SNR. We believe that this is the case because the training database contains only three kinds of reverberation for each RT level, which are, in turn, different from those used in testing. It is likely that the small sample of reverberation types available for training would result in a lack of generalizability of the system to unseen reverberation types.

The results presented in this section show the performance of the system as a nuisance prediction system. Nevertheless, our ultimate goal in this paper is not to predict the kind of nuisance present in the signal but to use this information to improve the performance of a speaker recognition system. Section 3 shows one way in which the vector of posteriors generated by the system can be used for this purpose.

| | Detected class | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| True class | mic | phn | rt0.3 | rt0.5 | rt0.7 | snr08dB | snr15dB | snr20dB |
| mic | 83.19 | 4.16 | 0.19 | 0.67 | 0.32 | 0.2 | 0.05 | 11.21 |
| phn | 0 | 99.73 | 0 | 0 | 0.01 | 0.07 | 0.03 | 0.16 |
| rt0.3 | 22.71 | 1.1 | 0 | 0.17 | 0 | 4.92 | 31.53 | 39.58 |
| rt0.5 | 35.51 | 0.08 | 5.42 | 0.17 | 0 | 3.39 | 44.92 | 10.51 |
| rt0.7 | 3.64 | 0 | 0.17 | 45.85 | 50 | 0.34 | 0 | 0 |
| snr08dB | 1.02 | 1.53 | 0 | 0 | 0 | 47.54 | 26.36 | 23.56 |
| snr15dB | 1.78 | 1.69 | 0 | 0 | 0 | 4.49 | 48.14 | 43.9 |
| snr20dB | 1.95 | 2.29 | 0 | 0 | 0 | 0.93 | 12.03 | 82.8 |

Table 2: Confusion matrix when using the proposed audio characterization system for detection of the classes found in training on a held-out set of signals.

# 3. Application to Calibration of Speaker Recognition Systems

Speaker recognition is the task of deciding whether the speaker present in a test signal is the same as the speaker present in a certain enrollment signal. Speaker recognition samples, comprised of these two signals, are usually called *trials*.

Adaptation to the detected audio characteristics can occur at many different stages of a speaker recognition system. In this paper, we choose to do the adaptation at the final stage, taking the scores produced by the system and calibrating them with a function that depends on the posteriors generated by the audio characterization system.

## 3.1. Calibration Using Metadata

In previous work we have proposed the use of metadata (or high-level information) about the signal to affect the parameters of the fusion or calibration stage [6, 7]. In both of those papers, the metadata was required to be discrete. Since the audio characterization posteriors are continuous measures and we believe that valuable information would be lost if we discretized it (by, for example, choosing the class with the highest posterior), in this work we choose to use the approach implemented by the Bosaris toolkit [17]. In this approach, the calibrated log-likelihood-ratio output for a trial among signals $i$ and $j$ is

$$\ell_{ij} = \alpha + \beta s(i,j) + \mathbf{q}(i)'\mathbf{W}\mathbf{q}(j), \quad (2)$$

where $s(i,j)$ is the score generated by the system for the trial and $\mathbf{q}(i)$ and $\mathbf{q}(j)$ are vectors of metadata for the two signals in the trial, where the vector is augmented by appending a 1. The fusion parameters are the offset $\alpha$; weight $\beta$; and the bilinear combination matrix $\mathbf{W}$, constrained to be symmetric. Note that, in this functional form, the metadata affects the final score only through a bias. It does not affect the weight given to the scores. While this might be suboptimal, it is a good first approach for testing the effect of the audio characterization posteriors when used as metadata for calibration.

The parameters $\alpha$, $\beta$, and $\mathbf{W}$ are trained through maximization of a cross-entropy objective function (as described in [18]) using cross-validation on trials from all conditions available in the PRISM evaluation set described below. For this, the speakers in the trials are split in two lists. Given one of these lists, the trials involving only these speakers are used for training the calibration parameters. These parameters are then used to calibrate scores for the trials involving only speakers from the other list. The process is then reversed to get scores on the first set of trials. The concatenated set of scores is then used to compute the final performance measures shown in this paper. This procedure discards all trials involving a speaker from one of the lists and a speaker from the other list, reducing the number of impostor trials to around half of those available in the original PRISM evaluation set.

## 3.2. Experimental Setup

We test our proposed approach on the PRISM speaker recognition evaluation database [1] also used to train and test the audio characterization system. Several sets are defined within the PRISM database aimed at assessing the effect of different types of nuisance variability on speaker recognition systems. Different conditions are defined within each of these sets to allow for comparisons. For the results in this paper we focus on a small subset of conditions that are a good indicator of the effect of each type of nuisance variability on the system's performance. Since our system is symmetric with respect to the two signals involved in the trial (enrollment and testing), we define the conditions by specifying the characteristics of the two signals in the trials, regardless of whether they are used for enrollment or testing. The conditions are the following:

- **telp:** English telephone calls over telephone channel for both signals in the trial. Corresponds to condition "tel vs tel, phn vs phn" in Table VI in [1].
- **tela:** English telephone calls over either telephone or microphone channels for both signals in the trial. Corresponds to condition "tel vs tel, all vs all" in Table VI in [1].
- **int:** English interviews over microphone channels for both signals in the trial. Corresponds to condition "int vs int, mic vs mic" in Table VI in [1].
- **vel:** Normal vocal effort English conversations versus normal, low and high vocal effort English conversations. Corresponds to condition "normal vs all" in Table V in [1].
- **lan:** Trials where both signals are telephone conversations in the same language, which can be either English, Chinese, Russian, Arabic or Thai. Corresponds to condition "lang X vs lang X" in Table IV in [1].
- **noi:** Clean and noisy microphone interview signals with different SNR levels tested against each other. Corresponds to condition "all vs all" in Table II in [1].
- **rev:** Clean and reverberated microphone interview signals with different RTs tested against each other. Corresponds to condition "all vs all" in Table III in [1].

We show results for calibration of a speaker recognition system based on MFCC features. Nineteen MFCCs and the energy with appended deltas and double deltas are used as features. iVectors of dimension 600 are then extracted as described in Section 2.1.2. The universal background model used to obtain the statistics that are the input to the iVector extractor is a gender-dependent 2048-component diagonal-covariance model. Linear discriminant analysis (LDA) is used to reduce dimensionality of the iVectors from 600 to 250. The resulting iVectors are then mean-normalized using the mean over the LDA training data. Finally, the iVectors are length-normalized as explained in [19]. For each verification trial, the resulting iVectors are compared by means of a probabilistic linear discriminant analysis (PLDA) [20] model to obtain verification scores, where the variability present in the signal is described by full rank matrices of eigenvoice and eigenchannel bases.

The training data used by the system to train the UBM, iVector extractor, LDA, and PLDA is the same as the one used to train the audio characterization system. For PLDA and LDA, only data from speakers with at least six sessions is used. A small set of 66 held-out speakers from SRE10 is added to the LDA/PLDA training data. These speakers are not used in the trials considered for the experiments in this paper. All models are trained and applied separately for each gender.

The set of audio characterization posteriors obtained for the experiments in this paper are extracted using the same MFCC iVectors as the ones used in the speaker recognition system. For the audio characterization system, though, we use the iVectors of dimension 600 as they are generated by the iVector extractor, without applying LDA, mean-normalization or length-normalization.

Results will be shown in terms of equal error rate (EER) and decision cost function (DCF) as recently defined by NIST for the core condition of 2010 SRE [21]. Even though all processing (including calibration) occurs by gender, results are shown on trials from both genders.

### 3.3. Results

Calibration with a linear function and without metadata does not affect the EER or DCF since those measures are immune to linear transformations. On the other hand, when metadata is used to affect the parameters of the calibration (either scale or shift), the performance of the system might change with respect to that obtained with the original uncalibrated scores or those calibrated without metadata. If the metadata corresponds to a nuisance factor that creates a bias in the scores, using it as input to the calibration process allows the system to compensate for this bias, aligning the distributions for the different types of metadata and, as a consequence, improving overall system performance.

Figure 2 shows the performance on the different PRISM conditions for the scores calibrated without metadata and the scores calibrated using the audio characterization posteriors as metadata. Table 3 shows the relative gains for each condition.

We can see that in four out of the seven conditions there is a significant gain obtained from using the audio characterization scores in the calibration process, with the biggest gains in DCF for the *tela* condition. This condition is formed by signals from the two classes from Table 1 with the best prediction performance (Table 2): clean telephone and clean microphone. The gain from using metadata can then be explained by the fact that the calibration procedure can choose different shifts for each of these two classes, successfully compensating for any existing bias across them. A similar explanation can be given for the gain observed in the *noi* condition, which is formed by three classes in the audio characterization system: noisy 8 dB, 15 dB and 20 dB. Note that this is the case even though the performance of the audio characterization system for these classes is not as good as for the clean classes.

The gains in conditions *int* and *rev* were rather surprising considering that, in the first case, all signals in this condition belong to the same audio characterization class (clean microphone) and, in the second case, the prediction of the classes within this condition was very poor (Table 2). Nevertheless, we can interpret these gains if we consider that the true class is not always the best representation for a certain signal in terms of the effects that the nuisance characteristics have on the corresponding iVector. That is, in many cases, the "wrong" class, as detected by the audio characterization system in a soft way

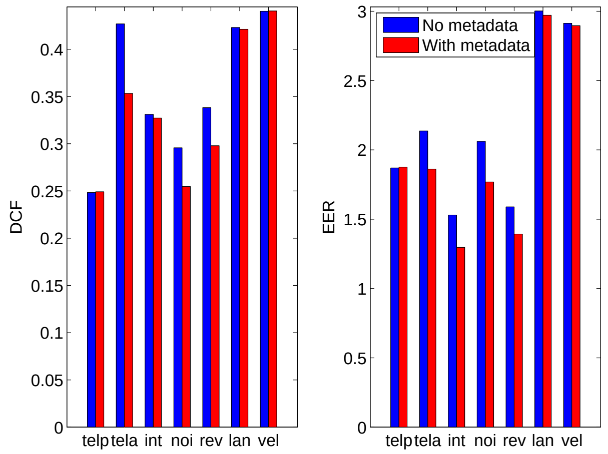


Figure 2: Comparison of performance for the original MFCC scores and the scores calibrated using audio characterization metadata.

is a better predictor of the bias that will be found in the scores involving a certain signal.

The lack of gain in the *telp*, *lan*, and *vel* conditions is simply explained by the fact that signals within these conditions come from clean telephone data, which is very reliably labeled by the audio characterization system. Hence, a single bias is applied to all trials within these conditions explaining the lack of change in performance with respect to not using metadata in calibration.

Table 3: Relative gains per condition when using the audio characterization posteriors as metadata for calibrating the MFCC system with respect to the result obtained without the use of this metadata.

| System | Rel. Gain in DCF | Rel. Gain in EER |
|---|---|---|
| telp | -0.28 | -0.34 |
| tela | 17.24 | 12.87 |
| int | 1.18 | 15.24 |
| noi | 13.87 | 14.23 |
| rev | 11.92 | 12.33 |
| lan | 0.45 | 1.01 |
| vel | -0.07 | 0.59 |

## 4. Discussion

We propose a method for determining the nuisance characteristics present in an audio signal. The method relies on the extraction of iVectors over the signal, an approach borrowed from the speaker recognition literature. Given a set of audio classes in the training data, a Gaussian model is trained to represent the iVectors for each of these classes. During testing, these models are used to obtain the posterior probability of each class given the iVector for a certain signal. This framework allows for a unified way of detecting any kind of nuisance characteristic that is properly encoded in the iVector used to represent the signal.

We show results when using this method for prediction of the same classes defined over the training data for a held-out set of signals. Results show excellent performance in detecting clean microphone and telephone data and noisy data, even though, in this case, different SNR levels are sometimes confused with each other. Reverberated data is not effectively detected by this system. We believe this is mainly because too few kinds of reverberation are used in training, not allowing for proper generalization.

The proposed system was conceived as a way to detect the nuisance characteristics in a signal that might be affecting the performance of a speaker recognition system (or some other speech processing system). If the type of nuisance in a certain signal is known, the system can somehow adapt to it, probably improving performance. We show one approach for the use of the output generated by the audio characterization system by a speaker recognition system. The information is used at the last stage of the speaker recognition system, when calibration of the scores is performed. A modified logistic regression approach is used that takes into account the vector of posteriors for each audio class generated by the audio characterization system, adapting the parameters of the calibration as a function of this vector's values. The idea can be trivially extended for fusion of several speaker recognition systems using the same logistic regression method.

We show that this approach leads to significant gains in calibration of a state-of-the-art MFCC speaker recognition system. Gains are obtained over a variety of nuisance effects, including noise, reverberation, and channel variability with relative gains in EER of up to 15%.

The described system is only one particular implementation of a more general idea in which vectors that represent the waveforms (or even segments within them) are modeled using a certain trainable distribution that is then used to obtain posteriors for a new waveform. The classes into which the training data is divided can be given by labels, as described here, but they can also be inferred from the training iVectors using clustering techniques. This is a promising direction we plan to pursue in the near future.

Finally, as part of the posterior computation, the system first computes the likelihoods for the different classes given a waveform. If all likelihoods are very small, the system could then output a warning to the user that the waveform does not match the training data well. This is useful since, in many cases, such a waveform would result in unpredictable performance of the classification system of interest. For example, if the ultimate goal is to detect the speaker identity and the observed waveform has a type or a level of noise that has not been observed during training, it is reasonable to expect that the score generated by the speaker identification system will be unreliable on that waveform.

## 5. Acknowledgement

This work was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U. S. Government.

## 6. References

[1] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, and N. Scheffer, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proceedings of SRE11 Analysis Workshop*, Atlanta, Dec. 2011.

[2] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Proceedings of the Interspeech Conference*, Brisbane, Sept. 2008.

[3] M. Graciarena, S. Kajarekar, A. Stolcke, and E. Shriberg, "Noise robust speaker identification for spontaneous arabic speech," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Apr. 2007.

[4] M. Graciarena, H. Franco, G. Myers, and V. Abrash, "Robust feature compensation in nonstationary and multiple noise environments," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech 05)*, Lisbon, Sept. 2005.

[5] B.E.D. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Apr. 1997.

[6] L. Ferrer, K. Sönmez, and S. Kajarekar, "Class-dependent score combination for speaker recognition," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech 05)*, Lisbon, Sept. 2005.

[7] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg, "System combination using auxiliary information for speaker verification," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Apr. 2008.

[8] S. Sagayama, K. Shinoda, M. Nakai, and H. Shimodaira, "Analytic methods for acoustic model adaptation: A review," in *Proceedings of ISCA Workshop on Adaptation Methods (Sophia Antipolis, France)*, Aug. 2001, pp. 67–76, Invited Paper.

[9] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proceedings of the Interspeech Conference*, Brisbane, Sept. 2008.

[10] R. Ratnam, D. L. Jones, B. C. Wheeler, Jr. W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[11] D. Gonzalez Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," in *Proceedings of the Interspeech Conference*, Florence, Italy, Aug. 2011.

[12] N. Dehak, P.J. 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 2011.

[13] P. Matejka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-Vector speaker verification," in *Proceedings of the Interspeech Conference*, Florence, Italy, Aug. 2011.

[14] "Freedsound," http://www.freesound.org.

[15] G. Hirsch, "Fant," http://dnt.kr.hs-niederrhein.de/download.html.

[16] S. G. McGovern, "A model for room acoustics," http://www.2pi.us/rir.html.

[17] E. de Villiers and N. Brummer, "Bosaris toolkit," .

[18] N. Brummer, L. Burget, P. Kenny, P. Matejka, E. de Villiers, M. Karafiat, M. Kockmann, O. Glembek, O. Plchot, D. Baum, and M. Senoussauoi, "ABC system description for NIST SRE 2010," in *Proceedings of NIST 2010 Speaker Recognition Evaluation*. 2010, pp. 1–20, National Institute of Standards and Technology.

[19] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of the Interspeech Conference*, Florence, Italy, Aug. 2011.

[20] S.J.D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of the International Conference on Computer Vision*, 2007.

[21] "NIST SRE10 evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf.