ACCENT RECOGNITION USING I-VECTOR, GAUSSIAN MEAN SUPERVECTOR AND GAUSSIAN POSTERIOR PROBABILITY SUPERVECTOR FOR SPONTANEOUS TELEPHONE SPEECH

Mohamad Hasan Bahari[⋆] *Rahim Saeidi*[†] *Hugo Van hamme*[⋆] *David Van Leeuwen*[†]

[⋆]Center for processing speech and images, KU Leuven, Belgium
{mohamadhasan.bahari,Hugo.Vanhamme}@esat.kuleuven.be
[†]Center for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
{r.saeidi,d.vanleeuwen}@let.ru.nl

ABSTRACT

In this paper, three utterance modelling approaches, namely Gaussian Mean Supervector (GMS), i-vector and Gaussian Posterior Probability Supervector (GPPS), are applied to the accent recognition problem. For each utterance modeling method, three different classifiers, namely the Support Vector Machine (SVM), the Naive Bayesian Classifier (NBC) and the Sparse Representation Classifier (SRC), are employed to find out suitable matches between the utterance modelling schemes and the classifiers. The evaluation database is formed by using English utterances of speakers whose native languages are Russian, Hindi, American English, Thai, Vietnamese and Cantonese. These utterances are drawn from the National Institute of Standards and Technology (NIST) 2008 Speaker Recognition Evaluation (SRE) database. The study results show that GPPS and i-vector are more effective than GMS in this accent recognition task. It is also concluded that among the employed classifiers, the best matches for i-vector and GPPS are SVM and SRC, respectively.

Index Terms— Accent Recognition, i-vector, Gaussian Posterior Probability Supervector, Gaussian Mean Supervector

1. INTRODUCTION

A fundamental challenge of using Automatic Speech Recognition (ASR) systems in real world markets such as telephone networks and personal computers is their significant performance drop for non-native speakers [1, 2]. Consequently, accent/dialect recognition, has received an increased attention during the last years due to its importance for the enhancement of ASR performance [2]. It has also a wide range of commercial applications such as targeted advertising, service customization and forensics software. Although different methods have been suggested to solve this problem during the last decade, it still remains a challenging task.

Accent/dialect recognition techniques can be divided into phonotactic and acoustic approaches [3]. Since phonotactic features and acoustic (spectral and/or prosodic) features provide complementary cues, state-of-the-art methods usually apply a combination of both through a fusion of their output scores [3]. A phone recognizer followed by language models (PRLM) and parallel PRLM (PPRLM) techniques developed within the language recognition area, are successful phonotactic methods focusing on phone sequences as an important characteristic of different accents [4].

This work is supported by the European Commission as a Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008), namely Bayesian Biometrics for Forensics (BBfor2), under Grant Agreement number 238803.

The acoustic approaches, which are the main focus of this paper, enjoy the advantage of requiring no specialized language knowledge [3]. One effective acoustic method for accent recognition involves modeling speech recordings with Gaussian mixture model (GMM) mean supervectors before using them as features in a support vector machine (SVM) [3]. Similar Gaussian mean supervector (GMS) techniques have been successfully applied to different speech analysis problems such as speaker recognition [5]. While effective, these features are of a high dimensionality resulting in high computational cost and difficulty in obtaining a robust model in the context of limited data. Another effective approach for modeling the utterances is Gaussian posterior probability supervector (GPPS), which entails a lower dimension compared to GMSs [6, 7]. Recent studies show that the GPPSs carry complimentary information to GMSs [6, 8]. Consequently, incorporating them in the recognition system might increase the overall accuracy. A similar GPPS framework was effectively applied to the problem of age and gender recognition [6, 9, 10]. In the field of speaker recognition, recent advances using i-vectors have increased the recognition accuracy considerably [11]. An i-vector is a compact representation of an utterance in the form of a low-dimensional feature vector. The same idea was also effectively applied to spoken language recognition and speaker age estimation [12, 13].

In this paper, we apply GMSs, GPPSs and i-vectors to recognize the native language of speakers from English spontaneous telephone speech recordings (L1 recognition problem). To find out a suitable classifier for each modeling method, three different pattern recognition approaches, namely the Support Vector Machine (SVM), Naive Bayesian Classifier (NBC) and the Sparse Representation Classifier (SRC), are tested. The evaluation database is formed by using English utterances of speakers whose native languages are Russian, Hindi, American English, Thai, Vietnamese and Cantonese. These speech signals are extracted from the National Institute of Standards and Technology (NIST) 2008 Speaker Recognition Evaluation (SRE) corpus.

The rest of this paper is organized as follows. Section 2 presents the related work and contributions of this paper. In Section 3, the developed accent recognition systems are elaborated in detail. Section 4 explains our experimental setup. The evaluation results are presented and discussed in section 5. The paper ends with conclusions in section 6.

2. RELATED WORK AND CONTRIBUTIONS

Different acoustic approaches developed in the area of language recognition have been suggested to reach a desirable accent recog-

nition accuracy [3, 14, 15, 16]. Recently, Hanani et al. reported results of applying GMM-UBM, GMM-SVM (which is labeled as GMS-SVM in the rest of this paper) and GMM tokenization followed by n-gram language model methods to recognize 14 accents in the British Isles [3]. They used the Accents of the British Isles (ABI-1) corpus in their research. Their evaluation results show that GMS-SVM is more accurate compared to their other acoustic-based accent recognition systems.

DeMarco and Cox take this a step further by applying i-vectors to the same task [15]. They tested six different classification algorithms such as SVM and Linear Discriminant Analysis (LDA) and concluded that similar results as those of GMS-SVM can be obtained in the i-vector framework. Their results show no advantage for using i-vectors instead of GMSs.

In this paper, we investigate the effectiveness of GMS and i-vector for native language recognition on a spontaneous and real speech database instead of the ABI-1 corpus, which consists of clean and read speech signals. Consequently, we formed a database of non-native accents of English by extracting English utterances with Russian, Hindi, American English, Thai, Vietnamese and Cantonese accents from the NIST 2008 SRE database. For each utterance modeling method, three different classifiers, namely SVM, NBC and SRC are employed to further investigate the role of classifiers in this task. Unlike SVM and NBC, sparse representation classification techniques have never been tested on accent recognition problems. On the other hand, recent studies show the effectiveness of GPPS in other speech technology problems such as speaker adaptation and speaker age group recognition [6, 8]. Consequently, we test GPPS along with i-vectors and GMS in our investigations on accent recognition too.

3. SYSTEM DESCRIPTION

3.1. Problem Formulation

In the accent or dialect recognition problem, we are given a training data set $S^{\text{tr}} = \{(x_1, y_1), \cdots, (x_n, y_n), \cdots, (x_N, y_N)\}$, where x_n denotes the n^{th} utterance of the training data set and y_n denotes a label vector which shows the correct accent of the utterance. Each label vector contains a one in the i^{th} row if x_n belongs to the i^{th} class and zeros elsewhere. The goal is to approximate a classifier function (g), such that for an unseen observation x^{tst}, $\hat{y} = g(x^{\text{tst}})$ is as close as possible to the true label.

The first step for approximating function g is converting variable-duration speech signals into fixed-dimensional vectors suitable for the classification algorithms. Three approaches, namely GPPS, GMS and i-vectors are widely used for this purpose. These methods are described in section 3.2.

3.2. Utterance Modelling Approaches

In this section, the underlying idea of GMS, GPPS and i-vector is explained in more details.

3.2.1. Gaussian Posterior Probability Supervector

Consider a Universal Background Model (UBM) with the following likelihood function.

$$p(\mathbf{o}_t|\lambda) = \sum_{j=1}^{J} \omega_j p(\mathbf{o}_t|\mu_j, \mathbf{\Sigma}_j)$$
$$\lambda = \{\omega_j, \mu_j, \mathbf{\Sigma}_j\},\ j = 1, \cdots J \tag{1}$$

where $\mathbf{o}_t$ is the acoustic vector at time t, ω_j is the mixture weight for the j^{th} mixture component, $p(\mathbf{o}_t|\mu_j, \mathbf{\Sigma}_j)$ is a Gaussian probability density function with mean μ_j and covariance matrix $\mathbf{\Sigma}_j$ and J is the total number of Gaussians in the mixture (2048 in this work). Given an utterance, the occupancy posterior probability for the j^{th} mixture component is calculated as follows:

$$\kappa_j = \frac{1}{T} \sum_{t=1}^{T} \frac{\omega_j p(\mathbf{o}_t|\mu_j, \mathbf{\Sigma}_j)}{\sum_{j=1}^{J} \omega_j p(\mathbf{o}_t|\mu_j, \mathbf{\Sigma}_j)} \tag{2}$$

where T is the total number of frames in the utterance. Finally, the GPPS of the given utterance is formed as follows.

$$\mathbf{k} = [\kappa_1, \cdots, \kappa_j, \cdots, \kappa_J] \tag{3}$$

Assuming the UBM components represent the acoustic space of all accents in the training dataset, each element in the GPPS supervector of a sufficiently long utterance shows the existence level of the corresponding component in the utterance accent. This information facilitate in the identification of accents.

3.2.2. Gaussian Mean Supervector

Given an utterance, different approaches can be applied to adapt a Universal Background Model (UBM) to the speech characteristics of the new speaker [17, 5]. Then, the Gaussian means of the adapted GMM are extracted and concatenated to form a GMS for the given utterance. In this research, we apply Maximum-A-Posteriori method to adapt the Gaussian means of the UBM [5].

3.2.3. i-vector

GMSs described in section 3.2.2 have been shown to provide a good level of performance. In the related field of speaker recognition, GMSs are commonplace. Recent progress in this field, however, has found an alternate method of modeling GMM supervectors that provides far superior speaker recognition performance [11]. This technique is referred to as total variability modeling. Total variability modeling assumes the GMM mean supervector, $\mathbf{M}$, that best represents a set of feature vectors can be decomposed as

$$\mathbf{M} = \mathbf{u} + \mathbf{T}\mathbf{v} \tag{4}$$

where $\mathbf{u}$ is the mean supervector of the UBM, $\mathbf{T}$ spans a low-dimensional subspace (400 dimensions in this work) and $\mathbf{v}$ are the factors that best describe the utterance-dependent mean offset $\mathbf{T}\mathbf{v}$. The vector $\mathbf{v}$ is commonly referred to as the i-vector and has a standard normal distribution $N(0, \mathbf{I})$. Subspace $\mathbf{T}$ is estimated via factor analysis to represent the directions that best separate different speech recordings in a large development data set. An efficient procedure for training $\mathbf{T}$ and MAP adaptation of i-vectors $\mathbf{v}$ can be found in [18]. In the total variability modeling approach, i-vectors are the low-dimensional representation of an audio recording that can be used for classification and estimation purposes.

3.3. Classifiers

In this section, the applied classifiers are briefly described.

3.3.1. Naive Bayesian Classifier

Bayesian classifiers are probabilistic classifiers working based on Bayes' theorem and the maximum posteriori hypothesis. They predict class membership probabilities, i.e., the probability that a given test sample belongs to a particular class. The Naive Bayesian classifier (NBC) is a special case of Bayesian classifiers, which assumes class conditional independence to decrease the computational cost and training data requirement [19]. In this paper, class distributions are assumed to be Gaussian.

3.3.2. Support Vector Machine

Support Vector Machines (SVM) is a supervised, binary and discriminative classifier initially introduced by Cortes and Vapnik [20]. Given a set of training examples, an SVM attempts to find the maximum margin separation hyperplane between two classes of data such that it generalizes well to the test data points. The basic SVMs are binary and discriminative classifiers, however, an effective multi-class and probabilistic extension have also been developed by Wu et al. based on pairwise coupling strategy [21].

3.3.3. Sparse Representation Classifier

Sparse representation classification techniques have received a great deal of attentions in recent years. In sparse representation classification, first we search for a sparse representation of a test sample in terms of a linear combination of training samples. Then, the residuals for each class are calculated. These residuals show the level of similarity of the test sample with each category [22].

In our experiments, the dimension of feature vectors, i.e., the dimension of the GPPS, GMS or i-vector, is greater than the number of training samples which leads to an over-determined sparse representation problem. Therefore, to achieve the sparse representations of the test samples, we applied an l_1-minimization approach [22].

3.4. Training and Testing

The principle of the proposed accent recognition approach is illustrated in Figure 1. As it can be interpreted from this figure, in the training phase, each utterance in the train data set is converted to a high dimensional vector using one of the three utterance modeling approaches (GPPS, GMS or i-vector) described in Section 3.2. Then, the obtained high dimensional vector along with their corresponding accent label are used to train one of the three classifiers described in Section 3.3.

In the testing phase, the utterance modeling approach applied in the training phase is used to extract a high dimensional vector from the utterance of an unseen speaker. Then the trained classifier uses the extracted vector to recognize the accent of the test speaker.

4. EXPERIMENTAL SETUP

4.1. Database

The National Institute for Standard in Technology (NIST) have held annual or biannual speaker recognition evaluations (SRE) for the past two decades. With each SRE, a large corpus of telephone (and more recently microphone) conversations are released along with an evaluation protocol. These conversations typically last 5 minutes and originate from a large number of participants for whom additional meta data is recorded—including participant age, language and smoking habits. The NIST databases where chosen for this work due to the large number of speakers and because the total variability subspace requires a considerable amount of development data for training. The development data set used to train the total variability subspace and UBM includes over 30,000 speech recordings and was sourced from NIST 2004–2006 SRE databases, LDC releases of Switchboard 2 phase III and Switchboard Cellular (parts 1 and 2).

The NIST 2008 SRE database includes, many English utterances from speakers whose native languages are Spanish, Russian, Hindi, etc. The native language of speakers usually affects their English pronunciation, i.e., accented speech, due to transferring the phonological rules from their native language into their English speech and creating innovative pronunciations for English sounds

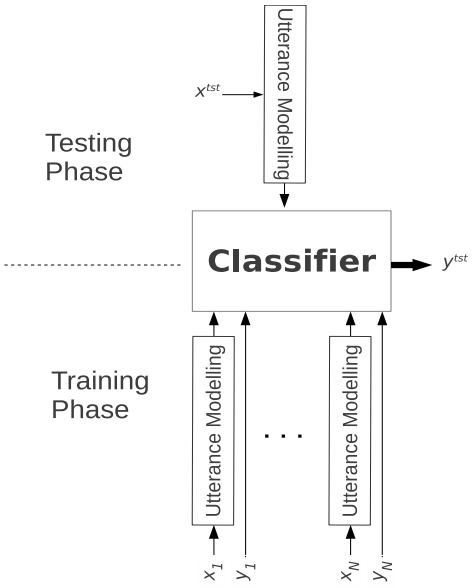

Fig. 1. *The block diagram of the accent recognition systems in training and testing phases.*

which do not exist in their mother tongue [23]. Unfortunately, the number of utterances in some accents is not high enough to perform our recognition experiments. Consequently, only five accents — Russian (RUS), Hindi (HIN), American English (USE), Thai (THA) and Vietnamese-Cantonese (VIE-YUH)— with enough available recordings are chosen for our experiments. These utterances are extracted from telephone recordings of the core protocol, short2-short3, of the NIST 2008 SRE database. Note that since a fraction of Vietnamese Americans consists of Hoa people whose native language is Cantonese, Vietnamese and Cantonese are considered as one category in our experiments. Table 1 lists the number of utterances and speakers for each accent.

4.2. Performance Measure

The effectiveness of the proposed method is evaluated using the percentage of correctly classified utterances ($P_{\rm cc}$) and minimum log-likelihood-ratio cost ($C_{\rm llr}^{\rm min}$) [24, 25]. This section briefly describes the applied performance measure methods.

4.2.1. Percentage of Correctly Classified Utterances

$P_{\rm cc}$ is a simple performance measure which can be calculated using the following relation.

$$P_{\rm cc} = \frac{N_{cc}}{N_T} \tag{5}$$

Table 1. *The number of utterances and speakers for each accent category.*

Accent	Number of Utterances	Number of Speakers
USE	84	84
THA	63	41
RUS	49	32
HIN	62	39
VIE-YUH	101	69
Total	**359**	**265**

Table 2. *Comparison of various i-vector, GPPS and GMS based systems. The results are given in $P_{\rm cc}$ and $C_{\rm llr}^{\rm min}$.*

Classifier	Feature	$P_{\rm cc}(\%)$	$C_{\rm llr}^{\rm min}$
SVM	GMS	53	2.03
	GPPS	58	1.92
	i-vector	56	1.77
NBC	GMS	47	2.12
	GPPS	48	2.05
	i-vector	52	1.97
SRC	GMS	49	2.00
	GPPS	56	1.63
	i-vector	41	2.08

where N_{cc} and N_T denote the number of correctly classified utterances and the total number of utterances in the test data set respectively.

4.2.2. Log-Likelihood Ratio Cost

Log-Likelihood Ratio Cost ($C_{\rm llr}$) is an application-independent performance measure for recognizers with soft decisions output in the form of log-likelihood-ratios. This performance measure, which has been adopted for use in the NIST SRE, was initially developed for binary classification problems such as speaker recognition. It is extended to multi-class classification problems such as language recognition later in 2006 [24]. $C_{\rm llr}^{\rm min}$ represents the minimum possible $C_{\rm llr}$ which can be achieved for an optimally calibrated system [24]. In this research we apply FoCal Multiclass Toolkit [26] to calculate $C_{\rm llr}^{\rm min}$.

5. RESULTS

In this section, the performances of nine developed systems are evaluated and compared. The acoustic feature consists of 20 Mel-Frequency Cepstrum Coefficients (MFCCs) including energy appended with their first and second order derivatives, forming a 60 dimensional acoustic feature vector. This type of feature is very common in state-of-the-art i-vector based speaker recognition systems. To have more reliable features, Wiener filtering, speech activity detection [27] and feature warping have [28] been considered in front-end processing.

For the evaluation, a one speaker hold out training-testing strategy is adopted so that test speaker utterances are never included in the training set. In other words, 265 (total number of speakers in the database) independent experiments have been run. In each experiment, all utterances of a new speaker are used as testing and the rest of the utterances are used for training.

Table 2 lists the $P_{\rm cc}$ and $C_{\rm llr}^{\rm min}$ for all nine developed systems. For the SVM classifier different kernels have been tested and Table 2 shows only the best results obtained by the linear kernel. As it can be seen from Table 2, both classifier types and utterance modelling methods influence the recognition accuracy. While in SVM and NBC classification algorithms the i-vector framework leads to the most accurate recognition, for the SRC algorithm, GPPS provides the best results.

The results also show that the NBC algorithm is not effective in this case. It can be due to high dimensionality of input features which increases class conditional dependency violating the naive assumption of the NBC(class conditional independence).

Table 2 also illustrates that the GPPS and the i-vector utterance modelling approaches are more effective than the GMS method in this non-native accents recognition task.

Table 3. *Comparison of NBC, SVM and SRC after feature level fusion. The results are given in $P_{\rm cc}$ and $C_{\rm llr}^{\rm min}$.*

Classifier	Feature	$P_{\rm cc}(\%)$	$C_{\rm llr}^{\rm min}$
NBC	i-vector-GPPS-GMS	50	2.07
SVM	i-vector-GPPS-GMS	56	1.84
SRC	i-vector-GPPS-GMS	58	1.63

5.1. Feature Level Fusion

Many literatures reported the effectiveness of score level fusion [3, 29]. However, this type fusion requires a development data set which is not available in this task due to the limited number of utterances per accent. In this paper, we employed feature level fusion requiring only one learning stage while taking advantage of mutual information [30]. In this type of fusion, the extracted i-vector, GPPS and GMS of each utterance are concatenated to form a high dimensional supervector representing the utterance. Table 3 lists the results of NBC, SVM and SRC after feature level fusion. It shows that the accuracy of accent recognition increases after the fusion when SRC is applied for the classification. However, this improvement is not observed when NBC or SVM are employed.

Table 4 illustrates the results of i-vector-GPPS-GMS-SRC system as a confusion matrix. As it can be interpreted from this table, the recognition accuracy for all accents is noticeably higher than the chance level which confirms the efficiency of the proposed approach.

6. CONCLUSIONS

In this paper, we have investigated the effectiveness of the GMS, GPPS and i-vector utterance representation approaches for accent recognition on a spontaneous and real speech database formed by extracting English utterances with Russian, Hindi, American English, Thai, Vietnamese and Cantonese accents from the NIST 2008 SRE database. For each utterance modeling method, three different classifiers, namely SVM, NBC and SRC, have been employed to find out suitable matches between the utterance modelling schemes and the classifiers. The study results show that GPPSs and i-vectors are more effective than GMS in this accent recognition task. Among the employed classifiers, the best matches for i-vector and GPPS are SVM and SRC respectively. Furthermore, feature level fusion was found to be marginally effective in increasing the accent recognition accuracy, when SVM or SRC were applied as classifiers.

Table 4. *The confusion matrix of accent recognition for i-vector-GPPS-GMS-SRC system. The results are given in percentage*

		Predicted				
		USE	THA	RUS	HIN	VIE-YUH
Actual	USE	65	4	7	6	18
	THA	14	46	2	3	35
	RUS	27	0	43	14	16
	HIN	8	5	3	60	24
	VIE-YUH	15	14	6	7	58

7. REFERENCES

[1] A. Hanani, "Human and computer recognition of regional accents and ethnic groups from british english speech," *University of Birmingham*, July 2012.

[2] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," *Columbia University*, 2011.

[3] Hanani A., Russell M.J., and Carey M.J., "Human and computer recognition of regional accents and ethnic groups from british english speech," *Computer Speech and Language*, vol. 27, no. 1, pp. 59–74, 2013.

[4] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[5] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[6] M. Li, K.J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech and Language*, vol. 27, no. 1, pp. 151 – 167, 2013.

[7] X. Zhang, S. Hongbin, Z. Qingwei, and Y. Yonghong, "Using a kind of novel phonotactic information for svm based speaker recognition," *IEICE TRANSACTIONS on Information and Systems*, vol. 92, no. 4, pp. 746–749, 2009.

[8] X. Zhang, K. Demuynck, et al., "Latent variable speaker adaptation of gaussian mixture weights and means," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4349–4352.

[9] M.H. Bahari and H. Van hamme, "Speaker age estimation and gender detection based on supervised non-negative matrix factorization," in *Proc. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, 2011, pp. 1–6.

[10] M.H. Bahari and H. Van hamme, "Speaker age estimation using hidden markov model weight supervectors," in *Proc. 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, 2012, pp. 517–521.

[11] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[12] N. Dehak, P.A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via ivectors and dimensionality reduction," in *Proc. Interspeech*, 2011, pp. 857–860.

[13] M.H. Bahari, M. McLaren, H. Van hamme, and D. Van Leeuwen, "Age estimation from telephone speech using i-vectors," in *Proc. Interspeech*, 2012.

[14] F. Biadsy, J. Hirschberg, and D.P.W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *Proc. Interspeech*, 2011.

[15] A. Demarco and S.J. Cox, "Iterative classification of regional british accents in i-vector space," in *Proc. Machine Learning in Speech and Language Processing*, 2012.

[16] M.K. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4398–4401.

[17] M.H. Bahari and H. Van hamme, "Speaker adaptation using maximum likelihood general regression," in *11th International Conference on Information Science, Signal Processing and their Applications*, 2012, pp. 29–34.

[18] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[19] R.R. Yager, "An extension of the naive bayesian classifier," *Information Sciences*, vol. 176, no. 5, pp. 577–588, 2006.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[21] T.F. Wu, C.J. Lin, and R.C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *The Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[22] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[23] M. MacDonald, "The influence of spanish phonology on the english spoken by united states hispanics," *American Spanish pronunciation: Theoretical and applied perspectives*, 1989.

[24] N. Brummer and D.A. van Leeuwen, "On calibration of language recognition scores," in *Proc. IEEE Odyssey Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[25] N. Brummer, "Application-independent evaluation of speaker detection," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[26] Niko Brummer, "Focal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition scores," *Tutorial and User Manual. Spescom DataVoice*, 2007.

[27] M. McLaren and D. van Leeuwen, "A simple and effective speech activity detection algorithm for telephone and microphone speech," in *Proc. NIST SRE Workshop*, 2011.

[28] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," pp. 213–218, 2001.

[29] E. Wong and S. Sridharan, "Fusion of output scores on language identification system," 2003, pp. 1–11.

[30] S. Planet and I. Iriondo, "Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition," in *Proc. 7th Iberian Conference on Information Systems and Technologies (CISTI)*, 2012, pp. 1–6.