SWISS FRENCH REGIONAL ACCENT IDENTIFICATION

Alexandros Lazaridis[1]*, Elie Khoury*[1]*, Jean-Philippe Goldman*[2]*,*
Mathieu Avanzi[3]*, Sébastien Marcel*[1] *and Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]University of Geneva, Geneva, Switzerland
[3]LLF, UMR 7110, Paris Diderot, France

alaza@idiap.ch

ABSTRACT

In this paper an attempt is made to automatically recognize the speaker's accent among regional Swiss French accents from four different regions of Switzerland, i.e. Geneva (GE), Martigny (MA), Neuchâtel (NE) and Nyon (NY). To achieve this goal, we rely on a generative probabilistic framework for classification based on Gaussian mixture modelling (GMM). Two different GMM-based algorithms are investigated: (1) the baseline technique of universal background modelling (UBM) followed by maximum-a-posteriori (MAP) adaptation, and (2) total variability (i-vector) modelling. Both systems perform well, with the i-vector-based system outperforming the baseline system, achieving a relative improvement of 15.3% in the overall regional accent identification accuracy.

Index Terms— Accent Identification, French Regional Accents, GMM Modelling, i-vectors, SVM

1. INTRODUCTION

In verbal communication, the para-linguistic aspects of speech convey information about gender, age, emotions, emphasis, contrast and even the regional and social accents of the speaker [1]. Humans over the years learn, to some extent, to identify and interpret most of these aspects of speech. Over the last decades a lot of effort has been made to automatically, with the help of machines, identify this kind of information from speech, such as emotion recognition [2], gender and age recognition [3].

One of the para-linguistic aspects, embodied in speech, is the accent/dialect information. In contrast to accent variations, dialect variations are characterized by differences mainly in word selection and use of the grammar in a language. However, accent variations are defined by diversities in pronunciation (phone sequence) and speaking style (rhythm, variation in pitch) [4, 5]. Furthermore, accent variations can be divided in two subcategories: foreign accents and regional accents. The former characterizes the variations in speech uttered by non-native speakers speaking a foreign language. In this case, the pronunciation of a word might vary a lot depending on the native language of the speaker and level of the foreign language proficiency of the speaker. The latter case, regional accents, refers to the changes in pronunciation but mainly in speaking style [5, 4, 6, 7] among native speakers of a language, which makes it even harder to differentiate them and identify the origin/region of the speaker.

Over the last years, a lot of research has been done in the field of automatic foreign accent and dialect recognition [8, 9, 10, 11, 12, 13]. The main purpose for this, is to build robust automatic speech recognition (ASR) systems which are not influenced by the foreign accent of the speaker or are adapted to the dialect of the speaker [14, 15]. On the other hand regional accent identification (RAI) can help in personalizing synthetic speech of a text-to-speech (TTS) system according to a speaker of a specific regional accent. Consequently, RAI can also be beneficial for personalizing a speech-to-speech translation (S2ST) system for synthesizing the recognized and translated speech from one language to a specific regional accent in another language [16]. To the extent of our knowledge, no previous work has been done on the regional accent identification task of French or Swiss French accents.

This paper is a preliminary work on attempting to automatically recognize the speaker's accent among regional Swiss French accents from four different regions of Switzerland, i.e. Geneva (GE), Martigny (MA), Neuchâtel (NE) and Nyon (NY). Among these regional accents, the variations in speech occur in both segmental and suprasegmental domains. These differences are subtle and thus can not be considered as phonological differences. For instance, some typical attested variations lie with the realisation of the primary accent. In the segmental side, some differences mainly concern the realisation of /o/, /R/ or some nasal vowels, but are very sporadic. In other words, the variations are mainly focused on the speaking style, i.e. different rhythm and pitch variations, rather than on the pronunciation of the words [5, 4, 6], making the task of regional accent identification even more difficult. The goal of this work is to investigate if speaker identification techniques, along with the acoustic features used in these approaches, can help to distinguish regional accents in order to be used for improving ASR and TTS systems. For achieving this goal, we believe we can cast this task as a biometric identification problem, relying on techniques which were first introduced in speaker recognition and then successfully applied for several audio processing problems (e.g. speaker diarization [17], language identification [18]). In this paper, we implement a generative probabilistic framework for classification based on *Gaussian mixture modelling* (GMM). Two GMM-based algorithms are investigated: (1) the baseline technique of universal background modelling (UBM) followed by maximum-a-posteriori (MAP) adaptation [19] and (2) the total variability (i-vectors) modelling [20].

The rest of the paper is organized as follows. Section 2 presents the related work in the field. In section 3, the Swiss French speech database is described, along with a human evaluation of the accent degree of the speakers of the database. In section 4, the proposed GMM-based system is presented. The experimental protocol and results are described in section 5. Finally the conclusions are given in section 6.

2. PRIOR WORK

In the last years, a lot of research has been done on the foreign accent identification (FAI) task [8, 10, 11, 21], following mainly the techniques from the dialect/language identification (DID/LID) task [12, 13, 22]. There are two main approaches to tackle this problem: phonotactic and acoustic [22]. The former technique is based on the accent-dependent variabilities in the sequences in which the speech sounds occur, whilst the latter, acoustic methods, exploit differences in the distributions of sounds in different accents [22]. In [8] hidden Markov models (HMM) were used to identify 6 foreign accents of English. A parallel competing sub-nets topology was used. Each sub-net was composed of an ergodic net of the full set of HMM phone models trained on the corresponding accent of English, selecting the sub-net with the highest likelihood. In contrast to acoustic-based systems, in [23] phone labels and segmentation were used to constrain the acoustic models. GMM-supervectors were extracted for each phone type after obtaining phone hypotheses using a phone recognizer. Finally an SVM classifier was trained to identify foreign accents of English.

Recently in [24] three utterance level modelling techniques, i.e. Gaussian mean supervector (GMS), i-vectors and Gaussian posterior probability supervector (GPPS), were evaluated using three different classification algorithms, i.e. support vector machine (SVM), naive bayesian classifier (NBC) and sparse representation classifier (SRC), in the task of foreign accent identification on English utterances spoken by speakers having Russian, Hindi, American English, Thai, Vietnamese and Cantonese as native languages. The main conclusion drawn from this work was that i-vectors combined with SVM and GPPS combined with SRC yield the best results.

By contrast, the research done on the task of regional accent identification (RAI) is very limited. In Hanani et al. [11], Gaussian mixture model - universal background model (GMM-UBM), GMM-SVM and GMM tokenization combined with n-gram language model (LM) were used for identifying 14 British English regional accents. The results from the evaluation show that GMM-SVM achieved the highest identification accuracy score. In [25], in the same task, using the same database as the previous work, i-vectors were compared to GMM-SVM concluding that no advantage was gained from the use of i-vectors.

To the extent of our knowledge no work has been done on RAI in French or Swiss French regional accents. In [6], a study was made on human accent identification and how the background of the listeners affects their perception of the accents of 6 Francophone regions, i.e. Normandy, Vendée, Romand Switzerland, Languedoc and Basque Country. The listeners from two different regions (Paris and Marseille) achieved an average of approximately 43% of accuracy on human regional accent identification, verifying the difficulty of the RAI task in French accents.

3. SWISS FRENCH ACCENT DATABASE

The data used in this work is a part of the PFC database [26]. They were processed for a previous study dealing with prosodic variation in Swiss French [27]. For each of the 4 sites (regional accents), 4 female and 4 male speakers, born and raised in the city in which they were recorded, were selected. Based on ANOVA[1] tests, the age of the speakers is similar among the 4 groups of speakers ($F(3, 32) = 0.308$, n.s.), between male and female speakers ($F(1, 32) = 0.04$, n.s.) and between male and female speakers across the 4 groups ($F(5, 32) = 0.32$, n.s.).

The speakers were asked to read carefully a journalistic text (22 sentences, 398 words) and additionally to speak freely for 20-25 minutes in pairs. The entire reading text, and 3 minutes of monologue continuous speech for each speaker, were orthographically transcribed and automatically with the usual HMM technique used in forced alignment mode and implemented within EasyAlign tool [28], a plugin of the Praat software [29]. Alignments were manually checked and corrected by inspecting waveforms and spectrograms. All in all, the corpus is approximately 3 hours and 20 minutes long. Both types of speech, i.e. reading and free style, were used as unified speech for each speaker in our experimental setup. In Table 1 more information concerning the database is presented.

Furthermore, an experiment was conducted online to rate

[1] http://en.wikipedia.org/wiki/Analysis_of_variance

Table 1. DATABASE SUMMARY. *Ranges and mean age, duration of speech, total number of phones, syllables and tokens for each of the 4 groups of speakers.*

Accent Groups	Age ranges	Mean Age (s.d.)	Dur. (s.d.)	Total Phones (s.d.)	Total Sylls (s.d.)	Total Tokens (s.d.)
GE	21-59	44.3 (17.9)	2946 (25.6)	23564 (129)	10386 (63)	7296 (82)
MA	22-79	48.8 (27.6)	2773 (31.2)	22421 (131)	9863 (57)	6845 (61)
NE	25-78	52.5 (24.1)	3289 (27.1)	22150 (135)	9679 (57)	6740 (58)
NY	30-70	46.2 (17.1)	3002 (27.1)	22243 (118)	9721 (44)	6799 (44)

Age: in years, Duration: in seconds

Table 2. ACCENTS' DEGREE. *Mean and standard deviation of degree of accent (on a scale from 1 to 5) rated by 37 subjects.*

Accent Groups	Mean Accent Degree (s.d)
GE	2.57 (1.09)
MA	3.32 (0.97)
NE	3.37 (0.89)
NY	3.75 (0.86)

these 32 speakers with respect to their degree of regional accent. One sentence was chosen within the text and the corresponding audio was extracted for the 32 speakers. Theses extracts were randomly presented, in a website[2], to 37 subjects who were asked to rate the degree of accent of each speaker from *No accent* to *Marked accent* on a slider (with hidden values from 1 to 5). The mean value and standard deviation are shown by sites (accents) in Table 2. The degree of accent is different for the 4 groups ($F(3, 668) = 47.22, p < 0.001$). Post-hoc tests show significant difference between all groups except for the MA-NE pair.

4. PROPOSED ACCENT IDENTIFICATION

In this section the features which were used in our models are presented along with the two evaluated systems.

4.1. Acoustic features

Acoustic features are extracted at equally-spaced time instants using a sliding window approach. First, a simple energy-based voice activity detection (VAD) is performed to discard the non-speech parts. Second, 19 MFCC and log energy features together with their first- and second-order derivatives are computed over 20 ms Hamming windowed frames every 10 ms. Finally, cepstral mean and variance normalization (CMVN) is applied on the resulting 60-dimensional feature vectors. Let $\boldsymbol{O}$ denote the set of K acoustic feature vectors ($\boldsymbol{O} = \{\boldsymbol{o}^1, \boldsymbol{o}^2, \cdots, \boldsymbol{o}^K\}$) that is extracted from each utterance $\boldsymbol{\mathcal{O}}$.

4.2. Gaussian Mixture Modelling

We assume that acoustic feature vectors can be modelled by a Gaussian mixture model (GMM), which is a parametric probability density function represented as a weighted sum of C multivariate Gaussian components [30]:

$$p(\boldsymbol{o}|\boldsymbol{\Theta}_{\text{gmm}}) = \sum_{c=1}^{C} \omega_c \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right) , \tag{1}$$

where $\boldsymbol{\Theta}_{\text{gmm}} = \{\omega_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=\{1,\ldots,C\}}$ are the weights, the means and the covariances of the model. To meet our problem of regional accent recognition, each accent class ($i \in \{\text{GE}, \text{MA}, \text{NE}, \text{NY}\}$) is represented by a GMM $\boldsymbol{\mathcal{G}}_i$. The parameters $\boldsymbol{\Theta}_{\text{gmm}}$ can be estimated from training data using either (1) iterative expectation-maximization (EM) algorithm or (2) maximum a posteriori (MAP) adaptation from a well-trained prior model (known as universal background model). The latter is used in our work since it is more effective when the training suffers from the lack of data [19].

The universal background model $\boldsymbol{\mathcal{G}}_{ubm}$ is learned using the EM algorithm that aims to maximize the likelihood estimate of the parameters $\boldsymbol{\Theta}_{\text{ubm}}$. The enrolment of each accent-specific model $\boldsymbol{\mathcal{G}}_i$ is done using MAP adaptation, where only the means of the UBM are updated using the accent-specific samples.

At the test phase, the goal is to compute the similarity between the test utterance $\boldsymbol{\mathcal{O}}_t$ and each of the four accent-specific GMMs. This similarity is estimated using the log-likelihood ratio (LLR):

$$L\left(\boldsymbol{\mathcal{O}}_t, \boldsymbol{\mathcal{G}}_i\right) = \sum_{k=1}^{K_t}\left[\ln\left(p\left(\boldsymbol{o}_t^k \mid \boldsymbol{\mathcal{G}}_i\right)\right) - \ln\left(p\left(\boldsymbol{o}_t^k \mid \boldsymbol{\mathcal{G}}_{ubm}\right)\right)\right] . \tag{2}$$

The higher the value of $L\left(\boldsymbol{\mathcal{O}}_t, \boldsymbol{\mathcal{G}}_i\right)$, the greater is the probability that $\boldsymbol{\mathcal{O}}_t$ is produced by a speaker who has the accent i. Practically, a linear approximation [31] of Eq. 2 is used in our experiments.

4.2.1. *Total Variability Modelling*

Recently Dehak et *al.* [20] have proposed the total variability (TV) modelling technique, that achieved state-of-the-art performance on the task of speaker recognition. This framework is built on top of the GMM approach and relies on the definition of a single subspace that contains both within-class and channel variabilities. TV aims to extract low-dimensional vectors $\boldsymbol{w}_{i,j}$, so-called i-vectors, which are assumed to follow a normal distribution $\mathcal{N}\left(0, \boldsymbol{I}\right)$. More formally, this approach can be described in the GMM mean super-vector space by:

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{i,j} , \tag{3}$$

where $\boldsymbol{T}$ is the low-dimensional total variability subspace and $\boldsymbol{m}$ is the mean super-vector of the UBM $\boldsymbol{\mathcal{G}}_{ubm}$. $\boldsymbol{m}$ is obtained by concatenating the means $\boldsymbol{\mu}_c$ of all its components: $\boldsymbol{m} = \left[\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_c^T, \cdots, \boldsymbol{\mu}_C^T\right]^T$.

[2] http://www.labguistic.com

Table 3. PERFORMANCE SUMMARY. *This table reports the accuracy of the GMM and TV-SVM systems.*

System	GE	MA	NE	NY	Total Accuracy
GMM	23.7%	**38.5%**	19.6%	54.6%	33.4%
TV-SVM	**35.1%**	32.9%	**25.7%**	**63.4%**	**38.5%**

Furthermore, a set of preprocessing algorithms has been proposed to map the i-vectors into a more adequate space such as: (1) whitening that consists of normalizing the TV space such that the covariance of the i-vectors is turned into identity matrix, (2) length-normalization that aims to reduce the mismatch between training and testing i-vectors, (3) within-class covariance normalization (WCCN) [20] that normalize the within-class covariance matrix of training i-vectors. These three preprocessing techniques are used in our experiments.

Since the task of accent identification is a closed set problem, we applied a multi-class support vector machine (SVM) [32] classifier on the preprocessed i-vectors. In our preliminary experiments, we have found that the SVMs technique outperforms the simple cosine distance with more than 5% of absolute gain.

5. EXPERIMENTS

In this paper we are interested in validating two hypotheses. Firstly, whether biometric identification approaches could be used to tackle the regional accent identification task. Secondly, whether i-vectors, as a more discriminative representation of the feature space, could outperform the baseline GMM technique.

The experimental evaluation is conducted on the PFC dataset using a cross-validation technique: out of the eight speakers of a specific regional accent, seven of them are selected for the training, and the remaining one is used for the testing. This selection is done iteratively until all the possible combinations ($8^4 = 4096$ folds) are tested with the GMM and TV-SVM systems. GMMs are composed of 128 Gaussian components (no significant gain was shown with higher dimensionality), and the TV subspace has a rank of 100. The two systems were developed using Spear[3] [33], an open-source toolbox based on Bob [34].

In Table 3, the accuracy of the two systems is shown along with the total accuracy of each system. As can be seen, the TV-SVM system outperforms the GMM-based system in the 3 out of 4 accents. A relative improvement of 15.3% in the overall accent identification accuracy was achieved by TV-SVM over the GMM-based system. More precisely, the highest improvement can be seen in the GE accent where a 48.1% relative improvement was achieved. For the NE and NY accents, the TV-SVM system outperforms the GMM-based one by a relative improvement of 31.1% and 16.1% respectively. By contrast, in the case of the MA accent, the TV-SVM system cannot improve the accuracy over the GMM-based one, showing a 14.5% relative decrease in the performance. These results are confirmed by a Wilcoxon signed rank test [35] which was performed for each of the four regional accents and for the overall accuracy. The test showed that in all cases the difference in accuracy between the two systems is statistical significant at the 5% significance level.

	GE	MA	NE	NY
GE	21	**26**	17	24
MA	20	**33**	22	11
NE	**19**	7	17	18
NY	15	7	13	**42**

(a) GMM

	GE	MA	NE	NY
GE	**31**	17	21	19
MA	21	**28**	26	10
NE	**22**	4	**22**	19
NY	11	4	13	**48**

(b) TV-SVM

Fig. 1. AVERAGE CONFUSION MATRIX. *This figure illustrates the average confusion matrix among the four accents for both GMM and TV-SVM systems in terms of coutns.*

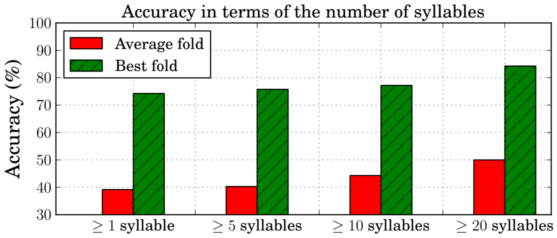

Fig. 2. ACCURACY IN TERMS OF NUMBER OF SYLLABLES. *This figure shows the accuracy rates of the TV-SVM system with the average and the best fold, evaluating utterances with equal or bigger number of syllables than a threshold.*

In Fig. 1 the two average confusion matrices are shown respectively for the two systems. In the case of the GMM-based system, it can be seen that the GE accent is mainly confused with NY and MA accents and the NE accent is confused in a high degree with GE and NY accents. These confusions are clearly overcome by the TV-SVM system as can be seen in the Fig. 1(b). For the NY accent which has a high accent degree (see Table 2), both systems managed to achieve very high identification rates. On the other hand, in the case of the GE accent, which has the lowest accent degree, only the TV-SVM system manages to perform well.

Fig. 2 shows the accuracy of the TV-SVM system for the best and average folds in respect to the size of the test utterances. In this figure, the accuracy in respect to the number of syllables can be seen for the cases of evaluating on utterances with equal or more syllables than: one (all test utterances of each fold, approximately 310 utterances), five (approximately 280 utterances), ten (approximately 210 utterances) and twenty (approximately 80 utterances). It is clearly shown that as the number of the syllables increases *i.e.* not using utterances with less syllables than a threshold, the accent identification accuracy improves. This can be contributed to

[3] https://pypi.python.org/pypi/bob.spear/1.1.2

the fact that the small utterances do not convey enough accent information so as to be correctly identified by the system.

The experimental results confirmed our two hypotheses showing that biometric identification approaches can be used to cast the RAI task and furthermore, total variability modelling outperforms the baseline GMM technique.

6. CONCLUSIONS AND DISCUSSION

The objective of this paper was to automatically recognize the speaker's accent among 4 regional Swiss French accents by using biometric identification techniques. To achieve this task, two different modelling techniques were investigated, the GMM baseline technique and the TV-SVM modelling. The results have validated our first hypothesis of using speaker identification techniques in the task of regional accent identification. It was shown that the TV-SVM system outperform the GMM baseline confirming our second hypothesis that i-vectors could create a more discriminative feature space and achieve a higher performance. The results were statistical significant according to Wilcoxon test. Furthermore, the accuracy of the TV-SVM system was gradually increasing when utterances with more syllables were used for the evaluation.

As noted in the introduction, this is a preliminary work on Swiss French regional accent identification, that the authors are interested in further developing. In the future we intent to deal with the problem of the lack of data by incorporating additional speakers to the database as soon as they are available in PFC project. Additionally, a more in depth analysis of the differences among the regional accents, concerning the phonetic and mainly prosodic characteristics of speech, will be conducted. In this way, we can identify and focus more on these differences, taking advantage of them for identifying the different regional accents. Furthermore, more appropriate features like shifted delta coefficients and prosody-specific ones could be used for this task, along with prosody-specific techniques, investigating their importance in respect to the acoustic-based ones used in this paper. Finally, the acoustic-based TV-SVM framework which was used in this work could be combined with more prosody-specific techniques in order to take advantage of the benefits of both of these approaches.

7. ACKNOWLEDGMENT

The research leading to this work has received funding from the Swiss National Science Foundation (SNSF) under the SIWIS and LOBI projects.

8. REFERENCES

[1] J. Laver, *Principles of Phonetics*, Cambridge University Press, Cambridge, 1994.

[2] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *IEEE ICASSP*, 2003.

[3] T. Bocklet, A. Maier, J. Bauer, F. Burkhardt, and E. Noth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," in *IEEE ICASSP*, 2008.

[4] I. Racine, S. Schwab, and S. Detey, "Accent(s) suisse(s) ou standard(s) suisse(s)? approche perceptive dans quatre régions de Suisse romande," in *La perception des accents du français hors de France*, A. Falkert, Ed. Mons: Éditions CIPA, 2013.

[5] A. Lodge, *French. From Dialect to Standard*, London, Routledge, 1993.

[6] C. Woehrling and P. Boula de Mareüil, "Identification of regional accents in french: perception and categorization," in *INTERSPEECH*, 2006.

[7] A. Leemann, *Comparative Analysis of Voice Fundamental Frequency Behavior of Four Swiss German Dialects*, Ph.D. thesis, Bern Universität, 2009.

[8] C. Teixeira, I. Trancoso, and A. Serralheiro, "Accent identification," in *INTERSPEECH*, 1996.

[9] F. Biadsy, *Automatic dialect and accent recognition and its application to speech recognition*, Ph.D. thesis, Columbia University, 2011.

[10] M.K. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," in *IEEE ICASSP*, 2010.

[11] A. Hanani M.J. Russell and M.J. Carey, "Human and computer recognition of regional accents and ethnic groups from british english speech," *Computer Speech and Language*, 2013.

[12] R. Huang, J.H.L. Hansen, and P. Angkititrakul, "Dialect/accent classification using unrestricted audio," *IEEE Trans. on Audio, Speech and Language Processing*, 2007.

[13] I. Mporas, T. Ganchev, and N. Fakotakis, "Phonotactic recognition of greek and cypriot dialects from telephone speech," in *SETN 2008, Advances in Artificial Intelligence, Lecture Notes in Computer Science, Springer Berlin/Heidelberg*, 2008.

[14] J.J. Humphries and P.C. Woodland, "Identification of foreign-accented french using data-mining techniques," in *INTERSPEECH*, 1997.

[15] F. Biadsy, J. Hirschberg, and M. Collins, "Dialect recognition using a phone-gmm-supervector-based svm kernel," in *INTERSPEECH*, 2010.

[16] H. Liang, J. Dines, and L. Saheer, "A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for hmm-based speech synthesis," in *IEEE ICASSP*, 2010.

[17] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and bic for speaker diarization," in *INTERSPEECH*, 2005.

[18] P.A. Torres-Carrasquillo, D.A. Reynolds, and J.R. Deller, "Language identification using gaussian mixture model tokenization," in *IEEE ICASSP*, 2002.

[19] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 2000.

[20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, 2011.

[21] B. Vieru-Dimulescu, P. Boula de Mareüil, and M. Adda-Decker, "Identification of foreign-accented french using data-mining techniques," in *International Workshop on Paralinguistic Speech*, 2007.

[22] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. on Speech and Audio Processing*, 1996.

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

[24] M.H. Bahari, R. Saeidi, H. Van hamme, and D. Van Leeuwen, "Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech," in *IEEE ICASSP*, 2013.

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

[26] J. Durand, B. Laks, and C. Lyche, *Phonologie, variation et accents du français*, Paris, Hermés, 2009.

[27] M. Avanzi, S. Schwab, P. Dubosson, and J.-P. Goldman, *La prosodie de quelques variétés de français parlées en Suisse romande*, in Simon, A. C. (éd.). La variation prosodique régionale en français, Bruxelles, De Boeck/Duculot, 2012.

[28] J.-P. Goldman, "Easyalign: an automatic phonetic alignment tool under praat," in *INTERSPEECH*, 2011.

[29] P. Boersma and D. Weenink, "Praat, v. 5.3.," http://www.fon.hum.uva.nl/praat/, 2012.

[30] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, 1995.

[31] O. Glembek, L. Burget, N. Dehak, N. Brümmer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *IEEE ICASSP*, 2009.

[32] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc., 1995.

[33] E. Khoury, L. El Shafey, and S. Marcel, "Spear: An open source toolbox for speaker recognition based on Bob," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[34] A. Anjos, L. El Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *ACM International Conference on Multimedia*, 2012.

[35] J.S. Milton and J.C. Arnold, *Introduction to probability and statistics*, McGraw-Hill, Inc., 3rd edition, 1995.