ON THE USE OF PHONE LOG-LIKELIHOOD RATIOS AS FEATURES IN SPOKEN LANGUAGE RECOGNITION

Mireia Diez, Amparo Varona, Mikel Penagarikano, Luis Javier Rodriguez-Fuentes and German Bordel

GTTS (http://gtts.ehu.es), Department of Electricity and Electronics,
University of the Basque Country UPV/EHU, 48940 Leioa, Spain.
e-mail: mireia.diez@ehu.es

ABSTRACT

This paper presents an alternative feature set to the traditional MFCC-SDC used in acoustic approaches to Spoken Language Recognition: the log-likelihood ratios of phone posterior probabilities, hereafter Phone Log-Likelihood Ratios (PLLR), produced by a phone recognizer. In this work, an iVector system trained on this set of features (plus dynamic coefficients) is evaluated and compared to (1) an acoustic iVector system (trained on the MFCC-SDC feature set) and (2) a phonotactic (Phone-lattice-SVM) system, using two different benchmarks: the NIST 2007 and 2009 LRE datasets. iVector systems trained on PLLR features proved to be competitive, reaching or even outperforming the MFCC-SDC-based iVector and the phonotactic systems. The fusion of the proposed approach with the acoustic and phonotactic systems provided even more significant improvements, outperforming state-of-the-art systems on both benchmarks.

Index Terms— Spoken Language Recognition, Phone Posterior Probabilities, Log-Likelihood Ratios, iVectors

1. INTRODUCTION

Spoken Language Recognition (SLR) tasks are commonly carried out using two main complementary approaches, based on *low-level* acoustic and *high-level* phonotactic features, respectively [1].

High-level phonotactic approaches use counts of phone n-grams to build a feature vector which feeds a classifier, typically Support Vector Machines (SVM) [2]. In *low-level* acoustic systems, the target language is modeled with information taken from the spectral characteristics of the audio signal. Among acoustic systems, Joint Factor Analysis (JFA) [3], which had been previously used in speaker recognition, was successfully applied to spoken language recognition in [4]. Recently, an approach derived from JFA, known as *Total Variability Factor Analysis* or, more briefly, *iVector* approach has been successfully applied to Language Recognition [5, 6].

Initially, the iVector technology was applied to language recognition using the concatenation of Mel-Frequency Cepstral Coefficients (MFCC) and Shifted Delta Cepstrum (SDC) features as input. Lately, other sets of features have been tested, as in [7], where prosodic features (pitch, energy and duration) were introduced, or in [8], where features of a Subspace Gaussian Mixture Model were used.

In this work, we propose a simple yet effective idea: using log-likelihood ratios of phone posterior probabilities produced by a phone decoder as features for an iVector system. The study of the effectiveness of these features has been carried out using two different benchmarks: the NIST 2007 and 2009 LRE narrow-band, conversational telephone speech databases [9] [10]. The system has been compared to and then fused with two state-of-the-art approaches: an iVector acoustic system and a phonotactic system.

The rest of the paper is organized as follows. Section 2 describes the proposed approach including details about the computation of the phone log-likelihood ratios used as features and the flavor of iVectors applied in this work. Section 3 describes the experimental setup and Section 4, the datasets used in the experiments. Section 5 presents results and compares the performance of the proposed approach to that of state-of-the-art approaches. Finally, conclusions are given in Section 6.

2. PHONE LOG-LIKELIHOOD RATIOS AS FEATURES FOR AN IVECTOR APPROACH

Given a phone decoder with a set of N phone units, each of them represented typically by means of a model of S states, the acoustic posterior probability of each state s $(1 \leq s \leq S)$ of each phone model i $(1 \leq i \leq N)$ at each frame t, $p_{i,s}(t)$, is directly provided by the phone decoder. Then, the acoustic posterior probability of a phone unit i at each frame t, is computed by adding the posteriors of its states:

This work has been supported by the University of the Basque Country, under grant GIU10/18 and project US11/06, by the Government of the Basque Country, under program SAIOTEK (project S-PE11UN065), and the Spanish MICINN, under *Plan Nacional de I+D+i* (project TIN2009-07446, partially financed by FEDER funds). Mireia Diez is supported by a 4-year research fellowship from the Department of Education, University and Research of the Basque Country.

$$p_i(t) = \sum_{\forall s} p_{i,s}(t) \quad (1)$$

These $p_i(t)$ are taken as likelihoods and log-likelihood ratios are computed assigning 0.5 prior to the target phoneme (the remaining 0.5 being distributed evenly among the $j \neq i$ phonemes). Therefore, the log-likelihood ratios at each frame t are computed as follows:

$$LLR_i(t) = \log \frac{p(x_t|i)}{\frac{1}{(N-1)} \sum_{\forall j \neq i} p(x_t|j)} \qquad i = 1, ..., N \quad (2)$$

The resulting N log-likelihood ratios per frame are the new Phone Log-Likelihood Ratio (PLLR) features. In our approach, an iVector system is trained on these features.

The total variability space (iVector) approach [11] has become state-of-the-art in speaker and language verification. Under the iVector modeling assumption, an utterance GMM supervector (stacking GMM mean vectors) is defined as:

$$M = m + Tw \quad (3)$$

where M is the utterance dependent GMM mean supervector, m is the utterance independent mean supervector, T is the total variability matrix (a low-rank rectangular matrix) and w is the so called iVector (a normally distributed low-dimensional latent vector). That is, M is assumed to be normally distributed with mean m and covariance TT^t. The latent vector w can be estimated from its posterior distribution conditioned to the Baum-Welch statistics extracted from the utterance and using the Universal Background Model (UBM). The iVector approach maps high-dimensional input data (a GMM supervector) to a low-dimensional feature vector (an iVector), hypothetically retaining most of the relevant information.

3. SYSTEM CONFIGURATION

In this work, two state-of-the-art SLR baseline systems are considered: (1) an acoustic iVector approach trained on MFCC-SDC features, which shares the modeling part with the PLLR-based system; and (2) a Phone-Lattice-SVM phonotactic approach based on expected counts of n-grams, computed from the same posterior probabilities used to compute PLLR features. In this section we address the configuration of the PLLR iVector system, along with the two baseline systems, the backend and fusion modules applied on the resulting scores and the evaluation measures that will be used in this work.

3.1. PLLR iVector system

As a first step to get the PLLR features, the open software Temporal Patterns Neural Network (TRAPs/NN) phone decoders, developed by the Brno University of Technology (BUT) for Czech (CZ), Hungarian (HU) and Russian (RU) [12] were applied. The BUT decoders feature 45, 61 and 52 phonetic units for Czech, Hungarian and Russian, respectively. For each unit, a three-state model was used, so three posterior probabilities per frame were calculated.

The BUT decoders produce a sequence of numbers representing the posterior probabilities $p_{i,s}(t)$ for each one of the three states s of each phone model i at each frame t, encoded in the following way:

$$x_{i,s}(t) = \sqrt{-2 \log p_{i,s}(t)} \quad (4)$$

Thus, the posterior probability $p_{i,s}(t)$ can be obtained as follows:

$$p_{i,s}(t) = e^{-\frac{(x_{i,s}(t))^2}{2}} \quad (5)$$

Before computing log-likelihood ratios, the non-phonetic units *int* (intermittent noise), *pau* (short pause) and *spk* (non-speech speaker noise) were integrated into a single non-phonetic unit model. Then, a single posterior probability was computed for each phone i ($1 \leq i \leq N$), by adding the posterior probabilities of all the states in the corresponding model (Equation 1). Finally the log-likelihood ratio for each phone was computed according to Equation 2, getting 43 (CZ), 59 (HU) and 50 (RU) log-likelihood ratios which we call PLLR features.

Voice activity detection was performed by removing the feature vectors whose highest PLLR value corresponded to the non-phonetic unit.

For the iVector system, a gender independent 1024-mixture UBM was estimated by the Maximum Likelihood criterion on the training dataset, using binary mixture splitting, orphan mixture discarding and variance flooring. The total variability matrix T was estimated as in [11], but using only data from target languages, according to [6].

A generative modeling approach was applied in the iVector feature space [6], the distribution of iVectors of each language being modeled by a single Gaussian set. Thus, the iVector scores were computed as follows:

$$score(f, l) = N(w_f; \mu_l, \Sigma) \quad (6)$$

where w_f is the iVector for target signal f, μ_l is the mean iVector for language l and Σ is a common (shared by all languages) within-class covariance matrix.

3.2. MFCC-SDC iVector system

In this case, the concatenation of MFCC and SDC coefficients under a 7-2-3-7 configuration was used as acoustic representation. Voice activity detection, GMM estimation and the total variability matrix training and scoring were performed as in the PLLR iVector approach.

3.3. Phonotactic systems

The three phonotactic systems presented in this work have been developed under the phone-lattice-SVM approach [2]. As in [13], given an input signal, an energy-based voice activity detector was applied in first place, which split and removed long-duration non-speech segments. Then, the BUT TRAPs/NN phone decoders for CZ, HU and RU [12] were applied.

BUT decoders were configured to produce phone posteriors (see Equation 4) that were converted to phone lattices by means of HTK [14] along with the BUT recipe [12]. Then, expected counts of phone n-grams were computed using the *lattice-tool* of SRILM [15]. Finally, a SVM classifier was applied, SVM vectors consisting of expected frequencies of phone n-grams (up to $n = 3$), weighted as in [16]. A sparse representation was used, which involved only the most frequent features according to a greedy feature selection algorithm [13]. L2-regularized L1-loss support vector regression was applied, by means of LIBLINEAR [17].

3.4. Backend/Fusion parameters

The backend serves as a precalibration stage that transforms the space of scores to get a reliable computation of the true class probabilities. Besides, when the set of languages for which models have been trained does not match the set of target languages, the backend maps the available scores to the space of target languages (the trained languages being used as *anchor* models) [18]. In the case of NIST LRE datasets, different models can be trained for dialects of a target language or for different sources (e.g. telephone conversational speech vs. radio broadcast speech), and non-target languages can be modeled as well.

Based on preliminary experiments on the development set of each database, a zt-norm and a discriminative Gaussian backend were applied to scores. The FoCal toolkit was used to estimate and apply the calibration/fusion models [19] [20].

3.5. Evaluation measures

In this work, systems will be compared in terms of: (1) the average cost performance C_{avg} as defined by NIST, and (2) the so called C_{LLR} [19], an alternative performance measure used in NIST evaluations.

4. DATASETS

4.1. NIST 2007 LRE

The NIST 2007 LRE [9] defined a spoken language recognition task for conversational speech across telephone channels, involving 14 target languages (see Table 1). Some languages featured various dialects or accents. French was the only non-target language.

Training and development data used in this work were limited to those distributed by NIST to all 2007 LRE participants: (1) the Call-Friend Corpus[1]; (2) the OHSU Corpus provided by NIST for the 2005 LRE[2]; and (3) the development corpus provided by NIST for the 2007 LRE[3]. A set of 23 languages/dialects was defined for training. For development purposes, 10 conversations per language were randomly selected, and the remaining conversations (amounting to around 968 hours) were used for training. Development conversations were further divided into 30-second speech segments (see Table 1 for more details). Results reported in this paper have been computed on the subset of 30-second speech segments of the test set for the closed-set condition (2158 segments), which was the primary task in the NIST 2007 LRE.

Table 1. 2007 NIST LRE core condition: training data (hours), development and evaluation data (number of 30s speech segments), disaggregated for target and non-target languages.

	Hours	**# 30s segments**	
Language	Train	Devel	Eval
Arabic	52.59	179	80
Bengali	5.0	76	80
Chinese	166.12	567	398
English	143.7	288	240
Farsi	46.22	225	80
German	57.03	173	80
Industani	64.35	243	240
Japanese	79.11	141	80
Korean	72.86	150	80
Russian	5.0	66	160
Spanish	117.35	531	240
Tamil	58.18	165	160
Thai	5.0	64	80
Vietnamese	46.7	205	160
Non-Target	48.74	-	-
TOTAL	967.95	3073	2158

4.2. NIST 2009 LRE

The NIST 2009 LRE featured 23 target languages [10], involving 12 target languages for which Conversational Telephone Speech (CTS) was available in the NIST 2007 LRE dataset, plus 11 new target languages (see Table 2 for more details). For this evaluation, Broadcast Narrow-Band Speech was provided consisting mostly of telephone calls included in Voice of America (VOA) broadcasts.

Training and development data used in this work were limited to those distributed by NIST to all 2009 LRE participants. A set of 64 languages/dialects was defined for training models. Each of them was mapped either to a target language of the NIST 2009 LRE or to the set of non-target languages (including 26 languages/dialects).

For languages appearing in VOA recordings, the longest speech segment out of each file was added to the train dataset.

[1] See http://www.ldc.upenn.edu/.

[2] OHSU Corpora, http://www.ohsu.edu/.

[3] See http://www.itl.nist.gov/iad/mig/tests/lre/2007/.

Table 2. 2009 NIST LRE core condition: training data (hours), development and evaluation data (number of 30s speech segments), disaggregated for target and non-target languages.

	Hours		**# 30s cuts**			
	Train		Devel			Eval
Language	2007 (CTS)	2009 (VOA)	2007 (CTS) devel	2007 (CTS) eval	2009 (VOA)	2009
Amharic	-	58.31	-	-	262	398
Bosnian	-	5.63	-	-	259	355
Cantonese	5.0	2.45	83	80	104	378
Creole	-	7.21	-	-	256	323
Croatian	-	6.45	-	-	190	376
Dari	-	69.05	-	-	276	389
EngAmerican	130.7	9.01	204	80	230	896
EngIndian	13.0	-	84	160	-	574
Farsi/Persian	46.22	25.16	225	80	294	390
French	48.74	67.91	222	80	293	395
Georgian	-	4.32	-	-	166	399
Hausa	-	48.31	-	-	274	389
Hindi	59.35	10.06	174	160	178	667
Korean	72.86	5.70	150	80	250	463
Mandarin	151.1	32.4	331	158	230	1015
Pashto	-	184.3	-	-	281	395
Portuguese	-	25.74	-	-	240	397
Russian	5.0	147.76	66	160	299	511
Spanish	117.35	45.44	531	240	242	385
Turkish	-	6.67	-	-	289	394
Ukrainian	-	5.59	-	-	281	388
Urdu	5.0	36.60	69	80	299	379
Vietnamese	46.7	9.5	205	160	240	315
Non-Target	266.92	408.82	-	-	-	-
TOTAL	967.95	1222.39	2344	1518	5433	10571

For development, some materials taken from the development and evaluation datasets of the NIST 2007 LRE were used (see Table 1). For languages appearing in VOA, besides the audited segments provided by NIST, additional randomly extracted speech segments lasting around 30-seconds (specifically, between 25 and 35 seconds) were used. Evaluation was carried out on the NIST 2009 LRE evaluation corpus, specifically on the 30-second, closed-set condition (primary evaluation task) (see Table 2 for more details).

5. RESULTS

To evaluate the performance of the iVector system trained on PLLR features, a detailed study was carried out for the NIST 2007 LRE dataset. In order to further evaluate the proposal, a second series of experiments was then carried out on the NIST 2009 LRE dataset.

5.1. Results on NIST 2007 LRE

To optimize the performance of the iVector system trained on PLLR features, we first tested on the PLLR features some state-of-the-art techniques applied to other parameterizations.

5.1.1. Dynamic coefficients

Dynamic coefficients of acoustic representations have proven to be effective in speech, speaker and language recognition, so we first evaluated the usefulness of this technique when applied to PLLRs. Three different parameterizations based on the PLLR features were tested. The first one comprised only PLLR static features. In the second one, the feature vector was augmented with first-order dynamic coefficients (PLLR+Δ). In the third one, the feature vector was augmented with first and second order dynamic coefficients (PLLR+Δ+$\Delta\Delta$). Tests were carried out using the BUT phone decoders. Similar results and conclusions can be drawn with any of the decoders. Therefore, for the sake of clarity, only the results for the BUT HU decoder are shown in Table 3.

Table 3. $C_{avg} \times 100$ and C_{LLR} performance for iVector systems using HU PLLR, PLLR+Δ and PLLR+Δ+$\Delta\Delta$ features, on the LRE07 primary evaluation task.

System	$C_{avg} \times 100$	C_{LLR}
PLLR	3.45	0.564
PLLR+Δ	2.66	0.382
PLLR+Δ+$\Delta\Delta$	3.60	0.506

As shown in Table 3, adding first order dynamic coefficients improved significantly the performance of the system, whereas second order deltas had the opposite effect, leading to degraded performance. Therefore, in the following experiments, PLLR+Δ were used as features in the PLLR-iVector approach.

5.1.2. Parameterization techniques

Degradation of SLR performance is related in most cases to some kind of language independent variability of audio utterances (typically channel/noise variability). Several techniques can be applied to minimize the effect of such variability. Among them, Feature Normalization [21] and Feature Warping [22] are two of the most effective (the latter, commonly used in speaker recognition). However, as shown in Table 4, no improvement in performance was attained when applying any of these techniques under the PLLR-iVector approach.

Table 4. $C_{avg} \times 100$ and C_{LLR} performance for iVector systems using HU PLLR+Δ features, with: (a) no noise reduction technique, (b) Feature Normalization (FN) and (c) Feature Warping (FW), on the LRE07 primary evaluation task.

System	$C_{avg} \times 100$	C_{LLR}
PLLR+Δ	2.66	0.382
PLLR+Δ +FN	2.95	0.436
PLLR+Δ +FW	3.21	0.435

5.1.3. Overall results

Results obtained under the PLLR-iVector approach with the three BUT decoders, using PLLR+Δ features are shown in Table 5. The performance of state-of-the-art acoustic and phonotactic systems is also presented and compared to that of the PLLR-iVector systems, along with all the possible fusions, to study their complementarity.

Table 5. $C_{avg} \times 100$ and C_{LLR} performance for iVector baseline systems using acoustic features (MFCC-SDC), iVector systems with PLLR+Δ features, phonotactic baseline systems and the fusion of them, for each of the BUT decoders, on the LRE07 primary evaluation task.

System			$C_{avg} \times 100$	C_{LLR}
iVector MFCC-SDC		(a)	2.85	0.407
CZ	Phonotactic	(b1)	2.94	0.440
	iVector PLLR+Δ	(c1)	4.18	0.550
Fusion	(a)+(b1)		1.22	0.189
	(a)+(c1)		1.95	0.280
	(b1)+(c1)		1.79	0.257
	(a)+(b1)+(c1)		1.24	0.176
HU	**Phonotactic**	**(b2)**	**2.08**	**0.310**
	iVector PLLR+Δ	(c2)	2.66	0.382
Fusion	(a)+(b2)		1.08	0.152
	(a)+(c2)		1.40	0.215
	(b2)+(c2)		1.20	0.166
	(a)+(b2)+(c2)		**0.82**	**0.124**
RU	Phonotactic	(b3)	2.69	0.383
	iVector PLLR+Δ	(c3)	4.08	0.549
Fusion	(a)+(b3)		1.13	0.182
	(a)+(c3)		1.72	0.265
	(b3)+(c3)		1.76	0.240
	(a)+(b3)+(c3)		1.10	0.163

If we focus on single-systems, the best performance was yielded by the HU phonotactic system, which attained 2.08 $C_{avg} \times 100$, followed by the HU PLLR+Δ iVector system, which outperformed the acoustic iVector based on MFCC-SDC features. The RU and CZ phonotactic systems performed worse that the HU phonotactic system. The performance of the RU PLLR+Δ and CZ PLLR+Δ systems degraded in the same proportion as the corresponding RU and CZ phonotactic systems.

Regarding the fused systems, similar conclusions can be drawn from any of the CZ/RU/HU sets of results. The fusion of the acoustic iVector system and the PLLR+Δ iVector system yielded very good results. As may be expected, the fusion of the phonotactic and the acoustic iVector systems provided very competitive performance, closely followed by the fusion of the phonotactic and PLLR+Δ systems, which, though sharing a common source (the phone posterior probabilities provided by BUT decoders), proved to be quite complementary. Finally, the best performance was always yielded by the fusion of the three systems (remarkably, for HU: 0.82 $C_{avg} \times 100$ and 0.124 C_{LLR}), meaning that PLLR features provide complementary information to both acoustic and phonotactic approaches.

5.2. Results on NIST 2009 LRE

Based on the performance attained on the NIST 2007 LRE dataset, for the NIST 2009 LRE dataset results will be only reported for the HU BUT decoder. Note that better results could probably be reached for this dataset using a different decoder, as previous experiments with phonotactic systems suggest [18].

Table 6 shows the performance of the baseline systems and the proposed approach on the NIST 2009 LRE primary evaluation task. First, note that the three systems attained similar results. Fusion performance was consistent with the results obtained for the NIST 2007 LRE dataset. The fusion of the phonotactic and MFCC-SDC iVector systems provided the best pairwise performance, followed by the fusion of the PLLR+Δ iVector and phonotactic systems. The fusion of the PLLR+Δ iVector and MFCC-SDC iVector systems was also competitive. Once again, the best performance (outperforming other state-of-the-art systems reported in the literature for the same dataset) was achieved when fusing the three systems: 1.48 $C_{avg} \times 100$.

Table 6. $C_{avg} \times 100$ and C_{LLR} performance for the baseline phonotactic and iVector systems, the PLLR+Δ iVector system and the fusion of them, on the LRE09 primary evaluation task.

System			$C_{avg} \times 100$	C_{LLR}
iVector MFCC-SDC		(a)	2.70	0.535
HU	Phonotactic	(b)	2.49	0.502
	iVector PLLR+Δ	(c)	2.42	0.505
Fusion	(a)+(b)		1.67	0.346
	(a)+(c)		1.79	0.392
	(b)+(c)		1.69	0.357
	(a)+(b)+(c)		**1.48**	**0.321**

6. CONCLUSIONS

In this paper, a new set of features suitable for iVector language recognition, the so called Phone Log-Likelihood Ratios (PLLR), has been presented and evaluated. The study of the features performed on the NIST 2007 LRE database showed that best results were attained when using the features and their first order dynamic coefficients. The analysis of the usefulness of these features has been extended to the NIST 2009 LRE dataset, and compared to state-of-the-art phonotactic and acoustic approaches. The PLLR+Δ iVector system outperformed an MFCC-SDC-based iVector system on both benchmarks.

The fusion of the acoustic and PLLR-based iVector systems has proved to be effective, yielding (in terms of C_{avg}) between 33% and 50% relative improvement with regard to the acoustic iVector system. Moreover, though sharing the source of features (phone posteriors provided by phone decoders), the fusion of the PLLR iVector system with a phono-

tactic system was also fruitful, providing between 32% and 42% relative improvement with regard to the phonotactic system. Finally, the fusion of the three approaches yielded the best results in experiments on both datasets, outperforming the fusion of two state-of-the-art (phonotactic and MFCC-SDC iVector) systems. These results suggest that the proposed PLLR iVector approach contributes complementary information to both acoustic and phonotactic approaches.

Current work involves evaluating the PLLR features on other benchmarks and exploring the performance that can be attained under other modeling approaches and/or configurations.

7. REFERENCES

[1] P.A. Torres-Carrasquillo, E. Singer, T. Gleason, A. McCree, D.A. Reynolds, F. Richardson, and D.E. Sturim, "The MITLL NIST LRE 2009 language recognition system," in *Proceedings of ICASSP 2010*, pp. 4994–4997.

[2] W. M. Campbell, F. Richardson, and D. A. Reynolds, "Language Recognition with Word Lattices and Support Vector Machines," in *Proceedings of IEEE ICASSP*, Honolulu, Hawaii, USA, 2007, pp. 15–20.

[3] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep. Technical Report CRIM-06/08-13, CRIM, 2005, [Online]. Available: http://www.crim.ca/perso/patrick.kenny/.

[4] F. Castaldo, S. Cumani, P. Laface, and D. Colibro, "Language Recognition Using Language Factors," in *Proceedings of Interspeech*, Brighton, UK, September 2009, pp. 176–179.

[5] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *Proceedings of the Interspeech 2011*, Firenze, Italy, 2011, pp. 857–860.

[6] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space," in *Proceedings of Interspeech*, Firenze, Italy, 2011, pp. 861–864.

[7] D. Martínez, L. Burget, L. Ferrer, and N.S. Scheffer, "iVector-based Prosodic System for Language Identification," in *Proceedings of ICASSP*, Japan, 2012, pp. 4861–4864.

[8] O. Plchot, M. Karafiát, N. Brümmer, O. Glembek, P. Matejka, and E. de Villiers J. Cernocký, "Speaker vectors from Subspace Gaussian Mixture Model as complementary features for Language Identification," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 2012, pp. 330–333.

[9] A. F. Martin and A. N. Le, "NIST 2007 Language Recognition Evaluation," in *Proceedings of Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 016*, 2008.

[10] A. Martin and C. Greenberg, "The 2009 NIST Language Recognition Evaluation," in *Odyssey 2010 - The Speaker and Language Recognition Workshop, paper 030*, Brno, Czech Republic, 2010, pp. 165–171.

[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, May 2011.

[12] P. Schwarz, *Phoneme recognition based on long temporal context*, Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.

[13] M. Penagarikano, A. Varona, L.J. Rodriguez-Fuentes, and G. Bordel, "Dimensionality Reduction for Using High-Order n-grams in SVM-Based Phonotactic Language Recognition," in *Proceedings of Interspeech 2011*, Firenze, Italy, August 28-31 2011, pp. 853–856.

[14] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Lui, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Entropic, Ltd., Cambridge, UK, 2006.

[15] A. Stolcke, "SRILM - An extensible language modeling toolkit," in *Proceedings of Interspeech*, November 2002, pp. 257–286.

[16] F. Richardson and W. Campbell, "Language recognition with discriminative keyword selection," in *Proceedings of ICASSP*, 2008, pp. 4145–4148.

[17] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008, http://www.csie.ntu.edu.tw/∼cjlin/liblinear.

[18] M. Penagarikano, A. Varona, M. Diez, L.J. Rodriguez Fuentes, and G. Bordel, "Study of Different Backends in a State-Of-the-Art Language Recognition System," in *Interspeech 2012*, Portland, Oregon, USA, 9-13 September 2012.

[19] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection," *Computer, Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[20] FoCal, *Toolkit for Evaluation, Fusion and Calibration of statistical pattern recognizers*, 2008, http://sites.google.com/site/nikobrummer/focal.

[21] Fu-Hua Liu, Richard M. Stern, Xuedong Huang, and Alejandro Acero, "Efficient cepstral normalization for robust speech recognition," in *Proceedings of the workshop on Human Language Technology*, Stroudsburg, PA, USA, 1993, HLT '93, pp. 69–74.

[22] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, 2001, pp. 213–218.