Regularized Subspace n-Gram Model for Phonotactic iVector Extraction

Mehdi Soufifar[1,2], *Lukáš Burget*[1], *Oldřich Plchot*[1], *Sandro Cumani*[1,3], *Jan "Honza" Černocký*[1]

[1] Brno University of Technology, BUT Speech@FIT and IT4I Centre of Excellence, Czech Republic
[2] Department of Electronics and Telecommunications, NTNU, Trondheim, Norway
[3] Politecnico di Torino, Italy

qsoufifar@stud.fit.vutbr.cz, {burget,cumani,cernocky,iplchot}@fit.vutbr.cz

Abstract

Phonotactic language identification (LID) by means of n-gram statistics and discriminative classifiers is a popular approach for the LID problem. Low-dimensional representation of the n-gram statistics leads to the use of more diverse and efficient machine learning techniques in the LID. Recently, we proposed phototactic iVector as a low-dimensional representation of the n-gram statistics. In this work, an enhanced modeling of the n-gram probabilities along with regularized parameter estimation is proposed. The proposed model consistently improves the LID system performance over all conditions up to 15% relative to the previous state of the art system. The new model also alleviates memory requirement of the iVector extraction and helps to speed up subspace training. Results are presented in terms of C_{avg} over NIST LRE2009 evaluation set.

Index Terms: Language identification, Subspace modeling, Subspace multinomial model

1. Introduction

State–of–the–art approaches to language identification (LID) can be mainly divided into two main categories: phonotactic LID and acoustic LID [1]. The phonotactic approach comprises techniques that use linguistic abstraction in speech modeling, while acoustic models try to infer the language of an utterance by directly modeling the spectral content of the utterance. This paper focuses on the phonotactic approach.

A successful representation of the phonetic content of utterances are n-gram statistics, which are often used as features for different language classifiers. However, the huge size of n-gram statistics poses some serious limitations on the choice of the LID backend classifier. Many solutions have been proposed to deal with the problem of n-gram vectors dimensionality. In [2], discriminative selection of the n-grams was proposed to discard less relevant n-grams. Many other phonotactic LID systems use principal component analysis (PCA) to reduce the dimensionality of the n-gram vectors [3, 4, 5]. We recently proposed a feature extraction technique based on subspace modeling of multinomial distribution parameters [6], where we showed that our approach outperforms former state of the art techniques based on n-gram statistics. This technique is inspired by the idea of iVector in acoustic speaker identification (SID) [7], where a low–dimensional vector is used to represent an utterance–dependent GMM supervector. In our context, we use a low–dimensional vector to represent the parameters of an utterance–dependent n-gram model.

The iVector extraction procedure presented in our previous work [6] was based on simpler Subspace Multinominal Model (SMM), where we assumed that n-grams are independent events generated from a single multinomial distribution and iVectors were computed as to maximize the likelihood of the observed n-grams. While this approach allows to obtain good results, the corresponding objective function is not directly related to the likelihood of the observed phoneme sequences. This is because the n-grams observed in a phoneme sequence are not independent. Using an n-gram model, likelihood of a phoneme sequence can be calculated as a product of the conditional probabilities of the individual phonemes given their histories. Such likelihood function is maximized in order to extract phonotactic iVectors using the Subspace n-Gram Model (SnGM) proposed in this work.

We found SnGM to be prone to over-fitting especially for short sequences, where only few different n-grams were observed. This was also one of the reasons for the former use of SMM, which is more robust to over-fitting. We show that this problem can be mitigated using regularization applied for both the subspace training and iVector extraction, which results in the superior performance of the newly proposed SnGM technique.

The paper is organized as follows: Section 2 describes the multinomial subspace model and details the subspace training and iVector extraction procedure. Section 3 describes our experimental setup. and compares the proposed method with PCA–based techniques and the multinomial model in [6]. An analysis of the model parameters is given in Section 4. Experimental results are reported in Section 5 and conclusions are drawn in Section 6.

2. Subspace multinomial model

In phonotactic LID, every speech utterance is tokenized to a sequence of phoneme labels. The n-gram model assumes that the probability of observing a phoneme is dependent only on the $n-1$ previous observed tokens. The log–likelihood of a

This work was partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015, by Czech Ministry of Trade and Commerce project No. FR-TI1/034, and by European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070). Sandro Cumani was supported by The European Social Fund (ESF) in the project Support of Interdisciplinary Excellence Research Teams Establishment at BUT (CZ.1.07/2.3.00/30.0005).

sequence of phonemes $l_1 \dots l_M$ can therefore be computed as

$$\log P(l_1 l_2 l_3 ... l_M) = \sum_i \log P(l_i | l_{i-n+1} l_{i-n+2} \dots l_{i-1}) \tag{1}$$

In order to model the phoneme generation process, we assume that the conditional distribution of a phoneme l given a history h is a multinomial distribution with parameters ϕ_{hl}, i.e.

$$\log P(l|h) = \log \phi_{hl}, \tag{2}$$

with $\phi_{hl} > 0$ and $\sum_l \phi_{hl} = 1$. The joint log–likelihood of a sequence of phonemes $l_1 \dots l_M$ can then be computed as

$$\log P(l_1 l_2 l_3 ... l_M) = \sum_i \log P(l_i | h_i) = \sum_i \log \phi_{h_i l_i}, \tag{3}$$

where $h_i = (l_{i-n+1} l_{i-n+2} \dots l_{i-1})$ denotes the history for the observed phoneme l_i. The ν_{hl} denotes number of times the n-gram hl (i.e. phoneme l with history h) appears in the phoneme sequence, we can rewrite (3) as

$$\log P(l_1 l_2 l_3 ... l_M) = \sum_h \sum_l \nu_{hl} \log \phi_{hl}. \tag{4}$$

It is worth noting the difference between (3) and the objective that was maximized to obtain iVectors in [6] as:

$$\sum_{i=1}^{M} \log P(h_i, l_i) = \sum_h \sum_l \nu_{hl} \log \hat{\phi}_{hl}, \tag{5}$$

where n-grams were assumed to be generated independently from a single multinomial distribution (i.e. $\sum_h \sum_l \hat{\phi}_{hl} = 1$). This objective allows to obtain good performance. However, the corresponding iVectors do not maximize the likelihood of the observed phoneme sequence. In the following, we show how to build a phonotactic iVector extractor where iVectors are estimated in order to maximize the likelihood of the observed phoneme sequences under the n-gram model assumptions.

Our first step towards the phonotactic iVector extractor is to make assumption that, phoneme sequence from each utterance s was generated from an utterance–specific n-gram distribution. Next, we assume that, the parameters of the corresponding multinomial distributions $\phi_{hl}(s)$ can be represented as

$$\phi_{hl}(s) = \frac{\exp(m_{hl} + \mathbf{t}_{hl}\mathbf{w}(s))}{\sum_i \exp(m_{hi} + \mathbf{t}_{hi}\mathbf{w}(s))}, \tag{6}$$

where m_{hl} is the log-probability of n-gram hl calculated over all the training data, $\mathbf{t}_{hl}$ is a row of a low–rank rectangular matrix $\mathbf{T}$ and $\mathbf{w}(s)$ is utterance–specific low–dimensional vector, which can be seen as low–dimensional representation of the utterance–specific n-gram model. The parameters $\{m_{hl}\}$ and the matrix $\mathbf{T}$ are the parameters of the proposed SnGM. Given these parameters, $\mathbf{w}(s)$ maximizing log-likelihood in (3) can be taken as the phonotactic iVector representing the an utterance s. Before iVectors can be extracted, however, the SnGM parameters have to be trained on a set of training utterances. This is done in an iterative EM-like process alternating between maximum likelihood (ML) updates of vectors $\mathbf{w}(s)$ (one for each training utterance s) and ML updates of SnGM parameters.

In the case of standard GMM based iVectors, the utterance–dependent parameters similar to $\mathbf{w}(s)$ are treated as latent random variables with standard normal priors. The subspace parameters are then trained using standard EM algorithm, where the M-step integrates over the latent variable posterior distributions from the E-step. Unfortunately, calculation of posterior distribution for $\mathbf{w}(s)$ is intractable in the case of SnGM. Instead, SnGM parameters are updated using only $\mathbf{w}(s)$ point estimates, which can negatively affect the robustness of SnGM parameter estimation. To mitigate this problem, we propose to regularize the ML objective function using L2 regularization terms for both the subspace matrix $\mathbf{T}$ and the vectors $\mathbf{w}(s)$. This corresponds to imposing an isotropic Gaussian prior on both the SnGM parameters and $\mathbf{w}(s)$, and obtaining MAP rather than ML point estimates. This is in contrast to our previous work [6], where only ordinary ML estimates of SnGm parameters and iVectors were used. In order to train our model, we maximize the regularized likelihood function

$$\sum_{s=1}^{S} \sum_h \sum_l \nu_{hl}(s) \log \phi_{hl}(s) - \frac{1}{2}\lambda \left\|\mathbf{t}_{hl}\right\|^2 - \frac{1}{2}\lambda \|\mathbf{w(s)}\|^2), \tag{7}$$

where the sum extends over all S training utterances. The term λ is the regularization coefficient for both the model parameters $\mathbf{T}$ and for $\mathbf{w}(s)$. Notice that we should regularize both $\mathbf{T}$ and $\mathbf{w}$ since limiting magnitude of $\mathbf{T}$ without regularizing $\mathbf{w}$ would be compensated by a dynamic range increase in $\mathbf{w}$.

2.1. Parameter estimation

The model parameters m_{hl} are shared for all utterances and can be initialized as the logarithm of the conditional probability of a phoneme given its history computed over all training utterances:

$$m_{hl} = \log \left(\frac{\sum_s \nu_{hl}(s)}{\sum_s \sum_i \nu_{hi}(s)} \right). \tag{8}$$

In the following, we assume that the terms m_{hl} do not require retraining. In order to alternately maximize the objective function (7) with respect to $\mathbf{T}$ and $\mathbf{w}$, we adapt the approach proposed in [8]. For a fixed $\mathbf{T}$, Newton Raphson-like update of $\mathbf{w}(s)$ is given by:

$$\mathbf{w}(s)^{new} = \mathbf{w}(s) + \mathbf{H}^{-1}_{w(s)} \boldsymbol{\nabla}_{w(s)}, \tag{9}$$

where the $\boldsymbol{\nabla}_{w(s)}$ is the gradient of the objective function (7) with respect to $\mathbf{w}(s)$

$$\boldsymbol{\nabla}_{w(s)} = \sum_h \sum_l \mathbf{t}^T_{hl} (\nu_{hl}(s) - \phi^{old}_{hl}(s) \sum_i \nu_{hi}(s)) - \lambda \mathbf{w}(s), \tag{10}$$

where the terms $\phi^{old}_{hl}(s)$ are the model parameters computed from the current estimate of $\mathbf{w}(s)$. $\mathbf{H}_{w(s)}$ is an approximation to the Hessian matrix proposed in [8] as

$$\mathbf{H}_{(\mathbf{w}(s))} = \sum_h \sum_l \mathbf{t}^T_{hl} \mathbf{t}_{hl} \max(\nu_{hl}(s), \phi^{old}_{hl}(s) \sum_i \nu_{hi}(s)) - \lambda \mathbf{I}. \tag{11}$$

Similarly, to update the $\mathbf{T}$ matrix, we keep all $\mathbf{w}(s)$ fixed and update each row of $\mathbf{T}$ as

$$\mathbf{t}^{new}_{hl} = \mathbf{t}_{hl} + \boldsymbol{\nabla}_{t_{hl}} \mathbf{H}^{-1}_{hl}, \tag{12}$$

where $\boldsymbol{\nabla}_{t_{hl}}$ is the gradient of the objective function (7) with respect to the row $\mathbf{t}_{hl}$ of $\mathbf{T}$

$$\boldsymbol{\nabla}_{t_{hl}} = \sum_s (\nu_{hl}(s) - \phi^{old}_{hl}(s) \sum_i \nu_{hi}(s)) \mathbf{w}(s)^T - \lambda \mathbf{t}_{hl}, \tag{13}$$

and

$$\mathbf{H}_{t_{hl}} = \sum_s \max(\nu_{hl}(s), \phi_{hl}^{old}(s) \sum_i \nu_{hi}(s))\mathbf{w}(s)\mathbf{w}(s)^T - \lambda\mathbf{I}. \quad (14)$$

Notice that in (13), since we only need the n-gram statistics corresponding to n-gram history h, there is no need to load the whole vector of n-gram statistics. This reduces memory overhead of the $\mathbf{T}$ matrix update. Moreover, update of $\mathbf{T}$ rows belonging to different histories are completely independent, which simplifies parallel estimation of $\mathbf{T}$.

In our experiments Matrix $\mathbf{T}$ is initialized with small random numbers. Update of $\mathbf{T}$ or $\mathbf{w(s)}$ may fail to increase the objective function in (7). In that case, we keep backtracking by halving the update step. In case the objective function did not improve after certain number of backtracking, we retain the value of $\mathbf{t}_{hl}$ or $\mathbf{w}(s)$ from the previous iteration. The iterative parameter estimation continues until the change in the objective function becomes negligible. Once the SnGM is trained and fixed, it can be used to extract iVectors from new utterances by iteratively applying $\mathbf{w}(s)$ update formulas (9)-(11).

3. Experimental setup

To keep the results comparable to previously reported ones in [6], we report performance of the system over NIST LRE2009. We briefly explain the system description and the tuning. Interested readers are referred to the corresponding detailed system description [9].

3.1. Data

The LRE09 task comprises 23 languages. The EVAL set contains telephone data and narrowband broadcast data. The training data is divided into two sets denoted as TRAIN and DEV, both of which comprises data from 23 languages corresponding to the target list of the NIST LRE09 task [10]. The TRAIN set is filtered in order to keep at most 500 utterances per language as proposed in [9], resulting in 9763 segments (345 hours of recording). This allows to have almost balanced amounts of training data per language, thus avoiding biasing the classifiers toward languages with lots of training data. The DEV set contains 38469 segments mainly from the previous NIST LRE tasks plus some extra longer segments from the standard conversational telephone speech (CTS) databases (CallFriend, Switchboard, etc.) and voice of America (VOA). The TRAIN and the DEV sets contain disjoint sets of speakers. The DEV set is used to tune parameters and score calibration in the backend. A full description of the used data is given in [9].

3.2. Vector of n-gram counts

The n-gram counts were extracted using the Brno university of technology (BUT) Hungarian phone recognizer, which is an ANN/HMM hybrid [11]. The Hungarian phoneme list contains 51 phonemes. We map short and long variations of similar phonemes to the same token, obtaining 33 phonemes. This results in $33^3 = 35937$ 3-grams. Since neither 2-grams nor 1-grams improved the system performance we use only 3-gram counts. The 3-gram expected counts are extracted from phone lattices generated by the Hungarian phone recognizer.

3.3. Back end

We showed in [12] that iVector normalization is necessary to good LID performance using phonotactic iVectors. For this work, after mean removal, length normalized iVectors are used to train 23 logistic regression (LR) classifiers in one-vs-all configuration using LIBLINEAR[1]. The scores generated by 23 LR classifiers are calibrated on DEV data by means of a linear generative model followed by a multi-class LR as described in [13].

Table 1: $C_{avg} \times 100$ for different systems on NIST LRE09 Evaluation task over 30s, 10s and 3s conditions.

System	Reg. Coef.	30s	10s	3s
PCA	-	2.93	8.29	22.60
SMM	-	2.81	8.33	21.39
SnGM	-	2.68	8.63	23.15
RSnGM	0.01	**2.52**	**7.06**	**19.11**

4. Analysis of the model parameters

Optimizing the objective function in (7) with $L2$ regularizer can be seen as obtaining MAP point estimate of the model parameters $\mathbf{T}$ and $\mathbf{w}$ with Gaussian priors. In Figure 2, the histogram of 10 random dimensions of $\mathbf{w}$ over TRAIN set and histogram of 10 random rows of the matrix $\mathbf{T}$ are depicted. The y axis in both cases is the frequency of the bin. It can be seen from Figure 2 that the values in case of $\mathbf{w}$ are Gaussian distributed, which confirms assumption of the Gaussian priors over $\mathbf{w}$ vectors is appropriate. On the other hand, in the case of $\mathbf{T}$ rows, values seem to be Laplace distributed. This is mainly because the subspace matrix $\mathbf{T}$ is expanding the iVector space to the sparse original space of n-gram log-probabilities. Intuitively, this suggests use of an $L1$ regularizer that corresponds to the assumption of Laplace prior over estimation of the $\mathbf{T}$ matrix.

5. System evaluation & analysis

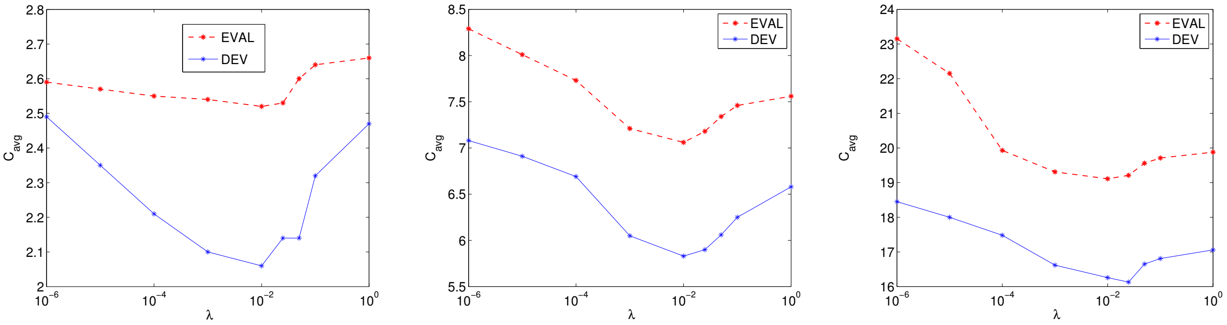
We showed in [12] that 600 is a reasonable choice for the subspace dimension over LRE2009 task. A 600 dimensional subspace and 5 iterations of parameter estimation is used since the value of the objective function over TRAIN set seems to converge after 4 iterations.

In Table 1, performance of the proposed SnGM (without regularization) is compared with subspace multinomial model (SMM) [6] and PCA-based feature extraction that is developed according to the recipe from [4]. The PCA system was widely used by the participants of NIST LRE11 as a phonotactic state of the art system. Aside from marginal degradation for $10s$ condition, the SMM outperforms PCA.

The SnGM system shows notable improvement over the baseline for the 30s condition. However, it also shows performance degradation over shorter conditions. We also noticed big dynamic range for the iVectors corresponding to the short utterances. Intuitively, for utterances with only few n-grams, there can be subspace basis (columns of $\mathbf{T}$) that do not (significantly) affect multinomial distributions corresponding to the seen histories. When estimating iVectors, its coefficients corresponding to such basis can take ”arbitrary” values without affecting the likelihood of the observed n-grams. Note that SMM with single multinomial distribution does not suffer from this problem, and as such can be more robust to over-fitting.

To address the problem with over-fitting, we proposed SnGM with regularized parameter estimation (RSnGM). We

[1] http://www.csie.ntu.edu.tw/ cjlin/liblinear

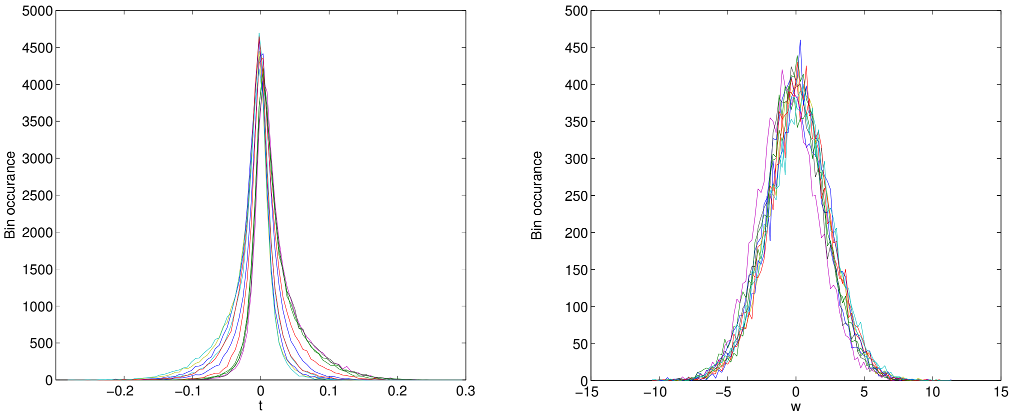


(a) 30 seconds

(b) 10 seconds

(c) 3 seconds

Figure 1: Effect of λ on $C_{avg} \times 100$ over DEV and EVL set for 30s, 10s and 3s conditions on NIST LRE09



(a) Distribution of values in $\mathbf{T}$ rows

(b) Distribution of $\mathbf{w}$ over TRAIN set

Figure 2: Distribution of the values in the model parameters $\mathbf{T}$ and $\mathbf{w}$

use a grid search with logarithmic scale to tune the regularizer coefficient λ. This is depicted in Figure 1. The λ value is tuned over the DEV set. Figure 1 shows that the best LID performance in terms of the C_{avg} over DEV set is obtained with $\lambda = 0.01$. We also depicted the system performance on the held out EVAL set to study generalization of the λ tuning to other unseen data. Interestingly, Figure 1 shows that the tuning of λ over the DEV set generalizes well to the LRE09 EVAL set since the best performance on the NIST LRE09 EVAL set over all conditions are also obtained with $\lambda = 0.01$.

Table 1 shows effect of the regularized parameter estimation on the overall system performance. Results show that the RSnGM system shows significant improvement over the other state of the art systems.

6. Conclusion & future works

We proposed an enhanced phonotactic iVector extraction model over the n-gram counts. In the first step, a subspace n-gram model is proposed to model conditional n-gram probabilities. Modeling different 3-gram histories with separated multinomial distributions shows promising results for the long condition however, we observed model over-fitting for the short duration conditions.

Dealing with the model over-fitting problem, a regularized parameter estimation is proposed. Comparing the effect of the regularized and non-regularized parameter estimation on the overall system performance shows that the regularized parameter estimation is necessary to avoid over fitting of the subspace to the TRAIN set particularly for the short utterances. The proposed regularized subspace n-gram model shows consistent and significant improvement compared to the state of the art phonotactic systems as our baseline over all conditions. To the very best knowledge of the author, this is the best result reported on this task.

The Subspace n-gram model also reduces memory requirement for the parameter estimation and simplifies parallel parameter estimation that leads to a faster model training.

Our experiment with the proposed model shows importance of the numerical optimization during the parameter estimation. Since the $\mathbf{T}$ matrix is expanding iVector to a huge sparse space of the n-gram log-probabilities, use of an *L1* regularizer for estimating the $\mathbf{T}$ matrix may give us a better subspace model and will be explored in future.

7. References

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

[2] F. Richardson and W. Campbell, "Language recognition with discriminative keyword selection," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4145–4148.

[3] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 271–284, 2007.

[4] T. Mikolov, O. Plchot, O. Glembek, P. Matějka, L. Burget, and J. Černocký, "Pca-based feature extraction for phonotactic language recognition," in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010, pp. 251–255.

[5] S. M. Siniscalchi, J. Reed, and T. Svendsen, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language Language*, 2013.

[6] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen, "ivector approach to phonotactic language recognition," in *Proceedings of Interspeech 2011*, Florence, IT, 2011.

[7] N. Dehak, P. Kenny, R. eda Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, pp. 1–23, Jul 2009.

[8] D. Povey, L. Burget, M. Agarwal, P. Akyazi, A. Ghoshal, O. Glembek, K. N. Goel, M. Karafiát, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model-a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, April 2011.

[9] N. Brümmer, L. Burget, O. Glembek, V. Hubeika, Z. Jančík, M. Karafiát, P. Matějka, T. Mikolov, O. Plchot, and A. Strasheim. But-agnitio system description for nist language recognition evaluation 2009. [Online]. Available: http://www.fit.vutbr.cz/research/groups/speech/publi/2009/brummer_BUT_AGNITIO_LRE09_SYSD.pdf

[10] "The 2009 NIST Language Recognition Evaluation Plan (LRE09)," http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf.

[11] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," *Proceedings of ICASSP 2006, Toulouse*, pp. 325–328, Mar 2006.

[12] M. Soufifar, S. Cumani, L. Burget, and J. Černocký, "Discriminative classifiers for phonotactic language recognition with ivectors," in *Proc. International Conference on Acoustics, Speech, and Signal Processing 2012*. Kyoto, Japan: IEEE Signal Processing Society, 2012, pp. 4853–4857.

[13] Z. Jančík, O. Plchot, N. Brummer, L. Burget, O. Glembek, V. Hubeika, M. Karafiát, P. Matějka, T. Mikolov, A. Strasheim, and J. Černocký, "Data selection and calibration issues in automatic language recognition - investigation with BUT-AGNITIO NIST LRE 2009 system," in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*. Brno, CZ: International Speech Communication Association, 2010, pp. 215–221.