



Phonemes Frequency based PLLR Dimensionality Reduction for Language Recognition

Saad Irtza^{1,2}, Vidhyasaharan Sethu¹, Phu Ngoc Le^{1,2}, Eliathamby Ambikairajah^{1,2}, Haizhou Li³

¹ School of Electrical Engineering and Telecommunications, UNSW Australia

² ATP Research Laboratory, National ICT Australia (NICTA), Australia

³ Institute for Infocomm Research, A*STAR, Singapore

s.irtza@student.unsw.edu.au

Abstract

This paper presents a new approach to reduce the dimensionality of Phone Log likelihood Ratio (PLLR) features, which have been shown to be effective for language recognition, by removing the likelihoods corresponding to less frequent phonemes. In this work, phoneme frequencies are estimated using a suitable phoneme recogniser. Following this, an i-vector framework is used to represent the total variability in the reduced dimensional PLLR feature space. This paper also proposes the use of Gaussian probabilistic linear discriminant analysis (GPLDA) as a backend for Language Recognition Evaluation (LRE) tasks. The suitability of both, the proposed dimensionality reductions technique and the GPLDA back-end has been evaluated on NIST 2007 and 2011 LRE tasks. The results show that the novel dimensionality reduction method outperforms PCA based dimensionality reduction by 7%. Further the results also show that GPLDA outperform generatively trained Gaussian back-ends, which have previously been used in conjunction with PLLR feature, by 14.6%.

Index Terms: Language identification, Phone log likelihood ratio, i-vector, Dimensionality reduction

1. Introduction

In Language Recognition Evaluation (LRE) tasks, a variety of information can be used to distinguish between different languages. Among these, the most widely used approaches are those that use acoustic and phonotactic information [1, 2, 3]. In phonotactic systems, speech is tokenized into phoneme sequence which can then be used to extract n-grams counts as features [1, 2]. While in acoustic systems, speech signals are represented by a sequence of short term spectral or prosodic feature vectors. Longer term information is then captured through the use supervector representations of utterances and their total variability factor analysis (i-vector framework). The supervector and subsequently i-vector framework have both been used in speaker verification tasks, and have also shown to be effective in LRE tasks [4]. The i-vector approach was initially used with Mel Frequency Cepstral Coefficients (MFCC) [5] and more recently with Phonotactic features [4]. Phone Log Likelihood Ratio (PLLR) was proposed as a short term features and has been used for LRE tasks [6, 7, 8, 9]. PLLR features are obtained by computing the log likelihood ratio of phone posteriors computed using suitable phone decoders such as Temporal Pattern Neural Network (TRAPs/NN) [10]. Consequently, PLLR features dimension depends on number of phonetic units considered which in turn

depends on the language of the phone decoder. Among the TRAPs/NN phone decoders, Hungarian (HU), Czech (CZ) and Russian (RU) are most frequently used decoders which output 59, 43 and 50 features per frame (corresponding to the number of phonemes in these three languages) respectively.

A recent modification to PLLR features involves the log likelihood ratio being computed over phoneme states instead of entire phonemes, and is referred to as state based PLLR [8]. However, the dimensionality of state based PLLR is three times higher than original PLLR based on entire phonemes. Also, PLLR and state based PLLR features computed from three different phone decoders (HU, CZ, RU) may be concatenated to form a single feature set [9]. Supervised and unsupervised dimensionality reduction approaches have been used to determine the most appropriate low dimensional features to improve the performance and computational cost of language recognition systems [11, 12]. However, a drawback of supervised approaches is that knowledge of each language is required to define appropriate phone sets [11]. Among the unsupervised methods, PCA has been a prominent technique for dimensionality reduction.

In this work, a new dimensionality reduction technique based on phoneme frequency in all target languages is proposed. This approach aims to reduce the dimensionality of PLLR features by dropping the phone log likelihood ratios corresponding to the least frequent phonemes that are common to all languages. The proposed method is based on the assumption that phonemes that are not frequent but are common to all languages can be expected to contain the least amount of discriminatory information.

This paper, in addition to the proposing a dimensionality reduction methods for a PLLR front-end, also evaluates the suitability of Gaussian probabilistic linear discriminant analysis (GPLDA) as an alternative back-end for language recognition systems. Length normalized GPLDA and heavy tailed PLDA (HTPLDA) approaches have previously been introduced to directly model the speaker and channel variability in i-vector based speaker verification systems where they are popular for their simplicity and computationally efficiency [13, 14]. To the best of the authors' knowledge, this is the first paper that investigates the suitability of GPLDA for an LRE task.

2. System Description

2.1. PLLR based i-vector System

The system described in this paper utilises an i-vector framework built on PLLR features [6] as shown in Figure 1.

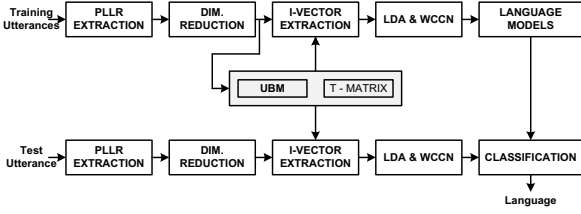


Figure 1: Overview of Language Recognition System

The PLLR feature extraction process estimates phoneme state posteriors, using a phone decoder of N phone units and S states per phone, at each frame t . At a given time frame t , posterior probability of state s ($1 \leq s \leq S$) corresponding to phoneme i ($1 \leq i \leq N$) is given by $p(i|s, t)$. The phoneme posterior is then obtained by summing the states posteriors of each phoneme i at each frame t as given by equation (1):

$$p(i|t) = \sum_{s=1}^S p(i|s, t) \quad (1)$$

Under the assumption of flat priors, the PLLR, is computed from phone posteriors, $p(i|t)$, as per equation (2).

$$PLL R(i|t) = \log \frac{p(i|t)}{\frac{1}{N-1}(1-p(i|t))}, \quad i = 1, \dots, N \quad (2)$$

These PLLR features sets are then used for UBM and total variability matrix training using the approach outlined in [16]. Within the i-vector framework, LDA is used to reduce the i-vector dimensionality further while maximizing the discrimination between classes. Additionally, within class covariances normalization (WCCN) is used to reduce the within class covariance of i-vectors.

2.2. Proposed Dimensionality Reduction

As mentioned previously, the proposed dimensionality reduction method operates by identifying the least frequent phonemes common to all target languages and dropping the phone likelihood ratios corresponding to these from the PLLR feature vector. In order to determine the least frequent phoneme, estimates of relative phone frequencies for all phonemes for each language l is computed as:

$$R(i, l) = \frac{\sum_{u=1}^{U_l} c(i, u, l)}{\max_j \sum_{u=1}^{U_l} c(j, u, l)} \quad (3)$$

Where, $R(i, l)$ is an estimate of the relative frequency of phoneme i in language l ; $c(i, u, l)$ is the number of occurrences of phoneme i in utterance u from language l ; and U_l is total number of utterances available from language l . The relative frequencies are estimated (equation 3) as the phoneme counts normalised to have values between 0 and 1 for all languages. Figure 2 shows the relative phoneme frequencies for Arabic language.

The least frequent phonemes across all languages are then removed by setting a threshold, θ , on the relative phoneme frequencies and dropping the phonemes whose relative frequency is below the threshold for all languages. i.e.,

$$\hat{I} = \{i: \max_l R(i, l) < \theta\}, \quad 1 \leq l \leq L \quad (4)$$

Where, \hat{I} is the set of least frequent phonemes; $R(i, l)$ is the relative frequency of phoneme i in language l ; L is the total number of languages. By changing the threshold θ , the

number of phonemes dropped from the PLLR representation can be varied.

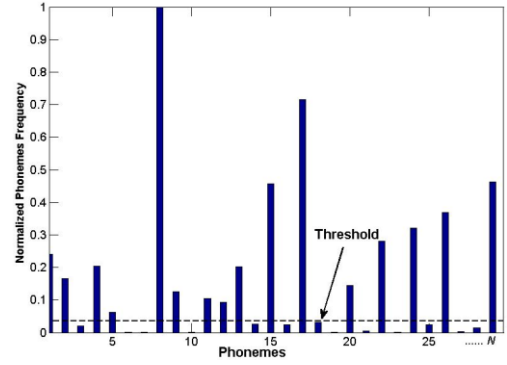


Figure 2: Normalized phoneme frequency bar graph for Arabic (N=59 in the system reported in this paper)

Figure 3 shows an overview of the proposed PLLR dimensionality reduction technique. Specifically, the log likelihood ratios, $LLR(i|t)$, corresponding to the least frequent phonemes, \hat{I} , are removed to give the reduced dimensional PLLR feature vector.

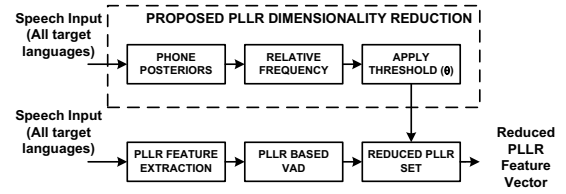


Figure 3: Proposed dimensionality reduction system

2.3. GPLDA Back-End

Gaussian PLDA is currently employed as the back-end in state-of-the-art speaker verification systems to model the speaker and channel variations in an i-vector space [13]. In this paper, we evaluate the suitability of a similar GPLDA back-end for language recognition systems. GPLDA directly models session and language variability within the i-vector space. This idea is similar to that of joint factor analysis (JFA) of i-vectors instead of GMM supervectors. Specifically, the GPLDA models the i-vectors estimated from PLLR features as:

$$\mathbf{w}_u = \bar{\mathbf{w}} + \mathbf{L}\mathbf{x}_u + \boldsymbol{\epsilon}_u \quad (5)$$

Where, \mathbf{w}_u denotes the i-vector corresponding to utterance u ; \mathbf{L} is the eigenlanguage matrix; \mathbf{x}_u denotes the language factors corresponding to utterance u ; $\boldsymbol{\epsilon}_u$ denotes the utterance specific within language variability. In the PLDA approach, the language specific part is modelled by $\bar{\mathbf{w}} + \mathbf{L}\mathbf{x}_u$ and contains the discriminatory information for language recognition.

The GPLDA parameters are estimated via maximum likelihood estimation (MLE) as described in [17]. GPLDA scoring is performed using batch likelihood between test and target i-vector as follows:

$$LLR = \log \frac{P(\mathbf{w}_{target}, \mathbf{w}_{test} | H_1)}{P(\mathbf{w}_{target} | H_0)P(\mathbf{w}_{test} | H_1)} \quad (6)$$

Where H_1 is the hypothesis that i-vectors represent the true language and H_0 does not.

In this paper we compare the GPLDA back-end to a generatively trained Gaussian back-end, which has previously been used with PLLR i-vector systems for language recognition [6].

3. Experimental Setup

3.1. NIST 2007 & 2011 LRE Datasets

The NIST 2007 LRE task dataset consists of conversational telephonic speech (CTS) involving total 15 languages out of which French is only non-target language [19]. In this experiment, all the training and development data distributed by NIST to LRE participants have been used which consists of data from 1) Call-Friend corpus 2) NIST 2005 LRE task 3) NIST 2007 LRE task. Total duration of training data is approximately 968 hours. Development set consists of 10 conversations randomly selected from each target language. Results are reported on 30, 10 and 3 sec test set for the primary task in NIST 2007 LRE. Training, development and evaluation data of each target language is same as described in [6].

The NIST 2011 LRE task involved 24 target languages which includes all 15 from NIST 2007 and 9 new languages. NIST has distributed 100, 30 seconds segments for each of the new languages (except for Lao which has only 93 segments). No additional development data had been distributed for this evaluation. In all experiments conducted on the NIST 2011 dataset, data from previous NIST LRE 2009 was added to the training set [20]. The LRE 2009 dataset consists of, telephonic speech segments from NIST LRE 1996, 2003, 2005, 2007 and 2009 evaluations and segments from VOA broadcasts. Training, development and evaluation data of each target language is identical to the setup described in [21]. For development purposes, 10 segments are randomly selected from each target language. Results are reported on 30, 10 and 3 sec test set for the primary task in NIST 2011 LRE.

3.2. System Configuration

The baseline PLLR i-vector system was developed as described in [6]. A Hungarian TRAPs/NN phone recognizer [10] is used to extract phoneme state posteriors corresponding to 61 phonetic units. State posteriors are then summed as per equation 1 to obtain phoneme unit posteriors. Further, 3 non speech units (from the 61 phonetic units) are merged into a single non-speech unit to give a total of 59 units whose posterior probabilities are estimated by the recogniser. PLLR values are then estimated from the phoneme posteriors as per equation (2). Following this, voice activity detection (VAD) is implemented by removing the frames whose highest PLLR value corresponds to the non-speech unit. Consequently, each speech frame is represented by a 59 dimensional phone log likelihood ratio (PLLR) vector.

In systems involving dimensionality reduction, either via PCA or via the proposed method, these methods are applied directly on the PLLR feature vectors at this stage. Following this, dynamic and shifted delta coefficients (SDC, 42-1-5-1) are estimated from the reduced dimensional PLLR vector.

All Universal Background Models (UBM) were estimated using Maximum Likelihood criteria (ML) using 1024 mixtures employing binary mixture splitting. Total variability matrix (T-Matrix) is estimated as in [16]. I-vectors of 400 dimensions are used since they have shown promising results for language recognition [5].

The evaluation criteria defined by NIST [16] are employed in the work reported in this paper. Namely, (a) average cost performance (C_{avg}); and (b) log likelihood ratio cost (C_{llr}).

4. Results

Detailed study of PLLR i-vector system with reduced PLLR features were initially carried out on the NIST 2007 database. Following this, the best system was validated on the NIST 2011 database.

4.1. Results on NIST 2007 LRE

Table 1 shows the results of the PLLR i-vector baseline system as well as the systems using the proposed dimensionality reduction. All systems were evaluated with two back-ends, namely a generative Gaussian backend previously employed in PLLR based systems [6] which does not use LDA or within class covariance normalization (WCCN) on the i-vectors; and LDA+WCCN followed by GPLDA on 30, 10 and 3 seconds on NIST 2007 test segments. This initial experiment was carried out to determine the optimal number of dimensions in the reduced set obtained by the proposed dimensionality reduction technique.

Table 1: Results on NIST 2007 LRE Primary Task to determine optimal feature dimensionality

PLLR Features Dimensions	$C_{avg} \times 100 / C_{llr}$					
	Gaussian Backend			LDA+GPLDA		
	30s	10s	3s	30s	10s	3s
59 (Baseline – from [6])	3.45/ 0.564	Not Reported in [6]				
46	3.36/ 0.52	7.27/ 0.802	15.02/ 1.42	2.57/ 0.49	7.18/ 0.78	14.32/ 1.37
42	3.34/ 0.44	6.85/ 0.76	14.5/ 1.3	2.44/ 0.32	6.18/ 0.70	13.75/ 1.23
39	3.63/ 0.48	7.01/ 0.79	15.73/ 1.35	2.64/ 0.35	6.33/ 0.77	15.83/ 1.31

Different threshold values were used in dimensionality reduction technique, as described in section 3, to reduce the baseline (59 dimensional) frame-level PLLR features to 46, 42, and 39 dimensional vectors. It can be seen (Table 1) that reducing the feature dimensions from 59 to 42 leads to a 3.18% improvement. However, reducing the PLLR feature dimensionality further below 42 led to an increase in the error rate, which may be due to the removal of discriminatory information in those dimensions.

It can also be observed that LDA+WCCN with GPLDA scoring outperforms the Gaussian back-end by 26.9% when using reduced dimensional (42 dimensions) PLLR features in the 30s test condition. In all further experiments, the proposed dimensionality reduction method is used to reduce the baseline 59 dimensional PLLR vectors to 42 dimensional vectors.

Table 2 shows the results of a second experiment using reduced dimensional PLLR i-vector system augmented with delta (Δ) and shifted delta coefficients (SDC) with both generative Gaussian; and LDA+WCCN followed by GPLDA back-ends on 30, 10 and 3 seconds on NIST 2007 test segments. In this experiment, the proposed dimensionality reduction technique is compared to baseline systems that employ PCA for dimensionality reduction. In the case of PCA, the target feature dimensions are 23 and 13 when employing deltas and SDCs respectively. These were the optimal feature dimensions reported [11,12].

It can be seen that the proposed dimensionality reduction technique outperforms the baseline PCA approaches in both cases, i.e., when using Deltas and when using SDCs. Further, it can be seen that the GPLDA based systems significantly outperforms the generative Gaussian back-end based systems.

Table 2: Results on NIST 2007 LRE Primary Task with reduced dimensional PLLR augmented with Δ and SDC

PLLR Feature Dimensions	$C_{avg} * 100 / C_{LLR}$					
	Gaussian Backend			LDA+GPLDA		
	30s	10s	3s	30s	10s	3s
PCA (23 + Δ) (From [11])	2.17/ 0.32	Not Reported in [11]				
Proposed (42+ Δ)	2.12/ 0.31	5.45/ 0.67	11.2/ 1.1	1.95/ 0.29	5.34/ 0.6	11.02/ 1.05
PCA (13+SDC) (From [12])	1.71/ 0.26	Not Reported in [12]				
Proposed (42+SDC)	1.64/ 0.24	4.62/ 0.53	10.44/ 0.95	1.41/ 0.23	4.48/ 0.477	9.98/ 0.88

It should be noted that the baseline PCA system utilized 13-2-3-7 (N-D-P-K) SDC configuration, resulting in a 104 dimensional representation while in the proposed approach, 42-1-5-1 SDC configuration is used to obtain an 84 dimensional representation. The best performance on the NIST 2007 LRE primary task is achieved by the system using the proposed dimensionality reduction augmented with SDC followed by LDA+WCCN and GPLDA scoring resulting in C_{avg} of 1.41% (17.5% improvement over PCA with SDC.).

4.2. Results on NIST 2011 LRE

In the final experiment the suitability of the best performing systems, namely the reduced dimensional PLLR augmented with deltas and SDCs, was validated on the NIST 2011 LRE task. Table-3 shows the performances of the systems employing deltas and SDCs with both the generative Gaussian backend, and LDA+WCCN followed by GPLDA back-end on NIST 2011 test segments.

Table 3: Results on NIST 2011 LRE Primary Task with reduced dimensional PLLR augmented with Δ and SDC

PLLR Features Dim	$C_{avg} * 100 / C_{LLR}$					
	Gaussian Backend			LDA+GPLDA		
	30s	10s	3s	30s	10s	3s
PCA(23)+ Δ (Baseline) [9]	4.48/ 0.188	Not Reported				
42+ Δ	4.16/ 0.106	8.84/ 0.245	13.54/ 0.327	4.08/ 0.105	7.52/ 0.194	13.36/ 0.284
PCA(13)+SDC (Baseline) [10]	4.10/ 0.826	Not Reported				
42+SDC	3.94/ 0.097	7.79/ 0.208	12.65/ 0.314	3.81/ 0.094	6.81/ 0.184	11.58/ 0.24

Once again it can be seen that the proposed dimensionality reduction method outperforms the standard PCA based approaches in all test conditions. It can also be seen that GPLDA outperforms the Gaussian back-end under all test conditions. The best performance on the NIST 2011 LRE task is achieved by the system employed the proposed dimensionality reduction on the PLLR feature vector, augmented with SDCs and employing a GPLDA based back-end.

5. Conclusions

A novel dimensionality reduction technique based on relative phoneme frequencies has been proposed for PLLR features and evaluated on NIST LRE corpora. Experimental results suggest that the proposed method outperforms traditional PCA based dimensionality reductions methods. It is likely that this is due the fact that the proposed method estimates the least discriminatory phonemes across all target languages and removes them from the PLLR feature vector. The results also suggest that the performance of the proposed system is further improved when the reduced PLLR features are augmented with dynamic and shifted delta coefficients.

This paper also evaluated the suitability of GPLDA as a back-end for language recognition systems. Experimental results in this paper show that GPLDA consistently outperforms generatively trained Gaussian back-ends, which were previously used with PLLR front-ends. This suggests that similar to speaker verification systems, GPLDA based back-ends may constitute a new baseline system for language recognition.

6. References

- [1] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language Identification: A Tutorial", in *IEEE Circuits and Systems Magazine*, Vol. 11, No. 2, 2011, pp.82 – 108
- [2] H. Li, K. A. Lee, and B. Ma, "Spoken Language Recognition: From Fundamentals to Practice", in *Proceedings of the IEEE*, Vol. 101, No. 5, 2013, pp. 1136 – 1159
- [3] D.A. Reynolds, W.M. Campbell, W. Shen and E. Singer "Automatic language recognition via spectral and token based approaches" *Springer Handbook of Speech Processing. Springer Berlin Heidelberg*, 2008. 811-824.
- [4] M. Souffifar, K. Marcel, B. Lukás, P. Oldrich, G. Ondrej and S. Torbjørn, "iVector Approach to Phonotactic Language Recognition." in *INTERSPEECH 2011, Florence, Italy*, 2011, pp. 2913-2916.
- [5] N. Dehak, P.A. Torres-Carrasquillo, D.A. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *INTERSPEECH 2011, Florence, Italy*, 2011 pp.857–860.
- [6] M. Diez, A. Varona, M. Penagarikano, Luis J. Rodriguez-Fuentes and G. Bordel, "On the use of phone log-likelihood ratios as features in spoken language recognition" *Spoken Language Technology Workshop (SLT), 2012 IEEE*, vol., no., 2-5 Dec. 2012, pp.274, 279.
- [7] M. Diez, A. Varona, M. Penagarikano, Luis J. Rodriguez-Fuentes and G. Bordel, "New Insight into the Use of Phone Log-Likelihood Ratios as Features for Language Recognition" in *INTERSPEECH 2014, Singapore, 14-18 sep.*, 2014.
- [8] L.F. D'haro, R. Cordoba, C. Salamea and J.D. Echeverry, "Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition," *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE*, vol., no., 4-9 May, 2014, pp.5342,5346.
- [9] L.F. D'haro, R. Cordoba, C. Salamea and J. Ferreiros., "Language Recognition using Phonotactic-based Shifted Delta Coefficients and Multiple Phone Recognizers" in *INTERSPEECH 2014, Singapore, 14-18 sep.*, 2014.
- [10] P. Schwarz, "Phoneme recognition based on long temporal context," *Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology*, <http://www.fit.vutbr.cz/>, Brno, Czech Republic, 2008.
- [11] M. Diez, A. Varona, M. Penagarikano, L. J. Rodríguez-Fuentes, and G. Bordel, "Dimensionality Reduction of Phone Log-Likelihood Ratio Features for Spoken Language Recognition," in *INTERSPEECH 2013 Lyon, France, August 2013*.

- [12] M. Diez, A. Varona, M. Penagarikano, L.J. Rodriguez-Fuentes, and G. Bordel. "Optimizing PLLR Features for Spoken Language Recognition." *22nd International Conference on Pattern Recognition (ICPR 2014), Stockholm, Sweden, 2014*, pp. 779-784.
- [13] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in Proc. Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic, 2010.
- [14] Kanagasundaram, Ahilan, D. Dean, J. Gonzalez-Dominguez, S. Sridharan, D. Ramos, and J. Gonzalez-Rodriguez. "Improving the PLDA based speaker verification in limited microphone data conditions." *In the Proceedings of 14th Annual Conference of the International Speech Communication Association 2013 (ISCA), Israel, 2013*, pp. 3674-3678.
- [15] Free PLLR computation software. [Online]. Available: <https://sites.google.com/site/gttspllrfeatures/home>, accessed on November, 2014.
- [16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification" *Audio, Speech, and Language Processing, IEEE Transactions on 19, no. 4, 2011*, 788-798.
- [17] Prince, Simon JD, and J.H. Elder. "Probabilistic linear discriminant analysis for inferences about identity." *In IEEE 11th International Conference on Computer Vision, ICCV 2007, 2007*, pp. 1-8.
- [18] 2011 Language Recognition Evaluation, <http://www.nist.gov/itl/iad/mig/lre11.cfm>
- [19] A. F. Martin and A. N. Le, "NIST 2007 Language Recognition Evaluation," in *Proceedings of Odyssey 2008 – The Speaker and Language Recognition Workshop, 2008*.
- [20] A.F. Martin and C.S. Greenberg, "The 2009 NIST Language Recognition Evaluation," in *Odyssey 2010 – The Speaker and Language Recognition Workshop, Brno, Czech Republic, 2010*, pp. 165–171.
- [21] Singer, Elliot, P. Torres-Carrasquillo, D. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. Sturim. "The MITLL NIST LRE 2011 language recognition system" *In Acoustics, Speech and Signal Processing, ICASSP 2007, 2012*, pp. 209-215.