ViVoLab UZ Language Recognition System for Albayzin 2010 LRE

David Martínez, Jesús Villalba, Antonio Miguel, Alfonso Ortega and Eduardo Lleida

Voice Input Voice Output Laboratory (ViVoLab)
Aragon Institute for Engineering Research (I3A)
University of Zaragoza (UZ), Spain
(david | villalba | amiguel | ortega | lleida)@unizar.es

Abstract

This paper describes the 2 systems submitted by ViVoLab UZ for the Albayzin 2010 Language Recognition Evaluation (LRE) [1]. Both submissions are a fusion of 5 phonotactic and 3 acoustic subsystems. The only difference between them is the normalization and fusion of the scores. We have investigated the state-of-the art methods for Language Recognition (LR) in the KALAKA-2 database [2]. Our group obtained the first position in the evaluation.

Index Terms: language recognition, phonotactic LRE, acoustic LRE, channel compensation, discriminative training

1. Introduction

Language Identification (LID) has experimented a huge development in the last years. To compare the quality of the different LR systems around the world, NIST has coordinated several evaluations (1996, 2003, 2005, 2007 and 2009) [3]. In year 2008, the Spanish Network on Speech Technology coordinated a similar one, for research groups of the Iberian Peninsula [4]. Albayzin LRE 2010 is the second edition. The main difference with NIST is that the languages to be recognized are Spanish, Catalan, English, Basque, Galician and Portuguese, and they are extracted from multi-speaker TV broadcast recordings.

The systems submitted by ViVoLab UZ are a fusion of 5 phonotactic subsystems and 3 acoustic subsystems. Both systems are identical except for the normalization and fusion methods used at the back-end. In the first submission, we make a t-norm of scores and perform a discriminative fusion. In the second, we make a zt-norm of scores, and follow a generative gaussian backend by a discriminative calibration. Our submission includes closed- and open-set condition for the clean speech task, for the 3s, 10s and 30s tests.

The thresholds for each submission are set separately for each condition and for each duration of file, detecting the length of each one by counting the number of frames. For the closed- and open-set conditions, we have used the same systems, but setting the threshold to different values in order to minimize C_{avg}.

The rest of the paper is organized as follows: Section 2 specifies the data used for training; Section 3 describes the acoustic, phonotactic and fusion methods; Section 4 indicates the computational cost of the systems; in Section 5, results obtained in the evaluation are analysed; and Section 6 gives the conclusions and comments some next steps.

2. Data and Performance Measurement

The data used for training our system come from the training part of KALAKA-2 database, with the exception of the training of the phone recognizers. We have used phoneme recognizers trained in Czech, with the Czech SpeechDat-E database [18], in Hungarian, with the Hungarian SpeechDat-E database [19], in Russian, with the Russian SpeechDat-E database [20], in English, with the TIMIT database [21], and in Spanish, with the Albayzin [22] and Speech Dat Car [23] databases.

Calibration of the results was performed with the development data of KALAKA-2 database.

3. System Description

The 2 submitted systems are a fusion of 8 subsystems: 3 acoustic and 5 phonotactics

3.1. Features

The features used for the acoustic systems are MFCC concatenated to their Shifted Delta Cepstra Coefficients (SDC) [5]. We obtain 6 MFCCs plus energy, perform cepstral mean normalization, and then we calculate the SDC with a 7-1-3-7 configuration. After that, we transform the features with a Short Time Gaussianization (STG) [6].

3.2. Acoustic Systems

The 3 acoustic subsystems are a GMM Maximum Likelihood (ML) subsystem, a GMM Maximum Mutual Information (MMI) subsystem and a GMM Factor Analysis (FA) subsystem which performs channel compensation.

3.2.1. GMM ML subsystem

The ML GMM subsystem is based on a calculation of one ML GMM model for each language using the EM algorithm. We perform 10 iterations to obtain a 2048 gaussians model. This method tries to maximize the likelihood of the data for each class.

3.2.2. GMM MMI subsystem

Starting from the GMM ML model, we perform a discriminative re-training based on MMI to obtain the final models. 10 iterations of MMI re-training are run. Unlike the ML training, this method tries to maximize the posterior probability of recognizing all training utterances given the labelled data. The objective function is [7]

$$F_{\mathrm{MMI}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(\chi_r|s_\mathrm{r})^{K_\mathrm{r}} P(s_\mathrm{r})^{K_\mathrm{r}}}{\sum_{\forall s} p_\lambda(\chi_r|s)^{K_r} P(s)^{K_\mathrm{r}}} \quad (1)$$

where $p_\lambda(\chi_r|s_\mathrm{r})$ is the likelihood of r-th training segment, χ_r, given the correct language identity of the segment, s_r, and model parameters λ. R is the number of training segments, and the denominator represents the overall probability density, $p_\lambda(\chi_r)$. We consider the prior probabilities of all classes equal and drop the prior terms $P(s_\mathrm{r})$ and $P(s)$. Usually, segment likelihood $p_\lambda(\chi_r|s)$ is computed as simple multiplication of frame likelihoods incorrectly assuming statistical independence of feature vectors. Factor $0 < K_r < 1$, which is increasing the confusion between hypothesis represented by numerator and denominator, can be considered as a compensation for underestimating segment likelihoods caused by this incorrect assumption.

3.2.3. GMM FA subsystem

This system is based in a FA for the mean of the models based on [8], in which we have defined two factors, one for the language and one for the channel. Thus, we can obtain a channel compensated model for each language. This is a two-level hierarchy model, since we assume a different GMM that generates every speech segment, and we also assume that for every speech segment, this GMM has been generated by a sub-model. Then, for the speech segment *s*, we have

$$M_s = t_{l(s)} + Ux_s \quad (2)$$

where $t_{l(s)}$ are the *language location vectors*, x_s is a vector of C segment-dependent *channel factors*, and U is a 56-by-C *factor loading matrix*, which translates the channel factors from their low dimensional space to the high dimensional space where the model M_s lies.

The $t_{l(s)}$ matrix is obtained by MAP adaptation from a UBM model with mean m_0 and covariance matrix Σ, in the following way

$$t_{l(s)k} = \frac{\sum_s f_{sk}}{\tau + \sum_s n_{sk}} \quad (3)$$

being n_{sk} and f_{sk} the zero and first order statistics respectively, for the *k*th gaussian component.

The U matrix and the channel factors are calculated according to a ML criterion, using the EM algorithm iteratively, in a similar way to [8].

The scoring of each utterance is made via a linear scoring, as proposed in [9].

3.3. Phonotactic Systems

5 PRLM sub-systems [12] in different languages have been fused. 4 of them use the Brno University of Technology (BUT) phoneme recognizer, based on ANN/HMM and Temporal Patterns (TRAPS) with Split Temporal Context (STC) [10], and are trained on Czech, Hungary, Russian and English. The other one uses the phoneme recognizer of the UZ, which is based on GMM/HMM with conventional MFCC and is trained on Spanish. In this one, the phonemes are taken with right and left context, so we will call the recognition unit subphoneme instead. However, we will keep only the central unit for the posterior step, that is, the phoneme without context. The output phonemes are used to train a langauge model (LM) for each one of the target languages with the SRILM tool [11].

All LMs are built with an interpolated Witten-Bell discounting method. We use 4-grams for building them in all cases, and for testing we also use 4-grams for all cases except for the Spanish LM, in which we test with 3-grams. The reason is that we empirically experimented a better performance with this configuration. In addition, for the four phoneme recognizers based on GMM/ANN, we make use of lattices [13] to get more information out of the acoustic signal. Specifically, we create a 100-best list for the train and a 5-best list for the test.

3.4. Fusion for the Primary Submission

In our primary submission, the results coming from each system are T-Normalized [15], fused, and for the closed-set condition, another T-Normalization is applied after the fusion. The fusion is also a calibration [14] and the fused log-likelihood vector is

$$\mathbf{l}'(x_t) = \sum_{k=1}^{K} \alpha_k \mathbf{l}_k(x_t) + \beta \quad (4)$$

where the coefficients α_k and β are calculated via a discriminative Linear Logistic Regression (LLR) training, using the FoCal Multi-class toolkit [14], and $\mathbf{l}_k(x_t)$ is the output of system k when input in time t is x_t.

3.5. Fusion for the Alternative Submission

In this submission, we investigate the ZT-Normalization [15] technique, combined with a Gaussian Back-End (GBE) followed by a discriminative LLR fusion [14], as the one in the primary submission. In the closed-set condition, results after the GBE and after the LLR fusion are again T-Normalized.

In a GBE, the likelihood scores are obtained from multivariate Gaussians, with target language specific means and one common full covariance matrix. As explained in [16], the GBE can be seen as an affine transformation. The linear part of the transform is the same as a Linear Discriminant Analysis (LDA) transform, which tries to maximize the ratio of between-class to within-class variance. The translation part of the affine transform is equivalent to the calibration task of setting language dependent thresholds. The decision made by the GBE corresponds to the following normal distribution function:

$$\delta_l(\mathbf{x}) = (\mathbf{x} - \mu_l)^t \Sigma^{-1} (\mathbf{x} - \mu_l) \quad (5)$$

where μ_l is the mean for language l, Σ is the common covariance matrix, $\mathbf{x}$ is the input score and $\delta_l(\mathbf{x})$ is the transformed output.

The posterior discriminative LLR was added to give further calibration to the system, and we could check a further improvement in the results.

4. Computational Cost

Real time factor was approximately 0.9xRT for both submissions, since the normalization and calibration technique does not alter the overall processing time.

5. Analysis of Results

In this section we will show the results obtained in the evaluation for the described systems in terms of C_{avg}. We can see how accurate is the calibration on the DET curves comparing C_{avg} (marked with 'x') with C^*_{avg} (marked with 'o'). To analyse results, we will focus on the the primary system, since the results of the alternative system follow the same trend but with higher error rates.

5.1. Primary System - Closed Clean (CC)

In Fig. 1 we have the results of our primary system for the clean speech, closed set condition, for 30, 10 and 3 s of duration of utterance. C_{avg} is 0.0184, 0.0418 and 0.0943, respectively. The 30s test of this condition is the one used to rank systems in the evaluation. Our system was the best.

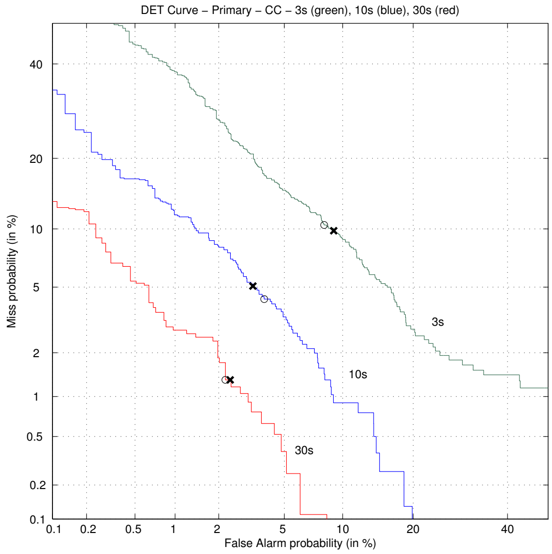

Figure 1: *DET Curves for Primary System CC condition*

$P_{fa}(L_t, L_n)$	Target Language L_t					
L_n	SPA	CAT	ENG	BAS	GAL	POR
SPA	–	0.016	0.000	0.000	0.232	0.000
CAT	0.020	–	0.000	0.000	0.007	0.007
ENG	0.000	0.000	–	0.000	0.000	0.000
BAS	0.000	0.000	0.000	–	0.000	0.000
GAL	0.413	0.016	0.000	0.000	–	0.000
POR	0.000	0.000	0.000	0.000	0.000	–
$P_{miss}(L_t)$	0.008	0.013	0.000	0.008	0.050	0.000
Avg. $P_{fa}(L_t)$	0.087	0.006	0.000	0.000	0.048	0.001
Avg. P_{miss} = 0.0131						
Avg. P_{fa} = 0.0237						

Table 1: *Error Rates for CC 30s condition in the primary system. We show the target language L_t in the columns and the segment language L_n in the rows. Labels of languages are SPA=Spanish, CAT=Catalan, ENG=English, BAS=Basque, GAL=Galician and POR=Portuguese*

If we analyse Table 1, we can see a very good performance recognizing all languages for the 30s CC condition, having a global P_{miss} of 0.0131. The highest figure is for Galician with a P_{miss} of 0.050. However, if we look at the false-alarm probabilities, we can check in general a good performance, but a low one when discriminating between Spanish and Galician. The false-alarm probability of saying that the language transmitted is Galician when it is really Spanish is 0.232 and of saying that it is Spanish when it is really Galician is 0.413. Several reasons could be considered for this behaviour, but we think that the most dramatic is, after listening some of the recordings, the fact that many Galician speakers are Spanish-native speakers. Therefore, their Galician accents are very influenced by the Spanish language.

The problem caused by people who speak several languages is a general one for language recognizers, and it would be beneficial to have into account this information when training LID systems. One solution could be to train systems for native and non-native speakers as different languages. Another approach to this problem could be to apply discriminative training techniques which place more gaussians (in GMM systems) at the borders between these languages for a better characterization of these areas of the vector space.

For the rest of languages, we can see small confusion rates, especially for English and Basque, which are 0. This is due to the highly different acoustic and phonotactic nature of these languages with regard to the others.

We detail the results for each individual subsystem for the 30 s condition in Table 2. We can see that the subsystem that performs the best is the FA. On the other hand, the PRLM_ES and PRLM_EN do not give good results by themselves. After evaluation, we could check that the back-end was not optimum, and with only a GBE we obtained great improvements in all subsystems.

Subsystem	C_{avg}
JFA	0.0357
ML	0.0855
MMI	0.0598
PRLM_CZ	0.0569
PRLM_HU	0.0501
PRLM_RU	0.0547
PRLM_EN	0.2618
PRLM_ES	0.1474

Table 2: *C_{avg} for the individual subsystems of the primary submission for the CC 30s condition*

5.2. Primary System - Open Clean (OC)

In Fig. 2 we have the results of our primary system for the clean speech, open set condition, for 30, 10 and 3 s of duration of utterance. C_{avg} is 0.0307, 0.0644 and 0.1202, respectively.

In Table 3 we can check that the performance of the system for the 30s OC condition has slightly dropped, compared to the CC condition. This is due to the introduction of out-of-set (OOS) languages. The average P_{miss} drops to 0.0193, but more dramatic is the decrease in the average P_{fa}, which drops to 0.0422, i.e. a relative decrease of 79% compared to the 30s CC condition.

We still observe the problem when distinguishing Spanish of Galician. And now, we can see that the OOS languages are mainly confused with English, with a $P_{fa}(EN, OOS)$ of 0.188.

The results for the individual subsystems for the 30 s condition are in Table 4. Again, the system that performs the best is the FA. In this case, the difference with the rest is much higher. However, we have to take into account that our system was tuned for the fusion of all subsystems and not for each individual one. If we tuned it for each one, we would obtain better results. As in the CC case, we could check after evaluation that the back-end was not optimum.

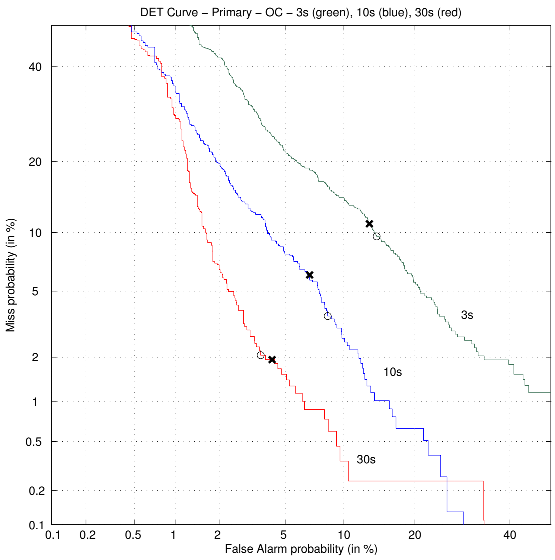

Figure 2: *DET Curves for Primary System OC condition*

Subsystem	C_{avg}
JFA	0.0510
ML	0.3658
MMI	0.3262
PRLM_CZ	0.4499
PRLM_HU	0.3998
PRLM_RU	0.4293
PRLM_EN	0.4991
PRLM_ES	0.4920

Table 4: C_{avg} *for the individual subsystems of the primary submission for the OC 30s condition*

$P_{fa}(L_t, L_n)$	Target Language L_t					
L_n	SPA	CAT	ENG	BAS	GAL	POR
SPA	–	0.008	0.000	0.000	0.208	0.000
CAT	0.0134	–	0.000	0.000	0.000	0.000
ENG	0.000	0.000	–	0.000	0.000	0.000
BAS	0.000	0.000	0.000	–	0.000	0.000
GAL	0.446	0.008	0.000	0.000	–	0.000
POR	0.000	0.000	0.000	0.000	0.000	–
OOS	0.026	0.064	0.188	0.052	0.003	0.094
$P_{miss}(L_t)$	0.024	0.013	0.007	0.008	0.050	0.013
Avg.$P_{fa}(L_t)$	0.092	0.003	0.000	0.000	0.042	0.000
Avg.$P_{fa}(L_t + L_o)$	0.066	0.028	0.075	0.021	0.026	0.038
Avg. P_{miss} = 0.0193						
Avg. P_{fa} = 0.0422						

Table 3: *Error Rates for OC 30s condition in the primary system. We show the target language* L_t *in the columns and the segment language* L_n *in the rows. Labels of languages are SPA=Spanish, CAT=Catalan, ENG=English, BAS=Basque, GAL=Galician, POR=Portuguese and OOS=Out-Of-Set*

5.3. Alternative System - CC

In Fig. 3 we have the results of our alternative system, for the clean speech, closed set condition, for 30, 10 and 3 s of duration of utterance. C_{avg} is 0.0238, 0.0498 and 0.1087, respectively. The detail analysis of results is similar to the primary system CC condition, but the error rates are higher.

5.4. Alternative System - OC

In Fig. 4 we have the results of our alternative system, for the clean speech, open set condition, for 30, 10 and 3 s of duration of utterance. C_{avg} is 0.0373, 0.0635 and 0.1309, respectively. The detail analysis of results is similar to the primary system OC condition, but the error rates are higher.

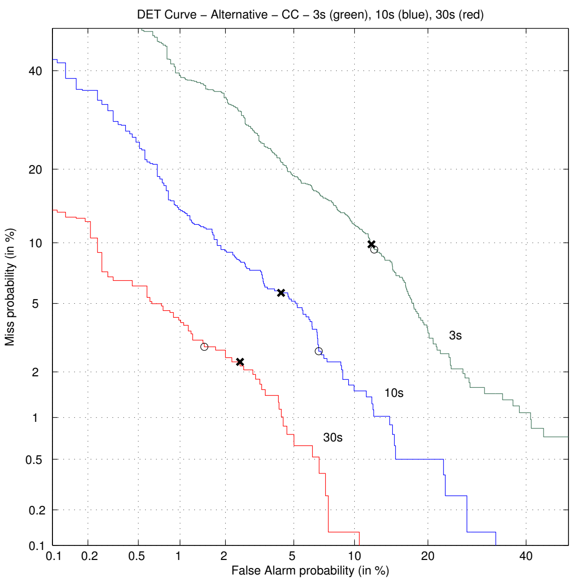

Figure 3: *DET Curves for Alternative System CC condition*

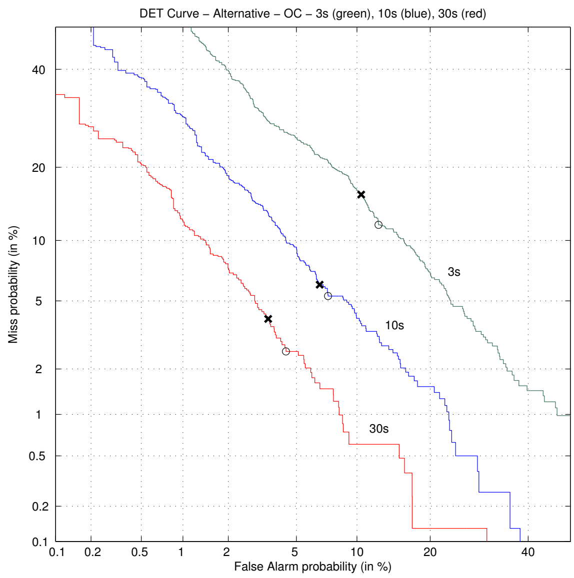

Figure 4: *DET Curves for Alternative System OC condition*

6. Conclusions and Future Work

In this edition of Albayzin LRE, ViVoLab UZ participates for the first time in a Language Recognition Evaluation. We have built several state-of-the-art systems that have been tested in the KALAKA-2 database. All these systems are fusioned into one, in which the characteristics of each are combined to get a higher performance. For the ranking of systems only the primary submission of the CC 30s condition was considered and our system was the best of all participant sites.

First, we built 5 phonotactic systems trained in different languages using 2 phoneme recognizers, one of BUT and one of UZ. Then, we combined them with 3 acoustic systems, ML, MMI and FA, coming from the Speaker Identification (SPKID) ideas, where our group has big experience.

Finally, we normalized and fused the results in two different ways, what generates our two submissions. One is based on a T-Norm and a discriminative LLR fusion, while the other is based on a ZT-Norm and a GBE followed by a discriminative LLR fusion.

Analysing the results, and focusing on the 30s CC condition, we can see very low P_{miss} values and, in general, low P_{fa} values. However, we detect a great confusion between Spanish and Galician, mainly caused by the fact that many Galician speaker are non-native speakers and their accent is influenced by the Spanish language. To solve this, we should think of having into account if the speaker is native or non-native when training the systems, and of turning our efforts to build a discriminative algorithm able to differentiate properly the borders between these languages.

As next steps we consider including several other approaches to our final system. First, the introduction of Vocal Tract Length Normalization (VTLN) should make the features more independent of the speaker. Secondly, we will experiment with a Probabilistic Linear Discriminant Analysis (PLDA) system on LID, since the performance on SPKID has shown to be excellent [17]. And finally, we would like to continue investigating new normalizing techniques, as one only based on the length of the files under test. In addition, we checked in a post-evaluation of the systems, that the selected configurations for the back-end were not optimal, and that only a GBE could have given a better performance.

7. Acknowledgements

We would like to thank the organizers of Albayzin 2010 LRE for his effort in preparing all the data of KALAKA-2 database, and also to all the organization of Fala 2010 for supporting the Albayzin 2010 LRE.

This work was funded by the Spanish Ministry of Science and Innovation under project TIN2008-06856-C05-04.

8. References

[1] The Albayzin 2010 Language Recognition Evaluation Plan (Albayzin 2010 LRE).

[2] KALAKA-2. Speech database created for the Albayzin 2010 Language Recognition Evaluation, organized by the Spanish Network on Speech Technology. Produced by the Software Technologies Working Group (GTTS, http://gtts.ehu.es), University of the Basque Country.

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

[4] Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Germán Bordel, and Amparo Vasrona, "The Albayzin 2008 Language Recognition Evaluation", in Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, 28 June - 1 July 2010.

[5] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstal Features", in Proc. International Conferences on Spoken Language Processing (ICSLP), Sept. 2002, pp.89-92.

[6] J. Pelecanos and S. Sridharan, "Feature Waroing for Robust Speaker Verification", Proc. Speaker Odyssey 2001 conference, June 2001.

[7] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition", Ph.D. thesis, Cambridge University, July 2004.

[8] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms", Technical Report CRIM-06/08-13, CRIM, 2005, http://www.crim.ca/perso/patrick.kenny/FAtheory.pdf.

[9] O. Glembek, L. Burget, N. Dehak, N. Brümmer, and P. Kenny, "Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis", in Proc. ICASSP, pp.4057-4060, Taipei, Apr. 2009.

[10] P. Schwarz, "Phoneme Recognition Based on Long Temporal Context", Ph.D. Thesis, Brno University of Technology, 2009. http://speech.fit.vutbr.cz/cs/software/phoneme-recognizer-based-long-temporal-context.

[11] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", in Proc. ICSLP, pp. 901-904, 2002. http://www.speech.sri.com/projects/srilm.

[12] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telepone Speech", IEEE Trans. Acoust., SPeech Signal Processing, vol. 4, no. 1, pp. 31-44, 1996.

[13] J.L. Gauvain, A. Messaoudi, and H. Schwenk, "Language Recognition using Phoneme Lattices", in Proc. International Conferences on Spoken Language Processing (ICSLP), Sept. 2004, pp. 1283-1286.

[14] N. Brümmer, "FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores - Tutorial and User Manual-. http://sites.google.com/site/nikobrummer/focalmulticlass.

[15] R. Auckenhalter, M. Carey, and H. Lloyd-Thomas, "Dcore Normalization for Text-Independent Speaker Verification Systems, Digital Signal Processing, 10(1), 42-54.

[16] D.A. van Leeuwen and N. Brümmer, "Channel-Dependent GMM and Multi-Class Logistic Regression Models for Language Recognition", 2006 IEEE Odyssey: The Speaker and Language Recognition Workshop.

[17] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors", in Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, 28 June - 1 July 2010.

[18] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/czech.html

[19] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/hungarian.html

[20] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/russian.html

[21] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[22] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.-B. M. no, and C. Nadeu, Albayzin speech database: Design of the phonetic corpus, in Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech-Interspeech), Berlin, Germany, September 1993, pp. 175178.

[23] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, Speech Dat Car. A large speech database for automotive environments, in Proceedings of the II Language Resources European Conference, Athens, Greece, June 2000.