The Albayzin 2012 Language Recognition Evaluation

Luis Javier Rodríguez-Fuentes[1], Niko Brümmer[2], Mikel Penagarikano[1], Amparo Varona[1], Germán Bordel[1], Mireia Diez[1]

[1] GTTS, Department of Electricity and Electronics, ZTF/FCT
University of the Basque Country UPV/EHU, Barrio Sarriena, 48940 Leioa, Spain
[2]AGNITIO Research, South Africa

luisjavier.rodriguez@ehu.es, niko.brummer@gmail.com

Abstract

The Albayzin 2012 Language Recognition Evaluation (LRE), carried out from June to October 2012, was the third effort made by the Spanish/Portuguese community for benchmarking language recognition technology. As in previous Albayzin 2008 and 2010 evaluations, the task consisted on deciding whether or not a target language was spoken in a test utterance. The primary condition involved 6 target languages for which there was plenty of training data: English, Portuguese and the four official languages in Spain (Basque, Catalan, Galician and Spanish). A new challenging condition was defined involving 4 target languages for which no training data were available: French, German, Greek and Italian. In both cases, other (Out-Of-Set) languages were also recorded to allow open-set verification tests. An innovative feature of this evaluation, not common to other evaluations, was that audio data for system development and evaluation were extracted from YouTube videos. Also, a new performance metric was proposed, the so called Multiclass Cross-Entropy, summarizing in a single figure the information provided by system scores, without the need to take hard decisions. This paper presents the main features of the evaluation and analyses the performance of the submitted systems on the different conditions, including the confusion among target languages.

Index Terms: Language Recognition Evaluation, YouTube audio, Multiclass Cross-Entropy.

1. Introduction

As in previous editions, the goal of Albayzin 2012 Language Recognition Evaluation (LRE) was to promote the exchange of ideas, to foster creativity and to encourage collaboration among research groups worldwide working on language recognition technology. To this end, a language recognition evaluation was proposed, similar to those carried out in 2008 and 2010 [1, 2], but under more challenging conditions. The application domain moved from TV Broadcast speech to any kind of speech found in the Internet, and no training data was available for some of the target languages (aiming to reflect a common situation for low-resource languages).

The change in the application domain pursued two objectives: first, the task should reflect a practical application (in this case, indexing of multimedia content in the Internet); and second, the task should be challenging enough for state-of-the-art systems to yield a relatively poor performance. Results attained in the Albayzin 2010 LRE showed that a possible key to define such a challenging task may be acoustic variability (channel, noise, music, overlapping speakers, etc.), which is inherent to some media (such as the videos posted by people in the Internet) [3].

Audio signals for development and evaluation, extracted from YouTube videos, were heterogeneous regarding duration, number of speakers, ambient noise/music, channel conditions, etc. Besides speech, signals may contain music, noise and any kind of non-human sounds. Each signal contained from five up to 120 seconds of speech in a single language, except for signals corresponding to Out-Of-Set (OOS) languages, which might contain speech in two or more languages, provided that none of them were target languages.

Overall, the Albayzin 2012 LRE introduced some interesting novelties with regard to previous editions (see [1, 2] for reference) and NIST Language Recognition Evaluations[1]. The task could be still described as spoken language recognition, but the type of signals used for development and test, the number and identity of target languages, some of the conditions for system development (such as the lack of training data) and the evaluation criterion were significantly different. There was an international call for participation and many expressions of interest. Finally, seven groups registered and submitted their systems to the evaluation: two from China, one from France, one from Portugal and three from Spain.

The rest of the paper is organized as follows. The task and the evaluation conditions are described in Section 2. The datasets provided for system development and evaluation and the performance measure specifically defined for this evaluation are described in Sections 3 and 4, respectively. Results are presented and discussed in Section 5, with special attention to the confusion among languages. Finally, conclusions are given in Section 6.

2. Task definition and test conditions

The language recognition task was defined as follows: *given a segment of speech and a set of n languages of interest (target languages), produce a likelihood score for each target language plus an additional score for the Out-Of-Set (OOS) language class, based on an automated analysis of the data contained in the segment*. Although hard language classification decisions were not required, the likelihood scores were required to be well-calibrated so that they could be used to make Bayes decisions. In closed-set language recognition tests, the last score was not used to compute performance. System performance was evaluated with a calibration-sensitive, multi-class cross-entropy criterion, which is explained in Section 4.

2.1. Test conditions

2.1.1. Closed-set vs Open-set

Depending on the set of languages that were allowed to appear in the audio signal, two types of recognition tests were defined:

[1] http://www.nist.gov/itl/iad/mig/lre.cfm

- In *closed-set recognition*, only target languages were expected to appear in the audio signals. In this case, system performance was computed on the subset of test segments containing speech in one of the target languages.
- In *open-set recognition*, the audio signals may contain any language, either a target language or OOS languages. In this case, system performance was computed on the whole set of test segments, including those containing OOS languages.

As we explain in Section 3, whereas the training set did not provide data for OOS languages, both the development and evaluation sets included segments in OOS languages (with different distributions). The set of OOS languages was not disclosed to participants.

2.1.2. Plenty of Training vs Empty Training

Two different conditions were defined depending on the availability of training materials for target languages, in order to check to what extent the availability of training materials (and thus specific models) for target languages affected system performance. In fact, two separate tasks were defined depending on this condition, since they involved disjoint sets of target languages:

- The first condition, called *Plenty of Training*, involved 6 target languages (those used for the Albayzin 2010 LRE): Castilian Spanish, Catalan, Basque, Galician, Portuguese and English. For all of them, a large amount of training data (around 18 hours of speech per language) was supplied, specifically speech signals recorded from TV broadcasts used to build the Kalaka-2 database [4]. Development signals (YouTube audio) were also supplied, both for target languages (around 150 signals per language) and for Out-Of-Set languages (around 500 signals), to allow tuning systems for open-set tests.
- The second condition, called *Empty Training*, involved 4 target languages: French, German, Greek and Italian, for which no training materials were supplied. Only development signals (YouTube audio) were supplied, both for target languages (around 150 signals per language) and for Out-Of-Set languages (around 500 signals). Under this condition, the training and development data supplied for target languages in the *Plenty of Training* condition could be also used. Note also that development signals provided for OOS languages were shared by both conditions.

Though development signals were provided for tuning purposes, participants were free to use part of them for training models. In any case, participants were only allowed to use the data provided for this evaluation, which should be seen as a common starting condition, necessary for the comparison of systems to depend only on the applied technologies. The only exception to this rule and for the sole purpose of preventing some approaches to be penalised, was that auxiliary subsystems trained on external data (e.g. phonetic decoders) were allowed.

2.1.3. Evaluation tracks

Unlike previous editions of the Albayzin LRE, neither the duration nor the acoustic conditions (presence of background noise or music, etc.) of test segments were taken into account to define different evaluation tracks. There were just 4 tracks, combining the two tasks described in Section 2.1.2 and the two recognition modalities described in Section 2.1.1:

- Plenty of Training, Closed-Set Recognition (briefly, PC)
- Plenty of Training, Open-Set Recognition (briefly, PO)
- Empty Training, Closed-Set Recognition (briefly, EC)
- Empty Training, Open-Set Recognition (briefly, EO)

The first track (PC) was mandatory, meaning that participants were required to submit at least one complete set of recognition results for that condition. The PO, EC and EO tracks were optional. Participants could submit multiple sets of recognition results (each corresponding to a different system) for each track, but they were required to specify one of them as *primary system*, the remaining being *contrastive systems*. To determine the ranking in each track (in terms of the evaluation measure, as defined in Section 4), only primary systems were taken into account.

3. Data

3.1. Training data

Training data provided for this evaluation amounted to around 108 hours of speech, with 18 hours on average for each one of the 6 target languages considered in the Plenty-of-Training condition. Speech files were extracted from the materials used to produce KALAKA-2 (the database created for the Albayzin 2010 LRE) [2]. All of them were continuous excerpts (of different durations) from multi-speaker TV broadcast recordings, featuring various speech modalities and diverse environment conditions.

The training dataset consisted of two disjoint subsets, including clean speech (around 86 hours) and noisy speech (around 22 hours), respectively. Clean-speech segments were high SNR speech signals. Noisy-speech segments may include different and variable types of noise. Telephone-channel speech was not included in any case. In all cases, each training segment contained speech in a single language.

3.2. Development and evaluation data

The development and evaluation datasets were similar in size and structure. As noted above, audio contents were extracted from YouTube videos. The goal was to have around 300 videos validated for each target language and around 100 videos validated for each Out-Of-Set language. A preliminary study was carried out using a few words in Spanish and considering different video categories, as defined by the provider. Eventually we focused our efforts on the six categories more likely to contain speech: (1) Education; (2) News; (3) Entertainment; (4) Howto; (5) Nonprofit; and (6) Technology. Then, a large list of YouTube videos was created for each language, based on different criteria, the most important being the existence of a *Creative Commons* license, but also the occurrence of words in the language of interest (as tags) and the geographical location of the publisher (constraining the search to a certain radius around a major city increased the chances of finding audio in the language of interest). A length criterion was also applied: all the videos in the list were between 30 and 120 seconds long.

Videos in the list were sequentially audited by human experts until the target number of validated videos was attained, according to the following criteria:

1. the amount of speech should be enough to make possible the recognition of the language;
2. speech was found only in the target language, or in several OOS languages; and
3. videos with very noisy or poor-quality speech were discarded.

Note that speech signals collected in this way may still feature quite challenging background and/or channel conditions. Audio files were given random names, so that language labels were kept undisclosed.

At the end of the process, there were 4.168 validated videos out of 21.860 audited videos. The 2.059 videos used for development were mostly extracted from the News, Education and Howto categories, whereas the 2.109 videos used for evaluation were primarily extracted from Entertainment, Nonprofit and Technology. Among OOS languages, all the videos in Czech, Croatian, Polish and Romanian were used for development and all the videos in Bulgarian, Finnish, Slovak and Serbian were used for evaluation, whereas the videos in Hungarian, Russian and Ukrainian were split fifty-fifty between development and evaluation.

4. Evaluation of system performance

In this evaluation, a new metric —based on a form of empirical *multiclass cross-entropy* criterion— has been used for the first time as primary metric to evaluate SLR performance. This new metric measures the information provided by a SLR system through a set of log-likelihoods and does not require making hard decisions (see [5] for details).

To compute the new metric, a prior distribution over language classes must be specified, so that Bayes' rule can be used to map the submitted log-likelihoods to language class posteriors. The goodness of these posteriors is then evaluated by means of a logarithmic cost function. A weighted average of the logarithmic cost over all audio segments forms the cross-entropy criterion.

In order to facilitate comparison of the cross-entropies across different tasks, which have different perplexities, we show how to present cross-entropy in the form of *relative confusion*, a measure closely related to perplexity.

Prior: let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$ represent a prior distribution over the $m = n + 1$ language classes, so that $\pi_i = P(L_i|\boldsymbol{\pi})$. We specify:

$$\boldsymbol{\pi} = \left(\frac{1-\pi_m}{n}, \ldots, \frac{1-\pi_m}{n}, \pi_m \right) \tag{1}$$

For the closed-set condition, we specify $\pi_m = 0$. For the open-set condition, we specify $\pi_m = \frac{1}{m}$.

Posterior: given a log-likelihood-vector, $\boldsymbol{\ell}_t = (\ell_{1t}, \ell_{2t}, \ldots, \ell_{mt})$, the posterior distribution is calculated:

$$P(L_i|\boldsymbol{\ell}_t, \boldsymbol{\pi}) = \frac{\pi_i \exp(\ell_{it})}{\sum_{j=1}^{m} \pi_j \exp(\ell_{jt})} \tag{2}$$

The mapping (2) is the familiar softmax function. In what follows, we refer to the whole posterior distribution as:

$$\boldsymbol{\Pi}_t = \big(P(L_1|\boldsymbol{\ell}_t, \boldsymbol{\pi}), \ldots, P(L_m|\boldsymbol{\ell}_t, \boldsymbol{\pi}) \big) \tag{3}$$

Logarithmic cost function: for every audio segment, s_t, the system under evaluation submits the log-likelihood-vector, $\boldsymbol{\ell}_t$. The evaluator has access to the *true class label* for segment s_t, which we denote $L_{\mathrm{true}(t)} \in \{L_1, \ldots, L_m\}$. This allows the evaluator to compute a measure of goodness for $\boldsymbol{\ell}_t$, in the form of the *logarithmic cost function*:

$$C_{\log}(\boldsymbol{\Pi}_t | L_{\mathrm{true}(t)}) = -\log P(L_{\mathrm{true}(t)}|\boldsymbol{\ell}_t, \boldsymbol{\pi}) \tag{4}$$

Multiclass cross-entropy: we form our *evaluation criterion*, known as *multiclass cross-entropy*, by a weighted average of the logarithmic cost:

$$\mathcal{C}_{\text{mce}} = \sum_{i=1}^{m} \frac{\pi_i}{\|\mathcal{T}_i\|} \sum_{t \in \mathcal{T}_i} -\log P(L_i|\boldsymbol{\ell}_t, \boldsymbol{\pi}) \tag{5}$$

where $\mathcal{T}_i$ is the subset of indices for segments of class i. By $\|\mathcal{T}_i\|$ we mean the number of segments of language class i. Note that for the closed-set case, when $\pi_m = 0$, all segments of class L_m are effectively ignored[2].

The default system: the one that cannot make up its mind about the language class and outputs $\ell_{it} = k_t$ for every t. This gives $P(L_i|\boldsymbol{\ell}_t, \boldsymbol{\pi}) = \pi_i$ for every i, t.

$$\mathcal{C}_{\text{def}} = \sum_{i=1}^{m} -\pi_i \log \pi_i \tag{6}$$

which is just the prior entropy[3]. If a submitted system has $\mathcal{C}_{\text{mce}} \geq \mathcal{C}_{\text{def}}$, then it does not improve upon the default system.

Confusion: to facilitate interpretation of cross-entropy, we define the *confusion* of the system under evaluation as:

$$F_{\text{mce}} = \exp(\mathcal{C}_{\text{mce}}) - 1 \tag{7}$$

Similarly, the *prior confusion* (confusion of the default system) is:

$$F_{\text{def}} = \exp(\mathcal{C}_{\text{def}}) - 1 \tag{8}$$

Since cross-entropy is non-negative, a perfect system would have zero confusion. To get an intuitive feeling for confusion, consider the prior confusion for the closed-set case where we have a flat prior over n classes, so that $\mathcal{C}_{\text{def}} = \log n$ and $F_{\text{def}} = n - 1$. This can be interpreted as the number of *wrong* alternatives. Notice that confusion is closely related to perplexity (the *total* number of alternatives).[4] Here we choose to use confusion, rather than perplexity, in order to facilitate comparison across different tasks with different prior perplexities. We do this by defining *actual relative confusion*:

$$F_{\text{act}} = \frac{F_{\text{mce}}}{F_{\text{def}}} \tag{9}$$

The relative confusion is the factor by which the system has changed (hopefully reduced) the prior confusion. The reference value for relative confusion is 1. Badly calibrated systems that have relative confusion greater than one are doing worse than the default system. Good systems must have relative confusion below 1. A perfect system would have relative confusion of zero.

5. Results

Seven groups from four different countries submitted systems to the Albayzin 2012 LRE. Overall, 95 different submissions were made (33 to the PC condition, 22 to the PO condition, 20 to the EC condition and 20 to the EO condition), including late submissions (those made between the established deadline and the release of results and keyfiles) and some post-key submissions (those made after the release of results and keyfiles). These latter were allowed to groups that detected bugs in their submissions and wanted to share the results attained by the fixed systems.

[2] When $\pi_m = 0$, $P(L_m|s_t, \boldsymbol{\pi}) = 0$ and $-\log P(L_m|s_t, \boldsymbol{\pi}) = \infty$, but we can use the limit: $\lim_{\pi_m \to 0} \pi_m \log P(L_m|s_t, \boldsymbol{\pi}) = 0$.

[3] Shannon's entropy.

[4] perplexity = confusion + 1

Most of the submitted systems followed state-of-the-art approaches, including Total Variability Factor Analysis (iVectors) [6] followed by SVM or linear generative classifiers [7], and Parallel PR-SVM [8, 9] based on high-performance phone decoders [10]. In particular, iVector systems were trained on different sets of features, including SDC [11], trigrams of posteriorgram counts [12] and prosodic features [13]. Most systems were built by fusing various independent subsystems, commonly by applying the FoCal toolkit [14, 15].

Table 1 presents the results attained by primary systems and some late primary systems on the four tracks of the Albayzin 2012 LRE, the best result being marked in boldface. Each number refers to a different site and results in a row correspond to systems based on the same technology. Unfortunately, these results cannot be compared to those attained in other evaluations, because the evaluation metric is different. However, we can still compare the performance across test conditions.

Table 1: Performance (in terms of the multiclass cross-entropy measure F_{act}) of primary systems (including some late submissions) in the four tracks of the Albayzin 2012 LRE.

Systems	PC	PO	EC	EO
1	**0.071**	**0.085**	–	–
2	0.078	0.120	0.498	0.516
3	0.113	0.114	0.711	0.796
4	0.121	0.160	0.626	0.676
5	0.122	–	–	–
6	0.141	0.184	–	–
7 (late)	0.407	0.216	–	–
1 (late)	–	–	**0.216**	–
6 (late)	–	–	0.310	**0.372**

As may be expected, best performance was found in the PC condition, because there were plenty of data for the target languages and no OOS trials. When moving to the PO condition, that is, when including OOS trials, performance degraded (around 20% for the best system), but not as much as expected (e.g. in the case of site 3, the performance was the same in both PC and PO). This may be due to the fact that OOS and target languages are too far each other, thus including OOS trials did not significantly increase the difficulty of the task.

On the other hand, when moving from PC to EC, the performance degraded drastically, the best figure in EC (0.216) being three times worse than the best figure in PC (0.071). Unfortunately, we don't have a reference for system 1 (late), but in the case of system 2, performance degraded in more than 500% and similar percentages result for systems 3 and 4. We expected that using the Plenty-of-Training data to estimate reference models (producing scores for input utterances) and then training a backend based on the development data for target languages (mapping reference scores to target scores) may allow to overcome this issue. But this large degradation suggests that the lack of training data is a really challenging condition.

Once again, when including OOS trials (moving from EC to EO), performance didn't degrade so much. In the case of system 6 (late), which yielded the best performance in EO (0.372), degradation with regard to EC (0.310) was of 20%.

5.1. Confusion among languages

Table 2 shows the confusion among languages for the best primary system in the PO condition. Target languages are identified as *eu* (Basque), *ca* (Catalan), *en* (English), *gl* (Galician), *pt* (Portuguese) and *es* (Spanish). The last two rows show the average false alarm probabilities for audio files containing target languages (AVG) and the false alarm probabilities for audio files containing OOS languages (OOS), respectively. In some cases, the AVG figure is similar (for Basque and Catalan) or even greater (for Galician and Spanish) than the OOS figure. This means that OOS languages used in this evaluation did not make the task more difficult, but just the opposite in some cases. This is consistent with the small degradations observed when moving from closed-set to open-set recognition.

Table 2: Confusion matrix for the best system in the PO condition of the Albayzin 2012 LRE. *Miss probabilities* (%) are shown in the diagonal and *false alarm probabilities* (%) out of the diagonal.

		Target language					
		eu	**ca**	**en**	**gl**	**pt**	**es**
Test audio	**eu**	4.00	8.00	0.00	5.33	1.33	5.33
	ca	5.70	5.70	0.63	6.33	1.27	9.49
	en	0.00	3.85	4.49	1.92	3.85	0.00
	gl	6.88	10.62	0.00	8.75	2.50	26.87
	pt	0.61	3.07	0.00	3.07	6.13	2.45
	es	7.14	14.94	0.00	20.13	0.65	4.55
	AVG	4.07	8.01	0.13	7.36	1.92	8.83
	OOS	5.59	10.61	3.91	3.72	11.73	6.70

Attending to these results, the most confused languages were, by far, Galician and Spanish, then Catalan and Spanish and, in third place, Catalan and Galician. This is consistent with previous findings in former editions of Albayzin LRE, and can be explained by the common origins and close evolution of the three languages, which share some features. Besides, most Catalan and Galician speakers have Spanish as their second language or even as their mother language.

6. Conclusions

In this paper, we have described the main features of the Albayzin 2012 LRE and have briefly analyzed the results attained by the submitted systems. A number of novelties made this evaluation specially interesting for the language recognition community: (1) YouTube audio was used for development and evaluation; (2) a condition involving four target languages with no training data was proposed; and (3) a new evaluation metric was defined, based on scores (log-likelihoods) and not on hard decisions, which allows for application-independent assesment of language recognition technology. After an international call for participation and many expressions of interest, seven sites from four different countries submitted their systems to this evaluation, including state-of-the-art approaches. The low performance attained in some conditions demonstrated that the proposed tasks were really challenging.

7. Acknowledgements

We thank all the members of the Organizing Committee of Iberspeech 2012 for their help and support. We also thank all the participants for their work and feedback.

This work has been supported by the University of the Basque Country, under grant GIU10/18 and project US11/06; and by the Government of the Basque Country, under program SAIOTEK (project S-PE12UN055); Mireia Diez is supported by a 4-year research fellowship from the Department of Education, University and Research of the Basque Country.

8. References

[1] L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, and A. Varona, "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, pp. 172–179.

[2] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "The Albayzin 2010 Language Recognition Evaluation," in *Proceedings of Interspeech*, Firenze, Italia, August 28-31 2011, pp. 1529–1532.

[3] L. J. Rodriguez Fuentes, A. Varona, M. Diez, M. Penagarikano, and G. Bordel, "Evaluation of Spoken Language Recognition Technology Using Broadcast Speech: Performance and Challenges," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 25-28, 2012.

[4] L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "KALAKA-2: a TV broadcast speech database for the recognition of Iberian languages in clean and noisy environments," in *Proceedings of the LREC*, Istanbul, Turkey, 23-25 May 2012.

[5] L. J. Rodriguez Fuentes, N. Brummer, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, *The Albayzin 2012 Language Recognition Evaluation Plan (Albayzin 2012 LRE)*, URL: http://iberspeech2012.ii.uam.es/images /PDFs/albayzin_lre12_evalplan_v1.3_springer.pdf.

[6] N. Dehak, P. 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.

[7] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Firenze, Italy, 2011, pp. 861–864.

[8] H. Li, bin Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 271–284, January 2007.

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

[10] P. Matejka, P. Schwarz, J. Cernocky, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Proceedings of Interspeech*, Lisboa, Portugal, September 2005, pp. 2237–2241.

[11] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to language identification using Gaussian mixture models and Shifted Delta Cepstral features," in *Proceedings of ICSLP*, 2002, pp. 89–92.

[12] L. F. D'Haro, O. Glembek, O. Plchot, P. Matejka, M. Soufifar, R. Cordoba, and J. Cernocky, "Phonotactic Language Recognition using i-vectors and Phoneme Posteriorgram Counts," in *Interspeech 2012*, Portland, Oregon, USA, 9-13 September 2012.

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

[14] N. Brümmer and D. van Leeuwen, "On calibration of language recognition scores," in *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

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