Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Description and analysis of the Brno276 system for LRE2011

Niko Brümmer,[2] Sandro Cumani,[3] Ondřej Glembek,[1] Martin Karafiát,[1] Pavel Matějka,[1]
Jan Pešán,[1] Oldřich Plchot,[1] Mehdi Soufifar,[1] Edward de Villiers[2] and Jan "Honza" Černocký[1]

(1) Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czech Republic
(2) AGNITIO, South Africa
(3) Politecnico di Torino, Italy

Abstract

This paper contains a description of data, systems and fusions developed by the joint team of Brno University of Technology (BUT), Politecnico di Torino (PoliTo) and AGNITIO for the NIST 2011 Language Recognition Evaluation. The primary submission was a fusion of one acoustic and three phonotactic systems, with extensive use of sub-space projections for both approaches. The results are analysed from the view-point of the new NIST measure involving the $N = 24$ worst language pairs. Some of the results are compared to the MIT-LL submission. As in our previous work, we conclude that having lots of carefully processed data is as important as having good algorithms.

1. Introduction

The goal of this paper is to describe the Brno276 system for the NIST 2011 LRE, which was created in a joint effort by BUT, Agnitio and PoliTo. The submission name, "Brno276", reflects to location where the work was done and the number of language pairs.

The Brno276 primary submission included four systems — one acoustic and three phonotactic:

1. i-vector-2048FG (acoustic i-vector extractor)
2. PHN-HU-i-vector (phonotactic i-vector extractor)
3. PHN-RU-PCA (PCA)
4. PHN-ENG-BT (binary decision tree)

schematically depicted in Figure 1. The details of the systems, calibration and fusion will be discussed later in the paper. We also submitted two contrastive systems—see section 7.3. Our systems make extensive use sub-space projections, mainly in the form of i-vectors [1].

This work has again confirmed that careful preparation and pre-processing of training and development data is crucial for a well performing system. The paper includes our own analysis and comparison of our results to the ones of MIT Lincoln Lab.[1]

The paper is organized as follows: Section 2 describes the construction of data-sets used for training, development and testing prior to the evaluation, as well as the actual evaluation data. Section 3 explains the acoustic and phonotactic front-end system types and section 4 explains their actual setup in our system. Sections 5 describes the recognizers based on logistic regression and section 6 covers the fusion, calibration and decision making. Section 7 summarizes the results of individual systems as well as fusions. The analysis (comparison with MIT-LL and investigation of phonotactic versus acoustic systems) is in section 8, and section 9 concludes the paper.

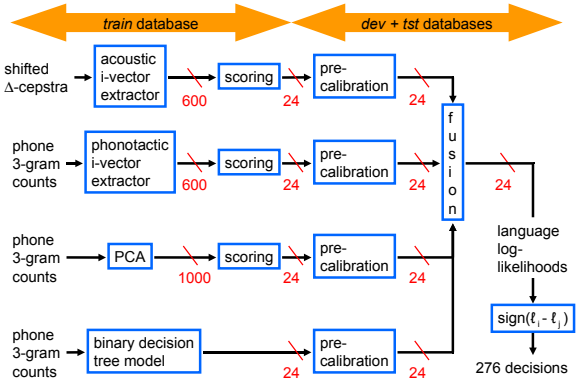

Figure 1: Block diagram of the submitted system.

2. Data

2.1. Development data

Table 1 lists the freely or commercially available data (distributed by the LDC and ELRA) used to train our systems.

There was insufficient data available for the following languages: Czech, Farsi, Arabic Maghrebi, Mandarin, Russian and Ukrainian, so we downloaded additional radio data from the following public sources: Radio Free Europe, Radio Free Asia, Czech Broadcasts, Voice of America (VOA). We performed phone-call detection [2] and for each language, ran automatic speaker labelling to prevent repetition of the same speaker in different sets, in a similar way to our work for NIST LRE 2009 [3].

LDC development data distributed for LRE 2011 was used with caution:

- on *annotated conversations*, automatic speaker labelling was run (like was done for radio-data) to prevent overlapping speakers in different sets. We have denoted this set as `Lre11d1` or "trusty data".
- *Entire conversations* were split into 30, 10 and 3 sec-

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, by Czech Ministry of Education project No. MSM0021630528 and by European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070).

[1] Published with permission of MIT-LL

Table 1: *Publicly available databases used to create our TRAIN, DEV and TEST databases.*

CF	CallFriend
F	Fisher English Part 1. and 2.
F	Fisher Levantine Arabic
F	HKUST Mandarin
SRE	Mixer (data from NIST SRE 2004, 2005, 2006, and 2008)
LRE	development and evaluation data from previous NIST LRE
OGI	OGI-multilingual
OGI22	OGI 22 languages
FAE	Foreign Accented English
SpDat	SpeechDat-East (`http://www.fee.vutbr.cz/SPEECHDAT-E` or the ELRA catalog)
SB	SwitchBoard
VOA	Voice of America radio broadcast
RFEL	Radio Free Europe broadcast
AR-IR	Iraqi Arabic Conv. Tel. Speech (LDC2006S45)
AR-MSA	2003 NIST Rich Transc. Eval Data (LDC2007S10)
AR-MSA	Arabic Broadcast News Speech (LDC2006S46)

ond segments. All splits from one conversation were assigned together to one of the TRAIN/DEV/TEST set. This database (`Lre11d2`) therefore had more data, but was less reliable. From the results of our 1st contrastive system, which did not use `Lre11d2`, we will see that `Lre11d2` did help.

The data was separated into three independent subsets, denoted TRAIN, DEV, and TEST. They all contained data from the 24 target languages — see Table 2. The TRAIN subset had about 60 000 segments, the DEV subset about 38 000 segments and the TEST subset about 26 000 segments in total. The DEV and TEST subsets were split into balanced subsets having nominal durations of 3s, 10s and 30s.

The DEV set is based on our LRE09 development set [4], and contains data from previous LRE evaluations up to and including 2007. The data for the new languages include additional segments extracted from longer files from CTS and radio data (which were not contained in the TRAIN set).

The TEST set consisted mainly of NIST LRE09 evaluation data, plus data for the languages new in the 2011 evaluation.

2.2. Evaluation data

The evaluation data contained 7k to 9k segments, depending on the condition—the exact numbers per language are given in Table 3.

3. Front-end types

We used two types of front-end: *acoustic* and *phonotactic*. Here, we give general descriptions of both types, followed by details of each front-end.

3.1. Acoustic

The acoustic system is based on MFCC/SDC acoustic features. This section provides a brief summary of the acoustic feature extraction and UBM training. For more detail, see our previous work [5, 6].

The inputs to the language recognizer are segments of recorded speech of varying duration. The voice activity detection (VAD) is performed by our Hungarian phone recognizer, with all the phoneme classes linked to the 'speech' class.

The acoustic system used the popular shifted-delta-cepstra (SDC) [7] feature extraction. The feature extraction is similar to the BUT LRE 2005 system [6]. Every speech segment is mapped to a variable-length sequence of feature vectors as follows: After discarding silence portions, every 10ms speech-frame is mapped to a 56-dimensional feature vector. The feature vector is the concatenation of an SDC-7-1-3-7 vector and 7 MFCC coefficients (including C0). VTLN, Cepstral mean and variance normalization and RASTA filtering are applied before SDC.

Vocal-tract length normalization (VTLN) performs simple speaker adaptation. We used MAP adaptation from the UBM (single GMM with 32 diagonal Gaussians trained on Switchboard) to derive specific models for each warping factor [8]. Models are retrained using an MMI (Maximum Mutual Information) criterion. The reference warping factors were generated by an LVCSR system. The models are trained only on English data.

A 2048-component, language-independent, maximum-likelihood GMM was trained with the EM-algorithm on the pooled acoustic feature vectors of all 24 languages in the TRAIN data-set. We follow speaker recognition terminology and refer to this language-independent GMM as the *universal background model*, or UBM [9].

3.2. Phonotactic

The phonotactic systems were based on two kinds of phone recognizers: left-context/right-context hybrids and one based on GMM/HMM context dependent models. All the recognizers are able to produce phone strings as well as phone lattices. In case of lattices, posterior-weighted counts ("soft-counts") were used in the subsequent processing [10].

3.2.1. Hybrid phone recognizers

The phone recognizer is based on a hybrid ANN/HMM approach, where artificial neural networks (ANN) are used to estimate posterior probabilities of phonemes from Mel filter bank log energies using the context of 310ms around the current frame. Hybrid recognizers were trained for Hungarian and Russian on the SpeechDat-E databases. For more details, see [11].

Table 2: *Amounts of data in our three databases.*

Language	TRAIN		DEV		TEST	
	#files	#hours	#files	#hours	#files	#hours
arir	476	17.01	579	2.63	585	2.84
arle	3442	158.74	576	2.52	585	2.66
arma	212	6.83	399	1.77	438	2.05
arms	201	4.72	435	2.71	391	1.96
bang	4084	88.54	633	2.32	563	1.91
czec	2279	19.49	1015	5.49	694	4.02
dari	2410	78.83	579	1.79	1167	3.33
engi	1444	4.72	1174	3.82	1167	3.85
engl	14523	423.33	8170	24.81	2615	8.80
fars	2477	96.63	1718	5.53	1540	4.89
hind	1113	41.79	2222	6.68	1922	6.26
laot	147	3.68	357	1.82	336	1.82
mand	2370	100.55	6281	18.43	2976	9.94
pash	6317	102.35	588	1.87	1185	3.52
pjbc	160	4.36	360	2.17	348	2.10
poli	2098	17.63	772	4.02	489	3.30
russ	8792	122.45	3014	10.26	2247	7.68
slvk	1776	13.55	505	3.26	513	3.41
span	2624	115.08	4784	14.44	1155	3.44
tami	623	19.59	900	2.49	888	2.67
thai	267	7.52	943	3.05	595	2.08
turk	262	9.77	579	1.90	1182	3.43
ukra	967	24.07	572	1.91	1535	4.65
urdu	1266	68.65	1016	3.53	1133	3.41
total	60330	1549.91	38171	129.23	26249	94.00

3.2.2. GMM/HMM phone recognizers

The second type of phone recognizer was based on GMM/HMM context dependent state clustered triphone models from an English LVCSR system. The models were trained using 2000 hours of English telephone conversational speech data from the Fisher, Switchboard and CallHome databases. The features are 13 PLP coefficients augmented with their first, second and third derivatives projected into 39-dimensional space using an HLDA transformation. The models are trained discriminatively using fMPE [12] and MPE criterion [13]. VTLN and CMLLR adaptation are used for both training and recognition in a similar manner to Speaker Adaptation Transform (SAT). Triphones were used for phone recognition with a bigram phonotactic model trained on English data only.

4. Front-End descriptions

This section lists details of all the different front-end variants.

4.1. i-vector-2048FG

This is an acoustic system inspired by our speaker recognition system [14] following the popular i-vector paradigm. We used a full covariance UBM to generate zero and first order statistics which are used for training the i-vector extractor. The output is a 600-dimensional vector for every file.

4.2. PHN-RU-PCA

This phonotactic system makes use of n-gram modelling with the dimensionality of the vector with trigrams soft-counts reduced by PCA to 1000. For more details, see [15]. We used a Russian phone recognizer for generation of trigram soft counts.

4.3. PHN-ENG-TREE

Binary decision tree language modelling was based on creating a single language independent tree (referred to as the "UBM") and adapting its distributions to individual language training data, as described in Navratil's work [16]. We used the English phone recognizer to generate 3-gram lattice counts. The output is a 24-dimensional score vector representing likelihoods for all target languages.

4.4. PHN-HU-i-vector

For this system, a low-dimensional multinomial subspace over the trigram counts in the TRAIN set is trained using the approach described in [17]. We use the multinomial subspace model along with hard pruning of the low-frequency trigrams to overcome the problem of the data sparsity (also explained in [17]). The i-vectors are the point estimates of the latent variables describing the coordinates of count vectors in the new low-dimensional sub-space model. The output is a 600-dimensional vector for every file.

5. Logistic regression recognizers

Our i-vector and phonotactic-PCA recognizers were discriminatively trained (on the TRAIN set) via regularized multiclass logistic regression [18]. The input vectors (of dimension 400–1000) were not reduced by LDA prior to logistic regression, but were conditioned by within-class covariance normalization (WCCN). The PCA vectors were length-normalized, i-vectors were not.

The logistic regression optimization was performed using automatic differentiation and a trust-region Newton conjugate-

Table 3: *Scored segments in evaluation database categorized according to language and duration.*

language	#3sec	#10sec	#30sec	total
arir	308	308	308	924
arle	308	308	308	924
arma	305	305	305	915
arms	306	306	306	918
bang	447	447	412	1306
czec	358	358	261	977
dari	398	399	267	1064
engi	416	416	387	1219
engl	452	452	221	1125
fars	405	405	404	1214
hind	416	416	213	1045
laot	158	158	62	378
mand	432	432	360	1224
pash	401	401	383	1185
pjbc	308	308	299	915
poli	381	380	267	1028
russ	441	441	441	1323
slvk	314	314	280	908
span	419	419	419	1257
tami	413	414	414	1241
thai	403	403	375	1181
turk	472	472	276	1220
ukra	186	186	170	542
urdu	478	478	478	1434
total	8925	8926	7616	25467

gradient optimizer [19]. Each recognizer uses an affine transform to convert the D-dimensional vector, $\mathbf{v}_t$, for trial t, into a K-dimensional score-vector, $\mathbf{s}_t$:

$$\mathbf{s}_t = \mathbf{A}\mathbf{T}\mathbf{v}_t + \mathbf{b}, \tag{1}$$

where $K = 24$ is the number of target languages. $\mathbf{T}$ is a D-by-D matrix which effects within-class covariance normalization, such that the mean class-conditional sample covariance matrix over the training data becomes identity. The logistic regression parameters are $\mathbf{A}$, a K-by-D matrix, and $\mathbf{b}$, a K-dimensional vector, and they are trained by minimizing the regularized objective function:

$$\begin{aligned} &\frac{\lambda}{N}\,\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) \\ &-\frac{1}{K\log K}\sum_{i=1}^{K}\frac{1}{N_i}\sum_{t\in\mathcal{S}_i}\log\frac{\exp(s_{it})}{\sum_{j=1}^{K}\exp(s_{jt})}, \end{aligned} \tag{2}$$

where $N = \sum_i N_i$, and s_{it} is the ith component of $\mathbf{s}_t$ and $\mathcal{S}_i$ is the set of N_i training examples for language i. The regularization weight, λ, was set to:

$$\lambda = \left(\frac{1}{N}\sum_{t\in\mathcal{S}}\sqrt{\mathbf{v}_t^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{T}\mathbf{v}_t}\right)^2, \tag{3}$$

where $\mathcal{S} = \bigcup_i \mathcal{S}_i$. We refer to the two terms of (2) as the *regularization penalty* and the *multiclass cross-entropy* and we note that they were almost equal at the minimum, suggesting that regularization plays an important role.

6. Fusion, calibration and decision taking

6.1. Pre-calibration

Each of the recognizers was independently pre-calibrated with an affine transform, trained on the independent DEV data-set (for the testing prior to the arrival of evaluation data) and on DEV+TEST for the "hot evaluation":

$$\mathbf{r}_t = \mathbf{C}\mathbf{s}_t + \mathbf{d} \tag{4}$$

where $\mathbf{C}$ is a full K-by-K matrix and $\mathbf{d}$ is a K-dimensional vector. Note that the pre-calibration does not change the score-vector dimensionality. These parameters were again trained by regularized logistic regression, but here, no WCCN was applied.[2]

After pre-calibration, zero vectors, $\mathbf{r}_t = \mathbf{0}$, were inserted for those segments for which the basic recognizers failed to produce scores or input vectors.

6.2. Fusion

Let $\mathbf{r}_{ti}$ denote the outputs of the ith pre-calibrated recognizer. These outputs were fused as:

$$\boldsymbol{\ell}_t = \sum_i \alpha_i \mathbf{r}_{ti} + \boldsymbol{\beta} \tag{5}$$

where each α_i is a scalar weight and $\boldsymbol{\beta}$ is a K-dimensional vector. These parameters were again trained by multiclass logistic regression on the DEV subset (and later on DEV+TEST for the evaluation). Here, neither WCCN, nor regularization, were applied.

The output vector of the fuser can be interpreted as 24 language log-likelihoods.

6.3. Pair scores and decisions

The 276 pair scores were formed as differences between pairs of the language log-likelihoods output by the fuser. The decisions were made by thresholding at zero.

7. Results

All results are presented with the new NIST metric: average of actual decision operating point cost function values over the most difficult 24 language pairs (determined by the greatest values of minimum cost operating points for 30-second segments)—see the evaluation plan [20] for details.

7.1. Acoustic system

Table 4 presents the results for the submitted acoustic system (with full covariance) in the first line. In the post-analysis, we performed a comparison with a diagonal covariance system. Although the performance of the full-covariance system was better on the development data (for 30s), the classical (and simpler) diagonal covariance model outperformed the full covariance one on the evaluation data, but their performances were similar in absolute numbers.

We have also experimented with training the UBM as well as i-vector extractor on more languages than the target 24 (54 languages from our work for NIST 2009 LRE submission [4]) but found the training on target languages to produce the best results.

[2]The decision whether to apply WCCN in the particular steps was based on running experiments on our development set. The chosen configuration was the one which gave the best performance.

Table 4: *Results of the acoustic system.*

	Development			Evaluation		
NIST 24 [%]	30s	10s	3s	30s	10s	3s
Full covariance UBM, ivec 600	**4.19**	9.70	19.03	10.39	17.64	28.16
Diagonal covariance UBM, ivec 600	4.33	**9.65**	**18.72**	**10.35**	**17.20**	**27.70**

7.2. Phonotactic systems

Table 5 shows the results of our phonotactic systems. While we have seen the superiority of i-vector and decision tree systems in the development, the evaluation turned these results upside-down: of the three phonotactic systems, the relatively simple PCA system with Russian phone recognizer performed the best on the eval data, and its hit compared to development data was in the same range as that of the acoustic system. Although slightly worse, the i-vector performed predictably on the eval data.

On the other hand, the decision tree system failed badly on the eval data. To investigate the cause and rule out possible problems with the phone recognizer, we performed tests with the same phone recognizer (Hungarian) with all three modelling techniques. The results in Table 6 suggest that the decision tree system suffered the biggest hit while changing from development to evaluation data. The exact causes are still unclear to us and will have to be investigated.

7.3. Fusions

Figures 2 to 4 compare three values of the average cost computed over the worst 30 language pairs: *actual* (with threshold zero, as stated in section 6.3), *min* with threshold optimized for every language pair on the eval data and *star** with an offset optimized for every language on eval data. The left bar in the figure denotes results on the development data and the right one on the eval data.

The results are shown for the individual systems and their fusions:

- **primary**: including all four systems
- **3sys**: excluding the misbehaving PHN-ENG-TREE system
- **Contrastive1**: a fusion of all four systems but with the (presumably less reliable) `Lre11d2` set was excluded from the calibration.
- **Contrastive2**: included only two systems — English decision tree one (which was a bad choice as we have seen) and acoustic i-vectors.

Table 7 contains detailed results of these fusions for all conditions.

As for the individual systems, we have seen a big deterioration in minDCF for evaluation versus development. There were no calibration disasters, but 30s could have been better. The hit of the decision tree system is obvious also on the figures. On average, the acoustic system outperform phonotactic systems for all durations, but this is not true for all language pairs (see section 8.2). As expected, the fusion helped.

8. Analysis

8.1. Comparison with MIT-LL

The new NIST metric is different from the previous years in that the list of the most difficult pairs is *different* for different sites.

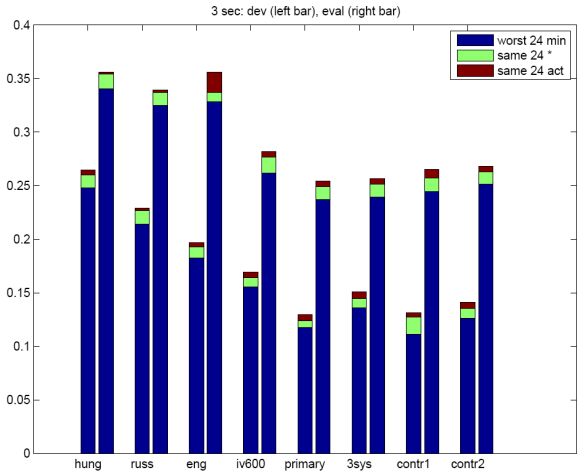

Figure 2: Results of individual systems and fusions for 3s condition.

Therefore, we were eager to compare our system with the better performing submission from MIT-LL [21].

The analysis we ran on the 30s condition has shown that there is only a weak correlation between sites in the difficulty of pairs and that:

- both sites are very similar when minDCF is averaged over all language pairs.
- Brno276 is slightly worse than MIT-LL for minDCF estimated on the site-dependent worst 24 pairs.
- Brno276 is significantly worse on the site-dependent worst 24 pairs in terms of actDCF, suggesting a calibration hit.

To further study the differences, we generated a plot showing the five worst pairs for each site (Figure 5). It is obvious that the results presented correspond largely to the efforts the different sites devoted to different types of the data: while MIT-LL performs very well on the Arabic and English dialects, Brno276 is much better on Slavic languages where the data collection and pre-processing efforts were concentrated.

This proves again that the preparation of training and calibration data is crucial in the system development and that having sub-optimal resources for a language or group of languages can severely impair the system especially for the new NIST metric.

8.2. Phonotactics better than acoustics?

In the previous sections, we have noted the superiority of acoustic system over the phonotactic ones even for the 30s condition where the systems have been traditionally on par. Based on a question "are there language pairs where phonotactics is superior?" raised by George Doddington at the evaluation workshop,

Table 5: *Results of the phonotactic systems.*

	Development			Evaluation		
NIST 24 [%]	30s	10s	3s	30s	10s	3s
ENG-BT	**6.70**	**12.24**	**22.58**	22.28	27.75	35.58
RU-PCA	7.76	14.58	26.13	**14.32**	**23.90**	**33.91**
HU-i-vector	**6.68**	15.00	27.95	15.42	24.59	35.61

Table 6: *Investigation of three modelling techniques with the same (Hungarian) phone recognizer.*

	Development			Evaluation		
NIST 24 [%]	30s	10s	3s	30s	10s	3s
BT	7.91	13.26	24.08	21.87	26.17	33.45
PCA	6.30	15.00	27.91	16.12	25.03	35.99
i-vector	6.68	15.00	27.95	15.42	24.59	35.61

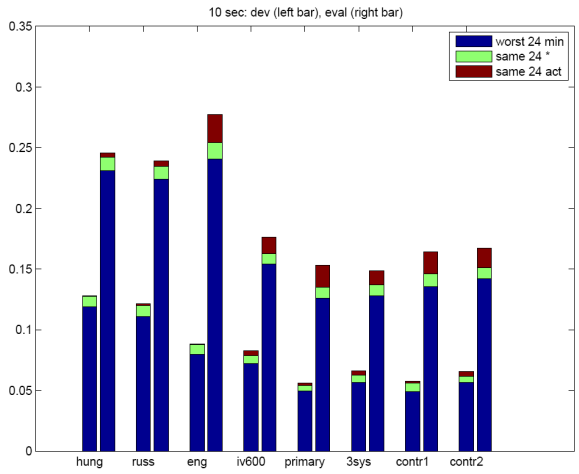

Figure 3: Results of individual systems and fusions for 10s condition.

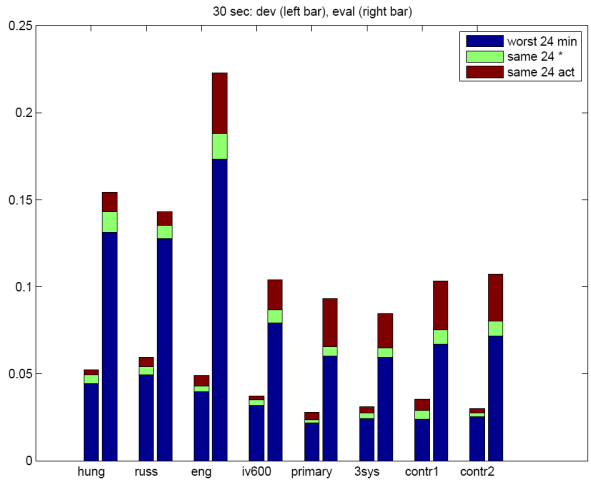

Figure 4: Results of individual systems and fusions for 30s condition.

we have generated a list of pairs where the phonotactic systems indeed beat the acoustic system — see Figure 6. We have found

- 19 pairs where PHN-RU-PCA performs better than the acoustic system.
- 9 pairs where PHN-HU3-i-vector performs better than the acoustic system.
- 5 pairs where both phonotactic systems perform better.

The results are difficult to interpret. A first comparison with the sizes of data (Table 2) would suggest that phonotactic systems outperform the acoustic system in cases with little training data; this is however not valid for the Ukrainian/Russian pair with abundant data. We might also suspect the labelling of data — for example Ukrainian is very similar to Russian for East Ukrainian speakers and the performance might depend heavily on the region in which the speakers were sampled. Similar cases are likely to occur in other language pairs. Also, we cannot rule out over-training of the acoustic system on a particular transmission channel (especially for Indian and Pakistani languages): in this case, phonotactics should provide better performance.

9. Conclusions

The paper describes the four systems that were included in the Brno276 submissions to NIST 2011 LRE and discusses their results from the viewpoint of the new NIST metric. Generally, we have seen that the acoustic system outperformed the phonotactic ones, although exceptions exist and deserve further investigation. Of the phonotactic systems, an i-vector and a simple PCA one performed well in the evaluation and largely helped the acoustic one in the fusion. The decision tree system that provided the best performance on our development data failed in the evaluation and further investigation is needed to find causes. We must still decide what our next development steps should be, but it seems that (1) special detectors for selected pairs of languages and (2) better development set design are strong candidates.

10. References

[1] Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of Interspeech*, Brighton, UK, Sept.

Table 7: *Results for fusions.*

	Development			Evaluation		
NIST 24 [%]	30s	10s	3s	30s	10s	3s
primary	3.11	6.19	13.87	9.33	15.34	25.41
3sys	3.01	6.49	14.96	8.47	14.84	25.68
contr1	4.14	6.68	14.60	10.33	16.44	26.50
contr2	3.42	7.47	15.73	10.74	16.74	26.83

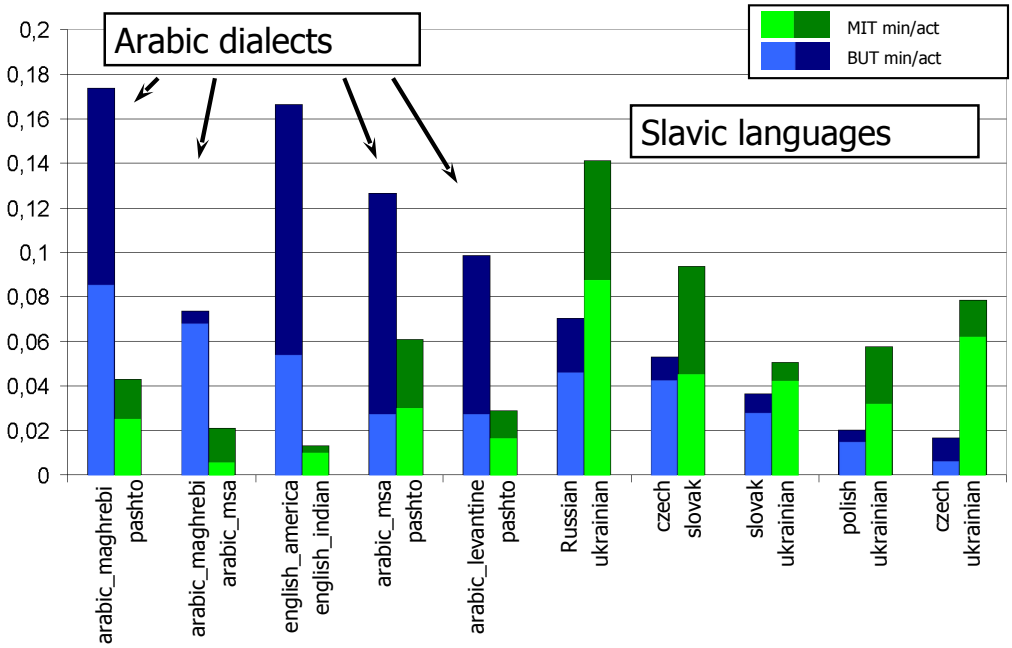

Figure 5: Comparison of MIT-LL and Brno276 on selected language pairs.

2009, pp. 1559–1562.

[2] Oldřich Plchot, Valiantsina Hubeika, Lukáš Burget, Petr Schwarz, and Pavel Matějka, "Acquisition of telephone data from radio broadcasts with applications to language recognition," in *Proceedings of the 11th International Conference on Text, Speech and Dialogue*, Brno, Czech Republic, 2008, pp. 477–483.

[3] Zdeněk Jančík, Oldřich Plchot, Niko Brümmer, Lukáš Burget, Ondřej Glembek, Valiantsina Hubeika, Martin Karafiát, Pavel Matějka, Tomáš Mikolov, Albert Strasheim, and Jan "Honza" Černocký, "Data selection and calibration issues in automatic language recognition — investigation with BUT-AGNITIO NIST LRE 2009 system," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[4] Niko Brümmer, Lukáš Burget, Ondřej Glembek, Valiantsina Hubeika, Zdeněk Jančík, Martin Karafiát, Pavel Matějka, Tomáš Mikolov, Oldřich Plchot, and Albert Strasheim, "BUT-AGNITIO system description for NIST language recognition evaluation 2009," in *Proceedings of the NIST-LRE Workshop*, Baltimore, Maryland, June 2009, Online: `http://www.fit.vutbr.cz/research/groups/speech/publi/2009/brummer_BUT_AGNITIO_LRE09_SYSD.pdf`.

[5] Valiantsina Hubeika, Lukáš Burget, Pavel Matějka, and Petr Schwarz, "Discriminative training and channel compensation for acoustic language recognition," in *Proceedings of Interspeech*, Brisbane, Australia, Sept. 2008.

[6] Pavel Matějka, Lukáš Burget, Petr Schwarz, and Jan "Honza" Černocký, "Brno University of Technology system for NIST 2005 language recognition evaluation," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, June 2006, pp. 57–64.

[7] Pedro A. Torres-Carrasquillo, Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds, and John R. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, Colorado, USA, Sept. 2002, pp. 89–92.

[8] Lutz Welling, Stephan Kanthak, and Hermann Ney, "Improved methods for vocal tract normalization," in *Proceedings of the International Conference on Acoustics,*

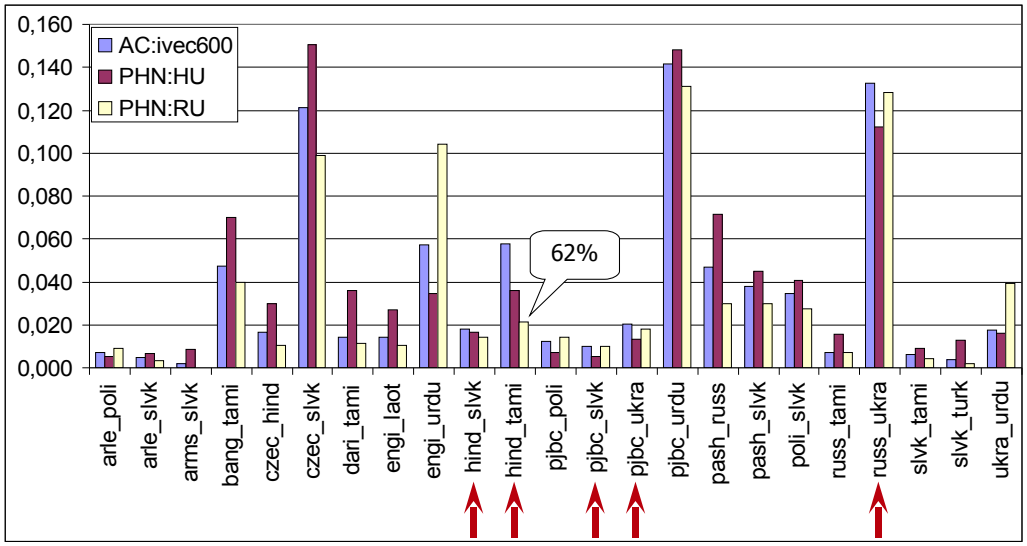

Figure 6: Comparison of phonotactic and acoustic systems on the 30s condition.

Speech, and Signal Processing, Phoenix, Arizona, USA, Mar. 1999, pp. 761–764.

[9] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[10] Jean-Luc Gauvain, Abdel Messaoudi, and Holger Schwenk, "Language recognition using phone lattices," in *Proceedings of the 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, Oct. 2004, pp. 1283–1286.

[11] Petr Schwarz, Pavel Matějka, and Jan "Honza" Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proceedings of the 31st International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, pp. 325–328.

[12] Daniel Povey, "Improvements to fMPE for discriminative training of features," in *Proceedings of the 9th European Conference on Speech Communication and Technology, EUROSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 2977–2980.

[13] Daniel Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, University of Cambridge, Cambridge, UK, Mar. 2003.

[14] Pavel Matějka, Ondřej Glembek, Fabio Castaldo, Jahangir Alam, Oldřich Plchot, Patrick Kenny, Lukáš Burget, and Jan "Honza" Černocký, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.

[15] Tomáš Mikolov, Oldřich Plchot, Ondřej Glembek, Pavel Matějka, Lukáš Burget, and Jan "Honza" Černocký, "PCA-based feature extraction for phonotactic language recognition," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010.

[16] Jiří Navrátil, "Recent advances in phonotactic language recognition using binary-decision trees," in *Proceedings of Interspeech*, Pittsburgh, Pennsylvania, USA, Sept. 2006.

[17] Mehdi Soufifar, Marcel Kockmann, Lukáš Burget, Oldřich Plchot, Ondřej Glembek, and Torbjørn Svendsen, "I-vector approach to phonotactic language recognition," in *Proceedings of Interspeech*, Florence, Italy, Aug. 2011.

[18] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics. Springer, 2007.

[19] Jorge Nocedal and Stephen J. Wright, *Numerical Optimization*, Springer, second edition, 2006.

[20] The National Institute of Standards and Technology, "The 2011 NIST language recognition evaluation plan," `http://nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev1.pdf`, Mar. 2011.

[21] Najim Dehak, Alan McCree, Douglas Reynolds, Fred Richardson, Elliot Singer, Doug Sturim, and Pedro Torres-Carrasquillo, "MITLL 2011 language recognition evaluation system description," in *Proceedings of the NIST-LRE Workshop*, Atlanta, Georgia, USA, Dec. 2011.