TRAINING CANDIDATE SELECTION FOR EFFECTIVE REJECTION IN OPEN-SET LANGUAGE IDENTIFICATION

Qian Zhang, John H.L. Hansen

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering
University of Texas at Dallas, Richardson, Texas, U.S.A.
{qian.zhang, john.hansen}@utdallas.edu

ABSTRACT

Research in open-set language identification (LID) generally focuses more on accurate in-set modeling versus improved out-of-set (OOS) rejection. Unknown or OOS language rejection is a challenge, since research developers seldom commit equivalent OOS corpus development effort versus the desired in-set languages. To address this, we propose an OOS candidate selection method for universal OOS language coverage. Since effective selection always requires abundant knowledge of inter-language relationships, three broad measurements across world languages are considered. Finally, the advanced OOS selection method is evaluated on a database derived from a large-scale corpus (LRE-09) with a state-of-the-art i-Vector system followed by two back-ends. The baseline system is realized using a random selection of OOS candidates. With the proposed selection method and probabilistic linear discriminative analysis (PLDA) back-end, the OOS rejection performance is improved with false alarm and miss rates achieving a relative reduction of 32.6% and 4.4%, respectively. In addition, the overall classification performance are relatively improved 8.4% and 7.5% according to the two back-ends based on an average cost function.

Index Terms: Open-set language identification, Out-of-set identification, language distance, candidate selection, LRE-09, i-Vector

1. INTRODUCTION

Recently, language identification (LID) has experienced substantial attention in the speech processing community [1,2]. Due to the contribution in targeting interested languages, LID plays an essential role in audio pre-processing which is typically followed by automatic speech recognition (ASR). It is also critical for effective diarization and dialog system in spoken language. In recent years, robust feature extraction [3] and various discriminative modeling techniques which include both acoustic models (GMM-UBM [4], JFA [5], i-Vector [6]) and phonotactic models(PPRLM [7], PRSVM [8]) have been proposed with great success for language identification/recognition. However, most research addresses closed-set LID, where all the target languages are known. In real scenarios, open-set LID is more general, where training data might not cover all possible OOS languages. A system is still required to recognize in-set languages and effectively reject unknown/out-of-set(OOS) languages. Here, our study is mainly focused on effective rejection for open-set LID.

Notably, the key problem in dealing with open-set LID is to select effective and universal training data for OOS language modeling. Subsequently, system OOS rejection performance would be reduced by refining the boundary between in-set and OOS languages. However, due to the financial constraints and geographical limitations in data collection, it is essential to select the most efficient languages as candidates representing the entire OOS. The same issue also exists in speaker identification (SID) which is explored using cohorts [9] to leverage a wealth of available data as the impostor set. However, it is generally easier to find OOS speakers for open-set SID than OOS languages for open-set LID. In this study, we propose a method for compact OOS candidate language selection to boost LID system classification performance.

Since accurate selection always requires knowledge concerning subject content and inter-language relationships, three measurements for the distance across world languages are considered here. The most fundamental is from a linguistic perspective, which involves the language origin and geographical factors. Also, selection of language require aspects related to prosody which contains patterns of stress, rhythms and intonation. Accordingly, pitch pattern distribution analysis is investigated here. Finally, a language tree based on engineering perspective is an effective tool to express classifier distance. With this knowledge, an efficient OOS candidate selection method will be proposed.

The main purpose of this study is to seek an intelligent method for OOS data selection through a comprehensive data analysis, which has not been investigated before. The proposed method is evaluated on a state-of-the-art i-Vector system followed by two classifiers: Gaussian back-end (GB) and probabilistic linear discriminative analysis (PLDA) back-end [10]. Performance analysis will be based on two criteria.

This paper is organized as follows: Sec. 2 elaborates on the design database and the baseline system set-up. Three types of distance based language tree partitions are detailed in Sec. 3. Based on this estimated and combined knowledge, the proposed selection method is described in Sec. 4. Sec. 5 analyzes the results and illustrates performance. Finally, conclusions are summarized in Sec. 6.

2. EXPERIMENT SET-UP

2.1. Database

With no loss of generality, the data set design in this study is derived from the large-scale corpus NIST LRE-09 [11] (40 languages in total). In order to include a second corpus test, and to be consistent with NIST LRE-09, the DARPA-sponsored Robust Automatic Transcription of Speech (RATS) [12] is used where five similar languages are assigned as in-set targets and the remaining thirty-five languages are defined as OOS candidates. Since the goal is to explore an efficient method for OOS candidate selection, at this point, any single language and combinations are accessible as training data for modeling. To be fair, since we only have five in-set targets, the quantity of OOS representatives used for training is also set to be five, which is the same as the amount of in-set languages. In other words, the proposed method seeks to find the best five out of thirty-five OOS languages as candidates for modeling in order to achieve optimal performance. The data partition and distribution are well organized, especially for OOS data, which in theory would be approximately uniform in distribution. The detailed information are showed in Table 1.

Table 1: *Corpus statistics.*

	In-set	OOS
Language	Dari Farsi Hindi Pashto Urdu	Amharic Arabic : Ukrainian Uzbek
Total number of languages	5	35
Average duration(*sec/file*)	20.9	20.7
Count of training files	6566	3361
Count of test files	3550	1343

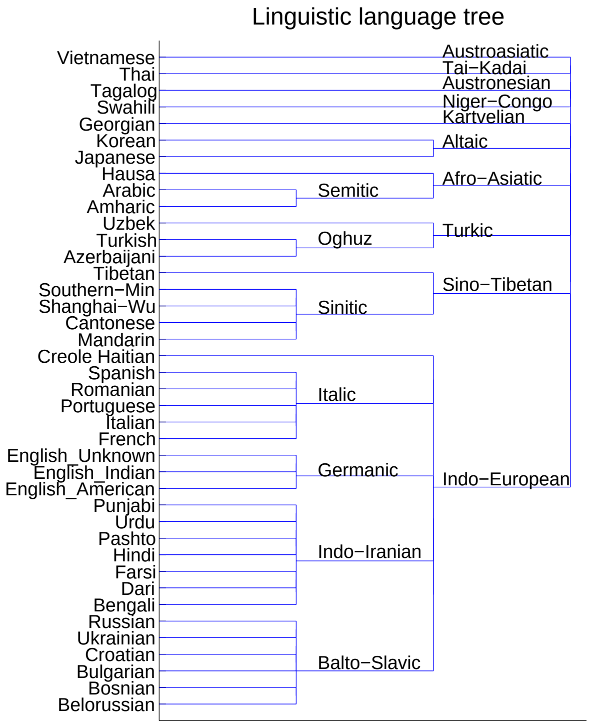

Fig. 1: *Language family based on linguistic knowledge.*

Table 2: *Frequency of pitch patterns for English (%).*

Contour	—	╱	╲	——	—╱	—╲	╱╱	╱—	╱╲	╲—	╲╱	╲╲
American	37.7	24.2	38.1	20.6	7	11.5	7.3	8.9	8.4	12.6	0	15.9
Indian	40.3	24.4	35.4	23.1	7.7	10.8	7.8	8.8	8.1	12.1	0	13.6
Unknown	42.6	20.2	37.2	25.6	7	11.6	7.8	7.3	6.3	13.5	0	15.1

2.2. System set-up

In this study, modeling is based on a state-of-the-art i-Vector framework. Voice activity detection [13] is performed prior to the extraction of Mel-frequency cepstral coefficients (MFCCs). In addition, shifted delta cesptra [14] based processing [7-1-3-7], are applied for all original MFCCs features to derive a 56 dimensional feature set for further i-Vector extraction. After feature extraction, we follow the i-Vector system paradigm for language recognition presented in [6,15]. All training data is used to build a 1024-mixture Universal Background Model (UBM) and estimate the total variability matrix using EM algorithms for Eigenvoice as presented in [16]. To suppress redundant information, a dimensionality reduction based on linear discriminative analysis (LDA) is applied to the 400 dimension i-Vectors. In the end, two types of classifiers, Gaussian back-end and PLDA, are explored.

3. LANGUAGE DISTANCE

Since the main purpose of this study is to select more effective OOS candidates for modeling, the process is to determine whether a language is typical in representing the entire OOS for discriminating the in-set targets. Compared with target languages, the measurement of confusability is needed for a reasonable decision. Therefore, knowledge concerning the relationship or distance between various world languages is essential for addressing this issue. Three types of distance criteria are investigated: (i) fundamental linguistic perspective, (ii) classical prosodic knowledge, and (iii) engineering perspective.

3.1. Distance criteria #1: Linguistic language tree

Based on the origin of each language and the corresponding geographical position, a linguistic language tree for these 40 languages is shown in Fig. 1 which includes two-tier language family information. More specifically, the name of the corresponding language family /sub-family is indicated at the end of each cluster. It can be noted that all in-set targets we designed belong to the same branch named Indo-Iranian.

3.2. Distance criteria #2 : Prosody based language tree

Prosody is one of most effective features previously employed for speech/language related analysis and classification [17]. The patterns of stress, rhythms and intonation across different languages vary greatly. In addition, some tonal languages employ tone contours to distinguish different meanings, are included in our data set. To implement prosodic analysis, N-gram pitch pattern distributions are considered [18]. Here, we extract pitch contours of each utterance using WaveSurfer [19], then voiced islands that consist of nonzero F0 values are detected. Next, a median filter is applied for contour smoothing. For exploring subtle variations on pitch, each island is processed by a sliding window of length ($T_{win} = 50ms$) with a step ($T_{step} = 25ms$). We also employ a linear regression strategy to obtain the actual pitch slope within each window. However, for simplicity, only directions (rising/flat/falling) are retained by comparing each to a given threshold. Theoretically, the statistics of the N-gram pitch patterns in a language could express the unique prosodic characteristics. In this study, only unigram and bigram based frequency of patterns are considered as prosodic features. For example, it is shown in Table 2, that even for the same language (English) with different accents results in quite different pitch pattern distributions. This is also consistent with human perception. A prosody based language tree (similar to Fig. 1) is also developed. Here, hierarchical clustering is based on the cosine distance (Eq. 1) between prosodic feature vectors,

$$Cosine\ distance = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

where A and B are feature vectors. The cosine distance between any two different languages is measured with normalization. The sum distance across each in-set language provides a general distance between every OOS language versus the whole in-set group. Finally, a prosody feature based OOS confusability rank (shown in Table 3) is generated according to the general distances.

3.3. Distance criteria #3 : Engineering language tree

Compared with the prosodic strategy, a more efficient and directive way to investigate world language relationships is by evaluating each

OOS Rank	Prosody	Acoustic
1	Georgian	Spanish
2	Azerbaijani	Creole_haitian
3	Hausa	Ukrainian
4	Portuguese	Bosnian
5	French	Russian
6	Turkish	Portuguese
7	Bulgarian	Punjabi
8	Tagalog	English_Unknown
9	English_Indian	Amharic
10	Croatian	Tagalog
11	Arabic	Turkish
12	Spanish	Azerbaijani
13	Uzbek	Japanese
14	Romanian	Arabic
15	Korean	Vietnamese
16	English_American	Bulgarian
17	Italian	Cantonese
18	Creole_haitian	Croatian
19	Cantonese	Bengali
20	Bengali	Hausa
21	Bosnian	Georgian
22	Belorussian	Italian
23	Russian	Tibetan
24	Tibetan	Thai
25	Swahili	Swahili
26	English_Unknown	Shanghai-wu
27	Ukrainian	English_Indian
28	Punjabi	Belorussian
29	Vietnamese	English_American
30	Southern-min	Uzbek
31	Japanese	Southern-min
32	Thai	Korean
33	Amharic	French
34	Shanghai-wu	Romanian
35	Mandarin	Mandarin

Table 3: *OOS confusabilty rank.*

language based on performance of the development data (here, this is the same as training data).

To be specific, the confusabilities among languages which belong to the same language family are different. For example, the Chinese languages Mandarin and Cantonese, are quite different based on pronunciation, even though they share the same written form. Therefore, it is easy to distinguish them base on acoustic features. However, the Indian languages Hindi and Urdu, represent one of most confusable language pairs in LRE-09 [11]. In a similar manner, some languages belong to different language families according to Wikipedia, however might be very similar, such as Vietnamese and Thai. In brief, an assessment solution for restricted OOS language selection that only depends on linguistic knowledge may not provide consistent performance.

In addition, it is arbitrary to quantize the pair-wise confusability, which is essential for both automatic LID performance and OOS candidate selection. Since all in-set languages in our study are from the same language sub-family, a more comprehensive and dedicated relationship analysis is needed for effective OOS candidate selection. In addition, the ideal OOS model should be discriminative from in-set models, while capturing a variety of unknown factors as much as possible. From an engineering perspective, the score/probability assigned for each utterance according to different models are expected to be the best explanation on their mutual relationship.

In this study, the system was based on a state-of-the-art i-Vector system followed by two types of back-ends, GB and PLDA. More specifically, for back-end processing, each in-set language possesses a corresponding individual model, and one general model is used for evaluating all OOS languages. Subsequently, each test utterance is assigned 6 scores (score vector) according to each model. For simplicity, only the average score vector is counted as the new feature for each language. Instead of single score, the score patterns are adopted for distance calibration. Similar to Sec. 3.2, the cosine distances across score vectors are employed as the strategy for clustering and pair-wise distance calculation. A language tree similar to Fig. 1 is form based on engineering distance which shows cluster relationships between each OOS and in-set language. A pair-wise relationship also was explored for precise information used for OOS candidate selection. Similarly, the derived OOS confusability rank table (shown in Table 3) is sorted according to the ascending general distance.

4. PROPOSED SELECTION METHOD

This section proposes an efficient OOS candidate selection method. Of course, any five out of thirty-five combination is possible. However the main purpose is to find an intelligent way to address this selection issue. A good candidate combination is required to be discriminative from each in-set target, meanwhile, it also needs to be general to cover OOS diversity. The optimal five languages should complement each other with a minimum redundant coverage. Accordingly, three requirements are proposed as follows:

- Linguistic family restriction: in general, all candidates should come from different linguistic language families/sub-families.
- Prosodic restriction: from a prosodic perspective, all the candidates should scatter in terms of the prosody based OOS confusability rank.
- Engineering restriction: according to the engineering OOS confusability rank , some candidates should be close to in-set languages; while others far away to reflect more general OOS properties.

To avoid selecting candidates that are too intensive in terms of the general distance to the in-set languages, a prosodic restriction is proposed to building a more diverse model set that covers the entire OOS languages. In addition, some candidates may have similar properties with in-set languages, so we need to refine the decision boundary. Notably, if two languages are quite similar, it is very difficult to classify them accurately. Therefore, for better performance, the most troublesome OOS languages should potentially be included in model training. However, if the entire OOS model only represents these in-set confusing languages, more general languages could be mislabeled. Therefore, leveraging close/far trade-offs could reduce both false alarm and miss rates in real scenarios. In order to demonstrate the impact of three restrictions, a set of experiments are designed for comparison.

5. RESULT ANALYSIS

This section analyses all open-set/closed-set LID system performance for the particular scenario designed in Sec 2. To elaborate experiments on different OOS training set-ups, two types of measurement criteria were adopted here. The first one is evaluating the

Table 4: *Experiments set-up.*

Restriction	OOS training/test lang. No.	Comments on OOS candidate selection	Candidates^{rank}	Task type	Abbr.
No restriction	35/35	all data		closed-set	35 vs. 35
	5/35	close to in-set languages	Spanish^{1}, Creole_haitian^{2}, Ukrainian^{3}, Bosnian^{4}, and Russian^{5}	open-set	5 vs. 35(near)
	5/35	far to in-set languages	Southern-min^{31}, Korean^{32}, French^{33}, Romanian^{34}, and Mandarin^{35}	open-set	5 vs. 35(far)
Only linguistic family restriction	5/5	random selection	Ukrainian^{3}, Swahili^{25}, French^{33}, Mandarin^{35}, and Turkish^{11}	closed-set	5 vs. 5
	5/35	random selection	same as above	open-set	5 vs. 35(random)
	5/35	close to in-set languages	Spanish^{1}, Creole_haitian^{2}, Ukrainian^{3}, Punjabi^{7}, English_Unknown^{8}	open-set	5 vs. 35(near_1R)
	5/35	far to in-set languages	Belorussian^{28}, English_American^{29},Uzbek^{30}, Romanian^{34}, and Mandarin^{35}	open-set	5 vs. 35(far_1R)
Linguistic family & prosodic restriction	5/35	close to in-set languages	Spanish^{1}, Creole_haitian^{2}, Ukrainian^{3}, Amharic^{9}, Turkish^{11}	open-set	5 vs. 35(near_2R)
	5/35	far to in-set languages	Swahili^{25}, Belorussian^{28}, English_American^{29}, Romanian^{34}, and Mandarin^{35}	open-set	5 vs. 35(far_2R)
All three restrictions	5/35	proposed method (4 near & 1 far-away)	Spanish^{1}, Creole_haitian^{2}, Ukrainian^{3}, Amharic^{9}, and Mandarin^{35}	open-set	5 vs. 35(proposed)

overall classification performance using the standard criterion average cost function (C_{avg}) [11]. While, to better analyze In-set/OOS classification performance, a binary confusion matrix is employed which illustrates details about false alarm and miss rates on OOS rejection performance.

In addition, two close-set benchmark systems were evaluated for comparison. More details are shown in Table 4, where any selection scheme utilizes the optimal option defined by the OOS confusability rank (see Table. 3). For an instance, the selection scheme is "near to in-set languages only with linguistic family restriction". According to confusability rank, the optimal candidates are the top 5 languages which are not belong to same linguistic family. Similarly, random selection was semi-supervised random generation only with linguistic family restriction.

To illustrate all the experiments' performance systematically, we analyzed them in three perspectives as follows:

5.1. Analysis on linguistic family restriction

According to the linguistic language tree mentioned in Sec. 3.1, the first requirement in proposed OOS selection is linguistic family related restriction. Candidates without linguistic family restriction might contain redundant information. Therefore, how does that impact OOS selection? Two groups of comparable experiments are shown in Fig. 2. It is interesting to note that linguistic family restriction is effective on near selection package (comparing "near" and "near_1R") while relative futile on far selection package (comparing "far" and "far_1R"). The near package is focused on making the OOS model more discriminative from in-set languages, while the far package could cover more diverse properties across the entire OOS. Therefore the language family related redundancy is more sensitive to near package. However, the impact of linguistic family restriction is non-negative.

Table 5: *System performance (100*C_{avg}).*

Experiment abbr.	GB	PLDA
35 vs. 35	15.1	13.4
5 vs. 5	16.6	14.3
5 vs. 35 (random)	17.8	16.1
5 vs. 35 (near_2R)	17.0	15.6
5 vs. 35 (far_2R)	16.8	15.1
5 vs. 35 (proposed)	16.3	14.9

5.2. Analysis on prosodic restriction

Similarly, with the same distance based selection scheme, we compare the performance between with and without prosodic restriction. It can not noted that overall classification performance are relatively improved 2.75% and 2.6% according to near and far-away selection method, respectively. Therefore, the prosodic restriction can benefit the performance with relatively diverse and comprehensive coverage on prosody.

5.3. Benefits on proposed selection scheme

All the performance in this section are based on the non-negative impact of first two requirements which have already been proved. The main focus here is analyzing the benefits of proposed selection scheme, especially compared with other engineering distance based selection methods.

From overall classification performance perspective, Table 5 shows the corresponding performances based on two closed-set benchmark systems and the system using four different OOS candidate selection schemes. Generally speaking, PLDA outperforms GB across the different experiments settings. In addition, it can not noted that the proposed selection scheme with 4 near and 1 far-away

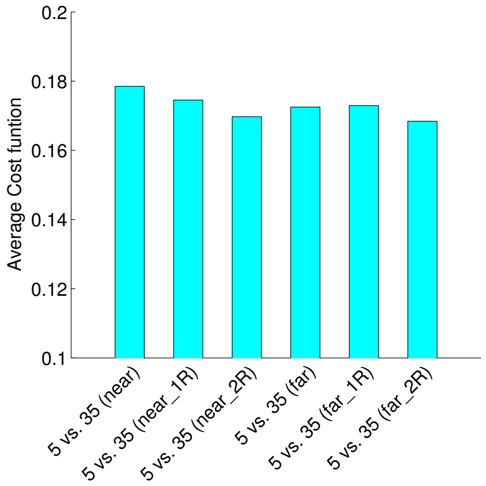

Fig. 2: *Analysis on Restrictions (near/far means distance based selection scheme; suffix 1R represents the selection only with linguistic family restriction, and 2R means the selection with both linguistic family and prosodic restriction).*

Table 6: *Confusion matrix based on PLDA(%).*

35 vs. 35		
true pred.	in-set	OOS
in-set	88.5	11.5
OOS	18.4	81.6

5 vs. 5		
true pred.	in-set	OOS
in-set	80.1	19.9
OOS	15.8	84.2

5 vs. 35 (random)		
true pred.	in-set	OOS
in-set	84.0	16.0
OOS	36.8	63.2

5 vs. 35 (near_2R)		
true pred.	in-set	OOS
in-set	83.7	16.3
OOS	27.5	72.5

5 vs. 35 (far_2R)		
true pred.	in-set	OOS
in-set	85.4	14.6
OOS	29.2	70.8

5 vs. 35 (proposed)		
true pred.	in-set	OOS
in-set	84.7	15.3
OOS	24.8	75.2

candidates achieves the best performance for the open-set experiment. Assuming the open-set task based on random OOS candidates selection is our baseline system, the best proposed method achieve relatively 8.4% and 7.5% improvements according to GB and PLDA, respectively.

Furthermore, instead of overall classification performance, the OOS rejection performance also need to be optimize for effective candidate selection. To better analyze binary (In-set/OOS) classification task, a confusion matrix is employed to express false alarm and miss rates. The performance based on PLDA is shown in Table 6. It is noted that the false alarm rate increases significantly from 15.8% (closed-set: 5 vs. 5) to 36.8% (open-set: 5 vs. 35(random)) when a five language trained model was evaluated on the entire OOS languages. Again, this is the reason why this study is so meaningful. Our goal is to optimize the open-set performance (5 vs. 35) to approach the benchmark system (35 vs. 35) by choosing the most effective OOS training data. The best proposed candidate selection method reduces the false alarm from 36.8% to 24.8% and the miss rate from 16.0% to 15.3%, with relatively 32.6% and 4.4% improvement respectively.

6. CONCLUSION

This study has focused on addressing a particular open-set LID problem: given data from several in-set languages and limited/none OOS languages, how do you select the best limited number of OOS languages to achieve effective OOS language rejection? Solving this problem is effectively asking which language is more important and worth the time and resource to invest for data collection? We proposed an OOS candidate selection method based on knowledge of world-language distance. The baseline system was realized by a random selection of OOS candidates. With the proposed OOS selection method, OOS rejection performance false alarm and miss rates are relatively reduced by 32.6% and 4.4%, respectively. In addition, the overall classification performance are relatively improved 8.4% and 7.5% according to two back-ends based on average cost function.

7. REFERENCES

[1] Najim Dehak, Alan McCree, Douglas Reynolds, Fred Richardson, Elliot Singer, Doug Sturim, and Pedro Torres-Carrasquillo, “Mitll 2011 language recognition evaluation system description,” in *Proc. NIST 2011 Language Recognition Evaluation Workshop, Atlanta, USA*, Dec. 2011.

[2] Niko Brümmer, Sandro Cumani, Ondrej Glembek, Martin Karafiát, and Pavel Matejka, “Description and analysis of the brno276 system for lre2011,” in *Proc. Odyssey : The Speaker and Language Recognition Workshop, Singapore*, Jun. 2012, pp. 216–223.

[3] Leena Mary and B. Yegnanarayana, “Extraction and representation of prosodic features for language and speaker recognition,” *Speech Communication*, vol. 50, no. 10, pp. 782 – 796, 2008.

[4] Pedro A Torres-Carrasquillo, Douglas A Reynolds, and JR Deller Jr, “Language identification using gaussian mixture model tokenization,” in *Proc. ICASSP, Orlando, USA*, May 2002, vol. 1, pp. 757–760.

[5] Najim Dehak, Pierre Dumouchel, and Patrick Kenny, “Modeling prosodic features with joint factor analysis for speaker verification,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2095–2103, 2007.

[6] David Martınez, Oldrich Plchot, Lukás Burget, Ondrej Glembek, and Pavel Matejka, “Language recognition in ivectors space,” in *Proc. Interspeech, Firenze, Italy*, Sep. 2011, pp. 861–864.

[7] Wade Shen, William Campbell, Terry Gleason, Doug Reynolds, and Elliot Singer, “Experiments with lattice-based pprlm language identification,” in *Proc. Odyssey: Speaker and Language Recognition Workshop, Puerto Rico*, 2006, pp. 1–6.

[8] Qian Zhang, Hynek Boril, and John HL Hansen, “Supervector pre-processing for prsvm-based chinese and arabic dialect identification,” in *Proc. ICASSP, Vancouver, Canada*, May 2013, pp. 7363–7367.

[9] Z.N. Karam, W.M. Campbell, and N. Dehak, “Towards reduced false-alarms using cohorts,” in *Proc. ICASSP, Prague, Czech Republic*, May 2011, pp. 4512–4515.

[10] Gang Liu, Taufiq Hasan, Hynek Boril, and John HL Hansen, “An investigation on back-end for speaker recognition in multisession enrollment,” in *Proc. ICASSP, Vancouver, Canada*, May 2013, pp. 7755–7759.

[11] Alvin Martin and Craig Greenberg, “The 2009 nist language recognition evaluation,” in *Proc. Odyssey, Brno, Czech Republic*, Jun. 2010, pp. 165–171.

[12] Sibel Yaman, Jason W Pelecanos, and Mohamed Kamal Omar, "On the use of non-linear polynomial kernel svms in language recognition.," in *Proc. Interspeech, Portland, USA*, Sep. 2012.

[13] Lee Ngee Tan and Abeer Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech," *Speech Communication*, vol. 55, no. 7, pp. 841–856, 2013.

[14] Mary A Kohler and M Kennedy, "Language identification using shifted delta cepstra," in *Proc. 45th Midwest Symposium on Circuits and Systems, MWSCAS*, 2002, vol. 3, pp. 69–72.

[15] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas A Reynolds, and Reda Dehak, "Language recognition via i-vectors and dimensionality reduction.," in *Proc Interspeech, Florence, Italy*, Aug. 2011, pp. 857–860.

[16] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[17] David Martinez, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel, "Prosodic features and formant modeling for an ivector-based language recognition system," in *Proc. ICASSP, Vancouver, Canada*, May 2013, pp. 6847–6851.

[18] Hynek Boril, Qian Zhang, Pongtep Angkititrakul, John HL Hansen, Dongxin Xu, Jill Gilkerson, and Jeffrey A Richards, "A preliminary study of child vocalization on a parallel corpus of us and shanghainese toddlers," in *Proc. Intersppech, Lyon, France*, Aug. 2013.

[19] Kåre Sjölander and Jonas Beskow, "Wavesurfer-an open source speech tool.," in *Proc. Interspeech, Beijing, China*, Oct. 2000, pp. 464–467.