INTEGRATING ACOUSTIC, PROSODIC AND PHONOTACTIC FEATURES FOR SPOKEN LANGUAGE IDENTIFICATION

Rong Tong[1,2]*, Bin Ma*[1]*, Donglai Zhu*[1] *, Haizhou Li*[1] *and Eng Siong Chng*[2]

[1]Institute for Infocomm Research, Singapore
[2]School of Computer Engineering, Nanyang Technological University, Singapore
[1]{tongrong,mabin,dzhu,hli}@i2r.a-star.edu.sg, [2]aseschng@ntu.edu.sg

ABSTRACT

The fundamental issue of the automatic language identification is to explore the effective discriminative cues for languages. This paper studies the fusion of five features at different level of abstraction for language identification, including spectrum, duration, pitch, *n*-gram phonotactic, and *bag-of-sounds* features. We build a system and report test results on NIST 1996 and 2003 LRE datasets. The system is also built to participate in NIST 2005 LRE. The experiment results show that different levels of information provide complementary language cues. The prosodic features are more effective for shorter utterances while the phonotactic features work better for longer utterances. For the task of 12 languages, the system with fusion of five features achieved 2.38% EER for 30-sec speech segments on NIST 1996 dataset.

1. INTRODUCTION

Automatic language identification (LID) is a process of determining the language identity corresponding to a given spoken query. It is an important technology in many applications, such as spoken language translation, multilingual speech recognition and spoken document retrieval.

Recent studies have explored different levels of speech features which include articulatory parameters [1], spectral information [2], prosody [3], phonotactic [2] and lexical knowledge [4]. It is generally believed that spectral feature and phonotactic feature provide complementary language cues to each other [5]. Human perception experiments also suggest that prosodic features are informative language cues [1]. However, prosodic feature has not been fully exploited in LID [6]. In general, LID features fall into five groups according to their level of knowledge abstraction as shown in Figure 1. Lower level features, such as spectral feature, are easier to obtain but volatile because speech variations such as speaker or channel variations are present. Higher level features, such as lexical/syntactic features, rely on large vocabulary speech recognizer, which is language and domain dependant. They are therefore difficult to generalize across languages and domains. Phonotactic features become a trade-off between computational complexity and performance. It is generally agreed that phonotactics, i.e. the rules governing the sequences of admissible phone/phonemes, carry more language discriminative information than the phonemes themselves. They are extracted from output of a phoneme recognizer, which is supposed to be more robust against effects such as speaker and channel than spectral features. For practicality, research has been focused on acoustic-prosodic-phonotactic features. In this paper, we study how the three levels of language cues, n-gram LM, *bag-of-sounds*, spectral feature, duration and pitch complement in LID tasks.

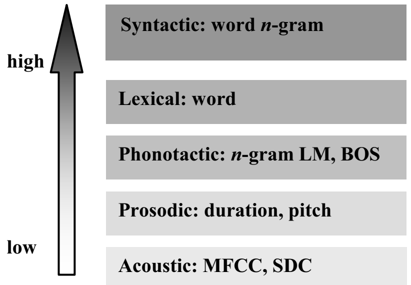

Figure 1 Five levels of LID features

We typically represent a speech utterance as a collection of independent spectral feature vectors. The collection of vectors can be modeled by a Gaussian mixture model, known as GMM [7], that captures the spectral characteristics of a language. The prosody of speech can be characterized mainly by energy, pitch and duration among others. They can be modeled in a similar way as that for spectral feature. Phonotactic features capture the lexical constraint of admissible phonetic combination in a language. One typical implementation is the P-PRLM (Parallel Phone Recognition followed by Language Model) approach that employs multiple phoneme recognizers that tokenize a speech waveform into phoneme sequences and then characterizes a language by a group of *n*-gram language models (LM) over the phoneme sequences [2]. A new phonotactic model, known as *bag-of-sounds* was proposed recently to model utterance level phonotactics collectively. Its language discriminative ability is comparable to that of the *n*-gram LM [8][9].

In this paper, we study five LID features: *n*-gram LM in P-PRLM, *bag-of-sounds*, spectral feature, pitch and duration. In Section 2, the development and evaluation databases are introduced. In Section 3, the feature fusion LID system is described. In Section 4, we report the experiment results. Finally we conclude in Section 5.

2. DEVELOPMENT AND EVALUATION DATA

The NIST 1996 and 2003 language recognition evaluation (LRE)

1-4244-0469-X/06/$20.00 ©2006 IEEE

sets are used to evaluate the performance of the LID systems. There are 12 target languages in both sets: Arabic, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, Vietnamese and English. Dialects of English, Mandarin and Spanish are also included in 1996 test set. In 2003 test, there are test segments in Russian. Each language consists of test segments in 3 length groups: 30, 10 and 3 seconds.

The development data come from CallFriend corpus [10]. We use the same 12 languages and 3 dialects as the target languages specified in the NIST LRE. In CallFriend corpus, data for each language are grouped into 3 parts: 'train', 'devtest' and 'evaltest'. We are using 'train' and 'devtest' as our development data.

All the development and test data are pre-processed by a speech activity detection program to remove silence. In the development process, we treat the dialects of English, Mandarin and Spanish as different languages. Therefore, there are 15 languages in the training process. For our results to be comparable with other reports in the literature, in the test process, we only measure the LID performance of the 12 primary languages by grouping the dialect labels into their respective primary language.

3. SYSTEM DESCRIPTION

One of the solutions to fuse multiple features is the ensemble method. An ensemble of classifiers is a set of classifiers whose individual decisions are combined in the classification process. Our five-feature fusion LID system is formulated in this way. In this section, we discuss five member classifiers in the ensemble.

3.1 *n*-gram LM in P-PRLM

Following the P-PRLM formulation as in [2], seven phoneme tokenizers are used in our system: English, Korean, Mandarin, Japanese, Hindi, Spanish and German. English phonemes are trained from IIR-LID [11] database. Korean phonemes are trained from LDC Korean corpus (LDC2003S03). Mandarin phonemes are trained from MAT corpus [12]. Other phonemes are trained from OGI-TS corpus [13]. 39-dimensional MFCC features are extracted from each frame. Utterance based cepstral mean subtraction is applied to the MFCC features to remove channel distortion. Each phoneme in the languages are modeled with a HMM of 3-state. The English, Korean and Mandarin states are of 32 mixtures each, while others are of 6 mixtures considering the availability of training data. Based on the phoneme sequence from each tokenizer, we train up to 3-gram phoneme LM for each tokenizer-target language pair, resulting in $105 = 15 \times 7$ LMs. For each input utterance, 105 interpolated language scores are derived to form a vector. In this way, a set of training utterances are represented by a collection of 105-dimensional score vectors. The score vectors are normalized by subtracting the mean of their competing languages.

The P-PRLM classifier consists of 15 pairs of Gaussian mixture models (GMMs), known as the backend classifier. For each target language, we build two GMMs $\{m^+, m^-\}$. m^+ is trained on the score vectors of target language, called positive model, while m^- is trained on those of its competing languages, called negative model. The confidence of a test utterance O is given by the likelihood ratio $\lambda_{PPRLM} = \log(p(O \mid m^+) / p(O \mid m^-))$.

3.2 Bag-of-Sounds (BOS)

The *bag-of-sounds* method uses a universal sound recognizer to tokenize an utterance into a sound sequence, and then converts the sound sequence into a count vector, known as *bag-of-sounds* vector [8]. The *bag-of-sounds* method differs from the P-PRLM method in that it use single universal sound recognizer, with this universal sound recognizer, one does not need to carry out acoustic modeling when adding new language capability to the classifier. Although the sound inventory for the universal sound recognizer can be derived from unsupervised learning [8], in this paper, the universal sound inventory is a combined phoneme set from 6 languages: English, Mandarin, Hindi, Japanese, Spanish and German. There are 258 phonemes in total. The phoneme labeled training corpus of these 6 languages are come from same sources as described in P-PRLM system.

For each sound sequence generated from the universal sound tokenizer, we count the occurrence of bi-phones. A phoneme sequence is then represented as a vector of bi-phone occurrence with 66,564 = 258 $\times$ 258 elements. A Support Vector Machine (SVM) is used to partition the high dimensional vector space [14]. As SVM is a 2-way classifier, we train pair-wise SVM classifiers for the 15 target languages, resulting in 105 SVM classifiers. The linear kernel is adopted when using SVM-light tool.

A training utterance is classified by the 105 SVM classifiers to derive a 105-dimensional score vectors. The collection of training score vectors are used to train a backend classifier in the same way as it is used in P-PRLM. The likelihood ratio for a test utterance can be given by the backend classifier as λ_{BOS} .

3.3. SDC Feature in GMM

Gaussian mixture models are used to model acoustic characteristics of a language, known as GMM acoustic in [5]. We use the shifted delta cepstral (SDC) features [7] to capture long time spectral information across successive frames. The parameter 7-3-1-7 is used as in [5]. We build a set of GMMs to form a classifier. First, a 2,048-mixture Gaussian Mixture model is trained from all the SDC feature vectors of 15 languages, this is the universal background model (UBM). Then, we adapt the UBM towards each target language amounting to 15 language dependent GMMs. We further adapt the language dependent GMM by gender resulting in 30 gender-language dependent GMMs. In summary, we obtain 30 gender-language dependent GMMs, 15 language dependent GMMs and 3 UBMs.

An utterance is evaluated on the 45 GMMs and 3 UBMs to generate 45 language dependent scores in a 45-dimensional vector. The score vectors are normalized by their respective UBM scores. The collection of training score vectors are used to train a backend classifier in the same way as it is used in P-PRLM. The confidence of a test utterance can be given by the backend classifier as λ_{SDC} .

3.4. Duration

As one of the prosodic features, we believe that the phoneme duration statistics provide language discriminative information.

Early research has found that duration is useful in the speaker recognition study [15].

We use the same universal sound recognizer as in *bag-of-sounds* classifier. After tokenization, we obtain duration statistics for each phoneme. The duration feature vector has 3 elements representing the duration of 3 states in a phoneme. For each phoneme in a target language, we train a 16-mixture language-dependent GMM model using the collection of duration features. For each phoneme, we also train a 16-mixture language-independent GMM model as the negative model using the collection of duration features from all its competing phonemes. As a result, we arrive at $3,874 = 258 \times 15$ positive models and 258 negative models.

For each utterance, the likelihood ratios from the 258 positive-negative model pairs are multiplied to generate a score for each language, resulting in a score vector of 15 dimensions representing 15 languages. The collection of training score vectors are used to train a backend classifier in the same way as it is used in P-PRLM. The confidence of a test utterance can be given by the backend classifier as λ_{DUR}.

3.5. Pitch

Pitch feature is another important prosodic feature. It has been used in some speaker recognition tasks [16], but has not successfully used in LID task yet. We initially design pitch features for Chinese dialect identification as Chinese dialects are largely differentiated by different intonation schemes. We have seen promising results [17]. Here we adopt pitch features to build one member classifier in the ensemble.

For given utterance, 11 dimensional pitch features are extracted from each frame [17]. A Gaussian mixture model, i.e. universal background model (UBM), is trained using feature vectors from all languages. Then a GMM model is adapted from the UBM model for each target language. As a result, we build 15 GMM models and one UBM model. All models have 16 Gaussian mixtures each.

An utterance is evaluated on the 15 GMMs and 1 UBM to generate 15 language dependent scores in a 15-dimensional vector. The score vectors are normalized by the UBM score. The collection of training score vectors are used to train a backend classifier in the same way as it is used in P-PRLM. The confidence of a test utterance can be given by the backend classifier as λ_{PIT}.

4. EXPERIMENTS

We conduct experiments on NIST 1996 and 2003 LRE datasets. We use NIST 1996 LRE development data for fine-tuning of the ensemble. With the same resulting setting, we run the test on both 1996 and 2003 datasets.

To investigate how different levels of discriminative features complement each other, we use our P-PRLM classifier as the baseline, and then fuse other classifiers one by one into the ensemble. The fusion is carried by multiplying the likelihood ratio score from individual member classifiers. In the case of 5-feature fusion, we have $\lambda = \lambda_{PPRLM} + \lambda_{BOS} + \lambda_{SDC} + \lambda_{DUR} + \lambda_{PIT}$. Table 1 & 2 show the results for incremental fusion of ensemble with the last row being extracted from Singer et al [5] for comparison. The performance of individual languages and confusion matrix among 12 languages on NIST 1996 30-sec data are shown in Table 3 and Table 4. Figure 2 shows the DET plots on 3-sec NIST 2003 LRE data. The proposed ensemble system significantly outperforms previous reported results on the 3-sec short test utterances and compare favorably on longer test utterances except 30-sec in 2003 LRE.

Method	**30-sec**	**10-sec**	**3-sec**
P-PRLM	2.92	8.23	18.61
P-PRLM+BOS	2.61	7.11	16.98
P-PRLM+BOS+SDC	2.38	6.80	15.70
P-PRLM+BOS+SDC+Duration	2.38	6.35	14.55
P-PRLM+BOS+SDC +Duration+Pitch	2.38	6.26	14.31
MIT fused system [5]	2.70	6.90	17.40

Table 1 EER% of system fusion on NIST 1996 LRE data

Method	**30-sec**	**10-sec**	**3-sec**
P-PRLM	4.54	11.31	20.37
P-PRLM+BOS	4.17	10.03	18.64
P-PRLM+BOS+SDC	3.27	8.55	16.66
P-PRLM+BOS+SDC+Duration	3.27	8.37	15.94
P-PRLM+BOS+SDC+ Duration+Pitch	3.27	7.97	15.54
MIT fused system [5]	2.80	7.80	20.30

Table 2 EER% of system fusion on NIST 2003 LRE data

Language	**EER%**	**#test utterances**
French (FR)	1.30	80
Arabic (AR)	1.76	80
Farsi (FA)	3.15	80
Geman (GE)	3.80	80
Hindi (HI)	7.92	76
Japanese (JA)	1.20	79
Korean (KO)	3.51	78
Tamil (TA)	4.70	73
Vietnamese (VI)	4.38	79
Mandarin (MA)	1.86	156
Spanish (SP)	2.03	153
English (EN)	1.56	478

Table 3 EER% for individual language on NIST 1996 LRE data (30-sec)

	FR	**AR**	**FA**	**GE**	**HI**	**JA**	**KO**	**TA**	**VI**	**MA**	**SP**	**EN**
FR	77	0	0	0	0	0	0	0	1	0	0	2
AR	1	72	3	0	0	0	0	0	0	0	2	2
FA	0	0	74	2	0	0	0	0	2	1	0	1
GE	0	0	5	74	0	0	1	0	0	0	0	0
HI	2	1	3	0	57	0	6	2	1	0	3	1
JA	0	0	0	0	0	76	2	0	0	1	0	0
KO	0	0	2	0	2	1	70	1	0	2	0	0
TA	0	1	1	0	3	1	2	65	0	0	0	0
VI	1	0	0	0	1	0	1	0	72	1	1	2
MA	0	1	0	0	0	0	1	0	0	151	0	3
SP	1	0	0	0	3	2	0	0	1	0	145	1
EN	1	0	5	1	0	0	1	0	2	1	0	467

Table 4 Confusion matrix of NIST 1996 LRE data (30-sec)

To look into the contribution of each member classifier in the ensemble, we break down the EER reductions by individual classifier when it is added into the ensemble, as in Table 5.

As both P-PRLM and BOS systems capture phonotactic features in different way, by fusing the two systems, we gain average 10.2% EER reduction evenly across the board. The P-PRLM classifier extracts phoneme 3-gram statistics and uses perplexity measure to evaluate similarity between languages. The BOS classifiers extract bi-phone statistics, which is similar to phoneme bigram, but projects the statistics into a high dimensional space for SVM to carry out discrimination [8][9].

	30-sec		10-sec		3-sec	
	1996	2003	1996	2003	1996	2003
P-PRLM	-	-	-	-	-	-
BOS	10.6	8.1	13.6	11.3	9.2	8.5
SDC	8.8	21.6	4.4	14.7	7.5	10.6
Duration	0.0	0.0	6.6	2.1	7.3	4.3
Pitch	0.0	0.0	1.4	4.8	1.6	2.5

Table 5 EER reduction (%) by member classifiers in the ensemble

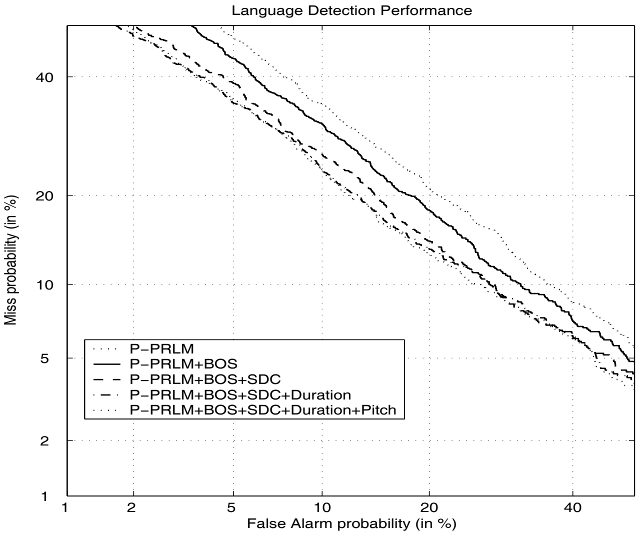

Figure 2 DET curve of fused system on NIST 2003 LRE (3-sec)

The SDC classifier captures low level acoustic information. The results also show that it also significantly contributes to EER reduction across the board. However, the effect is more obvious in 2003 LRE than in 1996 LRE. As for the prosodic based classifiers, we only see effect in 3-sec and 10-sec test cases.

5. CONCLUSIONS

We have proposed an effective ensemble method for LID. The ensemble fuses different levels of discriminative features. We have shown that different levels of information provide complementary language identification cues. It is found that P-PRLM and *bag-of-sounds* features complement each other to fully explore both n-local phonotactics and utterance level collective phonotactic statistics. The P-PRLM and bag-of-sounds classifiers form the backbone of the ensemble. The spectral feature also consistently contributes to the LID tasks. It is found that fusing the lower level acoustic information and high level phonotactic information greatly improves the overall system. We have also successfully integrated the prosodic features into the LID task. The experiment results show that even the simple prosodic feature as pitch and phoneme duration are useful, especially for short speech segments.

The performance of proposed ensemble LID system on NIST 1996 and 2003 LRE datasets are comparable with the best system reported in the literature. The experiments in this paper also re-affirm, from a different angle, the findings in other reports [5] that spectral and phonotactic features are the most effective features for LID.

REFERENCES

[1] Y.K. Muthusamy, N. Jain and R.A. Cole, “Perceptual benchmarks for automatic language identification.” Proc. *ICASSP* 1994, pp. 333-336.

[2] M.A.Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Transactions on Speech and Audio Processing*, 4(1), pp. 31-44, 1996.

[3] A. E. Thymé-Gobbel and S. E. Hutchins, "On using prosodic cues in automatic language identification", *ICSLP 96*, Philadelphia, USA, October 1996

[4] D. Matrouf, M. Adda-Decker, L. Lamel and J. Gauvain, "Language Identification Incorporating Lexical Information", ICSLP 98, Sydney, Australia, December 1998.

[5] E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, and D.A. Reynolds, “Acoustic, Phonetic, and Discriminative approaches to Automatic Language Identification,” in *Proc. Eurospeech 2003*, pp. 1345–1348, Sept. 2003

[6] T. J. Hazen and V.W. Zue. “Recent improvements in an approach to segment-based automatic language identification”. *ICSLP 1994*

[7] Pedro A. Torres-Carrasquillo, et al. “Approaches to Language Identification using Gaussian Mixute Models and Shifted Delta Cepstral Features”, *ICASSP 2002*

[8] B. Ma, H. Li, “A Phonotactic-Semantic Paradigm for Automatic Spoken Document Classification”. *SIGIR2005,* Salvador, Brazil. August 15-19, 2005,

[9] H. Li and B. Ma, "A Phonotactic Language Model for Spoken Language Identification", *ACL05*, Ann Arbor, USA. 2005

[10] Linguistic Data Consortium (LDC), “The CallFriend corpra”, http://www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone

[11] Language Identification Corpus of the Institute for Infocomm Research

[12] H.-C. Wang, MAT-a project to collect Mandarin speech data through networks in Taiwan, in: Int. J. Comput. Linguistics Chinese Language Process. 1 (2) (February 1997) 73-89.

[13] http://cslu.cse.ogi.edu/corpora/corpCurrent.html

[14] SVM-light,http://svmlight.joachims.org/

[15] L. Ferrer et al., “Modeling Duration Patterns for Speaker Recognition”, *Proc. Eurospeech*, Geneva, pp.2017-2020, 2003.

[16] D.A.Reynolds, et al., “The 2004 MIT Lincoln Laboratory Speaker Recognition System”, *ICASSP* 2005

[17] B. Ma, D. Zhu, R. Tong, “Chinese Dialect Identification using Tone Features Based on Pitch Flux”, Submitted to *ICASSP* 2006