LANGUAGE DIARIZATION FOR CONVERSATIONAL CODE-SWITCH SPEECH WITH PRONUNCIATION DICTIONARY ADAPTATION

Dau-Cheng Lyu[1], Eng-Siong Chng[1,2], and Haizhou Li[1,2,3,4]

[1] Temasek Laboratories, Nanyang Technological University, Singapore 639798
[2] School of Computer Engineering, Nanyang Technological University, Singapore 639798
[3] Institute for Infocomm Research, 1 Fusionopolis Way, Singapore 138632
[4] The University of New South Wales, Sydney, NSW 2052, Australia
dclyu@ntu.edu.sg, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

ABSTRACT

Language diarization is the task to perform automatic language segmentation and recognition in a code-switch speech. Towards this task, we developed a conversational Mandarin-English code-switch corpus spoken by Singaporean/Malaysian speakers. We also developed a Singapore accent specific pronunciation dictionary, with which we built a Singapore accent phone recognizer to extract long term context phonotactic feature. Our experiment shows that accent-specific phone recognizer is essential to improve language diarization performance. Specifically, the language diarization experiment, the phonotactic features generated by the Singapore accent phone recognizer has a 6.5% relative frame error rate reduction over the phone recognizer using the CMU dictionary. In addition, the ASR performance using this dictionary on the Singapore English corpus achieved 21% relative word error rate reduction over the system using the American accent CMU dictionary.

***Index Terms*—** language diarization, language recognition, code-switched speech, pronunciation adaptation

1. INTRODUCTION

Code-switch refers to the switching of languages in speech, and is a common occurrence among multilingual speakers [1]. For example, in the following transcription extracted from our Singapore corpus [2] :"通常是他們來了 confirm 了那個 date everything 我們才會知道 (we will know after they come here to confirm the date and everything)", the example shows 4 language turns between Chinese and English in a single utterance. In countries such as United States and Switzerland, studies show that a mixture of Spanish and English or French and Italian are commonly spoken [3]. Cantonese-English code-switched speech is also very common in colloquial Cantonese [4] in Hong Kong, and in Taiwan, Mandarin-Taiwanese code-switched speech is also widespread [5].

In this paper, we propose a system to automatically segment and identify languages in a code-switched utterance. As this task is similar to the speaker diarization task [6] where the objective is to automatically segment and cluster speakers from given utterances, we will call this task 'language diarization'. Language diarization is different from language recognition. In language recognition, the input is a mono-lingual utterance [7] and the task is simply to identify the language identity. However, in language diarization, both language identity and boundary of the language transition are unknown. In addition, we found that the duration of mono-lingual segments of code-switched speech are much shorter than those traditionally studied in the language recognition research [8].

From our previous analysis on the SEAME corpus, we find that code-switch often occurs between word, phrase or sentence and the duration of mono-lingual segment is very short [2]. We hence consider phone boundaries as the natural candidates for language transition and propose an architecture using phonotactic features extracted from phone recognizer. To robustly analyze very short segment, our feature for back-end classifier contains long term context information across several phones from phonotactic features extracted from accent specific phone recognizers.

Phonotactic feature has been widely used in language recognition system such as phone recognizers followed by language models (PRLM) and PPR-SVM [7-8]. The phonotactic feature, representing the phonetic constraints in a language, can be extracted from an utterance using a phone recognizer. It is well known that more accurate phonotactic feature result in better performance of language recognizer [9]. The study of phone recognizer optimization for language recognition has attracted much attention [9, 10].

The accent of Singapore English (SE) is different from that of American English (AE) [11]. Hence, the CMU (Carnegie Mellon University) English pronunciation dictionary would not be suitable to train the accent specific phone recognizer for the language recognition/diarization task. In order to build a phone recognizer to extract more accurate phonotactic feature on collected speech data we adapted the CMU dictionary to Singapore accent. Our experiment shows that the performances of language recognition and diarization are improved by using the Singapore accent dictionary to train the accent specific phone recognizer.

2. LANGAUGE DIARIZATION SYSTEM FOR CODE-SWITCHED SPEECH

Fig. 1 shows the proposed diarization system. The system first determines the identified phone segments and then classifies each segment using a CRF backend classifier system. The input to the CRF classifier is a phonotactic feature with long term context information.

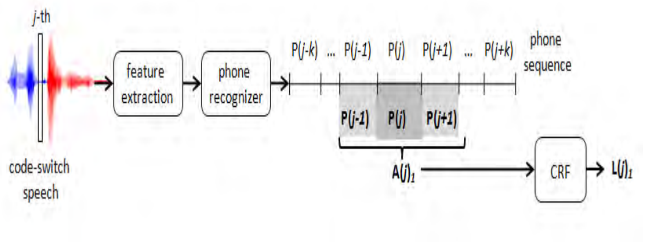

Figure 1. Diagram of proposed language diarization system.

The symbol P(j) denotes the identified phone segment. To capture temporal information, the neighboring recognized phone are combined to form feature vector $A(j)_K$ where the subscript k denotes number of left and right segment information. For example, $A(j)_1$ is formed by concatenating recognized phones of segment (j-1) to (j+1) to form a 3-element vector, i.e. $A(j)_1 = \{ P(j-1), P(j), P(j+1)\}$ To evaluate the effect of segment duration to diarization performance, we examine various length of $A(j)_K$ are sent to a CRF to classify the language identity of segment j. The output of the CRF classifier for $A(j)_K$ is $L(j)_K$.

CRF [12] is known to be a discriminative undirected probabilistic graphical model used for structured prediction of sequential data. In our case, the CRF is used to learn the language transition in the code-switched speech given the long term context recognized phones.

3. PRONUNCIATION DICTIONARY ADAPTATION

The initial pronunciation lexicon of English we used to train the acoustic model of phone recognizer is based on the CMU dictionary. It is well known that Singapore English is different from American English [11]. For example, the pronunciation of the two words "three" and "tree" in American speaking style are /th r iy/ and /t r iy/. However, the Singaporean pronounced both of them as /t r ih/, or sometimes even as /ch r ih/. In order to obtain more accurate phone sequence on Singaporean accent English speech, we adapted the CMU dictionary to Singaporean English pronunciation.

To adapt the dictionary, we examine the use of linguistic knowledge and data-driven approach. In the linguistic knowledge approach, the pronunciation modification rules used to generate the Singaporean pronunciation from the CMU dictionary are based on the two out of the twenty one rules proposed by Chen et al [11]. These 2 rules listed below are chosen as they represent the most common pronunciation changes effected by Singaporean speakers:

- Syllable-final voiceless plosive omitted if preceded by another consonant: /p/, /t/, /k/ could be deleted.
- Word-final /t/,/d/ omitted if preceded by another consonant: /t/, /d/could be deleted.

In the data-driven approach, 800 Singaporean accent utterances were transcribed at the phone level. These transcriptions are used to generate data driven pronunciation adaptation rules. For example, the word "moment" in CMU dictionary is transcribed as /m ow m ah n t/ while the manually transcribed Singaporean speakers' pronunciation of this word is /m ow m ah n/. To discover new pronunciation modification rules, dynamic programming is used align these two pronunciations. In this case, the possible new rule is the deletion of phoneme t in a word final, i.e., /n t #/→/n - #/, where "#" denotes the end of the word or syllable boundary and "-" denotes deletion. By this process, we extracted 302 pronunciation rules. To reduce the number of spurious rules, we removed rules which have less than 3 occurrences to derive the final list of 160 pronunciation rules. Table 1 lists the top ten rules found.

Rank	American English	S'pore English	Rank	American English	S'pore English
1	n_t_#	n_-_#	2	l_iy_#	l_ih_#
3	n_d_#	n_-_#	4	ah_z_#	ah_s_#
5	r_iy_#	r_ih_#	6	er_z_#	er_s_#
7	ao_r_#	ao_-_#	8	aa_r_t	aa_-_t
9	#_t_r	#_ch_r	10	n_z_#	n_s_#

Table. 1 The top ten pronunciation adaptation rules from American English (AE) speaking style to Singaporean English (SE) speaking style.

To verify the performance of the system using the adapted Singaporean accent pronunciation dictionary, we perform experiments on a) Aurora 4 corpus (American speakers) and b) Singaporean speakers reading Aurora 4 text. Each set has a separate training and test set.

The training and test is the Singaporean-read SE corpus which is a close-talk microphone recorded read speech database by Singaporean speakers reading the Aurora4 text.

The corpus consists of 15 male and 24 female university educated speakers, between the ages of 19-25. The training set consists of 31 speakers (11 males and 20 females) with 7137 utterances and 16.8 hours in total. The test set consists of the rest of the 8 speakers (4 males and 4 females) recordings with 330 utterances and 0.7 hours in total. The speech data is recorded at 16KHz, 16 bit format.

The performance using the CMU (AE) AE-Dict and Singaporean English (SE) SE-Dict on SE-Test and AE-Test with SE-train and AE-train is shown in Table 2. As expected, the direct application of acoustic model trained from AE-Train and AE-Dict to decode the SE-Test will result in very bad performance (word error rate: 75.19%). The use of acoustic models trained from SE-Train with AE-Dict achieved a decent 21.42% word error rate. The performance using adapted SE-Dict evaluating on SE-Test outperforms that using AE-Dict and obtains about 21% relative word error reduction. This experiment confirms that using an accented English dictionary is essential for ASR task on accent specific speech.

Training Data	Pronunciation Dictionary	Word error rate (%)	
		Test data	
		AE-Test	SE-Test
AE-Train	AE-Dict	12.16	75.19
SE-Train	AE-Dict	NA	21.43
SE-Train	SE-Dict	NA	16.89

Table 2. The word error rate on the AE and SE corpus using AE-Dict and SE-Dict, where AE and SE mean American English and Singaporean English, respectively

4. CORPUS AND EXPERIMENTS

To examine code-switched speech, we developed a 63-hour conversational South-East-Asia Mandarin/English (SEAME) code-switched corpus collected from Singaporean and Malaysian speakers. The average language intervals in monolingual Mandarin and English segments are about 0.81 second and 0.67 second, respectively. The average number of language changes within a code-switched utterance is about 2.2 [2]. In Table 3, we summarize the data use to train/develop and test our language diarization systems:

To measure our systems' performance, we evaluated both the language diarization and language recognition task. For the language diarization experiments, we evaluate the proposed system without any language boundary information. In the language recognition task, we first extracted all test monolingual segments and perform language recognition using our systems as well as other published state-of-the-art approaches.

The experimental setup is as follows: The training set is used to train the phone recognizer. The development set is used to train the SVM and CRF classifiers. For the phone recognizer, the features are MFCC-based temporal feature extracted from a 110 ms speech segment - this feature is named as LDA42 as it uses 42 coefficients per frame extracted by linear discriminant analysis dimensional reduction from a window of 11 frames which contains 13th-order MFCC features. LDA provides a linear transformation to reduce the dimensionality while preserving the discriminative power of features. The number of class in LDA is 75 which is language dependent and context independent phoneme in Mandarin and English. The acoustic model of the phone recognizer is a HMM-based tri-phone tied-state system. Each HMM contains three states, and each state contains 32 GMM then a free phone loop is used for decoding.

In the language diarization evaluation for code-switched speech, we only detect English and Mandarin segment. For other categories such as silence and others (filler pause, noise, discourse particles and other languages), there are not taken into consideration - this implies that such segments when they occur will be miss-classified. The evaluation result used is frame error rate (FER). As the language recognition decisions are made on phone segments, we convert the current language identity for each phone segment into frame level for evaluation. For the language recognition task, we use the equal error rate (EER) to evaluate the performance.

	Training Set	Dev. Set	Test Set
# of speakers	133	11	13
# of utterances	44,524	3,505	4,116
# of hours	52.38	5.26	5.21

Table 3. The statistics of training, development and testing set of SEAME corpus.

4.1. Language recognition on mono-lingual speech

To compare the proposed methods with state-of-the-art language recognition approaches on mono-lingual segments, we build three systems 1) PPRLM+SVM; 2) PR+CRF and 3) PR(SG-Dict)+CRF where 2) and 3) are new systems proposed in this paper.

PPRLM+SVM: This is a state-of-the-art PPRLM system with SVM to identify language [7, 13]. The front-end classifiers are English and Mandarin phone recognizers. The acoustic models for both recognizers are a HMM tri-phone tied-state system. The back-end stage is two tri-gram mono-phone language models trained from the training set of English or Mandarin mono-phone transcription in the SEAME corpus. The decision stage is a SVM which identifies language identity given the likelihood scores of the mono-lingual speech segment.

PR+CRF: This is the proposed system described in section 2 using the CMU dictionary to train the phone recognizer. In the decision stage after processing CRF, a majority voting system is used to determine the language identity. E.g., if the majority segments are classified as Mandarin, the whole given monolingual speech segment is assigned to Mandarin.

PR(SG-Dict)+CRF: This is similar to the previous system except that the Singaporean English pronunciation dictionary was used to build the phone recognizer as described in section 3.

Systems	Speech duration in seconds			
	0.1-0.5	0.5-1	1-3	3-9
PPRLM+SVM	20.4	16.2	10.7	5.1
PR+CRF	18.1	13.9	8.7	4.1
PR(SG-Dict)+CRF	17.3	13.1	8.1	3.8

Table 4. Language recognition performance (EER) on monolingual speech segment of different duration.

The evaluation results of language recognition in EER are shown in Table 4. We divided the data into four groups according to their durations, e.g. 0.1 to 0.5 sec., 0.5 to 1 sec., 1 to 3 sec. and 3 to 9 sec. The results show that better performance is achieved on longer monolingual segment, which confirms the benefits of using additional temporal information in our proposed systems. Finally, the system using Singaporean English pronunciation dictionary obtained the best performance among these systems.

4.2. Language diarization on code-switched speech

The language diarization performance on code-switched speech is shown in Table 5. The results show that the frame error rate decreases as the length of temporal features increase. This suggests shows that temporal features offer complementary information to discriminate one language from another. Second, the language diarization system using Singaporean English pronunciation dictionary outperforms that using American English pronunciation dictionary. This shows that the phonotactic feature extracted from phone recognizer which uses Singaporean English pronunciation dictionary is more discriminative on Mandarin-English code-switched speech. The best system achieved 14.4% frame error rate with a 5-phone context setting.

	L_0	L_1	L_2	L_3	L_4
PR+CRF	18.6	16.6	15.4	15.9	16.3
PR(SG-Dict)+CRF	16.8	15.2	14.4	14.9	15.1

Table 5. The language diarization performances (FER) of our two proposed system on code-switched speech using the CMU and the Singaporean English dictionary.

5. CONCLUSION

This paper examines phonotactic-based language diarization system on code-switched utterances. The proposed framework using long term context phonotactic feature across several phone-based segments outperforms state-of-the-art language recognition system. Our results also show that phonotactic feature extracted from phone recognizer which uses adapted Singaporean English pronunciation dictionary improves language diarization performance.

6. REFERENCES

[1] Barbara E. Bullock and Almeida J. Toribio, The Cambridge Handbook of Linguistic Code-switch, Cambridge University Press, 2009

[2] Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li, "SEAME: a Mandarin-English Code-switch Speech Corpus in South-East Asia," In Proc. of Interspeech, Japan, 2010

[3] P. Auer, Code-switch in Conversation: Language, Interaction and Identity, London: Routledge, 1998

[4] Joyce Y. C. Chan, P.C. Ching, Tan Lee and Helen M. Meng, "Detection of Language Boundary in Code-switch utterances by Bi-phone Probabilities," In proc. of ISCSLP 2004

[5] Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang and Chun-Nan Hsu, "Speech Recognition on Code-switch Among the Chinese Dialects," In Proc. of ICASSP, 2006

[6] Anguera X, Bozonnet Simon, Evans Nicholas W D, Fredouille Corinne, Friedland O, Vinyals O, "Speaker Diarization A Review of Recent Research," in IEEE Transactions on Audio, Speech and Language Processing, Vol 20, No. 2, 2012

[7] M.A. Zissman, "Comparison of four approaches to automatic LID of telephone speech," IEEE Trans. on Acoustic., Speech, Signal Processing, Vol. 4, No. 1, pp. 31-44, 1996

[8] Haizhou Li, Bin Ma, and Chin-Hui Lee, "A Vector Space Modeling Approach to Spoken Language Identification", in IEEE Transactions on Audio, Speech and Language Processing, Vol 15, No. 1, 2007

[9]C.P. Santhosh Kumar, Haizhou Li, Rong Tong, Pavel Matˇejka, Luk'aˇs Burget, Jan ˇCernock'y, "Tuning Phone Decoders For Language Identification", In Proc. of ICASSP, 2010

[10] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell and D. A. Reynolds, "Acoustic, phonetic and discriminative approaches to automatic language recognition," in Proc. Eurospeech, 2003.

[11] Chen Wenda, Tan Ying Ying, Chng Eng Siong, Li Haizhou, "The development of a Singapore English CALL Resource", In Proceedings of O-COCOSDA, Kathmandu, Nepal, Nov. 2010

[12] Lafferty, J., McCallum, A., Pereira, F. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In Proc. on Machine Learning, 2001

[13] Haizhou Li, Bin Ma, and Kong Aik Lee, "Spoken Language Recognition: From Fundamentals to Practice", Accepted in Proceedings of the IEEE