LANGUAGE DIARIZATION FOR CODE-SWITCH CONVERSATIONAL SPEECH

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

This paper examines the process of language diarization, the process to perform language segmentation and recognition, in a code-switched speech. Towards this task, we have developed a 63 hours conversational code-switch corpus recorded from Singapore/Malaysia speakers. We show that code-switching can occur frequently and the average language interval may be as short as one second. As such, language diarization is challenging task. To process such short segments, we propose a language diarization system using long term context feature across several phone-based segments and the combination of acoustics and phonotactic information. We achieved a frame error rate of 14.7% for language diarization performance on a Mandarin-English code-switch corpus. To evaluate our system, we measured the language recognition performance on monolingual segments extracted from the code-switch corpus against published techniques of LID systems - we obtained a relative equal error rate reduction of 5.2%, 13.8%, 15.1% and 17.9% on speech durations of 0.1 to 0.5 sec., 0.5 to 1 sec., 1 to 3 sec. and 3 to 9 sec respectively.

Index Terms— language diarization, language recognition, code-switch, conversational speech

1. INTRODUCTION

Code-switch refers to the switching of languages in speech, and is a common occurrence among multilingual speakers [1]. For example, in United States and Switzerland, studies show that a mixture of Spanish and English or French and Italian are commonly spoken [2]. In Hong Kong, Cantonese-English code-switch speech is also very common in colloquial Cantonese [3-4]. In Taiwan, Mandarin-Taiwanese code-switch speech is also widespread [5-6].

Fig. 1 shows an excerpt of code-switch utterance from a Mandarin-English database recorded in Singapore [7]. The transcription of the speech is 通常是他們來了 confirm 了那個 data everything 我們才會知道. There are five language turns in this utterance: Mandarin, English, Mandarin, English and Mandarin. Each of the monolingual segments has a short duration of 1.2, 0.5, 0.7, 0.9 and 0.9 seconds, respectively.

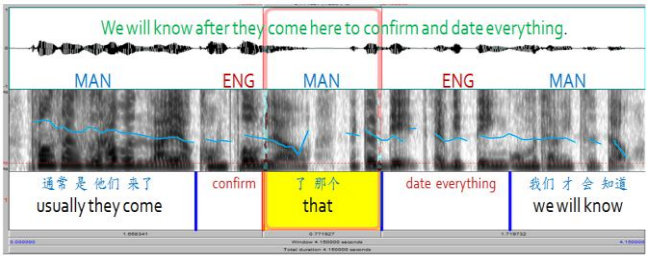

Figure 1. An example of Mandarin-English code-switch utterance.

In this paper, we propose a system to automatically segment and identify languages in a code-switch utterance. As this task is similar to the speaker diarization task [8] 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 [9] where the language boundary is known while the language identity is unknown. However, in language diarization, both language identity and boundary of the given speech are unknown. In addition, the mono-lingual segments of code-switch speech are much shorter than those traditionally studied in the language recognition research [11-12].

Language recognition (LID) methods can be categorized into two groups: acoustic versus phonotactic approaches [11-13]. By acoustic approach, we mean that the features used are derived from the spectral characteristics of the speech utterance. For phonotactic approach, the feature is the phone sequence statistics.

In this paper, we propose an architecture which combines acoustic and phonotactic features for the language diarization task. To robustly process very short mono-lingual segment in code-switch speech, our feature contains long

term context information (across several phones) from both acoustics and phonotactic features. From our previous analysis on the SEAME corpus, we found that code-switch often occurs between word, phrase or sentence [14]. To exploit this, a code-switch LVCSR system would be required. As this is often not available, we will instead first identify phones and use the identified phone boundaries as candidates for language transitions.

An immediate application of language diarizaton result is to improve LVCSR performance when the input contains code-switch speech. The identified language segments likelihood score can be cleverly integrated to multilingual speech recognition system; as oppose to making hard decisions on individual segments, and then use mono-lingual LVCSR engine to process them.

2. LANGAUGE DIARIZATION SYSTEM FOR CODE-SWITCH SPEECH

This section describes our proposed language diarization system, which is a combination of an acoustic language diarization system and a phonotactic-based system. Our system is targeted only for Mandarin-English code-switch, i.e, only these two language classes will be identified; languages outside these two will be mis-classified. In section 2.1, we first describe the acoustic system, section 2.2 describes the phonotactic system, and finally section 2.3 introduces our proposed combined system.

2.1. Acoustic language diarization system

Fig. 2 shows the acoustic language diarization system. The system first determines the identified phone segments and then classifies each segment using a backend classifier system. The inputs to the classifier are likelihood scores generated by two-class GMMs processing acoustic-based features.

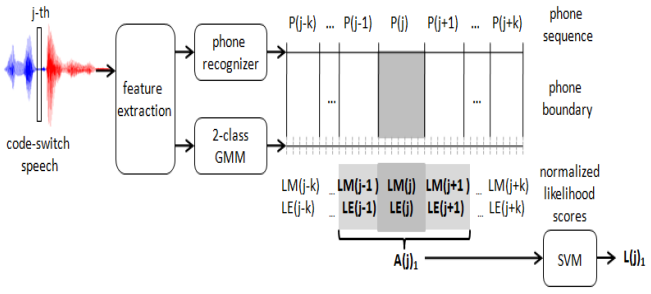

Figure 2. Diagram of acoustic language diarization system.

For the identified phone segment j, two normalized log likelihood scores, LM(j) and LE(j), one for each target class (Mandarin/English), is computed. To capture temporal information, the likelihood scores from k number of left and right neighboring segments are combined to form feature vector $A(j)_K$. For example, $A(j)_1$ is formed by concatenating the average log likelihood score of segment (j-1) to (j+1) to form a 3-element vector. To evaluate the effect of segment duration to diarization performance, we examine various length of temporal information: $A(j)_0$, $A(j)_1$, $A(j)_2$, $A(j)_3$ and $A(j)_4$. The output of the SVM classifier for $A(j)_K$ is denoted as $L(j)_K$.

2.2. Phonotactic language diarization system

Phonotactic information has been widely used for the LID task. Example include: PRLM (Phoneme Recognizer followed by Language Model), P-PRLM (Parallel PRLM), and PPR-VSM system which uses phoneme recognizer to transcribe input speech into phoneme strings and then use a statistic language model to estimate the likelihood [8,10].

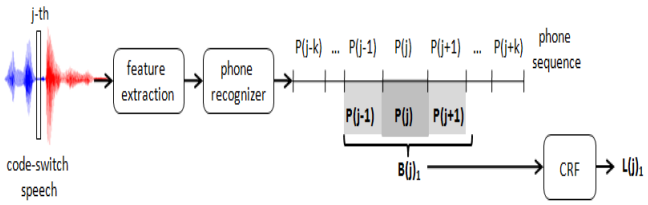

Figure 3. Framework of phonotactic language diarization system.

Our proposed phonotactic language diarization framework is shown in Figure 3. As in our acoustic language diarization system, the system exploits the fact that phoneme boundaries are candidates of language turns. In the phonotactic framework, the phoneme identities across several phoneme segments are combine as features to robustly process identify very short mono-lingual segment. For example, to evaluate the language identity of the current phone with index j with K left and right context, the feature is a 2k+1 feature vector $B(j)_K$ contains as its element [P(j-k), …, P(j-1), P(j), P(j+1), ..., P(j+k)]. The corresponding output of the back-end conditional random fields (CRF) classifier is $L(j)_K$.

CRF is introduced in 2001 by Laffertyetal [15], and is a novel 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-switch speech given the recognized phone sequence with the corresponding language identity.

2.3. Fusion language diarization system

Acoustic and phontactic features have been shown to provide complementary effect for language recognition [12]. In this section, a fusion language diarization system combines acoustic and phonotactic features. The difference to the phonotactic system is the inclusion of the acoustic

likelihood scores of each segment to the CRF classifiers as illustrated in Figure 4.

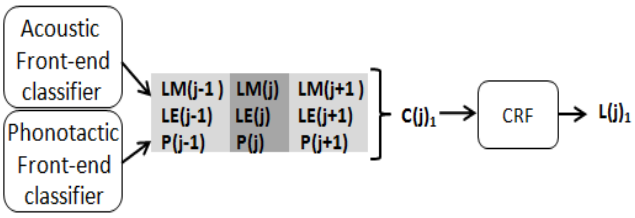

Figure 4. The fusion language diarization system.

3. CORPUS AND EXPERIMENTS

To examine code-switch speech, we developed a 63-hour conversational South-East-Asia Mandarin/English (SEAME) code-switch corpus collected from Singapore and Malaysian speakers. We extract code-switch utterances if the utterance contains both Mandarin and English segments and the utterance is self-contained semantically or separated by an obvious pause. The average utterance length is around 4 seconds. 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-switch utterance is about 2.2 [14]. In Table 1, we summarize the data use to train/develop and test our language diarization systems:

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

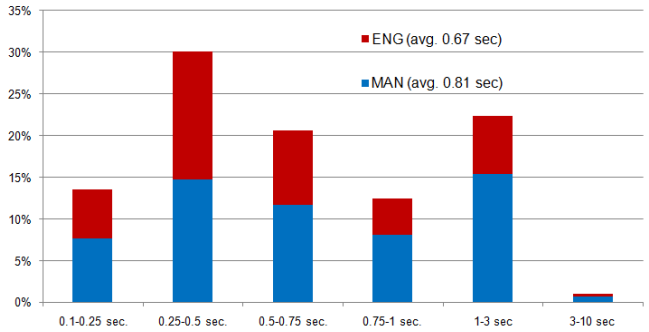

Figure 5. The distribution of average monolingual English and Mandarin segments of the SEAME corpus.

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

The experimental setup is described in the followings. The training set is used to train the two-class GMM and phone recognizer. The development set is used to train the SVM and CRF classifiers.

For the phone recognizer, a standard setup using free phone loop and MFCC-based temporal feature is used. The phone recognizer is a HMM-based tri-phone tied-state system. Each HMM contains three states, and each state contains 32 GMM. A phone bigram language model is used for speech decoding.

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. Each frame is processed as 13th-order MFCC features using a frame size of 25 ms, with a step size of 10 ms. 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 LDA42 feature is used to train both GMM-based classifier and phone recognizer. The two-class GMM classifier consists of two individual GMMs, one for each language, Mandarin and English. Each GMM generates a log likelihood score for each frame and the number of Gaussians in each GMM is 4096.

In the language diarization evaluation for code-switch 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 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. In addition, the equal error rate (EER) is used for language recognition evaluation on mono-lingual speech segment.

3.1. 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 the following systems (S1-S3). The last three systems (P1-P3) are the proposed frameworks described in section 2.1, 2.2 and 2.3, respectively. The main differences between our proposed acoustic (P1) and phonotactic system (P2) as compared to traditional main stream systems (S1 and S2) is in the exploitation of identified segments as candidates and the use of long temporal information across segments.

S1) GMM4096+SVM: The language recognition classifier is a two-class, Mandarin and English, GMMs with 4096 Gaussians in each GMM. The decision stage is a SVM which identifies language identity given the likelihood scores of the mono-lingual speech segment [12].

S2) P-PRLM(ENG+MAN): This is P-PRLM system where 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 [10].

S3) GMM+PPRLM+SVM: This is a fusion system which integrates S1 and S2 systems with a SVM decision system [12].

P1) GMM+SVM: This is the acoustic system described in section 2.1. In the decision stage, a voting system is used to determine the language identity. For example, as majority segments which their language identities belong to Mandarin, the whole given monolingual speech segment is assigned to Mandarin.

P2) PR+CRF: This is phonotactic system described in section 2.2. A decision stage is the same with that in P1 system.

P3) GMM+PR+CRF: This is the combined system described in section 2.3. A decision stage is the same with that in P1 system.

Systems	Speech duration in sec.			
	0.1-0.5	0.5-1	1-3	3-9
S1) GMM4096+SVM	24.6	20.2	15.2	6.8
S2) P-PRLM(MAN+ENG)	20.4	16.2	10.7	5.1
S3) GMM+PPRLM+SVM	17.3	11.6	7.3	3.9
P1) GMM+SVM	22.8	18.1	11.1	5.4
P2) PR+CRF	18.1	13.9	8.71	4.1
P3) GMM+PR+CRF	16.4	10.0	6.2	3.2

Table 2. Language recognition performance (EER) on monolingual speech segment.

The evaluation results of language recognition in EER are shown in Table 2. 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 the better performance is achieved as longer monolingual segment is used. By combining P-PRLM and GMM framework, GMM+PPRLM+SVM (S3), performs better than GMM and PPRLM individual systems (S1 and S2).

Although both S1 and P1use the GMM front-end classifier, our proposed system contains temporal information and hence resulted in better performance. In addition, our proposed system, PR+CRF (P2), also outperforms the P-PRLM system (S2). Finally, for comparison of two fusion systems (S3 and P3), the proposed system, GMM+PR+CRF (P3) obtains relative equal error rate reduction of 5.2%, 13.8%, 15.1% and 17.9% on speech durations of 0.1 to 0.5 sec., 0.5 to 1 sec., 1 to 3 sec. and 3 to 9 sec., respectively.

3.2. Language diarization on code-switch speech

The language diarization performance on code-switch speech is shown in Table 3. To validate the effect of the length of temporal features to language diarization performance, we set variant length of context phone from zero to four (L_0, L_1 L_2, L_3 and L_4) in our experiment. For example, L_2 means that we use right and left neighboring phones and GMM likelihood normalized scores then combine to form the temporal feature for CRF classifier. The results show that the frame error rate decreases as the length of temporal features increase. This suggests shows that the temporal features offer sufficient language information for language diarization system to discriminate one language from another. Second, the performance of the phonotactic language diarization system outperforms acoustic-based language diarization system. This shows that phonotactic information is a crucial feature for detecting language changes on code-switch speech. Finally, the fusion language diarization system which combines acoustic and phonotactic features performs the best and it achieves 14.7% frame error rate as using 5-phone segments temporal information.

	L_0	L_1	L_2	L_3	L_4
P1) GMM+SVM	26.2	17.4	16.7	16.5	16.4
P2) PR+CRF	18.6	16.6	15.4	15.9	16.3
P3) GMM+PR+CRF	17.8	15.9	14.7	15.6	16.1

Table 3. The performance (FER) of the proposed language diarization framework on code-switch speech.

4. CONCLUSION

This paper introduces the language diarization task on code-switch utterances. Our proposed language diarization system uses combined acoustic and phonotactic cues with variant length of temporal information. The results show that the system performs well as using longer length of temporal information. The best language diarization performance is 14.7% FER with the length of five phone long temporal cues. In addition, the performance of language recognition on very short monolingual speech also outperforms the state-of-the-art LID system.

5. REFERENCES

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

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

[3] David C.S. Li, Cantonese-English code-switch research in Hong Kong: a Y2K review, World Englishes, 19-3, 2000

[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] C.-M. Chen, "Two types of code-switch in Taiwan," paper presented in Sociolinguistics Symposium15, Newcastle, 2004

[6] 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

[7] 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

[8] 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

[9] H. Li, B. Ma, and K. A. Lee, "Spoken Language Recognition: from Fundamentals to Practice", to appear in Proceedings of the IEEE

[10] 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

[11] 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

[12] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language Identification: A Tutorial", in IEEE Circuits and Systems Magazine, Volume: 11, Issue: 2, pp.82-108, 2011

[13] Muthusamy, Y.K., Bernard, E., Cole, R.A, "Automatic LID: A Review/Tutorial," IEEE Signal Processing Magazine, Vol.11, No.4, pp.33-41, 1994

[14] Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li, "An Analysis of a Mandarin-English Code-switching Speech Corpus: SEAME," In Proc. of O-COCOSDA, Nepal. 2010

[15] Lafferty, J., McCallum, A., Pereira, F. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In Proc. of 18th International Conf. on Machine Learning. pp. 282–289, 2001