# INTEGRATION OF LANGUAGE IDENTIFICATION INTO A RECOGNITION SYSTEM FOR SPOKEN CONVERSATIONS CONTAINING CODE-SWITCHES

Jochen Weiner[1], Ngoc Thang Vu[1]
*Dominic Telaar[1], Florian Metze[3], Tanja Schultz[1,3], Dau-Cheng Lyu[2], Eng-Siong Chng[2], Haizhou Li[2]*

[1]Cognitive Systems Lab, Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT)
[2]School of Computer Engineering, Nanyang Technological University (NTU), Singapore
[3]Language Technologies Institute, Carnegie Mellon University (CMU), Pittsburgh, PA, USA
jochen.weiner@student.kit.edu, thang.vu@kit.edu

## ABSTRACT

This paper describes the integration of language identification (LID) into a multilingual automatic speech recognition (ASR) system for spoken conversations containing code-switches between Mandarin and English. We apply a multistream approach to combine at frame level the acoustic model score and the language information, where the latter is provided by an LID component. Furthermore, we advance this multistream approach by a new method called "Language Lookahead", in which the language information of subsequent frames is used to improve accuracy. Both methods are evaluated using a set of controlled LID results with varying frame accuracies. Our results show that both approaches improve the ASR performance by at least 4% relative if the LID achieves a minimum frame accuracy of 85%.

***Index Terms***— code-switching, multi-stream combination, language lookahead

## 1. INTRODUCTION

Code-switching speech is defined as speech which contains more than one language within an utterance. It is a very common phenomenon in multilingual communities [1]. Code-switching is a phenomenon of spoken communication and usually not written. Therefore, it is hard to find large text data for reliable language modeling, which puts code-switching speech into the same category as under-resourced languages. Furthermore, code-switching affects co-articulation and context dependent acoustic modeling. As a consequence, it is very challenging to build accurate automatic speech recognition (ASR) systems for code-switch speech. Despite the clear need for ASR systems which can handle code-switches, there is so far only little evidence of research and development in this interesting field.

In this paper we describe the integration of language information into our multilingual system that recognizes code-switch speech. The language information is provided by a language identification (LID) component. In general, there exist two different strategies to handle code-switch speech recognition. The first strategy is to apply two monolingual ASR systems and a dedicated LID component which takes the input code-switch speech, decides for the language spoken in each speech segment and passes the segment on to the respective monolingual ASR system [2, 3]. This strategy has two major drawbacks, first it heavily depends on the accuracy of the LID module since errors cannot be recovered. Second, it assumes that code-switch speech segments are independent from each other and can be easily separated. Both assumptions are not correct. Due to these shortfalls, we prefer the second strategy, i.e. the use of a multilingual ASR system, as presented in section 3. A multilingual ASR system consists of a multilingual acoustic model, a pronunciation dictionary which combines the word entries of both languages, and a multilingual language model which allows switching between languages. Such a multilingual system does not require any explicit language identification since language information is implicitely integrated. However, the accurate time information, i.e. at which point in an utterance the speaker switches from one language to another could be a quite valuable additional knowledge source during decoding. Therefore, we investigate two different approaches to integrate language information into decoding, i.e. the "Multistream" approach and a new method called "Language Lookahead". Furthermore, we study the impact of LID performance on the error rate of the code-switch ASR system when applying these two approaches.

The paper is organized as follows: Section 2 describes the SEAME data corpus which we used for our experiments. In section 3 we introduce the baseline ASR system and its performance on the development set. Section 4 describes our approaches to integrate the language information into the decoding process. Section 5 presents the experiments and results when the language identification performance is controlled. A summary in Section 6 concludes the paper.

## 2. SEAME DATA CORPUS

SEAME is a conversational Mandarin-English code-switching speech corpus recorded from Singaporean and Malaysian speakers [4]. The corpus is designed for multiple research purposes which include language boundary detection, language identification studies and multilingual LVCSR. For these purposes, manual transcriptions at word-level including language boundary alignment are provided. To take regional language variations into account, data was collected in two countries: Singapore and Malaysia. As the corpus was developed for research on spontaneous code-switching speech, the recordings consist of interviews and conversations. Compared to our earlier paper [4], the corpus was extended to about 63 hours of audio data. Considering the particular speaking styles in Singapore and Malaysia, we classify transcribed words into four categories for language identification: English words, Mandarin words, Silence, and Others (discourse particle, other languages, and hesitations). The ratio of Mandarin words is 44%, English words 26%, Silence 21% and Others 7%. The average number of code-switches within each utterance is 2.6 when counting only switches between Mandarin and English. In total, the corpus contains 9,210 unique English and 7,471 unique Mandarin words. The duration of monolingual segments is very short: More than 82% English and 73% Mandarin segments are less than 1 second long while the average duration of English and Mandarin segments is only 0.67 seconds and 0.81 seconds, respectively. Further details and analyses on the 25-hrs corpus can be found in [4]. The corpus is divided into three sub-sets (training, development and test set) and is balanced for criteria like gender, speaking style, ratio of Singaporean and Malaysian speakers, ratio of the four categories, and recording lenght. Table 1 summarizes the statistics of the SEAME corpus for the three sub-sets.

**Table 1**: *Statistics of the SEAME corpus*

| | Train set | Dev set | Eval set |
|---|---|---|---|
| # Speakers | 139 | 8 | 10 |
| Duration(hours) | 58.4 | 2.1 | 2.3 |
| # Utterances | 48,040 | 1,943 | 2,162 |

To measure the performance of the ASR systems we adopted the Mixed Error Rate (MER) which applies word error rates for English and character error rates for Mandarin. The presented MER is the weighted average over all English and Mandarin portions of the speech recognition output. By applying character based error rates for Mandarin, the performance does not dependent on the word segmentation algorithm applied to Mandarin and thus allows for comparison across different segmentations for future investigations.

## 3. BASELINE CODE-SWITCH SYSTEM

Based on the SEAME corpus we developed an initial baseline speech recognition system. For the dictionary, the CMU English [5] and the Mandarin pronunciation dictionary [6] are merged into one bilingual pronunciation dictionary, with a lexical size of 135K entries for English and 130K for Mandarin. We applied some rules in [7] for phone deletion and substitution to generate pronunciation variants for Singaporean English. For language modeling we used the SRI Language Modeling Toolkit [8] to build trigram language models (LMs) from the SEAME training transcriptions consisting of the complete 16K-vocabulary of the transcriptions. These LMs were interpolated with two monolingual language models. Both monolingual language models were created from 350K English sentences from NIST (*EN-mono*) and 400k Mandarin sentences from the GALE project (*CH-mono*) which had been collected from online newspapers. The 30K-vocabulary contains all words of the transcriptions and the most frequent words of both monolingual corpora. Furthermore, *CH-mono* and *EN-mono* serve as basis for the generation of artificial code-switch texts. For this purpose, we analyze the characteristics of code-switching from the SEAME training transcriptions and extract monolingual English segments from the SEAME training transcriptions, translate them to Mandarin and search the translations in a large monolingual Mandarin text. When translations are found, we create an artificial code-switch text by replacing the (Mandarin) translation with its English counterpart [9]. The resulting best language model has a perplexity of 483.9 and an out-of-vocabulary (OOV) rate of 1.21% on the SEAME development set.

For signal preprocessing we apply a Hamming window of 16ms length with a window overlap of 10ms. A 143 dimensional feature vector is extracted by stacking 11 adjacent frames with 13 MFCC coefficients each. An LDA transformation reduced this vector to 42 dimensions. The acoustic model uses a fully-continuous 3-state left-to-right HMM. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. The phone set contains English, Mandarin, and bilingual phones which share the same International Phonetic Alphabet [10] (IPA) symbols. Special models for +noise+, +breath+ and +laugh+ are added to cope with continuous speech. Since particles are heavily used in the Singaporian and Malaysian languages, a phone +particle+ is also added. For context dependent acoustic modeling, we apply a polyphone decision tree splitting process which is stopped at 3,500 quintphones. We then apply merge&split training with a maximum of 64 mixtures per state and a global Semi-Tied Covariance (STC) matrix [11] to all the acoustic models followed by three iterations of Viterbi training.

Our baseline speech decoder is a 2-pass system which consists of two different acoustic models. The first acoustic model AM1 is speaker-independent. The second AM2 is trained by applying Speaker Adaptive Training (SAT) with Feature Space Adaptation (FSA). In addition, we perform boosted Maximum Mutual Information Estimation (bMMIE) [12] to improve performance. The best system has a mixed error rate (MER) of 36.9% on the SEAME development set.

## 4. INTEGRATE LANGUAGE IDENTIFICATION INTO ASR

We describe two different approaches to integrate language information into the decoding process of the multilingual baseline system: The multistream approach was proposed in [13] and can be used to integrate an additional LID stream into decoding. We advance this approach to further improve the performance of the integrated system and call this new method "*Language Lookahead*".

### 4.1. Multistream Approach

The multistream approach, a "relatively simple approach to combining several information sources in ASR" [13], is a straight-forward way to combine the language (LID score) and the acoustic model (AM score) information. We use it to integrate the frame-level LID into the decoding process based on the architecture described in [13]. This approach operates on the acoustic level: In addition to the stream of scores from the acoustic model and the phonetic decision tree, it introduces an additional stream of scores and an additional decision tree for the language information. The additional decision tree decides between English and Mandarin, and depending on its decision, selects the appropriate LID score of the current frame from the stream. After that a linear combination $(1 - w) \cdot score_{am} + w \cdot score_{lid}$ of the AM score and LID score is created, where $w$ is a weight factor for the LID stream. Based on the weighting done in [14] we experimented with factors around 0.05 and chose 0.1 as weighting factor for all experiments. This combined score replaces the AM score in the subsequent decoding steps (Figure 1).

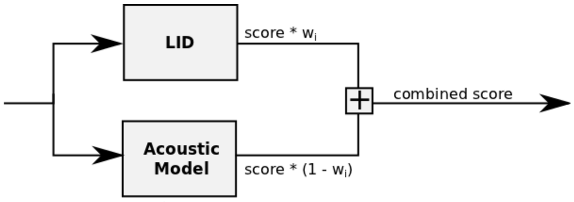

**Fig. 1**: The Multistream Approach

### 4.2. Language Lookahead Approach

The "Language Lookahead" is an extended version of the multistream approach which also integrates LID at the frame-level (Figure 2). While the multistream approach integrates only one single stream for the LID of the current frame, the lookahead approach integrates two additional streams of LID scores, one corresponding to the current frame and one for the LID of the $n$ subsequent frames. Both streams use the same decision tree, and their scores are determined the same way as described above. The scores are linearly combined $(1 - w_i - w_j) \cdot score_{am} + w_i \cdot score_{lid} + w_j \cdot score_{lookahead}$ and the final combined score is used for decoding. The aim of integrating language information of future frames is to support the decoder to discovering falsely classified frames early, making corrections into the right direction, and thereby increasing the robustness of the recognition.

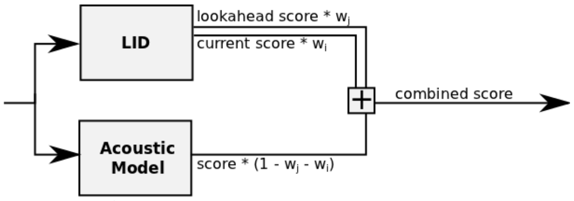

**Fig. 2**: The Language Lookahead Approach

## 5. EXPERIMENT AND RESULTS

The aim of our experiments is to investigate the usefulness of integrating language information into the ASR decoding process. Also, we want to determine which of the above described approaches performs better. However, intuitively we expect that the performance of the LID component impacts its usefulness. For this reason we perform the experiments in two steps. First, we modify the LID component output such that we can control the LID accuracies. Second, we compare ASR error rates of the multistream with the language lookahead approach based on these controlled LID frame accuracies.

### 5.1. Controlling LID Performance

The simplest experiment would be to take the actual LID frame accuracies and evaluate the ASR performance. This could then be compared to an oracle experiment where the language information is assumed to be always correct. However, the actual LID frame accuracy achieves 70.64% which leaves ample space for improvement. Furthermore, a post-processing of the LID output is rather straight-forward, for example by smoothing over subsequent frames to avoid rapid language switches. Therefore, we decided to control the LID output such that the frame accuracies cover the whole range between the 70% baseline and the 100% oracle performance.

Starting with the actual LID output, we modified the results step-by-step to generate artificial LID results with different frame accuracy. The LID system outputs the language identity along with a confidence score on a frame-by-frame basis. Using the same features as in ASR, an HMM-based voice activity detector separates speech and non-speech segments in each utterance. The speech segments are then evaluated by two GMM acoustic-based LID classifiers to produce two log likelihood scores for each speech frame. A post-processing step on each $i$-th frame eliminates rapid language changes by averaging the log likelihood scores generated from the Mandarin GMM and English GMM from the $(i - w)$-th frame to the $(i + w)$-th frame, where $w$ describes the window length. We used Hamming windows to emphasize the weight of the current frame over the log likelihood scores. The frame error rate for voice activity detection and language identification on the development set is 5.88% and 70.64%, respectively.

The controlled LID accuraries are then performed by the following procedure: (1) Pick incorrect frames randomly, (2) correct the frames until a certain frame accuracy is achieved. We generated controlled LID results with frame accuracies 75%, 80%, 85%, 90%, 95%, and 100%.

### 5.2. Impact of LID Performance on ASR

In the following experiments we investigate the impact of the varied LID accuracies on the ASR code-switch performance and we compare the above described two approaches for integrating language information into the decoder. The results evaluated on the SEAME development set are depicted in Figure 3, plotting the MER over the frame accuracies of the LID component. If no language information is provided, MER corresponds to the baseline result of 36.9%. The blue line shows the performance of the Multistream approach, the red line shows the performance of the Language Lookahead approach. The results show that MER improves when language information is added. Also, as expected, MER varies with the LID frame accuracy. As the frame accuracy improves, the MER improves almost linearly. For the oracle experiment the code-switch recognizer achieves 34.2% MER with Language Lookahead approach and 34.4% with Multistream approach.

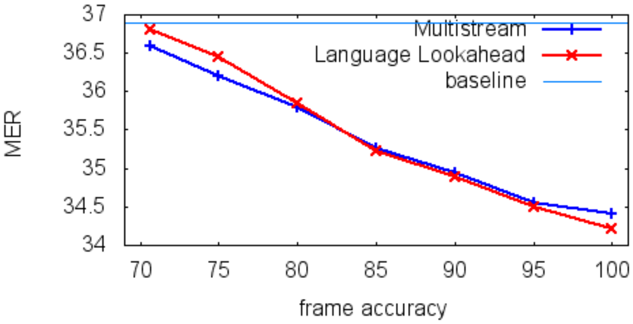

**Fig. 3**: MER of applying Multistream (blue) and Language Lookahead (red) approaches on the SEAME development set

The weights for the LID scores in both approaches are experimentally optimized. The weight for the LID stream is set 0.1, the weight for the Language Lookahead score is set to 0.01. Within the range of 70% to 95% LID frame accuracies, we do not see significant differences between both approaches in terms of MER performance. Both approaches achieve around 4% relative improvements when the LID frame accuracies are higher than 85%.

## 6. CONCLUSIONS

The paper describes the integration of LID into a multilingual ASR system for code-switching speech. We introduced and compared two approaches, the Multitream and the "Language Lookahead" method. Furthermore, we investigated the impact of LID performance on the ASR performance by creating a set of controlled LID accuracies. Both approaches give improvements over the baseline results. Also, in both approaches the ASR performance varies with the LID frame accuracy. Higher LID accuracies result in better ASR performance on code-switch speech. We achieve at least 4% relative improvement when the LID has a minimum frame accuracy of 85%, which is feasible to achieve as indicated by [15, 16].

## 7. REFERENCES

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

[2] K. Bhuvanagiri and S. Kopparapu, " An Approach to Mixed Language Automatic Speech Recognition ", Oriental COCOSDA, Kathmandu, Nepal, 2010.

[3] D.C. Lyu, R.Y. Lyu, Y. Chiang and C.N. Hsu, "Speech Recognition on Code-Switching Among the Chinese Dialects," In Proceedings of ICASSP, Toulouse, France, May. 2006.

[4] D. Lyu, T. Tan, E. Chng and H. Li, "An Analysis of a Mandarin-English Code-switching Speech Corpus: SEAME ", Interspeech, Makuhari, Japan, 2010.

[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[6] R. Hsiao, M. Fuhs, Y. Tam, Q. Jin and T. Schultz, "The CMU-InterACT 2008 Mandarin Transcription System ", Interspeech, Brisbane, Australia, 2008.

[7] W. Chen , Y. Tan , E. Chng , H. Li. "The development of a Singapore English call resource", Oriental COCOSDA, Nepal, 2010.

[8] A. Stolcke, SRILM an Extensible Language Modeling Toolkit, ISCSLP, 2002.

[9] N.T. Vu, D.C. Lyu, et al. "A First Speech Recognition System for Mandarin-English Code-switch Conversational Speech", in Proc. ICASSP, Japan, 2012.

[10] Handbook, IPA: Handbook of the International Phonetic Association, 1999.

[11] M. Gales, "Semi-tied covariance matrices for hidden Markov models", IEEE Transactions Speech and Audio Processing, vol. 7, pp. 272-281, 1999.

[12] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for Model and Featurespace Discriminative Training" in Proc. ICASSP, USA, 2008.

[13] F. Metze, "Discriminative speaker adaptation using Articulatory Features", Speech Communication 49(5), 2007.

[14] F. Metze, A. Waibel, "A flexible stream architecture for ASR using articulatory features", in Proc. of the International Conference on Spoken Language Processing, Denver, 2002.

[15] D.C. Lyu, R.Y. Lyu, Y.C. Chiang, C.N. Hsu, "Speech Recognition on Code-Switching among the Chinese Dialects", in Proc. ICASSP, France, 2006.

[16] Joyce Y C Chan, P C Ching, Tan Lee, Houwei Cao, "Automatic recognition of Cantonese-English code-mixing speech", in Computational Linguistics and Chinese Language Processing, vol. 14, pp. 281-304, 2009.