# DETECTION OF LANGUAGE BOUNDARY IN CODE-SWITCHING UTTERANCES BY BI-PHONE PROBABILITIES

*Joyce Y. C. CHAN*, P. C. CHING*, Tan LEE*, Helen M. MENG***

*Department of Electronic Engineering
**Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong
{ycchan, pcching, tanlee}@ee.cuhk.edu.hk, hmmeng@se.cuhk.edu.hk

## ABSTRACT

In this paper, we present an effective method to detect the language boundary (LB) in code-switching utterances. The utterances are mainly produced in Cantonese, a commonly used Chinese dialect, whilst occasionally English words are inserted between Cantonese words. Bi-phone probabilities are calculated to measure the confidence that the recognized phones are in Cantonese. Two sets of context-independent mono-phone models are trained by monolingual Cantonese and monolingual English data separately. Both knowledge-based and data-driven model selection approaches are studied in order to retain the language-dependent characteristics and to merge duplicated phone sets between the two languages. The LB detection accuracy is 75.12% for utterances that contain one single code-switching word or phrase.

## 1. INTRODUCTION

Code-switching involves the use of words from two different languages within a single discourse or within a single utterance. It is frequently used, in particular, in bilingual communities [1]. In Hong Kong, code-switching between Cantonese and English is used widely in daily conversation [2].

Automatic speech recognition (ASR) in Cantonese and English has already been studied by many researchers [3], [4], but code-switching between these two languages is seldom applied in ASR [5]. The main difference between ASR of code-switching utterances and monolingual or multilingual utterances is that there are multiple languages within a single utterance. If we know when the language is switched, higher word-based recognition accuracies can be obtained since the search space is greatly reduced. We can recognize the code-switching utterances by two monolingual speech recognizer and switch between them at the LB. Another way is to recognize the code-switching utterances using a speech recognizer with bilingual dictionary, then apply the LB information to obtain the best match hypothesis from the N-best recognition result. The term "code-switched word" here refers to a single word or a phrase.

In this paper, detection of LB between Cantonese and English in code-switching utterances will be investigated. Cantonese is in majority (the matrix language) and English is the embedded language. Monolingual Cantonese corpus CUSENT [6] and English corpus TIMIT [7] will be used for phone model training since large amount of code-switching training data is not available. Differences between the two languages will be studied and bi-phone likelihood of Cantonese will be applied as confidence measurement for phone based language identification. Recognized phone sequences with low Cantonese bi-phone likelihood are likely to be English.

## 2. LANGUAGE-DEPENDENT CHARACTERISTICS

For the acoustic models, we define a universal phone set for the two languages. Some phones across the two languages may be similar enough to be equated, thus we do not need the whole set of English phone models for recognizing the code-switching utterances. Since Cantonese is in dominant within the code-switching utterances, the Cantonese phones will have a higher priority than the English phone models that are similar to them. Hence model selection is applied on the English phone models to obtain the optimum universal phone set.

Phonotatic knowledge will be explored by calculating the bi-phone probability of Cantonese. This bi-phone probability will be regarded as confidence measurement that the detected phones are Cantonese.

### 2.1 Acoustic Models (Model Selection)

Both knowledge-based and data-driven method has been explored to cluster the English phones and Cantonese phones together. Phones from the two languages that are similar to each other will share the same phone model.

#### *2.1.1 Knowledge-based model sharing*

The English phones are divided into three categories according to their similarity to Cantonese phones.

Category 1 (Low distance): The phones from Cantonese and English that share the same IPA symbols. [8]

Category 2 (Middle distance): The phones from the two languages that do not share the same IPA symbol but have some degree of similarity due to the Cantonese accent. When the speaker is code-switching, the English words usually carry heavy Cantonese accent, hence some English phones may be pronounced as Cantonese phones.

Category 3 (High distance): The phones that are unique to English and they are not easily confused with the Cantonese phones, such as the retroflex (r, er, axr). The phones are

categorized according to the confusion between English phones and Cantonese phones. The confusion is measured by recognizing the CUSENT testing data with the English phone models. Since the speech data in CUSENT is monolingual Cantonese that contains no English at all, if an English phone appears in the hypothesis phone sequence, it must be an error. English phones that seldom appear in the hypothesis are regarded as high distance. The three categories of the English phone models are listed in Table 1.

*Table 1: Categories of the English phone models*

| Cat. | Models | no. |
|---|---|---|
| 1 | l, m, n, ng, w, f, s, p, t, k, eh, ih, ix, iy, uw | 15 |
| 2 | ch, sh, th, dh, zh, z, v, b, d, g, aa, ae, ah, ay, ow, oy | 16 |
| 3 | r, er, axr, cl, vcl, dx, hh, hv, jh, nx, y, ey, ao, aw, ax, axh, el, em, en, eng, uh, ux | 22 |

*2.1.2 Data-driven model sharing*

Data-driven model sharing assumes all the English models have equal distance from the Cantonese models. Different combinations were tried so as to obtain an optimum result for the detection of LB.

There are N=53 English models from the TIMIT corpus. For the initial trial, N-1 English models and all the Cantonese models will be used for phone recognition and then LB will be detected. The initial trial will iterate N times such that all combinations are considered. The one with the highest accuracy will be examined, and one more English model will be removed from this model set in the next trial. N-2 English models will next be used and N-1 iterations will be tried. English phone models are removed in each iteration until eventually the optimum result is obtained.

LB detection accuracy will drop when too many English phone models are removed. It is because those phones with language-dependent characteristics are clustered to the Cantonese phones and hence the language detection becomes less accurate.

**2.2 Language Model (Phonotatic Knowledge)**

The differences between the matrix language (Cantonese) and embedded language (English) will be reviewed before performing language identification.

Cantonese and English come from two different language families, so their phonological structures are quite different. Cantonese is a major Chinese dialect, which is Sino-Tibetan language. It is monosyllabic in nature and all the Cantonese syllables are of the canonical V, CV, CVC, VC forms, where V is vowel and C is consonant [9]. English is Indo-European language and the phonological structure is much more complex than Cantonese. In English discourse, over 80% of the syllables are of the canonical form mentioned above, and the remaining are C, CC, CCV, VCC, CCCV, CCCVCC, etc. [10]

Apart from the differences between phonological structures of the two languages, we can also make use of the intra-syllable bi-phone likelihood of Cantonese for language identification.

Many Chinese characters are homographs which have multiple meaning and pronunciations. To calculate the probability that the Chinese character exist in a particular phone sequence, Cantonese lexicon database CULEX is used [11]. The same character may have different phone sequences when it has different meanings. For example, the character 行 can be pronounced as /haang/, /hong/ and /hang/ in different phrases. The following example is to calculate the probability that 行 is pronounced as /haang/.

$$P(\text{行} = /haang/) = \frac{\sum_{i=1}^{N} b(phrase_i \sim \text{行}, \text{行} = /haang/)}{\sum_{i=1}^{N} b(phrase_i \sim \text{行})} \quad (1)$$

Where

- $N$ is the total number of phrases in the lexicon database
- $b(phrase \sim \text{行}, \text{行}=/haang/)$ means the phrase contains the character "行" and the phone sequence of "行" is /haang/, output 0 for false, output 1 for true
- $b(phrase \sim \text{行})$ means the phrase contains the character 行, output 0 for false, output 1 for true

Apart from the lexicon database, the character frequency database is also applied from which we can know the probability that the character appears in an utterance. A spoken Cantonese character frequency database with 1,646,000 characters is collected from three local newspapers. General equation for intra-syllable bi-phone probability is given by:

$$P_{LM}(XY) = \sum_{i=1}^{N} P(character_i) \times P(XY \mid character_i) \quad (2)$$

Where

- $X$ and $Y$ are Cantonese phones
- $XY$ is the phone sequence
- $N$ is total number of unique Chinese characters in the character frequency database

Intra-syllable bi-phone probability is calculated only for Cantonese since it is the matrix language. The intra-syllable bi-phone probability is for measuring the likelihood that the bi-phone is in Cantonese. All the Cantonese phone models and some of the English phone models will be used for phone recognition. Zero probability will be assigned for hypothesis bi-phone which involves English phone since it is likely to be English. The language model being used is mono-gram, thus there is no information on inter-syllable bi-phone probability. Hence, all the inter-syllable bi-phone probabilities are assumed to have equal likelihood.

English phones that have large distance from Cantonese phones can be identified by the English phones themselves. For English phones that have small distance from Cantonese, they will share the acoustic features with the Cantonese phones, thus bi-phone probability is used.

## 3. EXPERIMENT SETUP

Two sets of phone models are trained for each of the languages by monolingual speech corpus. Code-switching testing utterances are then passed to the phone recognizer and the bi-phone probability will be applied for language confidence measurement (likelihood that the phones are in Cantonese). The hypothesis LB is then compared with the hand-aligned reference. If the errors at both boundaries are less than the threshold, the LB detection will be regarded as correct.

### 3.1 Training of phone models

Two sets of phone models are trained individually with HTK [12] with two monolingual speech corpora – CUSENT and TIMIT.

*3.1.1 Cantonese phone model set*

Cantonese continuous speech corpus CUSENT is used for training the Cantonese phone model set. 59 context-independent mono-phone HMM models have been trained and they are listed in Table 2. Acoustic features used are mel-frequency cepstral coefficients (MFCC) which includes 12 cepstral coefficients and normalized energy. The first and second derivatives of the parameters are also included.

*Table 2: Cantonese phone model list*

| Phone type | Cantonese Phone models (LSHK) |
|---|---|
| Consonants | w, b, p, f, d, t, z, c, s, z(yu), c(yu), s(yu), l, j, h, k, g, kw, gw, initial_m, initial_n, initial_ng, final_m, final_n, final_ng, m, ng |
| Vowels | a, aa, e(i), e, u, u(ng), yu, i, i(ng), o, o(u), oe, eo |
| Vowel-consonant | ap, at, ak, aap, aat, aak, ek, ut, uk, yut, ip, it, ik, ot, ok, eot, oek |
| Others | ShortPause, Silence |

The Cantonese phone recognizer is evaluated by the development testing data in CUSENT. When no dictionary and language model are used, mono-phone accuracy is 72.05%.

*3.1.2 English phone model set*

English continuous speech corpus TIMIT is used for training the English phone model set. Here, 53 context-independent mono-phone models have been trained and they are listed in Table 1. Acoustic features are the same as the Cantonese phone models.

The English phone recognizer is evaluated by development test data in TIMIT. When no dictionary and language model are used, mono-phone accuracy is 50.07%.

### 3.2 Selection of phone models for phone recognition

The phone recognizer is constructed by all the Cantonese phone models and a selected English phone set. Data-driven and knowledge based approach are studied in order to obtain the optimum set of English phone models. Several experiments are set up and the details are listed in table 3 and table 4.

*Table 3: Experiment settings for data-driven model selection*

| Exp | removed Eng. models | # of models |
|---|---|---|
| 1 | nil | 101 |
| 2 | ah, eh | 99 |
| 3 | ah, eh, ae, ih | 97 |
| 4 | ah, eh, ae, ih, oy | 96 |
| 5 | ah, eh, ae, ih, oy, ng | 95 |
| 6 | ah, eh, ae, ih, oy, ng, ow | 94 |
| 7 | ah, eh, ae, ih, oy, ng, ow, ix | 93 |
| 8 | ah, eh, ae, ih, oy, ng, ow, ix, nx | 92 |

*Table 4: Experiment settings for Knowledge-based model selection*

| Exp | Eng. models | # of models |
|---|---|---|
| 1 | nil | 59 |
| 2 | Cat. 3 | 74 |
| 3 | Cat. 3+2 | 80 |
| 4 | Cat. 3+2+1 | 102 |

### 3.3 Code-switching data

Since no Cantonese-English code-switching speech data is available, we have initiated a task to compile code-switching speech data for this project. 940 code-switching utterances are collected and they all come from the same female speaker who's mother tongue is Cantonese. Each utterance contains one English word or phrase. 650 different English words and phrases are involved. All the Cantonese words are spoken Cantonese and the English words are pronounced with Cantonese accent.

520 utterances are for development testing, such as selecting English phone models and fine-tuning the threshold of bi-phone probability for language identification. The remaining 420 utterances are for final testing. Sample of code-switching data: "乜你冇 apply 嗰份工咩？" (Oh, you didn't apply for that job?)

### 3.4 Phone Recognition

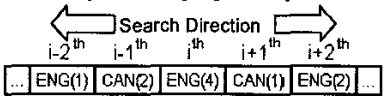
The selected phone models for each experiment will be used for phone recognition. No language model or dictionary will be used for the recognizer such that the recognition result can reflect the real phone sequence of the testing utterances.

### 3.5 Bi-phone Language Identification

Bi-phone probability of Cantonese is applied to the phone recognition hypothesis in order to measure the confidence that the bi-phone is Cantonese. If the bi-phone probability is larger than the threshold, the bi-phone will be identified as Cantonese bi-phone; otherwise, it will be identified as English bi-phone.

The bi-phone language identification is not reliable enough and therefore they are merged to a longer phrase. The assumption is that English content will only appear once in each utterance. It can be a single word or a phrase. The following example shows the procedure to merge the bi-phone language identification labels, "ENG" stands for English bi-phone and "CAN" stands for Cantonese bi-phone, the number inside the blanket is the number of consecutive bi-phone.

*Figure 1: Merging bi-phone language identification results into phrase language identification*




The part with maximum number of consecutive English bi-phone will be located (the $i^{th}$ language section), then search at the neighbouring language sections. If the previous $(i-1)^{th}$ or the next $(i+1)^{th}$ Cantonese section has more than one consecutive bi-phone, then the $(i-1)^{th}$ or the $(i+1)^{th}$ section has a higher probability to be Cantonese and the merging stops. Otherwise, if the next language section $(i+2)^{th}$ or the previous section $(i-2)^{th}$ has more than 1 consecutive English bi-phone, both of them are probably English and they will be merged into a longer English section. The final

language identification result will contain at most 3 language sections; only 1 of them is English. The assumption is that code-switching occurs once only in the utterance.

### 3.6 Measurement of LB detection error

The reference LB are marked manually. The hypothesis LB is compared with the reference one and the conditions are listed in table 5.

*Table 5: LB error measurement*

| Condition | Result |
|---|---|
| No overlap time for English phrases | Incorrect |
| Errors at both the two boundaries > threshold | Incorrect |
| Errors at both the two boundaries < threshold | Correct |

The threshold is set to be 0.3s, where the averaged time for each phone is around 0.1s in code-switching training data. The error is acceptable since each Cantonese character usually has 2 to 3 phones only.

### 3.7 Evaluation of the LB detector for monolingual speech

Two experiments have been devised to evaluate the performance of LB detection for monolingual English and monolingual Cantonese speech data. 200 monolingual English utterances and 300 monolingual spoken Cantonese utterances are collected for the evaluation test. All the speech data are recorded in the same condition as the code-switching data and the speaker is also the same.

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

The accuracy of LB detection for both the training data and testing data are listed below:

*Table 6: Accuracy on LB detection (Knowledge based)*

| | Threshold | LB detection rate (%) | | | |
|---|---|---|---|---|---|
| | | Dev. Test | Testing | English | Cantonese |
| Exp 1 | 0.00575 | 57.69 | 55.45 | 1.00 | **88.33** |
| Exp 2 | 0.00575 | 63.27 | 65.17 | 7.50 | 73.00 |
| Exp 3 | 0.00950 | **69.62** | **68.48** | 14.50 | 64.00 |
| Exp 4 | 0.00950 | 66.73 | 68.25 | **17.00** | 57.00 |

*Table 7: Accuracy on LB detection (Data-driven)*

| | Threshold | LB detection rate (%) | | | |
|---|---|---|---|---|---|
| | | Dev. Test | Testing | English | Cantonese |
| Exp 1 | 0.00575 | 67.50 | 74.88 | **17.00** | 57.67 |
| Exp 2 | 0.00575 | 68.08 | **75.12** | 16.50 | 59.67 |
| Exp 3 | 0.0055 | **76.54** | 74.64 | 16.00 | 60.33 |
| Exp 4 | 0.0055 | 76.35 | 74.17 | 16.00 | 60.67 |
| Exp 5 | 0.0055 | 76.35 | 73.46 | 16.00 | 61.00 |
| Exp 6 | 0.0055 | 75.77 | 73.46 | 16.00 | 62.33. |
| Exp 7 | 0.0055 | 76.15 | 73.70 | 16.00 | **63.00** |
| Exp 8 | 0.0055 | 75.78 | 73.70 | 16.00 | 63.00 |

The LB detector is designed for code-switching data that Cantonese is in dominant. Hence it performs better for monolingual Cantonese and code-switching data, and fail for monolingual English data. For monolingual Cantonese data, the LB detector performs best when no English phone models is used. For monolingual English data, the LB detector performs best when more English phone models is used. For code-switching utterances, data-driven model selection performs better than knowledge based model selection, which aims to select the English phone models having a lower similarity to Cantonese phone models. The remaining English phone models share the acoustic features of the Cantonese phone models. Each English word is recognized as a sequence of English and Cantonese phones. The present of English phones and low bi-phone probability mean that the phone sequence is likely to be English.

This paper describes some preliminary result in our study on detection of language boundary in code-switching utterances. This underlying problem is complicated and more work is necessary, for instance, to build an N-gram spoken Cantonese language model and to compile more training and testing data, before automatic recognition of speech utterances that contain more than one languages can be facilitated.

## 5. REFERENCES


[1] Ping Li, "Spoken Word Recognition of Code-Switched Words by Chinese-English Bilinguals", *Journal of Memory and Language*, 35, pp. 757 – 774, 1996

[2] Brian Hok-Shing Chan, "Code-mixing in Hong Kong Cantonese-English bilinguals: Constraints and process. *CUHK Papers in Linguistics*, 4, pp. 1-24, 1993

[3] P. C. Ching., et. al., "From phonology and acoustic properties to automatic recognition of Cantonese,"*ISSIPN-94*, pp. 127 – 132, Hong Kong, 1994.

[4] K.F. Chow, Tan Lee and P.C Ching, "Sub-syllable acoustic modeling for Cantonese speech recognition", *Proc. of ISCSLP98*, pp. 327 – 342, Singapore, 1998

[5] Pascale Fung, Liu, Xiaohu, and Cheung, Chi Shun, "Mixed-language Query Disambiguation". *Proc. of ACL 1999*, pp. 333 - 340, Maryland, 1999

[6] Tan Lee, W.K. Lo, P.C. Ching and Helen M. Meng, "Spoken language resources for Cantonese speech processing", in *Speech Communication*, Vol.36, No.3-4, pp.327 - 342, March 2002.

[7] P.C. Ching, K.F. Chow, Tan Lee, Alfred Y.P. Ng and L.W. Chan, "Development of a Large Vocabulary Speech Database for Cantonese", *Proc. of 1997 ICASSP*, pp. 1775-1778, Germany, 1997

[8] International Phonetic Association, *Handbook of the International Phonetic Association : a guide to the use of the International Phonetic Alphabet.*, Cambridge University Press, New York, NY, 1999

[9] Fisher, W.; Zue, V.; Bernstein, J. & Pallet, D., "An acoustic-phonetic data base", *Journal of the Acoustical Society of America, 81*, S92-3, 1987.

[10] Mirjam Wester, "Syllable Classification using Articulatory-Acoustic Features", *Proc. of Eurospeech 2003*, pp. 233-236, Geneva, Switzerland, 2003

[11] CULEX, http://dsp.ee.cuhk.edu.hk, 1999

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.1)*, Microsoft Corporation, 2000