Automatic Speech Recognition of Cantonese-English Code-mixing Utterances

CHAN Yeuk Chi Joyce

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Philosophy
in
Electronic Engineering

July 2005

Abstract of thesis entitled:

Automatic Speech Recognition of Cantonese-English Code-mixing utterances

Submitted by **Chan Yeuk Chi Joyce**

for the degree of **Master of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong**

in **July 2005**

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. In Hong Kong, code-switching between Cantonese and English is used widely in daily conversation. Cantonese is in majority, and English words or phrases are embedded in the utterances. Code-switching in Hong Kong tends to be intra-sentential and switching involving linguistic units above the clause level is rare, hence the preference for the term "code-mixing" in many studies.

Solutions to code-mixing speech recognition are a bit different from monolingual or multilingual speech recognition in both acoustic modeling and language modeling. In this thesis, performance of monolingual and cross-lingual acoustic models is compared and both monolingual and code-mixing training corpora are involved. Code-mixing speech data is collected for both training and evaluation purpose. Although the speech data is in fact read speech, the speaking style tends to be spontaneous since code-mixing mainly occurs in conversation rather than read speech such as news report.

Difficulties specific to code-mixing such as accents and lack of training data are tackle by several methods. Accents in the code-switch words are handled by modifications in pronunciation dictionary and clustering of acoustic models in the two languages. Four different types of language models are proposed in order to solve the problem on lack of code-mixing training text data. Language boundary detection based on algorithms in language identification as well as word lattice is studied and applied on the speech recognizer.

The proposed Cantonese-English code-mixing speech recognition system achieve 56.04% overall accuracy for the two languages, while accuracy on Cantonese characters is 56.37% and accuracy on English lexicons is 52.99%.

摘要

在一段談話或一句句子中使用兩種不同的語言，可稱爲語碼轉換(code-switching)。語碼轉換在雙語的社會中是常見的現象。在香港，廣東話和英語的語碼轉換在日常的對話中經常出現。廣東話是主要的語言，而英文單字及短句會加插其中。香港流行的語碼轉換通常在句子內出現，子句以上的轉換較爲少見，因此，很多研究都以語碼混合(code-mixing)來形容此現象。

語碼混合的語音辨認方法，與單語(monolingual)及多語(multilingual)時採用的方法，在聲學建模(acoustic modeling)和語言建模(language modeling)兩方面都有一定的差異。在這篇論文中，我們會比較以單語及語碼混合語音數據訓練出來的單語及誇語 (cross-lingual) 聲學模型。我們亦收集了一些語碼混合的語音數據用作訓練及系統測試。雖然這些語音數據是朗讀語料(read speech)，但其風格卻接近自發性口語語料 (spontaneous speech)，原因是語碼混合大多出現在對話而並非朗讀語料如新聞報導。

我們提出了多個方案以解決語碼混合獨有的難題，如口音(accents)及缺乏訓練數據等問題。我們修改了發音字典，並群集兩個語言的聲學模型，以解決嵌入語言字辭的口音問題。我們亦提出了四個語言模型的方案，以解決缺乏語碼混合文字的訓練數據之問題。我們亦會探討兩個分別以語言識別(language identification)演算法及以詞格(word lattice) 爲基楚的語言段點偵測(language boundary detection)系統，並將之應用在此語音辨認系統當中。

我們提出的廣東話及英語語碼混合語音辨認系統，總準確度爲 56.04%，其中廣東話單字的準確度爲 56.37%，而英語詞彙的準確度則爲 52.99%。

Acknowledgments

There are lots of people that I must recognize for their help towards completing this thesis. First, I would like to thank my supervisor, Professor P. C. Ching for his guidance and insightful advice throughout this research. He has given me lots of advices and suggestions on my thesis. I would also like to thank Professor Tan Lee, Professor William Wang and Professor Y. T. Chan for their valuable comments and suggestions.

Special thanks are given to Professor Helen Meng and Dr. Frank Soong for fruitful discussion and encouragement.

During these two years, I have had the fortune to work with groups of excellent and friendly colleagues and friends in DSP lab. I would like to appreciate Qian Yao for her knowledge sharing and technical assistance. Also, thanks are given to Zhu Yu and K. F. Chow for their help and comments. Some members have kindly donated speech data for the research as well; they are Athena Tam, Ada Ngan, Gogo Tsui and Natalie Tsang. Other members in DSP lab are always around here to give me their kind help, they are Li Yujia, Yang Chen, Yvonne Lee, Lousia Tao, Yuen Meng, Zhang Wei, Zheng Neng Heng, S. L. Oey, Qin Chao, Michael Zhang, Dexter Chan, Ouyang Hau and Gong Tao. Thanks very much Arthur Luk for his technical support.

Finally, I would like to express my deepest gratitude to my parents, my sister and Chan Chi Hong for their patience, love and encouragement throughout this research.

Table of Content

List of Figures

List of Tables

Abbreviations

The following abbreviations have been used throughout this work

ASR	Automatic Speech Recognition
C	Consonant
CDHMM	Context-dependent Hidden Markov Model
CFG	Context-free Grammar
CMU	Carnegie Mellon University
DAT	Digital Audio Tape
DCT	Discrete Cosine Transform
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
GWPP	Generalized Word Posterior Probability
HMM	Hidden Markov Model
IPA	International Phonetic Association
LBD	Language Boundary Detection
LID	Language Identification
LPC	Linear Predictive Coding
LSA	Latent Semantic Analysis
LSHK	Linguistic Society of Hong Kong
LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum a Posteriori
MFCC	Mel-Frequency Cepstral Coefficient
MI	Mutual Information
MLF	Master Label File
NLU	Natural Language Understanding
OOV	Out-of-Vocabulary
pdf	Probability Density function
PLP	Perceptual Linear Predictive
PLU	Phone-like Unit
POS	Parts-of-speech
PTT	Phonetic-to-text
SCSI	Small Computer System Interface
SIL	Silence
SLM	Stochastic Language Models

SP	Short Pause
TIMIT	Texas Instrument / Massachusetts Institute of Technology
V	Vowel
VQ	Vector Quantization
WER	Word Error Rate
WPP	Word Posterior Probability

Chapter 1

Introduction

1.1 Background

With the advances in computer technology, the use of computers has become an essential part of human life. The traditional way of human-computer communication depends mainly on keyboard, mouse and monitors, which is quite unnatural to human.

On the other hand, speech is the most convenient and natural way of communication among people. Human learn speaking in early ages and use speech every day. With speech recognition technology, human are able to communicate with computers in their most natural way.

Speech recognition technologies have been improved in the recent decades. In the past, the speech recognition systems were mainly speaker dependent, but now most of them are speaker independent. Users no longer need to provide large amount of speech data to train the system before using it. The vocabulary size also increased from tens of words to large vocabulary system that supports thousands of words. The content of speech utterance can be simple command, natural speech, or even spontaneous speech and conversation. Language being involved in speech recognition system is no longer limited to single language. Many systems already support multilingual speech recognition that user can select any one of the languages supported by the system [1]. In this case, either language identification (LID) is performed automatically [2], or the user has to specify which language to be used before the start of the interaction [3].

With the advances in speech recognition algorithms and computation power, the efficiency of man-machine communication is greatly improved and real-time applications can be deployed. These include enquiry systems [3], booking systems [4], call routing systems [5], dictation [6] [7], speech dialing [8] and many others.

Integrated with natural language processing system, information retrieval system, or translation system can also be realized [9-11].

There is a recent trend that more and more bilinguals often mix two languages in their daily conversation. For example, we can easily find that people in Hong Kong are so used to mixing Cantonese and English together while they talk. It is therefore necessary to develop speech recognition systems that are able to handle this type of multi-lingual speech input. The behavior to switch between two languages in conversation is called "code-switching". Matrix language is defined as the language with the higher frequency of morphemes in a discourse sample in which code-switching occurs [12], while the language with lower frequency of morphemes is the embedded language. If the switching of languages is intra-sentential, it can be called "intra-sentential code-switching" or "code-mixing", while the words in the embedded language are called "code-switch words". Code-switching occurs frequently in bilinguals' discourse [13] [14], and it is considered to be a very critical issue in bilingualism by many linguists [14] [15]. In the speech technology area, instead of using the term "code-switching", the phenomenon is usually generalized to "mixed-lingual" or "mixed language", and often literature can be found in this interesting field [16-18].

1.2 Previous Work on Code-switching Speech Recognition

Code-switching speech recognition can be realized by several approaches. It can be considered as a keyword spotting problem that the keyword may be a code-switch word [16]. It can also be regarded as a translation problem, such that the code-switch word is translated to the matrix language of the utterance. Appropriate lexicon in the matrix language is selected, such that the language model likelihood is maximized [17]. In code-switching utterances, several languages are involved, therefore automatic language identification is necessary [18]. The language boundary information can be treated as one of the confidence scores, in addition to the acoustic scores and language model scores, for decoding the hypothesis word string.

1.2.1 Keyword Spotting Approach

The keyword spotting approach for code-switching ASR was applied on the auto attendant telephone routing system [16]. The languages involved are Chinese and Taiwanese-accented English. Auto attendant telephone routing system routes the call according to the name in the input speech, which is either in Chinese or English, such that the name is the keyword to be spotted.

The acoustic score, verification score and tone score are referenced for the final decision of the hypothesis Chinese-keyword. However, English is a non-tonal language such that no tone score can be applied for the hypothesis English-keyword. The dynamic range on the combination scores is therefore different for the two languages. A normalization process is proposed to solve the scaling problem. A score-mapping function is obtained based on the relationship between false rejection rate and combination score. The combination score of one language can be projected to another language through using the same false rejection rate as shown in Figure 1-1. The dash-line shows the example that projects the combination score with false rejection rate 0.33.

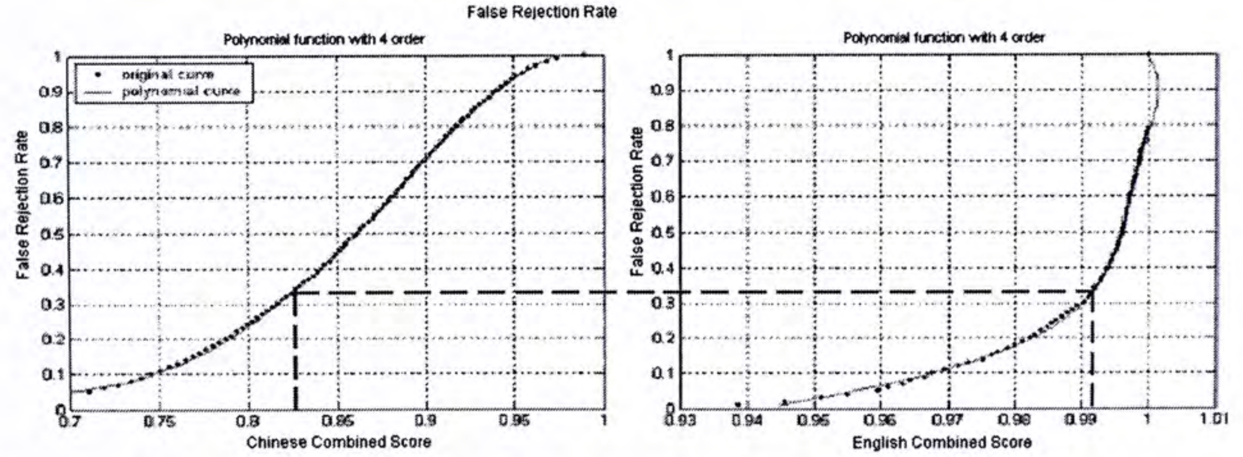


Figure 1-1 The relation curve of false rejection rate and combination score

The baseline recognition system without score mapping obtain 90.61% keyword spotting accuracy for the CEDB-1 database, which includes both English and Chinese person names. The proposed method achieves Top-1 recognition rate of 96.71% such that there is 6.10% absolute improvement [16].

The keyword spotting approach is domain specific that all the keywords have the same parts-of-speech (POS). A single language model with simple grammar can be applied as shown in Figure 1-2.

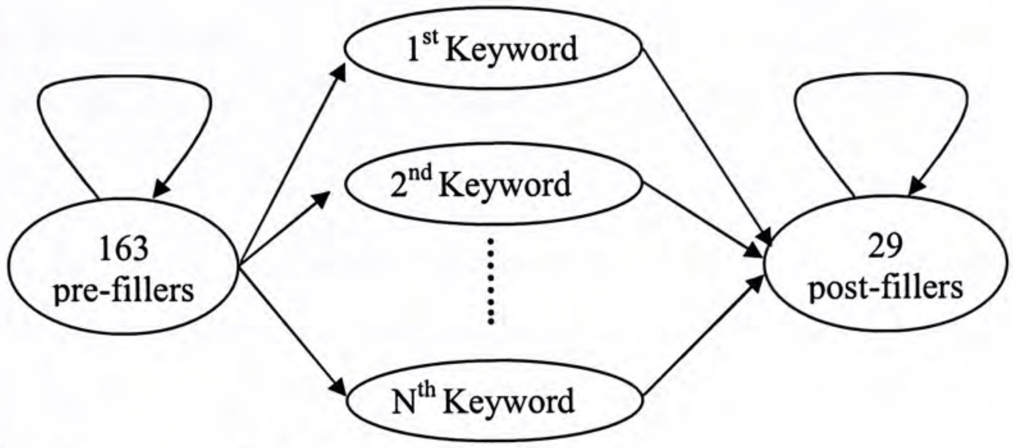


Figure 1-2 Language model of the keyword spotting system

The keyword spotting approach is not suitable for Large Vocabulary Continuous Speech Recognition (LVCSR) which includes thousands of code-switch words. There will be lots of pre-fillers and post-fillers, and position as well as the number of code-switch words may not be fixed. If it is code-switching between Chinese and English, where Chinese is the matrix language, the number of keywords may be even more than the number of filler words when non-tonal syllables are considered. The language model will be complicated and the concept of "keyword spotting" may be changed to "code-switch word spotting".

On the other hand, the idea of combination score may be applicable to LVCSR. The code-switch words may contain accent from the matrix language, such that the scale of acoustic likelihood may be different from the words in matrix language.

1.2.2 Translation Approach

On the other hand, mixed language query disambiguation approach by using co-occurrence information from monolingual data only has also been used to tackle the problem inherent to code-switching speech recognition [17]. The queries in mixed language are translated to monolingual queries in the matrix language. Their target is to build a multilingual spoken language interface to the Web, called SALSA [19-21]. Users can use speech to surf the internet via various links as well as issue search commands in natural language sentences. Commands and queries in English, Mandarin and Cantonese, as well as mixed language sentences of Mandarin-English and Cantonese-English are supported. Since English still constitutes 88% of the web pages, they regard English as the matrix language, Mandarin and Cantonese as the embedded language.

Each code-switch word C_i in the utterance is translated to translation candidates $\{E_{C_i}\}$ via an online bilingual dictionary. The translation candidates are then weighted based on co-occurrence information. The co-occurrence statistics is collected from monolingual English database. It provides mutual information (MI) between any two words in the same sentence:

$$MI(E_i, E_j) = \log \frac{f(E_i, E_j)}{f(E_i) f(E_j)} \tag{1.1}$$

where

$f(E)$ = frequency of the word E in the database

E_i = the i^{th} word in the utterance

In order to select the target translation, a particular E_c should be chosen from the translation candidates $\{E_{C_i}\}$. This pruning process is called "translation disambiguation". Three unsupervised statistical methods were proposed, 1) The baseline system that only considers the MI between the code-switch word and the word next to it; 2) The voting method that all contextual words vote for the best translation word; and 3) The 1-best contextual word approach that the most discriminative contextual word is applied for MI calculation to select the translation word. Among the three methods, the 1-best contextual word approach achieves the highest translation accuracy, which is about 90% when percentage of code-switch words is 10% in each utterance.

The mixed language query disambiguation approach can efficiently translate the code-mixing utterances into monolingual. However, it needs a large text database to obtain the mutual information and the contribution weight, which may not be available for all languages, especially for languages that the written form and spoken form are fairly different. Cantonese is one of these languages, which the written form is standard Chinese that Mandarin is based on. There are differences in lexicon and grammar between the written form and spoken form. Although spoken Cantonese can also be found in some text materials, the percentage is low and domain specific [22]. Moreover, there is no word boundary in Chinese written text, such that the word-based co-occurrence statistics changes with the segmentation method. On the other hand, each Chinese character can form different words that have diverse meanings, such that character-based co-occurrence statistics may not be reliable. Therefore, if the speech utterances are Cantonese-English code-switching that Cantonese is the matrix language, the proposed method may not be applicable. Additionally, not all the code-switch terms can be translated to the matrix language and the word order may sometimes be changed. In some cases, the reason behind code-mixing is that no equivalent is available in the matrix language, such that people have to switch the words to another language [23]. Moreover, the meaning of the code-switch word in code-switching speech is sometimes different from those being applied in monolingual speech. The standard bilingual dictionary is not applicable, such that a tailor-made dictionary should be built.

To apply the mixed language query disambiguation approach on Cantonese-English code-mixing LVCSR, several parts should be modified:

a) If large Cantonese text database is not available for mutual information computation, text database in other language (e.g. standard written Chinese) may also be considered.

b) Tailor-made bilingual dictionary should be built

The purpose of translation of the code-switch word is changed, no longer for information retrieval, but to verify the hypothesis code-switch word. The mutual information can be considered as a confidence measure on the hypothesis code-switch word.

1.2.3 Language Boundary Detection

It has been proposed to use maximum a posteriori (MAP) estimation to detect and identify the language boundaries in mixed-language speech [18]. A speech utterance S with N_s feature vectors and q language boundaries, can be segmented into $q+1$ speech segments, such that $S = (S_1, S_2, ..., S_{q+1})$. The positions of the language boundaries are denoted as $R = (r_1, r_2, ..., r_q)$, where $1 < r_1 < r_2 < ... < r_q < N_s$. The corresponding language sequence is $\tilde{L} = (L_{S_1}, L_{S_2}, ..., L_{S_{q+1}})$, where $\forall L_{s_i} \in \{L_1, ..., L_K\}$, K = number of languages in the utterance, such that L_{s_i} represent the language of the speech segment S_i.

The proposed MAP based approach jointly segment and identify an utterance with mixed languages. The boundary number is at first determined by the MAP estimation. Latent Semantic Analysis (LSA) [24] based Gaussian Mixture Model (GMM) and Vector Quantization (VQ) based bi-gram language model are then proposed to characterize a language and used for language identification. Finally, a likelihood ratio test approach is utilized to determine the optimal number of language boundaries.

This approach shows a promising performance on the boundary detection rate and the LID rate (76.2%). The language boundary information is important for code-mixing speech recognition system that words in the wrong language will rarely be chosen during the decoding process. Improvement can be made on both the matrix and embedded language. However, the training process of the LSA-based GMM is complicated and a large database is necessary. Hence, to simplify the case, the traditional method for LID in monolingual speech will be applied in this thesis for language boundary detection. Phone-based and syllable-based approaches are proposed, and likelihood in the matrix language unit is utilized for LID.

1.3 Motivations of Our Work

Code-switching is common in bilingual societies. The languages involved can be Spanish / English, Swiss-German / Italian, French / Italian, Hebrew / English, Cantonese / English, etc [25]. People mainly speak in their primary language (matrix language), and embed words from a secondary language (embedded language). However, the code-switch words in embedded language usually contain accent of the matrix language, which make automatic speech recognition difficult, if not impossible.

Hong Kong is an international city where most people are Cantonese and English bilinguals. They also like mixing English words in their daily conversation, where Cantonese is mainly used. In order to realize automatic speech recognition for applications in bilingual societies, the speech recognizer should be able to handle code-switching speech properly. With the natural language understanding (NLU) system, human speech can be understood by computers, no matter the speech is in monolingual or code-switching. The code-switching speech recognizer can be applied for query systems, booking systems, speech dialing systems, etc. Integrated with NLU systems, much more applications can be realized. For example, automatic ticket booking systems can employ code-switching speech recognizer and NLU system. The automatic speech recognition system apply acoustic-phonetic analysis techniques to translate the audio input to text, while the NLU system try to understand human language and generate appropriate response. Feedback mechanism between the NLU and acoustic-phonetic stages is essential since continuous speech is normally filled with acoustic ambiguity which can only be resolved through the use of higher source of knowledge. NLU plays two important roles in translating speech input to useful computer commands. The first role is extracting the correct meaning of the speech such that the computer gets the right message. The second role is reducing acoustic phonetic ambiguity in normal speech based on "understanding" of meanings [26]. With code-switching speech recognizer, users are able speak naturally, no matter it is monolingual or code-switching speech, to interact with the system.

The switching between languages can be inter-sentential (code-switching) or intra-sentential (code-mixing). In Hong Kong, people usually code-switch in word-based, such that code-mixing is more common than code-switching. Hence, our research will focus on Cantonese-English code-mixing instead of code-switching.

This thesis will focus on methodologies of LVCSR for Cantonese-English code-mixing utterances. The following problems will be considered and tackled:

a) Accent in the embedded language
b) Phone change and syllable fusion in Spoken Cantonese
c) Language boundary detection
d) Lack of speech data for training and evaluation
e) Lack of text data for training the code-mixing language model

1.4 Methodology

It is proposed to use one cross-lingual speech recognizer with bilingual dictionary to solve the code-mixing problem [27]. To deal with the Cantonese accents in the English code-switch words, the pronunciation dictionary includes the phone sequence of both the native and accented version of English words. The speech recognizer is word-based for English, and syllable-based for Cantonese, and a word graph is generated for each speech file. The word graph is syllable-based for Cantonese, and word-based for English, and we call it syllable lattice. A separated language boundary detector will be applied, such that the time of language change can be detected. Instead of using the MAP estimation and GMM mentioned in [18], the language identification will be based on bi-phone or bi-syllable probabilities of the N-best hypothesis. Duration information of the English hypothesis in the syllable lattice is considered as language identification measurement as well. The language boundary information, which acts as confidence measure of the hypothesis words, will be applied on the syllable lattice. Another pronunciation dictionary which contains the mapping of Cantonese syllables and Cantonese characters will then be applied, such that the syllable lattice can be converted to character lattice (character for Cantonese, word for English). The translation approach [17] will be applied on the code-switch word *E*, therefore mutual information will be measured between the translated candidates $\left\{C_{E_i}\right\}$ of the code-switch word and the neighbouring Cantonese words. The maximum mutual information will be regarded as the language model score for that code-switch word. Generalized Word Posterior Probability (GWPP) will be calculated for each arc in the character lattice, and the weight of acoustic likelihood and language model will be turned, hence the hypothesis word string with lowest expected word errors can be obtained. Figure 1-5 summarizes the Cantonese-English code-mixing ASR system.

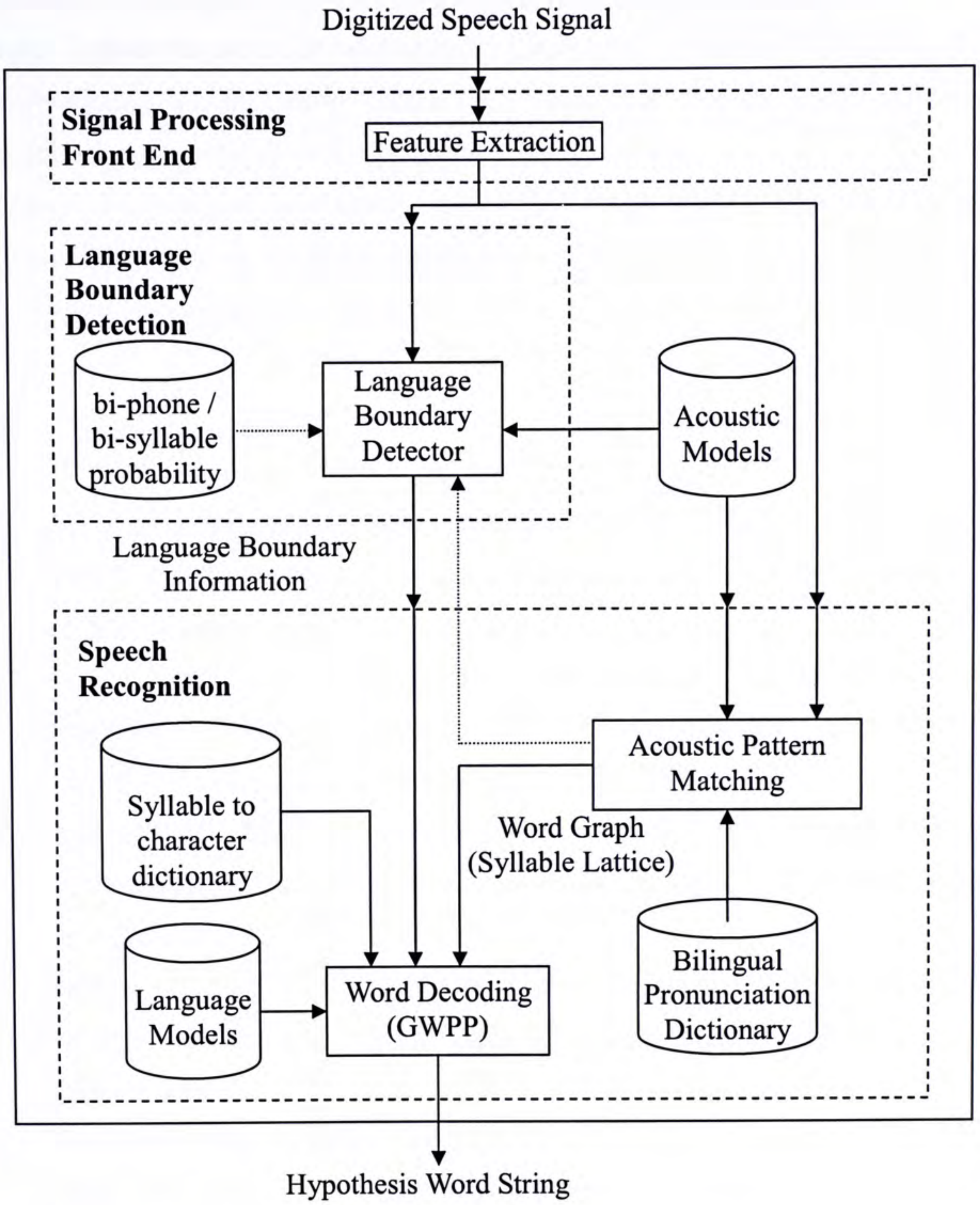

Digitized Speech Signal
Signal Processing Front End
Feature Extraction
Language Boundary Detection
bi-phone / bi-syllable probability
Language Boundary Detector
Acoustic Models
Language Boundary Information
Speech Recognition
Acoustic Pattern Matching
Syllable to character dictionary
Word Graph (Syllable Lattice)
Language Models
Word Decoding (GWPP)
Bilingual Pronunciation Dictionary
Hypothesis Word String

Figure 1-3 Block diagram of the Cantonese-English code-mixing ASR system

1.5 Thesis Outline

The thesis includes 6 chapters. Chapter 1 is introduction which mentions the background and motivation for the research. Chapter 2 includes the fundamentals of Large Vocabulary Continuous Speech Recognition (LVCSR) for Cantonese and English. Basic theories of automatic speech recognition and characteristics of the two involved languages – Cantonese and English will also be introduced.

Chapter 3 gives the definition of code-mixing and code-switching. Characteristics of code-mixing will be centered around Cantonese and English. Difficulties for code-mixing speech recognition are analyzed and discussed.

Chapter 4 describes the code-mixing speech data being used for both training and testing, while experimental setup will be given in Chapter 5. Finally, results and analysis will be elaborated in Chapter 6.

1.6 References

[1] Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, M. Woszczyna, "Multilinguality in speech and spoken language systems", in *Proc. of the IEEE*, Vol. 88, Issue 8, pp. 1297-1313, 2000

[2] N. Udhyakumar, R. Swaminathan and SK Ramakrishnan, "Multilingual speech recognition for information retrieval in Indian context", in *Proc. of the Student Research Workshop*, HLT/NAACL, Boston, USA, 2004

[3] Helen M. Meng, Steven Lee, Carmen Wai, "CU FOREX: A Bilingual Spoken Dialog System for Foreign Exchange Enquiries", in *Proc. of ICASSP 2000*, pp. 1229-1232, 2000

[4] Peabody, Mass., Ghent, *ScanSoft Powers Premier Travel Inn UK Speech-Based Booking System,* http://www.scansoft.com/news/pressreleases/20050512_pretravel.asp, 2005

[5] Valentine C. Matula, "Improved LSI-Based Natural Language Call Routing Using Speech Recognition Confidence Scores", *Technical Report of Avaya Labs Research (ALR-2004-023)*, 2004

[6] Philips Dictation Systems, http://www.dictation.philips.com

[7] IBM ViaVoice Release 10 – Standard Edition, http://www.scansoft.com/viavoice/standard/

[8] Lucent Technologies, *AnyPath® Voice Portal Solution for Service Providers*, http://www.lucent.com/livelink/0900940380043da2_Brochure_datasheet.pdf, 2004

[9] Michael Witbrock and Alexander G. Hauptmann, "Speech Recognition and Information Retrieval: Experiments in Retrieving Spoken Documents", in *Proc.*

of the 1997 DARPA Speech Recognition Workshop, Chantilly, VA, 1997

[10] Helen Meng, Sanjeev Khudanpur, Gina Levow, Douglas W. Oard, Hsin-Min Wang, "Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval", in *Proc. of the HLT 2001*, San Diego, 2001

[11] Ryosuke Isotani, Kiyoshi Yamabana, Shin-ichi Ando, Ken Hanazawa, Shin-ya Ishikawa, Tadashi Emori, Ken-ichi Iso, Hiroaki Hattori, Akitoshi Okumura, Takao Watanabe. "An Automatic Speech Translation System on PDAs for Travel Conversation", in *Proc. of ICMI 2002*, p. 211, 2002

[12] Helena Halmari, *Government and Codeswitching: Explaining American Finnish*, Amsterdam; Philadelphia: J. Benjamins, 1997

[13] Domingue, N, "Bi- and multilingualism: Code-switch, interference and hybrids", *Trends in Linguistics: Studies and Monographs,* Vol. 48, New York: Mouton de Gruyter. 1990

[14] Myers-Scotton, C. , *Social motivations for codeswitching*, Oxford: Oxford University Press, 1993

[15] Romaine, S. , *Language in society*, Oxford: Oxford University Press, 1994

[16] Shan-Ruei You, Shih-Chieh Chien, Chih-Hsing Hsu, Ke-Shiu Chen, Jia-Jang Tu, Jeng Shien Lin, Sen-Chia Chang, "Chinese-English mixed-lingual keyword spotting", in *Proc. of ISCSLP 2004*, pp. 237-240, Hong Kong

[17] Pascale Fung, Liui Xiaohu, Cheung Chi Shun, "Mixed language query disambiguation", in *Proc. of the ACL 1999*, pp.333-340, USA, 1999

[18] Chi-Jiun Shia, Yu-Hsien Chiu, Jia-Hsin Hsieh, Chung-Hsien Wu, "Language boundary detection and identification of mixed-language speech based on MAP estimation", in *Proc. of ICASSP 2004*, Vol. 1, pp . 381-384, 2004

[19] Pascale Fung, Cheung Chi Shun, Lam Kwok Leung, Liu Wai Kat, Lo Yuen Yee, "SALSA Version 1.0: A Speech-Based Web Browser for Hong Kong English", in *Proc. of ICSLP 1998,* Vol. 4, pp. 1615-1619, Sydney Australia, 1998

[20] Pascale Fung, Cheung Chi Shun, Lam Kwok Leung, Liu Wai Kat, Lo Yuen Yee and Ma Chi Yuen, "SALSA, A Multilingual Speech-Based Web Browser", in *Proc. of The First AEARU Web Technology Workshop*, pp. 16-21, Kyoto, 1998

[21] Amélie Plu, Ma Chi Yuen, Pascale Fung, "SALSA Version 3.0: A Single Recognizer-based Multilingual Speech-based Web Browser", in *Proc. of RAIO 2000*, Paris, France, 2000

[22] Don Snow, *Cantonese as Written Language: the Growth of a Written Chinese Vernacular*, Hong Kong University Press, Hong Kong, 2004

[23] David C. S. Li, "Cantonese-English code-switching research in Hong Kong: a Y2K review", *World Englishes*, Vol. 19, No.3, p.305-322, Blackwell Publishers Ltd., 2000

[24] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, Vol. 41, pp. 391-407, 1990

[25] Peter Auer, *Code-Switching in Conversation: Language, Interaction and Identity*, London and New York: Routledge, 1998

[26] George M. White, "Natural Language Understanding and Speech Recognition", *Communications of the ACM archive*, Vol. 33, Issue 8, pp. 72-82, August 1990

[27] Ma Chi Yuen, *Multilingual Speech Recognition and its Application in a Multilingual Voice Browser*, MPhil. Thesis, The Hong Kong University of Science Technology, 2001

Chapter 2

Fundamentals of Large Vocabulary Continuous Speech Recognition for Cantonese and English

The underlying principles for Large Vocabulary Continuous Speech Recognition (LVCSR), Word Posterior Probability (WPP) and Generalized Word Posterior Probability (GWPP) will be discussed briefly in this chapter. Characteristics of the two languages involved in code-mixing - Cantonese and English – will also be introduced.

2.1 Basic Theory of Speech Recognition

The goal of a speech recognition system is to convert the input speech utterance to its written form. Acoustic features will be extracted from the speech input to form the observation sequence. Pattern matching will then be performed to find a word string that has maximum likelihood to produce such an observation sequence.

2.1.1 Feature Extraction

The purpose of feature extraction is to compress the speech signal to provide a compact representation on it. There are many features commonly used for speech recognition, such as Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC) [1], Perceptual Linear Predictive (PLP), etc [2]. In this research, MFCC will be used and a brief description is given below.

Mel frequency scale is a logarithmic scale in the frequency domain that models the non-linearity of human perception. MFCC represents the short-time spectral feature by coding the speech signal on a frame-by-frame basis with overlap. Each window frame of speech is first transformed to frequency domain, and then triangular filter banks with equal bandwidth in the Mel scale are applied to the spectrum. The resultant spectral coefficients are decorrelated using Discrete Cosine Transform (DCT). The first N cepstral components after liftering form the desired N-dimension MFCC vectors. Frame energy may also be considered as one of the parameters of the feature vector. The first and second time derivatives are usually calculated as the dynamic features [3]. Figure 2-1 shows the block diagram of the MFCC coding scheme.

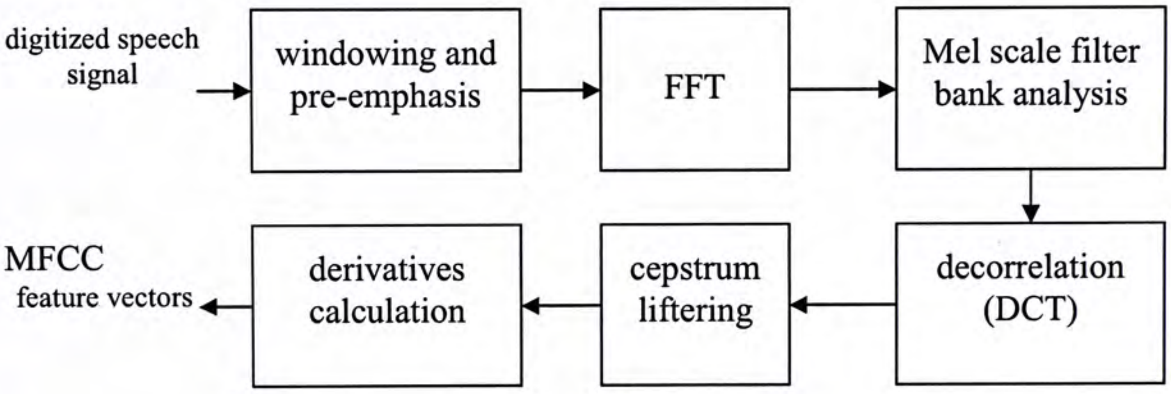


Figure 2-1 Block diagram of the MFCC coding scheme

2.1.2 Maximum a Posteriori (MAP) Probability

Speech recognition is assumed to a simple probabilistic model of speech production. A specified word sequence *W* produces an acoustic observation sequence *O*, with probability *P(W,O)*. Based on the acoustic observation sequence, a word string $\hat{W}$ with maximum a posteriori (MAP) probability will be decoded [4].

$$\hat{W} = \arg\max_{W} P(W \mid O) \qquad (2.1)$$

When Bayes' Rule Equation is applied to equation (2.1), the equation becomes:

$$\hat{W} = \arg\max_{W} \frac{P(O \mid W) \cdot P(W)}{P(O)} \qquad (2.2)$$

Since *P(W)* and *P(O)* are independent, the denominator *P(O)* can be neglected and the equation can be simplified to

$$\hat{W} = \arg\max_{W} P(O \mid W) \cdot P(W) \qquad (2.3)$$

P(O|W) is the acoustic model which estimates the probability of a sequence of acoustic observations, conditioned on the word string. For LVCSR, the acoustic models are usually statistical models in sub-word speech units. Sub-word speech units enable the recognition of words that had not been encountered in the training data. There are several possible choices for sub-word units that can be used to model speech, such as phone-like units (PLUs), syllable-like units and demisyllable-like units [4]. Large amount of speech data will be collected for training of the acoustic models, and observed features *O* of the specified speech unit *W* will be modeled by P(O|W).

The second term P(W) is the language model which describes the probability associated with a postulated sequence of words. The language models are built from text database, such that both syntactic and semantic constraints of the language and the recognition task can be incorporated. Stochastic language models (SLM) are one of the commonly used language model, which take a probabilistic viewpoint of language modeling. There are two major SLM: 1) Context-free grammar (CFG); 2) N-gram models. CFG models the syntax and semantics of the language. It is powerful to describe most of the structure in spoken language, and it is restrictive enough to have efficient parsers. On the other hand, N-gram assumes the occurrence of the current word only depends on the previous N-1 words. N-gram models are powerful for domain-independent applications [3].

2.1.3 Hidden Markov Model (HMM)

Acoustic modeling can be performed by Hidden Markov Model (HMM) approach [5][6]. A speech signal can be regarded as an observation sequence of feature vectors derived from the samples. An HMM is a state machine that can generate the observation sequence. The term "hidden" refers to the fact that we can observe the output sequence of an HMM but the actual state transitions are unknown. In acoustic modeling, the feature vectors of a sound unit will be considered as the output generated by an HMM as in Figure 2-2. The HMM can be treated as the template of that sound unit.

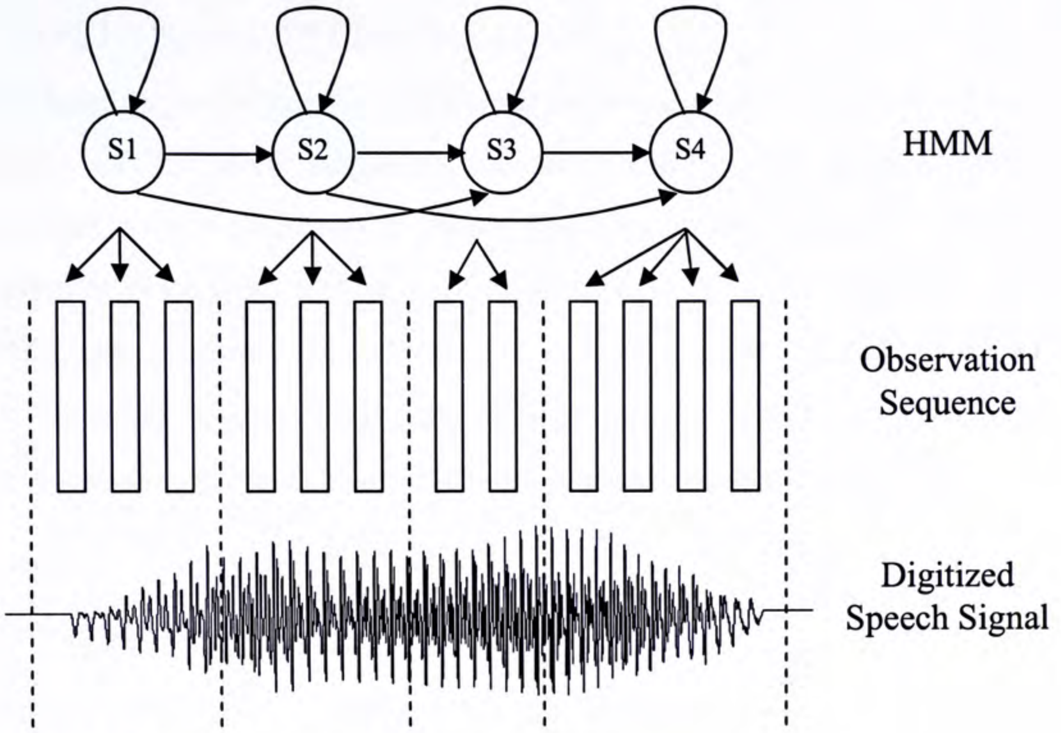


Figure 2-2 An illustration of HMM for acoustic modeling

2.1.4 Statistical Language Modeling

Language model defines the constraints in the language. It determines the most possible word sequence of the input speech signal. LVCSR involves thousands of words, thus statistical approach is usually adopted. Let the word sequence *W* be $w_1w_2w_3...w_M$, where M is the number of words in the word sequence. The probability of occurrence of the word sequence *W* can be computed:

$$P(W) = P(w_1w_2...w_M) = P(w_1)P(w_2 \mid w_1)...P(w_M \mid w_1w_2...w_{M-1}) \qquad (2.4)$$

where

w_i = the ith word in the word string *W*

M = number of words in the word string *W*

Given the previous *M-1* words $w_1w_2...w_{M-1}$, the probability of the current word w_M has to be estimated. However, it is impossible to have a reliable estimate of $P(w_M \mid w_1w_2...w_{M-1})$ for all value of *M*. The size of training data is limited that cannot include all the word combinations for LVCSR systems. Hence, the conditional probability in equation (2.4) is usually approximated that each word only depends on the previous N-1 words:

$$P(w_M \mid w_1 w_2 \ldots w_{M-1}) \approx P(w_M \mid w_{M-N+1} \ldots w_{M-1}) \qquad (2.5)$$

Zero probability may occur for some word sequence when N or the word size is large, i.e. some of the word sequences do not exist in the training data. These probabilities will be smoothed to ensure that zero probabilities will not occur. The estimation of those unseen word sequences depends on the (N-1)gram. The idea of backoff N-gram is that whenever there is an unseen event, the probability is backoff to a lower order N-gram. The backoff N-gram (N=3) is given in equation (2.6), where $\tilde{P}$ is the smoothed N-gram, γ_1 and γ_2 are the backoff weight [7].

$$\tilde{P}(w_i \mid w_{i-2} w_{i-1}) = \begin{cases} P(w_i \mid w_{i-2} w_{i-1}) & \text{if } C(w_{i-2} w_{i-1} w_i) > 0 \\ \gamma_1 P(w_i \mid w_{i-1}) & \text{if } C(w_{i-2} w_{i-1} w_i) = 0 \\ & \text{and } C(w_{i-2} w_{i-1}) > 0 \\ \gamma_2 P(w_i) & \text{otherwise} \end{cases} \qquad (2.6)$$

If the (N-1)gram probability is also zero, (N-2)gram probability is considered. There are many methods to determine the value of backoff weight γ_1 and γ_2, such as Good Turing discounting, Witten Bell discounting, Absolute discount, Linear discounting [7-9], etc. All these methods have their own advantages, and a comparison of these methods for language modeling for Mandarin Chinese can be found in [10].

Good Turing discounting is applied in this research, and the details will be described in chapter 5.

2.1.5 Search Algorithm

The set of HMMs trained for different sound units is used as the acoustic model. Since there are variations in speaking rate, the time duration is not always the same even for the same word. To align the observation sequence with the HMM as well as the states, searching algorithms are necessary. Viterbi algorithm [6] which is a dynamic programming technique is applied to solve the optimum sequence and time alignment of the acoustic models. All the paths will be processed at the same time, and the maximum score will be stored. The score here is the cumulative probability density of the observations given the HMM. There is at least one state in each HMM, and likelihood of state transitions is stored in the model. For continuous speech, all the HMMs are connected together, according to the language model constraints if provided, to form a large HMM, and Viterbi search is then performed. The decoded word sequence is found by back tracing the best state sequence and recording every inter-HMM transition. Figure 2-3 shows an example of using Viterbi algorithm to find the best state sequence. The black dots in the figure store the maximum score at time *t* at state *j*. The best state sequence in this example is {1,1,1,2,2,3,3,3}. The last black dot carries the total score of this HMM for generating the observations.

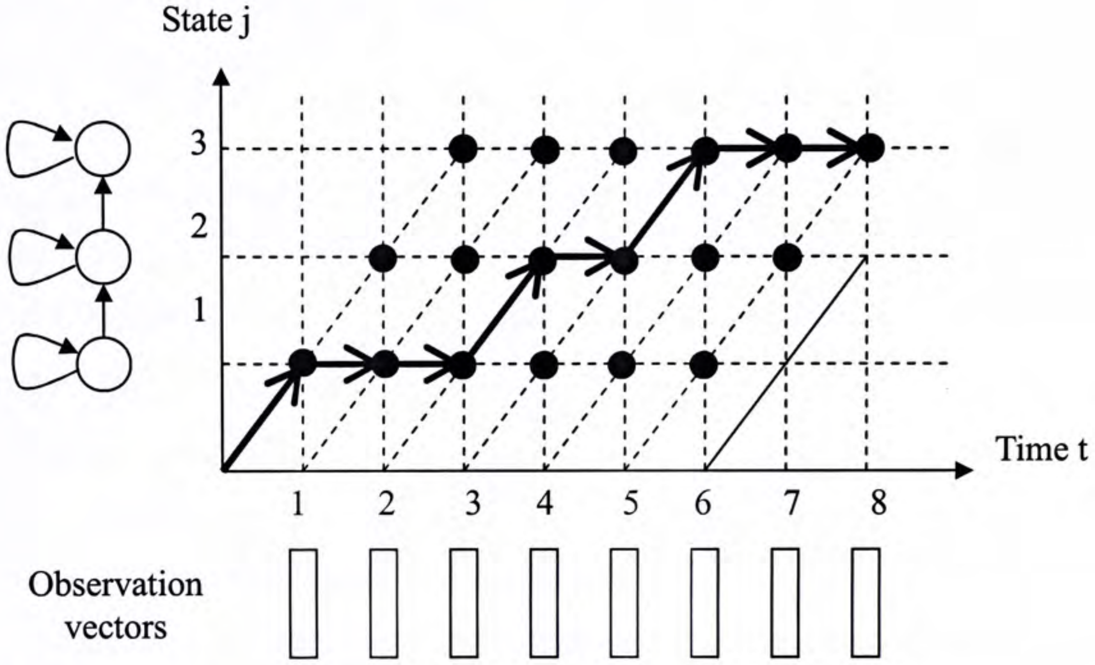


Figure 2-3 Using Viterbi algorithm to find the best state sequence

2.2 Word Posterior Probability (WPP)

The string posterior probability in equation (2.3) measures the likelihood of a recognized string, $\hat{W}$, given the observation sequence *O*. It is hypothesized with its corresponding time segmentations by the Viterbi search, i.e.

$$[w;s,t]_1^M = [w_1;s_1,t_1]...[w_M;s_M,t_M] \qquad (2.7)$$

where

s = starting time frame of the word w

t = ending time frame of the word w

$s_1 = 1$ (the first frame)

$t_M = T$, which is the last frame

$t_m + 1 = s_{m+1}$ for $1 \le m \le M-1$

Assume that the acoustic observations $x_{s_m}^{t_m}$ which starting time frame is s_m and ending time frame is t_m, depends solely on the corresponding word w_m, equation (2.3) can be rewritten as equation (2.8) [11].

$$\begin{aligned}\hat{W} = \hat{w}_1^M &= \arg\max_{M,w_1^M} P(w_1^M \mid x_1^T) \\ &= \arg\max_{M,w_1^M} \frac{P\left(x_1^T \mid [w;s,t]_1^M\right)\cdot P\left([w;s,t]_1^M\right)}{P(x_1^T)} \\ &= \arg\max_{M,w_1^M} \frac{\prod_{m=1}^{M} P(x_{s_m}^{t_m} \mid w_m)\cdot P(w_m \mid w_1^{m-1})}{P(x_1^T)} \qquad (2.8)\end{aligned}$$

The string posterior probability is now decomposed into a product of all the acoustic and language model probabilities of the corresponding word w_m at the corresponding segmentation points s_m and t_m. The dependency between the current word w_m and preceding words is addressed in the language model $P(w_m \mid w_1^{m-1})$.

The performance of a speech recognizer is usually measured by word error rate (WER). However, the standard MAP decoding approach mentioned in equation (2.3) and (2.8) does not necessarily minimize the WER even given optimal models [12]. The goal of the standard MAP approach is to find the sentence hypothesis that maximizes the posterior probability *P(W|O)* of the word sequence *W* given the acoustic information *O*, which has already been mentioned in equation (2.1) and (2.3):

$$\hat{W} = \arg\max_{W} P(W \mid O) \qquad (2.1)$$

$$\hat{W} = \arg\max_{W} P(O \mid W)\cdot P(W) \qquad (2.3)$$

Bayesian decision theory [13] tells that the maximizing sentence posteriors minimize the sentence level error, which is the probability of having at least one error in the sentence string. However, performance of speech recognizers is usually evaluated by word error, i.e. the Levenshtein distance [14] between the hypothesis and the reference string. Levenshtein distance is defined as the number of substitutions, deletions and insertions in the hypothesis relative to the reference under an alignment of the two strings that minimizes a weighted combination of these three types of errors. This performance metric is a more forgiving that it gives partial credit for correctly recognized portions of sentences:

$$WER = \frac{S+D+I}{N} \times 100\% \qquad (2.9)$$

where

S = Number of substitution errors

D = Number of deletion errors

I = Number of insertion errors

N = Number of words in the reference

It is assumed that sentence error and word error rates are highly correlated, such that minimizing one would tend to minimize the other. However, according to the empirical result from [12], there is a significant difference between optimizing for sentence vs. word error rate. An example is shown in Table 2-1.

Hypothesis (H)			P(H|O)	P(w1|O)	P(w2|O)	P(w3|O)	E[correct]
w1	w2	w3					
I	DO	INSIDE	*0.16*	0.34	0.29	0.16	0.79
I	DO	FINE	0.13	0.34	0.29	0.28	0.91
BY	DOING	FINE	0.11	0.45	0.49	0.28	*1.22*
BY	DOING	WELL	0.11	0.45	0.49	0.11	1.05
BY	DOING	SIGHT	0.10	0.45	0.49	0.10	1.04
BY	DOING	BYE	0.07	0.45	0.49	0.07	1.01
BY	DOING	THOUGHT	0.05	0.45	0.49	0.07	0.99
I	DOING	FINE	0.04	0.34	0.49	0.28	1.11
I	DON'T	BUY	0.01	0.34	0.01	0.01	0.36
BY	DOING	FUN	0.01	0.45	0.49	0.01	0.95

Table 2-1 Example illustrating the difference between sentence and word error measures

The 10-best list of hypotheses produced by the recognizer is shown in the first column. The corresponding joint posterior probabilities *P(H|O)* is in column 2. Column 3 gives the posterior probabilities *P(w|O)* for individual word. The posterior word probabilities follow from the joint posterior probabilities by summing over all hypotheses that share a word in a given position. The expected number of correct words *E[correct]* is shown in column 6, is computed simply by summing up the individual word posterior probabilities:

$$\begin{aligned} E[correct] &= E[\text{words correct}(w_1 w_2 w_3) \mid O] \\ &= E[correct(w_1) \mid O] + E[correct(w_2) \mid O] + E[correct(w_3) \mid O] \\ &= P(w_1 \mid O) + P(w_2 \mid O) + P(w_3 \mid O) \end{aligned} \tag{2.10}$$

From the above example, although the hypothesis in row 3, "BY DOING FINE" does not have the highest posterior, it has the highest expected number of correct words, i.e. minimum expected word error. The correct answer for this example is "I'M DOING FINE", such that the MAP hypothesis "I DO INSIDE" has misrecognized all the words (WER=3), whereas the hypothesis in row 3 recognized incorrectly only one word (WER=1). From this example, we see that optimizing overall posterior probability (sentence error) does not always minimize expected word error. This is because words with high posterior probability did not have high posterior probability when combined [12].

Hence, instead of considering the whole utterances (MAP), word posterior probability is preferred, such that confidence can be measured in the same unit as the performance metric. Word posterior probability can measure the word reliability by summing all the posterior probabilities of strings consisting of the specific word *w*, at given starting and ending time frames, *s* and *t*. Equation (2.11) describes the word posterior probability.

$$P([w;s,t] \mid x_1^T) = \sum_{\substack{\forall M,[w;s,t]_1^M \\ \exists n, 1 \le n \le M \\ w=w_n, s=s_n, t=t_n}} \frac{\prod_{m=1}^{M} P(x_{s_m}^{t_m} \mid w_m) \cdot P(w_m \mid w_1^{m-1})}{P(x_1^T)} \tag{2.11}$$

where

$[w;s,t]$ = word hypothesis

x_s^t = sequence of acoustic observation

M = number of words in a string hypothesis

$P(s_1^T)$ = probability of the acoustic observations

T = the length of the complete acoustic observations

In order to measure the word confidence effectively, some practical issues of word posterior probability need to be investigated. Therefore, Generalized Word Posterior Probability (GWPP) was proposed [11].

2.3 Generalized Word Posterior Probability (GWPP)

GWPP uses the N-best list or word graph to compute the acoustic probability $P(x_1^T)$ in equation (2.11). Only a subset of strings is considered, such that the search space is greatly reduced. Besides, the time registration of the word hypothesized is relaxed to include also the words that intersect with the time interval [*s*,*t*], not only the words with exact start and end frame time *s* and *t*. After releasing the time registration, equation (2.11) is modified to:

$$P([w;s,t]\,|\,x_1^T)=\sum_{\substack{\forall M,[w;s,t]_1^M\\ \exists n,1\le n\le M\\ w=w_n\\ [s,t]\cap[s_n,t_n]\neq\phi}}\frac{\prod_{m=1}^{M}P(x_{s_m}^{t_m}\,|\,w_m)\cdot P(w_m\,|\,w_1^{m-1})}{P(x_1^T)} \qquad (2.12)$$

Besides, acoustic and language models are re-weighted since some of the assumptions in HMM based continuous ASR are not quite accurate or obviously wrong [11]. They are:

a) The assumption that neighbouring acoustic observations are statistically independent when acoustic likelihoods are computed;
b) The dynamic range between acoustic models and language models is different. The probability density function (pdf) used in the continuous Gaussian mixture based acoustic model with unlimited range, while N-gram language model with probability between 0 and 1;
c) The frequency of computation is different, acoustic model likelihood is computed at every frame interval, while language model likelihood is computed only once at every hypothesized word boundary;

Therefore, the acoustic model likelihood and language model likelihood are scaled by α and β respectively, and equation (2.12) becomes:

$$P([w;s,t]\,|\,x_1^T)=\sum_{\substack{\forall M,[w;s,t]_1^M\\ \exists n,1\le n\le M\\ w=w_n\\ [s,t]\cap[s_n,t_n]\neq\phi}}\frac{\prod_{m=1}^{M}P^{\alpha}(x_{s_m}^{t_m}\,|\,w_m)\cdot P^{\beta}(w_m\,|\,w_1^{m-1})}{P(x_1^T)} \qquad (2.13)$$

GWPP will be applied in this research, in order to find the optimum word sequence in Cantonese and English for the code-mixing utterances.

2.4 Characteristics of Cantonese

Cantonese is one of the popular Chinese dialects, which belongs to the Sino-Tibetan language family. Cantonese is spoken by about 55 millions people all over the world [15], and there are several sub-groups in Cantonese, and we refer Cantonese to Guangfu in this thesis. There are about 5,000 commonly used Cantonese characters, some of them are unique to Cantonese but not standard written Chinese or Mandarin.

2.4.1 Cantonese Phonology

Cantonese is a monosyllabic and tonal language, which means every written character in Cantonese corresponds to one syllable sound. The tone of a syllable carries lexical meaning, such that the variation in the pitch profile of a syllable changes it to another word. There are 6 tones in Cantonese, such that there are about 1,800 tonal syllables. If the tone is ignored, the number of syllable reduces to about 665.

A Cantonese syllable can be further divided into two components, namely an Initial and a Final. There are 19 Initials and 53 Finals in Cantonese, and each Final can be further decomposed to a nucleus and a coda. Both the Initial and coda are optional consonants, and the nucleus is vowel. From the Table 2-2, we can see that tones can be separated from the base syllable, which means they can be recognized individually. Consequently, we will only focus on recognition of the base syllable but not the tonal syllable.

Tonal Syllable (1,800)			
Base Syllable (665)			Tone (6)
Initial (19)	Final (53)		
[Onset] (19) ~ consonant	Nucleus (20) ~vowel	[Coda] (6) ~ consonant	

Table 2-2 Syllable structure of Cantonese

The syllable structure in Cantonese is simple that the possible combinations of sounds are severely restricted. There are no consecutive consonants in syllables, and the syllables typically have the form $(C_1)V(C_2)$, where C is consonant and V is vowel (Monophthong or Diphthong).

Initial Consonants

There are 19 initial consonants in Cantonese, which is optional but must occur at the beginning of a syllable. Linguistic Society of Hong Kong transcription scheme (LSHK) is commonly used for labeling Cantonese syllables. The initial consonants in LSHK and the corresponding International Phonetic Alphabet (IPA) symbols are shown in Table 2-3.

LSHK	IPA
b	p
d	t
g	k
gw	k^w
p	p^h
t	t^h
k	k^h

LSHK	IPA
kw	kw^h
l	l
w	w
j	j
m	m
n	n
ng	ŋ

LSHK	IPA
s	s / ʃ
f	f
h	h
z	ts / tʃ
c	ts^h / tʃ^h

Table 2-3 Initial Consonants in Cantonese

The Cantonese stops /b/ and /p/ are distinguished by aspiration, i.e. whether or not a burst of air is emitted immediately after oral release in the process of articulation [16]. The labiovelar consonants /gw/ and /kw/ are coarticulated stops, that means the velar sound /g/ or /k/ is simultaneously with the bilabial /w/. However, many Hong Kong people fail to articulate the /g/ or /k/ sound simultaneously with /w/ when the velar sound are followed by /o/ or /u/, such that there is a tendency to simplify /gw/ and /kw/ to /g/ and /k/ respectively. Examples for such simplification includes /gwok3/ (國) becomes /gok3/ (各) and /kwong3/ (礦) becomes /kong3/ (抗). The pronunciation of the initials /s/, /z/ and /c/ may be affected by the following vowel. When they are followed by the high front rounded vowel /yu/, the palatalization of /s/, /z/ and /c/ tends to change from [s], [ts], [ts^h] to [ʃ], [tʃ] and [tʃ^h] respectively. One of the examples is that the IPA of the syllable /syu/ is [ʃy] while the IPA for the syllable /si/ is [si].

Unreleased Consonants

The consonant stops in the final position of Cantonese syllables are unreleased [16]. No air is released for the final stop consonants /p/, /t/ and /k/. Some speakers do not distinguish between /t/ and /k/ in words like /baak3/ (百) and /hok6/ (學), such that they are pronounced as /baat3/ (八) and /hot6/ (no Chinese character for this syllable) respectively. Since the stop consonants are unreleased, the sound of them is mainly reflected in the vowel instead of the consonants themselves.

Nasal Consonants

The three nasals /m/, /n/ and /ng/ can appear in initial and final position of syllables. However, the initial /ng/ is usually pronounced as null initial, i.e. the initial /ng/ is not pronounced, especially for younger speakers [16]. The most common example is the character 我 is pronounced as /o3/ instead of /ngo3/. Similarly, the initial /n/ are usually pronounced as /l/, such that 你 is pronounced as /lei3/ instead of /nei3/. If the /ng/ sound appears at the final position, they are usually pronounced as /n/ instead. The example that includes several phone changes in nasal is the name of the local bank (Heng Sang Bank) "恆生銀行" /hang4 sang1 ngan4 hong4/, which is commonly pronounced as /han4 san1 an4 hon4/ (痕 - 身- no character for the syllable an4 - 寒).

Vowels

There are 9 vowels [17] and 11 diphthongs [18] (sounds consisting of a combination of two vowel sounds) in Cantonese. All the vowels are shown in Table 2-4.

Monophthong vowel	
LSHK	IPA
aa	a
i	i, ɪ
u	u, ʊ
e	ɛ
o	ɔ
yu	y
oe	œ
a	ɐ
eo	ɵ

Diphthong vowel	
LSHK	IPA
ai	ɐi
aai	ai
au	ɐu
aau	au
ei	ei
eu	ɛu
eoi	ɵy
iu	iu
ou	ou
oi	ɔi
ui	ui

Table 2-4 Monophthong and diphthong vowels in Cantonese

The vowel /i/ is usually pronounced as [i] unless the final consonant is /ng/ or /k/. When the final consonant is /ng/ or /k/, /i/ are pronounced as [ɪ]. Similarly, the pronunciation of the vowel /u/ also depends on the final consonant. It is pronounced as [u] unless the final consonant is /ng/ or /k/, such that it become [ʊ] instead.

2.4.2 Variation and Change in Pronunciation

There is variation in many aspects of pronunciation of Cantonese. Cantonese is dialect such that there is a lack of widely recognized standard form of the language, in particular of a phonetically based written form, to limit the variation [16]. Some of the sounds are confused due to hypercorrection that speakers using what they perceive to be correct or prestige pronunciation. There are patterns for the sound changes, but its happening appears to be random as they affect individual words differently. The main sound changes are summarized in Table 2-5.

Position	Original sound	Changed sound
initial	gw	g
initial	kw	k
initial	k	h
initial	n	l
initial	ng	null initial
final	k	t
final	ng	n
single consonant	ng (五)	m (唔)

Table 2-5 Main sound changes in Cantonese (labeled in LSHK)

The change in pronunciation causes degradation in speech recognition accuracy. If the changes occur in training data, observations from other sounds will be used for training the particular sound. If the changes occur in testing data, the words may be recognized as words in other syllables.

2.4.3 Syllables and Characters in Cantonese

The basic unit of written Cantonese is Chinese character, while spoken Cantonese is syllable based. Cantonese is monosyllabic, such that every character is pronounced as a syllable. However, Cantonese is homophonic and homographic, such that the mapping between a character and its pronunciation is not always one-to-one. Many Chinese characters share the same syllable. On the other hand, the same character will be pronounced as different syllables when it has different meaning or context. Some examples on homophones and homographs in Cantonese are shown in Table 2-6 as follows.

Homophones	/si1/	司、思、斯、屍、私、施、詩…
Homographs	重	/zung6/, /cung4/, /cung5/

Table 2-6 Examples on homophones and homographs in Cantonese.

2.4.4 Spoken Cantonese vs. Written Chinese

Spoken Cantonese is very different from standard written Chinese. The main difference is lexicon that different words may be chosen for the same meaning. The difference can be observed from the statistics on the most common written Chinese [19] and spoken Cantonese listed in Table 2-7.

	Written Chinese	Spoken Cantonese
1	的	嘅
2	一	一*
3	是	係
4	不	人*
5	人	佢
6	有	唔
7	在	有*
8	了	好*
9	我	個*
10	中	大*
11	大	都
12	個	會*
13	這	我*
14	會	就
15	他	到
16	爲	不*
17	上	話
18	來	喺
19	以	同
20	時	日

(*: exist in Top-20 of both languages)

Table 2-7 Top-20 characters in written Chinese and Spoken Cantonese

Moreover, some of the spoken Cantonese does not have standard written form that people use different characters with the syllable instead [20]. These non-standard characters lead to problems when text data is collected for building language models. Different characters are used in for same word, such that tagging may be required to cluster them. Table 2-8 includes some commonly used non-standard Cantonese characters.

Cantonese words	Equivalent in written Chinese
尋日、噚日、擒日、琴日	昨天
唔洗、唔使、唔駛	不用
仲有、重有	還有

Table 2-8 Cantonese characters without standard written form

Furthermore, there are many colloquial words in Cantonese that do not have a written form, i.e. there is no Chinese character pronounced as that particular syllable. People sometimes use English words with similar sound instead. Therefore, we should distinguish these Cantonese words from English code-switch words.

Examples of the colloquial are listed in Table 2-9 [21].

Pronunciation	Usual written form
/wu1 we5/	烏 where
/ping1 ling1 paang1 laang1/	平 ling 彭蘭, ping 拎攀躝
/soe4 waat6 tai1/	sir 滑梯
/npap1 dap1/	up 耷, 噏耷

Table 2-9 Examples of Cantonese colloquial without written form

2.5 Characteristics of English

English is one of the Top-3 popular languages that spoken by about 309 million people in the world. It is also the second most common first language in the world. It is a West Germanic language that belongs to the Indo-European language family [15]. The vocabulary of English is vast, but there is no Academy to define officially accepted words. The Oxford English Dictionary (2nd edition) includes over 500,000 headwords, following a rather inclusive policy that it embraces not only the standard language of literature and conversation, whether current at the moment, or obsolete, or archaic, but also the main technical vocabulary, and a large measure of dialectal usage and slang [22].

2.5.1 English Phonology

The syllable structure of English is more complicated than Cantonese. Cantonese only has four categories of syllables: V, CV, VC, CVC, while the syllable structures of English can various from a single vowel (V) to syllables including as many as seven consonants [23]. However, for most English syllables (> 80% in TIMIT), the syllable structure is of the canonical form (C)V(C), where the consonant is optional [24].

In isolated English words, there is a fixed position of stressed syllables. The stressed syllables are characterized by power level, pitch, duration and vowel quality [25]. Two words can differ only by the position of the stress, and it usually occurs on the noun/verb pairs. The examples of noun/verb pairs with different stress position are shown in Table 2-10.

Word	Noun	Verb
record	/r eh' k er d/ [‘rɛkɔːd]	/r ah k ao' r d/ [rɪ'kɔːd]
present	/p r eh' z ah n t/ [‘prɛzənt]	/p r iy z eh' n t/ [prɪ'zɛnt]
produce	/p r ow' d uw s/ [‘prɒdjuːs]	/p r ah d uw' s/ [prəʊ'djuːs]

(The pronunciations are labeled in ARPABET from CMU pronunciation dictionary [26] and IPA from Oxford English Dictionary [22])

Table 2-10 Examples of English noun/verb pairs with different stress position

2.5.2 English with Cantonese Accents

If the Cantonese accents are adopted to English, the syllable structure of the English words will usually be modified to the (C)V(C) structure of Cantonese syllables [27][28]. The way of changes in syllable structures can be classified as the following three categories:

<u>Softening or Dropping the Second Consonants in a CC Sequence</u>

Since there are no consecutive consonants in Cantonese syllable, the $C_1C_2VC_3$ syllables in English are often changed to C_1VC_3. The second consonant C_2 is either softened or dropped [27]. This usually happens for the /r/ sound and /l/ but seldom if C_1 is the fricative /s/. Examples include the word “plan” /p l ae n/ becomes /p ae n/ and “fresh” /f r eh sh/ becomes /f eh sh/. However, for words with C_1=/s/ such as “stay” /s t ey/ and “slide” /s l ay d/, the C_2 usually does not change.

<u>Softening or Dropping the Final Stop Consonant</u>

As mentioned in the previous section, the final stop consonants in Cantonese are all unreleased. There are only six final consonants in Cantonese; three of them are stop consonants - /p/, /t/ and /k/. For English, there are six stop consonants - /p/, /b/, /t/, /d/, /k/, /g/. Cantonese speakers usually adopt these stop consonants in English to unreleased sound, such that the stop consonants are either softened or dropped [27]. Examples: “ask” /ae s k/ becomes /ae s/, “meet” /m iy t/ becomes /m (it)/. (The sound /it/ is the one in LSHK)

Adapting a Monosyllabic Word with Fricative Endings to Produce a Disyllabic

When the consecutive consonants occur after the vowel, such as $C_1V_1C_2C_3$, the third consonants are either dropped (for final stop consonants, such that the syllable structure becomes $C_1V_1C_2$) or another vowel will be added after the third consonants (for monosyllabic words with final fricative consonants, such that the syllable structure becomes $C_1V_1C_2$ C_3V_2) [27]. This happens to English words that commonly used for daily life, such as "notes" /n ow t s/ ⇒ /n ow t s iy/, "F" /eh f/ ⇒ /eh f uh/.

Apart from syllable structure, the Cantonese accent can also affect the pronunciation of English words by sound change, i.e. the sounds in English are adapted to similar sounds in Cantonese. It mainly happens to sounds that appear in English but not in Cantonese, details will be described in the next chapter.

2.6 References

[1] J. W. Picoone, "Signal Modeling Techniques in Speech Recognition", in *Proc. of the IEEE*, Vol. 81, No, 9, pp. 1215-1247, September 1993

[2] H. Hermansky, "Perceptual Linear Predictive Analysis of Speech", *Journal of Acoustics Society America*, Vol. 87, No, 4, pp. 1738-1752, April 1990

[3] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, *Spoken Language Processing, a Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, 2001

[4] Lawrence Rabiner, Bing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1999

[5] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, Dec. 2001

[6] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", in *Readings in Speech Recognition (A. Waibel and K.F. Lee eds.)*, pp.267 – 296, 1990.

[7] Slava M. Katz, "Estimation of Probabilities from Sparse Data for Language Model Component of a Speech Recognizer", *IEEE Transactions on Acoustics,*

Speech and Signal Processing, Vol. ASSP-35, pp.400-401, March 1987

[8] Ian H. Witten and Timothy C. Bell, "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression", *IEEE Transactions on Information Theory*, Vol. 37, No. 4, July 1991

[9] H. Ney, U. Essen and R. Kneser, "On Structuring Probabilistic Dependencies in Stochastic Language Modeling", *Computer Speech and Language*, Vol. 8, pp. 1-28, 1994

[10] H. Y. Lenug, C. Y. Choy and H. C. Leung, "Characteristics of Chinese Language Models for Large Vocabulary Telephone Speech", in *Proc. of Eurospeech 1999*, Vol. 4, pp. 1775-1778, 1999

[11] Frank K. Soong, Wai-Kit Lo, Satoshi Nakamura, "Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words", Special Workshop in Maui, 2004

[12] Lidia Mangu, Eric Brill, Andreas Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks", *Computer Speech and Language,* Vol. 14, pp. 373-400, 2000

[13] R. O. Duda, P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973

[14] Levenshtein, V. I., *Binary codes capable of correcting spurious insertions and deletions of ones*, (original in Russian: Russian Problemy Peredachi Informatsii 1), pp. 12–25., January 1965

[15] Raymond G. Gordon, Jr., *Ethnologue: Languages of the World*, 15th Edition, 2005

[16] Stephen Mattews and Virginia Yip, *Cantonese: a Comprehensive Grammar*, Routledge, London; New York, 1994

[17] 香港語言學學會粵語拼音字表編寫小組,《粵語拼音字表》，香港語言學學會, 第 1 版., 1997

[18] International Phonetic Association, *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*, Cambridge University Press, 1999

[19] Ho Hsiu Hwang, "Basic Character Frequency Statistical Tables, 80s to 90s, Hong Kong", *Hong Kong, Mainland China & Taiwan: Chinese Character Frequency – A Trans-Regional, Diachronic Survey*, 2001

[20] Don Snow, *Cantonese as Written Language: the Growth of a Written Chinese Vernacular*, Hong Kong University Press, Hong Kong, 2004

[21] 彭志銘,《次文化語言：香港新方言概論》, 次文化有限公司, 香港, 1994

[22] John Simpson, Edmund Weiner, *Oxford English Dictionary*, third edition, Clarendon Press, 1989

[23] Kazunori Imoto, Yasushi Tsubota, Antoine Raux, Tastsuya Kawahara, Masatake Dantsuji, "Modeling and Automatic Detection of English Sentence Stress for Computer-Assisted English Prosody Learning System", in *Proc. of ICSLP 2002*, pp. 749-752,

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

[25] M. Sugito, *English spoken by Japanese*, chapter 1-4, Izumishoin, 1996

[26] The CMU Pronouncing Dictionary v0.6, The Carnegia Mellon University, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

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

[28] F. Grosjean, "Exploring the recognition of guest words in bilingual speech", *Language and Cognitive Processes*, Vol. 3, 233–274, 1988

Chapter 3

Code-mixing and Code-switching Speech Recognition

3.1 Introduction

Code-mixing between Cantonese and English is very common in Hong Kong, where people use Cantonese in their ordinary conversation, with a mixed of English words, phrases or utterances. Most existing speech recognizers are monolingual, which do not support the automatic speech recognition (ASR) of two different languages at the same time.

In this chapter, we will first give the definition of monolingual, multilingual and code-mixing speech. Details on code-mixing in Hong Kong will be discussed. Then the difficulties of constructing a code-mixing ASR will be analyzed.

3.2 Definition

3.2.1 Monolingual Speech Recognition

For the situation that the speech input contains only one particular language and the recognizer is capable to perform ASR of that language, it is referred to as monolingual speech recognition. The pronunciation dictionary in this case includes phone sequence of one language only, and also for the language model.

3.2.2 Multilingual Speech Recognition

On the other hand, there are situations that the speech input may come from a selection of different languages. For example, a tourist enquiry system may have users from different countries who will use their own mother language. The user can choose to use any one of the predefined supporting languages in their conversation, but they are expected to use the same language throughout their enquiry [1]. Most of these ASR systems will perform language identification [2] and users need not inform the system which language is being used. Some systems run parallel recognizers for language identification [3-6], while the others may apply universal phone set for all languages [3][7].

3.2.3 Code-mixing and Code-switching

According to John Gumperz [8], the definition of code-switching is "the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-system". In Hong Kong, code-switching tends to be intra-sentential and switching involving linguistic units above the clause level is rare, hence the preference for the term "code-mixing" in many studies [9]. Although there is admittedly a grey area between (inter-sentential) code-switching and (intra-sentential) code-mixing, "code-mixing" is a more preferable term to describe the typical language behaviour of the average Hong Kong bilinguals [10]. Table 3-1 lists some examples on code-mixing in several languages.

Involved languages	Example
Spanish / English	I started going like this. Y luego decía (and then he said), look at the smoke coming out of my fingers [4]
French / English	Va chercher Marc (go and fetch Marc) and bribe him avec un chocolat chaud (with a hot chocolate) with cream on top [5]
Russian / French	Imela une femme de chambre. (She had a chambermaid) [6]
Cantonese / English	質素唔算太差，不過鬼馬處係影相時會發出「咔嚓」一聲，都幾鬼相機 Feel。 (The quality is not bad, but the interesting part is that it produces shutter sound when taking photos, which make it has the feeling of camera) [7]

Table 3-1 Examples on code-switching and code-mixing in several languages

Inter-sentential code-switching is similar to multilingual, but users can switch between different languages in different utterances. Two monolingual speech recognizers can be applied after the language identification system, or a single multilingual speech recognizer can be used instead.

Code-mixing is intra-sentential, which involves at least two languages in a single utterance. The language in majority is the primary language, or the matrix language, and the other language is secondary language, or the embedded language. For code-mixing between Cantonese and English, the grammar mainly depends on the matrix language (Cantonese), but sometimes word order may change due to the parts-of-speech (POS) of the embedded words. Language identification is difficult since the embedded words may be mono-syllabic, which is the same as the phonological structure of Cantonese. Several approaches will be discussed in the next chapter to tackle this problem.

3.3 Conversation in Hong Kong

Hong Kong was a British colony until 1 July 1997, and it was handed back to the People's Republic of China according to the Sino-British Joint Declaration signed in 1984. English was the only official language until 1974 [15], that's why English is widely used by the executive authorities, the legislature, the judiciary as well as the legal, professional and business sectors in Hong Kong [16]. According to the basic law, both Chinese and English are the official languages in Hong Kong now.

Over 98% of Hong Kong's population is Chinese, and most of them are Cantonese native speakers. Cantonese is mainly used in daily conversation, and English is taught in schools. Majority of the residents are bilingual in Cantonese and English, and the language choice depends on situation.

3.3.1 Language Choice of Hong Kong People

People in Hong Kong use different languages under different situation and domain. The details are discussed below.

Monolingual Cantonese

Cantonese is the primary language of most residents in Hong Kong. Cantonese is used in daily conversation, as well as most of the mass media such as TV and radio. Most Chinese families use Cantonese at home, some may use other Chinese dialects or English. In most primary schools, teachers use monolingual Cantonese as media of instruction for all subjects except the English lessons. Chinese is the media of instruction for most of the secondary schools since the 1998/99 schools year [17]. The language policy is to enable students to learn effectively, and mother-tongue teaching is encouraged.

Monolingual English

The objective of the government's education policy is to enhance students' biliterate (Chinese and English) and trilingual (Cantonese, Mandarin, English) abilities [17]. English is one of the major subjects in primary and secondary education, thus most people in Hong Kong are able to speak English. It is the media of instruction in some secondary schools and all the universities. However, monolingual English is seldom used in daily life between local people, as most of them are native Cantonese speakers. Monolingual English are used in situation that some of the speakers in the conversation do not understand Cantonese and English is the only language in common. Since the official language of most universities is English, professors usually use monolingual English for lectures, and students use it for formal presentation or oral exam.

Code-switching and Code-mixing

In code-switching or code-mixing, Cantonese is usually the matrix language. People intersperse English terminology in their conversations. There are many technical terms in professional communication, but not all of them have Cantonese equivalent. Besides, most documents of business sectors and government departments, as well as the textbooks, are in English. Using English terminology allows a parallelism between the written and spoken language [18].

Code-switching and code-mixing are regarded as informal, and thus people usually do not use it in formal speech or presentation. Besides, the elderly usually have lower English proficiency, so people tend to use monolingual Cantonese when they are speaking to their parents or the seniority.

Code-mixing is more commonly then code-switching in Hong Kong since the code-switch words involves are mainly terminologies instead of sentences. The frequency of code-mixing mainly depends on the conversation domain and the language proficiency of the speakers. According to previous research on code-mixing in Hong Kong, code-mixing is wide spread in the following six domains: 1) computer discourse; 2) business discourse; 3) fashion discourse; 4) food discourse; 5 showbiz discourse [19]. Code-mixing is commonly used in domains that have more interference with other cultures or languages. Many of the terms are new lexicon that does not appear in Cantonese, people translate it or just use the English term directly [9].

Code-mixing is common among young people. In order to show the importance of research on code-mixing ASR, we have collected text data from some online diary [20] to study the frequency of code-mixing. The statistics are shown in Table 3-2.

Index	Author Information		Monolingual Cantonese		Cantonese-English Code-mixing	
	Age	Gender	No. of utterances	Percentage	No. of utterances	Percentage
1	22	F	620	49.64%	629	50.36%
2	19	M	335	57.66%	246	42.34%
3	15	F	244	36.69%	421	63.31%
4	28	M	143	44.00%	182	56.00%
5	19	F	425	73.15%	156	26.85%
Total			1767	51.96%	1634	48.04%

Table 3-2 Frequency of code-mixing among young people in general domain

3.3.2 Reasons for Code-mixing in Hong Kong

In some situation, people are able to choose between monolingual Cantonese and Cantonese-English code-mixing. With the understanding on the reasons behind code-mixing in Hong Kong, appropriate speech materials can be collected to reflect the real situation.

Euphemism

One of the motivations for using English words is euphemism. When people found themselves in the unenviable position of speaking out those terms in Cantonese, they usually use English words instead. For most cases, the equivalents in both 'high' Cantonese and 'low' Cantonese exist, but people tends to use English words to avoid embarrassment. Examples: underwear, bra, toilet, washroom.

Specificity

The meaning of English words and their Cantonese equivalent may have some differences. When people wish to specify or generalize something, but fail to find the suitable lexicon in Cantonese, they will switch to English.

The English term is more specific – If we wish to make a reservation for which no money or deposit is required, we usually use the verb "book". The nearest equivalent in Cantonese is 訂, (訂位、訂造) cannot tell clearly whether deposit is required or not, so people usually prefer to use the more specific word "book".

The English term is more general – Many entertainers have multiple talents, they may be singer, actors, disc jockey, etc. The word "fans" is more general then the Cantonese terms "歌迷"、"影迷"、"球迷"。If people just wish to refer to the supporters of a particular person, they may use the more general word "fans" such that the reasons of supporting need not be mentioned [9].

Principle of Economy

Sometimes the English expression is shorter, thus requires less linguistic effort compared with its Cantonese equivalent. For example, university students usually use the English expression "grant and loan" instead of its Chinese equivalent (本地專上學生資助計劃, Local Student Finance Scheme). People also prefer the English abbreviations such as GPA (Grade Point Average, 學期平均積點)、LCD (Liquid Crystal Display, 液晶顯示)、FYP (Final Year Project, 畢業習作)。Moreover, people use the first one or two syllables only instead of the whole word for some English terms so as to further reduce linguistic effort, such as "con" (Contact Lens, 隱形眼鏡)、"mon" (Monitor, 顯示器)、"reg" (Register, 登記 / 註冊)。

Change of Attitude or Relationship

People have different code choice when relationship between participants or functions changes. 'Low' Cantonese and code-mixing are regarded as less formal and so if people wish to make the situation relax, they may use some English words in their Cantonese conversation.

3.3.3 How Does Code-mixing Occur?

The code-mixing in Hong Kong is usually intra-sentential, and the code-switch words are mainly nouns. Besides, there are many new words that compose of both Cantonese and English. Some of the English terms have been borrowed to Cantonese and form new Cantonese words. The details of this situation will also be discussed as we have to define clearly the difference between code-switch words and borrowing terms.

Words vs. fragment

According to previous research [9-10], a single English word surrounded by Cantonese is the most frequent form of Cantonese-English code-mixing. The switching is mainly intra-senentential, thus code-mixing is a better description on the situation in Hong Kong.

Word Class

According to the word class distribution, most of the English words involved are nouns, verbs, and adjectives. Table 3-3 is the word class distribution of the data collected by Brian Chan [21]:

	No. of utterances	Percentage
Noun-mixing	260	41.27%
Verb-mixing	148	23.49%
Adjective / Adverb-mixing	84	13.33%
Preposition / conjunction-mixing	11	1.75%
Fragment	127	20.16%

Table 3-3 Word class distribution of code-mixing utterances

In most situations, there is only one code-switch word in each utterance. According to the code-mixing text data collected from "The Open Diary" [20], the number of code-switch terms in each utterance is shown in Table 3-4.

No. of code-switch terms	No. of utterances	Percentage
1	398	63.99%
2	156	25.08%
3	57	9.16%
4	11	1.77%

Table 3-4 Number of code-switch terms per utterance

It is found that most code-mixing utterances (88%) include one to two code-switch terms. Therefore, our research will focus on utterances with one to two code-switch words.

Compromise forms

Some code-switch terms are compromised form of Cantonese and English. The English words may be adopted to the Cantonese grammar. Table 3-5 includes some example of these forms:

Compromised form	Meaning
mind 唔 mind	Mind or not?
un 唔 understand	Understand or not?
On 返 line	Online again
阿 sir	Sir

Table 3-5 Examples on compromised forms of Cantonese and English

True Code-switch and Borrowing

According to the pronunciation of the code-switch words, code-mixing can be classified into true code-switch and borrowing.

True Code-switch

If the pronunciation of the code-switch word is those in the embedded language, it is regarded as true code-switch. This usually happens in bilingual speakers with high language proficiency in the embedded language.

Borrowing

If the pronunciation of the code-switch word is adapted to the matrix language, it is regarded as borrowing. Borrowing is commonly used by most local people in Hong Kong. For some common borrowing terms, the English words can be written by Cantonese characters with similar sounds, such as "快勞" (file), "吉士" (guts), "柯打" (order), etc. However, it is difficult to distinguish whether these words should be treated as Cantonese or English, since some of the words are so common that people don't realize that they come from English. Table 3-6 lists some common borrowing words [22].

Some common borrowing words	
Borrowing term in Cantonese	English words
巴士	bus
的士	taxi
結他	guitar
爹哋	daddy
甫士	pose
餐士	chance
Terms that people may regard them as Cantonese	
杯葛	boycott
嘉年華	carnival
啤(酒)	beer
啤梨	pear
車里子	cherry
摩打	motor

Table 3-6 Examples on borrowing words

To simplify the case, if the borrowing term can be written in Cantonese, and it can be found in Cantonese text database, it will be regarded as Cantonese in this research. On the other hand, if there is no well-accepted Cantonese written form, the borrowing term will be regarded as English.

3.4 Difficulties for Code-mixing – Specific to Cantonese-English

Difficulties on code-mixing ASR will be discussed in this section. Code-mixing involves two languages, and the phoneme inventories as well as phonological structures are usually different. Accents of the primary language may be involved in the code-switch words, such that the pronunciation dictionary will be inaccurate. Some of the code-switch words are pronounced with syllables in the primary language and can be written in the primary language, which lead to grey area in classifying them as borrowing words or code-switch words. The lexicons and grammar for monolingual and code-mixing speech are not exactly the same. New lexicons unique to code-mixing may be generated and these words should also be included in the pronunciation dictionary. Lack of appropriate speech corpus for training and testing also cause difficulties to code-mixing speech recognition.

In this section, difficulties for code-mixing speech recognition will be discussed, and the details will be focus on Cantonese-English code-mixing.

3.4.1 Phonetic Differences

Although words in the embedded language are in minority of the code-mixing utterances, the phonemes from the embedded language cannot be neglected as the phoneme inventory is different for every language. For code-mixing LVCSR, the problem becomes more complex that we have to consider the problems on context-dependent acoustic models, accents, as well as differences in phonology structures.

Missing Phone

Every language has its own phoneme inventory so as to contrast meaning. Some phones appear in several languages, while the other does not. For example, English involves 24 consonants, 11 monophthongs and 3 diphthong. In Cantonese, there are 22 consonants, 11 monophthongs and 11 diphthongs [23]. If we only consider the phones in the matrix language, there will be missing phones since the phones in the matrix language and embedded language are not identical. Figure 3-1 shows the phoneme inventories of Cantonese and English. Cantonese phones are in the left circle, and English phones are in the right circle. The phones that exist in both Cantonese and English are in the intersection of the two circles. There are 25 phones unique to Cantonese, 19 phones unique to English, and 19 phones exist in both languages.

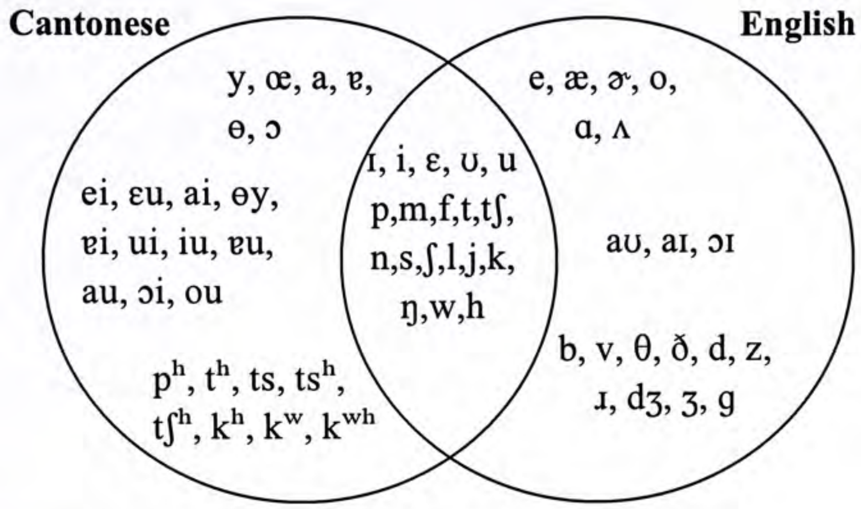


Figure 3-1 Phoneme inventories of Cantonese and English

Allophone

An allophone is one of several similar phones that belong to the same phoneme in a language. The mapping between phonemes and phones is not one-to-one but rather context -specific. For example, the phoneme /p/ in /pin/ and /spin/ are allophones in the English language. The /p/ in /spin/ is unaspirated, which sounds a little more like the /b/ of English. The preceding /s/ is the usual context for the unaspirated allophone.

Speakers of a particular language perceive a phoneme as a single distinctive sound in that language. They may be allophones which are considered as members of one single phoneme, in which they are actually different phones.

Similar phone

Some phonemes occur in more than one language, but the qualities of them are a bit different. They are similar but not exactly the same. In code-switching, the speaker may pronounce the code-switch words with accent of the matrix language. The phones in the embedded language will be pronounced as the similar phones in the matrix language. For example, the English word "back" /b ae k/ will be pronounced as /p^h ae k/ by most Cantonese speakers. The reason is that in English, both [p] and [b] are bilabial plosive, and the difference between them is voiced vs. unvoiced. However, in Cantonese, there is no [b] sound, but only [p^h]. The difference between [p] and [p^h] is aspirated vs. unaspirated.

People learn their mother language in their early ages. When they adapt to a new language, they usually map the phones in the new language with their mother language for the ease of memory. Both of [b] and $[p^h]$ are phoneme /b/ in the two languages, so Cantonese speakers usually mix up these two phones and regard them as identical if they are not well-trained.

Although some of the phonemes are identical or similar in the two languages, the distance between them is unknown; hence using speech recognizer with phoneme inventory only from the matrix language is inadequate. Both the Cantonese and English phoneme inventory should be considered, and the phonemes with high degree of similarity can be clustered so as to reduce the number of acoustic models involved. The phoneme inventory of Cantonese and English are listed below, such that we can have a clear picture on the differences and similarities on the phonemes. Figure 3-2 and Figure 3-3 includes both the monophthong and diphthong vowels in Cantonese and English. The consonants of Cantonese and English are also listed in Table 3-7 and Table 3-8 respectively.

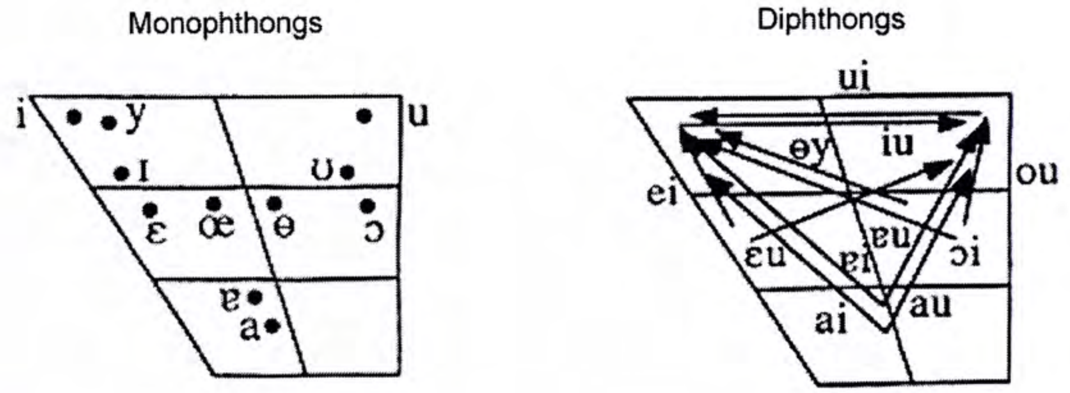


Figure 3-2 IPA Table, Vowel (Cantonese)

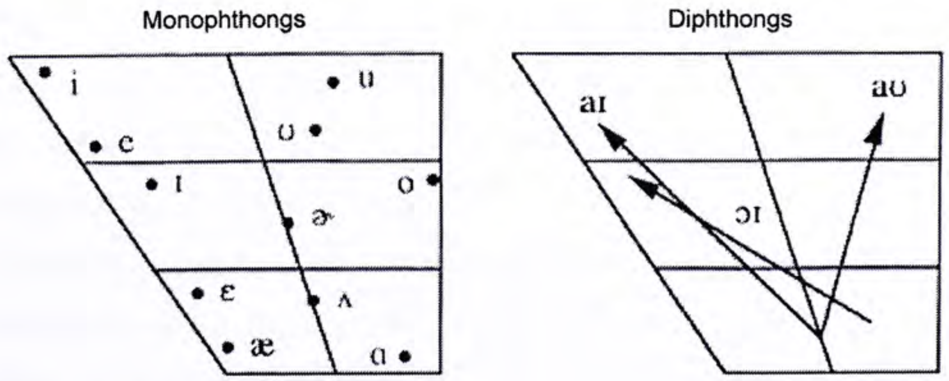


Figure 3-3 IPA Table, Vowel (English)

	Bilabial	Labiodental	Dental	Alveolar	Post-alveolar	Palatal	Velar	Labial-Velar	Glottal
Plossive	p pʰ			t tʰ			k kʰ	kʷ kʷʰ	
Affricate				ts tsʰ	tʃ tʃʰ				
Nasal	m			n			ŋ		
Fricative		f		s	ʃ				h
Approximant						j		w	
Lateral Approximant			l						

Table 3-7 IPA Table, Consonant (Cantonese)

	Bilabial	Labiodental	Dental	Alveolar	Post-alveolar	Palatal	Velar	Labial-Velar	Glottal
Plossive	p b			t d			k g		
Affricate					tʃ dʒ				
Nasal	m			n			ŋ		
Fricative		f v	θ ð	s z	ʃ ʒ				h
Approximant				ɹ		j	w		
Lateral Approximant				l					

Table 3-8 IPA Table, Consonant (English)

3.4.2 Phonology difference

Apart from the difference at the phonetic level, there is also a great difference in phonological level among different languages. Cantonese is a Sino-Tibetan language, which is monosyllabic in nature. It has a general syllable structure C_1VC_2, where C_1 and C_2 are optional consonants and V is either monophthong or diphthong. Hence, the all Cantonese syllables are of the 4 canonical forms V, CV, CVC or VC [24].

On the other hand, English is an Indo-European language and the phonological structure is much more complicated than Cantonese. In English discourse, over 80% of the syllables are of the canonical form of Cantonese, and the remaining are C, CC, CCV, VCC, CCCV, CCCVCC, etc [24]. Consecutive consonants may occur which are not found in Cantonese.

In continuous speech, the pronunciation of a sound depends on the neighbouring sounds due to co-articulation. The acoustic models are "context-dependent" if the effect of neighbouring sound units are also considered, or else, they will be called "context-independent". When there exist large amount of training data, context-dependent acoustic models perform better than the context-independent one. If the amount of Cantonese data and English data is different in large proportion, the consecutive consonants context-dependent models may be under-trained, such that the overall recognition accuracy may be degraded.

3.4.3 Accent and Borrowing

Accent differences introduce degradation into speech recognition accuracy since the phone sequence is greatly different from those in the pronunciation dictionary. The effect of accent is serious especially in code-switching since it may mislead the recognizer that it remains with the primary language. For example, people sometimes pronounce the English word “file” /f ay l/ as /f ay l ow/, which is similar the Cantonese character 快 /faai3/ and 佬 /lou2/. It is difficult to justify whether it is Cantonese or English, as it is actually an English word, but borrow the pronunciation in Cantonese. This phenomenon is called “lexical borrowing” [21].

3.4.4 Lexicon and Grammar

In code-mixing utterances, the grammar mainly based on the matrix language, such that the code-switch words seem just to replace its Cantonese equivalent. However, it is not true for all cases. The code-switch words sometimes duplicate the meaning which has already been mentioned by the Cantonese words in the same utterance. This phenomenon usually occurs when the code-switch words are related to time, for example, people may say “before 星期五之前”, the code-switch word “before” have the same meaning with the Cantonese word “之前”.

Moreover, the verbs in English have tense and there are plural or singular forms for nouns. When code-mixing occurs, the English verbs are usually in present tense. For nouns, it depends on the habit of the speaker, but usually in singular form.

There will also be some new lexicons that do not appear in traditional English speech, but unique to code-mixing. Some of these words exist in English, but have a totally different meaning [25]. For example, there are abbreviations for some English words which have more than one syllable, such as the word “mon” (monitor), “con” (contact lens), “reg” (register), etc.

3.4.5 Lack of Appropriate Speech Corpus

There is no existing Cantonese-English code-mixing corpus, none for text data or speech data. Code-mixing text data is necessary in order to facilitate analysis of the phenomenon as well as to build up language model. Speech data, on the other hand, is necessary for evaluating the performance of the code-mixing speech recognizer. Large amount of both text data and speech data are collected for this research and the details will be discussed in chapter 4 and chapter 5.

3.5 References

[1] Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, M. Woszczyna, “Multilinguality in speech and spoken language systems”, in *Proc. of the IEEE,* Vol. 88, Issue 8, pp. 1297-1313, 2000

[2] N. Udhyakumar, R. Swaminathan and SK Ramakrishnan, “Multilingual speech recognition for information retrieval in Indian context”, in *Proc. of the Student Research Workshop*, HLT/NAACL, Boston, USA, 2004

[3] Eddie Wong and Sridha Sridharan, “Three approaches to multilingual phone recognition”, in *Proc. of ICASSP 2003*, Vol. 1, pp. 44-47, April 2003

[4] T. Nagarajan and H. A. Murthy, “Language identification using parallel syllable-like unit recognition”, in *Proc. of ICASSP 2004*, Vol. 1, pp. 401-404, May 2004

[5] K. V Sai Jayram, V. Ramasubramanian and T. V. Sreenivas, “Language identification using parallel sub-word recognition”, in *Proc. of ICASSP 2003*, Vol. 1, pp. 32-35, April 2003

[6] Shizhen Wang, Jia Liu and Runsheng Liu, “Language identification using discriminative weighted language models”, in *Proc. of ISCSLP 2004*, pp. 53-56, Hong Kong, Dec 2004

[7] Zhirong Wang, Umut Topkara, Tanja Schultz, Alex Waibel, “Towards Universal Speech Recognition”, in *Proc. of ICMI 2002*, pp. 247-252, 2002

[8] John Gumperz, *Discourse Strategies*, p.59, Cambridge University Press, 1982

[9] David C. S. Li, “Cantonese-English code-switching research in Hong Kong: a Y2K review”, *World Englishes*, Vol. 19, No.3, p.305-322, Blackwell Publishers

Ltd., 2000

[10] A. Tse, "Some observations on code-switching between Cantonese and English in Hong Kong", *Working Papers in Languages and Linguistics*, Vol. 4, p.101-108, Department of Chinese, Translation and Linguistics, City Polytechnic of Hong Kong, 1992

[11] Valdés Fallis G, J Amastae, L Elías-Olivares, *Spanish in the United States*, pp. 209-229, Cambridge University Press, 1982

[12] Grosjean F, *Life with Two Languages*, Harvard University Press, Cambridge, MA, 1982

[13] Vivian Cook, *Codeswitching By Secondary Language Users*, http://homepage.ntlworld.com/vivian.c/SLA/codeswitching.htm

[14] "細架又細價 NEC N190$1,480 登場", 電腦/手機, *《蘋果日報》 (Apple Daily)*, Hong Kong, 6 June 2005

[15] Alberto Costa, "The British Experience: the Cases of Malaysia, Singapore and Hong Kong", *Contribution to Defining a Policy for Macau's Law in the Light of Other European-Style Experiences in the Region*, http://www.rjmacau.com/english/rjm1996n3/ac-mary/hk.html

[16] *The Basic Law of the Hong Kong Special Administrative Region of the People's Republic of China*, Chapter I, Article 9, Hong Kong

[17] Education and Manpower Bureau, The Government of the Hong Kong Special Administrative Region, Kindergarten, Primary and Secondary Education, http://www.ed.gov.hk/index.aspx?nodeID=2063&langno=1

[18] Bertha Du-Babcock, Richard D babcock, Patrick Ng, Rachel Lai, *A Comparison of Use of L1 and L2 in Small-Group Business Decision-Making Meetings*, Department of English, Research Monograph No. 6, City University of Hong Kong, 1995

[19] David C. S. Li, *Issues in Bilingualism and Biculturalism: a Hong Kong Case Study*, Peter Lang, New York, 1996

[20] The Open Diary, http://opendiary.com

[21] Brian H. S. Chan, *Code-mixing in Hong Kong Cantonese-English Bilinguals: Constraints and Processes*, M.A. in English Language Thesis, The Chinese University of Hong Kong, 1992

[22] Mimi Chan, Helen Kwok, *A Study of Lexical Borrowing from English in Hong*

Kong Chinese, Centre of Asian Studies, University of Hong Kong, Hong Kong, 1990

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

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

[25] John Simpson, Edmund Weiner, *Oxford English Dictionary*, third edition, Clarendon Press, 1989

Chapter 4

Data Collection

Since there is no existing Cantonese-English code-mixing speech database, the CUMIX speech corpus [1] is therefore developed for this research. In order to compare the performance of the speech recognizer on code-mixing data and monolingual data, baseline system is necessary.

4.1 Data Collection

A Cantonese-English code-mixing speech database, called, CUMIX is designed, complied and validated in the course of this study. The purpose of contracting CUMIX is to provide speech data for the training of Cantonese accented English acoustic models, and to evaluate the performance of the Cantonese-English code-mixing speech recognition system. The data can also be used to make a through study on automatic language identification within code-mixing utterances.

4.1.1 Corpus Design

The speech data is divided to two main types: 1) training data; and 2) testing data, details are as follows:

Training Data

The training set includes four types of speech data: 1) Cantonese-English code-mixing utterances, 2) monolingual English words, 3) English numbers and 4) English alphabets. Each code-mixing utterance includes one English segment only. Table 4-1 gives the summary of the training data set:

Type of speech data	Number of speech files per speaker
Cantonese-English code-mixing utterances	200
Monolingual English words	100
English numbers	1 (one to nine, zero)
English alphabets	4 (ABCDEFG / HIJLMN / OPQRST / UVWXYZ)

Table 4-1 Training data in the CUMIX speech corpus

The monolingual English words are those commonly used in code-mixing. Large proportion of the speech data are Cantonese-English code-mixing, so that the pronunciation of the English words encompass more Cantonese accents. It is because when people uttering with code-mixing speech, they usually adapt the code-switch words in embedded language to their mother tongue. However, if they are speaking in monolingual English, they will try to speak every word clearly, so the pronunciation of the words will be different from the code-mixing one. For example, when people speak the word "notes", if it is in an English sentence, it will be pronounced as /n ow t s/. However, if it is in code-mixing, they may adapt it to disyllabic form, such that the pronunciation become /n ow t s iy/. This type of adaptation occurs in code-mixing only but not in monolingual English, even for the same speaker. Hence, English words with different amount of accents can be obtained and we can train Cantonese accented English phone models with these code-mixing data.

The speech data is recorded by 20 male and 20 female speakers, and there are only 8000 code-mixing utterances in total. The Cantonese data is inadequate for training good Cantonese acoustic models, hence other corpus is needed. At CUHK, we have developed the CUSENT Cantonese speech corpus, which mainly contains read newspaper content. It is phonetically rich under various contexts, and the recording environment is the same as this corpus [2]. The CUSENT corpus mainly involves written Cantonese, which is a bit different from spoken Cantonese in lexicon choice. The code-mixing data in CUMIX can be applied for adapting the written Cantonese to spoken Cantonese.

Apart from acoustic model training, the training data can also be used for studying algorithms for language boundaries detection in code-mixing speech data. A listing of the utterances used in the training set can be found in Appendix A.

Testing Data

The testing data is for performance evaluation of the code-mixing speech recognizer or language identification system. Part of the data can be used for development purpose, such as acoustic model adaptation and parameters tuning. Each speaker read 120 code-mixing utterances, and about 90 monolingual Cantonese utterances. Among the 120 code-mixing utterances, 110 of them include single English segment only, and the remaining 10 utterances include two English segments. Each English segment can contain a single word, a vocabulary with multiple words, or a phrase. Table 4-2 shows the examples on code-mixing utterances with single English segment and two English segments.

Code-mixing utterances with single English segment	
Single word	我覺得今年有 bonus 嘅機會好渺茫。
Vocabulary with multiple words	好多 contract terms 我都睇唔明。
Phrase	by the way，我唔介意你食煙。
Code-mixing utterances with two English segments	
Single word + Single word	過埋今日，我 suppose 可以落鋪做 sales。
Single word + Vocabulary	今日 lunch 我哋又去咗 Pizza Hut 食嘢。

Table 4-2 Examples on code-mixing utterances with single English segment and two English segments

The monolingual Cantonese utterances are identical to the code-mixing utterances except that the code-switch words are replaced by their Cantonese equivalent. If Cantonese equivalent does not exist, the monolingual version of the utterance will not be recorded, thus the number of monolingual Cantonese utterance is less then those in code-mixing. The following as shown in Table 4-3 is an example of the code-mixing utterance and monolingual Cantonese utterance, where the code-switch word "bonus" is replaced by its Cantonese equivalent "花紅" in the monolingual case:

Code-mixing utterance								
ngo5	gok3-dak1	gam1-nin4	jau5	B OW N AH S	ge3	gei1-wui6	hou2	miu5-mong4
我	覺得	今年	有	bonus	嘅	機會	好	渺茫。
Monolingual Cantonese utterance								
ngo5	gok3-dak1	gam1-nin4	jau5	faa1-hung4	ge3	gei1-wui6	hou2	miu5-mong4
我	覺得	今年	有	花紅	嘅	機會	好	渺茫。

Table 4-3 Example of code-mixing utterance and monolingual Cantonese utterance

The testing data involves 20 male speakers and 20 female speakers. The utterances can be found in Appendix B. Table 4-4 summarizes the type of speech data found in CUMIX:

Type of speech data	Number of speech files per speaker
Cantonese-English code-mixing utterances (1 English content)	110
Cantonese-English code-mixing utterances (2 English content)	10
Monolingual Cantonese utterances	About 90

Table 4-4 Testing data in the CUMIX speech corpus

Code-mixing Text Material

Appropriate text materials are needed to reflect and cover the most prominent code-mixing scenario. The conventional way of obtaining these materials is to extract them directly from generally accessible text database such as local newspapers and books. However, this is not the case for Cantonese. Cantonese is a dialect and spoken Cantonese is noticeably different from written Cantonese. The grammar is similar but the lexicon selection is quite different. Table 4-5 shows an example.

Written Cantonese:	我	明天	不用	上學。
Spoken Cantonese:	我	聽日	唔駛	返學。
English Translation:	I	tomorrow	need not	go to school
	(I need not go to school tomorrow)			

Table 4-5 Examples on different lexicon selections in spoken Cantonese and written Cantonese

The text materials for recording are in spoken Cantonese, which is a 'low' language, such that most of the local newspapers and books do not use it. In order to collect adequate text materials that encompasses code-mixing of spoken Cantonese and English, other sources such as newsgroup and online diary are also included. Utterances being used in the previous researches related to Cantonese-English code-mixing / code-switching are also considered [3]. The sources of code-mixing text data for the CUMIX corpus are listed in Table 4-6:

Type of text	Details
Online diary	The Open Diary (http://www.opendiary.com/)
Local newsgroup	News://news.cuhk.edu.hk/cuhk.forum News://news.hkpcug.org/hkteen.university.CUHK News://news.hkpcug.org/hkpcug.mobile-phone News://news.hkux.net/edu.electronic
Research materials	Brian Hok-Shing Chan, Code-mixing in Hong Kong Cantonese-English bilinguals : constraints and processes, M. A. Thesis, The Chinese University of Hong Kong, 1992

Table 4-6 Source of code-mixing text data for the CUMIX corpus

In order to reflect and cover the code-mixing scenario, the followings criteria are considered when selecting the code-mixing text material:

a) frequency of occurrence of the code-switched words
b) part-of-speech (POS) of the code-switch words
c) the 'length' of the code-switch words, e.g. single alphabet, abbreviations, single word, multiple words, phrase

POS distributions and word length distributions has been taken into account in the speech corpus. There are 1047 unique code-switch words in the training data and 1069 in the testing data. Each code-switch word appears 4 to 12 times in the training data, depending on the frequency of occurrences in daily conversation. There are in total 2087 unique code-mixing utterances in the training data and 2256 in the testing data. Table 4-7 and Table 4-8 list the POS distributions and word length distributions of the code-switch words.

Parts-of-speech	No. of utterances	Proportion
Noun / Noun Phrase	7723	62.28%
Verb / Verb Phrase	2778	22.40%
Adjectives / Adverbs	1771	14.28%
Phrases / Others	128	1.03%

Table 4-7 POS distribution of the code-switch words

Word length	No. of utterances	Proportion
1	9295	74.96%
2	2197	17.72%
3	94	0.76%
4	16	0.13%
abbreviation	692	5.58%
ENG-CAN-ENG	106	0.85%

Table 4-8 Word length distribution of the code-switch words

Some of the code-switch words in abbreviation or "short form" of the English words. For example, the word "register" /r eh jh ah s t er/ is usually modified to monosyllabic and become "reg" /r eh/ in Cantonese-English code-mixing utterances. Besides, there are also English-Cantonese-English patterns in Cantonese-English code-mixing utterances. The Cantonese words involved are usually single character word, such as "唔" /ng4/ (not). For example, people will say "un 唔 understand" (understand or not?), or "mind 唔 mind"(mind or not?) [3].

Background of Speakers

This speech corpus includes speech data from 80 speakers, 40 male and 40 female. 20 male and 20 female speakers participated in the training session, and the remaining took part to provide the testing data. All the speakers are native Cantonese speakers and they can also speak fluent English. Most of them are undergraduate or graduate students from the Chinese University of Hong Kong, aged between 19 and 26.

4.1.2 Recording Setup

The speech data are collected in a closed silent recording room. The recordings are expected to be good quality clean speech data. The speakers are requested to read the utterances naturally and fluently, and use their normal pronunciation for the English words at a normal speaking rate. Therefore, borrowing may occur in the English words, and phone change or syllable fusion may also take place in the Cantonese words. The speakers have to check if they know the pronunciation of all the code-switch words before recording. Correct pronunciation will be told to ensure the quality of the code-switch words. The speakers were left alone in the recording room to do the recording themselves without assistance or intervention. They read the utterance prompted on a computer screen. The recording of each utterance is terminated automatically with silence detection. The speakers were requested to record the utterances again if the noise level is too high or the pronunciation of the code-switch words is too far away from the correct one.

The recording is collected using a high quality, unidirectional dynamic head-worn microphone. The microphone position is about 1 inch from the corner of the mouth, so as to eliminate explosive breath sounds. The signal passes through a pre-amplification mixer and sampled by a DAT recording at 48kHz and 16bit. The sampled digital data is then down sampled to 16kHz, and transferred through SCSI interface to computer and stored as disk files. The data collection set-up is shown in Figure 4-1.

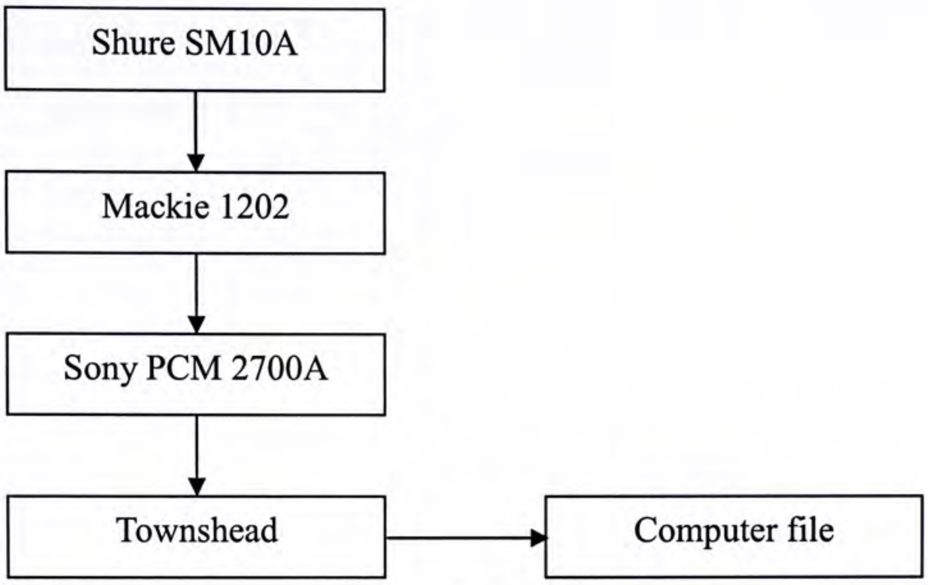


Figure 4-1 Block diagram for data collection set-up

4.1.3 Post-processing of Speech Data

The speech data is verified after compilation. Verification is performed in two stages. Stage one is done by generally trained assistants who mark out all incorrectly spoken and noisy data. The speakers are requested to record these data again.

Stage two of the verification process is performed by experts in phonetics. Pronunciation of the code-switch words is labeled and the Cantonese transcriptions will be corrected. If deemed uncorrectable, the corresponding data will be discarded.

Language boundaries (in ms) are also labeled manually for language identification. It uses the MLF format of HTK [4] that includes the time alignment for silence, Chinese content and English content. For data annotation, orthographic transcriptions in BIG5 code and phonemic transcriptions are provided. Phonemic transcriptions of English words are based on the CMU Pronouncing Dictionary version 0.6 [5]. ARPABET [6] is used and the phoneme set has 39 phonemes, lexical stress is ignored. Cantonese phonemes are labeled with LSHK [7]. Spoken Cantonese includes Hong Kong-specific characters which are not contained in the BIG5 standard character set, thus the Hong Kong Supplementary Character Set (HKSCS) [8] must be installed in order to view the text. Details of the CUMIX corpus are summarized in table 4-9 as follows:

Speech data format	NIST SPHERE
Sampling rate	16kHz
Precision	16 bit per sample
Orthographic transcription	BIG5 code (HKSCS)
Phonemic transcription	Cantonese: LSHK English: ARPABET
No. of speakers	Training: 20M, 20F Testing: 20M, 20F
No. of code-switching / monolingual Cantonese utterances	Training: 8000 Testing: 4800 / 3600
Corpus size (include silence time)	Training: 9 hr Testing: 8 hr

Table 4-9 Summary on the CUMIX corpus

4.2 A Baseline Database

A baseline database is required to compare the performance of the code-mixing speech recognizer. Monolingual Spoken Cantonese speech data is used for comparison.

4.2.1 Monolingual Spoken Cantonese Speech Data (CUMIX)

The monolingual Cantonese testing data in CUMIX is uttered by 20 male and 20 female speakers, and there are about 3600 utterances in total. These utterances are identical to the code-mixing utterances except that the code-switched words are replaced by their Cantonese equivalent. The same speaker utters the utterance pairs in order to minimize the speaker-dependent variation. Besides, the utterance pairs only different in the code-switched word or its Cantonese equivalent, such that effect from the language model is almost identical to the two utterances.

4.3 References

[1] Joyce Y. C. Chan, P. C. Ching and Tan Lee, “Development of a

Cantonese-English Code-mixing Speech Corpus", to be presented in *Proc. of Eurospeech 2005*, Lisbon, Portugal, September 2005

[2] 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 ICASSP 1997*, Vol.3, pp. 1775-1778, Munich, Germany, 1997

[3] Brian Hok-Shing Chan, *Code-mixing in Hong Kong Cantonese-English bilinguals : constraints and processes*, M. A. Thesis, The Chinese University of Hong Kong, China, 1992

[4] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, Dec. 2001.

[5] *The CMU Pronouncing Dictionary v0.6*, The Carnegia Mellon University, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[6] Shoup, J. E., "Phonological Aspects of Speech Recognition", in *Trends in Speech Recognition*, edited by W. A. Lea (Prentice-Hall, New York), pp. 125-138, 1980

[7] Thomas Hun-tak Lee, *Research on Chinese linguistics in Hong Kong* (《香港漢語語言學研究文集》), Linguistic Society of Hong Kong, Hong Kong, China, 1992

[8] *Hong Kong Supplementary Character Set – 2001*, Information Technology Services Department & Official Languages Agency, Government of the Hong Kong Special Administrative Region, Hong Kong, China, December 2001

Chapter 5

System Design and Experimental Setup

This chapter describes the details on the system design of the Cantonese-English code-mixing speech recognizer. Experimental setup includes a detail description on the resources and tools being used. The process of training and evaluation will also be elaborated.

5.1 Overview of the Code-mixing Speech Recognizer

In order to implement a Cantonese-English code-mixing speech recognizer, several resources and tools are required. Speech corpus is necessary to provide speech data to train the acoustic models, as well as to evaluate the whole system. Appropriate acoustic units and models should be selected, such that sound units in the two languages can be covered. For the purpose of constructing the N-gram language model, a large amount of spoken Cantonese text data has been collected. Appropriate training tools are sourced so as to train the acoustic models and language models. Language boundaries are detected so as to provide additional confidence measures on the hypothesis words. The whole system is finally evaluated by the Cantonese-English code-mixing and monolingual spoken Cantonese speech data.

5.1.1 Bilingual Syllable / Word-based Speech Recognizer

The recognition process is essentially a two-pass system. Acoustic features are extracted from the input speech signal; which are then recognized by the acoustic models. Bilingual pronunciation dictionary includes Cantonese syllables and English words are applied, and the language model will be implemented in the next stage. The bilingual speech recognizer will generate a word graph for each input speech file. The word graph is primarily syllable-based for Cantonese and word-based for English, and we call this "syllabie lattice". Details of the acoustic modeling will be described in Section 5.2, and Figure 5-1 gives a block diagram of the proposed bilingual speech recognizer.

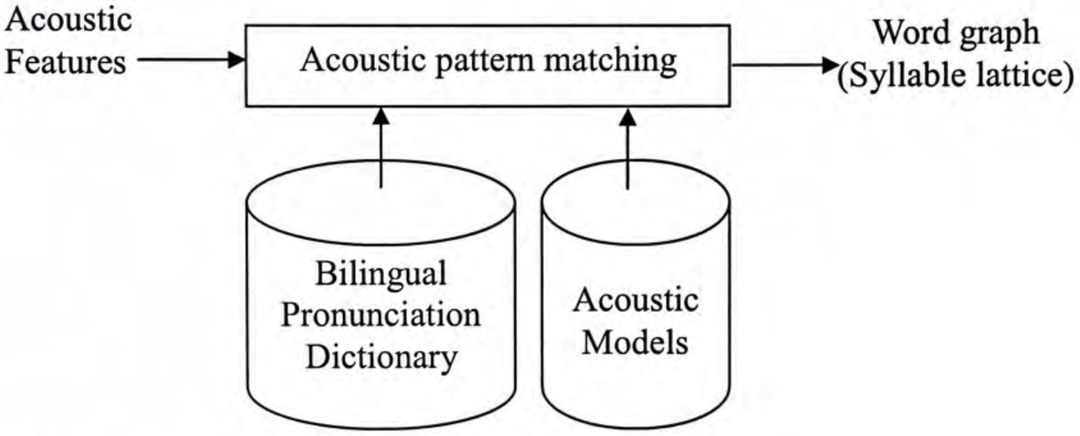

Acoustic Features
Acoustic pattern matching
Word graph (Syllable lattice)
Bilingual Pronunciation Dictionary
Acoustic Models

Figure 5-1 Block Diagram for the bilingual speech recognizer

5.1.2 Language Boundary Detection

Language Boundary Detection (LBD) on code-mixing utterances is achieved by automatic Language Identification (LID) techniques. The language boundaries act as an additional confidence measurement on the hypothesis word strings. The features vectors as mentioned previously will be passed to the language detector for language confidence measurement. Two language detectors are proposed: 1) the phone-based language boundaries detector; and 2) the syllable-based language boundaries detector. Details of the language boundaries detectors will be included in section 5.4. The language information is integrated to the syllable lattice by modifying the acoustic scores of the hypothesis words. If the detected language is the same as the language of the hypothesis word in the word graph, the acoustic scores are increased, such that the words in "correct" language will have higher probabilities to be selected during best path searching. On the other hand, if the detected language is different from the hypothesis word in the word graph, the acoustic scores will be deducted as a penalty. Figure 5-2 shows how language boundary information is integrated into the syllable lattice.

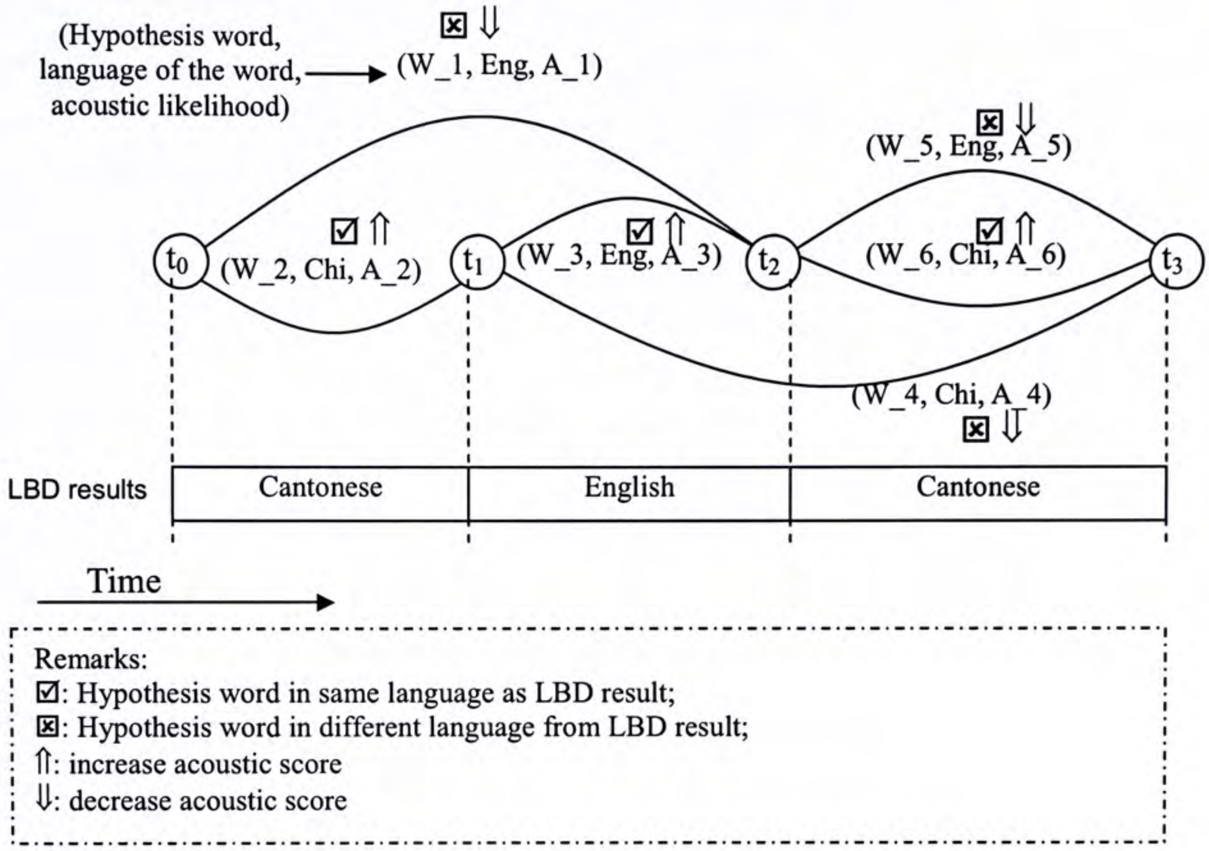

(Hypothesis word, language of the word, acoustic likelihood)
☒ ⇓
(W_1, Eng, A_1)
☒ ⇓
(W_5, Eng, A_5)
☑ ⇑
(W_2, Chi, A_2)
☑ ⇑
(W_3, Eng, A_3)
☑ ⇑
(W_6, Chi, A_6)
t_0
t_1
t_2
t_3
(W_4, Chi, A_4)
☒ ⇓
LBD results
Cantonese
English
Cantonese
Time
Remarks:
☑: Hypothesis word in same language as LBD result;
☒: Hypothesis word in different language from LBD result;
⇑: increase acoustic score
⇓: decrease acoustic score

Figure 5-2 Example on the integration of the language confidence score to the word graph

5.1.3 Generalized Word Posterior Probability (GWPP)

The language model scores are integrated to the modified syllable lattice by Generalized Word Posterior Probability (GWPP). GWPP is calculated, where time registrations in the word graph are released, and acoustic models scores and language models are re-weighted. Optimum weight of the acoustic model scores and language model scores (α, β) are derived from the database - CUMIX. Since there are homographs and homophones in Cantonese, the convention from syllable to character mainly depends on the language model score. The syllable-based word graph is converted to character-based word graph (character lattice), and word sequence with the highest GWPP score is then searched. Finally, the hypothesis word string is compared to the reference transcriptions, which allows the word error rate to be computed. Figure 5-3 shows the block diagram of the decoding process based on GWPP. Details of the integration process will be discussed in Section 5.5.

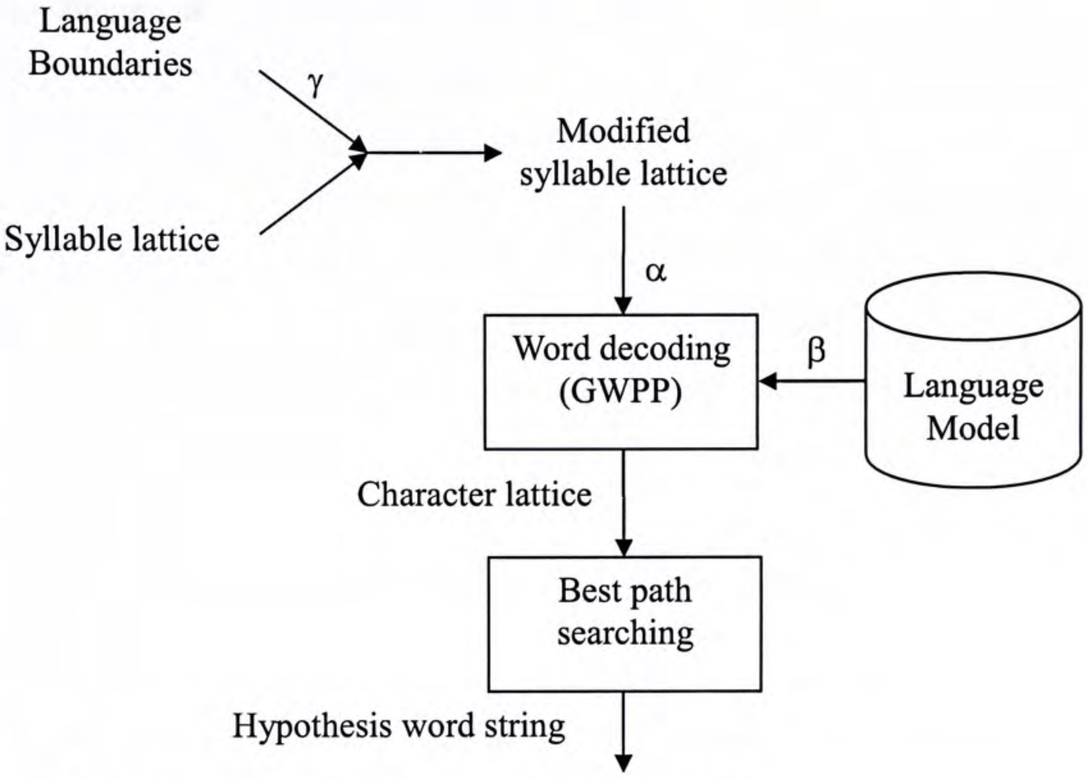


Figure 5-3 Block diagram of the decoding process based on GWPP

5.2 Acoustic Modeling

Speech recognition is basically a pattern matching process. Given the feature vectors extracted from the input testing speech, the sequence of acoustic units having maximum likelihood is searched. Figure 5-4 illustrates the block diagram of the training and evaluation process of acoustic modeling.

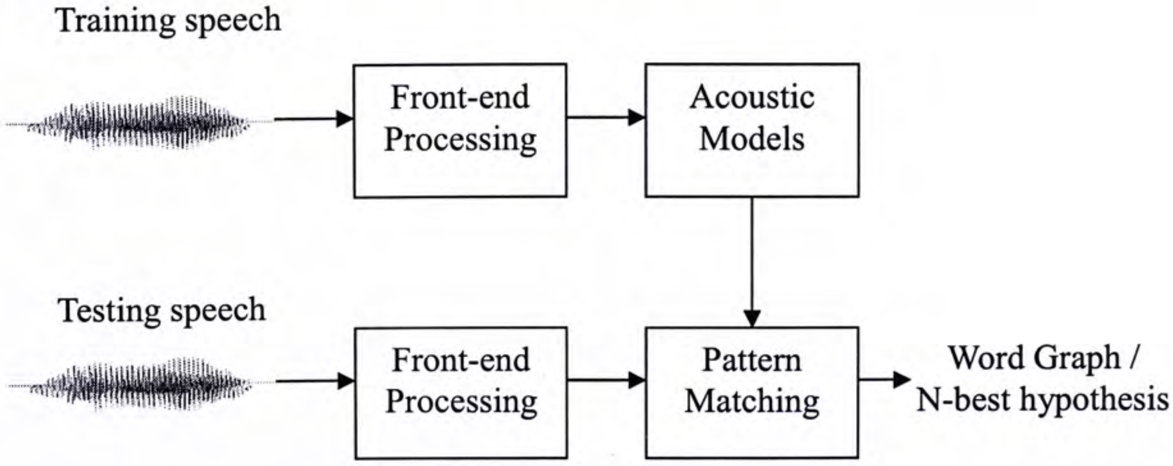


Figure 5-4 Block diagram of the training and evaluation process of acoustic modeling

Parameter distributions of the acoustic units are estimated in the training process.

Given the feature vectors and the transcriptions, the parameters of each acoustic unit are estimated and then stored in the HMMs.

The acoustic models used should be appropriately chosen, such that they are in suitable size and able to cover phonemes in the two languages. For LVCSR, the acoustic unit is usually sub-word based since it enables the recognition of words that had not been encountered in the training data. Three speech corpora are involved in order to study the performance of language-dependent and language-independent acoustic models.

5.2.1 Speech Corpus for Training of Acoustic Models

Speech corpus provides speech data to train the acoustic models. Three speech corpora are involved: 1) TIMIT; 2) CUSENT; 3) CUMIX. The purpose of using different combinations of training data is to compare the performance of the acoustic models, such that they can have minimum syllable (Cantonese) and word (English) error rate with the code-mixing testing data.

TIMIT

The Texas Instruments / Massachusetts Institute of Technology (TIMIT) corpus of read speech contains speech from 630 speakers representing 8 major dialect divisions of American English [1-3]. The TIMIT corpus was designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development as well as evaluation of automatic speech recognition systems. It was so designed to maintain a balance between applicability and manageability through the inclusion of a variety of utterances from a relatively diverse speaker population and a large range of phonetic environments.

The corpus contains a total of 6300 utterances, 10 sentences spoken by each speaker. The 10 sentences represent roughly 30 seconds of speech material per speaker, such that the corpus contains approximately 5 hours of speech. The entire corpus includes 2342 distinct texts, 6099 distinct words and 52 distinct phonemes. In order to synchronous the phonemes with those being used in CUMIX, the number of phonemes are reduced to 39 in the training and evaluation process.

CUSENT

CUSENT is one of the speech corpora in the CUCorpora speech database developed in the Chinese University of Hong Kong (CUHK). It is the world's first collection of large-scale Cantonese speech databases commenced in 1997 [4]. CUSENT is a large collection of read Cantonese sentence that is designed to be phonetically rich. There are 5100 distinct sentences for training and 600 distinct sentences for testing. It covers different combinations of phonetic units, so as to provide an abundant amount of data for context dependent phonetic unit modeling. The training data involves speech data from 34 male and 34 female speakers, 300 sentences per speaker [4]. It provides manually verified phonemic transcription, and it includes 637 unique non-tonal syllables.

The speech data are in NIST format, sampling rate is 16kHz, 16-bit per sample [5]. Speech data in CUSENT and CUMIX is collected in the same recording condition, by the same set of equipment.

CUMIX

CUMIX is a Cantonese-English code-mixing speech corpus. It involves read speech in both Cantonese and English, and the English words carry Cantonese accents. The read data are spoken Cantonese, and the details have already been described in the previous chapter.

The main difference between CUSENT and CUMIX is in the read content. Although both the two corpora are in Cantonese, the one in CUSENT is based on the standard written Chinese, while that in CUMIX is based on spoken Cantonese. The average speaking rate of CUMIX is higher than that of CUSENT, and the frequency of phone change [6] and syllable fusion [7] are also higher. When recording the CUMIX speech data, the speakers are requested to speak naturally such that the speaking style is similar to fluent spontaneous speech [8] or read spontaneous speech [9]. The speech data in CUMIX does not contain filled pause, but sometimes includes word fragments, overly stressed function words and mispronunciations. A summary and comparison of the three corpora are shown in Table 5-1.

Details on the training corpus	TIMIT	CUSENT	CUMIX
No. of speakers	462	68	40
No. of speech files	4620	20K	12.2K
No. of Cantonese syllables	-	216K	75K
No. of English words	40K	-	13.5K
Total Length (hours)	5	20	9
Average sentence length (No. of Cantonese syllables / English words)	8.6	10.6	10.4
Cantonese syllable rate	-	3.94 / sec	4.89 / sec

Table 5-1 Comparison of training set of the TIMIT, CUSENT and CUMIX corpora

5.2.2 Features Extraction

Mel Frequency Cepstral Coefficient (MFCC) is used as the acoustic features. Each feature vector represents features in the 25ms frame, and the frame shift is 10ms. 12MFCCs and the speech energy of the frame, as well as their first and second order derivatives are used, such that the dimension of each feature vector is 39. Cepstral mean normalization is also incorporated, which is estimated by computing the average of each cepstral parameter across each input speech file [10].

5.2.3 Variability in the Speech Signal

Acoustic modeling should take into account speaker variations, pronunciation variations, environmental variations, and context-dependent phonetic co-articulation variations [11]. To deal with speaker variations, speech data from speakers with different gender, age and background is collected. The same sound unit from different speakers is modeled by a single HMM with multiple Gaussian mixture densities, such that the acoustic models are speaker-independent. Pronunciation variation is also solved by similar approach as speaker variation by utilizing speech data from different speakers. To reduce the variations due to environment differences, the training data and testing data should be recorded in similar environment with background noise. To tackle the context-dependent phonetic co-articulation variations, context-dependent HMM (CDHMM) is implemented.

Sub-word acoustic models are used such that phone sequences unseen in training data can also be modeled.

Context-Dependent HMM (CDHMM)

Due to co-articulation, speech uttered in isolation is very different from those uttered naturally as in ordinary conversation [11]. To tackle this problem, context-dependent HMMs are employed. Tri-phone HMMs are utilized, such that effects from both the left and right context can be considered. Since training data is insufficient that not all the tri-phones appear in the speech, the HMM states are clustered according to their statistic distributions by the decision-tree based clustering algorithm [10].

Phone-based Models

Instead of using demi-syllable (initials and finals) models commonly used by other researchers on ASR in Cantonese [13-15], phone-based models are selected. In the research on Mandarin [12], it is found that context-dependent phone-based models achieve similar syllable accuracy as demi-syllable models.

For speech recognition on English, it usually employs syllable-like models [16-18], phone models [19][20] or both of them [21]. The syllable structure of English is more complicated than Cantonese, such that there are thousands of syllables in English [17]. In the TIMIT corpus, there are already 3K syllables, although 78.3% of the canonical syllables have a CV, CVC, VC or V structure [22]. Even if demi-syllable models are utilized, the number of sound units is still numerous.

Therefore, phone is the most suitable acoustic unit for speech recognition for English and Cantonese. The same acoustic unit can be used for the two languages, and similar sounds can be clustered. Phone models are highly reusable across the two languages, and each acoustic model can be represented by more Gaussian mixtures due to larger amount of training data for each acoustic model.

In this thesis, each HMM involves 3 emitting states if it is consonant and 5 emitting states if it is vowel or vowel-constant. Short pause (SP) and silence (SIL) models are also employed, which has 1 and 3 emitting states respectively.

5.2.4 Language Dependency of the Acoustic Models

In monolingual speech recognition, language-dependent set of acoustic units is usually used. However, this monolingual approach shows serious limitations, since for each new language, a new set of sub-word units would have to be defined, thus calling for a growing number of costly database [23]. Instead, cross-language or language-independent tri-phones are usually applied for multilingual speech recognition, or to boostrap models for a new language [23-27].

In order to study the performance of the language-dependent and language-independent acoustic units on code-mixing speech, both types of models are trained. They are trained by combinations of speech corpora mentioned in section 5.2.1. Different phoneme inventories are applied, and the details are as follows:

Language-dependent Acoustic Unit

The two languages involve in code-mixing are different in many ways. Some sounds in the two languages are in common, but the duration and context may be different. To retain the language-specific features, language-dependent acoustic unit can be considered. Language-dependent acoustic units always achieve higher recognition accuracy for monolingual speech recognition [28]. However, the case for code-mixing is different from monolingual that the code-switch words may contain accents of the matrix language. Sounds in the embedded language are usually pronounced as similar sounds in the matrix language, and the importance of the language-specific information will be lost.

The language-dependent acoustic models are phone-like models for both Cantonese and English, and they are listed below in Table 5-2 and Table 5-3. There are in total 56 acoustic models for Cantonese and 39 acoustic models for English.

Phone Type	No. of models	Cantonese Phone models (labeled in LSHK)
Consonant	25	f, h, k/kw, g/gw, l/initial_n, m, initial_m, final_m, final_n, ng, initial_ng, final_ng, initial_null, b, p, s, s(yu), z, z(yu), c, c(yu), d, initial_t, w, j
Vowel	11	a, aa, o, e, eo, i, i(ng), oe, u, u(ng), yu
Vowel-stop consonant	20	ap, at, ak, aap, aat, aak, ep, et, ek, ut, uk, yut, ip, it, ik, op, ot, ok, eot, oek

Table 5-2 Acoustic models for Cantonese

Phone Type	No. of models	English Phone models (labeled in ARPABET)
Consonant	24	dh, th, f, v, w, z, zh, s, sh, t, d, b, p, ch, g, h, jh, k, l, m, n, ng, y, r
Vowel	15	aa, ae, ah, ao, aw, ay, eh, er, ey, ih, iy, ow, oy, uh, uw

Table 5-3 Acoustic models for English

Cantonese Phone Models

The acoustic models are mainly phone-based for Cantonese and phoneme-based for English. Since the final stop consonants in Cantonese are unreleased, their sound mainly occurs at the vowel instead of the consonant. Therefore, the stop consonants are coded with the nucleus, such that the acoustic unit becomes vowel-stop consonant. Some of the easily confused initial consonant phones in Cantonese are clustered. They are: 1) /k/ and /kw/; 2) /g/ and /gw/; and 3) /l/ and /n/ [29].

English Phone Models

The pronunciation dictionary of TIMIT in fact contains 52 phones instead of 39. There is no allophone in the original TIMIT pronunciation dictionary [1]. However, in order to synchronize with the CMU pronunciation dictionary [30], the one used for transcribing CUMIX, the 39 ARPABET [31] phonemes are employed instead. The CMU pronunciation dictionary is a machine-readable pronunciation dictionary for North American English that contains over 125,000 words and their transcriptions. It has mappings from words to their pronunciations in the given phoneme set. The phoneme set contains 39 ARPABET phonemes, for which the vowels may carry lexical stress. ARPABET is designed for transcribing American English phonetic sounds, and only ASCII characters are used. However, it fails to distinguish between all the phonemes as IPA does, therefore modification is necessary if it is to be applied to languages other than English. The 13 extra phonemes in TIMIT are clustered to the 39 ARPABET, and the details are listed in Table 5-4 [1].

Original symbol in TIMIT		Clustered symbol		Example
ARPABET [IPA]	Description	ARPABET [IPA]	Description	
dx [ɾ]	voiceless tap (allophone of /t/)	t [t]	voiceless alveolar stop	muddy
q [ʔ]	voiceless glottal stop (allophone of /t/), or initial vowel, or vowel-vowel boundary	t [t]	voiceless alveolar stop	bat
em [m̩]	syllabic	ah m [ʌm]	-	bottom
en [n̩]	syllabic	ah n [ʌn]	-	button
eng [/]	syllabic	ih ng [ɪŋ]	-	Washington
el [ɫ̩]	syllabic	ah l [ʌl]	-	bottle
nx [n]	nasal flap	n [n]	voiced alveolar nasal	winner
hv [/]	voiced /hh/	hh [h]	voiceless glottal fricative	ahead
ux [ʉ]	fronted /uw/	uw [u]	high back tense rounded	toot
ax [ə]	middle central lax (unstressed)	ah [ʌ]	middle central lax (stressed)	about
ix [ɪ]	unstressed /ih/	ih [ɪ']	high front lax	debit
axr [ɚ']	high central lax r-colored (unstressed)	er [ɚ]	high central lax r-colored (stressed)	butter
ax-h [/]	devoiced-schwa, short devoiced vowel	ah [ʌ]	middle central lax (stressed)	suspect

Table 5-4 ARPABET phonemes being merged

When the TIMIT corpus is utilized for training the English phone models, clustering is required. However, when the code-mixing corpus CUMIX is applied, the clustering is not necessary since the transcriptions of English words in CUMIX already follow the CMU pronunciation dictionary.

Language-independent Acoustic Unit

Although the phoneme inventory of each language is different, some of the phones may be in common, and the acoustic features in this case can be shared. In code-mixing speech, if the embedded words carry accents of the matrix language, it may be pronounced as similar phones in the matrix language. If the alike phones in the two languages share the same acoustic units, more training data can be utilized for each acoustic model. There will be less acoustic models and the number of Gaussian mixture densities of each model is increased. Efficiency of the speech recognizer can be enhanced and accuracy should accordingly be increased as well. However, if the sounds from the two languages are greatly different, but clustered to the same acoustic model, recognition accuracy will be degraded. Hence, the clustering of phone models should be decided carefully. Universal phone set based on IPA [32] is employed for the language-independent acoustic models. All the Cantonese phones are used, and English phones that are very different from Cantonese phones are also included. Table 5-5 to Table 5-7 shows the universal phone set for Cantonese and English.

Consonants					
Model name	**LSHK label**	**ARPABET label**	**IPA**	**Cantonese Example**	**English Example**
C_F1	f-	f, v, th, dh	f,v,θ,ð	faa1 花	function
C_H1	h-	hh	h	haa1 哈	hall
C_KH1	k-, kw-	k	k^{h}, k^{wh}	kaa1 卡	keep
C_K1	g-, gw-	g	k, k^{w}, g	gaa1 加	graph
C_L1	l-, n	l, n	l, n	laa1 啦	last
C_ML	m	-	m	m4 唔	-
C_M1	m-	m	m	maa1 嗎	make
C_M3	-m	m	m	gaam1 監	nickname
C_N3	-n	n	n	gaan1 奸	offline
C_NGL	ng	-	ŋ	ng5 五	-
C_NG1	ng-	ng	ŋ	ngaa4 牙	singing
C_NG3	-ng	ng	ŋ	ling4 零	spelling
C_NULL1	null-	-	-	aa1 丫	-
C_P1	b-	b	p	baa1 巴	bar
C_PH1	p-	p	p^{h}, b	paa1 扒	plan
C_SH1	s-	s	s	si1 詩	say
C_SY1	s(yu)-	sh, zh	ʃ, ʒ	syu1 書	publish, visual
C_Ts1	z-	-	ts	zi1 之	-
C_TsY1	z(yu)-	jh	tʃ, dʒ	zyu1 朱	homepage
C_TsH1	c-	-	ts^{h}	ci1 癡	-
C_TsHY1	c(yu)-	ch	$tʃ^{h}$	cyu3 處	lunch
C_T1	d-	d	t	daa1 打	dancer
C_TCH1	t-	t	t^{h}, d	taa1 他	target
C_W1	w-	w, v	w, v	waa1 娃	win, value
C_Y1	j	y	y	jaa3 也	year
E_d3	-	d	d	-	background
E_k3	-	k	k	-	park
E_t3	-	t	t	-	art
E_s1	-	s, z	s, z	-	abstract
E_r1	-	r	ɹ	-	drop

Table 5-5 Universal Phone Set for Cantonese and English (Consonants)

Vowels					
Model name	LSHK label	ARPABET label	IPA	Cantonese Example	English Example
C_A2	a	ah	ɐ, ʌ	man4 文	assignment
C_AA2	aa	er, ae, ah, aa, etc.	a, æ, ɚ, ɑ, ɒ	maa4 麻	grammar, academic
C_AO2	o	ao, aa	ɔ	do1 多	draw, desktop
C_EH2	e	eh	ɛ	se1 些	chairman
C_EO2	eo	ah	ø	seon4 純	conclusion
C_IY2	i	iy	i	si4 時	busy
C_IH2	i(ng)	ih	ɪ	ting4 停	casting
C_OE2	oe	er	œ	soeng1 相	concern
C_UW2	u	uw	u	fu1 夫	full
C_UH2	u(ng)	uh, ow	ʊ	sung1 鬆	bonus
C_YU2	yu	y uw, ch	y	syu1 書	unique
C_iu	iu	y uw, uw, iy, y ah, etc.	iu	liu1 撩	annual, coupon
E_ay2	aai	ay	ɐi	waai1 歪	supervisor
C_ai	ai	ai, ay	ai	wai1 威	writer
C_au	au	au	ɐu	sau1 收	consultant
E_ow2	ou	ow	ou	ou3 奧	crossover
E_oy2	oi	oy	ɔi	oi1 哀	disappointed
E_ey2	ei	ey	ei	pei1 砒	display
E_ah2	-	ah	ʌ	-	number
E_el2	-	el	ɛl	-	parallel

Table 5-6 Universal Phone Set for Cantonese and English (Vowels)

Vowel-consonants					
Model name	**LSHK label**	**ARPABET label**	**IPA**	**Cantonese Example**	**English Example**
C_AP2	ap	ah b, ah p, aa p, ah, etc.	ɐp	sap6 十	submit, support
C_AT2	at	ah, ah k	ɐt	bat1 不	puzzling
C_AK2	ak	ah k	ɐk	bak1 北	Starbuck
C_AAP2	aap	ae p, ae b	ap	aap3 鴨	abstract
C_AAT2	aat	aa r t, aa r d	at	aat3 壓	card
C_AAK2	aak	aa r k, aa r t, ae k, etc.	ak	baak3 百	park
C_EP2	ep	eh p, ah p, ae p, eh, etc.	ɛp	gep3 夾	accept, apply
C_ET2	et	ae t, ae k, etc.	ɛt	-	at, activity
C_EK2	ek	ae k, ae t, etc	ɛk	tek3 踢	technical
C_UT2	ut	uw t, uh d, etc.	ut	wut6 活	boot, touchwood
C_UHK2	uk	uh k, uw p	ʊk	buk1 卜	book, group
C_YT2	yut	uw b	yt	tyut3 脫	tube
C_IYP2	ip	iy p, ih p	ip	cip3 妾	cheap, skip
C_IYT2	it	ih, ah t, iy t, etc.	it	bit1 必	admission, credit
C_IHK2	ik	ih k, ih g	ik	jik1 益	curriculum, exactly
E_aop2	op	aa, ah b, aa p, etc.	ɔp	-	hip-hop, objective
C_AOT2	ot	ao r t	ɔt	got3 割	port, support
C_AOK2	ok	ao k, ao g	ɔk	gok3 角	talk, analog
C_EOT2	eot	-	øt	zeot1 卒	-
C_OEK2	oek	er t, er, etc	œk	zoek3 雀	alert, birthday

Table 5-7 Universal Phone Set for Cantonese and English (Vowel-consonants)

Cantonese phones are labeled in LSHK (The Linguistic Society of Hong Kong Cantonese Romanization Scheme) [33][34] and English phones are labeled in ARPABET. The naming of the HMMs is based on the LSHK label, and language tag "C_" (Cantonese) and "E_" (English) is applied as well. The IPA symbols in column 4 mainly based on the pronunciation in Cantonese phones, and those for English phones are included if they are different from the Cantonese phones. The words in the English examples includes Cantonese accents, such that the sounds may be different from the native pronunciation

Although the Cantonese final stop consonants are unreleased [29], the one in English is released. In code-mixing utterances, the final stop consonants of the code-switch words may be released or unreleased, it depends on the speakers' habit. Hence, the optional released stop consonants /E_d3/, /E_k3/ and /E_t3/ are used. Besides, the [s] and [z] sounds in English are similar, but only [s] exist in Cantonese, such that some speakers pronounce the [z] sound, as [s]. On the other hand, there are [s] and [ʃ] in Cantonese, but they are seldom confused. Hence, the [s] and [z] in English are clustered to one acoustic unit (E_s1), but separated from the [s] (C_SH1) and [ʃ] (C_SY1) in Cantonese. The [ɹ] in English is treated as optional sound when it is the second consonants C_2 in the $C_1C_2VC_3$ or C_1C_2V structure. Diphthongs are also considered as phone models if they exist in both languages, so there are 70 phone models in total.

The mapping from English phones to the Cantonese phones is not one-to-one because the pronunciation of the phonemes is context-dependent. Phoneme is the smallest speech unit in the formation of a particular spoken language. It forms the smallest set of unambiguous symbols that altogether will be sufficient for representing a language [35]. On the other hand, phones are the smallest sound-building units that are physically differentiable. Different phones are formed physically by changing the manner and place of articulation during speech production. They are the fundamental sound categories that describe the range of acoustic features. [35]

There exist allophones in every language that several similar phones may belong to the same phoneme [36][37]. The allophones in one language may be regarded as different phonemes in another language. The problem is less significant for monolingual speech recognition that all the phones in the same phoneme can be represent by CDHMMs. In code-mixing speech recognition, the allophones in each language should be represented by appropriate phones instead of phonemes in order to reduce the confusion. In this research, all the Cantonese acoustic models are phone models instead of phonemes. The language-dependent English models are phonemes, while the language-independent English modes are changed to phones.

However, the mapping of phonemes to phones is not one to one but context dependent. For example, the phoneme /aa/ in the word "context" /k aa n t eh k s t/ is pronounced as [ɒ], but the one in another word "hard" /h aa r d/ is pronounced as [ɑ]. The same phoneme can be realized as different phones when it has different context. The allophone /aa/ in the word "context" is similar to /ao/ in Cantonese, but the one in "hard" is similar to /aa/. Therefore, the same phoneme in English will be mapped to different phones according to their contexts.

5.2.5 Pronunciation Dictionary

Pronunciation dictionary here refer to the dictionary for speech recognition by acoustic models. It breaks down the English words and Cantonese syllables to phone sequences, which is used to transform the word-based transcriptions to phone-based transcriptions.

Since the code-switch words may contain Cantonese accents, both the native pronunciation and the accented pronunciation should be included in the pronunciation dictionary. Three pronunciation dictionaries are prepared in order to train the language-dependent and language-independent acoustic models, they are:

a) Language-dependent bilingual dictionary without Cantonese accents (DICT01);
b) Language-dependent bilingual dictionary with Cantonese accents (DICT02).
c) Language-independent bilingual dictionary with Cantonese accents (DICT03).

The main difference of the three pronunciation dictionaries is on the labeling of English words. For the Cantonese syllables, the same phone sequence and labeling scheme is employed.

<u>Cantonese Syllables in the Pronunciation Dictionary</u>

The Cantonese pronunciation dictionary being employed is the electronic syllabary of the Cantonese dialect "Cantonese Pronunciation Dictionary" (《粵音韻彙》), which based on the work of Professor Wang Shih-ling. The pronunciation dictionary was developed in 1996 by the Research Centre for Humanities Computing of the Research Institute for the Humanities (RIH), Faculty of Arts, The Chinese University of Hong Kong [38]. The Cantonese syllables are transcribed by the LSHK labeling scheme, and tones are ignored. These Cantonese syllables are mapped to the 56 Cantonese phones being shown in Table 5-2. No English phone is utilized for the labeling of Cantonese syllables.

Phone change occurs frequently in spoken Cantonese, especially for the speech of young people [29]. Phone change of the phone pairs listed in Table 5-8 is handled by clustering the easily confused phones models.

Clustered Phones
k-, kw-
g-, gw-
l-, n-

Table 5-8 Phone changes in Pronunciation Dictionary (Cantonese)

There exist homographs in spoken Cantonese such that the same character may have different pronunciation when it has different meaning. However, some of the characters are pronounced differently even when it has the same meaning and context. These characters are commonly applied in spoken Cantonese but the lexicons that contain them seldom appear in standard written Chinese. Pronunciations of them are not taught in schools since the official language is standard written Chinese but not spoken Cantonese. People learn the pronunciation from their parents or through conversations, such that they may be variations. However, the several pronunciations of the same character are usually similar to each other, that the difference between them is generally one phone only.

To solve this problem, theses characters are modeled separately by HMMs with 8 emitting states, but not the phone-based HMMs. After several iteration of training, the phone sequences of these words are obtained by "force alignment". The phone-based HMMs are connected together according to the syllable-based transcriptions and the pronunciation dictionary. State sequence with highest likelihood is calculated, such that time alignment of each acoustic unit is obtained. If there are multiple entries for the same word in the pronunciation dictionary, the phone sequence with highest likelihood will be selected, such that the phone-based transcriptions of these words can be obtained [10]. These specially treated words are listed in Table 5-9 as follows:

Model Symbol	Character	LSHK Syllable	Example
C000A	呢	lik, li, le, ji, nik, ni, ne	呢個
C000B	咪	mai, mei	係咪
C000C	佢	heoi, keoi	佢哋
C000D	嚟	lai, lei	嚟咗未
C000E	面	min, bin	入面
C000F	拎	ling, lik	拎住
C000G	成	seng, sing	成日

Table 5-9 Specially treated homographs

Errors due to syllable fusion will be dealt with by the character pronunciation dictionary which converts the Cantonese syllables to Cantonese characters. The occurrence of syllable fusion is context-dependent, and mainly occurs in high-frequency words [7]. Hence, problems due to syllable fusion will be solved in the word level instead of the phone level.

Cantonese Accents in English Words

To include the Cantonese accents in English words, the pronunciation dictionary is modified manually. The modifications are based on the phone change and phonological change mentioned in Chapter 3. It includes insertion, deletion and substitution of phones. Both the native phone sequence from the CMU pronunciation dictionary and the accented phone sequence are included in the pronunciation dictionary such that each English word may now have multiple entries.

Language-dependent Bilingual Dictionary Without Cantonese Accents (DICT01)

English words in the pronunciation dictionary DICT01 do not contain Cantonese accents. It is prepared for training the language-dependent phone models – Model Set A. In Model Set A, all the English words are trained from the TIMIT [1] corpus. Since the English words in the TIMIT corpus are all pronounced by native speakers, there should be no Cantonese accents in the English words.

All the Cantonese syllables are labeled by the 56 Cantonese phones in Table 5-2.

For the English words, they are all labeled by the 39 English phones listed in Table 5-3. The phone sequence of English words is mainly based on the pronunciation dictionary in the TIMIT corpus but lexicon stress is ignored. Some of the English words in testing data do not appear in the TIMIT corpus, hence, phone sequence of these unseen words is based on the CMU pronunciation dictionary [30]. The pronunciation dictionary utilizes the 39 ARPABET phonemes for labeling and the vowels may carry lexical stress. To simplify the case, lexical stress is ignored in this research.

Language-dependent Bilingual Dictionary With Cantonese Accents (DICT02)

The pronunciation dictionary DICT02 which contains Cantonese accents is prepared for training the Model Set B, which CUSENT and CUMIX are applied to train the language-dependent acoustic models. The English words in CUMIX contain Cantonese accents, such that the pronunciation dictionary should also include them. On other hand, some of the speakers may pronounce the words in native phone sequence, such that the native pronunciation is included in DICT02 as well.
Again, we use the CMU pronunciation dictionary [30] and lexicon stress is ignored. The English words are labeled by the 39 language-dependent English models listed in Table 5-3.

Language-independent Bilingual Dictionary With Cantonese Accents (DICT03)

The pronunciation dictionary DICT03 is prepared for training the Model Set C, which CUSENT and CUMIX are employed to train the language-independent acoustic models. Both the native and accented pronunciation of the English words is included. Similar to DICT02, the pronunciation of all English words is based on the CMU pronunciation dictionary [30] without lexicon stress. The difference from DICT02 is that, the English words are labeled by the 72 language-independent acoustic models listed in Table 5-5 to Table 5-7. 1136 code-switch words are included and each of them has an average 2.267 different pronunciations.

5.2.6 The Training Process of Acoustic Models

Three set of acoustic models are trained in order to select the appropriate acoustic model set for Cantonese-English code-mixing speech recognition. The three acoustic model sets are listed in Table 5-8, which involve different combination of the training data.

Model Set	TIMIT	CUSENT	CUMIX	No. of phones	Language dependency	Pronunciation Dictionary
A	✓	✓		97	YES	native
B		✓	✓	97	YES	native + accented
C		✓	✓	72	NO	native + accented

Table 5-10 Acoustic model sets trained by the three Speech Corpus

Training Tools and the Training Process

To train the acoustic models, appropriate training tools are required. In this thesis, all the training and evaluation tools used for acoustic modeling come from HTK version 3.1 [10]. Parameter distributions of each acoustic unit are modeled by Hidden Markov Model (HMM). MFCC feature vectors are first extracted from the digitized speech signal by the tool HCOPY [10]. The mean and variance of each Gaussian component in a HMM is then initialized to the global mean and variance of all the speech by the tool HCOMPV [10]. The phone-based transcriptions identify the sequence of phones, and no phone boundary information is needed for this flat-start training.

Parameter distributions of the HMMs are then trained by the tool HEREST [10]. Training files which contain the MFCC feature vectors are processed by HEREST in turn. A composite HMM is constructed by concatenating instances of the phone-based HMMs corresponding to each label in the transcriptions. The Forward-Backward algorithm [39][40] is then applied and the sums needed to form the weighted averages accumulated in the normal way. After processing all the training files, new parameter estimates are formed from the weighted sums and the updated HMM set is output. It is assumed that the parameters are Gaussian distributions, such that the mean and variance of each parameter in the feature vectors are stored in the HMMs. There is one Gaussian mixture in each state of the HMMs, and state transition probabilities are stored in the transition matrix of each HMM.

Baum-Welch re-estimation and forward-backward algorithms are applied in order to estimate the mean and variance of each state, as well as the state transition probability. After several re-estimations, the parameters are converged. Then the context-independent mono-phones will be converted to context-dependent tri-phones that depends both the left and right context. After several re-estimations again, states in tri-phones will be clustered and tied according to phonetic question which is a decision-tree based clustering method [10]. It classifies all phonetic contexts into two groups buy asking yes/no questions. The phonetic questions are based on 1) manner of articulation of the adjacent consonant; 2) place of articulation of the adjacent consonant; and 3) the vowel identity of the adjacent vowel. The decision-tree based clustering is realized by the tool HHED [10]. It reads the statistic file which includes state information of each tri-phone HMM. Outlier states will be removed during clustering. The same state of all the tri-phones with the same center phone are clustered and tied. Top down clustering is performed, which start by placing all items in a single root node and then choosing a question from the phonetic question set to split the node in such a way as to maximize the likelihood of a single diagonal covariance Gaussian at each of the child nodes generating the training data. The splitting continues until the increase in likelihood falls below the threshold, or no questions are available which do not pass the outlier threshold test [10]. An example of the decision-tree based clustering is illustrated in Figure 5-5.

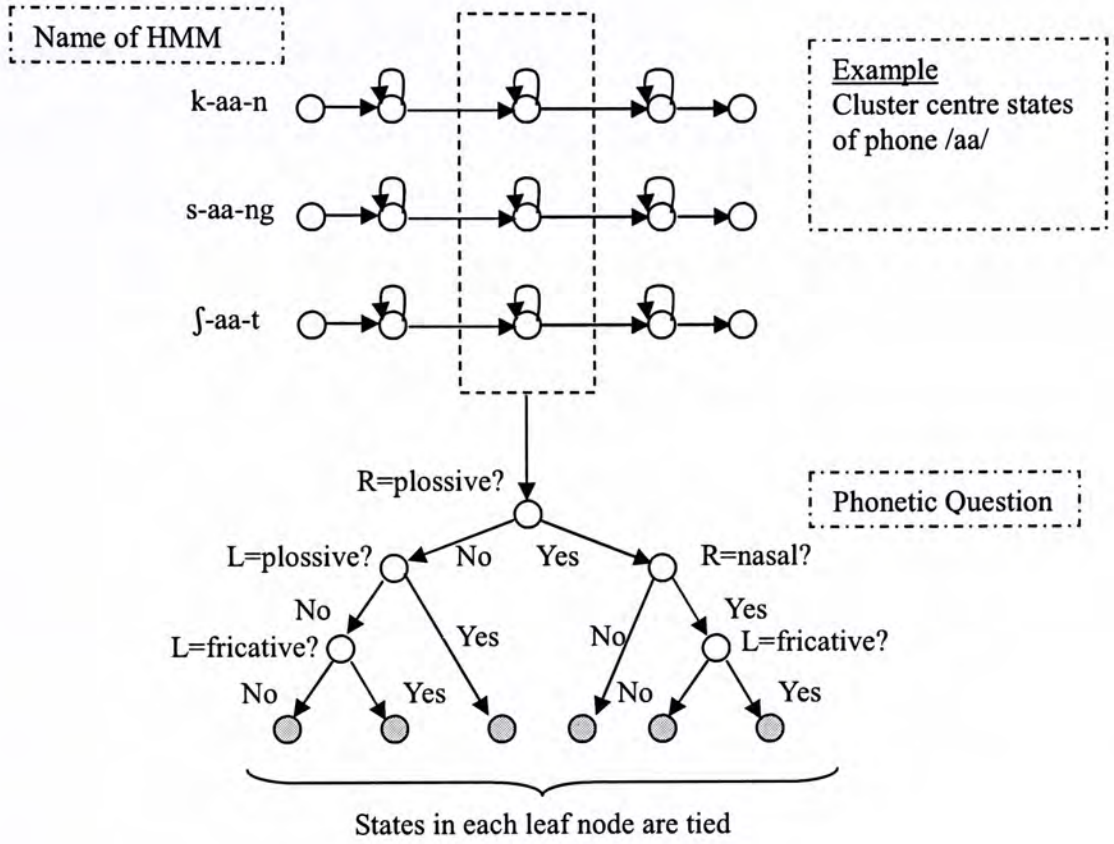


Figure 5-5 Decision-tree based clustering

The tri-phones that do not appear in the training data but possibly appear in the testing data are modeled by the logical models [10]. These logical models are also tied by the decision-tree based clustering to the physical models which really appear in the training data.

Up to this stage, the states in HMMs are still single Gaussian pdf. Since there are different types of variations, the observations associated with a particular state will not in general conform to a single Gaussian pdf, but a linear combination of Gaussian densities [41][42]. Therefore the HMMs are spitted into several Gaussian mixture densities step by step with iterations of HEREST. The number of Gaussian mixtures of each HMM depends on the amount of available training data. If more data is available and the variance is large, more Gaussian mixtures will be applied.

The Three Acoustic Model Sets

The training of HMMs is divided into three stages: 1) Language-dependent acoustic models without accents; 2) Language-dependent acoustic models with accents; and 3)

Language-independent acoustic models with accents.

Stage one – Model Set A

In stage one, all the HMMs are language-dependent, and they are trained by monolingual speech data TIMIT and CUSNET. The pronunciation dictionary DICT01 being used only includes the native pronunciation of English words and Cantonese syllables. Tri-phone models with 16 Gaussians mixtures are trained.

Stage two – Model Set B

In stage two, the HMMs are still language-dependent, but trained by CUSENT and CUMIX. The pronunciation dictionary DICT02 involves native English, accented English and Cantonese syllables. Since there are multiple pronunciations for most of the code-switch words due to allophones and Cantonese accents, the "correct" phone sequence is obtained by force alignment [10]. Phone-based transcriptions are obtained by applying Model Set B to the word-based transcription. The phone-based transcriptions are then utilized for training the new HMMs. The new HMMs are trained from mono-phone to tri-phone, from single Gaussian mixture to 16 Gaussian mixtures.

Stage Three – Model Set C

In stage three, the phone models are language-independent. CUSENT and CUMIX are employed for training, and pronunciation dictionary DICT03 is utilized. Similar to DICT02, English words in DICT03 also includes Cantonese accent, but they are labeled by the language-independent models. The new phone-based transcriptions are obtained from force alignment as well by the HMMs in the previous stage. The HMMs are trained again from mono-phone to tri-phone, from single Gaussian mixture to 32 Gaussian mixtures. The language-independent phone models are cross-lingual; such that speech data from the two languages can be utilized to train the same phone model. Due to the sharing of training data, the HMMs can be represented by more Gaussian mixture densities, such that variations due to language difference can also be modeled.

5.2.7 Decoding and Evaluation

The syllable accuracy is obtained by passing the testing speech input to the bilingual speech recognizer. The syllable and word-level hypothesis is computed by the decoder, and the hypothesis is then compared with the reference transcriptions in order to find the syllable and word error rate.

Two sets of testing data are involved for the bilingual speech recognizer. The first set is Cantonese-English code-mixing utterances, and the second set is monolingual spoken Cantonese utterances. For details, please refer to the previous chapter.

Decoding

The decoder for the bilingual speech recognizer is the tool HVITE in HTK [10]. The decoding is controlled by a recognition network compiled from a word-level network, a dictionary and a set of HMMs. The recognition network consists of a set of nodes connected by arcs, and it is shown in Figure 5-6. It is assumed that the following information is known:

a) Language of the utterance, either Cantonese-English code-mixing or monolingual spoken Cantonese;
b) If it is code-mixing, each utterance consists of one and only one English segment; otherwise, there is no English segment.

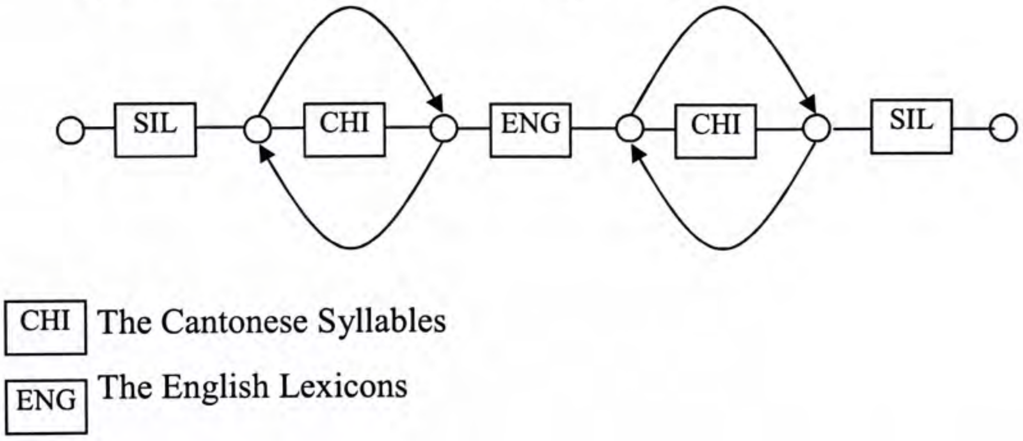


Figure 5-6 Recognition network for HVITE

Each Cantonese syllable and English lexicon is itself a network consisting of HMMs connected by arcs, and each HMM is a network consisting of states connected by arcs. The recognition network levels are illustrated in Figure 5-7.

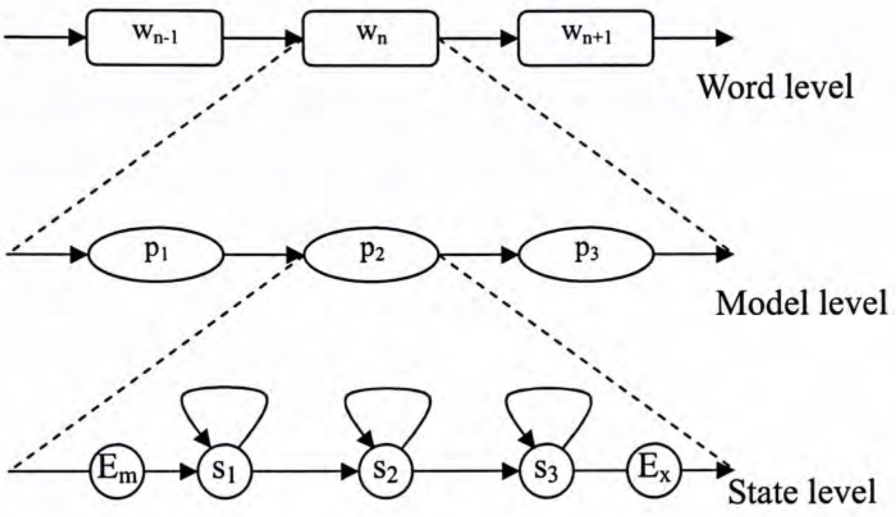


Figure 5-7 Recognition network levels

The job of the decoder HVITE is to find those paths through the recognition network which have the highest log probability. These paths are found using a token passing algorithm, which a token represents a partial path through the network extending from time 0 through to time t. In each time step, the tokens are propagated along connecting transitions stopping whenever they reach an emitting HMM state E_m. When there are multiple exits from a node, the token is copied so that all possible paths are explored in parallel. When the token passes across transitions and through nodes, its log probability is incremented by the corresponding transition and emission probabilities. The network node will hold at most N tokens, such that at the end of each time step, all but the N best tokens in any node are discarded [10]. In this thesis, the value of N is set to 5. If N is large, more memory is required and time for decoding increase due to more computations. If N is small, the average word graph density (WGD) decreases such that only a few alternative syllables or words are available. Only the top-best syllables or words are retained such that the correct syllable and word may be pruned out. A syllable / word lattice that uses nodes and arcs to store the paths is exported. The acoustic likelihood in log scale, start time and end time of each hypothesis word are stored, and the result will be integrated with the language model in the next step.

In order to reduce computation, pruning is applied to keep a record of the best 5 tokens and de-activating all tokens whose log probabilities fall more than a beam-with below the best. For efficiency reasons, primary pruning is implemented at the model rather than the state levels. The pruning beam-width is set to 500 in this thesis.

Word insertion penalty [10] is applied such that a fixed value is added to each token when it transits from the end of one word to the start of the next. The insertion penalty is set to -30 in this thesis.

Evaluating recognition results

The tool HRESULTS from HTK is implemented in order to compare the transcriptions output by HVITE with the original reference transcriptions. Optimal alignment is found, and statistics on different types of errors are output. The syllable / word accuracy is obtained by the following equation:

$$Accuracy = \frac{N-D-S-I}{N} \times 100\% \qquad (5.1)$$

where

N = number of syllables / words in the reference transcriptions

D = number of deletion errors

S = number of substitution errors

I = number of insertion errors

The model set that obtains the highest top-best syllable / word accuracy will be selected for the further process.

5.3 Language Modeling

The details of the language model will be discussed in this section. N-gram language model with N=3 will be built. Code-mixing text data is difficult to be collected, and the problem is serious especially in Cantonese-English code-mixing. Lack of code-mixing text data means the code-switch words in the hypothesis word string may have zero occurrences in the training text data, such that back-off also fails to represent the probabilities of them. In order to tackle this problem, three language models are considered.

5.3.1 N-gram Language Model

Stochastic language model is realized by statistic from large amount of text data. N-gram language models are formulated as a probability distribution *P(W)* over word strings *W* that reflects how frequently a string *W* occurs as a sentence. It assumes that the current word only depends on the previous N-1 words.

5.3.2 Difficulties in Data Collection

It is difficult to collect large amount of text data to train the Cantonese-English code-mixing language mode. The reasons are as follows:

a) Spoken Cantonese is not often used in written text
b) Mixing of spoken Cantonese and standard Chinese
c) Frequency of code-mixing is domain-specific

Spoken Cantonese is Avoided to be Used in Written Text

Cantonese is a dialect rather than a written language, whilst standard Chinese is the official written language in Hong Kong [43]. Notwithstanding that Cantonese is not a standard written language, it can still be found in newspapers, magazines, novels, advertisements, newsgroups and web pages. Nevertheless, it is only regarded as a colloquial and is not used in any official documents. Therefore, they are usually for soft news and quotations, or the diary-like articles [43]. The domain is limited and it is difficult to collect a large amount of text for LVCSR purpose.

Mixing of Spoken Cantonese and Standard Chinese

The spoken Cantonese and standard Chinese are often mixed in written text, such that lexicons and grammars that seldom used in speech may appear. These types of text are not suitable for building the code-mixing language model. Examples can be found in Table 5-11.

Example:
市面上*的*套裝食具，有啲一套有幾個飯碗、有啲有埋筷子前菜碟、中式嘅亦有碗羹套裝，甚至合家歡套裝都有！[44]
閒時最愛*喝*紅酒、食靚嘢*的他*，經常下廚鑽研太史公食譜。[45]
但我哋相反，如果我哋喺十樓俾人投訴，看更嚟到搵唔到人，自然*下*去九樓，但我哋其實喺十一樓。[46]

Table 5-11 Mixing of spoken Cantonese and standard Chinese in written text

Frequency of Code-mixing is Domain-specific

Code-mixing mainly happens in the domains that have frequent interference with other cultures or languages. In Hong Kong, code-mixing is wide spread in the following five domains: 1) computer discourse; 2) business discourse; 3) fashion discourse; 4) food discourse; 5) showbiz discourse [47].

Besides, code-mixing in text often involves standard Chinese but not spoken Cantonese. It is difficult to collect text data that Cantonese-English code-mixing must occur, or it is the only "language" in the text. Thus instead of searching for code-mixing text data, spoken Cantonese data is considered. Among the collected text data, about 10% of them include code-switch words.

5.3.3 Text Data for Training Language Model

The training data for the language models is collected from three main sources: 1) local newspaper; 2) local magazines; and 3) online diaries. The target is to collect text in spoken Cantonese and avoid standard Chinese. Spoken Cantonese is informal written language that people avoid to use in written text, hence not all the text from newspapers and magazines are suitable. At the beginning, we have to search the sections, column names and authors that spoken Cantonese appear. The searched characters are based on the character frequency of standard Chinese [48], such that spoken Cantonese equivalents of the high frequency characters are searched. In order to avoid standard Chinese, 28 characters that commonly adopted in spoken Cantonese but rare in standard Chinese are chosen. The characters in Table 5-12 are spoken Cantonese words being searched. They are searched in OR-based, chosen.

Spoken Cantonese	Sample lexicons (Standard Chinese)
揸	揸車 (駕駛)
噚	噚日 (昨天)
喇	(助語詞)
咩	咩事 (什麼事)
攞	攞獎 (領獎)
俾	俾錢 (付錢)
咗	食咗飯 (吃過飯)
㗎	(助語詞)
諗	我諗 (我想)
乜	乜事 (什麼事)
搵	搵工 (找工作)
嘢	食嘢 (吃東西)
畀	畀人 (被人)
喺	喺邊 (在哪兒)
啲	多啲 (多些)
冇	冇事 (沒有事)
搞	搞錯 (弄錯)
睇	睇書 (看書)
咁	咁多 (這麼多)
佢	佢哋 (他們)
嘅	好嘅 (好的)
係	係唔係 (是不是)
唔	唔好 (不好)
噃	(助語詞)
嚟	返嚟 (回來)
喎	(助語詞)
吓	聽吓歌 (聽歌)
嗱	嗱嗱聲 (快點)

Table 5-12 Spoken Cantonese characters for collecting training data for language model

Subsequent to the searching process, text from sections, column names and authors that spoken Cantonese is commonly used are collected. In order to have contents that cover more other domains, online diaries are also considered, as they are very close to what we speak in daily conversations. The information such as names of newspaper or magazine, column names, author names, type of content, as well as number of characters, is listed in Table 5-13.

Type of publication	Publication	Section / Column name / author	Date	Text size
Local Newspaper	Apple Daily 《蘋果日報 》	《G G細語》左丁山 《隔牆有耳》李八方	01/01/2002 to 29/2/2004	1,646,000
	Oriental Daily News 《東方日報》	《功夫茶》功夫茶		
	The Sun 《太陽報 》	《特區日記》上大人		
	Apple Daily 《蘋果日報 》	Entertainment news (Quotations only)	01/02/1999 to 31/03/2005	2,858,000
Local Magazine	Easy Finder 《壹本便利》	Entertainment news、《青雲路》	Jan 2004 – Mar 2004	300,000
	Hi Tech Weekly 《電腦科技》	Technical News		
	Eat & Travel Weekly 《飲食男女》	Entertainment news		
	Monday 《新 Monday》	Entertainment news		
	Next Magazine 《壹週刊》	Local news and Entertainment news		
	Weekend Weekly 《新假期》	Entertainment news		
	Sudden Weekly 《忽然一周》	Entertainment news	Jan 2002 – Apr 2004	1,984,000
Online diary	The Open Diary http://www.opendiary.com/	9 authors		20,000
			Total size:	6,808,000

Table 5-13 Source of text data for language modeling

To handle the problem that some of the Cantonese characters do not have standard written form [43], all these characters are changed to the most frequent character in the collected data or the one with single pronunciation. This modification is also applied on the training and testing transcriptions to enable the characters being used to be synchronized. Table 5-14 lists some of the modified non-standard characters.

original character	modified to
重(有),、重(好)	仲
尋(日)、擒(日)	噚
(唔)洗，(唔)使	駛
食(左)、瞓(左)	咗
依(家)	而

Table 5-14 Modified non-standard characters

5.3.4 Training Tools

The N-gram language models are trained by the CMU-Cambridge Statistical Language Modeling toolkit Version 2.05 [49]. The toolkit is a suite of UNIX software tools to facilitate the construction and testing of statistical language models. Several discounting schemes are supported and it provides support for N-gram of arbitrary size of N.

In this thesis, tri-gram character-based language model with good turning discounting is trained. The training data involves about 4,600 unique Chinese characters; total size is 6.8M characters.

5.3.5 Training Procedure

After collecting the text data, modifications mentioned in section 5.3.3 are performed. All the titles, headers, footers, reference ID and punctuations are removed from the text. In some of the text data, only words in quotations are used, since standard Chinese are mainly used but the words in quotation are in spoken form. Word segmentation is performed and about 60,000 Cantonese and English lexicons are found from the 6.8M text. Although there are 4.6K unique characters in the training text database, the data in CUMIX only includes 1.2K unique characters. Hence, the Cantonese character list for training only includes the 1.2K characters; all the remaining characters are regarded as Cantonese Out-of-Vocabulary (OOV). To deal with the problem of inadequate code-mixing training data, four language models are proposed. The English words are handled differently in the four language models, and the details are as follows.

Monolingual Language Model (CAN_LM)

The monolingual language model (CAN_LM) only includes N-gram word probabilities of Cantonese. No English words are included in the training text data. All the code-switch words are considered as OOV during evaluation. This method makes the complicated cases simple. Since there is no English word in the training text data, the OOV will have zero-probability. It means that the code-switch words will have much less likelihood than the Cantonese words, and the code-switch words may be missing during recognition.

Code-mixing Language Model (CS_LM)

Due to limited text data, not all the code-switch word in the speech data will appear in the training text, hence there must be OOV for code-mixing language model. One of the solutions is to assume equal probability for all the code-switch words. All the code-switch words in the training text are mapped to the same word ID during the training process. This method can solve the problem of zero occurrences of the code-switch words. Even if the code-switch words do not appear in the training data, their probabilities of occurrence can still be estimated. However, all the code-switch words share the same word ID means that the likelihood of code-switch words will be higher than the other Cantonese characters. Cantonese words may be misrecognized as English words, i.e. there will be higher "false alarm" rate of code-switch words.

Class-based Language Model (CLASS_LM)

The other solution is to cluster the embedded words according to their parts-of-speech (POS), or even, meaning of the words. This method requires tools to label the POS of the words, and may lead to other errors if the labeling is incorrect. In the class-based language model, all the English words are classified into 15 classes as mentioned in Table 5-15:

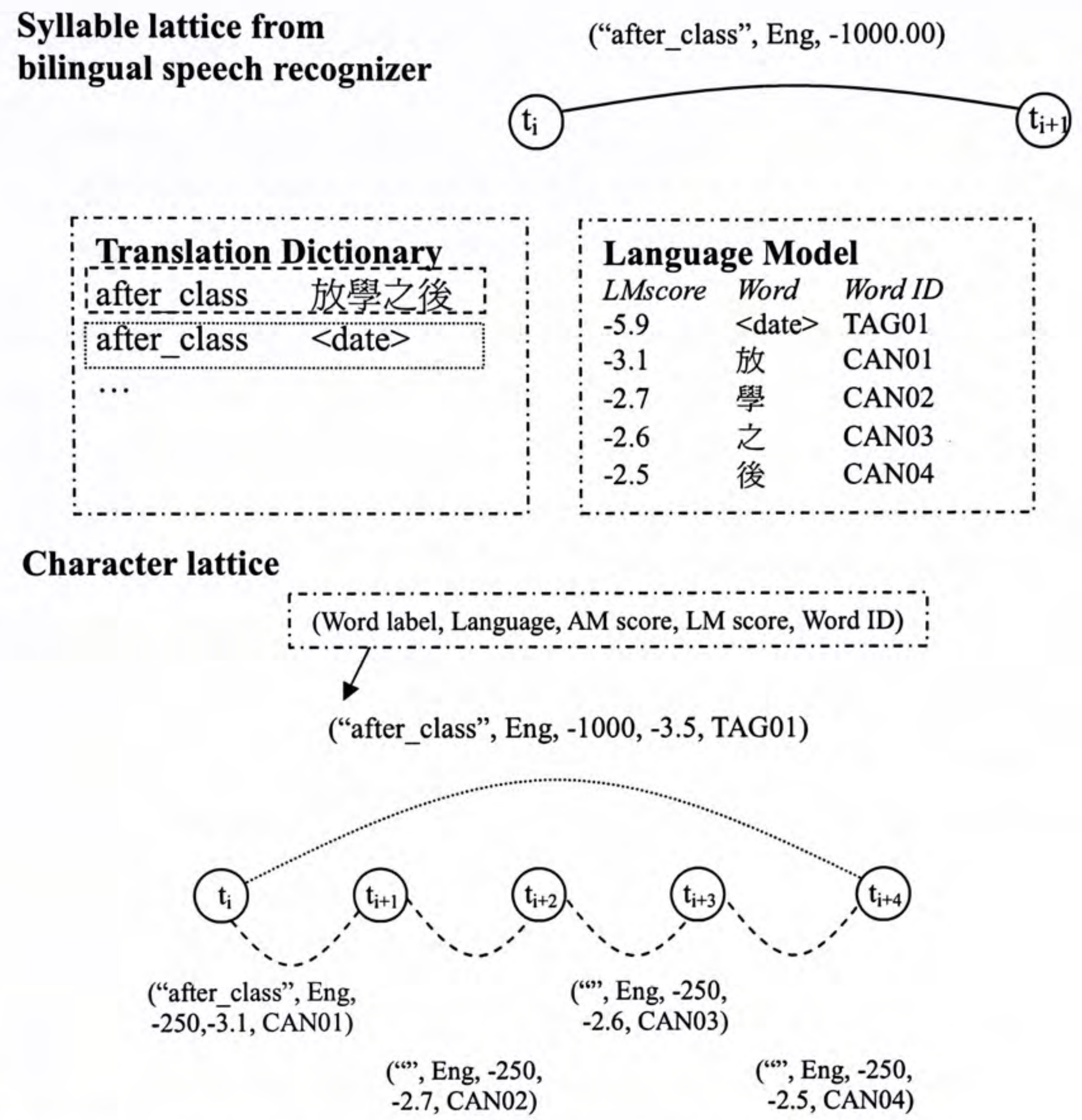
Tag ID	Tag name	Meaning
1	<adj>	adjectives / adverbs
2	<company>	name of companies / organizations
3	<date>	date / time
4	<event>	activities / event
5	<fashion>	terms related to fashion
6	<food>	food / drinks
7	<noun_brand>	brand name
8	<obj>	objects / appliances / tools
9	<people>	human name / title of people
10	<place>	name of place
11	<sent>	sentence / phrase
12	<shop>	name of shops / restaurants
13	<software>	name of software / OS
14	<verb>	verb
15	<noun>	all the remaining nouns

Table 5-15 Tags for class-based language model

The nouns are classified with finer details than the other classes since majority of the code-switch words are nouns [50]. They are categorized into more classes, such that probabilities of each class will not be much higher than the others.

Translation-based Language Model (TRANS_LM)

The forth proposed language model involves English-to-Cantonese translation. However, since not all the code-switch terms have Cantonese equivalent, the classes being used in CLASS_LM will be considered as well. The language model is character-based and the word graph is syllable-based but the Cantonese equivalent may have multiple characters, so some modifications on the syllable lattice are necessary. The arc of hypothesis English word will be divided into *m* arcs, where *m* equals to the number of characters of the Cantonese equivalent. The acoustic likelihood of each of the *m* arcs is the averaged acoustic likelihood of the hypothesis English word. Figure 5-8 shows an example of the translation.


Syllable lattice from bilingual speech recognizer
("after_class", Eng, -1000.00)
t_i
t_{i+1}
Translation Dictionary
after_class 放學之後
after_class <date>
...
Character lattice
(Word label, Language, AM score, LM score, Word ID)
("after_class", Eng, -1000, -3.5, TAG01)
t_i
t_{i+1}
t_{i+2}
t_{i+3}
t_{i+4}
("after_class", Eng, -250,-3.1, CAN01)
("", Eng, -250, -2.7, CAN02)
("", Eng, -250, -2.6, CAN03)
("", Eng, -250, -2.5, CAN04)

Language Model

LMscore	*Word*	*Word ID*
-5.9	<date>	TAG01
-3.1	放	CAN01
-2.7	學	CAN02
-2.6	之	CAN03
-2.5	後	CAN04

Figure 5-8 Integrating the translation-based language model to the word graph

5.3.6 Evaluation of the Language Models

The evaluation method is utilized to measure the performance of the language models: 1) character-based perplexity; and 2) phonetic-to-text (PTT) conversation rate.

Character-based Perplexity

Character-based perplexity is proposed specifically for Chinese language models [51]. Instead of word perplexity, Character-based perplexity is adopted since there is no standard word boundary in Chinese, such that the same piece of text will be treated as different if the vocabulary list is changed. The definition of character-based perplexity B_C for an N-gram model is expressed as:

$$B_C = \exp\left\{-\frac{1}{N_C}\sum_{i=1}^{Q}\log P(w_i \mid w_{i-N+1}...w_{i-1})\right\} \qquad (5.2)$$

where

Q = total number of words in testing data

N_c = total number of characters in testing data

Instead of referring to perplexity as the average number of words following by a word string, character-based perplexity is considered as the average number of characters following by a character string. The character-to-character transition probability within a word is embedded by lexicon, such that it always equals to one. Hence, character-based perplexity shows not only the syntactic constraints by the language model, but also the lexical constraints from the vocabulary. In addition, character-based perplexity normalizes the probability by the number of characters but not the number of words. Therefore, performance of different language models of different vocabulary size can be measured from the same piece of testing text.

Phonetic-to-text Conversation Rate

Phonetic-to-text (PTT) conversation rate measure the character accuracy when the syllables are converted to Cantonese characters. There are homophones in Cantonese therefore each syllable can be mapped to many characters. The mapping mainly depends on the language model, such that given the syllable transcriptions, the character string with the highest language model likelihood is searched. In the decoding process of speech recognition systems, syllables in the word graph are at first mapped to all the possible characters, then the best path is searched according to the acoustic model likelihood and language model likelihood. The PTT conversation rate measurement is very similar to the decoding process, hence it can reflect the performance of the language model.

5.4 Language Boundary Detection

Language identification (LID) is usually performed on multilingual speech recognizer, such that suitable language-specific information such as pronunciation dictionary and model sets can be included. For multilingual speech recognizer, LID is usually implemented by phone-based approach [52-55]. LID is useful for code-mixing speech recognition that efficiency and accuracy can be raised. For efficiency, if language boundaries are known, code-mixing speech recognition can be implemented by switching between two monolingual speech recognizers at the language boundaries. Search space is greatly reduced and thus less computation is necessary for the decoding process. For accuracy, the language score can be regarded as a confidence measurement on the hypothesis words.

Unlike the LID systems for multilingual speech, the problem is complicated for code-mixing speech. For code-mixing speech, the LID problem can be considered as language boundary detection (LBD) problem. The code-switch word may consist of accents and the speech segment of each language is relatively short. The speech segment for the embedded language may only involve a single word or a single syllable, and it is difficult to identify the language with such small amount of data.

In order to detect the language boundaries, three approaches are proposed: 1) phone-based LBD; 2) syllable-based LBD and 3) LBD based on syllable lattice.

5.4.1 Phone-based LBD

LID in multilingual speech is usually realized by phone-based approaches [52-55]. For code-mixing speech, LBD by phones can be realized by applying bi-phone likelihood to phone hypothesis. Bi-phone probability of the matrix language is calculated from monolingual training text data. By applying the bi-phone likelihood on the hypothesis phone sequence, confidence of the languages can be obtained. The likelihood is a measurement that the phone pairs are in Cantonese in this thesis. Low bi-phone likelihood means the phone pairs may be in the embedded language, or at the boundary of the two languages.

Two methods are proposed for phone-based LBD: 1) language-dependent acoustic models with model selection; and 2) language-independent acoustic models.

Language-dependent Acoustic Models with Model Selection (PHONE_LBD01)

Language-dependent acoustic models are trained by two monolingual speech corpora TIMIT [1] and CUSENT [4], and they are all context-independent mono-phones.

The acoustic models in the two languages are clustered by data driven method, and Cantonese acoustic models are used when the two models are similar.

Model Selection by Data-driven Method

It is assumed that all the English models have equal distance from the Cantonese models. Different combinations were tried in order to obtain the optimum LBD result.

There are in total N = 53 English phone models. In the initial trial, N-1 English models and all the Cantonese models are used for phone recognition, and LBD is then performed. The initial trial will iterate N times, such that all combinations are considered. The phone set with the highest LBD accuracy will be examined, and one more English model will be removed from the model set in the next trial. N-2 English models will next be used and N-1 iterations will be tried. One English phone model is removed in each iteration until eventually the optimum LBD result is obtained [56].

Inter-syllable Bi-phone Likelihood for Cantonese

The inter-syllable bi-phone likelihood for Cantonese is calculated from monolingual spoken Cantonese text and pronunciation dictionary. It is the combination of Chinese character probability, and the phone sequence probability of the particular Chinese character.

As there are homographs in Cantonese, the same Chinese character may have different pronunciation when they have different meaning or context. The phone sequence probability of each Chinese character is first calculated, based on the Cantonese lexicon database CULEX [57]. To calculate the probability that a character is pronounced as a particular syllable:

$$P(CH,SYL)=\frac{\text{Total no. of } CH \text{ pronounced as } SLY \text{ in CULEX}}{\text{Total no. of } CH \text{ in CULEX}} \qquad (5.3)$$

where

CH = Cantonese character

SYL= Cantonese syllable

The character frequency is obtained from the 1.6M spoken Cantonese text database, which contains articles from four authors in three local newspapers (Apple Daily, Oriental Daily, The Sun). The language model is mono-gram, thus there is no information on inter-syllable bi-phone probability. It is therefore assumed that all the inter-syllable bi-phone probabilities share the same likelihood. The general equation for intra-syllable probability is given by:

$$P_{LM}(XY) = \sum_{i=1}^{N} P(character_i) \times P(XY \mid character_i) \tag{5.4}$$

where

X and Y= Cantonese phones

XY= the phone sequence that Y follows X

N = total number of unique Chinese characters in the character frequency database

Intra-syllable bi-phone probability is calculated only for Cantonese but not English, since it is the matrix language which large proportion of words in the code-mixing utterances is in Cantonese. All the Cantonese phone models and the selected 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 phone in the English words will either be recognized as English phones or Cantonese phones. They will be recognized as Cantonese phones if the original phone is closed to Cantonese phones. In this way, the English word will be detected due to the low bi-phone likelihood. On the other hand, if there is a large distance between the original phone and the Cantonese phones, no clustering will be applied on the English phone, which means that they will be recognized as English phones.

Development Data Set

The development data set DEV_PHONE involves 600 utterances from three native Cantonese female speakers. The English words in the speech data of these speakers only contain a few Cantonese accents, which are mainly phone change but seldom on phone deletion or insertion. They are part of the training data in CUMIX.

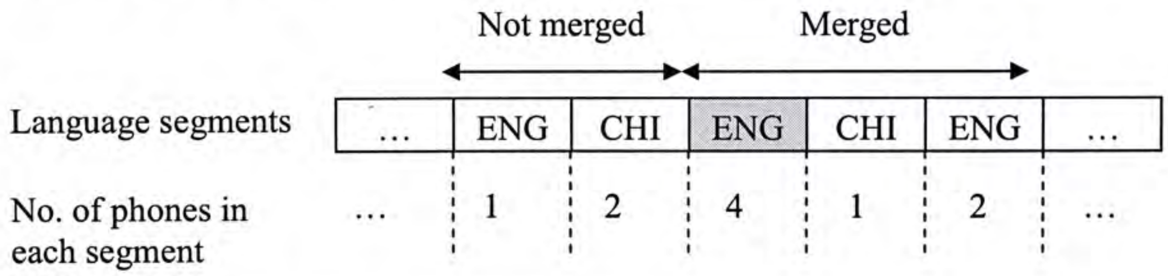
The purpose of the development data is to select the appropriate phone set for phone recognition. They are also used to select the threshold in bi-phone likelihood for language boundary detection.

Language-independent Acoustic Models (PHONE_LBD02)

Since the model selection process is time-consuming, we proposed the use of language-independent acoustic models instead of language-dependent acoustic models. Similar to PHONE_LBD01, the new method employs a phone recognizer and compute bi-phone likelihood of the hypothesis phone sequence. Development data is utilized in order to obtain the threshold for language identification. No model selection is necessary such that it is more efficient than PHONE_LBD01.

Merging of Language Segments

Phone is the smallest unit for acoustic distinguishes. The duration of each phone is short, so there may be many false alarms of language boundaries. It is assumed that each code-mixing utterance only contain one English segment. The hypothesis language segments will be merged according to the number of phones in each segment. The English segment with the largest number of phones or longest time duration will be regarded as the hypothesis English segment. If the neighbouring Cantonese segment only contains a single phone, while the next English segment contains more than one phone, the English-Cantonese-English segments will be merged. Figure 5-9 shows the example of the merging of language segments.



Language segments	…	ENG	CHI	ENG	CHI	ENG	…
No. of phones in each segment	…	1	2	4	1	2	…

Figure 5-9 Example on merging of language segment

Bi-phone probability of the embedded language is not applied since there may be accents from the matrix language. The produced phone sequence in the code-switch words may be different from those in the pronunciation dictionary, therefore the calculated bi-phone probability may be inaccurate.

The difficulty for phone recognition is that if the embedded words include heavy accents from the matrix language, the phone sequence will be similar to those in the matrix language. It will mislead the phone recognizer that it is remained in the matrix language. Besides, it is inefficient to use tri-phone models for reorganization since no constraints can be applied. It needs longer time to decode the phone sequence. On the other hand, if mono-phone is used instead, the efficiency can be greatly improved, but the phone recognition accuracy may be degraded.

The phone models are trained by TIMIT and CUSENT. The phone models are context-independent and language-dependent. Reference language boundaries are obtained by force alignment of the HMMs with correct transcriptions.

5.4.2 Syllable-based LBD

The main difference between LID on multilingual speech and code-mixing speech is that the latter usually includes accents. Accents in the code-switch words may mislead the LID system that it is still in the matrix language. There may be phone change in the code-switch word, and the phonological structure is adapted to the one in matrix language. LBD by phone-based approach may fail to identify the code-switch words that contain accent since the units being considered is short in time duration. To tackle the accent problems, larger units should be considered, such that LBD based on syllable recognition is proposed.

Bi-syllable probability of the matrix language is considered as the confidence measure on the language of the hypothesis words. If the embedded words are pronounced with accents of the matrix language, it will have high acoustic likelihood, but a relatively lower bi-syllable probability. If there is no accent from the matrix language, the acoustic likelihood should be relatively low since the syllable structure of English is quite different from those in Cantonese. That means, if either the acoustic model score, or the bi-syllable score is low, it has higher probability that the syllable pair is in the embedded language, or at the language boundary.

Bi-syllable probability of the matrix language can be computed from monolingual spoken Cantonese text data being used for constructing the language model. The advantage of syllable-based language identification is that more constraints are applied, such the reliability of the hypotheses sound units are higher. However, if the syllable recognizer is inaccurate, the hypothesis syllable sequence may have incorrect bi-syllable likelihood due to the wrong recognition result.

The syllable recognizer only includes syllables from Cantonese. It utilizes the Model Set C trained by CUSENT and CUMIX, which is mentioned in section 5.2.6.

The HMMs are language-independent, context-dependent tri-phone models. Bi-syllable likelihood obtained from text data will be implemented on the hypothesis syllable sequence. If the likelihood of the syllable pairs is large, it has higher probability to be Cantonese; otherwise, they may be in English or at the code-mixing language boundary. A threshold is set for language identification by considering the bi-syllable probability, such that larger than the threshold means it is likely to be Cantonese, otherwise, it is likely to be English. It is assumed that each code-mixing utterance consists of one and only one English segment. If more than one English segment is obtained, the one with longer duration wins. On the other hand, if no English segment is obtained, both the bi-syllable likelihood and acoustic likelihood are considered. The language confidence score is linearly combination of the bi-syllable likelihood and the acoustic likelihood. The threshold is increased incrementally until the English segment includes at least two bi-syllable pairs (single syllable in the English segment). Figure 5-10 shows the block diagram for the bi-syllable based language boundary detector.

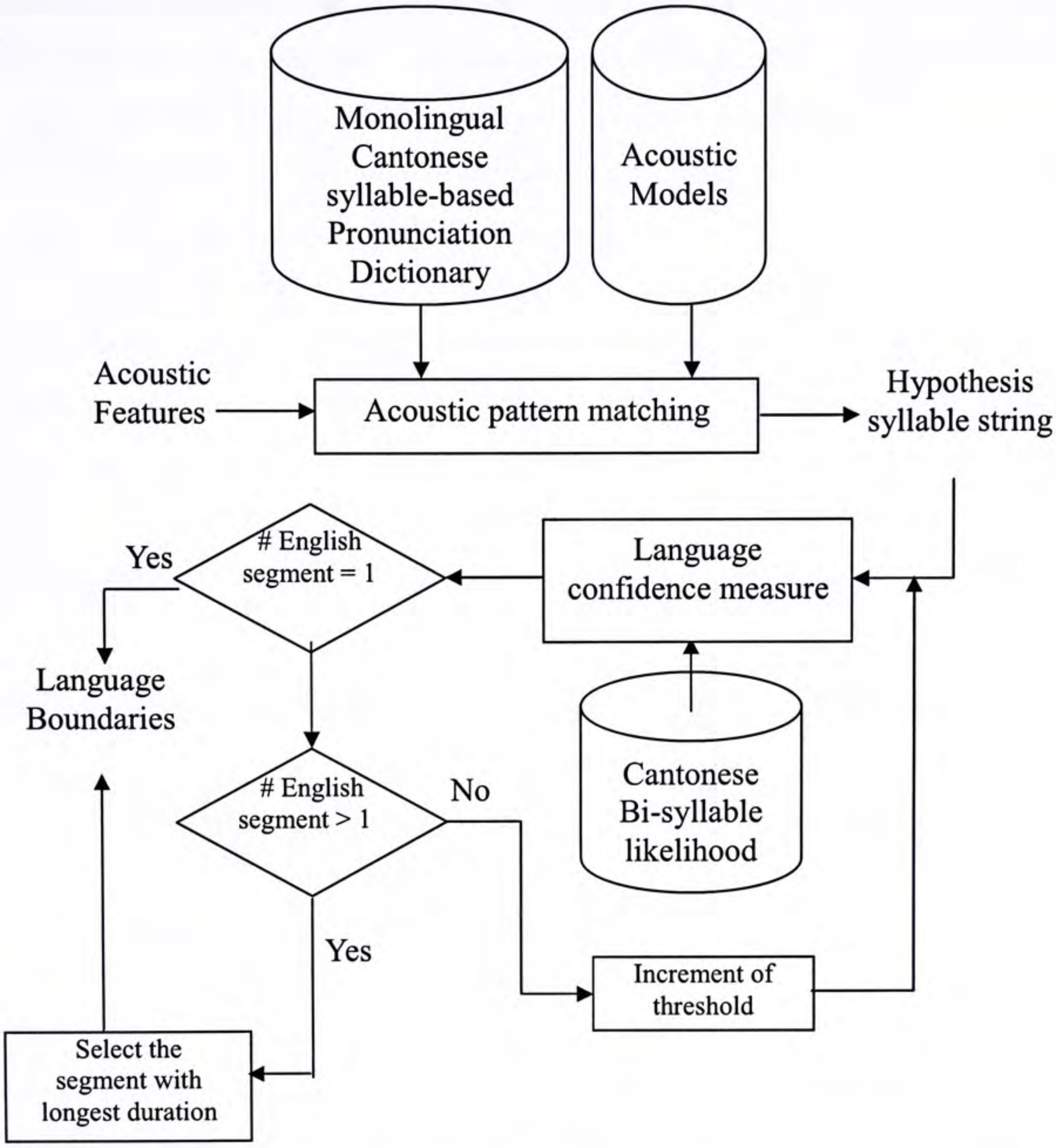

Monolingual Cantonese syllable-based Pronunciation Dictionary
Acoustic Models
Acoustic Features
Acoustic pattern matching
Hypothesis syllable string
Yes
# English segment = 1
Language confidence measure
Language Boundaries
# English segment > 1
No
Cantonese Bi-syllable likelihood
Yes
Increment of threshold
Select the segment with longest duration

Figure 5-10 Language boundaries Detector based on bi-syllable likelihood

5.4.3 LBD Based on Syllable Lattice

The third proposed LBD system does not require any additional speech recognizer but just to extract information from the syllable lattice generated by the bilingual syllable / word based speech recognizer mentioned in Section 5.1.1. Since tokens are utilized in the decoding process, the paths apart from the top-best are also included in the syllable lattice. There is usually more than one English arc in the syllable lattice, hence we can make use of this information as language boundary detection. Time duration is regarded as the confidence measure of the hypothesis English word. We know that English words in average have a longer duration than

Cantonese since they usually contain multiple syllables. With the assumption that each utterance has one and only one English content, the English word with the longest duration in the syllable lattice is likely to be correct. If it is a wrong recognition, it leads to many insertion errors since the English word may be misrecognized as a sequence of Cantonese syllables. Insertion penalty is therefore to avoid this type error by adding a negative constant score to each hypothesis word and syllable. Even the English hypothesis and the sequence of Cantonese syllables have identical acoustic score, the English hypothesis will be selected since the final score includes the insertion penalty. That means hypothesis with longer duration has a higher priority. The start and end time of the English word that has the longest duration is regarded as the boundary of the English content.

5.5 Integration of the Acoustic Model Scores, Language Model Scores and Language Boundary Information

The acoustic model scores, language model scores and language boundary information are integrated in two steps. The first step integrates the acoustic model scores and language boundary information by modifying the acoustic scores in the syllable lattice. The second step integrates the modified acoustic scores with the language model scores to form a character lattice, and the best path is searched by forward-backward algorithm.

5.5.1 Integration of Acoustic Model Scores and Language Boundary Information

After passing the acoustic features to the bilingual syllable-based speech recognizer, a word graph is generated. The word graph includes starting and ending time as well as the acoustic likelihood of each hypothesis word. Since the bilingual dictionary is syllable-based for Cantonese and word-based for English, we call the word-graph "syllable lattice". An example of syllable lattice is shown in Figure 5-11.

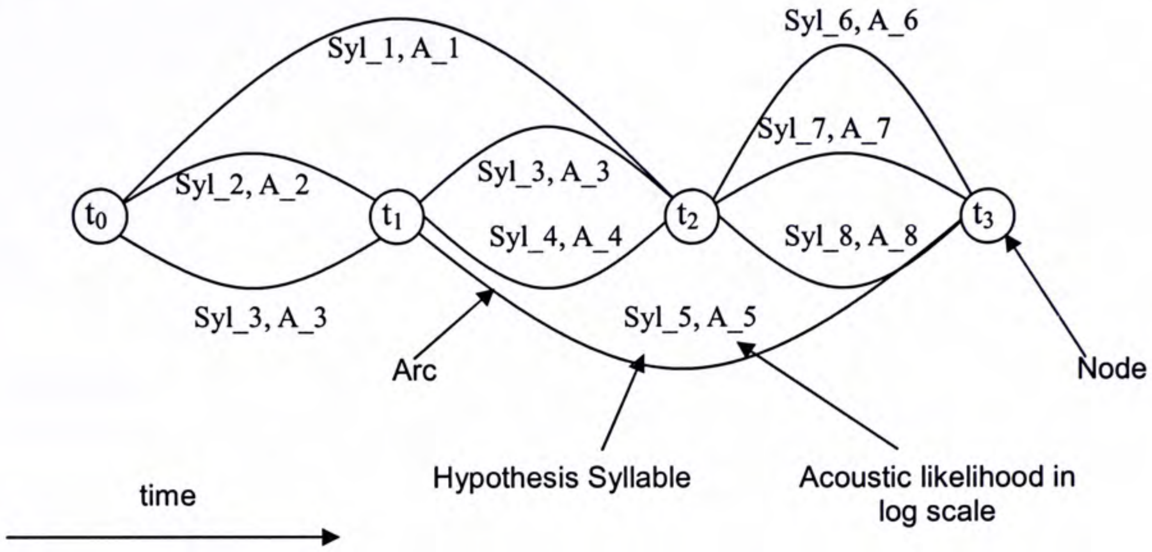


Figure 5-11 Example of syllable lattice

The language boundary information is integrated to the syllable lattice by modifying the acoustic likelihood of hypothesis words. If the hypothesis word is in the same language as the recognized language, a constant score is added to the acoustic likelihood; otherwise, the same constant score is deducted from the acoustic likelihood. The optimum value of the constant score is obtained from the development data such that highest character-based accuracy is performed. An example on the acoustic score modification is illustrated in Figure 5-12.

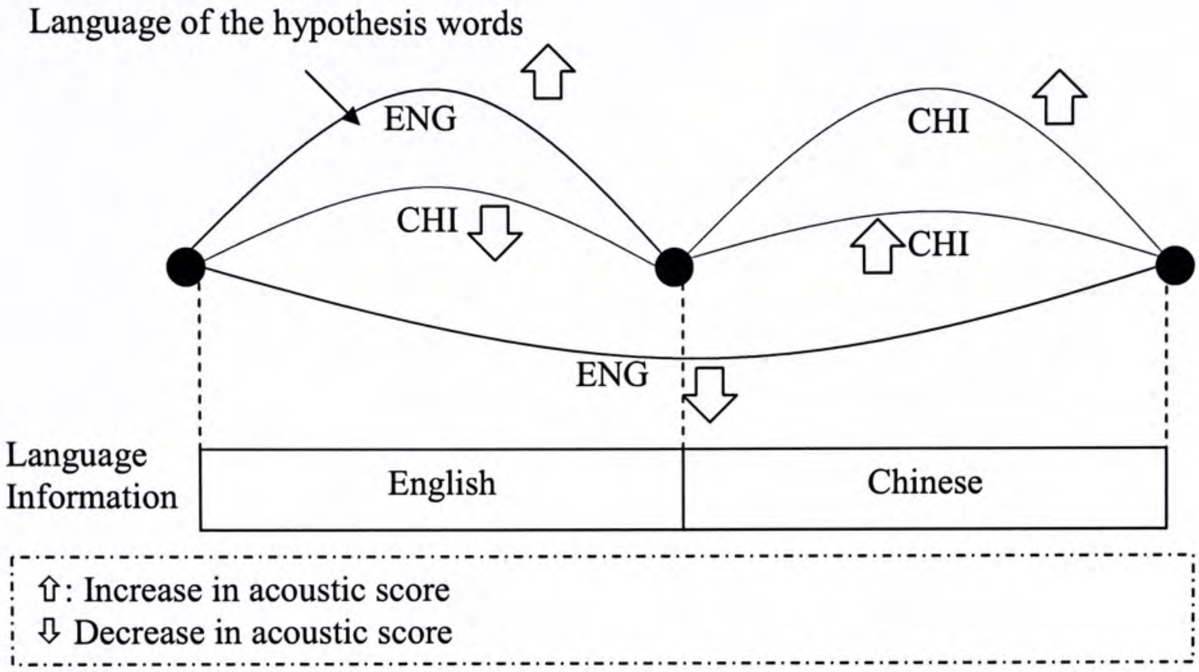

Language of the hypothesis words
ENG
CHI
CHI
CHI
ENG
Language Information
English
Chinese
⇧: Increase in acoustic score
⇩ Decrease in acoustic score

Figure 5-12 Integration of the language information to the word lattice

5.5.2 Integration of Modified Acoustic Model Scores and Language Model Scores

To integrate the language model scores to the modified acoustic model scores, the word lattice is at first converted to character lattice. Cantonese consists of homophones such that one syllable can be mapped to different characters. A character pronunciation dictionary is applied for the syllable-to-character conversion. The pronunciation dictionary is based on the "Cantonese Pronunciation Dictionary" (《粵音韻彙》) [38]. Syllable fusion is considered by including the pronunciation of both the original syllables and the modified syllables. Syllable fusion is considered for the words shown in Table 5-16.

Word	Original Pronunciation	Modified Pronunciation
今日	gam1-jat6	gam1-mat6
噚日	cam4-jat6	cam4-mat6
已經	ji5-ging1	ji5-(g)ing1
冇人	mou5-jan4	mou5-(j)an4
出去	ceot1-heoi3	ceoi1
可唔可以	ho2-m4-ho2-ji5	ho2-m4-mo2-ji5
如果	jyu4-gwo2	jyu4-(gw)o2
成日	seng4-jat6	seng4-(j)at6
而家	ji4-gaa1	ji4-(g)aa1
自己	zi6-gei2	zi6-(g)ei2
但係	daan6-hai6	dai6
即刻	zik1-hak1	zik1-ak1 / zik1-kak1
呢個	nei1-go3	nei1-(g)o3
放學	fong3-hok6	fo3-(h)ok6
知道	zi1-dou6	zi1-(d)ou6
唔好	m4-hou2	m4-mou2
唔係	m4-hai6	m4-mai6
真係	zan1-hai6	zan1-(h)ai6
淨係	zing6-hai6	zing6-(h)ai6
就係	zau6-hai6	zau6-(h)ai6
鍾意	zung1-ji3	zung1-(j)i3
點解	dim2-gaai2	dim2-(g)aai2
邊度	bin1-dou6	bi1-(d)ou6

Table 5-16 Syllable fusion words

For the English words, the conversion depends on the language model being applied. Apart from the word name, each arc also stores a word ID, which is used for searching the N-gram likelihood from the language model. The word name represents the hypothesis word, while the word ID represents the identity of the word in the language model, such that one hypothesis word can be mapped to different word IDs. The mapping depends on which language model is used, such that the word ID of an English word may be the ID of a class, Cantonese character(s), or

OOV.

The syllable-to-character mapping converts the syllable lattice to character lattice. The character lattice is the expanded version of the syllable lattice, and the score in each arc is updated to the GWPP score. Generalized Word Posterior Probability (GWPP) is computed to ensure that word error rate is minimized [58]. The acoustic model score and language mode score are re-weighted, while the GWPP in log scale gives a linear combination of these two scores. Equation 5.5 gives the formula for calculating the GWPP score.

$$P([w;s,t] \mid x_1^T) = \sum_{\substack{\forall M,[w;s,t]_1^M \\ \exists n, 1 \le n \le M \\ w = w_n \\ [s,t] \cap [s_n,t_n] \ne \phi}} \frac{\prod_{m=1}^{M} P^{\alpha}(x_{s_m}^{t_m} \mid w_m) \cdot P^{\beta}(w_m \mid w_1^{m-1})}{P(x_1^T)} \qquad (5.5)$$

where

$w =$ hypothesis word

$s =$ start time of the hypothesis word

$t =$ end time of the hypothesis word

$x =$ observation

$\alpha =$ weight of the acoustic model

$\beta =$ weight of the language model

$P(x_{s_m}^{t_m} \mid w_m)$ is the acoustic model

$P(w_m \mid w_1^{m-1})$ is the language model

The path that has the maximum GWPP score is searched by the forward-backward algorithm. The two weights α and β are tuned by using the development data with a goal to achieve the optimum error rate.

5.5.3 Evaluation Criterion

Performance of the speech recognizer is evaluated by two ways: 1) Syllable accuracy for Cantonese and word accuracy for English; as well as 2) Character accuracy for Cantonese and word accuracy for English.

Output of the bilingual speech recognizer is a word graph, which is syllable-based for Cantonese and word-based for English. Generating word graph is time consuming since the computation is complicated. It has to store the history information of each token, such that more time and memory is needed. To save time, the top-best hypothesis is generated by using single token for each model set. The model set with the highest syllable / word accuracy is utilized for generating the syllable lattice This syllable lattice is then converted to character lattice by the character-based pronunciation dictionary. GWPP score is calculated for each arc of the character lattice, and the best path is searched. The character / word accuracy is finally computed from the top-best hypothesis.

The effect of the code-switch words are studied by comparing the character / word accuracy of the code-mixing speech recognizer to the monolingual Cantonese speech recognizer. For the monolingual Cantonese speech recognizer, the same acoustic models, language models and decoder are employed, but the recognition network for the decoder is different. The recognition network only consists of the Cantonese syllables but not the English lexicons. Character accuracy is compared between the monolingual Cantonese utterances from CUMIX and the code-mixing utterances in the same corpus.

5.6 References

[1] John S. Garofolo, Lori F. Lamel, Willim M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*, NIST Speech Disc1-1.1, NISTIR 4930, 1993

[2] Fisher, William M., Doddington, George R., Goudie-Marshall, Kathleen M., "The DARPA Speech Recognition Research Database: Specifications and Status", in *Proc. of the DARPA Speech Recognition Workshop*, pp. 93-99, Palo Alto, February 1986

[3] Lamel, Lori F., Kassel, Robert H., Seneff, Stephaine, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus", in *Proc. of the DARPA Speech Recognition Workshop*, pp. 100-109, Palo Alto, February 1986

[4] *CUCorpora(TM) - Cantonese Spoken Language Corpora*, Speech Technology Laboratory, Department of Electronic Engineering, The Chinese University of Hong Kong, http://dsp.ee.cuhk.edu.hk/speech/cucorpora

[5] W. K. Lo, Tan Lee, P. C. Ching, "Development of Cantonese spoken language corpora for speech applications," in *Proc. of ISCSLP 98*, pp. 102-107,

Singapore, 1998

[6] Patgi Kam, Tan Lee , "Modeling Pronunciation Variation for Cantonese Speech Recognition", in *Proc. of PMLA 2002,* pp. 12-17, Colorado, USA, September 2002

[7] Wai Yi Peggy Wong, "Syllable fusion and speech rate in Hong Kong Cantonese", in *Proc. of Speech Prosody 2004*, pp. 255-258. Nara, Japan, March 2004

[8] John Butzberger, Hy Murveit, Elizabeth Shriberg, Patti Price, "Spontaneous Speech Effects In Large Vocabulary Speech Recognition Applications", in *Proc. of DARPA Speech and Natural Language Workshop*, pp. 339-343, M. Marcus (ed.), Morgan Kaufmann, 1992

[9] Jared Bernstein, Denise Danielson, "Spontaneous speech collection for the CSR Corpus", in *Proc. of the Fifth DARPA Workshop on Speech and Natural Language*, 1992

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, Dec. 2001

[11] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, *Spoken Language Processing: a Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, 2001

[12] Jim Jian-XiongWu, Li Deng, and Jacky Chan, "Modeling Context-dependent Phonetic Units in a Continuous Speech Recognition System for Mandarin Chinese", in *Proc. of ICSLP 1996*, pp. 2281-2284, PA, USA, October 1996

[13] Wai-Kit Lo, Helen M. Mend and P. C. Ching, "Sub-syllabic Acoustic Modeling Across Chinese Dialects", in *Proc. of ISCSLP 2000*, pp. 97-100, Beijing, China, 2000

[14] Sheng Gao, Tan Lee, Y.W. Wong, Bo Xu, P. C. Ching, Taiyi Huang, "Acoustic Modeling for Chinese Speech Recognition: a Comparative Study of Mandarin and Cantonese", in *Proc. of ICASSP 2000*, Vol. 3, pp. 1261 – 1264, Istanbul, 2000

[15] Gang Peng, William S-Y. Wang, "Parallel Tone Score Association Method for Tone Language Speech Recognition", in *Proc. of ICSLP 2004*, pp. 664 – 667, Jeju Island, Korea, 2004

[16] L Zhang, WH Edmondson, "Speech Recognition Based on Syllable and Pseudo-articulatory Features", in *7th Annual CLUK Research Colloquium*, University of Birmingham, UK, January 2004

[17] Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington and J. Picone, "Syllable-Based Large Vocabulary Continuous Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 4, pp. 358-366, May 2001

[18] M. Hunt, M. Lennig, P. Mermelstein, "Experiments in syllable-based recognition of continuous speech", in *Proc. of ICASSP '80*, Vol. 5, pp. 880-883, April 1980

[19] Chin-Hui Lee, Lawrence R. Rabiner, Roberto Pieraccinit and Jay G. Wilpon, "Acoustic Modeling of Subword Units for Large Vocabulary Speaker Independent Speech Recognition", in *Proc. DARPA Speech & Nat. Lang. Workshop*, pp. 280-291, Cape Cod, Oct 1989

[20] L. Deng, D. Sun, "Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds", in *Proc. of ICASSP-94*, Vol. 1, pp. 45-48, April 1994

[21] A. Sethy, B. Ramabhadran, S. Narayanan, "Improvements in English ASR for the MALACH project using syllable-centric models", in *Proc. of ASRU '03*, pp. 129-134, 2003

[22] Mirjam Wester, "Syllable Classification using Articulatory-Acoustic Features", in *Proc. of Eurospeech 2003*, 2003

[23] A. Constantinescu, G. Chollet, "On Cross-Language Experiments and Data-Driven Units for ALISP (Automatic Language Independent Speech Recognition Processing", in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 606-613, Dec 1997

[24] C. Nieuwoudt and E. C. Botha, "Cross-language adaptation of acoustic models in automatic speech recognition", in *Proc. of AFRICON*, Vol. 1, pp. 181-184, Cape Town, South Africa, 1999

[25] Lorin Netsch and Alexis Bernard, "Automatic and Language Independent Triphone Training Using Phonetic Tables", in *Proc. of ICASSP 2004*, Vol. 5, pp. 297-300, Montreal, Canda, 2004

[26] Lori F. Lamel and Jean-Luc Gauvain, "Cross-Lingual Experiments with Phone

Recognition", in *Proc. of ICASSP 1993*, pp. 507-510, 1993

[27] Joachim Kohler, "Multi-Lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds", in *Proc. of ICSLP 1996*, Vol. 4, pp. 2195-2198, Oct 1996

[28] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Large Vocabulary Speech Recognition", in *Proc. of ICSLP 98*, Sydney, Australia, 1998

[29] Stephen Mattews and Virginia Yip, *Cantonese: A Comprehensive Grammar*, London ; New York : Routledge, 1994

[30] *The CMU Pronouncing Dictionary v0.6*, The Carnegia Mellon University, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[31] Shoup, J. E., "Phonological Aspects of Speech Recognition", in *Trends in Speech Recognition*, edited by W. A. Lea (Prentice-Hall, New York), pp. 125-138, 1980

[32] International Phonetic Association, *Handbook of the International Phonetic Association : a guide to the use of the International Phonetic Alphabet*, Cambridge University Press, 1999

[33] Thomas Hun-tak Lee, *Research on Chinese linguistics in Hong Kong* (《香港漢語語言學研究文集》), Linguistic Society of Hong Kong, Hong Kong, 1992

[34] Gu Yang, *Studies in Chinese linguistics*, Linguistic Society of Hong Kong, Hong Kong, 1998

[35] Robert S. Bauer, Paul K. Benedict, *Modern Cantonese phonology* (《摩登廣州話語音學》), Mouton de Gruyter, Berlin, New York, 1997

[36] K.-F. Lee, S. Hayamizu, H.-W. Hon, C. Huang, J. Swartz, and R.Weide, "Allophone clustering for continuous speech recognition", in *Proc. ICASSP 1990*, pp. 749–752, 1990

[37] S. Sagayama, "Phoneme environment clustering for speech recognition", in *Proc. of ICASSP 89*, Vol.1, pp. 397-400, May 1989

[38] 黃錫凌, 范國, 《粵音韻彙》電子版, 香港中文大學 人文學科研究所 人文電算與人文方法研究室, http://www.arts.cuhk.edu.hk/Lexis/Canton/, October 1998

[39] Arthur Nadas, "Hidden Markov Chains, the Forward-Backward Algorithm, and

Initial Statics", in *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. 2, pp. 504-506, April 1983

[40] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer, "Estimating Hidden Markov Model Parameters So As To Maximize Speech Recognition Accuracy", in *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 1, pp. 77-83, January 1993

[41] B. H. Juang and L. R. Rabiner, "Mixture Autoregressive Hidden Markov Models for Speech Signals", in *IEEE Transactions ASSP-33*, Vol. 6, pp. 1404-1413, December 1985

[42] J. R. Bellagarda and D. Nahamoo, "Tied Mixture Continuous Parameter Modelling for Speech Recognition", *IBM Research Report RC 14516 (#64989)*, March 1989

[43] Don Snow, *Cantonese as Written Language: the Growth of a Written Chinese Vernacular*, Hong Kong University Press, Hong Kong, 2004

[44] "農曆年家中設宴全新餐具款客得體" ,《蘋果日報》，精明消費, 31 Jan 2005

[45] "星級花王　辦盡富豪紅白事", 《壹週刊》，財經, 30 Dec 2004

[46] "2005 避雷大法　揭有線劣爆推銷術", 《壹週刊》，時事, 6 Jan 2005

[47] David C. S. Li, *Issues in Bilingualism and Biculturalism: a Hong Kong Case Study*, Peter Lang, New York, 1996

[48] Ho Hsiu-hwang, *Hong Kong, Mainland China & Taiwan: Chinese Character Frequency - A Trans-Regional, Diachronic Survey*, Research Institute for the Humanities, The Chinese University of Hong Kong, http://humanum.arts.cuhk.edu.hk/Lexis/chifreq, 2001

[49] P.R. Clarkson and R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit", in *Proc. of Eurospeech 1997*, Vol. 1, pp. 2707–2710, 1997

[50] Brian H. S. Chan, *Code-mixing in Hong Kong Cantonese-English Bilinguals: Constraints and Processes*, M.A. in English Language Thesis, The Chinese University of Hong Kong, Hong Kong, China, 1992

[51] H. Law and C. Chan, "Ergodic Multigram HMM Integrating Word Segmentation and Class Tagging for Chinese Language Model", in *Proc. of ICASSP 1996*, Vol. 1, pp. 196-199, 1996

[52] Lori F. Lamel and Jean-Luc Gauvain, "A Phone-based Approach to Non-Linguistic Speech Feature Identification", *Computer Speech and Language*, Vol. 9, No.1, pp. 87-103, January 1995

[53] Kay Margarethe Berkling, *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters*, PhD Thesis, Oregon Graduate Institute of Science & Technology, October 1996.

[54] Yonghong Yam and Etienne Barnard, "Experiments for an Approach to Language Identification with Conversational Telephone Speech", in *Proc. of ICASSP 1996*, Vol. 2, pp. 777-780, 1996

[55] L. F. Lamel, J. L. Gauvain, "Language Identification Using Phone-based Acoustic Likelihoods", in *Proc. of ICASSP 1994*, pp. 293-296, 1994

[56] Joyce Y. C. Chan, P. C. Ching, Tan Lee and Helen M. Meng, "Detection of Language Boundary in Code-switching Utterances by Bi-phone Probabilities", in *Proc. of ISCSLP 2004*, pp. 293-296, Hong Kong, China, December 2004

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

[58] Frank K. Soong, Wai-Kit Lo, Satoshi Nakamura, "Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words", *Special Workshop in Maui*, 2004

Chapter 6

Results and Analysis

This chapter discusses the results of the experiments mentioned in the previous chapter. Experimental results will be analyzed and conclusion will be given.

6.1 Speech Data for Development and Evaluation

All the development data and testing data come from CUMIX [1]. There are three sets of speech data: 1) Cantonese-English code-mixing speech data (CS); 2) monolingual Spoken Cantonese speech data (CAN) and 3) monolingual English words speech (ENG). Usage of the speech data from individual speakers can be found in Appendix C.

6.1.1 Development Data

Speech data from 5 male and 5 female speakers in the testing set of CUMIX are utilized for system development. The development data includes 1100 code-mixing utterances (DEV_CS) and 900 monolingual Cantonese utterances (DEV_CAN). These data allow the optimum weights of language boundary information, acoustic model score and language model score to be derived.

Another set of development data (DEV_PHONE) is employed for selecting thresholds of the phone-based language boundary detector PHONE_LBD01. It involves speech data from 3 female speakers in the training set of CUMIX.

6.1.2 Testing Data

The testing data involves speech data from 10 male and 15 female speakers in the testing set of CUMIX. All the speakers are different from those in the development and training data sets. The testing data includes 2750 code-mixing utterances

(TEST_CS) and 2250 monolingual Cantonese utterances (TEST_CAN). Each of the utterances in TEST_CS contains one and only one English segment. The monolingual testing data TEST_CAN is regarded as the baseline set that enables the effect of code-switch words on recognition accuracy to be eliminated.

The monolingual English words speech data (TEST_ENG) are the English segments in the data TEST_CS. The English segments are extracted from the code-mixing data, while the language boundaries are obtained from force alignment of the transcriptions. The purpose of TEST_ENG is to find the upper bound of accuracy on English words.

6.2 Performance of Different Acoustic Units

Three sets of acoustic models are trained, and their performance is measured by the syllable and word accuracy. Details of the model sets are listed in Table 6-1. Two sets of testing data (TEST_CS and TEST_CAN) are involved, and the results are given in Figure 6-1.

Model Set	Training Data	Type / no. of models	Pronunciation Dictionary
A	TIMIT, CUSENT	Language-dependent (97)	Native
B	CUSENT, CUMIX	Language-dependent (97)	Native and accented
C	CUSENT, CUMIX	Language-independent (73)	Native and accented

Table 6-1 Details of the model sets

Performance of the three model sets are compared in accuracy for the code-mixing utterances and monolingual Spoken Cantonese utterances. Since two languages are involved in code-mixing, the overall accuracy and the accuracy on individual language are also included for comparison.

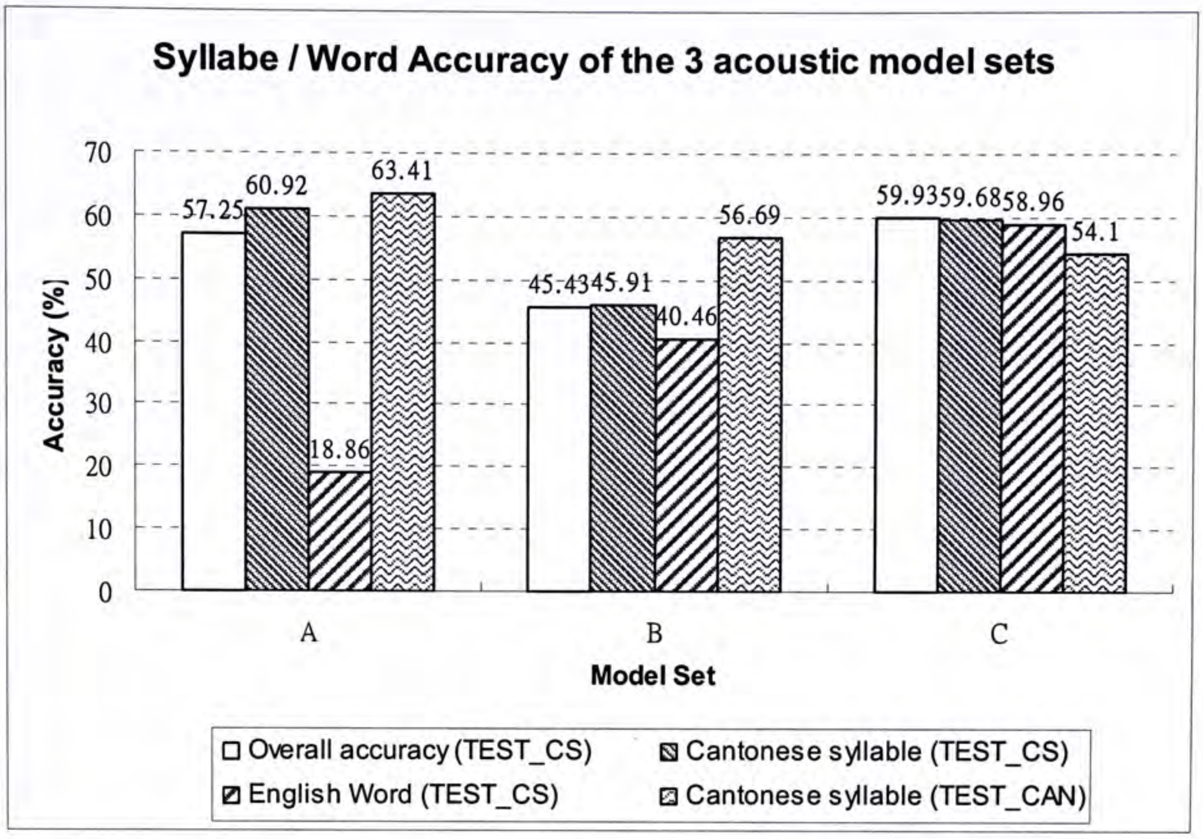


Figure 6-1 Syllable and Word accuracy of the 3 acoustic model sets

6.2.1 Analysis of Results

From the syllable and word accuracy shown in Figure 6-1, it is found that model set C obtains the highest overall accuracy for the code-mixing utterances; whilst model set A attains the highest syllable accuracy for monolingual Cantonese utterances.

<u>Model Set A</u>

Model set A is language-dependent tri-phone HMMs trained by CUSENT [2] and TIMIT [3]. The pronunciation dictionary includes phone sequence of native English and Cantonese syllables. It achieves the highest Cantonese syllable accuracy for both code-mixing utterances and monolingual Spoken Cantonese utterances among the three model sets. The acoustic models are language-dependent, which preserve the language-specific features of Cantonese syllables. However, there is an obvious mismatch of English data between the training and testing corpora because Cantonese accents in testing data are not considered. Due to the mismatch, the word accuracy of English words is very low (18.86%). In code-mixing speech recognition, the accuracies of the matrix language and embedded language are both important. Therefore, modifications should be made on the English phone models in order to handle the accents in English words properly.

Model Set B

Model set B is generated from language-dependent tri-phone HMMs trained by CUSENT and CUMIX. The pronunciation dictionary includes both the native and accented version of English words, therefore the word accuracy of English is higher than model set A. However, the Cantonese syllable accuracy of code-mixing utterances is comparatively low, with about 15% degradation compare with model set A. The reason for such poor performance comes from the language-dependent characteristic of the HMMs. The English words in CUMIX consist of Cantonese accents, in which some of the English phone models are similar to the Cantonese phone models. Similar features are captured by two different phone models, hence the Cantonese syllables are easily misrecognized as English words, and the English words are also easily misrecognized as Cantonese syllables. Due to the above reasons, the overall accuracy for code-mixing utterances is degraded. However, the monolingual Spoken Cantonese utterances are not affected since the pronunciation network only includes Cantonese syllables. The Cantonese syllables will not be misrecognized as words in another language, which leads to a higher Cantonese syllable accuracy than those in code-mixing utterances.

Model Set C

Model set C is obtained from language-independent tri-phone HMMs trained by CUSENT and CUMIX. The pronunciation dictionary includes both the native and accented version of English words. The word accuracy of English is the highest among the three model sets since the accent problems are handled and the similar phones in the two languages are clustered. Each phone model can be modeled by more Gaussian mixture densities due to data sharing between the two languages, therefore the Cantonese syllable accuracy and English word accuracy are higher than those in model set B.

The language-independent models perform less satisfactory than language-dependent models for monolingual speech. The allophones in one language may be considered as different phones in another language, which means if the phone models are not defined properly, the language-independent phone models may be trained by observations that belong to another phone. In this thesis, the clustering is based on IPA and the similarity of phones in manner and place of articulation. Clustering of phone models improve the accuracy of the embedded language, but may degrade the accuracy of the matrix language. Hence, we should strike a balance between accuracy on the matrix language and the embedded language.

6.3 Language Boundary Detection

Language Boundary Detection (LBD) is performed by a phone recognizer, a syllable recognizer, or a bilingual speech recognizer. The hypothesis language boundaries are compared to the reference language boundaries obtained by force alignment [4] on the testing data. If the differences of the two reference boundaries and hypothesis boundaries (Δt_1 and Δt_2) are within the threshold T, it will be regarded as a correct LBD. Either Δt_1 or Δt_2 larger than T will be regarded as wrong LBD. Besides, the hypothesis English segment must overlap with the reference English segment. In this thesis, the value of T is set to 0.3 second, which is about time duration of 1.5 syllables. The definition of LBD accuracy is illustrated in equation (6.1) and Figure 6-2.

$$\text{LBD is correct if } \Delta \mathrm{t}_1 < T \text{ and } \Delta t_2 < T \tag{6.1}$$

where

$\Delta \mathrm{t}_1 = |t_{sr} - t_{sh}|$

$\Delta \mathrm{t}_2 = |t_{er} - t_{eh}|$

t_{sr} = start time of the reference English segment

t_{er} = end time of the reference English segment

t_{sh} = start time of the hypothesis English segment

t_{eh} = end time of the hypothesis English segment

T = time threshold for LBD

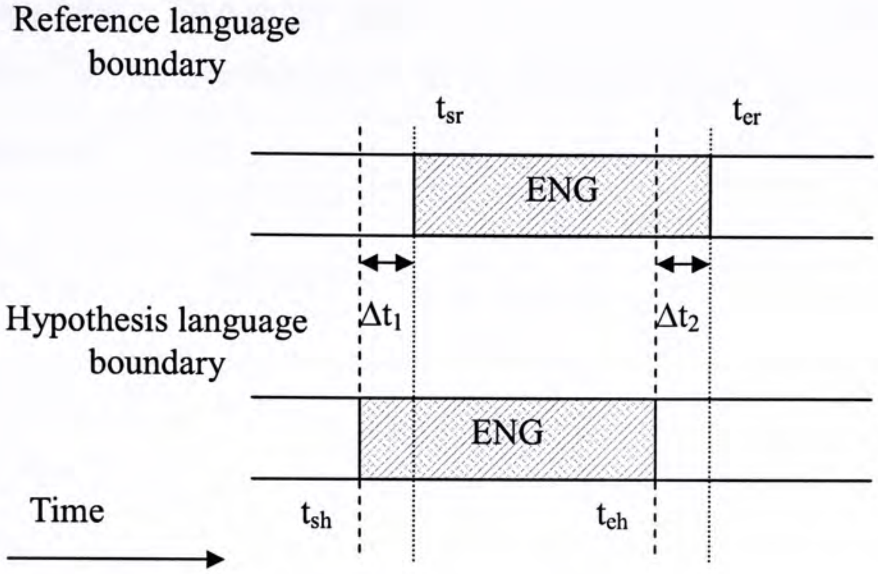


Figure 6-2 LBD accuracy definition

6.3.1 Phone-based Language Boundary Detection

Two sets of acoustic models are involved in the phone-based language boundary detector. The first model set (PHONE_LBD01) employs language-dependent mono-phones trained by two monolingual speech corpora; while the second model set (PHONE_LBD02) employs language-independent mono-phones trained by monolingual and code-mixing speech corpora.

Language-dependent Acoustic Models (PHONE_LBD01)

The acoustic models for phone recognition are selected by data-driven method, and they are trained by two monolingual speech corpora TIMIT and CUSENT. There are 32 Gaussian Mixtures in each state. The phone-based LBD is evaluated by two sets of testing data: 1) true code-switch code-mixing data (CS_TRUE); and 2) borrowing code-mixing data (CS_BORROW). The threshold for the intra-syllable bi-phone LBD and the phone model sets are selected by the same set of development data.

Threshold and the Selected Phone Set

In order to obtain the threshold for language identification and the phone set for LBD, development data DEV_PHONE is applied to the LBD system. The code-switch words in DEV_PHONE only contain a few Cantonese accents so they are true code-switch. One phone is removed in each trial and the details are shown in Table 6-2. The experimental results on the development data are illustrated in Figure 6-3.

Iteration	Removed Phone		Iteration	Removed Phone		Iteration	Removed Phone
1	eh		12	v		23	uh
2	y		13	hh		24	aa
3	ae		14	g		25	axh
4	l		15	n		26	s
5	ay		16	zh		27	z
6	ng		17	er		38	el
7	aw		18	eng		29	nx
8	ih		19	en		30	ax
9	d		20	axr		31	em
10	dx		21	ao			
11	eh		22	uw			

Table 6-2 Removed phones in each iteration for phone-based LBD

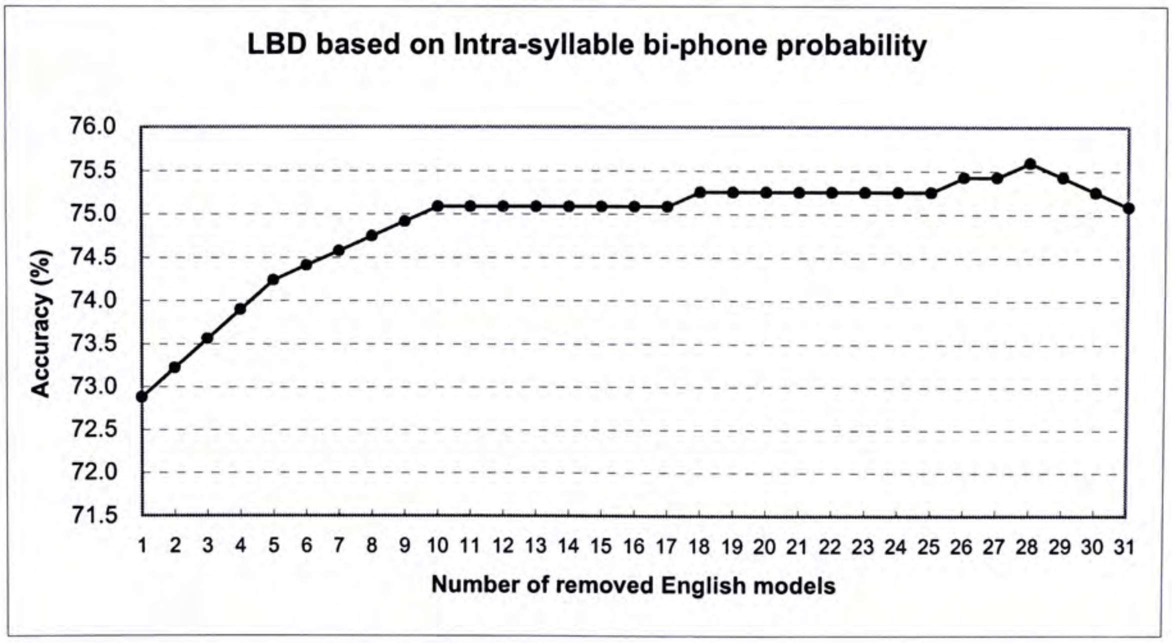


Figure 6-3 Phone-based LBD accuracy (Development data)

The optimum LBD accuracy for the development is 75.6% at the 28th iteration, and we call the model set "PHONE_28". The threshold for bi-phone likelihood is 0.00475, which means phone pair with bi-phone likelihood higher than 0.00475 will be regarded as Cantonese phones. In the first 10 iterations, the improvement in LBD accuracy is significant since the highly confused English phones are removed. The Cantonese phones are no longer misrecognized as English phones and lead to great improvement. The LBD accuracy decreases after the 28th iteration since the phones that carry language-specific features are removed. English phones are misrecognized as Cantonese phones which have relatively higher bi-phone likelihood and lead to more errors.

Evaluation by True Code-switch Data

English words in the testing data CS_TRUE are pronounced with less Cantonese accents relative to CS_BORROW. The accents involved are mainly on phone change and seldom on deletion or insertion of phones. CS_TRUE involves Cantonese-English code-mixing speech data from 3 female speakers in the training set of CUMIX and there are in total 600 testing utterances. The LBD accuracy of CS_TRUE with model set PHONE_28 and threshold 0.00475 is 74.6%.

Evaluation by Borrowing Data

English words in the testing data CS_BORROW are pronounced with more Cantonese accents relative to CS_TRUE. The CS_BORROW data is in fact identical to the CS_TEST data which is used for overall evaluation. The accents involved lead to phone change as well as deletion and insertion of phones. CS_BORROW involves Cantonese-English code-mixing speech data from 15 female speakers and 10 male speakers in the test set of CUMIX. The LBD accuracy of CS_BORROW with model set PHONE_28 and threshold 0.00475 is 52.7%. Table 6-3 summarizes the LBD accuracy of the development data, true code-switch data and borrowing data.

Data Set	No. of utterances	LBD accuracy
Development data (DEV_PHONE)	600	75.6%
True code-switch data (CS_TRUE)	600	74.6%
Borrowing data (CS_BORROW)	2750	52.7%

Table 6-3 Summary on the phone-based LBD accuracy

Language-independent Acoustic Models (PHONE_LBD02)

The model selection of the first approach, PHONE_LBD01, is time consuming and the accents in embedded language are not considered. The training time increase exponentially with the size of development data and number of phone models. Hence, the language-independent phone models are proposed. The language-independent phone models are trained by CUSENT and CUMIX. They are based on the language-independent phone models Model Set C that involves 73 phones, but are mono-phones instead of tri-phones. There are 64 Gaussian mixtures per state for each acoustic model. Since the models are language-independent, no model selection is required and the threshold for language score is the only parameter needed to be optimized. The threshold is selected by code-mixing development data DEV_CS so as to maximize the language boundary detection accuracy.

Threshold Selection

Threshold for language identification is selected by development data CS_DEV. Value of threshold is selected hence the LBD accuracy of the development data is maximized. LBD accuracy versus different value of threshold is shown in Figure 6-4 and Table 6-4.

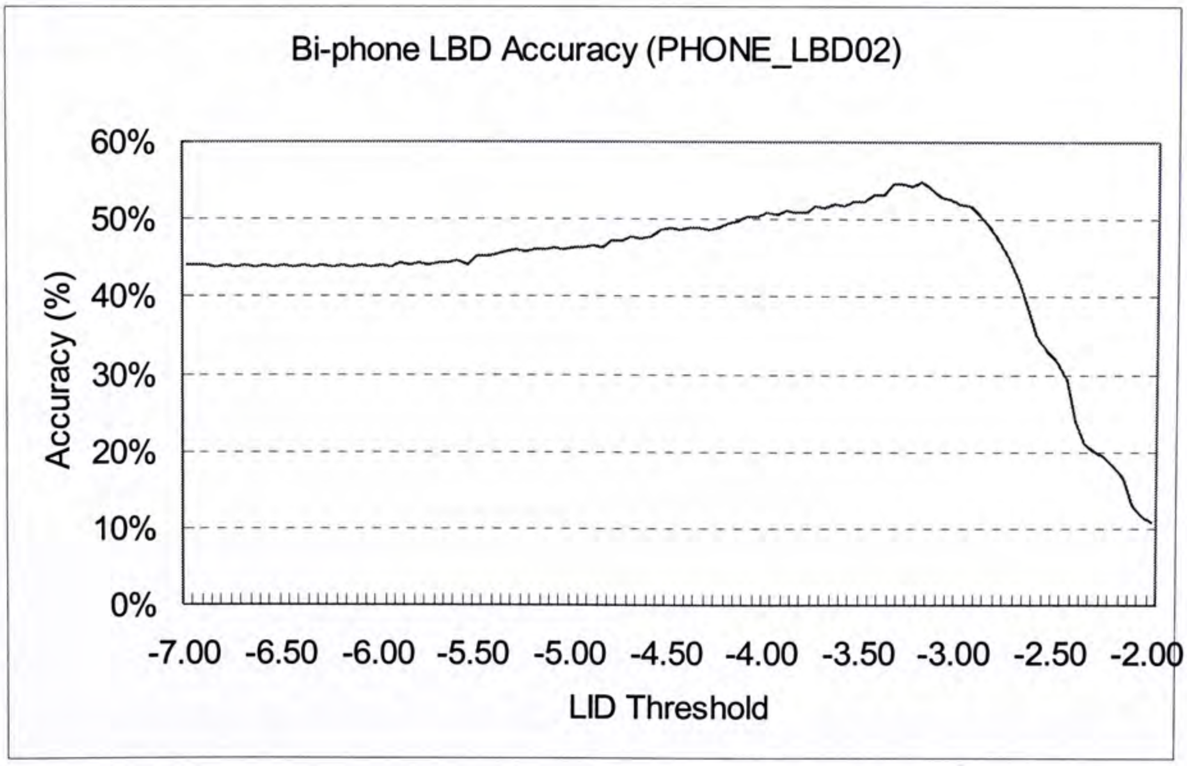


Figure 6-4 LBD accuracy versus different value of threshold (PHONE_LBD02)

Data	Threshold	LBD accuracy
Development Data (CS_DEV)	-3.50	52.5%
	-3.45	53.2%
	-3.40	53.3%
	-3.35	54.6%
	-3.30	54.7%
	-3.25	54.2%
	-3.20	*54.8%*
	-3.15	54.0%
	-3.10	53.0%
	-3.05	52.6%
	-3.00	52.1%
Testing Data (CS_TEST)	-3.20	55.3%

Table 6-4 LBD accuracy versus different value of threshold (PHONE_LBD02)

6.3.2 Syllable-based Language Boundary Detection (SYL_LB)

The acoustic models employed in the syllable-based language boundary detector are language-independent tri-phone models trained by CUSENT and CUMIX. The syllable-based language boundary is evaluated by the code-mixing testing data in CUMIX. The threshold for the syllable-based language identification is selected by the development data in CUMIX. The development data DEV_CS involves speech from 5 female and 5 male in CUMIX, while the testing data involves speech from 15 female and 10 male speakers.

Threshold for the Syllable-based LBD

The threshold for syllable-based language identification is obtained from the language boundary detection accuracy of the development data (DEV_CS). If the probability of the bi-syllable pair in the hypothesis is larger than the threshold, they will be regarded as Cantonese syllables; otherwise, they will be regarded as English syllables or language boundary. Since it is not a must that English segments can be obtained, the threshold will increase incrementally until an English segment with at least one syllable (two bi-syllable pairs) is detected. The bi-syllable likelihood is in log scale, and the increment is a constant = 0.1. Increase in the threshold means more syllable pairs will be regarded as English. The development data is used to tune the initial value of the threshold. The LBD accuracy and the corresponding thresholds are shown in Table 6-5.

Data	Threshold	LBD Accuracy
Development Data	-6.50	65.6%
(CS_DEV)	-6.45	65.3%
	-6.40	65.6%
	-6.35	65.4%
	-6.30	*65.7%*
	-6.25	65.5%
	-6.20	*65.7%*
	-6.15	65.4%
	-6.10	65.6%
Testing data (CS_TEST)	-6.30	65.9%

Table 6-5 Syllable-based LBD accuracy

The maximum LBD accuracy is 65.727%, but it exist at two thresholds -6.30 and -6.20 therefore we have to consider the details on the time alignment difference. Apart from the threshold *T*, the time alignment difference can be further divided into 2 more classes: $\frac{T}{3}$ and $\frac{2T}{3}$. Distribution of the time alignment difference at threshold -6.30 and -6.20 are listed in Table 6-5. The time alignment difference at threshold -6.30 is less than those of threshold -6.20, thus threshold = -6.30 is applied in the further process.

Threshold	Δt1 & Δt2 < T/3	T/3 < Δt1 & Δt2 < 2T/3	2T/3 < Δt1 & Δt2 < T
-6.30	89 (12%)	379 (53%)	255 (35%)
-6.20	88 (12%)	368 (51%)	267 (37%)

* Number of Utterances (% out of 1100 utterances)

Table 6-6 Distribution of the time alignment difference at threshold -6.30 and -6.20

6.3.3 Language Boundary Detection Based on Syllable Lattice (BILINGUAL_LBD)

LBD based on syllable lattice make use of the duration information of the hypothesis English words. The syllable lattice is generated by the bilingual word / syllable based speech recognizer with Model Set C. The start and end time of the English hypothesis with the longest duration is regarded as the boundary of the English content. The LBD accuracy as well as the word accuracy of the English words is listed in Table 6-7.

	LBD accuracy	English word accuracy
Development Data (CS_DEV)	83.2%	60.30%
Testing Data (CS_TEST)	82.3%	60.96%

Table 6-7 LBD accuracy and word accuracy of English words by BILINGUAL_LBD

6.3.4 Observations

From the experimental results, it is found that if the code-switch words only contain a few Cantonese accents, phone-based language boundary detector is preferred since it is more efficient than the syllable-based language boundary detector. The phone-based language boundary detector requires less computation for decoding the hypothesis string. Among the two phone-base LBD system, PHONE_LBD01 requires huge amount of computation for model selection and iterations, hence PHONE_LBD02 is preferred. Although PHONE_LBD01 has good performance on true code-switch data, it is not applicable to the general code-mixing situation since the code-switch words usually contain accents from the matrix language. When the code-switch words consist of accents, its syllable structure is adapted to the matrix language, which means the bi-phone likelihood of the code-switch words is close to those in matrix language. Therefore, syllable-based LBD is preferred since it considers longer sequence of sound units and accents from matrix language can be detected by the low syllable likelihood. This is the reason why the syllable-based language boundary detector gets a relatively higher LBD accuracy than the phone-based language boundary detectors for borrowing speech data.

The boundary detection approach based on syllable lattice attains the best LBD result. It does not require any addition process but just to extract the duration information of English words from the syllable lattice. However, the decoding process is the most complicated one since there are hundreds of syllables and about one thousand English words. The recognition network includes 205K nodes and 562K links, thus more computation time and resources are required. On the other hand, we only need to recognize the utterances once and process on the syllable lattice, so this approach is the most efficient among the four language boundary detectors. The LBD accuracy for the four language boundary detectors is summarized in Table 6-8.

LBD system	LBD Accuracy%	
	Development data	Testing data
Phone-based, monolingual phone models, model-selection	75.6⋆	52.7
Phone-based, cross-lingual phone models	54.8*	55.3
Syllable-based, cross-lingual phone models	65.7*	65.9
Word / Syllable-based, cross-lingual, phone models	83.2*	82.3

⋆ True code-switch data DEV_PHONE

* Borrowing and true code-switch data DEV_CS

Table 6-8 Summary on the LBD accuracy for borrowing speech data

6.4 Evaluation of the Language Models

4 language models are proposed to tackle to problem on lack of code-mixing training data. Character perplexity and phonetic-to-text (PTT) conversion rate are utilized to evaluate the language models.

6.4.1 Character Perplexity

Character perplexity is applied so as to measure the association of the language models and the testing data. It can be considered as the average number of characters following by a character string. The character perplexity of code-mixing testing data in CUMIX is shown in Table 6-9.

Language Model	Character Perplexity
Monolingual Cantonese (CAN_LM)	*74.41*
Code-mixing (CS_LM)	76.39
Class-based (CLASS_LM)	140.44
Translation-based (TRANS_LM)	141.07

Table 6-9 Character perplexity of the 4 language models

6.4.2 Phonetic-to-text Conversion Rate

Another index phonetic-to-text (PTT) conversion rate can also be employed for characterization. The bilingual speech recognizer being used is syllable-based for Cantonese and word-based for English. The language model plays an important role to select appropriate characters for each hypothesis syllable. The more powerful the language model, the more accurate the PTT conversion. Syllable-based transcriptions of TEST_CS are applied to test the PTT conversion rates, and the details are illustrated in Table 6-10.

Language Model	PTT conversion rates
Monolingual Cantonese (CAN_LM)	88.84%
Code-mixing (CS_LM)	89.28%
Class-based (CLASS_LM)	*91.52%*
Translation-based (TRANS_LM)	86.14%

Table 6-10 PTT conversion rates of the language models

6.4.3 Observations

The monolingual Cantonese language model obtains the lowest character perplexity, which means it should be the most similar to the testing data. On the other hand, the PTT conversion rates of the four language models are close to each other since their differences are mainly on the code-switch words. The syllable-based transcriptions already contain the correct code-switch words. Consequently, even the language models treat the code-switch words differently, the code-switch word must be converted correctly. However, probability between the code-switch word and the neighbouring Cantonese character is different for the four language models, therefore different characters may be selected and lead to differences in PTT conversion rate.

The translation approach (TRANS_LM) obtains the lowest PTT conversion rate since some of the translated Cantonese characters originally do not present in the character list of the language models. These new characters have a relatively lower frequency so the probabilities between them and the neighboring characters of the code-switch words are low. The low likelihood affects the decision on the neighboring characters and leads to degradation in the overall PTT conversion rate. Moreover, the code-switch words are translated to Cantonese, while the number of characters of each translated term is usually larger than one. The length of the hypothesis string is changed, which may also affect the PTT conversion rate. A better performance may be obtained if the language models are word-based, such that the translation will have less effect on the length of the hypothesis string.

In real applications, the language model not only converts the syllables to characters, but also makes decisions between Cantonese characters and English words. In the word graph, there are usually multiple arcs which may start or end at the same time, while both the acoustic model score and language model score are employed in the decoding process. Hence, to decide which language model is the most appropriate for code-mixing speech recognition systems, PTT rather than character perplexity is preferred since it can better describe the correlation between the language models and the testing data.

6.5 Character Error Rate

The overall error rate of the Cantonese-English code-mixing speech recognizer refers to the character error rate of Cantonese and word error rate of English. The acoustic models set and language model with the highest accuracies are employed. The best two language boundary detectors are compared as well, and the details are summarized in Table 6-11. Apart from the acoustic model scores and language model scores, there is no constraints in the GWPP decoding process, thus the hypothesis character / word string may contain zero or more than one English contents.

	Name	Description
Acoustic Model	Model Set C	Language-independent tri-phones
Language Boundary Detector 1	SYL_LBD	Based on bi-syllable likelihood
Language Boundary Detector 2	BILINGUAL_LBD	Based on syllable lattice
Language Model	CLASS_LM	Class-based

Table 6-11 Optimum acoustic model sets and language boundary detector for the Cantonese-English Code-mixing Speech Recognition System

6.5.1 Without Language Boundary Information

The language boundary information is not included in the preliminary trial. Only acoustic models scores and language model scores are included in order to compute the GWPP. The GWPP score helps to choose the local optimum in each arc, which means overall word and character error rate can be minimized. The development data CS_DEV is utilized to select the weights of acoustic model score and language model score. The class-based language model is applied and the character and word accuracy of the development data are shown in Table 6-12.

LM AM	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
0.004	54.67	54.50	54.14	54.10	54.11	54.01	53.82	53.75	53.72	53.75
0.005	54.84	54.73	54.58	54.43	54.34	54.11	54.11	53.98	53.97	53.92
0.006	54.82	54.91	54.85	54.73	54.64	54.49	54.52	54.35	54.11	54.12
0.007	54.75	54.79	54.84	54.95	54.85	54.73	54.72	54.57	54.42	54.52
0.008	54.81	54.76	54.74	54.83	54.82	54.97	54.85	54.75	54.76	54.61
0.009	54.57	55.05	54.83	54.82	54.72	54.86	54.85	54.91	54.88	54.79
0.010	54.72	54.88	54.72	54.72	54.73	54.76	54.72	54.64	54.57	54.52

Table 6-12 Overall accuracy (%) for different AM and LM weight (No language boundary information)

From the development data, optimum weight of acoustic model score and language model score are 0.009 and 1.10 respectively. These two weights are then employed for the code-mixing testing data and monolingual Spoken Cantonese testing data. The overall accuracy, Cantonese syllable accuracy and English word accuracy are illustrated in Table 6-12 and Figure 6-5.

	Overall Accuracy (%)	Cantonese Character Accuracy (%)	English Words Accuracy (%)
Development data (CS_DEV)	55.05	55.74	48.41
Code-mixing testing data (CS_TEST)	55.29	56.01	48.40
Monolingual testing data (CAN_TEST)	50.31	50.31	-

Table 6-13 Accuracy of the development data and testing data (no language boundary information)

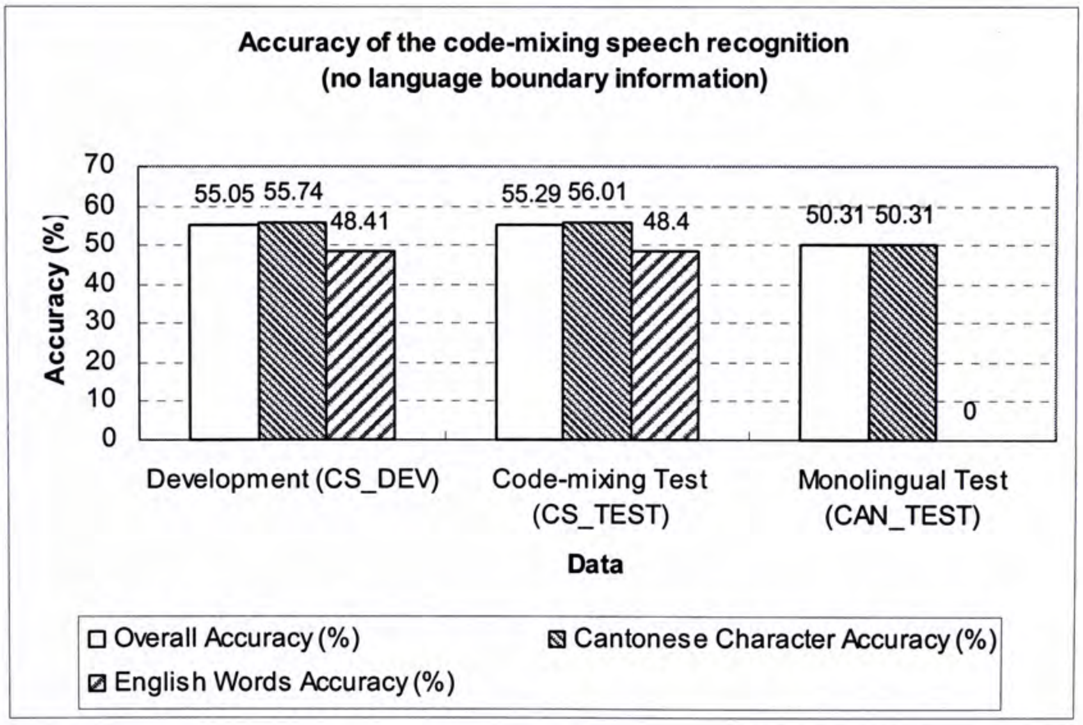


Figure 6-5 Accuracy of the development data and testing data (no language boundary information)

6.5.2 With Language Boundary Detector SYL_LBD

Since the accuracy of the SYL_LBD is relatively low, its effect to the whole system is in doubt. If the language boundary detector can obtain a higher accuracy, it can have positive improvement, but if it is not reliable, it may add adverse effect to the system. Therefore, the weight of the language boundary information should be related to the confidence of the language boundary detector. To simplify the case, a global weight is obtained from the development data CS_DEV for the language boundary information. The fixed weight is added to the acoustic model score in the syllable lattice if the language of hypothesis word is identical to the detected language; otherwise, the fixed weight is subtracted from the acoustic model score. The accuracy of development data with different language information weight are shown in Table 6-13. The weights of acoustic model score and language model score are still 0.009 and 1.1 respectively.

LBD weight	Overall accuracy (%)	Absolute Improvement
80	55.16	+0.11%
70	55.18	+0.13%
60	55.17	+0.12%
50	55.18	+0.13%
40	55.20	+0.15%
30	55.24	+0.21%
20	55.29	+0.24%
10	55.41	*+0.36%*
0	55.05	0

Table 6-14 Overall accuracy (%) and absolute improvement for different LBD weight (SYL_LBD)

From the above result, it is found that the optimum weight of language boundary information is 10. The language boundary information is then included in the syllable lattice for further process. The code-mixing speech recognition system is finally evaluated by Cantonese-English code-mixing testing data, and the accuracies are illustrated in Table 6-15.

	Overall Accuracy (%)	Cantonese Character Accuracy (%)	English Words Accuracy (%)
Development data (CS_DEV)	55.41 (+0.36)	56.27 (+0.52)	47.14 (-1.27)
Code-mixing testing data (CS_TEST)	57.00 (+1.71)	57.56 (+1.54)	49.02 (+0.62)

Table 6-15 Accuracy (%) and absolute improvement of the development data and testing data when language boundary information from SYL_LBD is embedded

From the above experimental results, it is found that the language boundary information brings improvement to the code-mixing speech recognition system but the significance is low. The weight of the language boundary information (±10) is low relative to the insertion penalty (-30). The reason is that accuracy of the language boundary detection system is low, thus heavy weight will lead to wrong decision in the decoding process and produce adverse effect.

The word accuracy of the English words decrease from 58.96% to 48.40% (no language boundary information) and 49.02% (with boundary information) since there is no language constraint in the decoding process. English words are sometime removed or inserted in the decoding process, so the word accuracy degrades.

For Cantonese characters, the character accuracy is close to expectation. The syllable accuracy of Cantonese is 59.68% and the PTT conversion rate is 91.52%. Consequently, the expected character accuracy can be estimated from the following equation:

$$\begin{aligned}&Expected\_character\_accuracy\\&= syllable\_accuracy \times PPT\_conversion\_rate\\&= 59.68\% \times 91.52\% = 54.62\%\end{aligned} \qquad (6.2)$$

From the experimental result, character accuracy is 56.01% when no language boundary information is included, which is close to the expected accuracy.

6.5.3 With Language Boundary Detector BILINGUAL_LBD

No matter for the language boundary accuracy or the English word accuracy, BILINGUAL_LBD achieves the most satisfactory results. We can apply the information from BILINGUAL_LBD to the character lattice by two methods: 1) use the language boundary information to increase likelihoods of all English words in the hypothesis English content; and 2) increase the likelihoods of the particular hypothesis English word so it is more likely to be selected in the decoding process.

As Language Boundary Information

Since the LBD accuracy of BILINGUAL_LBD is relatively higher than those from SYL_LBD, a heavier weight can be applied on the language boundary information. The overall word and character accuracy with different LBD weight are given in

Table 6-16.

LBD weight	Overall accuracy (%)	Absolute Improvement
80	56.12	+0.07%
70	56.18	+0.13%
60	56.26	+0.21%
50	56.35	+0.30%
40	56.43	+0.38%
30	*55.49*	*+0.43%*
20	55.34	+0.29%
10	55.23	+0.18%
0	55.05	0

Table 6-16 Overall accuracy (%) and absolute improvement for different LBD weight (BILINGUAL_LBD)

From the above result, the optimum LBD weight is 30. The accuracy on individual languages is shown in Table 6-17.

	Overall Accuracy (%)	Cantonese Character Accuracy (%)	English Words Accuracy (%)
Development data (CS_DEV)	55.49 (+0.43)	55.81 (+0.07)	52.31 (+3.90)
Code-mixing testing data (CS_TEST)	56.04 (+0.75)	56.37 (+0.36)	52.99 (+4.59)

Table 6-17 Accuracy (%) and absolute improvement of the development data and testing data when language boundary information from BILINGUAL_LBD is embedded

Although the improvement of the boundary information from BILINGUAL_LBD is not significant, the word accuracy of English obtains 4.59% absolute improvement for the testing data. Among the errors in English words, 39.0% of them are deletion error whilst 44.2% of them are substitution error. Deletion error means no English hypothesis appears in the top-best hypothesis string and the English word is missed. Substitution errors are mainly caused by wrong language boundary thus the hypothesis English word and the reference English word has no or just a few overlap in time duration. Sometimes the English hypothesis can only cover part of the reference word, for example, the word "evening" is recognized as "even", and the word "around" is recognized as "round".

Acoustic Score Modification on Hypothesis English Words

The acoustic likelihood of the English hypothesis with the longest duration is set to a negative value close to zero. The acoustic likelihood per frame (10ms) is set to -1.00, hence the total acoustic score of the English hypothesis is much lower than the Cantonese syllables. Acoustic scores of all the other English hypothesis are set to -9999999.00 thus it will less likely be selected in the best path searching. No threshold is necessary and the overall accuracy is shown in Table 6-18.

	Overall Accuracy (%)	Cantonese Character Accuracy (%)	English Words Accuracy (%)
Development data (CS_DEV)	54.14 (-0.92)	53.85 (-1.89)	56.85 (+8.44)
Code-mixing testing data (CS_TEST)	54.95 (-0.34)	56.01 (-1.24)	57.07 (+8.67)

Table 6-18 Accuracy (%) and absolute improvement of the development data and testing data when the longest English hypothesis is selected

It is found that the accuracy on English words is greatly improved when the English hypothesis with the longest duration is selected. However, it is still a bit lower than the English word accuracy in the top-best syllable-based hypothesis. Although modifications have been made on the acoustic scores of the English words, in some cases, the best path still contains no English hypothesis, or more than one English hypothesis. Further constraints are necessary in order to limit the number of English hypothesis in the decoding process. The syllable-based hypothesis is generated by the HVITE [4] decoder from HTK, while the character-based hypothesis is generated by the GWPP decoder. The GWPP decoder can only output the top-best hypothesis and no recognition network can be applied, thus lead to differences in the constraints in the decoding process.

6.5.4 Observations

Although the character and word accuracy is still low, the result is not much worse than the monolingual case. Many of the errors come from the matrix language and the accuracy on English words can sometimes be better than Cantonese characters. Table 6-19 summarizes the syllable, character and word accuracy of all the proposed solutions.

Experiment	Overall accuracy (%)	Syllable accuracy (%)	Character accuracy (%)	English word accuracy (%)
(1) Bilingual speech recognizer with model set C	59.93	59.68	-	58.96
(2) Class-based LM, no LBD	55.29	-	56.01	48.40
(3) Class-based LM & LB info. from SYL_LBD	*57.00*	-	*57.56*	49.02
(4) Class-based LM & LB info. from BILINGUAL_LBD	**56.04**	-	**56.37**	**52.99**
(5) Class-based LM & Longest English word	54.95	-	54.77	*57.07*

Table 6-19 Summary on the experiment results with code-mixing testing data

It is found that when English word accuracy increases, Cantonese character accuracy decreases. To strike a balance between the accuracy in the two languages, the solution that employs language-independent acoustic model, class-based language model and language boundary information from BILINGUAL_LBD (experiment 4) is preferred.

In order to derive the upper bound accuracy on the two languages without the effect of code-mixing, the system is also evaluated by monolingual testing data. In monolingual English case, data set TEST_ENG is applied. When acoustic model set C and monolingual English language models are employed, the word accuracy is 81.07%. It means that if the language boundaries are actuate, the upper limit of word accuracy of English is 81.07% when model set C is utilized. For monolingual Cantonese, a different approach is employed. Although in most cases there is only one English segment in code-mixing speech, we cannot ensure that there must be code-mixing in a real conversation. Therefore, we apply the code-mixing speech recognition system to TEST_CAN in order to test its performance on monolingual

Cantonese speech data. The acoustic models are still model set C, while the language model is CLASS_LM. Language boundary detection system is the one based on start and end time of the longest English word in the word-lattice (BILINGUAL_LBD). Parameters and model sets in experiment 4 are applied on the monolingual Cantonese data, summary on the character accuracy are listed in Table 6-20.

Language	Weight			Character Accuracy		Absolute
Boundary Detector	LBD	AM	LM	Dev. data	Test data	Improvement
Nil	0	0.009	1.1	51.45%	52.87%	-
BILINGUAL_LBD	30	0.009	1.1	51.54%	52.98%	0.12%

Table 6-20 Summary on character accuracy of monolingual test data TEST_CAN

It is found that the character accuracy on monolingual Cantonese speech data is more or less the same no matter the language boundary information are applied or not. Among the errors due to wrong language, 51.98% of the Cantonese characters are misrecognized as monosyllabic English words, while 45.72% of them are misrecognized as disyllabic English words. Since Cantonese is syllable based, the monosyllabic English hypothesis leads to one substitution error while disyllabic English hypothesis leads to one substitution error and one deletion error.

Character accuracy of code-mixing speech data is 56.37% while monolingual Cantonese speech data is just 52.98%. However, it does not lead to the conclusion that the speech recognition system is specially designed for code-mixing data and perform worse for monolingual Cantonese speech. From the syllable accuracy mentioned in Figure 6-1, Cantonese syllable accuracy are 59.68% and 54.10% for code-mixing speech and monolingual speech respectively. The recognition network for Cantonese syllable is already monolingual, that means all the errors come from Cantonese speech itself. It tells that the code-mixing and monolingual speech data are different and it is the baseline accuracy of them. The syllable accuracy already has 5.58% absolute difference and therefore there are 3.39% absolute difference in character accuracy is acceptable.

From the experiments, it is found that language-independent acoustic units trained by monolingual and code-mixing speech data achieve better performance than language-dependent acoustic units for code-mixing speech recognition. Majority of the training data is in the matrix language therefore accents in the embedded language can also be captured by the language-independent acoustic units. Since code-mixing usually occurs in conversation, the speech data is close to spontaneous speech even it is in fact read speech. The high speaking rate and mispronunciation in spontaneous speech lead to additional errors to the code-mixing speech recognition system. Several approaches such as modification in pronunciation dictionary and training transcriptions are proposed in order to tackle the homographs and syllable fusion in spontaneous Cantonese. Since the focus on the thesis is ASR in code-mixing rather than spontaneous Cantonese, we just propose some possible but not optimum solutions to the problems to eliminate their effect.

Instead, one of our main concerns is to tackle the accents problem in the embedded language. The properties of these accents are analyzed and modifications are made in the pronunciation dictionary. The solutions to phone deletion and insertion are automatic, while those for phone change are done manually. It is a preliminary study on the Cantonese accents in English words, hence only the commonly used code-switch words are analyzed.

The main function of applying language model to the syllable lattice is to convert the Cantonese syllables to characters. It no language boundary information is provided, the decoding process often fails to include the English hypothesis thus the word accuracy of English words is relatively lower. In the decoding process, which English hypothesis will be selected mainly depends on the language boundary information instead of the language model. It is because the language model is classed-based that can only tell there should be an English word and the POS of it, but not exactly which word it should be.

To realize code-mixing speech recognition for real-life applications, more spontaneous code-mixing speech data should be collected. Besides, more text data is necessary in order to build language models that are suitable for code-mixing speech recognition.

6.6 References

[1] Joyce Y. C. Chan, P. C. Ching and Tan Lee, "Development of a Cantonese-English Code-mixing Speech Corpus", to be presented in *Proc. of*

Eurospeech 2005, September 2005

[2] W. K. Lo, Tan Lee, P. C. Ching, “Development of Cantonese spoken language corpora for speech applications,” in *Proc. of ISCSLP98*, pp. 102-107, Singapore, 1998.

[3] Fisher, William M., Doddington, George R., Goudie-Marshall, Kathleen M., "The DARPA Speech Recognition Research Database: Specifications and Status", in *Proc. of the DARPA Speech Recognition Workshop*, pp. 93-99, Palo Alto, February 1986

[4] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, Dec. 2001

Chapter 7

Conclusions and Suggestions for Future Work

This chapter gives the conclusion on the overall system as well as suggestions for future work on code-mixing speech recognition.

7.1 Conclusion

Code-switching and code-mixing are common in many bilingual societies and different combinations of languages are involved. Problems due to accents such as phone change, phone insertion and phone deletion should be handled properly as speakers usually include accents from their primary language to the embedded language. In this thesis, several solutions are proposed in order to tackle the problems on automatic speech recognition of Cantonese-English code-mixing utterances. For Cantonese and English, code-mixing is more common than code-switching as the language switching is usually at word level but seldom above clause level. Since studies on code-mixing just started in the recent decades, not much speech and text resources are available. Therefore, we developed the speech corpus CUMIX [1] to provide Cantonese-English code-mixing speech data for the training of acoustic models and analysis on Cantonese accents in English words. Text data with 6.8M characters are also collected for the construction of language models. Pronunciation dictionary is modified as well in order to include the

Cantonese accents in the English words. Language boundary detectors based on different acoustic units are proposed and therefore language information can be obtained. Finally, different combinations of the sub-systems are evaluated by code-mixing and monolingual speech data from CUMIX. The optimum accuracy on Cantonese characters and English words are 56.39% and 52.99% respectively.

7.1.1 Difficulties and Solutions

Code-mixing speech is quite different from monolingual speech. Apart from language switching, accents and grammar difference lead to additional problems. Moreover, code-mixing in text materials are sometimes quite different from those in speech, which may lead to lack of training data for language models.

Acoustic Modeling and Pronunciation Modeling

Difficulties

The phonetic inventory and the phonological structure are different for each language. Allophones in one language may be considered as different phones in another language, and therefore universal phone set should be applied. Besides, the code-switch words usually contain accents from the matrix language and therefore phone change, phone insertion and phone deletion may occur. For phone change, the changes may not be one-to-one but sometimes speaker and context dependent. The code-switch words will be difficult to be recognized if the accents are not considered in acoustic modeling and pronunciation modeling.

Solutions

In order to cover all the phonetic inventories in the two languages, language dependent acoustic models are proposed. However, accents in the embedded language may lead to merging of similar phones, and therefore language independent acoustic models are then suggested. The units of acoustic models are based on IPA [2], which is universal phone set that suitable for many languages. If there is no code-mixing speech data for training, we can use monolingual speech corpora instead (CUSENT [3] and TIMIT [4]). However, if there are mismatch between the training and testing data, the performance may be worse. Hence, we proposed the use of monolingual Cantonese corpus CUSENT and Cantonese-English code-mixing corpus CUMIX to train the acoustic models. Both language dependent and language independent approaches on acoustic modeling are studied in order to obtain the optimum solution.

To recognize English words with Cantonese accents, modifications are necessary for the English entries in the pronunciation dictionary. The current approach is to handle phone deletion and phone insertion automatically since their occurrences have regular patterns. The English syllable structures are usually adapted to the (C)V(C) structure in Cantonese. For phone change, we now handle it manually as we have not yet found a regular mapping between the phone pairs.

Results and Observations

From the experiments, it is found that language independent acoustic model set trained by CUSENT and CUMIX outperform the other acoustic models. Its performance on both Cantonese syllable and English words is similar, which the accuracy is around 59%. If language boundary information is provided, accuracy on English words can be up to 81.07% when language independent acoustic models are utilized. If monolingual speech corpora are employed, accuracy on English words is very low since the English training data in TIMIT are provided by native speakers and therefore Cantonese accents are not included. It is obvious that there is mismatch between training and testing data, so the use of code-mixing speech data is preferred.

Language Modeling

Difficulties

Code-mixing is domain specific that mainly occurs in areas that have high frequency of interaction with the western culture. Moreover, Cantonese is a dialect which is not often used in written text, and therefore the size of collected text data is quite small. Mixing between written Chinese and spoken Cantonese may sometimes occur in text material that lead to combination of words that seldom appear in speech. Besides, some of the colloquial words in Cantonese cannot be written in Chinese characters and people usually use English words with similar sound instead. Hence, these words should be distinguished from code-switch words as we just borrow their pronunciation but not the meaning.

Solutions

In order to collect code-mixing text data between spoken Cantonese and English, several sources such as newspapers, newsgroups and online diaries are considered. Since the lexicons of written Chinese and spoken Cantonese are quite different, the selection of articles is based on the high frequency spoken Cantonese characters. In order to remove the written Chinese that rarely appear in speech, word segmentation is performed on the collected data. If the characters unique to written Chinese are found, the whole utterance will be removed.

Since code-mixing is domain specific but there is no available list on code-switch words, the data collection is based on Cantonese instead of English. Among the collected data, 10% of the utterances are code-mixing. Although the collected data involves 8,000 code-switch words, they still fail to cover all the code-switch words in the testing data which is collected mainly from online diaries. Data from online diaries are the text version of speech that we use in daily life, while data from newspapers are mainly reporting or introducing new information. Code-switch words involved are a bit different and therefore zero occurrences may appear for some terms. To solve this problem, several approaches are proposed to handle the code-switch words. The most efficient solution is to classify the code-switch words to 13 categories according to their part-of-speech. Code-switch words in the same class share the same tri-gram likelihood in the language model.

Results and Observations

No matter for phonetic-to-text (PTT) conversion rate or the character accuracy, the class-based language model obtains the optimum result. The PTT conversion rate is 91.52% for code-mixing test data in CUMIX. For the recognition hypothesis, syllable accuracy is 59.68% while character accuracy is 56.01%, which means the PTT conversation rate is up to 93.85%. The improvement mainly comes from the character lattice since other possible hypotheses with lower acoustic likelihood are available as well. The decoding process is based on GWPP score which consider both the acoustic scores and language model scores, where their weights are tuned in order to obtain an optimum accuracy. Therefore, a higher PTT conversation rate can be realized from the character lattice.

Language Boundary Detection (LBD)

Difficulties

Two languages are involved in code-mixing utterances. If language boundary information is provided, overall accuracy can be improved since the hypothesis will less likely to be in the wrong language. However, switching between languages is mainly in word level for code-mixing speech, therefore time duration of the embedded language is usually short. Thus, it is difficult to perform language identification in code-mixing utterances.

Solutions

Several approaches are proposed on LBD by employing different acoustic units. Among the proposed solutions, the one that based on word and syllable lattice achieves the highest LBD accuracy. The lattice is generated by a bilingual speech recognizer which is word based for English and syllable based for Cantonese. The English hypothesis with the longest duration in the lattice is searched and its start and end time is regarded as the hypothesis language boundaries. The acoustic likelihood in the lattice will then be re-scored according to the hypothesis language boundaries. If the hypothesis word in the lattice is in the same language as the hypothesis language information, positive score will be added, hence the arc will be more likely to be selected. Otherwise, the same score will be deducted from the arc and therefore the arc will have lower chance to be selected in the best path searching.

Results and Observations

The lattice based language boundary detector brings absolute improvement of 0.36% and 4.59% to Cantonese character and English word accuracy respectively. The main reason of the improvement is that there is no language constraint in the best path searching process and therefore many of the code-switch words are missed. With language information, likelihood of the code-switch words is increased and therefore they have better chance to be selected in the decoding process.

7.2 Suggestions for Future Work

The proposed speech recognizer in this thesis is based on language-independent acoustic models, word-based (English) and character-based (Cantonese) pronunciation dictionary, character and class based language models as well as language boundary detector based on bilingual lattice.

7.2.1 Acoustic Modeling

For acoustic modeling, more studies can be applied on clustering of phone models across the two languages. If the distance between two phones is large but they are clustered to the same phone, acoustic data from different phones are utilized to train the same model. Parameter variations increase and it may lead to more errors. Different sub-word units can also been considered so as to maximize the reliability of the acoustic models. Language-independent acoustic models are preferred since each of the models can be trained by more data. However, the clustering should be correct and therefore the models will not be trained by wrong data.

7.2.2 Pronunciation Modeling

The current pronunciation dictionary is word based for English and syllable based for Cantonese in the first pass. Other units such as syllable-based for both the languages or word-based for Cantonese can be considered as well. Syllable based recognition with language model can reduce decoding time. The language model constraints the combination of syllables, and therefore number of entries in the pronunciation dictionary can be reduced. Although there are thousands of syllables in English, when Cantonese accents are included, the number can be greatly reduced. There will be about one thousand syllables for the two languages for the CUMIX corpus.

Besides, the current approach to include Cantonese accents in the English words is partially manual, which is inefficient if the vocabulary size increases. More studies are necessary in order to find the characteristics of phone changes in English words due to Cantonese accents.

7.2.3 Language Modeling

The current language model is based on a 6.8M character database. Although the PTT conversation rate is quite high when character based language model is applied, more data is still necessary if word based language models are built. Moreover, larger text database means more code-switch words can be involved, so more categories can be applied on them. The class based approach is in fact a mixture of N-gram language model and Context-free Grammar (CFG). When more text data is available, CFG can also be applied on Cantonese, and therefore the classes can be trained by more data.

7.2.4 Speech Data

The code-mixing speech data being applied in this thesis only contain one English segment in each utterance. In fact, the CUMIX speech corpus also includes code-mixing speech data that involves two English segments. For further research, different types and frequencies of code-mixing speech data should be considered. Moreover, spontaneous speech should be collected as code-mixing mainly occurs in conversation but not read speech. Therefore, spontaneous speech should be considered.

7.2.5 Language Boundary Detection

All the language boundary detection methods in this thesis are based on acoustic recognition results. In fact, there are also other solutions, for example, MAP estimation based on LSA-based GMM and VQ based bi-gram [5]. Studies on alternative solutions may be considered for further research.

7.3 References

[5] Joyce Y. C. Chan, P. C. Ching and Tan Lee, "Development of a Cantonese-English Code-mixing Speech Corpus", to be presented in *Proc. of Eurospeech 2005*, Lisbon, Portugal, September 2005

[6] International Phonetic Association, Handbook of the International Phonetic Association : a guide to the use of the International Phonetic Alphabet, Cambridge University Press, 1999

[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", in *Proc. of ICASSP 1997*, Vol.3, pp. 1775-1778, Munich, Germany, 1997

[8] John S. Garofolo, Lori F. Lamel, Willim M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*, NIST Speech Disc1-1.1, NISTIR 4930, 1993

[9] Chi-Jiun Shia, Yu-Hsien Chiu, Jia-Hsin Hsieh, Chung-Hsien Wu, "Language boundary detection and identification of mixed-language speech based on MAP estimation", in *Proc. of ICCASSP 2004*, Vol. 1, pp. 381-384, 2004

Appendix A

Code-mixing Utterances in Training Set of CUMIX

1. 用 inkjet printer 印數碼相得唔得㗎？
2. 其實 inkjet printer 印出嚟啲嘢係咪靚啲㗎？
3. 佢 year one 嘅時候啲頭髮好長。
4. 我今年讀 year one。
5. 加上轉眼間又完結 year two。
6. 你 year two 嗰陣唔係去過交流咩？
7. 邊度有得學畫 3D 嘅電腦動畫呀？
8. 佢部機嘅開機畫面係一幅 3D 嘅公仔。
9. 一部 3G 手機要幾多錢？
10. 聽講 3G 啲電話成日斷線喎。
11. 佢會考有九個 A㗎！
12. 你今次考試又攞咗幾多個 A 呀？
13. 用縮寫之前記得寫番 abbreviation。
14. 我睇睇吓 abbreviation 又瞓着咗 。
15. 你有冇足夠 ability 去應付㗎？
16. 有時我都會懷疑自己嘅 ability。
17. 你嘅做法好 abnormal。
18. 我不知幾正常，你就 abnormal！
19. 佢尋日 absent。
20. absent 超過兩堂就一定肥硬。
21. 通常 abstract 係簡要地將最重要嘅嘢寫出嚟。
22. 寫公函嘅時候要具體啲同埋避免 abstract。
23. 佢接受你嘅申請嗰陣已經知道你嘅 academic background。
24. 叫得你去面試，已經証明咗你嘅 academic background 合乎佢哋要求。
25. 我 accept 之前都已經問過好多人意見。
26. 大家都一致贊成 accept，希望唔會後悔。
27. 我 access 唔到佢部機。
28. 點樣先至可以 access 到宿舍部機？
29. 佢都睇咗唔少 accessories。
30. 我會用多啲 accessories 嚟襯托。
31. 可以喺邊度攞到 accommodation 嘅資料？
32. accommodation 係自己俾定學校俾？
33. 我同房讀 account。
34. 我個 account 好似有啲問題。
35. 我都冇開到 account。
36. 請喺呢度寫銀行名同 account number 吖。
37. 你張提款卡上面咪有 account number 囉。
38. 你個 accuracy 係幾多呀？
39. 佢個 accuracy 都唔知點計出嚟嘅。
40. 你嘅推測好 accurate 喎。
41. 其實呢個方法好唔 accurate，不過好過冇咁啦。
42. 論文入面要寫 acknowledgment㗎。
43. 佢居然唔記得喺 acknowledgment 入面寫老細個名。
44. 你哋平時有啲咩 activity㗎？
45. 佢哋啲 activity 其實唔係咁啱我。
46. actually 我唔知發生緊咩事。
47. actually 呢啲都唔係咩秘密。
48. 今日一開有好多小朋友 add 我。
49. 當然仲有啱啱 add 咗冇耐嗰啲。
50. admin 通知可以收工。
51. 我唔係 admin 所以好多嘢都做唔到。
52. 你嗰科嘅 admission grade 係幾多？
53. 今年嘅 admission grade 已經出咗，原來我係最尾嗰幾個。
54. 係咪間間公司嘅寬頻都係 ADSL 嚟㗎？
55. 我屋企個寬頻係咪 ADSL 嚟㗎？
56. 你估人人都 afford 得起住宿舍㗎？
57. 啲居民 afford 唔起買公屋。
58. 要 after 復活節之後先至有期。
59. 佢要 after 十月先至得閒喎。
60. 我打算星期五 after class 再去。
61. 我哋次次約食飯都係 after class㗎啦。
62. 我好 agree 佢嘅講法。
63. 唔一定要佢 agree 嘅。
64. 大家一定要睇我個 album。
65. 呢個 album 幾好用喎。
66. 事實上係真係唔夠 alert。
67. 往往都係問題來嘅時候先至 alert。

68. 想當年，等考 A-Level 真係好辛苦。
69. 我考 A-Level 嗰陣都冇乜點溫書。
70. 我好 amazing 佢會咁樣做。
71. 佢啲設計不嬲都咁令人 amazing㗎啦。
72. 我點知而家係咪用緊 analog？
73. 咩係 analog 呀？
74. 點樣去 analyze 個故仔？
75. 你會點 analyze 呢啲資料？
76. and then 我哋就去咗買嘢。
77. and then 咪繼續開工囉。
78. 換個 angle 再試過。
79. 如果由另一個 angle 睇呢？
80. 呢個消息其實係未 announce㗎。
81. 佢上星期唔記得 announce。
82. 想知道最新消息可以留意我哋嘅 announcement。
83. 呢個 announcement 公佈咗好耐㗎喇！
84. 一年有十四日 annual leave。
85. 佢一次過請晒十日 annual leave 去旅行。
86. 從來冇試過全情投入咁做有關 anthro 嘅嘢。
87. 其實 anthro 係讀咩㗎啫？
88. 我個 antivirus 好似費費哋咁。
89. 裝完個 antivirus 之後部機好似慢咗咁。
90. anyway，返到嚟之後我就打咗俾佢。
91. anyway 到咗信和之後我哋就租屋啦。
92. anyway，我就喺度睇電視，連續睇咗幾套電影。
93. anyway，多謝晒咁多位先，啲朱古力我已經食晒。
94. 呢幾頁資料可以當做 appendix。
95. 佢個 appendix 有成三十頁咁多。
96. 你有冇收到 application 又或者其他查詢？
97. 我哋係唔會接受你嘅 application。
98. 點樣 apply 宿舍？
99. 但我仲未有時間 apply。
100. 你有睇過佢份 appraisal 咋。
101. 佢份 appraisal 寫得好好。
102. 咁樣我反而仲 appreciate 添。
103. 你會 appreciate 佢啲寓意。
104. 佢用咗好多唔同嘅 approach。
105. 我覺得你個 approach 唔得。
106. 佢唔肯 approve 係因爲想公報私仇。
107. 佢唔 approve 我都冇符㗎。
108. 我哋嘅研究集中喺呢幾個 area。
109. 初初就喺度玩啤牌，後嚟玩玩下個 area 越嚟越大。
110. 本書有好多 argument。
111. 呢個 argument 好好吖。
112. 佢平時 around 六點半至走。
113. 我 around 九點嘅時候見到佢攞住好多嘢。
114. 如果打風會有啲咩 arrangement？
115. 我覺得佢嘅 arrangement 好差囉。
116. 我 arrive 嘅時候已經五點。
117. 佢十點先 arrive 梗係冇人啦！
118. 今日又有 art 堂，又要去參觀喇。
119. 原本諗住成班人唔上 art。
120. 訪問每個 artist 都要三四日。
121. 一個 artist 有成四五個人跟住。
122. 你上晒 assembly 未？
123. 佢去親 assembly 都瞌眼瞓。
124. 題目係教授 assign 嘅。
125. 我 assign 咗五個練習俾個學生做。
126. 我唔識做 assignment。
127. 點解啲 assignment 做極都做唔完㗎？
128. 老師 assume 我已經學咗呢啲嘢。
129. 佢 assume 你知道晒之前發生嘅嘢。
130. at first 佢好積極㗎。
131. 個個 at first 都對前途充滿憧憬㗎啦！
132. 不過 at last 都係衰咗。
133. 佢 at last 都入咗兩球，算係將功補過咁啦。
134. 我自己嘅目標係 at least 入一球！
135. 原來香港人好好㗎喇，at least 佢哋識得咩叫排隊。
136. at the end 就係整理實驗結果。
137. at the end 個結局都令人滿意。
138. 你講嘢嘅時候應該望住啲 audience。
139. 呢啲 audience 嘅質素好差。
140. 兩個細路都仲係叫佢做 auntie。
141. 佢廿幾歲但係個樣成個 auntie 咁。
142. 呢本書嘅 author 係邊個？
143. 個 author 好好人，仲幫我簽咗個名添。
144. 佢已經有一個 bachelor degree。
145. 冇 bachelor degree 都可以讀碩士㗎咩？
146. 我想將 background 嘅顏色改做黑色。
147. 你知唔知 background 嗰首歌係邊個唱㗎？
148. 朝頭早成班 backstage 喺台上唱歌跳舞。
149. 你未做過 backstage，有好多嘢都唔知。
150. 我有 backup，如果大家想要就問我攞啦。
151. 你 backup 晒啲重要資料先至叫我啦。
152. 有冇諗過點解次次都係安排你做 bad guy 呢？
153. 佢做 bad guy 直頭入形入格啦。
154. 我都覺得自己好鬼 bad taste。
155. 佢好 bad taste，成日群埋呢班人。
156. 每年除夕都會去 ball。
157. 一年用過百萬買衫去 ball。
158. 你又俾人 ban 咗呀？
159. 我講乜佢都會 ban㗎啦。
160. 今日我返學校夾 band。

161. 要夾 band 所以補課要早走。
162. 好在我都唔係志在夾 band。
163. 當然唔少得嗰兩塊 banner 啦。
164. 你唔知佢整 banner 好叻㗎咩？
165. 中大係咪有得 barbecue？
166. 次次去完 barbecue 我都唔舒服。
167. 之後去石澳 BBQ，真係好開心。
168. 我好鍾意 BBQ㗎！
169. before 下星期一我要寫好份講稿。
170. 如果 before 星期五做好就最好啦。
171. beginning 嘅時候主要係搜集資料。
172. 如果 beginning 嗰陣知道佢係咁我就唔同佢一組啦。
173. 我唔識，不過唔會唔合格，最多 below mean。
174. 次次都 below mean，咁唔係辦法。
175. 最好 between 兩個考試。
176. between 佢兩個之間，好難揀啫。
177. 或者有人追唔到我喺外面 big mouth 呢。
178. 你唔知佢好 big mouth㗎咩？
179. 有冇啲網頁係教人改 BIOS㗎？
180. 我點知個 BIOS 有問題啫？
181. 阿強嘅 birthday 係喺下個星期一。
182. 佢 birthday 嗰日收到好多禮物。
183. 我哋傾咗無限咁多嘢，有鬼故、black magic 同埋旅行等。
184. 你識唔識玩 black magic？
185. 我咁講唔係想 blame 任何人。
186. 如果佢係想 blame 你就唔會咁講啦。
187. 其實 bluetooth 係咩嚟㗎？
188. 你部電腦用唔用到 bluetooth㗎？
189. 個女人淨係叫我哋俾銀行簿同 BNO 佢睇。
190. 我本 BNO 過咗期好耐啦。
191. 係塊 board 近咗囉，咁其實又唔係唔好嘅。
192. 出面塊 board 係你整㗎？
193. 你屋企有冇用 boardband？
194. 邊間公司嘅 boardband 口碑最好？
195. 返到去之後我就匯合我阿媽去 body check。
196. 原來佢一直以嚟都冇做 body check。
197. 你唔駛旨意今年有 bonus！
198. 我覺得今年有 bonus 嘅機會好渺茫。
199. 我想 book 壁球場。
200. 明明噚日約咗我哋九點而且 book 埋房。
201. 你一陣幫我打電話 book 檯食飯吖！
202. 我之前同阿媽 book 咗今日同你慶祝。
203. 呢本書 book store 有得賣。
204. 你知唔知呢本書 book store 賣幾多錢？
205. 有時成日一個 booking 都冇。
206. 你今日有三個 booking。
207. 尤其有幾個 brand 一直俾我好高嘅價錢。
208. 好多香港品牌都好受歡迎，不過可以 brand 香港地位嘅就冇乜。
209. 放完 break 之後我先至做完份功課。
210. 你想個 break 放耐啲定早啲落堂？
211. 我起碼半年有多未食過 breakfast 囉。
212. 佢成日都喺呢度食 breakfast㗎。
213. 請你好 brief 咁介紹一下自己。
214. 你 brief 啲得㗎喇，其他我自己睇都得。
215. 個 briefing 完咗之後大家可以一齊食飯。
216. 一陣個 briefing 大概幾長？
217. 你個 budget 係幾多？
218. 我預咗三千蚊 budget，可以買邊啲款？
219. 呢度嘅 buffet 好出名㗎！
220. 我從來唔食 buffet。
221. 邊度啲 buffet 最正？
222. 點樣先至可以 build up 自信呢？
223. 我好辛苦先至 build up 到公司嘅信譽。
224. 考試前永遠係最 busy 嘅。
225. 佢咁 busy 你咪早啲約佢囉。
226. 佢好 buy 我嘅建議。
227. 雖然扮嘢但我又 buy 喎。
228. 我對 buying 幾有興趣嘅。
229. 雖然我對 buying 唔係好熟，但係好有興趣。
230. by the way 我好鍾意呢間餐廳。
231. by the way，你知唔知佢其實唔鍾意你㗎？
232. 上網速度講嗰啲係咪 byte 嚟㗎？
233. 我一路講緊都係 byte 喎。
234. 你一幫襯佢就會送隻 cable modem 俾你。
235. 我個 cable modem 唔知係咪壞咗。
236. 報埋警 call 埋白車。
237. 媽咪啲奪命追魂 call 係咁催。
238. 枉我次次都係第一個就 call 佢報平安。
239. 星期五晚我喺度勁 call 人。
240. 所以收線後不停 call back。
241. 我講緊電話，一陣先 call back 你。
242. 我笑到死死下，好難先 calm down 到。
243. 我哋 calm down 咗之後就去食嘢。
244. 去 camp 嘅時候大家都鍾意通頂。
245. 你下星期唔係入 camp 咩？
246. 下晝個會 cancel 咗喇。
247. 唔得閒咪 cancel 咗佢囉。
248. 我通常喺 canteen 食飯。
249. 我最憎食 canteen 啲飯。
250. 我突然間好想買帽，係 cap 帽。
251. 你頂 cap 帽好靚喎，係邊度買㗎？
252. 第一次操旗同埋第一次做 captain 位。
253. 佢以前係校隊，仲係 captain 嚟㗎。
254. 邊度有 car park？
255. 呢個 car park 唔係架架車都停得。

256. 個 card reader 係咪一插入部電腦就用得？
257. 你買個 card reader 咪得囉。
258. 我都唔 care 依啲嘢嘅。
259. 佢會好 care 每個細節。
260. 你有冇聽過 career talk？
261. 佢成日走堂去聽 career talk。
262. 好多天長地久嘅 case 都係咁開始。
263. 其實呢啲 case 唔係咁常見。
264. 我記得嗰日有幾十個女仔 casting。
265. 否則連 casting 嘅機會都冇。
266. 你可以着得 casual 啲。
267. 佢問我點解今日着得咁 casual。
268. 唔係個個 catholic 都係好人。
269. 佢話自己係 catholic。
270. 燒一隻 CD 要幾耐？
271. 上堂照書讀，咁我買隻 CD 聽好過喇。
272. 用唔同速度嘅 CD writer 要嘅時間都會唔同。
273. 我個 CD writer 好慢，得四速。
274. 如果因爲考試而唔同佢 celebrate，好似怪怪哋咁。
275. 年年都有同佢 celebrate，但係今年就冇。
276. 你生日會有啲咩 celebration？
277. 你唔知佢哋幫你搞咗個 celebration 咩？
278. 一早就喺 centre 等，睇吓啲老師用嘅嘢有冇事。
279. 個 centre 地方唔係好大，不過好骨子。
280. 冇張見得人嘅 cert 真係唔得㗎！
281. 你張 cert 係人見到都會請啦。
282. 去邊度可以申請 certifying letter？
283. 你有 certifying letter 就會快好多。
284. 佢個 chairman 講明今年冇人工加。
285. 唔係人人都適合做 chairman。
286. 其他人 challenge 你係應分嘅。
287. 你要先預計人哋會 challenge 邊啲地方。
288. 要譯晒成篇，真係幾 challenging。
289. 咁 challenging 嘅嘢留番俾你做。
290. 我哋今年攞咗啦啦隊 champion。
291. 佢哋已經攞咗兩次 champion，今次贏埋就三連冠。
292. 所以過嚟 chapel 搵我。
293. 我去到 chapel 但係冇人喎。
294. 今日講到邊個 chapter？
295. 我溫咗成日書都淨係睇咗一個 chapter 咋。
296. 我而家就 charge 你「行爲不檢」。
297. 佢一落 charge 我哋就麻煩。
298. 目的唔係要靚，而係要 charming。
299. 佢有時真係好 charming㗎。
300. 話我有其他事要做，唔同佢哋 chat。
301. 只係 chat 咗一晚，交換咗電話啫。
302. 我覺得佢好 cheap。
303. 點知條友勁 cheap。
304. 乜你咁 cheap㗎？
305. 點樣可以 check 到自己借咗幾多本書？
306. 點樣可以 check 到部機有幾快？
307. 記得 check spelling 之後先至好印出嚟。
308. 記事簿係冇得 check spelling㗎！
309. 尤其係趕住去第八個 checkpoint 途中。
310. 過咗第二個 checkpoint 之後就唔見咗佢。
311. 洗完頭，搽啲 chemical，局一陣，再洗，然後吹乾啲頭髮。
312. 我都係想休息吓啫，無謂再放咁多 chemical 上塊面度啦。
313. 問我哋將語言學擺落第幾個 choice。
314. 我將電子工程放喺第一個 choice㗎。
315. 你仲有冇唱 choir？
316. 佢以前唱 choir 嘅時候識咗好多人。
317. 你張相 chop 咗我個頭喎！
318. 我想 chop 走最頭嗰個字母。
319. 佢 claim 自己係最掂嗰個。
320. 呢啲嘢係佢自己 claim 嘅，真唔真冇人知。
321. class participation 係其中一個同學嘅好提議，佢話唔夠互動。
322. 都唔知個 class participation 係點計分嘅。
323. 今季品牌嘅設計好 classic。
324. 佢哋會玩啲大家耳熟能詳嘅 classic 歌。
325. 用完啲嘢要 clean up 番。
326. 佢攞晒啲嘢出嚟又唔 clean up 番。
327. 我 click 完個掣之後等咗好耐都冇反應。
328. 你 click 入「我的電腦」就可以見到剩番幾多位。
329. 唔舒服就去 clinic 睇醫生啦。
330. 佢依家喺政府 clinic 做藥劑師。
331. 甚至唔想咁 close。
332. 細個嘅時候同爸爸 close 啲。
333. 最正就係佢喊嗰個 close up。
334. 呢個 close up 影到佢好靚。
335. 原來 cocktail 唔難整㗎！
336. 我嗰杯 cocktail 啲味好怪。
337. 我想 collect 一啲關於植物嘅資料。
338. 佢 collect 咁多資料唔知有咩用。
339. 唔同 college 有咩分別？
340. 我同佢都唔係同一個 college 嘅。
341. 呢度賣嘅多數係 colourful 嘅少女款式。
342. 見到一支支咁 colourful 嘅顏料我就想買。
343. 你對我嘅網頁有啲咩 comment？
344. 我係唔會 comment 人哋嘅意見嘅。
345. 佢哋 commit 好晒先至通知我。
346. 你哋 commit 好未？
347. 呢單嘢真係唔 common，真係好想睇吓。
348. 呢啲咁 common 嘅嘢自己睇啦。

349. 去完 common room 之後入房打機。
350. 你想去 common room 食定係喺房食？
351. 題題都好似 common sense 咁，就算冇讀過都可以答。
352. 佢話呢啲係 common sense，係基本求生技能，一定要學。
353. 而家啲家長動不動就 complain。
354. 你唔驚俾人 complain 咩？
355. 帶 con 要注意啲咩？
356. 配 con 會唔會好麻煩？
357. 我仲有時間化妝同帶 con 添。
358. 我對呢套戲嘅 concept 有所保留。
359. 我覺得佢個 concept 唔係好啱。
360. 佢好 concern 我哋嘅表現。
361. 我絕對明白你嘅 concern。
362. 幾年冇喺香港開 concert。
363. 我從來冇聽過 concert。
364. 最後要做 conclusion。
365. 你個 conclusion 做得唔錯喎。
366. 唔係所有 condition 都啱用。
367. 呢個 condition 好特別喎。
368. 佢講嘢好有 confidence。
369. 我對佢冇乜 confidence。
370. 表現得流暢同好有 confidence 咁囉。
371. 凡係 confidential 嘅文件都唔可以攞走。
372. 呢份嘢係 confidential 唔好俾其他人睇。
373. 你用邊個 config㗎？
374. 我部機同你部機個 config 一樣㗎咩？
375. 但係香港嗰啲就一啲都唔 confine。
376. 好多時外國嘅紅燈區都會 confine 喺某啲地區。
377. 下星期個會 confirm 咗未？
378. 佢要遲啲先至可以 confirm 邊日得。
379. 夜晚去咗酒樓 confirm 圍數同菜單。
380. 有自己嘅 connection 先至可以搵到明星去首映。
381. 你無非都係想識人，建立 connection 啫。
382. 希望大家可以 consider 一下其他人嘅感受。
383. 當你應承嘅時候就會 consider 好多嘢。
384. 但係佢做嘢好唔 consistent。
385. 佢講啲嘢次次都唔同，好唔 consistent。
386. 呢個 constraint 真係幾大。
387. 呢個可能唔係影響我最大嘅一個 constraint。
388. 有時大學會幫一啲公司做 consultant。
389. 啱啱畢業出嚟做嘢都可以做 consultant？
390. 佢話冇你 contact 喎。
391. 我晏啲再 contact 你吖。
392. content 要寫埋頁數同埋副標題。
393. 我淨係睇咗個 content，其他未睇喎。
394. 要簽十八個月 contract，但係有好多嘢送。
395. 你張 contract 幾時完呀？
396. 呢幅相嘅 contrast 唔夠大。
397. 其實個 contrast 大同細有咩分別？
398. 你有咩 contribution 呀？
399. 講真，我唔覺佢有 contribution 喎。
400. 唔係樣樣嘢都 control 到。
401. 我 control 唔到佢幾時笑幾時喊。
402. 個飯堂嘅位置好 convenient。
403. 呢度喺正地鐵站上面，夠晒 convenient。
404. 佢講嘅嘢好 convincing。
405. 有啲人講嘢好 convincing，好似催眠咁。
406. 啲 cookies 其實有冇用㗎？
407. 我需唔需要定期刪除啲 cookies 呀？
408. 呢個袋沉實穩重有型又夠 cool。
409. 佢仍舊保持一貫 cool 到痺嘅形象。
410. 唔該幫我將呢幾幅圖 copy 落個電話度吖。
411. 我幫我嗰班某啲人 copy 咗啲相。
412. 要做嘅嘢就係 copy and paste。
413. copy and paste 呢啲咁簡單嘅嘢駛乜搵大學生做呀？
414. 你講錯嘢佢會 correct 你。
415. 有時唔係樣樣嘢都要 correct 嘅。
416. 個 correlation 冇想像中咁強。
417. 我都唔知原來個 correlation 咁大㗎。
418. 個 cost 唔只咁細㗎喎！
419. 呢個 cost 已經係最低㗎喇。
420. 去完 count down 之後一直飲到天光。
421. 你今年會唔會去 count down？
422. 記得帶埋張 coupon 去食飯喎。
423. 爲咗嗰兩張半價 coupon，諗都唔諗就去咗西九龍。
424. 咁就讀完成個 course 喇。
425. 呢個 course 冇乜人讀。
426. 其實 CPU 嘅原理你又知唔知呢？
427. 點解啲 CPU 成日都唔減價嘅？
428. 呢個蛋糕上面有好多 cream。
429. 佢好鍾意食 cream，但係又怕肥喎。
430. 香港學生唔夠 creative。
431. creative 呢樣嘢比較難學。
432. 我已經讀咗四十個 credit。
433. 一年最多讀四十二個 credit。
434. 有啲嘢有時好難去 criticize。
435. 佢淨係識得 criticize。
436. 現階段已經構思好幾件 crossover 作品。
437. 其實 crossover 係幾時開始流行㗎呢？
438. 呢間公司嘅 culture 好特別。
439. 每個地區都有自己獨特嘅 culture。
440. 我想搵一啲 curriculum。
441. 麻煩你幫我攞本 curriculum 吖。
442. 佢講嘢嘅時間俾人 cut 咗。

443. 張相晒出嚟會唔會有啲嘢 cut 咗㗎？
444. 於是我 cut 咗佢線。
445. 俾人又 cute 又可愛嘅印象。
446. 有個靚仔好 cute。
447. 仲有我個頭，好 cutie，因爲個頭今年多咗人搵我影相。
448. 佢咁 cutie 你咪讚吓佢囉。
449. 款式按日本 cutting 做。
450. 件衫嘅 cutting 好靚，不過唔知我着唔着到。
451. 同 daddy 啱傾啲。
452. 細個嗰陣成日同 daddy 玩騎膊馬。
453. 咁我咪搵兩個 dancer 落嚟表演跳舞囉。
454. 佢十幾歲就開始做 dancer，到依家已經跳咗十幾年舞。
455. 呢個 database 有晒你要嘅嘢。
456. 我而家學緊點用 database。
457. 今日係辛苦咗六日之後嘅 day off，我真係覺得好攰。
458. 但係我咁多 day off，冇理由蝕俾佢㗎。
459. 我聽完呢番說話，淨係証實咗佢係唔識用 DC。
460. 部 DC 跌咗落地，好彩冇事。
461. 我哋喺度嘈，之後就 dead-air 囉。
462. 但 dead-air 嗰陣我其實真係好嬲。
463. 我唔係 deal with 呢啲嘢㗎！
464. 最怕 deal with 啲講極都唔明嘅人。
465. 你哋個 dean 叫咩名？
466. 佢下一個目標係做 dean。
467. 點樣先至有得攞 dean list？
468. 我嘅 dean list 夢好快就會粉碎。
469. 撩邊個嚟可以 debug？
470. 佢叫我幫佢 debug，我鬼得閒理佢。
471. 你自己 decide 點樣做啦。
472. 我 decide 到就唔駛問你啦。
473. 我哋就拎咗啲啤牌出嚟鋤 Dee。
474. 佢最叻就係玩鋤大 Dee。
475. 點樣 define 個問題？
476. 你都冇 define 清楚。
477. 佢個 degree 係喺外國讀㗎！
478. 一年咁多大學生畢業，大把人有 degree 啦。
479. 可唔可以 delay 呀？
480. 佢起機嗰陣仲 delay 咗一個鐘。
481. 我已經 delete 咗嗰啲資料。
482. 我好想 delete 呢度嘅某啲舊文章。
483. 應承得人就一定要 deliver 到。
484. 應承咗人但係 deliver 唔到，咪失信於人囉。
485. 一陣有得睇 demo？
486. 我想睇最新嗰隻相機嘅 demo。
487. 你邊個 department㗎？
488. 我哋 department 得五個人。
489. 成個 department 啲人一齊放假。
490. 我 design 咗一個新嘅網頁。
491. 我同 design 啲人唔熟。
492. 啱啱睇完人哋嘅 design。
493. 你個 desktop 上面有好多嘢喎。
494. 你做乜放咁多嘢喺 desktop 上面嘅。
495. 但我又唔想喺度講得咁 detail。
496. 講得咁 detail 有啲人可能會唔高興。
497. 我會同佢哋 develop 一個融洽嘅關係。
498. 興趣係可以慢慢 develop 嘅。
499. 今日要諗吓俾啲乜嘢由 DHL 過嚟嘅新同事做。
500. 佢叫我用 DHL 寄份文件俾佢。
501. 我仲係用緊 dial-up 咋。
502. 佢用 dial-up 先至連到返嚟喎。
503. 點解我可以打到呢篇 diary？
504. 估唔到本 diary 都未到一周年，就已經要斷纜，實在太遺憾。
505. 我啱啱買咗部新嘅 dig cam。
506. 呢部 dig cam 同我嗰部有咩唔同？
507. 你要 diligent 啲先至得。
508. 佢話我唔夠 diligent，但我自問已經盡咗力。
509. 呢間餐廳嘅 dinner 好好食。
510. 睇嚟我哋只能夠「係咁易」食餐 dinner。
511. 竟然識埋啲咁 dirty 嘅人。
512. 不過佢好 dirty 呀，襯我出咗埠就出去四圍玩。
513. 今次我哋會 focus 喺邊方面？
514. 你影啲相張張個 focus 都唔啱喎。
515. 咁我又唔會好 disappointed。
516. 你同佢講咗啲乜呀？佢好 disappointed 喎。
517. 佢唔係有心打擊你，你唔駛咁 disappointed 嘅。
518. 我部 discman 買嘅時候要成二千幾蚊㗎！
519. 而家買 discman 應該平過以前好多。
520. 我知呢度有好多 discount 同更好嘅牌子。
521. 我簽咭冇 discount 都有得攞分吖嘛。
522. 不過 discover 咗一樣嘢。
523. 我 discover 唔到呢個錯誤，係我唔夠小心。
524. 男裝可以利用門口嚟做 display。
525. 點改電話 display 張相？
526. 我唔識 distinguish 佢地有咩唔同。
527. 我 distinguish 到就唔駛搵你幫手啦。
528. 超級市場有冇 DIY 蛋糕套裝賣？
529. 依家咁興 DIY，我咪食住個勢囉。
530. 其實上網係咪要用 DNS㗎？
531. 我唔知咩係 DNS。

532. 老細叫我做完啲 documentation 先至好走。
533. 我最怕就係做 documentation㗎喇。
534. 你以爲我會幫你？don't waste my time 啦！
535. 我唔想再花時間喺呢個問題上，don't waste my time。
536. 一早就出嚟同我 double check 有冇漏嘢。
537. 你出門口之前再 double check 多次啦。
538. 你 double click 佢就會自動爆開。
539. 喺呢度 double click 就得㗎喇。
540. 你係咪讀 double E㗎？
541. 我細細個嘅志願就係做工程師，所以決定讀 double E。
542. 我永遠都唔會 down、唔會輸。
543. 我嘅心情都好 down。
544. 邊度有得 download 嗰封信？
545. 佢叫我上網 download 嗰幅圖喎。
546. 你可以 draft 咗個大概先。
547. 你俾份 draft 我睇住先啦。
548. 我 drag 唔到個快勞喎。
549. 你撳實個掣就可以 drag 到㗎喇。
550. 放學之後一齊練 drama。
551. 返到學校，我哋就喺度練 drama。
552. 我嘅 dream house 唔駛好大，但係一定要望到海。
553. 希望買間 dream house 俾媽咪。
554. 呢度一杯 drink 要幾多錢？
555. 五十蚊包兩杯 drink，花生同果盤價錢另計。
556. 我已經 drop 低咗佢講嘅嘢。
557. 我只會講一次，你記唔到就 drop 低佢啦。
558. 你有冇 drop notes 嘅習慣？
559. 佢上親堂都 drop notes，不停咁抄。
560. 同朋友攞出 duty list 睇睇其他人教咩科。
561. 張 duty list 冇我個名喎。
562. 你 easter 嗰陣冇去旅行咩？
563. 我打算 easter 嘅時候去澳洲。
564. 我舊年聖誕收到好多 e-card。
565. 邊度仲有免費嘅 e-card？
566. 我讀咗三年 Econ，次次都好高分。
567. 你咁憎計數就唔好讀 Econ 喇。
568. 我有個讀 economics 嘅朋友。
569. 其實 economics 喺好多方面都可以應用到。
570. 呢個方法好 efficient。
571. 真係咁 efficient 先算啦！
572. 我已經讀咗兩個 elective。
573. 我揀晒嗰啲「易碌」嘅 elective。
574. 你識唔識發 email？
575. 你個 email 唔係免費㗎咩？
576. 呢本書係教人點樣控制自己嘅 emotion。
577. 佢 emotion 有啲問題，成日發脾氣。
578. 佢好 encourage 我去做自己想做嘅嘢。
579. 我唔會 encourage 你去接受任何嘢。
580. 見到呢個鎖即係表示已經做咗 encryption。
581. 有冇啲書係教 encryption 嘅原理㗎？
582. 原來《四葉草》嚟拍 ending 呀！
583. 個 ending 同預料嘅一樣，一啲驚喜都冇。
584. 我都係讀 engine 嘅。
585. 架車個 engine 壞咗，冇排整喎。
586. 以前做 engineer 嘅時候要輪班。
587. 我想搵份 engineer 工。
588. 佢好 enjoy 一個人行街。
589. enjoy 同珍惜而家所有嘅嘢。
590. 我幾 enjoy 做大爺㗎！
591. 主持呢個節目唔能夠暢所欲言，所以唔 enjoy。
592. 有任何 enquiry 都可以打呢個電話。
593. 你有 enquiry 可以直接搵我。
594. 你 ensure 自己準時交份功課？
595. 我唔可以 ensure 自己可以完成賽事。
596. 你唔駛 entertain 我㗎喎。
597. 我唔慣 entertain 人。
598. 做媽媽嘅 EQ 一定要好。
599. 我諗你要提升一下自己嘅 EQ。
600. 我寫 essay 成日都會寫得太多字。
601. 考試要寫 essay 嘅話我通常會做唔晒。
602. 呢個數你係點樣 estimate㗎？
603. 佢 estimate 啲嘢都好準。
604. 和記嘅 ethernet 係咪穩定啲？
605. 用 ethernet 係咪會快啲㗎？
606. even 你唔返工都冇人知。
607. even 同佢分析咗好多，但最後都係唔明白。
608. 佢好鍾意喺 evening 嘅時候打波。
609. 如果 evening 嘅時候坐喺度就正咯。
610. 突然見到個女仔好熟口面，原來係以前個 ex。
611. 我個同事嘅男友原來係我個 ex。
612. 佢嘅諗法 exactly 同我一樣。
613. 我哋嘅興趣 exactly 一樣，都係咁好動。
614. exam 要帶啲咩？
615. 下個禮拜開始 exam 喇。
616. 佢喺 exam 嘅時候遲咗起身。
617. 你嘅 examiner 係邊個？
618. 我都未知邊個係我 examiner。
619. 你有學過點用 excel㗎咩？
620. 佢用 excel 用到好熟。
621. 舊年暑假我去咗英國 exchange。
622. 佢 exchange 嗰陣識咗個泰國女仔。
623. 成個課室都係得一個 exchange student。
624. 我隔離房住咗個德國嚟嘅 exchange

student。

625. 能夠同佢哋合作我覺得好 exciting。
626. 真係咁 exciting 嘅話我都要去試吓。
627. 佢 exclude 咗我嘅意見。
628. 你其實 exclude 咗好多人。
629. 我唔會接受呢啲 excuse。
630. 呢啲係你嘅 excuse 啫。
631. 跟住佢俾咗啲 exercise 我做。
632. 佢淨係識得叫人做 exercise，自己就乜都唔做坐喺度等落堂。
633. 我冇 expect 會有人重視我嘅作品。
634. 冇人會 expect 自己唔掂嘅。
635. 結果同我所 expect 嘅差唔多。
636. 我冇 expect 到會有人影相。
637. 你要自己 explore 一下呢個問題。
638. 佢叫我 explore 一下先，唔明先至搵佢。
639. 你嘅缺點係唔識得 express 自己。
640. 文學教一個作家 express 一啲嘢。
641. 份功課 extend 咗。
642. 我想準時畢業，唔想 extend。
643. 而我亦聽佢話唔戴 face mask。
644. 佢沙士嗰陣買咗好多 face mask。
645. 唔得閒溫書，因爲我要做 facial。
646. 星期六我一早起身就去咗做 facial，啲皮膚又愈來愈差。
647. 你邊個 faculty㗎？
648. 我個 faculty 成日搞好多講座。
649. 當初嘅細心體貼慢慢 fade out。
650. 啲音樂慢慢 fade out，之後主角就出場。
651. 佢對 family 好重視。
652. 我哋個 family 又多咗幾個人。
653. 佢係一個好顧家嘅 family man。
654. 喺戲入面佢係一個 family man，不過現實中就剛好相反。
655. 睇到佢十分熱愛唱歌，熱愛佢嘅 fans。
656. 就連師奶 fans 都有一大堆。
657. 嗰日我同我老豆 farewell。
658. 突然 farewell 所以帶咗飯都唔食(lu)。
659. 以爲可以同自己班有 farewell party。
660. 啲同事同我搞咗個 farewell party。
661. 邊度有 fax 機？
662. 份嘢 fax 咗未？
663. 點樣 fax 嘢去外國？
664. 你個 feedback 好慢喎。
665. 你仲等緊佢嘅 feedback。
666. 我 feel 到有啲唔妥。
667. 雖然俾蚊針死，但都幾有 feel。
668. 不過好好 feel。
669. 佢襯晒我哋酒吧個 feel 所以先至請佢嚟唱歌。
670. 我唔係話想 fight 啲咩。
671. 個老細又幫唔到佢 fight 到啲乜嘢喎。
672. 呢個 file 可能有毒。
673. 早兩日喺巴士執到個 file。
674. 你做乜唔做 file compression 呀？
675. 做咗 file compression 之後先至好交俾我。
676. 我想將全部資料入面最大嘅數 filter 出嚟。
677. 我買咗個新嘅 filter，用嚟過濾啲水。
678. 佢今年係 final year。
679. 佢 final year 嗰年先至住宿。
680. 你個 final year project 關於咩㗎？
681. 我唔係好明佢個 final year project 做咩。
682. 你即刻就會覺得啲皮膚 firm 咗。
683. 搽咗之後皮膚滑咗但唔算 firm。
684. 點嘅成績先至可以攞到 first hon？
685. 佢 first hon 畢業之後就入咗政府做政務官。
686. 時間又好 fit。
687. 你咁 fit，我怕我跟唔上喎。
688. 我個網頁分咗 flame 之後清楚咗好多。
689. 你張相個 flame 係咪自己整㗎？
690. 你部相機用 flash memory㗎？
691. 我張 flash memory 燒咗。
692. 今次份功課要用 floppy 交。
693. 你可唔可以借隻 floppy 俾我？
694. 每個 folder 都可以用唔同嘅公仔去代表。
695. 個個 folder 嘅名都要唔同。
696. 佢講嘢太快，我 follow 唔到。
697. 佢要求咁高，我唔想再 follow 佢。
698. 呢本書係 for 初學者睇嘅。
699. 今日係 for 文、理同社會科學學院嘅學生。
700. 我一定有講過㗎，for sure。
701. 我會盡最大努力去爭個冠軍返嚟，for sure。
702. 你可唔可以 forgive 我？
703. 佢肯 forgive 我，要我請食飯又點話？
704. 張 form 幾時要交？
705. 上到去佢已經攞咗飛仔填緊 form。
706. 由我 form five 嗰年嘅經歷，我悟出一個道理。
707. 我 form five 未考完會考已經返緊工。
708. 個個 form one 嘅時候都係咁㗎啦！
709. 我 form one 嗰陣已經有一米七。
710. 記得上次已經係 form three 嗰年啦！
711. 我 form three 之後就冇讀過中史。
712. 你有冇論文嘅 format 呀？
713. 你隻碟 format 咗未㗎？
714. 有好幾題唔記得啲 formula。
715. 咁多 formula 點記呀？
716. 麻煩你一陣 forward 嗰啲資料俾我。
717. 你 forward 封電郵俾我啦！
718. 我唔認爲咁樣都叫做 free 囉。

719. 我覺得咁樣係最 free 嘅。
720. 老細唔喺度我哋好 free。
721. 乜都可以講，好 free。
722. 約返啲 friend 出嚟要靠你喇。
723. 本來約咗個 friend 行街。
724. 同班 friend 吹水真係好開心。
725. 你嘅 frustration 咁大我都唔知點安慰你好。
726. 咁嘅話佢嘅 frustration 會好大。
727. 點解我用唔到 FTP 嘅？
728. 我想用 FTP 駛唔駛辦咩手續？
729. 我驚我啲同學 fulfil 唔到你嘅要求喎。
730. 我實 fulfil 唔到佢個要求啫。
731. 我想讀嗰個興趣班 full 咗。
732. full 咗咪報過第二班囉。
733. 其實我哋如果 full team 齊人，係唔會輸波。
734. 就算 full team 佢哋都唔係我哋對手。
735. 佢依家讀緊 full time 嘅學位課程。
736. 我之前曾經 full time 做過一年嘢。
737. 呢部電話有啲咩 function？
738. 佢連部相機有乜 function 都未知就俾錢。
739. 唔駛同我講 functionality，我淨係想知要幾耐先至起到貨。
740. 單係計 functionality 嘅話就梗係買呢部啦。
741. 不過佢好 funny，仲肯同我哋玩埋一份。
742. 佢真係好 funny。
743. 用盡各種 fusion 方式炮製青口。
744. 而最特別嘅就要數師傅自創嘅法越 fusion 菜。
745. 但因 FYP 在身唔可以放下組員唔理 。
746. 我個 FYP 都仲未開始。
747. 唔知 ga 唔 gather 到幾個好朋友呢？
748. ga 唔 gather 到人就要睇你號召力喇。
749. 最近有啲咩新 game 好玩㗎？
750. 呢隻 game 啱啱出嗰陣賣成四百幾蚊。
751. 佢仲話下次我哋又玩返呢隻 game。
752. 但係裝完隻 game 竟然玩唔到。
753. 第一段婚姻多數會 game over。
754. 本雜誌做唔夠三個月就 game over。
755. 下星期個 gathering 有幾多人會去？
756. 今次 gathering 有成二十人出席。
757. 咁樣講好 general。
758. 問埋啲咁 general 嘅問題，都唔知佢點揀人。
759. 你有冇讀呢個 general education 呀？
760. 你以前讀大學嗰陣已經有 general education㗎嗱？
761. 我同佢 geographically 相隔好遠。
762. geographically 其實唔係好遠，只係轉車用多咗時間啫。
763. 我 get 到嘅嘢真係幾有意思。
764. 我 get 到佢想講咩。
765. 一直向上行，行行下 get lost 呀！
766. 唔知點解行商場都可以 get lost 嘅。
767. 佢係唔會 give up 嘅。
768. 我係唔會咁快 give up 嘅。
769. 你可以 go on。
770. 你有你 go on，佢有佢講電話。
771. 你好快咁 go through 一次啦！
772. 因爲太多字所以唔想逐個咁 go through。
773. 但我覺得佢俾我嘅印象係 good 咗。
774. 佢淨係同我講 good，冇講其他嘢。
775. 今日終於都 good show 咯。
776. 佢同我講咗句 good show 就走咗。
777. 可能佢哋睇化咗，當 GPA 唔係一回事啦。
778. 聽講佢 GPA 有四點零，不過我好懷疑。
779. 反正我都唔會再對 grad din 有咩期望。
780. 爲咗準備今日嘅 grad din，我特登等到今日先去剪髮。
781. 要檢查清楚 grammar 啱唔啱。
782. 佢啲 grammar 錯得好離譜。
783. 利希慎音樂廳係全中大最 grand 嘅場地之一。
784. 呢度咁 grand 襯晒你啦。
785. 你有冇申請 grant loan？
786. 食完就落去交埋份 grant loan 申請表。
787. 我見你啲 graphic 畫得好靚。
788. 佢都冇打算整過個 graphic。
789. 黑同白中間總會有 grey area。
790. 好多嘢都係 grey area，做唔做你自己決定。
791. 成個 group 等你一個咋。
792. 就算俾你 group 到一班人都未必會成事。
793. 我已經覺得我哋個 group 贏咗。
794. 做 group leader 要照顧吓啲新人。
795. 做 group leader 你估易㗎？
796. 呢個係 group project。
797. 做 group project 如果個個組員都肯做嘢咪好囉。
798. 有屋企人喺可以安排佢住喺 guest house。
799. 我老師依家暫時住喺 guest house。
800. 教授會 guide 你點做。
801. 我 guide 咗佢好耐佢都未學識。
802. 心裡面有啲 guilty。
803. 你覺得 guilty 咪唔好做囉。
804. 好耐都冇去做過 gym。
805. 頂佢唔順就走咗去做 gym。
806. 佢得閒就會研究吓點樣 hack 人。
807. 你廿四小時上網唔驚俾人 hack 咩？
808. 你住 hall 嘅時候有冇煮嘢食？
809. 住喺 hall 唔係方便啲咩？
810. 我哋一齊參與嘅最後一個 hall function。

811. 你平時都唔參加 hall function，梗係識唔到人啦。
812. 而家揸車講電話唔用 hand free 係咪犯法㗎？
813. 你點解唔用原裝嗰個 hand free？
814. 點樣 hand in 呢份計劃書呀？
815. 你見唔到我就唔駛 hand in 份功課㗎啩？
816. 我睇中咗一個新嘅 handbag。
817. 佢個 handbag 同我嗰個好似樣。
818. 我部電腦 hang 咗點算。
819. 估唔到連手提電話都會 hang 機。
820. 一陣去邊度 happy？
821. 佢笑晒好 happy 咁。
822. 連續兩日都係睇愛情片，次次都係 happy ending。
823. 唔一定要 happy ending 先至好睇㗎。
824. 你有冇玩過 happy run？
825. 我跑埋 happy run 都唔夠分住宿。
826. 而家啲 hard disk 平咗好多。
827. 我想買一個唔好太貴嘅 hard disk。
828. 當唔當係 hard time 睇你點睇㗎啫。
829. 每個人都會有 hard time。
830. 連阿 head 都未做過就分俾妳做？
831. 唔知阿 head 俾唔俾我放假呢？
832. 個 heading 應該寫咩？
833. 你喺 heading 度寫番清楚你個題目喎。
834. 呢次我要做到一百分有 heart 嘅演唱會。
835. 有冇 heart 我一睇就知啦。
836. 呢句 hello 對我嚟講實在係太沉重喇。
837. 佢只係同我講咗句 hello 咋。
838. 今次佢買咗條工人褲同一對 Hello Kitty 拖鞋。
839. 地鐵出咗套 Hello Kitty 飛喎。
840. 佢好幫得手好 helpful。
841. 佢咁 helpful，咪次次都叫佢幫手囉。
842. 佢哋兩個即刻 high 到忘晒形。
843. 我哋又喺度勁大聲咁嗌，個個都 high 到死。
844. 之後去咗間好 high class 嘅餐館食嘢。
845. 呢度好似好 high class 咁喎。
846. 做 highlight 會唔會傷頭髮？
847. 佢用咗好多時間先至整到呢個 highlight。
848. 仲有唔少 hip hop 衫同日本名牌衫賣。
849. 好多時大家都對 hip hop 有誤解。
850. 係咪喺 history 可以搵返之前去過嘅網頁？
851. 而且啲 history 功課掂都未掂過。
852. 雖然並唔 hit 但好好聽。
853. 我哋個個都曉得 hit 波。
854. 唉，究竟邊個 hold 住我本書呀？
855. 過咗一個禮拜我已經續借本書，但都難逃被 hold 嘅命運。
856. 擅長焗製蛋糕同 homemade 嘅甜品。
857. 無論係朱古力味定栗子味，都係 homemade 嗰隻好食啲。
858. 你知唔知中大個 homepage 點去？
859. 佢個 homepage 整得好靚。
860. hopefully 佢能夠及時趕到。
861. hopefully 佢可以快啲好番。
862. 邊度可以查到啲 hotkey 呀？
863. 你成日都用 hotkey，唔怪得快我咁多啦。
864. how come 佢會做成咁？
865. 真係唔明 how come 佢變咗咁多。
866. 乜你整網頁嘅時候係打 HTML㗎？
867. 我想學寫 HTML 應該睇邊本書？
868. 我想買隻 hub 喺屋企用。
869. 一隻 hub 可以俾幾多部電腦用？
870. 佢從來都唔覺得自己 hurt 緊人。
871. 我知我可能 hurt 到你。
872. 佢嘅一啲話 hurt 到人。
873. 好搞笑，I mean 我哋傾啲話題好搞笑，好似啲交流活動咁。
874. 其實初初我諗住實見佢唔到，I mean 今日。
875. 我終於成功裝咗 i-cable 喇。
876. 有啲人話 i-cable 唔好用喎。
877. 我可唔可以自己設計 icon？
878. 呢啲 icon 全部都係我自己整嘅。
879. 噚晚啲人 ICQ 講嘅嘢好抵死無聊。
880. 今日喺屋企，悠手好閒地 ICQ。
881. 學校入面有冇得打 IDD？
882. 邊間公司打 IDD 會平啲？
883. 你有冇啲新嘅 idea？
884. 我哋係喺星期五先真真正正傾 idea。
885. 唔係成日都有咁好嘅 idea。
886. 成日都咁 ideal 就好啦。
887. 唔會次次都咁 ideal 嘅。
888. 你封信唔可以同個樣本 identical。
889. 你抄還抄都冇理由連個名都 identical㗎嘛。
890. 我唔會 ignore 任何機會。
891. 你哋可以 ignore 佢。
892. 所有 illegal 嘅嘢我哋都係唔會做嘅。
893. 你未經人同意就用人哋啲嘢係 illegal㗎。
894. 呢個 image 啱晒佢啦！
895. 一向大家嘅 image 都唔同㗎啦。
896. 邊啲嘢係 important㗎？
897. 好 important㗎，千祈唔好漏口風喎。
898. 佢呢個方法簡直係 impossible。
899. 有時候你以爲 impossible 嘅嘢喺十幾年後就會成真。
900. 面試嘅時候俾人哋嘅第一個 impression 係好重要嘅。
901. 佢嘅打扮會令人對佢嘅 impression 唔係

咁好。

902. 佢有一句說話，令我好 impressive。
903. 我好 impressive 佢嗰日嘅髮型。
904. 本來我都冇諗住去 in，因爲我知實好辛苦。
905. 我到目前爲止已經 in 咗五份工。
906. in case 有意外點算？
907. in case 有事都唔係你同我負責啦。
908. in general 嚟講大學生係應該醒目啲。
909. 其實 in general 我哋會請大學生。
910. 僱主希望啲學生可以 independent 一啲。
911. 如果你 independent 一啲，依家就唔會搞成咁啦。
912. 太過 individual 未必係好事。
913. 如果係 individual 咁計分，佢哋又點會唔做嘢呀？
914. 點樣可以攞到多啲 information？
915. 呢啲 information 係邊個提供㗎？
916. 你應該喺呢度 insert 一啲嘢。
917. 喺呢度 insert 會唔會有問題。
918. 我啱啱 install 咗一個電子字典。
919. 其他軟件我可以自己 install。
920. instead of 呢個方法你仲可以試吓其他方法。
921. 其實 instead of 飲茶，你可以試吓去西餐廳。
922. 你係咪跟住個 instruction 嚟做㗎？
923. 呢個 instruction 都講得唔清楚。
924. 佢嘅諗法好 interesting。
925. 佢有時都幾 intersting。
926. 所以 intern 係必須嘅。
927. 你舊年係咪有去 intern 呀？
928. 大學會邀請一啲 international 嘅教授嚟演講。
929. 有得去 international 嘅大公司做嘢就梗係好啦。
930. internet 上面可以搵到好多有用嘅資料。
931. 有咗 internet 之後，啲人會唔會少咗同人面對面溝通呢？
932. 我有得去上海做 internship。
933. 好多人都係去同一個地方做 internship。
934. 好多人 interview 嘅時候都犯咗呢啲錯誤。
935. 順順利利過晒啲 interview。
936. 星期一仲要 interview㗎！
937. 想約我聽日去 interview。
938. 個 intro 得十頁都唔夠，但我已經查咗字典不下數十次。
939. 除咗 intro，其他都做得幾好。
940. 喺 introduction 入面要寫咩？
941. 你咁講即係你同意佢 introduction 入面講嘅嘢？
942. 我唔記得咗 invisible 佢。
943. invisible 即係你見到人，但人哋見唔到你。
944. 點樣知道自己個 IP 係幾多？
945. 我見唔到你個 IP 喎。
946. 你知唔知 IP phone 係咩嚟？
947. 有啲公司而家已經轉用咗 IP phone。
948. 以下嘅 IQ 題，答中有獎。
949. 而家啲人對 IQ 已經冇以前咁重視。
950. 目的地係李惠利 IVE。
951. 自問讀咗 IVE 之後見多咗唔同嘅人。
952. 唔知點解近排成日 jam 紙。
953. 部影印機一印親雙面就 jam 紙。
954. 由依家開始我會學聽 jazz。
955. 唔係人人都鍾意 jazz㗎嘛。
956. 如果冇把握做得好，我係唔會接 job。
957. 有時成個月一個 job 都冇，收入好唔穩定。
958. 佢 join 我哋嘅時候我哋都差唔多走喇。
959. 之後就去維園 join 阿文。
960. 我個同房係讀 journal 嘅。
961. journal 對學生嘅中英文成績要求都好高。
962. 所有 junior 嘅職員入職嘅時候都要去上堂。
963. 我係成間公司最 junior 嗰個。
964. 好耐冇唱過 K 喇。
965. 我好耐無睇戲同唱 K，係時候要約大家多啲出嚟玩。
966. 唔該幫我 keep 住本書先。
967. 反而 keep 到呢個好習慣。
968. 我想 keep 住日日跑步。
969. 佢可能一路都有 keep 住睇香港啲戲。
970. 而家興嘅瘦身已經超過咗 keep fit 嘅定義。
971. 做運動係希望 keep fit。
972. 屢敗屢試係成功嘅 key point。
973. 佢成日帶我遊花園，硬係唔講 key point。
974. 呢個 keyboard 好啱我手形。
975. 我買咗個新嘅 keyboard。
976. 食啲唔係喺 KFC 買嘅嘢，你話幾過份。
977. 之後去咗 KFC 食下午茶。
978. 會唔會拍 kiss 戲嘅時候借位？
979. 男主角 kiss 女主角。
980. 陳教授好 knowledgable。
981. 佢俾人一個好 knowledgable 嘅感覺。
982. 今朝知道咗個 Korean 死喺武裝份子手下。
983. 佢男朋友係 Korean 嚟㗎。
984. 我一陣會去 lab。
985. 唯有喺 lab 做嘢避開佢哋。
986. 好彩我都借到 label 紙貼住先。

987. 啲 label 真係好靚。
988. 你嘅 language skill 如何？
989. 入咗大學之後我嘅 language skill 似乎越嚟越差。
990. 你有冇去過 language table？
991. 我覺得如果你唔講嘢嘅話，去 language table 棧嗺時間。
992. 你嗰部 laptop 喺邊度買㗎？
993. 我想買部新嘅 laptop。
994. 話晒最 last 一次，唔緊要啦。
995. 最 last 嗰堂喺度劲畫公仔。
996. 人哋都 last day 啦，放肆多一陣就冇。
997. last day 嗰日我哋留到九點幾至走。
998. 今日一早返咗沙田攞部 LCD Mon 去屯門。
999. 屋企個 LCD Mon 用咗五年都冇壞過。
1000. 佢會係一個出色嘅 leader。
1001. 我哋啲 leader 又喺度鬧人。
1002. 講番 leadership 先啦！
1003. leadership 呢樣嘢唔係人人都有。
1004. 你印咗 lecture notes 未？
1005. 呢份 lecture notes 寫得好詳細。
1006. 如果嗰堂去 lecture room 要行到氣咳嘅話都走。
1007. 入到 lecture room 發現一個人都冇。
1008. 你 leisure 嘅時候會做啲咩？
1009. 佢雖然好忙，但 leisure 嘅時候都會全情投入咁去玩。
1010. 你同佢嘅 level 差好遠喎。
1011. 我發現我啲英文 level 係咁向下滑。
1012. 我好唔 like 佢。
1013. 好唔 like 嗰個人囉，死人冧甩佬。
1014. 通常唔同 line 會搵唔同嘅代言人。
1015. 呢條 line 嘅對象係三十歲以上嘅女士。
1016. 大家可以去呢條 link 度睇下。
1017. 你可唔可以俾你個網頁條 link 我呀？
1018. 我會 list 晒所有要用嘅嘢出嚟。
1019. 佢已經 list 咗要注意嘅地方出嚟。
1020. 嗰間嘢有 live 嘅表演。
1021. 佢靠個樣又唔係靠把聲搵食，所以啲歌迷都唔介意佢係咪唱 live。
1022. 今年仲有 live band 現場彈奏添。
1023. live band 會重複地唱務求你唔識唱都識哼。
1024. 我 load 唔到隻碟喎。
1025. 你將隻碟 load 落部電腦度咪得囉。
1026. 呢間宿舍個 lobby 成個酒店大堂咁。
1027. 個 lobby 整得咁靚，但係入面就爛溶溶。
1028. local 學生佔全部學生幾多？
1029. 我哋主要收 local 學生。
1030. 有咗呢啲資料我哋就可以 locate 到你嘅位置。
1031. 我 locate 唔到自己嘅位置。
1032. 你而家嘅 location 喺邊？
1033. 呢個電話好劲，喺任何 location 都打到。
1034. 我臨走嗰陣唔記得 lock 門呀，好彩咩事都冇。
1035. 我部電腦俾佢 lock 咗嚟用。
1036. 我去鎖 locker，點知一返轉頭佢已經走咗。
1037. 我搵吓啦，應該喺是但一個人個 locker 度嘅。
1038. 你個 login name 一定要全部細楷。
1039. 我唔記得咗個 login name 係咩添。
1040. 你學校個 logo 好靚喎。
1041. 呢個 logo 係咪好特別呢？
1042. 你今日個 look 好得喎。
1043. 佢話我個 look 唔似大學生。
1044. 今年只有星期四早放 lunch。
1045. lunch 竟然食咗少少就番去。
1046. 我記得上年大家嘅 lunch 時間唔同。
1047. lunch 食完飯再飲多包嘢。
1048. 平時冇咩人嘅 M 記，今日竟然大排長龍喎。
1049. 一返到去見到好多人都食緊 M 記。
1050. 你嘅 main point 究竟係乜嘢？
1051. 講咗好多費話，完全講唔到 main point。
1052. 你 major 讀咩？
1053. 唔理你 major 邊科，大家都係考同一個試。
1054. majority 嘅人都唔知道抑鬱症嘅成因。
1055. 其實 majority 嘅人都支持佢連任。
1056. 駛唔駛 make appointment 呀？
1057. 冇 make appointment 嘅話可能要等好耐。
1058. 做啲嘢都唔 make sense 嘅。
1059. 你啲問題好唔 make sense。
1060. 你 make sure 自己已經交齊所有資料？
1061. 你唔 make sure 就唔好咁大聲。
1062. 不過，話明 management trainee，應該捱完就有光明嘅前途。
1063. 但係好似乜嘢系畢業嘅人都可以做 management trainee。
1064. 我個 manager 係一個女人。
1065. 個 manager 話咁就咁。
1066. 唔知幾時先至升到 manager 呢？
1067. 嘩！淨係本 manual 都有排睇。
1068. 份 manual 全部都係我寫㗎。
1069. 本記事簿入面有學校嘅 map。
1070. 我畫完幅所謂嘅 map 就收工。
1071. 個 margin 要幾闊呀？
1072. 我已經留咗好多 margin 位。
1073. 呢個 market 好細。
1074. 你唔覺得好有 market 咩？
1075. 你識唔識人讀 marketing？

1076. 我覺得都係讀 marketing 好啲。
1077. 如果搽大量 mascara 就唔需要畫眼線。
1078. 點樣搽 mascara？
1079. 我見佢做緊 mask。
1080. 應該買邊個牌子嘅 mask 好呢？
1081. 佢依家讀緊 master。
1082. 讀完 master 之後有咩打算？
1083. 到咗全晚焦點所在，就係最佳服裝獎同最 match 服裝獎。
1084. 都幾好，兩個啲話題都好 match。
1085. 有啲咩 material 可以用呀？
1086. 唔係全部 material 都咁易搵到。
1087. 你要嘅 material 我已經準備好。
1088. 結果 MC 被迫要取消個遊戲。
1089. 全部都係 MC，唔識做都可以撞吓。
1090. 文學唔係睇句子結構而係睇 meaning。
1091. 佢講嘅每句說話都有 meaning。
1092. 呢套電影除咗好笑之外仲好 meaningful。
1093. meaningful 唔一定要用好多字㗎。
1094. 你可唔可以教我 measure 個長度嘅方法呀？
1095. 一次唔得咪 measure 多次囉。
1096. 今次考試個 median 係幾多分？
1097. 今次個 median 高過上次。
1098. 幾時會再有 meeting？
1099. 一陣個 meeting 有幾多人出席？
1100. member 嘅質素好與壞。
1101. 啲 member 又真係好似跳得好咗咁喎，唔知係咪心理作用。
1102. 你張 memo 打錯字喎。
1103. 佢連打張 memo 都唔識，點做嘢㗎？
1104. 個腦好亂，mentally 好攰。
1105. mentally 好精神，不過體力支持唔到。
1106. 我部電話入面啲 message 冇晒。
1107. 有兩個 message 同大家分享。
1108. 呢個 message 係咩意思？
1109. 幾時 mid-term 呀？
1110. 下個星期就要 mid-term。
1111. 但係我 mind 喎。
1112. 佢自己話唔 mind 嘅。
1113. 你 mind 唔 mind 我借你本書？
1114. 佢問我 mind 唔 mind 佢食煙。
1115. 佢 minor 日本研究。
1116. 你以前有冇 minor 過任何科目？
1117. 佢呢個表達方式好 misleading 喎。
1118. 佢講嘢有時都幾 misleading。
1119. Miss「無啦啦」叫我哋睇兩本書，然後寫一篇六百字嘅感想。
1120. Miss 本來叫我哋四個派一個代表起身講。
1121. 跟住 Miss 叫我哋自我介紹。
1122. 我 miss 咗上星期嘅講座。
1123. 佢上個星期 miss 咗一堂。
1124. 我最討厭 missed call。
1125. 我哋一傾就一個鐘，期間無限個 missed call。
1126. 有時都會挑選一啲平嘅牌子自己 mix and match。
1127. 淨色衫可以容易啲 mix and match。
1128. 去美國做咗幾年全職 model。
1129. 好多人以爲做 model 搵錢好容易。
1130. 佢喺份 model answer 上面畫公仔。
1131. 擺住份 model answer 照抄，咁做功課仲有咩意義？
1132. 原來 modem 係兩個字嚟㗎！
1133. 我個 modem 壞咗。
1134. 呢度嘅裝修好 modern。
1135. 呢個古董手袋個款好 modern，一啲都唔老土。
1136. 喺嗰個 moment 我真係好感動。
1137. 呢個 moment 唔適合講呢樣嘢住。
1138. 而家啲 mon 係咪會少啲幅射？
1139. 你想買邊種 mon 呀？
1140. 近排冇乜 mood 做嘢。
1141. 跟住又冇晒 mood 溫書。
1142. 瞓咗一陣就俾 morning call 嘈醒咗。
1143. 要唔要我 morning call 叫你起身？
1144. 隻火牛嘅風扇壞咗會唔會影響到塊 motherboard㗎？
1145. 我想換過塊 motherboard。
1146. 點樣可以將個硬盤 mount 落部電腦度呀？
1147. 佢想買可以 mount 上櫃嗰啲電腦。
1148. 我隻 mouse 好襟用。
1149. 九十八蚊一隻 mouse 係咪好貴？
1150. 你有冇啲舊歌嘅 MP3 呀？
1151. 好想要部新嘅 MP3。
1152. 你部 MP3 機係咪有得錄音㗎？
1153. 等我睇下邊個有 MP3 先。
1154. 佢依家讀緊 MPhil。
1155. 佢男朋友都係讀緊 MPhil。
1156. 第一次玩 MSN 玩到唔停手。
1157. 你識唔識用 MSN？
1158. 真係好想知 MT 究竟係咪我想像中嗰回事。
1159. 原本請咗五個 MT。
1160. 呢份卷係 multiple choice。
1161. 有得揀我梗係想做 multiple choice 啦。
1162. 你做人唔好成日都咁 negative 啦！
1163. 點知邊面係 negative 呀？
1164. 佢唔會 neglect 任何結識異性嘅機會。
1165. 佢就係 neglect 咗呢樣嘢所以先至錯。
1166. 又爲 neighbour 祈禱。
1167. 主席叫我哋同 neighbour 手牽手。

1168. 我已經「的」起心肝，申請咗 Netvigator。
1169. Netvigator 雖然係貴啲但勝在穩定。
1170. 一等到 neway 開門，佢哋就空群而出。
1171. 我都係鍾意去 neway 多啲。
1172. 去到 neway city，今日竟然十點九都未開門。
1173. 旺角都有 neway city 啦。
1174. 點樣睇 newsgroup？
1175. 點解喺屋企睇唔到中大嘅 newsgroup 嘅？
1176. 佢份人好 nice。
1177. 佢咁 nice 一定肯幫你嘅。
1178. 你有冇 nickname？
1179. 佢個 nickname 好得意。
1180. 今日放假，出咗去睇 Nike 有冇新貨。
1181. 晏晝就過咗旺角睇人哋 Nike 試範。
1182. 你嘅做法好 normal。
1183. 個角度係咪由 normal 嗰邊計起？
1184. 佢部 notebook 又攞咗去整。
1185. 邊隻牌子嘅 notebook 最襟用？
1186. 今次終於有 notes。
1187. 落咗堂去試下點用影印機，印頭先份 notes。
1188. 我而家用緊 now 嘅有線電視服務。
1189. 係咪一定要用網上行先至可以睇到 now？
1190. 你張單個 number 好唔清楚喎。
1191. 可唔可以俾你嘅電話 number 我。
1192. 你識唔識人讀 nursing？
1193. 我家姐都係讀 nursing。
1194. 組媽打電話通知我 O camp 時間。
1195. 其實我都冇去 O camp，所以冇乜感受。
1196. 佢好 objective 所以大家都會問佢意見。
1197. 其實佢都係想你 objective 啲啫。
1198. occasionally 佢會遲到。
1199. 佢都只係 occasionally 過嚟搵你啫。
1200. 今日好開心因爲我去咗 ocean park。
1201. 之前去 ocean park 嘅時候。
1202. 佢女朋友個 offer，咁啱又係去上海。
1203. 佢一共有三個 offer。
1204. 你有冇佢 office 嘅內線？
1205. 你駛唔駛日日坐喺 office㗎？
1206. office 啲人又走埋嚟搵我。
1207. 下個星期係 week 幾？
1208. 喺 office 做嘢都可以做到索索地氣㗎！
1209. 你 office hour 再打過嚟啦。
1210. 點解 office hour 一個人都冇㗎？
1211. 佢話要瞓覺要 offline 喎。
1212. 聯絡唔到佢，佢 offline 咗。
1213. 照平日返工嘅步速就 OK。
1214. 其實咁都 OK 驚，好彩唔駛我企埋出去講。
1215. 我都 OK 佩服佢。
1216. 我都 OK 驚下，點可以咁兒戲？
1217. 呢度嘅食客都係斯文嘅 OL。
1218. 呢度長時間都會有一班 OL 捧場。
1219. 你出去比賽係 on behalf of 學校。
1220. 你依家係 on behalf of 邊個先？
1221. 着住最新嘅花花雪紡 one piece 裙。
1222. 呢條 one piece 襯晒你啦。
1223. 你通常幾時 online㗎？
1224. 更加唔想見到佢 online。
1225. 係咪個個後生仔都玩 online game㗎？
1226. 依家啲 online game 其實有冇錢賺？
1227. 今日終於 on 返 line 喇。
1228. 斷完線之後 on 返 line，一個人都冇。
1229. 圖書館幾時 open 呀？
1230. 佢冇你諗得咁 open。
1231. 呢份卷係 open book 嘅。
1232. 多數考試都係 open book。
1233. 呢份卷係 open notes 嘅。
1234. open notes 要抄貓紙，好嘥時間。
1235. 今朝返完 open U 就行咗一陣街。
1236. open U 畢業，啲公司承唔承認㗎？
1237. 問題係我冇其他 option。
1238. 如果有另一個 option 呢？
1239. 日文堂係咁玩 oral。
1240. 之後上日文，兩堂都係咁玩 oral。
1241. 意粉係接到 order 先至淥。
1242. 要積極啲唔好坐喺度等 order。
1243. 未來嘅工作性質會係成績 oriented。
1244. 好多時讀書都係考試 oriented。
1245. 不過唯一慘情就係要幫佢通頂裝好個 OS。
1246. 你用開邊個 OS？
1247. 不過我都唔覺佢開 OT。
1248. 仲未做晒啲嘢，所以要開 OT。
1249. 今日教會 outing 去墓地。
1250. 其實我哋好少去 outing。
1251. 我唔識用 outlook 嚟睇新聞組。
1252. 咁你識唔識用 outlook 呀？
1253. 今年冇咩好 outstanding 嘅人嚟參加面試。
1254. 佢咁 outstanding，去到邊度都掂啦。
1255. 你有冇參加過 outward bound？
1256. 估唔到佢都去到 outward bound。
1257. 集齊兩張分別 over 廿蚊嘅收據寄返嚟俾我哋。
1258. 如果 over 八千蚊我就唔買囉。
1259. 多唔多人去 overseas 升學㗎？
1260. 去 overseas 唔係唔好，不過我都係鍾意香港多啲。
1261. 你 pack 咗你啲行李未？
1262. 啲行李 pack 得好啲就可以放多啲嘢。
1263. 呢個酒店加機票 package 都只係二千幾

蚊。

1264. 有冇啲平平哋嘅 package 包埋早餐㗎？

1265. 之前幾次都只係得一 pair 拍緊拖。

1266. 依家就變成三 pair。

1267. 我想買部 palm 送俾人。

1268. 我跌咗部 palm 落地。

1269. 見到檔賣 pancake 嘅，好好食。

1270. 我食咗好多嘢：燒雞串、燒賣、雞蛋仔同埋 pancake。

1271. 寫一份 paper 要用幾多時間呢？

1272. 我又要開始趕我份 paper。

1273. 我想問咩係 parallel port 呀？

1274. 係咪得一個 parallel port㗎咋？

1275. 噚日係 Parent's Day，我覺得自己真係好叻。

1276. 每年都會有兩次 Parent's Day。

1277. 普通話歌唱比賽，有一 part 係估歌仔。

1278. 特別係中間嗰 part。

1279. 我真係覺得好好睇喎，尤其是鄭中基嗰 part 真係好搞笑。

1280. 呢 part 精彩呀，上火車嗰陣真係亂到吖。

1281. 我又覺得唔係咁好，驚累咗個 partner。

1282. 今日打排球終於有 partner，球技同我差唔多。

1283. 佢喺度做 part-time 已經做咗五年。

1284. 你以前做過啲咩 part-time？

1285. 如果表現得好，有機會可以長做 part-time。

1286. 去 party 撞到咪打個招呼囉。

1287. 感恩會之後去咗開 party。

1288. 我做完會 pass 俾你。

1289. 幾時可以 pass 啲文件俾我。

1290. 你用完先 pass 俾我啦！

1291. 仲辛苦過大學讀書，因爲 passing rate 極低。

1292. passing rate 咁高，唔駛擔心啦。

1293. 之後「拿拿林」去灣仔拎 passport。

1294. 本 passport 過咗期幾個月。

1295. 佢好鍾意用生日日期嚟做 password。

1296. 個 password 要有數目字同英文字母。

1297. 坐車坐到腳又痛 pat pat 又痛。

1298. 成日坐地下，坐到 pat pat 好痛。

1299. 佢份人好冇 patient，成日都好趕咁。

1300. 佢啲 patient 好多都好有錢㗎。

1301. 自己設計一啲 pattern。

1302. 推出以經典格仔 pattern 爲題嘅餐具系列。

1303. 仲要係冇得 pay 嘅。

1304. 有得 pay 咪落力啲囉。

1305. 有邊啲地方需要 pay attention？

1306. 除咗開始嘅時候要 pay attention，其他時候都唔應該鬆懈。

1307. 邊隻牌子嘅 PDA 最好用呀？

1308. 我想買部有埋藍芽嘅 PDA。

1309. 你幾時上 PE？

1310. 跟住嗰兩堂 PE 堂我都冇上到。

1311. 有五個 percent 嘅人幫襯你都發達啦！

1312. 產生咗感情就會一百 percent 投入。

1313. 佢樣樣嘢都要做到 perfect。

1314. 呢個世界冇咁 perfect。

1315. 我信得過波子嘅 performance。

1316. 佢噚日嘅 performance 好好。

1317. 我個組爸係讀 pharmacy 嘅。

1318. pharmacy 畢業好易就搵到工。

1319. 我有個朋友讀緊 PhD。

1320. 我都冇諗住讀 PhD。

1321. 唔該晒你俾本 photo album 我。

1322. 個 photo album 可以放幾多相？

1323. 我已經 photocopy 咗成份嘢。

1324. photocopy 同正本要分開放好。

1325. 皆因小弟一向都冇乜 physical。

1326. 加上我又冇乜 physical。

1327. 佢去佢嘅 Pizza Hut，我返我嘅屋企。

1328. 我屋企樓下開咗間 Pizza Hut。

1329. 星期日 plan 咗去大嶼山同人慶祝生日。

1330. 原先 plan 咗去泰國，依家可能冇得去。

1331. 你嘅所謂 planning 用咗幾多時間做？

1332. 做嘢有 planning 唔係唔好。

1333. 能夠參加今次嘅面試已經係我嘅 pleasure。

1334. 我覺得見到佢已經係我嘅 pleasure。

1335. 呢幅圖點 plot㗎？

1336. 份功課要自己 plot 圖喎。

1337. 佢講嘢都冇 point 嘅。

1338. 一個 point 先至得五分。

1339. 問吓你美國加息對香港嘅影響，同埋香港政府嘅 policy 咁。

1340. 工作包括制定政府嘅 policy。

1341. 而家最 popular 嘅歌星係邊個？

1342. 依家邊個歌星最 popular？

1343. 佢話我個 port 可能有問題。

1344. 我想開返用嚟上網嗰個 port。

1345. 所擺嘅 pose，所有嘅表情，都冇晒。

1346. 呢個 pose 好有型。

1347. 冇法子視之爲一件有 positive 影響嘅事吧？

1348. 影響係咪咁 positive 要遲啲先知。

1349. 呢個 post 嘅工作量好龐大。

1350. 有人 post 嘢上去喎。

1351. 你張 poster 喺邊度買㗎？

1352. 貼 poster 要注意啲咩？

1353. 佢係 postgrad 嚟㗎

1354. 讀 postgrad 係咪要好好成績㗎？

1355. 我唔介意 postpone 個表演。

1356. 個測驗 postpone 咗，可以多啲時間準備。
1357. 你好有 potential 成爲天王巨星。
1358. 可能佢嘅 potential 好大。
1359. 呢科教啲嘢好 practical。
1360. 我有個好 practical 嘅問題。
1361. 第一日返就係 practice 用鎖匙，係咪笑死呢？
1362. 今次咁差都係因爲 practice 得唔夠啫。
1363. 我當時真係唔記得咗啲 prefect 會上嚟。
1364. 以前啲 prefect 好惡㗎！
1365. 我 prefer 第二個建議。
1366. 佢其實 prefer 我負責帶嘢。
1367. 可能我心裡已經有 preference。
1368. 佢哋一早有 preference，個面試只係做吓樣。
1369. 你嘅 preparation 做好未？
1370. 如果夠時間做 preparation 就唔會搞成咁啦。
1371. 你 prepare 咗聽日要用嘅嘢未？
1372. 我當慳番啖氣，所以唔 prepare。
1373. 總算爲佢 prepare 好呢個生日喇。
1374. 你幾時做 present 呀？
1375. 我一陣 present 用中文冇問題呵？
1376. 我下晝 present 完搵你吖。
1377. 佢可能 pressure 太大。
1378. 如果冇 pressure 你會咁落力咩？
1379. 你 print 咗份功課未？
1380. 你 print 咗一陣開會用嘅嘢未？
1381. 而家啲 printer 可以直接印相。
1382. 呢部 printer 可唔可以印到雙面？
1383. 真係要學會點樣編排自己嘅 priority。
1384. 喺我心目中，你嘅 priority 梗係高啲啦。
1385. 你哋咁樣影相係侵犯我嘅 privacy。
1386. 拉埋個簾落嚟有完全嘅 privacy。
1387. 有個超 pro 嘅城大生，個樣已經殺晒。
1388. 喺側邊聽啲 pro 嘅影相佬講嘢。
1389. 呢個 probability 好低。
1390. 唔理 probability 係幾多都好，盡咗力就算。
1391. 你做嘢嗰度過咗 probation 未？
1392. 過咗 probation 有得加人工。
1393. 你應該喺份報告入面寫埋個 procedure。
1394. 你都唔跟個 procedure，梗係做唔到啦。
1395. 試過我哋公司嘅 product 你一定會滿意。
1396. 手提電話真係一種令我又愛又恨嘅 product。
1397. 對上一次真真正正嘅 production 係上年。
1398. 我哋淨係做 production，其他嘢由第二個部門負責。
1399. 今日好 productive，唔似溫書咁，懶懶閒。
1400. 佢好 productive，一個人可以做兩個人嘅嘢。
1401. 佢好似好 professional 咁。
1402. 佢啲口吻好唔 professional。
1403. 你懶係 professional 咁。
1404. 呢個 professor 好受學生歡迎。
1405. 會唔會令個 professor 留低壞印象㗎？
1406. 你個 program 寫得好好。
1407. 你寫啲 program 好唔掂喎。
1408. 個 project 點樣交呀？
1409. 爲咗呢幾個 project，我連續頂咗四晚通宵。
1410. 個 project 點樣交呀？
1411. 識到一班好朋友，一齊做 project。
1412. 我已經唔敢 promise 話唔轉工。
1413. 我 promise 咗下年會打俾佢。
1414. 佢成日 promote 呢間公司嘅產品。
1415. 我用開嗰隻洗頭水而家做緊 promote。
1416. 佢啲英文嘅 pronuciation 好正。
1417. 唔好話 pronuciation 啦，好多字我連讀都唔識讀。
1418. 如果個賠償係 proportional 咁計，我咪發達？
1419. 如果要 proportional 咁計，咪要好多錢囉。
1420. 份 proposal 幾多人一組呀？
1421. 佢叫我寫份 proposal 俾佢睇先。
1422. 佢 propose 我下次先至講埋淨番嘅嘢。
1423. 我可以同老細 propose，不過佢未必贊成㗎。
1424. 呢份 prospectus 喺邊度攞㗎？
1425. 今年本 prospectus 靚過舊年嗰本。
1426. 佢 proud of 自己係貴族嘅身份。
1427. 佢希望爸爸會好 proud of 自己。
1428. 今日成日都喺度玩 PS2。
1429. 今次抽獎嘅大獎係一部 PS2。
1430. 聽日 public holiday，所以佢唔俾多啲時間我哋。
1431. 除咗 public holiday 之外，差唔多日日都要返工。
1432. 本書幾時 publish㗎？
1433. 佢公司係幫人 publish 年報嘅。
1434. 你最多可以做到幾多下 push up？
1435. 我諗佢一下 push up 都做唔到。
1436. 我見佢個樣好 puzzling，所以我將個問題再解釋多一次。
1437. 如果大家都咁 puzzling，不如再講多一次。
1438. 我哋班竟然撻 Q，你話係咪豈有此理呀？
1439. 但係條 Q 跑唔出。
1440. Q&A 嘅時候居然冇人問問題。
1441. 有幾多時間做 Q&A？
1442. 何況佢都唔夠 quali 同我撼。
1443. 佢 quali 咁好，大把人爭住要。
1444. 你要咁高 quality 嚟做咩？

1445. 我同佢影嗰張相 quality 太低，晒唔返出嚟。
1446. 今個 quarter 考得好好。
1447. 一年有幾多個 quarter？
1448. 最後一次同佢哋 quick change 喇。
1449. 你估佢會幾時 quit 吖啍？
1450. 如果呢個月都係咁，我真係想唔 quit 都唔得啦！
1451. 今次個 quiz 嘅問題同上次差唔多。
1452. 個 quiz 得廿二分，死梗。
1453. 雖然唔計 quota，但我都要做啲成績俾人睇。
1454. 反正我哋用唔晒啲 quota，咪分啲俾人囉。
1455. 而家啲 RAM 平咗好多。
1456. 我想問啲 RAM 係咪有分唔同速度㗎？
1457. 親自爲大碟撰寫 rap 詞。
1458. 呢首歌嘅 rap 由佢自己負責番。
1459. 我隨時 ready。
1460. 我隨時 ready。
1461. 只係等待 ready 嘅一刹那。
1462. 當時的確係未 ready。
1463. 唔用 real time scan 可能中咗毒都唔知。
1464. 開咗 real time scan 之後部機好似慢咗咁。
1465. 我咁講梗有我自己嘅 reason 啦。
1466. 你梗有自己嘅 reason。
1467. 你中咗六十秒自動 reboot 嗰隻毒？
1468. 佢部電腦唔知點解久不久就會自己 reboot。
1469. 如果有 recall 你就要提早還書。
1470. 我借嗰本書又有 recall 喇。
1471. 要攞住張 receipt 入場。
1472. 佢唔見咗張 receipt。
1473. 你有冇 receive 過有關電腦病毒嘅電郵？
1474. 佢一 receive 嘢就死機。
1475. 佢 recommend 我睇呢幾本書。
1476. 我唔會 recommend 大家嚟呢度食嘢。
1477. 唔知幾時先可以 recover 呢？
1478. 年紀大咗要多幾日至可以 recover。
1479. 我公司最近 recruit 咗幾個新人。
1480. 佢地會唔會 recruit 新會員呀？
1481. 操完真係要搵 red cross 睇下。
1482. 我中學嗰陣都做過 red cross。
1483. 點樣自動 redirect 啲人去另一個網頁？
1484. 我俾人 redirect 咗去第二個網頁。
1485. 我想搵呢本 reference book。
1486. reference book 唔一定要買，去圖書館借都得。
1487. 我 reg 唔到個普通話。
1488. 今個禮拜又 reg 唔到，所以多咗兩堂空堂。
1489. 邊度可以搵到圖書館嘅 regulation 呀？
1490. 有啲咩 regulation 我係要遵守㗎？
1491. 踢到四點就返咗去 rehearsal。
1492. 個 rehearsal 一個鐘搞唔搞得掂？
1493. 佢份建議俾人 reject 咗。
1494. 我唔會亂咁 reject 人嘅。
1495. 你哋咩 relation？
1496. 佢哋嘅 relation 唔係想像中咁簡單。
1497. 燃點薰衣草香薰油可以 relax 同減壓。
1498. 呢種感覺的確好舒服，好 relax。
1499. 好心你唔好咁緊張，relax 一吓啦！
1500. 邊度可以搵到 relevant 嘅資料？
1501. 佢講啲嘢同今日嘅主題一啲都唔 relevant。
1502. 我用嚟用去都係嗰幾個 vocab。
1503. 佢講咗好多我未聽過嘅 vocab。
1504. 對我嘅熱情已經開始減退，冇咗之前嘅 rely。
1505. 佢以前好 rely 我㗎。
1506. 上晝同下晝都有人 remind 我聽日唔駛返工。
1507. 佢已經 remind 咗我，係我自己唔記得啫。
1508. 係屋企用唔用到 remote desktop㗎！
1509. 點解我用唔到 remote desktop 嘅？
1510. 一本書可以 renew 兩次。
1511. 我本書已經 renew 咗兩次，一定要還。
1512. 有啲好聽啲嘅我咪 repeat 囉。
1513. 佢中五嗰陣已經 repeat 過兩次。
1514. 果然記憶係受 repetition 影響。
1515. 係咪 repetition 越多就越入腦先？
1516. 有邊啲書係教人寫 report㗎？
1517. 呢份 report 係自己做定抄人哋㗎？
1518. 因爲返工 require 我着黑色衫，於是我又買咗兩件。
1519. 係佢 require 你咁做定係你自己想咁做？
1520. 上莊有啲咩 requirement？
1521. 你一定要乎合份工嘅 requirement 先至有機會。
1522. 你 research 做啲咩㗎？
1523. 你做晒 research 之後先至去見工，會令人覺得你好有誠意。
1524. 做 research assistant 要咩條件？
1525. 我一畢業就做 research assistant，之後先至出去搵工。
1526. 呢度係冇得 reserve㗎！
1527. 請問兩位有冇 reserve 到檯㗎？
1528. 點樣將個 resolution 改低啲呀？
1529. 我真係好想問下佢知唔知乜嘢係 resolution。
1530. 如果 resources 唔夠嘅話就乜都冇得講。
1531. 都冇 resources，叫人點做嘢喎。
1532. 佢覺得自己俾咗錢所以唔駛 respect 啲老師。

1533. 佢識得 respect 人㗎咩？
1534. 佢好 responsible 所以啲嘢可以交晒俾佢。
1535. 你咁 responsible 咪幫佢做埋份功課囉。
1536. result 同我預料嘅差唔多。
1537. 以前會將 result 貼喺壁報板上。
1538. 做完個 review 之後就俾政府參考。
1539. review 完之後仲要等幾耐？
1540. 係咪隻隻碟都可以 rewrite㗎？
1541. 可以 rewrite 嗰啲碟幾錢隻？
1542. 唔怪之得味道咁 rich。
1543. 有好 rich 嘅水果味。
1544. 你 right click 就得㗎喇。
1545. 點解我 right click 完之後冇反應嘅？
1546. 我發現世界真係好細，原來我 roomate 同新同事係同學。
1547. 佢 roomate 今朝五點幾至返。
1548. 再嚟多 round 有冇問題？
1549. 佢第一個 round 就俾人飛咗出局。
1550. 抑或買 router 好呢？
1551. 唔用 router 得唔得？
1552. 今晚都係嚟睇位同睇佢個 rundown。
1553. 因爲知道佢成個 rundown，所以影出嚟唔錯。
1554. 呢期好忙，搞 SA 選舉啲嘢已經忙到死咗。
1555. 玩 SA 其實唔係咩大問題。
1556. 我都已經死忍，成日諗住啲 salary，千祈唔好臨尾香。
1557. salary 之外，表現好仲有額外獎金。
1558. 我哋一見到有 sale 字眼嘅鋪頭就入去。
1559. 你等到 sale 嗰陣至入貨囉。
1560. 據店鋪 sales 姐姐講。
1561. 我會有耐性咁聽個 sales 講解。
1562. 我覺得做 sales 嘅女仔好多都唔乖。
1563. 佢有十幾間分店，仲要唔計 salon，真係記死我。
1564. 佢兩個星期就去一次 salon 整頭髮。
1565. 個訊號再 sampling 之後會慳位啲。
1566. 做完 sampling 之後會點？
1567. 咁我咪同佢 say bye 囉。
1568. 我一 say bye 佢就到。
1569. 我諗住走去 say hi 啦，點知原來認錯人。
1570. 我同佢揮手 say hi，點知佢好似冇咩反應咁。
1571. 你叫到我又點會 say no 呢？
1572. 雖然明知做阿四但我都唔會 say no。
1573. 佢唔 say sorry 不特旨，態度仲好差。
1574. 條友見我超住佢先至即刻 say sorry。
1575. 我想 scan 呢幅圖。
1576. 唔該幫我 scan 咗呢幾頁佢吖。
1577. 根據而家嘅 schedule 我哋可以好快教晒啲嘢。
1578. 每日嘅 schedule 都排得密麻麻。
1579. 我以前係讀 science 嘅。
1580. 多數人都覺得讀 science 叻啲。
1581. 你做嘢 scientific 啲得唔得？
1582. 佢連食飯都好 scientific。
1583. 你個 screen saver 好靚喎。
1584. 我用自己影嘅相做 screen saver。
1585. 你自己上網 search 啦！
1586. search 完間公司嘅資料，我就即刻瞓覺。
1587. 呢個 season 唔太熱又唔太凍，最適合去行山。
1588. 自從第三個 season 之後，我都冇再睇囉。
1589. 幾時會有 seating plan？
1590. 邊度可以搵到個 seating plan 呀？
1591. 今日一早就去 second in。
1592. 以爲已經冇機會 second in。
1593. 你會聽邊個 section？
1594. 一陣間嗰個 section 由另一位老師負責。
1595. 我哋同公司但係唔同 section。
1596. 呢個 section 嘅同事都好好人。
1597. 學校嘅 security 似乎唔係好夠。
1598. security 唔係冇，不過唔夠啫。
1599. 你唔可以 selective 咁揀嚟讀㗎喎。
1600. 我唔會 selective 咁淨係問某幾個人問題囉。
1601. 係咪我唔識 sell 嘢，所以唔想俾我落鋪？
1602. 朝九晚六又唔駛企，唔駛 sell，係我夢寐以求嘅工作。
1603. 其實今個 sem 想讀返好啲書。
1604. 我今個 sem 最唔捨得嘅係佢。
1605. 佢哋啱啱參與完 seminar。
1606. seminar 之後可以一齊食飯。
1607. 我一陣會將嗰個會嘅資料 send 俾你。
1608. 人地 send 嘢俾我有時我會收唔到。
1609. 問我點解咁耐都唔 send 俾佢。
1610. 原來我會見兩個 senior。
1611. 佢唔係最高級，但係係最 senior 嗰個。
1612. 我 sense 到嘅時候已經太遲。
1613. 佢 sense 唔到咁大鑊。
1614. 我想買條 serial port 用嘅線。
1615. 佢都唔識咩係 serial port。
1616. 傾完啲好 serious 嘅嘢之後就傾閒計。
1617. 佢好 serious 咁問咗我一個問題。
1618. 點知之後要每次 serve 兩個。
1619. 我慣咗要人 serve 我。
1620. 初初都係每一次 serve 一個人。
1621. 我屋企部電腦係 server 嚟㗎！
1622. 佢要買 server，唔係買普通電腦。
1623. 有啲咩 service 係可以用㗎？
1624. service 絕對係一流。
1625. 我唔識 set 嗰部機。
1626. 好想啲 set 穩穩陣陣唔會爛。

1627. 好想個 set 靚到啲人見到會嘩一聲。
1628. 新版同舊版嘅 setting 係咪唔同㗎？
1629. 呢個 setting 同以前嗰啲唔同喎。
1630. 其實自己 setup 都唔係咁難啫。
1631. 我自己都識 setup 啦。
1632. 佢啲膚色真係好 sexy 呀。
1633. sexy 啲我咪鍾意多啲囉。
1634. 個餐好大份，兩個人 share 都夠食。
1635. 選擇「共用資料夾」就可以將啲嘢 share 俾人。
1636. 佢喺學校入面唔算太 sharp。
1637. 離遠望見個色已經 sharp 到爆。
1638. 點解有咁嘅 shift？
1639. 有啲人 shift 咗去隔離嗰行坐。
1640. 鍾唔鍾意同佢哋一齊 shopping 又係另一回事。
1641. 唔係個個女仔都咁鍾意 shopping㗎嘛。
1642. 佢 short㗎唔駛理佢。
1643. 佢都 short 嘅。
1644. 你明知佢 short㗎啦！
1645. 份卷係咪淨係得 short question？
1646. 話明係 short question，答幾句就夠。
1647. 其他啲二打六又想我哋幫佢哋整咗先，都 short short 地嘅。
1648. 我個電話 short short 地，間中會打唔到。
1649. 通常過幾日就會出 shortlist㗎喇。
1650. 滙豐筆試嘅 shortlist 出咗未？
1651. 呢個 shot 要再拍過。
1652. 上個 shot 拍得好好。
1653. 我想 show 呢幅圖俾佢睇。
1654. 開 show 時佢哋比噚日更緊張。
1655. 我見有人 show 佢，所以就幫下佢手。
1656. 話晒都係最後一次一齊玩，但又有人 show 我，所以都係走先。
1657. 或者佢係特登 show off㗎呢。
1658. 我即刻 show off 我個新銀包，佢都話靚呀。
1659. 呢次 singing contest 可能係我最後一次喇。
1660. 決定下年 singing contest 分兩組出賽。
1661. 咁啱阿 sir 今日唔駛我哋跑圈，唔係嘅話就大鑊。
1662. 阿 sir 仲話今堂係最後一堂教下手，之後就會學上手同開波。
1663. 我成日都去 sit 堂。
1664. 你會唔會去 sit 晒全部堂先？
1665. 輸咗要做十下 sit up。
1666. 我覺得做 sit up 有效啲。
1667. 都話我正處於「瘦田冇人耕，耕開有人爭」嘅 situation 啦。
1668. 呢個 situation 我都未見過。
1669. 一客嘅 size 足夠兩個人食。
1670. 迷你 size 嘅麵包可以一啖一個。
1671. 除咗呢啲 skill 之外你仲識啲咩？
1672. 我哋贏嘅可以話真係 skill。
1673. 佢地想請一個 skillful 嘅人。
1674. 一個 skillful，一個乜都唔識，你會請邊個呀？
1675. 今年生日考試，所以 skip 一年啦，下年繼續。
1676. 睇唔明咪 skip 咗佢先囉。
1677. 你 smart 少少得唔得？
1678. 呢個 smart 啲嘅都係咁，真係好唔掂。
1679. 一早收到你嘅 SMS。
1680. 我好怕打中文嘅 SMS。
1681. 你哋又要五日之後先再見到我㗎喇，so 五日後見啦。
1682. 你哋都唔想見到我俾人話㗎喇，so 唔該晒咁多位！
1683. 啲同事好好，係 so far 最好嘅一次。
1684. so far 我覺得佢表現唔錯。
1685. 一疊 soc 紙要幾多錢？
1686. 今年件 soc 褸都係以白色爲主。
1687. 我要學會 social。
1688. 我諗你要去 social 下至得。
1689. 我本來想讀 social work 嘅。
1690. 讀 social work 咪仲難搵工。
1691. 唔知點解啲相變到好似用咗 soft 鏡咁。
1692. 呢啲嘢俾人一個好 soft 嘅感覺。
1693. 你用邊隻 software 執相㗎？
1694. 燒唔到碟關唔關個 software 事？
1695. 你計到個 solution 未？
1696. 佢計嚟計去都計唔到個 solution。
1697. 今日落咗去又一城間豐澤訂部 Sony 八二八數碼相機。
1698. 買音響就買 Sony 啦，佢做收音機起家㗎嘛。
1699. sorry 我唔覺得喎。
1700. sorry 呀，其實你係邊個？我認唔出你把聲。
1701. 佢同我講 sort 完啲資料就走得。
1702. 我想將啲資料跟大細 sort 好。
1703. 我張 sound card 好似燒咗。
1704. 一張 sound card 要幾多錢呀？
1705. 好多時上網都會搵到 source code。
1706. 我有 source code，你唔駛寫得咁辛苦。
1707. 佢帶咗一份好特別嘅 souvenir 比我。
1708. 每次去旅行都會買好多 souvenir。
1709. 你得一套工具有 spare㗎？
1710. 你有冇預多啲做 spare㗎？
1711. 今日冇咩 special，都係咁悶。
1712. 影印嗰度冇咩 special，拎相嗰度先值得一提。
1713. 一陣李教授會有一個 speech 係關於選科

嘅。
1714. 我做 speech 已經做咗三年。
1715. 你估樣樣嘢都有得 sponsor㗎咩？
1716. 佢搵到好多公司 sponsor 呢個演唱會。
1717. 星期一係 Sports Day，要好早就去到。
1718. 我哋學校嘅 Sports Day 分兩日舉行。
1719. 我決定返屋企瞓陣先去學 squash。
1720. 佢個個星期都打 squash。
1721. 下一個 stage 幾時開始？
1722. 而家呢個 stage 仲未需要呢啲嘢住。
1723. 佢一點就已經 stand by 等我。
1724. 佢要我哋喺公司 stand by。
1725. 工作坊 stand for 咩嘢？
1726. 食環署 stand for 食物環境衛生署。
1727. 呢個係咩 standard 嚟㗎？
1728. 佢個 standard 你估人人都做到㗎？
1729. 有冇 standing 的確要靠關係。
1730. 估唔到佢咁有 standing 都會偷嘢。
1731. 邊度有就業資料嘅 statistics？
1732. statistics 顯示香港出生率下降，所以幼稚園首先會受到影響。
1733. 每個人都會有自己嘅 story。
1734. 佢構思緊一個關於海嘅 story。
1735. 我希望成隻碟都可以反映到 street culture。
1736. 致力推動香港 street culture。
1737. 你一定要 strong 過個女仔。
1738. 佢咁 strong 駛乜人保護呀？
1739. 大學嘅 structure 係點㗎？
1740. 唔同公司嘅 structure 都會唔同。
1741. 登入嘅時候要輸入 student ID。
1742. 佢連自己個 student ID 都唔記得。
1743. student union 有咩用？
1744. student union 係冇得退會㗎。
1745. 帶啲親友一齊去 studio 影畢業相。
1746. 我好想影 studio，話晒一世人一次吖嘛。
1747. 啲人今日去埃及 study aboard(lu)。
1748. study aboard 要好多錢㗎，所以我都係香港讀·
1749. 我想參加 study trip。
1750. 去 study trip 要用好多錢㗎。
1751. 唔知話佢 stupid 定咩好。
1752. 唔係話佢 stupid，只係大家諗法唔同啫。
1753. 你知係我 style㗎啦！
1754. 只不過係着件唔啱 style 嘅衫啫。
1755. 我唔係 stylist 唔會教你點襯衫。
1756. 仲送咗幅俾烏龍茶廣告嘅 stylist。
1757. 份功課 submit 去邊度？
1758. 佢已經 submit 咗份報告。
1759. 圖書館 subscribe 咗邊啲雜誌？
1760. 你有冇 subscribe 到呢份期刊？
1761. 係咪要行 subway 先至可以過到對面？
1762. 唔行 subway 過唔過到去？
1763. 佢 suddenly 撞門入嚟。
1764. 佢 suddenly 大叫一聲，跟住就暈咗。
1765. 我 suggest 佢聽日下晝先至嚟。
1766. 佢 suggest 我遲啲先至搵人。
1767. 我買咗套深灰色嘅 suit，係用嚟返工着嘅。
1768. 一季有兩套 suit 已經好夠。
1769. 佢係咪最 suitable 嘅人先？
1770. 我覺得自己唔係咁 suitable 做呢份工。
1771. 個 summary 大約要寫幾多字？
1772. 個 summary 一定要到肉。
1773. 要等到 summer holiday 先得，點算？
1774. 每年 summer holiday 佢都會返香港探屋企人。
1775. 好想搵份 summer job。
1776. 我之前做 summer job 見過佢。
1777. 主管同帶我個 supervisor 都冇話有問題。
1778. 佢係一個好好嘅 supervisor。
1779. 無論咩事你都知佢哋會 support 自己。
1780. 我部電腦唔 support 無線上網。
1781. 呢部機係咪 support 藍牙㗎？
1782. suppose 你唔應該喺度。
1783. 呢啲嘢你 suppose 做咗㗎喇喎。
1784. 你部電話 sup 唔 support 藍牙㗎？
1785. 我唔知個電話 sup 唔 support 視象短訊喎。
1786. 你 sure 我哋一定得？
1787. 你 sure 佢一定得？
1788. 佢令我好 surprise。
1789. 呢啲 surprise 係驚定喜真係見仁見智。
1790. 我好 surprise 佢係呢個時候出現。
1791. 俾唔到乜嘢 surprise 大家。
1792. 佢叫我幫佢做個 survey，有四百蚊賺。
1793. 你可唔可以幫我做個 survey？
1794. 喺我哋最 sweet 嘅時候。
1795. 你話幾感人幾 sweet 呢。
1796. 我想要嘅係真正嘅 sweet。
1797. 有冇人想買 switch 呀？
1798. 我想買隻 switch 將兩部機駁埋。
1799. 交埋份 symbols，又完咗兩年大學生活。
1800. 點樣可以打 symbols 打得快啲？
1801. 呢個 system 係用嚟查成績嘅。
1802. 我最鍾意公平嘅 system。
1803. 佢話你做嘢唔夠 systematic 喎。
1804. 我爸爸做嘢好 systematic㗎！
1805. 我都唔係好明點解要請個 TA 嚟。
1806. 有啲大學會請全職嘅 TA。
1807. 我舊年 take 咗一個通識。
1808. 乜你唔係上個學期已經 take 咗喇咩？
1809. 等佢出聲先至 take action。
1810. 幾時可以 take action？
1811. 駛唔駛 take attendance 呀？

1812. 我哋上堂唔駛 take attendance㗎？
1813. 冇人會 take care 你嘅感受。
1814. 你自己 take care 啦！
1815. 我依家先知原來可以 take notes。
1816. 我通常都會 take notes。
1817. 佢令人覺得佢好 talent。
1818. 每一個人都有自己嘅 talent。
1819. 幾時再有類似嘅 talk？
1820. 我對呢啲 talk 好有興趣。
1821. 我九點入到去，原來已經講緊 talk。
1822. 你呢個講座嘅 target 係乜嘢人？
1823. 你嘅 target 係咩先？
1824. 個 task 係譯一句英文，然後講下點解咁譯。
1825. 做晒三個 task 就可以走。
1826. 可以反映出屋主嘅 taste。
1827. 冇 taste 都唔會揀你做朋友啦。
1828. 食 tea 唔好食太多煎炸嘢。
1829. 好彩主管好好人，放我食 tea。
1830. tea set 仲包括兩個雪糕球。
1831. 我食唔晒一個 tea set。
1832. 我同佢係同一個 team 嘅。
1833. 同一 team 都唔會日日見到。
1834. 做嘢好多時都係 team work。
1835. 唔係得你一個人做，我哋 team work㗎嘛。
1836. 而家仲有好多 technical problems 要解決。
1837. 咁多 technical problems 唔怪得遲遲都起唔到貨啦。
1838. 個 technical staff 明明冇同我講過，不過，我都會繼續等。
1839. 好彩個 technical staff 肯幫我咋。
1840. 有嘢唔識可以問 technician。
1841. 我完全唔明個 technician 講乜。
1842. 大學會教一啲最新嘅 technology。
1843. 呢個世界咁多新嘅 technology，唔識並唔出奇。
1844. 身穿 tee 恤牛仔褲嘅五呎八吋高男子。
1845. 呢件 tee 係獨一無二嘅。
1846. 無奈我只係個 temp。
1847. 依家政府請人多數都係 temp，好少請長工。
1848. 呢個 term，我唔識點譯，唔知點擺啲次序。
1849. 佢係咁撻啲完全唔明係乜嘅 term 出嚟。
1850. 你唔係下星期有 test 咩？
1851. 呢個 test 咁淺，實合格啦。
1852. 今晚返學又有 test。
1853. 呢啲情況往往就 test 到自己都唔知嘅嘢。
1854. 我好想當面 thank 佢。
1855. 我唔知佢係咪想 thank you 我。
1856. 呢個 theory 有咩用？
1857. 呢個咩 theory 嚟㗎？
1858. 你可唔可以借份 thesis 俾我睇呀？
1859. 佢因爲 thesis 嘅問題未能參加呢次旅程。
1860. time management 對我嚟講非常重要。
1861. 我諗我要學下點樣做 time management。
1862. 你出咗 timetable 未？
1863. 我個 timetable 已經出咗。
1864. timing 係非常可惡嘅把戲。
1865. 如果 timing 啱嘅都冇所謂。
1866. 最尾嗰堂通常係講考試 tips。
1867. 食飯駛唔駛俾 tips㗎？
1868. tiramisu 原來本身係冇酒嘅。
1869. 我真係好鍾意個 tiramisu 呀！
1870. 百佳嘅 tissue 而家大減價。
1871. 而家買報紙仲有冇 tissue 送？
1872. 佢去咗 toilet，麻煩你等一等。
1873. 全場遲到，行幾個圈，再去埋 toilet 先至有人嚟。
1874. toilet 裡面有個婆婆猛咁同我哋傾計。
1875. 之後由課室玩到入 toilet。
1876. 佢係呢間公司最 top 嗰幾個。
1877. 要做到最 top 嗰個，一啲都唔容易。
1878. 每次一睇到嗰啲 topic 都會記番起裡面嘅內容。
1879. 呢個 topic 好悶。
1880. 唔講得，係 top-secret 嚟㗎。
1881. 呢啲係 top-secret，唔好同人講。
1882. 我哋 total 會有一百人參加呢個活動。
1883. 佢哋依家 total 請三個。
1884. 可能係我暫時未遇到一啲可怕回憶吧，touch wood。
1885. 除非 touch wood 咁你發生一件更大嘅事。
1886. 有三首歌好好聽，好 touching。
1887. 聽講個結局好 touching。
1888. 聽講係想我哋學識再 train 其他人。
1889. 都冇人 train 我，係我自己邊做邊學之嘛。
1890. 你哋會有啲咩 training？
1891. 公司都冇任何 training。
1892. 我明白 training 係袋錢落我袋。
1893. 好辛苦，但仲要打 transcript。
1894. 我發現啲 transcript 入面有好多錯字。
1895. translate 咗少少俾佢知啦。
1896. 可唔可以將啲日文字 translate 做中文？
1897. 係我第一次做 translation。
1898. 我都想試下做 translation。
1899. 仲要排期到八月至做完晒成個 treatment。
1900. 已經完成咗成個 treatment。
1901. 佢嘅打扮好 trendy。
1902. 佢哋嘅設計一直好 trendy。
1903. 可能要俾我經歷一啲事，先可以再次 trigger 起我。

1904. 我諗都好難有嘢 trigger 到佢。
1905. 呢個 trip 係我人生中一件重要嘅事。
1906. 出一次 trip 要用好多錢。
1907. 你可以 try 吓呢個方法。
1908. 你唔 try 下又點知自己唔得？
1909. 原來我屋企一件黑色短袖 T-shirt 都冇。
1910. 呢件 T-shirt 同我嗰件差唔多。
1911. 佢唔係第一次俾人 turn down㗎喇。
1912. 去醫院傳福音成日俾人 turn down。
1913. turn out 有二百個家長參加。
1914. 雖然佢都幾努力但係turn out嘅成績都係差。
1915. 有嘢唔識可以去問 tutor。
1916. 邊個 tutor 負責收功課？
1917. 好多學生都冇去上 tutorial。
1918. 我唔介意你上 tutorial 嘅時候食嘢。
1919. 我要去 U gym 上堂。
1920. U gym 裝修過之後好靚。
1921. 因爲要等人，所以蒲 U lib 蒲足六粒鐘。
1922. 之後我哋入 U lib 等佢哋落堂。
1923. 佢 ultimately 可以炒咗你。
1924. ultimately 佢炒我先算。
1925. 老細而家 unavailable 你遲少少再嚟過啦。
1926. 我暫時 unavailable，不如你遲啲再搵我吖？
1927. 電子工程學系係 under 工程學院嘅。
1928. 佢話 under 三百萬像素嘅相機質素唔夠好。
1929. 我唔係好 understand 佢講咩。
1930. 大家 understand，我都無謂講啦。
1931. 有啲嘢你 understood 就算，唔好四圍講。
1932. 有啲嘢 understood 就算。
1933. 遲啲會拍一個 underwear 廣告。
1934. 只係着住 underwear 喺片場四圍走。
1935. 如果做人都有得 undo 就好啦！
1936. 你試下可唔可以 undo 番之前做過嘅嘢？
1937. 有好多嘢都係 unexpected㗎啦！
1938. 今次 unexpected 得嚟好過上次好多。
1939. 一個掂過下學生組織嘅大學生，就係需要呢一點 uniqueness。
1940. 冇咗呢點 uniqueness，我諗佢都唔會請我。
1941. 你學 unix 學咗幾耐？
1942. 我唔識 unix 點算呀？
1943. 你 un 唔 underatand 我頭先講嘅嘢？
1944. 你 un 唔 underatand 都出句聲吖。
1945. 你幾時會 update 個網頁？
1946. 你俾我啲資料係咪最 update㗎？
1947. 而家我冇得 upgrade，唯有怪自己蠢，成日借功課俾人抄。
1948. 部電腦執到好靚，幾年都可以唔駛 upgrade。
1949. 我已經將啲歌 upload 咗上網。
1950. 我會自己 upload 啲課文上網。
1951. 我唔知佢今日好 upset 嘅。
1952. 佢成日都係好 upset 咁。
1953. 我部電腦個 USB 壞咗。
1954. 佢問可唔可以唔用 USB？
1955. 佢 usually 五點半收工。
1956. usually 會咁做，但間中都有例外。
1957. 我哋既 value 唔係淨係要嚟背㗎！
1958. 係咪都反映咗當代人嘅 value？
1959. 等咗好耐 van 仔，差一分鐘就遲到。
1960. 點知上咗架 van 仔一直係咁等客。
1961. 坐喺上面就睇住其他 van 仔係咁走。
1962. 有冇限最多可以用幾多個 variable？
1963. 我將個 variable 名打錯咗。
1964. 點解隻 VCD 冇聲嘅？
1965. 而家啲 VCD 平到冇人有。
1966. 究竟 vegetarian 食唔食雞蛋㗎呢？
1967. 有啲航空公司會提供埋 vegetarian 嘅飛機餐。
1968. 呢個字嘅 verb 係咩呀？
1969. 佢用 verb 嘅時候成日都用錯。
1970. 呢個係第幾個 version？
1971. 你用嗰個 version 唔係最新㗎咩？
1972. 我仲影咗相同 video 留念。
1973. 但開場前都驚餐飽，個 video 好唔穩定。
1974. 佢哋仲喺裡面個排球場打，個 view 勁正囉。
1975. 呢度夠晒高，仲有馬場 view。
1976. 整 virus 係咪犯法㗎？
1977. 呢輪有好多 virus 所以我都係唔上網住。
1978. 呢位係新嚟嘅 visiting professor。
1979. 唔知做 visiting professor 一個月有幾多錢呢？
1980. 開放日會有好多 visitor 嚟參觀。
1981. 你咁樣答個 visitor 會覺得你好冇自信。
1982. 點解我會選修 visual art？
1983. 我都唔知咩係 visual art。
1984. 我一上 VPN 就中咗毒。
1985. 喺屋企去中大啲網頁係咪要用 VPN㗎？
1986. 最慘係張地氈都中埋招，搞到啲 waiter 勁黑面。
1987. 冇理由大學畢業走去做 waiter㗎嘛。
1988. 呢度啲 waitress 好有禮貌。
1989. 佢哋請 waitress 係以貌取人嘅。
1990. 今次唔係搵 walkie-talkie，係搵電話。
1991. 如果行山有 walkie-talkie，可以慳番唔少電話費。
1992. 我部 walkman 自從會考之後都冇用過。
1993. 要有 walkman 先至聽到啲錄音帶㗎嘛。
1994. 你用緊嗰張 wallpaper 係咪自己整㗎？
1995. 係咪任何圖片都可以做 wallpaper㗎？

1996. 你有冇收到 warning？
1997. 佢好驚會俾人 warning。
1998. 嗰日我就會俾人記我兩次名，拎 warning letter㗎喇。
1999. 收到張 warning letter，我呢排做咩呀，咩都好黑呀。
2000. 佢去咗 washroom。
2001. 唔該 washroom 喺邊度？
2002. 佢隻表有 water proof 跌落水都唔驚。
2003. 化全套眼裝要買 water proof。
2004. 我買咗隻新嘅 web cam。
2005. 隻 web cam 係咪一插入部機度就用到？
2006. 但個 web site 都係麻麻地至嗎。
2007. 呢個 web site 好似有問題嗎。
2008. 今個學期一共有十四個 week。
2009. well，喺度祝各位情人節快樂先，雖然已經過咗。
2010. well，或者人真係唔會有完美㗎呢。
2011. 佢做嘢都唔 well organized 嘅。
2012. 你個 nized 嘅。
2013. 有啲人鍾意飲 whiskey，我就鍾意飲紅酒多啲。
2014. whiskey 好貴，但係有時紅酒可以仲貴。
2015. 不過我哋下場唔知會唔會 win 到呢？
2016. 我買咗二千蚊 win。
2017. 用 window explorer 可以睇到晒你部機有啲咩？
2018. 「檔案總管」係咪即是 window explorer？
2019. 我仲係用緊 windows 九八咋！
2020. 我阿爸連 windows 都唔識用。
2021. 你有冇用過 wireless LAN？
2022. 其實喺屋企用 wireless LAN 有冇用呢？
2023. 我都有張 wish list。
2024. 你有咩想要，寫落張 wish list 度啦。
2025. 你 within 一個星期起唔起到貨？
2026. within 一個月做唔做得晒啲嘢？
2027. 我 wonder 我自己適唔適合讀文學。
2028. 我 wonder 佢其實明唔明我講咩？
2029. 我對 word 唔係好熟。
2030. 我喺 word 入面搵到啲有用嘅嘢。
2031. 如果佢個方法唔 work 仲有冇其他方法？
2032. 係咪咁 work 要試過至知。
2033. 你個建議係咪 workable㗎？
2034. 雖然未必 workable，但係好有創意。
2035. 下堂就要做 workshop，唔知難唔難。
2036. 一陣個 workshop 你會唔會去？
2037. 佢做啲嘢都唔知 work 唔 work 嘅。
2038. 你有冇試過 work 唔 work㗎？
2039. 今次考試淨係得 written。
2040. 今次 written 我十拿九穩。
2041. 睇咗張 X 光片。
2042. 照完 X 光，見埋醫生就走得。
2043. 你邊個 year㗎？
2044. 三個 year 都係咁辛苦。
2045. 你交咗 year plan 未？
2046. 個 year plan 寫得幾好，不過唔知做唔做到。
2047. 幾經辛苦喺人群中穿插先到 YMCA。
2048. YMCA 好似有得食自助餐。
2049. 記得 zip 咗份功課先至好交。
2050. 我 zip 咗啲資料然後燒碟俾你。
2051. 你嗰個係電視機定係 LCD 嚟㗎？
2052. 佢睇中咗一個十七吋 LCD 電腦顯示屏。
2053. 我舊年讀咗一個 psycho 課程。
2054. 中五會考有得考 psycho㗎咩？
2055. 佢個 client 好鬼死麻煩。
2056. 好多有錢佬都係佢嘅 client。
2057. 今次考試個 mean 係幾多？
2058. 有一題三十分滿分，但係個 mean 得五分咋！
2059. 要覺得個 DJ 親切先至會每日聽佢嘅節目。
2060. 呢啲身爲聽眾嘅感受，無疑成爲佢做 DJ 時嘅取向。
2061. 佢係呢方面嘅 expert 嚟㗎！
2062. 講到食，你就係 expert。
2063. 你要 realize 自己嘅優點同缺點。
2064. 你 realize 到佢想請咩人，咪將自己表現成嗰種人囉。
2065. eye-contact 好重要㗎，可以令人以爲你好有信心。
2066. 佢講嘢淨係望住張紙，一啲 eye-contact 都冇。
2067. 一部 plasma 電視好似要幾萬蚊㗎！
2068. plasma 咁大部，我屋企邊放得落？
2069. 佢 undergrad 嗰陣已經開始搞生意。
2070. undergrad 畢業一個月可以搵到幾多錢？
2071. 佢咁有 experience，你咪叫佢做囉。
2072. 唔一定要有 experience 先至做得㗎。
2073. 你一個一個 step 跟住做就唔會錯。
2074. 唔知邊個 step 做錯咗。
2075. 我買咗件黑色 V 領衫。
2076. 對住鏡頭豎起 V 字手勢。
2077. 咁樣着先至夠 man！
2078. 佢話你太斯文，唔夠 man。
2079. 居然有人叫我做 uncle。
2080. 有啲 uncle 同我講，我先至知呢件事。
2081. 佢係嗰間 pub 嘅老板。
2082. 得閒會同幾個朋友落 pub 飲嘢。

Appendix B

Code-mixing Utterances in Testing Set of CUMIX

1. 幾時開始可以 add drop？
2. 你唔知有得上網 add drop㗎咩？
3. 又唔係用好耐，cheap cheap 地用住先啦。
4. 件衫 cheap cheap 地，但係好暖喎。
5. 或者放佢入 ignore list。
6. 可能佢放咗我入 ignore list 呢。
7. 用 inkjet printer 印數碼相得唔得㗎？
8. 其實 inkjet printer 印出嚟啲嘢係咪靚啲㗎？
9. 好灰呀，竟然連做咗一個月嘅 junior sales 都可以成日串我。
10. 佢由 junior sales 開始做起，依家已經係經理。
11. 今日終於整負離子，真係有 student discount。
12. 去到先知星期日冇 student discount。
13. 如果有一日，我唔介意任何 test result，我會寫一啲好出位嘅嘢。
14. 無論個 test result 係點，你盡咗力就得啦。
15. 我寫咗個程式防止人哋 view source code。
16. 點樣先至可以唔俾人 view source code？
17. 佢 year one 嘅時候啲頭髮好長。
18. 我今年讀 year one。
19. 你 year two 嗰陣唔係去過交流咩？
20. 我 year two 先至識佢。
21. 冇 relevant experience 都唔緊要，可以邊做邊學。
22. 有 relevant experience 梗係會着數啲啦。
23. 好多 contract terms 我都睇唔明。
24. 啲 contract terms 對我好不利。
25. 唔知打風 training center 開唔開呢？
26. 去完 training center 上堂仲要返工。
27. 我都好耐冇去過 K lunch 喇。
28. K lunch 可以唱幾耐呀？
29. 我唔見咗張信用卡，所以要打去 card centre cut card。
30. 打電話去 card centre cut card 之後，仲有咩要做？
31. 你搬宿可以 call van 仔，唔駛自己搬得咁辛苦。
32. 我夠想 call van 仔啦，不過好貴吖嘛。
33. 邊度有得學畫 3D 嘅電腦動畫呀？
34. 佢部機嘅開機畫面係一幅 3D 嘅公仔。
35. 一部 3G 手機要幾多錢？
36. 聽講 3G 啲電話成日斷線喎。
37. 佢會考有九個 A㗎！
38. 你今次考試又攞咗幾多個 A 呀？
39. 搭 A12 搭咗啱啱一個鐘就到機場。
40. 其實你可以搭 A12 返屋企。
41. 我睇睇吓 abbreviation 又瞓着咗 。
42. 呢啲 abbreviation 你自己作出嚟㗎？
43. 你有冇足夠 ability 去應付㗎？
44. 有時我都會懷疑自己嘅 ability。
45. 你嘅做法好 abnormal。
46. 我不知幾正常，你就 abnormal！
47. absent 超過兩堂就一定肥硬。
48. 遲到超過三十分鐘就當 absent。
49. 通常 abstract 係簡要地將最重要嘅嘢寫出嚟。
50. 寫公函嘅時候要具體啲同埋避免 abstract。
51. 叫得你去面試，已經証明咗你嘅 academic background 合乎佢哋要求。
52. 佢唔係想知你嘅 academic background，而係想睇吓你係個點樣嘅人。
53. 我 accept 之前都已經問過好多人意見。
54. 大家都一致贊成 accept，希望唔會後悔。
55. 我 access 唔到佢部機。
56. 點樣先至可以 access 到宿舍部機？
57. 佢都睇咗唔少 accessories。
58. 我會用多啲 accessories 嚟襯托。
59. 可以喺邊度攞到 accommodation 嘅資料？
60. accommodation 係自己俾定學校俾？
61. 我都冇開到 account。

62. 跟住返咗沙田幫 account 整電腦。
63. 請喺呢度寫銀行名同 account number 吖。
64. 你張提款卡上面咪有 account number 囉。
65. 佢個 accuracy 都唔知點計出嚟嘅。
66. 個 accuracy 咁高會唔會有蠱惑？
67. 其實呢個方法好唔 accurate，不過好過冇咁啦。
68. 我就唔信佢呢個方法會 accurate 過我嗰個。
69. 佢居然唔記得喺acknowledgment入面寫老細個名。
70. 人哋幫過你，咪喺 acknowledgment 度多謝人囉。
71. 佢哋啲 activity 其實唔係咁啱我。
72. 放工之後會唔會有 activity？
73. actually 我唔知發生緊咩事。
74. actually 呢啲都唔係咩秘密。
75. 當然仲有啱啱 add 咗冇耐嗰啲。
76. 其實係我特登唔 add 佢㗎。
77. 我唔係 admin 所以好多嘢都做唔到。
78. 係咪淨係得 admin 可以安裝軟件㗎？
79. 你嗰科嘅 admission grade 係幾多？
80. 今年嘅 admission grade 已經出咗，原來我係最尾嗰幾個。
81. 係咪間間公司嘅寬頻都係 ADSL 嚟㗎？
82. 我屋企個寬頻係咪 ADSL 嚟㗎？
83. 你估人人都 afford 得起住宿舍㗎？
84. 啲居民 afford 唔起買公屋。
85. 佢要 after 十月先至得閒喎。
86. 佢 after 八月就可以返工。
87. 我哋次次約食飯都係 after class㗎啦。
88. 佢諗住 after class 去剪頭髮。
89. 我好 agree 佢嘅講法。
90. 唔一定要佢 agree 嘅。
91. 呢個 album 幾好用喎。
92. 啲相已經放咗上 album，你哋自己去睇啦。
93. 往往都係問題嚟嘅時候先至 alert。
94. 除咗會好好 alert 以上嘅問題。
95. 我考 A-Level 嗰陣都冇乜點溫書。
    A- Level 同會考嘅程度差好遠㗎。
96. 佢啲設計不嬲都咁令人 amazing㗎啦。
97. 佢好似好 amazing 咁喎。
98. 我點知而家係咪用緊 analog？
99. 咩係 analog 呀？
100. 點樣去 analyze 個故仔？
101. 你會點 analyze 呢啲資料？
102. and then 咪繼續開工囉。
103. 我食完飯 and then 去睇戲。
104. 換個 angle 再試過。
105. 如果由另一個 angle 睇呢？
106. 呢個消息其實係未 announce㗎。
107. 佢上星期唔記得 announce。
108. 呢個 announcement 公佈咗好耐㗎喇！
109. 你冇睇到個 announcement 咩？
110. 佢一次過請晒十日 annual leave 去旅行。
111. 你唔知 annual leave 同病假係分開計㗎咩？
112. 其實 anthro 係讀咩㗎啫？
113. 我個組仔都係讀 anthro㗎喎。
114. 裝完個 antivirus 之後部機好似慢咗咁。
115. 你當個 antivirus 係萬能㗎咩？
116. anyway，最後唱到十點幾我哋就鬆人(lu)。
117. anyway，都係費事俾佢知啦，陣間佢睇完又話我，咁就唔好啦。
118. anyway，等我講講呢幾日啲嘢啦。
119. anyway，冇咗呢十個人，真係會玩唔成。
120. 佢個 appendix 有成三十頁咁多。
121. 呢啲放喺 appendix 咪得囉。
122. 封信係話已經收到你個 application，遲啲會叫你去面試。
123. 收 application 嗰陣第一件事就係睇學歷。
124. 但我仲未有時間 apply。
125. 如果你哋有興趣，一於 apply 吓。
126. 你冇睇過佢份 appraisal 咋。
127. 佢份 appraisal 寫得好好。
128. 咁樣我反而仲 appreciate 添。
129. 你會 appreciate 佢啲寓意。
130. 佢用咗好多唔同嘅 approach。
131. 我覺得你個 approach 唔得。
132. 佢唔 approve 我都冇符㗎。
133. 你肯 approve 就得㗎喇。
134. 我哋嘅研究集中喺呢幾個 area。
135. 初初就喺度玩啤牌，後嚟玩玩下個 area 越嚟越大。
136. 本書有好多 argument。
137. 呢個 argument 好好吖。
138. 我 around 九點嘅時候見到佢攞住好多嘢。
139. 今朝 around 八點半有人搵你。
140. 我覺得佢嘅 arrangement 好差囉。
141. 呢啲 arrangement 一啲都唔夠班。
142. 佢十點先 arrive 梗係冇人啦！
143. 你 arrive 之後就打俾我啦！
144. 到咗 art 堂嗰陣我就好俾心機咁做嘢。
145. 佢以前讀 art 嘅。
146. 一個 artist 有成四五個人跟住。
147. 身爲 artist 嘅就更應該以身作則。
148. 佢去親 assembly 都瞌眼瞓。

149. 你哋幾耐先上一次 assembly？
150. 我 assign 咗五個練習俾個學生做。
151. 佢 assign 得你做呢個位，你就應該對自己有信心。
152. 點解啲 assignment 做極都做唔完㗎？
153. 要改好所有 assignment。
154. 老師 assume 我已經學咗呢啲嘢。
155. 佢 assume 你知道晒之前發生嘅嘢。
156. 個個 at first 都對前途充滿憧憬㗎啦！
157. at first 嘅時候根本冇人留意到佢。
158. 佢 at last 都入咗兩球，算係將功補過咁啦。
159. at last 佢都係冇嚟。
160. 都好嘅，at least 自始會少咗個人爭呢塊田吖嘛。
161. 依家唔奢望啲咩喇，at least 特別啲、難忘啲。
162. at the end 個結局都令人滿意。
163. 佢一路講嘢一路食嘢，at the end 咪哽親囉。
164. 呢啲 audience 嘅質素好差。
165. 你要同啲 audience 有多啲交流至得。
166. 佢廿幾歲但係個樣成個 auntie 咁。
167. 佢 auntie 好錫佢，成日帶佢四圍去。
168. 個 author 好好人，仲幫我簽咗個名添。
169. 我唔記得個書名，淨係記得個 author 個名。
170. 冇 bachelor degree 都可以讀碩士㗎咩？
171. 而家通街都係大學生，淨係得 bachelor degree 係唔夠㗎。
172. 你知唔知 background 嗰首歌係邊個唱㗎？
173. 佢嘅 background 我一早就知啦。
174. 你未做過 backstage，有好多嘢都唔知。
175. 佢做 backstage 嗰陣一得閒就睇書。
176. 你 backup 晒啲重要資料先至叫我啦。
177. 久不久就 backup 吓，如果唔係部機死咗咪乜都冇？
178. 佢做 bad guy 直頭入形入格啦。
179. 我唔介意人哋當我係 bad guy。
180. 佢好 bad taste，成日群埋呢班人。
181. 你唔會咁 bad taste 揀呢個款呀？
182. 一年用過百萬買衫去 ball。
183. 佢設計啲衫去 ball 着就差唔多。
184. 我講乜佢都會 ban㗎啦。
185. 人哋一番心機就咁俾你 ban 晒。
186. 差唔多四點嗰陣，我哋就散 band。
187. 食完飯之後就散 band。
188. 你唔知佢整 banner 好叻㗎咩？
189. 張 banner 有冇限大細？
190. 次次去完 barbecue 我都唔舒服。
191. 佢搞咗個 barbecue，問你去唔去。
192. 我好鍾意 BBQ㗎！
193. 佢去親 BBQ 都會帶埋呢隻燒烤叉去㗎。
194. 如果 before 星期五做好就最好啦。
195. 佢想我 before 下個禮拜起貨。
196. 如果 beginning 嗰陣知道佢係咁我就唔同佢一組啦。
197. 佢 beginning 嗰陣都唔係咁衰㗎。
198. 次次都 below mean，咁唔係辦法。
199. 佢次次都好高分，從來未試過 below mean。
200. between 佢兩個之間，好難揀啫。
201. between 兄弟姊妹間又點會有隔夜仇呢？
202. 你唔知佢好 big mouth㗎咩？
203. 佢咁 big mouth 你仲講咁多嘢佢知？
204. 我點知個 BIOS 有問題啫？
205. 我知道係個 BIOS 有事就晨早自己搞掂咗啦。
206. 佢 birthday 嗰日收到好多禮物。
207. 下星期一係佢 birthday 喎。
208. 你識唔識玩 black magic？
209. 好多時一大班人一齊玩 black magic。
210. 如果佢係想 blame 你就唔會咁講啦。
211. 我要 blame 佢嘅話，佢一早已經唔係度啦。
212. 你部電腦用唔用到 bluetooth㗎？
213. 你估用 bluetooth 快啲定用紅外線快啲呢？
214. 我本 BNO 過咗期好耐啦。
215. 你唔知特區護照去好多地方都唔駛簽証，但係用 BNO 就要咩？
216. 出面塊 board 係你整㗎？
217. 塊 board 頭先俾風吹冧咗。
218. 你屋企有冇用 boardband？
219. 邊間公司嘅 boardband 口碑最好？
220. 原來佢一直以嚟都冇做 body check。
221. 上次做 body check 嘅時候醫生話我好健康㗎喎。
222. 你唔駛旨意今年有 bonus！
223. 我覺得今年有 bonus 嘅機會好渺茫。
224. 晏晝食完飯就打咗電話去 book 小巴。
225. 佢話 book 位要兩個禮拜前先至有位。
226. 你知唔知呢本書 book store 賣幾多錢？
227. 我搵均成間 book store 都搵唔到喎。
228. 你今日有三個 booking。
229. 上晝有兩個 booking，下晝就得一個。
230. 尤其有幾個 brand 一直俾我好高嘅價錢。
231. 好多香港品牌都好受歡迎，不過可以 brand 香港地位嘅就冇乜。
232. 你想個 break 放耐啲定早啲落堂？
233. 直落三個鐘一個 break 都冇。
234. 佢成日都嚟呢度食 breakfast㗎。

235. 一個 breakfast 要百幾蚊，唔好去搶！
236. 你 brief 啲得㗎喇，其他我自己睇都得。
237. 你唔駛講得咁複雜，可以 brief 啲。
238. 一陣個 briefing 大概幾長？
239. 我要上堂，唔去 briefing 喇。
240. 我預咗三千蚊 budget，可以買邊啲款？
241. 都冇 budget，邊度搵錢出嚟喎？
242. 學校有 buffet 食㗎咩？
243. 夜晚就去咗尖沙咀食 buffet。
244. 我好辛苦先至 build up 到公司嘅信譽。
245. 有啲嘢要慢慢 build up，急唔嚟嘅。
246. 佢咁 busy 你咪早啲約佢囉。
247. 佢成日都 busy，都唔知係咪嘅。
248. 佢好 buy 我嘅建議。
249. 雖然扮嘢但我又 buy 喎。
250. 雖然我對 buying 唔係好熟，但係好有興趣。
251. 你究竟知唔知 buying 即係做啲咩？
252. by the way，你知唔知佢其實唔鍾意你㗎？
253. by the way，我唔介意你食煙。
254. 我一路講緊都係 byte 喎。
255. 你知唔知每個 byte 代表咩先？
256. 我個 cable modem 唔知係咪壞咗。
257. 其實用 cable modem 係咪會快啲？
258. 周圍 call 吓人睇吓有邊個得閒。
259. 好彩最尾都有人 call 番我問我喺邊。
260. 初初諗住夜晚去睇煙花，點知好夜都冇人 call 我。
261. 踢完波俾大佬 call 咗我出去飲嘢。
262. 我講緊電話，一陣先 call back 你。
263. 唔該你晏啲 call back 我吖。
264. 我笑到死死下，好難先 calm down 到。
265. 我哋 calm down 咗之後就去食嘢。
266. 呢個 camp 嘅目的係咩？
267. 可惜你去咗 camp，我其實可以同你出街㗎。
268. 唔得閒咪 cancel 咗佢囉。
269. 做乜冇人通知我個會 cancel 咗㗎？
270. 我最憎食 canteen 啲飯。
271. 城大係咪得一個 canteen？
272. 你頂 cap 帽好靚喎，係邊度買㗎？
273. 佢屋企有好多帽，淨係 cap 帽都有幾十頂。
274. 佢以前係校隊，仲係 captain 嚟㗎。
275. 好難想像佢居然係 captain 囉。
276. 呢個 car park 唔係架架車都停得。
277. 除咗呢度之外，仲有冇第二個 car park？
278. 你買個 card reader 咪得囉。
279. 六合一嘅 card reader 會唔會好貴？
280. 佢會好 care 每個細節。
281. 我都唔 care 份工。
282. 你有冇聽過 career talk？
283. 佢成日走堂去聽 career talk。
284. 其實呢啲 case 唔係咁常見。
285. 佢係神探嚟㗎，破咗好多大 case。
286. 我記得嗰日有幾十個女仔 casting。
287. 否則連 casting 嘅機會都冇。
288. 你可以着得 casual 啲。
289. 佢問我點解今日着得咁 casual。
290. 唔係個個 catholic 都係好人。
291. 佢話自己係 catholic。
292. 燒一隻 CD 要幾耐？
293. 上堂照書讀，咁我買隻 CD 聽好過喇。
294. 我個 CD writer 好慢，得四速。
295. 點解隻碟喺 CD writer 會讀唔到嘅？
296. 年年都有同佢 celebrate，但係今年就冇。
297. 其實有冇人同我 celebrate 又有咩所謂？
298. 你唔知佢哋幫你搞咗個 celebration 咩？
299. 個 celebration 幾點開始？
300. 個 centre 地方唔係好大，不過好骨子。
301. 呢個 centre 唔係俾你喺度玩㗎。
302. 你張 cert 係人見到都會請啦。
303. 你見工做乜唔帶埋啲 cert 去？
304. 你有 certifying letter 就會快好多。
305. 你叫老細幫你寫 certifying letter 咪得囉。
306. 佢個 chairman 講明今年冇人工加。
307. 唔係人人都適合做 chairman。
308. 你要先預計人哋會 challenge 邊啲地方。
309. 如果你咁 challenge 我，我繼續自己搵囉。
310. 咁 challenging 嘅嘢留番俾你做。
311. 雖然幾難，但係好 challenging。
312. 佢哋已經攞咗兩次 champion，今次贏埋就三連冠。
313. 你呢個 champion 係實至名歸嘅。
314. 我去到 chapel 但係冇人喎。
315. chapel 前面有好多人，唔知喺度做咩呢？
316. 今日講到邊個 chapter？
317. 我溫咗成日書都淨係睇咗一個 chapter 咋。
318. 我而家就 charge 你「行為不檢」。
319. 佢一落 charge 我哋就麻煩。
320. 佢有時真係好 charming㗎。
321. 化咗妝即刻好 charming。
322. 話我有其他事要做，唔同佢哋 chat。
323. 只係 chat 咗一晚，交換咗電話啫。
324. 我間公司真係幾 cheap，勁怕蝕底。
325. 佢好 cheap，扮客人嗰陣係咁刁難人。
326. 我去嘅髮型屋都好 cheap，完全剪唔到我想要嘅髮型。
327. 睇嚟我今日都唔太應該嚟，攪到個場 cheap 晒。

328. 食完晏就 check 吓有冇得面試。
329. 返屋企 check 番資料，先知道原來係我今日整頭髮嗰間。
330. 記得 check spelling 之後先至好印出嚟。
331. 記事簿係冇得 check spelling㗎！
332. 過咗第二個 checkpoint 之後就唔見咗佢。
333. 一共有幾多個 checkpoint？
334. 我都係想休息吓啫，無謂再放咁多 chemical 上塊面度啦。
335. 嗰啲果汁入面有好多 chemical。
336. 我將電子工程放喺第一個 choice㗎。
337. 我仲有其他 choice 咩？
338. 你仲有冇唱 choir？
339. 佢以前唱 choir 嘅時候識咗好多人。
340. 我想 chop 走最頭嗰個字母。
341. 一隻碟放唔晒咪 chop 開佢囉。
342. 呢啲嘢係佢自己 claim 嘅，真唔真冇人知。
343. 我夠可以 claim 自己係行政總裁咯。
344. 都唔知個 class participation 係點計分嘅。
345. 如果唔係夠 class participation，我都好難合格。
346. 今季品牌嘅設計好 classic。
347. 佢哋會玩啲大家耳熟能詳嘅 classic 歌。
348. 佢攞晒啲嘢出嚟又唔 clean up 番。
349. 我一陣會 clean up 番啲嘢㗎喇。
350. 我 click 完個掣之後等咗好耐都冇反應。
351. 你 click 入「我的電腦」就可以見到剩番幾多位。
352. 佢依家喺政府 clinic 做藥劑師。
353. 由呢度行去 clinic 大約要十分鐘。
354. 細個嘅時候同爸爸 close 啲。
355. 若果連諗乜都要分享的話我驚會太 close。
356. 呢個 close up 影到佢好靚。
357. 我冇化妝，唔好影 close up 呀！
358. 我嗰杯 cocktail 啲味好怪。
359. 你識唔識溝 cocktail 呀？
360. 佢 collect 咁多資料唔知有咩用。
361. 你要 collect 幾多資料至夠？
362. 我同佢都唔係同一個 college 嘅。
363. 有啲係 college 搞嘅，有啲係大學搞嘅。
364. 呢度賣嘅多數係 colourful 嘅少女款式。
365. 見到一支支咁 colourful 嘅顏料我就想買。
366. 我係唔會 comment 人哋嘅意見嘅。
367. 咁你有啲咩 comment 吖？
368. 你哋 commit 好未？
369. 我唔想再嘥時間，你 commit 好話我知啦。
370. 呢啲咁 common 嘅嘢自己睇啦。
371. 我唔覺得咁都唔 common 喎。
372. 你想去 common room 食定係喺房食？
373. 我一陣會去 common room。
374. 題題都好似 common sense 咁，就算冇讀過都可以答。
375. 佢話呢啲係 common sense，係基本求生技能，一定要學。
376. 佢事後佢仲同主管 complain，話我串佢。
377. 我要 complain，無良顧主虐待我。
378. 冇化妝同帶 con (contact lens)，成個傻婆咁。
379. 夜晚去咗配 con (contact lens)，因爲開始做嘢，我決定要以靚樣示人。
380. 我覺得佢個 concept 唔係好啱。
381. 呢個 concept 唔錯喎。
382. 佢好 concern 我哋嘅表現。
383. 我絕對明白你嘅 concern。
384. 我從來冇聽過 concert。
385. 個 concert 幾時開始賣飛呀？
386. 你個 conclusion 做得唔錯喎。
387. 通常嚟講 conclusion 係最重要嘅。
388. 呢個 condition 好特別喎。
389. 喺唔同嘅 condition 梗有唔同嘅處理手法。
390. 表現得流暢同好有 confidence 咁囉。
391. 你自己有冇 confidence 先？
392. 凡係 confidential 嘅文件都唔可以攞走。
393. 呢份嘢係 confidential 唔好俾其他人睇。
394. 你用邊個 config㗎？
395. 我部機同你部機個 config 一樣㗎咩？
396. 好多時外國嘅紅燈區都會 confine 喺某啲地區。
397. 如果唔 confine，會好難管理。
398. 跟住落咗中環俾錢 confirm 機票同酒店。
399. 你話有乜理由連茶會時間都 confirm 咗，之後至話唔成團吖？
400. 你無非都係想識人，建立 connection 啫。
401. 要搞呢啲活動，少啲 connection 都唔得。
402. 希望大家可以 consider 一下其他人嘅感受。
403. 當你應承嘅時候就會 consider 好多嘢。
404. 佢講啲嘢次次都唔同，好唔 consistent。
405. 大家嘅意見都好 consistent。
406. 呢個 constraint 真係幾大。
407. 呢個可能唔係影響我最大嘅一個 constraint。
408. 啱啱畢業出嚟做嘢都可以做 consultant？
409. 你唔知保險經紀都係叫 consultant㗎咩？
410. 我晏啲再 contact 你吖。

411. 我好耐冇 contact 佢，唔係好清楚佢近況。
412. 我淨係睇咗個 content，其他未睇喎。
413. 本本書個 content 都差唔多㗎啦。
414. 脫離咗學校嘅 context 感覺自由咗好多。
415. 唔好淨係聽顧客說話嘅內容，而係要聽說話嘅 context。
416. 你張 contract 幾時完呀？
417. 原來張 contract 上面有寫，係你自己冇留意啫。
418. 其實個 contrast 大同細有咩分別？
419. 我覺得個 contrast 冇問題，不過佢唔同意。
420. 講真，我唔覺佢有 contribution 喎。
421. 呢啲都算係 contribution，咁我嗰啲叫咩？
422. 我 control 唔到佢幾時笑幾時喊。
423. 拍戲最難 control 嘅就係小朋友同動物。
424. 呢度喺正地鐵站上面，夠晒 convenient。
425. 咁 convenient 你又唔去住？
426. 有啲人講嘢好 convincing，好似催眠咁。
427. 講 convincing 我實唔夠佢嚟，所以唯有搞啲花巧嘢。
428. 啲 cookies 其實有冇用㗎？
429. 我需唔需要定期刪除啲 cookies 呀？
430. 佢仍舊保持一貫 cool 到痺嘅形象。
431. 我都唔敢同人傾偈，唯有坐埋一邊扮 cool。
432. 唔該幫我將呢幾幅圖 copy 落個電話度吖。
433. 我幫我嗰班某啲人 copy 咗啲相。
434. 要做嘅嘢就係 copy and paste。
435. copy and paste 呢啲咁簡單嘅嘢駛乜搵大學生做呀？
436. 你講錯嘢佢會 correct 你。
437. 有時唔係樣樣嘢都要 correct 嘅。
438. 我都唔知原來個 correlation 咁大㗎。
439. correlation 會唔會隨時間改變呢？
440. 呢個 cost 已經係最低㗎喇。
441. 我哋個 cost 一向都比人哋高，但係我哋勝在質素有保證。
442. 你今年會唔會去 count down？
443. count down 嗰陣我唔見咗個銀包。
444. 爲咗嗰兩張半價 coupon，諗都唔諗就去咗西九龍。
445. 有新貨我就可以用埋啲 coupon。
446. 呢個 course 冇乜人讀。
447. 有邊啲 course 又少功課又唔駛考試㗎？
448. 點解啲 CPU 成日都唔減價嘅？
449. 成部電腦最貴就係粒 CPU。
450. 佢好鍾意食 cream，但係又怕肥喎。
451. 咁多 cream 會唔會好滯㗎？
452. creative 呢樣嘢比較難學。
453. 佢諗啲嘢一啲都唔 creative，全部都係抄番嚟嘅。
454. 一年最多讀四十二個 credit。
455. 中六嘅尖子要讀一百二十三個 credit。
456. 有啲嘢有時好難去 criticize。
457. 佢淨係識得 criticize。
458. 其實 crossover 係幾時開始流行㗎呢？
459. 依家好多「唔喇更」嘅嘢都可以 crossover 埋一齊。
460. 每個地區都有自己獨特嘅 culture。
461. 我唔係好適應當地嘅 culture。
462. 麻煩你幫我攞本 curriculum 吖。
463. 出年嘅 curriculum 出咗未？
464. 最後又係佢 cut 我線囉。
465. 俾佢 cut 線我其實都嬲，但我真係頂唔順佢啲態度。
466. 個學生好 cute，好易教。
467. 支酒板好 cute㗎！
468. 佢咁 cutie 你咪讚吓佢囉。
469. 佢唔想人哋因爲佢 cutie 先至請佢。
470. 件衫嘅 cutting 好靚，不過唔知我着唔着到。
471. 你嘅設計唔錯，係 cutting 嗰度差少少。
472. 細個嗰陣成日同 daddy 玩騎膊馬。
473. daddy 好錫我，一有時間就帶我去玩。
474. 佢十幾歲就開始做 dancer，到依家已經跳咗十幾年舞。
475. 你咁鍾意跳舞，第時去做 dancer 囉。
476. 呢個 database 有晒你要嘅嘢。
477. 我而家學緊點用 database。
478. 今日係辛苦咗六日之後嘅 day off，我真係覺得好攰。
479. 但係我咁多 day off，冇理由蝕俾佢㗎。
480. 部 DC 跌咗落地，好彩冇事。
481. 佢買 DC 淨係揀款，都唔理功能嘅。
482. 我哋喺度嘈，之後就 dead-air 囉。
483. 但 dead-air 嗰陣我其實真係好嬲。
484. 我唔係 deal with 呢啲嘢㗎！
485. 最怕 deal with 啲講極都唔明嘅人。
486. 佢下一個目標係做 dean。
487. 我哋個 dean 好好人。
488. 點樣先至有得攞 dean list？
489. 我嘅 dean list 夢好快就會粉碎。
490. 佢叫我幫佢 debug，我鬼得閒理佢。
491. 我用咗成日但係都未 debug 到。
492. 我 decide 到就唔駛問你啦。
493. 佢幫我 decide 晒，咁我仲可以做咩？
494. 佢最叻就係玩鋤大 Dee。
495. 鋤 Dee 係我嘅強項。
496. 點樣 define 個問題？
497. 你都冇 define 清楚。

498. 佢個 degree 係喺外國讀㗎！
499. 一年咁多大學生畢業，大把人有 degree 啦。
500. 可唔可以 delay 呀？
501. 佢起機嗰陣仲 delay 咗一個鐘。
502. 我好想 delete 呢度嘅某啲舊文章。
503. 我唔小心 delete 咗啲嘢有冇得補救？
504. 應承咗人但係 deliver 唔到，咪失信於人囉。
505. 佢一定要我今日 deliver 到。
506. 一陣有得睇 demo？
507. 我想睇最新嗰隻相機嘅 demo。
508. 我哋 department 得五個人。
509. 成個 department 啲人一齊放假。
510. 佢 design 啲嘢水準好唔穩定。
511. 呢件衫係我自己 design㗎！
512. 你個 desktop 上面有好多嘢喎。
513. 你做乜放咁多嘢喺 desktop 上面嘅。
514. 講得咁 detail 有啲人可能會唔高興。
515. 正式上堂嘅時候會教得 detail 啲。
516. 興趣係可以慢慢 develop 嘅。
517. 一間公司嘅信譽，係要用時間嚟 develop 嘅。
518. 佢叫我用 DHL 寄份文件俾佢。
519. 領事館會用 DHL 送番本護照俾你。
520. 佢用 dial-up 先至連到返嚟喎。
521. 你唔覺用 dial-up 好慢㗎咩？
522. 點解我可以打到呢篇 diary？
523. 估唔到本 diary 都未到一周年，就已經要斷纜，實在太遺憾。
524. 呢部 dig cam 同我嗰部有咩唔同？
525. 佢成日拎住部 dig cam 都唔知有乜好影。
526. 佢話我唔夠 diligent，但我自問已經盡咗力。
527. 我已經好 diligent，唔通真係天資所限？
528. 最後飲茶、打機、食甜品同 dinner，全日不停食，肥死。
529. 我哋啱啱食完 dinner 佢就趕到。
530. 竟然識埋啲咁 dirty 嘅人。
531. 不過佢好 dirty 呀，襯我出咗埠就出去四圍玩。
532. 今次我哋會 focus 喺邊方面？
533. 你影啲相張張個 focus 都唔啱喎。
534. 我好 disaoppointed 呀，佢都對得我住啦。
535. 佢見完老細之後好 disappointed。
536. 我部 discman 買嘅時候要成二千幾蚊㗎！
537. 而家買 discman 應該平過以前好多。
538. 我知呢度有好多 discount 同更好嘅牌子。
539. 我簽咭冇 discount 都有得攞分吖嘛。
540. 我 discover 唔到呢個錯誤，係我唔夠小心。
541. 俾佢 discover 到又點？
542. 個電話嘅七色 display 夠晒吸引。
543. 得番件 display，仲要爛嘅，梗係唔要啦。
544. 我 distinguish 到就唔駛搵你幫手啦。
545. 呢個系統係用嚟 distinguish 佢哋嘅分別。
546. 依家咁興 DIY，我咪食住個勢囉。
547. 市面上咁多 DIY 書，但係咪本本睇完都可以即刻學識呢？
548. 其實上網係咪要用 DNS㗎？
549. 我唔知咩係 DNS。
550. 我最怕就係做 documentation㗎喇。
551. 佢淨係負責寫程式，documentation 啲嘢唔關佢事。
552. 你唔覺得呢啲嘢好無聊咩？don't waste my time。
553. 我啲時間好寶貴㗎，don't waste my time。
554. 你出門口之前再 double check 多次啦。
555. 我已經檢查過一次，但係都係 double check 多次穩陣啲。
556. 你 double click 佢就會自動爆開。
557. 喺呢度 double click 就得㗎喇。
558. 我細細個嘅志願就係做工程師，所以決定讀 double E。
559. 讀 double E 搵工係咪容易啲？
560. 我永遠都唔會 down、唔會輸。
561. 我嘅心情都好 down。
562. 邊度有得 download 嗰封信？
563. 佢叫我上網 download 嗰幅圖喎。
564. 你俾份 draft 我睇住先啦。
565. 我份正本唔見咗，好彩有份 draft 啫。
566. 你㩒實個掣就可以 drag 到㗎喇。
567. 佢唔俾我 drag 個快勞喎。
568. 返到學校，我哋就喺度練 drama。
569. drama 比賽，我哋攞咗個最佳演員獎。
570. 我嘅 dream house 唔駛好大，但係一定要望到海。
571. 希望買間 dream house 俾媽咪。
572. 五十蚊包兩杯 drink，花生同果盤價錢另計。
573. 一杯不含酒精嘅 drink 大約五十至六十蚊。
574. 我只會講一次，你記唔到就 drop 低佢啦。
575. 人哋講嘢你又唔 drop 低。
576. 佢上親堂都 drop notes，不停咁抄。
577. 我 drop notes 係唔想咁易瞓着啫。
578. 張 duty list 冇我個名喎。
579. 我睇過個 duty list，做嘅嘢同舊年一樣。
580. 我打算 easter 嘅時候去澳洲。

581. 其實 easter 得幾日假，去唔到好遠嘅地方。
582. 我可唔可以自己設計 e-card？
583. 一張 e-card 要收十蚊，唔好去搶？
584. 你咁憎計數就唔好讀 Econ 喇。
585. 我對 Econ 嘅認識唔多，但係好有興趣。
586. 其實 economics 喺好多方面都可以應用到。
587. economics 係好複雜嘅嘢。
588. 真係咁 efficient 先算啦！
589. 真係好 efficient，快到你唔信。
590. 我揀晒嗰啲「易碌」嘅 elective。
591. 同 elective 嘅人一齊上堂咪好着數？
592. 我個 email 成日收到好多廣告。
593. 我仲問佢哋拎埋 email 添，都話我「潮」㗎喇。
594. 佢 emotion 有啲問題，成日發脾氣。
595. 我唔覺得自己 emotion 有問題喎。
596. 佢好 encourage 我去做自己想做嘅嘢。
597. 我唔會 encourage 你去接受任何嘢。
598. 見到呢個鎖即係表示已經做咗 encryption。
599. 有冇啲書係教 encryption 嘅原理㗎？
600. 個 ending 同預料嘅一樣，一啲驚喜都冇。
601. 你估套套劇都會拍兩個 ending㗎咩？
602. 架車個 engine 壞咗，有排整喎。
603. 個 engine 壞咗，不如換過個啦。
604. 我想搵份 engineer 工。
605. 香港好難搵到識呢啲嘢嘅 engineer。
606. 好好 enjoy 你眼前嘅嘢。
607. 講真我真係 enjoy。
608. 你有 enquiry 可以直接搵我。
609. 你亦都可以將你嘅 enquiry 傳真俾我哋。
610. 我唔可以 ensure 自己可以完成賽事。
611. 你真係 ensure 先至好講。
612. 我唔慣 entertain 人。
613. 你去 entertain 其他人得㗎喇，我自己四圍行吓。
614. 我諗你要提升一下自己嘅 EQ。
615. 唔知係我好耐性定係佢 EQ 低。
616. 考試要寫 essay 嘅話我通常會做唔晒。
617. 我睇咗成個鐘都未睇完一篇 essay。
618. 呢個數你係點樣 estimate㗎？
619. 佢 estimate 啲嘢都好準。
620. 和記嘅 ethernet 係咪穩定啲？
621. 用 ethernet 係咪會快啲㗎？
622. even 你唔返工都冇人知。
623. even 同佢分析咗好多，但最後都係唔明白。
624. 如果 evening 嘅時候坐喺度就正咯。
625. 我 evening 唔得閒，不如再約晏啲吖？
626. 我個同事嘅男友原來係我個 ex。
627. 如果你見到以前個 ex，會同佢講咩？
628. 我哋嘅興趣 exactly 一樣，都係咁好動。
629. 佢講嗰啲嘢 exactly 同我之前聽過嘅一樣。
630. 真係好似 exam 咁。
631. 我有五個 exam，仲要完全未讀過。
632. 雖然上次 exam 我都係咁講，但係今次唔同。
633. 又話今日有 exam 要早啲返，點知返到嚟一個人都冇。
634. 我都未知邊個係我 examiner。
635. 好多人都做過會考嘅 examiner 啦。
636. 你冇學過點用 excel㗎咩？
637. 佢用 excel 用到好熟。
638. 佢 exchange 嗰陣識咗個泰國女仔。
639. 我好想去 exchange 呀，邊度可以搵到資料呀？
640. 我隔離房住咗個德國嚟嘅 exchange student。
641. 多啲同 exchange student 傾偈咪可以練好啲英文囉。
642. 真係咁 exciting 嘅話我都要去試吓。
643. 喺電視睇就好 exciting，不過玩落就唔覺。
644. 佢 exclude 咗我嘅意見。
645. 你其實 exclude 咗好多人。
646. 呢啲係你嘅 excuse 啫。
647. 不過都好嘅，因爲我哋有個 excuse 唔上堂。
648. 佢淨係識得叫人做 exercise，自己就乜都唔做坐喺度等落堂。
649. 呢個 exercise 好似難過之前嗰啲咁。
650. 我冇 expect 過咁快。
651. 今年，我都唔敢 expect 啲乜。
652. 佢叫我 explore 一下先，唔明先至搵佢。
653. 你要 explore 一下先至可以知道有咩問題。
654. 你嘅缺點係唔識得 express 自己。
655. 文學教一個作家 express 一啲嘢。
656. 我想準時畢業，唔想 extend。
657. 佢已經讀咗五年，唔可以再 extend。
658. 佢沙士嗰陣買咗好多 face mask。
659. 依家去醫院係咪要帶 face mask？
660. 唔得閒溫書，因爲我要做 facial。
661. 星期六我一早起身就去咗做 facial，啲皮膚又愈來愈差。
662. 我個 faculty 成日搞好多講座。
663. 唔知 faculty 今年請邊個做嘉賓呢？
664. 啲音樂慢慢 fade out，之後主角就出場。
665. 段片 fade out 之後就開燈。
666. 佢對 family 好重視。

667. 我哋個 family 又多咗幾個人。
668. 佢係一個好顧家嘅 family man。
669. 喺戲入面佢係一個 family man，不過現實中就剛好相反。
670. 三十幾個 fans 一早去到機場等接機。
671. 竟然遇到我一位 fans。
672. 嗰日我同我老豆 farewell。
673. 突然 farewell 所以帶咗飯都唔食(lu)。
674. 啲同事同我搞咗個 farewell party。
675. 呢個 farewell party 係爲今年走嘅同事而設。
676. 份嘢 fax 咗未？
677. 點樣 fax 嘢去外國？
678. 你嘅 feedback 好慢喎。
679. 你仲等緊佢嘅 feedback。
680. 個 feel 其實都幾正，好鬼死有成功感㗎。
681. 嗰種 feel 真係好鬼死正斗。
682. 好有嗰種心寒嘅 feel 囉。
683. 今日唔知點解有種好怪嘅 feel，真係好怪。
684. 個老細又幫唔到佢 fight 到啲乜嘢喎。
685. 我明知自己實唔夠人 fight。
686. 早兩日喺巴士執到個 file。
687. 咁啱 file 裡面又有佢聯絡方法。
688. 做咗 file compression 之後先至好交俾我。
689. 你唔做 file compression，一隻碟又點會放得晒啲嘢？
690. 嗰兩位出 film 嘅好同事。
691. 數碼相機係唔駛用 film㗎。
692. 我買咗個新嘅 filter，用嚟過濾啲水。
693. 佢 filter 過之後先至可以掉啲垃圾。
694. 佢 final year 嗰年先至住宿。
695. 我都 final year 喇，唔會搞咁多活動囉。
696. 我唔係好明佢個 final year project 做咩。
697. 其實 final year project 又唔一定要整新嘢嘅。
698. 你即刻就會覺得啲皮膚 firm 咗。
699. 搽咗之後皮膚滑咗但唔算 firm。
700. 佢 first hon 畢業之後就入咗政府做政務官。
701. 一年得幾個人 first hon㗎咋。
702. 你咁 fit，我怕我跟唔上喎。
703. 唔操 fit 啲又點應付繁重嘅工作呢？
704. 你張相個 flame 係咪自己整㗎？
705. 唔同主題就用唔同嘅 flame。
706. 你部相機用 flash memory㗎？
707. 我張 flash memory 燒咗。
708. 今次份功課要用 floppy 交。
709. 你可唔可以借隻 floppy 俾我？
710. 每個 folder 都可以用唔同嘅公仔去代表。
711. 個個 folder 嘅名都要唔同。
712. 佢要求咁高，我唔想再 follow 佢。
713. 俾你 follow 到，咪棧跟得辛苦。
714. 呢本書係 for 初學者睇嘅。
715. 今日係 for 文、理同社會科學學院嘅學生。
716. 我會盡最大努力去爭個冠軍返嚟，for sure。
717. 我一定有見過佢，for sure。
718. 佢肯 forgive 我，要我請食飯又點話？
719. 我一早已經 forgive 咗佢，係佢自己唔知啫。
720. 張 form 幾時要交？
721. 上到去佢已經攞咗飛仔填緊 form。
722. 我 form five 未考完會考已經返緊工。
723. 你只係 form five 畢業，點解可以爭贏份工？
724. 我 form one 嗰陣已經有一米七。
725. 佢讀 form one 嗰陣我只係讀緊幼稚園。
726. 我 form three 之後就冇讀過中史。
727. 你學校 form three 唔係要讀晒文科同理科啲嘢㗎咩？
728. 你有冇論文嘅 format 呀？
729. 你隻碟 format 咗未㗎？
730. 咁多 formula 點記呀？
731. 唔知呢條 formula 係咪所有情況都啱用呢？
732. 你 forward 封電郵俾我啦！
733. 佢 forward 俾我嗰封電郵都唔係呢封。
734. 乜都可以講，好 free。
735. 我哋就 free 好多。
736. 覺得班 friend 唔明自己。
737. 仲有一樣嘢，我係想同我嗰啲知情嘅 friend 講。
738. 起碼我冇一個 friend 見過我喊呀。
739. 好彩上次老細個 friend 證明我哋有鎖門。
740. 你嘅 frustration 咁大我都唔知點安慰你好。
741. 咁嘅話佢嘅 frustration 會好大。
742. 我想用 FTP 駛唔駛辦咩手續？
743. 啲相我 FTP 俾你啦。
744. 我實 fulfil 唔到佢個要求啫。
745. 你連最基本五科合格都 fulfil 唔到？
746. full 咗咪報過第二班囉。
747. 個講座唔駛半個鐘已經 full 晒。
748. 就算 full team 佢哋都唔係我哋對手。
749. full team 嘅話又點止贏咁少？
750. 我之前曾經 full time 做過一年嘢。
751. 讀 full time 返唔到工，你諗清楚至好喎。
752. 佢連部相機有乜 function 都未知就俾錢。
753. 首先學識開機同熄機，其他 function 可

以慢慢學。

754. 單係計 functionality 嘅話就梗係買呢部啦。
755. 如果唔理 functionality，就可以考慮呢部細機。
756. 不過佢好 funny，仲肯同我哋玩埋一份。
757. 佢真係好 funny。
758. 用盡各種 fusion 方式炮製青口。
759. 而最特別嘅就要數師傅自創嘅法越 fusion 菜。
760. 我個 FYP 都仲未開始。
761. 係咪個個做 FYP 都係一個人做㗎？
762. ga 唔 gather 到人就要睇你號召力喇。
763. 我人緣咁差，ga 唔 gather 到人都成問題。
764. 呢隻嘢原來都係格鬥 game。
765. 如果電腦有類似嘅 game，我都會試下玩。
766. 最後玩咗隻好似孖寶兄弟嘅 game。
767. 順便睇下有乜新 game。
768. 第一段婚姻多數會 game over。
769. 本雜誌做唔夠三個月就 game over。
770. 今次 gathering 有成二十人出席。
771. 咁下次 gathering 係幾時？
772. 問埋啲咁 general 嘅問題，都唔知佢點揀人。
773. 有時佢問啲 general 嘢，只係想睇吓你嘅說話技巧。
774. 我一個 general education 都未讀喎。
775. general education 嘅目的係想啲學生博學啲。
776. geographically 其實唔係好遠，只係轉車用多咗時間啫。
777. 雖然話係唔同區，但其實 geographically 好近。
778. 我 get 到嘅嘢真係幾有意思。
779. 我 get 到佢想講咩。
780. 唔知點解行商場都可以 get lost 嘅。
781. 明明睇住地圖但最後都係 get lost。
782. 佢係唔會 give up 嘅。
783. 我係唔會咁快 give up 嘅。
784. 你有你 go on，佢有佢講電話。
785. 我聽緊，你 go on 得㗎喇。
786. 你好快咁 go through 一次啦！
787. 因爲太多字所以唔想逐個咁 go through。
788. 佢淨係同我講 good，冇講其他嘢。
789. 你都去咁就 good 喇。
790. 佢同我講咗句 good show 就走咗。
791. 我都話今日一定 good show㗎啦！
792. 聽講佢 GPA 有四點零，不過我好懷疑。
793. 我 GPA 咁低實好難搵工啦。
794. 爲咗準備今日嘅 grad din，我特登等到今日先去剪髮。
795. 頒完獎，個 grad din 亦正式完結。
796. 佢啲 grammar 錯得好離譜。
797. 唔該你返去睇吓啲 grammar 書啦。
798. 呢度咁 grand 襯晒你啦。
799. 你件衫咁 grand，同今日嘅活動唔係好夾喎。
800. 你有冇申請 grant loan？
801. 食完就落去交埋份 grant loan 申請表。
802. 我見你啲 graphic 畫得好靚。
803. 佢都冇打算整過個 graphic。
804. 黑同白中間總會有 grey area。
805. 好多嘢都係 grey area，做唔做你自己決定。
806. 點解我唔練自己嗰 group 呢？
807. 不過都爲我哋個 group 增光，總算有個交代。
808. 做 group leader 要照顧吓啲新人。
809. 做 group leader 你估易㗎？
810. 做 group project 如果個個組員都肯做嘢咪好囉。
811. 個 group project 啲分點計先？
812. 我老師依家暫時住喺 guest house。
813. guest house 會唔會好似酒店咁貴㗎？
814. 教授會 guide 你點做。
815. 我 guide 咗佢好耐佢都未學識。
816. 你覺得 guilty 咪唔好做囉。
817. 佢少少 guilty 都冇，真係不知所謂。
818. 頂佢唔順就走咗去做 gym。
819. 放工就過咗銅鑼灣做 gym。
820. 佢得閒就會研究吓點樣 hack 人。
821. 你廿四小時上網唔驚俾人 hack 咩？
822. 住喺 hall 唔係方便啲咩？
823. 住咗一年 hall 之後我已經唔想再住。
824. 你平時都唔參加 hall function，梗係識唔到人啦。
825. 啲 hall function 唔係飲飲食食就係玩。
826. 你點解唔用原裝嗰個 hand free？
827. 其實 hand free 係咪都有輻射？
828. 你見唔到我就唔駛 hand in 份功課㗎嗱？
829. 你幾時 hand in 我咪幾時睇囉。
830. 佢個 handbag 同我嗰個好似樣。
831. 呢個 handbag 係邊個㗎？
832. 估唔到連手提電話都會 hang 機。
833. 一開呢個軟件就 hang 機喎。
834. 上到飛機之後我就周圍同啲團友影相，好 happy。
835. 呢幾日都幾 happy，不過首先都係講下啲衰嘢先。
836. 唔一定要 happy ending 先至好睇㗎。
837. 套套戲都係 happy ending㗎咩？
838. 我跑埋 happy run 都唔夠分住宿。

839. 跑 happy run 可以加兩分宿分。
840. 而家啲 hard disk 平咗好多。
841. 我想買一個唔好太貴嘅 hard disk。
842. 當唔當係 hard time 睇你點睇㗎啫。
843. 每個人都會有 hard time。
844. 唔知阿 head 俾唔俾我放假呢？
845. 做阿 head 有時係要有少少犧牲。
846. 你喺 heading 度寫番清楚你個題目喎。
847. 個 heading 用邊隻字體好呢？
848. 呢次我要做到一百分有 heart 嘅演唱會。
849. 有冇 heart 我一睇就知啦。
850. 佢只係同我講咗句 hello 咋。
851. 我只不過係想同佢講句 hello 啫。
852. 地鐵出咗套 Hello Kitty 飛喎。
853. 我就唔鍾意 Hello Kitty 喇。
854. 佢咁 helpful，咪次次都叫佢幫手囉。
855. 佢讚你 helpful 喎。
856. 佢哋兩個即刻 high 到忘晒形。
857. 我哋又喺度勁大聲咁嗌，個個都 high 到死。
858. 呢度好似好 high class 咁喎。
859. 有冇啲 high class 少少嘅選擇？
860. 佢用咗好多時間先至整到呢個 highlight。
861. 佢啲頭髮做咗金色 highlight。
862. 好多時大家都對 hip hop 有誤解。
863. 好多人會覺得 hip hop 係唔好嘅嘢。
864. 我淨係中三嗰年讀過一年 history 咋。
865. 啲同學噚日去溫 history。
866. 我哋個個都曉得 hit 波。
867. 呢期佢好 hit 呀，俾人當飯後話題咁勁講。
868. 唉，究竟邊個 hold 住我本書呀？
869. 過咗一個禮拜我已經續借本書，但都難逃被 hold 嘅命運。
870. 無論係朱古力味定栗子味，都係 homemade 嗰隻好食啲。
871. 呢度嘅糖果全部都係 homemade㗎。
872. 你知唔知中大個 homepage 點去？
873. 佢個 homepage 整得好靚。
874. hopefully 佢可以快啲好番。
875. hopefully 佢早日可以再返工。
876. 你成日都用 hot key，唔怪得快我咁多啦。
877. 有咁多 hot key，記唔晒㗎喎。
878. 真係唔明 how come 佢變咗咁多。
879. how come 你今日着成咁？
880. 我想學寫 HTML 應該睇邊本書？
881. 仲有得玩 HTML 添呢。
882. 一隻 hub 可以俾幾多部電腦用？
883. 我唔識點樣駁隻 hub 喎。
884. 我知我可能 hurt 到你。
885. 佢嘅一啲話 hurt 到人。
886. 其實初初我諗住實見佢唔到，I mean 今日。
887. 雖然我讀唔耐，但都唔係唔讀㗎嘛，I mean 未走嗰段時間。
888. 有啲人話 i-cable 唔好用喎。
889. 你宿舍睇唔睇到 i-cable？
890. 呢啲 icon 全部都係我自己整嘅。
891. 點樣可以轉用第二個 icon 呀？
892. 可以乜都傾，又可以晚晚 ICQ。
893. 我會主動咁同大家 ICQ㗎！
894. 邊間公司打 IDD 會平啲？
895. 你用手提電話打 IDD 呀？
896. 我哋係喺星期五先真真正正傾 idea。
897. 唔係成日都有咁好嘅 idea。
898. 唔會次次都咁 ideal 嘅。
899. 你諗得咁 ideal㗎！
900. 你封信唔可以同個樣本 identical。
901. 你抄還抄都冇理由連個名都 identical㗎嘛。
902. 我唔會 ignore 任何機會。
903. 你哋可以 ignore 佢。
904. 你未經人同意就用人哋啲嘢係 illegal㗎。
905. 有時我都唔識分點爲之 illegal。
906. 一向大家嘅 image 都唔同㗎啦。
907. 佢個 image 開演唱會用就啱。
908. 邊啲嘢係 important㗎？
909. 好 important㗎，千祈唔好漏口風喎。
910. 佢呢個方法簡直係 impossible。
911. 有時候你以爲 impossible 嘅嘢喺十幾年後就會成真。
912. 面試嘅時候俾人哋嘅第一個 impression 係好重要嘅。
913. 佢嘅打扮會令人對佢嘅 impression 唔係咁好。
914. 我好 impressive 佢嗰日嘅髮型。
915. 佢一入嚟已經令我好 impressive。
916. 我到目前爲止已經 in 咗五份工。
917. 最緊要有得去 in，咁就有機會俾人請。
918. in case 有事都唔係你同我負責啦。
919. 你咁都唔走，in case 真係火燭點算？
920. in general 嚟講大學生係應該醒目啲。
921. 其實 in general 我哋會請大學生。
922. 如果你 independent 一啲，依家就唔會搞成咁啦。
923. 佢係想你 independent 咁做晒啲嘢。
924. 如果係 individual 咁計分，佢哋又點會唔做嘢呀？
925. individual 嚟講表演唔錯，不過一夾埋就唔掂。
926. 呢啲 information 係邊個提供㗎？

927. 有冇多啲 information？
928. 你應該喺呢度 insert 一啲嘢。
929. 喺呢度 insert 會唔會有問題。
930. 其他軟件我可以自己 install。
931. 我自己都識 install 啦。
932. instead of呢個方法你仲可以試吓其他方法。
933. 其實 instead of 飲茶，你可以試吓去西餐廳。
934. 呢個 instruction 都講得唔清楚。
935. 佢都冇俾 instruction，叫我點做啫？
936. 佢有時都幾 intersting。
937. 我覺得呢個題目好 interesting。
938. 你舊年係咪有去 intern 呀？
939. 我去北京 intern，做咗三個月嘢。
940. 有得去 international 嘅大公司做嘢就梗係好啦。
941. 呢個會係 international㗎，好多大人物都會出席。
942. internet 上面可以搵到好多有用嘅資料。
943. 有咗 internet 之後，啲人會唔會少咗同人面對面溝通呢？
944. 我有得去上海做 internship。
945. 好多人都係去同一個地方做 internship。
946. 今次 interview 輸得好徹底。
947. 不過，我已經決定咗唔去 interview。
948. 前排掛住 interview，走咗好多堂。
949. 個個朋友都忙於 interview。
950. 做緊負離子中途，收到電話，叫我去 interview。
951. 我估呢個 interview 係我最開心嘅一個。
952. 佢真係巴閉，連去 interview 都遲到。
953. 去 interview 之前一定要起清楚份工嘅底。
954. 除咗 intro，其他都做得幾好。
955. 佢 intro 嗰陣講得唔錯㗎。
956. 你咁講即係你同意佢 introduction 入面講嘅嘢？
957. 今日本來想睇晒本書，但淨係睇完個 introduction。
958. invisible 即係你見到人，但人哋見唔到你。
959. 佢好多時都係 invisible㗎啦。
960. 我見唔到你個 IP 喎。
961. 你個 IP 有其他人用緊喎。
962. 你知唔知 IP phone 係咩嚟？
963. 有啲公司而家已經轉用咗 IP phone。
964. 而家啲人對 IQ 已經冇以前咁重視。
965. 有冇 IQ，做吓實驗咪知囉。
966. 自問讀咗 IVE 之後見多咗唔同嘅人。
967. 收到 IVE 寄嚟嘅課程資料。
968. 部影印機一印親雙面就 jam 紙。
969. 幾時再得閒一齊 jam 歌呀？
970. 唔係人人都鍾意 jazz㗎嘛。
971. 有人鍾意 jazz，有人鍾意流行曲，各有喜好啫。
972. 有時成個月一個 job 都冇，收入好唔穩定。
973. 一個 job 有成幾萬蚊㗎。
974. 佢 join 我哋嘅時候我哋都差唔多走喇。
975. 之後就去維園 join 阿文。
976. journal 對學生嘅中英文成績要求都好高。
977. 唔一定要讀過 journal 先可以做記者。
978. 我係成間公司最 junior 嗰個。
979. 佢咁 junior 都夠膽話我。
980. 之後我哋搭西鐵去元朗唱 K。
981. 唔係行街、睇戲、唱 K 就係食嘢㗎喇。
982. 希望我可以 keep 住呢條氣一路去。
983. 希望能夠 keep 住個個星期跑步呢個習慣。
984. 我依家呢個諗法係可以 keep 住守落去嘅心態。
985. 我覺得佢實會幫我 keep 住。
986. 做運動係希望 keep fit。
987. 行路都係運動，都可以 keep fit。
988. 佢成日帶我遊花園，硬係唔講 key point。
989. 佢講嚟講去到去唔到 key point。
990. 我買咗個新嘅 keyboard。
991. 我嗰陣係用 keyboard，但依家要用手掣。
992. 之後去咗 KFC 食下午茶。
993. 我就鍾意去 KFC 多啲喇。
994. 會唔會拍 kiss 戲嘅時候借位？
995. 男主角 kiss 女主角。
996. 佢俾人一個好 knowledgable 嘅感覺。
997. 我唔認爲佢 knowledgable 教呢科。
998. 佢男朋友係 Korean 嚟㗎。
999. 原來你同房係 Korean 嚟㗎？
1000. 我一陣會去 lab。
1001. 唯有喺 lab 做嘢避開佢哋。
1002. 好彩我都借到 label 紙貼住先。
1003. 啲 label 真係好靚。
1004. 入咗大學之後我嘅 language skill 似乎越嚟越差。
1005. 有冇啲課程可以提升 language skill㗎？
1006. 我覺得如果你唔講嘢嘅話，去 language table 棧嘥時間。
1007. 我舊年參加咗好多次 language table。
1008. 我想買部新嘅 laptop。
1009. 我屋企已經有兩部 laptop。
1010. 話晒最 last 一次，唔緊要啦。
1011. 最 last 嗰堂喺度勁畫公仔。

1012. last day 嗰日我哋留到九點幾至走。
1013. 通常 last day 嗰日會請同事食餅。
1014. 屋企個 LCD Mon 用咗五年都冇壞過。
1015. 我個 LCD Mon 壞咗，整咗成三百蚊。
1016. 佢會係一個出色嘅 leader。
1017. 我哋啲 leader 又喺度鬧人。
1018. leadership 呢樣嘢唔係人人都有。
1019. leadership 係可以學嘅。
1020. 呢份 lecture notes 寫得好詳細。
1021. 今次份 lecture notes 好多錯字。
1022. 如果嗰堂去 lecture room 要行到氣咳嘅話都走。
1023. 入到 lecture room 發現一個人都冇。
1024. 佢雖然好忙，但 leisure 嘅時候都會全情投入咁去玩。
1025. leisure 嗰陣我會睇吓書，聽吓音樂。
1026. 你同佢嘅 level 差好遠喎。
1027. 我發現我啲英文 level 係咁向下滑。
1028. 好唔 like 嗰個人囉，死人麻甩佬。
1029. 我真係好唔 like 結焦嗰舊嘢，搞到好明顯呀。
1030. 呢條 line 嘅對象係三十歲以上嘅女士。
1031. 你咁後生，唔應該用呢條 line 嘅化妝品。
1032. 政府網頁條 link 係乜嘢。
1033. 答案可以喺呢條 link 搵到。
1034. 佢已經 list 咗要注意嘅地方出嚟。
1035. 成個 list 得八十五個人咋喎。
1036. 佢靠個樣又唔係靠把聲搵食，所以啲歌迷都唔介意佢係咪唱 live。
1037. 佢唱 live 好正，絕對唔會走音。
1038. 今年仲有 live band 現場彈奏添。
1039. live band 會重複地唱，務求你唔識唱都識哼。
1040. 你將隻碟 load 落部電腦度咪得囉。
1041. 你俾我嗰隻碟一 load 就死。
1042. 個 lobby 整得咁靚，但係入面就爛溶溶。
1043. 十五分鐘後喺 lobby 集合。
1044. 我哋主要收 local 學生。
1045. local 畢業生對香港熟啲。
1046. 我 locate 唔到自己嘅位置。
1047. 有地圖都 locate 唔到依家喺邊。
1048. 呢個電話好勁，喺任何 location 都打到。
1049. 我都唔知自己依家嘅 location。
1050. 我部電腦俾佢 lock 咗嚟用。
1051. 佢 lock 咗個櫃，我依家開唔到。
1052. 我去鎖 locker，點知一返轉頭佢已經走咗。
1053. 我搵吓啦，應該喺是但一個人個 locker 度嘅。
1054. 你個 login name 一定要全部細楷。
1055. 我唔記得咗個 login name 係咩添。
1056. 呢個 logo 係咪好特別呢？
1057. 點解連個 logo 都可以整錯㗎？
1058. 你今日個 look 好得喎。
1059. 佢話我個 look 唔似大學生。
1060. lunch 嗰陣轉咗新地點食飯。
1061. lunch 之後嗰堂我哋落去咗書展。
1062. lunch 之後就到咗班際四乘一百接力。
1063. 放 lunch 山長水遠由長沙灣走入嚟沙田食飯。
1064. 一返到去見到好多人都食緊 M 記。
1065. 跟住落咗 M 記食早餐。
1066. 講咗好多費話，完全講唔到 main point。
1067. 佢好犀利，一針見血，一句說話就已經講到 main point。
1068. 唔理你 major 邊科，大家都係考同一個試。
1069. 好多時畢業之後做嘅嘢都同 major 無關。
1070. 其實 majority 嘅人都支持佢連任。
1071. 如果有幾個唔同意見，要有幾多人支持先至算係代表 majority 呢？
1072. 冇 make appointment 嘅話可能要等好耐。
1073. 我上個月已經同佢 make appointment。
1074. 做啲嘢都唔 make sense 嘅。
1075. 你啲問題好唔 make sense。
1076. 你唔 make sure 就唔好咁大聲。
1077. 佢 make sure 一定成團先至收錢。
1078. 但係好似乜嘢系畢業嘅人都可以做 management trainee。
1079. 我會申請 management trainee。
1080. 個 manager 話啲甜品唔係佢哋自己整嘅。
1081. 放工嘅時候，我入咗 manager 房，等佢派我落鋪。
1082. 份 manual 全部都係我寫㗎。
1083. 如果冇 manual，好難學得識㗎喎。
1084. 本記事簿入面有學校嘅 map。
1085. 我畫完幅所謂嘅 map 就收工。
1086. 我已經留咗好多 margin 位。
1087. 個 margin 最少要有一吋。
1088. 呢個 market 好細。
1089. 你唔覺得好有 market 咩？
1090. 我覺得都係讀 marketing 好啲。
1091. 第一科就係考 marketing。
1092. 點樣搽 mascara？
1093. 我唔識買 mascara 喎。
1094. 應該買邊個牌子嘅 mask 好呢？
1095. 買保濕定係美白 mask 好呢？
1096. 讀完 master 之後有咩打算？
1097. 我打算讀 master，不過學費好貴。
1098. 到咗全晚焦點所在，就係最佳服裝獎同最 match 服裝獎。
1099. 都幾好，兩個啲話題都好 match。

1100. 唔係全部 material 都咁易搵到。
1101. 你要嘅 material 我已經準備好。
1102. 全部都係 MC，唔識做都可以撞吓。
1103. 個 MC 都嚇到唔知講咩好。
1104. 佢講嘅每句說話都有 meaning。
1105. 我唔係好明佢嘅 meaning。
1106. meaningful 唔一定要用好多字㗎。
1107. 本書 meaningful 得嚟好多錯字。
1108. 一次唔得咪 measure 多次囉。
1109. 我唔識 measure 個電壓喎。
1110. 今次個 median 高過上次。
1111. median 即是最中間嗰個人嘅分數。
1112. 一陣個 meeting 有幾多人出席？
1113. 呢個 meeting 好悶呀。
1114. member 嘅質素好與壞。
1115. 啲 member 又真係好似跳得好咗咁喎，唔知係咪心理作用。
1116. 佢連打張 memo 都唔識，點做嘢㗎？
1117. 今朝收到張 memo，放咗喺你檯面。
1118. mentally 好精神，不過體力支持唔到。
1119. 有時做嘢做得耐，會 mentally 好攰。
1120. 點知佢留 message 俾我，話唔得閒嚟。
1121. 噚晚應該收到我個 message。
1122. 下個星期就要 mid-term。
1123. mid-term 之前我都唔會得閒。
1124. 佢自己話唔 mind 嘅。
1125. 我真係唔 mind 佢哋係外校生。
1126. 佢問我 mind 唔 mind 佢食煙。
1127. 你 mind 唔 mind 同我一間房？
1128. 佢 minor 日本研究。
1129. 你以前有冇 minor 過任何科目？
1130. 佢講嘢有時都幾 misleading。
1131. 呢本書嘅解釋好 misleading。
1132. 嗰個代課 Miss 好怪，好似精神分裂咁，傻傻地㗎。
1133. 原來教我哋個 Miss 今日仲度緊蜜月。
1134. 佢上個星期 miss 咗一堂。
1135. 通常都係順手先影，所以 miss 咗好多朋友嘅相。
1136. 我最討厭 missed call。
1137. 我哋一傾就一個鐘，期間無限個 missed call。
1138. 淨色衫可以容易啲 mix and match。
1139. 佢好耐之前已經開始 mix and match。
1140. 好多人以為做 model 搵錢好容易。
1141. 但係唔係人人都適合做 model。
1142. 攞住份 model answer 照抄，咁做功課仲有咩意義？
1143. 我都冇 model answer，全部都係自己做㗎。
1144. 我個 modem 壞咗。
1145. 可唔可以自己買過個 modem？
1146. 呢度嘅裝修好 modern。
1147. 呢個古董手袋個款好 modern，一啲都唔老土。
1148. 喺嗰個 moment 我真係好感動。
1149. 呢個 moment 唔適合講呢樣嘢住。
1150. 你想買邊種 mon(monitor)呀？
1151. 一個十七吋 mon(monitor)要幾多錢？
1152. 近排冇乜 mood 做嘢。
1153. 跟住又冇晒 mood 溫書。
1154. 邊個唔記得起身我可以 morning call 佢。
1155. 六點鐘阿政 morning call 我。
1156. 隻火牛嘅風扇壞咗會唔會影響到塊 motherboard㗎？
1157. 我想換過塊 motherboard。
1158. 佢想買可以 mount 上櫃嗰啲電腦。
1159. 將啲機 mount 晒上櫃咪可以慳番好多位囉。
1160. 我隻 mouse 好襟用。
1161. 九十八蚊一隻 mouse 係咪好貴？
1162. 最尾嗰兩份獎竟然係我發夢都想要嘅 MP3。
1163. 竟然要我貼錢送書劵、公仔同 MP3 俾人。
1164. 佢男朋友都係讀緊 MPhil。
1165. MPhil 要讀幾耐㗎？
1166. 你識唔識用 MSN？
1167. 用 MSN 駛唔駛錢㗎？
1168. 真係好想知 MT 究竟係咪我想像中嗰回事。
1169. 原本請咗五個 MT。
1170. 有得揀我梗係想做 multiple choice 啦。
1171. multiple choice 嗰份卷好難做。
1172. 點知邊面係 negative 呀？
1173. 佢嘅諗法好 negative。
1174. 佢唔會 neglect 任何結識異性嘅機會。
1175. 佢就係 neglect 咗呢樣嘢所以先至錯。
1176. 又為 neighbour 祈禱。
1177. 主席叫我哋同 neighbour 手牽手。
1178. Netvigator 雖然係貴啲但勝在穩定。
1179. 好多同學都係用 Netvigator。
1180. 我都係鍾意去 neway 多啲。
1181. neway 門口有好多人，所以都係去第二間。
1182. 旺角都有 neway city 啦。
1183. neway city 都普通嗰啲有咩唔同？
1184. 點樣睇 newsgroup？
1185. 點解喺屋企睇唔到中大嘅 newsgroup 嘅？
1186. 佢咁 nice 一定肯幫你嘅。
1187. 都唔知應該話佢 nice 定話佢蠢好。
1188. 佢個 nickname 好得意。
1189. 上網梗係用 nickname 啦，邊有人用真名

喋？

1190. 今日放假，出咗去睇 Nike 有冇新貨。
1191. 晏晝就過咗旺角睇人哋 Nike 試範。
1192. 個角度係咪由 normal 嗰邊計起？
1193. 同塊鏡垂直嗰條線係咪即係 normal？
1194. 邊隻牌子嘅 notebook 最襟用？
1195. 之後拎返部 notebook。
1196. 落咗堂去試下點用影印機，印頭先份 notes。
1197. 好勁咁將九課書溫完仲做埋 notes。
1198. 係咪一定要用網上行先至可以睇到 now？
1199. now 有啲咩台可以睇？
1200. 你張單個 number 好唔清楚喎。
1201. 可唔可以俾你嘅電話 number 我。
1202. 我家姐都係讀 nursing。
1203. 唔係人人都適合讀 nursing。
1204. 其實我都冇去 O camp，所以冇乜感受。
1205. 嗰日嘅節目有 O camp 回顧，而我哋又冇去過，所以驚去到會好悶。
1206. 其實佢都係想你 objective 啲啫。
1207. 你咁 objective，難怪老細咁鍾意問你意見啦。
1208. 佢都只係 occasionally 過嚟搵你啫。
1209. 呢啲嘢 occasionally 就可以，長時間就唔得。
1210. 之前去 ocean park 嘅時候。
1211. 我開始熟習咗 ocean park 嘅環境。
1212. 佢一共有三個 offer。
1213. 我放棄咗人工高啲嗰個 offer。
1214. 喺 office 做嘢都可以做到索索地氣喋！
1215. 你有冇佢 office 嘅內線？
1216. 過咗「試用期」，我哋就會返上 office，做自己想做嘅嘢。
1217. 成個 office 都無人返咁濟。
1218. 點解 office hour 一個人都冇喋？
1219. 雖然過咗 office hour，但佢都肯幫我。
1220. 聯絡唔到佢，佢 offline 咗。
1221. offline 之後我就出咗去。
1222. 只不過係想見下啲學生，都 OK 無聊。
1223. 我嗰陣真係好嬲，不過依家 OK 啦。
1224. 套戲俾我想像中差，得一、兩幕係 OK 咋。
1225. 返屋企煮生蠔煲，味道 OK 喇。
1226. 呢度嘅食客都係斯文嘅 OL。
1227. 呢度長時間都會有一班 OL 捧場。
1228. 你依家係 on behalf of 邊個先？
1229. 佢有咁多個身份，都唔知 on behalf of 邊便。
1230. 呢條 one piece 襯晒你啦。
1231. 唔知我着 one piece 裙好唔好睇呢？
1232. 近排好忙，所以冇乜點 online。
1233. 今晚又見到你 online。
1234. 依家啲 online game 其實有冇錢賺？
1235. 我都想學吓寫 online game。
1236. 斷完線之後 on 返 line，一個人都冇。
1237. on 返 line 之後首先係去自己個網頁。
1238. 最後一個建議比較 open。
1239. 啲日本人好極端，一係好保守，一係好 open。
1240. 多數考試都係 open book。
1241. 你估 open book 就唔駛溫書咩？
1242. open notes 要抄貓紙，好嘥時間。
1243. 今次測驗係 open notes，可以帶三張紙。
1244. open U 畢業，啲公司承唔承認喋？
1245. 有得揀我就唔會讀 open U 喇。
1246. 問題係我冇其他 option。
1247. 如果有另一個 option 呢？
1248. 我凡親上晝考 oral 就一定肚痛。
1249. 好彩會考同高考 oral 都喺下晝。
1250. 意粉係接到 order 先至淥。
1251. 要積極啲唔好坐喺度等 order。
1252. 八月中先至有 orientation camp。
1253. 好多人對 orientation camp 嘅印象都唔係咁好。
1254. 好多時讀書都係考試 oriented。
1255. 成績 oriented 太過短視喇。
1256. 你用開邊個 OS？
1257. 我都想試吓自己裝 OS。
1258. 仲未做晒啲嘢，所以要開 OT。
1259. 想睇下自己可以連續幾多晚 OT。
1260. 其實我哋好少去 outing。
1261. 一年先至得一次 outing。
1262. 我唔識用 outlook 嚟睇新聞組。
1263. 咁你識唔識用 outlook 呀？
1264. 佢咁 outstanding，去到邊度都掂啦。
1265. 最 outstanding 嗰個咪係我囉。
1266. 估唔到佢都去到 outward bound。
1267. outward bound 好辛苦喋。
1268. 如果 over 八千蚊我就唔買囉。
1269. 人工 over 二萬蚊一個月喋。
1270. 去 overseas 唔係唔好，不過我都鍾意香港多啲。
1271. 中大多唔多 overseas 嘅學生？
1272. 你 pack 咗你啲行李未？
1273. 啲行李 pack 得好啲就可以放多啲嘢。
1274. 有幾個 package，價錢都好大差別。
1275. 不過，仲有兩種 package 要揀。
1276. 之前幾次都只係得一 pair 拍緊拖。
1277. 依家就變成三 pair。
1278. 我跌咗部 palm 落地。
1279. 佢當部 palm 係遊戲機咁。
1280. 見到檔賣 pancake 嘅，好好食。
1281. 我食咗好多嘢：燒雞串、燒賣、雞蛋仔

同埋 pancake。
1282. 今日教我哋點睇 paper，但佢講得勁慢。
1283. 總共有兩份 paper，佢只講咗其中一份嘅四頁。
1284. 係咪得一個 parallel port㗎咋？
1285. 點知係咪個 parallel port 壞？
1286. 每年都會有兩次 Parent's Day。
1287. Parent's Day 多數係星期日，因爲多數人都唔駛返工。
1288. 呢 part 不得不提，我哋有個團友買咗枝巨型人參。
1289. 我連停咗係邊 part 都唔記得(lu)。
1290. 我又覺得唔係咁好，驚累咗個 partner。
1291. 今日打排球終於有 partner，球技同我差唔多。
1292. 最慘仲要問埋我哋有冇做過 part-time。
1293. 其他人都做過 part-time，好多仲係做倉務。
1294. 去 party 撞到咪打個招呼囉。
1295. 感恩會之後去咗開 party。
1296. 佢考車一次過 pass 咗。
1297. pass 唔到再考過好貴㗎。
1298. passing rate 咁高，唔駛擔心啦。
1299. 無論個 passing rate 係點，我都係照讀㗎啦。
1300. 本 passport 過咗期幾個月。
1301. 出門口之前檢查吓有冇帶 passport。
1302. 個 password 要有數目字同英文字母。
1303. 唔見咗個 password 點算？
1304. 成日坐地下，坐到 pat pat 好痛。
1305. 佢唔知點樣撞瘀咗個 pat pat。
1306. 平均一日要有三十個 patient 先至有錢賺。
1307. 唔夠 patient 唔通要喺街度派傳單咩？
1308. 自己設計一啲 pattern。
1309. 推出以經典格仔 pattern 爲題嘅餐具系列。
1310. 有得 pay 咪落力啲囉。
1311. 冇得 pay 咪求祈做完就算。
1312. 有邊啲地方需要 pay attention？
1313. 除咗開始嘅時候要 pay attention，其他時候都唔應該鬆懈。
1314. 邊隻牌子嘅 PDA 最好用呀？
1315. 我想買部有埋藍芽嘅 PDA。
1316. 第一次上 PE 堂上得咁鬼開心。
1317. 跟住就係 PE 堂，我哋今次玩手球。
1318. 產生咗感情就會一百 percent 投入。
1319. 超過三十 percent 收入來自股票投資。
1320. 當然唔可以一下做到咁 perfect 啦！
1321. 如果依家有埋紅酒就 perfect 喇。
1322. 我信得過波子嘅 performance。
1323. 佢噚日嘅 performance 好好。
1324. pharmacy 畢業好易就搵到工。
1325. 今年 pharmacy 收得好少人咋。
1326. 我都冇諗住讀 PhD。
1327. 香港邊度有咁多工要 PhD 先至做得吖？
1328. 個 photo album 可以放幾多相？
1329. 你用開邊個 photo album㗎？
1330. 我已經 photocopy 咗成份嘢。
1331. photocopy 同正本要分開放好。
1332. 皆因小弟一向都冇乜 physical。
1333. 加上我又冇乜 physical。
1334. 我屋企樓下開咗間 Pizza Hut。
1335. 我哋最後都係去 Pizza Hut。
1336. 約咗保險經紀傾保險同儲錢 plan。
1337. 啲嘢都 plan 好晒，唔好要我改假就得喇。
1338. 佢一向都咁有 planning，唔駛擔心喎。
1339. 有我負責做 planning，你可以放心。
1340. 我覺得見到佢已經係我嘅 pleasure。
1341. 有得嚟面試已經係我嘅 pleasure。
1342. 份功課要自己 plot 圖喎。
1343. 有冇啲軟件可以 plot 得快啲㗎？
1344. 一個 point 先至得五分。
1345. 個 point 唔啱會倒扣分㗎。
1346. 工作包括制定政府嘅 policy。
1347. 討論吓個新 policy 有咩問題。
1348. 依家邊個歌星最 popular？
1349. 請 popular 嗰啲好貴㗎。
1350. 我想開返用嚟上網嗰個 port。
1351. 咁我哋一怒之下咪去 port 佢囉。
1352. 呢個 pose 好有型。
1353. 佢影相成碌木咁，完全唔識擺 pose。
1354. 影響係咪咁 positive 要遲啲先知。
1355. 你諗嘢 positive 啲，個人咪會開心啲囉。
1356. 又有人回覆我個 post 喎。
1357. 我好似已經成四、五日 post 唔到嘢。
1358. 貼 poster 要注意啲咩？
1359. 定係未入戲院前喺 poster 上睇到？
1360. 你有冇諗過讀 postgrad？
1361. postgrad 啲宿舍貴好多㗎。
1362. 個測驗 postpone 咗，可以多啲時間準備。
1363. 冇得 postpone，咁我唯有通幾晚頂囉。
1364. 你好有 potential 成爲天王巨星。
1365. 可能佢嘅 potential 好大。
1366. 我有個好 practical 嘅問題。
1367. 有時 practical 啲都好嘅。
1368. 今次咁差都係因爲 practice 得唔夠啫。
1369. 我諗都係要 practice 多啲至得，熟能生巧吖嘛。
1370. 以前啲 prefect 好惡㗎！
1371. 我做 prefect 嗰陣都遇過類似嘅問題。
1372. 佢其實 prefer 我負責帶嘢。
1373. 我 prefer 去燒嘢食多啲。

1374. 可能我心裡已經有 preference。

1375. 佢哋一早有 preference，個面試只係做吓樣。

1376. 如果夠時間做 preparation 就唔會搞成咁啦。

1377. 你 preparation 都未做好就走去玩？

1378. 我覺得佢好似冇 prepare 就嚟上堂咁。

1379. 你 prepare 好一陣要講嘅嘢未？

1380. 我下晝 present 完搵你吖。

1381. 你去見工其實係去 present 自己。

1382. 佢可能 pressure 太大。

1383. 如果冇 pressure 你會咁落力咩？

1384. 你 print 咗份功課未？

1385. 你 print 咗一陣開會用嘅嘢未？

1386. 呢部 printer 可唔可以印到雙面？

1387. 今日買咗部細細部嘅 printer。

1388. 喺我心目中，你嘅 priority 梗係高啲啦。

1389. priority 就一定有，不過唔知有幾重。

1390. 你哋咁樣影相係侵犯我嘅 privacy。

1391. 拉埋個簾落嚟有完全嘅 privacy。

1392. 有個超 pro(professional)嘅城大生，個樣已經殺晒。

1393. 喺側邊聽啲 pro(professional)嘅影相佬講嘢。

1394. 唔理 probability 係幾多都好，盡咗力就算。

1395. 唔係冇可能，只係 probability 低啲啫。

1396. 你做嘢嗰度過咗 probation 未？

1397. 過咗 probation 有得加人工。

1398. 你都唔跟個 procedure，梗係做唔到啦。

1399. 有啲 procedure 係唔可以錯㗎！

1400. 試過我哋公司嘅 product 你一定會滿意。

1401. 手提電話真係一種令我又愛又恨嘅 product。

1402. 我哋淨係做 production，其他嘢由第二個部門負責。

1403. 設計完要俾老細睇過至可以做 production。

1404. 佢好 productive，一個人可以做兩個人嘅嘢。

1405. 呢幾日好 productive，所以有時間休息下。

1406. 佢啲口吻好唔 professional。

1407. 你懶係 professional 咁。

1408. 會唔會令個 professor 留低壞印象㗎？

1409. 我就鍾意坐喺度聽 professor 講書多啲。

1410. 我都唔識寫 program 嘅。

1411. 你識唔識寫 program？

1412. 冇計啦，要去做 project 嘛。

1413. 我竟然喺度幫佢哋做 project。

1414. 我已經唔敢 promise 話唔轉工。

1415. 我 promise 咗下年會打俾佢。

1416. 佢成日 promote 呢間公司嘅產品。

1417. 我用開嗰隻洗頭水而家做緊 promote。

1418. 唔好話 pronuciation 啦，好多字我連讀都唔識讀。

1419. 好似有啲課程係專門改善 pronuciation 㗎。

1420. 如果要 proportional 咁計，咪要好多錢囉。

1421. 工作時間同人工都唔係 proportional 嘅。

1422. 佢叫我寫份 proposal 俾佢睇先。

1423. 個 proposal 係我自己一個人做㗎。

1424. 我可以同老細 propose，不過佢未必贊成㗎。

1425. 都唔知係邊個 propose 要拆咗度門嘅。

1426. 今年本 prospectus 靚過舊年嗰本。

1427. 本 prospectus 我有份俾意見㗎。

1428. 佢 proud of 自己係貴族嘅身份。

1429. 佢希望爸爸會好 proud of 自己。

1430. 今次抽獎嘅大獎係一部 PS2。

1431. 其實我都未玩過 PS2。

1432. 除咗 public holiday 之外，差唔多日日都要返工。

1433. 我哋淨係放 public holiday 同埋星期日。

1434. 本書幾時 publish㗎？

1435. 佢公司係幫人 publish 年報嘅。

1436. 我諗佢一下 push up 都做唔到。

1437. 做 push up 係咪可以練到手臂嘅肌肉？

1438. 如果大家都咁 puzzling，不如再講多一次。

1439. 我好 puzzling，完全唔知做緊乜。

1440. 但係條 Q 跑唔出。

1441. 事前做足準備功夫，但係臨門一腳撻Q。

1442. Q&A 嘅時候居然冇人問問題。

1443. 有幾多時間做 Q&A？

1444. 佢 quali 咁好，大把人爭住要。

1445. 雖然 quali 唔好，但工作經驗搭夠。

1446. 你要咁高 quality 嚟做咩？

1447. 我同佢影嗰張相 quality 太低，晒唔返出嚟。

1448. 一年有幾多個 quarter？

1449. 呢個 quarter 唔掂，咪下次努力啲囉。

1450. 最後一次同佢哋 quick change 喇。

1451. 我地要跳兩隻，要 quick change！

1452. 如果呢個月都係咁，我真係想唔 quit 都唔得啦！

1453. 如果未夠一個月就 quit，一日得一百蚊。

1454. 今次個 quiz 嘅問題同上次差唔多。

1455. 個 quiz 得廿二分，死梗。

1456. 反正我哋用唔晒啲 quota，咪分啲俾人囉。

1457. 做唔夠 quota 會俾老細鬧。

1458. 而家啲 RAM 平咗好多。
1459. 我想問啲 RAM 係咪有分唔同速度㗎？
1460. 呢首歌嘅 rap 由佢自己負責番。
1461. 唔係人人都 rap 得起。
1462. 你鍾意我嗰時我未 ready。
1463. 我不嬲都 ready㗎喇。
1464. 唔用 real time scan 可能中咗毒都唔知。
1465. 開咗 real time scan 之後部機好似慢咗咁。
1466. 你梗有自己嘅 reason。
1467. 無論 reason 係咩，佢都當係藉口。
1468. 你中咗六十秒自動 reboot 嗰隻毒？
1469. 佢部電腦唔知點解久不久就會自己 reboot。
1470. 我借嗰本要又有 recall 喇。
1471. 佢要 recall 本書，下星期就要還。
1472. 要攞住張 receipt 入場。
1473. 佢唔見咗張 receipt。
1474. 你有冇 receive 過有關電腦病毒嘅電郵？
1475. 佢一 receive 嘢就死機。
1476. 我唔會 recommend 大家嚟呢度食嘢。
1477. 你會唔會向其他人 recommend 先？
1478. 年紀大咗要多幾日至可以 recover。
1479. 我諗要兩三日至 recover 到。
1480. 我公司最近 recruit 咗幾個新人。
1481. 佢地會唔會 recruit 新會員呀？
1482. 我中學嗰陣都做過 red cross。
1483. 以前間學校有 red cross 同埋交通安全隊。
1484. 點樣自動 redirect 啲人去另一個網頁？
1485. 我俾人 redirect 咗去第二個網頁。
1486. reference book 唔一定要買，去圖書館借都得。
1487. 邊度可以睇到本 reference book 賣晒未？
1488. 我 reg(register)唔到個普通話。
1489. 今個禮拜又 reg(register)唔到，所以多咗兩堂空堂。
1490. 有啲咩 regulation 我係要遵守㗎？
1491. regulation 都係人定出嚟㗎啫。
1492. 個 rehearsal 一個鐘搞唔搞得掂？
1493. 乜你唔駛做 rehearsal 咩？
1494. 佢份建議俾人 reject 咗。
1495. 我唔會亂咁 reject 人嘅。
1496. 佢哋嘅 relation 唔係想像中咁簡單。
1497. 我哋嘅 relation 就好似一家人咁。
1498. 呢種感覺的確好舒服，好 relax。
1499. 好心你唔好咁緊張，relax 一吓啦！
1500. 佢講啲嘢同今日嘅主題一啲都唔 relevant。
1501. 如果再有 relevant 嘅消息就通知我啦。
1502. 要學多啲 vocab，唔可以用死嗰幾個。
1503. 補習嗰陣學咗好多學校冇教嘅 vocab。
1504. 佢以前好 rely 我㗎。
1505. 你唔可以成日 rely 人哋。
1506. 佢已經 remind 咗我，係我自己唔記得啫。
1507. 你一陣 remind 我去開會喎。
1508. 係屋企用唔用到 remote desktop㗎！
1509. 點解我用唔到 remote desktop 嘅？
1510. 公共圖書館啲書可以 renew 五次。
1511. 張信用卡到期之後銀行會唔會幫你 renew？
1512. 佢中五嗰陣已經 repeat 過兩次。
1513. 佢講嘢成日都 repeat 番之前嘅說話。
1514. 係咪 repetition 越多就越入腦先？
1515. 如果 repetition 太多，會唔會好煩？
1516. 呢份 report 係自己做定抄人哋㗎？
1517. 佢寫 report 好快㗎，而且仲寫得好好添。
1518. 係佢 require 你咁做，定係你自己想咁做？
1519. 我冇 require 你八點鐘返工喎。
1520. 你一定要乎合份工嘅 requirement 先至有機會。
1521. 其實睇番佢個 requirement，你應該合資格㗎喎。
1522. 你做晒 research 之後先至去見工，會令人覺得你好有誠意。
1523. 我已經做過 research，知道晒市場需要。
1524. 我一畢業就做 research assistant，之後先至出去搵工。
1525. 你諗住做 research assistant 做到幾時？
1526. 呢度係冇得 reserve㗎！
1527. 請問兩位有冇 reserve 到檯㗎？
1528. 點樣將個 resolution 改低啲呀？
1529. 我真係好想問下佢知唔知乜嘢係 resolution。
1530. 都冇 resources，叫人點做嘢喎。
1531. 唔夠 resources 一定要出聲。
1532. 佢覺得自己俾咗錢所以唔駛 respect 啲老師。
1533. 佢識得 respect 人㗎咩？
1534. 你咁 responsible 咪幫佢做埋份功課囉。
1535. 你真係咁 responsible，就唔會做做下嘢走咗去啦。
1536. 以前會將 result 貼喺壁報板上。
1537. 無論 result 係點我都唔會留低。
1538. review 完之後仲要等幾耐？
1539. 每兩個月做一次 review。
1540. 係咪隻隻碟都可以 rewrite㗎？
1541. 可以 rewrite 嗰啲碟幾錢隻？
1542. 有好 rich 嘅水果味。
1543. 朱古力味 rich 得嚟又唔覺滯。

1544. 你 right click 就得㗎喇。
1545. 點解我 right click 完之後冇反應嘅？
1546. 佢 roomate 今朝五點幾至返。
1547. 你 roomate 成日都唔返嚟瞓㗎？
1548. 佢第一個 round 就俾人飛咗出局。
1549. 本來玩得好好哋，但到第二 round，佢就發爛渣。
1550. 唔用 router 得唔得？
1551. 我係咪應該用 router 駁住兩部電腦？
1552. 今晚都係嚟睇位同睇佢個 rundown。
1553. 因爲知道佢成個 rundown，所以影出嚟唔錯。
1554. 呢期好忙，搞 SA 選舉啲嘢已經忙到死咗。
1555. 玩 SA 其實唔係咩大問題。
1556. salary 之外，表現好仲有額外獎金。
1557. 除咗 salary，你冇其他嘢要考慮？
1558. 你等到 sale 嗰陣至入貨囉。
1559. 每隔一段時間就會有 sale，唔駛咁心急。
1560. 呢種女人真係好得，識扮嘢又識擦鞋，抵佢一世做 sales。
1561. 原來做 sales 係逐個各自食飯嘅。
1562. 一開始要落鋪做 sales，我估會幾辛苦。
1563. 俾個 sales 說服咗我買個袋。
1564. 佢兩個星期就去一次 salon 整頭髮。
1565. 唔知佢間 salon 生意點呢？
1566. 做完 sampling 之後會點？
1567. 我唔明 sampling 有咩用。
1568. 我一 say bye 佢就到。
1569. 我同佢 say bye 佢先至知我嚟咗。
1570. 見到咪大方啲 say hi 囉。
1571. 我過去同佢 say hi 先。
1572. 你叫到我又點會 say no 呢？
1573. 雖然明知做阿四但我都唔會 say no。
1574. 佢唔 say sorry 不特旨，態度仲好差。
1575. 條友見我超住佢先至即刻 say sorry。
1576. 我想 scan 呢幅圖。
1577. 唔該幫我 scan 咗呢幾頁佢吖。
1578. 根據而家嘅 schedule 我哋可以好快教晒啲嘢。
1579. 每日嘅 schedule 都排得密麻麻。
1580. 多數人都覺得讀 science 叻啲。
1581. 會考 science 啲人通常高分啲。
1582. 你做嘢 scientific 啲得唔得？
1583. 佢連食飯都好 scientific。
1584. 我用自己影嘅相做 screen saver。
1585. 你有冇嗰個好多魚游嚟淨去嘅 screen saver？
1586. 你自己上網 search 啦！
1587. search 完間公司嘅資料，我就即刻瞓覺。
1588. 自從第三個 season 之後，我都冇再睇囉。
1589. 知唔知《仁心仁術》做到第幾個 season？
1590. 邊度可以搵到個 seating plan 呀？
1591. 佢叫我畫個 seating plan 俾佢。
1592. 今日一早就去 second in。
1593. 以爲已經冇機會 second in。
1594. 我哋同公司但係唔同 section。
1595. 呢個 section 嘅同事都好好人。
1596. security 唔係冇，不過唔夠啫。
1597. 如果 security 真係咁好，就唔會有人入嚟偷到嘢啦。
1598. 我唔會 selective 咁淨係問某幾個人問題囉。
1599. 佢 selective 咁答學生嘅問題。
1600. 同人 sell 嘢簡直係我嘅死穴。
1601. 事實亦證明佢 sell 嘢嘅技巧好掂，唔請佢真係嘥晒。
1602. 其實今個 sem 想讀返好啲書。
1603. 我今個 sem 最唔捨得嘅係佢。
1604. seminar 之後可以一齊食飯。
1605. 原來下星期有 seminar。
1606. 我會 send 返我哋嘅合照俾你㗎。
1607. 總之佢肯 send 番啲相俾我就得喇。
1608. 佢唔係最高級，但係係最 senior 嗰個。
1609. 佢喺度三年，已經係最 senior。
1610. 我 sense 到嘅時候已經太遲。
1611. 佢 sense 唔到咁大鑊。
1612. 我想買條 serial port 用嘅線。
1613. 佢都唔識咩係 serial port。
1614. 佢好 serious 咁問咗我一個問題。
1615. 某啲情況係要 serious 啲嘅。
1616. 我慣咗要人 serve 我。
1617. 初初都係每一次 serve 一個人。
1618. 佢要買 server，唔係買普通電腦。
1619. 個 server 要廿四小時操作，唔可以死。
1620. 有啲咩 service 係可以用㗎？
1621. service 絕對係一流。
1622. 今日開始玩 set 波，我梗係唔掂啦。
1623. 個波 set 咗過隔離場，我真係睇唔到有任何機會合格。
1624. 一路 set 嘢一路俾人催。
1625. 我啲嘢都冇人幫我 set，點幫你呀？
1626. 呢個 setting 同以前嗰啲唔同喎。
1627. 點解個 setting 改咗嘅？
1628. 我自己都識 setup 啦。
1629. 佢唔係唔識，而係唔得閒做 setup。
1630. 佢啲膚色真係好 sexy 呀。
1631. sexy 啲我咪鍾意多啲囉。
1632. 選擇「共用資料夾」就可以將啲嘢 share 俾人。
1633. 啲筆記印唔夠數，所以要幾個人 share 嚟睇。
1634. 佢喺學校入面唔算太 sharp。
1635. 離遠望見個色已經 sharp 到爆。

1636. 有啲人 shift 咗去隔離嗰行坐。
1637. 佢份工係咪要 shift㗎？
1638. 唔係個個女仔都咁鍾意 shopping㗎嘛。
1639. 我對 shopping 冇乜興趣。
1640. 佢都 short 嘅。
1641. 你明知佢 short㗎啦！
1642. 話明係 short question，答幾句就夠。
1643. 多數都係 short question，唔係好難。
1644. 我個電話 short short 地，間中會打唔到。
1645. 部電腦 short short 地，成日都開唔到機。
1646. 通常過幾日就會出 shortlist㗎喇。
1647. 滙豐筆試嘅 shortlist 出咗未？
1648. 呢個 shot 要再拍過。
1649. 上個 shot 拍得好好。
1650. 擺到明係想 show 嘢，我仲可以點？
1651. 嗰個阿魚淨係掛住去睇 show。
1652. 我即刻 show off 我個新銀包，佢都話靚呀。
1653. 有啲人擺明背咗嘢入去，所以一定要講晒出嚟 show off。
1654. 決定下年 singing contest 分兩組出賽。
1655. 佢把聲超靚，singing contest 佢實係拎頭一、二名。
1656. 阿 sir 仲話今堂係最後一堂教下手，之後就會學上手同開波。
1657. 之後阿 sir 殺波，叫我哋去救。
1658. 你會唔會去 sit 晒全部堂先？
1659. 通常我 sit 到第五六個星期。
1660. 我覺得做 sit up 有效啲。
1661. 一分鐘做唔做到三十下 sit up？
1662. 呢個 situation 我都未見過。
1663. 個 situation 都唔同，唔可以用番以前嘅方法。
1664. 一客嘅 size 足夠兩個人食。
1665. 迷你 size 嘅麵包可以一啖一個。
1666. 除咗呢啲 skill 之外你仲識啲咩？
1667. 我哋贏嘅可以話真係 skill。
1668. 一個 skillful，一個乜都唔識，你會請邊個呀？
1669. 佢好 skillful，一陣就搞掂晒。
1670. 睇唔明咪 skip 咗佢先囉。
1671. 佢冇問我問題，淨係 skip 咗我。
1672. 你 smart 少少得唔得？
1673. 呢個 smart 啲嘅都係咁，真係好唔掂。
1674. 我好怕打中文嘅 SMS。
1675. 收 SMS 係唔駛錢㗎。
1676. 你哋又要五日之後先再見到我㗎喇，so 五日後見啦。
1677. 你哋都唔想見到我俾人話㗎喇，so 唔該晒咁多位！
1678. so far 我覺得佢表現唔錯。
1679. so far 冇咩大問題。
1680. 今年件 soc 褸都係以白色爲主。
1681. soc 房要下晝三點幾至開。
1682. 我諗你要去 social 下至得。
1683. 我都唔鍾意 social，所以都係唔去喇。
1684. 讀 social work 咪仲難搵工。
1685. 佢對 social work 好有興趣。
1686. 唔知點解啲相變到好似用咗 soft 鏡咁。
1687. 呢啲嘢俾人一個好 soft 嘅感覺。
1688. 你用邊隻 software 執相㗎？
1689. 燒唔到碟關唔關個 software 事？
1690. 佢計嚟計去都計唔到個 solution。
1691. 你唔知個 solution 錯㗎咩？
1692. 買音響就買 Sony 啦，佢做收音機起家㗎嘛。
1693. 但係 Sony 嗰部要四千幾蚊。
1694. sorry，我之前啲語氣可能重咗啲。
1695. sorry 呀，我撳錯掣！
1696. 佢同我講 sort 完啲資料就走得。
1697. 我想將啲資料跟大細 sort 好。
1698. 我張 sound card 好似燒咗。
1699. 一張 sound card 要幾多錢呀？
1700. 我有 source code，你唔駛寫得咁辛苦。
1701. 有成幾十頁 source code，有排先睇得晒。
1702. 佢帶咗一份好特別嘅 souvenir 比我。
1703. 每次去旅行都會買好多 souvenir。
1704. 你得一套工具冇 spare㗎？
1705. 你有冇預多啲做 spare㗎？
1706. 今日冇咩 special，都係咁悶。
1707. 影印嗰度冇咩 special，拎相嗰度先值得一提。
1708. 我做 speech 已經做咗三年。
1709. 佢一陣會有個 speech，會講下現時嘅教育制度。
1710. 佢搵到好多公司 sponsor 呢個演唱會。
1711. 一向好少 sponsor 呢類活動。
1712. 我哋學校嘅 Sports Day 分兩日舉行。
1713. 每年嘅 Sports Day 都會有好多啦啦隊打氣。
1714. 佢個個星期都打 squash。
1715. 我覺得打 squash 好危險。
1716. 下一個 stage 幾時開始？
1717. 而家呢個 stage 仲未需要呢啲嘢住。
1718. 佢要我哋喺公司 stand by。
1719. 星期日都要 stand by，所以唔可以走得太遠。
1720. 食環署 stand for 食物環境衛生署。
1721. 啲名簡化到都唔知 stand for 乜嘢。
1722. 佢個 standard 你估人人都做到㗎？
1723. 但係仲未到 standard 囉。
1724. 估唔到佢咁有 standing 都會偷嘢。
1725. 呢啲可能係病態，同 standing 無關。

1726. statistics 顯示香港出生率下降，所以幼稚園首先會受到影響。
1727. 呢啲 statistics 都唔知準唔準。
1728. 每個人都會有自己嘅 story。
1729. 佢構思緊一個關於海嘅 story。
1730. 我希望成隻碟都可以反映到 street culture。
1731. 致力推動香港 street culture。
1732. 佢咁 strong 駛乜人保護呀？
1733. 就係因為你太 strong，所以嚇走晒啲人。
1734. 唔同公司嘅 structure 都會唔同。
1735. 第一日返工就係了解公司 structure。
1736. 佢連自己個 student ID 都唔記得。
1737. 用 student ID 就可以查都個分數。
1738. student union 係冇得退會㗎。
1739. 我覺得個 student union 都唔係代表學生嘅。
1740. 帶啲親友一齊去 studio 影畢業相。
1741. 我好想影 studio，話晒一世人一次吖嘛。
1742. 好多同學都會 study aboard。
1743. 如果冇錢，申請到獎學金都可以 study aboard。
1744. 以前就算 study break，我都冇心向學。
1745. study break 要補課，唔夠時間溫書。
1746. 去 study trip 要用好多錢㗎。
1747. 唔知今年嘅 study trip 會去邊度呢？
1748. 唔係話佢 stupid，只係大家諗法唔同啫。
1749. 佢咁 stupid，講十次都唔會明。
1750. 你知係我 style㗎啦！
1751. 只不過係着件唔啱 style 嘅衫啫。
1752. 我唔係 stylist，唔會教你點襯衫。
1753. 仲送咗幅俾烏龍茶廣告嘅 stylist。
1754. 份功課 submit 去邊度？
1755. 佢已經 submit 咗份報告。
1756. 圖書館 subscribe 咗邊啲雜誌？
1757. 你有冇 subscribe 到呢份期刊？
1758. 唔行 subway 過唔過到去？
1759. 嗰條 subway 有好多人行，所以好安全。
1760. 佢 suddenly 大叫一聲，跟住就暈咗。
1761. 你 suddenly 跳出嚟，嚇咗我一跳。
1762. 我 suggest 佢聽日下晝先至嚟。
1763. 佢 suggest 我遲啲先至搵人。
1764. 一季有兩套 suit 已經好夠。
1765. 啲 suit 係有分冬天同夏天嘅。
1766. 我覺得自己唔係咁 suitable 做呢份工。
1767. 有好多人比我更 suitable 啦。
1768. 個 summary 一定要到肉。
1769. 我最怕就係寫 summary。
1770. 每年 summer holiday 佢都會返香港探屋企人。
1771. summer holiday 就梗係要四圍玩啦。
1772. 我之前做 summer job 見過佢。
1773. 你估搵 summer job 咁易㗎？
1774. 佢係一個好好嘅 supervisor。
1775. 唔係個個 supervisor 都咁衰嘅。
1776. 我哋俾咗好多 support 佢。
1777. 多謝你哋咁 support 我。
1778. suppose 你唔應該喺度。
1779. 呢啲嘢你 suppose 做咗㗎喇喎。
1780. 我唔知個電話 sup 唔 support 視象短訊喎。
1781. 唔理 sup 唔 support 都買咗先。
1782. 雖然個確實日子仲未 sure，不過我都係請定假先。
1783. 佢話有幾種讀法，唔係好 sure 點讀。
1784. 我聽完呢番說話之後好 surprise。
1785. 諗住遲啲先同你講俾個 surprise 你。
1786. 你可唔可以幫我做個 survey？
1787. 好多人扮做 survey，其實係想呃你個電話。
1788. 我想要嘅係真正嘅 sweet。
1789. 收到意想不到嘅禮物，即刻 sweet 到呢。
1790. 有冇人想買 switch 呀？
1791. 我想買隻 switch 將兩部機駁埋。
1792. 點樣可以打 symbols 打得快啲？
1793. 啲 symbols 可唔可以自己整㗎？
1794. 呢個 system 係用嚟查成績嘅。
1795. 我最鍾意公平嘅 system。
1796. 佢話你做嘢唔夠 systematic 喎。
1797. 我爸爸做嘢好 systematic㗎！
1798. 有啲大學會請全職嘅 TA。
1799. 唔單止要做 TA，仲有好多其他嘢要做。
1800. 我舊年 take 咗一個通識。
1801. 乜你唔係上個學期已經 take 咗喇咩？
1802. 等佢出聲先至 take action。
1803. 幾時可以 take action？
1804. 駛唔駛 take attendance 呀？
1805. 我哋上堂唔駛 take attendance㗎？
1806. 冇人會 take care 你嘅感受。
1807. 你自己 take care 啦！
1808. 我通常都會 take notes。
1809. 開會一路聽一路 take notes，好忙㗎。
1810. 佢令人覺得佢好 talent。
1811. 每一個人都有自己嘅 talent。
1812. 我入去聽 talk，只係見到一個男仔。
1813. 之後我哋去咗聽個選科嘅 talk。
1814. 一來唔係正日，二來都冇 target 俾我送。
1815. 我哋依家有 target 嘛，啲後勁上緊嚟喇。
1816. 做晒三個 task 就可以走。
1817. 一個 task 要做成個幾鐘㗎。
1818. 冇 taste 都唔會揀你做朋友啦。
1819. 我覺得佢好有 taste。
1820. 好彩主管好好人，放我食 tea。
1821. 我哋今日食 tea 傾咗好耐。

1822. 我食唔晒一個 tea set。

1823. 唔一定要叫 tea set，可以散嗌。

1824. 同一 team 都唔會日日見到。

1825. 唔同 team，好少合作。

1826. 唔係得你一個人做，我哋 team work㗎嘛。

1827. 咁多人一齊面試，其中一個目的就係想睇下你適唔適合 team work。

1828. 咁多 technical problems 唔怪得遲遲都起唔到貨啦。

1829. 呢啲 technical problems 應該有專人負責。

1830. 好彩個 technical staff 肯幫我咋。

1831. 佢唔係 technical staff，但係識嘅嘢比佢哋更多。

1832. 我完全唔明個 technician 講乜。

1833. 好彩嗰度都有個 technician 幫我手。

1834. 大學會教一啲最新嘅 technology。

1835. 呢個世界咁多新嘅 technology，唔識並唔出奇。

1836. 呢件 tee 係獨一無二嘅。

1837. 我哋會幫客人畫 tee，但係件衫就要個客自己提供。

1838. 依家政府請人多數都係 temp，好少請長工。

1839. 做 temp 今日唔知聽日事，梗係冇咁落力啦。

1840. 呢個 term，我唔識點譯，唔知點擺啲次序。

1841. 佢係咁撻啲完全唔明係乜嘅 term 出嚟。

1842. 今次個 test 死得。

1843. 派咗 test 嗰份成績表，俾阿媽鬧到一面屁。

1844. 我唔知佢係咪想 thank you 我。

1845. 佢一句 thank you 都冇，好冇禮貌。

1846. 呢個咩 theory 嚟㗎？

1847. 好多 theory 都未知有咩用。

1848. 你畀心機努力寫好份 thesis 啦！

1849. 係咪寫 thesis 寫得好辛苦呢？

1850. 我諗我要學下點樣做 time management。

1851. time management 做得好啲，就可以做多好多嘢。

1852. 我個 timetable 已經出咗。

1853. 你俾個 timetable 我，我再同你約時間啦。

1854. 遲早都要落鋪，只不過，我同其他人嘅 timing 唔同。

1855. 快門 timing 好咗，擺位都好咗，好滿意。

1856. 最尾嗰堂通常係講考試 tips。

1857. 食飯駛唔駛俾 tips㗎？

1858. tiramisu 原來本身係冇酒嘅。

1859. 我真係好鍾意個 tiramisu 呀！

1860. 而家買報紙仲有冇 tissue 送？

1861. 你借張 tissue 我吖。

1862. 走嗰陣我哋去咗 toilet。

1863. 屋企 toilet 盞燈燒咗，開唔到。

1864. 要做到最 top 嗰個，一啲都唔容易。

1865. 佢係全公司最 top 嘅經紀之一。

1866. 呢個 topic 好悶。

1867. 不如轉過第二個 topic 啦。

1868. 呢啲係 top-secret，唔好同人講。

1869. 咁 top-secret 都同我講，証明我都有番咁上下地位。

1870. 佢哋依家 total 請三個。

1871. 兩個人 total 萬五蚊，包七晚酒店連四程機票。

1872. 除非 touch wood 咁你發生一件更大嘅事。

1873. 如果呢一日真係咁快到，touch wood，咪當發個夢囉。

1874. 聽講個結局好 touching。

1875. 係咪咁 touching 真係見仁見智。

1876. 都冇人 train 我，係我自己邊做邊學之嘛。

1877. 佢叫我 train 班新人。

1878. 我明白 training 係袋錢落我袋。

1879. 今日嘅 training 朝九晚七，我真係好攰。

1880. 我發現啲 transcript 入面有好多錯字。

1881. 淨係打 transcript 都打咗幾日。

1882. 可唔可以將啲日文字 translate 做中文？

1883. 我想將簡體字 translate 做繁體字。

1884. 我都想試下做 translation。

1885. 讀 translation 要中英文都掂。

1886. 已經完成咗成個 treatment。

1887. 做 treatment 要成四個鐘。

1888. 佢嘅打扮好 trendy。

1889. 佢哋嘅設計一直好 trendy。

1890. 我諗都好難有嘢 trigger 到佢。

1891. 可能你可以 trigger 到佢呢。

1892. 出一次 trip 要用好多錢。

1893. 呢個 trip 我大部份時間都喺酒店。

1894. 你唔 try 下又點知自己唔得？

1895. try 下又唔會蝕底嘅。

1896. 呢件 T-shirt 都我嗰件差唔多。

1897. 可以着 T-shirt 牛仔褲返工。

1898. 佢唔係第一次俾人 turn down㗎喇。

1899. 去醫院傳福音成日俾人 turn down。

1900. turn out 有二百個家長參加。

1901. 雖然佢都幾努力但係 turn out 嘅成績都係差。

1902. 佢話搵唔到 tutor 所以搵我。

1903. 不過，tutor 可能都會教得幾好。

1904. 我唔介意你上 tutorial 嘅時候食嘢。

1905. tutorial 嗰陣啲學生都好乖。

1906. U gym 裝修過之後好靚。
1907. 新研宿咪即係 U gym 後面嗰幢囉。
1908. 之後我就入 U lib 繼續睇我本書。
1909. 我已經喺 U lib 撞過佢至少三次。
1910. ultimately 佢炒我先算。
1911. ultimately 佢可能會叫我返去收拾殘局。
1912. 我暫時 unavailable，不如你遲啲再搵我吖？
1913. 佢未必係 unavailable，知係唔想見你啫。
1914. 電子工程學系係 under 工程學院嘅。
1915. 佢話 under 三百萬像素嘅相機質素唔夠好。
1916. 大家 understand，我都無謂講啦。
1917. 如果你唔 understand 嘅話就要即刻問。
1918. 有啲嘢 understood 就算。
1919. 依啲嘢 understood喋喇。
1920. 只係着住 underwear 喺片場四圍走。
1921. 有時拍戲淨係得 underwear 都要照做喋啦。
1922. 你試下可唔可以 undo 番之前做過嘅嘢？
1923. 原來可以全部 undo，真係好彩。
1924. 有好多嘢都係 unexpected喋啦！
1925. 今次 unexpected 得嚟好過上次好多。
1926. 冇咗呢點 uniqueness，我諗佢都唔會請我。
1927. 個個都差唔多，冇乜 uniqueness 所以揀唔落手。
1928. 本 unix 書好易明喎。
1929. 份工話明要識 unix喋！
1930. 你 un 唔 underatand 都出句聲吖。
1931. 我唔知佢 un 唔 underatand，我估佢聽得明啩。
1932. 你俾我啲資料係咪最 update喋？
1933. 點樣知道幾時要做 update？
1934. 而家我冇得 upgrade，唯有怪自己蠢，成日借功課俾人抄。
1935. 部電腦執到好靚，幾年都可以唔駛 upgrade。
1936. 我會自己 upload 啲課文上網。
1937. 你 upload 唔到就打電話俾我啦。
1938. 我唔知佢今日好 upset 喎。
1939. 佢成日都係好 upset 咁。
1940. 佢問可唔可以唔用 USB？
1941. USB 二點零快過一點零好多。
1942. usually 會咁做，但間中都有例外。
1943. usually 會見晒全部人先至開會討論。
1944. 係咪都反映咗當代人嘅 value？
1945. 會影響到日後對有關事情嘅 value。
1946. 就算咁樣慢啲都唔俾頭先個 van 仔佬賺。
1947. 咁鬼貴梗係唔坐 van 仔喇。
1948. 有冇限最多可以用幾多個 variable？
1949. 我將個 variable 名打錯咗。
1950. 點解隻 VCD 冇聲嘅？
1951. 而家啲 VCD 平到冇人有。
1952. 究竟 vegetarian 食唔食雞蛋喋呢？
1953. 有啲航空公司會提供埋 vegetarian 嘅飛機餐。
1954. 佢用 verb 嘅時候成日都用錯。
1955. 我識嘅 verb 唔多，所以用嚟用去都係嗰幾個。
1956. 最新嗰個 version 好唔穩定。
1957. 每次出新 version 都買咪要用好多錢囉！
1958. 我仲影咗相同 video 留念。
1959. 但開場前都驚餐飽，個 video 好唔穩定。
1960. 呢度夠晒高，仲有馬場 view。
1961. 唔係咁多地方都睇到海景 view。
1962. 整 virus 係咪犯法喋？
1963. 呢輪有好多 virus 所以我都係唔上網住。
1964. 唔知做 visiting professor 一個月有幾多錢呢？
1965. 做得 visiting professor，應該都有番咁上下江湖地位。
1966. 開放日會有好多 visitor 嚟參觀。
1967. 你咁樣答個 visitor 會覺得你好冇自信。
1968. 我都唔知咩係 visual art。
1969. visual art 係咪要好立體喋？
1970. 我一上 VPN 就中咗毒。
1971. 喺屋企去中大啲網頁係咪要用 VPN 喋？
1972. 冇理由大學畢業走去做 waiter喋嘛。
1973. 佢就係唔甘心一世做 waiter。
1974. 佢哋請 waitress 係以貌取人嘅。
1975. 呢度嘅 waitress 係唔會同客人飲酒嘅。
1976. 如果行山有 walkie-talkie，可以慳番唔少電話費。
1977. 一對 walkie-talkie 都係幾百蚊之嘛。
1978. 要有 walkman 先至聽到啲錄音帶喋嘛。
1979. 依家都冇乜人買 walkman喋喇。
1980. 係咪任何圖片都可以做 wallpaper喋？
1981. 有好多張都可以做 wallpaper。
1982. 你有冇收到 warning？
1983. 佢好驚會俾人 warning。
1984. 嗰日我就會俾人記我兩次名，拎 warning letter喋喇。
1985. 收到張 warning letter，我呢排做咩呀，咩都好黑呀。
1986. 佢去咗 washroom。
1987. 唔該 washroom 喺邊度？
1988. 佢隻表有 water proof，跌落水都唔驚。
1989. 化全套眼裝要買 water proof。
1990. 我買咗隻新嘅 web cam。

1991. 隻 web cam 係咪一插入部機度就用到？
1992. 呢個 web site 好似有問題喎。
1993. 我去完呢個 web site 就中咗毒。
1994. 今個學期一共有十四個 week。
1995. 下個星期係 week 幾？
1996. well，或者人真係唔會有完美㗎呢。
1997. 後尾有個男嘅走咗過嚟，well，佢好似好好傾咁囉。
1998. 佢做嘢都唔 well organized 嘅。
1999. 你個故仔都唔 well organized 嘅。
2000. whiskey 好貴，但係有時紅酒可以仲貴。
2001. 我覺得 whiskey 加冰係最好飲嘅。
2002. 我買咗二千蚊 win。
2003. 咁我哋又唔係一定 win 嘅。
2004. 用 window explorer 可以睇到晒你部機有啲咩？
2005. 「檔案總管」係咪即是 window explorer？
2006. 我阿爸連 windows 都唔識用。
2007. windows 唔駛特登學，用用吓就會識。
2008. 你有冇用過 wireless LAN？
2009. 其實喺屋企用 wireless LAN 有冇用呢？
2010. 你有咩想要，寫落張 wish list 度啦。
2011. 我會將我今年未完成到嘅嘢寫落張 wish list 度。
2012. 你 within 一個星期起唔起到貨？
2013. within 一個月做唔做得晒啲嘢？
2014. 我 wonder 我自己適唔適合讀文學。
2015. 我 wonder 佢其實明唔明我講咩？
2016. 我喺 word 入面搵到啲有用嘅嘢。
2017. word 係咪有得對番邊度有嘢改咗？
2018. 係咪咁 work 要試過至知。
2019. 真係 work㗎，你唔信呀？
2020. 雖然未必 workable，但係好有創意。
2021. 你嘅構思咁天馬行空，可能唔 workable。
2022. 一陣個 workshop 你會唔會去？
2023. 開完會之後仲有 workshop。
2024. 佢做啲嘢都唔知 work 唔 work 嘅。
2025. 你有冇試過 work 唔 work㗎？
2026. 今次 written 我十拿九穩。
2027. 過完 written 之後仲有好多次面試。
2028. 照完 X 光，見埋醫生就走得。
2029. 駛唔駛照 X 光呀？
2030. 就算係唔同 year，都係照樣咁讀㗎啦。
2031. year 九九嘅同學幾時會有聚會呀？
2032. 個 year plan 寫得幾好，不過唔知做唔做到。
2033. 其實唔駛依足個 year plan 嚟做。
2034. YMCA 好似有得食自助餐。
2035. 成爲 YMCA 嘅會員，住青年旅舍會有優惠。
2036. 記得 zip 咗份功課先至好交。
2037. 我 zip 咗啲資料然後燒碟俾你。
2038. LCD 雖然貴但係慳位好多。
2039. 俾我就梗係要 LCD 啦，起碼對隻眼好啲。
2040. 其實讀 psycho(psychology)有冇用㗎？
2041. 做心理醫生係咪要讀 psycho(psychology)㗎？
2042. 見一個 client 要成幾個鐘㗎。
2043. 我仲以爲你係佢嘅 client 添。
2044. mean 係六十六分，有五個人滿分。
2045. 我嗰組啲人個個都高分，得我一個低過 mean 咋！
2046. 讀緊大學時已經開始兼職做 DJ。
2047. 做 DJ 講嘢着重咬字準確。
2048. 佢唔單止係 expert，簡直係權威。
2049. 佢呢啲 expert 級人馬又點會睇得起我哋呢啲初學者吖？
2050. 唔知佢真係 realize 唔到，定係扮唔知。
2051. 好彩你哋 realize 到呢個錯處咋！
2052. 佢好緊張，唔止冇 eye-contact，連少少笑容都冇。
2053. 有多啲 eye-contact，先至可以知道對方嘅反應。
2054. 依家連學校飯堂都有 plasma 電視。
2055. 我就唔會用咁多錢買個 plasma 電視喇。
2056. 唔理你係 undergrad 定係中七畢業，都係咁多人工。
2057. 如果我 undergrad 嗰陣勤力啲就好啦。
2058. 你如果冇 experience，咪由頭開始學囉。
2059. 我諗佢唔會當我有 experience。
2060. 每做完一個 step 都檢查一次。
2061. 佢應該係做漏咗其中幾個 step。
2062. 織一件 V 領冷衫會唔會好難？
2063. 佢好開心，仲舉起 V 字手勢添。
2064. 點樣可以 man 啲呀？
2065. 佢靚，但靚得嚟好 man。
2066. 唔想咁後生就俾人叫 uncle。
2067. 我要證明俾你睇，我唔係一個 uncle。
2068. 去 pub 又唔係啲唔見得光嘅嘢。
2069. 食完飯之後再去咗兩間 pub。
2070. 我 finish 咗三份功課，終於可以瞓喇。
2071. 無論幾點都好，你一 finish 就通知我。
2072. 你 register 咗雙週會未？
2073. 知唔知道有幾多人已經 register 咗？
2074. 我唔識 sketch 圖。
2075. 如果冇得用電腦，自己點識 sketch 呢幅圖喎。
2076. 你睇清楚啲 spelling 冇錯先至好交喎。
2077. 錯得最多就係 spelling，反而文法冇乜問題。
2078. 跟住下晝去咗 U2 訂制服。

2079. 而家 U2 大減價，一件恤衫都只係幾十蚊。
2080. 你以前有冇試過 unseen 嘅默書？
2081. 如果內容係未教過嘅，都可以算係 unseen。
2082. 呢度除咗 Prada，仲有唔少歐洲牌子賣。
2083. 嗰啲人用假 Prada 同我無關。
2084. 而家啲 game boy 全部都係彩色㗎。
2085. 我想買部 game boy 送俾細佬。
2086. 攞住部未出街嘅 Sony Ericsson 四圍捉人影相。
2087. 呢部 Sony Ericsson 有好多新功能。
2088. 我一向都係 Nokia 嘅忠實擁躉。
2089. 我鍾意用 Nokia 因爲有得換殼。
2090. 陪同父親出席上海 Gucci 新店開幕。
2091. 唔少廿幾歲住公屋嘅女仔，一樣有十幾個 Gucci 手袋。
2092. 一個人喺 Starbuck 坐咗成個鐘。
2093. 我好鍾意去 Starbuck 睇書。
2094. 今場有一定 chance 捧盃。
2095. 照計今次有一定 chance 贏馬。
2096. 我覺得上網俾錢好唔安全，所以都係寄 cheque。
2097. 收 cheque 都唔知過唔過到數，都係現金穩陣啲。
2098. 佢叫我幫佢整個 power point。
2099. 上堂用 power point 要帶電腦，好鬼麻煩。
2100. 讀埋 doctor 先至出嚟做嘢會唔會太遲？
2101. 佢同 doctor 張係小學同學。
2102. 你部相機可以 zoom 到幾多倍？
2103. zoom 到咁近，塊面有暗瘡到影到。
2104. 呢個係我個 wife。
2105. 我 wife 同我係大學同學。
2106. 佢有時會同佢 husband 一齊返工。
2107. 我 husband 都好鍾意跳舞。
2108. 聽到呢句說話即刻 warm 晒。
2109. 俾人一個好 warm 嘅感覺。
2110. 第一次行 catwalk 心情自然特別緊張。
2111. 爲一個時裝品牌擔任模特兒行 catwalk。
2112. 佢成日去 wet，冇心機做嘢。
2113. 得閒去 wet 番一晚半晚，都唔係好過份啫。
2114. 你諗住着呢件 see-through 去飲？
2115. 我哋唔打底玩濕身，玩 see-through！
2116. 從來冇因爲工作高底問題而影響佢同 Sammi。
2117. 問佢係咪約咗 Sammi 一齊買嘢。
2118. 舊年 Twins 同容祖兒一齊獲得傳媒大獎。
2119. Twins 早前聯同一啲歌手喺深圳演出。
2120. 已爲人父嘅 Eason 魅力依然。
2121. 心情大好嘅 Eason 自然係來者不拒。
2122. Maggie 喺書入面談及佢嘅演藝事業。
2123. 記者問 Maggie 係咪經常去蘭桂坊消遣。
2124. Edison 係演員，點會同劇組人員打交？
2125. 最近忙於錄音嘅 Edison，每日只係瞓得三、四個鐘。
2126. Kelly 更即場與啤梨合唱一曲。
2127. Kelly 嘅出現吸引唔少圍觀者。
2128. 經過上次嘅教訓後，Steven 而家提起桃花亦感怕怕。
2129. 一度暴肥嘅 Steven 前晚明顯收身。
2130. Leon 諗都冇諗就話冇難度。
2131. 新碟銷量好，Leon 係咪可以鬆一口氣？
2132. Joey 亦獲贈一條手鏈。
2133. 有雜誌又再重提 Joey 整容一事。
2134. 好多人都叫 Jacky，我點知你講緊邊個。
2135. 放咗工就同 Jacky 去旺角睇戲。
2136. 約咗 Eric 星期六去行山。
2137. 唔知 Eric 會送咩生日禮物俾我呢。
2138. 今日 Michael 冇去上堂。
2139. 下星期 Michael 會去日本旅行。
2140. William 今朝返工遲到。
2141. 一陣 William 會上嚟，你同我招呼住佢先。
2142. 頭先 Tony 打電話搵你。
2143. 好彩 Tony 唔知道呢件事。
2144. 設計得好差，一啲都唔 user friendly。
2145. 整得 user friendly 啲，就人人都識用。
2146. 着住黑色 tube top 同牛仔褲。
2147. 一身 Tube Top、短裙仔加四吋高踭鞋打扮。
2148. 呼朋叫伴一齊開 rave party。
2149. 結果，佢成爲 rave party 嘅發燒友。
2150. 佢同人講自己係 single。
2151. 大家都係 single，咪當識多個朋友囉。
2152. 我諗住 gather 四五個人就夠。
2153. 可能 gather 到十幾個人。
2154. 今年所有 counter 都會喺六樓。
2155. 我哋唔擺 counter，但係會搵好多人派傳單。
2156. 見到人唔開心，自己都會 down down 地。
2157. 我一直以來都想做啲有 public relaions 性質嘅工。
2158. 但始終 first impression 唔係咁好囉。
2159. 漸漸地本來已經狹小嘅 PR office 就變咗一個貨倉。
2160. 成日都低低 BB，白白痴痴。
2161. 上幾多個 course 幾多個 workshop 都冇用。
2162. 一心諗住 lunch 同 Eric 去食飯。
2163. 而 ICQ 亦都收到好多 message。
2164. 同佢哋 chat，聽佢哋嘅 story。

2165. 不過，我又想佢哋盡力完成 oral 呢 part。
2166. 打羽毛球打得好開心，雖然 partner 今日 absent。
2167. sorry，我唔係你啲男仔 friend。
2168. 聽日要 present，始終覺得我個 idea 唔夠其他人好。
2169. 其實你返嚟係 enjoy 個 holiday㗎嘛。
2170. 今日發現之前 reg 咗嘅《當代消費文化》cancel 咗。
2171. 佢啲畫真係好靚，畫得好 detail，好有 plan 咁畫。
2172. 如果你真係 delete 咗嘅話，就喺 ICQ 問我攞返啦。
2173. 班小朋友 feel 到我哋嘅 heart。
2174. 之後我哋就去咗 book K 房。
2175. 我個 friend 話交通安全隊同 red cross 互助互愛喎。
2176. 前幾日頂唔順，send 咗封 email 向佢訴苦。
2177. 不過有啲 show off 到嘅 point 冇講到。
2178. 頭先 ICQ 有個女仔想 add 我。
2179. 佢係做 sales 嘅，係 sell 波鞋呀手表呀嗰啲嘢嘅。
2180. 佢哋三 pair 今晚都好 happy 好恩愛。
2181. 不過都有一班 friend 一齊 BBQ。
2182. 嗰條就係部 MP3 用嘅 cable 喇。
2183. 但 internship 的確給予我一定嘅 experience 同優勢。
2184. 原來小時候嘅我已經有 potential 做一個好 leader。
2185. 今日呢個時間 U lib 當然 full 晒啦。
2186. 但每次 test 同埋 exam，都總會有事發生。
2187. 距離 last 一科嘅 marketing，仲有三日。
2188. 若果我冇 repeat 過 A-Level，我唔會學識點解要上進。
2189. Eric 急 call，叫我陪佢出旺角買鞋。
2190. 個 manager 好 nice，講咗好多嘢我哋聽。
2191. 返到去，send 咗個 SMS 俾阿強。
2192. 返到屋企 check 埋 email 先瞓。
2193. 每 sell 到一部 notebook 就多五十蚊。
2194. 仲 mean 嘅就係 training 都要用兩、三日。
2195. 今次係我第一次喺 interview 嘗試做 translation。
2196. 我竟然收到 call，話我聽日可以 second in。
2197. 佢哋同我講吓份 job 嘅 detail。
2198. 同埋 remind 我份 job 可能會有咩問題。
2199. 我同 manager 將百幾箱嘢搬入 office。
2200. 我同 Joey 一早就 stand by 要衝出去送花。
2201. 做 leader 同 captain 嘅，咪有啲紳士風度囉。
2202. 第二樣要做嘅嘢就係 upload 新歌落 MP3 機度。
2203. 大家 happy 完後我就返 hall。
2204. 之後去飲嘢，係我一向都好 buy 嘅 Starbuck。
2205. 一放 break 就走埋一齊傾計，好有中學 feel。
2206. 之後走去數 display 嘅 product，一共一千六百幾件。
2207. 俾個學生阻一阻我，我就同唔到 friend 食 lunch，唯有下次再約。
2208. 我唔係好信，硬係覺得佢是但搵個 reason 去 quit。
2209. 佢咁講已經 confirm 咗我 Maggie 係奸險小人。
2210. 我冇機會 quit 份 job，所以我會同佢死過。
2211. 等我 list 個 schedule 出嚟，等大家約我都方便啲。
2212. 我個阿 head 同 supervisor 喺房度傾緊嘢。
2213. 過埋今日，我 suppose 可以落鋪做 sales。
2214. 佢好強調我哋唔係 sales，所以我哋冇 quota。
2215. 間公司 in 咗三百幾人，但俾咗唔夠十個 offer 出嚟。
2216. 我最鍾意嘅 activity 就係 BBQ。
2217. 我唔怕熱、唔怕 dirty 又唔怕曬，所以 BBQ 最啱我。
2218. 大家生日一定要去 neway 唱 K，我覺得好抵。
2219. 我會 keep 住 remind 大家送禮物㗎喇。
2220. 原來要 follow 佢嘅 instruction 先有飯食。
2221. 又係一個 surprise，因爲我冇 expect 佢會送嘢。
2222. 打死都要今日上，因爲淨係得呢個 tutorial 係 William 親自教授。
2223. 一大班 friend 影相，自己張相喺朋友部 DC 裡面，都唔係好出奇。
2224. 如果 performance 好，就好快可以上 office。
2225. 我 accept 咗個 offer.
2226. 因爲佢俾嘅 information，我先過到 interview。
2227. 上兩個星期，我連續 miss 咗三個 interview。
2228. 個 interview 好短，around 十五分鐘。
2229. 我諗成個 trip 三千零蚊就 OK 喇。
2230. 我都唔知係大學 gathering、定係中學 gathering。
2231. 上到三樓，除咗見到 Eric，仲有個叫 Steven 嘅同我傾計。
2232. Miss 話做翻譯會有好多限制，唔可以成

日都咁 free 喎。

2233. 上次份 notes 得啲 heading，今次嗰份就厚好多。
2234. 所以 Leon 就打電話俾個叫 Kelly 嘅人報名。
2235. 佢將個 test 取消，將一半分撥入 class participation。
2236. 最後嗰堂 tutorial，派咗啲語言學嘅 notes。
2237. Last 嗰晚食飯佢仲同我哋啲團友隊啤，個個都好 high 呀。
2238. 六點幾嗰陣走咗去 neway 唱 K。
2239. 多咗好多 waitress，不過啲 quality 都係麻麻地。
2240. 佢可以咁 happy 係因爲佢多 friend。
2241. Miss 都話我哋好 disappointed。
2242. 不過我喺個 course 度都玩得好 happy。
2243. 入場睇戲啦，去 toilet 嗰陣先 funny。
2244. 雖然有 dead-air，但傾得都幾 happy。
2245. 今日 lunch 我哋又去咗 Pizza Hut 食嘢。
2246. 上個禮拜個阿 sir 咁正，呢個禮拜換番個 Miss 就好唔掂。
2247. 儲錢個 plan 就 OK 囉，不過保險就仲要諗諗。
2248. 俾番個 file 佢，都係諗住等佢唔駛唔見咗啲 notes 咁慘。
2249. 點知今年居然連 lunch 都無得供應，cheap 到爆。
2250. 睇嚟呢個 trip 都唔會同佢哋玩得 happy。
2251. 我明明 format 完，但都唔知點解仲有 virus。
2252. 食完 lunch 之後就開始我哋嘅 shopping。
2253. 最後個 result 都係我所 expect 嘅。
2254. 撞到 Sammi，佢返嚟做 board。
2255. 英文 oral 睇 demo，但冇幾耐又瞓咗。

Appendix C
Usage of Speech Data in CUMIX

CUMIX TRAINING DATA		
Speaker ID	PHONE_LBD01	PHONE_LBD02, SYL_LBD, Acoustic Model Set B and C, Overall Performance (no LBD), Overall Performance (with LBD), BILINGUAL_LBD
	CS	All Data
F01	Development - DEV_PHONE	Training of AM
F02	Development - DEV_PHONE	Training of AM
F03	Development - DEV_PHONE	Training of AM
F04	True-code switch test - CS_TRUE	Training of AM
F05	True-code switch test - CS_TRUE	Training of AM
F06	True-code switch test - CS_TRUE	Training of AM
F07		Training of AM
F08		Training of AM
F09		Training of AM
F10		Training of AM
F11		Training of AM
F12		Training of AM
F13		Training of AM
F14		Training of AM
F15		Training of AM
F16		Training of AM
F17		Training of AM
F18		Training of AM
F19		Training of AM
F20		Training of AM
M01		Training of AM
M02		Training of AM
M03		Training of AM
M04		Training of AM
M05		Training of AM
M06		Training of AM
M07		Training of AM
M08		Training of AM
M09		Training of AM
M10		Training of AM
M11		Training of AM
M12		Training of AM
M13		Training of AM
M14		Training of AM
M15		Training of AM
M16		Training of AM
M17		Training of AM
M18		Training of AM
M19		Training of AM
M20		Training of AM

CUMIX TESTING SET			
Speaker ID	PHONE_LBD01	PHONE_LBD02, SYL_LBD, BILINGUAL_LBD	PTT
	CS	CS	CS
F21		Development - DEV_CS	
F22		Development - DEV_CS	
F23		Development - DEV_CS	
F24		Development - DEV_CS	
F25		Development - DEV_CS	
F26	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F27	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F28	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F29	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F30	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F31	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F32	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F33	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F34	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F35	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F36	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F37	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F38	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F39	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
F40	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M21		Development - DEV_CS	
M22		Development - DEV_CS	
M23		Development - DEV_CS	
M24		Development - DEV_CS	
M25		Development - DEV_CS	
M26	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M27	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M28	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M29	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M30	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M31	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M32	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M33	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M34	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M35	Borrowing Test - CS_BORROW	Testing - TEST_CS	Testing - TEST_CS
M36			
M37			
M38			
M39			
M40			

CUMIX TESTING SET					
Speaker ID	Acoustic Model		Overall Performance (no LBD)		Overall Performance (with LBD)
	CS	CAN	CS	CAN	CS
F21			Development - DEV_CS		
F22			Development - DEV_CS		
F23			Development - DEV_CS		
F24			Development - DEV_CS		
F25			Development - DEV_CS		
F26	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F27	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F28	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F29	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F30	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F31	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F32	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F33	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F34	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F35	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F36	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F37	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F38	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F39	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
F40	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M21			Development - DEV_CS		
M22			Development - DEV_CS		
M23			Development - DEV_CS		
M24			Development - DEV_CS		
M25			Development - DEV_CS		
M26	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M27	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M28	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M29	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M30	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M31	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M32	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M33	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M34	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M35	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS	Testing - TEST_CAN	Testing - TEST_CS
M36					
M37					
M38					
M39					
M40					

Remarks:

Testing data from speakers M36 – M40 were not utilized in any experiments since the verification process was not yet completed when the experiments were done.