The State of the Art in Human-computer Speech-based Interface Technologies

LEE Tan BSc MPhil PhD MIEEE
Faculty of Engineering, The Chinese University of Hong Kong

Helen MENG SB SM PhD MIEEE
Faculty of Engineering, The Chinese University of Hong Kong

W K LO BEng MPhil PhD MIEEE
Faculty of Engineering, The Chinese University of Hong Kong

P C CHING BEng(Hons) PhD FHKIE FIEE SrMIEEE
Faculty of Engineering, The Chinese University of Hong Kong

This paper gives an overview of the state-of-the-art human-computer interface technologies based on automatic speech recognition and text-to-speech synthesis. It also describes the recent R&D activities at the Chinese University of Hong Kong (CUHK) in this challenging field. Speech is the most convenient and natural means of communication among human beings. By enabling the computer to listen and speak, speech technologies have empowered many important applications that improve our quality of life. Hong Kong is a trilingual society where people speak Cantonese, Putonghua as well as English. While a great deal of efforts have been spent on speech and language processing for English and Putonghua in Western countries and China, the Speech Research Group at CUHK is well known to be one of the few pioneers who initiated extensive study on Cantonese-focused speech technologies. Cantonese is a major Chinese dialect spoken by over 70 million people in South China and Hong Kong. We shall introduce the various speech recognition and speech generation technologies developed by our team during the past few years. These component technologies and related data resources are made publicly available with an aim to lower the entry barrier of any research institutes or private enterprises who wish to embark on this exciting area of research and development. Furthermore, a number of emerging technologies that integrate with speech-based interface and lead to many novel and innovative applications will be presented. Examples of such applications include a spoken document retrieval system and two spoken dialog systems for financial information inquiry. This article is intended to serve as a comprehensive reference for technology developers to explore the potential of speech technologies in the next-generation information and communication systems.

Keywords: *Human Computer Interface, Automatic Speech Recognition, Text-to-Speech Synthesis, Cantonese Dialect, Spoken Document Retrieval, Acoustic and Lanuage Modelling*

1. Introduction

Speech is an important means of communication among people, and the advent of speech technologies has extended the use of speech to communication between humans and computers. These technologies aim at making the use of computers more convenient and efficient. Speech technologies have also brought about many important applications that improve people lives. Examples include the command and control of machines by speech, computer dictation of spoken sentences, interactive speech-based interfaces with computers, and technological aids for people with visual or speech impairment.

Hong Kong is a trilingual society where people speak Cantonese, Mandarin, and English. To cater for human-computer communication in Hong Kong, these three languages must all be accommodated. There has been much research conducted in Europe, America and China on the two of the languages, English and Mandarin. The Chinese University of Hong Kong has pioneered in research and development in Cantonese speech technologies over a decade ago. Cantonese is a major Chinese dialect, predominant in Hong Kong, South China and many overseas Chinese communities. Our research team at CUHK has worked on both speech input technologies (mainly speech recognition that transforms input speech into text), speech output technologies (mainly text-to-speech synthesis that transform text into synthetic speech) and related technologies such as Chinese natural language processing and information retrieval. We have also integrated these component technologies into end-to-end systems such as spoken dialog systems and audio search engines.

In this paper, we will give an overview of the state-of-the-art human-computer speech-based interface technologies. First, we will outline some basic characteristics of Chinese and, in particular, the Cantonese dialect. We will then introduce the key component technologies that enable speech-based interactions between human and computer. In particular, we will elaborate on our focused effort in developing such technologies at CUHK for Cantonese. Integration of these component technologies have led to the emergence of several sophisticated systems surveyed in this paper. In addition, we will describe the rich set of speech-based technologies and resources developed in-house and have been made available to the public in order to lower the entry barrier of academic and industrial bodies who wish to embark on this exciting area of research and development.

2. Characteristics of Chinese

Computer processing of the Chinese language is difficult because of the number of homonyms and the lack of clarity of word definition. In Chinese, there are two types of characters: the simplified characters used in mainland China and the traditional characters used in Hong Kong and Taiwan. There are altogether around 6,000 simplified characters and 10,000 traditional characters. Within these characters, there are only 1,800 different syllable pronunciations in Cantonese which implies that many characters share the same pronunciation. Furthermore, many characters have more than one pronunciation and the choice depends on the context. For example, based on the statistics of the CULEX pronunciation lexicon [1], it is found that there are around 27,000 two-character words sharing around 22,000 different pronunciations. Table 2a shows a summary of the statistics of homophones in CULEX. Determination of the correct word based on the given pronunciation requires additional information about the context. Furthermore, homographs are also common in Cantonese. For example, there are five different pronunciations for the character 行 – /haang4/, /han4/, /hong4/, /hong2/, /hang6/. Additional information is needed to determine the appropriate pronunciation based on the context.

Word length (No of Characters)	Word Count	Pronunciation Count	Average no of Homophones
1	~10,000	~600	16.67
2	27,057	22,108	1.22
3	4,720	4,682	1.01
4	3,579	3,525	1.02
5	318	315	1.01

Table 2a – Homophone Statistics in the Cantonese Pronunciation Lexicon (CULEX)

In Chinese textual materials, there is no explicit delimiter placed at the boundaries of words. Chinese text is stored and presented as a sequence of characters. Word segmentation is required to locate word boundaries. However, word segmentation is an uncertain process. The example in Figure 2a shows that a Chinese character sequence can be segmented into three different word sequences. These sequences are all syntactically valid and semantically meaningful. The correct choice can only be determined with the help of the meanings of the adjacent sentences, if any. This kind of ambiguity has imposed an added degree of difficulty on Chinese language processing.

character sequence	**這一晚會如常舉行**
segmentation 1	**這一　晚會　如常　舉行**
(meaning)	**(This banquet will be held as usual)**
segmentation 2	**這一晚　會　如常　舉行**
(meaning)	**(Tonight an event will be held as usual)**
segmentation 3	**這一　晚會　如　常舉行**
(meaning)	**(If this banquet is held very often)**

Figure 2a – Illustration of Ambiguity in Word Segmentation of Chinese Character Sequence

From a phonetic point of view, Cantonese is a dialect from South China that is used in Hong Kong, Macau, the Guangzhou area, and in overseas communities. As a Chinese dialect, Cantonese is similar to Mandarin in that they are both monosyllabic tonal languages. 'Monosyllabic' means that every Chinese character is pronounced as a syllable. 'Tonal' means that the tone of the syllable carries lexical meaning. Cantonese also has a rich phonetic and tonal structure that imposes additional difficulties for computer processing.

Cantonese syllables have a similar structure (see Figure 2b) to Mandarin syllables. All syllables can be segmented into an optional syllable initial and a syllable final [2]. Phonetically, Cantonese initials can be consonants or semi-vowels. A Cantonese final consists of a vowel and may be followed

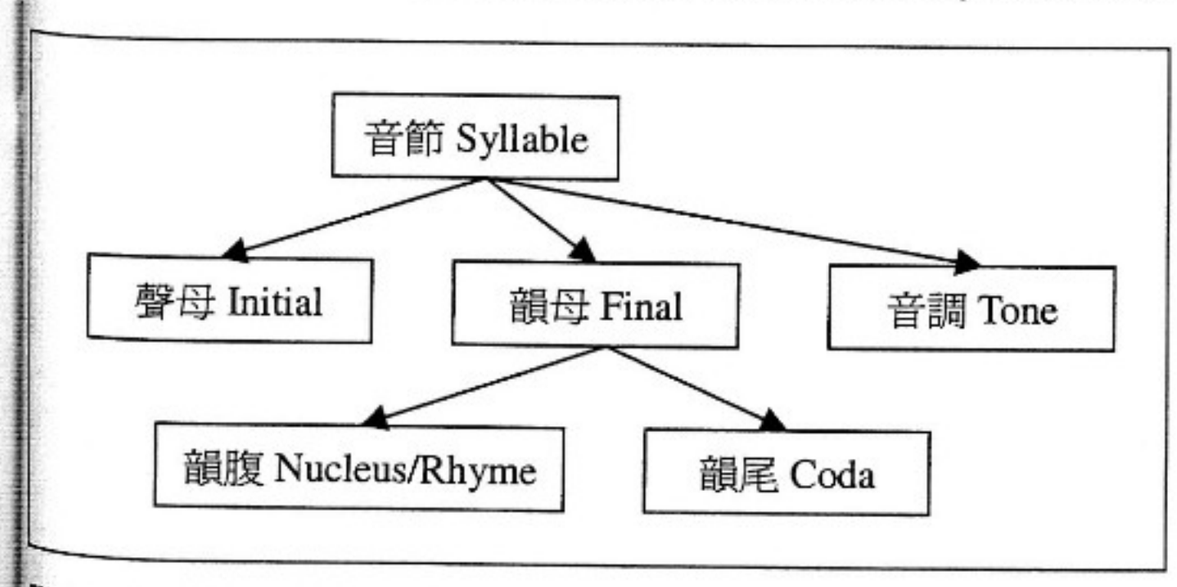


Figure 2b – Phonological Structure of Cantonese Syllables

Initial	Example	Character	Remark
	/aa1/	呀	Null initial
b	/baa1/	巴	Bilabial unaspirated unvoiced plosive
p	/paa3/	怕	Bilabial aspirated unvoiced plosive
m	/maa1/	媽	Bilabial nasal
f	/faa1/	花	Labial-dental unvoiced fricative
d	/daa2/	打	Alveolar unaspirated unvoiced plosive
t	/taa1/	他	Alveolar aspirated unvoiced plosive
n	/naa5/	那	Alveolar nasal
l	/laa1/	啦	Lateral
z	/zaa1/	渣	Alveolar unaspirated unvoiced affricative
c	/caa1/	查	Alveolar aspirated unvoiced affricative
s	/saa1/	沙	Alveolar unvoiced aspirated fricative
g	/gaa1/	加	Velar unaspirated unvoiced plosive
k	/kaa1/	卡	Velar aspirated unvoiced plosive
ng	/ngaa4/	牙	Velar nasal
gw	/gwaa1/	瓜	Velar unaspirated unvoiced lip-rounded plosive
kw	/kwaa1/	誇	Velar aspirated unvoiced lip-rounded plosive
w	/waa1/	蛙	Lip-rounded semi-vowel
j	/jaa5/	也	Alveolar semi-vowel
h	/haa1/	哈	Vocal fricative

Table 2b – List of the 19 Cantonese Initials Used in Hong Kong Pronunciation is Given in the Transcription Scheme Proposed by the Linguistic Society of Hong Kong - LSHK [3])

Nucleus	Coda								
	Null	Vowel		Nasal			Stop		
		i	u	m	N	ng	p	t	k
				m		ng			
aa	aa	aai	aau	aam	aan	aang	aap	aat	aak
a		ai	au	am	an	ang	ap	at	ak
i	i		iu	im	in	ing	ip	it	ik
yu	yu				yun			yut	
u	u	ui			un	ung		ut	uk
e	e	ei	eu			eng			ek
oe/eo	oe	eoi			eon	oeng		eot	oek
o	o	oi	ou		on	ong		ot	ok

Table 2c – List of the 54 Cantonese Finals Used in Hong Kong (in LSHK)

by a consonant coda. Tables 2b and 2c show all Cantonese initials and finals respectively.

Cantonese syllables also carry a lexical tone. A base syllable, in Cantonese, refers to the combination of an initial (optional) with a final. When a lexical tone is assigned to a base syllable, a tonal syllable is formed. There are six different tones [3] in Cantonese that are defined by the levels as well as the pitch profiles over the course of syllables. Figure 2c shows a generalised schematic of the six different tones in Cantonese.

Cantonese has the so-called entering tone syllables that always end with a stop coda (/p/, /t/, or /k/). They can only carry tones 1, 2, 3 and 6. Entering tone syllables are also characterised by a shorter duration when compared with non-entering tone syllables. Phonetically, Cantonese has the characteristic that all plosives and fricatives are unvoiced.[1] Syllables

[1] In Cantonese, the plosives /b/, /p/, /d/, t/, /g/, /k/ and fricatives /s/, /f/ are all unvoiced. The group /b/, /d/, /g/ and the group /p/, /t/, /k/ are differentiated by aspiration only. In contrast, the group /p/, /t/, /k/ in English is unvoiced and aspirated. The group /b/, /d/, /g/ in English is voiced and unaspirated.

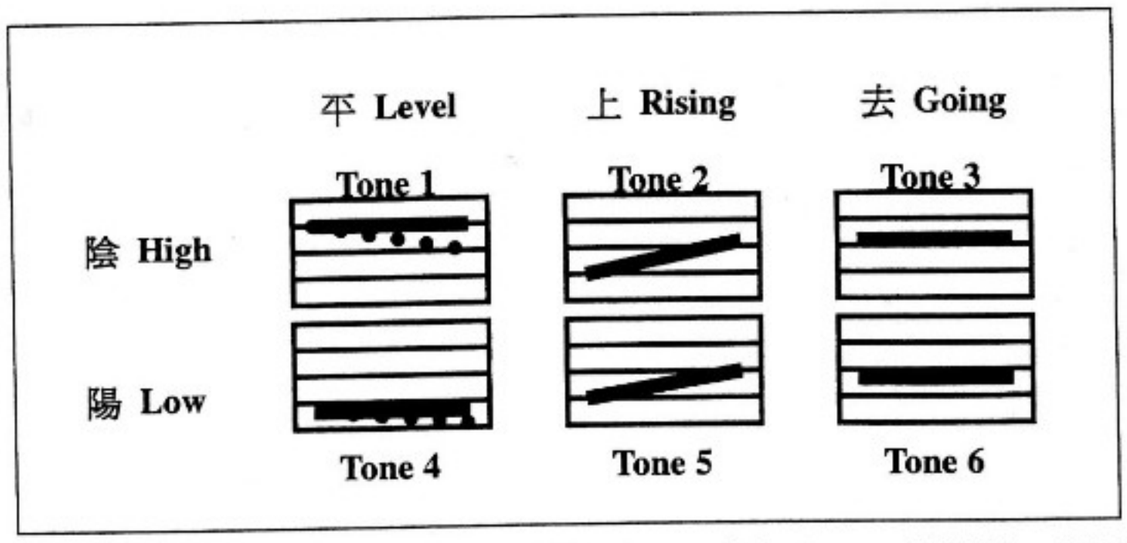


Figure 2c – The Six Different Tonal Structures of Cantonese Syllables. Bold Lines Show the Approximated Pitch Profiles of Syllables Carrying Different Tones. Dotted Lines Show Possible Alternative Profiles

	Mandarin	Cantonese
No of initials	22	19
No of finals	35	54
No of tones	5	6
Approx no of base syllables	400	600
Approx no of tonal syllables	1,200	1,800

Table 2d – Comparison of Some Statistics for Cantonese and Mandarin

carrying aspirated and unaspirated versions of a consonant have different meanings. In addition, there is no retroflex in Cantonese initials.

Among the large number of possible combinations of initials, finals and tones, there are approximately 1,800 tonal syllables, or around 600 base syllables, in Cantonese. Table 2d compares some basic characteristics of Cantonese and Mandarin.

3. Overview of speech recognition by computers

Speech recognition is a technology that transforms input speech into text. A speech recogniser is essentially a pattern recogniser for phonetic units (ie the collection of sounds in a language) associated with the recogniser's vocabulary (ie words that are in the computer's knowledge base). The recogniser may only be able to recognise isolated words. Alternatively, it may also be able to recognise continuous sentences. These are constrained by a grammar that defines the legitimate combinations of vocabulary items into speech.

In a computerised speech recognition system, the three components of phonetic units, vocabulary and grammar are modeled mathematically. For a given task of recognition, a vocabulary of recognisable items are first defined. Identification of the phonetic units in the input speech is handled by the acoustic model. Legitimate sentence structures defined by the grammar are handled by the language model. Figure 3a shows the general components of a speech recognition system.

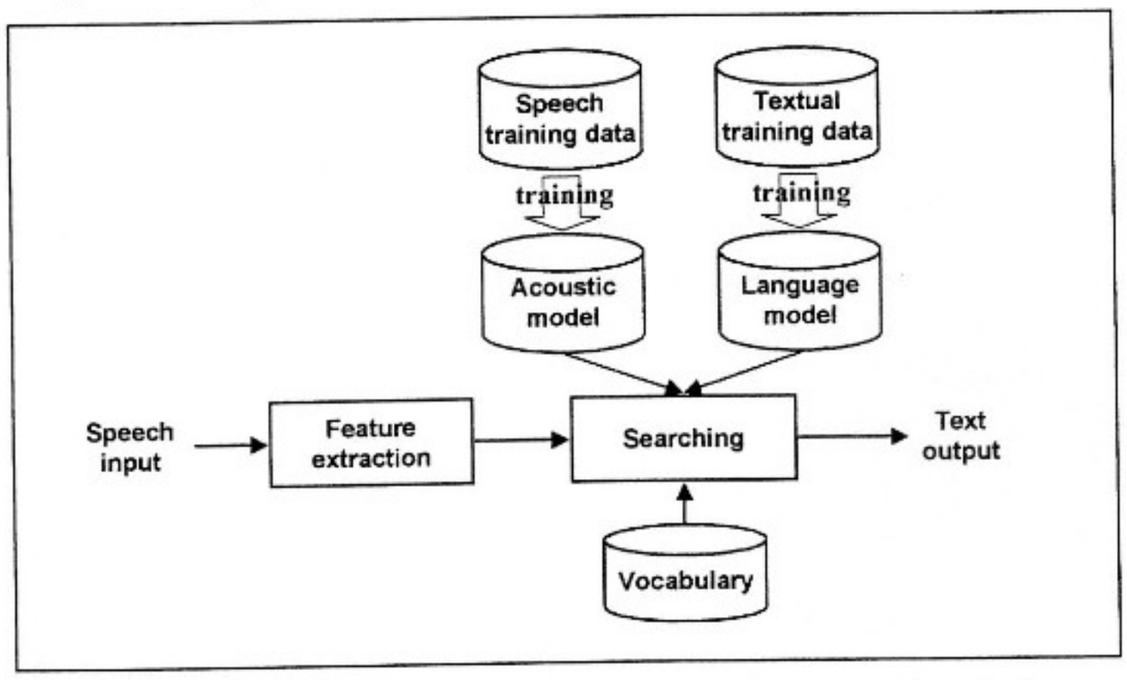


Figure 3a – Functional Block Diagram of a Speech Recognition System. The Training Processes for Acoustic and Language Models are Performed Once only until there is a Change in the Recognition Task

The Vocabulary

A recogniser's vocabulary constrains the words that a speech recognition system can recognise in the same way that a vocabulary determines the words a person could understand. In a speech recognition system, the size of the vocabulary also determines the complexity of the recognition problem. For example, there are usually fewer entries in a vocabulary for navigational command and control tasks. This kind of recognition systems are usually known as isolated word recognisers, where command words are issued by users as isolated words. However, for a human-computer dialog system or a voice dictation system, a mid- to large-size vocabulary is needed. The recogniser should also be capable of handling continuous streams of spoken words. These speech recognition systems are usually known as mid- or large- vocabulary continuous speech recognisers (the latter often referred to as LVCSR).

Acoustic Modeling

Computerised speech recognition is usually treated as a pattern recognition problem. Sample data is first collected and representative templates are derived or abstracted from this data during the training process. In the recognition process, the unknown input data is compared with the pre-stored templates to determine the best match as the output.

In the training and recognition process, representative features are always extracted from the speech data for processing, instead of using the raw speech data directly. Over the past decades, there have been many features proposed for speech recognition, eg linear predictive coefficients, filter bank coefficients, Mel-frequency cepstral coefficients. In general, a speech signal is assumed to be short-time stationary, and these features are extracted from the speech signal at regular time intervals. As an example, Figure 3b shows the process of feature extraction from the speech signal using overlapping windows (eg 25 ms) at a fixed interval (eg 10 ms).

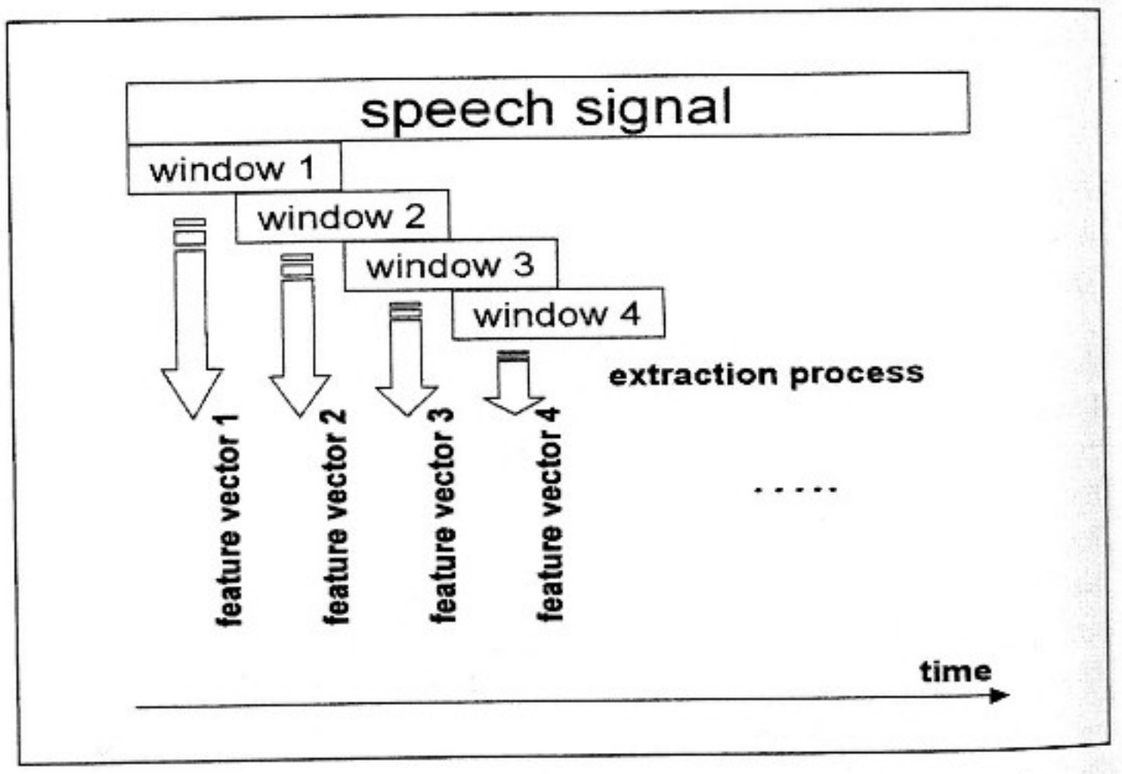


Figure 3b – Feature Extraction from Speech Signal Using Overlapping Windows at Fixed Time Intervals

Based on the features extracted from the speech data, acoustic models of the speech units are trained to create models for use during recognition. As a simple example, the mean of the feature vectors extracted can be used as the templates. However, the same speech unit can have large variations in duration when uttered by different people or even when uttered by the same speaker at a different time. Therefore, dynamic programming approaches are needed for aligning the speech signals of different speech units for both the training and recognition phase. Figure 3c is an illustration of the dynamic characteristics of speech signals.

Dynamic time warping (DTW) is a dynamic programming technique that warps the time scale so as to attain the best possible alignment between two sequences that have different lengths. This approach is most commonly used in isolated word recognisers and also in systems with limited computational power (eg embedded systems). Another commonly used approach for acoustic modeling in speech recognition is the hidden Markov model (HMM). Figure 3d illustrate the typical structure of an HMM. The

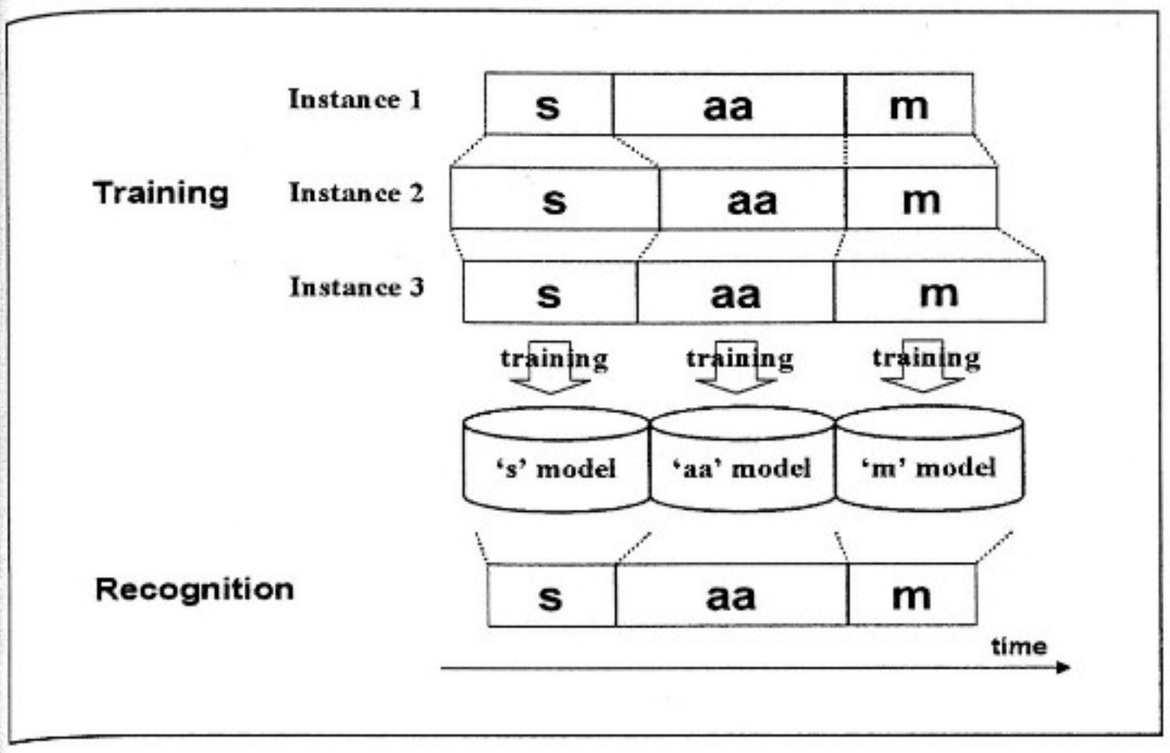


Figure 3c – An Illustration of the Dynamic Characteristics of Speech Signals for Different Instances of the Same Utterance and the Alignment Required for the Training and Recognition Phase

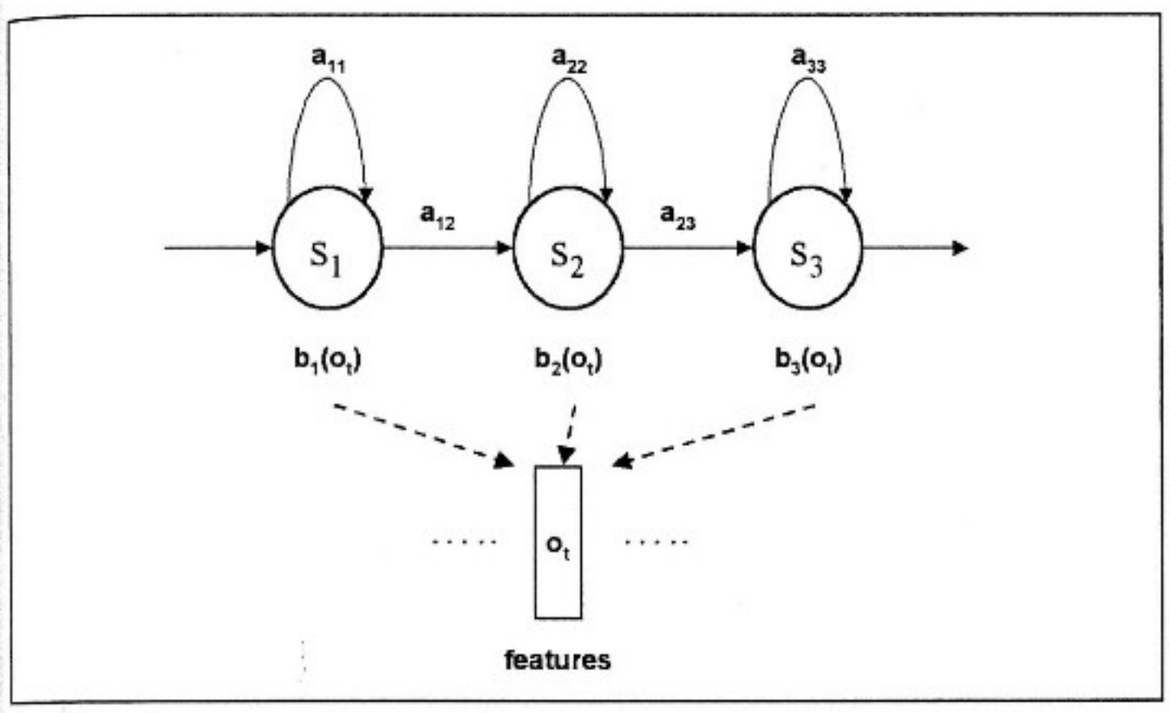


Figure 3d – Typical HMM Used for Acoustic Modeling in Speech Recognition

HMM is a statistical finite-state machine that captures the static and dynamic spectral properties of a speech unit.

A state in the HMM essentially represents certain acoustic event (eg the start, the end, the transition, or the stationary part of a speech sound). Transition from one state to another is a probabilistic event that is governed by the transition probability (a_{ij}) and also the probability of observing the features at time t at the particular state ($b_j(o_t)$). Application of the HMM for speech recognition has been studied for a long time and efficient techniques for maximum likelihood estimation of the parameters (a_{ij}, $b_j(o_t)$) are available (eg expectation maximisation).

There are several important considerations in the choice of speech units to be modeled. The units should be suitable for the targeted language and modeling method. Consideration should also be given to the need for sufficient training data for estimating model parameters. In a small vocabulary recognition task, word is the natural size of recognition units. When the vocabulary size increases, the amount of training data for each word may become insufficient. In this case, sub-word units are used. In speech recognition systems for English, phonemes[2] are commonly used sub-word units. There are a finite number of phonemes that can make up all words. For Chinese languages, including Mandarin and Cantonese, syllable initials and finals are used because initials and finals make up a finite set of phonological units that can fully represent the language. When sub-word units are used in a speech recognition system, entries in the defined recogniser vocabulary will then be defined as a concatenation of these units.

Language Modeling

The language model in a speech recogniser defines how words in the vocabulary can be combined. The language model can either be deterministic or probabilistic. A deterministic language model is as simple as a list of words that may feasibly be expected to appear as the input of the recogniser. More complicated language constraints may also be added through the use of a finite-state grammar. Figure 3e shows an example of a simple finite-state grammar.

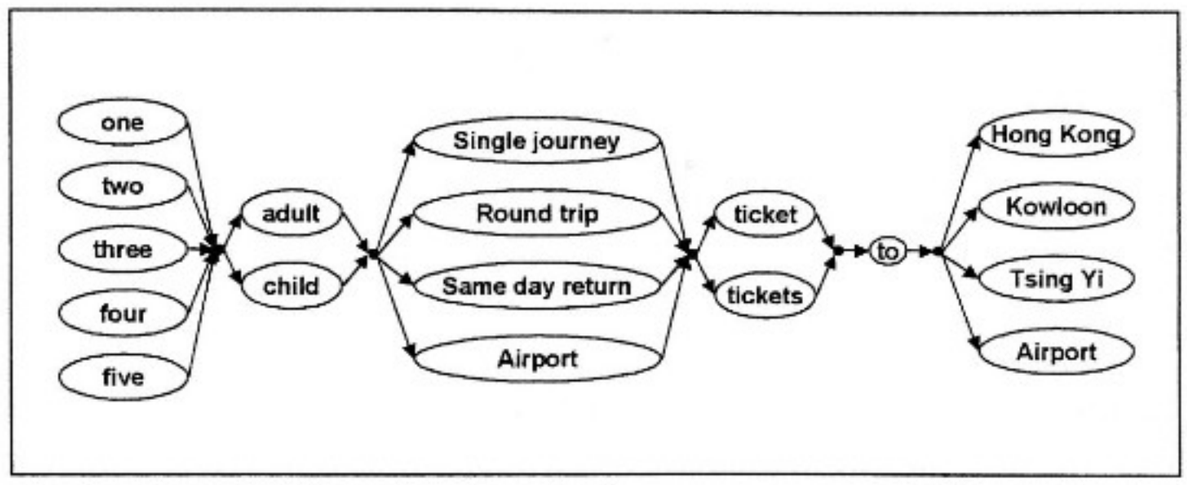


Figure 3e – Example of a Finite State Grammar Used in a Speech Recognition System

A deterministic language model is difficult to build because expert knowledge about the intended task is required to handcraft the grammar. This is feasible only for small-scale recognition tasks. Moreover, deterministic grammar is also prone to error due to the limited coverage of possible sentences. For large vocabulary continuous speech recognition tasks, full coverage of all possible sentences using a finite-state grammar will require a prohibitively huge grammar. In order to overcome these problems, a statistical language model can be used instead. Statistical n-gram is a common language model used for large vocabulary recognition tasks. In a statistical n-gram, the probability of occurrence of an entry in the vocabulary is determined by the joint probability of the n-1 previous words in the sentence. Probabilities for a statistical n-gram can be estimated using a large amount of training data. Such training data is usually a collection of textual materials from which the co-occurrence probabilities of entries in the vocabulary are estimated. In general, the value of n is 2 to 3.[3] Due to the lack of training data, statistical language model training usually evens out the language model for n-grams unseen in the training data. In addition, the statistical n-gram of word classes can also be used. Each word class typically includes words that have similar nature and/or grammatical function, eg belonging to the same part-of-speech category or carrying similar meanings. The advantage of using class-based grammar is that more data can be provided for the parameter estimation. It can also generalise the language model for unseen data falling into a trained class.

4. Overview of Speech Generation by Computers

Speech generation aims to present the computer's output in terms of synthetic speech for the user. The technology involved is often referred to as text-to-speech synthesis. When the information to be delivered is fixed and limited, playback of pre-recorded spoken messages can be used. When the information to be delivered is variable, more sophisticated generation approaches are needed. Figure 4a shows the basic components in the spoken language generation process.

Concepts (or information) in the message to be delivered are first verbalized into textual form. The text can then be converted into speech output through the text-to-speech synthesis. This usually involves conversion of textual data into pronunciation symbols. In addition, the appropriate prosody for the output speech needs to be generated. The pronunciation symbols, together with the prosodic information (if any), are used for

[2] Phonemes are defined as the minimal set of symbols representing all sounds in a language. From another point of view, when you change this sound unit, you change the language.

[3] For example, a statistical bigram (2-gram) for a vocabulary of 6,000 words will require at least 36,000 words in the training data for only one training sample. However, the occurrence of words in the training data is uneven and there should be more training data to achieve a reliable estimation of the probabilities. Therefore, hundreds of mega words and even giga words text data archives are needed for training purposes.

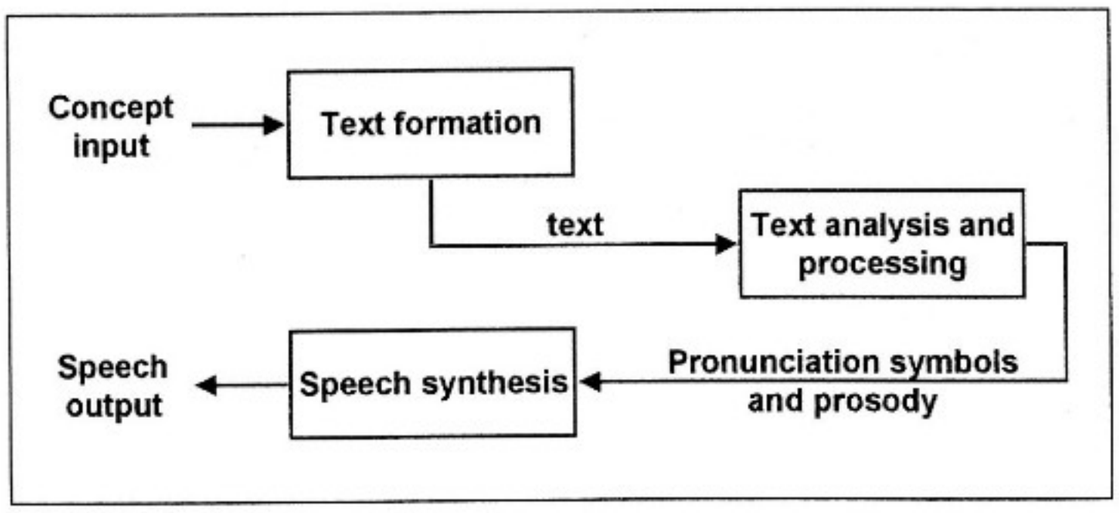


Figure 4a – Basic Components in the Spoken Language Generation Process

generation of speech output in the speech synthesis stage. The synthesised speech data is then delivered to the user through output channels such as loud speakers, telephone lines, computer files. In many cases, information to be delivered by the computer is already in textual form. Under such conditions, the process of verbalisation / text formation can be bypassed.

In a TTS system, the first step is to convert the textual input into pronunciation symbols via one of the two methods: dictionary look-up or rule-based generation. Dictionary look-up is usually used as the default conversion approach. Conversion of words in the textual input into pronunciation symbols can be as simple as a mapping task. For Chinese languages, if the textual information provided is a string of Chinese characters, word units have to be first extracted from the character sequence by a word segmentation process. If multiple pronunciations[4] are found during the dictionary look-up, the appropriate pronunciation is selected by analysing the contextual information. For example, each Chinese character is pronounced as a syllable. The mapping of characters to syllables is a many-to-many process as a character can have multiple pronunciations or different characters can have the same pronunciation. To handle the mapping ambiguity, segmented word entries are used during the pronunciation looking-up process because the contextual and lexical information can help to resolve the ambiguities.

To generate natural speech output, the intonation over the course of the whole sentence should also be controlled. For example, a declarative sentence has very different intonation from an interrogative sentence. Furthermore, there are many other aspects of variation, eg stress, focus, etc. The better the intonation, the more natural the generated speech sounds.

Having derived the pronunciations and the prosody of the targeted sentence, the acoustic signal is then generated. There have been many different approaches proposed for this task. The major difference is in the use of the sound units. For example, the smallest unit used is the phone unit. Phones may even be synthesised using information extracted (eg formants – resonant frequencies of the human vocal tract when generating the targeted sound) from speech templates. There are many cases in which intermediate-size units, such as syllables or words, are used. Larger units such as phrases, sentences, or variable length units can also be used. These speech templates are then selected, modified and concatenated according to the textual input to form the speech output. The selection process tries to choose the best possible units from the inventory of speech templates. The modification process tunes the selected units to cater for the acoustic and prosodic context whenever necessary. The final step is to put together the modified sound units to form the speech output. In general, the larger the units and the less modification done to them, the more natural will the generated speech sound. In the extreme case, output speech is simply played back as it is pre-recorded. This is a common approach for systems delivering fixed information.

Speech generation relies greatly on a pronunciation dictionary. An accurate pronunciation dictionary is one of the most important resources needed for speech generation. Sufficient and appropriate speech templates are also needed. The design and collection of speech templates for speech generation is a critical process. This process determines the number of speech templates needed and the resultant quality of the output speech.

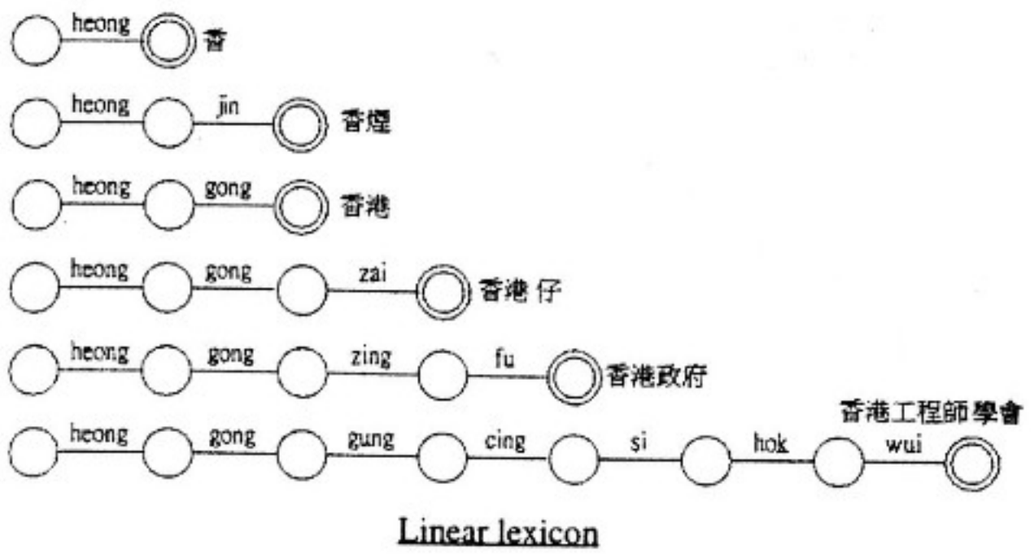
5. Advancement of Speech Recognition Technologies at CUHK

CUHK has pioneered research and development of Cantonese speech recognition technologies since over a decade ago. There has been investigation of the statistical modeling of speech sound units for Cantonese speech recognition. In [4], the neural network (NN) is used as the mathematical model for the representation of Cantonese syllables. The use of NN aims at capturing the non-linear relationship between the features and identities of syllables by the layers of neurons. In addition, recurrent NN is used to capture the temporally dynamic characteristics of speech sounds. The study of the use of NN for the recognition of isolated words in Cantonese was the early emphasis of CUHK's research on Cantonese acoustic modeling.

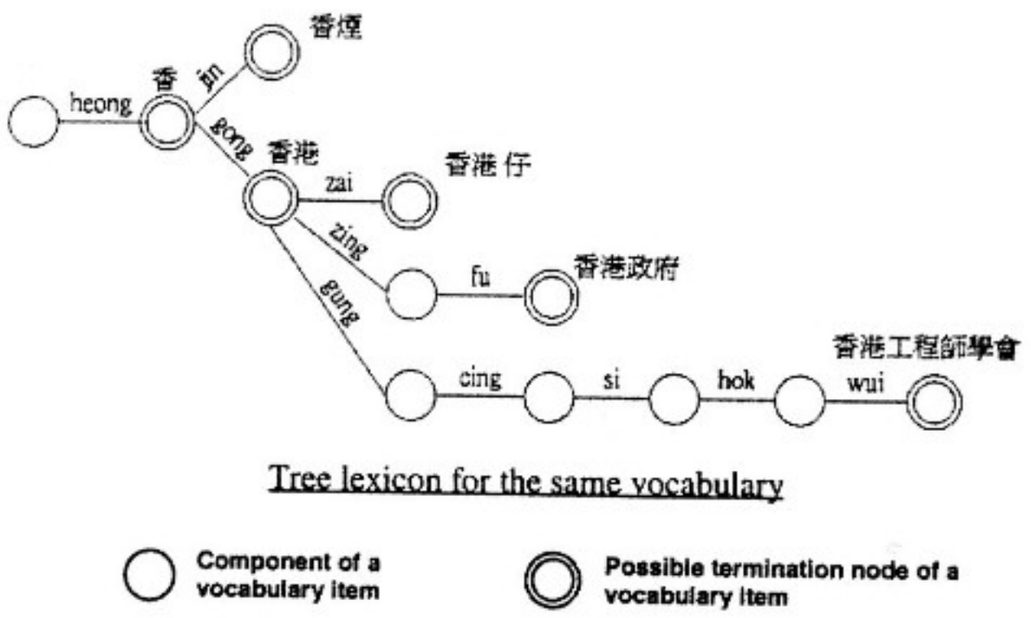
The use of HMM for acoustic modeling quickly became popular. The initial experience of using the HMM model for acoustic modeling in Cantonese speech recognition began with the work as described in [5]. In this study, the recognition task was speaker dependent, meaning that the training data was collected from the speaker who tested the system. Later on in [6], an investigation into the use of the HMM model for speaker-independent acoustic modeling was presented. A large amount of training data was systematically collected from many speakers so as to facilitate the estimation of the parameters in the acoustic models. The performance evaluation was done with the utterances from a set of different speakers. To deal with the problem of data sparseness, ie the training data is not sufficient for some of the sound units, a technique known as phonetic tree-based clustering of the HMM states was used. With this technique, HMM-states are clustered and merged when they are considered to be phonetically similar. The clustering is guided by expert knowledge about Cantonese phonetics and the similarity is determined by the change of statistical properties (likelihood) of the HMM states. In [7], HMM-based acoustic models were used in conjunction with statistical language models. The language model was established with the availability of a large amount of textual data. The combined use of speaker-independent acoustic models and statistical language models has made large-vocabulary continuous speech recognition (LVCSR) of Cantonese possible. However, as the vocabulary becomes larger, the computational requirement becomes greater. In [8] various techniques have been exploited to improve the speed of recognition. Particularly, tree-structured lexicon is used instead of linear lexicon inside the speech recogniser. In general, the vocabulary is a list of words with the corresponding pronunciations. These words are represented as parallel paths in a speech recogniser using linear lexicon. During the recognition process, there is some computational power wasted when multiple paths share the same prefix, as shown in a. The use of tree-structured lexicon can avoid unnecessary repetition of computation.

The recognition performance was further improved by applying the technique of language model look-ahead. In the recognition process, the language model is integrated with the acoustic model to search for the best-matching sequence of sound units. Language model look-ahead is a technique that attempts to take into account the contribution of the language model before actually confirming the word sequence. This aggressive application of the language model can improve recognition speed and accuracy because taking the language model into consideration ahead of time can lead to the decision to ignore unlikely paths at an earlier phase of the recognition process. More computational power can be saved by focusing on those more likely acoustic models. On the other hand, the class-based language model technique was also investigated in [8]. The use of this model is important, especially when there is a limited amount of training data for the intended recognition task. If the vocabulary items of the same class (eg company names, names of places)

[4] A word with multiple pronunciations is known as a homograph. Appropriate choice of pronunciation is necessary otherwise the output sounds unnatural to the users.



Linear lexicon



Tree lexicon for the same vocabulary

Component of a vocabulary item

Possible termination node of a vocabulary item

Figure 5a – Illustration of the Difference between a Linear Lexicon and a Tree-structred Lexicon

are treated as the same class entries, the amount of training data for language model would be increased effectively. In [9], the computational efficiency and recognition accuracy of the internal searching process of a speech recognition system was investigated. The efficacy of the improved searching algorithm has also been demonstrated in a practical recognition task in the domain of stock information enquiry.

For acoustic modeling of Cantonese speech, there are other possible choices of sound units for modeling instead of initial and final. In [11], we compared acoustic modeling using different sound units and found that the use of initial-final units gives the best balance between recognition accuracy and computational complexity (in terms of the number of HMM states after phonetic clustering). A similar result has also been found for Mandarin.

Tone recognition is a key issue in Chinese speech recognition. An early study of Cantonese tone recognition using supra-segmental features derived from the pitch contour can be found in [10]. Artificial neural network was used for the determination of tones in isolated Cantonese speech. In [12], the detection of tone in continuous Cantonese speech and the use of tone information in Cantonese speech recognition have both been studied.

Another issue in speech recognition is the variation of pronunciations from speaker to speaker. These variations may be due to the accent, speaker characteristics, speaking styles (eg spontaneous speech versus read speech). In [13], a probabilistic approach was applied to model the pronunciation variation in continuous Cantonese speech.

Speaker-dependent speech recognition is known to give better recognition accuracy than speaker independent speech recognition. For this reason, the speaker independent acoustic model can be adapted using speaker specific data to improve recognition accuracy in speaker-dependent case. This technique is particularly useful when the recognition performance of specific speakers needs to be improved and the speech data from those speakers is insufficient for training speaker-dependent models. In [14], it was shown that proper adaptation of speaker independent acoustic models would lead to improvement on the recognition accuracy.

6. Advancement of Speech Generation Technologies at CUHK

Our research portfolio at CUHK devotes a focused effort towards the research and development of Cantonese speech synthesis technologies. Cantonese is especially desirable for such research due to its large syllable inventory and rich tonal structure in comparison with Mandarin. As will be described later, the approaches that we have developed for Cantonese thus become directly portable for use in Mandarin. For synthesis of acoustic signals based on pronunciation symbols, [15] has investigated the use of NN for the synthesis of speech signals based on speech templates of Cantonese phones. This approach can generate speech output using a small number of templates when prosodic information is available. However, as detailed prosodic information may not always be available another approach that has been applied to Cantonese speech synthesis is syllable-based pitch synchronous overlap add (PSOLA). The advantage of using syllables as synthesis units is that the need for detailed prosodic control is eliminated and the recorded template can always be used as a default. This simplicity is obtained at the cost of storing a large number of speech templates. PSOLA is a special signal processing approach for speech synthesis which enables the prosodic modification of the speech template. In [16, 17], studies have been made of the use of this approach for Cantonese speech synthesis. Investigation has also been made of the use of cross-syllable [18] units in speech synthesis with PSOLA. A major disadvantage of PSOLA is that the collected speech templates must be manually pre-processed to locate the positions of pitch cycles. This is a time-consuming process that hinders the frequent change of speakers in the speech generation system.

With the advent of computer technology, a large amount of storage is often readily available at affordable price in most computer systems. [19] marks the first effort that demonstrates the use of a concatenative approach for synthesis of highly natural Cantonese speech. The engine developed is named CU VOCAL. A concatenative approach involves the design and collection of a speech corpus, extraction of phonetic units from this corpus for synthesis, as well as unit selection techniques to achieve high degrees of intelligibility and naturalness in the synthesis outputs. Unit selection in [20] considers both phonetic context and tonal context. This approach has been demonstrated to be directly portable for Mandarin in a domain-specific context. We have also developed a process for domain-specific optimisation to further improve the perceived quality of synthesised speech in restricted domain applications. More recent work synthesises the declination effect - a prosodic phenomenon where the fundamental frequency of the speaker lowers gradually over the course of a spoken sentence [21].

Prosodic control is also an important component of speech synthesis. Prosody includes the properties of speed, loudness and pitch of speech. It is well established that the intonation of an utterance gradually drops in a declaration and rises at the end in a question. However, more detailed prosodic control is necessary for natural sounding synthetic speech. [22] has made an initial study of the detailed effect of prosodic features in synthetic speech. In [23], the problem of prosody prediction for control in speech synthesis is further studied. This work tries to find out how word units are clustered and pronounced as a unit (prosodic phrase) and also predicts the values of these prosodic properties for better prosodic control in the speech synthesis process.

As mentioned before, text processing is one of the major components in a speech generation system. This is necessary in a text-to-speech system when the inputs are character sequences. The character sequences are analysed first and the intermediate control information (eg sequence of pronunciation, prosodic control information) is passed to the speech synthesis components. During the text analysis process, word segmentation is applied to extract words from the character sequence and to convert the character sequence into a syllable sequence. The word segmentation process can be achieved using the maximum matching approach. In a maximum matching process, a character sequence is repeatedly searched for the longest valid word based on a given vocabulary. The identified word is then extracted and the searching process is repeated for the

remaining portion of the character sequence. To illustrate, Figure 6a shows the idea of a left-to-right maximum matching process. Other versions are also possible such as the bi-directional maximum matching used in CU VOCAL. In addition, special objects in the input character sequences also need to be handled properly; for example, date, time, numbers, and proper names. In [20], these objects are first located in the character sequences and then processed properly for correct pronunciations of the textual material.

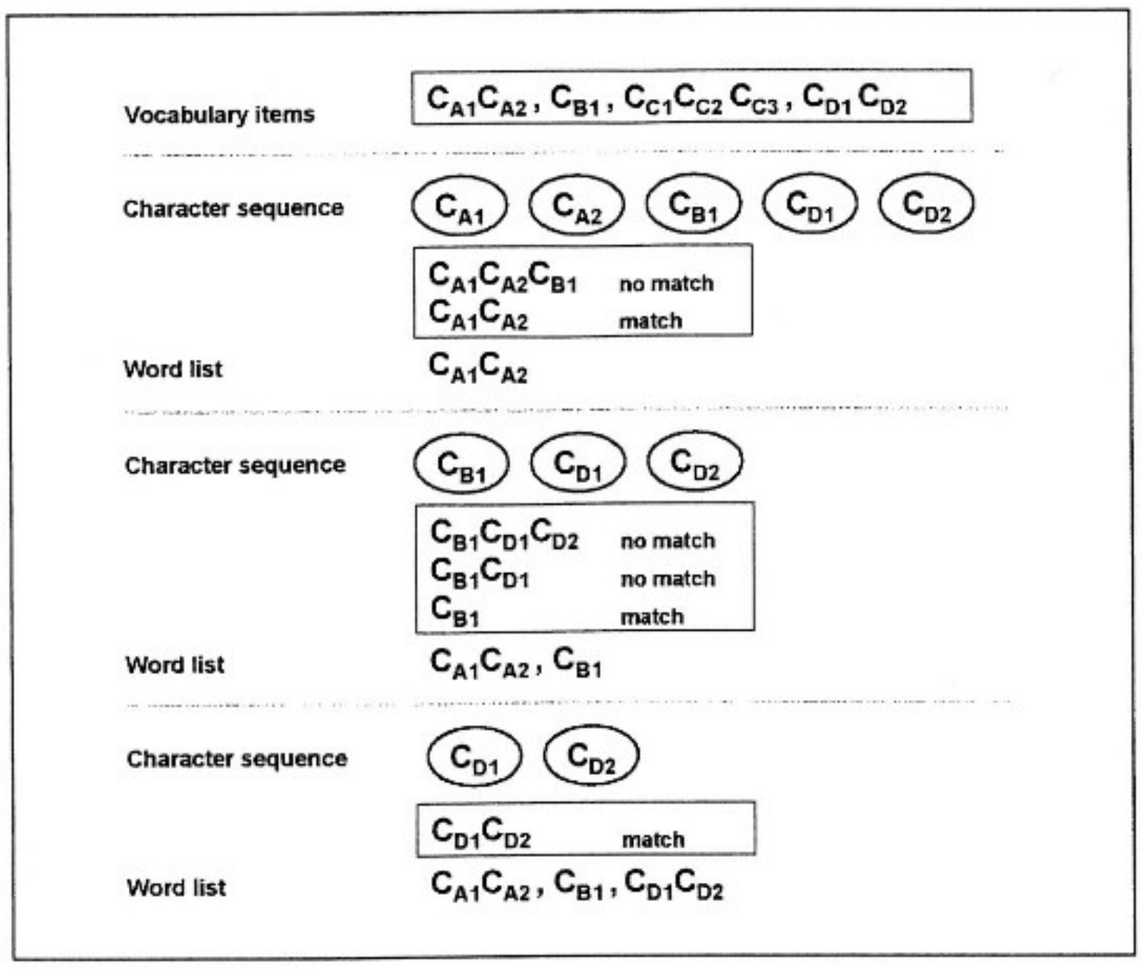


Figure 6a – Illustration of the Left-to-right Maximum Matching Process. Words are Segmented from the Input Character Sequence Based on a Vocabulary. The Results will be a List of Words

7. Processing Spoken Language

In addition to analysing the phonological characteristics of speech, computer processing of spoken language also involves a host of technologies that analyses the linguistic characteristics of speech, or rather, spoken language. Our research team at CUHK has developed a host of such spoken language technologies that can be integrated with speech recognition and/or speech generation to produce innovative interface technologies and applications. Several of these technologies are summarised as follows:

- Interpretation of the semantic content of a spoken/textual utterance is known as spoken/natural language understanding (SLU/NLU). This requires a process of linguistic analysis known as parsing. A parser refers to a grammar for analysis. The grammar is often written by a grammarian, which is a very tedious and time-consuming process. Furthermore, it is difficult to anticipate all the occurrences of disfluencies in spoken language for grammar writing. Typically, a grammar that governs legitimate word combinations in a restricted domain contains several hundred rules. As the complexity of the domain grows, the grammar size increases rapidly which poses efficiency concerns for parsing. We have developed a parsing process [24, 25, 26, 27, 28] that includes a parser and an algorithm that automatically partitions a large grammar into multiple smaller sub-grammars. Each sub-grammar works in conjunction with its specialised (sub-)parser. Another algorithm composes the outputs of the sub-parsers to produce a unified semantic interpretation. Furthermore, we have also developed a semi-automatic algorithm that can induce a grammar directly from un-annotated corpora of natural language sentences. This can greatly expedite the process of grammar development [29, 30].
- The ability to understand a spoken language query entails subsequent actions on the part on the computer. This includes endowing the computer with the intelligence to carry a coherent dialog with the user, to remember the context of the interaction and dialog history (also known as discourse) and to take initiative at appropriate times in the interaction flow to make suggestions for expediting the completion of a goal-oriented dialog. These capabilities are enabled by technologies in mixed-initiative dialog modeling, a new area at the forefront of speech and language research. The dialog is 'mixed-initiative' since the user and the system can both influence the dialog flow, and the interaction aims to help the user gather appropriate information while progressing towards task completion. Our team at CUHK is a pioneer in the use of Belief Networks for mixed-initiative dialog modeling. The use of Belief Network is a data-driven approach that can model domain-specific constraints and infer the communicative goal of the user based on his/her spoken concepts [31, 32, 33]. The data-driven nature of the approach also alleviates the tedium of hand-crafted heuristics and enhances the portability across application domains as well as scalability for increasingly complex application domains.
- Interactive dialogs involve the generation of coherent and succinct responses, and this requires the technology of natural language generation (NLG). NLG can verbalise a system's conceptual message into a fluent textual form. It can also be tightly integrated with text-to-speech synthesis to control the prosody of the spoken response. NLG is also an emerging area of active research. Our research team has embarked on this exciting area and designed one of the first NLG engines that is driven by both the task goal and the dialog act (or communicative goal) of the system responses [34, 35].
- The integration of NLU and NLG gives rise to a bi-directional English-Chinese machine translation system. An input English natural language query can be parsed and interpreted using NLU, and its meaning generated in the form of Chinese text by NLG. This constitutes automatic English-to-Chinese machine translation. Similarly, an input Chinese natural language query can be parsed and interpreted using NLU, and its meaning generated in the form of English text by NLG. This constitutes automatic English-to-Chinese machine translation using the same technology components [29, 36].
- Another technology of importance is information retrieval (IR). Information retrieval of textual documents has a much longer history of research and development than retrieval of spoken documents. Our research team is the first to develop a Cantonese spoken document retrieval system. This integrates speech recognition technology with information retrieval (IR) technology. We have developed two main IR engines - one based on the vector-space retrieval model and the other based on hidden Markov models (HMMs). These models have been adapted for retrieving transcribed spoken documents because transcription errors are inevitable in documents indexed by automatic speech recognisers. Our IR engines are robust to speech recognition errors [37] and can utilize multimedia fusion techniques to improve retrieval performance [38], as will be described in the next section.

8. Spoken Language Systems

By integrating the various component technologies described above, our group at CUHK has developed several spoken language systems that demonstrates the capabilities of state-of-the-art human-computer speech-based interfaces.

A. Spoken Document Retrieval

Spoken document retrieval is the task of finding information relevant to user specified queries from an archive of speech recordings. Speech documents include speech recordings of presentations, meetings, conferences, lectures, and news broadcasts. Spoken document retrieval can enable efficient retrieval of information from archives of these resources.

By integration of speech recognition with information retrieval technologies, spoken document retrieval can be achieved by converting speech recordings to textual transcriptions using LVCSR. These textual transcriptions can then be used for construction of indices of the spoken documents and when a user specifies a query, a search of the indices will identify relevant documents and return the result to the user. Figure 8a shows the block diagram of a general spoken document retrieval system.

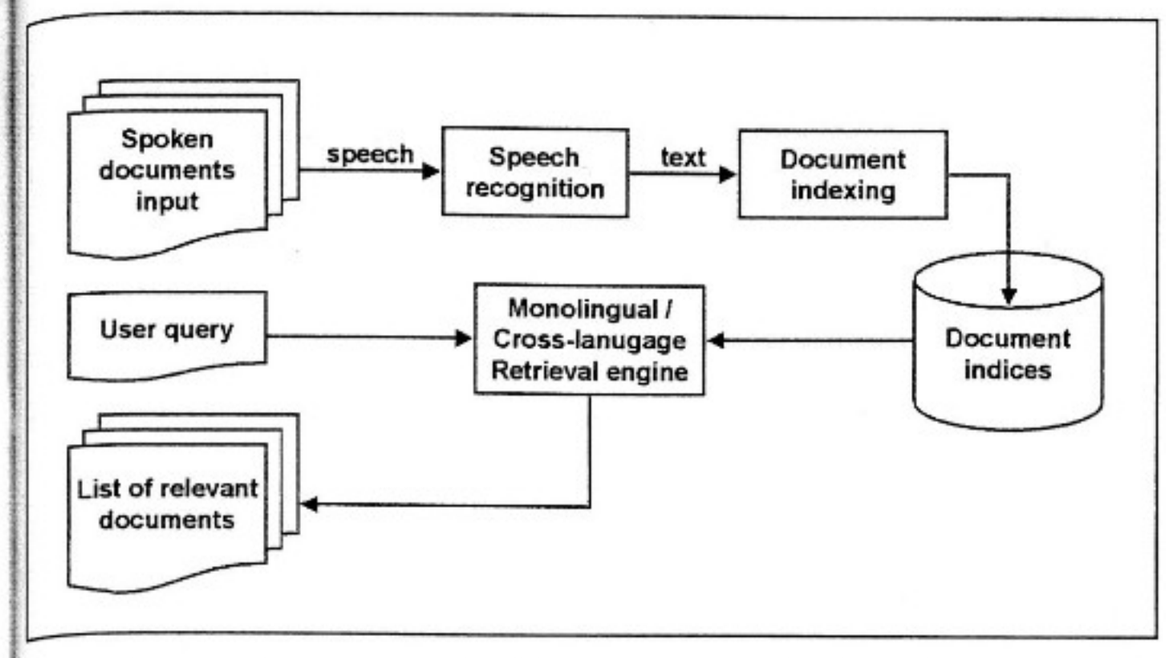


Figure 8a – Block Diagram of a General Spoken Document Retrieval System

In practice, speech recordings may be made in a language different from that used for making queries. Therefore, cross-language spoken document retrieval is needed to enable retrieval of spoken documents in other languages. The MEI project [39] has pioneered the study of CL-SDR by studying an English-Mandarin CL-SDR task where English queries are used to locate relevant news broadcast stories recorded in Mandarin. This project also studied the problem of automatic translation based on pronunciation - transliteration [40]. Transliteration is useful for translation of proper names because there are new names added every day and manually adding items to a translation dictionary is time-consuming.

We have also studied CL-SDR by combining a dictionary-based translation approach with an HMM-based retrieval model [41]. The dictionary-based translation of queries is used because of its simplicity. In addition, the use of the HMM-based retrieval model has the advantages of being probabilistically based and being able to integrate the statistical translation component. As a result, the same retrieval model can be used for both monolingual and cross-language SDR tasks.

The application of spoken document retrieval technologies for searching Cantonese news broadcasts has been studied since [42]. To cater for potential recognition errors, [37] has applied a query expansion approach that includes 'confusing terms' in the query before performing retrieval. 'Confusing terms' are frequently mis-recognised terms obtained from training data. This approach can improve the ability of the retrieval system to retrieve documents from automatic transcription with recognition errors. The disadvantage is that computational complexity is increased for there are more terms in the query that have to be processed. It has also been shown that performance improvement can be obtained from expansion of documents [43]. By adding more terms to documents in the archive, identification of relevant documents can be made easier because a document can now be retrieved by using more relevant terms. Investigation has also been made to the application of the document expansion technique to a CL-SDR task [44].

Another approach for improving retrieval performance is the application of multi-scale retrieval [42]. Multi-scale refers to the use of word and subword (eg character and syllable n-grams) units in retrieval. This has been demonstrated [42] as enabling the combination of the lexical information of words and the robustness of subwords to homophones and word segmentation ambiguity. Further detailed investigation has been made to multi-scale fusion in [45]. It was found that at the cost of using more indexing terms (combined use of words and subwords as indexing terms), consistent improvement in retrieval performance can be made in a Cantonese news broadcast retrieval task. This approach is also applicable to CL-SDR [41].

In addition, speech recordings are usually carried out continuously and the boundaries of documents are unknown. Story boundaries are important for segmentation of recordings into spoken documents. Manual annotation of story boundaries is time-consuming and therefore an automatic approach has been investigated. In [38, 46, 47, 48], information in the video field is employed for detection of possible boundaries. However, this approach is applicable only to recordings with video tracks. Therefore, [47, 48] has also investigated the use of a Gaussian mixture model (single state HMM with observation probability represented by mixtures of Gaussian) for identification of boundaries where the acoustic condition is changed. These boundaries give important clues in the determination of story boundaries especially in news broadcasts. Further improvement in boundary detection has been achieved by combining detection hypotheses based on video and audio tracks.

B. Spoken Dialogue Systems

A spoken dialogue system aims to support speech-based interactions between human and computer, in the form of an interactive dialog that seeks to complete a task in a restricted domain, eg information inquiries/ transactions. Speech recognition and speech generation are the basic component technologies in a spoken dialog system which facilitate the input and output via speech. In this section, we will describe two multilingual dialogue systems that are implemented by integrating several different modules, including third party modules available from the market. These systems aim at demonstrating the capability of the technologies and aggregating experience in developing technologies for real-world applications.

1. CUForex

CUForex is the first trilingual (Cantonese, Mandarin and English) financial information inquiry system made available for public trial [49]. This system integrates speech recognition components from SpeechWorks (now ScanSoft) as well as our concatenative speech synthesis technologies. It can provide real-time foreign exchange and interest rates captured from a direct satellite downlink provided by Reuters.

CUForex can operate in the mode of a directed dialog, where the user is guided stepwise to speak the appropriate keyword at the appropriate time, eg 'please say the currency you wish to buy', followed by 'please say the currency you wish to sell', etc. Directed dialogs are intended for new users to the system. In addition, CUForex can operate in a the mode of natural language shortcuts, where the user can specific in one dialog turn his informational goal, eg 'I am interested in the exchange rate between the greenback and the British pound.' Natural language shortcuts are intended for experienced or knowledgeable users.

These users' queries are interpreted by the system, which then retrieves the appropriate information and generates a textual response. This is passed to the speech generation component to deliver the information to the user over the telephone.

CUForex is a good demonstration of the capability of spoken language technologies to develop human-computer interfaces with enhanced usability. In a traditional interactive voice response system (IVRS), users need to map their options into the buttons of the keypad for touch-tone/ DTMF input. This often creates a tedious interaction, eg 'For US Dollars, press 1,' 'For British Pounds, press 2' ... 'For Thai Bhat, press 15', etc. In addition, users need to either memorise such code mappings or they need to listen to lengthy descriptions over the phone. Speech-based interface simplifies the interaction by allowing the user to speak their selected options directly.

2. ISIS

ISIS stands for Intelligent Speech for Information Systems for the stocks domain. It is also a trilingual system that behaves like a virtual stock broker. Aside from mixed-initiative interactions that involve stock quotes and management of simulated portfolios, users can also input mixed-modal requests. An example is: 'I'd like the transaction records of these two stocks <click1> <click2>', where the user clicks on the computer screen displaying a table with a stock listing. ISIS is also among the first systems to support asynchronous human-computer interaction by means of software agents. Users can request for information monitoring, eg 'Please notify me when the Hang Seng Index rises above nine thousand points'. Upon such requests, ISIS will launch software agents (software programs) that monitor the financial datafeed for the prescribed event(s). In the mean time, the user can carry an ongoing dialog with

the computer about other tasks. When the prescribed event occurs, the dialog model interrupts the ongoing dialog at an appropriate time to alert the user. In this way ISIS can support multilingual, mixed-modal and mixed-initiative dialogs with interleaving online interactions and offline delegations.

Figure 8b is an illustration of the components of the ISIS system [50].

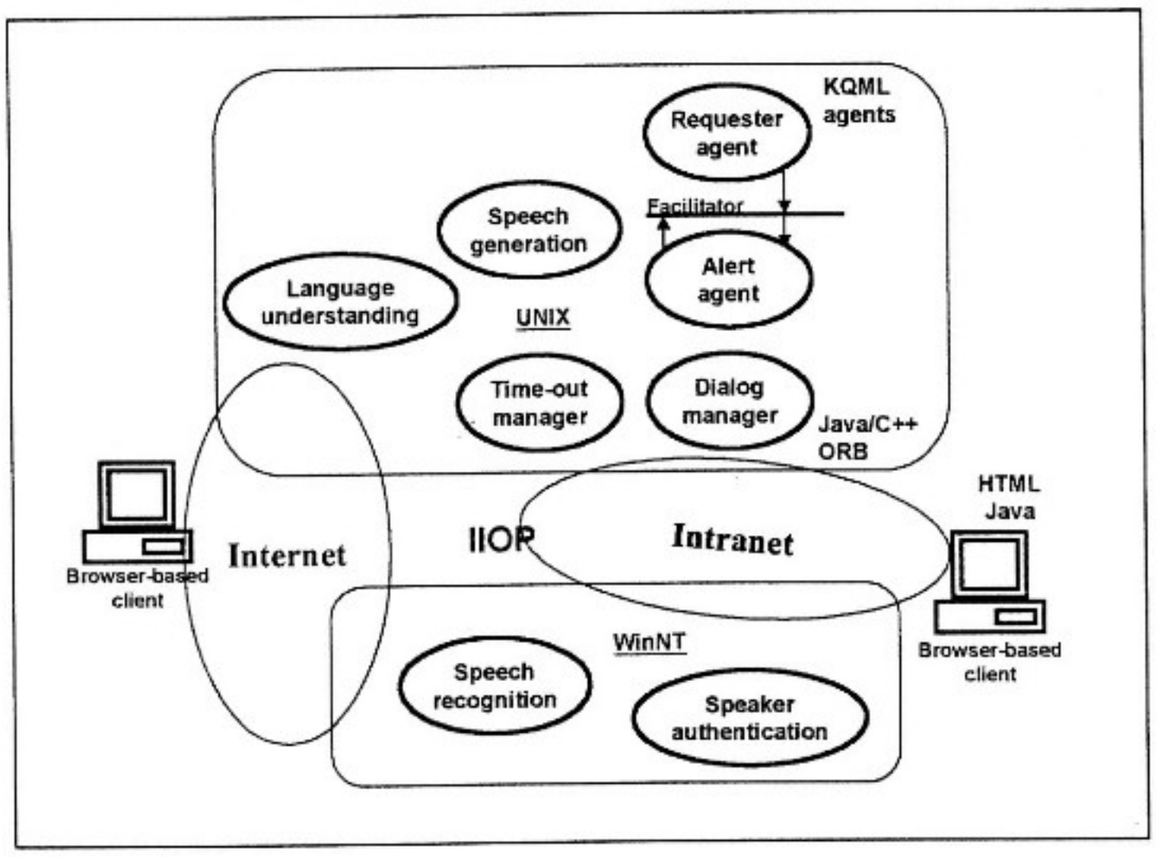


Figure 8b – System Architecture of the ISIS System

9. Resources and the Development of Technologies

Of the large number of component technologies, some are available for technology transferring to industrial parties. In addition, there is also a large amount of valuable data resources compiled during the research and development of these technologies. These resources are also made available to academic and industrial parties for further development of their own spoken language technologies. In this section, we will describe some of these resources and technologies.

A. Spoken Language Resources

Spoken language resources include speech data and pronunciation dictionaries. With the support of the Industrial Support Fund (ISF) of the Hong Kong SAR government, we have compiled the first large-scale collection of Cantonese spoken language corpora (CUCorpora) collected over the microphone channel. These corpora contain human speech collected from Cantonese speakers in a studio environment. The content of the speech data is carefully designed for several different purposes [51]. The speech data contains, for example, a multitude of Cantonese syllables annotated with pitch marks. Pitch marks are important information for speech data used in pitch-synchronous modification (such as PSOLA) during speech synthesis. For speech recognition applications, there is also in the speech data phonetically rich data in the form of isolated words and continuous sentences which are designed for training speaker independent acoustic models for use in continuous Cantonese speech recognition systems. In addition, there are also application-specific data for common application domains such as command and control, and digit string recognition. All of this speech data were fully annotated with transcriptions for the phonetic content collected. CUCorpora marks the significant beginning of our continuing effort to build up the infrastructure for research and development of spoken language technologies.

Like microphone input, spoken language interface is most relevant for situations where speech is the only means of communication. Interface to computer systems over telephone lines is one of these situations. Since there is difference between speech signals transmitted over microphone and telephone (eg bandwidth), speech data has also to be collected over telephone lines for training acoustic models used in speech recognition over telephone lines. Through the support of the Innovation and Technology Fund (ITF) of the Hong Kong SAR government, we have made a significant effort to collect speech data over telephone lines (both fixed-lines and mobile) from more than 1,000 speakers [52]. This large-scale collection of telephone speech data is known as CUCall. CUCall was designed to cover both phonetically rich data as well as domain specific data. This data can be used to develop speaker-independent speech recognition systems over telephone lines using the phonetically rich data. Customisation to specific application domains can be fine-tuned using the domain specific data (eg words for command and control, digit strings, name of places, stock names). CUCall has enriched the spoken language processing infrastructure to cover telephone applications as well.

Pronunciation dictionaries are also important data resources for computer speech processing. They are required for both speech recognition as well as speech generation systems. For Cantonese speech processing, we have compiled the full list of Chinese characters together with their pronunciations – the CUPDICT. In addition, there is also a long list (around 40,000 entries) of Chinese words (the CULEX) collected from various sources (such as newspapers, Internet) and pronunciations are also provided for all of these words [1]. These two pronunciation dictionaries are used extensively in speech recognition systems as recogniser vocabulary and also in speech generation systems for converting textual inputs into pronunciations.

B. Speech Synthesis Engines, both PSOLA-based and Concatenation-based

Text-to-speech (TTS) conversion is applicable to many systems (eg automatic information inquiries over telephones, public address systems). TTS is simple to integrate into existing systems because it is reliable. All textual input is converted into speech and returned as output. We have developed two different TTS systems based on the PSOLA and concatenation approach described above. CUTalk is a syllable based PSOLA TTS system implemented as API for both Windows and Linux platforms. Another concatenation based TTS is the CU VOCAL. CU VOCAL can synthesise natural sounding speech output and customisation for specific application domains is possible. In addition to the Windows API implementation, CU VOCAL is also compliant with Microsoft SAPI and can be delivered as a Web Service [53] for distributed system implementations. These TTS component technologies are being continuously developed to enhance naturalness and ease of integration.

C. Speech Recognition Software Building Blocks

The Chinese University Recognition Software Building Blocks (CURSBB) is a collection of developmental tools and Windows API for building custom-made speech recognition modules. The targeted users are developers of software systems that are interested in enhancing their products and services with speech input capability. Users of CURSBB can define the vocabulary of the recogniser, design the recognition grammar, and test the designed recognisers all using the provided graphical interface. A sample design flow is shown in Figure 9a. After testing and fine-tuning the designed speech recogniser, users can integrate the designed recognisers into their software systems by invoking the recognisers using provided Windows API. The designed recognisers can also be used over both microphone and telephone input channels simply by invoking the appropriate acoustic models. This architecture can significantly reduce the overhead of porting a speech recognition application between microphone and telephone channels.

D. Author Once, Present Anywhere Software Platform

The 'Author Once, Present Anywhere' (AOPA) [54] is a universally accessible software platform that supports Chinese Web content development for displayless voice browsers, mobile mini-browsers and regular Web browsers in E-business services provision. Universal accessibility refers to accessibility through displayless voice browsers for telephones or for the elderly/visually impaired; mobile mini-browsers for Internet-ready phones and PDAs, and regular Web browsers for desktop PCs. AOPA enables Web content/service providers, ISPs, and ASPs to author and maintain a single content repository, whose content automatically adopts usability-optimised presentation styles to reach the client devices of diverse form factors.

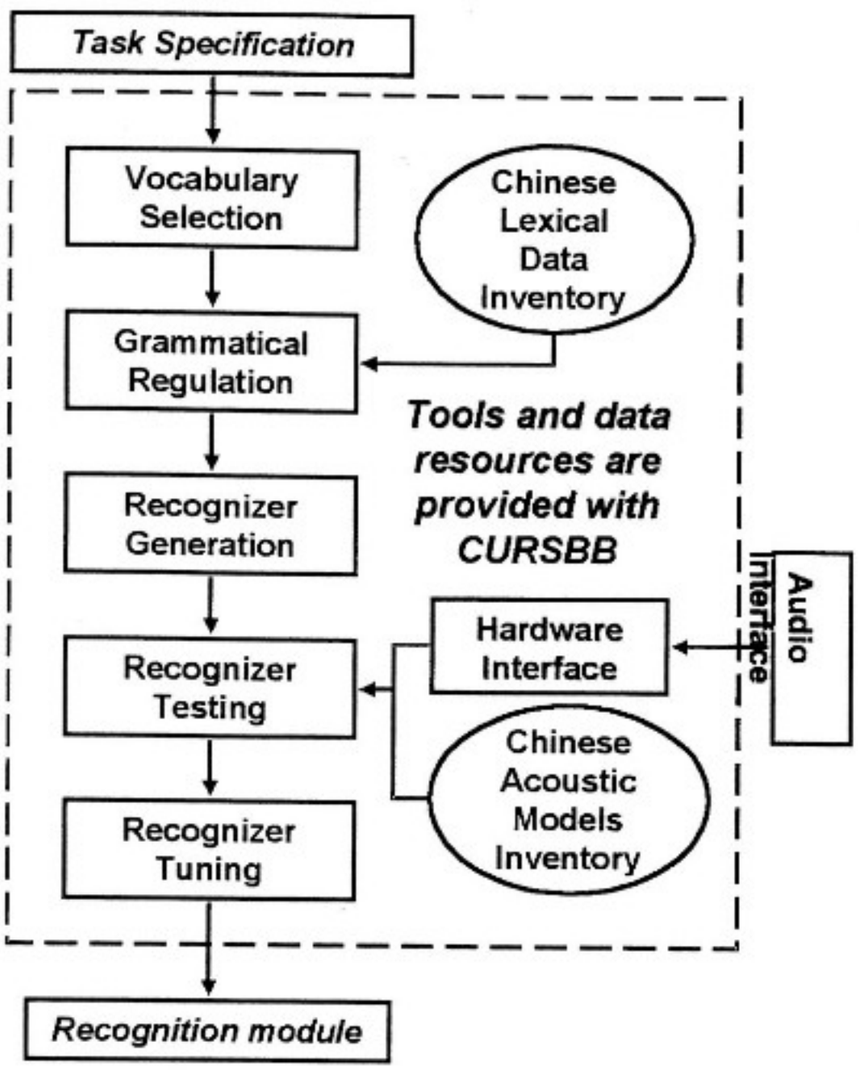


Figure 9a – Typical Design Flow of Speech Recogniser Using the CURSBB Tool Set

Web visitors using mobile handhelds or telephones will outnumber those using desktop PCs within three years. Universal accessibility enables information dissemination to a much wider audience. This is critical and beneficial to B2B/B2C E-commerce, M-commerce, and voice-enabled E-commerce.

AOPA eliminates major redundancies and inefficiencies in a common practice to achieve multiple accessibility - the same content is re-authored for every alternative form factor in client devices. AOPA leverages emerging standards from the W3C to decouple content specification (in XML) from presentation specification (in XSL). Platform components include:

(1) A HTML-to-XML transcoder to process existing Web content.

(2) Reference XSL stylesheets encoding usability-optimized presentation styles for specified content sources and client devices.

(3) The first reference implementation of Chinese VXML to enable displayless, Cantonese voice browsing, by integrating with Chinese text processing, Cantonese speech recognition and synthesis technologies developed at the Chinese University of Hong Kong.

Figure 9b shows an overview of the AOPA platform. For more details of this project, interested readers may refer to [54] or visit the project webpage at http://www.se.cuhk.edu.hk/AOPA.

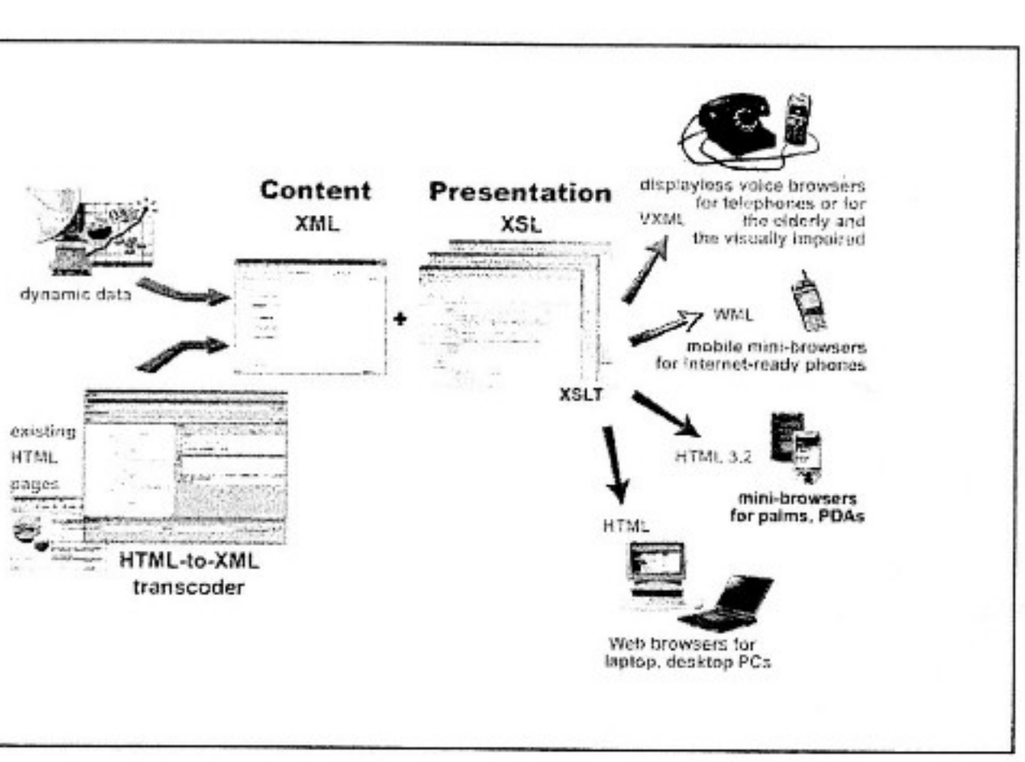


Figure 9b – An Overview of the AOPA platform. The Single, Unified XML Repository Can Have Its Content Displayed with Usability-optimised Presentation Styles on Client Devices of Diverse Form Factors

10. Summary

In this paper, we have presented the key technologies enabling Chinese speech-based human-computer interaction, namely, speech recognition and speech synthesis. We have also described the special processing required in these technologies for handling the phonological and linguistic characteristics of the Chinese language and especially the Cantonese dialect. This paper continues to describe the speech recognition and speech synthesis technologies developed in our research team at the Chinese University of Hong Kong, as well as other emerging technologies that can integrate with speech-based interface technologies to bring about novel and innovative systems. Illustrations of several of such systems ensue, including a spoken document retrieval system as well as two spoken dialog systems. We conclude with a description of the component technologies and resources that have been developed in-house and are made publicly available in order to lower the entry barrier of new academic/industrial parties who wish to embark on this exciting area of research and development.

Acknowledgements

This work was support in part by the Research Grants Council and the Innovation and Technology Support Fund of the HKSAR.

References

1. Description of CULEX and CUPDICT, http://dsp.ee.cuhk.edu.hk/speech/page/corpus/Documents/culex.pdf
2. Ching, P. C., Lee, Tan and Zee, Eric, From phonology and acoustic properties to automatic recognition of Cantonese, *Proceedings of the First International Symposium on Speech, Image Processing and Neural Networks*, Vol. 1, pp. 127-132, Hong Kong (1994).
3. 香港語言學會編，粵語拼音字表，香港語言學會出版 (1998).
4. Lee, Tan and Ching, P. C., Cantonese syllable recognition using neural networks. *IEEE Transactions on Speech and Audio Processing*. Volume 7, pp. 466-472 (July 1999).
5. Ng, Ying Pang Alfred et al., Automatic Recognition of Continuous Cantonese Speech with Very Large Vocabulary. *Proceedings of the 5th European Conference on Speech Communication and Technology*, pp.1551-1554, Rhodes, Greece (1997).
6. Chow, K. F. et al., Sub-syllable Acoustic Modeling for Cantonese Speech Recognition. *Proceedings of the 1998 International Symposium on Chinese Spoken Language Processing*, pp.75-79, Singapore (December 1998).
7. Wong, Y. W. et al., Acoustic Modeling and Language Modeling for Cantonese LVCSR. *Proceedings of the 6th European Conference on Speech Communication and Technology*, pp.1091-1094, Budapest, Hungary (1999).
8. Choi, W. N. et al., Lexical Tree Decoding with A Class-based Language Model for Chinese Speech Recognition. *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China (2000).
9. Choi, W. N. et al., Searching for the Missing Piece. Proceedings of the 2001 *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 230-233, Trento, Italy (2001).
10. Lee, Tan et al., Tone recognition of isolated Cantonese syllables. *IEEE Transactions on Speech and Audio Processing*. Volume 3, pp. 204-209 (May 1995).
11. Lo, Wai-Kit, Sub-syllabic acoustic modeling across Chinese dialects. *Proceedings of the 2nd International Symposium on Chinese Spoken Language Processing*, pp. 97-100, Beijing, China (2000).
12. Lee, Tan et al., Using tone information in Cantonese continuous speech recognition. *ACM Transactions on Asian Language and Information Processing*. Volume 1, pp. 83-102 (March 2002).
13. Kam, Patgi et al., Modeling Cantonese Pronunciation Variation by Acoustic Model Refinement. *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland (2003).
14. Kwan, Ka-Yan et al., Unsupervised N-best based Model Adaptation using Model-level Confidence Measures. *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 69-72, Denver, Colorado USA (2002).
15. Lo, W. K. and Ching, P. C., Phone-based speech synthesis with neural network and articulatory control. *Proceedings of the 4th International Conference on Spoken Language Processing*, pp. 2227-2230, Philadelphia, Pennsylvania USA (1996).
16. Chu, Min and Ching, P. C., A hybrid approach to synthesize high quality Cantonese speech. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 277-280, Seattle, Washington USA (1998).

17. Law, K. M. et al., Cantonese Text-To-Speech Synthesis using Sub-Syllable Units. *Proceedings of the 6th European Conference on Speech Communication and Technology*, pp. 991-994, Aalborg, Denmark (2001).
18. Law, K. M. and Lee, Tan, Using Cross-syllable Units for Cantonese Speech Synthesis. *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China (2000).
19. Fung, T. Y. and Meng, H., Concatenating Syllables for Response Generation in Spoken Language Applications. *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey (2000).
20. Meng, Helen M. et al., CU VOCAL: Corpus-Based Syllable Concatenation for Chinese Speech Synthesis Across Domains and Dialects. *Proceedings of the 7th International Conference on Spoken Language Processing*, pp.2373-2376, Denver, Colorado USA (2002).
21. Meng, H. et al., Recent Enhancement in CU VOCAL for Chinese TTS-Enabled Applications, *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland (2003).
22. Lee, Tan et al., Micro-prosodic Control in Cantonese Text-to-Speech Synthesis. *Proceedings of the 6th European Conference on Speech Communication and Technology*, pp.1855-1858, Budapest, Hungary (1999).
23. Li, Yujia et al., Acoustical F0 analysis of continuous Cantonese speech. *Proceedings of the 3rd International Symposium on Chinese Spoken Language Processing*, pp.127-130, Taipei, Taiwan (2002).
24. Luk, Po Chui et al., Grammar Partitioning and Parser Composition for Natural Language Understanding. *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China (2000).
25. Weng, F. et al., Parsing a Lattice with Multiple Grammars. *Proceedings of the Sixth International Workshop on Parsing Technologies (2000)*.
26. Xu, Kui et al., Multi-Parser Architecture for Query Processing. *Proceedings of the 7th European Conference on Speech Communication and Technology*, pp. 1077-1080, Aalborg, Denmark (2001).
27. Luk, P. C. et al., Automatic Grammar Partitioning for Syntactic Parsing. *Proceedings of the International Workshop on Parsing Technologies (2001)*.
28. Meng, Helen et al., GLR parsing with multiple grammars for natural language queries. *ACM Transactions on Asian Language and Information Processing*. Volume 1, pp.123-144 (June 2002).
29. Siu, K. C. and Meng, Helen M., Semi-Automatic Grammar Induction for Bi-directional English-Chinese Machine Translation. *Proceedings of the 7th European Conference on Speech Communication and Technology*, pp. 2749-2752, Aalborg, Denmark (2001).
30. Wong, C. C. and Meng, H., Improvements on a Semi-Automatic Grammar Induction Framework. *Proceedings of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 288-291, Trento, Italy (2001).
31. Meng, H. et al., Learning Belief Networks for Language Understanding. *Proceedings of the 1999 International Workshop on Automatic Speech Recognition and Understanding, Keytones*, Colorado USA (1999).
32. Meng, Helen M. et al., To Believe is to Understand. *Proceedings of the 6th European Conference on Speech Communication and Technology*, pp. 2015-2018, Budapest, Hungary (1999).
33. Meng, Helen M. et al., The use of Belief Networks for mixed-initiative dialog modeling. *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China (2000).
34. Chan, S. F. and Meng, H., Interdependencies among Dialog Acts, Task Goals and Discourse Inheritance in Mixed-Initiative Dialog. *Proceedings of the 2002 Human Language Technology Conference* (2002).
35. Meng, H. et al., Natural Language Response Generation in Mixed-Initiative Dialogs using Task Goals and Dialog Acts. *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland (2003).
36. Siu, K. C. et al., Example-based bi-directional Chinese-English machine translation with semi-automatically induced grammars. *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland (2003).
37. Li, Y. C. et al., Query Expansion using Phonetic Confusions for Chinese Spoken Document Retrieval. *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, pp. 89-93, Hong Kong, China (2000).
38. Meng, H. M. et al., Speech Retrieval with Video Parsing for Television News Programs. *Proceedings of the 2001 IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 1401-1404, Salt Lake City, Utah USA (2001).
39. Meng, H. et al., Mandarin-English Information: investigating translingual speech retrieval. *Final Report for NSF Summer Workshop 2000*, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, USA (October 2000).
40. Meng, H. M. et al., Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-language Spoken Document Retrieval. *Proceedings of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 311-314, Trento, Italy (2001).
41. Lo, W. K., Cross-language spoken document retrieval using HMM-based retrieval model with multi-scale fusion. *ACM Transactions on Asian Language Information Processing* (submitted).
42. Meng, Helen M. et al., Multi-Scale Audio Indexing for Chinese Spoken Document Retrieval. *Proceedings of the 6th International Conference on Spoken Language Processing*, pp. 101-104, Beijing, China (2000).
43. Li, Y. C. and Meng, Helen M., Document Expansion using a Side Collection for Monolingual and Cross-langauge Spoken Document Retrieval. *Proceedings of the ISCA Multilingual Spoken Document Retrieval Workshop (2003)*.
44. Lo, Wai-Kit et al., Multi-scale document expansion in English-Mandarin cross-language spoken document retrieval. *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland (2003)
45. Lo, W. K. et al., Multi-scale spoken document retrieval for Cantonese broadcast news. *International Journal on Speech Technology*. (April 2004).
46. Hui, Pui Yu et al., Automatic Story Segmentation for Spoken Document Retrieval. *Proceedings of the 10th IEEE International Conference on Fuzzy Systems*, pp. 1319-1322, Melbourne, Australia (2001).
47. Hui, Pui Yu et al., Two robust methods for Cantonese spoken document retrieval. *Proceedings of the ISCA Multilingual Spoken Document Retrieval Workshop*, pp. 7-12 (2003).
48. Hui, Pui Yu et al., Multimedia Fusion in Automatic Extraction of Studio Speech Segments for Spoken Document Retrieval. *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 725-728 (2003).
49. Meng, H. et al., CU FOREX: A Bilingual Spoken Dialog System for Foreign Exchange Inquiries. *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey (2000).
50. Meng, Helen et al., ISIS: A Multilingual Spoken Dialog System developed with CORBA and KQML agents. *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China (2000).
51. Lee, T. et al., Spoken Language Resources for Cantonese Speech Processing. *Speech Communications*. Volume 36, pp.327-342 (March 2002).
52. Lo, W. K. et al., Design, compilation and processing of CUCall: a set of Cantonese spoken language corpora collected over telephone networks. *Proceedings of Research on Computational Linguistics Conference XIV*, pp. 193-212, Tainan, Taiwan (2001).
53. Meng, Helen M. et al., CU VOCAL Web Service: A text-to-speech synthesis Web service for voice-enabled web-mediated applications. *Proceedings of the Twelfth International World Wide Web Conference 2003*, Budapest, Hungary (2003).
54. Meng, Helen M. et al., The "Author Once, Present Anywhere" (AOPA) software platform, *Proceedings of the 2003 Hong Kong International Computer Conference*, Hong Kong, China (2003).

LEE Tan

Lee Tan received his BSc and MPhil degrees in Electronics in 1988 and 1990 respectively, and his PhD degree in Electronic Engineering in 1996, all from the Chinese University of Hong Kong (CUHK). Since 1999, he has been an Assistant Professor at the Department of Electronic Engineering, CUHK. In 1997, Lee Tan was a guest researcher at the Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Sweden.

Lee Tan has been working on speech signal processing and related research topics for many years. He has published about 50 papers in international journals and conference proceedings. Tan Lee has initiated and coordinated a number of pioneer projects on the research and development of Chinese spoken language technologies. The projects have been receiving substantial funding support from the Hong Kong Research Grants Council (RGC) and the Innovation and Technology Fund (ITF). The project deliverables have been licensed widely for both academic research and commercial applications.

Lee Tan is a member of the Institute of Electrical and Electronic Engineers (IEEE) and a member of the International Speech Communication Association (ISCA). He is currently the Vice-Chairman of the IEEE Hong Kong Chapter of Signal Processing.

Helen MENG

Helen M Meng received the SB, SM and PhD degrees, all in electrical engineering, from the Massachusetts Institute of Technology. She has been a Research Scientist at the MIT Spoken Language Systems Group, where she worked on multilingual conversational systems. She is currently an Associate Professor in the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong, where she established and directs the Human-Computer Communications Laboratory. Her research interest is in the area of multimodal human-computer interaction via multilingual spoken language systems, and related core technologies in speech recognition, speaker authentication, natural language understanding, discourse and dialog modeling, language generation and speech synthesis. She also works on machine translation and translingual speech retrieval technologies.

Helen is an elected member and the Conference Board Representative of the IEEE Signal Processing Society Speech Technical Committee. She serves as an appointed member of the Hong Kong Information Technology Services Department Working Group on Chinese Information Processing. She is also the immediate past Chairman of the Association for Computing Machinery (Hong Kong Chapter) and currently serves on its Executive Committees as well as that of the IEEE Hong Kong Chapter of Signal Processing. Helen is an elected member of Sigma Xi and the International Speech Communication Association. She also serves on the editorial board of Computer Speech and Language journal.

W K LO

Lo Wai-kit received his BEng (1st Class Hons), MPhil and PhD degrees, all in electronic engineering, from the Chinese University of Hong Kong, in 1994, 1996 and 2002 respectively. He is currently a researcher at the Spoken Language Translation Laboratories, Advanced Telecommunications Research Institute in Japan.

Lo Wai-Kit was a Project Coordinator at the Department of Electronic Engineering, the Chinese University of Hong Kong from 1997 to 2002, where he engaged in research and development of spoken language corpora, speech synthesis, speech recognition and spoken document retrieval. In 2000, he participated in the Summer Research Workshop in Johns Hopkins University and engaged in the translingual speech retrieval project - MEI: Mandarin-English Information. From 2002 to 2003, he worked at the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong, as a Project Engineer and worked on the application of speech technologies to enrich Internet technologies and content management. His current research interest includes adaptive utterance rejection for speech recognition, cross-language and multimedia document retrieval.

Lo Wai-kit is a member of the Institute of Electrical and Electronic Engineers (IEEE) and a member of the International Speech Communication Association (ISCA).

P C CHING

P C Ching received the BEng (Hons) and PhD degrees in electrical engineering and electronics from the University of Liverpool, UK, in 1977 and 1981, respectively. From 1981 to 1982 he worked as a research officer at the School of Electrical Engineering, University of Bath, UK. During 1982-1984, he was a Lecturer in the Department of Electronic Engineering of the then Hong Kong Polytechnic University. Since 1984 he has been with the Chinese University of Hong Kong, where he is presently Dean of Engineering and a chair professor at the Department of Electronic Engineering. He has taught courses in digital signal processing, stochastic processes, speech processing, and communication systems. His research interests include adaptive filtering, time delay estimation, statistical signal processing, and hands-free speech communication.

Ir Dr Ching has actively participated in many professional activities. He was the Chairman of the IEEE Hong Kong Section in 1993-94. He has been a member of the Signal Processing Theory and Methods Technical Committee of the IEEE Signal Processing Society since 1996. He presently serves as an associate editor for the Signal Processing Letters, and was also an associate editor for the IEEE Transactions on Signal Processing from 1997 till 2000. He is a member of the Accreditation Board and Fellowship Committee of the Hong Kong Institution of Engineers and a Council member of IEE, UK. He has been involved in organizing many international conferences including the 1997 IEEE International Symposium on Circuits and Systems where he was the Vice-Chairman, and the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing where he served as the Technical Program Co-chair. Ir Dr Ching is a Fellow of IEE and HKIE, and a senior member of IEEE.