INITIAL DEVELOPMENT TOWARDS A TRILINGUAL SPEECH INTERFACE FOR FINANCIAL INFORMATION INQUIRIES

Helen M. Meng
Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China
hmmeng@se.cuhk.edu.hk

Abstract This paper describes our initial effort in developing a trilingual speech interface for financial information inquiries. Our foreign exchange inquiry system consists of: (i) monolingual and trilingual speech recognizers, which receive the user's spoken inputs in the form of microphone speech; (ii) a real-time data capture component which continuously updates a relational database from a financial data satellite feed; and (iii) a trilingual speech generation component, which generates English and Chinese text based on the raw financial data. The generated text is then transformed into spoken presentations. English text is processed by the FESTIVAL synthesizer system. Chinese text is sent to our syllable-based synthesizer, which employs a concatenative resequencing technique to produce spoken presentations in Putonghua or Cantonese. The speech interface is augmented with a visual display which aims to provide feedback to the user at all times during an interaction. Within the restricted scope of foreign exchange (FOREX), our recognition performance accuracies remain above 93%. Confusions across languages contributed significantly to our recognition errors, but most are confusions between the same currency/country names spoken in different languages. These errors are not detrimental with respect to data retrieval. Our concatenative re-sequencing technique reports the date, time and exchange rates of the input currency pair. A demonstration can be found at http://www.se.cuhk.edu.hk/hccl/demos/.

1. Introduction

This paper describes our initial effort in developing a trilingual speech interface for financial information inquiries. Our speech interfaces supports the languages of Hong Kong – Cantonese, Putonghua and English. (The first two are dialects of Chinese). We have selected foreign exchange (FOREX) as our application domain. Hong Kong has one of the largest foreign exchange trading centers in the world, and many in its populace are familiar with the globally traded currencies. However, the trilingual populace often refer to some currencies in English, and others in Cantonese or Putonghua. Intermixing the use of the three languages is common in the region's daily speech. We feel that a trilingual recognizer may offer greater flexibility than three monolingual recognizers, because the former does not restrict the user to a pre-specified language throughout the interaction. This project is an initial attempt in exploring the use of a trilingual speech interface for the FOREX domain.

Our foreign exchange inquiry system consists of: (i) monolingual and trilingual speech recognizers, which receive the user's spoken input in the form of microphone speech; (ii) a real-time data capture component, which continuously updates financial data from the Reuters satellite feed and retrieves the relevant data based on the user's request; and (iii) a trilingual speech generation component, which generates English and Chinese text based on the raw financial data. The generated text is then transformed into spoken presentations. English text is processed by the FESTIVAL synthesizer system. Chinese text is sent to our syllable-based synthesizers. We have devised a technique of concatenative resequencing to produce natural-sounding spoken presentations in Putonghua or Cantonese. Figure 1 shows the overall architecture of our system. At the beginning of an interaction session (which consists of a spoken input and a synthesized output), we provide the user with a selection that includes the trilingual recognizer and the three monolingual recognizers. If the user selects a monolingual recognizer, the system expects to receive spoken input which belong to the selected language, and produces synthesized output in the selected language as well. If the user selects the trilingual recognizer, the system allows the user to intermix the languages, and synthesizes in one of the input languages. This constitutes our trilingual speech interface (recognition and generation) for accessing foreign exchange data that is continuously captured from the Reuters satellite feed.

(Figure 1 here).

Similar (multilingual) speech interfaces have previously been developed at a number of research sites. Elaborate conversational systems have been demonstrated in several restricted domains that correspond to real applications. The domains range from air travel (e.g., the Air Travel Information System, ATIS) (Price 1990), train schedules (e.g. Railway Telephone Information Service, RAILTEL) (Lamel et al., 1998), restaurant guides (e.g., The Berkeley Restaurant Project, BeRP) (Jurafsky et al., 1994), ferry timetables (e.g., WAXHOLM) (Blomberg et al., 1993), weather (Zue et al., 1997) and and electronic automobile classifieds (Meng et al., 1996). The languages concerned include English and multiple European languages. Similar research was also initiated recently for Mandarin Chinese in the navigation (Wang et al., 1997) and telephone directory assistance domains (Yang and Lee, 1998). Our current work is an initial effort to use three languages (English, Putonghua and Cantonese) simultaneously for the foreign exchange domain.

2. The Visual Interface

Our speech interface is enhanced with a Java applet visual display. Push buttons enable users to make various kinds of selections. For example, users can select one of three monolingual recognizers (Cantonese, Putonghua or English), or the trilingual recognizer. Users can also select audio help instructions in any of the three languages. Another pair of push buttons (labeled 'First Currency' and 'Second Currency') invokes the recording process for each of the two currencies for which the user wishes to find the exchange rate. These currency push buttons prompt the user with an audio signal followed by a tone. Such prompting ease the process of end-point detection during recording. The layout and flow of the visual interface are depicted in the screen dumps in Figure 2. At every instant during the interaction, we utilize color highlighting to provide feedback to the user with regard to the processing status of the system, as well as the outcome of recognition.

(Figure 2 here).

3. Speech Recognition

As mentioned previously, we have developed four speech recognizers. The trilingual recognizer is derived from consolidating the vocabulary sets of the three monolingual recognizers. We have approximately 270 lexical entries in all, covering 36 international currencies. The lexical items include names of currencies, countries and their combinations. We also cover the common (colloquial) variations by which users may refer to the currencies, e.g., "Deutsche Marks," "German Marks" and "D-Marks." The recognizers are HMM-based (HTKBook, 1999), using word models with single Gaussian probability distributions. The HMMs are configured with 5 states per syllable for each Cantonese or Putonghua word model, and 8 states per syllable for each English word model. These state settings were determined by optimization based on training data. The trilingual recognizer encompasses all the models from the three languages. An excerpt from our vocabularies is shown in Table 1. Our experimental corpus consists of gender-balanced, microphone recordings collected from 60 speakers. Hence each lexical entry has sixty tokens, one from every speaker. Our recognizers are trained on the data from 50 speakers and tested on the remaining 10 speakers. The experiment is then repeated with another 10 speakers for testing and the remaining speakers for training. In this way, we can conduct 6 experimental runs. The overall performance accuracy is obtained from averaging over these 6 runs.

(Table 1 here).

3.1 Recognition Results

Figure 3 shows the performance of our speech recognizers. Results are based on the test sets for each monolingual recognizer. As we compare the performance between the monolingual and trilingual recognizers for each language, we observe performance degradation due to a larger number of vocabulary items present in the trilingual recognizer. Absolute degradation in performance was 1.4% for English, 2.1% for Cantonese and 5.0% for Putonghua. The degradation for Putonghua is particularly large, possibly because the monolingual recognizer only has a vocabulary size of 40, compared to the trilingual recognizer, which has 272. The Cantonese vocabulary size is 128, and the English vocabulary size is 104.

We have also performed another experiment investigating the effect of adding filler words to our system. A filler word is defined to be the set "um", "ah", "la" etc., which are characteristic of spontaneous speech. Speakers commonly insert them at the beginning or end of their utterances. Injection of filler words led to further performance degradations, ranging from 1 to 4% among the recognizers. Confusion between the filler words and the keyword was the main cause for such degradations.

3.2 Error Analyses

Table 2 shows some examples of recognition errors, organized according to confusions *within* a language (intra-language confusions), as well as *across* languages (cross-language confusions). In the table, we have included the language labels E*=English, C*=Cantonese, and P*=Putonghua. For Cantonese, we have also included the phonetic transcription based on the LSHK standard (LSHK, 1997). For example, the first reference currency is the Cantonese version of the South Korean Won, pronounced as “waan4”, where the number ‘4’ indicates the fourth tone, out of the 9 possible tones in Cantonese. Similarly, for Putonghua, we have included the phonetic transcription based on Han Yu Pin Yin (Han Yu Pin Yin, 1999), e.g., the fourth reference currency is the Putonghua version of the US dollar, pronounced as two tonal syllables – “mei3_chao1”. There are a total of 4 tones (plus a light tone) in Putonghua.

(Table 2 here).

The breakdown of these confusions in the errors of the trilingual recognizer is shown in Table 3.

(Table 3 here).

From the two tables, we observe that the additional vocabulary items in the trilingual recognizer led to cross-lingual confusions. From Table 3, we see that cross-lingual confusions account for approximately 33% of the errors in Cantonese recognition, 80% of the errors in Putonghua recognition, and 44% of the errors in English recognition. Cross-lingual confusions may be classified into two categories:

(I) The currency was correct, but the language was wrong – most of the examples (except for Row 3) in Table 2 belong to this category. Such errors are common, possibly because the pronunciation in one language is often derived from the pronunciation in another language.

(II) Both the currency and language were wrong – an example is row 3 in Table 2, where the Thai Baht (in Cantonese) was mistaken for the British Pound in English. The two pronunciations are similar.

3.3 Remarks on Trilingual Recognition

We feel that the trilingual recognizer offers much flexibility to the user in terms of intermixing languages, despite the degradation in performance due to the enlarged vocabulary size, when compared to the monolingual recognizers. Cross-language confusion errors in which the identity of the currency remains the same may be harmless towards the task of database retrieval for the exchange rates between two currencies. The exchange rates can then be displayed on screen, and / or presented in spoken form by means of speech synthesis. It should be noted, however, that the speech recognizers form the front end to receive the user's input, and when the trilingual recognizer is used, it also performs the task of implicit language identification. The identified language will be selected by the speech generation component for synthesis. Hence an error in language identification will affect the synthesized response as well.

4. Speech Generation

The outputs from speech recognition create a simple currency-pair representation, which consists of a language identifier, and the recognition outputs for the pair of currencies. The representation invokes the processes of real-time data retrieval (exchange rates), as well as speech generation. Our speech generation component performs the tasks of text generation (for both English and Chinese), as well as syllable-based concatenative resequencing for Putonghua and Cantonese. As for English, we send our generated English text directly to the FESTIVAL system (Black and Taylor, 1997).

Since we are performing synthesis for a restricted domain, our task complexity is lower when compared to synthesis for free-form running text. Hence we aim to produce synthesized speech with higher naturalness. The various processes in our speech generation component are described in the following subsections.

4.1 Text Generation

Our text generation procedure aims to present raw data in the form of a presentable sentence. The information provided includes the date, time, currencies, bid/ask prices and some system messages. Numerals were handled with special care for each of the languages. For example, the text for the year '1999' was generated with English groupings as "nineteen ninety nine" and serially in Chinese as "一九九九". This is different from decimals such as '3.456', generated serially for both languages – "three point four five six" in English, and "三點四五六" in Chinese. This, in turn, is different from a price of '123', generated as "one hundred and twenty three" in English, and "一百二十三" in Chinese. In the last case, words need to be inserted to indicate the order of magnitude of the number.

4.2 Concatenative Resequencing

As mentioned previously, the generated English text was fed directly to FESTIVAL, and we did not develop concatenative synthesis for English. However, we did include SABLE markings (Sproat et al., 1998) to provide better specification of the pronunciation of some vocabulary items in our domain.

Synthesis for Putonghua and Cantonese shared the same Chinese text generation output.[1] Our approach for concatenative synthesis consisted of the following steps:

(i) Each Chinese character in the text sentence was mapped to its appropriate *tonal syllable*, according to our Cantonese and Putonghua lexicons. However, in Putonghua, there are rules for *tone change* governing the tone realization in continuous speech. These are known as *tone sandhi rules* and many are well documented in (Lee et al., 1989). For example, consider a syllable with the third tone in Putonghua, such as the number *nine*, '九', pronounced as "jiu3". When a pair of tone 3 syllables occur in sequence, e.g., '九九' (nine nine), the first syllable will be changed from tone 3 to tone 2, i.e., "jiu3 jiu3" becomes "jiu2 jiu3" in continuous speech. Furthermore, if there are three or more tone 3 syllables occurring sequentially, e.g. '一九九九' (one nine nine nine), the situation becomes more complex – instead of "yi1 jiu3 jiu3 jiu3", we get "yi1 jiu3 jiu2 jiu3".

[1] We are aware that wordings in Putonghua and Cantonese do differ. However, for the sake of simplicity, and within the constraints of our domain, we find that using the same Chinese text generation output suffices for both dialects of Chinese. The generated text verbalizes the information requested by the user, and the text becomes the

Generally tone realization of sequential tone 3 syllables is complex, and may depend on the syntactic boundaries of a sentence. However, within the scope of the FOREX application, we found that the sandhi rules mentioned above are sufficient for our purposes.

(ii) For every tonal syllable in the sentence, we selected the "appropriate" syllable wave file from a previously prepared wave bank. Details of the syllable selection process are provided in the next subsection. The wave files were then concatenated in order to form the synthesized sentence.

4.2.1 Coarticulation and the Use of Distinctive Features

In designing an algorithm for syllable-based concatenation, special attention should be paid to the effect of coarticulation in continuous speech. It is widely known that context dependence heavily affects the acoustic realization of a syllable. Consider the character 七 (i.e., the number '7'), which is pronounced as 'cat1' in Cantonese. In the context of "六七八" (i.e., the number sequence '678', pronounced as "luk6 cat1 baat3"), the syllable 'cat1' has a *left velar* context, and a *right labial* context. Due to coarticulation, speakers tend to assimilate the alveolar closure of the syllable 'cat1' with the right labial, resulting in the production of 'cab1' (e.g., 輯). Hence if we were to extract the syllable wave file for 七 ('7') from the spoken phrase "六七八" ('678'), and use it to synthesize "八七六" ('876', correct pronunciation being "baat3 cat1 luk6"), the resulting waveform will sound like "八輯六" (i.e. "baat3 cab1 luk6", a nonsense syllable sequence), which sounds incorrect when compared to natural speech.

It is obvious that the contextual characteristics of a syllable are important for concatenation. However, if we were to consider both the left and the right syllable contexts simultaneously, we may need to store N^3 wave files for every syllable in our "wave bank", where N is the number of unique syllables in the language.[2] In order to minimize the size of our wave bank, we decided to consider *only* the place of articulation of the (optional) coda of the left syllable, and the (optional) onset of the right syllable. This facilitates more sharing and hence a smaller size for our wave bank. The left and right contexts are represented with a two-digit encoding which represents the left and right distinctive features (Stevens, 1972):

Right Context: LABIAL, ALVEOLAR, VELAR, GLIDE, LATERAL, PALATAL AND NEUTRAL (for the aspirant and syllables without onsets)

Left Context: ALVEOLAR, VELAR, NEUTRAL (for syllables without codas)

input to the speech generation components for the respective dialects (Cantonese or Putonghua).

We can see that the variation in the left context is much more constrained, which is characteristic of Chinese syllables.

4.2.2 Wave Bank Preparation and Concatenative Resequencing

Our wave bank stores a large set of syllable wave files which are extracted from continuous speech. We have designed a series of recording prompts that fully cover the possible context variants of the syllables in our vocabularies (for Putonghua and Cantonese). We collected the recordings from two female speakers, one for each language. The recordings were subsequently segmented to yield syllable wave files. Segmentation was achieved by forced alignment, and postprocessed by hand, with reference to the spectrograms. We have a total of 2400 syllable segments in all.

For a given textual input (with transcribed syllables), our synthesis algorithm concatenates the syllable wave files sequentially from left to right. The *unit selection* process ensures that the syllable variant with matching left and right contexts is chosen. We also inserted short pauses in between phrases, and long pauses in between sentences. Both the unit selection process and the insertion of pauses were found to be important contributing factors towards naturalness in the synthesized outputs.

5. End-to-end Interaction

Our speech recognizers form the front end of our foreign exchange inquiry system, to receive the user's spoken input. The system subsequently retrieves the relevant data, and responds by speech generation. The system is fully integrated end-to-end. Should the trilingual recognizer be selected at the beginning of an interaction session, and if both of the speaker's utterances (for the two currencies) are in the same language, the speech generation component will respond in the corresponding language. Alternatively, if the languages of the two utterances differ, the system simply selects one of the two languages to respond.

6. Summary and Future Work

This paper reports on our initial effort in developing a trilingual interface to support financial information inquiries. The three languages of interest are those most commonly used in Hong Kong, namely, Cantonese, Putonghua and English. Based on the foreign exchange domain, we have presented our speech recognition results for three monolingual speech recognizers, and a trilingual recognizer. Error analyses indicated that cross-language

[2] N^3 is an approximate upper bound. N is about 1800 for Cantonese, and about 1200 for Putonghua.

confusions generally occur for the same currency name, which is less detrimental from the perspective of task completion. We have also presented the design of our speech generation component. While English output is produced by the FESTIVAL system, synthesis of Cantonese and Putonghua are based on syllable concatenation. We have devised a procedure of syllable concatenative resequencing, which captures the left and right articulatory contexts using distinctive features to achieve enhanced naturalness in the synthesizer output. At present, we are continuing this work along the lines of natural language processing and telephone speech recognition. The use of a markup language, akin to VoiceXML (VoiceXML, 1999), should be helpful for such future development. Also, since Hong Kong has a fully digital telecommunications backbone and a high penetration of fixed-line and mobile phones (GSM, CDMA, PCS, and TDMA) and we may need to include speech data from the various networks when developing our speech recognizers.

Acknowledgments

The author wishes to thank the reviewers for their constructive comments in improving this paper. We acknowledge the undergraduate project students (Groups H and N, 1998-1999) of the Department of Systems Engineering and Engineering Management for various implementational assistance. Also, thanks to Professor P. C. Ching, Professor Tan Lee and Mr. Wai-Kit Lo of the Digital Signal Processing Laboratory, Department of Electronic Engineering at The Chinese University of Hong Kong for their support of this project. We are grateful to Reuters Hong Kong for donating the real-time satellite feed for this project.

References

Black, A. and Taylor, P. (1997). Festival Speech Synthesis System: system documentation (1.1.1) Technical Report HCRC/TR-83. Human Communication Research Center. Universities of Edinburgh and Glasgow.

Blomberg, M., Carlson, R., Elenius, K., Granstrom, B., Gustafson, J., Hunnicutt, S., Lindell R. and Neovius, L. (1993). An Experimental Dialogue System: WAXHOLM. *Proceedings of the European Conference on Speech Communication and Technology.* (pp. 1867-1870). Berlin: ESCA.

Han Yu Pin Yin. (1999). New Chung Wah Dictionary. Hong Kong: Chung Wah (Hong Kong) Book Company Ltd..

HTK Book (1999). (http://www.entropic.com/products/htk/HTKBook)

Jurafsky, D., Wooters, C. , Tajchman, G., Segal, J. , Stolcke, A., Fosler E. and Morgan, N. (1994). The Berkeley Restaurant Project. *Proceedings of the International Conference on Spoken Language Processin*g. (pp. 2139-2142). Yokohama: Acoustic Society of Japan.

Lamel, L., Bennacef, S., Gauvain, J., Dartigues, H. and Temem J. (1998). User Evaluation of the MASK Kiosk. *Proceedings of the International Conference on Spoken Language Processing*. Sydney: CD-ROM published by Causal Productions.

Lee, L. S., Tseng, C. Y., and Ming, O. Y., (1989). The Synthesis Rules in a Chinese Text-to-Speech System. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, No. 9, September 1989. (pp. 1309-1320).

LSHK Linguistic Society of Hong Kong (香港語言學會), Hong Kong Jyut Ping Character Table (粵語拼音字表). (1997). Hong Kong: Linguistic Society of Hong Kong Press.

Meng, H., Busayapongchai, S. , Glass, J. , Goddeau, D. , Hetherington, L., Hurley, E., Pao, C., Polifroni, J., Seneff , S. and Zue, V. (1996). A Conversational System in Automobile Classifieds Domain. *Proceedings of the International Conference on Spoken Language Processing*. Philadelphia: CD-ROM published by University of Delaware and Alfred I. DuPoint Institute.

Price, P. (1990). Evaluation of Spoken Language Systems: the ATIS Domain. *Proceedings of the DARPA Speech and Natural Language Workshop* (pp. 91-95). San Mateo, CA: Morgan Kaufman Publishers.

Wang, C., J. Glass, H. Meng, J. Polifroni, S. Seneff and V. Zue (1997). YINHE: A Mandarin Chinese Version of the GALAXY System. *Proceedings of the European Conference on Speech Communication and Technology.* Patras: Patras: CDRom published by ESCA

Sproat R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K. and Eddington, M. (1998). SABLE: a standard for TTS markup. *Proceedings of International Conference on Spoken Language Processing.* Sydney: CD-ROM published by Causal Productions.

Stevens, K. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David and P. B. Denes (Eds.), *Human Communication: A Unified View.* (pp. 51-66). New York: McGraw-Hill.

VoiceXML (1999). (http://www.voicexml.org).

Yang, Y. J. and Lee, L. S. (1998). A Syllable-based Chinese Spoken Dialogue System for Telephone Directory Services Primiarily Trained with a Corpus. *Proceedings of the International Conference on Spoken Language Processing.* Sydney: CD-ROM published by Causal Productions.

Zue, V., Seneff, S., Glass, J., Hetherington, L., Hurley, E., Meng, H., Pao, C., Polifroni, J., Schloming, R. and Schmid, P. (1997). From Interface to Content: Translinngual Access and Delivery of On-Line Information. *Proceedings of the European Conference on Speech Communication and Technology.* Patras: CDRom published by ESCA.

Footnotes:

[1] We are aware that wordings in Putonghua and Cantonese do differ. However, for the sake of simplicity, and within the constraints of our domain, we find that using the same Chinese text generation output suffices for both dialects of Chinese. The generated text verbalizes the information requested by the user, and the text becomes the input to the speech generation components for the respective dialects (Cantonese or Putonghua).

[2] N^3 is an approximate upper bound. N is about 1800 for Cantonese, and about 1200 for Putonghua.

Contact Information:

Helen Meng

Human-Computer Communications Laboratory

Department of Systems Engineering and Engineering Management

The Chinese University of Hong Kong

Shatin, N.T., Hong Kong, China

hmmeng@se.cuhk.edu.hk

Fax: +852.2603.5505

Phone: +852.2609.8327

INITIAL DEVELOPMENT TOWARDS A TRILINGUAL SPEECH INTERFACE FOR FINANCIAL INFORMATION INQUIRIES

Helen M. Meng

Human-Computer Communications Laboratory

Department of Systems Engineering and Engineering Management

The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China

hmmeng@se.cuhk.edu.hk

Abstract This paper describes our initial effort in developing a trilingual speech interface for financial information inquiries. Our foreign exchange inquiry system consists of: (i) monolingual and trilingual speech recognizers, which receive the user's spoken inputs in the form of microphone speech; (ii) a real-time data capture component which continuously updates a relational database from a financial data satellite feed; and (iii) a trilingual speech generation component, which generates English and Chinese text based on the raw financial data. The generated text is then transformed into spoken presentations. English text is processed by the FESTIVAL synthesizer system. Chinese text is sent to our syllable-based synthesizer, which employs a concatenative resequencing technique to produce spoken presentations in Putonghua or Cantonese. The speech interface is augmented with a visual display which aims to provide feedback to the user at all times during an interaction. Within the restricted scope of foreign exchange (FOREX), our recognition performance accuracies remain above 93%. Confusions across languages contributed significantly to our recognition errors, but most are confusions between the same currency/country names spoken in different languages. These errors are not detrimental with respect to data retrieval. Our concatenative re-sequencing technique reports the date, time and exchange rates of the input currency pair. A demonstration can be found at http://www.se.cuhk.edu.hk/hccl/demos/.

Keywords: speech recognition, multilingual, concatenative resequencing.

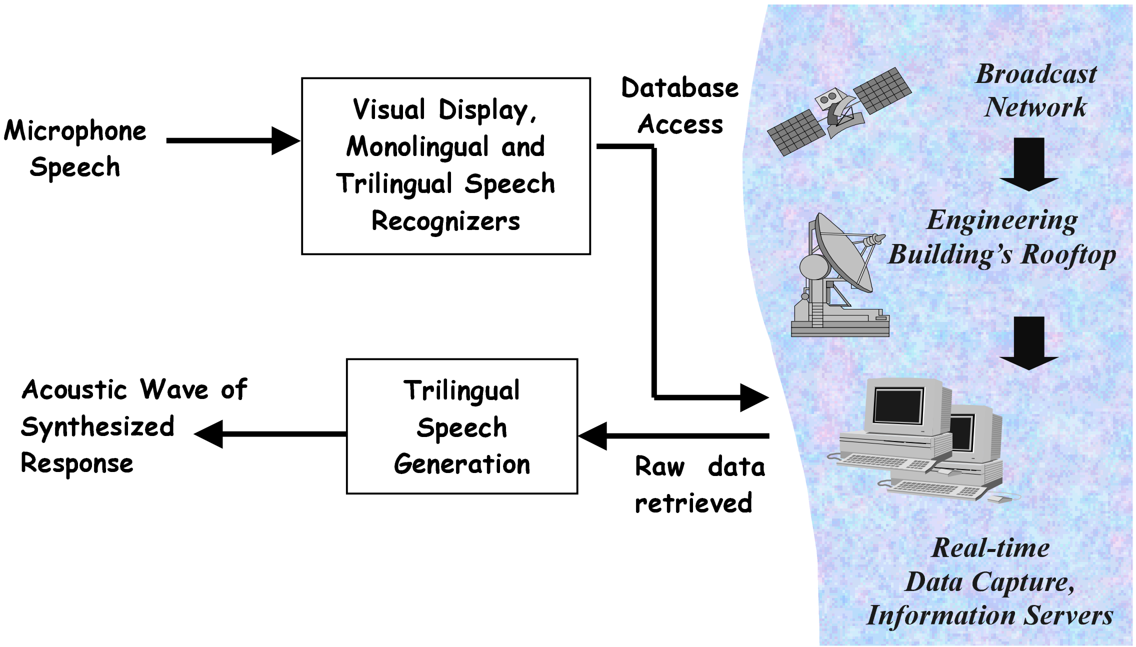
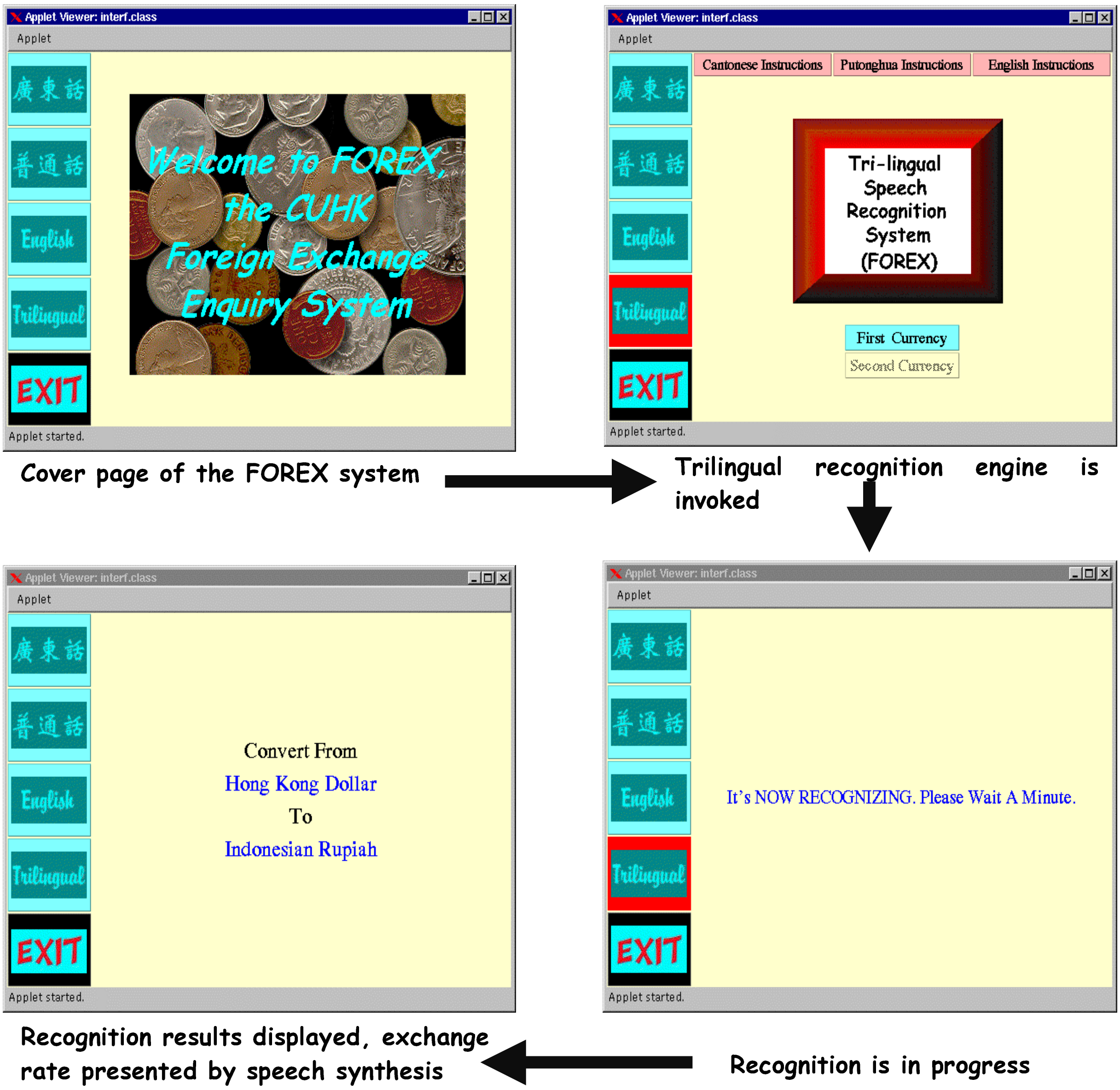

Microphone Speech
Visual Display, Monolingual and Trilingual Speech Recognizers
Database Access
Broadcast Network
Engineering Building's Rooftop
Acoustic Wave of Synthesized Response
Trilingual Speech Generation
Raw data retrieved
Real-time Data Capture, Information Servers

Figure 1. Overall system architecture of the foreign exchange inquiry system.



Cover page of the FOREX system

Welcome to FOREX, the CUHK Foreign Exchange Enquiry System

Trilingual recognition engine is invoked

Tri-lingual Speech Recognition System (FOREX)

Recognition is in progress

It's NOW RECOGNIZING. Please Wait A Minute.

Recognition results displayed, exchange rate presented by speech synthesis

Convert From
Hong Kong Dollar
To
Indonesian Rupiah

Figure 2. Interface layout and system flow of the FOREX inquiry system.

Table 1. An excerpt of our vocabulary showing the names of several countries and currencies, together with their pronunciations (and alternate pronunciations) across the languages of English, Cantonese and Putonghua.

Reuters Instrument Code	Country	Currency	Country and Currency	English Pronunciations	Cantonese Pronunciations	Putonghua Pronunciations
ATS	Austria 奧地利	Schilling 先令, 司令	Austrian Schilling 奧地利先令, 奧地利司令	ao s t r y ax, sh ih l ix ng, ao s t r y ax n sh ih l ix ng	ou3 dei6 lei6, sin1 ling6, si1 ling6, ou3 dei6 lei6 sin1 ling6, ou3 dei6 lei6 si1 ling6	ao4 di4 li4 xian1 ling4
AUD	Australia 澳洲	Dollar 元	Australian Dollar 澳洲元, 澳元, 澳幣, 澳洲紙	ao s t r ey l y ax, d aa l er, ao s t r ey l y ax n d aa l axr	ou3 zau1, ou3 zau1 jyun4 ou3 jyun4, ou3 bai6, ou3 zau1 zi2	ao4 zhou1 yuan2
BEF	Belgium 比利時	Franc 法郎	Belgian Franc 比利時法郎	b eh l jh ax m, f r ae ng k, b eh l jh ax n f r ae ng k	bei2 lei6 si4, faat3 long4, bei2 lei6 si4 faat3 long4	bi3 li4 shi2 fa3 lang2
CAD	Canada 加拿大元	Dollar 元	Canadian Dollar 加拿大元 加元, 加紙, 加幣, 加拿大紙	k ae n ax d ax, d aa l er, k ax n ey d iy ax n d aa l er	gaa1 naa4 daai6 jyun4, gaa1 jyun4, gaa1 zi2, gaa1 bai6, gaa1 naa4 daai6 zi2	jia1 na2 da4 yuan2
CHF	Switzerland 瑞士	Franc 法郎	Swiss Franc 瑞士法郎	s w ih t z ax l ax n d, f r ae ng k, s w ih s f r ae ng k	seoi6 si6, faat3 long4, seoi6 si6 faat3 long4	rui2 shi4 fa3 lang2
CNY	China 中國	Renminbi 人民幣	Chinese Renminbi 中國人民幣	ch ay n ax, r ah m ix n b iy, ch ay n iy z r eh n m ih n b iy	zung1 gwok3, jan4 man4 bai6, zung1gwok3 jan4 man4 bai6	ren2 min2 bi4
DEM	Germany 德國	Mark 馬克	Deutschmark D-Mark, German Mark 德國馬克, 德國幣, 西德馬克	jh er m ax n iy, d iy-m aa r k d oy ch m aa r k jh er m ax n m aa r k	dak1gwok3, maa5 hak1, dak1 gwok3 bai6, sai1 dak1 maa5 hak1	de2 guo2 ma3 ke4

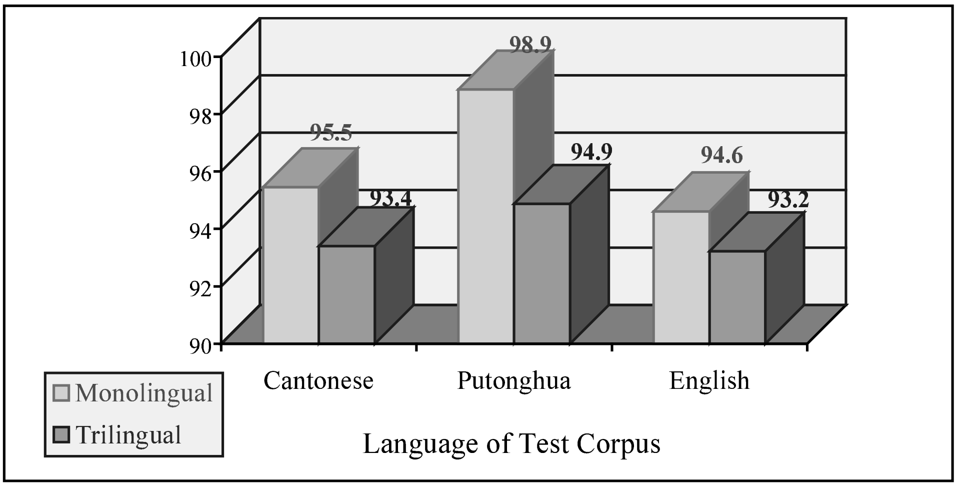


Figure 3: Performance comparison between the monolingual and trilingual recognizers.

Table 2. Intra-language and cross-language confusions based on our recognition results. (E*=English, C*=Cantonese, P*=Putonghua) The reference currency in Row 1 specifies that in Cantonese, the South Korean *Won* is pronounced as the syllable *waan* with the tone *4*. Confusion percentages are computed based on the recognition of 60 tokens per word.

Rows	Reference	**Intra-language Confusion**		**Cross-language Confusion**	
		Recognizer's Hypothesis	Percentage of Confusions	Recognizer's Hypothesis	Percentage of Confusions
1	waan4, (Won, C*)	toi4_waan1 (Taiwan, C*)	3.33%	won (Won, E*)	15%
2	sing1_gaa1_bo1 (Singapore, C*)	san1_gaa1_bo1 (Singapore, C*)	1.67%	singapore (Singapore, E*)	15%
3	pat1 (Baht, C*)	---	---	pound (Pound, E*)	3.33%
4	mei3_chao1 (US dollar, P*)	---	---	mei5_caau1 (US Dollar, C*)	36.67%
5	mei5_caau1 (US dollar, C*)	bei2_sok3 (Peso, C*)	1.67%	mei3_chao1 (US dollar, P*)	25%
6	ren2_min2_bi4 (Renminbi, P*)	---	---	renminbi (Renminbi, E*)	21.67%
7	ying1_bang4 (Pound, P*)	---	---	jing1_bong6 (Pound, C*)	8.33%
8	yuan (Yuan, E*)	won (Won, E*)	1.67%	yuan2 (Yuan, P*)	8.33%
9	renminbi (Renminbi, E*)	germany (Germany, E*)	1.67%	ren2_min2_bi4 (Renminbi, P*)	11.67%

Table 3: Distribution of correct and errorful test tokens of the trilingual recognizer.

	Percentage of the Test Set (%)		
	Cantonese Test Set	Putonghua Test Set	English Test Set
Correct Recognition	93.4	94.9	93.2
Intra-language Confusion	4.4	1.1	3.8
Cross-language Confusion	2.2	4.0	3.0