Intelligent Call Manager Based on the Integration of Computer Telephony, Internet and Speech Processing

Chung-Hsien Wu, Yeou-Jiunn Chen and Gwo-Lang Yan

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.
E-Mail:chwu@csie.ncku.edu.tw
Fax:(886)62747076

Abstract

In this paper, an Intelligent Call Manager which integrates the techniques of computer telephony, internet and speech processing is proposed. This system can answer an incoming call, transfer the call to an extension, send a voice mail and call the pager. Using natural speech interface, the call manager serves callers efficiently and courteously. The call manager is composed of two main subsystems, namely keyword spotting subsystem and text-to-speech subsystem. The keyword spotting subsystem was evaluated in a test set of 2400 conversational speech utterances from 20 speakers (12 males and 8 females). At 8.5% false rejection, the proposed keyword spotting subsystem resulted in 17.2% false alarm rate. The evaluation results for the text-to-speech (TTS) conversion subsystem indicated that the average correct rate was 95.7% for intelligibility, and that the mean opinion score (MOS) was 3.4 for naturalness.

Keywords : Intelligent call manager, multi-keyword spotting, utterance verification, text-to-speech

1. Introduction

The rapid growth and development of computer and telephone integration (CTI), or computer telephony represents tremendous potential and change in all areas of computing. Driven by integrating application and technological trends, the computer and telephone integration is incorporating automatic speech recognition and
text-to-speech into applications that have traditionally used a touch-tone interface. To realize the benefits of these new technologies, developers need to create applications with more natural user interfaces that provide speech communication. Rapid advances in speech recognition technology have been achieved in recent years. This enables speech recognition systems to migrate from laboratory to actual application. Thus, one application that cellular providers and local service providers now offer is name dialing. This feature enables a user to call a person by simply speaking the name and the developer can modify the application system more efficiently. Since a naive caller could not be expected, in general, to know how to speak to machine, the system is designed to be flexible in accepting a wide range of user response and behavior. For example, the users’ response may include disfluencies such as hesitation, false starts, and sounds like um’s and ah’s. As the technology improves, a user-friendly speech recognition system is equipped with a keyword spotting capability which allows users the flexibility to give a wide range of response and behavior [1,2,3,4,5,6,7,8].

This paper presents the fastest growing category of computer telephony applications: intelligent call management. The intelligent call manager proposed in this paper acts as a bridge between PBX switch and computer and adds programmed intelligence to the manner in which incoming calls are managed. Recently, computer speech processing applications have been found in areas ranging from telecommunications to computers and consumer products. By integrating telephone speech keyword spotting, text-to-speech conversion, Internet, voice mail, and pager, the intelligent call manager is designed to enhance and upgrade the key phone switch to an innovative voice processing system. Using natural speech interface, the intelligent call manager serves callers efficiently and friendly.

Chinese is a tonal language in which the same phonetic syllable when pronounced in different tones gives quite distinct meanings. Conventionally, there are 408 Mandarin base syllables, regardless of tones, which is composed of 21 INITIAL’s and 38 FINAL’s [9]. In this paper, a two-stage keyword recognition system is proposed. In the first stage, the conventional Viterbi algorithm is employed to find the scores of the N best keyword candidates and their corresponding subsyllable boundaries. Subsyllable boundaries are then used to extract the FINAL parts of Mandarin syllables, which contain the prosodic information. In the second stage, the prosodic features of these FINAL parts are fed to their corresponding prosodic Hidden Markov Models (HMM’s) and anti-prosodic HMM’s to output a prosodic verification
score. A keyword verification function combining acoustic score and prosodic-phase verification is proposed and used to make the final keyword acceptance/rejection decision.

Many studies have focused on Text-to-Speech (TTS) systems for different language [10,11,12,13,14,15,16,17,18,19]. Also, TTS systems and synthesis technology for the Chinese language have been developed in the last decade [14,15,16,17,18,19]. An important characteristic of Mandarin Chinese is that it is a tonal language based on monosyllables. There are five basic tones: the high-level tone (Tone 1), the mid-rising tone (Tone 2), the midfalling-rising tone (Tone 3), the high-falling tone (Tone 4), and the neutral tone (Tone 5). Besides, a word is the basic rhythmical pronunciation unit and its prosodic properties are generally affected by the tone combination, word length, part of speech (POS) of the word, and word position in a sentence. In this paper, a word prosody database recording the relationship between the linguistic features and the word prosodic patterns is established. Each word prosodic pattern contains the syllable duration, energy contour and pitch contour of the word. Finally, appropriate word prosodic patterns in a sentence are selected from the word prosody database according to the linguistic features.

The rest of this paper is organized as follows. A brief description of the whole system is presented in Section 2. The keyword spotting subsystem is described in Section 3. Section 4 introduces the text-to-speech conversion subsystem. The performance of the system is evaluated and discussed in Section 5. Some concluding remarks are given in the last section.

2. System Overview

The intelligent call manager is implemented on a PC platform with a Dialogic D/41ESC card as the telephone interface. The block diagram of the intelligent call manager is shown in Fig. 1. It is divided into tow main subsystems: keyword spotting subsystem and text-to-speech subsystem. They are described in detail in the following.

- **Keyword Spotting Subsystem**

  The phonetic and prosodic features are extracted from the input speech. HMM’s with continuous observation densities are adopted to model the phonetic and prosodic features. The N-best Viterbi algorithm is employed to fine the scores of the syllable lattice and their corresponding subsyllable boundaries. Subsyllable
boundaries are then used to extract the FINAL part of Mandarin syllables, which contains the prosodic information. A fuzzy search algorithm is proposed to spot keywords from syllable lattice. Finally, a keyword verification function combining phonetic-phase and prosodic-phase verification functions is investigated and used to reorder the ranks of the N-best keyword candidates.

- **Text-to-Speech Subsystem**
  The input text, which is an arbitrary Big-5 code string (for Chinese), is first performed to identify punctuation, Arabic numerals, and Chinese characters. Some contextual features, such as the phonetic structure and syntactical structure, are extracted and a Chinese word dictionary of about 80,000 entries is used for word identification. The phonetic transcription for each character is obtained by referring to a phonetic symbol-to-sound table. According to the information of the first two modules, word prosody patterns are selected from the database based on the linguistic features including tone combination, word length, POS of the word and word position in sentence. Using the outputs from the word prosody selection, prosody modification based on the PSOLA approach is carried out to produce synthesized speech. It adjusts the word prosody including the syllable duration, energy contour, and pitch contour.

  The system answers an incoming call using the speech output from the dialog manager and the text-to-speech conversion system. Using the keyword spotting subsystem, the system can detect the person’s name and discard other extraneous speech spoken by the caller. With a pre-defined name database in an organization, the system can automatically transfer the call to an extension of the called person. If an answering machine or a busy signal is encountered, the caller can leave a voice message or schedules a callback. If there are no answering machines and no one can answer the phone, the caller can leave a voice mail. This voice mail will be sent to a mail server via TCP/IP link and a message will be sent to the pager of the called person to notify him/her there is a voice mail in the server.

**3. Keyword Spotting Subsystem**

**3.1 Feature Extraction**

Two sets of features, phonetic and prosodic features, are used in this subsystem. For phonetic features, 12 Mel-Frequency Cepstrum Coefficient (MFCC), 12 delta...
MFCC, delta log energy, and delta delta log energy are adopted. The prosodic variations in human speech result from different speaking styles or emphasis of energies at different frequencies of the same utterance. These prosodic features are always encoded in duration, intensity, pitch contour, and spectral energy at fundamental frequency. A suitable representation of prosodic information is proposed to adequately model the different dynamic articulatory characteristics of that sequence when it is produced in different meaning contexts.

In prosodic feature extraction for the \( i \)th analysis frame of the \( j \)th FINAL part, four parameters in the prosodic feature vector are defined as follows:

\[
V_j^{(i)} = [v_j^{(i)}(1), v_j^{(i)}(2), v_j^{(i)}(3), v_j^{(i)}(4)]
\]

- Normalized pitch period

\[
v_j^{(i)}(1) = \begin{cases} \frac{P_j^{(i)} - \bar{P}_j^{(i)}}{r}, & \text{Pitch period} \neq 0 \\ r, & \text{Otherwise} \end{cases}
\]

where \( P \) is the logarithmic value of the pitch period of a FINAL part, \( \bar{P} \) is the average logarithmic value of the pitch period, and \( r \) is a small random value.

- Delta logarithmic pitch period

\[
v_j^{(i)}(2) = \Delta(P_j^{(i)} - \bar{P}_j^{(i)})
\]

- Spectral energy at the fundamental frequency

\[
v_j^{(i)}(3) = \begin{cases} \log \left( \frac{S_j^{(i)}}{S_{j,F}^{(i)}} \right), & \text{Pitch period} \neq 0 \\ -\log(S_j^{(i)}), & \text{Otherwise} \end{cases}
\]

where \( S_{j,F}^{(i)} \) is the spectral energy at the fundamental frequency, \( S_j^{(i)} \) is the average spectral energy.

- Spectral energy at the formant frequency

\[
v_j^{(i)}(4) = \begin{cases} \log \left( \frac{S_{j,\text{max}}^{(i)}}{S_j^{(i)}} \right), & \text{Pitch period} \neq 0 \\ -\log(S_j^{(i)}), & \text{Otherwise} \end{cases}
\]

where \( S_{j,\text{max}}^{(i)} \) is the spectral energy at the first formant frequency.

### 3.2 HMM for Keyword Spotting

In the keyword spotting subsystem, 59 context-independent subsyllables, i.e., 21
INITIAL’s, 38 FINAL’s, and a silence model are constructed. Each INITIAL HMM consists of 3 states and each FINAL HMM consists of 5 states, each with 10 Gaussian mixture densities. However, the silence model consists one state and 6 Gaussian mixture densities. In general, for every subsyllable model in the model set, a corresponding anti-subsyllable model is trained specifically for the verification task.

Since lexical tone is the most important feature of the prosodic information, prosodic model should be constructed based on lexical tone behavior. Earlier investigations showed that the tone behavior is very complicated in continuous Mandarin speech, although there are only 5 different tones in Mandarin. Therefore, we assume every kind of possible tone combination needs a context-dependent model, then a total of 175 prosodic HMM’s will be needed. For the construction of anti-prosodic models, the training data are divided into five groups according to their corresponding lexical tones. Five anti-prosodic HMM’s, each corresponding to one context-independent lexical tone, are constructed to enhance the discriminability among prosodic HMM’s. An anti-prosodic HMM can be considered as a lexical-tone-specific model. It is based on similar concept to the cohorts in speaker verification [20]. An anti-prosodic HMM is generally trained on the training data with all lexical tones but that with the corresponding lexical tone. Each prosodic HMM has 4 states and 6 mixtures.

3.3 Two-Stage Keyword Spotting

In this subsystem, a two-stage keyword spotting recognition scheme is used. Fig. 2 shows a block diagram of the keyword spotting subsystem. First, the phonetic and prosodic features are extracted. Hidden Markov models with continuous observation densities are adopted to model the phonetic and prosodic features. In the first stage, the N-best Viterbi algorithm is employed to decode the spontaneous speech into a syllable lattice. According to the syllable lattice, a fuzzy search algorithm is used to extract the possible keyword candidates. In the fuzzy search algorithm, the distances between every two subsyllables are calculated. The nearest neighbor for each subsyllable is determined and used to compensate the substitution, insertion, or deletion errors. Using the fuzzy search algorithm, we can find the most likely keyword $K_i$, where

$$K_i = s_i^1 s_i^2 \cdots s_i^L.$$ (6)
The subsyllable string $s_1^f s_2^f \cdots s_L^f$ is the subsyllable lexical representation of keyword $K_i$ and $L$ is the number of subsyllables. After the subsyllable string $s_1^f s_2^f \cdots s_L^f$ corresponding to the keyword $K_i$ has been determined, the prosodic features of FINAL part sequence in the subsyllable string $s_1^f s_2^f \cdots s_L^f$ are fed to their corresponding prosodic models for prosodic-phase verification.

### 3.4 Keyword Verification

Keyword verification can be treated as the problem of statistical hypothesis testing. Two types of errors can occur: false rejection (Type I) and false acceptance or false alarm (Type II) errors. In the keyword verification process, a two-phase verification scheme is employed.

In the phonetic-phase verification, the normalized confidence measure for a given subsyllable $s_n^{(i)}$ is defined as

$$LR(o_n^{(i)}; s_n^{(i)}) = \frac{1}{F_n^{(i)}} \left( \log \left[ L(o_n^{(i)}; s_n^{(i)}) \right] - \log \left[ L(o_n^{(i)}; \bar{s}_n^{(i)}) \right] \right)$$

(7)

where $\bar{s}_n^{(i)}$ is the anti-subsyllable model of $s_n^{(i)}$ and $F_n^{(i)}$ is the number of frames allocated for subsyllable $s_n^{(i)}$. For an $L$-syllable string $s_1^f s_2^f \cdots s_L^f$ corresponding to the most likely keyword $K_i$, the whole keyword phonetic verification function is defined as follows:

$$D(O; K_i) = \log \left[ \frac{1}{L} \sum_{n=1}^{L} \alpha_n^{(i)} \exp \left( - \eta \cdot LR(o_n^{(i)}; s_n^{(i)}) \right) \right]^{\frac{1}{\eta}}$$

(8)

where $\eta$ is a positive constant and set to 1. $\alpha_n^{(i)}$ is a subsyllable weighting empirically chosen as

$$\alpha_n^{(i)} = \begin{cases} 0.75 & \text{if } s_n^{(i)} \text{ is an INITIAL} \\ 1 & \text{if } s_n^{(i)} \text{ is a FINAL} \end{cases}$$

(9)

This measure weights the contributions of all the subsyllables within a given word based on the selected value of $\eta$. This is believed to be important since subsyllable strings with multiple putative errors would be more easily detected with this type of formulation.

The subsyllable weight for INITIAL is chosen smaller than that for FINAL. This is because that the INITIAL part in Mandarin syllable occupies just a short duration
compared to the FINAL part and the recognition accuracy or reliability for INITIAL is lower than that for FINAL part.

In the prosodic-phase verification, the corresponding lexical tone string $T_{K_i}$ with respect to the keyword $K_i$ is obtained using the sandhi rules [21] and written as

$$T_{K_i} = t_1' t_2' \cdots t_M'$$

where $M$ is the number of the FINAL subsyllables. Since most of the prosodic information is embedded in the FINAL part, the prosodic verification is only performed on the FINAL part. Given the prosodic feature vectors of a FINAL part corresponding to the lexical tone $t_j$, the prosodic confidence measure is written as

$$CM(P_j; t_j) = \log[G(p_{t_j}; t_j)] - \log[G(p_{\bar{t}_j}; \bar{t}_j)]$$

where $P_j = [p_{t_j}, p_{\bar{t}_j}]$ represents the verification feature vector, and $G(\cdot)$ is a Gaussian distribution of the verification feature vector. The parameters of the feature vectors $p_{t_j}$ and $p_{\bar{t}_j}$ are obtained by processing the prosodic feature vectors of the segmented FINAL part through prosodic model $t_j$ and anti-prosodic model $\bar{t}_j$, respectively. Therefore, $p_{t_j}$ forms a 21-dimensional vector consisting of the following:

- Coefficients representing the contour of the prosodic features of the segmented FINAL part. To be more specific, each prosodic feature in $\nu_j$ is represented by a smooth curve formed by orthonormal expansion with discrete Legendre polynomial [22]. The number coefficients used in this polynomial is up to the third order. The zero-th order coefficient represents the mean of the prosodic feature contour and the other three coefficients represent its shape. Given a 4-dimensional prosodic feature vector, the number of parameters is 16.
- Four parameters representing the state durations in number of frames normalized by the total frame duration of segmented FINAL part.
- The prosodic HMM likelihood $L(V_j | t_j)$.

Similarly, $p_{\bar{t}_j}$ is formed by processing $V_j$ using the anti-prosodic model $\bar{t}_j$ and computing the corresponding 21 parameters. For the whole word verification, the verification function can be decomposed into a series of FINAL part verification functions. Assuming independence, the whole word prosodic verification function is defined as follows:
\[ D(P; G_{K_i}) = \log \left[ \frac{1}{M} \sum_{j=1}^{M} \exp[-\kappa \cdot CM(P_j|t_j)] \right]^{\frac{1}{\kappa}} \]  

(12)

where \( \kappa \) is a positive constant. The outputs of the prosodic and phonetic verification functions are then combined as follows.

\[ D(O, P; K_i) = (1 - \beta)D(O; K_i) + \beta D(P; G_{K_i}) \]  

(13)

where \( \beta \) is a weighting. Finally, the word rejection/acceptance decision is made by comparing \( D(O, P, K_i) \) with a predefined threshold.

4. Text-To-Speech Conversion Subsystem

Fig. 3 shows a block diagram of the text-to-speech subsystem. In the TTS subsystem, the input text is an arbitrary Big-5 code string (for Chinese). Text analysis is first performed to identify punctuation, Arabic numerals, and Chinese characters. Some contextual features, such as the phonetic structure and syntactical structure, are extracted. Also, a Chinese word dictionary of about 80,000 entries, in which the words are organized in a tree-like structure to improve the search speed, is used for word identification. In phonetic transcription module, dictionary lookup of word pronunciation and rules for homonym disambiguation are performed. The phonetic transcription for each character is obtained by referring to a phonetic symbol-to-sound table. Based on the linguistic features of an input word, a word prosody database is first established from a large speech database. According to the information of the first two modules, word prosody patterns are selected from the database based on the linguistic features including tone combination, word length, POS of the word and word position in a sentence. Synthesis units with varied prosodic and spectral characteristics from a large, single speaker speech database are selected and concatenated to generate synthesized speech. The pitch-synchronous overlap-and-add (PSOLA) approach [13] is applied to adjust the word prosody including the syllable duration, energy contour, pitch contour, and pause duration.

4.1 Construction of Word Prosody Database

In the Chinese TTS system, some linguistic features are relevant to word prosodic information. They are tone combination, word length, POS of the word, and word position in a sentence. These features are discussed in more detail in the following.
**Tone combination:** In Mandarin Chinese, the word is the basic comprehensive pronunciation unit. A word with length $n$ consists of $n$ syllable(s) in which each syllable has a tone. However, the neural tone generally appears at the end of a word. As a result, there are $4^{n-1} \cdot 5$ tone combinations for an $n$-syllable word.

**Word length:** In Mandarin speech, the word is the basic pronunciation unit. There exist monosyllabic, disyllabic, and polysyllabic words. The intonational or prosodic relationship between syllables within a word is more obvious than that between two words. Therefore, word length of an $n$-syllable word is used to choose its corresponding word prosodic patterns with word length $n$.

**POS of the word:** POS is also an important linguistic feature to determine the word prosody. In this paper, POS is divided into 21 categories. The distance between two categories is defined by the distance of their corresponding average prosodic patterns in the training database, i.e., word pitch contour, word energy contour, and syllable duration in the word. This distance is then normalized to lie between 0 and 1. Therefore, a POS distance table was established.

**Word position in a sentence:** In general, the pitch contour and energy contour in a sentence will follow an intonation pattern. For example, the pitch contour and the energy contour will decline in a declarative sentence. This implies that the word position in a sentence will affect the word prosodic information. A word position ratio is defined as the order of the word position in the sentence divided by the number of words in the sentence.

Using the above linguistic features, a word prosody database is constructed. The linguistic features of a word, i.e., tone combination, word length, POS of the word and word position in a sentence, are associated with a set of prosodic patterns, i.e., syllable duration, energy contour, and pitch contour. To establish the word prosody database, a continuous speech database established by the Telecommunication Laboratories, Chunghwa Telecom Co., Taiwan, containing 655 reading utterances was used. The speech signals were digitized by a 16-bit A/D converter at a 20-kHz sampling rate. The syllable segmentation and phonetic labels were manually done. A total number of 38907 syllables and their phonetic labels were obtained. Using the text analysis, 9698 reference words (including 2-, 3-, and 4-syllable words) and their corresponding word prosodic patterns are obtained.

### 4.2 Selection of Word Prosodic Patterns
For each input word in a sentence, the tone combination and word length were first used to choose the corresponding word prosodic patterns. Second, the remaining linguistic features, POS and position of the word in a sentence, were used to calculate the linguistic distance between the input word and the reference words in the database. The prosodic pattern corresponding to the reference word with the minimum linguistic distance is chosen as the output using the following distance estimation method.

\[
j^* = \min_{j} \{d_C(C_i, C_j) + d_P(P_i, P_j)\}, \quad j = 1, \ldots, J
\]  

(14)

where \( J \) is the total number of words in a sentence corresponding to a given tone combination and word length. \( D_C(C_i, C_j) \) represents the linguistic distance between the POS of the input word \( C_i \) and that of the reference word \( C_j \) in the database. \( D_P(P_i, P_j) \) represents the absolute distance between word position ratio of the input word \( P_i \) and that of the reference word \( P_j \).

5. Experiments

In order to assess the Intelligent Call Manager, 1200 faculty names in National Cheng Kung University were selected as the keywords. A continuous telephone-speech database was employed to train the system. The database is part of the MAT (Mandarin Speech Across Taiwan) speech database and is composed of short spontaneous speech, numbers, syllables, words, and sentences. The total number of file is 12386. This database was pronounced by 295 speakers (192 males, 103 females). All speech data were recorded via public telephone lines in 8KHz using a Dialogic D/41D telephone card and a 16-bit Soundblaster card.

5.1 Experiments for the keyword spotting subsystem

We recorded 2400 utterances for testing spoken by a different group of 20 speakers (12 males, 8 females) responding to requests for a person’s name in our vocabulary. All test utterances were assigned to one of the following categories. The percentage for each category in the testing database is also listed below.

- In-vocabulary names, spoken in isolation (K): 25%
- In-vocabulary names, embedded before a phrase (K+N): 17%
- In-vocabulary names, embedded after a phrase (N+K): 26%
- In-vocabulary names, embedded in a sentence (N+K+N): 23%

In this database, only 91% of the users provided an isolated name or a name
embedded in a phrase or a sentence. 9% of user responses included no names at all. All of these responses need to be rejected. In our experiments, two types of errors, namely, false rejection and false alarm are used to evaluate the system performance. Several experiments were conducted to determine factors necessary to achieve the best performance.

5.1.1 Effect of the weighting parameter $\beta$

In the first experiment, the variation of the total Type I and Type II errors was evaluated as a function of the weighting parameter $\beta$. Fig. 4 shows that the combination of phonetic and prosodic information can improve the keyword spotting rate for $0.25 \leq \beta \leq 0.50$. When $\beta=0.375$, the system can achieve the best recognition performance.

5.1.2 Experiments for the effects of prosodic information

Two experiments were conducted to test the performance of the proposed verification method. In order to benchmark the verification performance, a baseline system that employs only phonetic-phase verification was established. Fig. 5 shows the verification performance of both the proposed and the baseline verification methods. It is clear that the proposed method outperforms the baseline system. For instance, at 8.5% false rejection, the proposed system resulted in 17.8% false alarm rate. This is compared with the baseline system, which results in 22.4% false alarm rate at the same 8.5% false rejection rate.

5.1.3 Experiments for the locations of keywords

The speech utterances divided into five categories were experimented upon to evaluate the effects of the locations of the keywords in an utterance. The experimental results are listed in Table 1. The false alarm rate for the first category (K) and the third category (N+K) was 14.5% and 19.8%, respectively, at a false rejection rate of 8.5%. They were better than that for other categories. This is because the FINAL part in these two categories can be easily detected. Consequently, we can obtain better performance in these two categories. It is reasonable that the first category (K) and the second category achieves the lowest and the highest false alarm rates, respectively. At 8.5% false rejection, the average false alarm rates were 14.5% and 20.9%, for isolated
and embedded keywords, respectively. Furthermore, for the fifth category (N), the proposed method was able to correctly reject 91.4% of nonkeywords.

5.2 Experiments for the TTS subsystem

The twenty users were also asked to subjectively evaluate this TTS system on the intelligibility test. They listened to an article without any prior knowledge of the content of them. Then, they wrote down what they heard sentence by sentence. By comparing the results with the original text, the average correct rate was about 95.7%, as shown in Table 2. According to the experimental results, some synthesized speech is too short and not loud enough due to the channel effect of the telephone line. This resulted in most of the misrecognized syllables in this test. Furthermore, the twenty subjects were further asked to evaluate the naturalness of the TTS system. For the synthesized speech, the subjects gave mean opinion scores (MOS) on a scale of 1 to 5, i.e., 5 for excellent level, 4 for good level, 3 for fair level, 2 for poor level, and 1 for unsatisfactory level. From Table 2, the average MOS is 3.4 for naturalness. The results indicate that a shorter word obtains a higher MOS since less linguistic information is needed. Furthermore, the MOS for short texts is lower than the average MOS. The reason is the lack of syntactic and semantic information for providing more prosodic information in this system.

In this experiment, a training text was chosen for the inside test of the proposed approach. Fig. 6 shows an example of the original speech and synthesized prosodic parameter sequences of the mean pitch period, average energy, initial duration, and final duration. For the mean pitch period of each syllable in the test text, panel (a) plots the synthesized and the original contours. It can be seen that the two contours are very alike for most syllables. The results of mean energy, initial duration, and final duration are shown in panel (b), (c), and (d), respectively. It can be seen that the contours of some syllables match quite well with their counterparts.

6. Conclusions

As the Internet’s popularity increase, the merger of computer telephony and the Internet presents tremendous business opportunity for individual and organizations from both industries. In this paper, the intelligent call manager shows a successful example of the integration of computer telephony and the Internet. At 8.5% false
rejection, the proposed keyword spotting subsystem resulted in 17.2% false alarm rate. The result for TTS conversion system indicated that the average correct rate was 95.7% for intelligibility, and that the mean opinion score (MOS) was 3.4 for naturalness.

**Acknowledgment**

The authors would like to thank the National Science Council, the Republic of China, for financial support of this work under contract No. NSC86-2622-E-006-003.

**References**


Fig. 1 Block Diagram of the Intelligent Call Manager

Fig. 2 The block diagram of the keyword spotting subsystem
Fig. 3. A block diagram of the TTS system

Fig. 4 Combined Type I and Type II errors as a function of the weighting factor $\beta$
Fig. 5 Utterance verification performance comparison of the proposed and the baseline methods.
Fig. 6. Example of the original (solid lines) and the synthesized (dotted lines) prosodic parameter sequences of: (a) mean pitch period, (b) average energy (c) initial duration, and (d) final duration. The x-axis represents the syllable positions corresponding to the Chinese characters in Big5 below.
Table 1 False alarm rates (%) for five speech utterance categories, at a false rejection rate of 8.5%.

<table>
<thead>
<tr>
<th>Speech utterance category</th>
<th>Isolated</th>
<th>Embedded</th>
<th>No Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>14.5</td>
<td>23.2</td>
<td>19.8</td>
</tr>
<tr>
<td>(25%)</td>
<td></td>
<td>(17%)</td>
<td>(26%)</td>
</tr>
<tr>
<td>K+N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(17%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N+K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(26%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N+K+N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(23%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td>8.6</td>
</tr>
<tr>
<td>(9%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

False alarms (%)  
Average (%)  

Table 2. Results for the intelligibility and the naturalness tests.

<table>
<thead>
<tr>
<th>Amount</th>
<th>Intelligibility</th>
<th>Naturalness (MOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monosyllable word</td>
<td>500</td>
<td>90.2%</td>
</tr>
<tr>
<td>2-syllable word</td>
<td>100</td>
<td>93.5%</td>
</tr>
<tr>
<td>3-syllable word</td>
<td>100</td>
<td>96.5%</td>
</tr>
<tr>
<td>4-syllable word</td>
<td>50</td>
<td>97.4%</td>
</tr>
<tr>
<td>Sentence</td>
<td>50</td>
<td>96.1%</td>
</tr>
<tr>
<td>Short text</td>
<td>5</td>
<td>N/A</td>
</tr>
<tr>
<td>Average</td>
<td>95.7%</td>
<td>3.4</td>
</tr>
</tbody>
</table>