Speech act modeling and verification of spontaneous speech with disfluency in a spoken dialogue system

Article *in* IEEE Transactions on Speech and Audio Processing · June 2005

DOI: 10.1109/TSA.2005.845820 · Source: IEEE Xplore

2 authors, including:

Chung-Hsien Wu
National Cheng Kung University
263 PUBLICATIONS **2,713** CITATIONS

SEE PROFILE

Speech Act Modeling and Verification of Spontaneous Speech With Disfluency in a Spoken Dialogue System

Chung-Hsien Wu, *Senior Member, IEEE,* and Gwo-Lang Yan

***Abstract*—This work presents an approach to modeling speech acts and verifying spontaneous speech with disfluency in a spoken dialogue system. According to this approach, semantic information, syntactic structure and fragment class of an input utterance are statistically encapsulated in a proposed speech act hidden Markov model (SAHMM) to characterize the speech act. An interpolation mechanism is exploited to re-estimate the state transition probability in SAHMM, to deal with the problem of disfluency in a sparse training corpus. Finally, a Bayesian belief model (BBM), based on latent semantic analysis (LSA), is adopted to verify the potential speech acts and output the final speech act. Experiments were conducted to evaluate the proposed approach using a spoken dialogue system for providing air travel information. A testing database from 25 speakers, with 480 dialogues that include 3038 sentences, was established and used for evaluation. Experimental results show that the proposed approach identifies 95.3% of speech act at a rejection rate of 5%, and the semantic accuracy is 4.2% better than that obtained using a keyword-based system. The proposed strategy also effectively alleviates the disfluency problem in spontaneous speech.**

***Index Terms*—Bayesian belief model, disfluency modeling, speech act modeling, spoken dialogue.**

I. Introduction

IN AN AGE of the information explosion, access to data is increasingly becoming essential in our daily lives. Spoken dialogue systems have crucial roles to play in human-computer interfaces. In this decade, several spoken dialogue systems have been demonstrated in real-world applications, such as air travel information services (ATIS), weather forecast systems, automatic call managers and ticket reservation services [1]–[3]. However, some problems remain to be solved before such dialogue systems are truly robust. Some research [4], [5] has focused on analyzing users' intentions, for example, in relation to call types or calling plans, and has found that they are strongly associated with grammar fragments. Additionally, speech acts are helpful in handling ill-formed or wrongly recognized utterances [6]–[8] and essential in determining the expressed intention of an input utterance. The dialogue system firstly detects a speech act to identify the intention associated with a query. Second, the identified semantic slots are verified to improve the accuracy of their identification. For speech act modeling, the keyword-based approach [9] is frequently used. This approach cannot extract the semantic information and syntactic structure of a sentence. Ambiguous results are typically obtained when indispensable relations are lost. Some other approaches that use statistical methods generally use bigram or trigram probability models [10] to select the appropriate speech act. These approaches encounter the problem that local syntactic characteristics (bigram/trigram) cannot specify the speech act very well without considering the semantic information of the entire sentence. The approach in [11] that used a large set of grammar rules to explain the syntactic and semantic possibilities for spoken sentences suffered from a lack of robustness when faced with the wide variety of spoken sentences that people really use. Generally, the derivation of syntactic and semantic rules is labor intensive, time consuming and tedious. Furthermore, because many of the various spoken sentences have different syntactic and semantic characteristics, it is indeed difficult to collect appropriate and complete rules to describe the syntactic and semantic diversities. Other approaches [12], [13] used latent semantic indexing in call routing to identify the goal of the utterance or Bayesian belief networks to infer the communicative intention in natural language understanding. This strategy of identifying intentions is similar to the proposed approach, but the above methods did not consider the syntactic structure of a sentence and were confront with the problems on the selection of discriminative terms of latent semantic indexing and the estimation of the probabilistic causal relationships among the nodes of Bayesian belief networks.

In spontaneous speech, disfluencies such as filled pauses, repetitions, lengthening, repairs, false starts and silence pauses are quite common. However, word repetitions (for example: I … I want) or filled pauses (ah, um) contribute little to the meaning. Moreover, they are not easily characterized by the language model. Many strategies, including language model adaptation techniques, the backoff model and the maximum entropy criterion, have been developed to solve the problem, yet disfluency modeling remains an important unsolved problem, especially for spoken language system [14].

During recent years, it has become essential to equip spoken dialogue systems with the ability to accommodate spontaneous speech input [9]. Some approaches [15], [16] verify word-level speech recognition outputs and out-of-domain requests using language model characteristics. In such a task, an attempt is often made to classify utterances as either in-domain or out-of-domain for communication. Such methods will generate semantic errors in accepted in-domain input. Some previous studies on verification strategies [9] have been based on the combined detection and verification of semantically tagged key-phrases in spontaneous speech. They detect key-phrases in speech directly and perform optimization jointly with applying the semantic constraints in a

Manuscript received March 19, 2003; revised March 15, 2004. This work was supported by the National Science Council of Taiwan, R.O.C., under Contract NSC91-2213-E-006-036. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Jerome R. Bellegarda.

The authors are with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan 70101, R.O.C. (e-mail: chwu@csie.ncku.edu.tw; yangl@csie.ncku.edu.tw).

Digital Object Identifier 10.1109/TSA.2005.845820

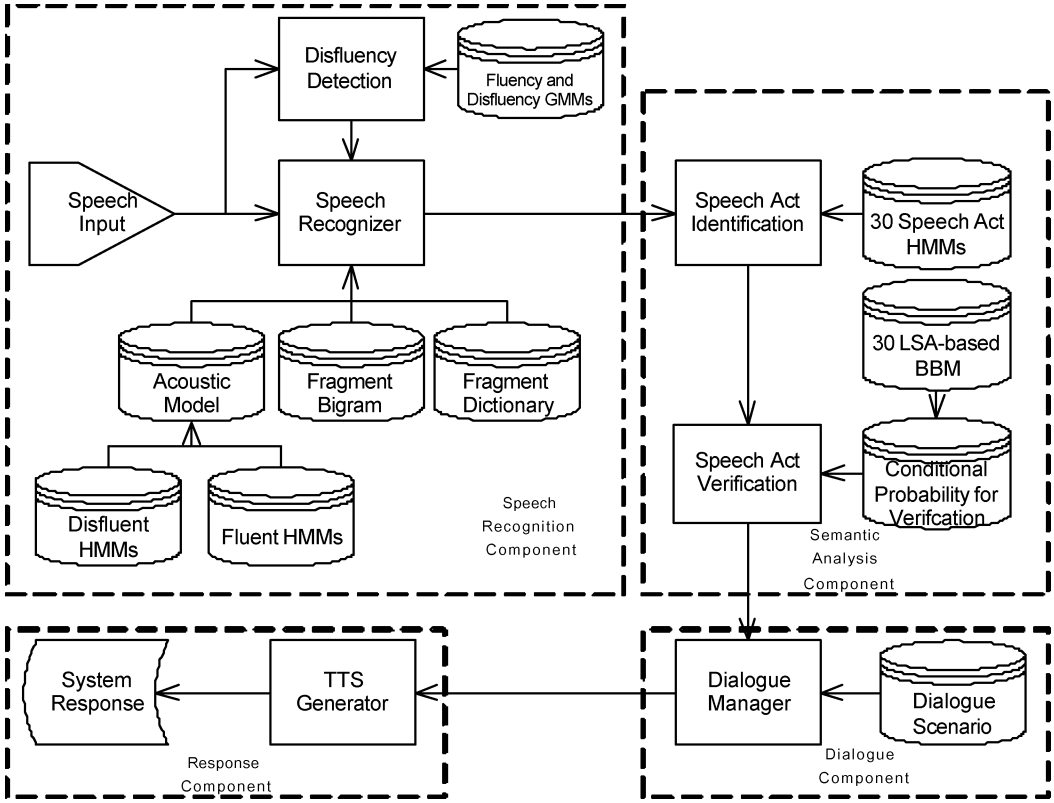


Fig. 1. Dialogue system architecture.

sentence parsing step. The semantic verification process determines whether the semantic representation in the output is complete or incomplete according to the remaining semantic slots. These studies on understanding speech or processing dialogue strictly formalize the user's input and are inconvenient in correcting the errors associated with semantic slots. Furthermore, the interactions between the key-phrases that fill the semantic slots cannot be accurately modeled.

The past approach in [17] adopted the hidden Markov model (HMM) for language modeling. The HMM's model the topic clusters as states and estimate the topic transition likelihood for the segmentation of speech into topics and sentences. In the proposed approach, a statistical speech act hidden Markov model (SAHMM) is employed as a language model to statistically characterize the semantic information, syntactic structure and fragment class of a speech act. The Bayesian belief model (BBM) [18] is adopted for verifying speech acts. This model estimates the verification score using the conditional probability derived from latent semantic analysis (LSA) [19]. The performance of the proposed method was conducted using a spoken dialogue system for an air travel information service. Fig. 1 shows the architecture of this system. In speech recognition component, Gaussian mixture models (GMM's) of fluent and disfluent speech [20], [21] are trained and applied to detect disfluency. The detection results are then integrated with the output of the speech recognizer, containing the fluency and disfluency HMM's, to determine the input speech in terms of possible fragment sequences, including disfluencies, according to a predefined fragment dictionary and a fragment bigram language model [21]. In the semantic analysis component, 30 speech acts are defined and their corresponding SAHMM's are trained to identify a potential speech act. The potential speech act is verified and accepted/rejected based on a threshold. Finally, the dialogue component and the text-to-speech component [22] respond to the identified speech act appropriately via speech.

This work is organized as follows. Section II describes the problems associated with disfluency and the SAHMM for speech act modeling. Section III elucidates a hybrid verification model, LSA-based BBM, for speech act verification. Section IV compare experimental results obtained using the proposed method to those obtained using the keyword-based system [9], which is applied to an ATIS. Finally, Section V summarizes the findings and draws a brief conclusion.

II. Hidden Markov Model for Speech Act Modeling

In a spoken dialogue system, people interact with a computer agent through speaking. Intelligent behavior depends primarily on identifying the speech act of a sentence. Speech acts are actions performed by speaking [23]. These actions such as people's desires or intentions communicated in language are called speech acts. Intuitively, each speech act can be decomposed into one or more combinations of meaningful fragments. For example, the fragments "Thanks" and "Good-bye" are two meaningful fragments for a speech act that could be called "Ending" in performing the task of the air travel information service. Such characteristics of fragment sequences can be modeled using the semantic information and the syntactic structure of the utterance. Accordingly, this work proposes a speech act hidden Markov model (SAHMM) of not only syntactic structure but also the semantic information

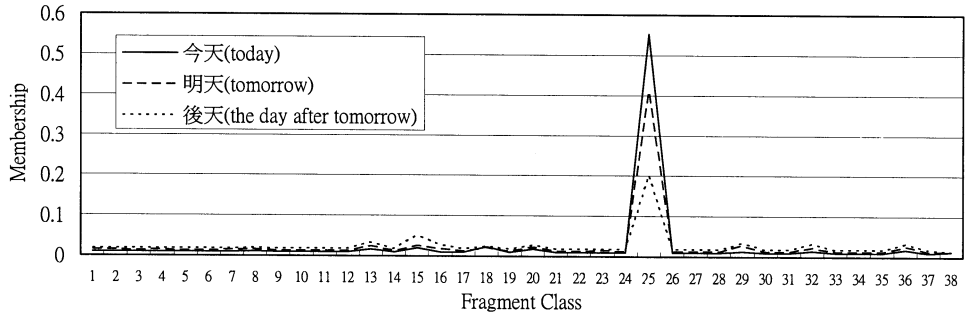


Fig. 2. Membership values of the three fragments, 今天 (today),明天 (tomorrow), and 後天 (the day after tomorrow), in the same class.

in a spoken utterance. In this approach, each SAHMM is defined to model a speech act.

A. Fragment Extraction and Clustering [8]

In this study, a fragment is defined as a combination of some words or characters that often appear together in a specific domain such as "我想知道 (I want to know)", which is combined by "我(I)","想 (want)" and "知道 (to know)". A fragment extraction algorithm in [8] is adopted to extract the fragments from the training corpus to generate a fragment dictionary. For an input fragment w, the bigram relations with all of the extracted fragments (Fg_i, $i = 1,2,\ldots,K$) in the fragment dictionary are employed as the characteristics in forming a bi-directional bigram vector $BBV(w)$. The element $f(Fg_i,w)$ in the bi-directional bigram vector $BBV(w)$ represents the frequency of the fragment Fg_i just preceding (or $f(w,Fg_i)$ for succeeding) the input fragment w in the training corpus. The bidirectional bigram vector for the fragment w is given by

$$BBV(w) = [f(Fg_1,w),\ldots,f(Fg_K,w),f(w,Fg_1),\ldots,f(w,Fg_K)] \quad (1)$$

where K is the total number of extracted fragments. The bidirectional bigram vector is used to cluster the fragments into fragment classes (*FCs*) using a fuzzy c-means algorithm [24] and each FC can be regarded as a syntactic constituent with similar syntactic characteristics. Fig. 2 shows the membership functions of fragments "今天(today)," "明天(tomorrow)," and "後天(the day after tomorrow)" in the same class. In fuzzy clustering, each fragment can be clustered into one or more classes and a fuzzy membership function is used to represent the proportion of the fragment that in a class. For example, the Chinese fragment "查詢 (inquire, inquiry)" includes verb-like and noun-like syntactic constituents. Conventionally, the verb-like fragment "查詢 (inquire)" is more frequently used than the noun-like fragment "查詢 (inquiry)" so that the membership of "查詢 (inquire)" exceeds that of "查詢 (inquiry)" By this approach, the center vector of the fragment class c_j is defined as follows:

$$BBV(\mu_j) = \frac{\sum_{i=1}^{K} u_{ji} BBV(w_i)}{\sum_{i=1}^{K} u_{ji}} \quad (2)$$

where u_{ji} is the fuzzy membership of the fragment w_i in the fragment class c_j. Each fragment class is then used to represent one state in an SAHMM.

B. Observation Probability in SAHMM

In estimating observation probability, three measures are defined to encapsulate semantic information, syntactic structure and fragment class to derive the observation probability for each SAHMM. The observation probability definition is described below.

1) Semantic Observation Probability: The Kullback–Liebler distance, typically applied in semantic clustering [25], is used to measure the semantic similarity between an input fragment and the center of the fragment class in the SAHMM. The Kullback–Liebler (KL) distance [25], defined as follows, is used:

$$KL(p\|q) = \sum_i p(i) \log \frac{p(i)}{q(i)} \quad (3)$$

where $p(\bullet)$ and $q(\bullet)$ are the probability mass functions (*pmf*). The KL distance is always nonnegative and $KL(p\|q) = 0$ if and only if p and q are equivalent. However, the measure does not satisfy the triangle inequality. The divergence measure is normally defined by (4) to represent the "distance" between two probability distributions as a symmetric distance measure [25].

$$Div(p,q) = \frac{KL(p\|q) + KL(q\|p)}{2}. \quad (4)$$

The distance measure is derived from the divergence between two distributions which are calculated from the occurrences of fragments. In order to estimate the semantic observation probability, the distance measure is then converted to a similarity measure

$$Sem\_Similarity(p,q) = \frac{1}{Div(p,q)}. \quad (5)$$

Speech acts treated as discrete random variables with values taken from a finite set $SA_X = \{SA_1, SA_2, \ldots, SA_H\}$, where H is the number of speech acts. The conditional probability of a fragment w with respect to all speech act SA_X is formulated as a *pmf* of SA_X and denoted as $p_w(SA_X)$, using the set of probabilities $\{P(w \mid SA_x), x = 1,2,\ldots,H\}$. $P(w \mid SA_x)$ represents the *a priori* probability that a fragment w occurs in

the xth speech act in the training corpus. Then the *pmf* of the fragment class c_j is defined as

$$p_{c_j}(SA_X) = \sum_{w_k \in c_j} p_{w_k}(SA_X) \times u_{jk} \tag{6}$$

where u_{jk} is the fuzzy membership of fragment w_k in fragment class c_j and obeys the probabilistic constraint $\sum_k u_{jk} = 1$ for any class c_j to normalize the *pmf*. In this approach, the similarity measure of an input fragment w_i in the fragment class c_j is defined as $Sem\_Similarity(p_{w_i}(SA_X), p_{c_j}(SA_X))$. The similarity measure represents the semantic relationship between fragment w_i and fragment class c_j for all speech act SA_X. To normalize observation measures, the semantic observation probability is defined as

$$p_{Sem}(w_i \mid S_\ell = c_j) = \frac{Sem\_Similarity(p_{w_i}(SA_X), p_{c_j}(SA_X))}{\sum_{m=1}^{C} Sem\_Similarity(p_{w_i}(SA_X), p_{c_m}(SA_X))} \tag{7}$$

where S_ℓ denotes the state at time ℓ in an SAHMM and C is the total number of fragment classes.

2) Syntactic Observation Probability: The syntactic observation probability is defined as the degree to which a fragment belongs to a syntactic constituent. This probability models the syntactic structure regarding the neighboring information of the considered fragment. In the proposed approach, this degree can be approximated as the similarity between the bi-directional bigram vectors $BBV(w_i)$ of the input fragment w_i and the center vector $BBV(\mu_j)$ of the fragment class c_j in the SAHMM. This probability is estimated by applying the cosine measure [4], [26] and is normalized as

$$p_{Syn}(w_i \mid S_\ell = c_j) = \frac{\cos(BBV(w_i), BBV(\mu_j))}{\sum_{m=1}^{C} \cos(BBV(w_i), BBV(\mu_m))} = \frac{\frac{BBV(w_i) \bullet BBV(\mu_j)}{\|BBV(w_i)\| \times \|BBV(\mu_j)\|}}{\sum_{m=1}^{C} \frac{BBV(w_i) \bullet BBV(\mu_m)}{\|BBV(w_i)\| \times \|BBV(\mu_m)\|}}. \tag{8}$$

3) Fragment Class Observation Probability: The frequency and the membership of the fragment in the fragment class are used to calculate the fragment class observation probability. This observation probability associated with fragment w_i in state S_ℓ in the SAHMM is defined as

$$p_{FC}(w_i \mid S_\ell = c_j) = \frac{N(w_i) \times u_{ji}}{\sum_{w_k \in c_j} N(w_k) \times u_{jk}} \tag{9}$$

where $N(w_i)$ represents the number of occurrences of fragment w_i in the speech act SA_x and u_{ji} is the membership of w_i in fragment class c_j.

For each input fragment, the observations are represented as $Y = \{y_1, y_2, \ldots, y_M\}$, in which y_1 refers to semantic observation; y_2 refers to syntactic observation, and y_3 refers to fragment class observation in an SAHMM. The observation probability associated with an input fragment w_i is defined as a linear combination of the three observations y_z ($z = 1, \ldots, M$ and $M = 3$) in state S_ℓ and estimated as

$$\begin{aligned} b_{c_j}(w_i) &= \sum_{Y, 1 \le z \le M} m_z P(w_i \mid y_z) \\ &= m_1 \times p_{Sem}(w_i \mid S_\ell = c_j) + m_2 \times p_{Syn}(w_i \mid S_\ell = c_j) \\ &\quad + m_3 \times p_{FC}(w_i \mid S_\ell = c_j) \end{aligned} \tag{10}$$

where m_1, m_2 and m_3 are the weights of semantic, syntactic and fragment class observation probabilities, respectively, satisfying the condition $m_1 + m_2 + m_3 = 1$.

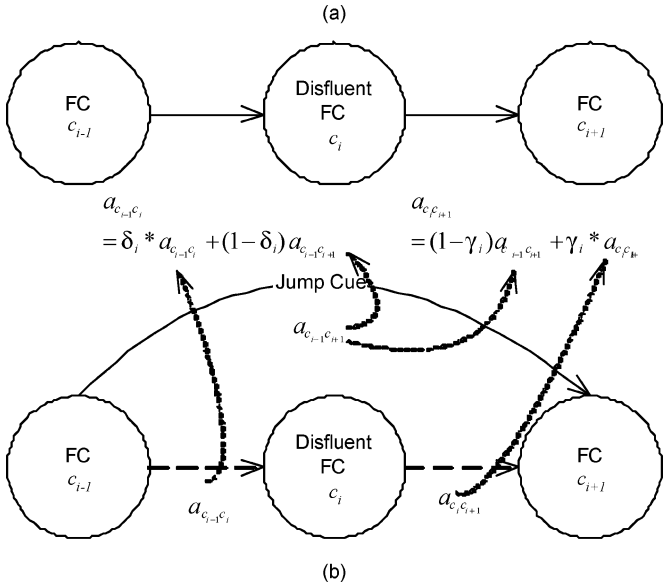


Fig. 3. Diagram of the interpolation method to estimate the transition probabilities $a_{c_{i-1}c_i}$ and $a_{c_i c_{i+1}}$ in SAHMM, using the jump cue if a disfluent FC occurs.

C. State Transition Probability

The state-transition probability from state $S_\ell = c_i$ at time ℓ to state $S_{\ell+1} = c_j$ at time $\ell + 1$ is estimated as

$$\begin{aligned} a_{c_i c_j} &= P(S_{\ell+1} = c_j \mid S_\ell = c_i) \\ &= \frac{N_c(S_{\ell+1} = c_j \mid S_\ell = c_i)}{N_c(S_\ell = c_i)}, 1 \le c_i, c_j \le N \end{aligned} \tag{11}$$

where $N_c(\bullet)$ is the number of occurrences.

Collecting the corpus of spontaneous conversations, which would cover all occurrences of disfluencies, would be impossible. Rather than collecting the diversified conversations exhaustively, the proposed approach estimates the probabilities from the sparse corpus. The approach exploits an interpolation method to re-estimate the state transition probability if a disfluency fragment c_i occurs, as indicated in Fig. 3(a). In this figure, the transition probabilities $a_{c_{i-1}c_i}$ and $a_{c_i c_{i+1}}$ associated with the fragment classes from c_{i-1} to c_i and c_i to c_{i+1}, respectively, cannot be effectively estimated because available data are too sparse. The only available relevant information is the jump cue $a_{c_{i-1}c_{i+1}}$, which represents a hypothetical transition probability without disfluency. This probability is estimated from the training corpus and then used to backoff the bigram probability

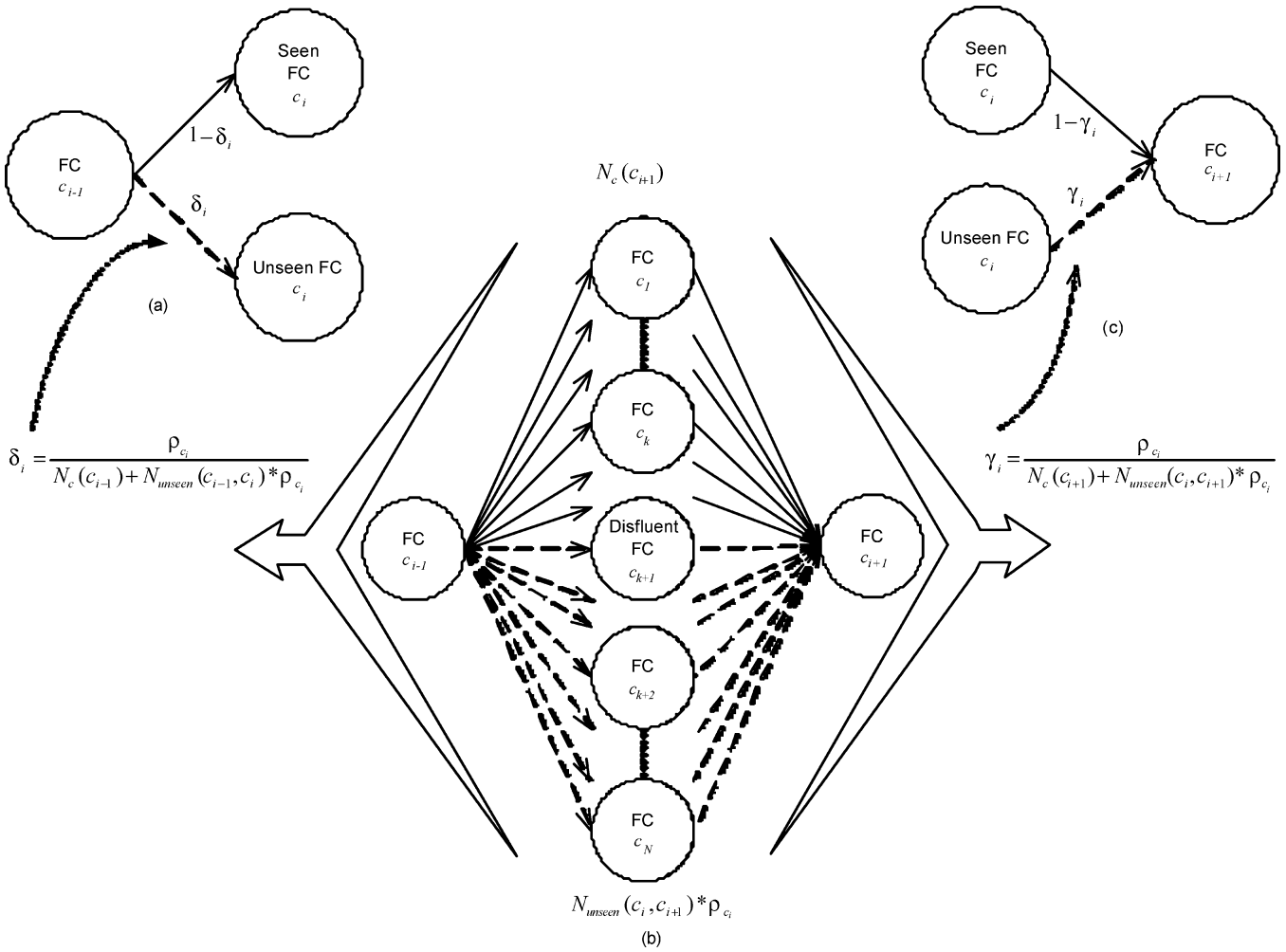


Fig. 4. Estimation of the weights δ_i and γ_i in the interpolation equation. (a) and (c) show the estimation of weights δ_i and γ_i using unseen and seen fragment classes and (b) shows the estimation of unseen and seen fragment classes with numbers N_{unseen} and N_c.

by interpolation, as shown in Fig. 3(b). These two transition probabilities are estimated as follows:

$$a_{c_{i-1}c_i} = \begin{cases} \delta_i \times a_{c_{i-1}c_i} + (1-\delta_i) a_{c_{i-1}c_{i+1}}, & \text{if } c_i \in Disfluency \\ a_{c_{i-1}c_i}, & \text{otherwise} \end{cases} \tag{12}$$

$$a_{c_i c_{i+1}} = \begin{cases} \gamma_i \times a_{c_i c_{i+1}} + (1-\gamma_i) a_{c_{i-1}c_{i+1}}, & \text{if } c_i \in Disfluency \\ a_{c_i c_{i+1}}, & \text{otherwise} \end{cases} . \tag{13}$$

The weights δ_i and γ_i in the interpolation equations are calculated by observing a disfluency fragment in state c_i in the training corpus as shown in Fig. 4. The weights are estimated as follows:

$$\delta_i = \frac{\rho_{c_i}}{N_c(c_{i-1}) + N_{unseen}(c_{i-1}, c_i) \times \rho_{c_i}} \tag{14}$$

$$\gamma_i = \frac{\rho_{c_i}}{N_c(c_{i+1}) + N_{unseen}(c_i, c_{i+1}) \times \rho_{c_i}} \tag{15}$$

where $N_c(\bullet)$ is the number of occurrences of a fragment class and ρ_{c_i} is the expected number of occurrences of an unseen fragment class c_i. $N_{unseen}(c_{i-1}, c_i)$ is the number of occurrences of the unseen fragment class c_i following c_{i-1}. If ρ_{c_i} in these two equations is set to one, the formula is similar to Laplace's law [27]. It is actually the Bayesian estimator if we assume the *a priori* probability of the unseen fragment class is uniform. The most widely used value of ρ_{c_i} is 0.5 [27]. This choice can be theoretically justified as being the expectation of the same quantity, which is maximized by maximum likelihood estimation. The original idea behind this method comes from the Jeferys-Perks law, or the Expected Likelihood Estimation. Herein, ρ_{c_i} is set to 0.5 [27].

D. Weight Determination Using the Expectation Maximization (EM) Algorithm

In the training process, the weights can be determined using the Expectation Maximization algorithm [28]. For a fragment sequence of N observations W: $w_1, w_2, \ldots, w_N$ and its corresponding state (fragment class) sequence *FCS:* $c_1, c_2, \ldots, c_N$ in the training corpus, the parameter set is estimated based on the maximum likelihood to obtain an optimal solution. The Q function [29], which is the auxiliary function for EM algorithm, is defined as shown in (16) at the bottom of the next page, where ϕ_{SA} is the parameter set of SAHMM and $\overline{P}(\bullet)$ is the re-estimated probability under parameter set $\overline{\phi_{SA}}$. The Q function

$Q(\overline{P}(w_n \mid c_n) \mid P(w_n \mid c_n))$ is of interest because the expectation is to be maximized with respect to the combination of weights. The function can be derived as shown in (17) at the bottom of the page, where Ks is the state number. The Q function must be maximized subject to $m_1 + m_2 + m_3 = 1$ ($\sum_z \overline{P}(w_n = y_z) = 1$). The Lagrange multiplier is added to the Q function, which is rewritten as shown in (18) at the bottom of the page. The parameter sets SA and η are partially differentiated, yielding the following equations:

$$\sum_{n=1}^{N}\sum_{i=1}^{Ks}\frac{P(c_n = S_i \mid w_n = y_z, \phi_{SA})}{P(w_n = y_z)} + \eta = 0 \tag{19}$$

$$\sum_{z} P(w_n = y_z) - 1 = 0. \tag{20}$$

Solving the algebraic (19) and (20) yields the optimal closed-form solution

$$m_z = P(w_n = y_z) = \frac{\sum_{n=1}^{N}\sum_{i=1}^{Ks} P(c_n = S_i \mid w_n = y_z, \phi_{SA})}{\sum_{z}\sum_{n=1}^{N}\sum_{i=1}^{Ks} P(c_n = S_i \mid w_n = y_z, \phi_{SA})}. \tag{21}$$

E. Speech Act Identification

Speech act identification seeks to identify the most probably occurring speech act intended by the input utterance U given the corresponding observation sequence W: $w_1, w_2, \ldots, w_N$ and state sequence *FCS*: $c_1, c_2, \ldots c_N$. If a uniform prior probability of speech act is being assumed, the Bayesian classifier decides in favor of speech act SA^* for which

$$\begin{aligned} SA^* &= \arg\max_{SA} P(SA \mid W, U) \\ &= \arg\max_{SA} P(W \mid FCS) P(FCS \mid SA) P(SA) \\ &\approx \arg\max_{SA} P(w_1 w_2 \ldots w_N \mid c_1 c_2 \ldots c_N) P(c_1 c_2 \ldots c_N \mid \phi_{SA}) \\ &= \arg\max_{SA} [b_{c_1}(w_1) b_{c_2}(w_2) \ldots b_{c_n}(w_N)] \\ &\quad \cdot [\pi_{c_1} a_{c_1 c_2} a_{c_2 c_3} \ldots a_{c_{N-1} c_N}] \\ &= \arg\max_{SA} \left[\pi_{c_1} b_{c_1}(w_1) \prod_{i=1}^{N-1} a_{c_i c_{i+1}} b_{c_{i+1}}(w_{i+1})\right]. \end{aligned} \tag{22}$$

Given a speech utterance U, the speech recognizer generates the candidate fragment sequences. The SAHMM is then adopted to identify the most probable speech act SA^* corresponding to

---

$$\begin{aligned} Q(\overline{\phi_{SA}}, \phi_{SA}) &= E\left\{\log P(W = w_1, w_2, \ldots, w_N, FCS = c_1, c_2, \ldots, c_N \mid \overline{\phi_{SA}}) \mid W, \phi_{SA}\right\} \\ &= \sum_n P(c_n \mid w_n, \phi_{SA}) \log P(w_n, c_n \mid \overline{\phi_{SA}}) \\ &= \sum_n P(c_n \mid w_n, \phi_{SA}) \log \left[P(w_n \mid c_n, \overline{\phi_{SA}}) P(c_n \mid \overline{\phi_{SA}})\right] \\ &= \sum_n P(c_n \mid w_n, \phi_{SA}) \left[\log P(w_n \mid c_n, \overline{\phi_{SA}}) + \log P(c_n \mid \overline{\phi_{SA}})\right] \\ &= \sum_n P(c_n \mid w_n, \phi_{SA}) \log P(w_n \mid c_n, \overline{\phi_{SA}}) + \sum_n P(c_n \mid w_n, \phi_{SA}) \log P(c_n \mid \overline{\phi_{SA}})] \\ &= Q(\overline{P}(c_{n+1} \mid c_n) \mid P(c_{n+1} \mid c_n)) + Q(\overline{P}(w_n \mid c_n) \mid P(w_n \mid c_n)) \end{aligned} \tag{16}$$

$$\begin{aligned} &Q(\overline{P}(w_n \mid c_n) \mid P(w_n \mid c_n)) \\ &= \sum_n P(c_n \mid w_n, \phi_{SA}) \log P(c_n \mid \overline{\phi_{SA}})] \\ &= \sum_{n=1}^{N}\sum_{i=1}^{Ks} P(c_n = S_i \mid w_n = y_z, \phi_{SA}) \log \left[\overline{P}(w_n = y_z) \overline{P}(w_n = y_z \mid c_n = S_i)\right] \\ &= \sum_{n=1}^{N}\sum_{i=1}^{Ks} P(c_n = S_i \mid w_n = y_z, \phi_{SA}) \log \overline{P}(w_n = y_z) + \sum_{n=1}^{N}\sum_{i=1}^{Ks} P(c_n = S_i \mid w_n = y_z, \phi_{SA}) \log \overline{P}(w_n = y_z \mid c_n = S_i) \end{aligned} \tag{17}$$

$$\begin{aligned} &Q(\overline{P}(w_n \mid c_n), \eta \mid P(w_n \mid c_n)) \\ &\quad = \sum_{n=1}^{N}\sum_{i=1}^{Ks}\sum_{z} P(c_n = S_i \mid w_n = y_z, \phi_{SA}) \log \overline{P}(w_n = y_z) + \eta \left\{\sum_z \overline{P}(w_n = y_z) - 1\right\} \end{aligned} \tag{18}$$

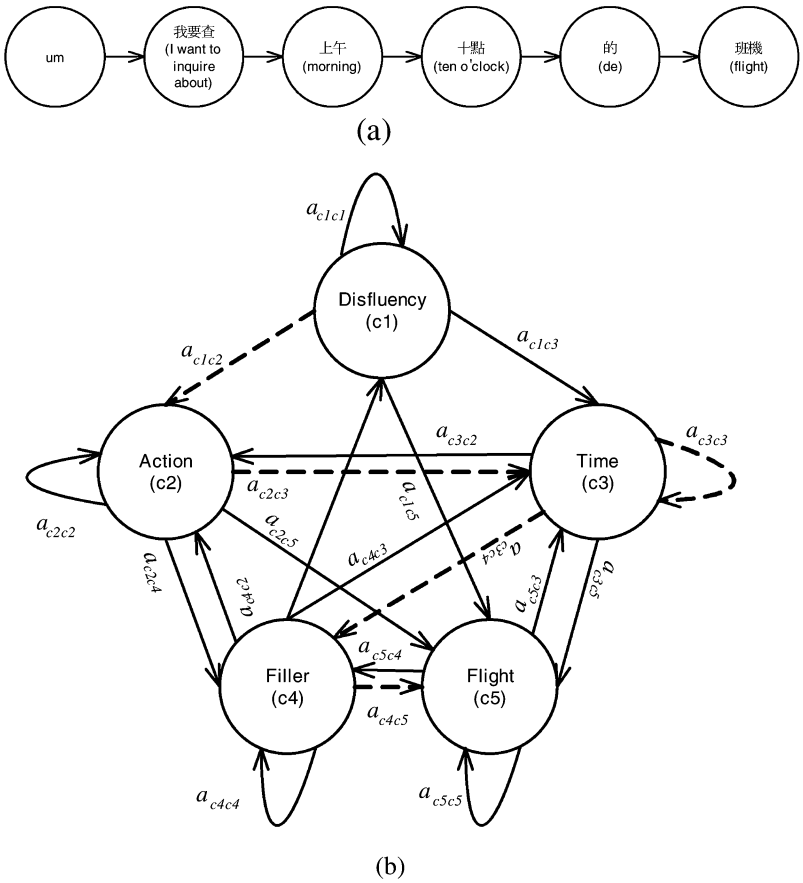


Fig. 5. (a) Fragment sequence. (b) State transitions in SAHMM and the best path (dotted line) of the fragment sequence in the "Inquire the flight around a specific departure Time" SAHMM.

the kth potential fragment class sequence W^k, according to the following equation:

$$\begin{aligned} SA^*_{W^k} &= \arg\max_{SA} P(SA \mid W^k, U) \\ &= \arg\max_{SA} P(W^k \mid FCS^k) P(FCS^k \mid \phi_{SA}) P(SA) \\ &\approx \arg\max_{SA} P(W^k \mid FCS^k) P(FCS^k \mid \phi_{SA}). \end{aligned} \quad (23)$$

The process will identify the most probable speech act $SA^*_{W^k}$ given the fragment class sequence W^k. The fragment sequence "um(disfluency), 我要查(I want to inquire about), 上午(morning), 十點(ten o'clock), 的(de), 班機(flight)" denoted in Fig. 5(a) is considered as an example. The dotted line in Fig. 5(b) represents the optimal state sequence in SAHMM given the fragment sequence. The corresponding fuzzy fragment classes of the optimal state sequence are c1, c2, c3, c3, c4 and c5, which correspond to the statistical classes "Disfluency," "Action," "Time," "Time," "Filler," and "Flight". In this case, the most likely SAHMM is the "Inquire the flight around a specific departure Time" SAHMM. The semantic information is then used to understand the dialogue, as indicated in Table I.

Eventually, the score $P(SA^*_{W^k} \mid W^k, U)$ estimated by the SAHMM is then combined with the acoustic score, as described in (24) at the bottom of the page, where β is a weight between zero and one. $BS(W^k \mid U)$ is the score obtained by the speech recognizer. In this approach, the fragment sequence W^*, corresponding to the identified speech act SA^*, is chosen according to the highest score determined by (24), given the utterance U; the fragment is then used as the input for further verification.

III. LSA-Based Bayesian Belief Model for Speech Act Verification

Usually, the speech act of a spoken utterance is characterized by descriptive information, which includes not only keywords but also the interactions between word sequences, especially words that will fill the semantic slots. This study presents an LSA-based Bayesian belief model of the above characteristics, which is then used to verify the determined speech act. The

$$(SA^*, W^*) = \arg \max_{(SA_{W^k}, W^k)} \{\beta \log P(SA^*_{W^k} \mid W^k, U) + (1-\beta) BS(W^k \mid U)\} \quad (24)$$

TABLE I

SEMANTIC INFORMATION FOR THE INQUIRY "UM 我要查上午十點的班機 (UM I WANT TO INQUIRE ABOUT THE FLIGHT DEPARTING AT TEN O'CLOCK IN THE MORNING)"

	SEMANTIC_INFORMATION	
ACTION	我要查 (I WANT TO INQUIRE)	
DATE	明天 (TOMORROW)	
TIME	PERIOD	上午 (MORNING)
	SPECIFIC TIME	十點 (TEN O'CLOCK)
AIRLINE_ COMPANY	(NULL)	
LOCATION	ORIGIN	(NULL)
	DESTINATION	(NULL)

LSA-based Bayesian belief model not only incorporates the relationship between keywords and a speech act, but also uses latent semantic analysis to discover the hidden interactions among keywords in a speech act.

A. LSA-Based Bayesian Belief Model

Given a fragment sequence $W: w_1, w_2, \ldots, w_N$, a speech act can be verified by considering the combination of fragments in a sentence. An arbitrary number of mutually exclusive and exhaustive speech acts $SA_x (x = 1, \ldots, H)$ are assumed to partition the speech act universe. The verification is generalized as

$$P(SA_x \mid w_1 w_2 \ldots w_N) = \frac{P(SA_x) \times P(w_1 w_2 \ldots w_N \mid SA_x)}{\sum_{i=1}^{H} P(w_1 w_2 \ldots w_N \mid SA_i) \times P(SA_i)}. \quad (25)$$

Even though (25) models the interaction between fragments in a speech act, the conditional probability $P(w_1 w_2 \ldots w_N \mid SA_x)$ is very difficult to calculate from the training corpus because the data are sparse. In practice, the Bayesian probability model [30] is often simplified by assuming that all fragments are statistically independent

$$P(w_1 w_2 \ldots w_N \mid SA_x) = P(w_1 \mid SA_x) \times P(w_2 \mid SA_x) \times \cdots \times P(w_N \mid SA_x). \quad (26)$$

Based on this assumption, the Bayesian belief model (BBM) [18] can be used to represent the probabilistic and causal relationship because such a model is a directed acyclic graph in which the nodes (fragments) represent distinct pieces of evidence; the arcs represent causal relationships between these pieces of evidence. This model is useful for providing a clean formalism for combining distinct evidences in support of the verification. However, applying the above independent assumption to the Bayesian probability model underestimates the latent semantic information (inter-dependencies) between fragments in a speech act. The verification ability can be improved by capturing possible inter-dependencies between fragments. Therefore, the latent semantic analysis (LSA) [19] is exploited to find the inter-dependencies between fragments in an input fragment sequence. The primary idea behind the exploration of latent semantic information is to convert the fragments and their corresponding sentences into a lower dimensional space. The starting point is the construction of an association matrix R between fragments and sentences for a speech act SA_x in the training corpus. Each sentence is associated with a row vector of dimension N (fragment number) in this matrix, and each fragment is associated with a column vector of dimension M (sentence number). The element R_{ij} in the association matrix R represents the number of occurrences of the ith fragment in the jth sentence. The association matrix R, constructed from the training corpus, is extremely large and typically very sparse. Singular value decomposition (SVD), a technique related to eigenvector decomposition and factor analysis [31], is applied to decompose the association matrix R into three components. Consider that only the D largest singular values of S are kept along with their corresponding columns in U and V. The resultant matrix R_D is the matrix of rank D which is close to the original matrix R in the least square sense

$$R \cong R_D = USV^T \quad (27)$$

where $U_{M \times D}$ is a left singular matrix with row vector u_i $(1 \leq i \leq M)$. It is the matrix of eigenvectors derived from the fragment-to-fragment correlation matrix given by $R^T R$; $V_{N \times D}$ is a right-singular matrix with row vectors v_j $(1 \leq j \leq N)$. This is the matrix of eigenvectors derived from the transpose of the sentence-to-sentence matrix given by RR^T; $S_{D \times D}$ is a $D \times D$ diagonal matrix of singular values. U and V are defined as orthonormal bases in the space of D dimensions; that is, $U^T U = V^T V = I_D$. Fig. 6 illustrates the decomposition. The relationship between any two fragments in the reduced space of dimensionality D can be obtained from the $R^T R$ matrix

$$\psi = {R_D}^T R_D = (USV^T)^T USV^T = VSU^T USV^T = VSSV^T = VS(VS)^T. \quad (28)$$

The element $\psi(i, j)$ quantifies the strength of the relationship between fragments w_i and w_j. The concurrent probability $P(w_i w_j \mid SA_x)$, which measures the inter-dependency between fragments w_i and w_j in the speech act SA_x, is defined as

$$P(w_i w_j \mid SA_x) = \frac{\psi(i, j)}{N_c(SA_x)} \quad (29)$$

where $N_c(SA_x)$ is the number of sentences in the speech act SA_x. The fragment pair $w_i w_j$ is then defined as a compound fragment when the condition $P(w_i w_j \mid SA_x) \geq P(w_i \mid SA_x) \times P(w_j \mid SA_x)$ is satisfied. The process is performed recursively when the newly defined compound fragments replace the column elements of R, as shown in Fig. 7.

In the proposed approach, for example, fragment "有沒有 (do you have)" and fragment 班機(flight)" will form a compound fragment in which the probability $P(\text{有沒有(do you have)}, \text{飛機(flight)} \mid SA_x)$ will replace the probabilities $P(\text{有沒有(do you have)} \mid SA_x)$ and $P(\text{飛機(flight)} \mid SA_x)$ as shown in (31) to assist in estimating speech act verification score more precisely. 1812 compound fragments that contain

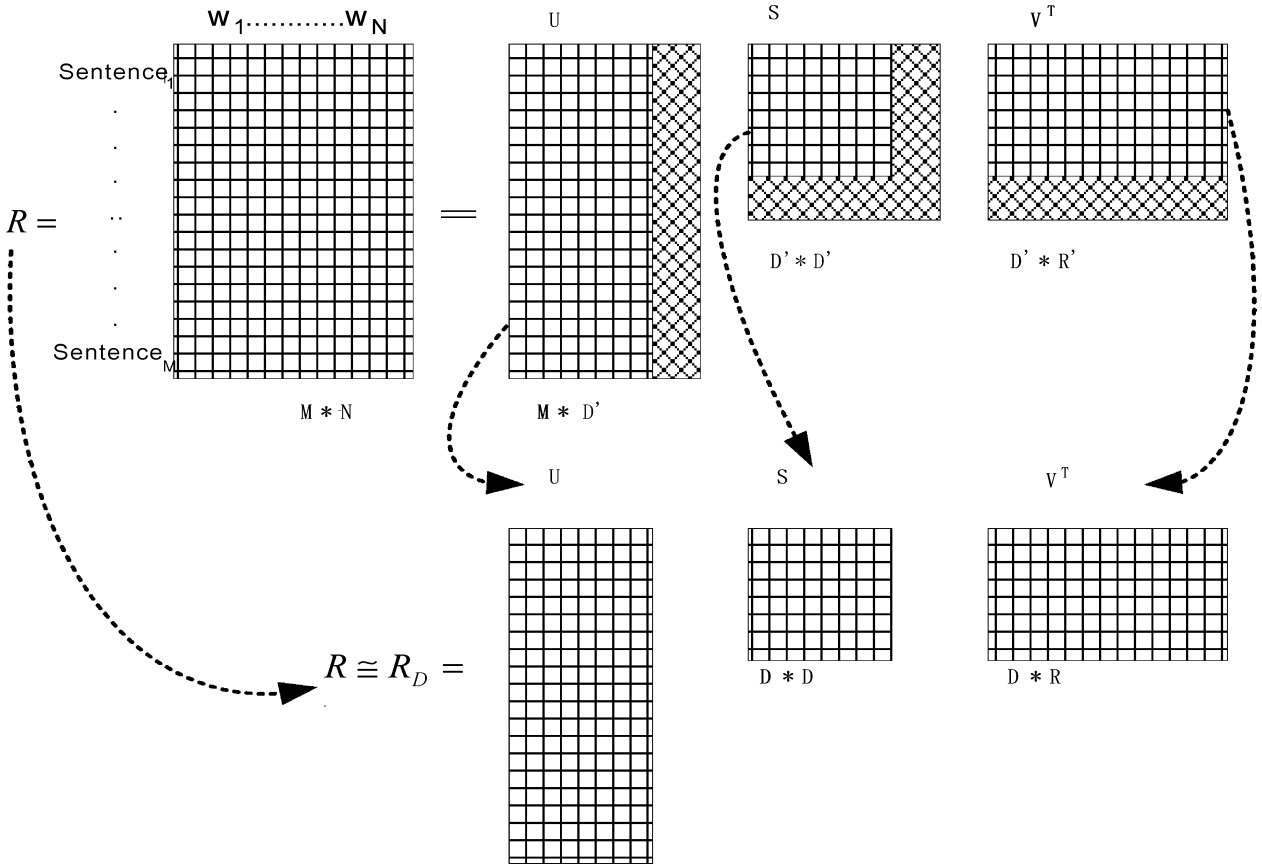


Fig. 6. Singular value decomposition for latent semantic information analysis.

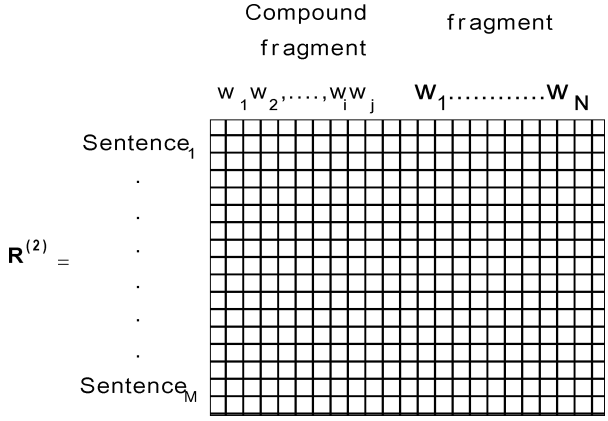


Fig. 7. Association matrix $\mathrm{R}^{(2)}$ for compound fragments.

two or three fragments are considered because a high-order fragment relationship is difficult to derive from a sparse corpus. These new relationships discovered by LSA are used to create concept nodes and form a new BBM. The following conditional probability is then redefined as

$$P(w_1 w_2 \dots w_N \mid SA_x) = P(v_1 v_2 \dots v_T \mid SA_x) = \prod_{t=1}^{T} P(v_t \mid SA_x) \quad (30)$$

where $v_1 v_2 \dots v_T$ refers to the relationship that corresponds to $w_1 w_2 \dots w_N$ for the speech act SA_x. Fig. 8 indicates an example. If compound fragments $v_1, v_2, v_3, \dots$, and v_T are derived from $w_1 w_2, w_3, w_4 w_5 w_6, \dots$, and w_N respectively, then the conditional probability is

$$\begin{aligned} &P(w_1 w_2 \dots w_6 \dots w_N \mid SA_x) \\ &= \prod_{t=1}^{T} P(v_t \mid SA_x) \\ &= P(w_1 w_2 \mid SA_x) \times P(w_3 \mid SA_x) \times P(w_4 w_5 w_6 \mid SA_x) \\ &\quad \times \dots \times P(w_N \mid SA_x). \end{aligned} \quad (31)$$

The estimated conditional probability is then integrated into the proposed system for speech act verification. The verification score from LSA-based BBM for the identified speech act SA^* and the fragment sequence W^* obtained from (24) is defined as follows:

$$\begin{aligned} &VScore(W^* \mid SA^*; U) \\ &\quad = \log(P(SA^* \mid w_1^* w_2^* \dots w_N^*)) \\ &\quad = \log\left(\frac{P(SA^*) \times P(v_1 v_2 \dots v_T \mid SA^*)}{\sum_{i=1}^{H} P(SA_i) \times P(v_1 v_2 \dots v_T \mid SA_i)} \right) \end{aligned} \quad (32)$$

where w_n^* is the nth fragment in the identified fragment sequence W^*. The speech act with the verification score $VScore(W^k \mid SA_{W^k}^*; U)$, above a selected threshold is regarded as the final output. Conversely, any speech act candidate with a score below the threshold is rejected.

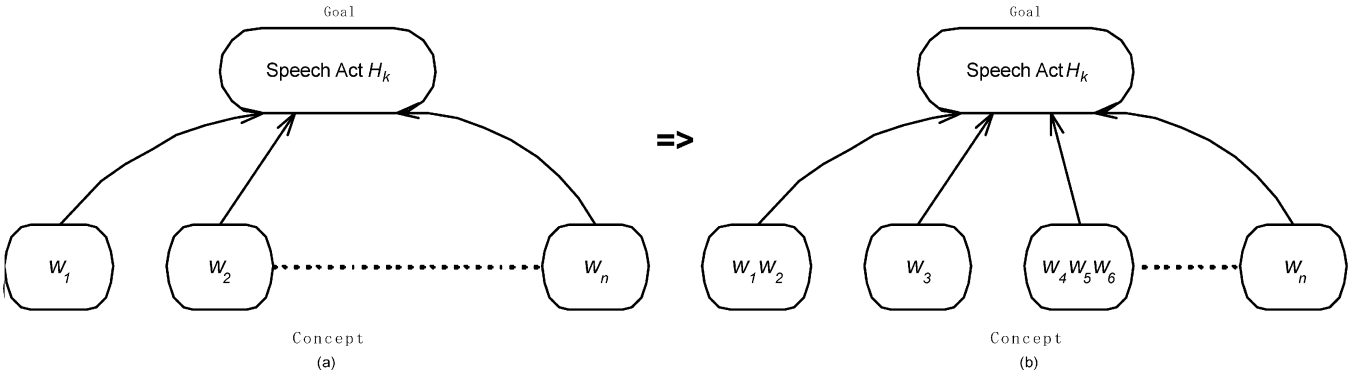


Fig. 8. (a) Topology of the Bayesian belief model. (b) Proposed topology of LSA-based Bayesian belief model.

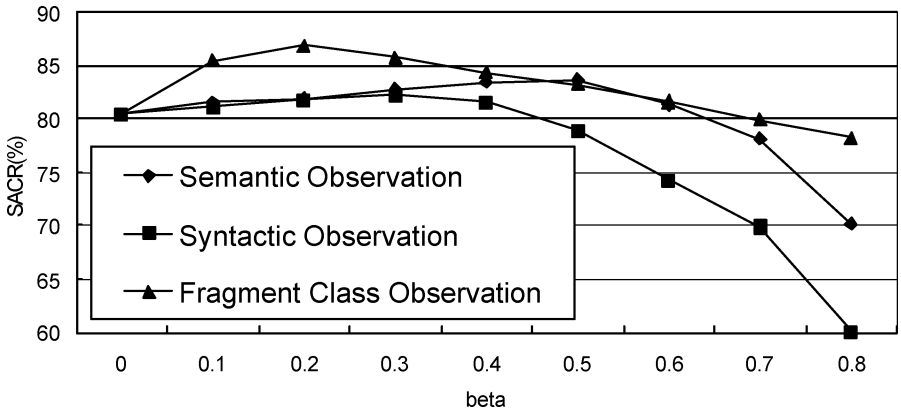


Fig. 9. SACR as a function of β for each observation in SAHMM.

IV. Experimental Results

Performance evaluation experiments were conducted on a spoken dialogue system for air travel information service. In this approach, the dialogue system accepts one sentence at a time and the experimental corpus involves one utterance (one sentence) in a dialog turn. The collected corpus of spoken language in human-machine interaction was developed in two steps. In the first step, an initial corpus was collected from a real air travel information service over approximately a month. This corpus contained 4250 spontaneous utterances collected in fluent spoken dialogue and real human interactions. These data were transcribed and used to develop an initial prototype air travel information service system. In the second step, the database was augmented in a wizard environment with subjects brought into the lab and given scenarios to work through. That is, there is a real person (the wizard) to control some of the functions to carry out spoken interactions with users, who are made to believe that they are interacting with a real spoken dialog system. Over 2200 utterances were collected from users of this prototype system. In practical human-machine interaction, the second corpus is helpful in improving the performance of the system because it concerns various habits and behaviors that people display when interacting with a computer service system [32]. The average length of the sentences and the perplexity in this corpus was smaller than that in the human/human corpus. This is because natural spoken language would be more complex in a sentence. If the user found that he/she was talking with a computer, he/she would speak shortly and simply [32]. This effect also affected the average turns to complete a dialog in air travel information service. The two corpora were then orthographically transcribed and tagged into a training text corpus containing fluent and disfluent sentences.

Two-hundred and six fragments were extracted using the collected corpus, and clustered into 38 fuzzy fragment classes [8]. According to the selection criterions of key-phrases proposed in [8], 12 groups of key-phrases, which included "Date Type I," "Date Type II," "Place," "Time Type I," "Time Type II," "Airline," "About Fragment," "Ending," "Greeting," "Inquiry," "Affirmative," and "Negative," were manually defined and employed to divide the corpus into 30 sub-corpora, each corresponding to a single speech act. Sentences in the same sub-corpus were consistent in the combination of key-phrase groups and defined as referring to the same speech act. These 30 types of speech acts are listed in the Appendix. For comparison, a speech-understanding framework based on key-phrase detection and verification [9] was implemented using 206 fragments as keywords. The key-phrase, which represents some task-related meaning in a sentence, was defined as the set of fragments to fill the semantic slots in air travel information service task. All the settings were prepared as required. The SAHMM's were then constructed and trained using the corresponding sub-corpus to model speech acts. The radius used in fuzzy c-means algorithm [8], [24] was set to 0.7. Experiments on fluent speech were conducted using a testing database established from 25 speakers (15 male and ten female). A total of 480 dialogues were used, containing 3038 sentences. For disfluent speech, the testing database included a total of 352 disfluent sentences.

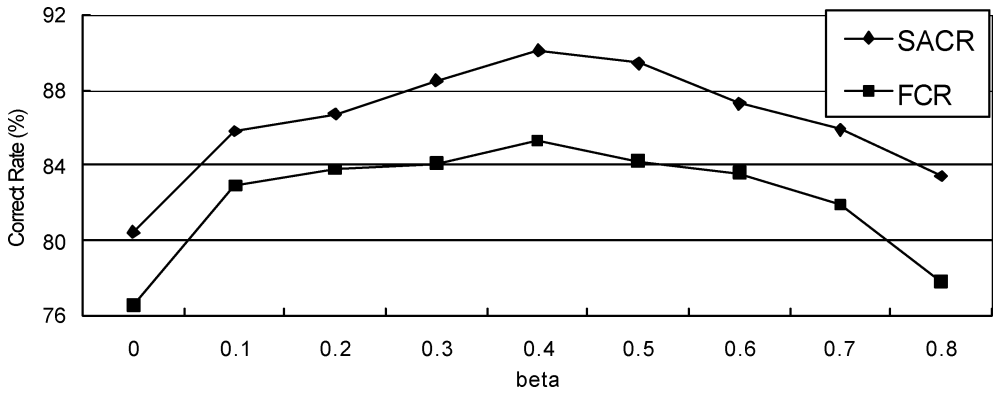


Fig. 10. SACR and FCR as functions of β for SAHMM.

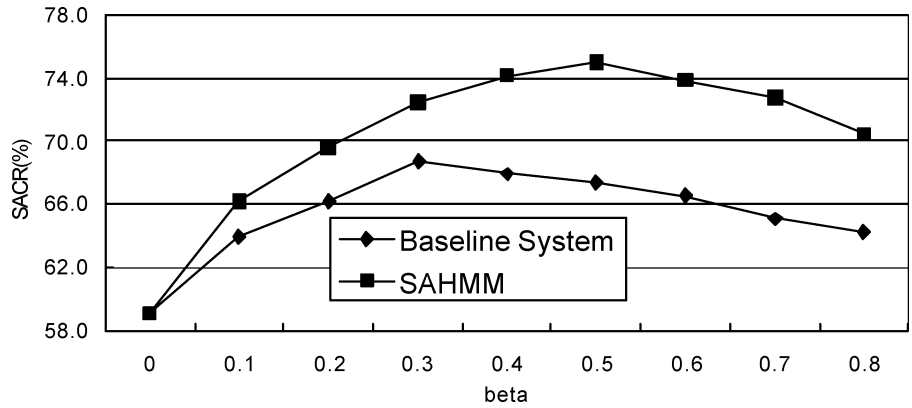


Fig. 11. SACR as a function of β for SAHMM on testing database with disfluency.

A. Experiment on the Performance of SAHMM

Thirty speech acts, each modeled by an SAHMM, were defined to evaluate the proposed approach. Given a fragment sequence, the SAHMM with the highest score obtained from (24) determines the speech act output. The fragment correct rate (FCR) and speech act correct rate (SACR) defined in [8] were adopted to evaluate the performance of the proposed approach. From the testing database for fluent speech, 3038 spoken sentences were fed into the speech recognizer to output the fragment sequences. An experiment on the SACR for various observations was conducted, as shown in Fig. 9. In this figure, the three characteristics: semantic information, syntactic structure and fragment class show their improvements in SACR, especially the fragment class. These performances yield the best results after the weights are determined by the EM algorithm. Fig. 10 shows the SACR and FCR on the proposed SAHMM. The best SACR achieved 90.1% with an 85.3% FCR when β was 0.4. Apparently, high performance in SACR and FCR is because the EM algorithm determines a closed-form solution for mixture weights. Additionally, the proposed approach adopting the semantic information, syntactic structure and fragment class also provides the ability to avoid confusion and ambiguity among speech acts.

B. Experiments on Disfluency Modeling

In this experiment, disfluencies such as "ah," "ung," "um," "em," and "hem" were investigated and their corresponding HMM's were trained and integrated into the recognition process [20], [21]. The input speech was recognized as possible fragment sequences using the acoustic fluency HMM's and five disfluency HMM's. The interpolation method was applied to disfluent speech. A baseline system, without interpolation, was established and the state-transition probability was estimated according to (11) to evaluate the efforts herein to solve the disfluency problem. The SAHMM's and the baseline system were trained by the same training corpus. The experimental results presented in Figs. 11 and 12 plot the SACR and FCR for different values of β. The best SACR and FCR of the baseline system were 68.8% and 63.2%, respectively, when $\beta = 0.3$. The performance is poorer than that obtained with fluent speech because ungrammatical sentences due to filled pauses must be modeled. The performance of the SAHMM with filled pauses shows a 75.0% SACR and a 70.4% FCR when $\beta = 0.5$. Obviously, the modeling of filled pauses is effective in solving the disfluency problem when SAHMM is used.

C. Experiment on the LSA-Based BBM for Speech Act Verification

In this experiment, two systems were conducted and compared to evaluate the performance of speech act verification with LSA-based BBM and classical BBM. For speech act verification with classical BBM, the verification probability of (25) is often simplified to Bayesian probability model [30] by assuming that all fragments are statistically independent as described in (26). The experimental results presented in Fig. 13 give the false acceptance rate (FAR) and false rejection rate (FRR) for various

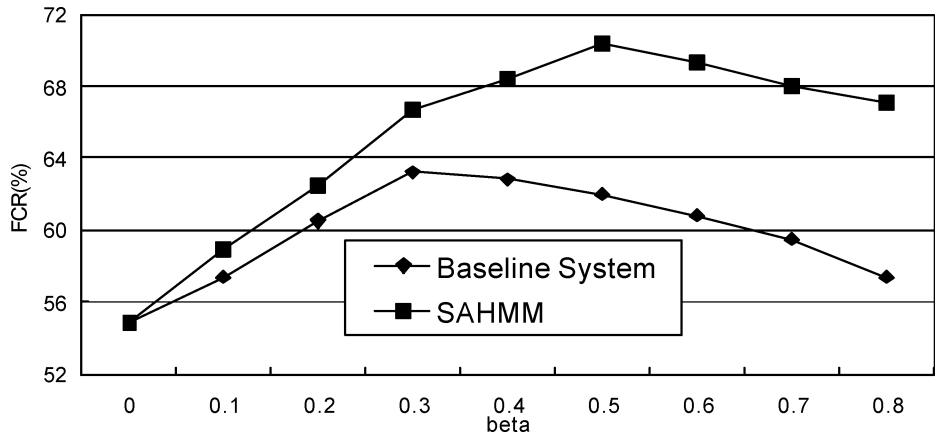


Fig. 12. FCR as a function of β for SAHMM on testing database with disfluency.

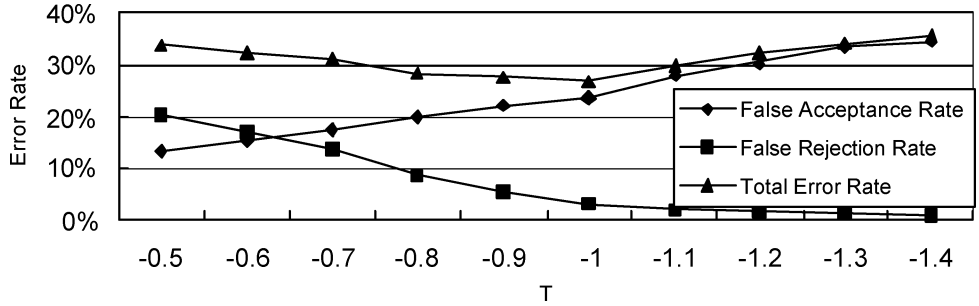


Fig. 13. Total error rate as a function of the threshold T in SAHMM with classical BBM for speech act verification.

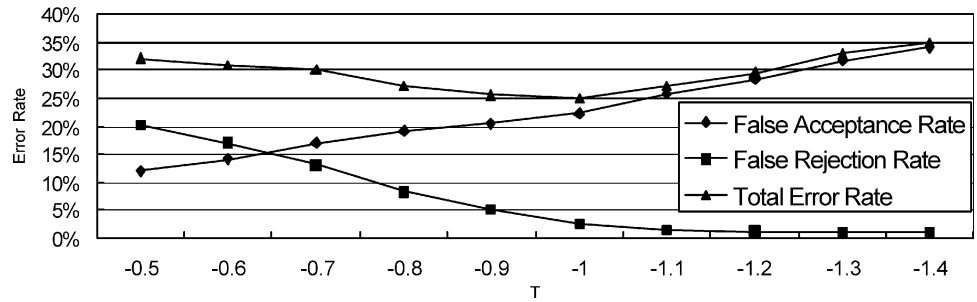


Fig. 14. Total error rate as a function of the threshold T in SAHMM with LSA-based BBM for speech act verification.

values of the threshold T. When the threshold T was chosen as -1, the sentence rejection rate was approximately 5.4% and the lowest total error rate achieved 26.9% with a false rejection rate of 3% and a false acceptance rate of 23.9%.

For speech act verification with LSA-based BBM, singular value decomposition of the association matrix R was performed. Various numbers of singular values were considered and $D = 35$ was found to yield an acceptable reconstruction error. The experimental results presented in Fig. 14 give the false acceptance rate (FAR) and false rejection rate (FRR) for various values of the threshold T. When the threshold T was chosen as -1, the sentence rejection rate was approximately 5% and the lowest total error rate achieved 25.1% with a false rejection rate of 2.6% and a false acceptance rate of 22.5%. Taking rejection into consideration, the modified FCR is defined as

$$MFCR = \frac{NA_C - NA_D - NA_I}{NA_{TPS}} \tag{33}$$

where NA_C is the correct number of fragments; NA_D is the number of deleted fragments; NA_I is the number of inserted fragments, and NA_{TPS} is the total number of fragments in the accepted sentences. The modified SACR is

$$MSACR = \frac{N_C(SA) - N_{FR}}{N_{TS} - N_R} \tag{34}$$

where N_{TS} is the total number of testing sentences; N_R is the number of rejected sentences; $N_C(SA)$ is the correct number of speech acts, and N_{FR} is the number of falsely rejected sentences.

The primary purpose of verification is to reject errors that have been accepted by the SAHMM. Sometimes, the accepted sentences are grammatical. However, that sentences are grammatical does not necessarily imply that they are meaningful in a specific speech act and so should sometimes still be rejected. The performance of the system yielded a 92.5% MSACR and an 87.1% MFCR after verification with classical BBM. The best

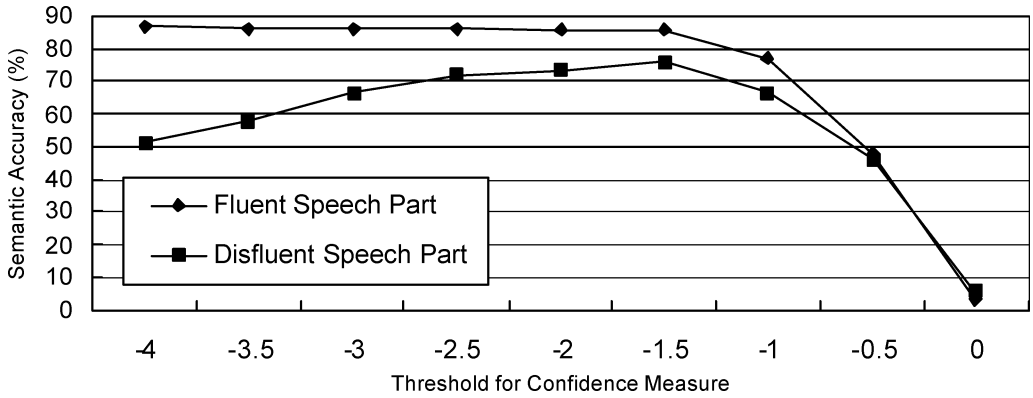


Fig. 15. Semantic accuracy versus the threshold for confidence measure.

TABLE II
COMPARISON OF PERFORMANCE OF KEYWORD-BASED SYSTEM AND SAHMM

	Fluent Speech	Disfluent Speech	Overall Performance
Keyword-based System	85.8	73.6	84.5
SAHMM	90.3	75.5	88.7

performance of the system yielded a 95.3% MSACR and an 89.5% MFCR after verification with LSA-based BBM. Notably, using LSA-based BBM for speech act verification further improved the performance of the system because more latent and salient dependencies between fragments were captured by the LSA-based BBM for speech act verification.

D. Comparison of SAHMM and a Keyword-Based System

Finally, the proposed approach was compared with a keyword-based speech understanding system, which has shown acceptable results when applied to a car reservation task and a movie locator task [9]. The keyword-based system herein was applied to the air travel information service task. To compare the performance of the two approaches, semantic accuracy which evaluates the ability of detecting the semantic slots is defined as shown in (35) at the bottom of the page where $N_C(Sem\_Slot)$, $N_I(Sem\_Slot)$, $N_D(Sem\_Slot)$, $N_S(Sem\_Slot)$ are the number of correct, inserted, deleted and substituted semantic slots respectively. $N_T(Sem\_Slot)$ represents the total number of semantic slots involved in the testing sentences. Fig. 15 shows the semantic accuracy obtained for fluent and disfluent speech, with respect to the threshold values. The best value of the threshold is between -1.5 and -2.0. The keyword-based system yielded semantic accuracies of 73.6% and 85.8% for the disfluent and fluent speech, respectively, corresponding to a threshold of -1.5. These results were compared with the performance presented in Section IV-C, which were transformed into the benchmark semantic accuracies of 75.5% and 90.3% for disfluent and fluent speech, respectively. Table II lists all results that pertain to performance of semantic accuracy. The keyword-based system under-performed by 4.5% on fluent speech and by 1.9% on disfluent speech. The former result follows from the fact that the proposed method correctly models the semantic information and syntactic structure of a speech act and provides a robust language model to be used by the speech recognizer. Furthermore, while the flexible understanding strategy of the keyword-based system identifies the desired fragments that fill the semantic slots and prevents problems of ungrammatical forms in the language model related to disfluency, the proposed interpolation method still outperforms the keyword-based system. This is because the proposed method reduces the disfluency problem in the language model, such that the updated language model exhibits approximately the same improvement as the language model applied to fluent speech. Although our proposed method indeed moderates the disfluency problem, the improvement in performance for disfluent speech over the keyword-based approach was not as good as was obtained for fluent speech because the keyword-based approach considers only keywords and disregards the extraneous part, including disfluencies. Accordingly, the proposed approach represents an insignificant improvement over the keyword-based approach for disfluent speech. In this experiment, the overall performance of the proposed SAHMM with LSA-based BBM verification was 4.2% better than that of the keyword-based system.

V. CONCLUSION

This work presents an SAHMM to characterize the semantic information and syntactic structure of a speech act and used to identify a speech act of the input sentence. The semantic information, syntactic structure and fragment class of an input sentence are statistically encapsulated in the proposed SAHMM to characterize the speech act. An interpolation method is applied to re-estimate the transition probabilities of the SAHMM concerning

$$Sem\_Accuracy = \frac{N_C(Sem\_Slot) - N_I(Sem\_Slot) - N_D(Sem\_Slot) - N_S(Sem\_Slot)}{N_T(Sem\_Slot)} \quad (35)$$

disfluencies. The verification strategy involves the LSA-based BBM, which not only captures possible inter-dependencies but also explores salient and latent semantic information. The verification score is then used to verify the speech act candidate and thus reduces the false acceptance rate caused by wide variations in utterances in the real-world environment. Most main design procedures in this paper are statistical and corpus-based, and so can be easily ported semi-automatically to other dialogue systems, whose portability can thus be improved. Experimental results show that for disfluent speech, the SAHMM yields a significant improvement of 6.2% in SACR and 7.2% in FCR over the baseline system without considering filled pauses. For fluent speech, the experimental results indicate that the performance can achieve 90.1% in SACR and 85.3% in FCR. When the LSA-based BBM is further applied to verify speech act, the best results yielded by the system were 95.3% MSACR and 89.5% MFCR with a rejection rate of 5%. This improvement follows from the fact that the LSA-based BBM for speech act verification captured the latent and salient dependencies between fragments. The semantic accuracy of the proposed approach was 4.2% higher than that of the keyword-based system.

APPENDIX

Speech Act 1:	Provide the information of Date for query.
Speech Act 2:	Provide the information of Location for query.
Speech Act 3:	Provide the information of Airline for query.
Speech Act 4~5:	Provide the information of Time for query.
Speech Act 6:	Provide the information of Date and Location for query.
Speech Act 7:	Provide the information of Date and Airline for query.
Speech Act 8~10:	Provide the information of Date and Time for query.
Speech Act 11~12:	Provide the information of Time and Location for query.
Speech Act 13~15:	Provide the information of Date, Time, and Location for query.
Speech Act 16~17:	Inquire arrival/departure time.
Speech Act 18~21:	Inquire arrival/departure time of a specific airline.
Speech Act 22~23:	Inquire the flight around a specific arrival/departure Time.
Speech Act 24~26:	Greeting.
Speech Act 27:	Ending.
Speech Act 28:	Acknowledgment.
Speech Act 29:	Confirmation “Yes”.
Speech Act 30:	Confirmation “No”.

REFERENCES

[1] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington, “JUPITER: A telephone-based conversational interface for weather information,” *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 85–96, Jan. 2000.

[2] E. Levin, R. Pieraccini, and W. Eckert, “A stochastic model of human-machine interaction for learning dialog strategies,” *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 11–23, Jan. 2000.

[3] G. Riccardi and A. L. Gorin, “Stochastic language adaptation over time and state in natural spoken dialog systems,” *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 3–10, Jan. 2000.

[4] K. Arai, J. H. Wright, G. Riccardi, and A. L. Gorin, “Grammar fragment acquisition using syntactic and semantic clustering,” *Speech Commun.*, vol. 27, no. 1, pp. 43–62, Feb. 1999.

[5] A. L. Gorin, A. Abella, T. Alonso, G. Riccardi, and J. H. Wright, “Automated natural spoken dialog,” *IEEE Computer Mag.*, vol. 35, no. 4, pp. 51–56, Apr. 2002.

[6] S. Rossato, H. Blanchon, and L. Besacier, “Speech-to-speech translation system evaluation: Results for French for the NESPOLE! project first showcase,” in *Proc. ICSLP*, 2002, pp. 1905–1908.

[7] H. U. Block, “The language components in VERBMOBIL,” in *Proc. ICASSP*, 1997, pp. 79–82.

[8] C. H. Wu, G. L. Yan, and C. L. Lin, “Speech act modeling in a spoken dialog system using a fuzzy fragment-class Markov model,” *Speech Commun.*, vol. 38, pp. 183–199, 2002.

[9] T. Kawahara, C. H. Lee, and B. H. Juang, “Flexible speech understanding based on combined key-phrase detection and verification,” *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 558–568, Nov. 1998.

[10] H. Kim, J. M. Cho, and J. Seo, “Fuzzy trigram model for speech act analysis of utterances in dialogues,” in *Proc. FUZZ-IEEE*, 1999, pp. 598–602.

[11] Y. Y. Wang and A. Acero, “Combination of CFG and N-gram modeling in semantic grammar learning,” in *Proc. Eurospeech*, 2003, pp. 2809–2812.

[12] L. Li and W. Chou, “Improving latent semantic indexing based classifier with information gain,” in *Proc. ICSLP*, 2002, pp. 1141–1144.

[13] H. Meng, W. Lam, and K. F. Low, “Learning belief networks for language understanding,” *Proc. ASRU*, 1999.

[14] G. Savova and J. Bachenko, “Designing for errors: Similarities and differences of disfluency rates and prosodic characteristics across domains,” in *Proc. EuroSpeech*, 2003, pp. 229–232.

[15] S. S. Rubén, B. Pellom, W. Ward, and J. M. Prado, “Confidence measures for dialogue management in the CU communicator system,” in *Proc. ICASSP*, 2000, pp. II1237–II1240.

[16] R. C. Rose, H. Yao, G. Riccardi, and J. Wright, “Integration of utterance verification with statistical language modeling and spoken language understanding,” in *Proc. ICASSP*, 1998, pp. 237–240.

[17] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, “Prosody-based automatic segmentation of speech into sentences and topics,” *Speech Commun.*, vol. 32, no. 1-2, pp. 127–154, 2000.

[18] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Reading, MA: Addison-Wesley, 1999, pp. 48–49.

[19] J. R. Bellegarda, “Exploiting latent semantic information in statistical language modeling,” *Proc. IEEE*, vol. 88, no. 8, pp. 1279–1296, Aug. 2000.

[20] C. H. Wu and G. L. Yan, “Acoustic feature analysis and discriminative modeling of filled pauses for spontaneous speech recognition,” *J. VLSI Signal Process.*, vol. 36, pp. 87–99, 2004.

[21] ——, “Discriminative disfluency modeling for spontaneous speech recognition,” in *Proc. EuroSpeech*, 2001, pp. 1955–1958.

[22] C. H. Wu and J. H. Chen, “Template-driven generation of prosodic information for chinese concatenative synthesis,” in *Proc. ICASSP*, 1999, pp. 65–68.

[23] J. Allen, *Natural Language Understanding*. New York: Benjamin/Cummings, 1994, pp. 542 and 554–557.

[24] H. J. Zimmermann, *Fuzzy Set Theory and Its Applications*. Norwell, MA: Kluwer, 1991, pp. 230–236.

[25] H. Meng and K. C. Siu, “Semiautomatic scquisition of semantic structures for understanding domain-specific nature language queries,” *IEEE Trans. Knowledge Data Eng.*, vol. 14, no. 1, pp. 172–181, Jan. 2002.

[26] F. Jelinek, R. Mercer, and S. Roukos, “Classifying words for improved statistical language models,” in *Proc. ICASSP*, 1990, pp. 621–624.

[27] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

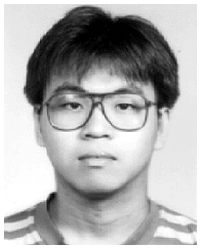
[28] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1999.

[29] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[30] D. W. Patterson, *Introduction to Artificial Intelligence & Expert System*. Englewood Cliffs, NJ: Prentice Hall, 1990, pp. 107–125.
[31] A. C. Rencher, *Multivariate Statistical Inference and Applications*. New York: Wiley, 1998.
[32] G. Riccardi and A. L. Gorin, "Stochastic language adaptation over time and state in natural spoken dialog systems," *IEEE Trans. Speech Audio Processing*, vol. 81, pp. 3–10, Jan. 2000.

Chung-Hsien Wu (M'94–SM'03) received the B.S. degree in electronics engineering from National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., in 1981, and the M.S. and Ph.D. degrees in electrical engineering from National Cheng Kung University, Tainan, Taiwan, in 1987 and 1991, respectively.

Since August 1991, he has been with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. He became a professor in August 1997. From 1999 to 2002, he served as the Chairman of the Department. He also worked at Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, in summer 2003 as a Visiting Scientist. He is currently the Editor-in-Chief for *International Journal of Computational Linguistics and Chinese Language Processing*. His research interests include speech recognition, text-to-speech, multimedia information retrieval, spoken language processing and sign language processing for hearing-impaired.

Dr. Wu is a member of International speech communication association (ISCA) and ROCLING.



Gwo-Lang Yan received the B.S. degree in information computer engineering from Chung-Yuan Christian University, Chung-Li, Taur Yuan, Taiwan, R.O.C., in 1995, and the M.S. degree in computer science information engineering from National Cheng Kung University in 1997. He is currently pursuing the Ph.D. degree in School of Computer Science Information Engineering, National Cheng Kung University, Tainan, Taiwan, and is the instructor at the Department of Information Management, Kao-Yuan Institute of Technology, Kaohsiung, Taiwan.

His research interests include digital signal processing, speech recognition, keyword spotting, and natural language processing.