2010 11th Int. Conf. Control, Automation, Robotics and Vision
Singapore, 7-10th December 2010

A Hybrid Computational Model for Spoken Language Understanding

Guangpu Huang and Meng Joo Er
School of Electrical and Electronics Engineering
Nanyang Technological University, Singapore
HU0002PU@ntu.edu.sg, EMJER@ntu.edu.sg

***Abstract*—This paper shows that the integration of statistical and connectionist methods can greatly enhance human-computer interaction through speech. The research approach is inspired by recent advances in high performance automatic speech recognition (ASR) systems and neurocognitive researches of natural language understanding (NLU). And a modest hybrid computational model is proposed and implemented to achieve intelligent spoken language understanding (SLU) in an information retrieval system.**

***Index Terms*—Human-computer interaction; spoken language understanding; speech recognition; natural language understanding.**

I. Introduction

The study of human-computer interaction through speech, commonly referred to as spoken language understanding (SLU), is derived from two parental research branches: automatic speech recognition (ASR), and natural language understanding (NLU). ASR research has traditionally been practiced in engineering departments to produce practical applications since the 1970s, for example, the ESPRIT SUNDIAL project in Europe, and the speech recognition and understanding project funded by the Advanced Research Projects Agency (ARPA) in the US [1]. One noticeable achievement is the hidden Markov model (HMM) based speech recognition systems developed during the ARPA project, where recognition accuracy of up to 95% has been obtained for large vocabulary continuous speech [1], [2]. On the other hand, NLU has grown mostly from symbolic researches in computer science, psycology and linguistic departments. This branch of research emphasizes more on theoretic studies of human intuition and cognition activities [3]. Differences in motivation, theoretical basis, techniques and applications have inhibited collaborations between these two departments, from which there are much to gain for both: ASR can provide efficient computational models and important information for prosodic, syntactic and semantic analysis; NLU can bring additional knowledge sources (KS) from the psychological and linguistic domains for human-like intelligent systems. Moreover, the integration affords the possibility of many other applications in artificial intelligence (AI) and holds great promises for man-machine interactions through conversational dialogues.

Currently there is a growing demand among SLU researchers to bridge the gap between these two schools of thoughts, NLU and ASR. At present, most of the developed SLU systems have very limited understanding ability and lack the contribution of cognitive psycholinguistic studies. In this direction, this work is carried out aiming to present a novel framework to address the issue of system integration.

The organization of this paper is as follows: in section II the relevant approaches on SLU systems will be briefly given. The framework of the proposed SLU system is described in section III. Section IV shows the detailed system implementation with its unique interactive process of statistical recognition and cognitive understanding. Discussion and conclusion of the project are given in section V and VI.

II. Related work

As shown in Fig. 1, a typical SLU system consists four units: speech analysis, speech recognition, language understanding, and the user interface (UI) to realize various interactions with the user. Though the field of speech processing has advanced enormously largely owning to the developments in computer technologies, SLU has turned out much more complicated than initially anticipated, and remained one of the puzzles in speech researches.

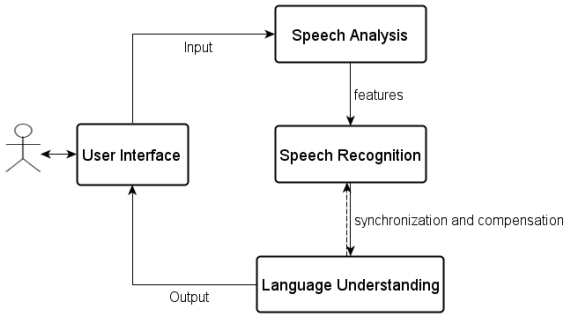

Fig. 1. Simplified overview of a typical SLU system.

During the 1980s , there emerged a series of useful speech recognition and understanding systems such as Carnegie Mellon University (CMU) Harpy, Hwim, System Development Corporation(SDC) and Hearsay-II, which are the ancestors of most of today's speech recognition systems used in telecommunication and personal computers [1], [4]. These systems achieve automatic speech recognition and understanding based on the statistical method, *hidden Markov modeling* (HMM), which computes language probability measures at phonetic, lexicon, syntactic, semantic and prosodic levels. More on

HMM is introduced in the section III, interested readers can also find useful information in [2], [5], [6], [7], [8].

In contrast, the branch of connectionist methods are based primarily on application of psycholinguistic theories of human language [9], [10], [11]. HMM-based systems are able to achieve high-accuracy speech recognition especially for small to medium vocabulary tasks. However, unlike the computer with its computational power, the human brain recognizes and understands speech rather in a cognitive way, which has motivated the statistical research to merge with the study of artificial neural networks (ANN), or connectionism [3], [4], [12], [13], [14]. And the last two decades have brought a number of hybrid HMM/ANN systems. He and Young proposed to encode semantic context into HMM system using hidden vector state (HVS) [15]; Friederici developed a neurocognitive model to demonstrate a parallel complementary process performed by syntactic and semantic analysis in human brain, indicating the possibility of unifying computational systems to represent the cognitive process [16], [17]; Kompe uses accent analysis, or prosodic information, to enhance the speech recognition performance, alongside similar works by others like Johnson and Gallwitz [18], .

III. Proposed SLU System Framework

This section introduces the two core components of the proposed SLU model, the statistical HMM-based speech recognition segment and the cognitive ANN-based language understanding segment.

A. HMM-based ASR: Theory and Structure

Statistical methods are by far the most popular algorithm for automatic speech recognition and understanding as it allows the machine to learn, directly from data, structure regularities in the speech signal. Basically, the hidden Markov model generates an "utterance" through a transaction from the entry state to the exit state. The likelihood of a certain output produced by a particular state sequence is calculated as the product of the probabilities for state transitions taken and the output probability densities for each speech vector from the corresponding HMM state. For automatic recognition the agenda is to find the likelihood that the model generated the input speech while the state sequence is unknown, or "hidden". In other words, it is the reversed calculation from the output to the input [2], [3].

Given the observed vector $Y = \vec{y}_1^T$ of certain acoustic features computed every 10 to 30 ms time interval, where vector $\vec{y}_1^T = [y_1, y_2, \cdots, y_T]$, statistical method employs HMM to find the most likely word sequence $\hat{W}$ with a maximization procedure, applying Baye's theorem:

$$\hat{W} = \arg\max_W (W|Y) = \arg\max_W \frac{p(Y|W)p(W)}{p(Y)}, \qquad (1)$$

where $\hat{W}$ corresponds to the candidate having maximum a-posteriori probability (MAP), and $p(Y|W)$, or $p(\vec{y}_1^T|W)$ is computed by acoustic models (AM) and $p(W)$ is computed

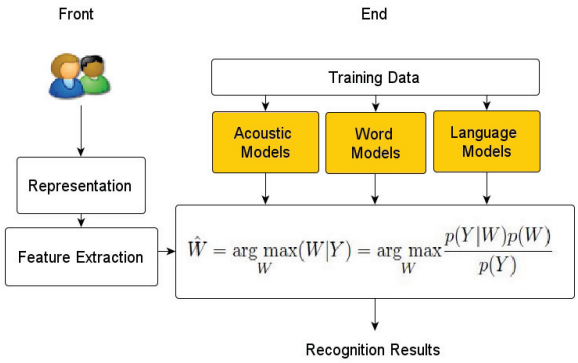

Fig. 2. An example of HMM-based speech recognition process.

by language models (LM). A typical HMM-based speech recognition process is shown in Fig. 2.

In Fig. 2, the *acoustic models* are the elementary probabilistic models of basic linguistic units e.g., phonemes which are used to build word representations. HMM consists of two stochastic processes, a *hidden* Markov chain, which accounts for *temporal* variability, and an observable process, which accounts for *spectral* variability. This combination has proven to be the most efficient to cope with speech ambiguity, and flexible enough to cope with large dictionaries [4], [8].

Secondly, *word models* represent words by networks of phonemes, each path in line with a pronunciation of a certain word. There exist many types of word models, for example, *allophone* models, *polyphones*, classification and regression tree (CART) *allophones* and so on. Word models can be constructed by sharing distributions taken from a pool of *senones*, or by concatenation of basic units, e.g., *fenones* and *multones*, made by states, transitions and probability distributions [8], [19].

Last but not the least, *language models* generate the probability $p(W)$ of a sequence of words $W = w_1, \cdots, w_L$, which is expressed as:

$$p(W) = p(w_1, \cdots, w_n) = \prod_{i=1}^{n} p(w_i|w_0, \cdots, w_{i-1}), \qquad (2)$$

where w_0 is an approximation of the initial condition. The probability of the next word w_i depends much on the history of words that have previously been spoken, thus the complexity of the model grown exponentially with the length of the history. In practice the history h_i is often modeled to simplify the process, and only certain length of h_i is used. The most popular method is the simple *n-gram* model, particularly the trigram model ($n = 3$), where only the most recent two words are used to compute the probability of the next words,

$$p(W) \approx \prod_{i=1}^{n} p(w_i|w_{i-2}, w_{i-1}). \qquad (3)$$

B. Neurocognitive NLU

The language theory by Chomsky states that speech is "understood" through the coherent analysis carried out by

the brain, where the Brodmann area (BA) is responsible for speech production and perception. And investigation in this area provides the basis for psycholinguistic studies of language understanding, which leads us to the neurocognitive model of auditory sentence processing proposed by Angela Friederici [15], [17], [20]. This model identifies the brain activities during speech perception and understanding, which consist of approximately five cognition phases: phonological segmentation and sequencing (up to 100ms); syntactic structure building (between 150 and 200 ms); semantic relations and role assignment definitions (between 300 and 500 ms); prosodic processing (between 500 and 800 ms); syntactic-semantic-prosodic integration (between 800 and 1000 ms). The simplified model is shown in Fig. 3.

The extraction and modeling of speech features in Friederici's model is in a way similar to the previously introduced the HMM-based method. However, the cognitive model differs from the HMM system in the sense that it relies on the use of biological/artificial neuron nets and knowledge sources to activate the understanding of speech. For conversational speech each has their own advantages and disadvantages, e.g. HMM excels in finite vocabulary speech-to-text recognition but it requires adequate training and fails to handle noise, out of vocabulary words and other unfamiliar events; cognitive model is more robust to the latter mentioned situations and generally outperforms HMM in deeper-level understanding tasks using *unit labeling*, *semantic segmentation*, and *meaning extraction*. Therefore, it is desirable to take advantage of both models to improve the performance of SLU systems.

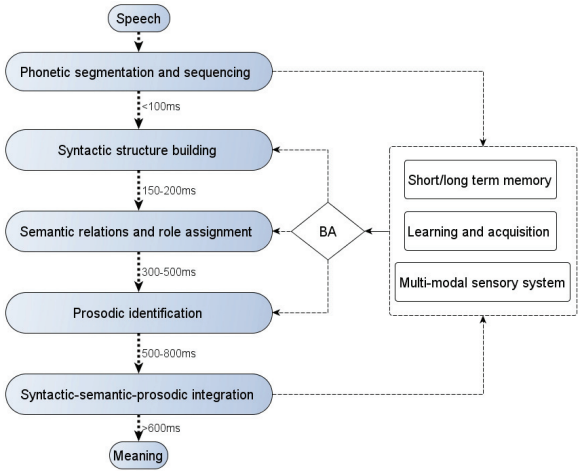

Fig. 3. Neurocognitive model of language understanding process [20].

IV. SLU Model Implementation.

The proposed SLU model intends to provide a framework to better serve the purpose of machine learning and interaction with human through the synchronization and compensation of three main speech processing units: pre-processing, speech recognition using CMU Sphinx-4 and neurocognitive speech understanding using neural nets. A schematic diagram of the spoken language understanding system is shown in Fig. 4.

A. Pre-processing

In the first step, the unknown speech is received/recorded by a recording device through a simple user interface (UI) using stereo (2 channel), 16 bit, 44100 Hz sampling rate format which is sufficient to retrieve the original speech signal information [21] . Then endpoint detection is applied to extract the speech segments between the beginning and ending points, where simple energy measure E is used to differentiate speech from silence, as the testing environment is inside the lab and noise is only caused by computer stations:

$$E(n) = \sum_{m=0}^{N-1} w(m)x(n-m)^2, \tag{4}$$

where N samples of $x(n)$ is selected through a weighing window $w(m)$. A suitable choice of N is on the order of 100-200 for a 10 kHz sampling rate [8]. The extracted speech segment is pre-emphasized using a simple first order digital filter with z-transform:

$$H(z) = 1 - az^{-1}, \tag{5}$$

where a takes a value in [0, 1], e.g., $a = 0.95$ is used here to compensate for the -6 dB spectral slope of the speech signal.. The step serves to reduce the distance variance of the parameters obtained in the autocorrelation analysis. The autocorrelation coefficients are calculated from overlapping frames of length $N = 300$ samples, or 45 ms using a Hamming window on the data. For the input sequence of speech data $x(n)$, linear predication or linear predictive coding (LPC) analysis is then applied,

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + Gu(n) \tag{6}$$

where $a_k, k = 1, 2, \cdots, p$ are the autoregressive parameters or the predictor coefficients, p is the order of the LPC analysis, and $u(n)$ is the driving function, with G as the gain parameter. And the set of parameters of a_k is then determined to best approximate the original signal $x(n)$ by minimization of the prediction error $e(n)$,

$$e(n) = x(n) - \tilde{x}(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k), \tag{7}$$

which is equivalent to $S(z) = E(z)/A(z)$ in the z-transform, where

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{8}$$

is a pth order polynomial. There are many different methods to obtain LPC coefficients through likelihood estimation, and detailed solutions for the autoregressive process are available in [2], [22]. These LPC coefficients are used for next-step processing and stored as reference patterns, i.e."user data" in the off-line unit.

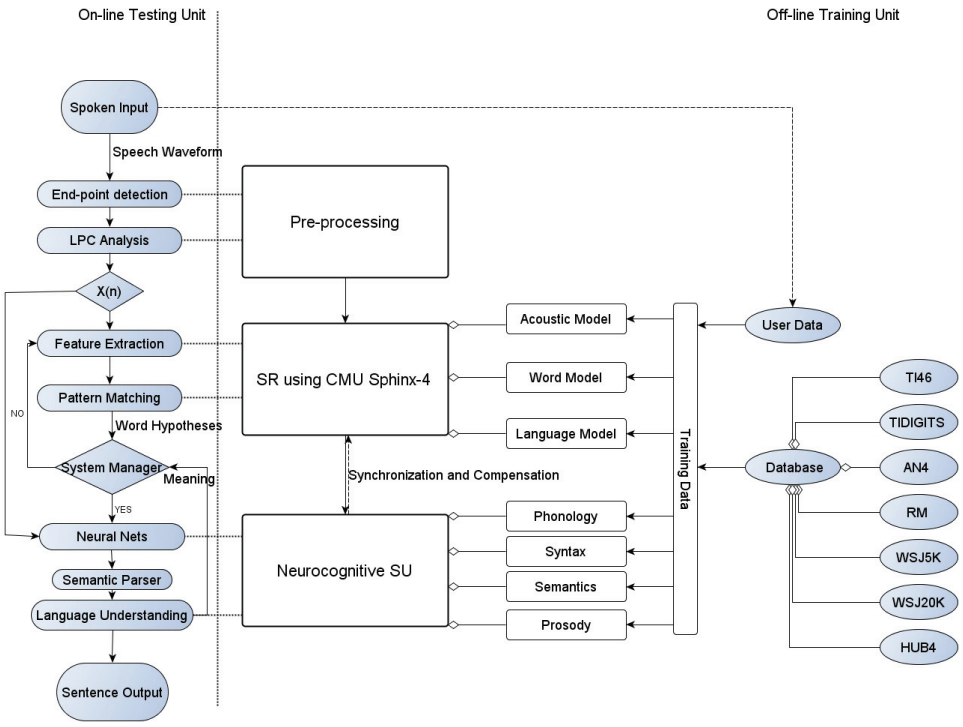

Fig. 4. Proposed structure of a spoken language understanding system.

B. Speech Recognition using CMU Sphinx-4

After pre-processing, the speech signal will be decoded using the HMM-based speech recognizer, CMU Sphinx-4, which is a state-of-the-art speech recognition system written in Java [21], [23], [24]. The recognizer user interfaces consists of the search manager, linguist, acoustic model, dictionary, language model, active list, scorer, pruner, and search graph. Speaker adaptation is realized through the adaptation and training of the user data when the user accesses the system interface. The two main acoustic models used by Sphinx-4 are TIDIGITS and Wall Street Journal; the associated language models used in Sphinx-4 were created with the CMU Statistical Language Modeling toolkit (SLM toolkit). In this step, the recognizer constructs the front end i.e., feature extraction, the decoder, and the linguist i.e., pattern matching, according to the user data configuration. Then the linguist will construct the acoustic model, the word model, and the language model. It will use the knowledge from these three components to construct a search graph that is appropriate for the recognition task. The recognizer, or decoder will construct the search manager, which in turn constructs the scorer, the pruner, and the active list [21], [24].

C. Neurocognitive Speech Understanding

This computational phase generates temporal registers and parsing trees through connectionist parsing in the SARDSRN-RAAM system, which is composed by three neural nets: sequential activation retention and decay network (SARDNET), simple recurrent network (SRN) and recursive auto-associative memory (RAAM) [15]. Then semantic and prosodic analysis are applied using the self-organizing map (SOM) to cluster similar speech patterns based on the obtained speech features. The system then computes the most likely semantic context for an interpreted sentence received from the previous step. SARDSRN-RAAM is in fact a shift-reduce parser with great capacity to generate parsing sequences serving as an artificial counterpart of the "Brodmann area" for decision making, where knowledge sources generated by NLU model are able to constrain the dynamic search space of the ASR engine into a domains defined by topics and dialogue history et cetera [25].

D. System Integration: Synchronization and Compensation

As shown in Fig. 4, the SLU system synchronizes the ASR and NLU procedures to simulate top-down statistical and bottom-up cognitive analysis of spoken input at phonetic, syntactic, semantic, and prosodic level, to obtain the meaningful message. Using unification grammar and rules the semantic parser identifies useful units such as noun phrases, verbs, function words, and conjunctions in the recognized word sequences generated by the SPHINX engine. More importantly, when the initial recognized sequence failed to give meaningful results, the recognizer is modified by the system manager to compensate and re-estimate the recognition process and give the top two most likely word sequences. Examined closely, speech knowledge sources such as acoustic cues, syntax, and semantics in the ASR step are modeled using probabilistic measures and then computed based on maximum likelihood estimation (MLE). In other words, the recognized sentence is no more than a hypothesis of the most likely word sequence, which may or may not make any sense in practice. For

instance, in continuous speech, coarticulation usually causes much degradation of ASR performances as observed in many experiments [26], [27]. An input query "*a single ticket to Singapore*" in normal speech would sound like "*a single tiki two Singapore*", where the stop sound /t/ of "*ticket*" is propagated into the following word "*to*" , in which case, the system needs to be aware of the variations of pronunciations in conversational speech. Encoding all these variations is unrealistic and unpractical considering the many sources of variations e.g., accents, cross-speaker, background speech et cetera. This is when it is useful to bring psychological and linguistic knowledge into the process to ensure meaningful construction of sentence hypothesis. Thus "*tiki*" is more likely to be matched with "*ticket*" since it is connected with the noun phrase"*destination_Singapore*" .

The system manager is defined with a set of parameters to best serve the specific task, for a general-purpose flight inquiry system (i.e., speaker-independent, medium vocabulary, continuous speech input), a syntactic error rate higher than 0.5 means a rejection of the corresponding semantic structure; a distance in the sentences map higher than 2 means a context rejection, and the system manager then points out the next best sentence of the semantic analysis. The iterative interaction between statistical speech recognition and cognitive speech understanding is also specified to have ≤ 5 cycles to ensure output accuracy of $\geq 95\%$ as well as computational efficiency of 2 seconds response time. In addition, manual keybords on a touch screen would be provided as an alternative option for the user to adapt to the system as well as to ensure smooth and accurate transactions.

E. Information Retrieval Task

DARPA resource management (RM1) vocabulary is applied for training and testing in this paper for the simple speech understanding task that has been attempted in Air Travel Information System (ATIS). In the system, the user speaks to the machine to obtain flight information about flight time, destination, fair, single/return trip and other general queries as encountered in ATIS. The SR engine will extract a sequence of words and passes them to the NLU unit to obtain semantic representations of the inquiry through parsing, as shown in Fig. 6 in an ideal recognition performance. In practical applications, more than one hypothesis of the word sequence are often analyzed, and the process is often prune to high error rates in many current SLU systems [28]. The next step is to cope with the system manager to decide which information is to be provided to the user; and speech synthesis is used to feedback the retrieved flight information in addition to the traditional text output on the screen; an example of dialogue simulation between the user (U) and the SLU system (S) is given below. The overall design of the information retrieval system is shown in Fig. 5.

- U: "Good Morning!"
- S: Good Morning! how may I help you?
- U: "I would like a single ticket to Singapore next Wednesday and non-stop."

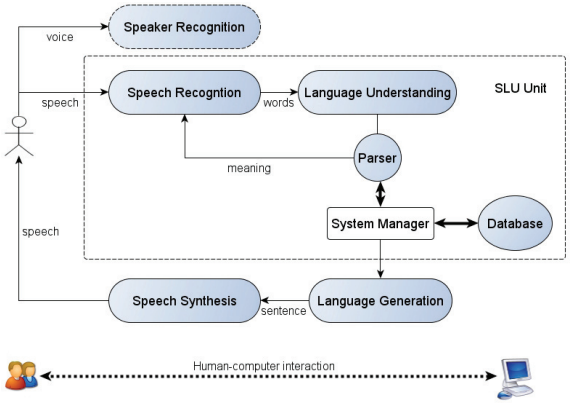

Fig. 5. Human-computer interaction of the information retrieval system.

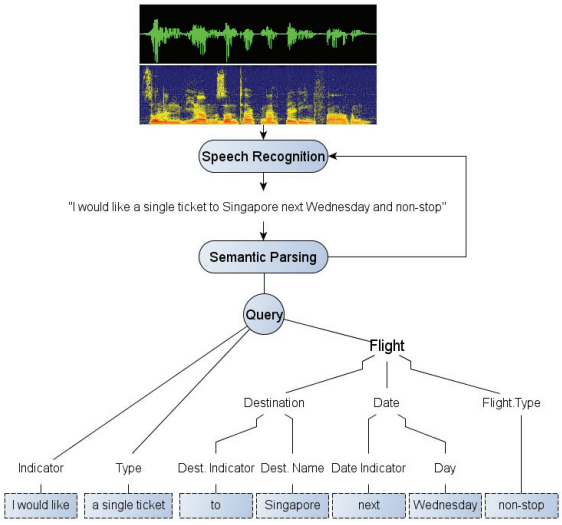

Fig. 6. Semantic representation of the user inquiry.

- S: Single ticket; destination Singapore; next Wednesday; non-stop; yes/no?
- U: "Correct."
- S: Tickets available at 1. $[18:00\text{pm}]$, 2. $[20:00\text{pm}]$ and 3. $[21:30\text{pm}]$.
- U: "Number three please."
- S: Please input your ID! ……

V. Discussion

The goal of this paper is to investigate the application of speech recognition and understanding methods to achieve robust and intelligent spoken language understanding (SLU). The main difficulty is the integration since ASR and NLU use two different sets of models and grammar rules. In order to fully exploit the advantages of these two units, the system manager needs to be further improved, for example, the threshold of recognition sequence acceptance and rejection rate based on the understanding results can be more flexible or strict depending on the specific testing condition. Such a task remains a major problem even in the heavily investigated AI

field, and requires intensive research efforts [28].

Another problem is the recognition of functional words such as "*and*","*to*" ,"*from*", "*for*" as in "*I would like a ticket to Singapore from London next Wednesday and non-stop*" . These words are important for the semantic analysis during information retrieval but are often difficult for the recognizer to identify especially in continuous speech [8]. One solution is to apply the keyword-spotting technique to extract meaningful messages from the input speech based on the assumption that actions are likely to be expressed in important keywords [22]. As shown in the dialogue in section IV-E, the recognizer would obtain key phrases like "*Singapore*", "*(next) Wednesday*" and "*non-stop*" which would significantly reduce the amount of searching in the database.

The next phase of this project will emphasis on the construction of speaker-independent acoustic models and task-oriented language models for more robust speech understanding. It is worthy noting that human conversations are more task/context-based than previously supposed, for example, the speaker's identity, facial expressions, gestures, topic, agenda, surrounding environment, and so on. Another direction is for the machine to have more human-like response during the tasks, for example, in the previous information retrieval task, instead of listing or telling the user all the available flights on "Wednesday", the machine could give more intelligent response like "flight available every four hours from $08:00$pm to $24:00$pm; additional flight available at $18:00$ pm".

VI. Conclusion

This paper proposed a modest SLU model to promote intelligent human-computer communication through conversational dialogues. The system incorporates statistical speech recognition technique with neurocognitive language understanding theories in a hybrid computational model. Speech recognition performance is improved through semantic parsing, and language understanding, and the system shows great potential in the flight inquiry task. Further system testing are to be finished to give detailed performance evaluation for future improvements. This paper is intended as a starting point for a complete dialogue-based spoken language processing systems for information retrieval tasks. The system will provide an excellent basis for design and development of more complexed SLU applications that can be embedded in various devices to improve the quality of daily lives and social services.

References

[1] D. Klatt, "Review of the arpa speech understanding project," *Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1345–1366, 1977.

[2] L. R. Rabiner, "Tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[3] R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, A. Zampolli, and V. Zue, "Survey of the state of the art in human language technology," 10 April 2009 1997.

[4] K. F. L. Alexander Waibel, *Readings in Speech Recognition*, M. B. Morgan, Ed. Morgan Kaufmann Publishers, Inc., 1992.

[5] L. R. Rabiner, B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of isolated digits using hidden markov models with continuous mixture densities," *AT&T Technical Journal*, vol. 64, no. 6 pt 1, pp. 1211–1234, 1985.

[6] L. R. Rabiner and B.-H. Juang, "Introduction to hidden markov models," *IEEE ASSP magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[7] L. R. Rabiner, B. H. Juang, and B. Keith, "Speech recognition: Statistical methods," in *Encyclopedia of Language & Linguistics*. Elsevier, 2006, pp. 1–18.

[8] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, United States Ed Edition, 1993.

[9] T. A. Harley, *The psychology of language : from data to theory*. Psychology Press, 2008.

[10] V. Zue, "The use of speech knowledge in automatic speech recognition," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1602–1615, Nov. 1985.

[11] V. Cook and M. Newson., *Chomsky's universal grammar: an introduction*. Blackwell Publishing, 2007.

[12] E. Trentin and M. Gori, "A survey of hybrid ann/hmm models for automatic speech recognition," *Neurocomputing*, vol. 37, no. 1-4, pp. 91 – 126, 2001.

[13] Y. C. Lee, H. H. Chen, and G. Z. Sun, "A neural network approach to speech recognition," *Neural Networks*, vol. 1, pp. 306 – 306, 1988.

[14] R. P. Lippmann, "Review of neural networks for speech recognition," *Neural Comput.*, vol. 1, no. 1, pp. 1–38, 1989.

[15] N. A. Muller N., de Siqueira L., "A connectionist approach to speech understanding," in *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, 0-0 2006, pp. 3790–3797.

[16] A. D. Friederici, "The developmental cognitive neuroscience of language: A new research domain," *Brain and Language*, vol. 71, no. 1, pp. 65 – 68, 2000.

[17] Y. Grodzinsky and A. D. Friederici, "Neuroimaging of syntax and syntactic processing," *Current Opinion in Neurobiology*, vol. 16, no. 2, pp. 240 – 246, 2006, cognitive neuroscience.

[18] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1 – 15, 1997.

[19] B. P. d. S. P. M. R. P. M. Bahl, L.R., "A method for the construction of acoustic markov models for words," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 443–452, 1993.

[20] A. D. Friederici, "Towards a neural basis of auditory sentence processing," *Trends in Cognitive Sciences*, vol. 6, no. 2, pp. 78 – 84, 2002.

[21] H. X. D. and L. K. F., "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, 1991, pp. 877–880 vol.2.

[22] F. S. Juang, H., "Automatic recognition and understanding of spoken language - a first step toward natural human-machine communication," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1142–1165, 2000.

[23] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language*, vol. 4, no. 2, pp. 127–165, 1990.

[24] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the sphinx speech recognition system," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 1, pp. 35–45, Jan 1990.

[25] S. R. Young, "Use of dialogue, pragmatics and sematics to enhance speech recognition," *Speech Communication*, vol. 9, no. 5-6, pp. 551 – 564, 1990.

[26] D. Reddy, "Speech recognition by machine: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 501–531, April 1976.

[27] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[28] R. D. Mori, F. Bechet, D. H. Tur, and G. T. Michael McTear, Giuseppe Riccardi, "Spoken language understanding," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 50–58, 2008.