Available online at www.sciencedirect.com

SCIENCE DIRECT

Knowledge-Based Systems 19 (2006) 153–163

Knowledge-Based SYSTEMS

www.elsevier.com/locate/knosys

Two-level speech recognition to enhance the performance of spoken dialogue systems

Ramón López-Cózar *, Zoraida Callejas [1]

Department of Languages and Computer Systems, Granada University, 18071 Granada, Spain

Received 24 September 2003; accepted 3 November 2005
Available online 13 December 2005

Abstract

Spoken dialogue systems can be considered knowledge-based systems designed to interact with users using speech in order to provide information or carry out simple tasks. Current systems are restricted to well-known domains that provide knowledge about the words and sentences the users will likely utter. Basically, these systems rely on an input interface comprised of speech recogniser and semantic analyser, a dialogue manager, and an output interface comprised of response generator and speech synthesiser. As an attempt to enhance the performance of the input interface, this paper proposes a technique based on a new type of speech recogniser comprised of two modules. The first one is a standard speech recogniser that receives the sentence uttered by the user and generates a graph of words. The second module analyses the graph and produces the recognised sentence using the context knowledge provided by the current prompt of the system. We evaluated the performance of two input interfaces working in a previously developed dialogue system: the original interface of the system and a new one that features the proposed technique. The experimental results show that when the sentences uttered by the users are out-of-context analysed by the new interface, the word accuracy and sentence understanding rates increase by 93.71 and 77.42% absolute, respectively, regarding the original interface. The price to pay for this clear enhancement is a little reduction in the scores when the new interface analyses sentences in-context, as they decrease by 2.05 and 3.41% absolute, respectively, in comparison with the original interface. Given that in real dialogues sentences may be out-of-context analysed, specially when they are uttered by inexperienced users, the technique can be very useful to enhance the system performance.

Keywords: Knowledge-based systems; Speech-based systems; Human–computer interaction; Dialogue systems; Automatic speech recognition; Speech understanding; Dialogue management

1. Introduction

Spoken dialogue systems can be considered knowledge-based systems developed to interact with users using speech in order to provide information or carry out a variety of simple tasks, such as travel information [14,26], language learning [7], car-driver assistance [2,4], weather information [20,31,34] and automatic call-routing [8,12], among others. Given that language is the most natural and efficient communication means for people, dialogue systems are developed to facilitate carrying out these tasks automatically, using eyes- and hands-free devices such as microphones and telephones. The initial systems were very limited regarding the user sentences and the types of task to carry out. However, the evolution of automatic speech recognition (ASR) and speech synthesis technologies in the last three decades has led to the development of sophisticated systems usable in real world conditions. Current systems offer a large potential for automation and increased functionality for telephone-based applications, allowing that users can talk more naturally, similarly as if they were talking to a human operator. Fig. 1 shows the typical structure of a current spoken dialogue system [19,21,23]. It is basically comprised of an input interface (speech recogniser and semantic analyser), an output interface (response generator and speech synthesiser), a dialogue manager between both interfaces, and some additional modules that provide knowledge to the previous modules.

The speech recogniser carries out the ASR process, i.e. receives the voice signal from the sentence uttered by the user and transforms it to a recognised sentence [3,29]. To do so, it uses acoustic models (AM) properly trained from a speech database, language models (LM) that determine the possible

* Corresponding author. Tel.: +34 958 240579; fax: +34 958 243179.
E-mail addresses: rlopezc@ugr.es (R. López-Cózar), zoraida@correo.ugr.es (Z. Callejas).

[1] Tel.: +34 958 240636; fax: +34 958 243179.

0950-7051/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2005.11.004

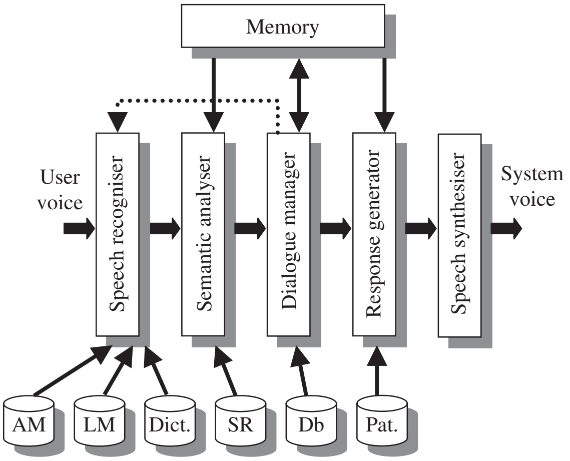

Fig. 1. Typical module structure of a spoken dialogue system.

sequences of words (sentences) and a Dictionary that contains all the possible words that can be recognised.

The semantic analyser finds out the meaning conveyed by the recognised sentences taking into account a set of semantic rules (SR) that map the syntactic and/or semantic structures found in the sentences to meaning representations, which are typically stored in the system memory in the form of semantic frames [1,5,6]. The work presented in this paper is concerned with the performance of the speech recogniser, while the experimental results shown in Section 3 are concerned with the performance of his module and the semantic analyser, given that both clearly influence the speech understanding process.

The dialogue manager implements the ‘intelligence’ of the system. It uses the semantic representation provided by the semantic analyser in the context of the dialogue and decides the next system response, which is typically built containing data extracted from a database (Db). The response can also be concerned with the dialogue management, such as prompting the user to confirm or rephrase the recognised sentence [13,32].

The response generator builds the system response, typically as a text sentence that must be syntactic and semantically correct [10,28]. To do so, it uses a set of patterns that contain some parts fixed and some others variable to be instantiated with the data extracted from the database. For example, the pattern ‘⟨company⟩ ⟨flight_id⟩ leaves at ⟨departure_time⟩ from gate ⟨gate_id⟩’ can be used to generate the sentence ‘American Airlines flight AA 1234 leaves at 18:30 h from gate B24’.

The speech synthesiser generates the system voice, using either a text-to-speech (TTS) conversion in the case of previously created text sentences [25], or playing pre-recorded segments (words and sentences). The TTS conversion is preferred when the vocabulary is very large, is unknown a priori or changes frequently. The recorded segments are used in small and fixed vocabulary applications since they generally provide more intelligibility, although some current TTS systems also provide excellent results.

The memory module supplies the dialogue context and historic information to several modules of the system. The semantic analyser uses contextual information to resolve possible anaphoric references in the user sentence (e.g. in the sentence ‘I want the first flight’, find the referent of ‘first’). The dialogue manager uses the historic information as a record of previous system and/or user actions, which can be useful to make the system behaviour more intelligent. For example, using this information the system can notice continuous misunderstanding of the user sentence and thus transfer the call to a human operator. The contextual information is also very useful for the response generator since it can make the system responses more human-like. Taking it into account the system can decide the use of anaphora (pronouns instead of nouns) and ellipsis (omission of unnecessary words) as humans do when uttering sentences within a context.

In despite of the advances made in the last years in terms of ASR and speech synthesis, spoken dialogue systems are still restricted to well-known application domains that provide very valuable knowledge about the words and sentences the users may likely utter. For example, in the Air Travel Information Service (ATIS) some likely words are airport and city names as well as travel dates, types, duration and fares. The domain knowledge is very important to create the system dictionary, since a word not included in it (called Out-Of-Vocabulary word) can never be recognised, and thus causes recognition errors (it can be either changed by another acoustically similar or discarded). The application domain also provides knowledge about the task the system must carry out. For example, this knowledge makes the system ask the user for the departure city if only the destination city was provided in a query to travel from one city to another.

1.1. Stochastic approach to ASR: acoustic and language knowledge

Several approaches have been developed to face the ASR problem, such as expert systems and artificial neural networks [24]. This problem can be stated as follows: ‘find the sequence of words uttered W, given a sequence of acoustic data A’. The technique presented in this paper is concerned with the stochastic approach, which is the one mostly used nowadays. According to this approach, W is the word sequence with the highest probability given the acoustic data, as shown in the following expression:

$$W = \max_{W} P(W|A) \tag{1}$$

Since it is not easy to calculate $P(W|A)$, the Bayes rule is used to ease the computation leading to the expression:

$$P(W|A) = \frac{P(A|W)P(W)}{P(A)} \tag{2}$$

Note that in this expression, the denominator is not necessary to compute $P(W|A)$ since it is independent of the word sequence W. Thus, Eq. (1) can be rewritten as follows

$$W = \max_{W} P(A|W)P(W) \tag{3}$$

which is the fundamental equation in the stochastic approach to the ASR problem. In this expression $P(A|W)$ is called