UNDERSTANDING SPONTANEOUS SPEECH

Enrico Bocchieri Esther Levin Roberto Pieraccini
Giuseppe Riccardi
Speech Research Department
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974, USA

1 Speech Understanding in the USA: the ATIS project

In 1989 ARPA (Advanced Research Projects Agency) started a new project aimed at developing a large vocabulary spontaneous speech understanding system. The project was called ATIS (Air Travel Information System) and the task was that of answering questions (in context) for retrieving information contained in a relational database consisting of a subset of the Official Airline Guide [2]. Several research centers (AT&T, BBN, CMU, MIT, SRI, NIST, MITRE, UNISYS, etc.) joined the project. Also, some of the sites teamed up for collecting a relatively large corpus [2] of spontaneous speech. The corpus was collected through a Wizard of Oz paradigm. Each subject was given a scenario and a travel planning problem to solve. The subjects were requested to solve the problem by interacting with a machine (all the transactions were actually supervised by a human *wizard*). The partial and the final responses of the machine were presented to the subjects via a display or a speech synthesizer. The sentences uttered by

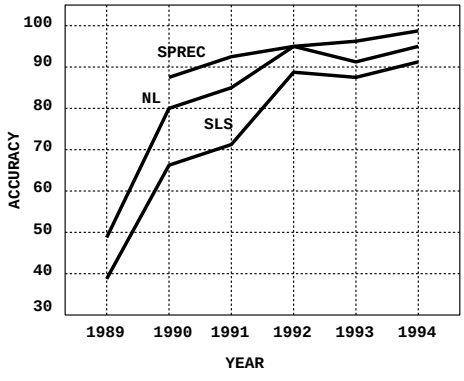

Figure 1: Highest accuracies in the ATIS tests.

the subjects were recorded, transcribed and annotated carefully. The ATIS corpus consists today of more that 20,000 sentences, including speech waveforms, transcriptions into words, sentence categories (e.g. context dependent, independent, or unanswerable), and *reference answers*, namely the set of tuples from the database that make a reasonable answer to the question posed by each sentence. The decision on whether a set of tuples makes a reasonable answer for a given question was carried out according to a document called *Principle of Interpretation*[2]. The Principle of Interpretation document was written and updated by a special committee within the ATIS project and used both by the corpus annotators and by the system developers. Roughly every year, since 1989, official tests were made to evaluate the performance of the systems in each lab. Three standard tests (*SPREC, NL,* and *SLS*) tests were generally performed at each evaluation. SPREC measured the accuracy of the speech recognizer as the overall percentage of correct words, NL measured the overall percentage of correctly answered sentences (i.e. those for which the given answer matched the reference answer) of the natural language component alone (i.e. the system input was the textual transcription of each answerable sentence), and SLS measured the overall percentage of correctly answered sentences for the complete systems (i.e. the system input was the actual speech). Fig 1 shows the highest accuracy reported among all participating sites in the three standard tests from the beginning of the project until today. The small drop in accuracy for the NL and SLS tests in 1993 was due to a change in the complexity of the task (the size of the relational database was increased considerably).

2 CHRONUS, the AT&T speech understanding system

CHRONUS, the AT&T speech understanding system, reported among the top scores in the three tests of 1994 ATIS evaluation. This version of CHRONUS was developed in less than one year by the authors of this paper based on the following principles:

- *Learnability.* Everything that can be learned from available data should. The corpus in an invaluable

source of knowledge that can be used for training stochastic models used at all levels of processing.

- *Separation of knowledge and algorithms,* allows for an easy incremental improvement of the system.

- *Separation of general and task specific knowledge,* eases the process of reuse and portability of the system.

- *Locality.* The various modules of the system uses local algorithms (i.e. that operate on independent portions of the sentence) that are generally easy to understand, modify, update, and improve. The non local linguistic phenomena are demanded to a single module (i.e. the *interpreter*) at a later stage of processing.

- *Habitability.* The system does not attempt to model complex linguistic events that are extremely rare in spontaneous speech. Rather it is extremely robust to *non linguistic* phenomena frequently occurring in spontaneous speech (e.g. ungrammaticalities, false starts, user and/or speech recognizer mistakes).

The functional diagram of CHRONUS is shown in Fig 2. In the rest of this paper we will be giving a short description of each module and a summary of the performance.

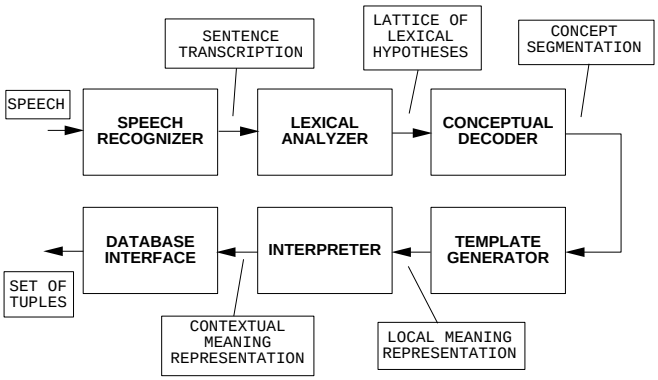

Figure 2: A functional diagram of CHRONUS, the AT&T speech understanding system

2.1 Spontaneous Speech Recognition

In general, the first operations performed by almost any recognizer are to sample the input speech (we used a 16kHz rate), and to extract from the sampled waveform a sequence of feature vectors. These are the acoustic data on the basis of which the recognizer decides which words are spoken. In our system the feature vectors are computed every 10 ms, and they have 39 components, namely twelve mel-frequency cepstrum coefficients and speech energy, with their 1^{st} and 2^{nd} derivatives [5].

Sentence recognition is the process of finding the most likely sequence of words conditioned by the available acoustic data. The conditioned word sequence probability is defined in terms of stochastic models, namely the phonetic acoustic models (which provide probabilities that a given feature vector sequence corresponds to the phonemes), the lexicon (that defines words in terms of phonemes), and the language model (it assigns a probability value to any word sequence). The recognition accuracy depends on the adopted models, and in particular we addressed two crucial issues concerning *spontaneous* speech:

1. **Speech model estimation problem.** Even the large speech data bases used for model training cannot *explicitly* represent all the characteristics of spontaneous speech, that includes disfluencies, false starts, breath noise, pauses, ungrammatical sentences, unexpected phrases, and an open and large vocabulary. Therefore, for state of the art recognition accuracy, it is necessary to improve the robustness of statistical estimates of events that are sparsely observed in the training data. Furthermore, it is important to use techniques to *generalize* the training data for the estimation of statistics of *unseen* events, that do not explicitly appear in the training data.

2. **Computational complexity** is also an important issue. For quick development of new ideas, we have designed a unified representation (by non deterministic stochastic finite state automata [7]) for n-gram language models of different orders. Therefore, we can test different types of models without changing the recognizer architecture.

Acoustic Models

The acoustic models consist of triphone (i.e. phoneme in left and right context) Hidden Markov Models (HMM) [5] with continuous state likelihood functions defined by Gaussian mixtures.

To make better use of the available training data, the state distributions of the HMM's are tied [6], i.e. corresponding states of different triphone models of the same base phone may share the same state likelihood function. With the available training data, without state tying, we can reliably estimate only the HMM's of a small number of the "most frequent" triphones (i.e. less than 2,000). Thanks to state tying we can now reliably estimate the HMM's of those triphones that appear only once in the training data. In the current system we defined 19,000 triphone models, based on 3,500 different likelihood functions.

For improved accuracy, the recognizer uses two sets of gender dependent (male and female) HMM's. However only half of the total amount of training data are available for the estimation of gender-dependent HMM's, which are therefore less robust than gender-independent

models. Moreover we interpolate the state likelihoods of the gender-dependent triphone HMM's with the corresponding state likelihoods of the gender-independent HMM's. Automatic speaker-gender classification is performed prior to recognition. Then the corresponding interpolated gender dependent triphone HMM's are used. The acoustic models were estimated using 20,000 ATIS training utterances with an additional 8,000 utterances from a different corpus. The lexicon consists of 1,530 words, with one pronunciation per word defined automatically by the AT&T text to speech system.

Language Model.

In the current speech recognizer we used a new language model representation based on non-deterministic stochastic finite state automata, including model estimation procedures [7]. These automata provide a unified representation for n-gram language models of different orders. They also implement an accurate and extremely efficient approximation of the back-off probabilities [1]. This representation allows both for a recognition algorithm independent of the language model order and type, and an efficient implementation.

Word classes are used to increase the robustness of the language model estimation. They correspond to sets of semantically similar words (e.g. city names, days of the week, etc..). Word classes allow for more robust probability estimates, because different words in the same class share the available training data. We also explicitly model *compound words*, that are word sequences that should be treated as single units (e.g. "I'D LIKE TO", "HOW MUCH", "THANK YOU", etc.). This implicitly lengthens the word *history* associated with the n-gram formulation.

The trigram model of word classes with compound words was estimated from a total of 20,000 ATIS sentences.

2.2 Natural Language Processing

Lexical Analysis

The function of the lexical analyzer is that of creating a lattice of word hypotheses out of a string of words (either from text input or recognizer output). The different hypotheses in the lattice correspond to possible interpretations of words and/or phrases according to predefined categories. Each category represents a set of words or phrases that are considered semantically equivalent by the successive conceptual representation. In general plural and singular forms of a noun constitute a category, as well as different verb inflections. This reduces the effective size of the lexicon enhancing the robustness of the stochastic conceptual representation [3], and the robustness against recognition errors (speech recognizers are likely to confuse among different inflections of the same word). Other categories correspond to items in the application, like numbers, city names, airport names, acronyms, etc.

Conceptual Decoding

Rather than attempting to parse sentences as a whole, the conceptual decoder detect phrases that are associated to units of meaning called *concepts*. The set of concepts is application dependent For instance, the set of concepts for the ATIS task roughly corresponds to the set of attributes of the relational database and includes concepts like `destination`, `origin`, `ground-transportation`, `aircraft-type`, `meal`, `departure-time`, `arrival-time`, etc. The output of the conceptual decoder is a segmentation of the original sentence into phrases labeled with the corresponding conceptual unit. The algorithm is based on a stochastic semantic model (conceptual HMM) first proposed in [3] that corresponds to HMM whose states are related to the concepts, and are described by *concept conditional* stochastic language models. In the current implementation both the semantic model and the concept conditional language models are bigram back-off networks [7], estimated automatically with a *training loop* procedure [4] from more than 7,000 sentences.

Template Generation and Interpretation

The template generator produces a representation of the sentence meaning in the form of a *template*, i.e. an unordered set of keyword/value pairs [4]. Each segment in the conceptual segmentation can produce one or more keyword/value pairs. The process of translating a segment into a keyword/value pair is performed by programmable finite state machines [9].

The function of the interpreter is to resolve ambiguities and to correlate non local information in the templates, both from other portions of the same sentence and from previous sentences. To deal with context dependent sentences the interpreter merges the template of the previous sentence in the same session with the current template. Merging rules are defined at the level of every single keyword/value pair. This module is implemented as a set of handwritten rules. A corpus and an evaluation procedure are very valuable for the development of the interpreter, since they allow to set up an automatic way of measuring the effectiveness and the consistency of each newly introduced rule. In the development of the interpreter for ATIS, every time new rules were introduce, the system was evaluated on more than 7,000 sentences for finding possible inconsistencies.

The Relational Database Interface

The relational database interface takes the meaning representation in the form of a template and extracts the requested information from the database. The conventional way of doing this consists in writing a set of tran-

scription rules that transform the template into an SQL statement. The complexity of this approach is high, and the resulting code is hard to maintain and to port to other applications because it depends on the structure of the database. In our approach the relational database is represented by a network that can be easily derived from the database schema. The network is navigated through a *constraint propagation algorithm* that extracts the proper tuples.

3 Results and conclusions

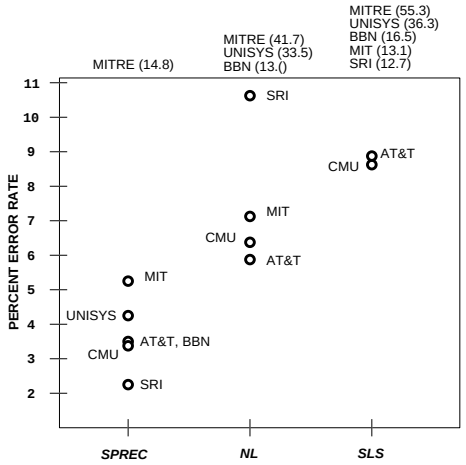

Figure 3: Error rates reported by different sites in the 1994 ATIS test

Fig. 3 shows the error rate in the three standard ATIS tests reported by different sites in the 1994 evaluation. The AT&T system obtained 3.5% of error in the speech recognition test (SPREC), 5.9% in the natural language test (NL) and 8.9% in the test of the complete system (SLS) [8, 9]. It was the best system in the NL test. An analysis of the errors showed that most of them can be attributed to the limitations of its flat, local, and syntax-free semantic representation that does not model complex relationships between subsets of constraints, other errors are due to the interpretation module that cannot handle long range contextual information, and of course general imperfectness of the system (and its developers). Overall, the AT&T system obtained better results than all the other more conventional systems that although able to handle more complex linguistic phenomena are probably harder to train and update.

References

[1] Katz, S. M., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. on Acoustics, Speech and Signal Proc.,* Vol. ASSP-35, pp 400-401, March 1987.

[2] *MADCOW*, "Multi-Site Data Collection for a Spoken Language Corpus,", *Proc. of Fifth Darpa Workshop on Speech and Natural Language*, Harriman, NY, Feb 1992.

[3] Pieraccini, R., Levin, E., "Stochastic Representation of Semantic Structure for Speech Understanding," *Speech Communication*, Vol.11 pp. 283-288, 1992.

[4] Pieraccini, R., Levin, E., "A learning approach to natural language understanding," NATO-ASI, *New Advances & Trends in Speech Recognition and Coding,* Springer-Verlag, Bubion (Granada), Spain, 1993.

[5] Rabiner L. R., Juang, B.-H., "Fundamentals of Speech Recognition," *PTR Prentice-Hall, Inc.*, 1993.

[6] Young, S.J., Woodland, P.C. "The Use of State Tying in Continuous Speech Recognition," Proc. Eurospeech-93, pp 2203-2206.

[7] Riccardi, G., Bocchieri, E., Pieraccini, R., "Non Deterministic Stochastic Language Models for Speech Recognition," *ICASSP 95.*

[8] Bocchieri, E.L., Riccardi, G. Anantharaman, J., "The 1994 AT&T ATIS CHRONUS Recognizer," Proc. of 1995 ARPA Spoken Language Systems Technology Workshop, Austin Texas, Jan. 1995.

[9] Levin, E., Pieraccini, R. "CHRONUS, The Next Generation," Proc. of 1995 ARPA Spoken Language Systems Technology Workshop, Austin Texas, Jan. 1995.