The Third International Conference "Problems of Cybernetics and Informatics"
September 6-8, 2010, Baku, Azerbaijan. Section #2 "Intellectual Technology and Systems"
www.pci2010.science.az/2/20.pdf

INTERACTIVE SPEECH UNDERSTANDING

Samir Rustamov[1], Elshan Mustafayev[2]

Cybernetics Institute of ANAS, Baku, Azerbaijan
[1]*samir.rustamov@gmail.com*, [2]*elshan.mustafayev@gmail.com*

Speech recognition stands as a medium, or a communications Ambassador between machines and people, ever promising to deliver "natural speech." As a result, Speech Recognition has the ability of unifying or incorporating many other current technologies, ultimately fusing many features and functions granted by today's technologies.

In the past few years, Speech Recognition has experienced many improvements that have enabled its machines, specifically computers, to perform elaborate tasks such as Dictation, and Command Recognition. Thanks to superior ways to capture and analyze sound files, even personal accents and speech impediments can now be taken into account. As Speech Recognition improves and advances, it is inevitably poised to directly compete with an old technology warrior, the keyboard.

Nonetheless, the keyboard and its means to communicate, typing, are not natural ways for people to convey information and ideas. The primarily vehicle for people to communicate is in fact language, or better said, speech itself. Therefore, the very moment that machines and humans can effectively communicate naturally, the keyboard will inevitable find a secondary role as a communications Ambassador. The promise of naturally speaking and naturally listening machines is an ever growing and exciting reality.

In view of the possible misinterpretations that natural language bestows, scientific fields such as Natural Language Processing (NLP) and Artificial Intelligence (AI) are increasingly being incorporated into Speech Recognition in hopes of overcoming this language paradox. NLP and AI sciences are simultaneously growing, evolving and experiencing many changes and improvements.

One of the most obvious and natural applications for speech technology is in providing a gateway to information services over the telephone network. Already a significant growth is occurring the provision of information from a centralized computing system using stored messages or synthetic speech derived from text files.

Most of these systems however currently rely on the use of touch tone input for selection of the information. The ability to recognize a small number of words or even the digits without the user requiring to train the system therefore has widespread application. As the scope of the information service expands so also does the need for more intelligent dialogues with much larger vocabularies for speech understanding.

Our computer systems will enable users to maintain telephone conversations about specific topics such as flight arrivals, schedules and reservations. The systems are planned to support a speaker independent vocabulary of up to 100 - 200 words in Azeri language.

The systems have been defined as computer systems with which humans interact on a turn-by-turn basis are called **spoken dialogue systems.** The main purpose of a spoken dialogue system is to provide an interface between a user and a computer-based application.

A common architecture of the system has 4 modules: Speech pattern processing, Linguistic processing, Dialogue management, Oral message generation.

Speech Pattern processing. This module carries out the acoustic-phonetic decoding of the incoming speech signal and produces a lattice or graph of word hypotheses. Techniques for handling the variable quality of telephone speech, mostly due to the use of different handsets and varying handset positions between calls, are also included in this module.

Particular objectives in the speech pattern processing module include:

- refinement of the acoustic/phonetic units used, in line with requirements for easy extension of the vocabulary, rapid speaker adaptation and handling fluent speech (the co articulation problem).

- improvement of speaker-independent phonetic rules and classification techniques.
- improvements in speech recognition over the telephone.
- fast lexical access.
- improvements in speech pattern processing techniques based on Artificial Neural Network.

The main part of this module is speech recognition of the uttered speech. The computing algorithms of speech features being the main part of speech recognition system are analyzed in the paper. From this point of view the determination algorithms of Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) coefficients expressing the basic speech features are developed. Combined use of cepstrals of MFCC and LPC in speech recognition system is suggested to improve the reliability of speech recognition system [1]. To this end the recognition system is divided into MFCC and LPC-based recognition subsystems. The training and recognition processes are realized in both subsystems separately, and recognition system gets the decision being the same results of each subsystems. This results in decrease of error rate during recognition.

The recognition process. One of main requirements in speech recognition system is reliability of recognition. To improve reliability of the system is offered combine using different structured or different features systems. This recognition systems can be work independently in one system and we called them conditionally subsystems of the main system.

Speech signal is trained by different mathematical models in each subsystem separately. The recognition results of the subsystems passed to decision making block in during recognition

As the subsystems of the our system are taken the neural networks trained from the different initial points. The results of the subsystems are compared in the decision making block and ASUS accepts this decision.

Linguistic processing. The major objectives of the linguistic processing module are:

- the modification of linguistic processing algorithms to meet the characteristics of spoken language such as hesitations and ill-formed or incomplete utterances. This may require, for example, the partial analysis of complete utterances;
- the development of analysis algorithms to handle indeterminacy within the linguistic analysis stage;
- the study and modelling of the spoken language sub-set used in the selected applications;
- exchange and evaluation of different parsing algorithms;
- development and evaluation of stochastic grammars in relation to rule based systems, with particular reference to computational efficiency, coverage, performance and extensibility;
- investigation and comparison of ways of integrating linguistic knowledge with pattern processing and with dialogue management;
- use of predictions from the linguistic processing module to improve hypotheses at the front end;

Dialogue management. Human-human dialogue is capable of robust handling of adverse conditions and recovery from communication failure. In human-machine communication which is speech based, it becomes necessary to repair recognition failure. Knowledge from a variety of sources: syntactic, semantic and pragmatic, and knowledge of the application domain, may be brought to bear both in understanding and recovery [3].

A major goal of the article is the development of an intelligent co-operative dialogue system. Most of the work done to date in the area of dialogue management has been in text based systems. Whilst speech-based systems can learn much from this work, particularly in the areas of architectures and the use of dialogue histories, oral dialogue contains specific phenomena not seen in text, such as hesitations and false starts. Knowledge is also used differently, for instance pragmatic knowledge is not only used to solve ambiguities but also to strengthen or weaken hypotheses generated from potentially error-prone acoustic-phonetic decoding. To progress in the modelling of oral dialogue, extensive use is being made of simulations and real applications in the selected information service domains such as flight enquiries, intercity train times and hotel information. A substantial amount of analysis is already

complete on a large corpus of spoken dialogues for the Azeri language. This analysis is providing information on:

- user requirements;
- oral dialogue strategies; trade-offs between user and system initiative in terms of user acceptability and constraining the domain (necessary to avoid over ambitious questions), confirmation and repair strategies;
- oral language requirements (grammar, lexicon, specific phenomena of spoken dialogue);
- the complexity of the semantic space of the domains chosen;
- generic rules for dialogue for information services domains where these can be abstracted from particular applications, including cross-language differences in dialogue strategies.

Oral message generation. Whilst research on natural language generation is more recent than natural language parsing and understanding, computer based language generators are rapidly expanding in response to growing needs for intelligent human-machine interaction. The emphasis hitherto has again been on written language.

The particular requirements of oral output in the course of a dialogue (intelligibility, reduced length, enumeration and requests for repetition etc) is being taken into account in the message generation component. Prosody (stress on words and melodic contour of an utterance) is affected by both linguistic and pragmatic constraints. A suitable symbolic description of the effects of these constraints on a particular utterance is passed, along with the text, to the Text to Speech Synthesis (TTS) system [4].

Speech synthesis. The two fundamental processes are performed by all TTS systems: text analysis and speech synthesis. The text analysis must determine from the input text following features:

1. Pronunciation of the text string: the text analysis process must decide on the set of phonemes, the degree of stress in speaking, the intonation of the speech, and the duration of each of the sounds in the utterance;
2. Syntactic structure of the sentence to be spoken: the text analysis process must determine where to place pauses, what rate of speaking is most appropriate for the text and how much emphasis should be given to individual words and phrases within the speech;
3. Semantic focus and ambiguity resolution: the text analysis process must resolve homographs and also must use rules to determine word etymology to decide on how best to pronounce foreign words and phrases.

The input data for the analysis is Azerbaijan text. The first stage of processing does text processing operations, including detecting the structure of the document containing the text, normalizing the text and performing a linguistic analysis. The text processing benefits from an online dictionary of word pronunciations along with rules for determining word etymology. The output of the basic text processing step is tagged text, where the tags denote the linguistic properties of the words of the input text string.

One other text normalization problem concerns the pronunciation of proper names of foreign languages.

The third step in the basic text processing block is a linguistic analysis of the input text, with the goal of determining the following linguistic properties [5]:

- The part of speech of the word
- The sense in which each word is used in the current context
- The location where a pause in speaking might be appropriate
- The word (or words) on which emphasis are to be placed, for prominence in the sentence
- The style of speaking, e.g., irate, emotional, relaxed, etc.

Ultimately, the tagged text obtained from the basic text processing block of a TTS system has to be converted to a sequence of tagged phones. The phonetic analysis block enables the TTS system to perform this conversion, with the help of a pronunciation dictionary. That is why the following operations are performed.

The homograph disambiguation operation must resolve the correct pronunciation of each word in the input string that has more than one pronunciation.

The second step of phonetic analysis is the process of grapheme-to-phoneme conversion, namely conversion from the text to speech sounds. Although there are a variety of ways of performing this analysis, the most straightforward method is to rely on a standard pronunciation dictionary, along with a set of letter-to-sound rules for words outside the dictionary.

Each individual word in the text string is searched separately. If the word exists, in its entirety, in the word dictionary, the conversion to sounds is straightforward and the dictionary search begins on the next word. If not the word is separate to the "root form" and affixes and the base search attempts to find both of them. If the "root form" or affixes are not present in the dictionary, a set of letter-to-sound rules is used to determine the best pronunciation of the root form or affixes of the word, again followed by reattachment of stripped out affixes. On the paper the edit distance method is applied for finding the roots of the words.

References

1. K.R. Aida-zade, C. Ardil and S.S. Rustamov. (2006). Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems. IJSP . International Journal of Signal Processing. Volume 3, pp. 105-111. http://www.enformatika.org/ijsp/v3/v3-2-14.pdf
2. K.R. Ayda-zade, S.S. Rustamov. (2005). Research of Cepstral Coefficients for Azerbaijan speech recognition system. Transactions of Azerbaijan National Academy of sciences. Informatics and control problems . Volume XXV, №3, pp. 89-94.
3. Jihyun Eun, Changki Lee, Gary Geunbae Lee. (2004). An Information Extraction Approach for Spoken Language Understanding.
   http://isoft.postech.ac.kr/publication/iconf/icslp04_eun.pdf
4. Ye-Yi Wang, Li Deng, and Alex Acero. (2005). An Introduction to Statistical Spoken Language Understanding. IEEE Signal Processing Magazine, vol. 22, no. 5, pp. 16-31.
   http://research.microsoft.com/pubs/75236/2005-Wang-Deng-Acero-SPM.pdf
5. Anton Batliner, Bernd Mobius, Gregor Mohler, Antje Schweitzer, Elmar Noth.(2001). Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground. Eurospeech 2001 – Scandinavia.
   http://www.smartkom.org/prosodic_models.pdf
6. C.H. Lee, B.H. Juang, F.K. Soong, L.R. Rabiner. (1989). Word recognition using whole word and subword models. ICASSP-89., International Conference on Volume. Acoustics, Speech, and Signal Processing. Volume 1, pp. 683-686.