ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 2011. No. 4(110)
ELEKTRONIKA IR ELEKTROTECHNIKA

T 121 SIGNAL TECHNOLOGY
SIGNALŲ TECHNOLOGIJA

Voice-based Human-Machine Interaction Modeling for Automated Information Services

R. Maskeliunas, K. Ratkevicius, V. Rudzionis
Speech Research Laboratory, Kaunas University of Technology
Studentų str. 65, LT-51369, Kaunas, Lithuania; phone: +370 37 354191, e-mail: rytis.maskeliunas@ktu.lt

Introduction

The main aim of telecommunications is to bring people thousands miles apart, anytime, anywhere together to communicate as if they were having a face-to-face conversation in a ubiquitous tele-presence way. One key component necessary to reach this main aim is the technology enabling usual communication by voice. This means the use of automatic speech recognition [1]. An IVR (Interactive Voice Response) based systems can be used to automate a wide range of services and data requests. These systems are used most often by the companies to provide the self-service abilities to customer. The system takes the input from the user and provides back the enterprise information in the form of recorded or synthesized voice, fax or even an email by connecting one or more online databases to the caller. Although there are several hundred million Internet-connected PCs in the world, this figure is dwarfed by the two billion fixed and mobile phones. The telephone is ubiquitous, increasingly mobile and could, in principle, provide a universal platform for accessing on-line services. To date efforts to harness this potential in the form of IVR systems have not proved especially popular with users. There's a wind of change blowing through the IVR world, impelled by advances in speech recognition technology and a transformation of the IVR programming environment [2].

The limitations of IVR approach could be easily observed: if the number of possible choices grows the IVR system soon becomes difficult to operate and navigate. And speech recognition could be the most appropriate and convenient solution. When implementing IVR systems using mobile devices the advantages of speech recognition based interfaces becomes even more evident. Mobile devices possesses small keyboards and screens what makes the use of traditional GUI (Graphic User Interface) interfaces even less appropriate for the human-machine interaction.

Very important characteristic of voice based interfaces is the dependability of the phonetic, syntactic and lexical properties of the language spoken by the user. This means that it is impossible to move technologies developed for the recognition of one language for the recognition of another automatically. Some sort of adaptation would be necessary. Since major developers of speech technologies aren't particularly interested in less spoken languages such as Lithuanian the need for adaptation in such cases is even more important. One of the possible solutions for some class of applications is the adaptation of foreign language based speech engines via the selection of proper phonetic transcriptions. In our previous studies the advantages of such method and its possible uses were established [3, 4].

Voice based HMI for automated information services

Speech recognition requires a very different approach from touch-tone to the telephone interface. The motivators and the technology behind the shift from traditional touch-tone input to speech recognition in telecommunications are described in [5].

Smart mobile devices support web technologies, audio and video playback and integrated extra features beyond telephony. Multimodal user interfaces are favorable in mobile environments, when the traditional standard modality (touch and keyboard) is supported by other modalities, primarily by speech input and output. Multimodality increases the usability of applications and makes them accessible for disabled people, although creating multimodal interfaces on mobile devices is a challenging task [6].

There are many industrial design samples. For example the IBM hotel information system was designed to make information, stored on Internet or enterprise backend servers, available in a natural dialog over the telephone. The system combines robust speech recognition with natural language and dialog components to allow the user to request the information in a human-human-like dialog [7].

In [8] a multi-agent production planning system designed for small and medium-sized enterprises with project-oriented production is presented. In order to make

the results of the system available even to users who are located away from the enterprise, it has been equipped with the possibility of remote access-a Web and telephony interface.

The development of a fully automatic multimodal information system for the consumer market is presented in [9]. The system will be able to provide information on a large number of topics via a single telephone number. The eventual system will integrate interactive voice response, speech recognition, speaker verification, direct dial in, calling line identification, facsimile and electronic mail.

A multi-modal interaction framework using speech recognition and computer vision to model a new generation of interfaces in the residential environment was developed in [10]. The design is based on the use of simple visual clues and speech interaction. The latter system incorporates video information processing block which moves this system to the class of multimodal systems. Further we will propose a dialog model for the Lithuanian spoken language based human-machine interaction system.

Human – machine dialog modeling

The dialog model was developed for the evaluation of a Lithuanian telephony interface targeting the use of various call centers: the chosen dialog model is capable of recognizing keyword phrases from the natural sounding sentences, while supporting additional modalities of OTA menus of choices, controlled by touch and keyboard (depending on a type of application and device used) allowing a user to enter the data using the means he proffers. The system was developed for the commercial GSM based IVR system. The system was adapted to a built-in processing server recognizers (due to security, licensing and compatibility reasons) based on the principles of foreign ASR engine adaptation to a Lithuanian language [3, 4]. In this case the traditional Spanish (SP-SP) recognizer was chosen for the base processing due to linguistic similarities and standard availability in server system used. The application framework was programmed to mimic the standard interface that Lithuanian medics use to enter and submit sick-list data of their patients to the Social security foundation of Lithuania. Looking at the general requirements to the content of such services we may observe that the essential component of them is the necessity to identify the person (the user, the provider, the patient, etc.). One of the well-known approaches is biometric approach used in [11]. But this approach requires user voice samples and it is inappropriate in many services were we can't have them in advance. In many cases it is possible to organize pronunciation of the personal identification number using spelling of the digit numbers. In this situation we obtain limited set of ten digit names that could be used to carry out core voice information for the public service. Using this approach we also could implement error – correction codes in some situations to obtain false recognitions and to use necessary measures to avoid misrecognition (e.g., to ask the user to repeat once more one digit instead of asking to repeat whole string of digit names). The algorithm of a dialog model capable of recognizing keywords out of the natural sentences is presented in Fig. 1.

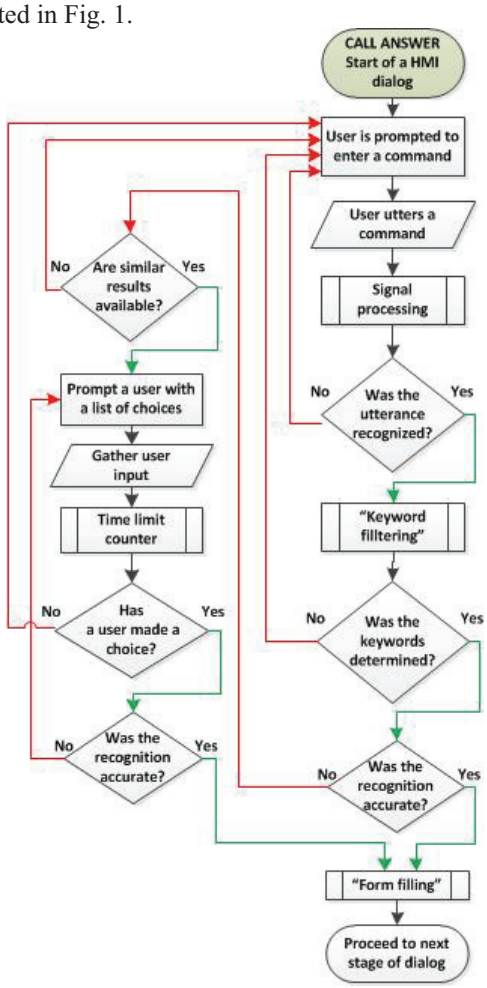

Fig. 1. The algorithm of an HMI dialog capable of recognizing keywords from the natural sounding sentences

The HMI dialog starts when an IVR system pickups a call. At the beginning of the dialog the user is prompted to enter a command (either by simple voice commands, or by the traditional means). After the person utters a command, the input signal is processed and the word is checked against the recognition vocabulary if such a command is possible. If so – the confidence value of the recognized phrase is measured and if it is high enough the semantic value is used in further processing. In case of an unclear recognition (system sees a few choices as similar) an n-best strategy might be used and a user might be offered to choose between the similar commands (the most similar results - i.e. "Did you say: Toma or Foma?"). After that the semantic value is processed and the application proceeds to the next stage of a dialog.

A system is preprogrammed to use a specific set of complex grammar rules, allowing keyword (the important words with a specific semantic value) spotting. This way a user can speak naturally (for example: "The FIRST number of my passport is FIVE") and a system only catches the important words (in this case "FIRST" and "FIVE"), assigns the appropriate semantic values and passes for further processing and finally jumps to a next stage in dialog. The biggest advantage of this approach

over the isolated words is the additional naturalness maintaining high enough recognition accuracy.

Experimental evaluation of speech recognition accuracy

The crucial question for the success of voice commands recognition based interfaces is the recognition accuracy. In our previous experiments we showed that proper selection of phonetic transcriptions enables to achieve high enough recognition accuracy of Lithuanian voice commands using foreign language speech engine. We also showed that selecting proper optimization procedure for the selection of phonetic transcriptions may lead to the significant improvement of the recognition accuracy. Another way to increase the recognition accuracy is to select better foreign language speech engine for the recognition of Lithuanian voice commands. Major providers of speech recognition engines provide about 20 speech engines designed to recognize some widely spoken languages. Besides of English engine it could be found engines for the recognition of French, German, Spanish, Italian and other languages. It could be easily hypothesized that not all of those engines could be of equal value when recognizing Lithuanian voice commands and that some of the engines developed for the recognition of foreign languages may lead to the better results than others. The primary reason of this lies in the phonetic structure of different languages: the sound systems of one language is more similar to Lithuanian speech than sound system of another language. From the other point of view there are no two different languages with identical phonemic and particularly allophonic structure. It could be felt intuitively that Spanish phonetics is more similar to Lithuanian phonetics than English but at the same time there are no in Spanish such sound as *'sh'* which is quite popular in Lithuanian. So even this short observation shows that can not be established 1-1 match between Lithuanian and any foreign language.

The main aim of these experiments was to establish the limits of possibilities to improve the recognition accuracy of Lithuanian voice commands selecting more proper foreign language engine. English *(Microsoft Speech Recognizer 9.0 for Microsoft Speech Server (English-US))* and Spanish *(Microsoft Speech Recognizer 9.0 for Microsoft Speech Server (Spanish-US))* engines were selected for the comparison. Transcription selection optimization procedure was used as in our previous experiments [4]. The algorithm of an HMI dialog, shown in Fig. 1, was used. The voice commands vocabulary consisted from 10 Lithuanian digit names. 20 speakers participated in the evaluation each of them pronouncing the same voice command 20 times (4000 utterances in total). The recognition accuracy of each voice command using English and Spanish engines is shown in Fig. 2.

It could be seen easily that Spanish speech engine provided significantly better results than English speech engine. It is important that Spanish engine allowed avoid such recognition accuracy "holes" as with the commands 3 and 4. The overall recognition accuracy increased from 77% using English speech recognizer to the 97% using Spanish one.

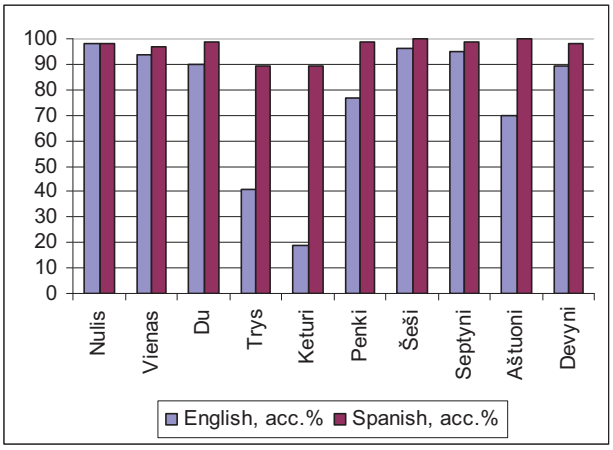

Fig. 2. The recognition accuracy of 10 Lithuanian voice commands using English and Spanish speech engines

Table 1 shows the recognition accuracy of voice command strings when Spanish or English recognizer was used. Three different experiments were carried on: in the first experiment string consisted from 4 commands, in the second from 8 while in the third from 11 commands randomly selected.

Table 1. Recognition accuracy of voice commands strings using English and Spanish engines

String size	4	8	11
English, acc. %	60.9	52.3	44.5
Spanish, acc %	90.4	87,7	80

It could be seen that advantages of Spanish engine could be seen even better when recognizing strings of voice commands (such as dictating digits in the personal identification code). Even long strings from 11 commands were recognized with rather high 80% accuracy while using English recognizer this accuracy felt to the in principle unacceptable 44% rate.

Conclusions

Voice-based human-machine interaction model for automated information services was proposed. This model allows recognize isolated commands together with some keywords. At the same model supports additional modalities such as OTA (Over the Air) menus of choices, controlled by touch and keyboard. Important characteristic of the model is the possibility to select a proposed choice (system of proposed selections).

Different foreign language speech engines have different capabilities to recognize Lithuanian voice commands. Spanish speech engine enabled to achieve significantly higher recognition accuracy than English engine: overall recognition accuracy increased from 77% for the English engine to the 97% for the Spanish engine.

Acknowledgements

This research was done under the grant by Lithuanian Academy of Sciences for the research project: "Dialogų modelių, valdomų lietuviškomis balso komandomis,

panaudojimo telefoninėse klientų aptarnavimo sistemose analizė" No.: 20100701-23.

References

1. **Juang B. H.** Ubiquitous speech communication interface // Automatic Speech Recognition and Understanding, ASRU '01, 2001. – P. 85–92.
2. **Dettmer R.** It's good to talk // Speech technology for on–line services access // IEE Review, 2003. – Vol. 49. – Iss. 6. – P. 30–33.
3. **Maskeliunas R., Rudzionis A., Rudzionis V.** Advances on the Use of the Foreign Language Recognizer // LCNS 5967, Development of Multimodal Interfaces: Active Listening and Synchrony. – Springer, 2010. – P. 217–224.
4. **Maskeliunas R., Rudzionis A., Ratkevicius K., Rudzionis V.** Investigation of Foreign Languages Models for Lithuanian Speech Recognition // Electronics and Electrical Engineering. – Kaunas : Technologija, 2009. – No. 3(91). – P. 37–42.
5. **Duerr, R.** Voice recognition in the telecommunications industry // ELECTRO '96, Professional Program, Proceedings, 1996. – P. 65–74.
6. **Toth B., Nemeth G.** Challenges of creating multimodal interfaces on mobile devices // ELMAR, 2007. – P. 171–174.
7. **Mast M., Gunther C., Kunzmann S., Ross T.** Multimodal output for a conversational telephony system // Multimedia and Expo (ICME 2000), 2000. – Vol. 1.– P. 293–296.
8. **Becvar P., Smidl L., Psutka J., Pechoucek M.** An Intelligent Telephony Interface of Multiagent Decision Support Systems Systems // Man, and Cybernetics, Part C, Applications and Reviews, IEEE Transactions, 2007. – Vol. 37. – Iss. 4. – P. 553–560.
9. **Damhuis M., Peeters M., Boves L.** A multimodal consumer information server with IVR menu // Interactive Voice Technology for Telecommunications Applications, Second IEEE Workshop, 1994. – P. 73–76.
10. **Macek T., Kleindienst J., Krchak J., Seredi L.** Multi–modal telephony services in hometalk // Intelligent Environments, 3rd IET International Conference, 2007. – P. 404–410.
11. **Šalna B., Kamarauskas J.** Evaluation of Effectiveness of Different Methods of Speaker Recognition // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 2(98). – P.67–70.

Received 2011 02 14

R. Maskeliunas, K. Ratkevicius, V. Rudzionis. Voice-based Human-Machine Interaction Modeling for Automated Information Services // Electronics and Electrical Engineering. – Kaunas: Technologija, 2011. – No. 4(110). – P. 109–112.

Voice based human-machine dialogs are becoming more and more important part of informative services. The implementation of voice dialogs enables to realize some of the aims of telecommunication services more successfully and efficiently. The main aim is to enable the communication according the principle "anytime-anywhere". The importance of voice dialogs is also caused by the fact that principle "anytime-anywhere" often could be realized only using mobile and portable devices. Those devices typically have small keyboards and screens and hence voice based interface has advantages over traditional keyboard and screen based interface. The paper presents the model of multimodal interface which core element is the recognition of voice commands. The model targets the informative services provided by the Lithuanian medical and social security enterprises. Paper shows that recognition accuracy of Lithuanian voice commands could be increased significantly if the foreign language which has closer to Lithuanian phonetic structure engine is adapted. Ill. 2, bibl. 11, tabl. 1 (in English; abstracts in English and Lithuanian).

R. Maskeliūnas, K. Ratkevičius, V. Rudžionis. Žmogaus ir mašinos balso dialogų modeliavimas automatinėms informacinėms paslaugoms // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2011. – Nr. 4(110). – P. 109–112.

Balso dialogai tampa vis svarbesne telekomunikacinių paslaugų dalimi. Jie leidžia sėkmingiau ir efektyviau atlikti daug pagrindinių telekomunikacinėms paslaugoms keliamų užduočių, kurių svarbiausia yra sujungti ir leisti tarpusavyje bendrauti asmenims bet kurioje pasaulio vietoje bet kuriuo metu. Balso dialogai labai svarbūs dėl to, kad bendravimą bet kurioje vietoje ir bet kuriuo metu gali užtikrinti tiktai mobilūs įrenginiai. Tokie įrenginiai turi nedideles klaviatūras ir nedidelius ekranus, todėl balsinė sąsaja daugeliu atveju yra pranašesnė. Straipsnyje pasiūlytas multimodalios sąsajos, kurios pagrindinė moda yra balso komandos, modelis medicinos ir socialinio draudimo paslaugas teikiančioms Lietuvos įmonėms. Parodyta, kad esminis sąsajos sėkmingo naudojimo veiksnys yra balso komandų atpažinimo tikslumas. Parodyta, kad balso komandų atpažinimo tikslumą galima labai pagerinti adaptuojant užsienio kalbos, kurios fonetinė struktūra yra artimesnė lietuvių kalbai, atpažinimo variklį. Il. 2, bibl. 11, lent. 1 (anglų kalba; santraukos anglų ir lietuvių k.).