ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2012, Vol. 41, No. 2

HMI Modelling for Multimodal Lithuanian Applications

Rytis Maskeliūnas, Kastytis Ratkevičius

Speech Research Laboratory, Institute of Information Technology Development,
Kaunas University of Technology,
Studentų St. 65-108, 51369 Kaunas, Lithuania,
e-mail: rytis.maskeliunas@ktu.lt

crossref http://dx.doi.org/10.5755/j01.itc.41.2.909

Abstract. Spoken dialogue based human-machine interfaces (HMI) are becoming more and more widely integrated in computer applications. Speech allows doing some task easier and faster. The combination with a more traditional means of inputs and outputs – i.e. the multimodality factor becomes more and more important allowing wider accessibility. It is important to model and design spoken language dialog trees to imitate the natural conversations in the human-computer interactions, especially in information retrieval systems and applications. The paper presents three algorithms of HMI dialogs and the results of their experimental evaluation. The results showed that it is possible to achieve about 97% recognition accuracy in simple phrase based dialog conversations and about 93% in a very naturally sounding keyword spotting based dialogs.

Keywords: speech recognition; voice dialog modeling; human-machine interfaces.

1. Introduction

All computer interface systems are in principle designed to control an application and perform several tasks. Such systems might be viewed as an interface between the user and the computer (HCI) or human-machine interface (HMI) targeted at gathering the user input and translating the perceived data into specific tasks. The most popular example of the spoken language dialog systems for information retrieval is the typical call center applications (i.e. 118, 1528, etc.) that enable a database research on the basis of user requests. For example, the user may use a spoken dialog to perform certain tasks such as inquiring the information, inputting the necessary identification data, accessing help and so on. In a multimodal application, a user can access the information either by using traditional means (i.e. keyboards, touchscreens, etc.) or by speaking the voice commands (do this, I need that, etc.).

A very important characteristic of the spoken language 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 the 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 to Lithuanian language in such cases is even more important. The development of spoken dialog models built upon Lithuanian recognizers is a very important factor, necessary to develop a successful voice driven, multimodal HMI application.

2. HMI dialog modeling

Initially the HCI dialogs were modeled on the air traffic control application simulators [1]. Its role was to simulate the aircraft movements in an air sector. Almost parallel a more advanced study on Flight traffic information was done targeting HMI dialogs, considering spontaneous speech effects, including disfluencies, hesitations, repeated words and repairs [2]. Other systems implemented novelties such as grammar formalism, for example, L'ATIS for air traffic [3] MASK [4] and ARISE for train traffic [5] information retrieval.

Most modern approaches on dialog modeling are based on the use of Belief Networks [6], Bayesian networks [7]. Some dialogs are modeled by combining n-grams and stochastic context-free grammars [8], others - by implementing a stochastic approach [9].

The dialog model provides a general description of the different application related situations: request for information, repetition, confirmation, etc. It also specifies the relations between these situations. Four classic dialog modeling approaches are recommended for HCI modeling [10]:

- Structural models have their origins in linguistics. The most established LOQUI system enables database access for call center

employees [11], based on a hierarchy of language acts, that were divided into: requests, assertions and comments. A later model, STANDIA was aimed at developing an intelligent telephone switchboard and to process written and spoken language dialogs [12], targeted at identifying the user intentions in order to respond appropriately to his requests;

- Plan-oriented models are mainly based on an artificial intelligence and employ the notions of plan, planification and plan recognition. For example, the Litman model is based on three plan categories: the domain plans that model the application, the language acts that model the elementary communication actions and, finally, the discourse plans that model the relations between utterances and domain plans [13]. Another one, ATR, was developed as a human-machine spoken language dialog system which predicts user utterances in different languages for a conference registration application [14];
- Logic models use a modal logic to represent the mental attitude of the interlocutor and the reasoning induced by these attitudes. ARGOT is a classic system based on the language act theory including planning and user modeling [15]. TENDUM is a system , in which the dialog is based on language acts which are functions acting on the context [16];
- Task-oriented models are closely related to the application. The knowledge about the dialog is combined with the task knowledge. MINDS is a spoken language dialog system for accessing a stock management database, developed at CMU [17]. VODIS (Voice Operated Database Inquiry System) is a system for database access via the telephone [18]. SYDOR is a spoken language dialog system that is driven by the task between the user and an application back-end [19].

Dialogue tasks [20, 21] in a HCI dialog can be classified in the following way:

- Learning tasks: knowledge acquisition, where the user is subsumed under teaching or educational tasks;
- Information tasks: the user asks for information in a specific domain (i.e. air traffic schedules);
- Command tasks: the aim of the user is to handle objects in a reference world (i.e. control of a wheelchair);
- Assistance tasks: in certain applications, the user needs to be assisted in decision processes (i.e. translation).

In the classical architecture of the typical dialog system the user generates an utterance, which is then recognized by the speech recognition component. In the next step that data is processed by the semantic analyzer. Depending on the syntactic and semantic knowledge contained in the case grammar, the semantic representation of the user utterance is generated in the form of a network of frames, stored in the dialog context. On the basis of this network, the task and the dialog model, other processes in the dialog management module are activated to establish a dialog, to send a command to the application backend and to generate a feedback to the user.

The dialog models presented in this article were built upon the architecture of three of the above task classifiers: information (user is prompted what and how to enter something, guided how to, etc.), command (for example, "turn left") and assistance (user is assisted in case of silence, incorrect recognitions, etc.).

3. The proposed algorithms

Three dialog models were developed for the evaluation purposes: the dialog model capable of recognizing isolated words only, the dialog model capable of recognizing keyword phrases from the natural sounding sentences and the dialog model capable of recognizing the natural speech. All the presented dialog models support additional modalities of touch (menus of choices) and keyboard (depending on a type of application) allowing a user to enter the date using the means he prefers.

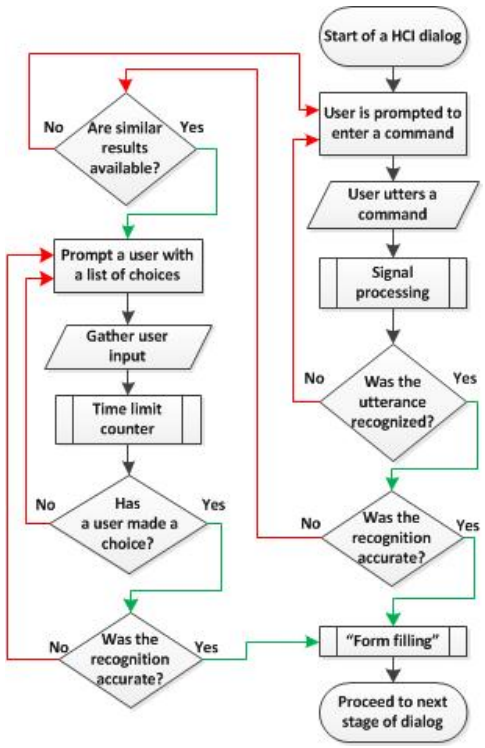

Figure 1. The algorithm of an HCI dialog capable of recognizing isolated words

The algorithm of a dialog model capable of recognizing isolated words is presented in Figure 1. At

the start of the dialog, a user is prompted (either by speech, graphically, or combined – depending on the type of application) to enter a command (either by simple voice commands or by the traditional means). After a 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 not to repeat the phrase, but to choose between the ones offered to him (the most similar results - i.e. "Did you say: Po vas or Ponas?"). After a successful gathering of the input (either way), the semantic value is processed and the application proceeds to the next stage of a dialog. The main advantage of this approach is the simplicity, as this implementation is based upon simple grammars (hopefully resulting in good recognition accuracy), but it is not a natural interface for the user.

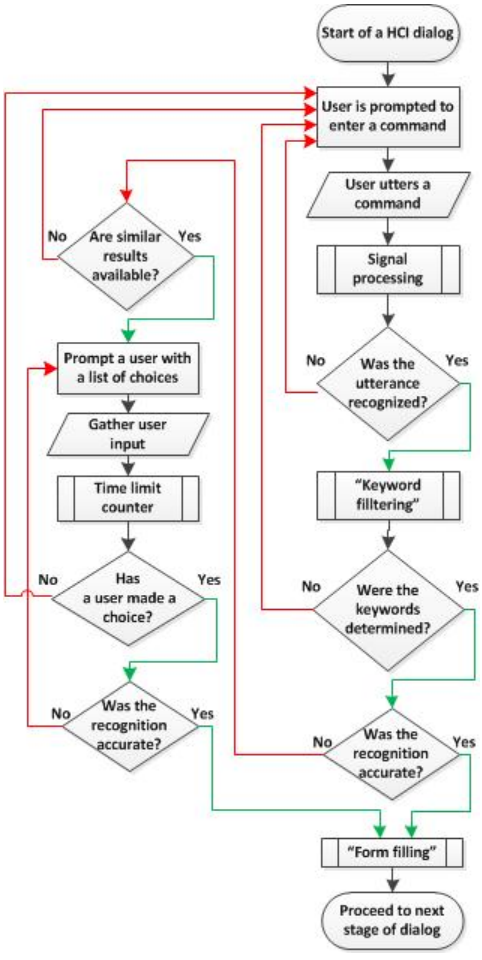

Figure 2. The algorithm of an HCI dialog capable of recognizing keywords from the natural sounding sentences

The algorithm of a dialog model capable of recognizing keywords out of the natural sentences is presented in Figure 2. The principle is quite similar, only this time 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. A prompt for self-correction is also possible in this case, and if available, a user is offered a list of selection (by voice or graphically). An error handling is done similarly as in the previous algorithm. The biggest advantage of this approach is the added naturalness, while still maintaining (hopefully) high enough recognition accuracy.

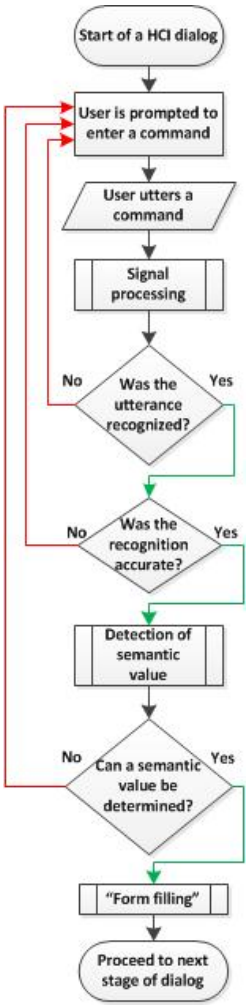

Figure 3. The algorithm of an HCI dialog capable of recognizing dictation

The algorithm of a dialog model capable of understanding dictation is presented in Figure 3. In this case – a system can understand a so-called dictation, where a user must speak a detailed operational instruction. The biggest disadvantage of such approach is a very complex set of grammar and a reduced accuracy of recognition. Another one – it is not possible to offer a self-correction list of choices, due to a very same reason – complex grammar rules.

This approach also uses a highest number of system resources and is the most sensitive to environmental factors (i.e. background noise, low quality transmission channel, etc.).

The advantages and disadvantages of each approach are analyzed in detail in the experimental evaluation section.

4. Technical realization

The whole HMI interaction model was realized as a server – client model. Depending on the type of application, a user might be offered a couple of interfaces: voice and traditional input only (target at standard telephones) and a more advanced web based interface with a graphical user interface (GUI) (target at smart-phone (currently not supported) and regular computer users). The whole application framework was programmed to mimic the standard interface that Lithuanian medical personnel uses to enter and submit sick-list data of their patients to the Social security foundation of Lithuania. A GUI version is a simplified copy of the same interface with added possibilities of multimodal input and feedback choices. A telephone version dialog goes through the same steps guided by voice (the user may enter data using his voice or his phone's keypad).

A simple illustration of system architecture (a popup window of a multimodal HCI software) is displayed in Figure 4. A user can enter the data using voice or by more traditional means (typing and clicking).

Figure 4. The view of a program implementing an HCI dialog capable of recognizing keywords

It is important to note that the recognition system is capable of recognizing a specific, preset set of complex rules of Lithuanian voice commands, phrases or dictation. 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 [22, 23]. 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 that we used (Microsoft Office Communications Server).

5. Experimental evaluation

5.1. Evaluation of recognition accuracy

Twenty speakers (equal number of males and females) took part in the experimental evaluation of HCI dialogs. Each speaker pronounced 10 phrases 11 times for each HCI dialog mode in Lithuanian. 11 phrases were used for dictation mode, because the additional phrase „The ELEVENTH number of patient's identification code is ONE" was used in this mode. The parameters of speech signal: sampling rate - 44.1 kHz, bit resolution – 16.

The phrases of "isolated words" mode varied from "ONE" to "NINE", the phrases of "dictation" mode – from "The FIRST number of patient's identification code is ONE" to - "The TENTH number of patient's identification code is ZERO". The phrases of "keywords" mode varied from the phrase of "dictation" mode to the phrase of "isolated words" mode (in Lithuanian):

- "The FIRST number of patient's identification code is ONE";
- "The SECOND number of identification code is TWO";
- "The THIRD number of code is THREE";
- "The FOURTH number is FOUR";
- "The FIFTH number - FIVE";
- "The patient's identification code is SIX";
- "The identification code is SEVEN";
- "The number is EIGHT";
- "The number – NINE";
- "ZERO".

Ordinal numbers were included in the first five phrases of "keywords" mode evaluation experiment. The averaged recognition accuracy of ordinal and cardinal numbers and the averaged confidence value are shown in Table 1.

Table 1. Accuracy and confidence value of Lithuanian digits recognition by Spanish recognizer in three modes

Dialog mode	Confidence value	Accuracy, %	
		Ordinal number	Cardinal number
Isolated words	**0.669**	-	93.9
Dictation	0.593	89.6	**96.9**
Keywords	0.474	84.9	93

Though the best recognition accuracy of Lithuanian digits was achieved in "dictation" mode, the mode "isolated words" outperforms it by the averaged confidence value, meaning that in theory this mode is more reliable. The worst recognition accuracy

results were obtained in "keywords" mode. These results were also confirmed by low confidence values.

The accuracy of Lithuanian digits recognition using a Spanish recognizer is shown in Figure 5.

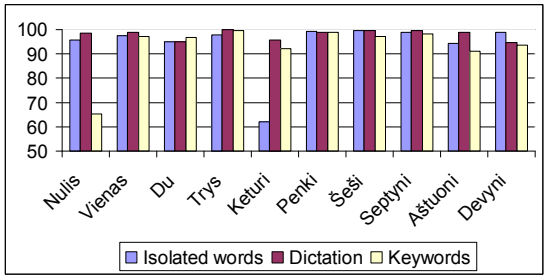

Figure 5. Accuracy of Lithuanian digits recognition by Spanish recognizer in three modes

The best recognition accuracy was achieved for the dictation dialogs. Long detailed sentences were recognized most accurately (96.9 %). Isolated words were recognized similarly to the keyword spotting (93.3 and 93 %).

The weak spots in the analyzed vocabulary were a few words. Obviously in the "keywords" mode the overall recognition accuracy was reduced by the bad recognition of the phrase "ZERO" – it is an unnatural phrase for this mode: the averaged recognition accuracy of nine digits (without "ZERO") is equal to 96.1 % in "keywords" mode. Another conspicuous result is a bad recognition of the phrase "FOUR" in "isolated words" mode. In order to find the reason of such degradation of these two digits recognition a more detailed analysis was made.

Table 2. Accuracy of Lithuanian digits "Four" and "Zero" recognition in three modes for males and females

Dialog mode	Four		Zero	
	Males	Females	Males	Females
Isolated words	87.3	**36.7**	96.9	94.5
Dictation	96.6	95.3	98.2	99.3
Keywords	89.7	94.9	89.5	**41.1**

From the results presented in Table 2, we may conclude that the bad recognition of the phrase "FOUR" in "isolated words" mode and the bad recognition of the phrase "ZERO" in "keywords" mode were determined by very low recognition accuracy of these digits pronounced by female speakers (respectively 36.7 % and 41.1 %).

The averaged overall deviation of the differences in speech recognition accuracy is shown in Figure 6. The results of the most accurately recognized female speaker (Female 1 – the averaged accuracy is 96%) and the worst accurately recognized female speaker (Female 2 – the averaged accuracy is 89.5 %) are presented. It is clear that more acoustical data should be used to train the recognizer (to develop better transcriptions) for the more poorly recognized words.

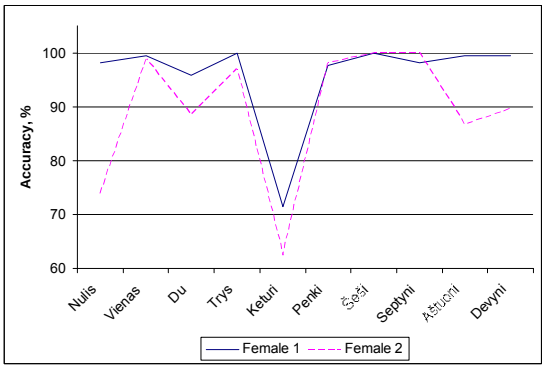

Figure 6. Averaged accuracy of Lithuanian digits recognition for two females

Speech corpora should be used in the future experiments of HMI modeling. Lithuanian speech corpus LTDIGITS which was successfully applied for phoneme classification [24] could be used in such experiments (the sequences of digits).

5.2. Subjective evaluation of dialogs by users

An end-user evaluation of all three human machine dialog systems has been performed. We have evaluated the performance (how fast and easy is the dialog flow - i.e. what time it takes to get to the desired goal), accuracy (how accurately the application responds to users input, how the utterances and situations are handled, etc.), naturalness (how natural the dialog flow is to end-user, comparing to real human person), recall (how easy it is to remember the control scheme of an application) and usability (overall usability, considering using such types of voice control in day to day application basis) aspects. The same speakers evaluated the dialogue systems, by rating from 1 (worst) to 10 (best). The results of subjective evaluation are presented in Table 3.

Table 3. Results of subjective evaluation of three dialog modes

Feature	Dialog mode		
	Isolated words	Dictation	Keywords
Performance	**9.6**	7.3	8.2
Accuracy	8.6	**8.8**	8.5
Naturalness	6.4	7.4	**8.9**
Recall	**9.8**	7.9	8.7
Usability	**8.6**	6.2	8

The performance aspect was rated best for the dialogs based on an isolated word recognition principle (scored 9.6/10) as it took a shortest time to say one or another short utterance (i.e. "do this" or "I need that"). Keywords were rated lower (8.2/10 as the users had the capability to use short commands if they wished), while dictation was rated as the slowest of them all (7.3/10 - the long and detailed instructions were used for each task).

All users decided that the interface worked most accurately in dictation mode (8.8/10) closely followed by isolated (8.6/10) words and keywords recognition (8.5/10) modes.

Users rated the keywords model as a most natural system (8.9/10), probably because the users were freely able to speak one way or the other, while logically simple isolated commands were considered to be the most unnatural (6.4/10) way of talking.

Recall function was the best in isolated recognition mode (9.8/10) as it is very easy to say simple voice commands. Keywords were rated second (8.7/10) due to the limitation in grammar rules not allowing the free flow of dialog. The dictation mode came last (7.9/10) due to forced condition to utter fully detailed instructions which were difficult to remember for some users.

The isolated words recognition model was rated as the most usable (8.6/10) due to simplicity and similarities with IVR call center services (simple menu type of interaction), keywords model was rated second (8/10) because the system was unable to accurately match any uttered keyword due to limited grammars at this phase of the development. The dictation model was rated worst (6.2/10) simply due to complexity of the utterances and the length of interaction itself.

6. Conclusions and future work

The results of an experimental evaluation of the three proposed "algorithms" of HCI dialogs showed that the best recognition accuracy was achieved for the dictation dialogs (96.9 %). Isolated words were recognized similarly to the keyword spotting (93.3 and 93 %).

The analysis of the recognition accuracy and the results of the subjective evaluation of all three models has shown that the implementation of a HMI dialog model, capable of recognizing keywords from the naturally sounding sentences is promising and definitely is the most suitable for real-life applications. This model ensured a reasonably high 93 % recognition accuracy of Lithuanian digits.

Further experiments with speech corpora should be done to prove the above mentioned conclusion.

Acknowledgements

This research was done under the grant by Lithuanian Academy of Sciences for the research project: No.: 20100701-23.

References

[1] **F. Marque, S. Bennacef, F. Néel, S. Trinh.** PAROLE: A Vocal Dialogue System For Air Traffic Control Training. *Applications of Speech Technology*, 1993.

[2] **W. Ward.** Extracting Information in Spontaneous Speech. *Proceedings of International Conference of Speech and Language Processing,* ICSLP, 1994, 83–86.

[3] **S. K. Bennacef, H. Bonneau-Maynard, J.L. Gauvain, L.F. Lamel, W. Minker.** A Spoken Language System For Information Retrieval. *Proceedings of International Conference of Speech and Language Processing,* ICSLP, 1994, 1271–1274.

[4] **J. L. Gauvain, S. Bennacef, L. Devillers, L. Lamel, S. Rosset.** Spoken Language Component of the MASK Kiosk. *K. Varghese, S. Pfleger, S., Human Comfort & Security of Information Systems*, Springer-Verlag, 1997, 93 – 103.

[5] **L. Lamel, S. Rosset, J. Gauvain, S. Bennacef, M. Garnier-Rizet, B. Prouts.** The LIMSI ARISE system. *Interactive Voice Technology for Telecommunications Applications*, IVTTA, 1998, 209–214.

[6] **H.M. Meng, C. Wai, R. Pieraccini.** The use of belief networks for mixed-initiative dialog modeling. *IEEE Transactions on Speech and Audio Processing*, Vol. 11, Issue 6, 2003, 757 – 773. http://dx.doi.org/10.1109/TSA.2003.814380.

[7] **F. F. Martinez, J. Ferreiros, R. Cordoba, J. M. Montero, R. San-Segundo, J. M. Pardo.** A Bayesian networks approach for dialog modeling: the fusion BN. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, 4789 – 4792. http://dx.doi.org/10.1109/ICASSP.2009.4960702.

[8] **K. Hacioglu, W. Ward.** Dialog-context dependent language modeling combining n-grams and stochastic context-free grammars. *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 1, 2001, 537 – 540.

[9] **E. Levin, R. Pieraccini, W. Eckert.** A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing,* Vol. 8, Issue 1 , 2000, 11 – 23. http://dx.doi.org/10.1109/89.817450.

[10] **B. G. Deutsch.** The structure of Task Oriented Dialogs. *In IEEE Symposium on Speech Recognition: Contributed Papers,* 1974, 13 p.

[11] **N. Ostler.** LOQUI: How Flexible can a Formal Prototype Be? *The Structure of Multimodal Dialogue*, Elsevier Science, 1989, 407–416.

[12] **B. Grau, A. Vilnat.** Cooperation in Dialogue and Discourse Structure. *IJCAI workshop on Collaboration, Cooperation and Conflict in Dialogue Systems*, 1997, 33-39.

[13] **D. Litman, J. Allen.** A Plan Recognition Model for Subdialogues in Conversations. *Cognitive Science*, 11, 1987, 163–200. http://dx.doi.org/10.1207/s15516709cog1102_4.

[14] **T. Yamoka, H. Iida.** A Method to Predict the Next Utterance using a Four-layered Plan Recognition Model. *European Conference on Artificial Intelligence*, ECAI,1990, 726–731.

[15] **J. Allen.** Natural Language Understanding. *Addison Wesley*, 1994, 654 p.

[16] **H. Bunt.** Information dialogues as communicative action in relation to partner modeling and information processing. *F.N. Taylor, D. Bouwhuis. Structure of Multimodel Dialogue*, North Holland, 1989, 47-73.

[17] **S. Young, G. Hauptmann, E. Smith, P. Werner.** High Level Knowledge Sources in Usable Speech

Recognition Systems. *Communications of the ACM*, 1989, 183–194.

[18] S. Young, C. Proctor. The Design and Implementation of Dialogue Control in Voice Operated Database Inquiry Systems. *Computer Speech and Language*, 3, 1989 , 329–353. http://dx.doi.org/ 10.1016/0885-2308(89)90002-8.

[19] A. Matrouf, J. Gauvain, F. Néel, J. Mariani. An Oral Task-Oriented Dialogue for Air-Traffic Controller Training. *SPIE Applications of Artificial Intelligence*, Vol. 1253, 1990, 826–837.

[20] R. Sarikaya, Gao Yuqing, H. Erdogan, M. Picheny. Turn-based language modeling for spoken dialog systems. *International Conference on Acoustics, Speech, and Signal Processing*, 1993, I, 27-30.

[21] S. Bangalore, G. Di Fabbrizio, A. Stent. Learning the Structure of Task-Driven Hu man–Human Dialogs. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.16 (7), 2008, 1249–1259. http://dx. doi.org/10.1109/TASL.2008.2001102.

[22] R. Maskeliunas, A. Rudzionis, V. Rudzionis. Advances on the Use of the Foreign Language Recognizer. *LCNS 5967: Development of Multimodal Interfaces: Active Listening and Synchrony*, Springer, 2010, 217–224. http://dx .doi.org/10.1007/978-3-642-12397-9_18.

[23] R. Maskeliunas, A. Rudzionis, K. Ratkevicius, V. Rudzionis. Investigation of Foreign Languages Models for Lithu anian Speech R ecognition. *Electronics and Electrical Engineering*, Kaunas, Technologija, 2009, VOL. 3(91), 37–42.

[24] K. Driaunys, V. Rudzionis, P. Zvinys. Implementation of Hi erarchical Phoneme Classification Approach on LTDIGITS Corpora. *Information Technology And Control*. 2009, V ol.38, No.4, 303 - 310.

Received March 2011.