Modeling Aspects of Multimodal Lithuanian Human - Machine Interface

Rytis Maskeliunas

Kaunas University of Technology, Speech Research Lab.,
Studentu 65-108, 51367 Kaunas, Lithuania
rytis.maskeliunas@ktu.lt

Abstract. The paper deals with modeling of multimodal human - machine interface. Multimodal access to the information retrieval system (computer) is possible by combining three different approaches: 1) Data input / retrieval by voice; 2) Traditional data input / retrieval systems; 3) Confirmation / rejection by recognizing and displaying human face expressions and emotions. A prototype of multimodal access for web application is presented by combining three modalities. Lithuanian language speech recognition experiment results on transcriptions and outcomes of discriminant analysis are presented.

Keywords: Multimodal interface, Lithuanian speech recognition.

1 Introduction

At the beginning of computer era the only way to interact with the machine was the use of perfect knowledge of programming, database systems and hardware. After developments and advances in microelectronics during 1970 – 1980's more modern computer technologies evolved, enabling new and inexperienced users to access and process the information without the need for special knowledge or the help of specialists. At the same time a human – machine interface began evolving and experiments were started on multimodal ways of interaction. The natural speech input and output allowed greatly enhance traditional communication modes, such as using a keyboard and the manipulation of icons and the text menus on computer display. The inclusion of speech recognition and synthesis greatly improved naturalness and accessibility of the information input and the retrieval, especially for inexperienced and disabled users. The experiments of Chapanis [2] showed the advantages of natural speech vs. the traditional ways: high performance and reliability of task execution and optimized interplay of the modalities make natural and spontaneous communication possible. Later on, the video based systems, such as a gesture, lips recognition, eye tracking, the electroencephalograph, etc. were added as additional mode of the human – machine interaction.

2 Main Human – Machine Interface Modalities

Multimodal systems offer many advantages over the traditional interfaces: the better error correction, improved handling in various situations and tasks and better accessibility [3].

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 75–82, 2009.

Humans perceive information by the senses known as modes, computers do that by using the artificial alternatives. To facilitate a more fluid human – machine interface it is better to enable machines to acquire the human – like skills, such as recognizing faces, gestures, speech, rather than to demand that humans should acquire the machine – like skills [4].

The best way in developing the most natural multimodal interface is to combine the main input / output types.

The direct manipulation, made popular by Apple and Microsoft graphical operating systems, has been based on the visual display of objects of interest, the selection by pointing, the rapid and reversible actions and a continuous feedback [16]. The virtual representation of information has been created, which can be operated by a user through the physical actions. [5]. It is important, that the traditional user interfaces have not been useful to disabled (blind, paralyzed, etc.) users. In the mobile devices those technologies don't work well either. It is not handy to type a text using a miniature (or simplified) PDA or a smart phone keyboard or a touch screen character recognition. The combining of speech and the traditional input might be a solution. However, there's still no real substitution for a keyboard and a mouse in the traditional computer use scenarios.

The speech plays main role in human communication. The main problem of an application of the speech human – machine interface has been the speech recognition (mostly accuracy in noisy and multi user environments, vocabulary size). Humans usually overestimate the capabilities of a system with the speech interface and try to treat it as another live person [6]. The speech technology has been maturing rapidly and the attention has been switching to the problems of using this technology in real applications, especially applications which allow a human to use the voice to interact directly with the computer-based information or to control the system. The systems therefore involve a dialogue and the discourse management issues in addition to those associated with prompting the user and interpreting the responses [18].

For the many reasons (the visibility, the relationship to a spoken signal, the ease of a graphical representation and the usefulness in speech-reading) the lips have been the most commonly modeled articulatory system [9], [13], [14]. As the speech recognition can degrade in noisy environments, the lip motion modality provides additional performance gains over those which are obtained by the fusion of audio and by the lip texture alone, in both cases of the speaker identification and the isolated word recognition scenarios [1]. Humans naturally use the faces as the organs of expression, reaction and query, which enrich the speech. The facial expression and the head movements are often used, when people express reactive states in conversations with others. The facial information (3D or 2D facial geometry, a skin pattern or a reflection, a facial thermogram, etc.) can also be used for the computer recognition of a personal identity (biometrics) [10], [11], [17].

2.1 Prototype Multimodal Web Access Application - FictionOIL

The multimodal access web application – FictionOIL has been developed to imitate a self service petrol station. A user can input and retrieve data by the voice (human – machine voice dialog), nod his head to confirm or reject a request or use traditional ways, greatly expanding usability and accessibility.

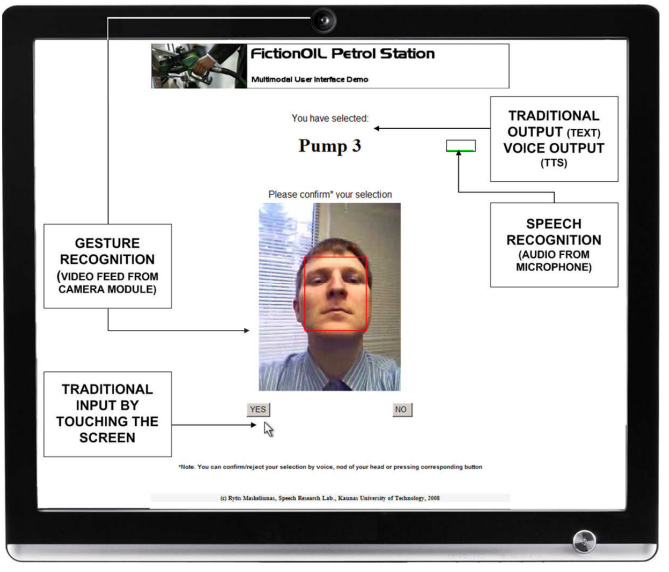

Fig. 1. User can confirm his selection by the voice (saying yes or no), nodding his head (up – down means yes, left – right means no) or pressing the correct button by touching the screen

The multimodal approach is utilized by combining ASP, the Speech Application Language Tags (SALT) or VoiceXML and the human face gesture recognition technologies (Fig. 1).

ASP (or similar markup languages, such as JAVA) can be used to create the traditional user interface. The information is displayed on the terminal monitor, the data are entered by a keyboard (specific actions are mapped to corresponding buttons) or directly manipulating the objects on a screen (using a mouse pointer or a finger with a touch screen display).

SALT or VoiceXML can add the second modality – the voice, which can greatly improve the usability and the accessibility. Therefore a user can input the data by simply saying what he wants and hearing the response. The voice Input can be easily controlled by using the simple syntax SALT elements, such as listen and grammar (W3C standard grammars are supported). The dynamic data (such as fuel prices) can be read by using the Lithuanian text – to – speech engine [7]. Again the prerecorded voice or the simple SALT elements, such as prompt can be used. The Lithuanian speech recognition has been developed on the base of regularized discriminant analysis and by using the transcriptions in the current commercial environments. The human – machine speech dialog has been created based on recommendations by Minker et al. [8].

The third modality can be added by recognizing and displaying a head movement (i.e. nodding to confirm), human face gestures, emotions and lip movements. This also extends the usability and the naturalness of the application. Video data (clusters of speech sounds mapped to the lips) used have been collected and processed by our colleagues at Image processing and analysis lab.

By analyzing the survey of the researchers testing this application, the following observations could be addressed:

The interpretation of the semantic text will be a difficult problem for some time and designers of voice based dialog systems should impose the restrictions on voice user interface and should use the voice commands with some prescribed meanings even if such approach is more difficult and less comfortable for a user. Looking from the speech technologist perspective the voice command recognition imposes own restrictions. Since the voice command does not have a grammatical content it is often difficult to use the statistical language models in such applications. Analyzing the causes of the recent progress in the speech recognition area it seems that the statistical language models have had a significant impact. However, the recognition of the simplified voice commands should be based on acoustic modeling.

The security problems arise when adding the speech and video modalities. The main problem is how to identify a speaking person (significantly important in remote control scenarios) and to prevent another person from intervening and entering incorrect data accidentally or on purpose. Partially this can be solved by using the biometrics technologies. Another problem is related to the poor speech recognition accuracy in noisy environments. This can be solved by limiting the size of the recognition vocabulary (decreasing dialog naturalness as fewer responses are accepted), by using the directional microphones (losing remote control possibility) or by adding the video technologies, such as the lips recognition.

3 Experiments on Lithuanian Speech Recognition

The experiments have been carried out by using the new Lithuanian speech corpus (basic segmentation, large sound units). This corpus has been expanding rapidly (more speakers and phonemic data added) and has been segmentated into small phonetical units. In the near future the experiments are going to be repeated by using the better quality speech data and the transitional parts of the diphones.

3.1 Discriminant Analysis Experiments

The modern automatic speech recognition systems possess the spoken language recognition capabilities that are far below the capabilities of human beings. One of the reasons is the fact that the human speech recognition and the automatic speech recognition have been performed in entirely different manner. This is particularly evident in the adverse environments or in the recognition of semantically meaningless sentences and phrases. These facts advocate that acoustic modeling of a speech signal has been carried out inefficiently in the automatic speech recognition systems.

The state-of-the-art of the speech recognition model which is based on the three state left-to-right continuous density hidden Markov model (CD-HMM) and the Mel – frequency cepstrum coefficients (MFCC) features, supplemented with the delta and delta-delta coefficients, enables to achieve serious progress in the acoustic modeling of the speech signal. However, the CD-HMM has had a lot of drawbacks, such as the relatively weak discrimination abilities. The linear discriminant analysis (LDA) is a method based on the discrimination capabilities with a good stability and a relative simplicity. However, this method needs large amounts of data for training and is sensitive to the data deviations in correlation matrix estimation procedure. To avoid these

shortages regularized discriminant analysis (RDA) has been applied [12]. The main idea of RDA is to modify the estimation procedure of correlation matrix and to provide the best recognition accuracy on a given training data set. This is achieved introducing a singular value in the decomposition of the correlation matrix. Then scaling and rotating of its estimates follows:

$$S^{RM} = T^{\alpha}(D + \lambda I)T^{\alpha'} \ . \tag{1}$$

where S^{RM} – the scaled and rotated estimate of a covariance matrix, T and D – are the singular values of the singular value decomposition, α and λ are the scaling and rotation coefficients. The optimal values of α and λ are evaluated empirically from the training data.

Previous experiments [15] showed the effectiveness of RDA approach comparing with other statistical classification methods (such as LDA, k-means, etc.) for the recognition of diphones composed from the nasal consonants or the semivowels and different vowels. These comparisons do not include modeling using CD-HMM for the same tasks. The motivation for it has been to extend phonetic data set and to evaluate the RDA approach efficiency with the results obtained by the Markov model.

A phrase which contains six diphones (ma, na, mi, ni, mu, nu) pronounced in isolation has been used. Each speaker (225 female speakers and 130 male speakers) pronounced these six diphones once. Since some pronunciations were of lower quality than the others, 220 female utterances and 125 male utterances have been used in the experiments. Utterances of 200 female and 110 male speakers were used for training, utterances of 20 female and 15 male speakers were used for testing.

The boundaries of the each diphone in recording and the boundary between a consonant and a vowel were fixed by an expert (human labeler).

An acoustic signal has been described using the MFCC features. They have been compiled using 20 msec analysis frame and 10 msec analysis step. In the recognition experiments with CD-HMM the MFCC feature vectors composed from 12 (cepstrum coefficients only) or 39 (cepstrum coefficients, delta and acceleration coefficients and energy) coefficients have been used. In the RDA discrimination experiments feature vectors composed of 24 coefficients (12 of them describing consonant part of diphone and 12 of them describing vowel part of diphone) have been used. These coefficients were the same MFCC coefficients as in CD-HMM recognition. The consonant part and vowel part descriptions were obtained by averaging MFCC coefficients of 5 consecutive frames from the stationary part of a consonant and a vowel.

For testing CD-HMM based recognition HTK provided tools have been applied. For RDA experiments proprietary software tools have been written in MATLAB.

Table 1 shows six diphones recognition results by applying CD-HMMs.

Table 1. Recognition results by applying the CD-HMM

Experiment type	male		female	
	Correct, %	Accuracy, %	Correct, %	Accuracy, %
12 MFCC, 1 set	78.57	78.57	69.14	68.00
12 MFCC, 2 set	84.29	84.29	68.42	64.66
39 MFCC, 1 set	92.86	92.86	78.57	77.98
39 MFCC, 2 set	90.00	88.57	75.00	74.40

Table 2. Discrimination results by applying the regularized discriminant analysis

Experiment type	Male, correct %	Female, correct %
4 classes, (ma, na, mi, ni)	89	79
4 classes, (ma, na, mu, nu)	92	80.5
6 classes, (ma, na, mi, ni, mu, nu)	83	72

Table 2 shows the results of the recognition experiments by applying the regularized discriminant analysis. Several experiments were performed: with 6 and 4 classes of diphones.

3.2 Deployed Commercial English Recognizers Adapted to the Lithuanian Language

The current commercial products don't support the Lithuanian voice recognition engines. The only way of currently deployed voice server's application for the Lithuanian language is to use the English transcriptions of the Lithuanian words.

The analysis has been carried out for improving the transcription effectiveness. First comparison of single transcriptions (i.e. word KETURI[1] has been changed to K EH T UH R IH) to multiple variations, based on preliminary automatic syllables recognition (i.e. k eh t uh r ih 1 (simple), k eh t uh d iy 1, k eh g ey r ih 1, g ah t uh r ih and so on) has been carried out. Syllables recognition has been conducted by using a set of 6 plosives C= {p, t, k, b, d, g}, a set of 16 vowel like IPA (International Phonetics Association) units V={ao, ah, aa, ih, iy, uh, uw, eh, ae, ey, ay, oy, aw, ow, ax, er}. Thus 96 consonant-vowel (CV) alternatives for the recognition have been processed.

The multiple transcriptions have been created in the following way:

1. A word has been divided into CV syllables (i.e. KETURI = KE + TU + RI);
2. Each syllable has been tested through the syllable recognition engine (100 times);
3. The most frequent answers have been selected (KE was recognized 69 times as g ah, 18 times as t ax);
4. The multiple transcriptions have been combined in most frequent answers (i.e. KETURI(15) = {k eh t uh r ih 1(simple), k eh t uh d iy 1, k eh g ey r ih 1, g ah t uh r ih 1, ...};
5. The best transcriptions have been selected by checking the predefined acceptable utterances against garbage model.

The experiments have been carried out in a noisy environment (street, computer rooms, outside areas, etc.), phoning the test application hosted on Microsoft Speech Server (Microsoft English US Recognition engine) by using a regular GSM mobile phone. Each word (10 Lithuanian digits) was spoken 100 times by 20 speakers.

In the first part of experiment only one transcription of chosen Lithuanian digit has been used. As recognition accuracy was poor for the most speakers (~ 47 %), 15 transcriptions using previously described method have been created. As more transcriptions of a word "keturi" were accessible to a recognizer, the recognition accuracy improved up to 98 %.

[1] Word "Keturi" means "four" in Lithuanian.

Table 3. Experiment results (excerpts) of comparison of the recognition accuracy using single and multiple transcriptions

Number of transcriptions	Word „DU“	Word „TRYS“	Word „KETURI“
1	83	77	47
12	84	80	60
13	88	87	84
14	90	91	95
15	97	99	98

The results proved (table 3) that the recognition accuracy depends on a number of transcriptions. The best accuracy is achieved when a larger number of transcriptions is used (allowing a recognizer to choose one from more variations of each word). By decreasing the number of transcriptions (removing the one that was chosen most times) the recognition accuracy decreased (sometimes almost 50% worse). The method of using the transcriptions works well only in limited vocabulary recognition scenarios, as it is impossible to generate qualitative transcriptions in real-time for dictations systems.

4 Conclusions

1. At the moment the Lithuanian speech recognition has been carried out on commercial speech recognition systems. A prototype of multimodal interface for Lithuanian language recognition has its own peculiarities. Because of differences in the English and Lithuanian pronunciations the recognition accuracy has not been satisfactory. To improve the situation the using of the English transcriptions for Lithuanian words have been adapted and recognition accuracy of about 96% for limited vocabulary scenarios has been achieved.
2. To achieve the good recognition accuracy and the performance for large vocabulary scenarios the native Lithuanian language recognition engine has been developed. The experiments based on the discriminant analysis proved an improvement in the recognition accuracy.
3. Another challenge was to introduce biometrical information in situations when identifying a speaking person acting in remote control scenarios. The security of operations can be increased by combining the information on voice recognition and face contours identification technologies.

References

1. Cetingul, H.E., Erzin, E., Yemez, Y., Tekalp, A.M.: Multimodal speaker/speech recognition using lip motion, lip texture and audio. In: Source Signal Processing, vol. 86(12), pp. 3549–3558. Elsevier North-Holland Inc., Amsterdam (2006)
2. Chapanis, A.: Interactive Communication: A few research answers for a technological explosion. In: Neel, D., Lienard, J.S. (eds.) Nouvelles Tendances de la Communication Homme-Machine, pp. 33–67. Inria, Le Chesnay (1980)

3. Cohen, P.R., Oviatt, S.L.: The Role of Voice Input for Human - Machine Communication. Proceedings of the National Academy of Sciences 92(22), 9921–9927 (1995)
4. Dauhman, J.: Face and gesture recognition: Overview. IEEE Transaction on Pattern Analysis and Machine Intelligence 19(7), 675–676 (1997)
5. Grasso, M.A., Ebert, D.S., Finin, T.W.: The integrality of speech in multimodal interfaces. ACM Transactions on Computer-Human Interaction (TOCHI) 5(4), 303–325 (1998)
6. Jones, D.M., Hapeshi, K., Frankish, C.: Design guidelines for speech recognition interfaces. Applied Ergonomics 20(1), 40–52 (1990)
7. Kasparatis, P.: Diphone databases for Lithuanian text – to – speech synthesis. Informatica 16(2), 193–202 (2005)
8. Minker, W., Bannacef, S.: Speech and human machine dialog. Kluwer academic publishers, Boston (2004)
9. Parke, F.I.: Parameterized models for facial animation. IEEE Computer Graphics and Applications 2(9), 61–68 (1982)
10. Pike, G., Kemp, R., Brace, N.: The Psychology of Human Face Recognition. In: IEE Colloquium on Visual Biometrics, Ref. no. 2000/018, pp. 11/1—11/6. IEE Savoy Place, London (2000)
11. Prokoski, F.J., Riedel, R.B., Coffin, J.S.: Identification of Individuals by Means of Facial Thermography. In: Proceedings of the IEEE 1992 International Carnahan Conference on Security Technology, Crime Countermeasures, October 14–16, pp. 120–125. IEEE Press, New York (1992)
12. Raudys, S.: Statistical and neural classifiers: An integrated approach to design. Springer, London (2001)
13. Reveret, L., Bailly, G., Badin, P.: MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In: Proceedings of the 6th International Conference of Spoken Language Processing, pp. 16–20. ISCA, Beijing (2000)
14. Reveret, L., Essa, I.: Visual Coding and Tracking of Speech Related Facial Motion. Georgia institute of technology technical report GIT-GVU-TR-01-16. Georgia institute of technology (2001)
15. Rudzionis, A., Rudzionis, V.: Phoneme recognition in fixed context using regularized discriminant analysis. In: Proceedings of EUROSPEECH 1999, vol. 6, pp. 2745–2748. ESCA, Budapest (1999)
16. Shneiderman, B.: Direct manipulation: A step beyond programming languages. In: Human-computer interaction: a multidisciplinary approach, pp. 461–467. Morgan Kaufmann Publishers Inc., San Francisco (1987)
17. Smith, W.A.P., Robles-Kelly, A., Hancock, E.R.: Skin Reflectance Modeling for Face Recognition. In: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR 2004), vol. 3, pp. 210–213. IEEE computer society, Washington (2004)
18. Young, S.: Speech understanding and spoken dialogue systems. In: IEE Colloquium on Speech and Language Engineering – State of the Art (Ref. No. 1998/499), pp. 6/1—6/5. Savoy Place, London (1998)