Technische Universität
Kaiserslautern

SCHRIFTEN ZUR

FUNKTIONALANALYSIS
UND GEOMATHEMATIK

G. Hebinger, V. Michel, M. Richter, A. Simon

Speech Recognition Support of Assisted Living

Bericht 40 – September 2008

FACHBEREICH MATHEMATIK

Speech Recognition Support of Assisted Living

Georg Hebinger[1], Volker Michel[1,2], Michael Richter[3], Andreas Simon[1]

[1] University of Kaiserslautern, Geomathematics Group

D-67653 Kaiserslautern, Germany
spracherkennung@mathematik.uni-kl.de

[2] currently at: Dept. of Applied Mathematics and Theoretical Physics, Univ. of Cambridge

Cambridge CB3 9AL, United Kingdom

[3] Dept. of Computer Science, University of Calgary

Calgary, AB, T2N 1N4, Canada
mrichter@.ucalgary.ca

Abstract. We present results and views about a project in assisted living. The scenario is a room in which an elderly and/or disabled person lives who is not able to perform certain actions due to restricted mobility. We enable the person to express commands verbally that will then be executed automatically. There are several severe problems involved that complicate the situation. The person may utter the command in a rather unexpected way, the person makes an error or the action cannot be performed due to several reasons. In our approach we present an architecture with three components: The recognition component that contains novel features in the signal processing, the analysis component that logically analyzes the command, and the execution component that performs the action automatically. All three components communicate with each other.

Keywords: Disabled persons, assistance, speech recognition, logical analysis.

1 Introduction

The percentage of persons in the EU older than 60 years of age will dramatically increase in the next 30 years and will rise to one third of the population, see [1]. These persons will need costly special attention to master their life, for instance in nursery homes.

In this paper we report about an ongoing project for supporting assisted living using speech recognition on the basis of wavelets for performing actions (supported by Stiftung Innovation Rheinland-Pfalz 15202-386261/773). Most parts of the approach are already working.

The goal is to perform actions automatically that elderly and/or handicapped persons are not able to do. There, we consider a room in a nursery home where the persons themselves cannot perform certain actions such as ”open the window”. Instead, the person will present a spoken command. A speech recognizer tries to understand the command that will be analyzed subsequently. In case of arising questions there is a spoken feedback; otherwise the command is transferred to an execution machine that will perform the action automatically.

There are several problems involved that prevent us from simply using a commercially available speech recognizer. The most important problems are:

a) Formulation of an erroneous command (e.g. one that is impossible to perform, dangerous or highly implausible)

b) Formulation of an incomplete command and, thus, creation of an ambiguity.

c) Utterance of words or phrases which are not in the vocabulary

d) Interference by other speakers (persons, radio, TV)

e) Occurrence of noise (e.g. due to open windows), etc.

In general, such errors can almost never be detected because the uttered command can refer to everything in the world. In human conversations, the background and the common sense knowledge are used to identify such errors. Some speech recognition systems have employed knowledge for this purpose with a very limited success.

In our approach, we use the fact that we can completely describe our scenario, i.e., we have a closed world scenario. This allows us to discuss all problems internally. There is, however, one part of the relevant world that is not represented in the scenario; this comprises the speaker and the internal states of the speaker. In order to get access to this, we realize a communication with the speaker in spoken language.

The system has the following components:

- The recognition component (microphone, signal analyzer and denoiser, speech recognizer)
- The analyzing component (discussing if the command can or should be executed)
- The execution component (performing the command if possible).

In addition, there is the speaker as an external participant.

In section 2 we present the general approach, sections 3, 4, and 5 describe the recognition, the analysis, and the execution.

2 The Approach

2.1 The Example Domain

The example scenario we are investigating consists of a living room in which an old or disabled person lives who is incapable of executing everyday actions.

The room contains several lights, a TV-set and a radio.. Furthermore, window blinds can be controlled.

The person denotes each action by a command that can, however, be formulated in several different ways. For instance, the command ”open the window” can be formulated as:

”open the window”, ”open the window now”, ”open the window right now”, ”please open the window”, ”open the window, I need fresh air”, etc.

2.2 The General Architecture

The architecture is built in order to realize the intended goals. It has three major components that are organized in a modular way so that they can be improved and/or modularized independently. The system has to perform the following tasks:

Task 1: Understand the spoken command
Task 2: Analyze the understood command with respect to possible errors
Task 3: Transform the recognized sentence into a formal and an executable form
Task 4: Execute the command

The general structure is shown in fig.1.

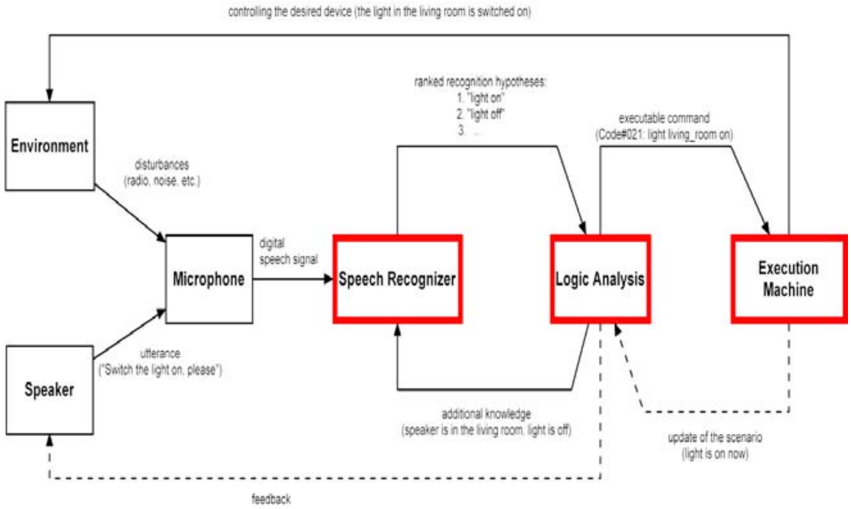
controlling the desired device (the light in the living room is switched on)
Environment
disturbances (radio, noise, etc.)
Microphone
digital speech signal
Speech Recognizer
ranked recognition hypotheses: 1. "light on" 2. "light off" 3. ...
Logic Analysis
executable command (Code#021: light living_room on)
Execution Machine
Speaker
utterance ("Switch the light on, please")
additional knowledge (speaker is in the living room, light is off)
update of the scenario (light is on now)
feedback

Fig. 1: Overall architecture of the project

The architecture shows three independent modules, that are connected by interfaces and communicate with each other:

a) The *recognition component* accepts spoken language that has to be understood.
b) The *analysis component* obtains an abstract command as input. It analyzes the commands with respect to correctness, errors, plausibility and related aspects.
c) The *execution component* performs the action intended by the human. It has access to all objects in the room to be manipulated and has a (formal) representation of the actions and processes.
d) In addition, it shows a sensor component but that is not realized yet.

For the spoken feedback we take the speech synthesis system Mary, see [2].

3 The Recognition Component

The recognition component contains the signal processing part and the proper recognition part. Because of the challenging demands from the application we had to equip the signal processing with some new features.

3.1 Signals

The signal analysis tool was created for the particular application. In a nutshell, the obtained acoustic signal is decomposed into different packages. These packages enable the extraction of characteristic features for the distinction of different utterances.

A classical toolbox of signal analysis is the (windowed) Fourier transform. In this case, the incoming signal is decomposed into its frequency spectrum, i.e. the original signal, which represents the volume with respect to the time, is transformed to a representation in terms of the amplitudes of the included frequencies. This information is helpful, if one wants to distinguish different pitches of sounds.

The windowed Fourier transform subdivides the time interval into small subintervals of equal length and performs the procedure on each subinterval separately. However, this procedure is too inflexible for speech recognition. It is known (see e.g. [3]) that different classes of phonemes need to be characterized in their time-frequency behavior in different ways. For instance, vowels correspond to relatively long time intervals and can be characterized by a precise measurement of the three dominant (i.e. associated to the highest energy) frequencies (see fig. 3), whereas plosives can be characterised by the point of time and the length of the time interval. Hence, vowels require a high resolution in the frequency domain and plosives require a high resolution in the time domain. Both cannot be achieved simultaneously due to Heisenberg's uncertainty principle. Consequently, the windowed Fourier transform with its fixed sizes of the time and the frequency windows appears not to be an ideal choice.

Fig. 2 shows the first three dominant frequencies (formants) of recorded vowels, calculated via the windowed Fourier transform, with different lengths of the time window: 12ms (left-hand side) yields a good separation, whereas 5ms (right-hand side) gives a poor distinction. Hence, the choice of the size of the window is very critical.

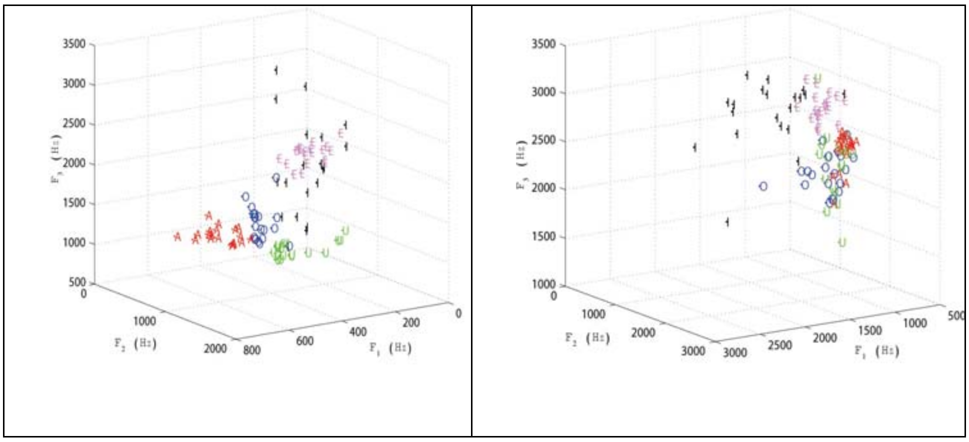

Fig. 2: First three formants of recorded vowels, calculated via the windowed Fourier transform with window size of 12ms (left) and 5 ms (right), respectively.

Therefore, we use a different approach based on the Local Trigonometric Transform (LTT) (see [4] for further details). The LTT has a particular advantage, similarly like the Wavelet Packet Transform (see [5], [6],). It enables a flexible adaptation of the sizes of the time and frequency windows. In subintervals where the frequency spectrum remains rather constant in time (as it is the case for vowels), the time window is adapted to a large size (low temporal resolution) in order to extract the occurring frequencies as precisely as possible (high frequency resolution), whereas in the opposite case a fine subdivision of the temporal interval is performed to focus on temporal changes of the spectrum rather than on the precise values of the frequencies. In fig. 2, one can see two long sections with rather constant frequency behavior each representing vowels (here ”o” from approximately 0.1 to 0.3 and ”u” from approximately 0.55 to 0.65), whereas another part at 0.40-0.42 shows a characteristic behavior of a short period of time (here the plosive ”t”). The extraction of characteristic features of this transform is a basis for the recognition process.

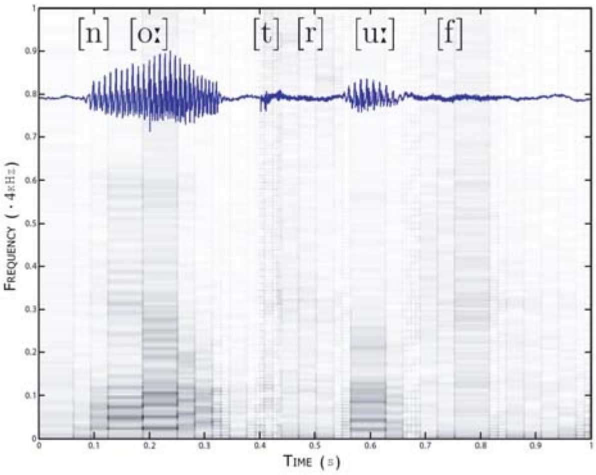

Fig. 3: LTT of the utterance ”Notruf” (”emergency”)

3.2 Recognition

The words in the commands are elements of the vocabulary of the system. The matching process itself is standard and will not be discussed here. A problem is that the commands may not be uttered in exactly the intended formal way. There are two ways to cope with this problem:

a) Top-down approach: Recognize the sentence with a recognizer that has linguistic capabilities and reconstruct the command with some method that is applied as a second step. This is the standard in existing systems.

b) Bottom-up approach: Recognize only key words from the commands and synthesize the command. Key words are e.g. “open” and “window”. From these key words the command is synthesized.

For the intended application scenarios the second approach has a number of other advantages:

- It requires a smaller vocabulary and, therefore, simplifies the training and increases the correctness.
- It does not require linguistic capacities.

For each command, we have a normal form which is identical with the formal command transferred to the execution component. The vocabulary has, however, different versions per one command stored that, e.g., take care of synonyms. That means there are usually several different human commands that are mapped to the same formal command because they have the same meaning. The key words are selected from the stored commands. E.g. for “turn the light on” the key words are “light” and “on”. The command is synthesized from these key words. In case of an ambiguity, all possible commands are given to the analysis component for resolution.

The bottom-up approach now is further supported by the special form of the signal analysis. This supports the detection of vowels and the sequence of the vowels determines to a large degree the discovery of the key words (the number of which is relatively small).

That means we have two representation forms for commands:

- Human expressions: Formulated by the human.
- Formal expressions: Understood by the computer.

The first of these expressions have to be translated to the second kind in a semantic preserving way. It means to transfer the recognized commands to the analysis component for further treatment.

4 The Analysis Component

Fig. 4 shows the general structure of the analysis component. It analyzes the action proposals obtained from the recognition component.

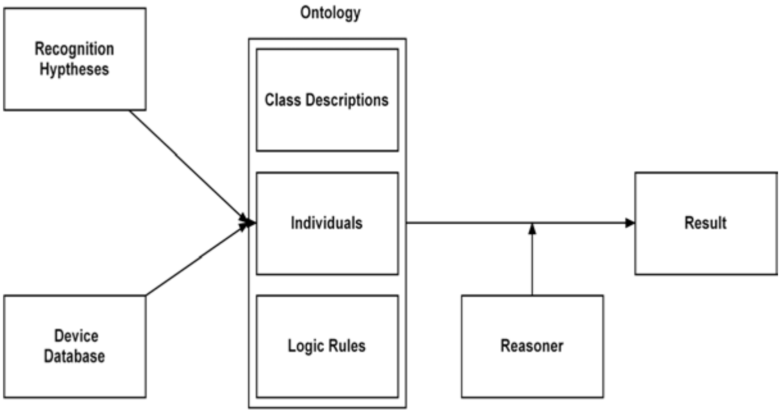

Fig. 4: Analysis component

The objects in the scenario are described by an ontology. For dealing with the dynamic character of the objects we use the attribute state where the values are changed by the actions. For instance, a window has the states {open, closed}.

The actions are also described in the ontology. They are mappings

A: states → states.

The actions have the form "action (object)" what says on which object the action is operating, e.g. action (window). For each action we have the following attributes:

- action_operations, e.g. action (light)_operations = {Switch On, Switch Off}.
- action_rating = {Possible, Impossible, Critical, Uncritical, Plausible, Implausible}.

The object operated by the action is Aop (). The main task of the analyzer is to determine the value of the attribute action_rating. Except for the first two values the remaining ones are in the first place fuzzy values what needs a defuzzification. In order to generate these values the reasoner is called. It has a rule package, e.g. with the rule:

IF the heater is on THEN the window cannot be opened.

For checking the preconditions of the rules, a query to the dynamic memory is send. After the values of action_rating are computed the result is sent again to the reasoner. A second rule package decides which of the following actions is taken:
a) Send a feedback to the user
b) Send the command to the execution component for executing the command.

5 The Execution Component

The execution component receives a command from the analysis component. At the same time it gets a message from the sensor component with additional information. Based on this, the component executes the command in the real world. This may be impossible for some reason that is unknown to the analysis component, e.g. if some part is broken. In each case, the component sends a message to the analyzer, either ”Executed” or ”Failed” in order to update the dynamic memory.
The execution component was realized by CIBEK technology + trading AG [7]. The CIBEK company has equipped 20 apartments in Kaiserslautern with “Senior-Displays“. This corresponds exactly to our scenario. In such apartments the speech recognition components will be installed.

6 Conclusion

This paper presents an example for assisted living in the form of a support for old and/or disabled persons. In a nursery home, a patient can express his/her commands verbally rather than perform an action like opening a window manually. This required an integrated architecture of speech recognition, logical reasoning and automated execution. In this respect, several novel elements have been integrated.

References

1. Statistisches Bundesamt Wiesbaden, http:// www.destatis.de/kontakt/
2. Mary: mary.dfki.de/documentation/maryxml
3. Hess, W. Fundamentals of Phonetics. Unpublished Manuscript. Universität Bonn, 2005.
4. Bittner, K. Biorthogonal local trigonometric bases. Handbook on Analytical-Computational Methods (ed. G. Anastassiou, CRC Press, 2000).
5. Coifman, R.R., Wickerhauser, M. V. Entropy-Based Algorithms for Best Basis Selection. IEEE Transactions on Information Theory 38, pp. 713-718, 1992.
6. Wickerhauser, M. V. Smooth localized orthonormal bases. C.R. Acad. Sci., Paris, 1993.
7. Cibek Company: http://www.cibek.de

Folgende Berichte sind erschienen:

2003

Nr. 1 S. Pereverzev, E. Schock.
On the adaptive selection of the parameter in regularization of ill-posed problems

Nr. 2 W. Freeden, M. Schreiner.
Multiresolution Analysis by Spherical Up Functions

Nr. 3 F. Bauer, W. Freeden, M. Schreiner.
A Tree Algorithm for Isotropic Finite Elements on the Sphere

Nr. 4 W. Freeden, V. Michel (eds.)
Multiscale Modeling of CHAMP-Data

Nr. 5 C. Mayer
Wavelet Modelling of the Spherical Inverse Source Problem with Application to Geomagnetism

2004

Nr. 6 M.J. Fengler, W. Freeden, M. Gutting
Darstellung des Gravitationsfeldes und seiner Funktionale mit Multiskalentechniken

Nr. 7 T. Maier
Wavelet-Mie-Representations for Soleniodal Vector Fields with Applications to Ionospheric Geomagnetic Data

Nr. 8 V. Michel
Regularized Multiresolution Recovery of the Mass Density Distribution From Satellite Data of the Earth's Gravitational Field

Nr. 9 W. Freeden, V. Michel
Wavelet Deformation Analysis for Spherical Bodies

Nr. 10 M. Gutting, D. Michel (eds.)
Contributions of the Geomathematics Group, TU Kaiserlautern, to the 2nd International GOCE User Workshop at ESA-ESRIN Frascati, Italy

Nr. 11 M.J. Fengler, W. Freeden
A Nonlinear Galerkin Scheme Involving Vector and Tensor Spherical Harmonics for Solving the Incompressible Navier-Stokes Equation on the Sphere

Nr. 12 W. Freeden, M. Schreiner
Spaceborne Gravitational Field Determination by Means of Locally Supported Wavelets

Nr. 13 F. Bauer, S. Pereverzev
Regularization without Preliminary Knowledge of Smoothness and Error Behavior

Nr. 14 W. Freeden, C. Mayer
Multiscale Solution for the Molodensky Problem on Regular Telluroidal Surfaces

Nr. 15 W. Freeden, K. Hesse
Spline modelling of geostrophic flow: theoretical and algorithmic aspects

2005

Nr. 16 M.J. Fengler, D. Michel, V. Michel
Harmonic Spline-Wavelets on the 3-dimensional Ball and their Application to the Reconstruction of the Earth's Density Distribution from Gravitational Data at Arbitrarily Shape Satellite Orbits

Nr. 17 F. Bauer
Split Operators for Oblique Boundary Value Problems

Nr. 18 W. Freeden, M. Schreiner
Local Multiscale Modelling of Geoidal Undulations from Deflections of the Vertical

Nr. 19 W. Freeden, D. Michel, V. Michel
Local Multiscale Approximations of Geostrophic Flow: Theoretical Background and Aspects of Scientific Computing

Nr. 20 M.J. Fengler, W. Freeden, M. Gutting
The Spherical Bernstein Wavelet

Nr. 21 M.J. Fengler, W. Freeden,
A. Kohlhaas, V. Michel, T. Peters
Wavelet Modelling of Regional and Temporal Variations of the Earth's Gravitational Potential Observed by GRACE

Nr. 22 W. Freeden, C. Mayer
A Wavelet Approach to Time-Harmonic Maxwell's Equations

Nr. 23 M.J. Fengler, D. Michel, V. Michel
Contributions of the Geomathematics Group to the GAMM 76th Annual Meeting

Nr. 24 F. Bauer
Easy Differentiation and Integration of Homogeneous Harmonic Polynomials

Nr. 25 T. Raskop, M. Grothaus
On the Oblique Boundary Problem with a Stochastic Inhomogeneity

2006

Nr. 26 P. Kammann, V. Michel
Time-Dependent Cauchy-Navier Splines and their Application to Seismic Wave Front Propagation

Nr. 27 W. Freeden, M. Schreiner
Biorthogonal Locally Supported Wavelets on the Sphere Based on Zonal Kernel Functions

Nr. 28 V. Michel, K. Wolf
Numerical Aspects of a Spline-Based Multiresolution Recovery of the Harmonic Mass Density out of Gravity Functionals

Nr. 29 V. Michel
Fast Approximation on the 2-Sphere by Optimally Localizing Approximate Identities

Nr. 30 M. Akram, V. Michel
Locally Supported Approximate Identities on the Unit Ball

2007

Nr. 31 T. Fehlinger, W. Freeden,
S. Gramsch, C. Mayer, D. Michel,
and M. Schreiner
Local Modelling of Sea Surface Topography from (Geostrophic) Ocean Flow

Nr. 32 T. Fehlinger, W. Freeden,
C. Mayer, and M. Schreiner
On the Local Multiscale Determination of the Earth's Disturbing Potential From Discrete Deflections of the Vertical

Nr. 33 A. Amirbekyan, V. Michel
Splines on the 3-dimensional Ball and their Application to Seismic Body Wave Tomography

Nr. 34 W. Freeden, D. Michel, V. Michel
Product Framelet Based Operator Decomposition

Nr. 35 M. Schreiner
The Role of Tensor Fields for Satellite Gravity Gradiometry

Nr. 36 H. Nutz, K. Wolf
Time-Space Multiscale Analysis by Use of Tensor Product Wavelets and its Application to Hydrology and GRACE Data

2008

Nr. 37 W. Freeden, M. Gutting
On the Completeness and Closure of Vector and Tensor Spherical Harmonics

Nr. 38 W. Freeden, K. Wolf
Klassische Erdschwerefeldbestimmung aus der Sicht moderner Geomathematik

Nr. 39 W. Freeden, T. Fehlinger, M. Klug, D. Mathar, K. Wolf
Classical Globally Reflected Gravity Field Determination in Modern Locally Oriented Multiscale Framework

Nr. 40 G. Hebinger, V. Michel, M. Richter, A. Simon
Speech Recognition Support of Assisted Living

TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Informationen:

Prof. Dr. W. Freeden
Prof. Dr. E. Schock
Fachbereich Mathematik
Technische Universität Kaiserslautern
Postfach 3049
D-67653 Kaiserslautern
E-Mail: freeden@mathematik.uni-kl.de
schock@mathematik.uni-kl.de