Towards an Evaluation Standard for Speech Control Concepts in Real-World Scenarios

Jens Maase[*], *Diane Hirschfeld, Uwe Koloska*[**], *Timo Westfeld, Jörg Helbig*[***]

[*] Bosch und Siemens Hausgeräte GmbH, Regensburg,
[**]voice INTER connect GmbH, Dresden, [***]MediaInterface Dresden GmbH
jens.maase@bshg.com, [hirschfeld,koloska]@voiceinterconnect.de, [westfeld,helbig]@mediainterface.de

Abstract

Speech control is still mainly evaluated through statistical performance measures (recognition rate, insertion rate, etc.) considering the performance of a speech recognizer under laboratory or artificial noise conditions. All these measures give no idea about the practical usability of a speech interface, since practical aspects concern more than operational aspects of the speech recognizer inside a product.
Since it was felt, that no evaluation standard so far fulfills the practical requirements for speech controlled products, this paper aims in the establishment of an open design- and evaluation standard for speech control concepts in real-world scenarios.
First, the behaviour of the users and the normal environmental conditions (typical noises) were evaluated in usability experiments. Data recordings were conducted, trying to capture these typical usage requirements in a special corpus (Apollo-corpus). Finally, a set of standard desktop-, as well as embedded speech recognizers were tested for their performance under these real world conditions.

1. Introduction

Bosch und Siemens Hausgeräte GmbH (BSH) planned to implement a speech control into an extractor hood. According to conventional user interfaces, the customer was ment to demand 100% functionality under all circumstances. Literature on speech recognition systems reported a recognition rate better than 90%, but first experiments with available solutions were felt to have even worse performance under real-world conditions.
An objective evaluation failed in adequate data, because the databases available for recognizer training and evaluation were mainly made for telephony applications [1], [3] or in-vehicle navigation [2], or suitable commands for control of home appliances were missing [4]. Only a few of the databases were containing real noises [3], the rest of the speech community reports real-world experiments with artificially added noise [7], [8].
So, BSH had to find an own solution to this problem.

2. Ergonomical aspects of product design for speech controlled devices

In order to find out more about a users acceptance level for recognition accuracy and about typical user behaviour, BSH built a usability lab with an implemented kitchen in L-shape, that can be modified by moveable walls into a small room (14 sqm) or a large room (24 sqm). In order to get a realistic acoustical environment with numerous reflections, the floor was covered with laminate and the walls were partly tiled.

A representative test group of persons was selected by age, sex, education and income. The behaviour of these subjects against the speech controlled device was tested in three situations (first contact, real cooking situation, recapitulation). In order to get reliable results concerning the acceptance of different recognition rates, a Wizard-of-Oz experiment was conducted by controlling the extractor hood manually over a computer programm that was designed to artificially reducing the human 100 % recognition accuracy down from 95 % to 70 % in 5 % steps. Each subject was confronted during the whole test with one constant recognition accuracy only.

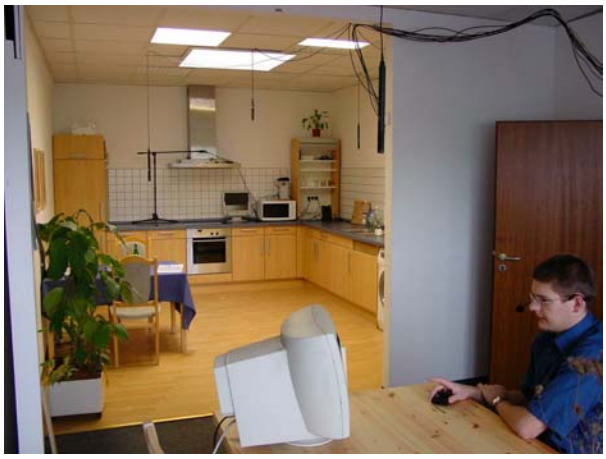

Figure 1: Usability Lab in Regensburg, view from far table.

A set of 48 test persons took part in the usability experiments. During the tests, the subjects could be observed through a semi-transparent mirror. DVD-video recordings with two cameras were made with timecode. There was the opportunity for 8 channel simultaneous audio recordings as well. After each experiment the opinion of the subjects was evaluated.
The findings of the experiments had a direct influence to the recording conditions of the test data for the evaluation process (cf. 3). All subjects expected the system (hood) working the right way. The recognition rate was accepted when better than 85%. This rate was wished to be constant independently from the subjects' position in the room. Some subjects even controlled the hood turning their back on it.
The subjects did not expect an acoustical answer (prompt) from the system. During cooking, the hood was not controlled. The preferred positions in relation to the hood were a) near the hearth (1 m distance) , b) at the sink beside the hood (3 m), c) from the near table (5 m) and d) from the far table (7 m).

3. Data acquisition and preprocessing – the Apollo corpus

Based on the experiences resulting from usability study, a real-world scenario for home speech control applications was elaborated, that defined the recording conditions for the evaluation data.

3.1. Requirements for evaluation data

First of all, a set of 17 connected word commands were choosen for the control of the device, all starting with the keyword "Apollo". These utterances were recorded in three repetitions under 5 different environmental conditions (silence, moderate and intense exhauster level, dishwasher and radio speaker background), in order to test typical noises with varying stationarity.

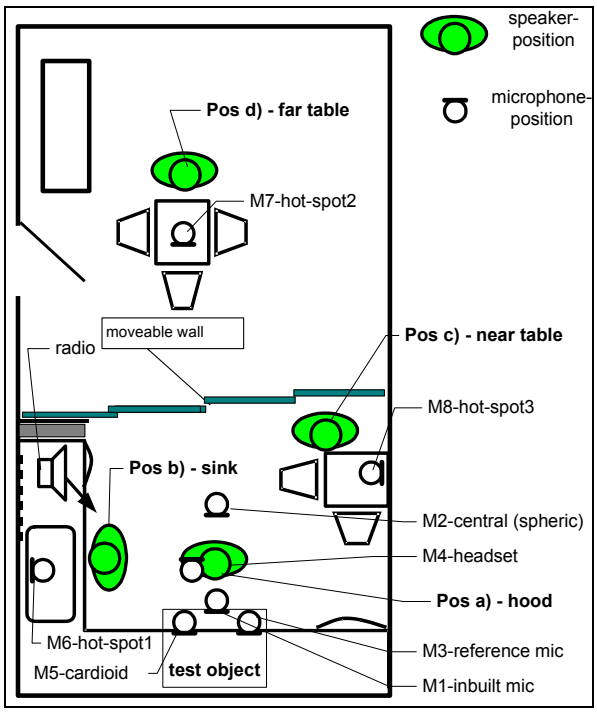

Figure 2: Plan of the usability lab together with microphone and subjects positions.

The subjects' positions in the laboratory were varied between position a) to d) from the previous usability experiments (view figure 2).
A total of 45 subjects were recorded, that were carefully chosen by age, sex and dialect. The speakers were asked to speak dialect. The recordings collected a total number of 1020 utterances per subject, resulting in a total recording time of 4 hours per session.

3.2. Data acquisition

The recordings collected simultaneously 8 audio-channels from 8 different microphones, positioned at different points in the kitchen. There were three microphones (M1, M3, M5) positioned at the hood, one inbuilt and one decoupled from the hood, with directional characteristics. A high quality directional microphone (M3) was placed on a stative as a reference directly above the hood. In the centre of the laboratory, a microphone with spheric characteristics was placed on the ceiling (M2). The subject was carrying a radio headset for the recording of one channel of clean reference data (M4). Finally, hotspot microphones with hypercardioid characteristics were placed above the three remaining speaker positions (M6-M8), in order to collect data for sound enhancement measures later on.
For each microphone, the sound level was calibrated and remained constant through the whole recordings. For compensation measures for bad SNRs at different positions, data was continuously recorded with 24 Bits data width at 48 kHz sampling frequency.
In order to capture the background conditions in the speech prompts, the recordings were made continuously, and a semi-automatic prompting over screen forced the subjects to sustain pauses of three seconds between the phrases, that allowed for later segmentation without cutting the echos and providing sufficient pauses before and after each utterance. In order to avoid learning effects, the order of the prompts was randomized. Quality control was carried out by the recording executive. Whenever an error was made, the prompt was re-recorded again.
Since the lab was placed in a business building and no special sound isolation was provided, the data is contamined with real world noises differing from the wanted ones (cellular phone, machine noise, hammer, traffic, and subjects noises like smacks, rustling of clothes, scratching etc.).

3.3. Data postprocessing

After recording, the raw data was converted to the target format (16 Bit, 44.1 kHz, mono, wav-Format) and cut into utterances. A set of automated tools allowed identical processing for all 8 channels.
Three control stages were carried out for errror localisation, documentation and categorization of unwanted noises as well as general characteristics of the speech data, and the detection and repair of processing errors (content of a file, length of pre- and post utterance pauses etc.). For each of the speakers, the following information were finally documented: age, sex, origin and dialect, and some general description. This was done to ease later selections at the database.
The data was burned to a set of 45 DVD. In all, a number of 293,662 files, and an amount of 180 GByte data were processed in a period of 6 weeks.
Additional data for insertion tests containing 4 h of speech (talks) and 10 h of radio-recordings (speech and music) were collected in the same recording conditions.

4. Evaluation environment

4.1. Test configuration and requirements

The test environment was designed to have universal conditions for a large variety of test objects – for embedded solutions as well as for PC-based recognizers. Therefor, the control PC and the test object are independent devices communicating only by standard cable interfaces.

A test object had to fulfill the following requirements:

- Always listening, no push-to-talk. The recognition result was transmitted via serial interface.

- The audio level on the cable (with the test object connected) was 100 mV peak voltage for a reference audio file (sinusoid, -6 dB peek value).
- Reply only complete commands, no parts of them. In our case the 17 German command sequences.

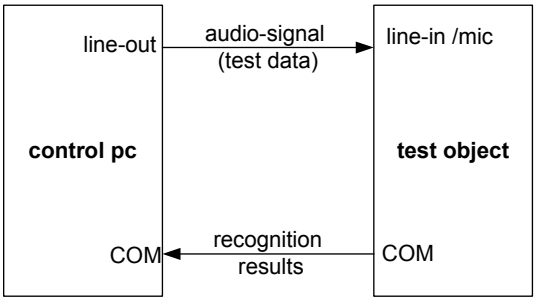

Figure 3: Test configuration

4.2. Test design and test management

In order to get detailed results for the typical usage situations, two kinds of tests were conducted:

Positive test: The normal way to evaluate the recognition rate. A defined subset of the test data (80%) was supplied to the participants before the test. The remaining 20% were not published to the participants.
The test data were - according to their characteristics – sorted in a way, that changes in environment and speaking position are similar to real-world conditions.

- The 45 speakers were not mixed.
- The 5 different background noise environments were changed as rarely as possible. Records having the same noise environment were grouped, subordinated to the speakers.
- The remaining characteristics "speaker position", "repetition", and the 17 different commands were randomly mixed.

The pause between two speech commands was always filled with the same noise like the background of the following spoken command. This measure supports adaptation processes (noise reduction units) in the recognizers and pays attention to the fact, that each speech recognizer has a certain output delay. In our test, the pause duration was fixed to 3 seconds (equal conditions for each test object).

Insertion test: To test the robustness (no unintended reactions) this special test evaluated the insertion rate (INS). 14 hour test data with a high speech portion but without complete spoken commands was collected. The material was not published to the participants. The insertion test data were provided without pauses between files.

The complete test was controlled and logged by the Testmanager software running on the control PC. It provided the audio samples at the sound card output according to test control scripts. In order to avoid clicks, gaps and other undesired effects at the concatenation point, a buffering strategy ensured a seamless connection of subsequent wave files. Further it was possible to smooth transitions between two audio samples by an adjustable cross fading (in our test 100 ms, marked by the x zones in Figure 4).

During the pause after a spoken command the Testmanager still accepted received strings as recognition results. Thus with the length of the pause wave (e), a defined reaction time period could be realized.
The Testmanager received the recognition results at the serial port of the control PC. Each string separated by <ASCII '013' '010'> was interpreted as a result and logged into a test protocol with the following parameters:

- ID of test object
- current audio file
- absolute start time of playing
- recognition result (received string)
- time delay (start time to end of result in ms)

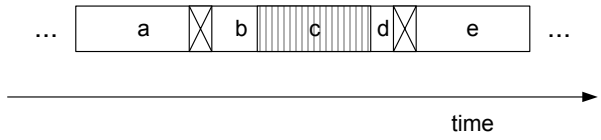

Figure 4: Structure of the audio signal in the positive test: pause wave with background noise I (a), test wave with background noise I (with pause before spoken command (b), spoken command (c), pause after spoken command (d)) and pause wave with background noise II (e).

5. Experimental results

Five software only and three hardware embedded recognizers took part in the evaluation. Altogether 45,900 spoken commands and 14 hour insertion data were provided to each test object. This corresponds to 5 days and 2 hours netto test duration per test object.

5.1. Evaluation criteria

For the evaluation, the logfile of the Testmanager (text format) was imported into a standard database. Each command, consisting of a three word sequence, was regarded as a "word" in the sense of the commonly used evaluation criteria like Word Error Rate [5], [6]. The answers of the recognizers were divided into four categories:

OK: Correct, the spoken command (c) was correctly recognized within the period (c) to (e).
SUB: Substitution, a wrong command instead of the spoken one was recognized within the period (c) to (e).
DEL: Missed, no command was recognized within the period (c) to (e) though a command was spoken.
INS: Insertion, an additional command was recognized. Thus in the positive test, an INS always had to be preceeded by an OK or a SUB. In the insertion test each result was an INS.

The recognition rate *RR* is calculated in percent (the correct recognitions related to the number of spoken commands):

$$RR = \frac{OK}{OK + SUB + DEL} \cdot 100\% \qquad (1)$$

The test duration is given in hours, thus the insertion rate *IR* results in INS / hour.

$$IR = \frac{INS}{test\_duration} \qquad (2)$$

5.2. Results

Though it was the original intention to determine the recognition rates of the test objects, the results had to be considered from many points of view. This was necessary in order to separate influences from different environmental conditions. Figures 5 to 7 demonstrate examples of typical results. The recognition rates varied in a wide range depending on background noise and speaker position. Only in silence and in position a) (close to the microphone) the recognizers achieved recognition rates comparable to those claimed in data sheets or literature. A dramatic decrease was found for far speaker positions and for speechlike noise (radio). This behaviour was common to all test objects.

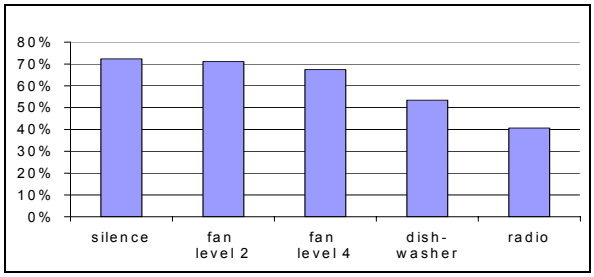

Figure 5: Dependency of the recognition rate on the background noise (recognizer 7, average of all speaker positions)

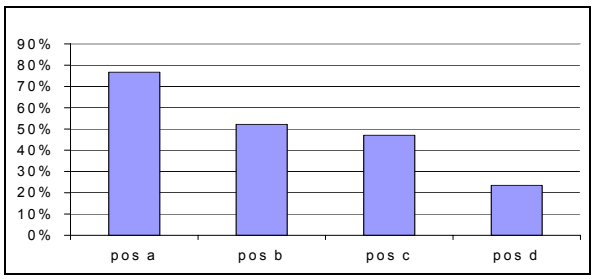

Figure 6: Dependency of the recognition rate on the speaker position according to Figure 2 (recognizer 6, average of all noise conditions)

The insertion rates were considerably different for the recognizers (Figure 7). The best test object had only 1.7 insertions per hour, the worst 141. There was a tendency towards a coherence of high recognition rates and high insertion rates.

The answer delay differed between the test objects. In the answer slot of ca. 4…5 seconds (Figure 4) the fastest test object was in average 1.7 seconds better than the slowest.

A little bit surprising was the fact, that state of the art embedded systems, which were adapted to the application, achieved recognition rates in the same range like the PC based systems.

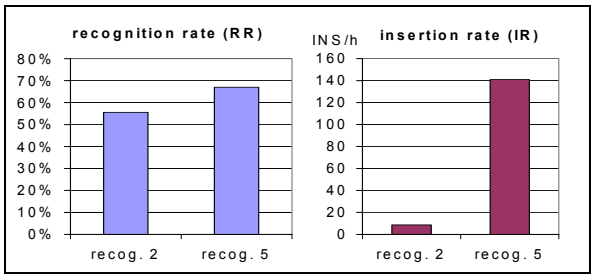

Figure 7: Comparison of two pairs of recognition and insertion rates (recognizer 2 and 5)

6. Conclusions and future outlook

From the technical side we found a big gap between recognition rates coming from measurements in an experimental environment and in reality.

The acceptable recognition rate of more than 85% stated in the usability experiment was never reached in real environments. This is mainly due to the transmission loss between the source of speech and the microphone. There is no lossless way for a later electronical repair of the disturbed signal, since there are many different signal- and room situations and there is no sufficient knowledge of all the kinds of interfering noises.

Therefor, effort will be spent to bring microphone and speaker together more tightly or to find a solution for speaker isolation (microphone arrays). Speaker presence can be detected by additional sensors (PTT-strategies). All these approaches will be subject of future activities.

7. References

[1] R. G. Leonard, “A database for speaker-independent digit recognition," in Conference Proc. IEEE ICASSP, 1984, pp. 42.11.1-42.11.4.

[2] Hansen, J.H.L., et. al. : "CU-Move : Analysis & Corpus development for Interactive In-vehicle Speech systems", Proceedings of EUROSPEECH 2001, Aalborg, 2001.

[3] Moreno, A., Lindberg, B., Draxler, C. et. al.: “SPEECHDAT-CAR. A large speech database for automotive environments”, in Proceedings of LREC2000, Athens, May 2000.

[4] Siegmund, R., Höge, H., Kunzmann, S., Marasek, K.: “SPEECON – Speech data for Consumer Devices”, Proceedings of LREC2000, Athens, May 2000.

[5] Gong, Y., “Speech recognition in noisy environments: A survey” *Speech Comm.,* (16):261-291, 1995.

[6] Zechner K., Waibel A., “Minimizing Word Error Rate in Textual Summaries of Spoken Language”, *Proceedings NAACL-ANLP-2000*, Seattle, WA, May, pp.186-193.

[7] Sivadas1, S., Jain1and, P., Hermansky, H.: “Discriminative MLPS in HMM-Based Recognition of Speech in Cellular Telephony”, Proceedings of ICSLP 2000, Beijing, Oct. 2000.

[8] Pearce, D., Hirsch, H.: “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions”, Proceedings of ICSLP 2000, Beijing, Oct. 2000.