Speech Input and Output Module Assessment for Remote Access to a Smart-Home Spoken Dialog System

Jan Kребер, Sebastian Möller, Alexander Raake

Institute of Communication Acoustics
Ruhr University Bochum, Germany
(jan.krebber sebastian.moeller alexander.raake)@ruhr-uni-bochum.de

Abstract

This paper presents the assessment of the speech input and output modules of a spoken dialog system which is accessed via a telephone connection. The spoken dialog system is being developed under the European INSPIRE project. It is used to control a smart home environment. The mentioned modules are assessed separately, with the help of a remote access simulation tool to model degradations which are typical for today’s telephone lines. This tool generates impairments in a controlled way, in order to quantify their impact on recognition performance and on speech output quality. First results for the effects of coding impairments on the input and output modules of the INSPIRE dialog system are presented and discussed.

1. Introduction

The core of the INSPIRE smart home system is a spoken dialog system. With the help of the system, a user can operate different home appliances (e.g. lamps, blinds, fan, answering machine, TV, VCR, electronic program guide) via speech. The spoken dialog system can be controlled either from inside the house or by remote access via a telephone connection. In this way, it is possible to control the home appliances from far distances, e.g. from the office or while driving a car.

The impairments which are introduced by the transmission in a wired or mobile phone network may lead to a considerable degradation of recognition performance, and consequently to problems in speech understanding and dialog flow. On the other side, the channel transfer characteristics will also impact the quality of the speech output. Both effects have to be taken into account for estimating the overall quality of the spoken dialog system, as they influence the conversation quality, system usability, and finally the acceptability of the voice operated system.

The transmission impairments found nowadays in analogous and digital telephone networks can be of different origin, like impedance mismatch, acoustic echo, low bit-rate codecs, or long distances for a call. These sources result in degradations like loss, circuit noise, quantizing and non-linear distortion, echo, and delay. During the last years, Voice over Internet Protocol (VoIP) has hit the market as an alternative to circuit-switched telephone networks. The VoIP technique introduces time-variant impairments like changing delay times or speech packet loss.

Besides of impairments introduced by the network, user interfaces (terminal equipment) nowadays can vary quite extensively, as there is no standard interface defined by telephone operators. Although traditional handsets telephones are still in use, mobile phones with bad acoustic characteristics, hands-free terminals and headsets become more and more popular. In a real telephone connection a combination of all such impairments will occur; consequently the joint effects have to be taken into account when the impact of the transmission channel on system performance has to be estimated [1].

This paper presents an approach for efficient assessment of the speech input and output modules of the INSPIRE system, for a remote-access situation. A real-time remote access simulation tool was developed at IKA [2][3]. It allows all relevant impairments to be generated in a defined way. The input parameters of the tool are planning values of traditional as well as of new (mobile or IP-based) networks. Section 2 describes the setup and the characteristics of the simulation system in more detail. The simulation system is then used to investigate the effects of transmission degradations on the speech recognition module performance and on the perceived quality of the speech output module. The experimental set-up and the results are presented in Section 3. In Section 4, a final discussion and an outlook on further work is presented.

2. Remote access simulation tool

The requirements for the remote access simulation tool have been defined as follows:

- Generation of impairments in a way which is realistic for (nearly) all kinds of telephone connections,
- Reproducibility of all transmission characteristics,
- Re-usability of the tool for the final evaluation of the entire dialog system,
- Extensibility towards new types of impairments caused by new transmission techniques.

As a result of these requirements, the tool has been specified to show the following features:

- The network topology has to contain all relevant impairments,
- All relevant transmission parameters have to be adjustable in a controlled way, and within the expected range of parameter values found in reality, and
- The tool has to show real-time capabilities.

The remote access simulation tool is based on a network topology proposed by the International Telecommunication Union (ITU-T) for estimating the end-to-end transmission performance of a network [4]. This topology considers instrumentally measurable characteristics of the acoustical or electrical paths of the transmission, in terms of so-called loudness ratings, i.e. scalar planning values which reflect to a certain extent the sensitivity of the human ear. The same
principle is adopted to different noise sources, which are represented by psophometrically or A-weighted power levels. To fulfill the given requirements, a programmable DSP hardware was chosen for model implementation. It operates in real-time, thus allows realistic man-machine interaction as well as off-line data processing. The implemented structure is shown in Figure 1. The triangles symbolize FIR filters or programmable attenuators. The rectangles represent delay lines \((T, T_a, \text{ and } T_r)\), codecs, the VoIP transmission path, and the channel band-pass filter. The following telephone line degradations can be simulated:

- Attenuation and frequency distortion of the transmission channel (send loudness ratings, SLR and receive loudness ratings, RLR)
- Continuous white circuit noise representing all possible noise sources, both on the channel \((N_c, \text{ band-pass filtered afterwards)}\) and at the receive side \((N_f,\text{ low-pass limited by the transmission characteristic of the user interface at the receiving side})\)
- Transmission channel bandwidth limitation \((\text{BP with 300 - 3400 Hz for narrow-band, or 50 - 7000 Hz for wide-band application})\)
- Absolute one-way delay \((T_a)\)
- Talker echo with round-trip delay \(2T\) and attenuation \(Le\)
- Listener echo with round-trip delay \(2T\) and attenuation \(WELP\)
- Listener sidetone with attenuation \(Lst\)
- Different speech codecs
- Sequences of lost packets, generated according to a specified loss distribution, and with a specified packet length \((40 \text{ ms for VoIP})\)

From Figure 1, it can be seen that all relevant speech paths (main transmission path, the talker sidetone path, talker echo path and listener echo path) are implemented in the remote access simulation tool. The structure allows a full duplex operation and an independent setting of parameters for all paths, thus unsymmetrical settings for both call directions. The terminal equipment is connected via line input and output junctions. For the purpose of the structure in Figure 1, this equipment can be described by the sensitivity in the sending direction (microphone) and the receiving direction (loudspeaker), expressed in loudness ratings, namely the loudness rating of the sending user interface \((SLRut)\) and the one of the receiving user interface \((RLRut)\), respectively. They represent the weighted attenuation on the line relative to a reference channel. In addition, each user interface is characterized by its difference in sensitivity between direct (speech) sound and diffuse (e.g. ambient noise) sound. Additional filters \((SLR, RLR)\) complete the sending \((SLR = SLRut * SLR)\) and receiving \((RLR = RLRut * RLR)\) characteristics. They can be adjusted to achieve a desired frequency response of the overall SLR and RLR.

3. Speech input and output module assessment

The described remote access simulation tool is now being used to assess the transmission impact on the speech input and output modules of the INSPIRE dialog system.

3.1. Speech input module assessment

3.1.1. Test setup

As shown in Figure 1 with dashed lines, the speech input module (ASR) is connected at the right side of the simulation tool. It is fed with speech data processed offline, by playing back pre-recorded utterances via a talking head/torso simulator (HATS), picking up the sound with a user interface (see left side of the simulation tool in Figure 1), processing it
through the tool, and re-recording it at the receive side (see right side, Figure 1).

The source utterances were generated on the basis of interaction scenarios with the smart home system. These scenarios have been read aloud by 10 native German speakers (9m, 5f) in a low-noise test cabinet, using a high-quality microphone. Overall, 1370 utterances were available for the test.

The 1370 source files have been processed by the remote access simulation tool with 24 different parameter settings, leading to 32,880 processed speech files. These files were analyzed by a commercial continuous speech recognizer for German (Loquendo Speech Suite). Because the same source files have been used for each transmission channel setting, the differences between the ASR results mainly depend on the impairments of the telephone transmission, and only to a minor extent on the characteristics of the (limited) source speech material.

3.1.2. Experimental results

The experiments took place in the initial phase of the system development, and hence the ASR for German language was not optimized. The low performance (68.7% word accuracy, WA) on the clean and unprocessed data is mainly caused by the absence of grammars which are still under development. For these tests, we are not interested in absolute recognition performance, but in relative recognition performance. Because of that, the missing grammars are tolerable.

![Figure 2: Word accuracy for different speech codecs and basic set-ups](image)

The transmission of the speech files via the HATS and a standard handset yields to a loss of the WA of approximately 14%, as the WA drops to 54.7%. Surprisingly, the use of logarithmic PCM coding (G.711) and a reference telephone line introduce an additional loss of 2% of the WA. The overall loss of 16% for the WA for the ideal telephone transmission line is unexpectedly high, as the Loquendo ASR is supposed to be optimized for telephone speech. As shown in Figure 2, there is no remarkable difference between the reference codec G.711 (52.8% WA) and the ADPCM codec G.726 (52.1% WA) with 32 kBit/s. Both codecs are used in ISDN networks. In mobile or VoIP networks, low bit-rate codecs such as the GSM-EFR or G729(A), respectively, are commonly used. A decrease of as much as 10% WA (GSM-EFR 42.4% WA, G.729 42.7% WA) for these codecs relative to the wireline codecs G.711 and G.726 was observed.

3.2. Speech output module assessment

3.2.1. Test setup

The dotted lines in Figure 1 show the processing of the speech data for the speech output module assessment. On the left side of the simulation tool, either naturally-produced or synthesized speech prompts were used as the source material. These files were processed through the tool and played back to the listening subjects via one of different user interfaces (handset telephone, hands-free terminal, headset).

The source prompts were taken from the vocabulary of the INSPIRE dialog system. The synthesized text-to-speech prompts were recorded from the AT&T speech synthesizer (AT&T Natural Voices, male voice “Rainer”). The human prompts were recorded from a male speaker. In former tests, this male speaker turned out to be best-rated in comparison to several other (male and female) speakers.

3.2.2. Rating procedure

The participants were asked to listen to each processed stimulus and to rate four characteristics on continuous rating scales, as described in [5]. Each of the scales is introduced by a question and labeled with two or five describing adjectives (translated from German):

- What is the overall impression of what you just heard? (excellent; good; fair; poor; bad)
- Which effort was necessary to understand the meaning of the utterance? (complete relaxation possible; no effort required; attention necessary; no appreciable effort required; moderate effort required; considerable effort required; no meaning understood with any feasible effort)
- How pleasant was the voice you just heard? (pleasant; unpleasant)
- How well does the voice fit to the described system? (excellent; good; fair; poor; bad)

The ratings were calculated on a scale from 0 to 6, 5 corresponding to the most positive label of the inner (thick) part of the scale, and 1 corresponding to the most negative label of that part. The mean rating of the overall impression will be abbreviated MOS (mean opinion score) in the following description.

3.2.3. Test subjects

24 subjects participated the test. Their age ranged between 20 and 70 years, with an mean of 31.6 years. 12 of them (50%) had experience with spoken dialog systems, only 2 (8%) with dictation systems, and 13 (54%) with synthesized speech. The participants were paid for their attendance.

3.2.4. Experimental results

As depicted in Figure 3, the medium and low bit-rate coding of the GSM-EFR, G.726 (32kBit/s), and G.729 results in a distortion effect and leads to a remarkable difference to the G.711 reference case, especially for synthetic speech. Up to
certain extent, the rating of the impairment of different codecs on human speech prompts reflects the rating by the impairment factors $I_e$, see Table 1.

![Figure 3: Mean overall quality ratings for different speech codecs.](image)

The impairment factors are established by the ITU-T and are recommended in ITU-T Rec. G.113 App.1 [6]. They describe the degradation of quality for human-to-human conversations over a telephone line due to coding. The point of reference is given by the G.711 codec.

### Table 1: Impairment factors $I_e$ and MOS ratings for human and synthesized speech prompts for different types of codecs.

<table>
<thead>
<tr>
<th>Codec Type</th>
<th>Impairment Factor $I_e$</th>
<th>human MOS</th>
<th>synth. MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.711</td>
<td>0</td>
<td>4.1</td>
<td>3.5</td>
</tr>
<tr>
<td>GSM-EFR</td>
<td>5</td>
<td>3.9</td>
<td>2.7</td>
</tr>
<tr>
<td>G.726</td>
<td>7</td>
<td>3.8</td>
<td>2.6</td>
</tr>
<tr>
<td>G.729/A</td>
<td>11</td>
<td>3.8</td>
<td>2.9</td>
</tr>
</tbody>
</table>

The main outcome of the experiment with synthesized speech is the perceived difference between the log PCM coding algorithm and the low and medium bit-rate codecs GSM-EFR, G.726 and G.729. Obviously, the perceptual features of the synthesis are overlaid by those introduced by the coding, leading to a large drop in quality.

### 4. Discussion

For the INSPIRE ASR tests, the VoIP codecs (GSM-EFR, G.729) lead to a higher loss of recognition performance than standard ISDN codecs (G.711, G.726). So far, the access by VoIP to the INSPIRE smart-home system needs to be assessed again, if the grammar is capable of catching up 10% WA difference of the ASR between ISDN and VoIP connections.

In contrast to the ASR assessment, the G.726 ADPCM codec is rated worse for synthesized speech than the other codecs. The overall impact of speech coding at medium and low bit-rates seems to be larger for the synthesized than for the naturally produced voice. This observation is somehow in contrast to findings reported in earlier investigations [7]. In that study, speech codecs seemed to introduce an "artificiality" dimension which did not show a strong impact on synthesized speech, but only on naturally produced one. It has to be noted that the former study used diphone synthesis instead of the unit selection approach used here. Thus the difference may be linked to the specific synthesizer being assessed.

Further experiments will investigate the impairment of time variant distortions like packet loss for VoIP, or the combination of several impairments on the ASR and synthesized speech as well as the overall quality of the INSPIRE spoken dialog system.

### 5. Acknowledgements

The presented work has been carried out in the frame of the EC-funded IST project INSPIRE (IST-2001-32746, see inspire-project.org). The authors would like to thank Alex Trunov (EPFL) for computing the word accuracy, and Rosi Pegam and Anders Krosh for their support in the experiments.

### 6. References