Performance of Speech Recognition and Synthesis in Packet-Based Networks

Sebastian Möller, Jan Kreber and Alexander Raake

Institut für Kommunikationsakustik (IKA)
Ruhr-Universität Bochum, Germany
sebastian.moeller@ruhr-uni-bochum.de

1. Introduction

In modern speech communication scenarios, speech is usually transmitted over a variety of networks differing with respect to the underlying technology, e.g., wireless analogue or digital networks, mobile GSM networks, or packet-based networks (Voice over Internet Protocol, VoIP). The different technologies give rise to a multitude of degradations of the transmitted speech. These degradations affect not only the communication between humans, but also the interaction between a human user and a spoken dialogue system – a scenario which increasingly gains importance in such networks.

The degradations which are introduced by packet-based transmission differ considerably from the ones found in traditional telephony. In particular, they show a strongly time-varying character, resulting from the loss of packets during the transmission. Packet loss is mainly due to the absence of prioritization in common IP-based networks, leading to different transmission times for individual packets (delay jitter). Speech transmission, however, requires a continuous data stream – a requirement which is contradictory to the asynchronous transmission enabled by a packet-based network. In order to compensate for a part of the transmission delay differences, a jitter buffer is typically applied at the receiver's side. On the one hand, this buffer should not be too long in order to avoid an excessive overall delay of the transmitted speech signal; on the other hand, packets arriving after the maximum delay corresponding to the length of the jitter buffer have to be discarded, leading to a loss of packets.

The perceptual effects of packet loss have largely been investigated for speech communication between humans, see e.g., [1] and [2]. It has been shown that the quality degradation depends not only on the percentage of lost packets, but also on the type of codec used before packetization, and on the concealment strategy applied to lost packets. For example, simple silence insertion is usually more annoying to the human ear than more refined strategies like the (pitch-synchronous) repetition of previous packets. In addition, the distribution of lost packets over time has a considerable effect on the perceived annoyance. Usually, packets get lost in “bursts” during a temporary overload of the network, and during other phases the packet loss is much less severe.

So far, little attention has been paid to the impact that packet loss exerts on speech directed to or originating from a spoken dialogue system (except recognition-related studies reported in [3] and [4]). This impact may, however, be decisive for the performance of spoken dialogue systems. Two modules of a spoken dialogue system are directly linked to the speech signal: The speech recognizer which is confronted with speech data corrupted by packet loss, and the speech synthesizer which generates a speech signal suffering from packet loss on its further way to the human listener.

It is the aim of the present study to quantify the impact of packet loss on (1) the performance of a speech recognizer and (2) on the quality of speech generated with a text-to-speech synthesizer. An online transmission simulation tool has been set up for generating realistic packet-loss impairments in a controlled way, see Section 2. With the help of the tool, speech data has been processed for an application system which allows voice control of smart home devices, developed under the EC-funded IST-project INSPIRE ("Intelligent management with Speech Interaction via Remote microphones and telephone interfaces"). Details on the experimental set-up are given in Section 3. Both speech recognition performance and synthesized speech quality are compared to the quality degradation which is expected for a communication scenario between two human users, when conversing over the same transmission channel. The results are analyzed in Section 4, and they show some differences between the two situations which are discussed in Section 5.

2. Transmission simulation tool

The tool simulating packet-based transmission is part of a larger simulation environment which has been set up at IKA in order to study the perceptive effects of speech transmission in an interconnected network environment. It consists of a simulation for wireline networks including their terminals, of mobile (e.g., GSM) networks, and of packet-based networks which are part of the overall transmission path. Details on the entire simulation environment can be found in [5] and [6].

For simulating VoIP transmission, the speech signal first has to be coded into a logarithmic PCM format (ITU-T Rec. G.711), which is the format usually required by gateways. After a voice activity detection (disabled for the present experiment), the speech signal is further transcoded and encapsulated into packets in a commercial gateway (Cisco 2611). The IP packets are then "transmitted" over the network.
emulation program NIST Net [7], which discards packets in a random way with a pre-defined loss probability. This results in a gaussian loss distribution. The remaining packets are then collected by the jitter buffer on the outgoing side of the gateway, the payload is extracted from the packet, and the signal is decoded into log. PCM format. Finally, the analogue signal is re-generated from the PCM waveform. The situation is depicted in Figure 1.

![Diagram of VoIP simulation process]

**Figure 1:** Structure of the VoIP simulation tool.

The input to the VoIP simulation is formed by signals which are collected with a typical user interface (usually a telephone handset, equalized to show an IRSmod-characteristic according to ITU-T Rec. P.830) and band-limited to 300-3400 Hz. To this signal, a circuit noise floor of ~70 dBm/Hz is added to reach realistic network conditions. At the output of the VoIP simulation, broad-band noise of level ~64 dBm/Hz is added, simulating the noises at the receiver side of the transmission line. These levels are defined as default values in [8]. The signals are finally played back to a human user via a telephone handset, or they are fed into a speech recognizer.

### 3. Experimental set-up

It is expected that the quality degradation due to packet loss which is perceived by a human listener is different from the performance degradation of an automatic speech recognizer (ASR) when being confronted with the same speech signals. Although speech intelligibility will be an important dimension of perceived quality, other dimensions like naturalness or speech sound quality may be as important for the human listener. Such dimensions may however be less decisive for the ASR, the sole task of which is to transcribe the contents of the received utterances. For the human listener, the perceived quality may additionally depend on whether the utterances stem from a human talker (e.g. with pre-recorded speech prompts) or from a speech synthesizer.

The aim of the investigations presented in the following paragraphs is to quantify and compare the degradations caused by packet loss for the following three situations:

1. A speech recognizer is confronted with human speech corrupted by packet loss
2. A human listener is confronted with human speech corrupted by packet loss
3. A human listener is confronted with synthesized speech corrupted by packet loss

All three situations are part of the INSPIRE scenario, where a human user controls different home appliances from a remote location (e.g. from his/her office) through a packet-based network. The following sub-sections describe the set-up of the speech recognition experiment (Section 3.1) and of the quality assessment by a human user, confronted either with naturally-produced or with synthesized speech (Section 3.2). The results are analyzed in Section 4.

#### 3.1. Speech recognition under packet loss

The ASR experiment has been carried out in two languages (German and Greek), with two different recognizers. For German, a number of typical user utterances have been generated from 10 dialogue templates, containing a total of 137 utterances. The dialogues have been read aloud by 10 native German speakers (5m, 5f) and recorded in a low-noise test cabinet, using a high-quality AKG C414-B-U.SS microphone. For Greek, 10 prototype dialogues, each containing 5 utterances, were read aloud by 10 native speakers (5m, 5f) and recorded. The resulting 1370+500 utterances have been played back through a head-end-torus simulator (Head-Acoustics HMS IL3) and re-recorded via a standard telephone handset (German) and a KONTRON 3911, in a standard office environment (dimensions 3.65m x 5.65m x 3.50m, reverberation time < 0.5 s in the frequency range of speech).

The speech files recorded in this way form the input to the VoIP simulation set-up, as described in Section 2. Two transcoding algorithms were selected: Either log. PCM with 64 kbit/s according to ITU-T Rec. G.711 with proprietary packet loss concealment (PLC), or a parametric CS-ACELP coding with 8 kbit/s according to ITU-T Rec. G.729A with built-in PLC. Both codecs are commonly used in VoIP applications. On the coded speech data, NIST Net generated random packet loss with a probability between 0 and 15%. Voice activity detection was disabled for the experiment.

The transmitted speech data was fed into one of two commercial speech recognizers. Because the experiments took place in the initial phase of system development, the German recognizer was not yet optimized; the topline recognition performance on the clean (unprocessed) data only reached 69% for this recognizer. This low performance will be mainly due to the absence of grammars which are still under development. As we are only interested in the relative degradation of recognition performance and not in absolute performance numbers, this fact is tolerable for the given purpose. For the Greek ASR, the topline performance reached 98.7%, using a grammar and dialogue-node-specific vocabularies.

#### 3.2. Quality of naturally-produced and synthesized speech under packet loss

In order to assess the quality of speech output transmitted over the packet-based network, an auditory listening-only experiment with 24 test subjects was carried out (20-70 years, mean 31.6 years; about half of the test subjects had prior experience with spoken dialogue systems). The stimuli consisted of typical prompts foreseen for the INSPIRE system. They contained either only static information (pre-defined according to the applications covered by the system), or they provided dynamic information which has to be generated on-line during system operation (e.g. TV program information, messages from the answering machine). The static prompts were pre-recorded with a male German speaker which turned out to be optimum amongst a selection of different voices. The dynamic prompts consist of synthesized speech, generated with a commercial TTS system.
concatenating units of different length (AT&T Natural Voices, male voice “Rainer”).

All prompts have been transmitted through the VoIP simulation in the way described above, with the same transmission characteristics (using only the G.729A codec). They were played back to the listening subjects in a quiet office environment through a telephone handset. The test subjects first had to indicate on a test sheet the device a particular speech prompt referred to; in this way, the subjects had to concentrate on the contents of the prompt (information to be conveyed) and not just on its surface form. Then, the subjects were asked to rate four aspects of system quality on continuous scales, labeled with two or five attributes, see ITU-T Rec. P.851 [9]:

- What is your 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)

Mean ratings over all subjects 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 ratings will be abbreviated with the terms “overall quality”, “listening-effort”, “voice pleasantness” and “voice adequacy”.

4. Experimental results and discussion

The results will now be analyzed with respect to the influence of packet loss in the three situations mentioned in Section 3, namely the ASR being confronted with corrupted speech, and a human listener being confronted either with naturally-produced or with synthesized speech corrupted by packet loss.

For the human listener judging the quality of naturally-produced speech degraded by packet loss, prediction models are available which estimate the mean overall quality rating on a five-point scale, on the basis of the parameters used by the VoIP simulation set-up. The most well known model for this purpose is the E-model described in ITU-T Rec. G.107 [8]. This model will be used in comparison to the subjective quality judgments.

4.1. Impact on speech recognition performance

Figure 2 shows the word accuracy degradation of the German ASR under packet loss, for the G.711 and the G.729A codec. In both cases, word accuracy degrades almost linearly with the percentage of lost packets by approximately 1% for each percent of lost packets. This rule of thumb holds true for both speech codecs, up to 15% packet loss. It is similar to the values observed in [3] for the ITU-T Rec. G.723 coder.

The Greek ASR seems to be less sensitive to packet loss, see Figure 3. With the G.711 codec, even 15% of lost packets only lead to a degradation of 1.5% in word accuracy. The impact is stronger for the G.729A codec, but still not comparable to the German ASR.

4.2. Impact on perceived speech quality

The overall quality ratings obtained for naturally-produced and synthesized prompts are depicted in Figure 4. In both cases, speech quality degrades more severely when the packet loss rate exceeds a certain threshold. For the natural voice, this threshold is around 5% packet loss, for the synthesized voice slightly lower. Above the threshold, perceived speech quality degrades quite strongly, by approximately 1 category with 15% when compared to the threshold value.

Figure 4: Overall quality degradation for naturally-produced and synthesized voices.

The overall quality ratings can be compared to the predictions of the E-model, cf. the dashed line in Figure 4. It
turns out that the E-model predictions are in relatively good agreement with the auditory test results for higher packet loss rates. However, there is a shift between the auditory rating curves and the model predictions. This shift is due to the fact that the absolute quality of the stimuli had to be rated by the test subjects, and not the relative quality degradation. The shift is related to the circuit and voice characteristics of the specific test; it turns out to be positive for the naturally-produced speech samples and negative for the synthesized samples. For low packet loss rates, the nearly flat curves are not well predicted by the model.

The picture is very similar for the other subjective ratings. As an example, Figure 5 shows the mean listening-effort ratings. Once again, a slight "threshold effect" can be observed: Up to a certain level of packet loss listening-effort degrades only by a small amount, and beyond this level the degradation is much stronger. It has to be noted that such a "threshold effect" has not been observed in the experiments reported earlier, e.g., in [1].

![Figure 5: Listening-effort ratings for naturally-produced and synthesized voices.](image)

When comparing the degradations on naturally-produced vs. synthesized voice, it seems that the natural voice is slightly more robust towards packet loss, in the sense that a more severe quality degradation starts at a higher loss rate. An explanation might be that synthesized voice contains less redundancy, which makes it more vulnerable in case of non-stationary degradations than naturally-produced speech. The finding is in contrast to the one for stationary degradations like continuous circuit noise, where synthesized speech was shown to be less vulnerable than naturally-produced one [10].

5. Conclusions

The experimental data show that the effects of packet loss on ASR performance differ from the ones on speech quality perceived by a human listener. Whereas ASR performance (measured in terms of word accuracy) degrades almost linearly with the percentage of lost packets, speech quality perceived by humans remains nearly constant up to a certain percentage of lost packets above this threshold, the degradation is much stronger. For the synthesized voice used in the auditory test (from a unit-selection synthesizer), the threshold is slightly lower than for the naturally-produced voice. A potential explanation may be that the synthesized voice contains less redundancy, which may render it more vulnerable in case of time-variant degradations — a hypothesis which requires a more thorough validation. It was shown that the E-model is able to predict the perceptive effects of packet loss, at least for higher loss rates.

So far, packet loss degradations have been generated in a randomized way, which is unrealistic for real-life networks. Further investigations will be necessary to analyze and potentially predict the effects of even stronger time-variant loss characteristics. It is expected that the type of codec and the applied loss concealment strategy will carry a strong influence on how such packet loss affects both ASR performance and the quality of synthesized speech.

6. Acknowledgements

The present work has been performed at IKA in the frame of the EC-funded IST-project INSPIRE (IST-2001-32746; see www.inspire-project.org). The authors would like to thank all INSPIRE partners for their support, in particular Alex Truin (EPFL) and Anastas Vovos (Knowledge SA) for processing the speech recognition results.

7. References