Diagnostic Assessment of Telephone Transmission Impact on ASR Performance and Human-to-Human Speech Quality

Sebastian Möller* and Ergina Kavallieratou†

* Institute of Communication Acoustics
Ruhr-University, 44780 Bochum, Germany
moeller@ika.ruhr-uni-bochum.de

†Wire Communications Lab
University of Patras, 26500 Patras, Greece
ergina@wcl.ee.upatras.gr

Abstract

This paper addresses the transmission channel impact on human-to-human speech communication quality as well as on ASR performance. Transmission channels include standard wireline or mobile telephone networks and IP-based networks, which can be operated via different types of user interfaces. In order to gain control over the transmission channel, a simulation model is developed. It implements all types of stationary impairments which can be found in the mentioned networks. Human-to-human speech communication quality in these situations is estimated using a network planning model. Experiments are carried out for assessing ASR performance over the same channel, with three different types of recognizers: two prototypical recognizers used in a telephone-based information server, and a standardized set-up developed under the AURORA framework for distributed ASR. It turns out that some interesting differences exist in behavior between the ASR system performance and speech quality in human-to-human communication. The differences should be taken into account by both developers of ASR systems and transmission network planners.

1. Introduction

Spoken dialogue systems are often accessed remotely over telecommunication networks, such as traditional analogue/digital telephones, mobile phones, or IP-based networks (voice over Internet Protocol, VoIP). In all of these cases, transmission channel degradations can have a severe influence on ASR performance, and subsequently on speech understanding performance and on overall dialogue system quality. Depending on the type and the characteristics of the transmission channel, the degradations are very diverse in magnitude and nature.

Some of these impairments have been investigated in detail with respect to their impact on speech recognizer performance, see e.g. the work performed by Euler and Zinke (1994), Lilly and Paliwal (1996), or Tucker et al. (1999). The investigations aim to develop recognition systems which are robust towards the specific impairment, e.g. by using preprocessing and adaptation techniques (Mokbel et al., 1993; Mokbel et al., 1997), or by training acoustic models with impaired speech data (e.g. Puel and André-Obrecht, 1997). Robust HMM architectures have also been proposed, e.g. for impairments which are to be encountered in GSM cellular networks (interruptions and impulsive noise) by Karray et al. (1998).

Unfortunately, there are nearly no possibilities to control the exact characteristics of an individual transmission line in operating networks. In principle, a combination of different degradations is to be expected in this situation. Thus, components of spoken dialogue systems should be assessed with respect to their robustness against this combination of degradations. This can be done efficiently by means of a simulation tool, as it has been proposed earlier (Möller and Bourlard, 2000; Möller and Bourlard, 2002). Such a tool allows all characteristics of the transmission channel to be generated in a controlled way. In this way, a diagnostic evaluation of the effects of specific impairments (e.g. of new speech codecs) on ASR becomes possible.

Speech transmission networks are normally set up to fulfill the quality requirements of human-to-human speech communication. It is therefore interesting to compare the ASR performance degradation to the quality degradation which occurs when humans converse over the same transmission channel. The comparison helps to decide whether the quality requirements according to which telecommunication networks are designed and set up, and which purely reflect the human-to-human dialogue, are also applicable to human-machine interaction.

In the present paper, we perform a comparative evaluation of three recognizers with respect to their sensitivity to transmission channel degradations. The degradations are generated by our simulation tool (see Section 2), which makes us independent of uncontrolled real-life networks. They include narrow-band and wide-band uncorrelated noise, signal-correlated noise, linear frequency distortions, and non-linear codec distortions. ASR performance degradation is compared to the degradation in speech quality between humans, as it is predicted by a network planning model (Section 3). This comparison is described in Section 5, and reveals some interesting differences in behavior between the ASR system performance and speech quality in human-to-human communication. The differences should be taken into account by both developers of ASR systems and transmission network planners.

2. Transmission Channel Simulation

The use of simulation techniques, in general terms, is not new in the development of ASR systems. E.g., Tarcisio et al. (1999) simulate the transmission channel by filtering with a measured impulse response and adding recorded background noise. A similar technique has been proposed by Giuliani et al. (1999) for modeling hands-free terminals. When artificially degraded data was included in

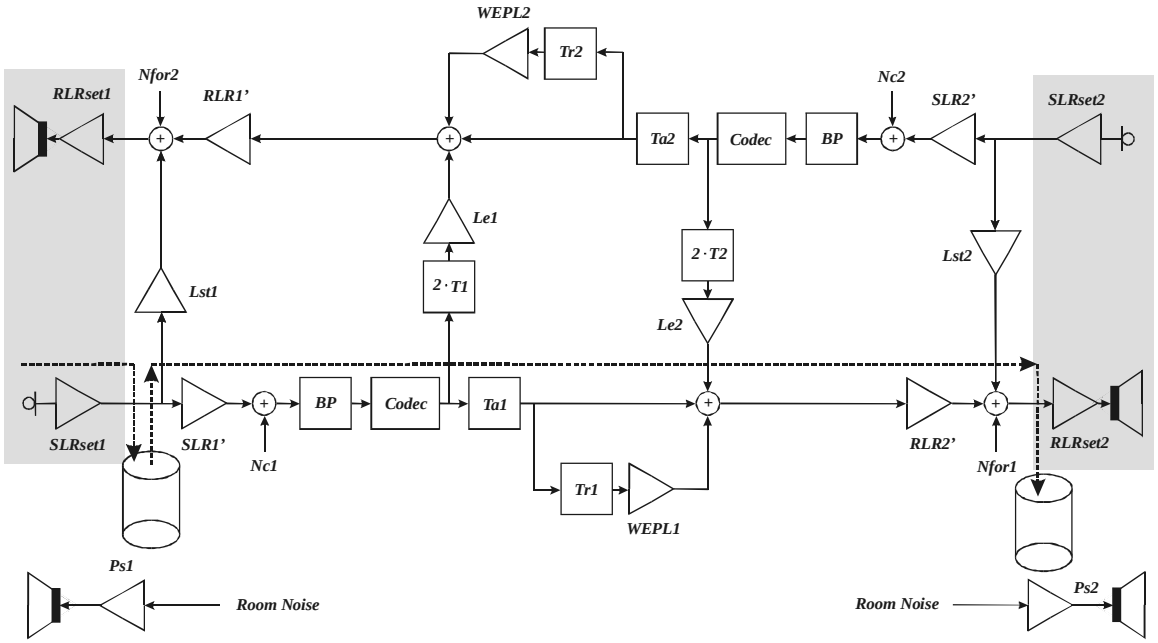

Figure 1. Transmission channel simulation.

the training material, the ASR performance improved significantly. The simulation of time-variant channel behavior (mobile GSM channels, ATM channels, voice over IP) has also been proposed, and partly been used for assessing ASR performance (e.g. in the ETSI STQ AURORA DSR working group).

In contrast to measuring specific channel characteristics, we base our simulation on planning values which are available already in the network planning phase, before a network has actually been set up. Such planning values are commonly used by telecommunication engineers. Due to their simplified nature, they only give a rough description of the channels involved in the transmission (direct speech path, talker and listener echo paths, sidetone path due to the coupling of one's own voice), e.g. via a frequency-weighted one-dimensional attenuation index (so-called loudness rating) and a corresponding mean delay time. Specifications for each of these paths can be found in the respective Recommendations given by the Telecommunication Standardization Sector of the International Telecommunication Union, ITU-T.

We chose a reference connection which is recommended by the ITU-T for estimating the transmission channel impact on the overall quality of the connection, mouth-to-ear, and in a conversational situation. This reference connection is given in ITU-T Rec. G.107 (2000). The reference connection has been transformed into the simulation model which is depicted in Figure 1, see Möller and Bourlard (2002). It includes all transmission paths and implements most of the time-invariant degradations occurring on these paths, namely:

- the attenuation and linear frequency distortion of the channel, both at the send (*SLR*) and receive (*RLR*) side; they partly stem from the acoustic-electrical conversion (*SLRset*, *RLRset*), and partly from the purely electrical paths (*SLR'*, *RLR'*)
- the channel bandwidth limitation (approx. 300-3400 Hz in the narrow-band case, and 50-7000 Hz in the wideband case)
- continuous (white) circuit noise as a model for all potential noise sources, both on the channel (*Nc*, narrow-band) and at the receive side (*Nfor*, wideband)
- non-linear speech coder-decoder pairs (several codecs standardized by the ITU-T or ETSI and North American codecs have been implemented so far)
- ambient room noise at the send (*Ps*) and receive (*Pr*) side (modeled by inserting different types of noise in the send and receive room, in order to include speaking style variations)
- pure delay of the connection (*Ta*)
- delay (*T*) and attenuation (*Le*) of the talker echo signal reaching the talker's ear
- average delay (*Tr*) and attenuation (*WEPL*) of multiply reflected signals reaching the listener's ear
- the attenuation and frequency distortion (*Lst*) of the sidetone path due to the coupling of the talker's own voice in the telephone handset

Figure 1 illustrates these elements with triangles indicating linear filters (*SLRset*, *SLR'*, *RLRset*, *RLR'*, *Le*, *Lst*, *WEPL*), and boxes indicating the channel bandpass filter (BP), the coder-decoder pair, or the delay lines (*T*, *Ta* and *Tr*). Indices 1 and 2 indicate the direction of the transmission. In the experiments described below, we are not interested in conversational features, but limit ourselves to the one-way transmission situation. For this reason, only one direction of the transmission model is used, and the pure delay, talker and listener echo elements are set to be without effect. Extensions of the simulation with respect to time-variant degradations (fading channels, IP packet loss) are underway, but they have not yet been addressed in our experiments.

The described transmission channel simulation can be used in different ways. E.g., the transmission channel impact on human conversation as well as on humans interacting with spoken dialogue systems over the phone can be investigated analytically. This is our aim for the present paper. On the other hand, the simulation permits to produce training material for speech recognizers which

shows defined transmission characteristics. In this way, existing training databases can be multiplied efficiently.

3. Human-to-Human Speech Quality

Speech communication quality between humans is a multi-dimensional feature (see e.g. Möller, 2000, for a discussion). It can ultimately only be assessed in realistic conversation scenarios, by performing auditory tests with human subjects. This is an expensive and time-consuming procedure, and therefore network planning experts make use of quality prediction models. Such models estimate speech communication quality on the basis of the above mentioned planning values.

The best-known and most complete network planning model is the so-called E-model (Johannesson, 1997), now recommended by the ITU-T in Rec. G.107 (2000). It predicts speech quality between humans in terms of a one-dimensional overall quality index (transmission rating factor, *R*, or a mean opinion score, MOS) as a function of the physical channel characteristics. The predicted MOS should ideally reflect the quality judgments of human test subjects – on a five-point absolute category rating scale – after conversing over a connection with the characteristics given by the planning values. Prediction accuracy of the model has been investigated in detail (Möller, 2000), and the predicted MOS values have been found to reach an agreement with subjective test results which is satisfying for the purpose of network planning.

As the input parameters of the transmission channel simulation (Figure 1) are mainly identical to the parameters the E-model bases its prediction on, we are able to perform a direct comparison between predicted speech quality for human-human conversation, and the recognition accuracy which can be reached over the same channel. The set-up for these experiments is described in the following section, and the results are discussed in Section 5.

4. Speech Data and Recognition Systems

The simulation model described in Section 2 allows us to realistically generate degradations occurring in real-life networks, using a schematic which is also underlying the E-model. The degradations are generated on pre-recorded clean speech files, so as to carry out recognition experiments on speech data which are identical with respect to language and speaker conditions. This, however, excludes to investigate the effects of background noise, where an adaptation of the speaking style occurs (e.g. due to the Lombard reflex). The test utterances were digitally recorded and then transmitted through the simulation model (cf. the dashed line in Figure 1). At the output of the simulator, the degraded utterances were collected and then processed by a recognizer.

Three different recognizers have been used for the experiments: The first one is a commercially available command-and-control recognizer for isolated words, which is part of a German dialogue system for accessing restaurant information. The second recognizer is part of a similar system in the same application domain, but for the Swiss-French language; it recognizes continuous speech using a standard n-gram language model. The third system is a more-or-less standardized HMM recognizer which has been defined in the framework of the ETSI AURORA project for distributed ASR in car environments. It has been built using the HTK toolkit and performs connected digit recognition for English. Training and test data for this system are available through ELRA (AURORA 1.0 database), whereas the German and the Swiss-French recognizer have been tested on specific speech data which stem from Wizard-of-Oz experiments in the restaurant information domain.

The Swiss-French recognizer is a large-vocabulary continuous system for the Swiss-French language. It makes use of a hybrid HMM/ANN architecture. ANN weights as well as HMM phone models and phone prior probabilities have been trained on the Swiss-French PolyPhone database (Chollet et al., 1996), using 4,293 prompted information service calls (2,407 female, 1,886 male speakers) collected over the Swiss telephone network. The recognizer's dictionary was built from 255 initial Wizard-of-Oz (WoZ) dialogue transcriptions on the restaurant information task. These dialogues have been carried out at IDIAP, Martigny, and EPFL, Lausanne, in the frame of the InfoVOX project. The same transcriptions were used to set up 2-gram and 3-gram language models. Log-RASTA feature coefficients (Hermansky, 1994) were used for the acoustic model, consisting of 12 MFCC coefficients, 12 derivatives, and the energy and energy derivatives. A 10th order LPC analysis and 17 critical band filters were used for the MFCC calculation.

The German recognizer is a partly commercially available small-vocabulary HMM recognizer for command and control applications. It can recognize connected words in a keyword-spotting mode. Acoustic models have been trained on speech recorded in a low-noise office environment and band-limited to 4 kHz. The dictionary has been adapted from the respective Swiss-French version, and contains 395 German words of the restaurant domain, including place names (which have been transcribed manually). Due to commercial reasons, no detailed information on the architecture and on the acoustic features and models of the recognizer is available to the authors.

The AURORA recognizer has been set up using the HTK software package (version 3.0; see HTK, 2000). Training and recognition parameters of this system have been defined in a way to compare recognition results when applying different feature extraction schemes (Hirsch and Pearce, 2000). Its task is the recognition of connected digit strings in English. Digits are modeled as whole-word HMMs with 16 states per word, simple left-to-right models without skips between states, and 3 Gaussian mixtures per state. Feature vectors consist of 12 cepstral coefficients and the logarithmic frame energy, plus their first and second order derivatives.

Because the German and the Swiss-French system are still in the prototype stage, test data is relatively restricted for these systems. We think that this is not a severe limitation, as we are only interested in the relative performance degradation, and not in absolute figures. The Swiss-French system was tested with 150 test utterances which were collected from 10 speakers (6m, 4f) in a quiet library environment ($Ps \sim 35$ dB(A)). 15 utterances that were comparable in dialogue structure (though not identical) to the WoZ transcriptions were solicited from each subject. Each contained at least two keyword specifiers, which are used in the speech understanding module of the dialogue system. Speakers were asked to read the utterances aloud in a natural way. The German

system was tested using recordings of 10 speakers (5m, 5f) which were made in a low-noise test cabinet (*Ps* ~ 35 dB(A)). Each speaker was asked to read the 395 German keywords of the recognizer's vocabulary in a natural way. All of them were part of the restaurant task context and were being used in the speech understanding module. In both cases recordings were made via a traditionally shaped wireline telephone handset. Training and test material for the AURORA system consisted of part of the AURORA 1.0 database which is available through ELRA. This system has been trained in two different settings: the first set consisted of the clean speech files only (indicated 'clean' in the following), and the second of a mixture of clean and noisy speech files, where different types of noise have been added artificially to the speech signals (so-called multi-condition training, see Hirsch and Pearce, 2000).

The source test speech material (not the training material) has been transmitted through the simulation model with 40 different connection settings. The parameters of each connection are given in Table 1, which indicates only the parameters differing from the default setting defined in Table 3 of ITU-T Rec. G.107 (2000). The connections include different levels of narrow-band or wideband circuit noise (No. 1-19), several codecs operating at bit-rates between 32 and 8 kbit/s (No. 20-26), signal-correlated quantizing noise modeled by means of a modulated noise reference unit at the position of the codec (MNRU, see ITU-T Rec. P.810, 1996, for details; No. 27-32), as well as combinations of non-linear codec distortions and circuit noise (No. 33-40).

It has to be mentioned that the tested impairments solely reflect the listening-only situation, and for the sake of comparison, they did not include background noise. In realistic dialogue scenarios, however, the simulation model also permits testing of conversational impairments.

5. Assessment Results

In this section, we will take the viewpoint of a transmission network planner, who has to guarantee that the transmission system performs well for both human-to-human and human-machine communication. A prerequisite for the former is an adequate speech quality, for the latter a good ASR performance. Thus, we will investigate the degradation in recognition performance due to the transmission channel, and compare it to the quality degradation which can be expected in human-to-human communication. This is a comparison between two unequal partners, which nevertheless have some similar underlying principles.

Speech quality is a subjective entity, and it is not completely determined by the acoustic signal reaching the listener's ear. Intelligibility, i.e. the ability to recognize what is said, forms just one dimension of speech quality. It also has to be measured subjectively, using auditory experiments. The performance of a speech recognizer, in contrast, is not a subjective entity, but it can be measured instrumentally. As for speech quality, it depends on the speech signal as well as on the 'background knowledge', which is mainly included in the acoustic and language models of the recognizer.

From a transmission point of view, comparing the unequal partners seems to be justified. Both are prerequisites for reasonable communication quality. Whereas speech quality is a direct, subjective quality measure, recognizer performance is only *one* quality element which contributes to the overall quality of the human-machine interaction. Unfortunately, there is no fixed relationship between recognition performance on the one hand, and human-machine communication quality on the other. Approaches to set up such a relation have been proposed by Walker et al. (1997) with the PARADISE framework, but they are not universal and have to be determined for each application anew.

For the planner of transmission networks, it is important that good speech quality as well as good recognition performance are provided by the network, because speech transmission channels are increasingly being used with both human *and* ASR back-ends. If, however, the aim is to have a close look at the underlying recognition mechanisms, it would be better to compare speech intelligibility to ASR performance, see e.g. Lippmann (1997). Intelligibility, however, is no longer a planning aspect of modern telecommunication networks.

In Figures 2 to 7, recognition results are presented in relation to the amount of transmission channel degradation, e.g. the noise level, type of codec, etc. Recognizer performance is first calculated in terms of the percentage of correctly identified words (*%corr*), and the corresponding error rates (substitutions, insertions and deletions; *%corr* = 100% - *%sub* - *%del*), which are not reproduced here. Because we are only interested in the relative recognizer performance with respect to the performance without transmission degradation (*topline*), an adjustment to a normalized performance range [$perf_{min}$;$perf_{max}$] has subsequently been performed. We used a linear transformation for this purpose:

$$\%corr_n = \frac{\%corr}{topline} \cdot (perf_{\max} - perf_{\min}) + perf_{\min}$$

The topline recognition rates – condition No. 0 without transmission – were 98.8 (clean training) and 98.6 (multi-condition training) for the AURORA recognizer, and 68.1 for the German recognizer. For the Swiss-French continuous recognizer, the calculation is carried out twice, both for all the words in the vocabulary (topline 57.4), as well as for just the keywords which are used in the speech understanding module (topline 69.5). The alignment was performed according to the NIST evaluation scheme, using the SCLITE software (see NIST, 2001). The German recognizer carries out a keyword-spotting, so the evaluation was performed uniquely on keywords. The AURORA recognizer was always evaluated with respect to the complete connected digit string.

Obviously, the recognizers differ in their absolute performance because the applications they have been built for are different. This fact is tolerable, as we are only interested in their relative degradation of recognition performance, as a function of the physically measurable channel characteristics. All recognition scores are thus normalized to a range which can be compared to the quality index predicted by the E-model. The normalization also helps to draw comparisons between the recognizers. For human-human communication, the E-model predicts speech quality in terms of a transmission rating factor *R* [0;100], which can be transformed via a non-linear relationship into estimations of mean users' quality judgments on a 5-point ACR scale, the mean opinion scores MOS [1;4.5] (see ITU-T Rec. G.107, 2000).

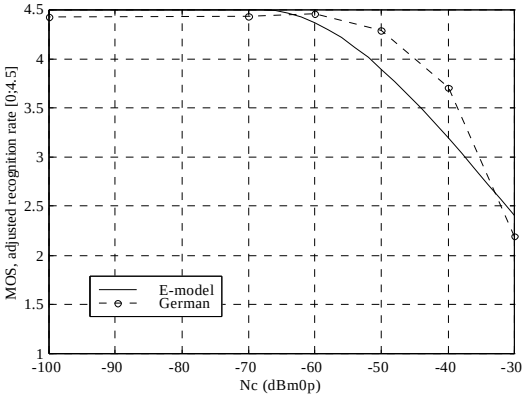

Figure 2. Comparison of E-model MOS prediction and adjusted recognition rate, German recognizer. Variable parameter: *Nc*. *Nfor* = -100 dBmp.

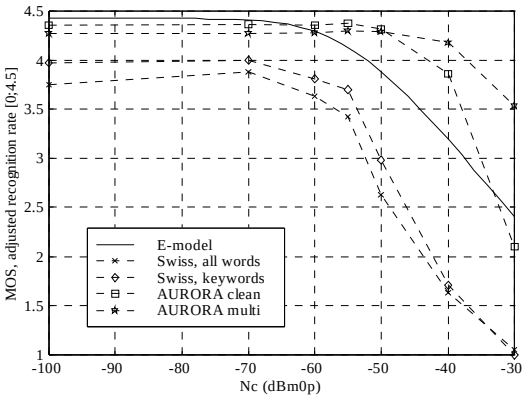

Figure 3. Comparison of E-model MOS prediction and adjusted recognition rate, Swiss and AURORA recognizers. Variable parameter: *Nc*. *Nfor* = -64 dBmp.

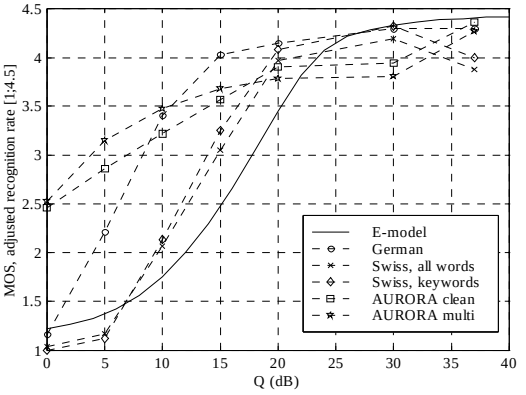

Figure 4. Comparison of E-model MOS prediction and adjusted recognition rate.
Variable parameter: Signal-to-quantizing-noise ratio

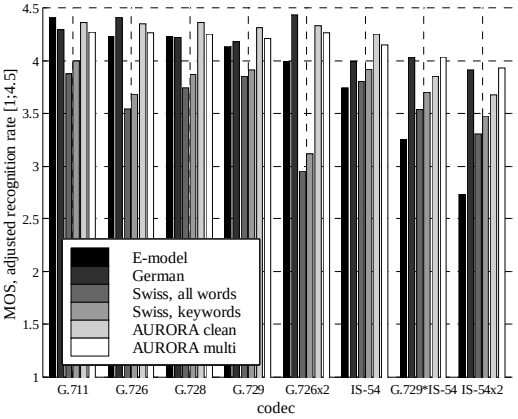

Figure 5. Comparison of E-model MOS prediction and adjusted recognition rate. Variable parameter: Codec.

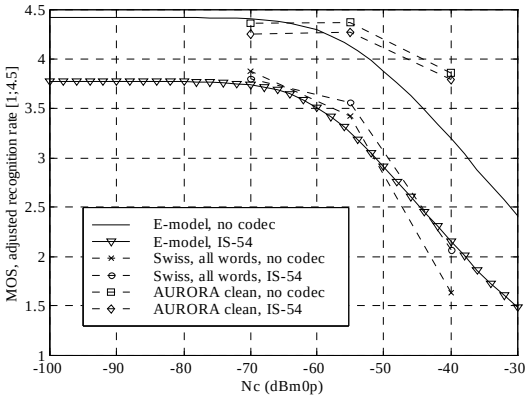

Figure 6. Comparison of E-model MOS prediction and adjusted recognition rate. Variable parameters: *Nc* and codec.

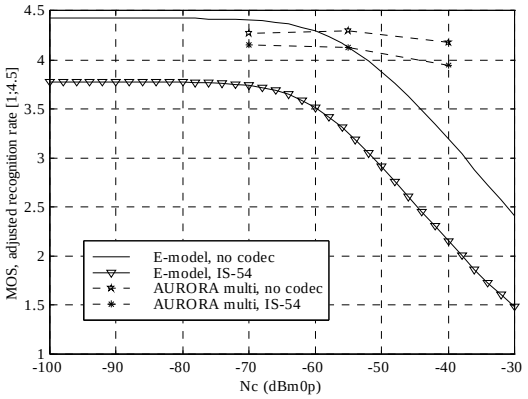

Figure 7. Comparison of E-model MOS prediction and adjusted recognition rate. Variable parameters: *Nc* and codec.

No.	*Nc* (dBm0p)	*Nfor* (dBmp)	Codec/MNRU	Note	Test Condition				
					Swiss, all w.	Swiss, keyw.	Ger.	AUR., clean	AUR., multi
0	-	-	-	no transmission	x	x	x	x	x
1	-100	-100	-	low noise, no codec	x	x	x	x	x
2	-100	-100	G.711	low noise	x	x	x	x	x
3	-70	-100	G.711	low noise			x		
4	-60	-100	G.711	moderate nb. noise			x		
5	-50	-100	G.711	moderate nb. noise			x		
6	-40	-100	G.711	high nb. noise			x		
7	-30	-100	G.711	high nb. noise			x		
8	-70	-70	G.711	low noise			x		
9	-70	-64	G.711	default connection	x	x	x	x	x
10	-70	-60	G.711	moderate wb. noise			x		
11	-70	-50	G.711	moderate wb. noise			x		
12	-70	-40	G.711	high wb. noise			x		
13	-70	-30	G.711	high wb. noise			x		
14	-100	-64	G.711	low nb. noise	x	x		x	x
15	-60	-64	G.711	moderate nb. noise	x	x		x	x
16	-55	-64	G.711	moderate nb. noise	x	x		x	x
17	-50	-64	G.711	moderate nb. noise	x	x		x	x
18	-40	-64	G.711	high nb. noise	x	x		x	x
19	-30	-64	G.711	high nb. noise	x	x		x	x
20	-70	-64	G.726	ADPCM coding	x	x	x	x	x
21	-70	-64	G.728	LD-CELP coding	x	x	x	x	x
22	-70	-64	G.729	CS-ACELP coding	x	x	x	x	x
23	-70	-64	IS-54	VSELP coding	x	x	x	x	x
24	-70	-64	G.726*G.726	ADPCM tandem	x	x	x	x	x
25	-70	-64	IS-54*IS-54	VSELP tandem	x	x	x	x	x
26	-70	-64	G.729*IS-54	mixed tandem	x	x	x	x	x
27	-70	-64	MNRU, *Q*=30 dB	low sign.-corr. noise	x	x	x	x	x
28	-70	-64	MNRU, *Q*=20 dB	moderate sign.-corr. noise	x	x	x	x	x
29	-70	-64	MNRU, *Q*=15 dB	moderate sign.-corr. noise	x	x	x	x	x
30	-70	-64	MNRU, *Q*=10 dB	moderate sign.-corr. noise	x	x	x	x	x
31	-70	-64	MNRU, *Q* = 5 dB	high sign.-corr. noise	x	x	x	x	x
32	-70	-64	MNRU, *Q* = 0 dB	high sign.-corr. noise	x	x	x	x	x
33	-100	-100	IS-54	VSELP, low noise			x		
34	-70	-100	IS-54	VSELP, low noise			x		
35	-60	-100	IS-54	VSELP, moderate noise			x		
36	-50	-100	IS-54	VSELP, moderate noise			x		
37	-40	-100	IS-54	VSELP, high noise			x		
38	-30	-100	IS-54	VSELP, high noise			x		
39	-55	-64	IS-54	VSELP, moderate noise	x	x		x	x
40	-40	-64	IS-54	VSELP, high noise	x	x		x	x

Table 1: Experimental conditions included in the tests

The comparison reveals some interesting differences. Figures 2 and 3 show the effects of uncorrelated narrow-band noise. For the German and the AURORA recognizers, higher levels of noise seem to be tolerable for ASR than for human-to-human speech quality predicted by the E-model. The performance of the Swiss-French recognizer drops at about the same noise level as the E-model quality estimation. All ASR performance curves seem to drop more dramatically than the E-model curve when noise levels become higher. Thus, a kind of threshold can be observed, above which recognition performance degrades dramatically. The exact position of the threshold depends on the recognizer and – as the AURORA recognizers show – on the training material as well.

Figure 4 shows the behavior with respect to signal-correlated quantizing noise of signal-to-noise level *Q* (in dB). The performance of the German and the Swiss-French recognizer degrade in a similar way than the E-model predicts, but they are more “robust” than humans, i.e. the decrease occurs at lower signal-to-noise ratios. The Swiss-French recognizer reaches its optimum performance not for the highest SNR, but around 30 dB – probably because this recognizer has been trained on telephone speech data with similar levels of quantizing noise. The performance of the AURORA recognizers degrades more gradually, both for clean and multi-condition speech data training. No explanation for this effect could be found so far. It has to be noted, however, that already low levels of quantizing noise significantly impact the performance of this recognizer (difference between 30 and 37 dB SNR, 37

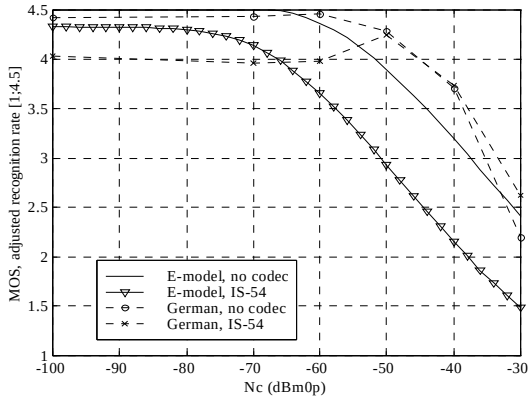

Figure 8. Comparison of E-model MOS prediction and adjusted recognition rate. Variable parameters: *Nc* and codec.

dB corresponding to logarithmic PCM coding according to ITU-T Rec. G.711).

Degradations originating from low bit-rate speech codecs are – with few exceptions – better "tolerated" by the recognizers than by humans, see Figure 5. Especially the German and the AURORA recognizers seem to be robust in this respect. The performance of the Swiss-French recognizer – although trained on telephone speech – is more affected by such degradations. In particular, this recognizer seems to be sensible to ADPCM coding (according to ITU-T Rec. G.726, both in simple operation and in double tandem). The high degradations of human-to-human speech quality predicted for the VSELP cellular codec (IS-54) by the E-model is not reflected in the recognition performance curves in the same way.

Figures 6 to 8 give an example of how combinations of different types of impairments – in this case uncorrelated narrow-band noise and VSELP coding – affect ASR performance and speech quality. The E-model curves are nearly parallel in the diagrams, indicating that the effects of noise and codec are reflected in a more-or-less additive way on the MOS scale. This is not the case for the Swiss-French recognizer ('x' and 'o' in both Figures), where the curves intersect. On the other hand, the AURORA recognizer shows parallel curves in both clean and multi-condition training conditions. The behavior of the German recognizer could not yet be explained. Especially the – reproducible – recognition rate at *Nc* = -50 dBm0p seems to be an outlier.

6. Conclusions and Outlook

The results have some implications for both developers of ASR systems as well as telephone network planners. Although in many cases transmission channels planned according to human quality considerations will also yield good ASR performance, the observed threshold effects for uncorrelated noise have to be taken into account when moderate-to-low quality transmission is encountered. In general, codecs operating at low bit-rates seem to have a lower impact on ASR performance than on human-to-human speech quality. Thus, networks planned according to human quality requirements may normally satisfy also the requirements set by the ASR. The ADPCM codec, on the other hand, shows that this rule is not without exception. From the experiments, it can unfortunately not be excluded that similar weak points may exist for other (or new) types of codecs.

On the basis of more extensive work in this domain, quality modeling approaches defined in the telephone community may become interesting also for application in the speech technology community. Thus, it might become possible to predict ASR performance in specific transmission channel configurations from the ASR system's topline performance. Together with prediction models for other quality aspects in human-computer interaction (like PARADISE), valuable information on the quality and usability of a system can be deduced at an early stage of system development. This will allow efficient and economical system set-up, and help to increase the success of spoken dialogue systems in the long term.

7. Acknowledgements

This work has been carried out at the Institute of Communication Acoustics, Ruhr-University Bochum (Prof. J. Blauert, PD U. Jekosch). It was partly performed under the EU-funded TMR-project "SPeech and HEARing" (SPHEAR). The simulation model was developed within a framework of a project funded by T-Nova Deutsche Telekom Berlin, the Swiss-French recognizer was developed under the Swiss CTI-funded project InfoVOX. The authors would like to thank S. Rehmann, M. Hilckmann and J. Riedel for processing part of the data material.

8. References

Chollet, G., J.-L. Cochard, A. Constantinescu, A. Jaboulet and Ph. Langlais, 1996. *Swiss French PolyPhone and PolyVar: Telephone Speech Databases to Model Inter- and Intra-Speaker Variability*. Technical Report RR-96-01, IDIAP, CH-Martigny.

Euler, S. and J. Zinke, 1994. The Influence of Speech Coding Algorithms on Automatic Speech Recognition. In: *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'94)*, AUS-Adelaide, 1:621-624.

Giuliani, D., M Matassoni, M. Omologo and P. Svaizer, 1999. Training of HMM with Filtered Speech Material for Hands-Free Recognition. In: *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'99)*, USA-Phoenix, 1:449-452.

Hermansky, H. and N. Morgan, 1994. RASTA Processing of Speech. *IEEE Trans. Speech and Audio Processing*, 2:578-589.

Hirsch, H.-G. and D. Pearce, 2000. The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions. In: *Proc. ISCA ITRW Automatic Speech Recognition: Challenges for the New Millennium (ASR 2000)*, F-Paris.

HTK, 2000. *Hidden Markov Model Toolkit*, Version 3.0. Cambridge University Engineering Department, UK-Cambridge. Available under http://htk.eng.cam.ac.uk.

ITU-T Rec. G.107, 2000. *The E-Model, a Computational Model for Use in Transmission Planning*. International Telecommunication Union, CH-Geneva.

ITU-T Rec. P.810, 1996. *Modulated Noise Reference Unit (MNRU)*. International Telecommunication Union, CH-Geneva.

Johannesson, N.O., 1997. The ETSI Computational Model: A Tool for Transmission Planning of Telephone Networks. *IEEE Comm. Mag.*, Jan:70-79.

Karray, L., A. Ben Jelloun and C. Mokbel, 1998. Solutions for Robust Recognition over the GSM Cellular Network. In: *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'98)*, USA-Seattle, 1:261-264.

Lilly, B.T., and K.K. Paliwal, 1996. Effect of Speech Coders on Speech Recognition Performance. In: *Proc. Int. Conf. Spoken Language Processing (ICSLP'96)*, USA-Philadelphia, 2344-2347.

Lippmann, R.P., 1997. Speech Recognition by Humans and Machines. *Speech Communication*, 22:1-15.

Möller, S., 2000. *Assessment and Prediction of Speech Quality in Telecommunications.* USA-Boston: Kluwer Academic Publ.

Möller, S. and H. Bourlard, 2000. Real-Time Telephone Transmission Simulation for Speech Recognizer and Dialogue System Evaluation and Improvement. In: *Proc. Int. Conf. Spoken Language Processing (ICSLP 2000)*, CHN-Beijing, 1:750-753.

Möller, S. and H. Bourlard, 2002. Analytic Assessment of Telephone Transmission Impact on ASR Performance Using a Simulation Model. *Speech Communication*, in press.

Mokbel, C., J. Monné and D. Jouvet, 1993. On-Line Adaptation of a Speech Recognizer to Variations in Telephone-Line Conditions. In: *Proc. 3rd European Conf. on Speech Communication and Technology (EUROSPEECH'93)*, D-Berlin, 1247-1250.

Mokbel, C., L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin and K. Bartkova, 1997. Towards Improving ASR Robustness for PSN and GSM Telephone Applications. *Speech Communication*, 23:141-159.

National Institute of Standards and Technology (NIST), 2001. *Speech Recognition Scoring Toolkit*. http://www.nist.gov/speech/tools.

Puel, J.-B. and R. André-Obrecht, 1997. Cellular Phone Speech Recognition: Noise Compensation vs. Robust Architectures. In: *Proc. 5th European Conf. on Speech Communication and Technology (EUROSPEECH'97)*, GR-Rhodes, 1151-1154.

Tarcisio, C., F. Daniele, G. Roberto and O. Marco, 1999. Use of Simulated Data for Robust Telephone Speech Recognition. In: *Proc. 6th European Conf. on Speech Communication and Technology (EUROSPEECH'99)*, H-Budapest, 6:2825-2828.

Tucker, R., T. Robinson, J. Christie and C. Seymour, 1999. Compression of Acoustic Features – Are Perceptual Quality and Recognition Performance Incompatible Goals? In: *Proc. 6th European Conf. on Speech Communication and Technology (EUROSPEECH'99)*, H-Budapest, 5:2155-2158.

Walker, M.A., D.J. Litman, C.A. Kamm and A. Abella, 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In: *Proc. 35th Ann. Meeting of the Assoc. for Computational Linguistics (ACL/EACL 97)*, USA-San Francisco: Morgan Kaufmann, 271-280.