IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Analytic Assessment of Telephone Transmission Impact on ASR Performance Using a Simulation Model

Sebastian Möller Hervé Bourlard

IDIAP-RR-01-17

May 2001

To be published in Speech Communication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

Analytic Assessment of Telephone Transmission Impact on ASR Performance Using a Simulation Model

Sebastian M*ÖLLER*

Institut für Kommunikationsakustik, Ruhr-Universität Bochum, D-44780 Bochum, Germany

and

Hervé B*OURLARD*

IDIAP – Institut dalle Molle d'Intelligence Artificielle Perceptive, CP 592,
CH-1920 Martigny, Switzerland

Communication address:

Sebastian Möller
Institut für Kommunikationsakustik
Ruhr-Universität Bochum
D-44780 Bochum
Germany

phone: +49 234 322 3979
fax: +49 234 321 4165
email: moeller@ika.ruhr-uni-bochum.de

Summary

This paper addresses the impact of telephone transmission channels on automatic speech recognition (ASR) performance. A real-time simulation model is described and implemented, which allows impairments that are encountered in traditional as well as modern (mobile, IP-based) networks to be flexibly and efficiently generated. The model is based on input parameters which are known to telephone network planners; thus, it can be applied without measuring specific network characteristics. It can be used for an analytic assessment of the impact of channel impairments on ASR performance, for producing training material with defined transmission characteristics, or for testing spoken dialogue systems in realistic network environments. In the present paper, we present an investigation of the first point. Two speech recognizers which are integrated into a spoken dialogue system for information retrieval are assessed in relation to controlled amounts of transmission degradations. The measured ASR performance degradation is compared to speech quality degradation in human-

human communication. It turns out that ASR shows a different behavior then expected human quality judgments for some impairments. This fact has to be taken into account in both telephone network planning as well as in speech and language technology development.

Zusammenfassung

Dieser Beitrag untersucht den Einfluss des Telefon-Übertragungskanals auf die Leistung von Spracherkennern. Zu diesem Zweck wird ein Simulationsmodell entwickelt, mit dessen Hilfe die Störungen von traditionellen und modernen (z.B. mobilen oder IP-basierten) Übertragungsstrecken gezielt generiert werden können. Das Modell verwendet Eingangsparameter, wie sie in der Netzwerkplanung üblich sind; es ist daher nicht notwendig, spezielle Messungen in realen Netzen durchzuführen. Drei Anwendungen des Modells bieten sich an: zum Einen die diagnostische Untersuchung des Einflusses verschiedener Störungen auf die Erkennungsleistung (Ziel dieser Untersuchung); des weiteren die Herstellung von Trainingsmaterial mit definierten Übertragungscharakteristika; oder die Beurteilung von Dialogsystemen in realistischen akustischen Situationen. Wir untersuchten die Leistungen von zwei Spracherkennern eines telefonbasierten Auskunftssystems in Abhängigkeit von den Übertragungseigenschaften. Es zeigte sich, dass sich die vom Telefonkanal hervorgerufenen Verschlechterungen teilweise anders auswirken, als dies für die Qualität direkter Mensch-zu-Mensch-Kommunikation zu erwarten ist. Im Beitrag werden die Auswirkungen sowohl für die Planung von Telefonnetzen als auch für die Entwicklung von Sprachtechnologie diskutiert.

Résumé

Dans ce papier, nous évaluons en détail l'influence du canal de transmission téléphonique sur les performances de systèmes de reconnaissance automatique de la parole (RAP). Un simulateur temps-réel est décrit et mis en œuvre, permettant une génération flexible et contrôlée des différentes perturbations habituellement rencontrées dans les réseaux téléphoniques, aussi bien traditionnels (fixes) que mobiles et IP. Le modèle utilisé est basé sur un ensemble de paramètres d'entrée qui sont connus des concepteurs de réseaux ; il est donc applicable sans devoir mesurer explicitement les caractéristiques du réseau. Ce modèle peut donc être utilisé pour mesurer analytiquement l'impact des perturbations du réseau sur les performances de la RAP, pour produire des données d'entraînement correspondant à des caractéristiques de transmission déterminées, ou encore pour tester des systèmes vocaux interactifs dans des conditions de réseau réalistes et multiples. Dans ce papier, nous nous

focalisons avant tout sur le premier point. La robustesse de deux systèmes de RAP, intégrés dans une application de recherche d'informations basée sur un dialogue vocal, est évaluée en fonction de la dégradation (contrôlée) du canal de transmission. Les dégradations résultantes du système de RAP sont ensuite comparées à celles observées dans le cas de la communication homme-homme. Il est alors intéressant de noter que les conclusions dépendent fortement du type de perturbations (résultant souvent en différents comportements). Ces conclusions sont pertinentes aussi bien dans le cadre de la conception des réseaux que lors du développement de technologies vocales.

1. Introduction

It is a well-known fact that the overall quality of spoken dialogue systems operated over telephone networks is largely affected by the quality of the transmission channel. On the one hand, the channel limits the performance of state-of-the-art speech recognition and speaker identification systems. This has an influence on the subsequent language processing stages, such as speech understanding and dialogue management. On the other hand, speech output – be it synthetically or naturally produced – is degraded on its way back to the human user. The quality degradation caused by the channel has to be taken into account both in the design of speech technology systems and components, as well as in the design of high-quality telecommunication networks. Speech technology providers try to produce systems which are robust or adaptive to noise and distortion, and telephone planning experts configure their networks according to transmission quality considerations.

The impairments which can be found in modern telecommunication networks are very diverse in nature. Due to the diversification of transmission techniques and user interfaces, and the liberalization of the telecommunication market, interconnected networks (wireline, mobile, IP-based, etc.) are common even for short-distance calls. Traditional – analogue as well as digital – wireline networks introduce loss, linear frequency distortion and noise, as well as quantizing distortion resulting from PCM-like waveform coding techniques. In modern networks, the channel is further degraded by the effects of non-linear codecs and time-variant transmission characteristics (transmission errors, voice-activity detection, clipping, fading, comfort noise). The transmission channel is often terminated by user interfaces with limited or poor acoustic properties, like short mobile handsets or hands-free terminals. Such user interfaces can easily pick up background noise, which is a serious problem in mobile communication scenarios.

Some of the impairments mentioned above 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). Making use of these approaches, the recognition performance can be improved for the addressed type of degradation, namely the one which was taken into account in the development of the system.

Flexible systems, however, should be able to cope with a variety of transmission channels. Unfortunately, it cannot be guaranteed that the recognizer will perform similarly well for new types of degradations, or for combinations of them. Wyard (1993) argues that the joint effects of different impairments have to be taken into account if systems are to developed successfully. Using databases recorded in real networks, corpus-based approaches include combined degradations in the training material. However, they do not provide any control over the type and amount of impairments, and they require large databases to be recorded in different networks and under different environmental (ambient noise) conditions. These databases are expensive and time-consuming to set up (e.g. Chang, 2000; Das et al., 1999; Höge et al., 1997, for the SpeechDat project; Hennebert et al., 2000, for speaker recognition).

Our ultimate aim is to investigate the overall performance of speech technology devices (here mainly speech recognition, but later also speaker recognition, dialogue management, and synthetic speech output) in relation to the transmission impairments. In order to be as flexible and economical as possible, the approach starts at an early stage of the system development. We develop a simulation model which generates all the relevant transmission channel degradations in a controlled way. The modeled degradations are those which are encountered in traditional and modern telecommunication networks. Several degradations can be implemented simultaneously; thus, it becomes possible to address the combined effects of different types of impairments (e.g. codecs operating on noisy speech) in a realistic way. Due to its real-time capability, the simulation model can be operated just as well in a one-way (transmission) or in a bi-directional (conversation) mode.

In contrast to proposals made by Tarcisio et al. (1999) or Guiliani et al. (1999), we do not apply a detailed filtering technique which necessitates the measurement of impulse responses in real-life networks. Input parameters to our simulation model are planning values, which are commonly used in the planning process of telecommunication networks. Such values are generally available to the planners of telecommunication networks, without the need for specific measurements. Based on these planning values, the ASR impact can easily be investigated *before* the respective network has been set up. In this way, it is not only possible to adapt the ASR system to a class of transmission channels. In addition, telecommunication network planners obtain diagnostic information on the impact of specific characteristics of their networks and may use this opportunity to select suitable components.

Speech communication quality between humans 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 quality index. In the future, however, telecommunication networks have to provide adequate quality in both human-to-human and human-machine communication. Thus, it is interesting to compare the quality index provided by the E-model (for human-to-human communication) with recognition performance (as one aspect of human-machine communication). This possibility is provided by using the simulation model, because identical transmission conditions can be guaranteed. The investigations will be presented in the second half of this paper. They will form a basis for quality network planning for ASR, and show limitations and future extension possibilities of network planning models like the E-model with respect to human-machine-communication.

The architecture and implementation of the simulation system are described in Section 2. In the first study that we present here, this system has been applied to the assessment of the impact of modern telephone channels without time-variant distortions on ASR performance. Two prototype ASR systems, which are part of a telephone-based information server, have been used for this purpose (Section 3). The results are given in Section 4, and they are compared to the quality degradation which can be expected for human-to-human communication, using the E-model. A discussion and an outlook on potential extensions of the simulation model, as well as on transmission quality aspects for human-to-human as well as human-machine communication, conclude the paper (Section 5).

Further applications of the simulation technique are planned: One consists of a controlled degradation of large databases of clean speech, which can be used for model training and adaptation. In this way, it is possible to multiply the amount of available data with respect to network characteristics which are expected to be representative for the later application situation. On the other hand, the availability of an on-line simulation tool makes it possible to assess spoken dialogue systems in realistic conversation scenarios, taking the transmission aspect into account. E.g., the influence of the ASR performance degradation on subsequent stages of speech understanding and dialogue management can be investigated, and appropriate adaptation techniques can be developed.

2. Telephone Transmission 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 Guiliani et al. (1999) for modeling hands-free terminals. When artificially degraded data was included in 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).

One main point of criticism against such techniques is that simulation systems first have to be validated for a given purpose, before they can be profitably used for describing and enhancing speech technology devices (Wyard, 1993). Because the number of potential impairment scenarios in real-life networks is almost infinite, such a verification can never be performed in an exhaustive way. In evolving networks, the actual connection characteristics will differ from connection to connection. As a consequence, a verification attempt is always limited to a few specific scenarios which have been measured by chance. Our simulation model has been verified with an intrusive network measurement system, showing that the characteristics of the generated transmission paths correspond to the measured ones (Raake and Möller, 1999), as well as during the assessment of the transmission impact in auditory experiments (Möller, 2000). By implementing a large number of perceptually diverse impairments (which are perhaps not diverse for a recognizer), we hope to overcome some of the limitations of simple filtering techniques and to reduce the concerns mentioned above. However, we admit that a more thorough validation of our modeling approach with respect to ASR in real-life network scenarios is necessary, e.g. by comparing the ASR performance in real-life networks to the results obtained with the transmission simulation. This is not an easy task, because of the lack of control over transmission characteristics in real-life networks.

In principle, telephone network characteristics can be measured either off-line with specific test signals (so-called intrusive measurements, because specific test calls are set up), or on-line, during normal system operation (so-called non-intrusive measurements). Depending on the type of measurement, different characteristics of the network can be quantified. The necessary measurement set-ups and algorithms are described by the International Telecommunication Union (ITU-T) in their P-Series Recommendations. All the measurements, however, require the presence of an operating network or its components. This is often not the case in the system development phase. At that stage, however, planning values do exist which reflect the instrumentally measurable characteristics of the transmission and terminal equipment. We will take these planning values as input parameters for our transmission simulation model.

[Figure 1]

The ITU-T recommends a simplified network configuration together with its computational model for network planners, in ITU-T Rec. G.107 (2000). This configuration is reproduced in Figure 1. It considers most of the impairments resulting from terminal, switching and connection elements which can be found in modern – analogue as well as digital – telephone networks. Specific time-variant impairments of mobile and IP-based networks have not yet been taken into account. This is a limitation of the configuration depicted in Figure 1. Our simulation model is currently being extended in this respect, namely for packet loss in IP-based networks, as well as for fading radio channels.

Input parameters in this configuration are scalar planning values, which can be obtained by measuring frequency responses of the transmission paths, or noise power spectra, and by using defined frequency-weighting algorithms. The characteristics of the transmission paths (main speech transmission path, echo path, sidetone path) are expressed in terms of so-called loudness ratings, which reflect the perceived loudness of the path (for a definition of loudness ratings see ITU-T Rec. P.79, 1999), and the corresponding mean delays. Noises are described by their psophometrically weighted power levels (for circuit noise and noise floor) or by standard A-weighted power levels (for ambient noise). Waveform codecs and effects of A/D-D/A conversion are expressed in terms of a signal-to-quantizing-noise ratio. The effects of non-linear codecs operating at medium or low bit-rates cannot easily be described by an instrumentally measurable parameter. For the purpose of network planning, they are covered by another scalar value, the so-called equipment impairment factor *Ie*. It describes the additional amount of degradation which is introduced by the coder-decoder pair, in comparison to other impairments. The exact description of all the planning values can be found in ITU-T Rec. G.107 (2000) and the corresponding ITU-T P-series Recommendations.

For setting up a bi-directional communication situation, the network configuration depicted in Figure 1 must now be implemented in a simulation system, which is able to generate all the impairments that potentially lead to a degradation in transmission performance. Inputs to the simulation model are all planning parameters given in Figure 1. They have to be adjustable in a controlled way, and within reasonable limits that are expected to occur in real-life networks.

In order to be usable in a bi-directional conversation mode, both transmission paths have to be implemented with minimal computational overall delay.

[Figure 2]

According to these requirements, the structure which is given in Figure 2 has been implemented on DSP hardware. We chose a signal processing hardware which can be wired and programmed via software. Software manipulation makes the system more flexible (e.g. for quickly changing parameter settings), while the hardware allows the simulation to run in real or close-to-real time. The triangles represent programmable filters (which can be used as attenuators as well), the rectangles delay lines (for *T*, *Ta* and *Tr*), external codecs, or the channel bandpass filter. Low bit-rate codecs have been implemented on another DSP hardware, and they can be cascaded up to three times. Different types of user interfaces can be connected to the simulation system in a four-wire mode (in contrast to Figure 1, which assumes two-wire/four-wire transitions at both user interfaces). We used standard wireline telephone handsets, short mobile handsets, as well as hands-free terminals and headsets. The electro-acoustic sensitivities of the user interfaces (*SLRset* and *RLRset*) were measured using an artificial head and subsequently adjusted via *SLR'* and *RLR'* to a desired frequency shape which is defined by the ITU-T (a so-called intermediate reference system, see ITU-T Rec. P.48, 1989).

The following degradations can be generated by the simulation model in a controlled way, using the corresponding planning values as input parameters (exact descriptions of all parameters are given e.g. in Möller, 2000; indices 1 and 2 in Figure 2 indicate the direction of the transmission):

- Attenuation and frequency distortion of the main transmission path (expressed in terms of loudness ratings, namely the send loudness rating, *SLR*, and receive loudness rating, *RLR*)
- Continuous white circuit noise, representing all the potentially distributed noise sources, both on the channel (*Nc*, narrow-band because it is filtered with the BP filter) and at the receive side (*Nfor*, wide-band restricted by the electro-acoustic coupling in the receiver handset)
- Transmission channel bandwidth impact: BP with 300-3400 Hz according to ITU-T Rec. G.712, 1996 (so-called "normal" telephone bandwidth), or a wide-band characteristic 50-7000 Hz according to ITU-T Rec. G.722 (1988)
- Impact of different speech codecs: Several low bit-rate codecs standardized by the ITU-T, as well as a North American cellular codec and proprietary codecs are implemented. They

include logarithmic PCM (ITU-T Rec. G.711), ADPCM (G.726), a low-delay CELP coder (G.728), a conjugate-structure algebraic CELP coder (G.729), and a vector sum excited linear predictive coder (IS-54). In the E-model, codecs are described by the corresponding equipment impairment factor, *Ie*, as given in ITU-T Rec. G.113 (1996). Alternatively, a generator for quantizing noise resulting from waveform codecs (log. PCM) or D/A-A/D conversions was implemented. The corresponding degradation is expressed in terms of the signal-to-quantizing-noise ratio *Q*, or in quantizing distortion units (*qdu*) (a fixed relation is given in ITU-T Rec. G.107 (2000): $Q = 37 - 15 \cdot \log_{10}(qdu)$)

- Ambient room noise of A-weighted power level *Ps* at the send side, and *Pr* at the receive side (*Pr = Ps2* in Figure 2)
- Pure overall delay (*Ta*) in ms
- Talker echo with one-way delay *T* (in ms) and attenuation *Le* (the corresponding loudness rating *TELR* can be calculated by: *TELR* = *SLR* + *RLR* + *Le*)
- Listener echo with round-trip delay *Tr* (in ms) and an attenuation with respect to the direct speech (corresponding loudness rating *WEPL* of the closed echo loop)
- Sidetone with attenuation *Lst* (loudness rating for direct speech: *STMR* = *SLRset* + *RLRset* + *Lst* – 1; loudness rating for ambient noise: *LSTR* = *STMR* + *Ds*)

Comparing Figures 1 and 2, it can be seen that all the relevant transmission paths and all the impairments in the planning structure are covered by the simulation model. There is a small difference to real-life networks in the simulation of the echo path: Whereas the talker echo normally originates from a reflection at the far end and passes through two codecs, the simulation only takes one codec into account. This allowance was made to avoid instability, which otherwise can result from a closed loop formed by the two echo paths. The simulation is integrated in a test environment which consists of two test cabinets (e.g. for recording or carrying out conversational tests) and a control room. Background noise can be inserted in both test cabinets, so that realistic ambient noise scenarios can be set up. This means that the speaking style variation due to ambient noise (Lombard reflex) as well as due to bad transmission channels is guaranteed to be realistic.

In the following study, the simulation has been used in a one-way transmission mode, replacing the second handset interface with a speech recognizer. For the pure transmission, it is not necessary to set up the whole system, and to make it run in real time. In that case it will be sufficient to implement only one transmission path (following the dashed line in Figure 2 and omitting the overall delay *Ta1*), and to use a pure software solution. Depending on the task, simplified solutions can easily be deduced from the full structure of Figure 2, and they can be implemented either using standard filter structures (as we did in our experiments) or specifically measured ones. When recording speech samples at the left bin of Figure 2, it is important to implement the sidetone path (*Lst1*) and in case of noticeable echo also the talker echo path (*Le1*), because the feedback they provide (of speech and background noise at the send side) might influence the speaking style – an effect which cannot be neglected in ASR. The real-time capability, on the other hand, is necessary for assessing conversational impacts in realistic dialogue situations, e.g. for a conversation with an adaptive spoken dialogue system.

3. Recognizer and Test Set-Up

The simulation model is now being used to assess the impact of several types of telephone degradation on the performance of speech recognizers (see very first results presented by Möller and Bourlard, 2000). Two recognizers are used for this purpose. Both are part of a spoken dialogue system which provides information on restaurants in the city of Martigny, Switzerland (Swiss-French version) or Bochum, Germany (German version). The spoken dialogue system is integrated into a larger server which enables voice and internet access, and which has been implemented under the Swiss CTI-funded project InfoVOX.

It has to be noted that neither of the recognizers are 'standardized' systems. They reflect typical solutions used in spoken dialogue systems. This means that the outcome of our experiments may be representative for similar application scenarios. Whereas we can obtain a reasonable estimation of the relative performance in relation to the amount of transmission channel degradation, the absolute performance of both recognizers is not yet competitive. This is due to the fact that the whole system is still in the prototype stage and has not been optimized for the specific application scenario. In the future, we hope to be able to repeat the experiments with both an optimized recognizer and some kind of 'standardized' recognizer set-up. The AURORA framework, established by the European Telecommunications Standards Institute (ETSI), provides useful definitions in this respect.

The Swiss-French recognizer (S) 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 (G) 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 proper 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. As we do not want to investigate the features of the specific recognizer, this fact is tolerable for the given purpose.

Because both systems are still in the prototype stage, test data is relatively restricted. We think that this is not a severe limitation, as we are only interested in the relative performance degradation, and not in absolute numbers. The Swiss-French system (S) was tested with 150 test utterances which were collected from 10 speakers (6m, 4f) in a quiet library environment (Ps ~ 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 (G) 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.

[Table 1; Table 2]

The test utterances were digitally recorded and then transmitted through the simulation model (cf. the dashed line in Figure 2). At the output of the simulator the degraded utterances were collected and then processed to the recognizer. All in all, 40 different settings of the simulation model were tested. The exact parameter settings are given in Table 1, which indicates only the parameters differing from the default setting. The connections include different levels of narrow-band or wide-band circuit noise (No. 2-19), 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). The other parameters of the simulation model, which are not addressed in the specific configuration, were set to their default values defined in ITU-T Rec. G.107 (2000), see Table 2. 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, conversational impairments can be tested as well.

4. Recognition Results and Discussion

In this chapter, 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 has been defined as the result of a perception and assessment process, in which the assessing subject establishes a relation between the perceived characteristics of the speech signal on the one hand, and the desired or expected characteristics on the other (see Jekosch, 2000). Thus, 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 also depends 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 communication. 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 systems, it is important that good speech quality as well as good recognition performance are provided by the system, 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 the following, 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:

$$\%corr_n = \frac{\%corr}{topline} \cdot \left(perf_{\max} - perf_{\min}\right) + perf_{\min} \tag{1}$$

For the Swiss-French continuous recognizer (S), the calculation is carried out twice, both for all the vocabulary, as well as for just the keywords which are used in the speech understanding module. The alignment was performed according to the NIST evaluation scheme, using the SCLITE software (see NIST, 2001). The German recognizer (G) carries out a keyword-spotting, so the evaluation was performed uniquely on keywords.

In order to estimate speech communication quality between humans, network planning experts use quality prediction models like the E-model (see Section 1). The E-model predicts speech quality in terms of a transmission rating factor R [0;100], which can be transformed via a non-linear S-shaped relationship into estimations of mean users' quality judgments on a 5-point ACR quality scale, the so-called mean opinion scores MOS [1;4.5]:

$$\begin{array}{rc} \text{for } R \le 0: & \text{MOS} = 1.0 \\ \text{for } 0 < R < 100: & \text{MOS} = 1 + 0.035 \cdot R + R(R-60)(100-R) \cdot 7 \cdot 10^{-6} \\ \text{for } R \ge 100: & \text{MOS} = 4.5 \end{array} \tag{2}$$

Based on the *R* values, classes of speech transmission quality are defined in ITU-T Rec. G.109 (1999), see Table 3. They indicate how the calculated *R* values have to be interpreted.

[Table 3]

E-model predictions are made for the network configuration which is depicted in Figure 1. Thus, it is possible to obtain speech communication quality estimates for all the tested transmission channels, based on the settings of the planning values which are used as an input to our simulation model. However, it has to be noted that both *R* and MOS values are only predictions, which do not necessarily correspond to user judgments in real conversation scenarios. Nevertheless, the validity of E-model predictions has been tested extensively (e.g. Möller, 2000; Möller and Raake, 2001), and it was found to be in relatively good agreement with auditory test data for most of the tested impairments.

In Figures 3-8, the E-model speech quality predictions in terms of *R* and MOS are compared to the normalized recognition performance of both the Swiss-French and the German recognizer. Because the transformation law between *R* and MOS is non-linear (see Formula 2), it is worth investigating both prediction outputs of the E-model. For *R*, the recognition rate has to be adjusted to a range of between $perf_{min}$ = 0 and $perf_{max}$ = 100, for MOS to a range of between $perf_{min}$ = 1 and $perf_{max}$ = 4.5. The *topline* parameter is defined as the recognition rate for the input speech material without any telephone channel transmission, collected at the left bin in Figure 2 (G: *topline* = 68.1%; S: *topline* = 57.4% for all words and 69.5% for keywords only). Because the default channel performance may be significantly lower than the topline performance, the normalized recognition performance curves do not necessarily reach the highest possible level (100 or 4.5). This fact can be clearly observed for the Swiss-French recognizer, where recognition performance drops by about 10 % for the default channel. The strict bandwidth limitation applied in the current simulation model (G.712 filter) seems to be responsible for the decrease, because this recognizer has been trained on a telephone database with very diverse transmission channels and probably diverse bandwidth limitations. Because only prototype versions of both recognizers were available at the time the experiments were carried out, this mismatch was foreseen. It seems to be significantly lower in the case of the German keyword recognizer.

The test results have been separated for the different types of transmission impairments and are depicted in Figures 3 to 8. In each Figure the left diagram shows a comparison between the transmission rating *R* and the normalized recognition performance [0;100], the right diagram between MOS and the corresponding normalized performance [1;4.5]. Higher values indicate better performances for both *R* and MOS. The discussion here can only show general tendencies in terms of the shape of the corresponding performance curves; a deeper analysis requires to define “acceptable” limits for recognition performance (which will depend on the system the recognizer is used in) and for speech quality (e.g. on the basis of Table 3).

[Figure 3]

Figure 3 shows the degradations due to narrow-band (300-3400 Hz) circuit noise *Nc*. Because two different settings of the noise floor *Nfor* were used, German recognition results (o) have to be compared to the solid E-model prediction, and Swiss-French results (x and diamonds) to the dash-dotted E-model prediction line. A considerable decrease in recognition performance occurs for $Nc \geq$ -55...-50 dBm0p[1]. Assuming an active speech level of -19dBm on the line, this corresponds to an SNR of 31...36 dB. The performance deterioration of the Swiss-French system occurs at lower *Nc* levels than in the German system. In comparison to the E-model predictions, the recognition performance decrease is much steeper than the *R* and MOS decrease. The agreement, however, is slightly better for MOS than for *R*. For the Swiss-French system, the performance curves for all the words and keywords only are mainly parallel, except for very high noise levels where they coincide. For both recognizers, the optimum performance is not reached at the lowest noise level, but for $Nc \sim$ -70...-60 dBm0p. This is due to the training material, which was probably recorded at similar noise levels.

[Figure 4]

When wide-band noise (*Nfor*) is added instead of channel-filtered noise, the agreement between recognition performance degradation and predicted speech quality degradation is relatively good, see Figure 4. The decrease in performance occurs at nearly the same noise level as was predicted by the E-model, though it is much steeper for high noise levels. Once again, the MOS predictions are closer to the recognition performance degradation than the transmission rating *R*.

[Figure 5]

Figure 5 shows the effect of signal-correlated noise, which has been generated by a modulated noise reference unit (MNRU) at the position of the codec. The abscissa parameter is the signal-to-noise ratio *Q*. Compared to the Swiss-French recognizer, the German system is slightly more robust, in that the recognition performance decrease occurs at lower SNR

---

[1] The unit dBm0p is commonly used in network planning. It describes the absolute power level (dBm) a signal has at a virtual 0 dB reference point in the network, behind the *SLR'* filter in Figure 2. The index p defines that the power level has to be weighted psyphometrically, e.g. using a psophometer as defined in ITU-T Rec. O.41.

values. The shape of both recognition performance curves is close to the E-model prediction for MOS, but the decrease occurs at lower SNR values. Thus, human-to-human communication seems to be more critical to this type of degradation. As for *Nc*, the optimum recognizer performance is not reached for the highest SNR, but for Q ~ 30 dB. This can clearly be observed for the Swiss-French system. It can be assumed that the nature of the training database (telephone call data) showed approximately the same level of signal-correlated noise.

[Figure 6]

Non-linear codecs which are commonly used in modern telephone networks introduce different types of impairment, which are often neither comparable to correlated or uncorrelated noise nor to linear distortions. In Figure 6, recognition performance degradation and E-model predictions are compared for the following codecs: logarithmic PCM at 64 kbit/s (G.711), ADPCM at 32 kbit/s (G.726), low-delay CELP coding at 16 kbit/s (G.728), conjugate-structure algebraic CELP at 8 kbit/s (G.729), vector sum excited linear predictive coding at 7.95 kbit/s, as it is used in the first generation North-American TDMA cellular system (IS-54), as well as tandems of these codecs. It can be seen that there is no close agreement between estimated speech quality and recognition performance, neither for MOS nor for R predictions. The Swiss-French recognizer seems to be particularly sensitive to ADPCM (G.726) coding. This type of degradation is similar to the signal-correlated noise produced by the MNRU (Figure 5), where the same tendency has been observed. The German recognizer, on the other hand, is particularly insensitive to this codec, resulting in high recognition performances for the ADPCM codec in single as well as tandem operation. This recognizer also seems to be quite insensitive to codec tandeming in general, whereas the Swiss-French recognizer’s performance deteriorates. This deterioration is also predicted by the E-model, with respect to speech quality. Lilly and Paliwal (1996) found their systems to be insensitive to tandeming at high (32kbit/s) bitrates, but more sensitive to tandeming at low bitrates; this is just the opposite of what we observed for the Swiss-French system. Apart from the ADPCM codec, the rank order between codecs predicted by the E-model is generally maintained. The overall amount of degradation, on the other hand, is smaller than predicted for speech quality. This may be a consequence of using RASTA coefficients, which are expected to be relatively insensitive to a convolution-type degradation.

[Figure 7; Figure 8]

In Figure 7 the effect of combined impairments is investigated for the German recognizer. In this study, channel-filtered noise (*Nc*) and the IS-54 cellular codec were used. A fundamental disagreement between E-model predictions (solid and dash-dotted lines) and recognition performance (dotted lines) can be observed: Whereas the E-model curves are nearly parallel (especially on the *R*-scale), there is an intersection for the recognition performance curves with and without codec. The same tendency is found for the Swiss-French recognizer, see Figure 8. For speech quality, the E-model assumes additivity of different types of impairments on the *R*-scale. Our result indicates that this additivity property might not be satisfied with respect to recognition performance. No explanation can be given for the surprisingly high German recognition rates at *Nc* = 50 dBm0p, when combined with the IS-54 codec. Neither the corresponding connection without codec nor the Swiss-French system show such high rates.

5. Conclusions

The comparison between recognition performance and E-model predictions for speech quality, which was made in Chapter 4, reveals similarities, but also differences between the two entities. On the one hand, the (normalized) amount of recognition performance degradation seems to be similar to what is predicted by the E-model, namely for uncorrelated white noise (narrow-band as well as wide-band) and for signal-correlated noise. The agreement is better for the MOS predictions than for the *R* predictions. However, for all these noises, the quality decrease is steeper than predicted by the E-model. This might be an indication of a threshold effect occurring in the recognizer: Recognition performance is acceptable up to a specific threshold of noise and drops quickly when the noise level exceeds this threshold. The exact level of the threshold has to be defined in terms of the recognition performance which is required for a specific application. Different values for such a minimum requirement have been mentioned by system developers.

On the other hand, the correlation between predicted speech quality degradation and recognition performance degradation is less clear when low bit-rate codecs are considered. This may indicate that the E-model puts emphasis on quality dimensions like naturalness or sound quality, which are perhaps not so important for good recognition performance. More experimental data is needed to justify this hypothesis. Whereas the German recognizer seems to be relatively insensitive to codec-produced distortions, the Swiss-French system is particularly sensitive to ADPCM coding. The combination of IS-54 coding and circuit noise has been tested as an example for combined impairments. The resulting recognition performance curves do not agree well with the E-model predictions. In particular, some “masking” between the two degradations seems to be present (the codec-originated degradation is masked by the noise degradation for higher noise levels), resulting in an intersection of the performance curves which cannot be observed for the E-model prediction curves. If this difference in behavior can be reproduced for other combinations of impairments, the whole principle underlying the E-model will be difficult to apply to predicting recognition performance. However, doubt has already been cast on this principle by auditory experiments combining background noise and codec distortions, see Möller (2000).

So far, we have tested the influence of the transmission channel on two specific speech recognizers with different languages and test/training material. For noise (*Nc*, *Nfor*, and signal-correlated noise) both recognizers behave similarly. Nevertheless, the German recognizer seems to be more robust, in the sense that a deterioration in performance occurs at a higher noise level, or lower SNR. The behavior of the recognizers is different for low bit-rate coded speech. It has to be emphasized that our experiments do not permit a quality comparison to be drawn between the two recognizers. Instead, our figures only have a relative significance with respect to the impairment-free case. Both recognizers have been assessed in an application-near scenario, but they cannot be considered to be optimized systems. In the future, we plan to repeat the experiments with some kind of “standardized” recognizer and training/test material.

6. Outlook

Our results have some implications, both for the development of speech technology devices (recognition, speech detection, speaker recognition, dialogue management), as well as for the planners of speech transmission networks. Speech recognizers may show weaknesses for certain types of transmission degradations, which are either typical for recognizers in general, or specific to a particular recognizer. The simulation model presented in Section 2 helps to identify these weaknesses and subsequently to enhance recognizer performance. E.g., specific training material can be produced for optimizing acoustic models. The recognition results for *Nc* and signal-correlated noise emphasize the need for training material which has characteristics similar to the later application scenario’s. Such training material can be produced very efficiently using the presented simulation model.

Speech technology aspects should also be considered by transmission planning experts. In particular, codec and combined degradations show that telephone networks, which are planned according to the needs of human-to-human communication, do not necessarily satisfy the requirements of modern speech technology devices. Thus, what is tolerable according to the E-model (which is the only planning tool for quality planning of telephone networks) is not always tolerated by speech recognizers. Fortunately, only in two of our experimental conditions (20 and 24) a remarkable decrease in performance was observed for the Swiss-French recognizer where the corresponding E-model predictions were far less pessimistic. One could argue that this is only a problem for speech technology developers. However, telecommunication networks are not static, but evolve very quickly due to a changing technical and economic background. As a consequence, speech technology which has been adapted for specific, current transmission equipment is not necessarily robust towards new types of speech processing devices (e.g. new codecs). The current standardization processes for new codecs only includes auditory speech quality tests, but no tests with speech recognizers.

Apart from assessing the transmission channel impact on speech recognition, other applications of the simulation system are expected. E.g., synthesized speech can be assessed under realistic transmission channel situations. Both speech recognizers as well as speech synthesis may be adapted to the current channel characteristics. The characteristics may be determined online, e.g. using in-service, non-intrusive measurement devices (specified in ITU-T Rec. P.561, 1996), and can then be mapped onto parameters which are identical to the ones used for the simulation model (for details on the mapping, see ITU-T Rec. P.562, 2000). By simply comparing the parameters describing the network characteristics, adequate acoustic models can be chosen for the speech recognizer, or Lombard speech can be generated by an adaptive speech synthesizer. The effect of the degraded recognition on dialogue flow can be assessed in realistic WoZ scenarios, by installing the simulation model between the speech recognizer and the test users' interface.

We are planning further extensions to the simulation model. One problem to face is wide-band systems which will become more common for IP-based networks. Another problematic topic is time-variant channel characteristics, like random bit errors, bursty error patterns, or lost frames. This type of impairment is common in mobile as well as IP-based networks. Until now, no adequate method, apart from auditory tests, has been able to predict the corresponding effects on speech quality. Tests showed that the speech material and the time distribution of errors in the speech sample have an influence on the quality perceived by humans – a fact which will similarly play a role in assessing speech recognizer performance. Another characteristic which has to be modeled more extensively is the user interface. With the exception of standard handset telephones, modern networks will be operated from hands-free and headset terminals. Such terminals have a different sensitivity to the received speech material (e.g. because of room acoustic properties) as well as to ambient background noise, when compared to handsets. The effect of background noise is taken into account by our simulation model when databases are produced. Variations in speaking style (e.g. Lombard reflex) will be reflected in the speech material, as long as the recordings are made with realistic user interfaces.

Acknowledgments

This study was carried out partly at IKA, Ruhr-University Bochum (J. Blauert, U. Jekosch), and partly at IDIAP, Martigny (H. Bourlard). It was supported by the EU-funded project SPeech and HEARing (SPHEAR). The telephone line simulation model was developed within the framework of a project funded by T-Nova Deutsche Telekom Berkom, Berlin, the dialogue system in the context of the Swiss CTI-funded project InfoVOX. The authors would like to thank the members of both institutes for their support, C. Hill and A. Raake for their comments of the manuscript, M. Hilckmann, S. Rehmann and J. Riedel for processing the data material, and their anonymous reviewers for useful comments and suggestions.

References

Chang, H. M., 2000. *Is ASR Ready for Wireless Primetime: Measuring the Core Technology for Selected Applications*. Speech Communication 31, 293-307.

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

Das, S., Lubensky, D. and Wu, C., 1999. *Towards Robust Speech Recognition in the Telephony Network Environment – Cellular and Landline Conditions*. In: Proc. 6th European Conf. on Speech Communication and Technology (EUROSPEECH'99), H-Budapest, Vol. 5, 1959-1962.

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

Giuliani, D., Matassoni, M., Omologo, M. and Svaizer, P., 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, IEEE Sign. Proc. Soc., Vol. 1, 449-452.

Hennebert, J., Melin, H., Petrovska, D. and Genoud, D., 2000. *POLYCOST: A Telephone-Speech Database for Speaker Recognition*. Speech Communication 31, 265-270.

Hermansky, H. and Morgan, N., 1994. *RASTA Processing of Speech*. IEEE Trans. Speech and Audio Processing 2, No. 4, 578-589.

Höge, H., Tropf, H. S., Winski, R., van den Heuvel, H., Haeb-Umbach, R. and Choukri, K., 1997. *European Speech Database for Telephone Applications*. In: Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'97), D-Munich, IEEE Sign. Proc. Soc., Vol. 3, 1771-1774.

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

ITU-T Recommendation G.109, 1999. *Definition of Categories of Speech Transmission Quality*. International Telecommunication Union, CH-Geneva.

ITU-T Recommendation G.113, 1996, and Appendix I, 1999. *Transmission Impairments.* International Telecommunication Union, CH-Geneva.

ITU-T Recommendation G.712, 1996. *Transmission Performance Characteristics of Pulse Code Modulation Channels*. International Telecommunication Union, CH-Geneva.

ITU-T Recommendation G.722, 1988. *7 kHz Audio-Coding within 64 kbit/s*. International Telecommunication Union, CH-Geneva.

ITU-T Recommendation P.48, 1989. *Specifications for an Intermediate Reference System.* International Telecommunication Union, CH-Geneva.

ITU-T Recommendation P.79, 1999. *Calculation of Loudness Ratings for Telephone Sets.* International Telecommunication Union, CH-Geneva.

ITU-T Recommendation P.561, 1996. *In-Service, Non-Intrusive Measurement Device - Voice Service Measurements*. International Telecommunication Union, CH-Geneva.

ITU-T Recommendation P.562, 2000. *Analysis and Interpretation of INMD Voice-Services Measurements*. International Telecommunication Union, CH-Geneva.

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

Jekosch, U., 2000. *Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*. Habilitation thesis, Universität/GH Essen, D-Essen.

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

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

Lilly, B. T. and Paliwal, K. K., 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 Machines and Humans*. Speech Communication 22, 1-15.

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

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

Möller, S., Bourlard, H., 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, Oct. 16-20, Vol. I, 750-753.

Möller, S., and Raake, A., 2001. *Telephone Speech Quality Prediction: Towards Network Planning and Monitoring Models for Modern Network Scenarios.* Accepted by Speech Communication.

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

Puel, J.-B. and André-Obrecht, R., 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.

Raake, A., and Möller, S., 1999. *Analysis and Verification of the Tektronix M366 GSM Network QoS Analyser's Measurements and Quality Predictions*. Unpublished report, Institut für Kommunikationsakustik, Ruhr-Universität, D-Bochum.

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

Tucker, R., Robinson, T., Christie, J. and Seymour, C., 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, Vol. 5, 2155-2158.

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

Wyard, P., 1993. *The Relative Importance of the Factors Affecting Recognizer Performance with Telephone Speech.* In: Proc. 3rd European Conf. on Speech Communication and Technology (EUROSPEECH'93), D-Berlin, 1805-1808.

Tables

Table 1: Parameter settings used for the experiments. An ‘X’ in the last two columns indicates that the circuit condition has been included in the corresponding test. All the other parameters were adjusted to the values given in Table 2.

Table 2: Default network planning parameter values for the simulation model.

Table 3: Relationship between the transmission rating factor *R* and categories of speech transmission quality (see ITU-T Rec. G.109, 1999).

Table 1

No.	*Nc* (dBm0p)	*Nfor* (dBmp)	Codec/MNRU	E-model *R*	E-model MOS	Note	Test Condition Swiss	Test Condition German
1	-100	-100	-	100	4.50	no noise/codec	X	X
2	-100	-100	G.711	100	4.50	no noise	X	X
3	-70	-100	G.711	100	4.50	low noise		X
4	-60	-100	G.711	91.3	4.37	narrow-band noise		X
5	-50	-100	G.711	76.7	3.89	narrow-band noise		X
6	-40	-100	G.711	61.8	3.19	narrow-band noise		X
7	-30	-100	G.711	46.9	2.41	narrow-band noise		X
8	-70	-70	G.711	99.2	4.49	low noise		X
9	-70	-64	G.711	93.2	4.41	default, see Table 2	X	X
10	-70	-60	G.711	88.1	4.29	wide-band noise		X
11	-70	-50	G.711	73.7	3.77	wide-band noise		X
12	-70	-40	G.711	58.8	3.04	wide-band noise		X
13	-70	-30	G.711	43.9	2.26	wide-band noise		X
14	-100	-64	G.711	94.1	4.43	low noise	X	
15	-60	-64	G.711	88.3	4.29	narrow-band noise	X	
16	-55	-64	G.711	82.9	4.13	narrow-band noise	X	
17	-50	-64	G.711	76.3	3.88	narrow-band noise	X	
18	-40	-64	G.711	61.8	3.19	narrow-band noise	X	
19	-30	-64	G.711	46.8	2.41	narrow-band noise	X	
20	-70	-64	G.726	86.2	4.23	ADPCM (32kbit/s)	X	X
21	-70	-64	G.728	86.2	4.23	LD-CELP	X	X
22	-70	-64	G.729	83.2	4.14	ACELP	X	X
23	-70	-64	IS-54	73.2	3.74	NA mobile	X	X
24	-70	-64	G.726*G.726	79.2	3.99	ADPCM tandem	X	X
25	-70	-64	IS-54*IS-54	53.2	2.74	mobile tandem	X	X
26	-70	-64	G.729*IS-54	63.2	3.26	mixed tandem	X	X
27	-70	-64	MNRU, Q=30 dB	90.0	4.34	signal correlated noise	X	X
28	-70	-64	MNRU, Q=20 dB	67.0	3.45	signal correlated noise	X	X
29	-70	-64	MNRU, Q=15 dB	48.0	2.47	signal correlated noise	X	X
30	-70	-64	MNRU, Q=10 dB	33.4	1.75	signal correlated noise	X	X
31	-70	-64	MNRU, Q = 5 dB	23.7	1.37	signal correlated noise	X	X
32	-70	-64	MNRU, Q = 0 dB	19.0	1.22	signal correlated noise	X	X
33	-100	-100	IS-54	89.4	4.33	mobile codec w/o noise		X
34	-70	-100	IS-54	83.9	4.16	mobile codec, low noise		X
35	-60	-100	IS-54	71.3	3.66	mobile codec + noise		X
36	-50	-100	IS-54	56.7	2.93	mobile codec + noise		X
37	-40	-100	IS-54	41.8	2.15	mobile codec + noise		X
38	-30	-100	IS-54	26.9	1.48	mobile codec + noise		X
39	-55	-64	IS-54	62.9	3.25	mobile codec + noise	X	
40	-40	-64	IS-54	41.8	2.15	mobile codec + noise	X	

Table 2

Parameter	Abbr.	Unit	Default
Send Loudness Rating	*SLR*	dB	+8
Receive Loudness Rating	*RLR*	dB	+2
Sidetone Masking Rating	*STMR*	dB	15
Listener Sidetone Rating	*LSTR*	dB	18
D-Value of Telephone, Send Side	*Ds*	–	3
D-Value of Telephone Receive Side	*Dr*	–	3
Talker Echo Loudness Rating	*TELR*	dB	65
Weighted Echo Path Loss	*WEPL*	dB	110
Mean One-Way Delay of the Echo Path	*T*	msec	0
Round Trip Delay in a 4-Wire Loop	*Tr*	msec	0
Absolute Delay in Echo-Free Connections	*Ta*	msec	0
Number of Quantization Distortion Units	*qdu*	–	1
Equipment Impairment Factor	*Ie*	–	0
Circuit Noise Referred to 0 dBr-Point (narrow-band)	*Nc*	dBm0p	–70
Noise Floor at the Receive Side (wide-band)	*Nfor*	dBmp	–64
Room Noise at the Send Side	*Ps*	dB(A)	35
Room Noise at the Receive Side	*Pr*	dB(A)	35

Table 3

Transmission Rating Range	Speech Transmission Quality Category
$100 \leq R \leq 90$	best
$90 < R \leq 80$	high
$80 < R \leq 70$	medium
$70 < R \leq 60$	low
$60 < R \leq 50$	poor
$R < 50$	not recommended

Figures

Figure 1: Reference telephone connection of the E-model for network planning (ITU-T Rec. G.107, 2000).

Figure 2: Telephone line simulation model.

Figure 3: Comparison of adjusted recognition rates and E-model prediction for speech communication quality. Variable parameter: narrow-band circuit noise, *Nc*.

Figure 4: Comparison of adjusted recognition rates and E-model prediction for speech communication quality. Variable parameter: wide-band noise floor, *Nfor*.

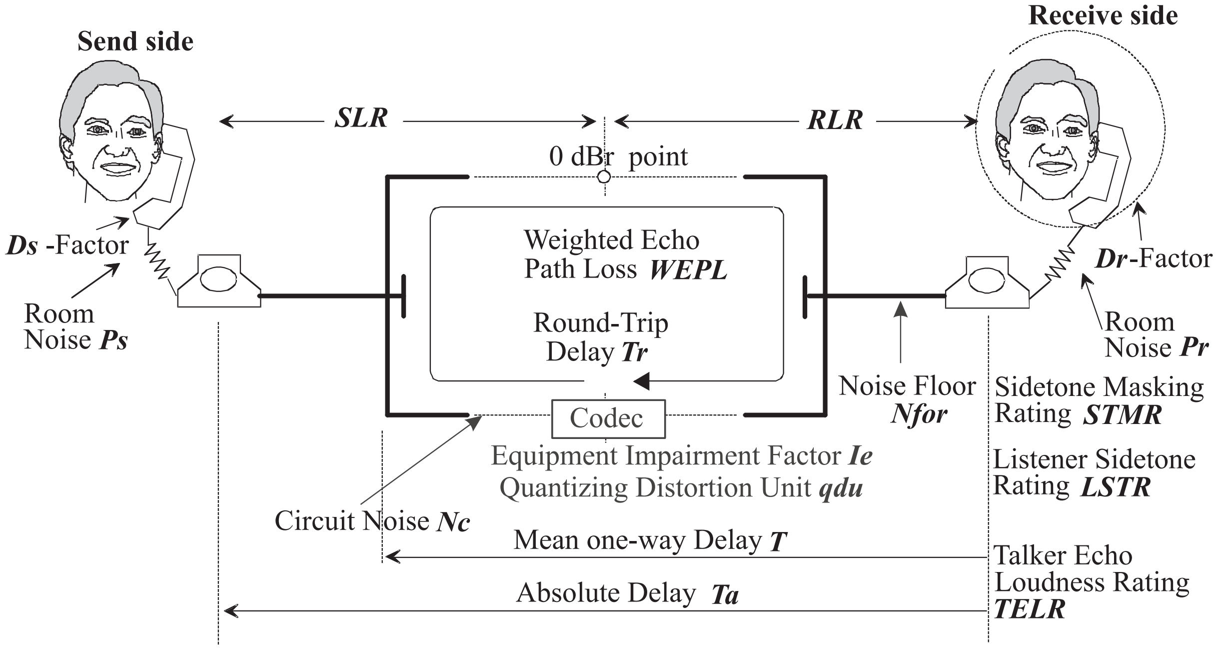
Figure 5: Comparison of adjusted recognition rates and E-model prediction for speech communication quality. Variable parameter: signal-to-quantizing-noise ratio, *Q*.

Figure 6: Comparison of adjusted recognition rates and E-model prediction for speech communication quality. Variable parameter: Codec.

Figure 7: Comparison of adjusted recognition rates and E-model prediction for speech communication quality. Variable parameter: Combination of codec and narrow-band circuit noise *Nc*. German recognizer, *Nfor* = -100 dBmp.

Figure 8: Comparison of adjusted recognition rates and E-model prediction for speech communication quality. Variable parameter: Combination of codec and narrow-band circuit noise *Nc*. Swiss-French recognizer, *Nfor* = -64 dBmp.

Figure 1



Send side
Receive side
SLR
RLR
0 dBr point
Ds -Factor
Room
Noise Ps
Weighted Echo
Path Loss WEPL
Round-Trip
Delay Tr
Codec
Dr-Factor
Room
Noise Pr
Noise Floor
Nfor
Sidetone Masking
Rating STMR
Listener Sidetone
Rating LSTR
Equipment Impairment Factor Ie
Quantizing Distortion Unit qdu
Circuit Noise Nc
Mean one-way Delay T
Talker Echo
Loudness Rating
TELR
Absolute Delay Ta

Figure 2

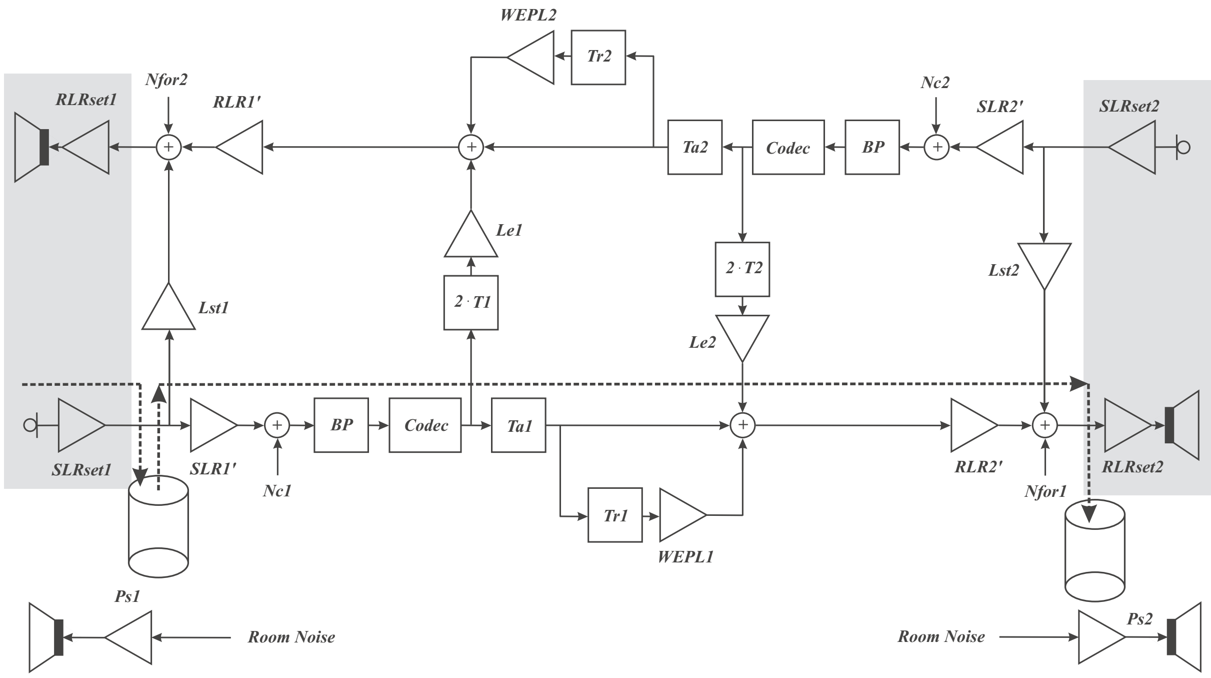

WEPL2
Tr2
Nfor2
RLRset1
RLR1'
Nc2
SLR2'
SLRset2
Ta2
Codec
BP
Le1
2 · T2
Lst2
2 · T1
Lst1
Le2
BP
Codec
Ta1
SLRset1
SLR1'
Nc1
RLR2'
RLRset2
Nfor1
Tr1
WEPL1
Ps1
Room Noise
Room Noise
Ps2

Figure 3

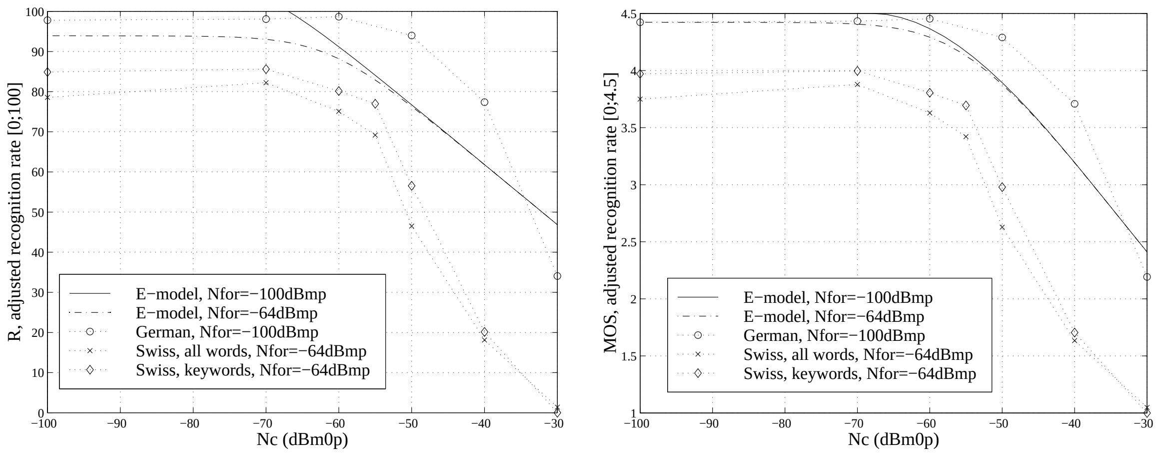
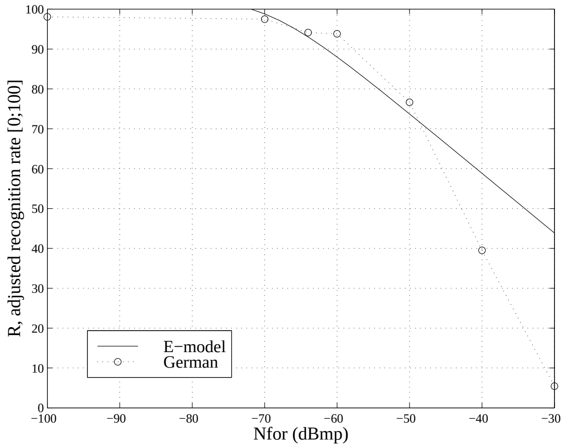
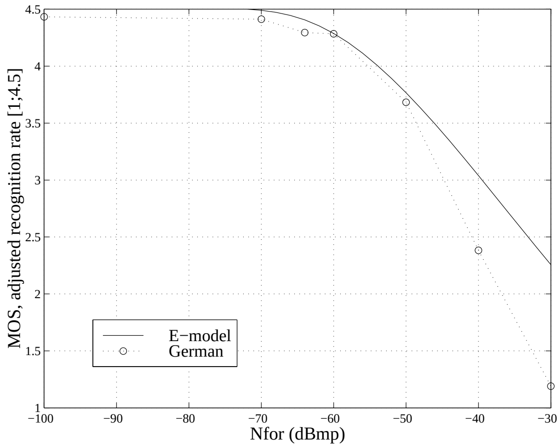

R, adjusted recognition rate [0;100]
MOS, adjusted recognition rate [0;4.5]
Nc (dBm0p)
E−model, Nfor=−100dBmp
E−model, Nfor=−64dBmp
German, Nfor=−100dBmp
Swiss, all words, Nfor=−64dBmp
Swiss, keywords, Nfor=−64dBmp

Figure 4



R, adjusted recognition rate [0;100]
Nfor (dBmp)
E−model
German



MOS, adjusted recognition rate [1;4.5]
Nfor (dBmp)
E−model
German

Figure 5

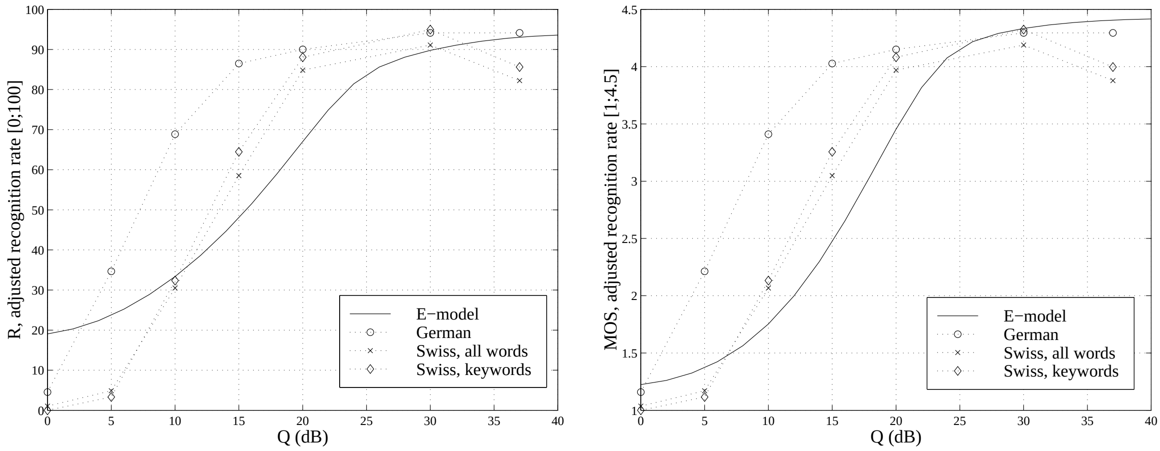
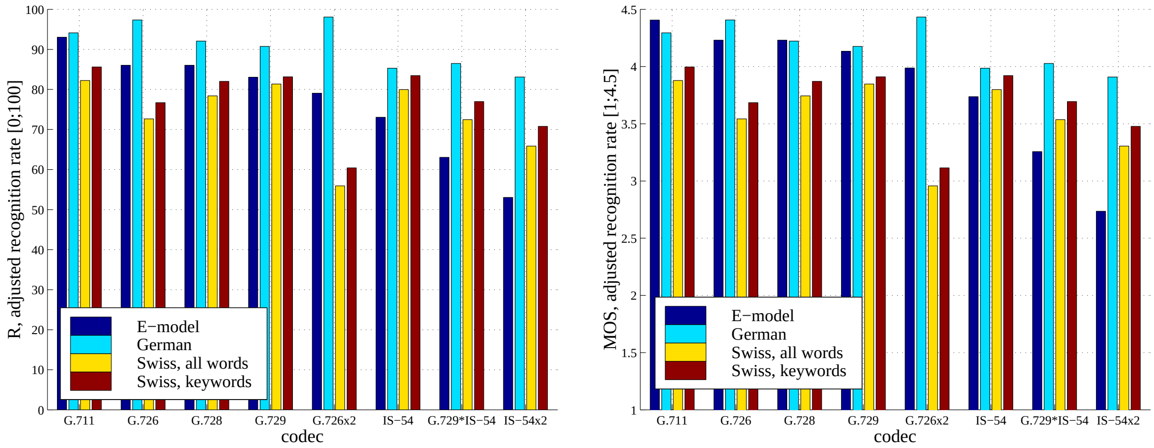

100
90
80
70
60
50
40
30
20
10
0
R, adjusted recognition rate [0;100]
0
5
10
15
20
25
30
35
40
Q (dB)
E−model
German
Swiss, all words
Swiss, keywords
4.5
4
3.5
3
2.5
2
1.5
1
MOS, adjusted recognition rate [1;4.5]
Q (dB)

Figure 6



R, adjusted recognition rate [0;100]
MOS, adjusted recognition rate [1;4.5]
E−model
German
Swiss, all words
Swiss, keywords
G.711
G.726
G.728
G.729
G.726x2
IS−54
G.729*IS−54
IS−54x2
codec

Figure 7

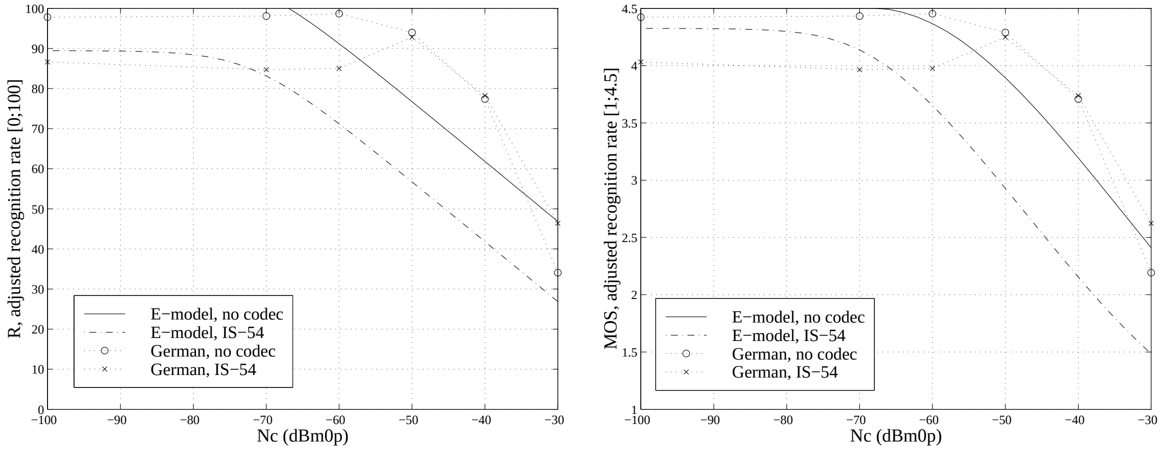
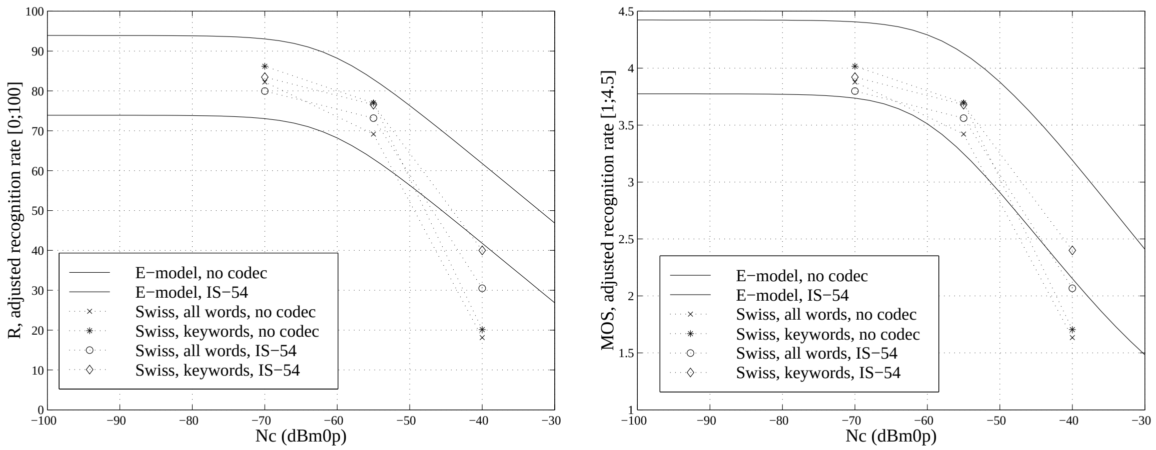

R, adjusted recognition rate [0;100]
100
90
80
70
60
50
40
30
20
10
0
−100
−90
−80
−70
−60
−50
−40
−30
Nc (dBm0p)
E−model, no codec
E−model, IS−54
German, no codec
German, IS−54
MOS, adjusted recognition rate [1;4.5]
4.5
4
3.5
3
2.5
2
1.5
1
−100
−90
−80
−70
−60
−50
−40
−30
Nc (dBm0p)
E−model, no codec
E−model, IS−54
German, no codec
German, IS−54

Figure 8


R, adjusted recognition rate [0;100]
Nc (dBm0p)
E−model, no codec
E−model, IS−54
Swiss, all words, no codec
Swiss, keywords, no codec
Swiss, all words, IS−54
Swiss, keywords, IS−54
MOS, adjusted recognition rate [1;4.5]
Nc (dBm0p)
E−model, no codec
E−model, IS−54
Swiss, all words, no codec
Swiss, keywords, no codec
Swiss, all words, IS−54
Swiss, keywords, IS−54