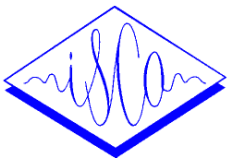

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Perceptual Dimensions of Speech Sound Quality in Modern Transmission Systems

Alexander Raake

Institute of Communication Acoustics (IKA), Ruhr-University Bochum, D-44780 Bochum, Germany
Tel. +49 234 322 3978, Fax +49 234 321 4165, E-mail: raake@ika.ruhr-uni-bochum.de

ABSTRACT

A study is reported which was carried out to determine factors governing the perception of *sound* quality of human, natural speech transmitted over telephone systems of different transmission-channel bandwidths, and their impact on perceived overall quality. The data collected in the described auditory tests is supposed to form the basis for instrumental modelling approaches of overall and especially *speech-sound* quality in wide-band systems. Measurement parameters of the listening-only test (LOT) were the lower and upper frequency limits of the transmission-channel as well as the terminal equipment used for speech presentation, with the aim of investigating the role of expectation and psychoacoustic reference for *speech-sound* quality perception. In this way, the terminal equipment best suitable for the assessment of speech-sound quality with respect to the channel frequency band was identified. With this approach, the test subjects' expectation towards different terminal equipment was taken into consideration, while the acoustical properties of the terminal were adjusted to an ideal reference. In the tests, speech samples were presented over an electrostatic headphone, monotic and diotic, as well as telephone handsets modified to allow wide-band presentation. Results show that an interdependence exists between the presentation method (handset vs. headphone) and the transmission bandwidth with respect to overall quality. The paper discusses reasons for this effect, which is assumed to be caused by the expectation listeners show towards specific types of terminal equipment.

1. INTRODUCTION

Modern telecommunication systems, such as IP based networks or ISDN, enable the wide-band transmission of telephone speech. Such systems are supposed to provide speech transmission of higher perceptual *speech-sound* quality than normal 3.1 kHz handset telephony.

On the one hand, the bandwidth and mid-frequency of the channel are relevant factors for the speech-sound quality. On the other hand, the terminal equipment used in a telecommunication system does not only influence the transmitted frequency band, but also impacts the users' expectation towards speech-sound and overall quality. Present instrumental models developed to predict or automatically assess transmission quality of specific telephone networks or network components do not consider the role of expectation adequately. Furthermore, such models mainly rely on degradations instead of improvements of the quality as compared to 3.1 kHz handset-telephone systems. Future transmission systems will not only allow wide-band transmission, but will be accessed using terminal equipment of significantly different acoustical properties as compared to classical handset telephones. Here, terminal related distortions of the speech signal spectrum transmitted to the listeners' ears will cause an altered perception of timbre and speech-sound quality. Such sound related effects, too, are only incompletely covered by the current instrumental models used in telephony (e.g. E-model, PSQM [5], [4]).

2. MODELS

Current models established to predict and assess transmission quality can be divided into three categories (see also [8]):

- Instrumental measures, mainly developed for automatically assessing speech quality of network components such as codecs, based on input and output signals of the component (e. g. PESQ [4]).
- Network planning models, for the instrumental prediction of mouth-to-ear transmission and communication quality in the planning process. The model predictions are derived from planning values of relevant network parameters (e.g. E-Model [5]).
- Monitoring models, based on planning model algorithms, where the input parameters are instrumentally measured.

The instrumental measures take the magnitude of the transfer function — more precisely the spectral shape of the transmission-channel — into consideration, but have to be enhanced to be applicable to wide-band systems. Furthermore, these models have not been trained to include terminal equipment other than handsets, so that effects of such terminal equipment on speech-sound quality are not included, especially in terms of user expectation. The current speech quality prediction models used in network planning were derived mainly from empirical studies concerning the influence of different network parameters on overall quality. These models include spectral properties of the network only indirectly. E.g., the E-model uses single numerical values (loudness ratings)

which only represent the loudness aspect of the spectral transmission characteristics. Other prediction models, on the other hand, use the transfer functions of the transmission lines as input parameter, but then translate it similarly to the loudness model. Especially for the case of wide-band transmission, where quality is improved as compared to 3.1 kHz telephony, no model exists which predicts overall quality or speech-sound quality. Therefore, it will not only be necessary to develop wide-band prediction and assessment models, but also to generally include the concept of speech-sound quality in speech quality related models. For this purpose, large amounts of subjective test data must be collected to base modelling approaches on the fundamental understanding of the underlying perceptual phenomena.

Several studies have been reported in the literature regarding quality assessment of linearly spectrally distorted speech, for telecommunication purposes as well as for other applications such as hearing aids. These studies were primarily performed in order to measure the influence of bandwidth, without focussing on the effects of the applied terminal equipment on quality (e.g. [1][3][9]). When both handset and headphones were used for presentation and the terminal was taken into consideration, the main emphasis was on the quality difference between wide-band (50-7000 Hz) and narrow-band (300-3400 Hz) transmission.

Therefore, a study is reported which was part of a series of auditory tests carried out at IKA aimed at collecting data as a basis for future approaches to modelling speech-sound quality. It was performed in order to systematically investigate the psychological role of different terminal equipment used for presentation, in combination with the filter bandwidth.

3. AUDITORY TEST SET-UP

As terminal equipment, electrostatic headphones and specially prepared handset-telephones with ideal-typical acoustical properties were employed. Three different transmission bandwidth and user-expectation related aspects and their impact on overall and speech-sound quality were investigated:

- The psychological difference between 'ideal' headphone (STAX LAMBDA PRO, monotic) and 'ideal' handset-telephone ('*Hi-Fi-Phone*'; this handset was built by using the left part of a STAX LAMBDA PRO headphone in combination with the lower part of a classical 'Type-7-handset', a model which was used in Germany for many years).
- The subjective effect of the acoustical difference between ideal and non-ideal handset due to imperfect acoustic coupling to the ear (*Hi-Fi-Phone* vs. handset, where the handset was modified by using an AKG-K-240-DF-headphone-loudspeaker).
- Comparison of monotic/diotic headphone-presentation.

The speech files for the corresponding listening only tests were recorded with six different speakers (3 f, 3 m) in an anechoic chamber. The microphone was placed 30 cm in front of their MRP, to reproduce telephone- as well as hands-free terminal-typical acoustical conditions. A shortened version of the Eurom-K [2] sentence material was read aloud by the speakers in a natural way. The recorded samples were further processed using the real-time telephone-line simulation established at IKA [7]. This DSP-simulation-tool allows all relevant transmission parameters of PSTN/ISDN-networks to be set in a defined way. The transmission-line parameters chosen for the reported tests were the default values as described in [5]. Table 1 Depicts the filter characteristics used for the tests.

Each filter-condition was applied to different sentences uttered by the six speakers. The active speech level of the samples was adjusted to –26 dBm0V to yield telephone typical conditions. The frequency characteristics of the different terminal equipment in the receive direction (and respective Receiving-Loudness-Ratings, RLRset) were determined according to the procedure described in ITU-T Recommendation P.64 (1997). The transmission characteristics of the 'normal' handset were adjusted to the Stax-characteristics, which were regarded as reference, using a corresponding correction filter RLR'. Although spectral correction of the transmission characteristics was applied, it is clear that imperfect coupling to the listener's ear — especially of lower frequencies — can occur due to different mechanical pressures when holding the hand-set and can differ from the measured properties. For the diotic presentation, the speech level was lowered by 6 dB to account for the effect of binaural loudness.

Terminal equipment	Band-width	Frequency-band.	Freq.-charact.
Handset	6950	50-7000	Flat
Hi-Fi-Phone	4900	100-5000	Flat
Stax (monotic)	3100	300-3400	Flat
Stax (diotic)	3100	300-3400	IRS
	2000	400-2400	Flat

Table 1: Tables of terminal equipment and filter-parameters. Four Terminals and 5 bandwidths yield 20 different conditions. As one sentence of each speaker is filtered and displayed under one of the twenty conditions, overall sample number is 120.

It has to be noted that no additional anchoring conditions were used during the test, as all presented qualities were corresponding to relatively good connections. Although presentation of additional 'bad' reference connections allows better comparison to other tests, the resulting quality ratings tend to narrow the range of good qualities so that differences cannot be assessed with the desired resolution. The test was divided into two sessions of 60 sample/terminal combinations each. During the tests, the speech files and the corresponding terminal equipment were randomly selected to reduce impacts of the presentation sequence on the listener ratings. Two different absolute category rating scales were employed for sound and quality ratings in order to reduce the visual similarity of the applied scales and prevent visual-driven ratings. Therefore, quality ratings were given on the ITU-T recommended 5-point MOS-scale (for possible usage of the

data within the standardization framework), while the sound quality or simply *sound* of the presented speech was rated on a one-dimensional, continuous 10-point ACR scale. 17 naive listeners with normal hearing abilities (2 female, 15 male) participated in the test. They were aged between 21 and 37 years and mostly recruited from the university's student body.

4. TEST RESULTS AND DISCUSSION

The results of the LOT show that an interdependence exists between applied bandpass-filter, terminal equipment used for presentation and perceived quality.

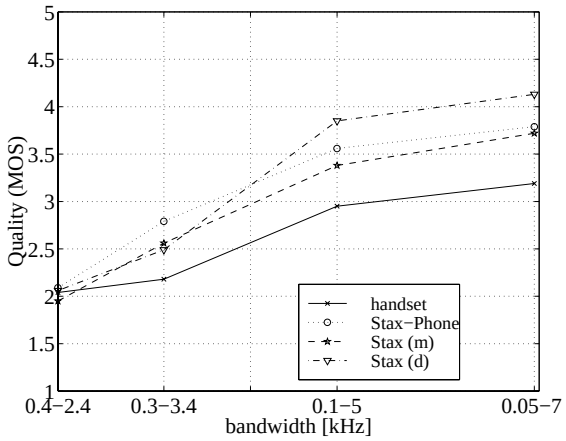

Figure 1: Overall Quality Ratings for the different terminal equipment as a function of the transmission band; all filters are flat within their transmission band.

The overall quality MOS for the flat filters show a clear preference for the Hi-Fi-Phone in case of narrow-band presentation (Figure 1). Only for higher bandwidths (higher than 4000 Hz), the diotic headphone was rated better than the Hi-Fi-Phone. With the diotic headset, wide-band quality ratings were more than 1.5 points MOS higher than for the respective narrow-band case. Comparison of monotic Stax presentation and Hi-Fi-Phone presentation reveals a small preference for the Hi-Fi-Phone throughout the whole bandwidth range, but decreasing for higher transmission bandwidth. The acoustical properties of the headphone (monotic presentation) and the Hi-Fi-Phone can be regarded as identical. During the acoustical characterization, it was found that the effect of different handset pressures and typical position variations among the listeners on the acoustic coupling of the Hi-Fi-Phone can be neglected. Thus, it can be concluded that the quality difference between monotic headphone presentation and Hi-Fi-Phone is related to the listeners' expectation towards the terminal equipment. Here, two influencing factors can be pointed out: On the one hand, monaural presentation does not correspond to common headphone usage, since the listeners expect diotic presentation. The Hi-Fi-Phone is perceived as handset-telephone and therefore — especially for lower bandwidths — the acoustic perception corresponds more to the expectation, resulting in higher quality (see [6]). It can also be concluded that listeners feel more involved in an actual telephone conversation than with headphones. On the other hand, the unusual design of the Hi-Fi-Phone will possibly have altered listener's expectation and make it appear to yield better quality than the headphones when used monotically.

A comparison of diotic and monotic headphone presentation shows clear preference for the diotic, especially for increasing transmission bandwidth. In case of wide-band transmission, the diotic presentation is rated 0.4 points MOS better than the Hi-Fi-Phone and monotic headphone presentation. This appears to be in line with the high expectation towards the acoustic properties of headphones as well as the advantage of listening diotically to high-quality signals.

The handset is rated worse than all other terminal equipment, apart from the case of the 400-2400 Hz band, where it is rated slightly better than the rest. Especially for wide-band transmission, the quality rating for the handset is significantly lower (about 1.0 points MOS as compared to diotic headphone presentation). The relatively poor performance of the handset for higher bandwidths can have three reasons: Firstly, transmission of lower frequency components might cause distortions due to mechanic resonance of the handset. Secondly, the handset does not couple low frequency components equally to all the listeners' ears due to differences in application pressure and handset position. Thirdly, wide-band speech does not match the specific expectation towards classical handsets. The latter argument is supported by another observation: When IRS characteristics were chosen for narrow-band presentation (300-3400 Hz), the quality ratings for the handset increased by 0.5 points MOS, whereas the quality ratings for the Hi-Fi-Phone and monotic headphone presentation were not affected. Here, only the diotic headphone was rated 0.2 points MOS better than without IRS, but the difference is not statistically significant (Figure 2).

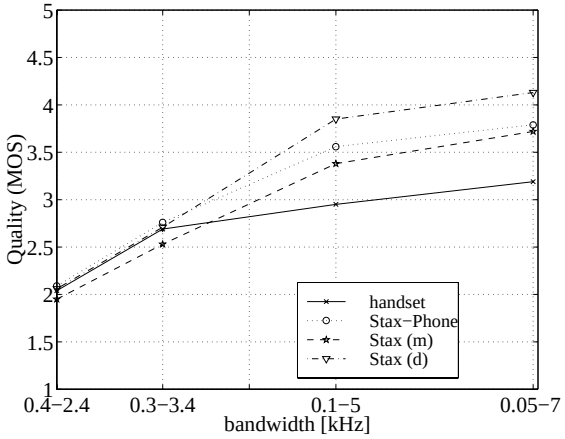

Figure 2: Overall Quality as a function of the transmission band. Here, IRS-filtering was used for 300-3400 Hz.

This implies that the acoustic perception of narrow-band speech as obtained after IRS filtering over a normal handset is closer to what the listeners expect for this specific terminal, and that sound or intelligibility are generally improved when favoring higher frequency components to a certain degree (see also [3]).

Looking at the results obtained from the sound scales, it can be stated that the ratings are comparable to those obtained for overall quality (Figure 3). Nonetheless, the test subjects seem to have used a relatively smaller range of the corresponding scale for their ratings. This might be due to the numeric annotations on both scales, leading to a tendency of rating within the range from 1 to 5 as on the 5-point MOS scale.

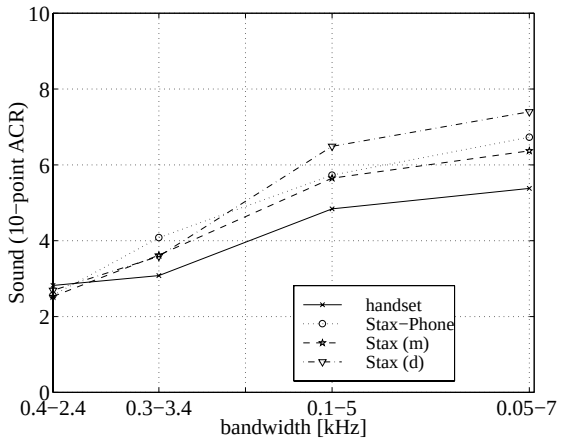

Figure 3: Sound ratings obtained for the flat passband filters.

It can be concluded that the differences in overall quality were obviously ascribed to the differences in speech-sound quality or speech-sound coloration, as no reference impairment conditions were presented. It was not the aim of this study to distinguish between the quality dimensions intelligibility and sound — 'vocal' — quality. As the mid-frequency of all filters used for processing of the samples was around 1 kHz, no priority was given to the lower nor the higher frequency components. Thus, a strong influence on intelligibility could not be expected. Nevertheless, it has to be noted that in spite of the described lower overall solution of the sound ratings the ratings for the different terminal equipment lie closer together at given bandwidths than the overall quality ratings do. This supports the interpretation of a measurable impact due to user expectation, which enters more into overall quality ratings than into speech-sound quality ratings.

5. CONCLUSION

It was shown that listeners' expectation towards the terminal equipment plays an important role for the assessment of *speech-sound* quality and overall quality, both for wide-band and narrow-band transmission. For the reported study, terminal equipment which does not introduce additional frequency distortion, as e.g. commercial hands free terminals do, was used. In this way, the focus could be laid on the basic interconnection between the transmission bandwidth, expectation linked to the terminal and resulting quality. The effect of expectation has to be taken into consideration when systematic auditory tests are performed in order to assess speech quality in wide-band transmission systems or systems with non-handset-telephone terminal equipment. Further detailed studies have to be carried out to enable modelling of the perceptual effects due to transmission bandwidth. Therefore, a terminal equipment such as the modified handset ('Hi-Fi-Phone') used in the tests seems to be appropriate to support the test subjects' impression of listening to an interlocutor during a telephone conversation. At the same time, the acoustic and expectation-related effects of the *sound quality* of non-ideal terminal equipment such as hands free terminals on speech communication quality — in receive as well as in send direction — have to be investigated. The corresponding data can then be modelled using a combined approach based on the perception of *speech-sound* quality.

6. ACKNOWLEDGEMENTS

The author would like to thank Professor Jens Blauert, head of the Institute of Communication Acoustics, for his support.

7. REFERENCES

[1] Gabrielsson, A., Schenkman, B. N., and Hagermann, B. (1985). "*The Effects of Different Frequency Responses on Sound Quality Judgements and Speech Intelligibility*". Report Technical Audiology No. 112, Karolinska Institutet, KTH, S-Stockholm.

[2] Gibbon, D. (1992). "*EUROM.1 German Speech Database*". ESPRIT project 2589 report (SAM, Multi-Lingual Speech Input/Output Assessment, Methodology and Standardization), Universität Bielefeld, D-Bielefeld.

[3] Gleiss, N. (1989). "*Desirable Sending Frequency Responses of Telephone Sets*". TELE (English edition), 1/89, 18-23, Swedish Telecommunications Administration, S-Stockholm.

[4] ITU-T Recommendation P.861 (1996). "*Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs*". International Telecommunication Union, CH-Geneva.

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

[6] Jekosch, U. (2000). "*Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*". Habilitation thesis. University Essen, D-Essen.

[7] Möller, S. (2000). "*Assessment and Prediction of Speech Quality in Telecommunications*". Kluwer Academic Publishers, USA-Boston.

[8] Möller, S., Jekosch, U., Raake, A. (2000). "*New Models Predicting Conversational Effects of Telephone Transmission on Speech Communication Quality*". To be published in Proc. ICSLP 2000, PRC-Beijing.

[9] Voran, S. (1997). "*Listener ratings of speech passbands*". Proc. IEEE Workshop on Speech Coding for Telecommunications. Back to Basics: Attacking Fundamental Problems in Speech Coding. IEEE. 1997, pp.81-2. USA-New York.