Speech quality: beyond the MOS score

Laetitia Gros, Noël Chateau, Virginie Durin

France Télécom Research & Development, 2, av. Pierre Marzin, 22307 Lannion Cedex, France,
+33 2 96 05 07 20, +33 2 96 05 13 16, {laetitia.gros, noel.chateau,
virginie.durin}@francetelecom.com

Abstract

New challenges of telecommunications operators include the development of methods and tools in order to improve customer knowledge and care. The diversity of usages, of commercial bids and the high degree of competitiveness in the voice services sector makes more complex than ever the relationship between customer satisfaction and speech quality. A review of how speech quality and its assessment are considered in literature is achieved. The conclusion of this review leads to the proposal of considering speech quality as a parameter influencing users' behaviours in telephonic communications rather than a perceptive auditory event. Two studies aiming at measuring the influence of speech quality on behavioural data in task-oriented protocols are presented.

Keywords

Speech quality, user satisfaction, assessment, methodology, behaviour.

1 Introduction

From a telecommunication operator's point of view, the understanding and the assessment of speech quality takes on two issues: on one part, it is one of the most important criteria either for improving, comparing or selecting one technology among several. This issue is particularly important in our days because, contrary to other domains in which quality reaches an upper limit, technologic evolutions in the domain of speech transmitted by telecommunication systems are not necessarily synonymous of speech quality improvement. Figure 1 shows the trends in the various transmission impairments which affect the end to end speech transmission quality of a telephone call. The diagram is intended to be illustrative rather than quantitative so the vertical axis should not be taken to be a direct measure of the end to end quality [1].

It can be clearly seen that there has been a steady reduction of the typical "analogue" impairments of noise, distortion and loss variation, largely due to the introduction of digital transmission systems, digital switching, and electronic subsets. Conversely it can be seen that there has more recently been a steady increase in "digital" impairments such as quantization distortion, delay, and jitter so that the end to end quality of service have worsen for more than ten years.

Thus, speech quality as a speech signal characteristic (*i.e.* with impairments or not) is still badly needed as criterion for improving, selecting, comparing technologies.

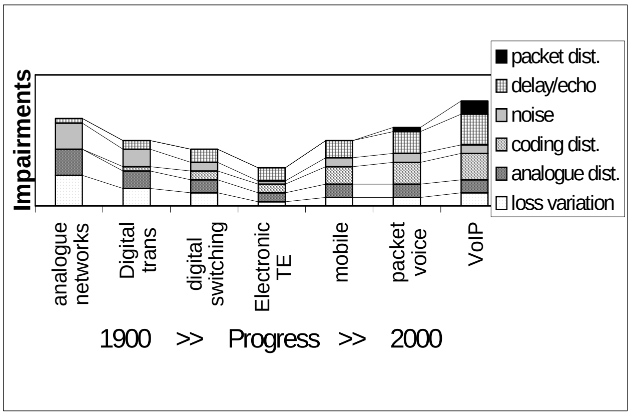

Fig. 1 Trends of end-to-end quality in terms of impairments according to technological progresses

In addition, if the relationship between speech quality and satisfaction was rather straightforward in the era of fixed RTC telephony, nowadays it is less obvious. The relationship satisfaction - speech quality has rapidly evolved with the arrival of fundamental criteria such as cost (for example low cost with VoIP), utility (for example mobility with GSM). Today, understanding user satisfaction, one needs to take into account utility factors, economic factors, contextual factors, not only environmental, but also aims and user tasks. In effect, the notion of context (physical environment, usefulness and aim of the communication, experience of the user, presence or not of secondary activities, etc.) strongly contribute to user satisfaction. Considering the diversity of services, the numerous possibilities of use, in particular geographic, etc., user satisfaction is not limited to the agreement and to the listening comfort related to a more or less degraded speech signal. A user could be satisfied with a communication of poor quality but that enables him to communicate (and give important information to his correspondent for example) in a context a priori not favourable to a communication.

In a strongly competitive context, the user satisfaction becomes crucial, and by extension, the impact of speech quality of the service on satisfaction becomes crucial too. For this issue, speech quality should not be limited to a characteristic of the transmitted speech signal, but should be more considered as a medium of communication, more or less efficient, and contributing or not to the user's satisfaction in a specific context.

.

2 Speech quality: some views

Globally researchers agree that sound quality is multi-dimensional ([2], [3], [4]), and that it can be expressed in terms of auditory sensations and in terms of pleasantness. The first aspect corresponds to a sound characterization without any emotional or hedonic judgment. In return, the second aspect of the sound quality adds an emotional and/or hedonic dimension to the simple characterisation of a sound. In the same way, speech quality is also considered as a multidimensional phenomenon [5] and can involve both speech characterisation (for example, I detect noise) and emotional/hedonic dimensions (for example, the noise disturbs me).

Parallel to the studies considering the sound quality as a sound attribute, an interesting point of view is the one of Jekosch ([6], [7]) referring to semiotics (sign theory): each sound can be seen as a sign characterized by a triadic relation between form (the significant, e.g. an acoustic form), content (the signified i.e. the meaning) and recipient (the interpreter, e.g. the listener), represented by the so-called semiotic triangle. According to this theory, there is no natural relationship between form and content, but it is the interpreter who associates a meaning to a form. Such interpretation of the perception of an auditory event as a sign introduces Jekosch to highlight that sound quality shouldn't be studied as an object but as a process. A sound doesn't have a quality in itself but a quality that one assigns to it. In other words, when one judges the quality of a sound, one does not judge an object "quality" but its balance between the perceived sound and the internal representation (implicit reference) that one has before experiment [8]. From this point of view, the quality of a sound is experienced, so dependant on a situation and on the subject who perceives and judges. In the case of telephony, quality will depend on the user expectation and on her/his experience relative to the service, as well as the motivation of the call [9]. This expectation will strongly influence the user perception and judgment. It was shown that demand is less extensive for mobile telephony and it resulted in quality judgments higher than for the fixed telephony. Conversely, for IP telephony, expectation is such as connections through Internet terminals were judged worse than the same connections (from a physical point of view), but through typical terminals. Thus one can easily imagine that nowadays internal references change with the high diversity of services (fixed, mobile, IP, uni- or multi-modal, etc.) and use contexts, and that there is not only one reference per individual any more.

However, although this speech quality view is closer to user experiences, speech quality is considered as the result of a judgement (not of the sound directly, but of its adequacy with a reference). But in everyday life, the speech quality of any service is not necessarily the object of such a judgement process from users and is not necessarily a conscious object. Generally, there is awareness and/or judgment when speech quality is as low as it disturbs the communication. In this case, the speech quality enters into the user consciousness. But most of the time, the audio quality of a telecommunication link does not raise user consciousness, unless impairments are too considerable. However, the influence of the speech quality on the realisation of the communication is strongly probable.

In effect, during the telephone communication, variations of speech quality influence users' behaviours in different ways (e.g. asking the far end user to repeat because the low speech quality did not allow comprehension, concentrating hard in order to catch some words in a noisy link, walking to get a better radio reception when communication with a mobile phone). We hypothesize that the modifications of behaviours which result from the variations of speech quality will impact overall users' satisfaction on the communication service: if the service requires many behavioural adaptations in order to handle low quality, users will be dissatisfied. As the second main speech quality issue for a telecommunication operator is the user satisfaction, it is essential to understand the relationship between speech quality and satisfaction.

3 Which methodologies for which speech quality issues?

With current methodologies, described in the ITU-T recommendation P.800 [10], the quality of the transmitted speech signal is assessed by groups of naive or expert listeners on different quality scales. The quality scales are chosen according to the different quality ranges presented within tests. These methodologies consider speech quality in its more common view: a characterization of the speech signal (Are there degradations, a lot or a few?). Sometimes the emotional/hedonic dimensions are officially involved and subjects are asked to judge the intrusiveness of degradations (DCR scale, P.835 [11], etc.), or the pleasantness of the voice (P.85 [12]). Anyway even not directly taken into account, the emotional and/or hedonic dimensions are inevitably involved, as well as the individual internal references. However, one does not know what internal references are considered by subjects, since the context is not taken into account in these methodologies. It appears that these methodologies are well suited for purposes of technologies improvement, selection or comparison.

However these subjective methods do not take into account the context that influences users' perception. Additionally, they are "intrusive" as listeners have to achieve the explicit task of judging the quality on subjective scales, activity not representative of real situations of use of telecommunications services. Merleau-Ponty tells us that what we measure with perceptive tests (perceptual events that we qualify/quantify in our consciousness) is a biased interpretation of what feeds our behaviours when we phone [13].

Based on the hypothesis that speech quality influences users' behaviour which, in turn, influences users' satisfaction, our proposition is to not ask subjects about their conscious perception of speech quality, but to observe, measure and characterize how speech quality influences his/her behaviours in task-oriented protocols. Our proposal is to study the speech quality through its impact on performances and behaviours in different communication tasks. This impact could be measured through behavioural criteria (reaction times, performances) and through electrophysiological criteria (skin conductance, heart rate, etc.) for different types of signals (natural voices, synthesis voices) and for different degradations (losses of signal packets, noise, echo, etc.).

Some studies give first elements: Mullin, Smallwood, Watson and Wilson [14] explored the electrophysiological way to measure the audio and the video quality. The considered criteria were the Galvanic Skin Response (GSR), the Heart Rate (HR) and the Blood Volume Response (BVP). Results show that different electrophysiological responses could be obtained for different degradations, and could be partially dependent of the task. Sonntag, Portele and Haas [15] measured the understanding of speech sequences uttered by six speech synthesis systems and with a natural voice, coded or not by the GSM network (so likely to be impaired) in a double-task. The primary and the secondary tasks both showed significant differences in reaction time for the different types of voices. In addition, for two of the seven voices, the differences between the coded and the non coded differences were significant.

4 First results

Two studies were conducted in order to understand how speech quality impacts performances and behaviours in different tasks involving auditory modalities with impaired speech signal. The first one [16] was based on a double-task and largely inspired by the study of Sonntag, Portele and Haas. The second one was based on a simple task, based on the dichotic listening paradigm.

4.1 Impact of speech quality in a double-task

4.1.1 Methodology

In the first experiment [16], the paradigm used in this experiment is a double-task: the primary task (a mental computing exercise) involved the auditory modality and using speech signals with different levels of quality. In the secondary task, subjects had to click with a mouse on the square, among four coloured ones appearing on a screen, whose colour matched a randomly displayed colour on the monitor. This secondary task, involving the visual and motor modalities, was used to overload the cognitive charge, and to create a stress state for the subject.

The instructions for the calculation task were given with different speech qualities: N = natural female voice, S = synthetic female voice, Nn = natural voice with MNRU at 15 dB, Sn = synthetic voice with MNRU at 15 dB and Nip = natural with packet losses (10% of packet losses, the signal was coded with a G.723.1 codec, the packet-loss law was uniform).

For each set of calculation task, subjects (twenty naive subjects) were asked to mentally compute and to orally give the answer, after each instruction, and as quickly as possible. At the end of each calculation set, subjects were asked to give their opinion on speech quality on the five-category MOS scale [2]: 5 = excellent, 4 = good, 3 = fair, 2 = poor, 1 = bad.

For each quality condition, twenty-four RTs were collected (twelve RTs per calculation set and two calculation sets per quality condition). In addition, RTs (time between the appearance of the coloured square and the subject's mouse click) were also measured. For the two tasks, the number of mistakes made by subjects in their calculations is computed.

4.1.2 Results

Figure 2 shows the mean RTs and the associated 95-% confidence intervals averaged on the twenty-four RTs per subject and on the twenty subjects, according to the quality conditions, for the primary task.

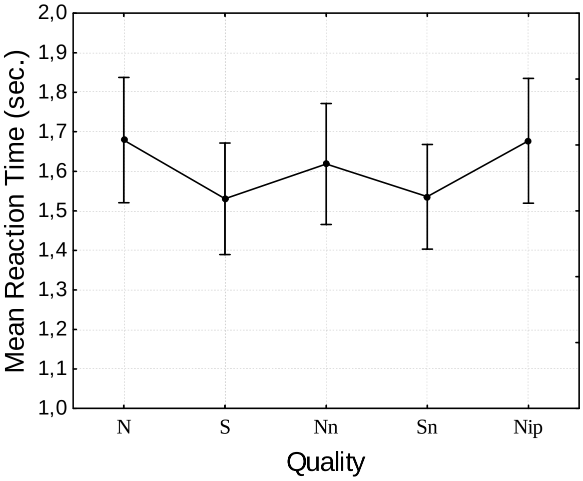

Figure 2: Mean reaction times according to the speech quality conditions, for the primary task.

It can be observed that RTs are not dependent on the speech quality of instructions. An ANOVA conducted on the RT data considering the quality (Q), Instruction (I), and Subjects (S) experimental factors shows that no significant effect of the Quality and Instruction factors were found; in return, a strong effect of the Subject factor was found (F(19, 240) = 29.01, $p<0.001$). In order to remove the variability due to the inter-subject differences, the individual RTs were centred and reduced. An ANOVA was conducted on the resulting RTs' *z*-scores. Again, no significant influence of the speech quality levels was observed (F(4, 960)=1.17, $p=0.32$). In addition, the distribution of the frequency of mistakes was not significantly influenced by the speech quality of instructions.

Now figure 3 shows RTs' *z*-scores and the associated 95-% confidence intervals for the five speech quality conditions, for the secondary task. Contrary to the primary task's RT's, it seems that secondary task's RT's are dependent on speech quality. An ANOVA was conducted and showed that there is a significant effect of the Quality factor (F(4,76)=3.84, $p<0.01$).

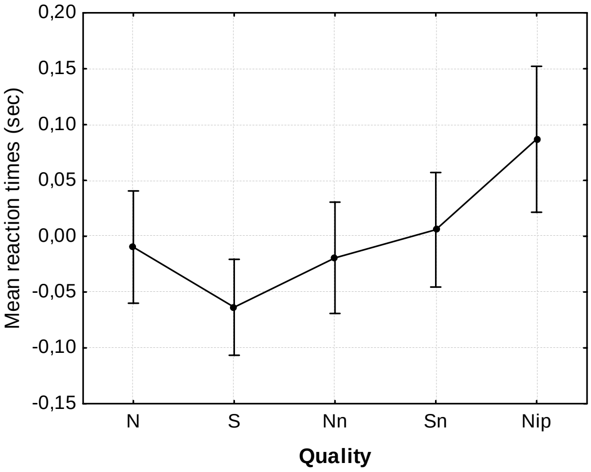

Figure 3: Reaction times' z-scores according to the speech quality conditions, for the secondary task.

Finally, figure 4 below shows the mean opinion scores (MOS) and the associated 95-% confidence intervals obtained for each quality conditions, averaged on all the subjects. As excepted, the quality judgements as measured by the typical MOS confirm that the different experimental conditions were perceived by subjects as distinct speech quality levels, from poor to good quality (confirmed by an ANOVA: F(4, 76)=28.3, $p<0.001$).

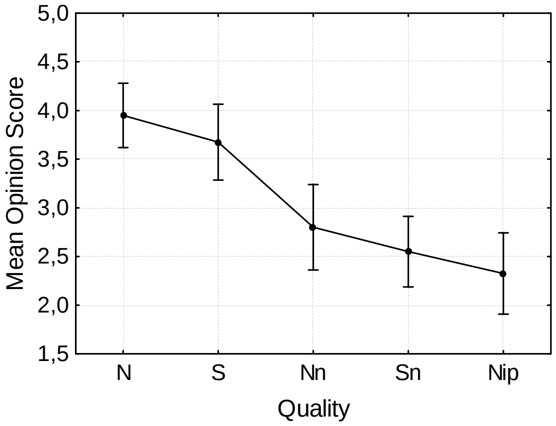

Figure 4: Mean Opinion Scores obtained for the different speech quality conditions

Contrary to [15], we did not find differences in RTs for the calculation task between natural and synthetic voices. However synthetic voices have improved during the last 7 years (diphone-based speech synthesis being replaced by corpus-based speech synthesis since 1998), and the TTS system used for our experiment provided high-quality speech signals (MOS = 3.95), close to that of natural voice (MOS = 3.68).

The speech quality factor did not influence RTs of the primary task but significantly influenced RTs of the secondary task. Concerning the primary task, mean RTs are long (as compared for instance to RTs measured for the detection of audio signals which are around 160 ms) as they result from complex and resource-demanding cognitive processes (mental calculation). Their variability is probably the consequence of the variability of the complexity of the various calculations subjects had to achieve (*e.g.*, 21-12 being more difficult than 21+1). Therefore, the potential influence of speech quality of audio instructions on RTs may have been hidden by such large variations.

However, for the secondary task, the worst speech quality (Nip, MOS = 2.3) highly lengthens the RTs. A detailed analysis of subjects' behaviour showed that, when mentally calculating (primary task), they tended to suspend their activity related to the secondary task till they orally gave their answer. This behaviour suggests that tasks were not treated in parallel but sequentially. It seems that for the worst speech quality, this suspension strategy is amplified. Thus it can hypothesized that speech quality more impacts the performances and management strategies of non-auditory subsidiary tasks than those of auditory main communication tasks. This hypothesis can be related to a bottleneck model of double-task processing. When the main task requires more cognitive resources (in our experiment, because of a worse speech quality), the attribution of additional resources is done to the detriment of the secondary task [17].

4.2 Impact of speech quality on the dichotic listening test.

4.2.1 Methodology

In this second experiment, subjects were placed in a dichotic listening situation, that presented the advantage to share the attentional process, and consequently to increase the cognitive load. Subjects listened to different lists of dissyllabic words presented on the two ears, with different and irregular rates (then overlapping between presentations on left and right ears was possible). Among all the presented words, twenty ones were pointed out as targets: they appeared successively on a PC screen face to the subject. Subjects had to detect these target words either on the right ear or on the left ear, and to indicate if the target word was on the right or on the left ear. Once detected, the following target word appeared on the screen.

Three levels of MNRU [18] (25, 15 and 5 dB) were applied to the words uttered by a male speaker. And three bandwidths (narrow band 300-3400 Hz, wide band 50-7000 Hz and super wide band 50-14000 Hz) were applied to the words uttered by a female speaker. Consequently four quality levels were considered for each type of degradation (MNRU or bandwidth), with the high quality level HQ (without any MNRU or filter) included.

The lists of words were presented at 48 kHz and 73 dB SLP through headphones to twenty-four right-handed subjects.

In addition to the two types of degradation and the four quality levels per degradation, they were two other factors: the symmetry or asymmetry of presentation of the degraded words (degradation only on one ear or on two ears), and the side of degraded presentation in asymmetrical situation (right or left).

For each condition (type of degradation x degradation level x situation symmetrical or asymmetrical x side), twenty reaction times RT (time between the end of the target word pronunciation and a keyboard touch pressing) were measured as well as eventual mistakes or absences of response.

4.2.2 Results

Figure 5 shows the reaction times z-scores averaged on all subjects and on all conditions, as a function of MNRU levels. It seems that the worse the quality, the longer the reaction times.

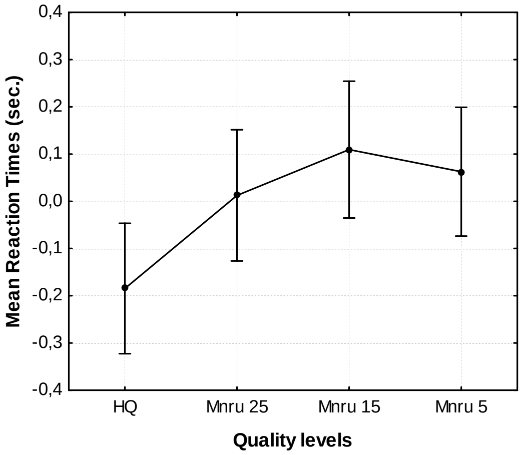

Figure 5: Averaged reaction times, as a function of quality levels.

A Variance Analysis on individual RT considering the factors Situation (Asymetrical / Symetrical), Quality levels (HQ High Quality, MNRU 25, MNRU 15, MNRU 5) and Ears (right or left), confirms the effect of quality (F(3,714)=18,56 *p<0,0001*). A HSD Tukey test shows that the mean RT for the HQ level is significantly longer that the three other mean RTs obtained for the three quality levels.

In addition, the Variance Analysis shows a significant effect (F(1,238) = 23,81 *p<0,0001*) of the Ear on which the sound is presented. Figures 6 and 7 show the mean RT respectively for the symmetrical and the asymmetrical reaction times, measured on each ear, for the group A (degradations presented on the right ear) and for the group B (degradations presented on the left ear). For the High Quality condition, it can be seen that RT for the right ear are shorter than for the left ear, whatever the situation.

In the symmetrical situation, whatever the group considered, the two ears are impaired in the same way (with the same quality levels). One observes that the right ears are globally more effective than the left ears (RT are shorter). In the asymmetrical situation, when degradations appeared on the right ear, the performances of this ear became similar to the left ear (figure 7, group A). In return, when degradations appear on the left ear, an increase of the difference between the two ears can be observed (figure 7, group B).

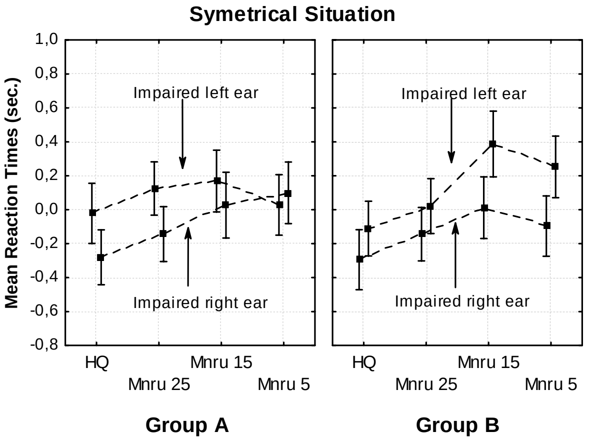

Figure 6: Averaged reaction times, as a function of quality levels, and ears, for the symmetrical situation

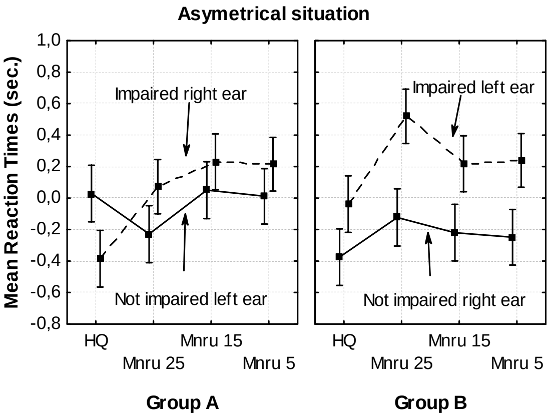

Figure 7: Averaged reaction times, as a function of quality levels, and ears, for the asymmetrical situation

Thus, right ear seems to be more effective, at least faster, that the left ear to detect the target words. In addition, it seems that the two are roughly independent (the degradations appearing on one ear seem not to disturb the other ear). Error rates confirm a predominance of the right ear on the left ear (less detection errors on the right ear) in the target words detection task in a dichotic listening situation. Similar effects, although weaker, are obtained for the bandwidths degradation type.

The predominance of the right ear for the word detection in dichotic listening corroborates the Wernicke-Geschwind model: speech is predominantly processed in the left brain area connected to the right ear.

5 Conclusion

New challenges of telecommunications operators include the development of methods and tools in order to improve customer knowledge and care. The diversity of usages, of commercial bids and the high degree of competitiveness in the voice services sector makes more complex than ever the relationship between customer satisfaction and speech quality. Perceived speech quality is a specific object of consciousness that is frequently assessed through subjective opinions in telecommunications operators' research labs. Although useful for many technological applications in labs, we suggest that these assessments are not of great use today for determining how speech quality influences customer satisfaction in real life.

In order to better characterize this influence, we suggest to observe, measure and characterize behaviours of subjects in task-oriented protocols depending on speech quality. Two first studies aiming at this goal were presented in this paper. They show that speech quality, as it can be characterized by physical parameters (percentages of packet losses, MNRU, etc.) influences behavioural data such as reaction times or percentages of stimuli detection. They also show that the data pattern of the dependent variables is highly dependent of the experimental protocols (single/double tasks), highlighting the extreme caution that has to be taken when building task-oriented protocols.

These results are encouraging as they show that the influence of speech quality on behavioural data can be systematically assessed. Further research shall concentrate on the definition of a series of task-oriented protocols allowing to measure the influence of speech quality on a set of behavioural data. These data should be representative of the modifications of customers' behaviours (cognitive and physical processes) due to speech quality variations in realistic communication scenes.

6 References

[1] ETSI EG 201 474 V1.1.1 (2000-04) "Speech Processing, Transmission and Quality Aspect (STQ); Future approaches to speech transmission quality across multiple interconnected networks."

[2] Guski, R. (1997). "Psychological methods for evaluating sound quality and assessing acoustic information". *Acustica* - *Acta Acustica*, 83, 765-774.

[3] Gabrielsson, A. (1979). "Perceived sound quality of sound-reproducing systems", *Journal of the Acoustical Society of America*, 65(4), 1019-1033.

[4] Susini, P., McAdams, S., Winsberg, S. (1999). "A multidimensional Technique for Sound Quality Assessment". *Acta Acustica*, 85, 650-656.

[5] Schroeder, M.R. (1968). "Reference Signal for Signal Quality Studies". *Journal of the Acoustical Society of America*, 44(6): 1735-1736.

[6] Jekosch, U. (1999-a). "Meaning in the Context of Sound Quality Assessment". *Acustica*, 85, 681-684.

[7] Jekosch, U. (1999-b). "The perception of product sound quality". *Acta Acustica,* Joint Meeting ASA/EAA/DEGA, Forum Acusticum, Berlin, 85, 352.

[8] Blauert, J., Jekosch, U. (1997). "Sound quality evaluation. A multi-layered problem." *Acustica* - *Acta Acustica*, 83, 747-753.

[9] Möller, S, Riedel, J. (1999). "Expectation in quality assessment of Internet telephony". *Acta Acustica*, Joint Meeting ASA/EAA/DEGA, Forum Acusticum, Berlin, 85, Suppl.1.

[10] ITU-T, P.800 (1996), "Methods for subjective determination of transmission quality." http://www.itu.int.

[11] ITU-T, P.835 (2003) "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm" http://www.itu.int.

[12] UIT-T P.85 (1996). "A method for subjective performance assessment of the quality of speech voice output devices." http://www.itu.int.

[13] Jekosch, U. (2005) "Voice and Speech Quality Perception. Assessment and Evaluation", Eds. Sringer. Signals and Communication Techology.

[14] Mullin, J., Smallwood, L., Watson, A., and Wilson, G. M. (2001) "New Techniques for Assessing Audio and Video Quality in Real-Time Interactive Communication," In: J.Vanderdonckt, A. Blandford & A. Derycke (eds.) Proceedings of IHM-HCI, pp221-222, Lille, France, September 10th - 14th.

[15] Sonntag, G.P., Portele, T. and Haas, F. (1998) "Comparing the comprehensibility of different synthetic voices in a dual task experiment", In: SSW3-1998, pp5-10.

[16] Gros, L., Chateau, N, Macé, A. (2005), "Assessing speech quality: a new approach", Forum Acusticum, Budapest 2005, 4th European Congress on Acoustics, 29 August-2 September 2005, Budapest, Hungary.

[17] Pashler, H. (1994) "Dual-task interference in simple tasks: Data and theory", *Psychological Bulletin*, 116, pp220-244

[18] ITU-T, P810 (1996), "Modulated noise reference unit (MNRU)", http://www.itu.int.