Redrawing the link between customer satisfaction and speech quality

Noël Chateau, Laetitia Gros, Virginie Durin, Amélie Macé

France Telecom, R&D Division
Technologies R&D Center
noel.chateau@orange-ft.com

Abstract

This paper analyzes the actual usage of Mean Opinion Scores (MOS) in the telecommunication industry and looks back to the tremendous evolutions that has encountered this sector since subjective test methodologies were standardized by ITU-T. It is pointed out that MOS have severe drawbacks and that test methods should evolve by taking into account valuable inputs from connected research fields such as the psychology of emotions, cognitive science, usability engineering and marketing research.

1. Introduction

In a highly competitive market, price and quality of services are two key issues that telecommunications operators have to take care of. Whereas price issues are generally only handled by marketing units, those on quality are the concern of many units, from marketing, networks, to R&D ones. R&D centers, which have a long experience in running subjective tests aiming at assessing speech or video quality of technological components (such as codecs, noise-suppressor or other signal-processing algorithms) are now asked by marketing units to produce data that reflect customers' opinions on quality and their subsequent satisfaction in their daily use of services. As a consequence, the 'famous' MOS (*Mean Opinion Score*) is not restricted any more to research labs, but is extensively used by marketers.

Figure 1 from ETSI [1] shows the trends of end-to-end quality in speech telecommunications as a function of technological progress. Till the end of the eighties, the different technological improvements in the field of telecommunications were synonymous of improved quality delivered to the users. Mobile networks and packet voice have definitely introduced two revolutions, one on usage, the other one on pricing and added services, but they have also introduced new uncontrolled speech impairments.

Before the introduction of these technologies and the apparition of competition on the European market that both occurred in the nineties, speech quality of voice communications was guaranteed and rather homogenous among users and over time. Moreover, the usage was restricted to land-line communications, only few terminals were available, and the pricing was fairly simple and the same for all users. Thanks to this homogeneity of quality, usage, terminals and pricing, the MOS measured in research labs were considered as good predictors of users' opinion and satisfaction of the service, since almost no other criteria would come into the scene. Today, we have to face a strong heterogeneity of quality, usages, terminals and pricing. In such a context, it is probable that the MOS measured in labs do not reflect as well as thirty years ago users' (who became customers) opinion on the quality of the voice communication services, and therefore their satisfaction.

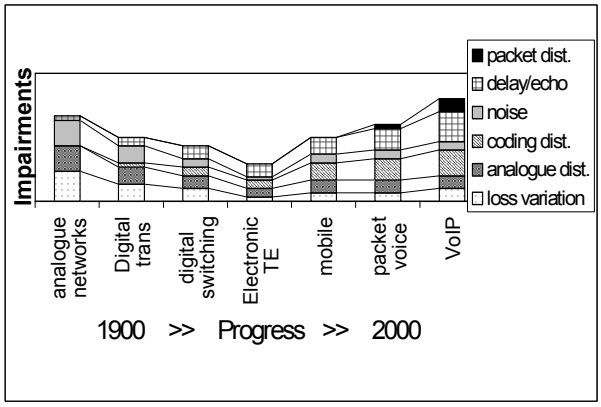

Figure 1: Speech quality impairments as a function of technological progress

This paper reviews some of the major drawbacks of the extensive use of MOS in today's context and proposes to consider speech quality in a broader context, including emotional reactions and behavioral adaptations of users of voice telecommunication services facing low speech quality periods. Inputs from the research field of the psychology of emotions, and notably from the appraisal theory developed by Scherer [2] are discussed in order to propose a refined definition of customer satisfaction in the field of voice telecommunication services. Methodological implications are discussed and directions for future work are proposed in the last section.

2. Three major drawbacks of the Mean Opinion Score

The MOS is undeniably a very useful tool for diagnosing and assessing technological components of voice communication services [3]. MOSs generally refer to Mean Opinion Scores obtained by averaging subjective data collected in ACR-listening tests [4]. However, MOS can also be established in conversation tests [4]. Additionally, the development of objective measurement tools such as PESQ [5] has extended the concept of MOS to the field of objective quality. Although

there is a wider and wider use of MOSs, we wish to point out three of their major drawbacks.

2.1. MOS and the quality of voice communication services

As discussed in the introduction, MOSs were rather convenient before the nineties for predicting users' opinion on the global quality of voice communication services. After the mobile and the VoIP revolutions, voice communication services have become complex services, including new functionalities (*e.g.* messaging, directory, chat), being accessible from a wide variety of terminals (fixed, cordless, mobile), through various speech qualities (associated to landline, mobile, VoIP networks, their specific codecs and all the tandeming possibilities), in various contexts of usage (at home, in the streets, in public/private transportation, etc.). This explosion of heterogeneity has blurred the link between speech quality, as measured in laboratories, and service quality, as perceived by users/customers. However, there is a growing demand of marketing and network units for producing speech quality data, as perceived by customers, in order to "market quality" for the first, and to monitor it on networks for the second. Such an extensive use of MOS data and their related extrapolations are really dangerous, since many non-experts use them "as the truth", reflecting end user's opinion, which is of course not the case. Although one way of limiting the problem is obviously communicating with pedagogy on the limits of the interpretation and extrapolation of MOS data, there is also an urgent need for developing new methods allowing to producing data that are closer to end users' perception of service quality.

2.2. Corpus dependence

MOS values are obtained by averaging ratings on a category scale. Such scales have a rather serious defect since the data they produce are dependent on the corpus distribution [6]. The bias comes from the fact that whatever the corpus is, subjects tend to assign stimuli to categories in such a way that all categories are used about equally often [6, p.108]. If the corpus contains an under or an over-representation of low or high-intensity stimuli, the psychophysical function is transformed by a non-linear function. In the case of speech quality assessment, the use of corpuses with only narrow band stimuli, or with mixed narrow and wideband stimuli typically exemplifies this problem. [7] have shown that narrow-band stimuli see a shift of their MOS between 0.01 and 0.63 when the narrow-band corpus is extended to a mixed narrow and wideband one. In the context of deployment of wideband speech in VoIP services (*e.g.* Orange HD and AOL in France, BT Broadband in UK, Skype and MSN worldwide), the instability of the MOS scale is rather critical to clearly compare quality levels. This topic is currently discussed at ITU-T, where [8] has proposed to used two different MOS scales (one for narrow-band corpuses, one for mixed corpuses).

2.3. Explicit assessment

Up to now, most of researchers have considered sound quality as an attribute of sound that can be expressed in terms of auditory sensations or by a hedonic judgment ([9]). In many studies, sound quality is either characterized by the qualification/quantification of the perception of the acoustic features of the stimuli (*e.g.* [10]) or by hedonic judgments such as the pleasantness of the stimuli (*e.g.* [11]) or their disturbance/annoyance (*e.g.*, [12]).

Jekosch [3] has introduced an interesting point of view based on semiotics: 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). 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. This implies that sound quality should not be studied as an object but as a process: the process of semiosis. In this process, the listener does not judge the object "sound quality" but its balance between the perceived sound and his/her internal representation (implicit reference). From this point of view, the quality of a sound is experienced, and therefore dependant on the context and on the listener. In the case of telephony, quality will depend on user's expectation and on her/his past experience relative to the service, as well as the motivation of the call [13]. This expectation will strongly influence user's perception and judgment.

Although Jekosch's point of view brings speech quality much closer to customer's experience, it is nonetheless based on an explicit judgment, which is not likely to take place in real life. In ecologic situations, speech quality is generally not a conscious object for users, excepted if it reaches a threshold below which it becomes difficult for them to communicate. To make it conscious and to be assessed, it has to become "a topic for thought" through the process of reflection [3, p. 55]. For Merleau-Ponty, this "intellectualization" of quality is a mistake [14, p.11] that he describes by the following sentence: "we consider quality as an object *of* consciousness whereas is it an object *for* consciousness."

As a consequence, it might be argued that if one wants to study the object "speech quality" as closer as possible to what it represents for an end user in an ecologic communication situation, he/she might not directly ask the user's opinion on speech quality but find alternative methods. We propose that speech quality can be indirectly analyzed through its impact on user's behavior. This issue will be discussed in the following section.

3. Speech quality, behaviors and emotions

3.1. Speech quality and behavior

When communicating with a telecommunication service, users pursue one or several goals (assuming that they also pursue parallel goals not related to the communication such as the basic one of surviving, or buying some food, etc.). As highlighted by Jekosch [3], they will have certain expectations about the quality of the service, dependent on their goals, the context and their past experience. According to ISO's 9001 definition of customer satisfaction [15], if the service meets their expectations and helps them to reach their goals, they will tend to be satisfied with it. Inversely, if the service does not meet their expectations and does not help them (enough) to reach their goals, they will tend to be unsatisfied. Speech quality is one of the elements of quality of the telecommunication system/service. If it is high, users will not encounter specific problems to orally communicate with their interlocutor. It is highly probable that they will not even pay attention to speech quality and that it will not be an

explicit object of their consciousness. On the contrary, if it is low, below a certain threshold, it will become a conscious object and will impact their behaviors. Typical behavioral adaptations to low speech quality will be for example increasing concentration, asking the interlocutor to repeat or to speak louder, walking away (from a noise source or to find a better network reception), terminating the communication and trying a few seconds later, etc. These behaviors can be essentially cognitive (*e.g.*, increasing concentration, that might only by expressed by a facial mimic such as frowning eye brows) or essentially motor (*e.g.* walking to find a better reception for mobile phones) or both (*e.g.*, deciding to interrupt the communication and dialing again immediately after). These few examples clearly show that the observation of users' behaviors might be a valuable source of information on speech quality and on its impact on subsequent users' satisfaction of the service. They also show that beyond considering speech quality as an element of quality of the telecommunication system, it might be considered as an *enabler* of the communication, enabling the user to reach his goal if it is high, and hindering him/her if it is low. Another enabler is for example the ease of use of the terminal/service that might be assessed through usability tests.

3.2. Emotion and behavior

The first theories on emotion, such as the *arousal theory* or the *discrete emotion theory* [16] mainly considered full-blown emotions (often referred as the big six after the work of Ekman [17]: anger, disgust, fear, sadness, surprise and happiness). These emotions are generally elicited by straightforward and strong stimulations. Classic examples are: seeing a snake triggers fear, eating aversive food triggers disgust, hearing a sudden loud sound behind triggers surprise, etc. Research on emotions in the past forty years has considerably refined the concept of emotion and its associated theories. Today, one of the most comprehensive and accepted emotion theory is the *appraisal theory*, which takes into account not only full-blown, but complex and blended emotions such as boredom, irritation, the mix of relief and fun, etc. [2, 18]. This theory considers that emotions are based on subjective evaluations of the signification of events which have a matter with the well-being of the concerned person. Lazarus and Folkman [19] have introduced the concepts of primary and secondary appraisals. The primary appraisals deal with the pleasantness/unpleasantness of the object/event under evaluation, and whether it helps or hinders satisfying a need or reaching one's goal. The secondary appraisals are related to what extent the person can cope with the consequences of the object/event under evaluation.

Full-blown emotions are considered as being associated to automatisms that can bypass cognition in certain conditions (*e.g.* seeing a snake => full blow emotion of fear => running away and shouting). Except these specific cases where little control seems to be left to the person who is subject to the full-blown emotion, it is generally admitted that emotions afford behavioral flexibility and play a central role in the regulation of cognitive and motor behavior [20]. Emotions are also essential for triggering attention on and memorizing objects/events [21]. Actually, emotion components include most of all parts of psychological functioning (feeling, physiological changes, motor expression, action tendencies, cognitive processing). Emotion processing is therefore a complex mechanism that takes place in the whole organism (the body and the brain) and which leads to consider the specificity of an *emotion episode* (as opposed to a global psychological functioning) [2] only as a question of threshold in intensity and duration of cognitive, physiological and motor activity.

3.3. Appraisal theory in the context of voice telecommunications

As discussed above, during a voice communication, users will generally not pay a specific attention to speech quality. However, for a given context (and the associated goals and expectations), if speech quality falls below a certain threshold, it will become a conscious object, grab their attention and trigger behavioral adaptations. From appraisal theorists' point of view, these behavioral adaptations can be considered as the consequence of the role that play emotions in the subjective evaluation of the signification of events affecting a person trying to reach a goal [2, 22]. Said differently, if the enabler "speech quality" does not fulfill its role (enabling the user to communicate easily with his/her interlocutor), it will be appraised as an object that triggers an emotional reaction which in turn will trigger a (cognitive, physiologic, motor) reaction and a behavioral adaptation (concentrating, making the interlocutor repeat, moving, etc.). As a consequence, it might be argued that studying users' behaviors *and* their emotional reactions might be a valuable source of information on speech quality and on its impact on customer satisfaction.

4. Customer satisfaction

4.1. Customer satisfaction and resources consumption

During a voice telecommunication, to reach their goal, users will spend four types of resources:

- *cognitive resources:* perception, attention, information processing, reasoning, decision making, etc.
- *physical resources:* physical resources required for cognitive processing, for speaking, for dialing and holding a handset, moving away from a noise source, etc.
- *telecommunication system's resources:* capacity of delivering and capturing a speech signal by the terminal, battery, display of information on the screen of the terminal, connection to the network, etc.
- *economic resources:* cost of the service for the user.

For a given context and a given goal, users have a certain expectation about the resources they need to spend to reach their goal and about the quality of the system they will use. For example, if, at the top of a mountain, a climber wants to communicate with a mobile phone to his mother that he reached the summit, he might expect to encounter a poor availability of the service and a poor speech quality (because of the low network reception at this location) and a difficult context for communicating vocally (*e.g.* presence of noisy wind) that will affect his consumption of resources: first, he might try to phone his mother three times before getting the communication; second, he might need to concentrate hard on her voice because of the low quality and to shout because of

the noisy wind around. Moreover, because of the low network reception, the consumption of the battery power of his mobile phone might also be important. Despite this high consumption of resources and the low speech quality encountered, if the climber reaches his goal, he will certainly feel very happy and rewarding to the telecommunication service. Let us take the same person, wanting to phone his mother for a little chat, sat in his sofa, at home, with a classic PSTN service and a cordless handset. His goal, context and expectations will be completely different and he will certainly not tolerate to encounter such a low speech quality and to spend as many resources as in the mountain scenario. It is probable that at the end of his chat, he will not be specially rewarding to the telecommunication service. Moreover, in the mountain scenario he expected speech quality to be low, consciously felt it was the case and adapted his behavior in consequence. On the opposite, it is highly probable that in the sofa scenario, he did not even consider speech quality during the communication and that speech quality did almost not affect his general behavior.

Between these two extreme scenarios, there is a wide variety of contexts, usages and encountered speech qualities by users that will determine the impact speech quality has on their behaviors, on the several resources consumptions and on their overall satisfaction of the service. From what has been discussed previously, it appears that if one wants to study in the lab the object "speech quality" as closely as possible to what this object means in ecologic situations, it implies to conduct task-oriented tests, where speech quality will be a controlled enabler to reach a given goal, and in which subjects' behaviors (measured resource consumptions and performance), emotions (expressed, measured and self-reported) and satisfaction will be recorded and analyzed. The development of such a methodology will be discussed in section 5.

4.2. A definition of customer satisfaction

ISO's 9001 definition of customer satisfaction is: "customer's perception as to whether the organization [service] has met his/her requirements". This general definition does not specify whether the service has allowed the customer to fully or partially reach his/her goals, and which and how many resources he/she had to spend during the usage of the service. If we go back to Lazarus' concepts of primary and secondary appraisals, we see that those appraisals can be applied to the usage of a service and that they allow an extension of ISO's definition. During a voice telecommunication, Lazarus' primary appraisals will concern the pleasantness/unpleasantness of the different objects/events that will encounter the user, and whether these objects/events will help or hinder him/her to reach his/her goal. Secondary appraisals will concern to what extent the user can cope with the consequences of these objects/events (*e.g.* how to manage a poor speech quality during an important commercial call on VoIP? Consequences of an important price to pay or of a battery-power failure resulting from a long-lasting communication on a mobile phone). In Lazarus' view, the outputs of these appraisals are complex emotions that will regulate user's behavior. We suggest here that the output of these appraisals also contribute to user's (customer's, if we consider the economic link between the user and the phone operator) satisfaction, since they are the reflection of his/her evaluation of the different elements of the service. The outputs of the appraisals are:

- primary appraisal: pleasantness and contribution of the service to reach user's goal. Pleasantness refers to the hedonic judgment of the elements of the service. For example, a voice telecommunication service with a high speech quality might offer the same contribution to a user to reach his/her goals but might be considered as more pleasant if the speech is wideband rather than narrow band. The contribution of the service to reach user's goal refers to the usability of the service and to the degree of completion of his/her goal resulting from this usability. For example, a complex terminal or a low speech quality service will complicate the communication and will hinder the user to reach his/her goal, therefore necessitating a behavioral adaptation of the user and an increase of resources consumption.
- secondary appraisal: extent to which the user can cope with the consequences of the elements of the service under consideration. As discussed in paragraph 4.1., the consequences of the variations of speech quality will be an increase of resource consumption. For a given context and a given goal, the user will tolerate only to a certain extent to spend more cognitive, physical, economic or system resources to compensate a low speech quality.

Considering the above elements, we propose the following definition of customer's satisfaction for voice telecommunication services:

For a given context of usage and a given goal, customer's satisfaction is the result of his/her appraisal of the service according to:

- *the degree of completion of his/her goal reached thanks to the service;*
- *the usability of the service and the amount of resources spent to reach this degree of completion;*
- *the consequences for him/her of the amount of spent resources to reach this degree of completion;*
- *the pleasantness of the elements of the service*

as regard to his/her expectations and to prior experience.

Speech quality influences the first, second and fourth item of this proposed definition. This influence might be assessed through user tests which will be discussed in the following section.

5. Methodological implications

In the growing context of heterogeneity (of quality, usages, terminals and pricing) described previously, considering the strong demand from marketing and network units for producing useful and reliable data on speech quality, and the three drawbacks affecting MOSs that were presented in the second section, there is an urgent need to develop test methodologies which furnish data that:

- are not dependent on the range of speech qualities under examination during the test;
- are obtained without explicit judgment from subjects on speech stimuli;
- better reflect the role that plays speech quality in user's global opinion on the quality and satisfaction

of telecommunication services where voice quality is an important matter.

5.1. Task-oriented protocols and motivation

As discussed in the third section, in a voice telecommunication context, speech quality appears to be an enabler that will influence user's emotional reactions, behavior and satisfaction. In order to evaluate and characterize these three dimensions, one needs to define and measure dependent variables that should vary according to speech quality levels under test. To obtain such a dependence, it is necessary to involve subjects in tests where they can develop a high motivation and where speech quality appears to be an important aspect for them (even if not explicitly described as such by the experimenter).

The problem of motivation in user tests is fundamental. In usability tests, when new services or new functionalities are tested, it is firstly tackled through a careful selection of subjects: only current or potential customers of the service under test are recruited. Second, task-oriented protocols are used where subjects have to achieve tasks which are similar or close to those they want to achieve when using the service (or an equivalent for potential customers) in real life. Our concern here is to develop test methodologies where subjects are not explicitly asked to give their opinion on speech quality (at least during the test). Consequently, we have to define a test where the goal and the motivation are not explicitly but implicitly connected to speech quality.

Research on the psychology of emotion shares the same concern of manipulating implicit variables in user tests where task-oriented protocols are used and where a high motivation of subjects is required in order to obtain realistic and spontaneous expressions of emotions [23, 24]. There are two families of protocols emerging from these studies. The first one is based on using an initial personal motivation of users, *e.g.* learning a new language. Aubergé et al. [25] used such a motivation in the e-Wiz experiment in order to obtain multimodal expressions of emotions. An application for learning a foreign language was proposed to motivated users who were manipulated in order to express emotions (by specific feedbacks of the application), without knowing that the aim of the experimenters was not to help them to learn this foreign language but to collect a multimodal emotional corpus. Such a protocol requires subjects to come regularly to the lab and is time consuming.

The other family of protocols is based on gaming. The natural tendency of many subjects to find a challenge and a motivation in a game where there is a competition (against the system or other participants) is used to obtain spontaneous emotional reactions [26]. The main advantages of game protocols are that they can be really challenging for users, that many test conditions can be tested in a relatively short period and that performance (the degree of achievement of the proposed tasks) can be rather easily measured. One of their major drawbacks is that performance is also dependent on prior competences of users in gaming. Simulating a competition where there is a financial reward for the winner can be an additional factor of motivation that is used in labs.

Gros et al. [27] used a simple game consisting in mentally computing as fast as possible without a mistake, while clicking also as fast as possible without a mistake on a specific colored square among four appearing on a computer screen. Computing instructions were given by playing back vocal instructions recorded at different speech qualities. Gros and her colleagues showed that speech quality did not influence the performance of the computing game, but that of the clicking one, according to a bottleneck model of double-task processing [28]. This result shows that the influence of speech quality might be studied in double-task protocols and observed in the non-auditory task.

5.2. Methods for user's emotional reaction assessment

Emotion is a psychological construct than can only accessed through the analysis of its expression by the person under consideration or by its own introspection. Research on emotion in the last decades has considerably intensified and has furnished a large range of methods for recording, analyzing and interpreting emotional signs expressed by humans. The first category of methods refers to the recording and the analysis of electro-physiological signals such as skin conductance, heart rate and blood volume pressure [29]. During emotion episodes, these signals reflect the activity of the sympathetic system and allow an interpretation of the emotional arousal. However, the valence of emotions can not be determined by the sole analysis of electro-physiological signals and other information are necessary [29].

A complementary source of information can be to realize audiovisual recordings of subjects during the test and to analyze them *a posteriori*, video (facial mimics, gestures, body postures) and audio (voice prosody, vocal quality, volume, lexical content) signals being a valuable source of information for characterizing humans' emotional states [30].

Audiovisual recordings of subjects' experience (the camera recording their point of view) during the test can also be made. These recordings are played back to subjects just after the test in order to immerse them back to what they just experienced. During this play back, they are asked to comment the recording and describe precisely the emotional experience they had. This method of self-confrontation has been used *e.g.* by [31].

A posteriori verbal reports can also be used. This method is less time consuming than the self-confrontation method, but cannot furnish a precise timing of subjects' emotional experience. The report can be free or guided, using a questionnaire.

Finally, pictorial methods are widely used, since the use of words for reporting emotional experiences presents several drawbacks [32]. Methods like the Self-Assessment Manikin (SAM, [33]) or the Product Emotion Measurement Tool (PrEmo [34]) are fast and easy to use, but again, can be only used *a posteriori* and furnish therefore only a global information on subjects' emotional experience, without a precise timing.

As pointed out by [35] and [36], no single method is sufficient and it is suggested to combine several methods to obtain a comprehensive and reliable analysis of subjects' emotional experiences during a test.

5.3. Methods for usability assessment

Several definitions of usability exist. ISO's 9241-11 [37] Guidance on Usability defines product usability as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. Nielsen [38], one of the most

famous authors on usability, considers that the usability of a system is based on five dimensions: learnability (*e.g.* intuitive navigation), efficiency of use (*e.g.* time required to accomplish a given task), memorability (capacity of the user to recall information about the system after its usage), errors (that should be few and easy to recover for the user) and subjective satisfaction. It has to be noticed that the definition used here for *satisfaction* is restricted to the subjective appreciation by the subject of the usage of the system under test and is different from the concept of *customer satisfaction* that was discussed in the previous section.

Dependent variables measured during a usability test using a task-oriented protocol will be the degree of completion of the task (overall performance), the time and number of actions required to reach this degree of completion (efficiency), the progression of these variables throughout the test (learnability), the *a posteriori* recall by the subject of different information concerning the system (memorability), the number of errors made by the system and the way the subject could cope with them (error management) and diverse *a posteriori* judgments on these dimensions, including satisfaction, made by the subject and obtained in a post-test interview or questionnaire [38].

5.4. Methods for user's satisfaction assessment

Satisfaction is a rather complex concept to characterize. Whereas *customer satisfaction* encompasses several criteria such as usability, *user satisfaction* is considered by several authors as being one of the criteria of usability. Our aim is to characterize subjects' satisfaction towards the enabler "speech quality" when they are involved in a challenging game and that their performances in this game are directly dependent on speech quality. Speech might be embedded in a service (*e.g.* by using a communication tool, like in [39]) or might just be used as a medium to convey instructions on the game, like in [27]. In order to obtain more realistic data on subjects' overall satisfaction, it seems preferable to simulate the use of a communication tool. Most of marketing studies suggest the use of interviews, questionnaires or likert scales for assessing customer satisfaction [40]. It seems preferable to use such scales in order to obtain numerical data that can be correlated to emotional and behavioral data in subsequent analyses.

6. Conclusion

This paper suggests that speech-quality test methodologies should move further in order to take into account the two revolutions of mobile telecommunications and VoIP, and to treat several drawbacks of Mean Opinion Scores obtained in classic tests recommended by ITU-T [4]. It is proposed to consider valuable inputs from the research fields on the psychology of emotions, cognitive science, ergonomics and marketing research in order to develop test methodologies that will furnish more comprehensive, realistic and reliable data on speech quality and its role in voice telecommunication services. Introducing new concepts in the field of speech quality has the obvious disadvantage of complicating actual test methodologies (mainly based on simple listening tests) which are still widely used. No magic recipes are given and the authors try to contribute to the debate [27], [41], as others like [42] and [43]. We hope that this paper will stimulate discussions and reflections on the development of new test methodologies helping researchers and engineers to produce data that will be more useful for all the actors of the quality chain in the industry of telecommunications.

7. References

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

[2] Scherer, K. R. Appraisal theory. In Dalgleish T and Power M (Eds.), *Handbook of Cognition and Emotion*, Chichester: Wiley. pp. 637-663, 1999.

[3] Jekosch, U. *Voice and Speech Quality Perception. Assessment and Evaluation*, Eds. Sringer. Signals and Communication Techology, 2005.

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

[5] ITU-T, P862. "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001. http://www.itu.int

[6] Gescheider, G.A., *Psychophysics: Method, Theory, and Application*, 2nd ed., Lawrence Erlbaum Associates, Inc., 1985.

[7] Barriac, V., Le Saout, J.Y. and Lockwood, C., "Discussion on unified objective methodologies for the comparison of voice quality of narrowband and wideband scenarios," Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction, Mainz, Germany, June 2004.

[8] France Telecom. "Discussion and proposal concerning the use of MOS scale and terminology in wide-band audio contexts," Delayed contribution 148, meeting of the SG12, June 2006. http://www.itu.int

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

[10] Gabrielsson, A. and Sjögren, H. "Perceived sound quality of sound-reproducing systems," *J. Acoust. Soc. Am. Vol. 65(4), 1979, 1019-1033.*

[11] Guyot F., Piron, C., Castellango, M. and Fabre, B. "Characterization of the sound quality of vacuum cleaners," ASA 129th meeting, Washington DC, USA, May1995.

[12] Namba, S. "On the psychological measurement of loudness, noisiness and annoyance. A review," *Journal of the Acoustical Society of Japan, Vol. 8, 1987, 211-222.*

[13] Möller, S, Riedel, J. "Expectation in quality assessment of Internet telephony". *Acta Acustica,* Joint Meeting ASA/EAA/DEGA, Forum Acusticum, *Suppl.1. 85,* Berlin, Germany, March 1998.

[14] Merleau-Ponty, M. *Phénoménologie de la perception.* Gallimard, 1945.

[15] BS EN ISO 9001:2000. "Quality Management Systems," 2000. http://www.iso-standards-international.com/

[16] Scherer, K. R., Johnstone, T. and Klasmeyer, G. "Vocal expression of emotion," In R.J. Davidson, K.R. Scherer & H.H. Goldsmith (Eds.), *Handbook of affective sciences.* New York: Oxford University Press, pp. 433-456, 2003.

[17] Ekman, P. "Basic emotion," In T. Dalgleish & M. J. Power (Eds.), *Handbook of cognition and emotion.* Chichester, England: Wiley, 1999.
[18] Frijda, N. *The Emotions*, Cambridge University Press, 1986.
[19] Lazarus, R. S., and Folkman, S. *Stress, Appraisal, and Coping*. New York: Springer, 1984.
[20] Scherer, K. R. "Cognitive components of emotion," In R.J. Davidson, K.R. Scherer & H.H. Goldsmith (Eds.), *Handbook of affective sciences.* New York: Oxford University Press, pp. 563-571, 2003.
[21] Schooler, J.W. and Eich E. E. "Memory for emotional events," In E. Tulving & F.I.M. Craik (Eds.) *Handbook of Memory*. New York: Oxford University Press, pp. 379-394, 2000.
[22] Ellsworth, P. C. and Scherer, K. R. "Appraisal processes in emotion," In R.J. Davidson, K.R. Scherer & H.H. Goldsmith (Eds.), *Handbook of affective sciences.* New York: Oxford University Press, pp. 563-571, 2003.
[23] Douglas-Cowie E., Campbell N., Cowie R. and Roach P. "Emotional speech: towards a new generation of databases". *Speech Communication - Speech and Emotion, Vol. 40, n. 1-2, 2003*
[24] Ververidis, D. and Kotropoulos, C. "Emotional speech recognition: Resources, features, and methods." *Speech Communication 1546, in press, 2006.*
[25] Aubergé, V., Audibert, N. and Rilliard A. "Why and how to control the authentic emotional speech corpora." Eurospeech, Geneva, Switzerland, September 2003.
[26] Scherer, K. R. Emotion. In M. Hewstone & W. Stroebe (Eds.). *Introduction to Social Psychology: A European perspective* (3rd. Ed.). Oxford: Blackwell, pp. 151-191, 2000.
[27] Gros, L., Chateau, N, Macé, A. "Assessing speech quality: a new approach", Forum Acusticum, 4th European Congress on Acoustics, August 2005, Budapest, Hungary.
[28] Pashler, H. "Dual-task interference in simple tasks: Data and theory." *Psychological Bulletin*, *116, 1994, 220-244.*
[29] Picard, R. W. *Affective computing.* The MIT Press, 1997.
[30] Keltner, D. and Ekman, P. "Introduction: Expression of emotion," In R.J. Davidson, K.R. Scherer & H.H. Goldsmith (Eds.), *Handbook of affective sciences* New York: Oxford University Press, pp. 411-414, 2003.
[31] Le Chenadec, G., Maffiolo, V., Chateau, N. and Colletta, J.M. "Creation of a Corpus of Multimodal Spontaneous Expressions of Emotions in Human-Machine Interaction," LREC 2006, Genoa, Italy, May 2006
[32] Wierzbicka, A. "Talking about emotions: Semantics, culture and cognition," *Cognition and Emotion*, *6, 1992, 285-319.*
[33] Bradley, M.M. and Lang, P.J. "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavioral Therapy and Experimental Psychiatry, 25 (1), 1994, 49-59.*
[34] Desmet, P.M.A., Hekkert, P., Jacobs, J.J. "When a car makes you smile: Development and application of an instrument to measure product emotions," In: S.J. Hoch, R.J. Meyer (Ed.), *Advances in Consumer Research*, 27, 111-117, 2000.
[35] Picard, R.W. and Bryant Daily, S. "Evaluating affective interactions: Alternatives to asking what users feel," CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches, Portland Oregon, USA, April 2005
[36] Chateau, N. and Mersiol, M. (2005), "AMUSE: A tool for evaluating affective interfaces," CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches, Portland Oregon, USA, April 2005
[37] ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs), 1998. http://www.iso-standards-international.com/
[38] Nielsen, J. *Usability Engineering.* Morgan Kaufmann, San Francisco. 1994. http://www.useit.com/
[39] Chateau, N., Maffiolo, V., Pican, N. and Mersiol, M.: "The Effect of Embodied Conversational Agents' Speech Quality on Users' Attention and Emotion," ACII: 652-659, Beijing, China, 2005.
[40] Hayes, B.E. *Measuring customer satisfaction: Development and Use of Questionnaires.* Milwaukee: ASQC Quality Press, 1993.
[41] Gros, L., Chateau, N. and Durin, V. "Beyond the MOS," MESAQIN, Prague, Czech Republic, 2006.
[42] Mullin, J., Smallwood, L., Watson, A., and Wilson, G. M. "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 2001.
[43] Sonntag, G.P., Portele, T. and Haas, F. "Comparing the comprehensibility of different synthetic voices in a dual task experiment," In: SSW3, pp5-10, 1998.