SPEECH AND AUDIO LOUDNESS DEPENDING ON TELEPHONE AUDIO BANDWIDTH AND CODEC – A SUBJECTIVE TESTING APPROACH

Idir Edjekouane[1, 2], Cyril Plapous[1], Catherine Quinquis[1], Sabine Meunier[2]

[1] Orange Labs – SVQ/MOV, 2 Avenue Pierre Marzin, F-22307 Lannion Cedex
[2] LMA, CNRS, UPR 7051, Aix-Marseille Univ, Centrale Marseille, F-13402 Marseille Cedex

ABSTRACT

In this paper, we propose a new approach for the subjective assessment of the loudness of complex audio signals such as speech or music. This two-stage approach makes it possible to study the influence on loudness of the frequency bandwidth and of different kinds of codecs. In the first stage, the individual loudness function of each subject is estimated using a specific 100-point response scale. In the second stage, the subject evaluates the loudness of each processed sample, by filtering or coding/decoding, using the same scale. The loudness obtained in terms of points is then converted in loudness levels in terms of phons using the estimated individual loudness function. Results show that loudness increases with the bandwidth extension up to super-wideband. Similar behavior is observed when codecs are applied.

Index Terms— *Loudness assessment, Loudness Ratings, Telephonometry, Speech.*

1. INTRODUCTION AND CONTEXT OF THE STUDY

Telephone systems have been created to replace face-to-face conversation; consequently, this situation is taken as a reference for the design of a telecommunication service. The information from the mouth of the speaker to the ear of the listener should be maintained. Loudness, *i.e.* perceived level, largely contributes to the overall quality of the transmitted speech [1] and is a very important perceptual factor necessary for the information to be transmitted. In the field of telephonometry [2], the loss in perceived loudness, due to the end-to-end transmission, is typically expressed as the loudness rating (LR) of the link. It can be decomposed into three parts: the LR of the sending device, that of the receiving device and that of the junction. The LR principle is based on the results of Fletcher (1937) on critical bands and masking effect [3-4]. An extensive description of LR can be found in [5]. The LR model used in telephonometry is published as ITU–T Rec. P.79 [6]. It was initially defined for narrowband (NB) handset terminals [6, Annex A] and has been generalized to wideband (WB) speech using a new set of weighting coefficients [6, Annex G]. However, experimental studies [7] showed large differences in the LR adjustment when WB terminals communicate with NB terminals. Indeed, the user experience in WB is significantly quieter (more than 6 dB) than in NB for a same calculated LR. Up to now, there has been no intention to adapt LR model to super wideband (SWB) and full band (FB) cases.

However, today, there is a real need for an objective model that can predict perceived sound level for end-to-end transmissions from NB to FB. This model must be consistent when switching from one bandwidth to another in order to keep a constant perceived level.

As the main concern is the estimation of the perceived sound level, we think that loudness models (such as [8-9]) are reliable and natural candidates to replace LR, as they work properly from NB to FB. In order to assess the behavior of those loudness models when different speech bandwidths (including when codecs are applied) are used, the loudness calculated using these models must be compared to the loudness evaluated by the listeners. The first step of such an approach is to estimate loudness from the perceptual tests on signals that are used for the assessment of telephone systems, in particular the test signal called P.501 (British-English single talk sequence described in clause 7.3.2 of ITU-T Rec. P.501 [10]). Thus, the aim of the work presented in this paper was to design such a subjective loudness test. No standard has been established yet for the measurement of the loudness of complex signals. However, some studies on loudness speech assessment exist [11-12] and they are based on methods derived from loudness matching or categorical loudness scaling (CLS) [13]. The limitation of loudness matching is that it is hardly usable for the long duration signals that are commonly used in telephonometry. The limitation of CLS is that little is known about how loudness in categorical units relates to the other established measures of loudness, *i.e.* sones and phons, and that it has been built to study the whole dynamic range of human hearing. That is why, in this paper, we propose a new method to assess the loudness of complex signals using a specific scale. We derived the

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

loudness level of different test signals from the known loudness of a band of noise.

2. METHODOLOGY

The test procedure included two stages. In the first stage, we estimated the individual loudness functions (ILF) of the subjects. In the second stage, the subjects evaluated the loudness of each of the test signals. All evaluations were made on a specific 100-point response scale. The stimuli were presented monaurally to simulate the use of a handheld telephone handset. The results were obtained in terms of points and thanks to the estimated ILF it was possible to convert the point scale into a phon scale.

2.1. Test signals

Audio samples with different contents (*cf.* Table 1) were selected. These samples can be speech in different contexts and languages, music or a mixture of speech and music. The so called P.501 signal is a speech test signal provided by ITU-T [10] and is widely used in telephonometry. It was of great interest in this study as it had already been used for the determination of LR. As the original P.501 signal was too long (34.5 s), it was cut out into 4 parts; each part containing 3 male or 3 female speakers leading to samples 5 to 8 (*cf.* Table 1).

These samples were processed according to the diagram in Figure 1 below. Thus, for each bandwidth (FB, SWB, WB or NB), the filtered samples were coded/decoded using 2 different families of codecs (*cf.* Table 2). The first family was made up of codecs mainly designed for speech content (referred to as "Speech codecs") whereas for the second one, the codecs were not content dependent (referred to as "Generic codecs"). The signals directly obtained after filtering or "filtering + coding/decoding" led to what we called the "Nominal" level (Gain at 0 dB in Figure 1). These signals were also amplified by 5 dB, which led to a "Nominal+5 dB" level, and attenuated by 10 dB, which led to a "Nominal-10 dB" level. These 2 additional conditions were introduced to test a wider range of levels. Finally, a total of 36 conditions were applied to the 9 samples, which resulted in a total of [(8+4) x 3] x 9 = 324 test signals.

	Content description	Duration (seconds)	Speech language
Sample 1	Rock Music	7.8	X
Sample 2	Music then Speech mixed with Music	12.4	French
Sample 3	Speech (voice announcement)	7.6	French
Sample 4	Speech mixed with Noise	10.2	French
Sample 5	Speech (P.501) Part 1	8.3	British-English
Sample 6	Speech (P.501) Part 2	9	British-English
Sample 7	Speech (P.501) Part 3	9.2	British-English
Sample 8	Speech (P.501) Part 4	10	British-English
Sample 9	Speech then Speech mixed with Music	8.5	French

Table 1: Description of samples

Bandwidth	Codec (bitrate)	
	Speech codec	Generic codec
Full Band (FB) codecs, sampled at 48kHz	OPUS (64 kb/s) [14]	G.719 (64 kb/s) [15]
Super Wideband (SWB) codecs, decimated to 32kHz	G.729.1 (32 kb/s) [16]	G.722.1 C (48 kb/s) [17]
Wideband (WB) codecs, decimated to 16kHz	AMR-WB (12.65 kb/s) [18]	G.722 (64 kb/s) [19]
Narrowband (NB) codecs, decimated to 8kHz	AMR (12.2 kb/s) [20]	G.711 (64 kb/s) [21]

Table 2: Description of codecs

2.2. Description of the response scale

After hearing a stimulus, the subject indicated how he/she perceived its loudness using a 100-point scale (Figure 2). The subject had 5 seconds to make his/her evaluation, the passage to the next stimulus being automatic so that the subject was pushed to give a spontaneous evaluation. The subject could see the chosen numeric value displayed on the scale. The three labels titled in French "Très fort" (very

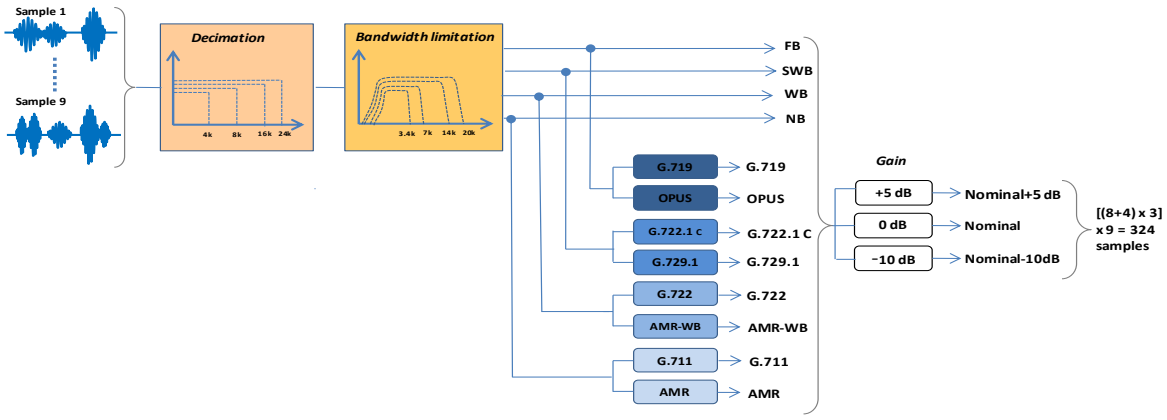

Fig. 1: Diagram describing the preparation of test signals for the subjective test

loud), “Moyennement fort” (averagely loud) and “Pas fort” (not loud) were used to help the subject by providing him/her with 3 reference points. These labels were chosen as they are common French language expressions related to loudness. The term “fort” (loud) was used in the three labels since the loudness range covering all test signals was relatively high.

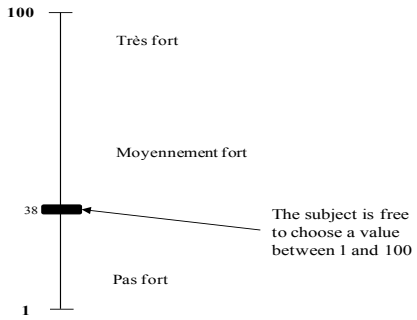

Fig. 2: Reproduction of the 100-point response scale

2.3. Subjects and apparatus

Eighteen subjects participated in this subjective loudness test. None reported having hearing problems. Before the beginning of the test, we asked each subject about his/her preferred ear (left or right) when making a phone call. The stimuli were then presented monaurally (left or right) to the subject via open back diffuse-field [22] corrected headphones (Stax SR-404). All stimuli were digitally processed at a sampling rate of 48 kHz, D/A-converted (SPL 2489) and amplified (Stax SRM-006t). The listening level of the setup was calibrated to ensure a comfortable level of 77 dB SPL (Sound Pressure Level) for FB signals at “Nominal” level.

2.4. First stage of the subjective test: ILF Measurement

The stimuli were constructed based on a critical band of noise with center frequency at 1 kHz and duration of 1 second. The stimuli were presented to each subject at different acoustical levels in a pseudo-randomized order. The dynamic range of acoustical levels for the band of noise had to be at least as large as the loudness dynamic range of the test signals, which were used in the second stage of the experiment. Thus, a small test was designed to determine this dynamic range.

2.4.1. Dynamic range determination

Among all the test signals, the ones with higher levels in dB SPL were related to the “FB and Nominal+5 dB” condition and the ones with lower levels were related to the “NB and Nominal-10 dB” condition. Thus, the determination of the dynamic range consisted in making a loudness-balance test between those signals and the critical band of noise presented in a large range of levels from 58 to 91 dB SPL with a step of 3 dB. At the end of this test, it was found that, on average, the test signals from the “FB and Nominal+5 dB” condition were as loud as the band of noise at 85 dB SPL and the test signals from the “NB and Nominal-10 dB” condition were as loud as the band of noise at 73 dB SPL. In order to be sure that the full dynamic range was covered, this dynamic range was increased to reach the range [61 dB SPL; 88 dB SPL]. This range was covered with a 3 dB step which led to a total of 10 possible stimuli. The enlargement of this dynamic range was non-symmetric because the critical band of noise was too loud over 88 dB SPL. This test was done once before the actual subjective test. It was conducted on ten colleagues working in Orange Labs.

2.4.2. Procedure for ILF Measurement

The assessment of ILF consisted in two phases (*cf.* Figure 3) in which the subject rated the loudness using the scale described in section 2.2. In the first phase (training phase), the subject heard a selection of samples covering the whole dynamic range of levels. This phase avoids biases caused by the first trials that do not cover the whole dynamic range. In the second phase, the 10 stimuli (critical-band of noise) were presented 6 times each, using 6 pseudo-random orders. To avoid judgments biased by the previous stimulus, the level difference between two successive stimuli was kept smaller than half of the dynamic range.

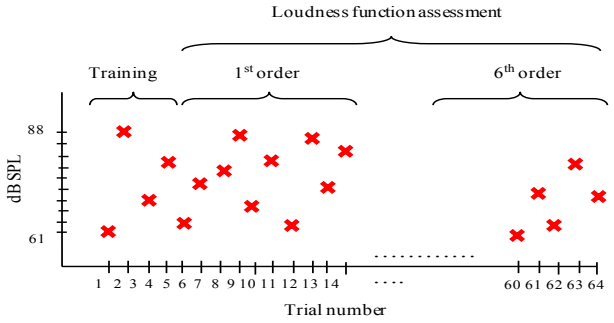

Fig. 3: Trials for the determination of ILF. The training is followed by 6 pseudo-random orders.

2.4.3. Results for ILF

Figure 4 shows the estimated ILF of the 18 subjects in terms of points. In general the curves are S-shaped, the two different saturated parts - the upper part [85 dB SPL; 88 dB SPL] and the lower part [61 dB SPL; 70 dB SPL] - being due to the saturation of the scale. Indeed, the subjects always judged the sound as “very loud” when the signal level was higher than 85 dB SPL, and as “not loud” when the signal level was lower than 70 dB SPL. In the middle part of the curve [70 dB SPL; 85 dB SPL], all the subjects used the scale efficiently. The conversion from loudness in terms of points to loudness in terms of phons was performed using this middle part (*cf.* section 2.6).

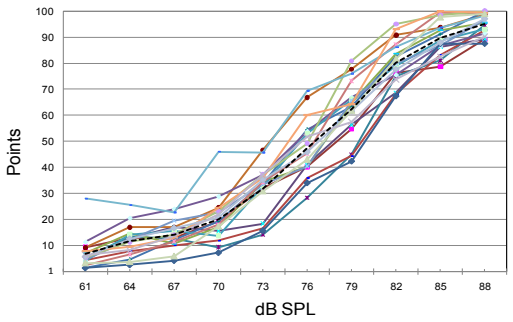

Fig. 4: Estimated ILF (in terms of points) obtained for the 18 subjects along with the overall average (dashed line).

2.5. Second stage of the subjective test: Loudness assessment of test signals

The assessment of the loudness of the test signals was performed in two phases in which the subject rated the loudness using the scale presented in Figure 2. The first phase was the training phase in which the subject heard a selection of samples covering the whole dynamic range of levels. This selection contained the softest and loudest conditions. All 9 samples were used in the training phase so that the subject could listen to all of them before the second phase. In the second phase, the 324 test signals were presented randomly and rated by the subject using the scale described in section 2.2.

2.6. Conversion from points to phons

The estimated ILF gives the relation between dB SPL and points for each subject (see Figure 4). The key to transform points to phons is that the phon scale is equal to the dB SPL scale for a critical band of noise with center frequency at 1 kHz [23]. Thus, it is possible from the estimated ILF to get the relation between points and phons. The estimated ILF showed a linear tendency in the middle part of the curve [70 dB SPL; 85 dB SPL]. A linear model function was then fitted to the individual data using a least-square fit:

$$N_{phons} = \alpha_i \times N_{points} + \beta_i \qquad (1)$$

where α_i and β_i are the fitting parameters determined for each subject i (i=1,2,..,18). N_{phons} and N_{points} are the loudness levels expressed in terms of phons and in terms of points, respectively.

Obviously, for each subject the point to phon conversion was based on his/her own loudness function. This is because each subject used the response scale in his/her own way. They created for themselves an internal reference system that could vary largely from a subject to another. However, as long as the subject kept the same internal reference system throughout the entire subjective test, it was possible to convert points to phons using his/her estimated ILF and equation (1).

3. RESULTS AND DISCUSSION

Figure 5 gathers loudness level results averaged over all samples. All conditions are represented in this figure, *i.e.* "Bandwidth", "Speech codecs", "Generic codecs" as well as the three levels, *i.e.* "Nominal+5 dB", "Nominal" and "Nominal-10 dB". These results are presented in terms of phons and come with a 95% confidence interval. These results are consistent with what could be expected, as loudness increases with bandwidth extension. There is a significant gap between loudness in NB conditions (*i.e.* NB, G.711, AMR) and WB conditions (*i.e.* WB, G.722, AMR-WB) and for all levels. After verifying that the data for each condition was normally distributed (Kolmogorov-Smirnov test), separate one-way ANOVAs (factors are conditions, significant level at 5%) were conducted for each level: "Nominal", "Nominal+5dB" and "Nominal-10dB", to investigate whether there was a significant difference between the loudness of WB conditions, SWB conditions (*i.e.* SWB, G.722.1 C, G.729.1) and FB conditions (*i.e.* FB, G.719, OPUS). For each level, the ANOVA results show a significant loudness difference between the three conditions. Post-hoc ANOVA tests show that there is a significant difference between the loudness of WB and SWB conditions. However, the difference between the loudness of SWB and FB conditions is not significant. The results are rather similar for "Speech codecs" and "Generic codecs", although the coding was handled differently.

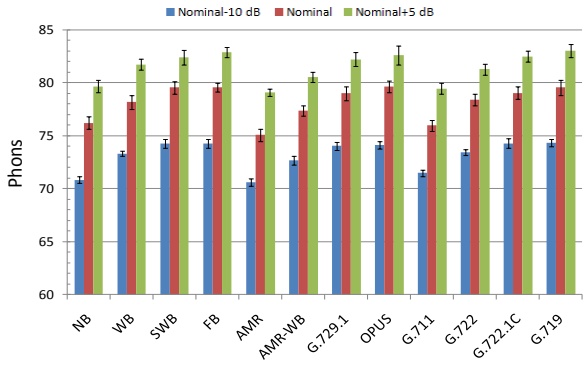

Fig. 5: Averaged results over all samples. All conditions are represented: "Bandwidth", "Speech codecs", "Generic codecs", "Nominal+5 dB", "Nominal" and "Nominal-10 dB".

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a subjective test to study loudness behavior depending on bandwidth frequency and codecs available on the market. The results show that loudness increases with bandwidth extension up to super wideband, including when codecs are applied. The next step will be to conduct objective loudness measurements on the same test signals in order to confront the behavior of existing algorithms to these subjective results.

5. REFERENCES

[1] N. Côté, V. Gautier-Turbin, and S. Möller, "Influence of Loudness Level on the Overall Quality of Transmitted Speech", 123rd Audio Eng. Soc. Conv., New YorkUS, Oct. 2007.

[2] ITU-T, Handbook on telephonometry, ITU-T, Geneva, 1992.

[3] H. Fletcher and W.A. Munson: "Relation between Loudness and Masking", J. Acoust. Soc. Am., vol. 9, issue 1, p. 1, New York, 1933.

[4] H. Fletcher and W.A. Munson, "Loudness, Its Definition, Measurement and Calculation", J. Acoust. Soc. Am., vol. 5, issue 2, p. 82, New York, 1933.

[5] S. Möller, "Assessment and Prediction of Speech Quality in Telecommunications", pp. 18-24, Springer 2000.

[6] ITU-T Recommendation P.79: "Calculation of loudness ratings for telephone sets", ITU-T, Geneva, 2000.

[7] K. Allen Woo, "Wide-band loudness ratings confusion (ref ITU-T P.79)", ETSI STQ(12), 10/2007.

[8] ISO 532 B, "Method for calculating loudness", International standard (1975).

[9] ANSI S3, 4-2007, "American National Standard Procedure for the Computation of Loudness of Steady Sound", 2007.

[10] ITU-T Recommendation P.501: "Test signals for use in telephonometry", 01/2012.

[11] H. Fastl, "loudness of running speech", J. Audiol. Technique 16 2-13, 1977.

[12] J. Rennies, J. L. Verhey, J. E. Appell, and B. Kollmeier, "Loudness of complex time-varying sounds. A challenge for current loudness models", J. Acoust. Soc. Am., vol. 19, number 1, Montreal, 2013.

[13] ISO 16832. "Acoustics – Loudness scaling by means of categories", 2006.

[14] JM. Valin, K. Vos, and T. Terriberry. Internet Engineering Task Force (IETF) RFC6716: "Definition of the Opus Audio Codec", September 2012.

[15] ITU-T Recommendation G.719: "Low-complexity, full-band audio coding for high-quality, conversational applications", June 2008.

[16] ITU-T Recommendation G.729.1: "G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729", May 2006.

[17] ITU-T Recommendation G.722.1 Annex C: "Low Complexity Coding at 24 and 32 kb/s for Hands-Free Operation in Systems with Low Frame Loss Annex C 14 kHz Mode at 24, 32,and 48 kb/s ", May 2005.

[18] 3GPP TS 26 290 V10.0.0, "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions", 2011.

[19] ITU-T Recommendation G.722: "7 kHz audio-coding within 64 kbit/s, Geneva, Switzerland", 09/2012.

[20] 3GPP TS 126 071 V3.0.1, "Mandatory speech CODEC speech processing functions; AMR speech Codec; General description", 2000.

[21] ITU-T Recommendation G.711: "Pulse code modulation (PCM) of voice frequencies", Geneva, Switzerland, Nov. 1988.

[22] ITU-T Recommendation P.58: "Head and torso simulator for telephonometry", ITU-T, Geneva, 2013.

[23] E. Zwicker and H. Fastl, "Psychoacoustics: Facts and models", 2nd Edition, Springer-Verlag, pp. 203, 1999.