Loudness of Speech Transmitted via Handsfree Telephone Systems - Perceptual Measurements and Loudness Models in Free Field Listening

Idir Edjekouane, Cyril Plapous, Catherine Quinquis, Sabine Meunier

Acta Acustica united with Acustica
Vol. 101 (2015) 1130 – 1144

DOI 10.3813/AAA.918906

Loudness of Speech Transmitted via Handsfree Telephone Systems - Perceptual Measurements and Loudness Models in Free Field Listening

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

[1] Orange Labs – SVQ/MOV, 2 Avenue Pierre Marzin, 22307 Lannion Cedex, France.
edjekouane_idir@hotmail.fr

[2] LMA, CNRS, UPR 7051, Aix-Marseille Université, Centrale Marseille, 4 impasse Nikola Tesla, CS400006, 13453 Marseille Cedex 13, France

Summary

The present study investigated the loudness of speech transmitted via a telephone system and the ability of existing loudness models to predict the perceived loudness. This study was focused on the case of handsfree telephony. To generate test signals for the experiment, twelve stimuli (mainly speech material) were selected and processed to simulate realistic telephone system paths. The processing included: filtering, coding/decoding and amplification/attenuation. A perceptual test was designed to measure the loudness level of the test signals. Results showed that loudness increases with bandwidth extension up to super wideband, including when codecs were applied. All tested models were variants of Zwicker's loudness model. Two models for stationary sounds and two models for non-stationary sounds were tested. In general, the models predicted the main trend observed in perceptual data and the increase in loudness with bandwidth extension. However, there was always a difference between prediction and measure, which depended on the sound pressure level (SPL). The models' behavior while varying the SPL was similar to what has been reported in many recent studies. Zwicker's loudness model yielded the best predictions, which did not support the hypothesis that using non-stationary loudness models improves the prediction of speech loudness.

PACS no. 43.66.-x, 43.66.Ba, 43.66.Cb

1. Introduction

Speech signals play a major role in voice telecommunication systems and loudness largely contributes to the overall quality of the transmitted speech [1]. It is a very important perceptual factor necessary for the information to be transmitted. In this study, we are interested in measuring loudness in different transmission conditions and in testing whether loudness models are suitable to predict the loudness of speech transmitted by telephone systems.

1.1. Context and motivation

Telephonometry studies the techniques used for instrumental and perceptual measurements of the voice quality of a telephone communication (see an overview in [2]). Loudness is one of the main parameters used for telephone network planning and has to be determined for all devices involved in telephone paths. In telephonometry, 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 and Munson on critical bands and masking effects [3, 4]. In practical terms, LR consists of applying weighting coefficients to the electroacoustic sensitivity of the equipment in third-octave bands before summing the values for each band; the obtained value (expressed in dB) reflects the loss in the perceived loudness compared with a reference telephone path. An extensive description of LR can be found in [5, 6, 7]. The LR model is published as ITU–T Recommendation P.79 [8]. It was initially defined for narrowband (NB) [300 Hz–3400 Hz] handset terminals [8, Annex A] and has been generalized to the case of wideband (WB) [50 Hz–7000 Hz] handset terminals using a new set of weighting coefficients [8, Annex G]. However, experimental studies reported some incoherences between the perceived loudness and the calculated LR in some situations [9, 10]. It was reported that there was a significant difference in the perceived loudness between WB and NB terminals when they communicated with each other. Indeed, with the same loudness rating for both NB and WB systems the user experience of the WB system is significantly quieter than in NB mode. At least 6 dB should be

Received 02 April 2015,
accepted 15 October 2015.

added to the WB signal to perceive the same loudness of the NB signal. It was concluded that there was a poor correlation between the LR calculated for WB and the perceived loudness. Despite the current tendency to increase the bandwidth of the transmitted speech, there has been no intention to correct LR for the WB case, or to adapt the LR model to super wideband (SWB) [50 Hz–14000 Hz] [11] and full-band (FB) [20 Hz–20000 Hz] [11]. Another issue comes out when telephone devices implement nonlinear and time variant speech enhancement functions, such as background noise cancellation, improved double-talk behavior, etc. All this has an influence on the computation of LR, because LR assumes linearity between the electrical signal and the acoustical signal in the terminal. A terminal used in a handsfree situation is an example of nonlinear processing. It usually includes automatic gain control (AGC) which introduces speech dynamic compression. Monfort *et al.* [10] showed that when a terminal is set at the same LR value in handset and handsfree modes, as defined by the standards, differences in loudness may be perceived when switching from handset mode to handsfree mode and vice versa.

Today, there is a real need for a model that can predict the perceived loudness for end-to-end transmissions from NB to FB and in handset or handsfree situations. This model must be consistent when switching from one bandwidth to another and from handset to handsfree in order to keep a constant perceived loudness. As the main concern is the estimation of the loudness, we think that state-of-the-art loudness models that are based on Zwicker's model [12] could be reliable and natural candidates to replace LR. The loudness calculated using these models, when different speech bandwidths and different codecs are used, should be comparable to the loudness evaluated by the listeners. The first step of such an approach is to estimate loudness from perceptual tests on signals that are usually used for the assessment of telephone systems, in particular the British-English single-talk sequence described in clause 7.3.2 of Recommendation ITU T P.501 [13]. This signal will be referred to as P.501 single-talk sequence in the rest of the document. Thus, in the present work we are interested in measuring loudness in different transmission conditions and in testing whether loudness models are suitable to predict loudness of speech transmitted by telephone systems.

1.2. Related work

The design of loudness models is based on studies about the auditory system. These studies are usually conducted with artificial sounds like pure tones and noise. Conversely, "real-world" sounds, especially speech signals, are less frequently used in research on loudness. However, some recent studies have focused on loudness of speech.

Since speech signals have complex acoustic properties (*e.g.*, spectrum, amplitude and/or frequency modulation), many studies have examined the effect of speech properties on the perception of the loudness of speech (see review in [14]). Brand and Hohmann [15] measured the loudness functions for speech and stationary speech-shaped noise. They found that the loudness functions were rather similar when compared at the same root-mean square (rms) level. Moore *et al.* [16] found small differences in rms levels at equal loudness between stationary speech-shaped noise and noise with the same spectrum but with speech-like temporal envelope modulations. Rennies *et al.* 2013 [14] studied the influence of speech-related properties on loudness. They used eight stimuli that had the same speech-like long-term spectrum, but differed in other speech-related properties, ranging from modulated stationary noise to real intelligible and unintelligible speech. They concluded that the long-term spectrum is the dominating factor for the loudness of speech. Rennies *et al.* [14] suggested that intelligibility had no effect on loudness. However, Warren [17] found that loudness functions differed between intelligible and unintelligible speech. Further studies investigated the effect of vocal effort on the loudness of speech [18, 19, 20]. They agreed that the loudness consistently increased when the speech was recorded at a higher effort and presented at the same sound pressure level. It was concluded that the effort a talker puts into producing a word is important in the perception of the loudness of this word.

In contrast to studying the role of speech properties in the perception of loudness, other studies have investigated the influence of severe modifications of speech signals. Fastl [21] examined the effect of bandwidth limitation on speech signal. He found that loudness is hardly diminished when the high frequencies are strongly attenuated, and by contrast it is sensitive to any attenuation at low frequencies. Other studies also addressed different types of speech distortion. For instance, Moore *et al.* [22] showed that the dynamic compression of speech leads to an increase in loudness for a fixed rms level. Fastl [21] also studied the loudness of speech distorted by peak clipping. He found that the clipping, even when significant, had a small effect on loudness. Pollack [23] studied the effect of white noise on the loudness of speech. He found that the effect of noise on the loudness of speech is a function of the speech-to-noise ratio rather than a function of the level of speech alone or of the noise alone.

As shown above, many aspects of loudness of speech have been studied in the last decades. These studies showed that the loudness of speech depends on various factors. Nevertheless, no loudness model has yet been established as an accurate measure of the loudness of speech material. The present study investigated the loudness of speech transmitted via a telephone system and the ability of existing loudness models to predict the perceived loudness. The perceptual test that was designed was an adaptation of the perceptual test presented in [32] related to handset telephony. In this test, the loudness levels of the signals were expressed in phon. The results were compared with predictions of four current loudness models for stationary and non-stationary sounds, namely the Zwicker model for stationary sounds [33], the Moore and Glasberg model for stationary sounds [34], the Fastl and Zwicker model for

Table I. Description of the stimuli used to generate the test signals.

	Content description	Duration (seconds)	Gender	Speech language
Sample 1	Speech	6	female	French
Sample 2	Rock Music	7.8	X	X
Sample 3	Speech (voice announcement)	7.6	female	French
Sample 4	Speech	10.2	male	French
Sample 5	Speech (P.501 single-talk sequence) Part 1	8.3	male	British-English
Sample 6	Speech (P.501 single-talk sequence) Part 2	9	female	British-English
Sample 7	Speech	8.4	male	French
Sample 8	Speech mixed with noise[1]	6	female	French
Sample 9	Speech then Speech mixed with Music	8.5	male	French
Sample 10	Speech mixed with noise[1]	8.4	male	French
Sample 11	Speech mixed with noise[1]	8.3	male	British-English
Sample 12	Speech mixed with noise[1]	9	female	British-English

Table II. Description of codecs.

Bandwidth	Speech codec (bitrate)	Generic codec (bitrate)
Full Band (FB) codecs, sampled at 48 kHz	OPUS (64 kb/s) [24]	G.719 (64 kb/s) [25]
Super Wideband (SWB) codecs, decimated to 32 kHz	G.729.1 (32 kb/s) [26]	G.722.1 C (48 kb/s) [27]
Wideband (WB) codecs, decimated to 16 kHz	AMR-WB (12.65 kb/s) [28]	G.722 (64 kb/s) [29]
Narrowband (NB) codecs, decimated to 8 kHz	AMR (12.2 kb/s) [30]	G.711 (64 kb/s) [31]

non-stationary sounds [35] and the Glasberg and Moore model for non-stationary sounds [36].

2. Experiment

The test procedure included three stages. In the first stage, the individual loudness function of the listener was estimated using a critical-band-wide noise (center frequency at 1 kHz) at different levels. In the second stage, the listener evaluated the loudness of 432 test signals (*cf.* Section 2.1.2). The third stage consisted of measuring again the individual loudness function of the subject, for validation purposes. All evaluations were made on a specific response scale of 100 points. The three stages of the test procedure were realized during one session. The results were obtained in terms of points and the estimated individual loudness function was used to convert the point scale into a phon scale.

2.1. Test signals

Our purpose was to study the loudness of speech in the specific case of the transmission through a telephone system. We thus selected 11 speech stimuli and 1 music stimulus that were processed to simulate realistic telephone system paths.

2.1.1. Stimuli

Audio samples with different contents (*cf.* Table I) were selected. These samples were speech in different contexts and languages, music or a mixture of speech and music. The so-called P.501 single-talk sequence [13] signal is a speech test signal provided by ITU-T that is widely used in telephonometry. It was of great interest in this study as it is already used for the determination of LR. This signal is a sequence of sentences in British-English uttered by 12 speakers – six different male speakers and six different female speakers – that lasts about 35.4 s. In order to keep the duration of the test reasonable, we decided to use only the first three male sentences (Sample 5, Table I) and the first three female sentences (Sample 6, Table I).

2.1.2. Generation of test signals

All samples were processed according to the diagram in Figure 1. For each bandwidth (FB, SWB, WB or NB), the filtered samples were coded/decoded using 2 different families of codecs (*cf.* Table II). The first family was made up of codecs mainly designed for speech content (referred to as "Speech codecs") whereas the second one was made up of codecs that were not content-dependent (referred to as "Generic codecs"). The signals directly obtained after filtering or "filtering + coding/decoding" led to the "Nominal" level (Gain at 0 dB in Figure 1). These signals were also amplified by 5 dB, which led to the "Nominal+5 dB" level, and attenuated by 10 dB, which led to the "Nominal-10 dB" level. These two additional conditions were introduced to test a wider range of levels. We can summarize the conditions as follows: [(8 types of codec + 4 bandwidths filtering) × 3 amplification levels] = 36 conditions. Finally, these 36 conditions were applied to the 12 samples, which resulted in a total of 432 test signals.

2.2. Description of the response scale

Loudness was evaluated using a scale of 100 points (Figure 2). After listening to the sound, the subject had 5 sec-

[1] Artificial noisy speech. Speech was mixed with Pub noise [37] at a signal-to-noise ratio (SNR) of 10 dB.

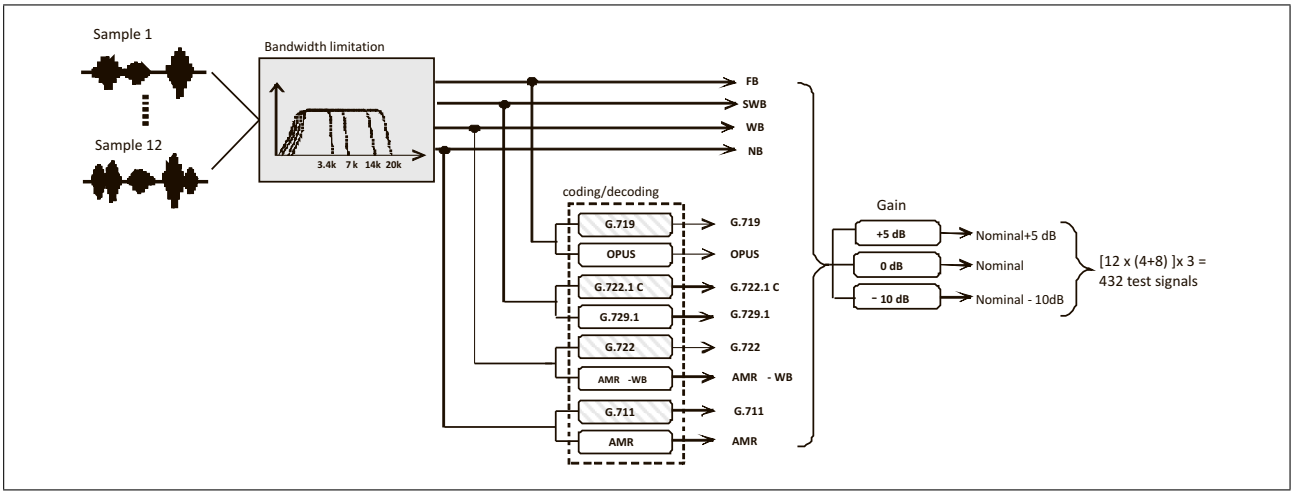

Figure 1. Diagram describing the processing of the stimuli in order to generate the 432 test signals.

onds to make his evaluation, the automatic passage to the next stimulus pushing them 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 loud) -situated at 85 points-, "Moyennement fort" (moderately loud) -situated at 50 points- and "Pas fort" (not loud) -situated at 15 points- were used to give the subject three 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 of all the test signals was relatively high.

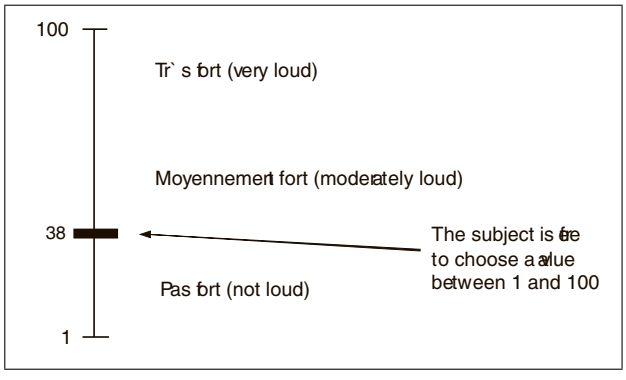

Figure 2. Reproduction of the response scale of 100 points.

2.3. Subjects and apparatus

Twenty-seven subjects participated in the experiment. None reported having hearing problems. They were 15 women and 12 men, with ages ranging from 19 to 50 years and an average age of 33 years. None of the subjects had previous experience in making loudness judgments. All were paid for their service.

The test was carried out in the Orange Labs anechoic chamber (8 m × 7.5 m × 8 m) so that the free field condition, as defined in [38, p. 12], could be achieved. The experiment setup is detailed in Figure 3. All stimuli were digitally processed at a sampling rate of 48 kHz and D/A-converted using a PreSonus FirePod soundcard. The test signals were presented to the subjects via one loudspeaker FOSTEX PM0.5n. This loudspeaker was equalized to have a flat frequency response (from 50 Hz to 20 kHz) at point M, defined in Figure 3, in the absence of listener. The equalization consisted of measuring the frequency response of the loudspeaker at the point M. Then, using an FIR filter of length 2048, we apply the inverse of the frequency response measured previously at the loudspeaker input. The point M represents the point bisecting the line joining the ear canal entrance points. All stimuli (*i.e.* Table I) were set to −24 dB FS before being processed (*i.e.* FB signals at "Nominal") according to the diagram in Figure 1. This Full Scale level corresponded to 65 dB SPL measured at point M using a B&K 4938-A-011 ¼-inch

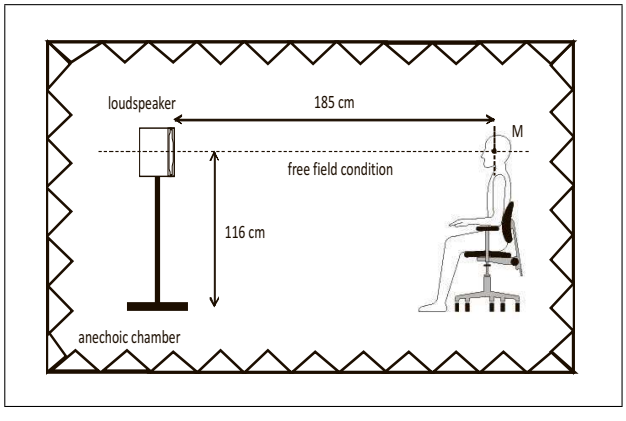

Figure 3. Experiment setup.

pressure-field microphone, a B&K 2636 measurement amplifier and a B&K 4231 sound calibrator. This measured level was judged as a comfortable listening condition by our perceptual test experts.

2.4. First stage: measurement of the individual loudness function

The individual loudness function describes the relation between the signal level (in dB SPL) and the corresponding loudness on the 100-point scale for each subject. This

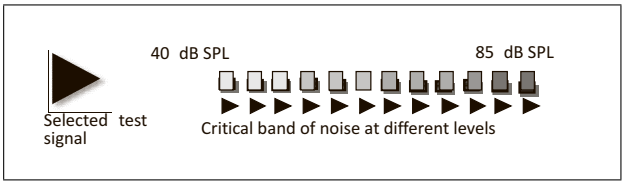

Figure 4. Reproduction of the graphical interface used in the preliminary test.

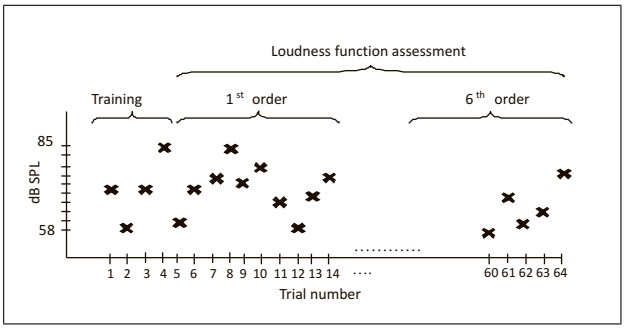

Figure 5. Trials for the determination of individual loudness function. Training is followed by 6 pseudo-random orders.

function was necessary in order to convert the evaluation of each test signal on the 100-point scale to its loudness level (in phon). To measure this function, stimuli were presented to the subject at different sound pressure levels (SPLs) in a non-systematic way (pseudo-randomized). The stimuli were frozen noise bands centered at 1 kHz with a bandwidth of one critical band (160 Hz).

In order to determine the range of levels to be used for the measurement of the loudness function, a preliminary test was run.

2.4.1. Preliminary test: dynamic range determination

This preliminary test was conducted on twenty colleagues working in our laboratory. This test consisted of measuring the loudness of a reduced number of test signals. Among all the test signals, the ones with higher SPLs were the ones processed in the "FB/Nominal+5 dB" condition and the ones with lower SPLs were the ones processed in the "NB/Nominal-10 dB" condition. These signals (12 for FB/Nominal+5dB and 12 for NB/Nominal-10dB) were all tested in order to determine the maximum and the minimum of the dynamic range. Using a graphical interface (*cf.* Figure 4), the listener could listen to one of the selected test signals and to critical-band-wide noise (centered at 1 kHz and 160 Hz wide) at different levels. The bands of noise as well as the test signals could be played as many times as desired. The bands of noise were presented in a large range of levels from 40 to 85 dB SPL with a step of 3 dB. The subject was asked to select the band of noise whose loudness matched best the loudness of the test signals.

At the end of this test, it was found that, on average, the test signals related to the "FB/Nominal+5dB" condition were judged as loud as the band of noise at 82 dB SPL and, on average, the test signals related to the "NB/Nominal-10dB" condition were judged as loud as the band of noise at 67 dB SPL. In order to be sure that the full dynamic range was covered, this dynamic range (*i.e.* [67 dB SPL, 82 dB SPL]) was increased to reach the range [58 dB SPL, 85 dB SPL]. This extension of the range was not symmetric since the noise was judged too loud at levels higher than 85 dB SPL (*e.g.* 88 dB SPL). Therefore, the stimuli used for the determination of the individual loudness function were made up of 10 levels of the noise ranging from 58 to 85 dB SPL in steps of 3 dB.

2.4.2. Measurement of the individual loudness function

The assessment of the individual loudness function was divided into two phases in which the subject rated the loudness using the scale described in Figure 2. The first phase was the training phase in which the subject heard a selection of samples covering the whole dynamic range. This phase was introduced to avoid biases caused by the first trials that did not cover the whole dynamic range [39, 40]. During the training phase, 4 stimuli were presented, one with the highest SPL, another with the lowest SPL and two stimuli with intermediate SPL.

In the second phase, the ten bands of noise (*cf.* Section 2.4.1) were presented 6 times each, using 6 pseudo-random orders. Attention was paid to keeping level difference between two successive stimuli smaller than half of the dynamic range. That way, context effects due to the tendency of many subjects to rate the current stimulus relatively to the previous one were reduced [41, 42]. All 64 trials (training plus 6 pseudo-random orders) are illustrated in Figure 5. Each subject heard the trials using their own pseudo-random order.

2.5. Second stage: assessment of the test signal loudness

The loudness of each test signal was evaluated on the scale described in Figure 2. First, as training, the subject heard a selection of signals covering the whole dynamic range of levels. This selection contained the test signals with the highest and the lowest SPL. All 12 samples in Table I were used in the training so that the subject could listen to all of them before the second phase.

In the second phase, the 432 test signals (*cf.* Figure 1) were presented randomly. Each subject heard the 432 test signals using their own random order. For the assessment of these test signals (including training), the subjects were asked to take into account the loudness over the overall signal as they were quite long (*cf.* Table I).

At the end of this test, we obtained, for each subject, the loudness assessment for the 432 test signals in terms of points. In Section 2.8 we explain in detail how the loudness values in points were converted into loudness level (phon) using the individual loudness functions.

2.6. Third stage: validation of the individual loudness function

The purpose of this third stage was to check the reliability of the measured individual loudness functions. The aim was to check if the subject kept using the response scale in

the same way throughout the session. It is expected that the subject keeps rating the test signals in the same way from the beginning to the end of the test. However, we had to reject a few subjects who modified their way to judge throughout the test (see an example in Figure 6). The mean absolute difference was calculated between the two loudness functions (*i.e.* the first and second measurements). In case the mean absolute difference was higher than a certain threshold, we rejected the subject. This threshold was determined by computing the mean confidence interval over all the 54 measured loudness functions (27 subjects × 2 loudness function measures = 54 measures). As a result, the mean confidence interval was equal to 8.14 points. Using the mean confidence interval as a criterion, we rejected 7 subjects and kept 20. Finally, for the conversion from points to loudness level, we used the average of both measured individual loudness functions (before and after loudness assessment) in order to get a more robust estimation of the individual loudness function.

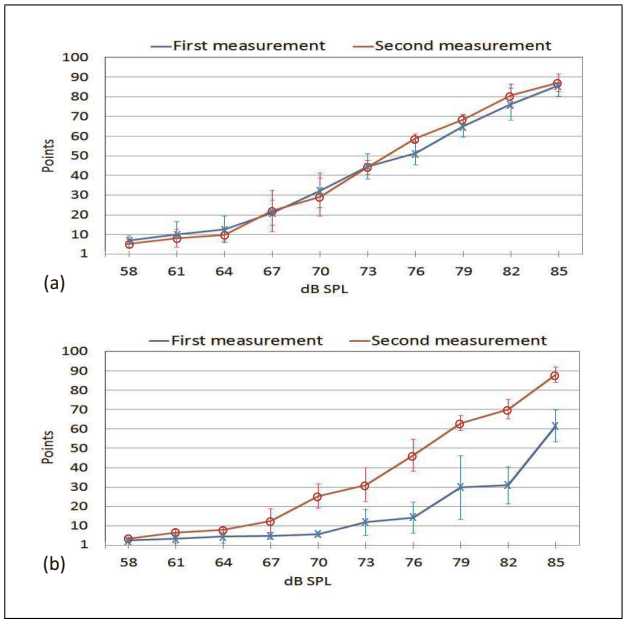

Figure 6. (Colour online) Individual loudness functions (in terms of points) before (crosses) and after (circles) the assessment of test signal loudness. The vertical bars represent CI at 95% over the 6 trials. Figure 6a shows the results of an accepted subject. Figure 6b shows the results of a rejected subject.

2.7. Results of the measurement of the individual loudness function

The individual loudness functions of the 20 accepted subjects are presented in terms of points in Figure 7. The overall average is also displayed for information (dashed line). We can observe that in general the curves have an S shape. Most curves (on average) can be divided into three principal parts: a linear part [70 dB SPL; 82 dB SPL], another linear part [82 dB SPL; 85 dB SPL] with a lower slope and finally a curved part [58 dB SPL; 70 dB SPL].

Based on the modeling of these individual loudness functions, the results obtained (in terms of points) in the second stage of the experiment were converted into loudness level. This will be described in the next Section (*cf.* Section 2.8).

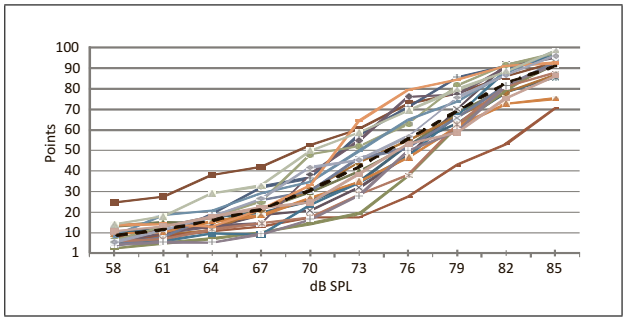

Figure 7. Averaged individual loudness functions (in terms of points) obtained for the 20 subjects along with the overall average (dashed line).

2.8. Conversion from points to loudness level

The individual loudness functions give the relation between SPL and points for each subject (see Figure 7). The key to transform points into loudness level is that the phon scale is equal to the SPL scale for a critical-band-wide noise with a center frequency of 1 kHz in a plane wave and frontal incidence [35, 43]. Thus, it is possible from the individual loudness functions to get the relation between points and phon. A cubic regression model was then fitted to the individual data using a least-square fitting,

$$N_{\text{phons}} = a_i \cdot N_{\text{points}}^3 + b_i \cdot N_{\text{points}}^2 + c_i \cdot N_{\text{points}} + d_1, \quad (1)$$

where (a_i, b_i, c_i, d_i) are the fitting parameters determined for each subject i ($i = 1, 2, .., 20$). N_{phons} is the loudness level expressed in phon and and N_{points} the loudness measured in points. For each subject the point to loudness level conversion was based on their own loudness function. This was because the subjects used the response scale in their own way. They created their own internal reference system which varied from one subject to another. However, as long as the subject kept the same internal reference system throughout the entire perceptual test, it was possible to convert points into loudness level using the estimated individual loudness function from equation (1).

2.9. Perceptual results

The loudness levels, averaged over all listeners, are presented in Figures 8, 9 and 10 respectively for the condition where only filtering was applied (no codec condition, called "bandwidth" in the following), when speech codecs and generic codecs were applied. The three amplification levels (*i.e.* "Nominal+5 dB", "Nominal" and "Nominal-10 dB") as well as the 12 samples (*cf.* Table I) are presented for each filtering/coding condition, *i.e.*, "Bandwidth", "Speech codecs", "Generic codecs".

The normality of the data distribution was verified using the Kolmogorov-Smirnov test (significance level of 0.05). An analysis of variance for repeated measures

Table III. Repeated ANOVA with three factors. The factors were the audio sample (12 levels, Sample 1 to Sample 12), the amplification level (3 levels: Nominal+5dB, Nominal, Nominal-10dB) and the Filtering/Coding condition (4 levels: NB, WB, SWB, FB for Filtering, AMR, AMR-WB, G.729.1, OPUS for the speech codecs, G.711, G.722, G.722.1C, G.719 for the generic codecs).

	Filtering				Speech codecs				Generic codecs			
df	$\sum x^2$	F	p	df	$\sum x^2$	F	p	df	$\sum x^2$	F	p
Audio samples											
11	2442	8.01	≤0.001	11	2678	11.14	≤0.001	11	1594	8.02	≤0.001
Amplification level											
2	75132	276.53	≤0.001	2	79611	282.27	≤0.001	2	74596	300.79	≤0.001
Filtering/Coding											
3	7918	67.88	≤0.001	3	13133	129.25	≤0.001	3	8909	90.19	≤0.001
Audio samples*Amplification level											
22	799	2.63	≤0.001	22	769	2.43	≤0.001	22	823	3.03	≤0.001
Audio samples*Filtering/Coding											
33	573	1.46	0.050	33	655	1.86	0.003	33	808	2.18	≤0.001
Amplification level*Filtering/Coding											
6	310	3.23	0.006	6	219	2.28	0.040	6	186	2.55	0.023
Audio samples*Amplification level*Filtering/Coding											
66	621	0.96	0.566	66	880	1.36	0.030	66	856	1.24	0.095

(ANOVA) was conducted on each coding/filtering condition -"Bandwidth", "Speech codecs", "Generic codecs"- independently. The factors were always the audio samples, the amplification level and coding/filtering condition. The level of significance was always set to 0.05 and the Bonferroni correction was taken into consideration as the analysis was done on separate data of the same experimental plan inducing the significance level to be reduced to 0.017. The results are reported in Table III.

As expected, the loudness level increased significantly with bandwidth extension, including when codecs were applied (from NB to FB as well as from AMR to OPUS and from G.711 to G.719). Furthermore, the loudness level depended significantly on the audio sample ($p < 0.001$). Obviously, changing the amplification level significantly changed the loudness level ($p < 0.001$). The ANOVA showed significant interactions between the audio samples and the amplification level, between the audio sample and the Filtering/Coding, except for the filtering condition ($p = 0.050$). No significant interaction was found between the amplification level and the Filtering/Coding except for the Filtering condition ($p = 0.006$). No interaction was found between the three factors, audio sample, amplification level and Filtering/Coding condition. These points will be discussed in the discussion paragraph.

Considering the measured loudness level averaged over the 12 samples, we can observe that the loudness level varied almost in the same way when only filters (from NB to FB) were applied and when generic codecs (from G.711 to G.719) were applied too; it increases as the bandwidth increases and reaches a plateau either for the WB (G.722 respectively) or the SWB (G.722.1C respectively) condition. Nevertheless, when speech codecs were applied after the filtering, the loudness increased continuously from the AMR to the OPUS conditions. We also note that the smallest difference in loudness level was always observed between the SWB and the FB filtering and the associated coding (between G.729.1 and OPUS for speech coding, between G.722.1C and G.719 for generic coding), except

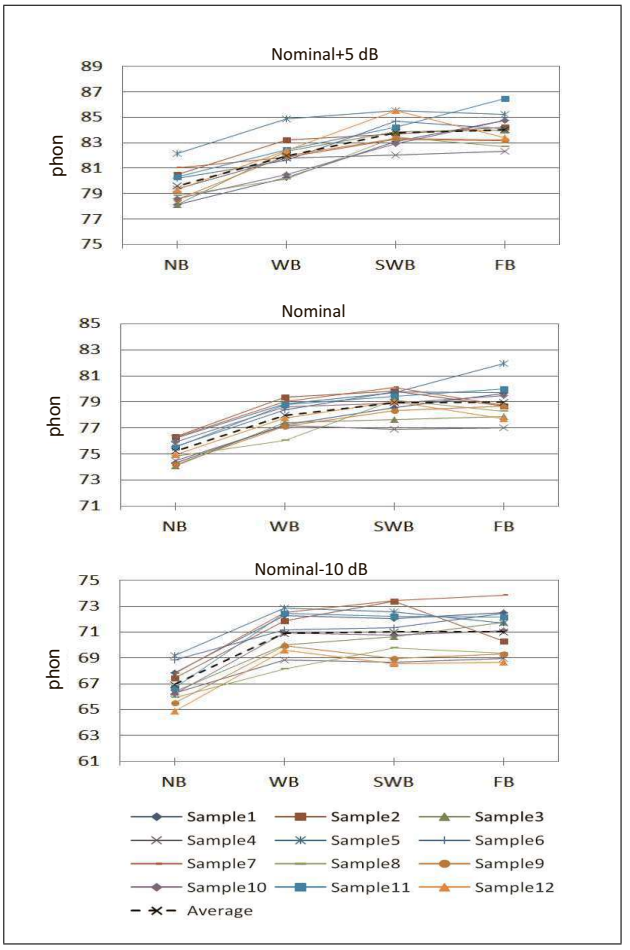

Figure 8. Measured loudness level per sample when no codec was applied (only filtering) for the Nominal +5dB, Nominal and Nominal-10 dB levels in solid lines. The dashed line indicates the measured loudness level averaged over all the samples.

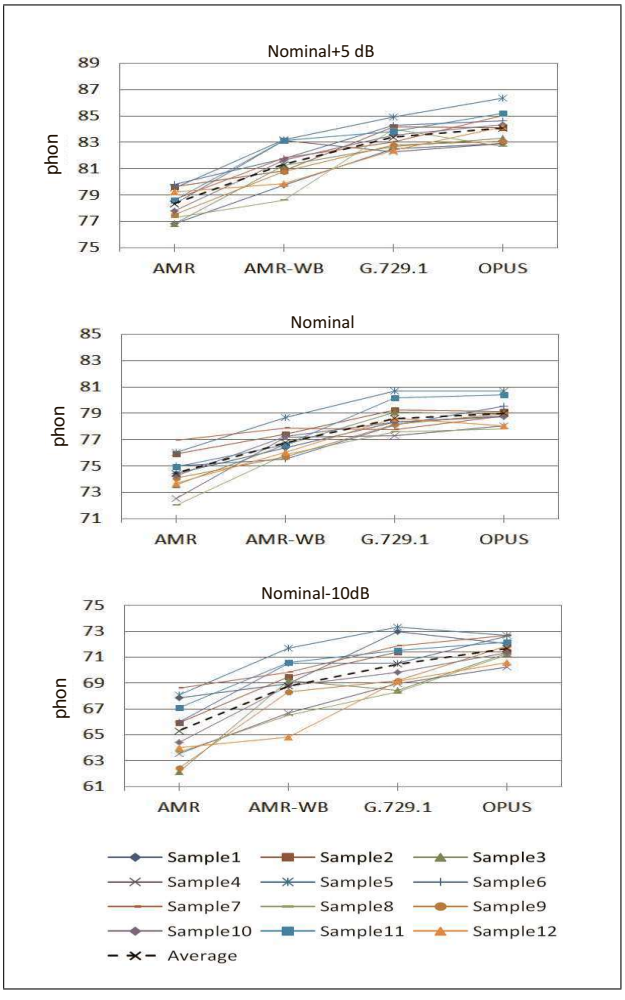

Figure 9. Measured loudness level per sample when speech codecs was applied (after filtering) for the Nominal +5dB, Nominal and Nominal-10 dB levels in solid lines. The dashed line indicates the measured loudness level averaged over all the samples.

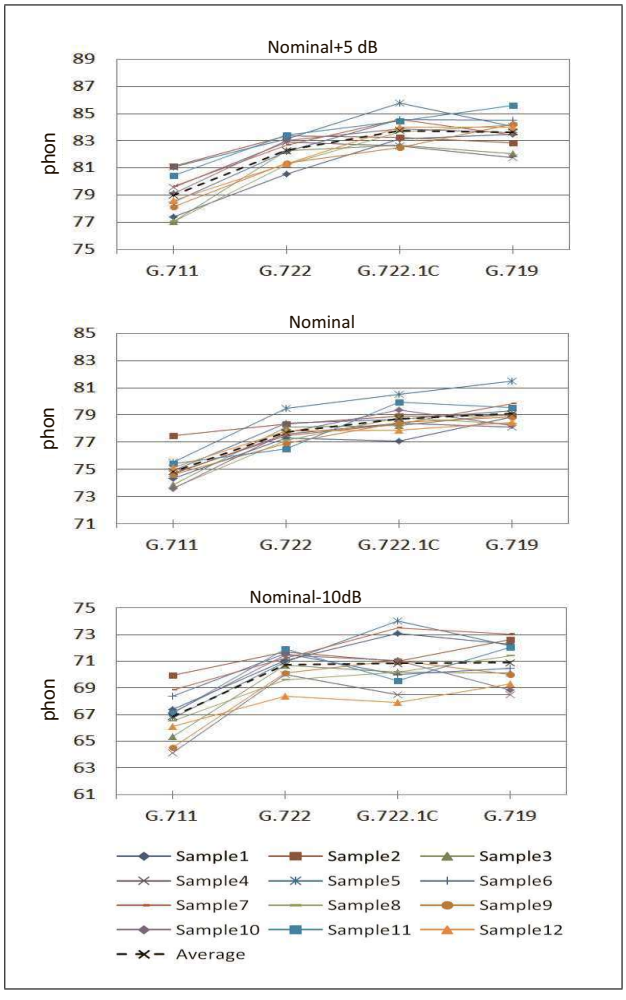

Figure 10. Measured loudness level per sample when generic codecs was applied (after filtering) for the Nominal +5dB, Nominal and Nominal-10 dB levels in solid lines. The dashed line indicates the measured loudness level averaged over all the samples.

for the case of G.729.1 and OPUS in the Nominal-10 dB level.

3. Predictions of loudness models

3.1. Loudness models

Many of the established loudness models are completely or partly based on the original Zwicker's loudness model [12, 44]. Zwicker proposed a sophisticated model that predicts average loudness judgments (in sone) not only as a function of intensity [3], but also depending on the spectral shape [4, 5, 45, 46] of a stationary sound using findings from both physiological acoustics and psychoacoustics [35, 47]. This model accounts for the hearing threshold, the change in loudness with level, the spectral masking of frequency components, and the effect of spectral loudness summation.

Most of the models use a similar structure, as shown in Figure 11. The general algorithm can be summarized as follows: (1) pre-filtering to account for outer and middle ear transmission, (2) construction of excitation patterns, (3) transformation of the excitation pattern into specific loudness (loudness in each auditory filter) and (4) summation of the specific loudness across the auditory frequency scale (Bark [48] or ERB [49]).

Two main families of models exist: one for stationary sounds and one for non-stationary sounds. Models for stationary sounds are based on the long-term spectrum of the signal and they do not account for the effects of signal duration or temporal modulation on loudness. For this family, there are mainly two models. The first one is the original Zwicker's model [12, 44]. It was adopted in the international standard ISO 532-B [50] and in a German standard DIN 45631 [33]. This model will be referred to as DIN 45631 in the rest of the document. The second one is the Moore and Glasberg's model for stationary sounds [51, 52]. Moore and Glasberg modified Zwicker's model to incorporate more recent findings in psychoacoustics, particularly the measurement of the auditory filters using

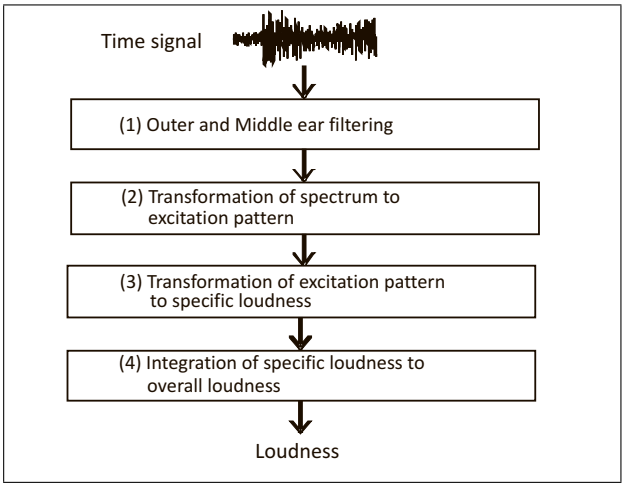

Figure 11. General structure of a loudness model based on the model proposed by Zwicker.

the "notched-noise" technique [53, 54, 55]. The three main differences compared to Zwicker's model are the calculation of the auditory filters, the outer and middle ear filtering and the calculation of the excitation pattern. This second model was adopted in the American standard ANSI S3.4 [34] and will be referred to as ANSI-S3.4 in the rest of the document. Furthermore, a new international standard ISO 532-2 for the determination of stationary sounds based on the American standard ANSI S3.4 is soon to be published.

For non-stationary sounds, the basic principle is the same as for stationary sounds. However, the models have been extended to better cope with time-varying sounds. This was done by modeling the post-masking effect and the temporal integration of loudness [35, 56, 57, 58], hence the loudness is calculated as a function of time and not in a global way. The model of Fastl and Zwicker [35] and the model of Glasberg and Moore [59] were developed to handle both spectral and temporal aspects of loudness for non-stationary sounds. Nevertheless, there exists a variant of the model of Fastl and Zwicker, called the dynamic loudness model DLM [60], with modifications proposed in [61] to consider the variation of spectral loudness summation with duration. In the present study, we chose to focus only on the original models [35, 59] since they are closer to the standards. The Fastl and Zwicker model for non-stationary sounds was adopted recently in the German standard DIN 45631/A1 [62]. A new international standard for the determination of non-stationary sounds is soon to be published as a revision of ISO 532. The future ISO 532-1 standard is based on the DIN 45631/A1 standard for both stationary and time varying sounds according to Zwicker. In order to derive a single value of loudness for the overall signal, Zwicker and Fastl recommend using a statistical indicator such as N4, N5 or N7, which are the loudness values reached and exceeded during 4, 5, or 7 percent of the time, respectively. Zwicker and Fastl recommend using N7 for speech signal [35, p. 319]. Thus, N7 was chosen for this study. This model will be referred to as N7 in the rest of the document.

Glasberg and Moore's model for non-stationary sounds is presented in [59]. They modified their model for stationary sounds to get loudness as a function of time, which they called instantaneous loudness. It would correspond to the overall activity inside the auditory nerve measured on a very short period of time. Then, they calculated the short-term loudness (STL) from the instantaneous loudness by taking into account the temporal masking and the temporal integration. STL corresponds to the loudness perceived during a short segment of sound (a syllable for example). They also derived from the STL the long-term loudness LTL, which is used to describe the loudness sensations that are built rather slowly. According to Glasberg and Moore [59], the loudness of brief duration sounds should be calculated as the maximum value of the STL time evolution; for the loudness of long sounds the averaged LTL well describes the speech loudness. Thus, averaged LTL was chosen for the current study. This model will be referred to as TVL (Time Varying Loudness) model in the rest of the document.

The implementation of the loudness models used in this study is a MatLab implementation developed by the company GENESIS. An evaluation of this implementation compared to the original implementations has been carried out in [63].

3.2. Comparison between evaluated and predicted loudness

Our protocol was designed to obtain the loudness level of the 432 test signals expressed in phon. Then the loudness levels predicted by the models were compared with the measured loudness levels. The signals at the input of the models were not modified since the loudspeaker in our experience was equalized to have a flat frequency response (from 50 Hz to 20 kHz) at the point M (*cf.* Figure 3).

In the following, we present both perceptual results and model predictions for loudness level. For each condition (36 conditions in total, *cf.* Section 2.1.2), the results are averaged over all listeners and all samples. The perceptual results come with confidence interval (CI) at 95%. The model predictions are presented with error bars indicating the standard deviation over the 12 samples. In Figures 12 and 13, all conditions are represented, *i.e.* "Bandwidth", "Speech codecs", "Generic codecs" as well as the three levels, *i.e.* "Nominal+5 dB", "Nominal" and "Nominal-10 dB". Note that the results presented in Figure 12 and 13 are averaged over the 12 samples (*cf.* Table I), because our purpose was to study the overall performance of models on a variety of audio signals.

Results in Figure 12 and 13 show that all models were able to predict the increase in loudness level as the bandwidth increases (from NB to FB, from AMR to OPUS, and from G.711 to G.719). The models were also able to detect the effect reported in Section 2.9, *i.e.*, the changes in loudness level when filtering was applied were similar to the changes in loudness level once generic codecs were

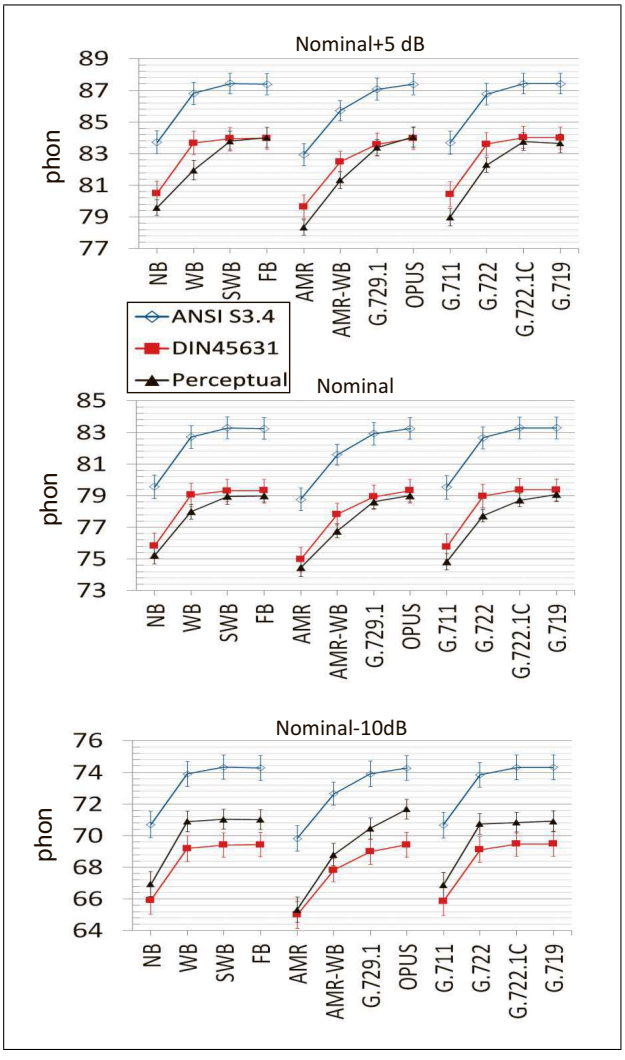

Figure 12. (Colour online) Loudness level averaged over all the samples for DIN 45631 (red lines, filled squares) and ANSI S3.4 (blue lines, open diamonds) models and measured loudness level averaged over all samples and listeners (black lines, filled triangles). The blue and red error bars represent +/- standard deviation over the 12 samples. The black bars represent CI at 95% over the samples and the listeners. The x-axis represents the filtering and coding/decoding conditions applied to the samples (*cf.* Table II).

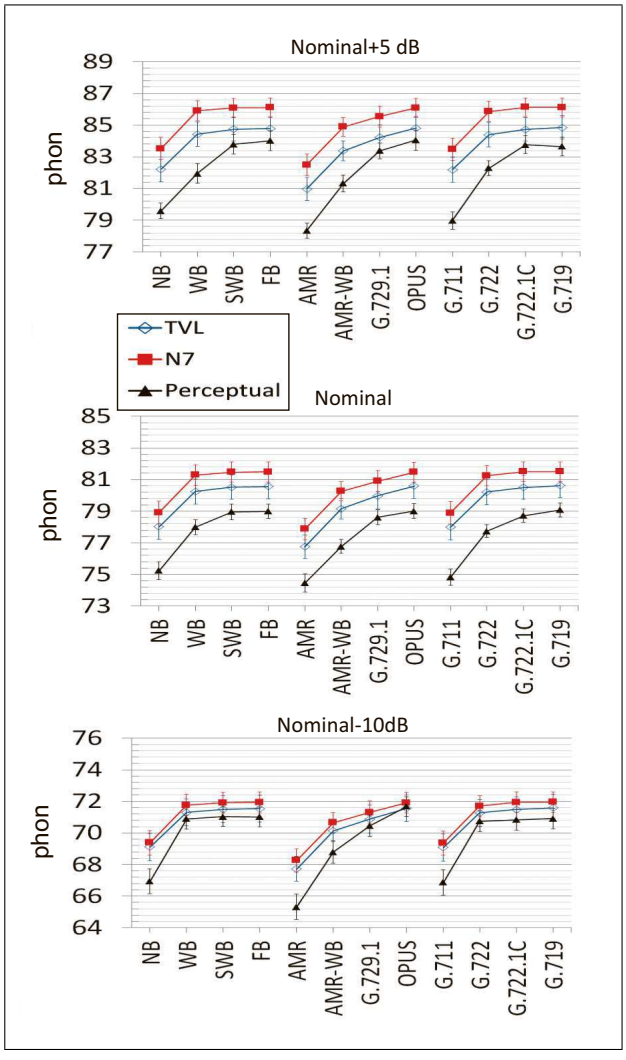

Figure 13. (Colour online) Loudness level averaged over all the samples for N7 (red lines, filled squares) and TVL (blue lines, open diamonds) models and measured loudness level averaged over all samples and listeners (black lines, filled triangles). The blue and red bars represent +/- standard deviation over the 12 samples. The black bars represent CI at 95% over the samples and the listeners. The x-axis represents the filtering and coding/decoding conditions applied to the samples (*cf.* Table II).

applied after filtering. However the changes in loudness level when speech codecs were applied after filtering were different.

Although the models predict rather well the changes in measured loudness level, an offset between predicted and measured loudness levels was observed depending on the model and on the tested amplification level. Different measures can be used to quantify the error of the model predictions. We decided to use the three following measures: the mean absolute error *MAE*, the residual mean R_{mean} and the residual standard deviation R_{STD}.

Assume that the calculated loudness level is noted $L_{\text{calculated}}$ and the measured loudness level is noted L_{measured}. The measures *MAE*, R_{mean} and R_{STD} were calculated over N cases of comparison (measurement / prediction).

The mean absolute error *MAE* is defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|L_{\text{calculated}}(i) - L_{\text{measured}}(i)\right|. \quad (2)$$

MAE is used to measure how close predictions are to the measures. As *MAE* is close to zero as the predictions are close to measures.

The residual mean R_{mean} is defined as

$$R_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N}\left(L_{\text{calculated}}(i) - L_{\text{measured}}(i)\right). \quad (3)$$

R_{mean} shows whether prediction errors are evenly distributed around a mean value. R_{mean} close to zero indicates that the errors of the model are evenly distributed around a

Table IV. The four loudness models (DIN 45631, ANSI 3.4, N7 and TVL) evaluated using three measures *MAE*, R_{mean} and R_{STD}.

Measures	Models	All levels	5dB	0dB	-10dB
MAE	DIN 45631	0.9	0.8	0.7	1.3
	ANSI S3.4	4.0	4.1	4.5	3.5
	N7	2.5	3.1	3.0	1.4
	TVL	1.6	1.7	2.1	1.0
R_{mean}	DIN 45631	0.0	0.7	0.7	-1.3
	ANSI S3.4	4.0	4.1	4.5	3.5
	N7	2.5	3.1	3.0	1.4
	TVL	1.6	1.7	2.1	1.0
R_{STD}	DIN 45631	1.1	0.7	0.7	1.1
	ANSI S3.4	0.6	0.3	0.4	0.6
	N7	1.1	0.7	0.7	0.9
	TVL	0.9	0.4	0.6	0.6

mean value. The sign of R_{mean} describes whether the errors are distributed above ($R_{\text{mean}} > 0$) or below ($R_{\text{mean}} < 0$) the measured loudness. Indeed, if the model has a systematic bias, *MAE* and R_{mean} would have the same value.

The residual standard deviation R_{STD} is defined as

$$R_{\text{STD}} = \left[\frac{1}{N-1}\sum_{i=1}^{N}\left[\left(L_{\text{calculated}}(i) - L_{\text{measured}}(i)\right) - R_{\text{mean}}\right]^2\right]^{1/2}. \quad (4)$$

R_{STD} allows us to know if the model has a systematic bias. R_{STD} close to zero indicates that the model suffers from a bias which value is given by R_{mean}. If it is possible to cancel this bias, then predictions should be accurate and close to the measured loudness level.

Table IV shows the results of *MAE*, R_{mean} and R_{STD} calculated in two cases. In the first case (*cf.* Table IV, column 3) the computing is done over the 36 conditions (all codecs/filtering and levels, $N = 36$). This can be seen as the global performance of the model. In the second case (*cf.* Table IV, column 4 to column 6), the computing is done over all codecs/filtering for each amplification level ($N = 12$). This can be seen as the performance of the model relative to each amplification level.

Table IV shows that R_{mean} is positive for all the models and all the amplification levels, except for DIN 45631 in the Nominal-10 dB level (where $R_{\text{mean}} = -1.3$ phon). This means that the models globally overestimate the measured loudness level, except DIN 45631, which underestimates the measured loudness level in the case of Nominal-10 dB level.

Regarding the overall performance of the models (*cf.* Table IV, column 3), we can make three findings: First, *MAE* indicates that models with the closest predictions on average to the measured loudness level were, ranked in order: DIN 45631, TVL, N7 and finally ANSI S3.4. DIN 45631 was the best model if no correction is *a priori* applied to the results. Secondly, *MAE* and R_{mean} had the same value for all the models except for DIN 45631. This means that the three models (ANSI S3.4, N7 and TVL) have a systematic error that can be eliminated by applying *a priori* the corresponding R_{mean} to the models predictions. For DIN 45631, $R_{\text{mean}} = 0$ phon, which shows that the errors of the model (*MAE* = 0.9 phon) are evenly distributed around the measured loudness. As already discussed, this is due to the underestimation of the data by the model at the level Nominal-10dB.

Finally, R_{STD} indicates that the models which best follow the changes in measured loudness level were, ranked in order: ANSI S3.4, TVL, DIN 45631 and finally N7. This means that the errors of prediction of ANSI S3.4 overestimate the loudness by almost a constant value of 4 phon whatever the level, the codec and the bandwidth, and if this error of 4 phon is corrected, ANSI S3.4 would be the model that predicts the best the measured loudness. Note that the N7 model was the worst model, as we consider that N7 showed the highest R_{STD} value with an R_{mean} value of 2.5 phon.

4. Discussion

The interaction between the audio sample and the amplification shown in Table III might be due to the fact that the loudness function depends on the spectrum of the signal. The different samples having different spectra (male or female voices, only speech, speech and noise, speech and music), their loudness functions might be different and thus the relationship between the loudness and the level (amplification) will depend on the audio sample. The same explanation can be drawn concerning the interaction between the amplification and the Filtering/Coding, as different Filtering/Coding might induce different loudness functions. But it does not explain why this interaction is only observed in the Filtering condition.

The interaction between the audio sample and the Coding conditions could be explained by the fact that the effect of the codec is different depending on the samples. Indeed, it was shown in [64] that some codecs have different frequency response depending on the input signal. The interaction was not significant when only filtering was applied, because it induces fewer changes to the signal spectrum.

The measured loudness levels are very similar for the cases where only bandwidth limitation is applied or when the generic codecs are also applied after filtering. This can be explained by the fact that the signal processing involved in generic codecs does not alter the speech spectrum significantly. However, speech codecs often introduce significant changes to speech spectrum because of the way they handle speech processing [65]. This can explain the fact that the changes in loudness level for the speech codec conditions were different from filtering and generic codec conditions. All tested models predict similar values for the loudness relative to the FB conditions (FB, OPUS, and G.719) and to the SWB conditions (SWB, G.729.1 and G.722.1C). This can be easily explained by the frequency limitation of the models. In fact, the frequencies considered by the Zwicker models range between 22 Hz

and 14030 Hz, and those considered by the Moore models range between 54 Hz and 15062 Hz. Thus, the larger FB frequency range compared with SWB is not taken into account by the models. It should be emphasized that this additional part did not bring large perceptual differences anyway (see Figures 12 and 13), which is consistent with the model behavior. Thus, the frequency limitation of the models might not be a problem for predicting loudness from NB to FB conditions.

The loudness of speech signals used in our experiment was always overestimated by the ANSI S3.4 model. This overestimation has already been reported in many studies [66, 67, 68] when the ANSI S3.4 model was evaluated on broadband sounds. Schlittenlacher *et al.* [67] measured the loudness of pink noise and compared it to ANSI S3.4 model predictions. They found that the predictions were always higher than the perceptual results (up to 5 phon at moderate and high sound levels). However, it was mentioned that predictions were better for low levels. This result is consistent with ours where the smallest differences between the measured and the predicted loudness were observed for the Nominal-10dB level. Schlittenlacher *et al.* [69, 70] mentioned two reasons for this overestimation: the first one is that the model used 40 equivalent rectangular bandwidth (ERB) filters to model peripheral filtering by the auditory system. In contrast, other models are based on 24 Bark channels. Despite the slight difference in the exponent of the compressive loudness transformation between the loudness models, we can suppose that using a higher number of auditory filters would lead to greater spectral summation. The second reason was a possible overestimation of the specific loudness when it is calculated around 3 kHz [70]. This is also supported by the fact that ANSI S3.4 predicts significantly greater loudness value than ISO226:2003 [43] around 3 kHz. Moore and Glasberg also noted this limitation for their model in [49].

This overestimation can also be partly explained by the phenomenon of binaural loudness summation. In fact, the ANSI S3.4 model calculates the loudness for one ear and then it is simply multiplied by the binaural-to-monaural loudness ratio for binaural presentation. Moore and Glasberg [52] assumed a perfect summation of loudness between the ears (binaural-to-monaural loudness ratio of 2). This assumption was based on earlier data [3, 4, 71, 72, 73]. However, more recent studies (see an overview in [74]) employing a greater variety of methodologies and sounds tend to obtain smaller binaural-to-monaural loudness ratios, in the range 1.2-1.5. A binaural-to-monaural loudness ratio of 2 means that the sound is 10 phon louder in binaural listening compared to monaural listening. A binaural-to-monaural loudness ratio of 1.5 gives a difference of 5.9 phon between monaural and binaural listening. Thus, in the model, if a loudness ratio less than 2 were used, the loudness level calculated would be lower and could correspond to the measured data. Moore and Glasberg recently updated their model to handle the binaural loudness summation more efficiently. They introduced a binaural inhibition model [75] and a binaural-to-monaural loudness ratio of 1.5 for binaural diotic listening.

Several studies showed a rather good loudness estimation of broadband sounds using DIN 45631. Schlittenlacher *et al.* [67] measured the loudness of pink noise and compared it to DIN 45631 model predictions. They found that the predictions were always within the interquartile range of the perceptual results. Meunier *et al.* [66] tested many broadband sounds including speech, and showed that the DIN 45631 model well estimated loudness for loud sounds (>70 phon) but underestimated loudness for sounds less than 70 phon. These results were consistent with what we obtained when comparing measured loudness with DIN 45631 predictions. Rennies *et al.* [14] showed good accuracy between measured loudness and loudness predicted by DIN 45631 for speech and speech-like signals at moderate sound levels. The authors suggested that the loudness of speech would be largely related to its long-term spectrum.

On the whole, loudness is overestimated by loudness models for time-varying sounds. As also shown in [14, 76], the TVL model has better predictions compared with the N7 model. A large overestimation by the N7 model was observed although the recommendation of Zwicker to use the percentile loudness N7 for speech signal [35, p. 319] instead of N5 for other time-varying signals was respected. For the Nominal-10dB level, however, measured and estimated loudness values agree rather well.

As recalled in Section 3.1, all loudness models were based on Zwicker's model. The stage of calculation of the specific loudness has the greatest importance. In this stage, the contribution of each critical band to the overall loudness is taken into account. The specific loudness is calculated from the excitation pattern and, according to Stevens's law [77], the relationship between excitation and specific loudness is a power function. Zwicker adjusted this function to better predict the empirical results of loudness growth functions for different types of sound. Several studies [78, 79, 80] have questioned the approximation of the loudness function by a simple power function. They advanced evidence that the slope (exponent) of the loudness function is smaller at moderate SPLs than at lower or higher ones. Consequently, they suggested the need to modify Stevens's simple power function with a more complex function. Florentine and Epstein described the revised power law as an "inflected exponential" or InEx law [80]. The overestimation and underestimation observed on our results, which depended on the level condition (Nominal-10dB, Nominal, and Nominal+5dB levels), lead us to think about the possible effect of using a simple power function (when the specific loudness is calculated) with a fixed exponent on the models predictions. We think that using a new formula of loudness function to calculate the specific loudness may enhance the model predictions.

Globally, whatever the model used to predict loudness, the difference between the prediction and the measure depends on SPL. The discrepancies between models and

data are very similar for "Nominal" and "Nominal+5dB" whereas at "Nominal-10dB" it does not follow the same tendency. In DIN 45631 and ANSI S3.4 the loudness is calculated over the long-term spectrum of the signal and in N7 and TVL the loudness is calculated over time, taking into account temporal integration. It is not obvious from our data that using temporal loudness models improves the model prediction for speech signals. Nevertheless, the TVL model resulted in the smallest errors by predicting loudness levels 1.6 phon higher than those measured. The absolute difference was about 4 phon when using ANSI S3.4. DIN 45631 predictions were the "least bad", the maximum reported deviation from the measured value was 1.3 phon. It should be noted that the stimuli were quite long (6 to 10.2s) compared to those used in Rennies *et al.* [14]. However, we found, as in [14], that the loudness is better estimated by a model that uses long-term spectrum than by a model considering the fluctuation of the sound over time.

5. Conclusion

In this article, we have investigated loudness of speech transmitted through realistic telephone system paths and the ability of existing loudness models to predict perceived loudness. The designed perceptual test allowed the loudness level of the test signals to be measured. The measured loudness was compared with predictions of four current loudness models for stationary and non-stationary sounds. These loudness models have already been standardized or are on the way to be standardized.

As expected, the measured loudness increased as the bandwidth increased (from NB to FB as well as from AMR to OPUS and from G.711 to G.719) and all models accurately predicted this effect. The measured loudness for FB conditions (FB, OPUS, and G.719) and SWB conditions (SWB, G.729.1 and G.722.1C) were almost identical. Thus, the frequency limitation of the tested loudness models was not an obstacle for predicting loudness from NB to FB conditions. However, the difference between the predicted and the measured loudness depends on SPL: (i) DIN 45631 overestimated loudness at Nominal+5 dB level and underestimated it at Nominal-10 dB level, (ii) ANSI S3.4 overestimated loudness at all levels, (iii) TVL and N7 overestimated loudness for Nominal and Nominal+ 5 dB levels. It is not obvious from our results that using temporal loudness models would improve the prediction of loudness for speech signals. DIN 45631 yielded the best predictions for Nominal and Nominal+5dB levels. This supports the hypothesis that the long-term spectrum is the main factor for the determination of speech loudness, as suggested in [14].

References

[1] N. Coté, V. Gautier-Turbin, S. Möller: Influence of loudness level on the overall quality of transmitted speech. Audio Engineering Society Convention 123. Audio Engineering Society, 10/2007.

[2] ITU-T Handbook: Handbook on telephonometry. International Telecommunication Union, Geneva, 1993.

[3] H. Fletcher, W. A. Munson: Loudness, its definition, measurement and calculation. Bell System Technical Journal **12** (1933) 377–430.

[4] H. Fletcher, W. A. Munson: Relation between loudness and masking. J. Acoust. Soc. Am. **9** (1937) 1–10.

[5] S. Möller: Assessment and prediction of speech quality in telecommunications. Springer Science & Business Media, 2000.

[6] D. L. Richards: Loudness ratings of telephone speech paths. Proceedings of the Institution of Electrical Engineers, IET Digital Library **118** (1971) 423–436.

[7] D. L. Richards: Telecommunication by speech. Halsted Press Division, Wiley, 1973.

[8] ITU-T Recommendation P.79: Calculation of loudness ratings for telephone sets. International Telecommunication Union, Geneva, 2011.

[9] K. A. Woo, R. Ceruti, J. Bareham: Wide-band loudness ratings confusion (ref ITU-T P. 79). ITU-T STQ (12) 107, 2007.

[10] J. Y. Monfort, C. Quinquis, L. Clarimon, J. F. Dolidet: Proposal for handsfree/handset when implemented in a single terminal. ITU-T STQ (10) 0134,, 10/2010.

[11] ITU-T Recommendation P.10 Amendment 3: Vocabulary for performance and quality of service. International Telecommunication Union, Geneva, 2011.

[12] E. Zwicker: Über psychologische und methodische Grundlagen der Lautheit. Acustica **8** (1958) Supplement 1, 237–258.

[13] ITU-T Recommendation P.501: Test signals for use in telephonometry. International Telecommunication Union, Geneva, 01/2012.

[14] J. Rennies, I. Holube, J. L. Verhey: Loudness of speech and speech-like signals. Acta Acustica united with Acustica **99** (2013) 268–282.

[15] T. Brand, V. Hohmann: Effect of hearing loss, centre frequency and bandwidth on the shape of the loudness function in categorical loudness scaling. Audiology **40** (2001) 92–103.

[16] B. C. Moore, D. A. Vickers, T. Baer, S. Launer: Factors affecting the loudness of modulated sounds. J. Acoust. Soc. Am. **105** (1999) 2757–2772.

[17] R. Warren: Anomalous loudness function for speech. J. Acoust. Soc. Am. **54** (1973) 390–396.

[18] J. Brandt, K. Ruder, T. Shipp: Vocal loudness and effort in continuous speech. J. Acoust. Soc. Am. **46** (1969) 1543–1548.

[19] M. Mendel, H. Sussman, R. Merson, M. Naeser, F. Minifie: Loudness judgments of speech and non-speech stimuli. J. Acoust. Soc. Am. **46** (1969) 1556–1561.

[20] G. D. Allen: Acoustic level and vocal efforts cues for the loudness of speech. J. Acoust. Soc. Am. **49** (1971) 1831–1841.

[21] H. Fastl: Loudness of running speech. J. Audiol. Technique **16** (1977) 2–13.

[22] B. C. Moore, B. R. Glasberg, M. A. Stone: Why are commercials so loud? Perception and modeling of the loudness of amplitude-compressed speech. J. Audio Eng. Soc. **51** (2003) 1123–1132.

[23] I. Pollack: The effect of white noise on the loudness of speech of assigned average level. J. Acoust. Soc. Am. **21** (1949) 255–258.

[24] J. M. Valin, K. Vos, T. Terriberry: Definition of the opus audio codec. Internet Engineering Task Force (IETF) RFC6716, 09/2012.

[25] ITU-T Recommendation G.719: Low-complexity, full-band audio coding for high-quality, conversational applications. International Telecommunication Union, Geneva, 06/2008.

[26] ITU-T Recommendation G.729.1: G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bit stream interoperable with G.729. International Telecommunication Union, Geneva, 05/2006.

[27] 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. International Telecommunication Union, Geneva, 05/2005.

[28] 3GPP TS 26.190: Speech codec speech processing functions; Adaptive multi-rate - wideband (AMR-WB) speech codec; Transcoding functions. The 3rd Generation Partnership Project, 2011.

[29] ITU-T Recommendation G.722: 7 kHz audio-coding within 64 kbit/s. International Telecommunication Union, Geneva, 09/2012.

[30] 3GPP TS 26.090: Mandatory speech codec speech processing functions; AMR speech codec; General description. The 3rd Generation Partnership Project, 2000.

[31] ITU-T Recommendation G.711: Pulse code modulation (PCM) of voice frequencies. International Telecommunication Union, Geneva, 11/1988.

[32] I. Edjekouane, C. Plapous, C. Quinquis, S. Meunier: Speech and audio loudness depending on telephone audio bandwidth and codec – subjective testing approach. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, 1325–1329.

[33] DIN 45631: Berechnung des Lautstärkepegels und der Lautheit aus dem Geräuschspektrum - Verfahren nach E. Zwicker (procedure for calculating loudness level and loudness). Deutsches Institut für Normung, 1991.

[34] ANSI S3.4-2007: American national standard. Procedure for the computation of loudness of steady sound. American National Standards Institute, 2007.

[35] H. Fastl, E. Zwicker: Psychoacoustics: Facts and models. 3rd ed. Springer, Berlin, 2007.

[36] B. R. Glasberg, B. C. Moore: A model of loudness applicable to time-varying sounds. J. Audio Eng. Soc. **50** (2002) 331–342.

[37] ETSI ES 202 396-1: Speech processing, transmission and quality aspects (STQ); Speech quality performance in the presence of background noise. ETSI, 2012.

[38] ITU-T Recommendation P.58: Head and torso simulator for telephonometry. International Telecommunication Union, Geneva, 2013.

[39] R. Teghtsoonian: Range effects in psychophysical scaling and a revision of Steven's law. The American journal of psychology (1973) 3–27.

[40] L. E. Marks, E. Warner: Slippery context effect and critical bands. Journal of Experimental Psychology: Human Perception and Performance **17** (1991) 986–996.

[41] ISO 16832:2006: Acoustics-loudness scaling by means of categories. International Organization for Standardization, Geneva, 2006.

[42] T. Brand, V. Hohmann: An adaptive procedure for categorical loudness scaling. J. Acoust. Soc. Am. **112** (2002) 1597–1604.

[43] ISO 226: Acoustics-normal equal-loudness-level contours. International Organization for Standardization, Geneva, 2003.

[44] E. Zwicker: Ein Verfahren zur Beredinung der Lautstärke. Acustica **10** (1960) Supplement 1, 304–308.

[45] E. Zwicker, G. Flottorp, S. S. Stevens: Critical bandwidth in loudness summation. J. Acoust. Soc. Am. **29** (1957) 548–557.

[46] B. Scharf: Loudness of complex sounds as a function of the number of components. J. Acoust. Soc. Am. **31** (1959) 783–785.

[47] E. Zwicker, R. Feldtkeller: The ear as a communication receiver. Acoustical Society of America, Woodbury, 1999.

[48] E. Zwicker, E. Terhardt: Analytical expressions for critical band rate and critical bandwidth as a function of frequency. J. Acoust. Soc. Am. **68** (1980) 1523–1525.

[49] B. C. Moore, B. R. Glasberg: A revision of Zwicker's loudness model. Acta Acustica united with Acustica **82** (1996) 335–345.

[50] ISO 532: Acoustics - Method for calculating loudness level. International Organization for Standardization, Geneva, 1975.

[51] B. C. J. Moore, B. R. Glasberg, T. Baer: A model for the prediction of thresholds, loudness and partial loudness. J. Audio Eng. Soc. **45** (1997) 224–240.

[52] B. R. Glasberg, B. C. J. Moore: Prediction of absolute thresholds and equal-loudness contours using a modified loudness model. J. Acoust. Soc. Am. **120** (2006) 585–588.

[53] R. D. Patterson, B. C. J. Moore: Auditory filters and excitation patterns as representations of frequency resolution. Frequency Selectivity in Hearing (1986) 123–177.

[54] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice: An efficient auditory filterbank based on the gammatone function. Meeting of the IOC Speech Group on Auditory Modelling at RSRE, 2(7), 1987.

[55] B. R. Glasberg, B. C. Moore: Derivation of auditory filter shapes from notched-noise data. Hearing research **47** (1990) 103–138.

[56] E. Zwicker: Procedure for calculating loudness of temporally variable sounds. J. Acoust. Soc. Am. **62** (1977) 675–682.

[57] S. Buus, M. Florentine, T. Poulsen: Temporal integration of loudness, loudness discrimination, and the form of the loudness function. J. Acoust. Soc. Am. **101** (1997) 669–680.

[58] J. L. Verhey: Psychoacoustics of spectro-temporal effects in masking and loudness perception. BIS Universität Oldenburg, 1999.

[59] B. R. Glasberg, B. C. Moore: A model of loudness applicable to time-varying sounds. J. Audio Eng. Soc. **50** (2002) 331–342.

[60] J. Chalupper, H. Fastl: Dynamic loudness model (DLM) for normal and hearing-impaired listeners. Acta Acustica united with Acustica **88** (2002) 378–386.

[61] J. Rennies, J. L. Verhey, J. Chalupper, H. Fastl: Modeling temporal effects of spectral loudness summation. Acta Acustica united with Acustica **95** (2009) 1112–1122.

[62] DIN 45631/A1: Berechnung des Lautstärkepegels und der Lautheit aus dem Geräuschspektrum - Verfahren nach E. Zwicker - Änderung 1: Berechnung der Lautheit zeitvarianter Geräusche (Calculation of loudness level and loudness from the sound spectrum - Zwicker method - Amendment 1: Calculation of the loudness of time-variant sound). Deutsches Institut für Normung, 2008.

[63] S. Molla, I. Boullet, S. Meunier, G.Rabau, B. Gauduin, P. Boussard: Calcul des indicateurs de sonie: revue des al-

gorithmes et implémentation. 10ème Congrès Français d'Acoustique, 2010.

[64] 3GPP TR 26.976: AMR-WB speech codec performance characterization. 3GPP Technical Report.

[65] J. H. Chen, J. Thyssen: Analysis-by-synthesis speech coding. – In: Springer Handbook of Speech Processing. Springer, Berlin Heidelberg, 2008, 351–392.

[66] S. Meunier, A. Marchioni, G. Rabau: Subjective evaluation of loudness models using synthesized and environmental sounds. Proc. Inter-Noise, Nice, France, 2000, 2205–2209.

[67] J. Schlittenlacher, T. Hashimoto, H. Fastl, S. Namba, S. Kuwano, S. Hatano: Loudness of pink noise and stationary technical sounds. Proc. Inter-Noise, 2011, 2314–2318.

[68] H. Fastl, F. Völk, M. Straubinger: Standards for calculating loudness of stationary or time-varying sounds. Proc. Inter-Noise, Ottawa, Canada, 2009, unpaginated.

[69] J. Schlittenlacher, W. Ellermeier, T. Hashimoto: Loudness model extension improving predictions for broadband sounds. Proc. Inter-Noise, 2012, 5495–5505.

[70] J. Schlittenlacher, H. Fastl, T. Hashimoto, S. Kuwano, S. Namba: Differences of loudness algorithms across the frequency spectrum. Tagungsband Fortschritte der Akustik-DAGA 2012, Darmstadt, 2012.

[71] L. E. Marks: Binaural summation of the loudness of pure tones. J. Acoust. Soc. Am. **64** (1978) 107–113.

[72] R. P. Hellman, J. Zwislocki: Monaural loudness function at 1000 cps and interaural summation. J. Acoust. Soc. Am. **35** (1963) 856–865.

[73] D. Algom, B. Ben-Aharon, L. Cohen-Raz: Dichotic, diotic, and monaural summation of loudness: A comprehensive analysis of composition and psychophysical functions. Perception & Psychophysics **46** (1989) 567–578.

[74] V. P. Sivonen, W. Ellermeier: Binaural loudness. – In: Loudness. Springer, New York, 2011, 169–197.

[75] B. C. Moore, B. R. Glasberg: Modeling binaural loudness. J. Acoust. Soc. Am. **121** (2007) 1604–1612.

[76] J. Rennies, J. L.Verhey, H. Fastl: Comparison of loudness models for time-varying sounds. Acta Acustica united with Acustica **96** (2010) 383–396.

[77] S. S. Stevens: On the psychophysical law. Psychological review **64** (1957) 153.

[78] M. Florentine, S. Buus, T. Poulsen: Temporal integration of loudness as a function of level. J. Acoust. Soc. Am. **99** (1996) 1633–1644.

[79] S. Buus, M. Florentine: Modifications to the power function for loudness. Fechner Day 2001, Berlin, 2001, 236–241.

[80] M. Florentine, M. Epstein: To honor Stevens and repeal his law (for the auditory system). Fechner Day 2006. Proceedings of the 22nd Annual Meeting of the International Society for Psychophysics, 2006, 37–42.