Lautheit von Sprachsignalen

Roland Sottek[1], Bianca Wiercinski[2]

[1] *HEAD acoustics GmbH, 52134 Herzogenrath, E-Mail: roland.sottek@head-acoustics.de*
[2] *Hochschule Düsseldorf, 40476 Düsseldorf, E-Mail: bianca.wiercinski@fh-duesseldorf.de*

Einleitung

Die Lautheit, eine der zentralen Empfindungsgrößen der Psychoakustik, wird als ein Kriterium für die Qualität von Telekommunikationseinrichtungen hinsichtlich der Sprachsignalübertragung immer bedeutender. Dennoch wurden in Messstandards für diesen Zweck bisher lediglich recht stark vereinfachte instrumentelle Berechnungen zur Auswertung verwendet (z. B. Loudness Rating nach ITU-T Empfehlung P.79). Die Lautheit stationärer sowie instationärer synthetischer Signale wurde bereits tiefgehend erforscht und lässt sich mittlerweile gut durch bestehende Berechnungsmethoden annähern. Es stellt sich jedoch die Frage, wie gut sich diese Lautheitsmodelle auch auf Sprachsignale mit komplexen zeitlichen und spektralen Strukturen anwenden lassen. Dies wird anhand von Sprachsignalen mit unterschiedlichen Bandbreiten untersucht. Die berechneten Lautheiten werden mit Ergebnissen aus Hörversuchen verglichen, um zu evaluieren, welche Methode die empirischen Daten am besten abbildet und worin potenzielle Vorhersagefehler bestehen. Verglichen wird neben dem Standard für zeitabhängige Lautheit ISO 532-1 [1] (entspricht weitestgehend der deutschen Norm DIN 45631/A1 [2]) und der „Time-Varying-Loudness“ nach Moore und Glasberg (TVL) [3] auch die Lautheit nach dem Gehörmodell von Sottek (HML) [4].

Die Lautheitsmodelle

Das von Zwicker entwickelte Verfahren [5] bildet die Grundlage für viele weitere Lautheitsmodelle, die alle im Wesentlichen aus den gleichen Bausteinen bestehen. Das Blockschaltbild in Abbildung 1 gibt einen Überblick.

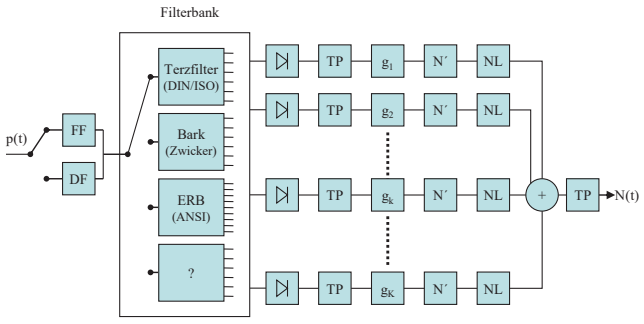

Abbildung 1: Blockschaltbild der Lautheitsberechnung

Die Lautheitsberechnung beginnt mit dem Schalldrucksignal p(t). Zunächst wird zwischen freiem und diffusem Schallfeld unterschieden (FF/DF). Dann folgt eine Zerlegung in Teilbänder durch eine Filterbank. Nach Gleichrichtung, Tiefpassfilterung, frequenzabhängiger Gewichtung g_k und nichtlinearer Transformation (Block N‘) ergeben sich die jeweiligen Teillautheiten. An dieser Stelle wird berücksichtigt, dass die Zeitkonstanten der zeitlichen Integration von der Signaldauer abhängen: Mit wachsender Signaldauer verläuft der Lautheitsabfall zunehmend flacher. Zudem werden Effekte der Maskierung miteinbezogen (Block NL, für die zeitabhängige Lautheitsberechnung nach ISO 532-1 und DIN 45631/A1). Nun werden die Teillautheiten zur Gesamtlautheit addiert. Ein weiterer Tiefpass simuliert abschließend, dass sich die Lautheit von Geräuschen etwa verdoppelt, wenn deren Dauer von 10 ms auf 100 ms ansteigt.

Ein zentraler Unterschied zwischen den einzelnen Berechnungsmodellen besteht im Aufbau der Filterbank und in der frequenzabhängigen Gewichtung g_k: DIN 45631/A1 und ISO 532-1 verwenden 28 steilflankige Terzfilter, die Filter mit konstanter Bandbreite auf der Bark-Skala approximieren (gemäß Zwicker) und zur Berechnung der Kernlautheiten dienen, die dann in einem weiteren Schritt um die sogenannten Flankenlautheiten ergänzt werden. Das TVL-Modell nutzt etwa 40 auditorische Filter konstanter Bandbreite auf der ERB-Skala. Statt mithilfe einer Filterbank werden die Teillautheiten durch sechs parallellaufende Fourier-Transformationen mit unterschiedlicher Auflösung erzeugt. So wird in den tiefen Frequenzen eine möglichst hohe spektrale und zugleich in den hohen Frequenzen eine hohe zeitliche Auflösung erreicht. Durch Integration der Teillautheiten wird zunächst die „Instantaneous Loudness“ berechnet, aus der sich dann durch Tiefpassfilterung 1. Ordnung mit unterschiedlichen Zeitkonstanten für steigende und fallende Lautheitsverläufe (22 ms bzw. 50 ms) die „Short Term Loudness (stTVL)“ ergibt. Die „Long Term Loudness (ltTVL)“ wird aus der stTVL durch Anwendung eines zweiten Tiefpassfilters mit den Zeitkonstanten 99 ms für steigende und 2 s für fallende Verläufe berechnet.

Zur Lautheitsberechnung nach dem Gehörmodell von Sottek werden auditorische Filter mit einer konstanten Bandbreite auf der Bark-Skala verwendet. Eine wesentliche Neuerung dieses Modells besteht vor allem in der Anpassung der Nichtlinearitäten. Es nutzt dabei Potenzfunktionen mit unterschiedlichen Exponenten für verschiedene Pegelbereiche. Außerdem können mittels der Auswertung der Autokorrelationsfunktion in den einzelnen Bändern tonale Komponenten separiert werden, um deren Lautheit gesondert zu bewerten [6].

Der Hörversuch

Es wurden Hörversuche zur Lautheit von Sprachsignalen unterschiedlicher Bandbreite durchgeführt. Das Prinzip des hier angewandten Hörversuchsverfahrens ist im Wesentlichen einer Methode von Edjekouane et al. [7] nachempfunden. Das Verfahren gliedert sich in drei Blöcke. Im ersten Teil wird eine sogenannte Lautheitsfunktion erstellt. Hierfür bewerten die Probanden die Lautheit eines schmalbandigen Rauschsignals (in verschiedenen Pegelstufen) auf einer 25-stufigen Skala.

Die so erhaltenen Zahlenwerte werden dann über die dazugehörigen Pegel aufgetragen. Im zweiten Teil werden Sprachsignale auf derselben Skala bewertet.

Werden die für die Sprachsignale erzielten Zahlenwerte in die Umkehrfunktion der beschriebenen Lautheitsfunktion eingesetzt, so lässt sich damit auf den Pegel des Rauschsignals schließen, das als gleichlaut empfunden wurde. Um die Konsistenz in der Beantwortung zu überprüfen, wird die Bewertung der Rauschsignale von jedem Probanden zweimal (vor und nach der Sprachsignalbewertung) durchgeführt.

Die Probanden

Am beschriebenen Hörversuch nahmen 17 normalhörende Probanden (< 15 dB Hörverlust im Frequenzbereich von 125 Hz bis 10 kHz) teil. Die sechs weiblichen und elf männlichen Versuchsteilnehmer waren zwischen 21 und 54 Jahre alt; der Altersdurchschnitt lag bei 28 Jahren. Vier Probanden wurden aufgrund inkonsistenter Bewertungen von der Auswertung ausgeschlossen. Kriterium für den Ausschluss war ein Unterschied zweier Bewertungen desselben Stimulus von mehr als acht Skalenstufen (entspricht mehr als zwei Attributstufen, siehe Kapitel „Die Durchführung“).

Die Stimuli

Zur Bewertung wurden den Versuchsteilnehmern verschiedene Sprachstimuli, ohne Stör- oder Hintergrundgeräusche, präsentiert. Hierzu wurden Signale aus der ITU-T P.501 auf die Länge eines gesprochenen Satzes gekürzt; für die einzelnen Stimuli ergaben sich dadurch Längen zwischen 2,1 und 3,7 Sekunden. Je drei der Stimuli wurden von weiblichen Sprechern, die übrigen drei von männlichen Sprechern gesprochen. Es waren sowohl deutsche als auch englische Sätze enthalten.

Um den Einfluss der Bandbreite auf das Lautheitsempfinden zu untersuchen, wurden die sechs unterschiedlichen Signale jeweils in drei Varianten bandbegrenzt: „Full Band“ (20 Hz – 20 kHz), „Wide Band“ (50 Hz bis 7 kHz) und „Narrow Band“ (300 Hz bis 3,4 kHz). Anschließend wurden die Pegel der nun 18 Stimuli auf die gewünschten Werte (45, 55, 65 und 75 dB SPL) angepasst, sodass alle Signale unabhängig von der Bandbreite in den gleichen Pegelstufen vorlagen. Insgesamt ergaben sich also 6 * 3 * 4 = 72 Stimuli für die Bewertung (Abbildung 2).

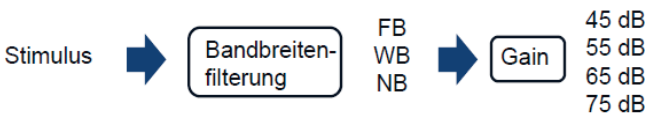

Abbildung 2: Generierung der Stimuli

Abbildung 3 zeigt den Zeitverlauf der sechs Stimuli beispielhaft für die „Full Band“ Variante und die Pegelstufe 65 dB SPL.

Als Referenzsignal wurde aus einem Weißen Rauschen mithilfe eines Bandpassfilters vierter Ordnung ein Schmalbandrauschen mit einer Breite von 3 Bark um die Mittenfrequenz von 1 kHz erzeugt (Dauer: 1 Sekunde).

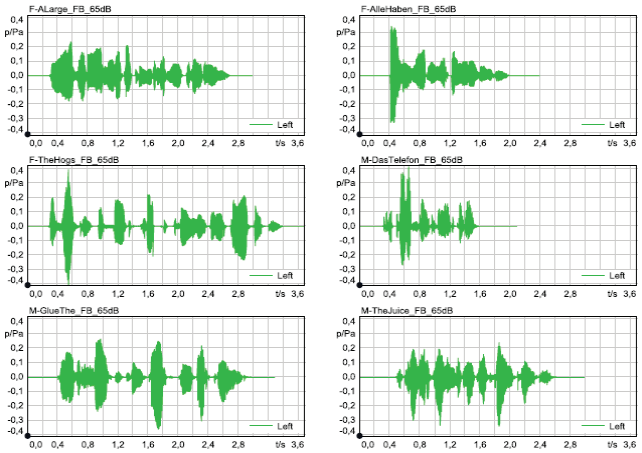

Abbildung 3: Zeitverlauf der Stimuli für die „Full Band“ Variante und den Pegel 65 dB SPL

Auch dieses Signal wurde anschließend auf die gewünschten Pegelstufen (in diesem Fall 40, 45, 50, 55, 60, 65, 70, 75, 80 und 85 dB SPL) gebracht. Dieses Referenzsignal wurde anstatt des von Edjekouane et al. [7] verwendeten steilflankig gefilterten Schmalbandrauschen der Bandbreite 1 Bark gewählt, da Vorversuche größere Unsicherheiten bei der Bewertung dieser stark tonalen Rauschsignale zeigten. Des Weiteren wurde im vorliegenden Versuch auch der Pegelbereich erweitert.

Die Durchführung

Wie eingangs erwähnt, begannen die Probanden mit der Bewertung des Referenzsignals. Dabei wurden die zehn Pegelstufen in pseudorandomisierter Reihenfolge präsentiert (die Pegeldifferenz zwischen zwei aufeinander folgenden Stimuli war nie größer als die Hälfte des gesamten untersuchten Pegelumfangs, also hier: max. 20 dB). Der Proband gab nach jedem präsentierten Stimulus eine Bewertung entsprechend einer 25-stufigen Skala ab. Zur Orientierung waren der Skala sieben Attribute zugeordnet. Abbildung 4 zeigt die verwendete Skala mit den Attributen.

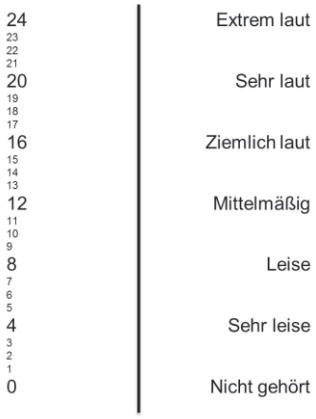

Abbildung 4: 25-stufige Bewertungsskala

Edjekouane et al. [7] verwendeten eine 100-stufige Skala mit nur drei Attributen („nicht laut = 15“, „mittelmäßig laut = 50“, „sehr laut = 85“), die sich im Vorversuch als unvorteilhaft erwies, da die Probanden kein einheitliches Verständnis hinsichtlich des Attributes „nicht laut“ hatten. Darüber hinaus wurde auch die feine Abstufung der Skala nicht genutzt.

Nach jeder Bewertung startete automatisch die Wiedergabe des nächsten Stimulus. Die Wiedergabe erfolgte in diesem Hörversuch über offene dynamische Kopfhörer des Typs HD 650 der Firma Sennheiser mit einer Freifeldentzerrung (PEQ V von HEAD acoustics), wohingegen im Versuch von Edjekouane et al. [7] über Lautsprecher in einem reflexionsarmen Raum wiedergegeben wurde. Jede Pegelvariante wurde fünfmal, bei Edjekouane et al. sechsmal, abgefragt.

Im zweiten Versuchsteil bewerteten die Versuchsteilnehmer die Sprachsignale. Das Verfahren blieb dabei unverändert: Ein Stimulus wurde wiedergegeben, der Proband gab seine Bewertung entsprechend der dargestellten Skala ab und es startete die Wiedergabe des nächsten Signals. Jedes Sprachsignal wurde pro Versuchsperson dreimal bewertet, die Reihenfolge war vollständig randomisiert. Den Probanden war es (ebenfalls im Unterschied zur Durchführung von Edjekouane et al. [7]) freigestellt, die Wiedergabe beliebig oft zu wiederholen.

Zur Auswertung der Hörversuchsergebnisse wurde zunächst eine Lautheitsfunktion aus den Resultaten der Rauschsignalbewertung aufgestellt. Hierbei boten sich zwei Optionen an: die Erstellung individueller Lautheitsfunktionen für jede Versuchsperson oder einer gemittelten Lautheitsfunktion für alle Probanden. Abbildung 5 zeigt beide Varianten. Sowohl die individuellen Lautheitsfunktionen als auch die gemittelte Lautheitsunktion wurden über sigmoidale Funktionen mit vier Parametern angenähert.

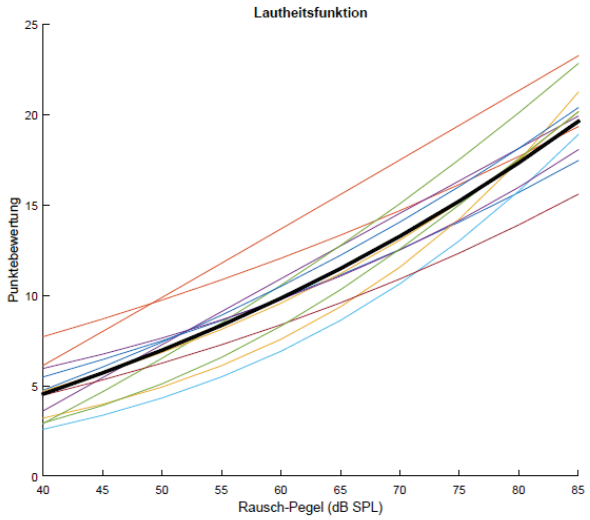

Abbildung 5: Lautheitsfunktionen (farbig: individuelle Lautheitsfunktionen für 13 Probanden, schwarz/fett: gemittelte Lautheitsfunktion aus den Daten aller Probanden)

In den Ergebnissen zeigten sich nur marginale Unterschiede (< 1 dB) zwischen den beiden Varianten. Daher sind in den folgenden Abbildungen 7 und 8 nur die Ergebnisse der Umrechnungen mit der gemittelten Lautheitsfunktion dargestellt. Die Umrechnung erfolgte, indem die für die Sprachsignale abgegebenen Punktzahlen in die Umkehrfunktion der sigmoidalen Lautheitsfunktion eingesetzt wurden. Abbildung 6 visualisiert den Zusammenhang.

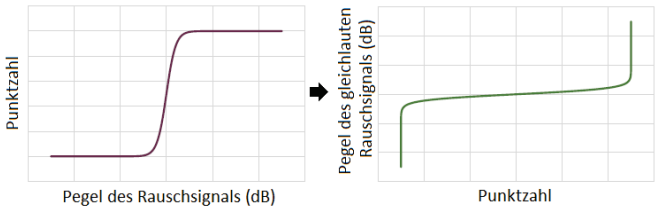

Abbildung 6: Schematischer Verlauf der Lautheitsfunktion und deren Umkehrfunktion

Es ist offensichtlich, dass bei der Bewertung der Rauschsignale in jedem Fall ein größerer Pegelbereich abgedeckt werden muss als bei der Bewertung der Sprachsignale, um mögliche Fehler in der Umrechnung zu vermeiden.

Ergebnisse

Probandenbewertungen

Abbildung 7 zeigt das Ergebnis des Hörversuchs. Dargestellt sind jeweils die Mittelwerte über alle Probanden und über alle Stimuli mit gleichem Pegel und gleicher Bandbreite. Zudem sind die jeweiligen Standardabweichungen eingezeichnet. Die betrachtete Größe ist derjenige Pegel des schmalbandigen Rauschsignals, das die gleiche Lautheit aufweist wie die jeweilige Stimulusvariante. So geht aus der Abbildung 7 beispielsweise hervor, dass das Referenzsignal einen Pegel von etwa 55,5 dB SPL benötigt, um die gleiche Lautheit zu erreichen wie die „Full Band“ gefilterten Sprachsignale bei 45 dB SPL.

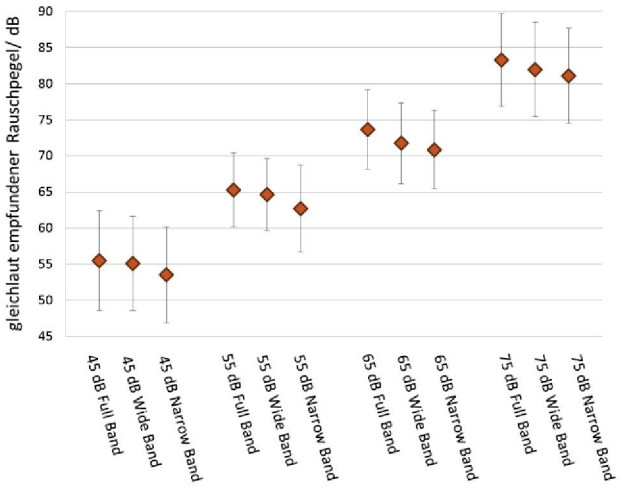

Abbildung 7: Pegel des Rauschens (3 Bark) bei gleicher Lautheit, Hörversuchsergebnisse: Mittelwerte und Standardabweichungen über 13 Probanden mit je 3 Bewertungen und 6 Stimuli

Es ist zu erkennen, dass die Pegeldifferenz zwischen gleichlautem Rausch- und Sprachsignal mit sinkender Bandbreite abnimmt, wodurch sich auf einen Abfall der Lautheit mit sinkender Bandbreite schließen lässt. Mit steigendem Pegel sinkt diese Pegeldifferenz leicht; dieser Effekt ist jedoch nicht statistisch signifikant. Die Standardabweichungen liegen in etwa konstant bei ± 5-6 dB.

Modellvergleiche

Um die Vorhersagegenauigkeit der Berechnungsmodelle zu überprüfen, wurde der beschriebene Hörversuch mit den ausgewählten Modellen nachgestellt. Hierzu wurden zunächst die Lautheiten der Sprachsignale und des Referenzsignals berechnet. Im Falle der „Long Term Loudness“ nach Moore und Glasberg wird der Mittelwert, bei den übrigen Modellen der Lautheitswert betrachtet, der in 5% der Zeit erreicht oder überschritten wird. Anschließend wurde das Referenzsignal im Pegel angepasst und erneut dessen Lautheit berechnet, bis sie der Lautheit des Sprachsignals entsprach.

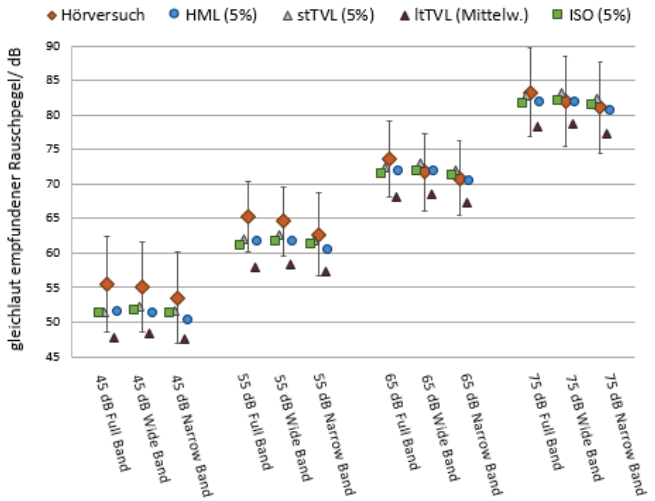

Abbildung 8: Pegel des Rauschens (3 Bark) bei gleicher Lautheit, Hörversuchsergebnisse (Mittelwerte und Standardabweichungen, siehe Abbildung 7) und Modellrechnungen

In Abbildung 8 sind die Berechnungsergebnisse zusammen mit den Ergebnissen aus dem Hörversuch dargestellt. Es fällt auf, dass drei der vier betrachteten Modelle recht ähnliche Vorhersagen liefern. Lediglich die Variante ltTVL fällt aus diesem Schema und unterschätzt die Lautheiten für alle Stimuli. Mit abnehmender Bandbreite sinkt in allen Modellen auch der Vorhersagefehler. Umgekehrt verhält es sich in Abhängigkeit von den Pegelstufen: bei höheren Pegeln treffen die Modellvorhersagen die empirischen Daten sichtbar genauer, wenngleich sie auch in den niedrigen Pegelstufen weiterhin innerhalb der Standardabweichungen liegen.

Fazit

Mit den aktuell bestehenden Berechnungsmethoden ergeben sich für hohe Pegelstufen (65 bzw. 75 dB SPL) bereits zutreffende Lautheitsvorhersagen. Für niedrige Pegelstufen (45 bzw. 55 dB SPL) hingegen wird die Lautheit von allen untersuchten Modellen leicht unterschätzt.

Die Resultate von Edjekouane et al. zeigen für alle Modelle noch größere Vorhersagefehler im Vergleich zu den Hörversuchsergebnissen (siehe [7]). Ein Grund hierfür könnte dabei im verwendeten Referenzsignal zu finden sein.

Im Versuch von Edjekouane et al. wurde ein deutlich schmalbandigeres Rauschen mit einer hohen Tonalität gewählt als das hier verwendete 3 Bark breite Rauschsignal. Scheinbar können die Berechnungsmodelle zurzeit die Lautheit dieser tonalen Komponenten nicht ausreichend berücksichtigen [6].

Offen bleibt überdies die Frage, ob der Klangcharakter der Sprecherstimme einen Einfluss auf die Lautheitseinschätzung ausübt, also ob ein laut eingesprochenes Sprachsignal beispielsweise eine höhere Lautheit suggeriert als sein späterer Wiedergabepegel eigentlich erwarten ließe. Ein solcher Effekt lässt sich mit den vorliegenden Daten nicht überprüfen, da lediglich in mediokrer Gesprächslautstärke eingesprochene Signale untersucht wurden. Ebenso sind für die Lautheitsbewertung extremer Variationen in der Stimmfarbe (wie Flüstern oder Schreien) weitere Hörversuche nötig, da sich hierbei nicht nur die jeweiligen Frequenzspektren messbar verändern, sondern vor allem davon auszugehen ist, dass die Wahrnehmung der Probanden stark durch die persönliche Hörerfahrung determiniert ist. Auch die Lautheitsempfindung von komprimierten Sprachsignalen bleibt insbesondere im Hinblick auf deren Anwendung in der Telekommunikation in einer möglichen weiteren Hörversuchsreihe näher zu untersuchen.

Literatur

[1] ISO 532-1: Acoustics – Methods for calculating loudness, Part 1: Zwicker method, in preparation, to be published in 2016.

[2] DIN 45631/A1:2010: Calculation of loudness level and loudness from the sound spectrum - Zwicker method - Amendment 1: Calculation of the loudness of time-variant sound, Beuth Verlag, 2010.

[3] Glasberg, B.; Moore, B.: A Model of Loudness Applicable to Time-Varying Sounds. J. Audio Eng. Soc. 50(5), pp. 331-341, 2002.

[4] Sottek, R.: Modelle zur Signalverarbeitung im menschlichen Gehör. Dissertation, 1993.

[5] Zwicker, E.: Procedure for calculating loudness of temporally variable sounds. J. Acoust. Soc. Am., vol. 62(3), pp. 675–682, 1977.

[6] Sottek, R.: Calculating tonality of IT product sounds using a psychoacoustically-based model, Proc. Internoise, San Francisco, 2015.

[7] Edjekouane, I.; Plapous, C.; Quinquis, C.; Meunier, S.: Loudness of Speech Transmitted via Handsfree Telephone Systems – Perceptual Measurements and Loudness Models in Free Field Listening. Acta Acustica united with Acustica, vol. 101(6), pp. 1130-1144, 2015.