Loudness Estimation Through Evoked Responses at Audiometric Frequencies

A Thesis presented by

Yonatan Sasson

to

The Department of Computer and Electrical Engineering

In partial fulfillment of the requirements for the degree of

Master of Science

in

Computer and Electrical Engineering

Northeastern University
Boston, Massachusetts

August 2011
# Table of Contents

ABSTRACT ................................................................................................................................. 4

1. Introduction .......................................................................................................................... 5

2. Theoretical Background ..................................................................................................... 6
   2.1 Models of Loudness ......................................................................................................... 6
   2.2 Subjective Loudness Estimation ...................................................................................... 8
   2.3 OAE – An Overview ........................................................................................................ 9
   2.4 ABR – An Overview ......................................................................................................... 10

3. Methods and Experiment ................................................................................................. 12
   3.1 Listeners .......................................................................................................................... 12
   3.2 Stimuli ................................................................................................................................ 12
   3.3 Hardware and Lab Setup ................................................................................................ 12
   3.4 Estimation via CMM ......................................................................................................... 13
   3.5 Estimation via ME ............................................................................................................ 15
   3.6 Estimation of Loudness via TBOAEs ............................................................................... 15
      3.6.1 Signal Averaging ........................................................................................................ 16
      3.6.2 Obtaining the Point Estimates ................................................................................ 16
      3.6.3 Polynomial Fitting .................................................................................................... 17
   3.7 Estimation of Loudness via TBABRs ............................................................................... 18
      3.7.1 Averaging ................................................................................................................ 19
      3.7.2 Segmentation ............................................................................................................ 20
      3.7.3 Point Estimates ......................................................................................................... 21
      3.7.4 Weighted Fitting ...................................................................................................... 21
3.7.5 Reducing Noise by Increasing the Number of Trials

4. Results

4.1 Psychoacoustic Procedures results

4.1.1 Cross-modality measurement correction factor

4.1.2 Comparison of results between CMM string length and computerized CMM

4.1.3 Comparison of results between ME and CMM (computerized version)

4.2 TBOAE results

4.2.1 General discussion

4.2.2 Results for 250 and 8000 Hz

4.2.3 Results for 500, 1000, and 2000 Hz

4.2.4 Results for 4000 and 6000 Hz

4.2.5 Coupler Measurements

4.3 TBABR results

4.3.1 General Discussion

4.3.2 Individual TBABR Results

4.3.3 Increasing the Number of Trials

4.3.4 Comparison with TBOAE Results

5. Conclusion

6. Acknowledgements

7. References
ABSTRACT

Procedures for evaluating individual loudness functions through tone burst evoked auditory responses - both Tone-Burst Otoacoustic Emissions (TBOAE) and Tone-Burst Auditory Brainstem Responses (TBABR) – have been investigated in the past and have yielded credible results for a limited stimulus frequency range across a wide range of levels. The aim of this work is to investigate this relationship further by expanding the frequency range examined to cover a more full audiometric range and expand the analysis parameters for an objective and thorough analysis process. Evoked-responses (TBOAE and TBABR) were recorded from six normal hearing listeners for 17 different levels and seven frequencies (250, 500, 1000, 2000, 4000, 6000 and 8000 Hz). The recordings were analyzed and the results compared to two psychoacoustic procedures: Cross Modality Matching (CMM) and Magnitude Estimation (ME). Results for TBOAEs show agreement with results from previous studies for 1-kHz and 4-kHz and support the assumption that recorded signals at 500 and 2000 Hz are strong enough for such analysis, while analysis of signals at 250 and 8000 Hz are feasible but suffer from weak OAE response compromising the robustness. Recordings at 4000 and 6000 Hz were too contaminated to produce consistent results due to ear canal resonance. Results for TBABRs also show agreement with results from previous studies for 1-kHz and 4-kHz, and also exhibit consistency throughout the selected frequency range.
1. Introduction

An objective and automated procedure for individual loudness estimation through the use of tone-burst otoacoustic emissions (TBOAE) and tone-burst auditory brainstem responses (TBABR) that yielded results consistent with common psychoacoustic procedures for the measurement of loudness at 1 kHz and 4 kHz was previously developed (Epstein & Silva 2009; Silva and Epstein 2010). The aim of the present work is to build further upon that procedure by expanding the scope of the procedure to allow for loudness estimation at frequencies across the audiometric range (specifically 200 Hz to 8 kHz) by applying new parameter sets and averaging/noise cancellation techniques. The automated, objective loudness estimation procedures themselves are compared to common, subjective psychoacoustic procedures for individual loudness estimation: Magnitude Estimation (ME) and Cross-Modality Matching (CMM).

Otoacoustic Emissions (OAE) and Auditory Brainstem Responses (ABR) are of particular interest because measurements of these responses do not require active patient participation. Thus, they could be considered perfect candidates for clinical use. Presently, they are already in use for neonatal hearing screening, differential diagnostics for hearing impairment, and identification of cochlear functionality. An automated and reliable method for loudness estimation would serve as a valuable tool mostly for hearing-aid fitting as the frequency-specific information could allow for careful, non-linear individual prescriptions taking into account not only the endpoints of the loudness functions (threshold and discomfort level), but also the shape across the entire dynamic range.

The organization of this thesis is as follows: Section 2 provides general theoretical background pertaining to this work and introduces the reader to its basic components such as loudness, models of loudness, psychoacoustic procedures, and evoked responses. Section 3 provides detailed explanations of the methods used to construct the experiments and analyze the results. This section also includes information on the hardware setup and listeners and is common to both the OAE and the ABR procedures (both were recorded simultaneously). Section 4 summarizes all of the results obtained from all of the experiments. Owing to the large quantity of
graphs and tables, the organization of section 4 is as follows: section 4.1 summarizes only the results pertaining to the psychoacoustic procedures (CMM, ME, Sting Length). Section 4.2 contains all the OAE results, using relevant results from section 4.1 for comparison purposes. Section 4.3 focuses on ABR results with reference to sections 4.1 and 4.2 for comparison.

2. Theoretical Background

Loudness is a subjective measure and is defined as the primary perceptual correlate of physical sound intensity. However, many parameters can affect loudness other than physical intensity. These include frequency, duration, and individual differences. The loudness of a sound can also be affected by the presence of other sounds, often called maskers. This section will provide some background outlining the early measurements of loudness and the development of models of loudness. Then, a short theoretical overview on otoacoustic emissions and auditory brainstem responses will be provided. Finally, a brief summary of the mechanics of loudness estimation using OAEs and ABRs, specifically TBOAEs and TBABRs, will be provided.

2.1 Models of Loudness

Direct experimental loudness measurements date back to at least 1933 (Fletcher and Munson, 1933; Geiger and Firestone 1933), in which the loudnesses of complex tones were measured monaurally and binaurally. The most commonly accepted first approximation for the modeling of loudness as a result of fits to empirical data was published by Stevens (1955), suggesting the loudness function can be modeled as a power function with an exponent of 0.6:

\[ L = kP^{0.6} \]

Where \( L \) is the loudness in sones and \( P \) is the sound pressure in pascals. Stevens recognized that this model is limited to higher sound pressure levels and did not suggest that this relationship represents levels closer to hearing threshold accurately. Zwicker and Scharf (1965) furthered this model by addressing lower levels. They made measurements of loudness for pure tones and noise and developed the following equation:
(2) \[ N' = N_g \left[ \left( \frac{2E_t}{E_0} \right)^k \cdot \left( \frac{E}{2E_t + 1} \right)^k - 1 \right] \]

where \( E_t \) is the excitation produced at threshold, \( E_0 \) is the excitation in power units, and \( N_g \) is an arbitrary constant that determines the loudness units; \( N' \) is the loudness. This model contained a few inherent problems that were not accounted for. To correct those, the model was later revised by Buus and Florentine (2001) who suggested this model:

(3) \[ N(SL) = k \left[ (1 + (snr_{th} \cdot 10^{10})^{SL/s_\infty})^{s_\infty/s_{-\infty}} - 1 \right] \]

where \( k \) is a scaling factor, \( snr_{th} \) is the signal-to-noise ratio at threshold, \( SL \) is the sensation level of the tone in dB and \( s_\infty \) and \( s_{-\infty} \) are the asymptotic local exponents of the loudness function at high and low level respectively.

Moore and Glasberg (2003) revised the same model to allow modeling for cochlear hearing loss:

(4a) \[ N = C \left[ (GE + A)^{0.2} - A^{0.2} \right] \]

(4b) \[ N = C \left( \frac{E}{1.115} \right)^{0.2} \]

(4c) \[ N = C \left( \frac{2E}{E + E_{TH}} \right)^{1.5} \cdot \left[ (GE + A)^{0.2} - A^{0.2} \right] \]

where \( E \) is the excitation in power units, \( E_{TH} \) is the excitation produced by a sinusoid at absolute threshold, \( C \) is a constant, \( A \) is frequency dependent constant and \( G \) is a term that represents the low-level gain of the cochlear amplifier at a specific frequency relative to 500 Hz (and will always be less than 1). An assumption is made that the product of \( G \) and \( E \) is constant. Equation (4a) is used for \( E_{TH} < E < 10^9 \), equation (4b) is used for \( E < 10^9 \), and equation (4c) is used for \( E < E_{TH} \).
The last two models still did not incorporate subtle changes at low and moderate levels, and a new model, the INEX, was proposed by Florentine and Epstein (2006), asserting that the loudness function is steeper at lower levels than at moderate levels. By introducing modifications to the classical power function, the INEX model function, for normal-hearing listeners can be expressed as the following polynomial:

\[
(5) \quad f(L) = 1.7058 \times 10^{-9}L^5 - 6.587 \times 10^{-7}L^4 + 9.7515 \times 10^{-5}L^3 - 6.6964 \times 10^{-3}L^2 + 0.2367L - 3.4831
\]

where \(f(L)\) is the loudness in sones and \(L\) is the level in dB SPL.

### 2.2 Subjective Loudness Estimation

In order to examine the validity of results arising from objective procedures – specifically from TBOAE and TBABR – results must be compared with the results of other established procedures for measuring loudness. It is common practice to measure individual loudness functions directly using subjective experiments that involve active listener participation. Stevens designed experiments comparing loudness to several continua of different modalities such as brightness, vibration, and force of handgrip (Stevens 1955, 1957, 1959). Numerous experiments were conducted, generally asserting that the most accurate results were yielded from magnitude estimation and production combined (Stevens and Greenbaum 1966; Hellman 1976). Another method that has been shown to be useful for measuring individual loudness functions is Cross Modality Matching (CMM), a procedure that compares judgments of change in sensation in substituted sensory modality with judgments of change in loudness (Hellman and Meiselman 1988, 1990, 1993; Serpanos et al. 1998). In this thesis, three methods are used: magnitude estimation (ME), cross-modality matching (CMM) with string length, and cross-modality matching using a line shown on a computer screen.
2.3 OAE – An Overview

Otoacoustic emissions (OAE) are sounds emitted involuntarily by the auditory system from within the cochlea which can be measured in the ear canal. These sounds are generated as a response to acoustic stimulation or spontaneously without stimulation. First reported measurements of OAEs date back to 1978 (Kemp 1978). The level of the emissions is generally very small, typically lower than the threshold of hearing, thus must be picked up by sensitive microphones (e.g., in the present experiments, responses are amplified by 40 dB prior to A/D conversion). There are many clinical applications for OAEs, as they can be measured quickly and noninvasively. Clinical uses of OAEs include hearing screening and differential diagnostics for hearing loss caused by administration of drugs, exposure to excessive sound, Meniere’s disease, tumors, and other disorders (Burkard and Don, 2007). Some researches argue that OAE activity is not influenced by age (when the influence of hearing level is considered), allowing the use of OAE for longitudinal experiments (Prieve and Falter 1995).

There are several types of OAEs. Otoacoustic emissions that are produced without any stimulus to evoke them are called spontaneous OAEs. Those emissions have narrow-band characteristics and would have a tonal quality if amplified. The sound pressure level is usually between -20 and -5 dB SPL, and rarely will exceed 0 dB SPL (Fastl and Zwicker 2007), making them rarely audible. The other types of otoacoustic emissions are produced as a response to stimulation, either to a continuous tonal or multi-tonal stimulation or as a response to short periodic sound impulses (transient evoked OAE); this is the type used in the present experiment. Evoked otoacoustic emissions include transient evoked otoacoustic emissions (TEOAE), stimulus frequency OAEs (SFOAEs), and distortion product OAEs (DPOAE). However, due to practical issues, TEOAEs and DPOAEs are more commonly used than SFOAEs (Burkard and Don, 2007). Distortion product OAEs are elicited by two pure tones (at frequencies f1 and f2), each presented by a different speaker to avoid system distortion. The largest DPOAE occur at the frequency $2f_1 - f_2$, although several other configurations are also used (e.g., $2f_2 - f_1$, $3f_1 - 2f_2$). The most commonly measured DPOAE feature is the level, which has been found to be largest for most human listeners when the ratio $\frac{f_2}{f_1}$ is about 1.22 (Burkard and Don, 2007).
In the present work, TEOAEs are used. A common way of eliciting the TEOAE is with a pure tone stimulus. Such responses are called Tone-Burst Otoacoustic Emissions (TBOAE). The stimulating tone is usually multiplied by a Gaussian window to reduce the spectral splatter caused by time windowing (Harris 1978, Fastl and Zwicker 2007). A typical presentation rate would be between 20 to 50 Hz and correspond to an analysis time of 50 to 20 ms, a suitable timeframe for allowing the responses to reach their maximum. Tone-burst OAEs are generally strongest around 1-2 kHz and decline for lower and higher frequencies, providing a typical range of 500 Hz – 5 kHz (Probst and Martin 1991; Robinette and Glattke 2000), presumably due mostly to the frequency response of the middle ear. Latency for responses of high frequencies is shorter, making the separation of stimulus energy and response energy difficult (Burkard and Don 2007). Recordings of OAEs between 3 and 5 kHz are often contaminated by ear canal resonance, which can overpower the evoked response (Ravazzani and Grandori 1993).

### 2.4 ABR – An Overview

Auditory Brainstem Response (ABR) is an electrical response generated by the auditory system. It was first reported by Jewett and Williston (1971) and has since been examined thoroughly and used for many applications. The most prevalent present use for ABRs is for newborn hearing screening. The recorded response has several distinct morphological characteristics (i.e. peaks), typically labeled with Roman numerals as waves I through wave VII, with waves I, III, V used most often clinically.

Several types of stimuli are used to elicit ABR responses. Broad spectra stimuli (e.g., clicks) are used by many researchers and for many clinical purposes (for a review, see Hall 2006). These are very useful for estimation of auditory functioning in the 1000-to-4000 Hz region and are suitable for hearing screening purposes (Hall 2006). However, information on auditory sensitivity across the audiometric range and especially the speech frequency region of 500 Hz to 4000 Hz is extremely important and calls for a more frequency-specific stimulus. There are two commonly used types of narrow-band stimuli: Frequency Following Responses (FFR) are responses elicited by frequency specific stimuli that are presented continuously. The FFR reflects sustained neural activity that is phase-locked to the cycles of the stimulus waveform (Burkard and Don 2007).
The second type is the tone-burst evoked response – evoked potentials elicited by a frequency-specific transient stimulus, with the response observed in a time frame following the offset of the stimulus.

Analysis of ABR is performed both in the time domain and frequency domain. For time-domain analysis of ABR, the morphological features of the waveforms are examined starting with the most prominent one: wave latency. Latency is defined as the time interval between stimulus onset and the appearance of a morphological feature (a peak or troughs). Additionally, sometimes the relative measure of inter-wave latency (time between peaks or troughs) is also measured. Latencies vary with stimulus intensity and frequency and generally increase as the intensity of the stimulus decreases. A second ABR feature that is commonly examined is the wave V amplitude. While other waves have also been examined, wave V is the most robust and easily identifiable in response to a wide range of stimulus levels. Many researchers have used both features for ABR analysis (Pratt and Sohmer 1977; Wilson and Stelmack 1982; Babkoff et al. 1984; Gallego 1999). One of the shortcomings of ABR time-domain analysis is the fact that analysis of ABR waves is done subjectively by clinicians. Thus, a “poor” morphology – one in which the peaks do not meet textbook expectations – may be difficult or impossible to analyze. This work, in part, aims to further a time-domain method of objective analysis of the ABR wave developed by Silva and Epstein (2010).

For frequency-domain analysis, the spectral content of the waveforms is investigated. An established result (Hall 1986; Abdala and Folsom 1994) asserts that the frequency content of the ABR resides in three prominent regions, a low-frequency region (below 200 Hz), a moderate-frequency region (900-1100 Hz) and a higher-frequency region with little spectral energy above 2000 Hz (Hall 2006). Due to that observation, it is common practice to filter the raw ABR recording through a bandpass filter with a lower cutoff frequency of about 30-50 Hz and a higher cutoff frequency of about 3000 Hz.
3. Methods and Experiment

3.1 Listeners
Six listeners with normal hearing, ranging in age from 20 to 31 years, participated in all parts of the experiment (ME, CMM, TBOAE, TBABR) and three out of the six participated in the CMM string length experiment. No listener had any history of hearing difficulties. Moreover, listeners’ middle ear were examined clinically and found to be normal. The right ear was arbitrarily chosen to be tested for all experiments.

3.2 Stimuli
Stimuli for all experiments consisted of tone bursts of 4 cycle pure tones, multiplied by a Gaussian window. Signals were padded with silence to create a 41.7 ms stimulus which was presented 5000 times back to back within a block of trials. This operation was repeated for each level and for each of the seven frequencies tested – ranging from 250 Hz to 8 kHz. For each frequency, 17 levels were examined, from 10 to 90 dB peak-equivalent sound pressure level (peSPL) in steps of 5 dB.

For the CMM, ME, and CMM string-length experiments, 12 concatenated basic units of 41.7 ms were used to create a tone burst train with an approximate length of 0.5 sec. This train was used in lieu of the single tone burst in order to compensate for any temporal-integration effects (Florentine, Buus et al. 1996; Zwicker and Fastl 2007) that may occur during the TBOAE/ABR experiments, due to rapid sequential presentations of tone-burst stimuli.

3.3 Hardware and Lab Setup
For all experiments, stimuli were generated in MATLAB (2010a) running on Windows XP and D/A converted using a 32-bit Lynx Two A soundcard with a 48-kHz sampling rate. The analog signal was then passed through TDT (Tucker Davies Technologies) HB7 buffer and presented monaurally using both channels of an Etymotic ER-10C transducer to a listener within a double-walled, sound-attenuating booth. Before each experiment, a system calibration was performed to ensure proper insertion of the ER-10C probe. Recordings of TBOAE and TBABR occurred
simultaneously for the full period of the stimulus (41.7 ms). TBOAE were recorded via the Etymotic ER-10C microphone and amplifier (+40 dB gain). TBABR were recorded using three electrodes: a non-inverting electrode placed at the forehead, an inverting electrode placed on the ipsilateral mastoid, and a ground electrode behind the contralateral mastoid (Hall 2006). The electrode signal was sent to a GRASS QP511 Quad AC Amplifier, where it underwent bandpass filtering from 30 to 3000 Hz and was amplified by a multiplicative factor of 50,000 (approximately +94 dB gain). The signals were then sent to the Lynx Soundcard for A/D conversion.

3.4 Estimation via CMM
Two similar CMM procedures were used to directly measure loudness for comparison with the TBOAE- and TBABR-derived estimates. The first procedure is the ‘string-length’ CMM procedure and the other, a ‘computerized’ CMM procedure. For the ‘string length’ procedure, the listener was given a thin string, presented with a randomized series of sound levels at a fixed frequency and asked to cut a string to be “as long as the sound is loud” for each stimulus presentation. After the string is cut, it is taped in a notebook, the page is turned, and the next stimulus is heard. This process is lengthy and requires the experimenter to perform post-measurements to determine string length. As an unestablished alternative, the present study introduces a similar, computerized procedure in which a clear white screen with a short and thin line in its center was presented to the listener from outside the sound-proof booth through a viewing window. A keypad interface allowed the listener to change length of the line using three different levels of sensitivity: coarse, medium and fine - corresponding to increments of 2.5%, 0.94%, and 0.124% of the screen’s width, respectively. A 17” LCD computer screen was used and was located 105 cm from the listener. The listener was asked to match the length of the line to the loudness of the stimulus. As with the string-length procedure, the listener does not receive any information about the actual length that is chosen and is unable to view prior lengths when making a new judgment. A “zero” key is also provided to allow indication that no sound was heard.
Figure 1: Computer interfaces for the psychoacoustical procedures. From left to right: ‘computerized’ CMM screen, in which the user matches line length on the screen, the ME screen, in which a listener enters a numerical estimate for loudness and the ‘string-length’ CMM, indicating the listener to cut the string and place it in the page in the book identified by the letter code. All three screens are preceded by a screen instructing the listener to expect the stimulus, after which the stimulus is heard.

For each of the seven frequencies, a block of trials was presented for which, listeners were presented with five repetitions for each of the 17 levels in random order. For each frequency, a training session was performed at the beginning of the experiment – one presentation for each level was presented and the listener performed the task to familiarize with the range of the stimuli and the general way that the task worked. If at least three of the five repetitions were heard and non-zero responses were given, the level was considered to be above threshold. The final loudness estimate for each of the levels for each task was calculated as the geometric mean of all non-zero estimates (geometric mean is used because we average the loudness estimates on a logarithmic scale). Some researchers have used a correction factor to adjust the CMM-derived loudness functions due to their tendency to produce shallower slopes than other loudness measures (Marks 1978; Epstein and Florentine, 2005). Such corrections factors are typically used to match the slope with that of a power function with an exponent of 0.3 that is widely used for such experiments. The estimates of the loudness slopes (the exponent
of the power function) were calculated by fitting an exponent of a power function to the data. The results from this experiment provided functions with loudness slopes close to the common Stevens power function, as will be discussed in section 4.1. Hence, no correction scaling factor was used.

3.5 Estimation via ME
A second psychoacoustic procedure was used in order to test the validity of the TBOAE/TBABR results. Listeners were presented with stimuli in blocks identical to the ones used for measuring CMM, but instead of matching loudness to a length, listeners were asked to assign a number that represented the loudness of the sound. Listeners were instructed to use any positive number and were encouraged to use decimal numbers to represent each train of tone bursts. An estimate of zero was given by the listener if and only if no sound was heard. As with the CMM procedures, levels were considered above threshold if at least three of the five repetitions were heard. For each frequency, a training session was performed at the beginning of the experiment – one presentation for each level was presented and the listener performed the task to familiarize with the range of the stimuli and the general way that the task worked. For each level and frequency, a final loudness estimate was calculated using the geometric mean of all non-zero estimates provided. Loudness slopes estimates were calculated in the same manner as they were calculated for CMM.

3.6 Estimation of Loudness via TBOAEs
Three stages were used in the process of estimating individual loudness function from the raw recordings. First, an averaging method was applied to the recorded data in order to achieve one estimated OAE signal for each level and frequency out of the 5000 recorded trials. Then, analysis methods were applied to calculate the loudness estimate for each level and frequency. Finally, a polynomial was fit to the group of level estimates for each frequency to generate a loudness-function estimate.
3.6.1 Signal Averaging
Four different averaging methods were tested, each using a different weighting algorithm for the trials: equal weighting, 5point-var, full-var, and correlation-averaging. Equal weighing of the trials (arithmetic averaging) was performed as a simplest, ‘worst-case’ average as a reference for the performance of the other weighting methods. In the 5point-var method, five equally spaced points in time were selected from the signal and variance for those points was calculated across all 5000 trials. Each trial was then weighted by the reciprocal of the sum of its variance at the selected points, normalized by the overall variance. Full-var is a method similar to 5point-var, but instead of selecting five equally spaced points in time, variance was calculated for every sample of the signal and the weight was the reciprocal of the sum of all variances. For the correlation averaging method, an autocorrelation function was calculated for each of the trials and compared to an average autocorrelation function, where the normalized reciprocal of the Euclidean distance from the average is the weight for the trial.

3.6.2 Obtaining the Point Estimates
After the signals were averaged and two independent averages were obtained, each an average of 2500 trials (first 2500 trials and latter 2500 trials), point estimates were calculated using a method proposed by Epstein and Florentine (2005) and refined by Epstein and Silva (2009). For each of the averaging methods described in section 3.7.1, three analysis parameters were varied: window delay from the stimulus onset: from 1 to 35 ms in 1-ms steps; window size: either 2.5, 10, 20, or 30 ms; and the frequency bandwidth of the analysis, referred to as F-ratio: either 1, 1.5, 2, or 3. A Hanning window was used throughout the whole analysis as suggested by the results of Epstein and Silva (2009). The F-ratio was defined by Epstein and Silva (2009) as the ratio by which the lower and upper bounds of the frequency bandwidth of the analysis region are related to the center frequency. An F-ratio of 2 indicates that the lower bound of analysis in one-half of the center frequency and the upper bound is twice the center frequency. As an example, an F-ratio of 2 for a frequency of 1 kHz indicates an analysis region from 500 Hz to 2000 Hz.
Once a specific parameter set was chosen (F-ratio, window size, and window delay), the loudness function was estimated using three steps. First, windowing was performed on the two individual averaged waveforms. In the second step, the FFT for each independent average was calculated with the analysis band determined by the F-ratio. Then, the logarithm of the sum of the real and positive components of the cross spectrum between the two FFT results is obtained as the point estimate. After a point estimate is calculated for each of the levels for a specific frequency, the final set of point-estimates is subtracted by an offset to yield a zero-mean set for comparison with other methods.

### 3.6.3 Polynomial Fitting

The weighted-polynomial fitting was then used to create a final loudness function by considering a ‘confidence measure’. For each level, the variance of the 5000 trials was calculated and was used as an estimate of the amount of noise in the recordings. Low variance indicated a consistent set of recordings and a high degree of confidence for that level. Contrary to that, a high variance indicated inconsistencies in the recording and likely, a less reliable point estimate.

\[
W = \begin{bmatrix}
\frac{1}{\sigma^2_{L_1}} & 0 & 0 & 0 \\
0 & \frac{1}{\sigma^2_{L_2}} & 0 & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & 0 & \frac{1}{\sigma^2_{L_n}}
\end{bmatrix}
\]

\[
H_{L \times P} = \begin{bmatrix}
1 & l_1 & \cdots & l_1^{p-1} \\
1 & l_2 & \cdots & l_2^{p-1} \\
\vdots & \vdots & \ddots & \vdots \\
1 & l_L & \cdots & l_L^{p-1}
\end{bmatrix}
\]
Using the weight matrix, $P_{\text{poly}}$ coefficients for the polynomial are calculated (Scharf 1991):

$$P_{\text{poly}} = (H'WH)^{-1}H'W_{\text{est}}$$

Where $W_{\text{est}}$ is our point estimates – data points to be fitted, $W$ as given in equation (6), and $H$ is a $L \times P$ matrix representing the independent variable (given in equation 7), in our case the stimulus level - $L$ being the number of levels and $P$ being the number of polynomial coefficients.

### 3.7 Estimation of Loudness via TBABRs

When analyzing ABR waveforms, several features must be considered beforehand. As mentioned in section 2.5, the ABR waveform undergoes systematic changes as a function of level: as the stimulus level increases, the latency of the waves in the signal decreases and the amplitude increases. These two features are important to consider when analyzing ABR waveforms through the expected morphology and indeed early works that tried to estimate loudness functions via ABR used either waves latency (Pratt and Sohmer 1977; Babkoff et al. 1984; Davidson et al. 1990) or amplitudes (Gallego et al. 1999). The present work, however, uses a different approach adopted from Silva and Epstein (2010). They presented two main analysis methods, ranging in effectiveness. In the present work, only the most effective method is used. Additionally, some modifications to the methods are suggested. Along with time-domain averaging of the waveforms, a few frequency-domain methods were tested, taking advantage of known ABR features in the frequency domain. The ABR analysis consisted of four stages. The first stage, similar to TBOAE analysis is **signal averaging**. For each level, the 5000 recorded trials were averaged and two independent averaged waveforms were obtained (each average from 2500 trials). The second stage, **segmentation**, selected specific regions of the waveform for which the analysis was to be applied. The third stage was the **point estimate**. A calculation of a single number representing the loudness for that specific level based on the regions selected in the former stage. The final stage was similar to the last stage of the TBOAE
analysis, **weighted polynomial fitting**, in which the set of loudness estimates were fitted using variance to obtain a final loudness-function estimate.

### 3.7.1 Averaging

Similar to the treatment described for TBOAE, the 5000 recorded trials were averaged in several ways in order to achieve two independent average representations of the ABR signal made up of 2500 trials each. Some of the methods used time-domain analysis while others used frequency-domain analysis.

#### 3.7.1.1 Time-Domain ABR Averaging

The four TBOAE averaging methods were used also for TBABR: Equal averaging, 5point-var, full-var, and correlation averaging. See section 3.6.1 for a detailed description.

#### 3.7.1.2 Frequency-Domain ABR Averaging

There are several frequency-domain properties of the ABR waveforms that can be utilized in averaging. Many researchers investigated the frequency content of ABR waves and reached similar conclusions (Abdala and Folsom 1994; Hall 1986). There are three main ranges in which most of the frequency content resides: A lower range of up to 200 Hz, a middle range of 900- to 1100 Hz and a higher region of frequencies above 2000 Hz. Prior to averaging the signals, trials were pre-filtered using different configurations and frequency ranges. The filtered trials were then averaged (arithmetic averaging with no weighting).

First, the lower-frequency region was considered: **frequency-50**, **frequency-100**, **frequency-200** and **frequency-250**, averaging only the spectral content contained in the bands of 1-100 Hz, 1-200 Hz, and 1-250 Hz respectively. A fifth averaging method – **frequency low-middle** – averaged both the 1-200 Hz band and the 900-1100 Hz band. For all five methods, signals were truncated and averaged only from the time after the stimulus onset. A sixth averaging method utilized the portion of the recording after the response was expected to have diminished well below the noise floor to characterize the recorded noise, assuming the spectral characteristics of
the noise are constant for the specific recording. The frequency components of the constructed ‘noise signal’ were subtracted from the actual recording to try and get a ‘noise-free’ recording. This averaging method was tried after frequency 200 (a band filter of 1-200 Hz) was applied.

3.7.2 Segmentation

Segmentation methods were adopted from Silva and Epstein (2010), with a few modifications. The first set of segmentation methods used were fullsync, abrsync, and amlrsync. In these procedures, the waveforms are segmented based on degrees of similarity between the waveform at a chosen level and a waveform at a level one step higher. As a first step, the two waveforms were time-aligned by calculating a time-lag and shifting the samples of one of the waveforms. The time-lag was selected such that it yielded maximum cross-correlation between the two waveforms, and was constrained to be within a 2 ms range; otherwise it was set to zero. After the waveforms were aligned, each waveform was divided to K sections of length w. A cross-correlation was then calculated for each section to create a “staircase function” f(k), where k represents the section’s index. The last step of the process was applying a pre-determined threshold on f(k): segments of f(k) above the threshold – indicating high degree of similarity between the two signals – were used for determining point estimates. Segments below threshold were discarded.

The difference between fullsync, abrsync, and amlrsync is the time ranges on which they operate: fullsync operates from 0.5 ms after offset throughout the whole recorded signal, abrsync operates from 0.5 ms after offset through 21 ms, and amlrsync operates between 20 and 41 ms after stimulus offset.

Based on these three methods, an additional approach was tried: instead of using two waveforms of different levels for the comparison, two independent averages of the same level were used. The segmentation itself was performed in the same manner, including the three different waveform ranges.
3.7.3 Point Estimates
After the segmentation of the waveform was complete, a point estimate was calculated as the logarithm (as the loudness is presented on a logarithmic scale) of the power of the segmented waveform. This process is repeated for each of the levels, creating a set of point estimates of loudness, later combined to form an estimated loudness function.

![Figure 2: Constructing the ‘step function’ of cross-correlations, for the wave segmentation](image)

3.7.4 Weighted Fitting
A loudness function was fit to the set of point estimates using a weighted fitting. The algorithm was similar to the fitting algorithm described for TBOAE in section 3.6.3. For each level, a “variance” measure was calculated during the averaging process and used as a measure of confidence for that estimate. The polynomial was calculated using the same equations 6 and 7.

3.7.5 Reducing Noise by Increasing the Number of Trials
Alongside the averaging methods described above, another method was tried. In order to test the hypothesis that signal-to-noise ratio increases as the number of averaged trials increases, one listener was used to record 50,000 trials at all levels for one frequency. The assumption was that ten times the number of trials, equally averaged, should result in a signal with an improved signal-to-noise ratio, which after analysis will yield improved MSE in comparison with any of the other methods. Although such method is not practical for several listeners and many frequencies due to the amount of time each listener would have to invest, it may, if successful, provide an approximate lower bound on the expected MSE.
4. Results

4.1 Psychoacoustic Procedures results

In this section, the results from the subjective psychoacoustical procedures are presented. First, a discussion regarding usage of the correction factor for the CMM procedure will be presented. Because a new interface for CMM was used, the average loudness slope was calculated and compared with results from other CMM experiments. Second, a comparison between the two proposed CMM procedures is presented to evaluate whether the new computerized procedure is comparable to the string-length procedure. Finally, a comparison between CMM and ME is provided, to determine both the degree to which the two procedures agree and the variability between the pair in order to perform comparisons with loudness estimates from TBOAEs and TBABRs.

4.1.1 Cross-modality measurement correction factor

Cross-modality matching and magnitude estimation have been used for many years, and thus an expected form for the loudness function as a result of CMM is well established in literature (Baired et al. 1980; Hellman and Meiselman 1988; Collins and Gescheider 1989; Serpanos et al. 1998; Epstein and Florentine 2005). Many researchers have used different correction factors in order to achieve the expected loudness slope of the estimated loudness function derived from CMM (Marks 1978; Collins and Gescheider 1989; Epstein and Florentine, 2005; Epstein and Silva 2009). However, as discussed in section 3.4, a new CMM procedure was used here, and therefore the need for a correction factor has been re-examined. As can be seen in table 1, the fitted power-function exponents (loudness slopes) from CMM and ME are close to those commonly found in the literature (power function exponent of 0.3), indicating that the procedure as applied, does not require a correction factor to obtain a loudness function with a loudness slope similar to that found using other methodologies. Further discussion of the results is presented in the following section (4.2.1) – a comparison between results of the string-length procedure and the computerized line-length procedure.
4.1.2 Comparison of results between CMM string length and computerized CMM

The two CMM procedures were nearly identical, allowing a direct comparison between results from the string-length and the new, computerized approach. The biggest and most obvious potential pitfall of the computerized version of the CMM procedure comes from edge effects at high levels when the user’s length domain is limited by the size of the screen and at low levels when the line length has a minimum quantized size and a step size that is close to total length, allowing relatively limited fine-tuning of the line length. The comparison between the string-length procedure and the computerized line-length procedure tested the assertion that a fixed correction factor applied to the result of the string length CMM will correct its loudness to a desired value.

Three listeners (picked from the listeners that participated in the complete experiment) were run using the string-length CMM using two different frequencies, 1 kHz and 4 kHz, and the results were compared and presented at Table 1. The third column presents MSE between the two methods. Columns four and five present the loudness slopes, and columns six and seven present the hearing thresholds for both methods.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Listener</th>
<th>MSE</th>
<th>slope comp.</th>
<th>slope string</th>
<th>TH comp.</th>
<th>TH string</th>
</tr>
</thead>
<tbody>
<tr>
<td>1k</td>
<td>L1</td>
<td>0.10</td>
<td>0.32</td>
<td>0.31</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>0.07</td>
<td>0.32</td>
<td>0.30</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>0.11</td>
<td>0.27</td>
<td>0.29</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>4k</td>
<td>L1</td>
<td>0.14</td>
<td>0.36</td>
<td>0.33</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>0.12</td>
<td>0.33</td>
<td>0.27</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>0.10</td>
<td>0.35</td>
<td>0.30</td>
<td>30</td>
<td>35</td>
</tr>
</tbody>
</table>

Table 1: comparison between string-length CMM and computerized CMM. Loudness slopes are consistent across all listeners for both 1 kHz and 4 kHz. Low MSE between the two methods confirms the hypothesis that the computerized interface can be used instead of the string-length interface.

4.1.3 Comparison of results between ME and CMM (computerized version)

Results of the two psychoacoustic subjective procedures are presented in Tables 2 through 4. Table 2 presents average results across all listeners for each of the seven test frequencies. The second column displays the average discrepancy between the two procedures. The third and fourth columns show the average thresholds for all listeners for the specific frequency. Columns
five and six show the calculated fitted-power-function loudness slopes for CMM and ME respectively. Columns seven and eight show the average standard deviation of the data from the slopes of columns five and six. Table 3 shows the hearing thresholds for all the listeners across all levels for both CMM and ME. The thresholds found by the two procedures agreed fairly well with 88% of the pairs (37 pairs) showing less than a 5-dB (one step in the experiment level scale) difference between CMM and ME, and only 5 pairs (11%) differing by 10 dB. When thresholds are examined as a function of frequency, there is a noticeable tendency for lower thresholds at lower frequencies, which does not match the common equal-loudness contour (Suzuki and Takeshima 2004). This may be explained by the temporal integration (Florentine et al. 1996) that occurs at lower levels, as the stimulus is considerably longer than at higher frequencies. For example, the stimulus at 500 Hz lasts 8 ms, whereas for 6 kHz lasts only 0.66 ms. The next parameter examined is the loudness slope – the exponent of a power function fit to the loudness data. Table 4 sums the individual loudness slope data across all listeners, for all frequencies. Mean values of these data are presented in Table 2 (columns 6 and 7). As can be seen in Table 4, the loudness slopes are all around the commonly cited value of 0.3 (see Serpanos et al. 1998 for review) and the mean values across all listeners and frequencies are 0.33 and 0.30 for CMM and ME respectively. Columns seven and eight of Table 2 show the standard deviation from these slopes, where the calculated standard deviation values of 0.23 and 0.17 indicate a fairly accurate slope estimates.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>MSE (CMM vs. ME)</th>
<th>CMM TH</th>
<th>ME TH</th>
<th>CMM slope</th>
<th>ME slope</th>
<th>CMM stdev</th>
<th>ME stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>0.190</td>
<td>23.333</td>
<td>28.333</td>
<td>0.363</td>
<td>0.353</td>
<td>0.273</td>
<td>0.233</td>
</tr>
<tr>
<td>500</td>
<td>0.183</td>
<td>20.000</td>
<td>25.000</td>
<td>0.330</td>
<td>0.267</td>
<td>0.307</td>
<td>0.193</td>
</tr>
<tr>
<td>1000</td>
<td>0.153</td>
<td>16.667</td>
<td>21.667</td>
<td>0.317</td>
<td>0.273</td>
<td>0.297</td>
<td>0.193</td>
</tr>
<tr>
<td>2000</td>
<td>0.123</td>
<td>28.333</td>
<td>31.667</td>
<td>0.357</td>
<td>0.340</td>
<td>0.232</td>
<td>0.157</td>
</tr>
<tr>
<td>4000</td>
<td>0.197</td>
<td>25.000</td>
<td>36.667</td>
<td>0.333</td>
<td>0.330</td>
<td>0.227</td>
<td>0.198</td>
</tr>
<tr>
<td>6000</td>
<td>0.157</td>
<td>31.667</td>
<td>28.333</td>
<td>0.340</td>
<td>0.280</td>
<td>0.140</td>
<td>0.131</td>
</tr>
<tr>
<td>8000</td>
<td>0.117</td>
<td>30.000</td>
<td>31.667</td>
<td>0.327</td>
<td>0.297</td>
<td>0.153</td>
<td>0.102</td>
</tr>
<tr>
<td>mean</td>
<td>0.160</td>
<td>25.000</td>
<td>29.048</td>
<td>0.338</td>
<td>0.306</td>
<td>0.233</td>
<td>0.173</td>
</tr>
</tbody>
</table>

Table 2: Results for CMM and ME procedures, averaged across all listeners. Columns 2 shows the averaged MSE between CMM and ME, columns 3 and 4 show averaged thresholds for each frequency, columns 5 and 6 show the calculated slopes and columns 7 and 8 show the standard error from the calculated slopes.
### Table 3: Loudness slopes for all listeners, across all frequencies.

<table>
<thead>
<tr>
<th>Frequency [Hz]</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CMM slope</td>
<td>ME slope</td>
<td>CMM slope</td>
<td>ME slope</td>
<td>CMM slope</td>
<td>ME slope</td>
</tr>
<tr>
<td>250</td>
<td>0.33</td>
<td>0.27</td>
<td>0.39</td>
<td>0.33</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>500</td>
<td>0.35</td>
<td>0.31</td>
<td>0.35</td>
<td>0.31</td>
<td>0.34</td>
<td>0.29</td>
</tr>
<tr>
<td>1000</td>
<td>0.32</td>
<td>0.31</td>
<td>0.33</td>
<td>0.31</td>
<td>0.31</td>
<td>0.3</td>
</tr>
<tr>
<td>2000</td>
<td>0.33</td>
<td>0.27</td>
<td>0.37</td>
<td>0.36</td>
<td>0.35</td>
<td>0.31</td>
</tr>
<tr>
<td>4000</td>
<td>0.27</td>
<td>0.24</td>
<td>0.35</td>
<td>0.3</td>
<td>0.31</td>
<td>0.29</td>
</tr>
<tr>
<td>6000</td>
<td>0.29</td>
<td>0.23</td>
<td>0.36</td>
<td>0.29</td>
<td>0.35</td>
<td>0.32</td>
</tr>
<tr>
<td>8000</td>
<td>0.31</td>
<td>0.24</td>
<td>0.37</td>
<td>0.31</td>
<td>0.32</td>
<td>0.31</td>
</tr>
</tbody>
</table>

### Table 4: Complete CMM and ME thresholds for all listeners, across all frequencies.

<table>
<thead>
<tr>
<th>Listener</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency [Hz]</td>
<td>CMM</td>
<td>ME</td>
<td>CMM</td>
<td>ME</td>
<td>CMM</td>
<td>ME</td>
</tr>
<tr>
<td>250</td>
<td>25</td>
<td>25</td>
<td>20</td>
<td>25</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>500</td>
<td>20</td>
<td>15</td>
<td>15</td>
<td>25</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>1000</td>
<td>20</td>
<td>25</td>
<td>15</td>
<td>25</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>4000</td>
<td>25</td>
<td>20</td>
<td>30</td>
<td>35</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>8000</td>
<td>15</td>
<td>30</td>
<td>40</td>
<td>35</td>
<td>25</td>
<td>35</td>
</tr>
</tbody>
</table>

### 4.2 TBOAE results

#### 4.2.1 General discussion

Due to the nature of this work, an enormous database of results was generated. It is not possible, therefore, to display a full, comprehensive data set. The following sections will present an abridged set of graphs and tables that capture the essence of both the complete analyses and also the optimal sets of parameters found. In this general discussion, several results are displayed. As mentioned in the background, TBOAEs are known to have strong responses in the range of 500 to 5000 Hz, with the strongest responses measured between 1000 and 2000 Hz (Probst and
Moreover, ear resonance will contaminate most responses between 3000 to 7000 Hz (Ravazzani and Grandori 1993). At lower frequencies than 500 Hz and higher than 5000 Hz, a very weak signal will be recorded as occurrences of otoacoustic emissions drop below 50% (Robinette and Glattke 2000). The left panel of Figure 3 shows the averaged results for the optimal parameters that were calculated, compared both to CMM and ME. The x-axis displays the frequency tested while the y-axis shows the averaged MSE (AMSE) between the loudness estimates to CMM and ME respectively, across all listeners (with the optimal analysis configuration – which will be discussed in the next sections). It can be seen that the frequencies 500, 1000 and 2000 Hz indeed yield best results compared with both CMM and ME. The second ‘tier’ includes 250 and 8000 Hz, while 4000 and 6000 Hz are much worse. The right panel of Figure 3 summarizes the actual calculated MSE for each listener. Results are presented as a 3D surface. The x-axis shows the frequency, the y-axis shows the listeners, and the surface shows the trend of the TBOAE MSE. Each peak represents large MSE, as can be seen for all listeners at frequencies 4000 and 6000 Hz (and also 250 Hz). A trough represents a more credible estimate, and is most obvious for 1000 Hz, but also for 500 and 2000 Hz.

Figure 3: On the left, averaged MSE between TBOAE and psychoacoustical procedures (CMM and ME). Results are averaged across all listeners. On the right a surface showing MSE information for each listener and each frequency
4.2.2 Results for 250 and 8000 Hz

As stated earlier, due to a high volume of data resulted from the many parameters of averaging and analysis, the following graphs will present only the results of the best sets of analyzed data. For 250 Hz, the averaging method selected was the full-variance with a window delay of 16 ms and analysis window size of 20 ms. The optimal parameter set for estimating loudness from TBOAEs at 250 Hz is a stimulus offset of 11 ms after the recording starts. Any analysis prior to that will fail and will present a linear response due to stimulus artifact, as can be seen in figure 4. The left graph shows an example of several raw, pre-averaged OAE recorded signals. The level of 80 peSPL was chosen to demonstrate the stimulus clearly. It can be seen that the offset is 11 ms after the recording starts. The middle graph is loudness estimation for that setting using a window delay of 5 ms, well within the range in which the stimulus exists. The expected linear response is clear visually. The graph to the right is loudness estimation using a window delay of 11 ms, a delay in which the stimulus dies out, but still a strong linear response can be seen.

Table 5 shows the chosen analysis parameters for all the frequencies together. Although robustness of analysis is fairly low, a reliable set of analysis parameters was found and proved to be rather consistent across all listeners. The chosen averaging method was full-variance with widowing time delay of 16 ms, window size of 20 ms and F-ratio of 2. The graphs of Figure 5 demonstrate the result of the chosen analysis parameters. The left panel shows the AMSE for the chosen averaging method while analysis parameters were allowed to change. It can be seen that the trend of the graph is consistent. However, only a window size of 20 ms was stable, a phenomenon that re-occurred for all the other listeners as well. Middle panel shows results for the chosen analysis parameters but with all four averaging methods.
<table>
<thead>
<tr>
<th>Frequency</th>
<th>Chosen Analysis Parameters</th>
<th>Analysis Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-ratio</td>
<td>Window delay (ms)</td>
</tr>
<tr>
<td>250 Hz</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500 Hz</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000 Hz</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000 Hz</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4000 Hz</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6000 Hz</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8000 Hz</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Optimal parameters chosen for TBOAE analysis, with MSE calculations for these configurations

effect of the linear region is quite noticeable making analysis possible starting from 11 ms. The last two graphs represent the loudness estimates using the optimal configuration for CMM and ME respectively.

Figure 5: results for 250 Hz. From left to right: averaged MSE as a function of analysis window delay (presented for four different window sizes), comparison between outcomes of the four different averaging methods, and examples of individual loudness functions for three listeners
Figure 6 shows the results for 8000 Hz: left and middle panels show the AMSE for a single listener, analyzed using correlation averaging and a comparison for all four averaging methods while keeping the analysis parameters fixed. The right panel shows individual loudness estimates for three different listeners, compared with CMM (upper graph) and ME (lower graph). The individual estimates show that all of the loudness function’s characteristics are lost below 60 peSPL. As mentioned before, for 8000 Hz, the recorded OAE signal is extremely weak and it is not surprising to see a response for higher levels only. Referring back to Table 2, the averaged threshold at 8000 Hz was 30 for CMM and 30.16 for ME, suggesting that this outcome resulted from the TBOAEs and not hearing thresholds.

![Figure 6: results for 8000 Hz. From left to right: averaged MSE as a function of analysis window delay (presented for four different window sizes), comparison between outcomes of the four different averaging methods, and examples of individual loudness functions for three listeners.](image)

4.2.3 Results for 500, 1000, and 2000 Hz

This section shows the results for the three most stable frequencies tested. As mentioned earlier, the strongest OAEs responses are recorded in these frequencies. Table 5 shows the optimal parameter set chosen for 1000 Hz. The parameters chosen agree with those chosen by Epstein and Silva (2009) for the same frequency and exhibit consistent results across all the sets that were tested.
Figure 7 shows the results graphically: the left figure, reconstructed from Epstein and Silva, shows four different curves, each one corresponding to a different window size used in the analysis. Window delays of 2.5 and 10 ms exhibited MSE values with high variance, possibly due to the fact that short analysis windows are more sensitive to background noise, and indeed as the window size reached 20 and 30 ms, the graph appears more stable. The middle panel compares the four different averaging methods that were tried, and shows similar trends and little differences across parameter choices.

![Graphs showing results for 1000 Hz](image)

Figure 7: results for 1000 Hz. From left to right: averaged MSE as a function of analysis window delay (presented for four different window sizes), comparison between outcomes of the four different averaging methods, and examples of individual loudness functions for three listeners.

Ultimately, the 5point-var method was selected due to providing the smallest averaged MSE across all the listeners. The right-side figure shows individual loudness estimates for three listeners, compared with CMM (upper figures) and ME (lower figures). The estimates (hollow circles) exemplify the expected characteristics of the loudness function. Averaged MSEs are higher than for 1000 Hz, but are still very low relatively to the other frequencies. Throughout the whole analysis, two parameters were chosen repeatedly as optimal for analysis: the size of the analysis window (20 ms) and the F-ratio (F-ratio of 2).
Figures 8 and 9 show graphic results for 500 and 2000 Hz respectively. For both frequencies, several common characteristics can be observed. First, similar to previous analyzed frequencies, window delay of 20 ms shows the most promise, while other window durations exhibit higher degrees of variance. Averaging methods (middle plots) are again, consistent and no method can be chosen visually, although when averaging across listeners, 5point-var showed lower MSEs for 500 Hz while correlation averaging showed lower MSEs for 2000 Hz. Looking at the right side plots of both figures, it can be visually seen that the estimations do not follow the CMM and ME estimates as closely as for 1000 Hz.

Figure 8: results for 500 Hz. From left to right: averaged MSE as a function of analysis window delay (presented for four different window sizes), comparison between outcomes of the four different averaging methods, and examples of individual loudness functions for three listeners.
Figure 9: results for 2000 Hz. From left to right: averaged MSE as a function of analysis window delay (presented for four different window sizes), comparison between outcomes of the four different averaging methods, and examples of individual loudness functions for three listeners.

4.2.4 Results for 4000 and 6000 Hz

As mentioned earlier, the contamination of the TBOAE affects the recordings of 4000 Hz in a severe manner. Figure 10 shows a typical analysis result. The plot on the left shows averaged MSE for four different analysis windows, as a function on window delay. It can be seen that AMSE values are markedly higher than those obtained at other frequencies. The plot on the right shows an individual loudness estimate at 4000 Hz. The linear response overpowering any TBOAEs is visually obvious.

Figure 10: Typical results for 4000 Hz. The plot on the left shows high averaged MSEs for four different window sizes. The plot on the right exemplifies the recorded linear response that overpowered the TBOAEs.
Results are consistent with Epstein and Silva (2009) who reported high MSEs for 4000 Hz analysis. There is reason to believe that recording at 6000 Hz will suffer from the same problem. The results are presented in figure 11. Left plot shows the averaged MSE for four different analysis methods, all as a function of window delay, and the plot on the right is an individual estimate, demonstrating the substantial linear effect.

4.2.5 Coupler Measurements

In order to ensure that the procedures used to estimate loudness from TBOAEs did not result in spurious functions, the procedures were performed on recordings made in an acoustic coupler. Therefore, the coupler measurements presented a form of a ‘sanity check’ for the TBOAE recording and analysis process. The coupler emulates a human ear canal and measurements made in the coupler should show no TBOAEs. If the results in a coupler yield similar or close results to the ones from a normal hearing human, the analysis used is likely strongly correlated to the stimulus rather than a TBOAE. Figure 12 shows the MSE between the analyzed coupler responses and an average human CMM: an analysis configuration that yielded optimal results for TBOAE was used to analyze the coupler recordings at each frequency, and was compared with the average CMM function (averaged across all listeners for the specific frequency). The MSE
was calculated between the two functions and four graphs are presented for four different frequencies chosen arbitrarily, 250, 1000, 4000, and 8000 Hz. For 250 and 8000 Hz, it is visually obvious that no configuration yields a small MSE value. The MSE is also highly variable. Results for 1- and 4 kHz look more consistent among different window choices. However, MSE values for these frequencies are much greater than those observed for TBOAE analysis, which is consistent with results obtained by Epstein and Silva (2009).

Figure 12: MSE between loudness estimated from a coupler and from average CMM. The plots show MSE as a function of analysis window delay and four different window sizes. Four different frequencies are presented. From the upper left corner, clockwise: 250 Hz, 1 kHz, 4 kHz, and 8 kHz.
4.3 TBABR results

4.3.1 General Discussion
This section previews the TBABRs results sections and provides general results and justifications for choosing particular parametric sets of analysis. An important observation regarding TBABRs is that, unlike TBOAEs that are recorded in the outer ear, the TBABRs recording via to the electrode setting – is not susceptible to contamination by ear-canal resonances or signal artifacts. As a result, it is possible that TBABRs would be a more robust measure than TBOAEs across the full range of audiometric frequencies. Figure 13 shows general results for TBABR analysis. The plot on the left shows optimal averaged MSE across all listeners, for each frequency. The surface on the right shows individual MSE data for each listener at each of the frequencies. To obtain these data, an optimal analysis set was chosen such that it yielded the lowest averaged MSE (as plotted on the left plot). This configuration was then applied to individual subject data, as the goal of the present work is to determine the best configuration for individuals within a group of listeners.

For TBOAEs, loudness estimation at 500, 1000, and 2000 Hz yielded the lowest averaged MSEs, whereas for the TBABRs, loudness estimation at 8000 Hz also yielded a low MSE. As expected, estimation at 4000 and 6000 Hz are more consistent with results of the other frequencies, and although they seem to produce higher error, it is lower than the error produced by estimation using TBOAEs (0.162 and 0.156 compared with 0.502 and 0.424 for 4000 and 6000 Hz respectively).
Figure 13: Averaged MSE across all listeners as a function of frequency (left), and individual MSE for each listener at each frequency.

The complete results are summarized in table 6:

<table>
<thead>
<tr>
<th>Frequency Listener</th>
<th>250</th>
<th>0.5</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
<th>6000</th>
<th>8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.220</td>
<td>0.123</td>
<td>0.106</td>
<td>0.125</td>
<td>0.152</td>
<td>0.173</td>
<td>0.099</td>
</tr>
<tr>
<td>L2</td>
<td>0.189</td>
<td>0.145</td>
<td>0.203</td>
<td>0.113</td>
<td>0.114</td>
<td>0.208</td>
<td>0.139</td>
</tr>
<tr>
<td>L3</td>
<td>0.239</td>
<td>0.185</td>
<td>0.129</td>
<td>0.341</td>
<td>0.168</td>
<td>0.131</td>
<td>0.151</td>
</tr>
<tr>
<td>L4</td>
<td>0.162</td>
<td>0.149</td>
<td>0.225</td>
<td>0.140</td>
<td>0.150</td>
<td>0.109</td>
<td>0.120</td>
</tr>
<tr>
<td>L5</td>
<td>0.423</td>
<td>0.485</td>
<td>0.290</td>
<td>0.133</td>
<td>0.164</td>
<td>0.363</td>
<td>0.363</td>
</tr>
<tr>
<td>L6</td>
<td>0.506</td>
<td>0.150</td>
<td>0.127</td>
<td>0.289</td>
<td>0.220</td>
<td>0.099</td>
<td>0.111</td>
</tr>
</tbody>
</table>

Table 6: Individual MSE for each listener at each frequency

As discussed in section 3.7.2, four parameters were allowed to vary: range of analysis region (defined as fullsync, abrsync, and amlrsync), the use of two independent averaged signals versus only one (defined as split average and whole average, respectively), sectional window size, and analysis threshold. Figure 14 shows average results across listeners comparing fullsync, abrsync, and amlrsync.
Visual observation reveals that the lowest MSE values were obtained using fullsync for all 50 parameter set configurations. The calculated AMSE values were 0.376, 0.425, and 0.419 for fullsync, abrsync and amlrsync respectively and are consistent with results from Silva and Epstein (2010). Due to a high volume of data yielded from the many parameters used in averaging and analysis, subsequent results in the following sections will not be presented for abrsync and amlrsync.
4.3.2 Individual TBABR Results

This section will present the analysis results for loudness estimation from TBABRs recorded at each test frequency. As the goal of this work is to find the best results (in terms of comparison to psychoacoustical procedures) among all listeners, rather than for specific listeners, optimal analysis parameter sets were chosen as the ones yielding lowest MSE on average for all listeners, compared with both CMM and ME. Loudness estimates for individual MSEs were compared with CMM and also for ME, and were averaged independently. The parameter set that yielded lowest MSE overall was chosen as optimal for the particular frequency. Figure 15 shows the average results across all listeners and procedures for each frequency. The x-axis shows the nine averaging methods used, each using 30 different analysis options, resulting in total of 270 different analysis combinations. (Although not displayed, full analysis was performed using abrsync and amlrsync as well, resulting in a total of 810 combinations.) The y-axis shows the average MSE. Figure 15 shows better overall results for time-domain averaging (averaging methods 1-4) compared with averaging methods that underwent pre-filtering (methods 5-9). An exception can be detected at 2000 and 4000 Hz, for which frequency-domain pre-filtering yields lower MSEs. Another observation can be derived from the figure regarding the consistency of the results. For 250 Hz, the overall MSEs are significantly higher than all the other frequencies. This could result from very long duration of the stimulus, 16 ms, that may have affected the analysis of the TBABR signal.
Figure 15: MSE values for all averaging methods and all analysis parameters – averaged across listeners, presented for each frequency separately
Table 7 summarizes the lowest MSE values calculated. Each row shows results for a specific frequency, including all averaging methods, but specific analysis parameters only for one chosen averaging method that yielded the lowest MSE. The first column shows the frequency, columns 2-5 show the analysis parameters that produced lowest MSE, and columns 6-14 show the lowest MSE for each averaging method individually. The highlighted rubric shows, for each row, the optimal parameter set across all the listeners. The table indicates that all optimal averaging methods selected were time-domain averaging. While full-var seemed to yield best results – four of the seven frequencies yielded lowest MSEs when using full-var averaging – and analysis section size of 40 ms was used for five of the seven frequencies, no specific threshold or full-versus-split averaging showed dominance. It is worthy to note that for such a vast database, several approaches for picking the optimal set exist, and the results in this section represent only the results of one optimality criteria set. Figure 16 shows individual loudness estimates from TBABRs at each test frequency made using the optimal parameter sets for three listeners compared with CMM and ME separately.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Analysis Parameters</th>
<th>Averaging Methods</th>
<th>freq-200</th>
<th>freq-50</th>
<th>freq-100</th>
<th>freq-250</th>
<th>freq mid-low</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>fullsync</td>
<td>Full</td>
<td>0.002</td>
<td>0.4</td>
<td>0.324</td>
<td>0.235</td>
<td>0.243</td>
</tr>
<tr>
<td>500</td>
<td>fullsync</td>
<td>Half</td>
<td>0.001</td>
<td>0.6</td>
<td>0.230</td>
<td>0.151</td>
<td>0.148</td>
</tr>
<tr>
<td>1000</td>
<td>fullsync</td>
<td>Half</td>
<td>0.004</td>
<td>0</td>
<td>0.136</td>
<td>0.144</td>
<td>0.146</td>
</tr>
<tr>
<td>2000</td>
<td>fullsync</td>
<td>Full</td>
<td>0.004</td>
<td>0</td>
<td>0.239</td>
<td>0.169</td>
<td>0.155</td>
</tr>
<tr>
<td>4000</td>
<td>fullsync</td>
<td>Full</td>
<td>0.004</td>
<td>0.4</td>
<td>0.187</td>
<td>0.172</td>
<td>0.162</td>
</tr>
<tr>
<td>6000</td>
<td>fullsync</td>
<td>Full</td>
<td>0.004</td>
<td>0.4</td>
<td>0.228</td>
<td>0.163</td>
<td>0.156</td>
</tr>
<tr>
<td>8000</td>
<td>fullsync</td>
<td>Half</td>
<td>0.004</td>
<td>0.8</td>
<td>0.180</td>
<td>0.178</td>
<td>0.132</td>
</tr>
</tbody>
</table>

Table 7: lowest MSE values for each averaging method, highlighting the lowest MSE value across all averaging methods. All results are averages across all listeners.
Figure 16: Individual loudness estimates through TBABR. Each plot shows a specific frequency, with three estimates compared with CMM (upper plots) and three estimates compared with ME (lower graphs).
4.3.3 Increasing the Number of Trials

Signal theory indicates that the signal-to-noise ratio increases as the number of averaged trials increases. In order to test this expectation 45,000 TBABR trials were recorded for two listeners (L2 and L6), for all levels at a single frequency. Due to the length of this experiment, only two listeners performed this task at a selected frequency of 2000 Hz. Figure 17 shows two examples of signals recorded and averaged over 2500, 22500 and 450000 trials respectively. As can be seen on the left plot, little change occurs for greater number of trials. However, the plot on the right exhibits major differences between an averaged signal of 2500 trials and an averaged signal of 22500 or 45000 trials. For both listeners, not much visual difference is noticeable between 22500 and 45000 trials. However, further testing with more listeners is needed to assess the ideal number of trials needed.

![Figure 17: two listeners recorded 45,000 trials of the same frequency to test the hypothesis that increasing the number of trials will decrease SNR.](image)

Figure 18 shows the expected decrease in MSE values as a function of the number of trials, calculated using the optimal analysis parameters chosen for 2000 Hz. Both listeners’ MSEs are plotted on the same graph, showing similar trends of lower MSEs as the number of trials increase. However, comparing the calculated MSEs of the optimal configuration (the MSEs calculated were not lower than the optimal MSE calculated for 2000 Hz for only 2500 trials (see 4.3.2). It appears as if more listeners are needed.
4.3.4 Comparison with TBOAE Results

Finally, a comparison of all loudness-estimation methods is presented. Figure 18 shows both objective estimations (through TBOAEs and TBABRs) compared with both subjective procedures (ME and CMM). For ease of comparison, each of the frequencies is plotted separately for only one listener, using the optimal sets of analysis parameters that were achieved earlier.

A general observation is regarding the lower levels, for which an obvious inconsistency can be seen, even between the two objective procedures. Some of the frequencies, e.g., 2000 Hz, do not exhibit the expected behavior, as this frequency showed promise for both TBOAEs and TBABRs. It is important to remember that there is individual variability in the results, as the optimal parameter set was chosen such that the averaged MSE is lowest. This may result in sporadic individual graphs that do not provide the expected MSE.

Figure 18: MSE as a function of recorded trials. This test was performed for two listeners at 2000 Hz
Figure 16: Comparison between estimates of TBOAEs, TBABR, ME, and CMM.

Results are presented for one listener, for all recorded frequencies.
5. Conclusion

Two different experiments were performed with the common goal of improving methods for loudness estimation using auditory evoked responses. This work was based on previous work by Epstein and Silva (2009) – Loudness estimation through TBOAEs, and Silva and Epstein (2010) – Loudness estimation through TBABRs, with specific goals in mind: expanding the tested frequencies from 1000 and 4000 Hz to a more full set of audiometric frequencies, ranging from 250 to 8000 Hz, and finding analysis parameter sets that yield optimal results for each frequency independently.

The first experiment examined the estimation of individual loudness functions through tone-burst otoacoustic emissions (TBOAEs). Four averaging methods were tested, and for each of those, 560 combinations of analysis parameter set configurations were examined, in order to get a comprehensive set of analyses and find the best one. The results for TBOAEs could be categorized into three ‘tiers’ of success, conforming to the known characteristics of the OAEs: elicited OAEs are known to be strongest for stimuli between 500 and 5000 Hz. However, due to ear canal resonance, recordings are contaminated severely between 3000 to 7000 Hz.

Considering the frequencies tested (250, 500, 1000, 2000, 4000, 6000, and 8000 Hz), a strong response would be expected for 500, 1000, and 2000 Hz, a contaminated response for 4000 and 6000 Hz, and a weak signal for 250 and 8000 Hz. Results for 500, 1000, and 2000 proved to be the most accurate and the most robust (least sensitive to small changes in analysis parameters) and are considered the first tier of success. The second tier includes 250 and 8000 Hz. Although relatively weak signals were measured at these frequencies, relatively accurate estimates were made using these data; these estimates were less robust to parametric choices than for the first tier. The third tier includes 4000 and 6000 Hz, for which the linear response arising from the ear canal resonance dominated the response resulting in high values of MSEs and poor loudness estimates. These frequencies could not be used to successfully estimate loudness functions via the methods described in this work.

The second experiment examined the estimation of individual loudness functions from the auditory brainstem responses (ABRs). Nine averaging methods were tested, and for each of those, 150 combinations of analysis parameter set configurations were utilized in order to get a comprehensive set of analyses and find the best one. For TBABR analysis, both time-domain
properties and frequency-domain properties were taken into consideration. Several averaging methods included the application of pre-filtering to the raw TBABR recordings, trying to take advantage of the assumption that the frequency content of ABRs resides in two main regions: a low region of 1-200 Hz and a middle region, 900-1100 Hz, with little-to-no content for frequencies above 2000 Hz. The frequency averaging methods had very limited success, with only substantial improvements for the 2000 Hz test frequency.

Another property of the ABR waveform that was considered is the change that ABR waves undergo with changes in level. This was used while testing several methods of wave segmentation with several combinations of analysis region choices, and the use of two independent averages of the same level versus one independent average for two different levels. The ABR-based loudness estimation, in general, resulted in low MSE values. The ABR recording is not contaminated at specific frequencies by resonance like OAEs. Thus, a preliminary expectation was to yield “good” loudness estimates across all the frequencies. This expectation was met almost fully. All the frequencies except 250 Hz demonstrated low MSEs between the estimates and the psychoacoustical procedures that were used as reference, and moreover, yielded consistent results with the previous work that examined 1000 and 4000 Hz. However, for 250, higher MSE values were received, questioning the feasibility of estimation through TBABR for that frequency.

6. Acknowledgements

The author would like to express his deep gratitude and appreciation to Professor Michael Epstein for the help throughout this work, and for the support throughout its process. Also, a warm thanks to all the listeners that participated in the experiment: Andre, Bina, Chelsae, Melissa, and Kathryn, and also to Kristen and Anna who helped with the technical work in the lab.
7. References


