CHAPTER 1

Ear-Recorded Speech

1.1 Introduction

The initial experiment was a data collection experiment aimed to create a small corpus of speech data recorded under very specific conditions for use in the following two experiments. Due to the variability of even the same speaker repeating the same sentence, speech critically needed to be recorded from two locations simultaneously: from the mouth, and from the inside of the ear canal. This would allow for a more accurate comparison of the two signals.

1.2 Background

WHEN FINISHED WITH LIT REVIEW, METHODS AND ANALYSIS, FINISH INTRO AND ADD TRANSITION/OPENNING SENTENCE The speech vibrations of a person's own voice will propogate throughout the head and body (cf. fig. 1.1). Of interest for the present study, these waves will pass through the tissue in the head, and enter into the ear canal, where they will be recorded.

Bone conduction of acoustic vibrations through a human head has been well studied (cf. Allen and Fernandez (1960), Håkansson et al. (1994), Stenfelt et al. (2000), Reinfeldt et al. (2010), etc); however most of these studies have involved attaching a mechanical vibration device to an animal head or a cadaver skull, or using a vibrating piston on a live human participant, allowing for precise manipulation of the input signal. Most of these studies, as well, are focused on audiometric bone conduction, i.e. the propogation of waves through the head and their effect specifically on the cochlea itself, which is not relevant to the present study.

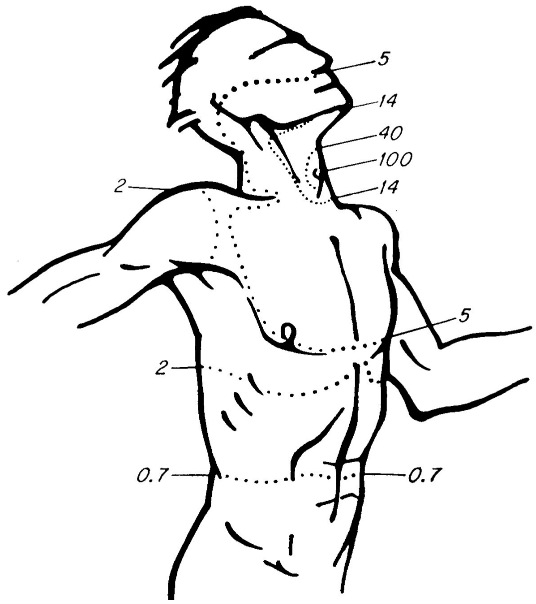


Figure 1.1: Diagram of the propogation of speech waves throughout the body. Numbers correspond to percentage of the original amplitude of the speech remaining when reaching the marked location. Taken from Békésy (1960).

Many of the early studies which were performed on cats do show that the sound generated by bone conduction propogating into the ear canal is dominated by low-frequency noise (Tonndorf (1972)). Normally, the open ear acts as a high pass filter, dampening these lower frequencies passing into it via bone condution. When occluded, this filter is non existant, and the lower frequencies are more noticeably present (discussed in depth further below).

Other more recent studies on human subjects have agreed with these findings. It has also been found that the acoustic response differs significantly depending on the location of the skull that is stimulated[1]. A few (cf. Békésy (1948), Hansen (1998), Pørschmann (2000), and Reinfeldt et al. (2010)) have investigated body conduction when the source of vibration (i.e. sound) is a person's own voice, not an artificial mechanical vibrator. These studies also record from the person's ear canal, and not another sensor on a different side of the skull.

Using speech as a source is inherently messy, because a) it is not as easily manipulated as a simple mechanical vibrator, b) it has far more frequency components than a simple mechanical vibrator, and c) it takes multiple pathways to get to the ear: from the vocal chords, through tissue, and into the ear canal, and also from vibrations in the air all along the vocal tract[2], through the solid medium of the

[1]Typically in these studies it is either the frontal bone or the mastoid process (Békésy (1960)).

[2]The speech sound is also filtered differently as it passes along the vocal tract

head[3], and back into the medium of air inside the ear canal. On top of this, the ear canal itself acts as a resonating chamber (Rosen and Howell (1991)), altering the signal beyond the distortion already caused by the passage through tissue and bone.

There has been much research on the resonating characteristics and amplitude response of the ear canal. One such project was performed by Stinson and Lawton (1989), which studied fifteen human ear canals. Their aim was to produce a model which can replicate the effect that the ear canal has on acoustics. One challenge in producing such a model is the considerable variability in the shape of the canal - both between subjects as well as between the right and left ear canal of a single subject (Stinson and Lawton (1989)). These differences are apparent in curvature, length, volume, and cross-sectional diameter throughout the ear canal. Stinson and Lawton (1989) created silicon ear molds for each of the ear canals, which were used to generate three different computational models: one following the contours and dimensions of their ear molds exactly, another following the dimensions of the ear mold, but straightening contours and curvatures as if along a central axis, and the third as if the ear canal were a uniform tube with the same length and volume of the ear canal molds (and the previous models). They noted that most significant differences between these models' spectral predictions of ear canal resonance occur above 6kHz (see fig. 1.2).

Since much of the acoustic information for distinguishing speech sounds is located below 6kHz, several (cf. Stinson and Lawton (1989), Hansen (1998) Stenfelt and Reinfeldt (2007)) who have made efforts to model the ear canal, have chosen to simply treat it as if it were a uniform tube.

Another challenge is to obtain the dimensions of the ear canal needed in order to treat it as a uniform tube in the first place. Immittance measurements are widely used in audiology, and involve emitting a chirp or tone into a pressurized ear canal. The chirp then bounces back from the tympanic membrane (assumed to have

[3] Although, of course, the head is composed of different tissues with different densities and acoustic resonances

infinite impedance in a pressurized canal) and can be recorded (Ballachanda (1997), 415): "The sound pressure developed inside a rigid cavity from a known sound source is directly related to the volume of the cavity". Therefore, the volume of the ear canal can be inferred for a subject using immittance testing without the need for invasive measurements (e.g. using a silicon mold). Making an assumption about either an 'average' diameter or an 'average' length of the ear canal[4] would allow for the approximate calculation of the other dimension, given the measured volume.

Once inside the ear canal with known approximate dimensions, it can either be modeled as an open-closed tube (if the ear is not plugged) or as a closed-closed tube (if the ear is plugged). This difference changes the resonance and reverberant structure of the ear canal. There have been many studies, a few in particular (c.f. Békésy (1948), Pørschmann (2000), Reinfeldt et al. (2010)) which use real human speech and measure the human ear as an open-closed tube.

Pørschmann (2000)'s study is generally looking at the perception of one's own voice, but in order to accomplish this devotes effort to looking at the bone conduction pathway separately. A general 900 Hz resonance (with subsequent harmonic resonances) was found in the collected bone-conduction speech, as well as a general broadband amplitude gain between 0.7 and 1.2 kHz. This correlates with the 800-1200 Hz range for the first resonance that others (cf. Håkansson et al. (1994)) have observed in mechanical-stimulated bone conduction studies. However, in this study only two phones were used (/s/ and /z/), and a masking threshold[5] technique was used to determine the frequency spectrum of the transfer function of body conduction. This is admittedly a rather subjective method of determining the spectrum.

Reinfeldt et al. (2010), on the other hand, use microphones to record the actual

[4] The average length of the ear canal has been cited from 23mm (Rosen and Howell (1991)) up to approximately 29mm (Stinson and Lawton (1989)) for a straight tube. The average diameter for the ear canal is approximately 7.1 mm (Salvinelli et al. (1991)).

[5] The masking threshold technique involves playing a pure tone at different frequencies and amplitudes while the participant is phonating. The participant indicates when the tone becomes audible over their own speech, which allows the spectrum of the sound transmission of their speech to be mapped.

sound pressure level (SPL) of both air and body conducted speech. Furthermore, Reinfeldt et al. (2010) used a more expansive and diverse set of phones. While a resonance was found in generally the same frequency region for /s/ (and other phones) as that found by Pørschmann (2000) (0.7 - 1.2 kHz), they discovered some interesting differences, which can be seen in fig. 1.3. Between each class that was used- voiceless sounds (/s/, /t/, /k/, and /tj/), nasals (/m/ and /n/), and vowels (/i/, /e/, /a/, /o/) - a moderately similar frequency response is seen, yet there are some interesting distinctions to note (see fig. 1.4.

In particular, as can be seen in fig. 1.3, there is much inter-speaker variation within the body conduction of the same sound. While it is difficult to track an individual speaker's relative spectral envelope within the figure, it appears that much of this difference, particular in the lower frequencies, originates from a difference in amplitude, and not necessarily from different resonance locations along the frequency axis. It is important to note that both Figs. 1.3 and 1.4 both contain *relative* spectral envelopes - i.e. the difference between the air conducted and body conducted components of speech, and do not contain an absolute frequency spectrum of body conducted speech.

An interesting observation is that the /e/ vowel has a relatively flat response up to 500 Hz, and only dips down to -5 dB around 1kHz. Contrast this with the phone /s/, which below 500 Hz has a fairly high (yet falling) response, and only dips below -5 dB near 4 kHz. However, compared with /e/, the body conducted to air conducted ratio for /s/ has a significant downward slope after 2000 Hz. This is likely due to the fact that there is relatively little energy produced by /s/ in the low frequencies, allowing for a high ratio, which drops as the general energy of the phone increases. This could indicate that most of the energy produced by a non-sonorant does not pass through to the ear canal.

More specific dichotomies can be found between sounds within the same class. For example, the low vowel /a/ is pronounced with a more open mouth vs the relatively closed mouth of the high vowel /i/; consequently, the dB SPL relative to the air conducted counterpart was much higher for /i/ than it was for /a/. An assump-

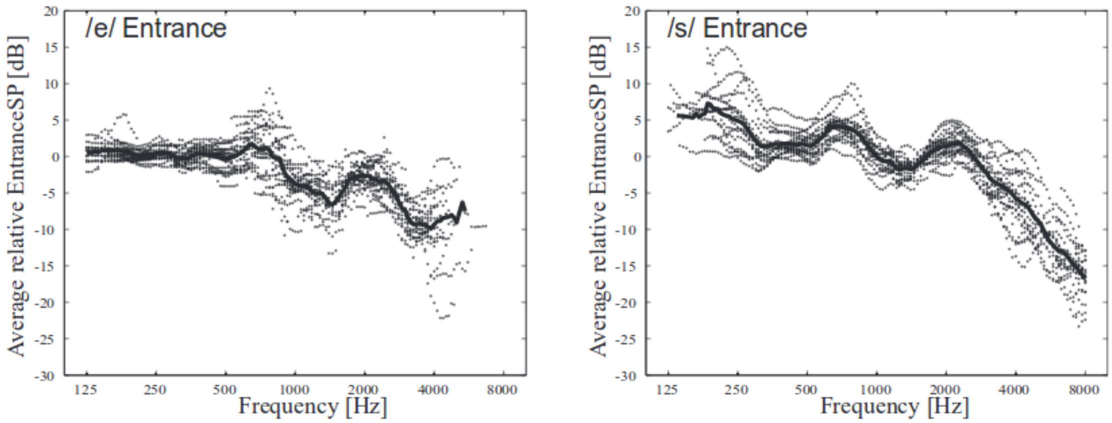


Figure 1.3: Body Conduction relative to Air Conduction for the phones /e/ and /s/. A value of less than zero indicates the amplitude of body conducted speech is less than that of air conducted speech, and a value greater than zero indicates a higher amplitude of body conducted speech. The solid line indicates the mean, and the remaining data points are from individual speakers. The signal was measured from the entrance of an open ear canal. Taken from Reinfeldt et al. (2010).

tion could be made from the data that the more open the mouth is, the more energy is transferred to the air conducted signal (cf. Fig. 1.4). This is supported by Békésy (1960), who also diagrammed the relative difference in amplitude in the ear canal between the air-conduction and body conduction of vowels (cf. Fig. 1.5), which also supports this hypothesis. There was much inter-speaker variability, but it appears that the bone-conducted phones with the least relative reduction in amplitude are the higher vowels. The more energy that is lost to air conduction during the production of low vowels, the less energy is transfered into the surrounding tissue.

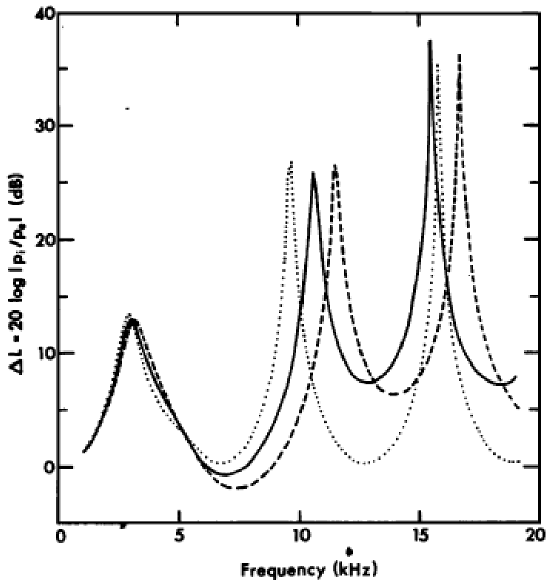


Figure 1.2: Stinson and Lawton (1989) diagrams three different models of the ear canal resonance. The bold line is based on their 3D canal molds from cadavers, the dashed line removes the curvature of the ear canal and acts as if the axis were straight, the dotted line assumes a constant diameter along a straight axis, with the same ear canal volume as the dashed and solid lines.

Additionally, Reinfeldt et al. (2010) found that the mid back vowel /o/ has a distinctive spectrum in relation to the other vowels, in which there is an amplitude peak near 2 kHz, rather than 1 kHz. Since the functions of other high back sonorants /u/ and /ŋ/ are not given, we cannot be certain if this is a phone-specific difference, or if it can be generalized to higher "back" articulations in which the articulatory constriction restricting the acoustic engergy is further back in the vocal tract.[6] It is also important to re-emphasize that these transforms are given as body conducted

[6]Békésy (1960) does not break down relative amplitude by frequency

amplitude *relative to* air conducted amplitude for the given phone, and do not reflect the absolute air- and body-conducted amplitude of phones compared with one another. For example, /a/ is a relatively loud air conducted sound due to its open articulation, and this loud air conducted component may cause its *relative* body conducted component to appear quieter than the other vowels, when in reality it is possible that the body conducted component of both vowels have the same absolute amplitude. Neither Békésy (1960) nor Reinfeldt et al. (2010) give information about body conducted components in relation to one another.

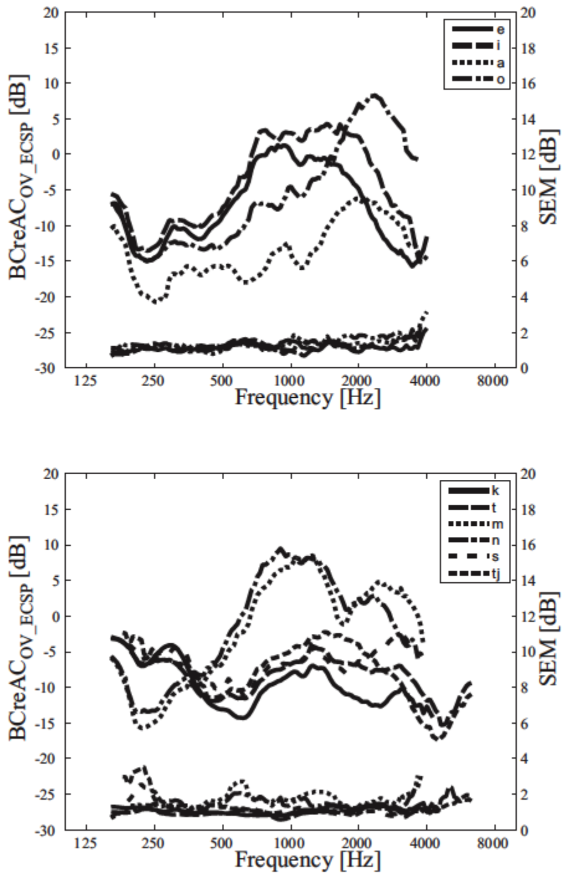


Figure 1.4: The mean relative amplitudes (left ordinate) of body conduction relative to air conduction for vowels (top plot) and other sounds (bottom plot). The set of lines along the bottom of each plot represent the standard error from the mean (SEM), measure on the right ordinate. Taken from Reinfeldt et al. (2010).

Modelling the ear as a closed-closed tube, however, results in a slightly different frequency response. This phenomenon, first noted by Wheatstone (1879), is termed the occlusion effect (OE). The occlusion effect[7] (OE) offers an amplitude

[7]The occlusion effect is the change in sound pressure level (SPL) resulting from body conducted vibrations emanating into, and reverberating within, a *closed* ear canal.

gain to certain frequencies and dampens others. This has been studied widely and extensively (cf. Wheatstone (1879), Kelly and Reger (1937), Littler et al. (1952), Goldstein and Hayes (1965), among many others). Generally, the occlusion effect (OE) results in a great increase in the amplitude of frequencies below 1-2kHz, acting as a low-frequency gain and that the amplitude of higher frequencies is dampened (as previously mentioned in the observation in studies of cats (Tonndorf (1972)).

As with bone conduction in general, most of the research of the occlusion effect (OE) has been conducted using controlled mechanical vibrations. Békésy (1960) reports that when the ear canal is closed, there is an increase in amplitude up to 2kHz, which afterwards vanishes quite suddenly (cf. fig. 1.6).

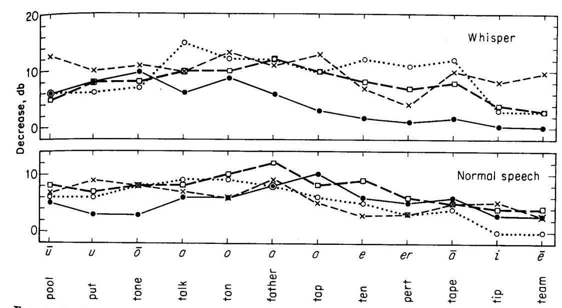


Figure 1.5: Demonstrated the different effect on amplitude that closing the ear canal has on the different vowels of English. Taken from Békésy (1960).

However, there is a variance in the OE - if using a mechanical vibrator - depending on the location of stimulation. This difference is most present in the lower frequencies (Dean and Martin (2000)), where the relative amplitude increase (of body-conducted sound versus air-conducted sound) appears to be the greatest, but tends to wash out when slightly higher frequencies are reached[8]. There are also differences based on the location of *occlusion* within the ear canal, i.e. how deep a plug is placed in the ear canal. Dean and Martin (2000)'s results indicate this difference does not disappear as frequency increases; the relative amplitude increase is greatest with supra-aural earmuffs at lower frequencies, and lowest with deep inserted earplugs[9]. However, at 1 kHz, the shallow-inserted earplug has a greater

[8]Dean and Martin (2000) found that the greatest relative amplitude increase occurs near 250 Hz, but the gain disappears when 1000 Hz is reached.

[9]Dean and Martin (2000) does not mention the explicit the depth for each condition, but from the article's description of the insersion procedure, it appears to be $\tilde{5}$mm.

relative amplitude gain than the supra-aural earmuff.

Stenfelt and Reinfeldt (2007) developed a model of an occluded ear using measurements generated from stimulating the skull separately at both the frontal bone and the mastoid process. Each site yielded a slightly different frequency response for the occlusion effect. Stimulation at the mastoid generally resulted in a greater increase in very low frequencies below 1 kHz. They also noted that the OE was greatest when using an ear plug near the opening of the ear canal, as opposed to supra-aural 'ear muffs' or a deep-insersion ear plug, though an OE was noticeable in each condition; this is in direct contrast with Dean and Martin (2000). Dean and Martin (2000) do not mention the size of earmuff used, but Stenfelt and Reinfeldt (2007) report the use of a large and small earmuff, with the latter providing a greater OE than the former, though both still below that of the shallow-insertion earplugs. With shallow insersion, their model estimates a gain in amplitude of frequencies below 2 kHz, and dampening of those above; all insersion depths, according to their model, will at minimum, slightly dampen frequencies above 2 kHz. As the plug is inserted deeper, the damping occurs on lower and lower frequencies. These results contrast slightly with Békésy (1960)'s in fig. 1.6 in that they predict higher, very low frequencies, as opposed with Békésy (1960)'s resonance around 1-2kHz.

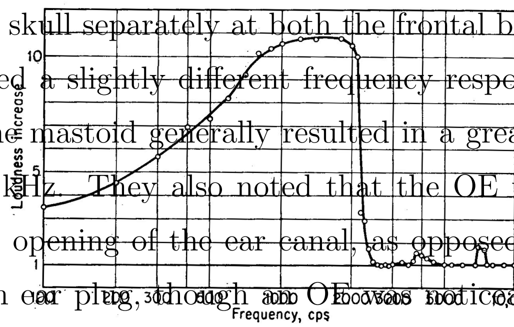


Figure 1.6: The frequency response inside the ear canal when taking a mechanical vibrator to a participant's forehead. Taken from Békésy (1960).

In contrast to the mechanical source studies above, Hansen (1998) tested the OE using one's own voice as the input source. Hansen (1998) presents a graph comparing three spectra calculated from continuous speech from three separate publications[10] (Fig. 1.7a). The study conducted its own tests (seen in Fig. 1.7b), which, by and large, agree with the previous studies. These represent the 'average' effect of occlusion

[10]From Wimmer (1986), Thorup (1996), and May and Dillon (1992)

on speech, and appears to resemble other (mechanical-source) estimations. Hansen (1998) developed a model of the OE which largely agrees with these measurements.

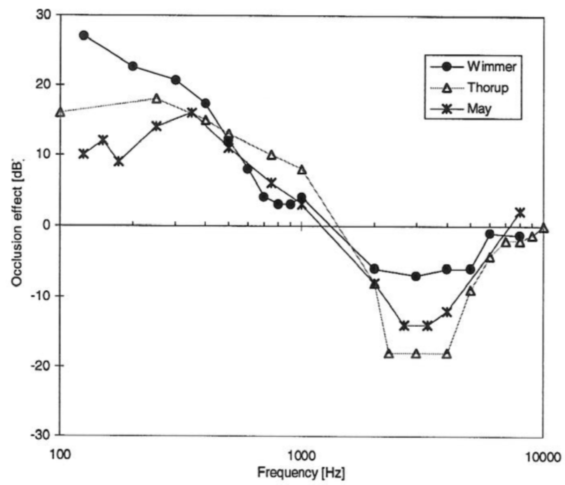
The studies looking at human speech result in similar spectral resonances as those dealing with simple mechanical vibrations, except real-speech studies are able to capture the different OE for different kinds of complex sounds in a real speech environment, such as vowels.

While Hansen (1998) found phone-specific differences in the occlusion effect (OE), it is ambiguous as to whether the differences are solely due to the differences in the transforms of phones as a result of body conduction (as seen in Reinfeldt et al. (2010)), or if there are sound-specific differences introduced within the ear canal or by the occlusion effect itself. Some have posited that variablility could critically stem from the placement of the jaw bone during speech next to the external auditory meatus, and as that changes as the jaw moves up and down[11] for 'higher' or 'lower' phones (e.g. /i/ vs /a/). Allen and Fernandez (1960) studied the OE on participants with a unilateral resection of the mandible (one side of the jaw has been removed), and found essentially no distinction between the OE in either ear (i.e. with a mandibular joint adjacent to the cartilage of the ear canal or without).

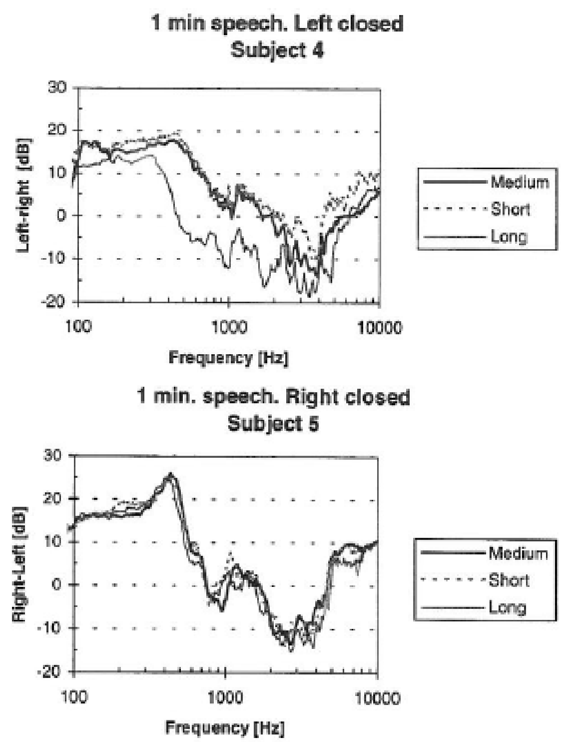
[11]Thereby changing the impedance characteristics of the vibration, (Békésy (1960))

Yet, Hansen (1998)) found that a change in shape of the ear canal due to different jaw positions can create an acoustic “leak” between the ear canal wall and the occlusion device; the occlusion effect, obviously, behaving differently for different levels of occlusion. Hansen (1998) diagrams cross sections of the ear canal with the jaw at different positions; between a closed jaw and 5mm of opening, there is relatively little difference between the shapes of the ear canal. Since Borghese et al. (1997) found that the jaw moves relatively little vertical distance during actual speech (max opening approx. 6 mm), it can be assumed that the ear canal changes shape negligibly during normal speech with a snug-fitting occlusion device.

In summary, it is important to emphasize the key difference between measurements from an occluded ear canal and those from an open ear canal, which can largely be seen between figs. 1.4 and 1.7a. There is a massive increase in the amplitude of the lower frequencies, which is not present from an open-closed ear canal, and a sizeable drop in amplitude after 2kHz, which similarly does not seem to manifest itself when the ear



(a)



(b)

Figure 1.7: In (a), three separate measured OE spectra. In (b), a comparison of the measured occlusion effect (OE) between two subjects and different sized ear molds that extend into the ear canal at different lengths. Taken from Hansen (1998).

canal is not occluded.

The aforementioned studies on body conduction and the occlusion effect, as would be expected, have indicated a fair deal of inter-person and inter-phoneme variability, and have shown the complexity involved in estimating the effect of body conduction and ear canal reverberance on speech entering the ear canal. However, the transfer function from the vocal tract to the ear canal does have some standard characteristics, namely, the body and (occluded) ear canal act as a low-pass filter on speech, removing many of the higher frequencies which are within range of containing critical components for speech intelligibility.

The methods described in this section, namely the models and transfer functions used by Hansen (1998), Stenfelt and Reinfeldt (2007), and Reinfeldt et al. (2010), predict a general rise in amplitude of lower frequencies (below 2 kHz), and a drop in frequencies above that range (with a few exceptions). This distortion is hypothesized to be predictable, unlike ambient noise from the environment, which is generally highly variable in both amplitude and form. Due to this, the technique of substituting unpredictable noise with the "predictable noise" of body conduction and the occlusion effect allows for greater confidence that a usable signal will be recovered. The "recovered" signal after undergoing minor transformations is hypothesized to perform better than a noisy signal collected from the mouth in both ASR and human speech perception tasks. Section 1.3 below describes the specific methods used to collect speech data from the mouth and the ear canal and recover an intelligible signal from the latter using the principles outlined in this section.

1.3 Experiment 1: Creating a dataset of ear-recorded speech

Due to the numerous constraints and requirement for the speech recordings required for this task it was necessary to create an original dataset for this study.

1.3.1 Design

The goal of this experiment was to create a dataset of recordings, both from the mouth, in noisy conditions, and from inside the ear in the same conditions. These recordings needed to demonstrate that (a) by recording speech from the ear, external noise was completely or largely eliminated, while simultaneously recorded speech from the mouth had a noisy background, (b) that the speech from the ear was more intelligible and recognizeable by humans than noisy speech, and (c) that the speech from the ear was more intelligible and recognizeable by an ASR system than noisy speech.

In alignment with the CHiME challenge[12] guidelines, this study uses different types of background noise at different noise levels. The noises used include the four sounds (bus, cafe, pedestrian area, & street) from the CHiMECHiME Challenge (2016) challenge, plus a 'factory' noise track. A short portion of the audio with relatively level amplitude was extracted from each sound file to be played in the background.[13]

Three noise levels were used. Since conversational speech is generally around 70 dB, the noise levels chosen were 60 dB, 70 dB, and 80 dB[14]. This would result in approximate SNR conditions of +10 (60 dB), 0 (70 dB), and -10 (80 dB). 80 dB was also chosen as a max loudness in order to leave a wide margin between it and any (albeit remote) possibility of hearing damage. A 'clean' condition was also utilized (no noise). Thus far this creates 16 different conditions (5 noise types * 3 noise levels + 1 'clean' condition).

---

[12] The CHiME challenge tasks researchers to improve upon or surpass the performance of a baseline automatic speech recognizer used on noisy speech data.

[13] **The exact portions of these sounds which were used are available online, along with the rest of the data at URL.COM**

[14] These were the 'averaged' dB levels over the course of the sound file

1.3.2 Stimuli

Thirty sentences were chosen from 3 Harvard Sentence lists[15]. Lists 14, 28, and 57 were used, and chosen semi randomly, eliminating lists with potentially unfamiliar or rare words. Each sentence occurs in all 16 conditions, resulting in 480 total stimuli.

1.3.3 Equipment

The experiment took place in a large soundbooth. To create the artificially noisy environment a Yamaha MS101 III loudspeaker was hooked to an HP ProBook 6470b laptop. A sound pressure level meter (SPL meter; Larson Davis Model 831) with a PCB Piezotronics Model 377B20 condenser microphone (omnidirectional) was placed 1 meter from the loudspeaker and measured the sound pressure to verify each of the three noise levels for each of the 5 noise types. A Grason-Stadler GSI Typstar Middle Ear Analyzer was used to measure the ear canal volume and test for plug leaks. Two Countryman B2D directional lavalier microphones with fixed XLR connections were used to record the mouth speech and the ear speech. These were hooked up to a PreSonus Digital Audio Firebox preamplifier, which was connected via TRS cables to a Zoom H6 Handy Recorder. A pair of 3M Professional Peltor Earmuffs with an NRR of -30 dB SPL were worn by the participant during the experiment.

1.3.4 Participants

Twenty participants were used in this study, ten female and ten male, all native speakers of English with normal hearing.

[15]The ‘Harvard Sentences’ is comprised of 72 lists, each 10 sentences long, where each list of 10 sentences is phonetically balanced, where the proportion of each phone in the list corresponds with it’s occurrence in the English language.IEEE (1969)

1.3.5 Procedure

The participant is initially asked a few preliminary demographic questions[16]. They are seated in front of the Middle Ear Analyzer. An otoscope is used to ensure the ear is mostly free of cerumen, to avoid blocking the microphone off from the rest of the canal or generally impacting the canal with cerumen. The chosen ear is fitted with an appropriate sized rubber clinical single-use ear tip, into which the Middle Ear Analyser hose is already plugged. An immittance test is performed, which involved playing a tone, and slightly and briefly pressurizing the ear. The Middle Ear Analyser checks that the ear plug solidly seals off the ear canal in order to be able to build up pressure, and alerts the researcher to a leak if the plug is not securely in place. This test results in an estimate of the volume in milliliters (mL) of the ear canal and of the middle ear, with precision to a tenth of a mL; additionally, a graph of middle ear function is given, which is checked for normalcy (cf. Appendix B**??**). Several other measures are given which are not used in this study.

The distance from the end of the ear plug to where it is enclosed by the ear canal is measured to determine how far the plug was placed in the ear canal (cf. Fig. **??** for diagram). Since the length of the plug is known, this was done by placing a measuring rod against the cavity of the concha to measure how far the plug was sticking out of the ear. The decision to treat the cavity of the concha as the “end” of the ear canal is taken from Stenfelt and Reinfeldt (2007), who made molds of ear canals, and treated the rapid increase in volume (where the cavity of the concha begins) as the end to the ear canal. This measure allows for the calculation of the depth of insersion of the earplug.

The Middle Ear Analyser hose is then taken out of the ear plug - which is carefully left in place to ensure a continuous seal. The participant then moves to a seat located in front of a computer monitor. The loudspeaker is on another table to the right of the participant, perpendicular to the direction the participant is facing (cf. Fig **??** for set-up diagram). The participant is then instructed as to the

[16] e.g. 2nd language (if any), etc. For a list of all information gathered, see Appendix A**??**.

proceedings of the rest of the experiment. One of the two microphones is taken, the wind-break foam removed, and is snugly inserted into the ear plug. A mark on the microphone cable was used to ensure the end of the microphone was fully inserted to the end of the earplug. Occasionally, the microphone was inserted deeper than, or just shy of, the end of the ear plug; the variance is within +/-1mm depth (cf. Appendix B**??**). The earmuffs are placed over both ears. Occasionally, a participant had glasses, or thick hair, which may have slightly compromised the seal. A note was taken of this.

A wooden rod was attached to the ear muffs, which extends forward, beside the participant's face. The second microphone was attached to this wooden rod via the lavalier clip at the level of the participant's mouth. The microphone was directed toward their mouth (cf. Fig **??**). The placement of the microphone on the wooden rod was adjusted to be exactly 10cm away from the participant's infra-nasal depression. At this point, the participant was asked to adjust the placement of their chair so that the microphone on the wooden rod was approximately 1 meter from the loudspeaker. Due to the length of the experiment (45min), no effort was made to discourage minor shifting in body position.

Both microphones were connected directly to the preamplifier through a fixed (non-changeable) XLR connection. Both channels were set to the same gain on the preamplifier. Two TRS cables took each microphone signal from the preamplifier to the recorder. Both channels were adjusted to appropriate (different) gain levels on the recorder itself to achieve a similar loudness for both signals and prevent clipping. These adjustments were made once the participant was situated, but before beginning the recording.

Once recording, an in-house computer program was used to display the stimuli sentences on a second monitor and play the background noises. For each sentence, the participant saw the clean-condition (no noise) first. The researcher was in the soundbooth with the participant listening through a pair of headphones connected to the preamplifier. The participant was asked to repeat the sentence twice to get a rhythm for it, at a normal, conversational loudness, with a normal, declarative intonation. The researcher asked the participant to repeat the sentence again in this condition if the rhythm or intonation of the two sentences did not match, or if the participant stumbled over a word. Each of the following 15 iterations of the sentence (one for each noise-type/noise-level combination) the participant was instructed to speak only once. If the rhythm of the sentence varied noticeably, or if a sentence was stumbled over, the researcher again asked the participant to repeat the sentence for that condition. The sentences were not randomized, i.e. all 16 iterations of a sentence occurred consecutively[17]. Within each sentence group (after the first, 'clean' condition, which always occurred first), all the noise conditions were randomized. Each stimulus is advanced by the researcher.

To help the participant notice when a sentence had been advanced[18], the number of the sentence condition was displayed underneath the stimulus (i.e. 1-16). This had the unindended consequence of occasionally producing a mild list-intonation.

After the recording was finished, the participant was asked to complete a short, 4 question survey[19] of their experiences during the experiment. They were instructed to give as basic or as detailed answers as they wished, but to answer truthfully. Extra credit in a Linguistics or other participating course was offered in exchange for participation in the experiment.

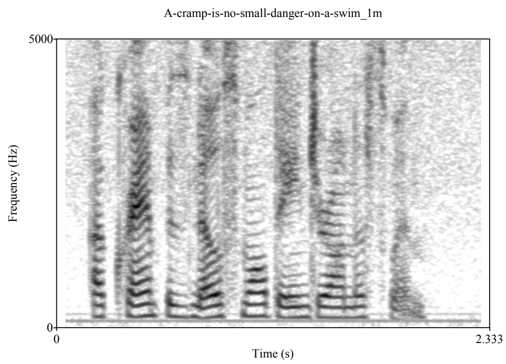
1.4 Experiment 1 Analysis

Each individual sentence was isolated in each recording with a Praat textgrid and extracted; this resulted in a sound file for each sentence, for each participant, for both the mouth-recorded and ear-recorded speech. Figures 1.8a, 1.8b, 1.9a, and 1.9b show the narrow and wide band spectrograms for ear- and mouth-recorded speech from participant 35, a female, for a "clean" example of the sentence "A cramp is no small danger on a swim". These two examples are fairly representative of the

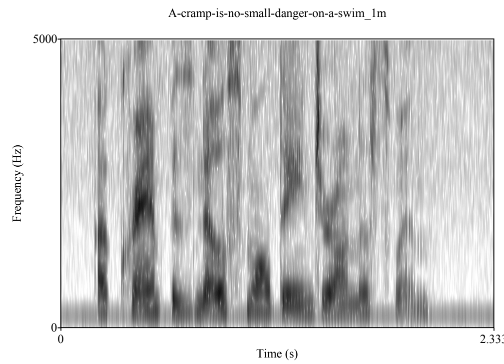
[17]This was done to enable to researcher to ensure a similar intonation and rhythm for each iteration of a given sentence.

[18]Wearing the ear muffs, they were often not aware when the noise condition changed.

[19]cf. Appendix **C??** for exact (non-coded) survey answers



(a) Narrow band spectrogram of speech recorded at the mouth.



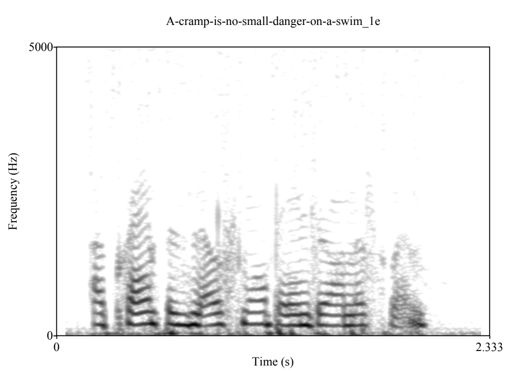
(b) Wide band spectrogram of speech recorded at the mouth.

Figure 1.8: Both (1.8a) and (1.8b) are the same sentence, "A cramp is no small danger on a swim", spoken by a female participant. This is the exact same sentence spoken at the exact same time as that in fig. 1.9.

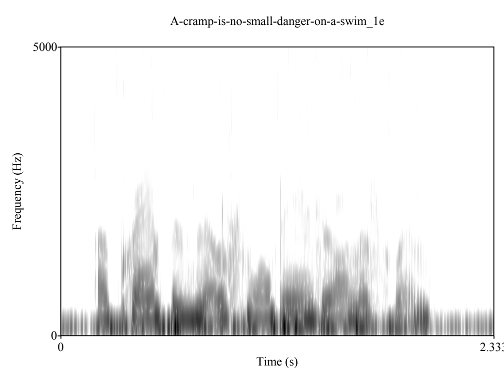
speech collected from each location.

As can be seen, the speech collected at the ear is heavily low-pass filtered, and the mouth speech by itself has much more speech information. However, there are still clear harmonics in the existing range in the ear-recorded speech, and most of the lower two formants can also be seen.

When noise is present, it can be seen in Figs. 1.10a and 1.10b that the noise

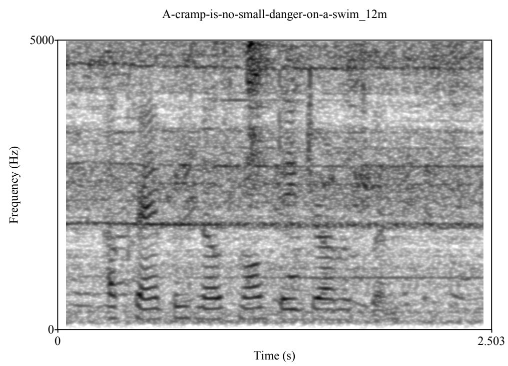


(a) Narrow band spectrogram of speech recorded from inside the ear canal.

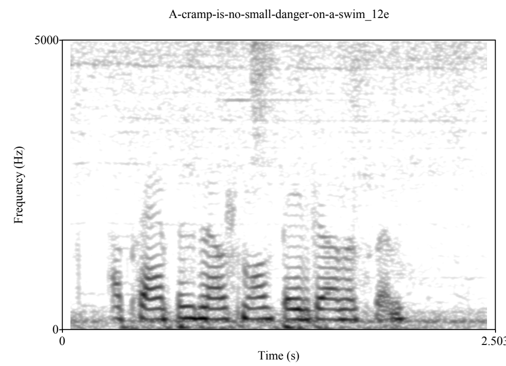


(b) Wide band spectrogram of speech recorded from inside the ear canal.

Figure 1.9: Both (1.9a) and (1.9b) are the same sentence, "A cramp is no small danger on a swim", spoken by a female participant. This is the exact same sentence spoken at the exact same time as that in fig. 1.8.



(a) Narrow band spectogram of speech recorded at the mouth, with 80dB bus noise playing in the background.



(b) Narrow band spectogram of speech recorded from inside the ear canal, with 80dB bus noise playing in the background.

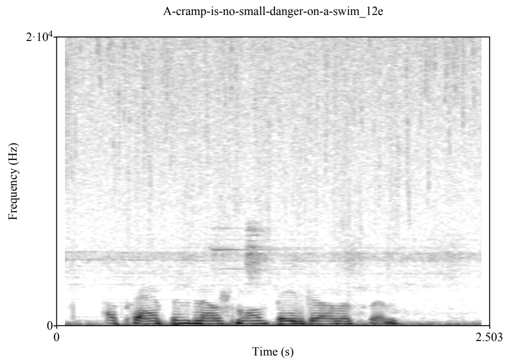
Figure 1.10: Both 1.10a and 1.10b were spoken by a female participant, recorded simultaneously, the sentence “A cramp is no small danger on a swim”.

does not affect the ear recorded signal nearly as much. There appears to be some of the louder noise (seen in the mouth-recorded speech in fig. 1.10a) present in the upper frequencies of the ear recorded speech, but it is significantly dampened and the signal has an overall higher speech to noise ratio (SNR).

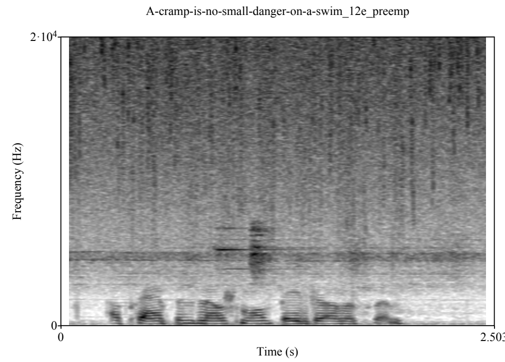
It should also be noted that the SNR in fig. 1.10a is much higher than originally intended. For this particular example, the speech was recorded with an 80dB noise background, with the intent of obtaining a -10dB SNR. Instead, the SNR is +6dB[20]. This is attributed to a) the participant speaking louder than anticipated, resulting in a higher speech theshold, and b) the directionality of the microphone used eliminated much more background noise than anticipated.

In an attempt to see if there are recoverable frequencies in the higher ranges, the spectrogram range of the non-premphasized ear signal was increased from 5kHz to 20kHz (see fig. 1.11a). There is certainly acoustic energy that makes it to the higher

[20] The SNR was calculated by using background noises recorded in isolation in the soundbooth. These were recorded at 60, 70, and 80 dB in the same soundbooth, with the same conditions and set up as a normal recording. The speech sound file was passed through a Hilbert Envelope, and a threshold was applied in order to extract just the speech data. The RMS values of both the speech and noise vectors were calculated, averaged, and then used in the SNR calculation. For explicit code, see Appendix E**??**



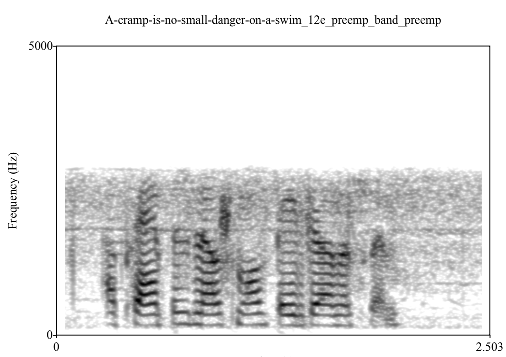
(a) Narrow band spectogram of ear recorded speech with 80dB bus background noise.



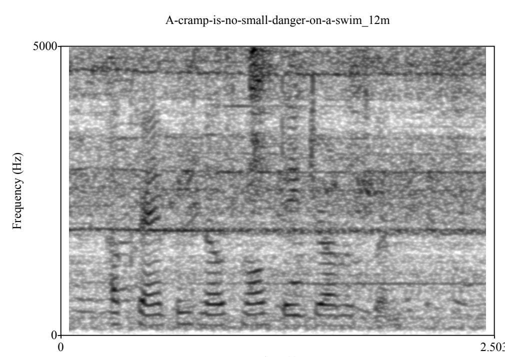
(b) Narrow band spectogram of ear recorded speech with 80dB bus background noise. The signal has been pre-emphasized.

Figure 1.11: Narrow band spectogram of ear-recorded speech from 0-20kHz to look for possible speech information in higher frequencies, of the sentence "A cramp is no small danger on a swim" spoken by a female participant.

frequencies, but there does not appear to be harmonics, nor does any of the visible acoustic energy appear to correlate with the speech seen in the lower frequencies. To be certain, the ear speech was pre-emphasized, seen in fig. 1.11b. It appears that while those fainter harmonics in the lower midrange frequencies have become more pronounced, there is no new speech information in the upper frequencies which



(a) The ear recorded speech which has been pre-emphasized, filtered at 2500 Hz, and pre-emphasized a second time.



(b) The noisy spectogram of the mouth-recorded speech; previously seen in 1.10a, repeated here for ease of comparison.

Figure 1.12: Narrow band spectogram of "A cramp os no small danger on a swim" recorded at the ear (1.12a) and the mouth (1.12b) and spoken by a female participant, with 80dB bus noise in the background.

has made it past the noise threshold.

This ear-recorded signal was then low-pass filtered at 2500 Hz with a 500 Hz smoothing slope. To further emphasize the higher frequencies in the available range (and to smooth over the 'muffled' attribute a bit), the sound was preemphasized a second time (after filtering). This can be seen in Figure 1.12a, next to the noisy mouth speech for comparison 1.12b.

1.5 Summary and Discussion

Here is my summary and discussion. **Mention the limitations - namely the directionality of the microphone affecting the lack of noisyness in the data**

REFERENCES

Allen, G. and Fernandez, C. (1960). The mechanism of bone conduction. *Ann Otol. Rhinol Laryngol*, 69:5–28.

Ballachanda, B. B. (1997). Theoretical and Applied External Ear Acoustics. *Jour. of the American Acadamy of Audiology*, 8:411–420.

Békésy, G. v. (1948). Vibration of the Head in a Sound Field and its Role in Hearing by Bone Conduction. *Jour. of the Acoust. Soc. of Am.*, 20(6):749–760.

Békésy, G. v. (1960). *Experiments in Hearing.* McGraw Hill Book Co., New York.

Borghese, N. A., Ferrigno, G., Redolfi, M., and Pedotti, A. (1997). Automatic integrated analysis of jaw and lip movement in speech production. *Jour. of Acoust. Soc. of Am.*, 101(1):482–487.

CHiME Challenge (2016). Chime speech separation and recognition challenge. `http://spandh.dcs.shef.ac.uk/chime_challenge/`. Online; accessed 02-18-2016.

Dean, M. S. and Martin, F. N. (2000). Insert Earphone Depth and the Occlusion Effect. *American Journal of Audiology*, 9(2):131–134.

Goldstein, D. P. and Hayes, C. S. (1965). The Occlusion Effect in Bone Conduction Hearing. *Journal of Speech, Language, and Hearing Research*, 8:137–148.

Hansen, M. (1998). *Occlusion Effects Part 2: A Study of the Occlusion Effect Mechanism and the Influence of the Earmould Properties.* PhD thesis, Technical University of Denmark.

Håkansson, B., Brandt, A., Peder, C., and Tjellstrm, A. (1994). Resonant Frequencies of the Human Skull in vivo. *Jour. of the Acoust. Soc. of Am.*, 95(3):1474–1481.

IEEE (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17.

Kelly, N. H. and Reger, S. N. (1937). The Effect of Binaural Occlusion of the External Auditory Meati on the Sensitivity of the Normal Ear for Bone Conducted Sound. *Jour. of Experimental Psychology*, 21(2):211–217.

Littler, T. S., Knight, J. J., and Strange, P. H. (1952). Hearing by Bone Conduction and the Use of Bone Conduction Hearing Aids. In *Proceedings of the Royal Society of Medicine*, volume 45, pages 783–790.

May, A. and Dillon, H. (1992). Comparison of physical measurements of the occlusion effect with subjective reports. Paper presented at the Audiologic Soc. Conference, Barossa National Valley Acoustic Laboratory, Sydney.

Pørschmann, C. (2000). Influences of Bone Conduction and Air Conduction on the Sound of One's Own Voice. *Acta Acustica*, 86:1038–1045.

Reinfeldt, S., Östli, P., Håkansson, B., and Stenfelt, S. (2010). Hearing ones own voice during phoneme vocalizationTransmission by air and bone conduction. *The Journal of the Acoustical Society of America*, 128(2):751–762.

Rosen, S. and Howell, P. (1991). *Signals and systems for speech and hearing*. Academic Press Inc., San Diego.

Salvinelli, F., Maurizi, M., Calamita, S., D'Alatri, L., Capellis, A., and Carbone, A. (1991). The external ear and the tympanic membrane. *Scand Auidiol*, 20:253–256.

Stenfelt, S., Hkansson, B., and Tjellstrm, A. (2000). Vibration characteristics of Bone Conducted sound in vitro. *Jour. of the Acoust. Soc. of Am.*, 107(1):422–431.

Stenfelt, S. and Reinfeldt, S. (2007). A Model of the Occlusion Effect with Bone Conducted Stimulation. *International Journal of Audiology*, 46:595–608.

Stinson, M. R. and Lawton, B. W. (1989). Specification of the geometry of the human ear canal for the prediction of sound-pressure level distribution. *Jour. of Acoust. Soc. of Am.*, 85(6):2492–2503.

Thorup, A. (1996). Okklusion (in danish). Master's thesis, Technical University of Denmark.

Tonndorf, J. (1972). *Foundations of Modern Auditory Theory*, chapter Bone Conduction, pages 195–238. Academic Press Inc., New York.

Wheatstone, C. (1879). *The Scientific Papers of Sir Charles Wheatstone*. Physical Society of London.

Wimmer, V. (1986). The occlusion effect from earmolds. *Hearing Instruments*, 37:19–57.