CHAPTER 1

Ear-Recorded Speech

1.1 Introduction

The accuracy of automatic speech recognition (ASR) has significantly improved over the past several years; similar, but less dramatic improvements have been made regarding the challenging task of recognizing speech in noisy situations (Zhang et al. (2017)). The performance of the latter still falls below the accuracy that is able to be reached by most human listeners. Even the human auditory system itself is only able to remove a limited amount of noise from a speech signal before that signal becomes completely unintelligible. The primary issue arises from the fact that speech normally passes from a speaker's mouth into either the ear (human speech perception) or a microphone (ASR), but the passage of speech through the medium of air allows corruption of the signal by noise of unknown loudness and unpredictable source.

This research proposes eliminating, by and large, the passage of speech through air by recording the speech in an unconventional location - from the inside of the *speaker's* ear canal. Speech is not only emitted from the oral cavity, but it is widely known that the vibrations pass throughout the human body. Using the ear canal as the source of speech adds the benefit that the entrance to the ear canal can be securely occluded behind the microphone (eg. with a noise reduction device such as an ear-plug). Thus, ambient noise from the air will be be largely filtered out by the occlusion device and the human skull.

Due to the very specific nature of the requirements for this study (namely, that the speech data recorded needs to be recorded from the ear), it is necessary to record data from scratch, rather than use an existing and more widely recognized corpus of speech. The goal for this new corpus is to offer speech data recorded from the inside

of the ear canal, as well as speech data recorded from the mouth, for comparison. Both clean and noisy speech are desired. Owing to the variability that occurs when even the same speaker repeats the same sentence, speech critically needs to be recorded from the two locations simultaneously. This allows for a more accurate comparison of the two signals. The initial experiment in Section 1.3 below is 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 (cf. Chapters **??** and **??**). The next section explains the theory behind the assumption that usable speech can be collected from the inside of one's ear canal.

1.2 Background

The speech vibrations of a person's own voice will propagate 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.

1.2.1 Bone Conduction

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. A few (cf. Békésy (1948), Hansen (1998), Pørschmann (2000), and Reinfeldt et al. (2010)) have investigated bone conduction when the source of vibration (ie. sound) is a person's own voice, not an artificial mechanically-created vibration. These studies also record using a microphone at the entrance to a speaker's ear canal, and not via another sensor on a different side of the skull.

The many studies that use a simple mechanically-created vibration as a stimulus do so in part because the use of speech as a source is inherently messy.

This is due to the fact that speech is a) is not as easily manipulated as a simple mechanically-created vibration, b) contains far more frequency components than a simple mechanically-created vibration, and c) 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[1], through the solid medium of the head[2], 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.

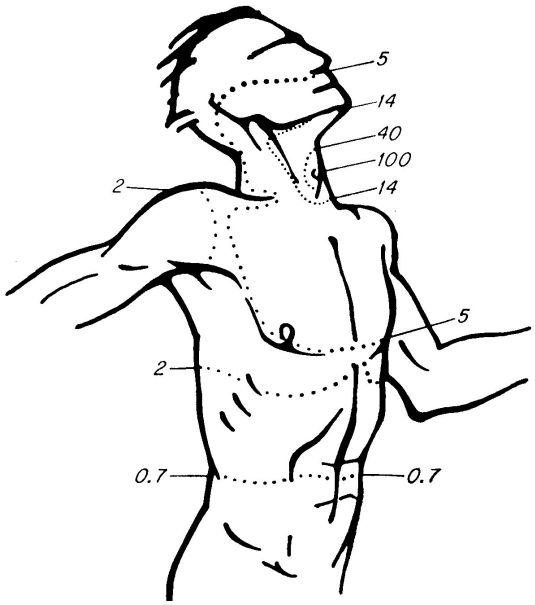


Figure 1.1: Diagram of the propagation of speech waves throughout the body. Numbers correspond to percentage of the original amplitude of the speech remaining when reaching the marked location. Taken from Békésy (1960).

1.2.2 Ear Canal Resonance

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 could 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

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

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

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 previous models (see Fig. 1.2). They noted that most significant differences between these models' spectral predictions of ear canal resonance occur above 6kHz.

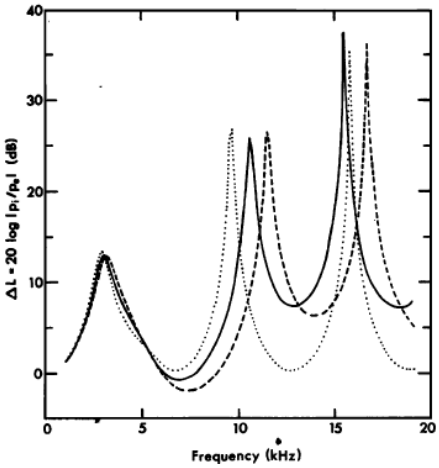


Figure 1.2: Stinson and Lawton (1989) diagrams three different models of the ear canal resonance. The bold line is based on their 3D canal molds from cadavers, the dashed line removes the curvature of the ear canal and acts as if the axis were straight, the dotted line assumes a constant diameter along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

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. Treating the ear canal model as a uniform tube, as opposed to incorporating the nuances of its diameter and curvature, will not have much effect on the output of a model.

Another challenge is to obtain the dimensions of the ear canal needed in order to

treat it as a uniform tube. 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 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 (eg. using a silicon mold). Making an assumption about either an 'average' diameter or an 'average' length of the ear canal would allow for the approximate calculation of the other dimension, given the measured volume. 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)).

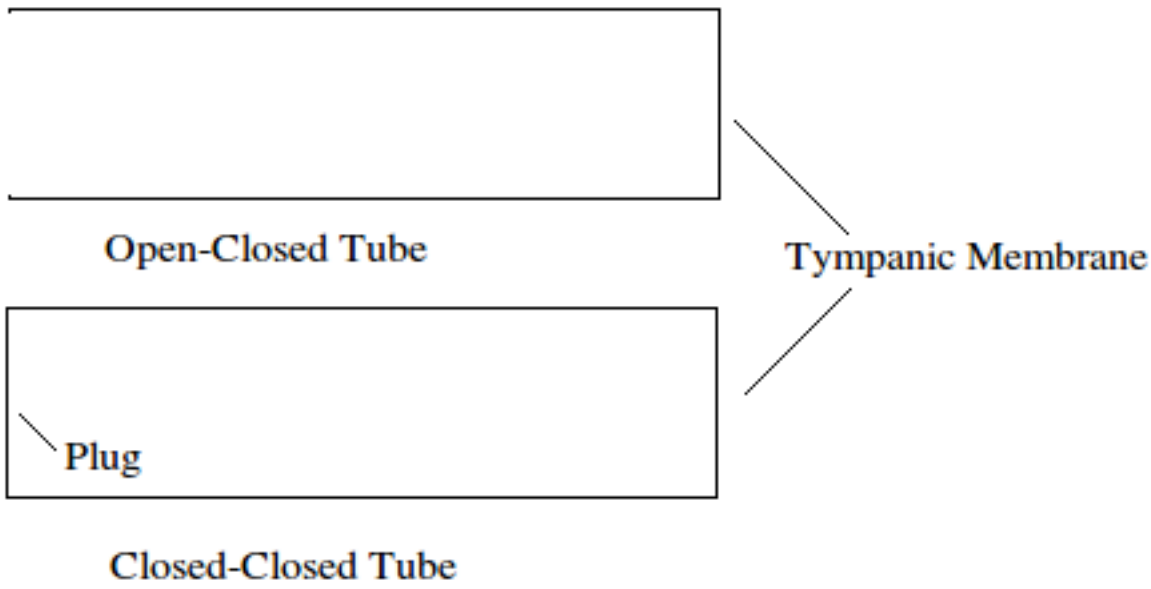


Figure 1.3: A diagram showing an example of an open-closed tube (top) and a closed-closed tube (bottom). Applied to the case at hand, the closed end (on the right) in both figures would be the tympanic membrane. The opening at the left (top) is the entrance to the ear canal, which can be plugged (bottom).

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, cf. Figure 1.3). This difference changes the resonance and reverberant structure of the ear canal, and will be discussed further in Section 1.2.4.

1.2.3 Ear Canal Resonance on Bone-Conducted Speech

There have been many studies pertaining to ear canal resonance from bone conducted speech. There are a few in particular (c.f. Békésy (1948), Pørschmann (2000), Reinfeldt et al. (2010)) which use real human speech as the sound source and measure the human ear as an open-closed tube. The acoustics of uniform tubes is usually thought of in terms of an open-open tube, an open-closed tube, or a closed-closed tube. Each of these have different resonating characteristics. The human vocal tract, for example, is normally modeled as an open-closed tube, where one end (the glottis) is generally considered 'closed' for modelling purposes, and the other (the mouth) is generally considered to be 'open'. Similarly, for the human ear, the tympanic membrane represents the 'closed' end of the ear canal tube, and the ear canal opening at the concha, or pinna, is the 'open' end (cf. Figure 1.3).

Pørschmann (2000)'s study is generally looking at the *self-perception* of one's own voice, but in order to accomplish this devotes effort to looking at the bone conduction pathway separately. A general 0.9 kHz resonance (with subsequent harmonic resonances) was found in the collected bone-conduction speech, with the amplitude gain generally present between 0.7 and 1.2 kHz. This correlates with the 0.8-1.2 kHz range for the first resonance that others (cf. Håkansson et al. (1994)) have observed in mechanical-stimulated bone conduction studies. This would mean, in terms of speech, that one would expect to find higher amplitudes near the first formant.

However, in this study only two phones were used (/s/ and /z/), and a masking threshold[3] technique was used to determine the frequency spectrum of the transfer

[3]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. Knowing the amplitude of the tone allows the researcher to know the amplitude of the speech as one becomes audible over the other. Having this knowledge, the

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 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 distinct differences, which can be seen in Fig. 1.4. 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. 1.5).

In particular, as can be seen in Fig. 1.4, 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, particularly 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. 1.4 and 1.5 both contain *relative* spectral envelopes - ie. 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 dips down to -5 dB around 1kHz. This is in contrast with the phone /s/, which has a fairly high (yet falling) response below 500 Hz, and does not dip below -5 dB until nearly 4 kHz. 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 obstruent does not pass through to the ear canal (Reinfeldt et al. (2010)).

---

spectrum of speech as it is perceived by the speaker can to be mapped.

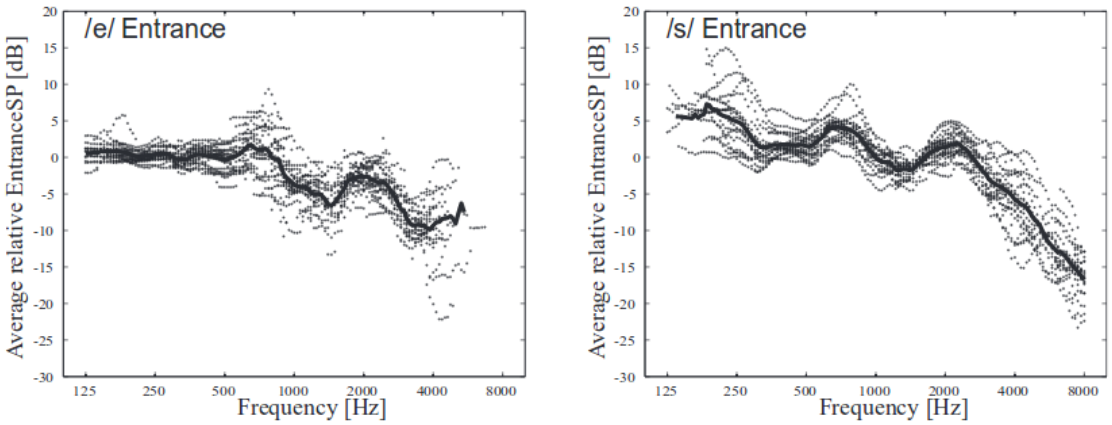


Figure 1.4: The amplitude of Body-Conducted speech relative to the amplitude of Air-Conducted speech as recorded in the ear canal for the phones /e/ and /s/. A value of less than zero indicates the amplitude of body-conducted speech is less than that of air-conducted speech, and a value greater than zero indicates a higher amplitude of body-conducted speech than air-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).

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 body-conducted amplitude relative to the air-conducted counterpart was much higher for /i/ than it was for /a/. An assumption 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. 1.5). Reinfeldt et al. (2010)'s findings are backed 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. 1.6), which also supports this hypothesis. There was much inter-speaker variability in the previous studies, but the bone-conducted vowels appeared to have the least relative reduction in amplitude are the higher vowels. Since the functions of other high sonorants (eg. /u/) are not given, we cannot be certain if this is a phone-specific difference, or if it can be generalized to other sound with a [high] articulation in which the tongue is close to the roof of the mouth. This latter generalization would

make sense, as the oral cavity is more 'closed, trapping energy inside the cavity resulting in move reverberations passing through the head and into the ear canal. On the surface, it appears that the more energy that is lost to air conduction during the production of low vowels (ie. from a more 'open' articulation), the less energy is transferred into the surrounding tissue.

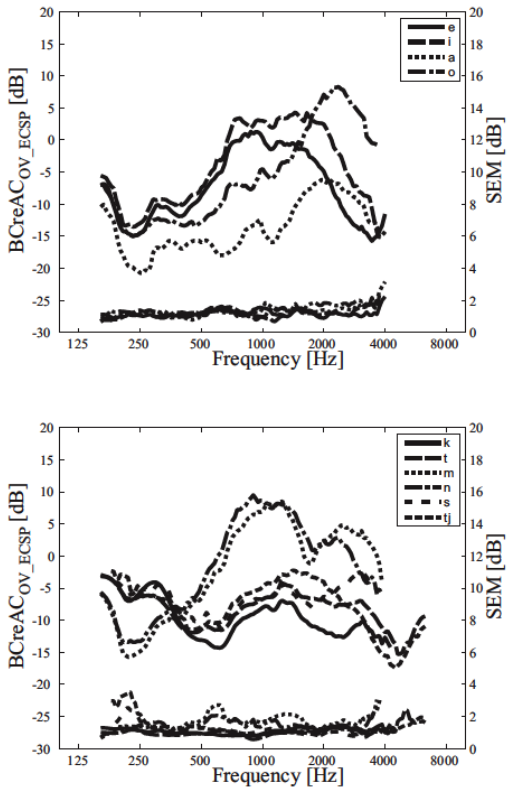


Figure 1.5: The mean relative amplitudes (left ordinate) of body conduction relative to air conduction for vowels (top plot) and other sounds (bottom plot). The set of lines along the bottom of each plot represent the standard error from the mean (SEM), measure on the right ordinate. Taken from Reinfeldt et al. (2010).

Here it is important to re-emphasize that these transforms are given as body-conducted 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, and so a direct comparison between two bone-conducted phones cannot be conducted.

1.2.4 The Occlusion Effect on Bone-Conducted Speech

Returning to the notion of tubes, if the ear were to be occluded at its one open end, the ear canal would shift from an open-closed tube to a closed-closed tube, and the frequency response would be altered accordingly. This phenomenon, first noted by Wheatstone (1879), is termed the occlusion effect (OE). The occlusion effect (OE) is the change in sound pressure level (SPL) resulting from body-conducted vibrations emanating into, and reverberating within, a *closed* ear canal. Objectively, inside the ear canal, the sound pressure level at lower frequencies is increased when the ear canal is occluded. This has been studied extensively (cf. Wheatstone (1879); Kelly and Reger (1937); Littler et al. (1952); Tonndorf et al. (1966), among many others).

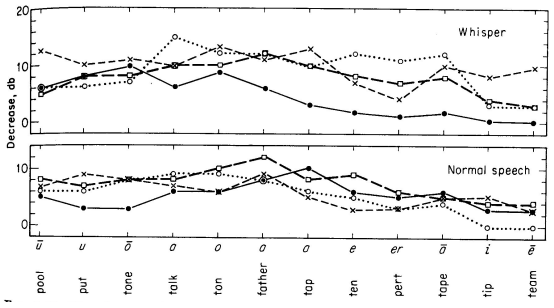


Figure 1.6: Demonstrated the different effect on amplitude that closing the ear canal has on the different vowels of English. Taken from Békésy (1960).

Of particular focus was the mechanism behind this shift in amplitude to the lower frequencies. Huizing (1960) proposed that this 'restructuring' of the frequency spectrum was due solely to the change in resonance characteristics of the ear canal when it is occluded. This proposal is supported the known physical phenomenon that the occlusion (or lack thereof) of ends of a uniform tube change the resonance properties of the tube. The resonance frequencies of an open-closed tube are proportional to 4x the length of the tube and a closed-closed tube produces resonant frequencies which are proportional to 2x the length of the tube. Using the known characteristics of tube resonance to describe an occluded ear can partially account for what is observed; it largely describes what happens in the upper frequencies (generally above 2 kHz, Stenfelt et al. (2003)).

Tonndorf et al. (1966), through many studies utilizing domestic cats, discovered

the mechanism behind the apparent low-frequency gain. When the ear canal is open, due to "the mass-effect of the air column in the ear canal together with the compliance with the air in the canal" it acts as a high-pass filter, dampening the lower frequencies (Stenfelt et al. (2003), 910). When the canal is occluded, this high-pass filter is gone, and the formerly filtered low-frequency energy is present within the canal.

As with bone conduction in general, most of the research of the occlusion effect (OE) has been conducted using controlled mechanical vibrations. Using a mechanical stimulus, Békésy (1960) reported that when the ear canal is closed, the increase in amplitude within the canal can be observed up to 2 kHz, and above 2 kHz, the increase in loudness 'vanishes' quite suddenly (cf. Fig. 1.7).

However, there is a variance in the OE - if using a mechanically-created vibration - 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 this difference tends to wash out when slightly higher frequencies are reached[4]. There are also differences based on the location of *occlusion* within the ear canal, ie. how deep a plug is placed in the ear canal. This affects the resonant frequencies that play a part in the shape of the higher frequencies, due to the different depths of plug insertion resulting in 'tubes' of different lengths. Another effect of plug depth is that, the deeper a plug is inserted into the ear canal, the less surface area there is in the canal for acoustic vibrations to enter.

Dean and Martin (2000)'s results indicate that the effect of plug depth on the amplitude of a signal does not disappear as frequency increases. The difference in relative amplitude between the different insertion depths is greatest at lower frequencies, with supra-aural earmuffs resulting in the greatest amplitude increase, and the deep inserted ear-plugs with the lowest. However, at 1 kHz, the shallow-inserted ear-plug has a greater relative amplitude gain than the supra-aural earmuff.

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

Dean and Martin (2000) does not mention the explicit depth used for each condition.

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 'earmuffs' or a deep-insertion 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 ear-plugs.

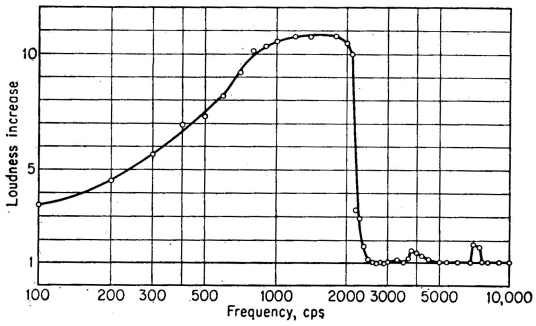
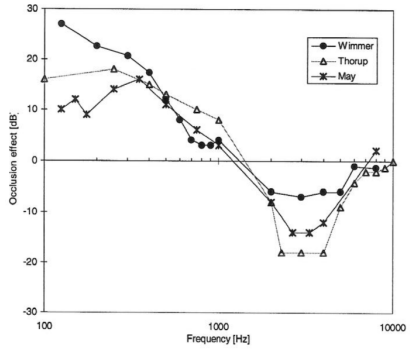


Figure 1.7: The frequency response inside the ear canal when taking a mechanical vibration as stimulus to a participant's forehead. Taken from Békésy (1960).

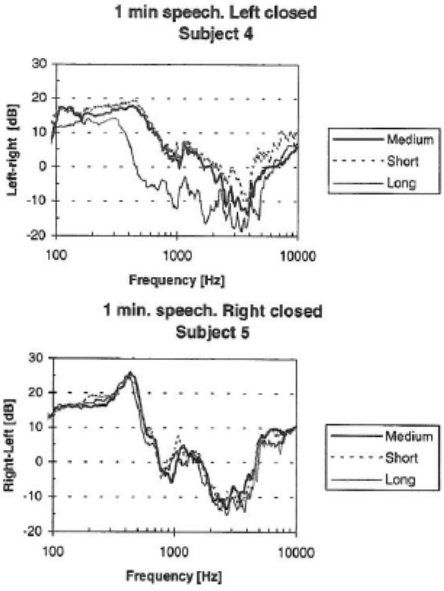
With shallow insertion, Stenfelt and Reinfeldt (2007)'s model estimates a gain in amplitude of frequencies below 2 kHz, and dampening of those above; all insertion depths, according to their model, will at minimum, slightly dampen frequencies above 2 kHz. As the plug is inserted deeper, the damping occurs on lower and lower frequencies. These results contrast slightly with Békésy (1960)'s in Fig. 1.7 in that they predict higher, very low frequencies, as opposed with Békésy (1960)'s bell-shaped resonance around 1-2kHz.

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

(From Wimmer (1986), Thorup (1996), and May and Dillon (1992), Fig. 1.8a). The study conducted its own tests (seen in Fig. 1.8b), which, by and large, agree with the previous studies. These represent the 'average' effect of occlusion on speech, and appear to resemble other (mechanical-source) estimations. Hansen (1998) developed a model of the OE which largely agrees with these measurements.



(a)



(b)

Figure 1.8: In (a), three separate measured OE spectra. In (b), OE of two subjects with ear molds extending into the canal at different lengths (Hansen (1998)).

The increase in amplitude of the voice in the ear due to occlusion can also influence the manner in which speech is spoken. This can be seen clearly by the OE's influence on the Lombard effect (Lombard; Lane and Tranel (1971)). The Lombard effect is the tendency of humans to speak louder in noise due to noise interrupting the normal feedback loop of speech, ie. speakers do not hear themselves nearly as well in noise as they do normally, and so they speak louder (and clearer) to compensate. However, recent research by Brungart et al. (2012) demonstrated that the Lombard effect disappears when a participant is wearing ear-plugs in a noisy environment. This is actually due to the presence of the OE, and the increase in amplitude of one's own voice by closing off the entrance to the ear canal. Instead of an interruption in, or lack of, feedback, the

occlusion provides the speaker with much louder feedback. Brungart et al. (2012) found that speakers wearing ear-plugs in a noisy environment actually speak *quieter* than those without ear-plugs, *regardless* of the actual noise-attenuation by the ear-plugs. This indicates that it isn't necessarily the lack of noise that prompts speakers to use a lower volume level, but the presence of occlusion and the consequent increase in amplitude of their own voice.

Yet despite an increase in overall amplitude, the OE does not alter all speech sounds equally. The studies looking at the OE of 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 variability 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 for 'higher' or 'lower' phones (eg. /i/ vs /a/). The placement of the jaw bone, as it moves changes the impedance characteristics of the vibration of the mandible against the temporal bone that houses the meatus (Békésy (1960)). 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 (ie. with a mandibular joint adjacent to the cartilage of the ear canal or without).

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, behaves differently when there are different sized 'leaks' resulting in different levels of partial 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 open ear canal and those from an occluded ear canal, which can largely be seen between Figs. 1.5 and 1.8a. There is a massive increase in the amplitude of the lower frequencies, which is not present from an open-closed ear canal, and a sizable drop in amplitude after 2kHz, which similarly does not seem to manifest itself when the ear canal is not occluded.

1.2.5 Summary

The aforementioned studies on body conduction and the occlusion effect, as would be expected, have indicated a large amount 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 resonance within a closed ear canal that cause relatively higher amplitude in the lower frequencies (below 2 kHz), and a drop in frequencies above that range (with a few exceptions). With this knowledge, it appears that it may be possible for recoverable speech information to be recorded from inside the ear canal. While the skull will prevent much ambient noise from reaching a microphone placed in the ear canal, an occlusion device would need to be placed at the opening of the canal to aid in dampening the noise. This occlusion results in the distortion

described above.

This distortion is hypothesized to be, by and large, predictable (Reinfeldt et al. (2010)), unlike ambient noise from the environment which is generally highly variable in both amplitude and form (Zhang et al. (2017)). This prediction is that the signal recorded from the ear canal is expected to be heavily low-pass filtered above approximately 2kHz. Minor transformations, such as pre-emphasizing the higher frequencies, are hypothesized to recover some of the information that is lost, resulting in speech that will perform better than a noisy signal collected at the mouth in both ASR and human speech perception tasks. Due to this, the technique of substituting potentially unanticipated, variable 'environmental' noise with the anticipated 'noise' of body conduction and the occlusion effect will allow for greater confidence that a usable signal could be recovered.

Section 1.3 below describes the specific methods used to collect speech data from the mouth and the ear canal, and the analysis of the collected speech in an attempt to recover an intelligible signal with the knowledge outlined in this section.

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

There are numerous constraints and requirements for the speech recordings necessary for this task, and it was essential to create an original dataset for this study. A small corpus was needed to be able to conduct ASR and Human speech perception experiments. Primarily, a corpus needed to be created because there is no known dataset of speech in which the recording location is inside the ear canal; speech recorded from this location was the primary focus of these studies. Secondly, in order to be comparable, the speech at the ear, and the speech at the mouth, needed to be recorded at the same time, in the same conditions. This was mainly to determine, when comparing the results of ear-recorded and mouth-recorded speech, whether the ear or mouth-recorded speech performed better in the ASR and human speech perception tasks and to avoid potentially confounding variables that would occur if the two sets of speech were recorded separately. This is discussed further

in Section 1.3.1 below.

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 were needed to test the following (previously mentioned) hypotheses: (a) whether recording speech from the ear, external noise would be completely or largely eliminated, while simultaneously recorded speech from the mouth had a noisy background, (b) whether the speech from the ear was more intelligible and recognizable by an ASR system than noisy speech, and (c) whether the speech from the ear is more intelligible and recognizable by humans than noisy speech.

In alignment with the CHiME challenge[5] guidelines, this study uses different types of background noise at different noise levels. The noises used include the four sounds (bus, café, pedestrian area, & street) from the CHiME Challenge (2016), 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. Relatively level amplitude was used to allow for a more accurate comparison of the different SNR levels. Furthermore, if a sound file varied in amplitude, it would confound the recognition tasks as to whether a certain word is difficult to recognize, or the noise level that occurred at that portion in the sentence was a hindrance to accurate recognition.[6]

Many existing works in ASR and human speech recognition in noise use multiple SNR levels to demonstrate the effectiveness of the technique of noise removal (eg. Braun et al. (2016)). This is generally done by adding a noise signal to already recorded clean speech, giving the researcher acute control over the SNR. Recording the noise while simultaneously recording the speech was necessary to demonstrate

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

[6]The exact portions of the sounds which were used are available online, along with the rest of the data at http://www.openslr.org/

the ability of the ear canal recording location to remove ambient noise. Due to this need, it was determined that noise would be played from a loudspeaker at pre-determined decibel levels.

Since human speech is variable in loudness, and amplitude will likely vary between and within speakers, only three, well-spaced noise levels were chosen to allow speakers' various 'loudnesses' to fall in the same, broad categories. Conversational speech is generally around 70 dB, and so the noise levels chosen were 60 dB, 70 dB, and 80 dB. These were the 'averaged' dB levels obtained by averaging over the duration of the sound file. This would result in approximate SNR conditions of +10 (60 dB), 0 (70 dB), and -10 (80 dB). A signal with +10 dB SNR is in the range of SNRs where ASR and human listeners have a very high recognition accuracy (cf. Braun et al. (2016); Gilbert et al. (2013)). A signal with 0 dB SNR occurs in a range in which ASR and human listeners are still able to make out most of a speech signal, but recognition begins to falter. At -10 dB SNR, recognition performance very noticeably suffers. As stated before, it was assumed that speakers will vary in how loud they speak, and so actual SNRs were expected to vary.

An amplitude of 80 dB was chosen as a max loudness, rather than a higher noise level to achieve a lower SNR, in order to leave a wide margin between it and any (albeit remote) possibility of hearing damage suffered by participants. A 'clean' (no noise) condition was also utilized for each sentence. This creates 16 different conditions (5 noise types * 3 noise levels + 1 'clean' condition).

1.3.2 Stimuli

Thirty sentences were chosen from 3 Harvard Sentence lists[7]. Sentences were chosen from the Harvard Sentence dataset due to being phonetically balanced (the distribution of phonemes in each list proportional to their occurrence in English), to their prolific use in speech science research, and specifically their history of serving

[7]The 'Harvard Sentences' dataset 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)).

as stimuli for many speech corpora (cf. Kabal (2002); Hu and Loizou (2007), the latter being a noisy speech corpus). 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 noise reduction rating (NRR)) of -30 dB SPL were worn by the participants during the experiment.

1.3.4 Participants

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

1.3.5 Procedure

The participants were initially asked a few preliminary demographic questions[8]. They were seated in front of the Middle Ear Analyzer. An otoscope was used to ensure the right ear was mostly free of cerumen, to avoid blocking the microphone off

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

from the rest of the canal and generally impacting the canal with cerumen. The ear was fitted with an appropriate sized rubber clinical single-use ear tip, into which the Middle Ear Analyzer hose was already plugged. An immittance test was performed, which involves playing a tone, and slightly and briefly pressurizing the ear. The Middle Ear Analyzer checked that the ear-plug solidly sealed off the ear canal in order to be able to build up pressure, and would alert the researcher to a leak if the plug were not securely in place. This test gave 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 was given, which was checked for normalcy (cf. Appendix A**??**). The Middle Ear Analyzer gave other measures which were not used in this study.

The distance from the end of the ear-plug to where it was enclosed by the ear canal was measured to determine how far the plug was inserted in the ear canal. 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, from which the insertion depth can be calculated. The decision to treat the cavity of the concha as the 'end" of the ear canal was 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 began) as the end to the ear canal. This measure allowed for the calculation of the depth of insertion of the ear-plug.

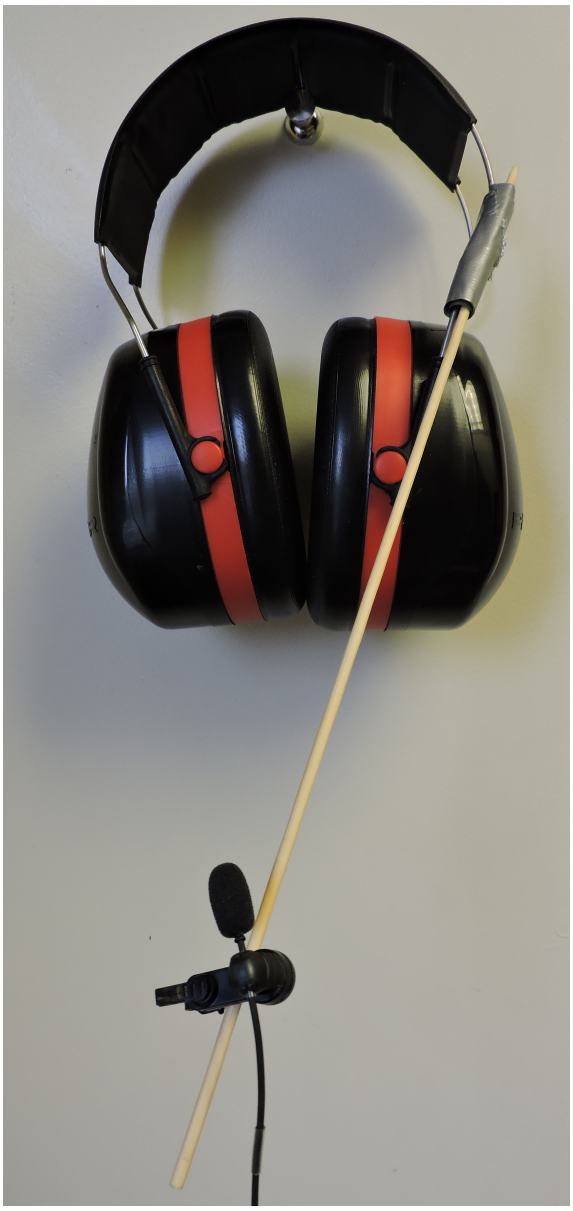


Figure 1.9: The Countryman B2D directional lavalier microphone, with with the wind-break foam removed, and inserted into a blue ear-plug.

The Middle Ear Analyzer hose was then removed from the ear-plug - which was carefully left in place to ensure a continuous seal. The participants then moved to a seat located in front of a computer monitor (cf. Fig 1.11 for set-up diagram). The

participants were then instructed as to the proceedings of the rest of the experiment. One of the two microphones was taken, the wind-break foam removed, and was 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 ear-plug (cf. Fig. 1.9). There were several instances where the microphone was inserted deeper than, or just shy of, the end of the ear-plug; the variance was within +/-2mm depth (cf. Appendix A**??**). The earmuffs were placed over both ears. Occasionally, participants had glasses, or thick hair, which may have slightly compromised the seal. A note was taken of this.

Figure 1.10: The 3M 30 NRR earmuffs, a wooden rod attached with adhesive tape, and a Countryman B2D directional lavalier microphone directed toward the location of the mouth.

A wooden rod was attached to the earmuffs, which extended forward, beside the participants' face. The second microphone was attached to this wooden rod via the lavalier clip at the level of the participants' mouth. The microphone was directed toward their mouth (cf. Fig 1.10). The placement of the microphone on the wooden rod was adjusted to be exactly 10cm away from the participants' infra-nasal depression. At this point, the participants were asked to adjust the placement of their chair so that the microphone on the wooden rod would be 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. The loudspeaker was on another table to the right of the participants, perpendicular to the direction of the microphone facing the mouth (cf. Fig. 1.11).

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 participants were situated, but before beginning the recording.

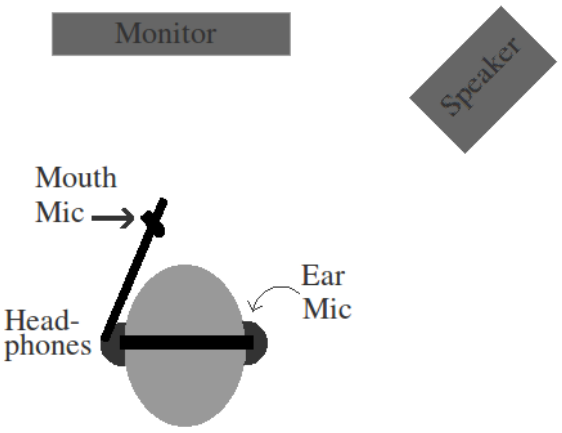


Figure 1.11: A diagram of the basic equipment set up for the experiment. The in-ear microphone is placed under the headphones in the right ear.

Once recording, an in-house computer program was used to display the stimuli sentences on a second monitor and to play the background noises. For each sentence, the participants saw the clean-condition (no noise) first. The researcher was in the soundbooth with the participants listening through a pair of headphones connected to the preamplifier. The participants were asked to repeat the sentence twice to develop a rhythm, at a normal, conversational loudness, with a normal, declarative intonation. The researcher asked the participants to repeat the sentence again in this condition if the rhythm or intonation of the two sentences did not match, or if the participants stumbled over a word. Each of the following 15 iterations of the sentence (one for each noise-type/noise-level combination) the participants were instructed to read the sentence aloud only once. Out of the loudspeaker, placed 1 meter from the participants, each

of the background noises were emitted to coincide with an iteration of the sentence. If the rhythm of the spoken sentence varied noticeably, or if the participant stumbled over a word, the researcher again asked the participants to repeat the sentence for that condition. The sentences were not randomized, ie. all 16 iterations of a sentence occurred consecutively[9]. *Within* each sentence group (after the 'clean' condition, which always occurred first), all the noise conditions were randomized. The researcher advanced each stimulus on the display for the participants.

To help the participants notice when a sentence had been advanced - as when wearing the noise reduction earmuffs, participants were often not aware when the noise condition changed - the number of the sentence condition was displayed underneath the stimulus (1-16). This had the unintended consequence of occasionally producing a mild list-intonation, as participants were aware of the final repetition of a given sentence (ie. the number '16' appeared beneath the sentence).

After the recording was finished, the participants were asked to complete a short, 4-question survey of their experiences during the experiment.[10] These included:

1. Can you describe your experience with wearing the experimental set-up?

2. Did you find the sound of your voice to be altered, annoying, or uncomfortable?

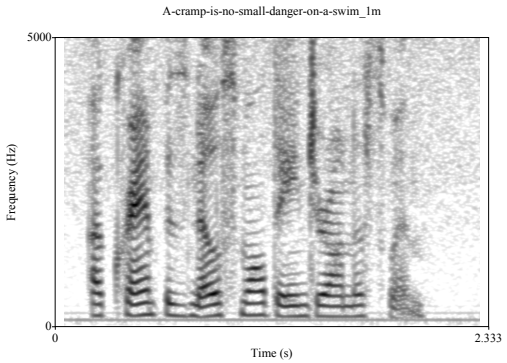
3. Would you consider wearing such a device if in a noisy workplace or environment in order to communicate?

4. Would you be more inclined to use such a device if earmuffs were not required?

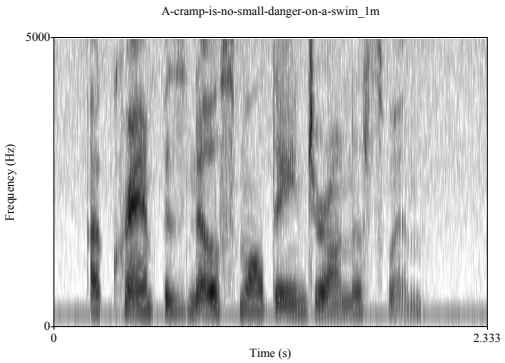
Participants were instructed to give as much detail in their answers as desired, and to answer truthfully. The answers to the first question that were provided by participants generally addressed the comfort of the experimental set-up (ie. the ear-mic and earmuffs). The researcher used the more specific question -"Did you find the experimental set-up (ie. the ear-mic and earmuffs) uncomfortable?" - to

[9]This was done to help the researcher ensure a similar intonation and rhythm for each iteration of the same sentence.

[10]cf. Appendix A**??** for coded answers.



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



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

Figure 1.12: Both (1.12a) and (1.12b) 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. 1.13.

code the answers. The answers were coded using a Likert scale with a range of 1-5, where '5' is 'yes - definitely' and '1' is 'no - definitely not'.

1.4 Observations of Collected Speech

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 1.12a, 1.12b, 1.13a, and 1.13b 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 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. It should be noted that while the two signals - from the mouth and from the ear - appear to have the same loudness, this is due only to the gain adjustment on the recording device during the experiment.

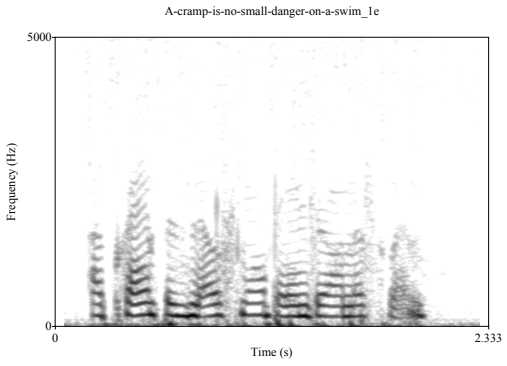
The speech from the ear was consistently *louder* than the speech from the mouth; on a scale of 0-10, the ear microphone gain was normally set anywhere between 2 and 4, while the mouth microphone gain was normally set anywhere between 4 and 6.

A cross-correlation similarity metric was run on each ear-recorded/mouth-recorded pair for each participant speaker (excluding those which contained noise in the background). The maximum value was obtained for each ear-recorded/mouth-recorded pair, was normalized, and then all values were averaged across speakers. This is given in Equation 1.1:

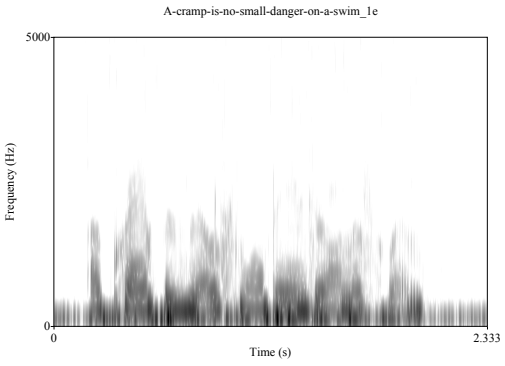
$$\sum_{k=1}^{K}\sum_{u=1}^{U}\frac{max(abs(xcorr(S_{uk}^{e},S_{uk}^{m})))}{\sqrt{\sum {S_{uk}^{e}}^{2}}*\sqrt{\sum {S_{uk}^{m}}^{2}}} \tag{1.1}$$

where K is the number of participants, U is the number of utterances, S^e refers to the ear-recorded signal, S^m refers to the mouth-recorded signal, and $xcorr$ is the cross-correlation function. If S_{uk}^{e} and S_{uk}^{m} are identical signals, the fractional portion of Equation 1.1 will return a '1'. The average normalized maximum cross-correlation value for ear-recorded speech compared with mouth-recorded speech is 0.358.

When noise is present, it can be seen in Figs. 1.14a and 1.14b that the noise does not affect the ear-recorded signal nearly as much. There appears to be some

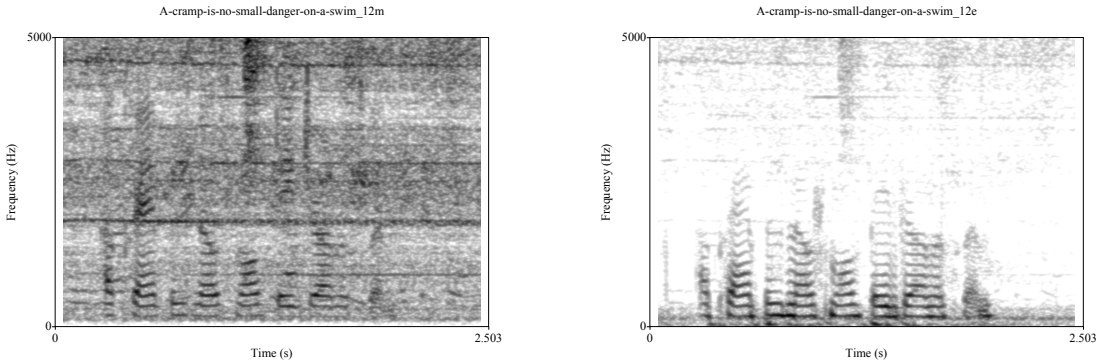


(a) Narrow band spectrogram of speech recorded from inside the ear canal.



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

Figure 1.13: Both (1.13a) and (1.13b) 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. 1.12.



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

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

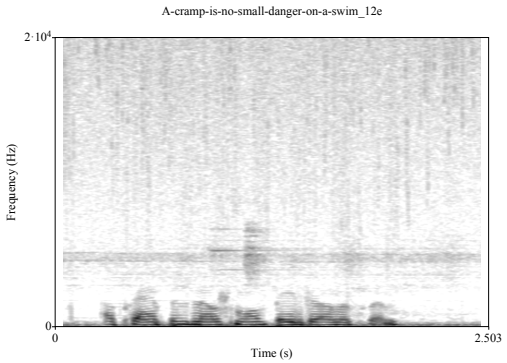
Figure 1.14: Both 1.14a and 1.14b were of the sentence "A cramp is no small danger on a swim", spoken by a female participant, recorded simultaneously.

of the louder noise (seen in the mouth-recorded speech in Fig. 1.14a) 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).

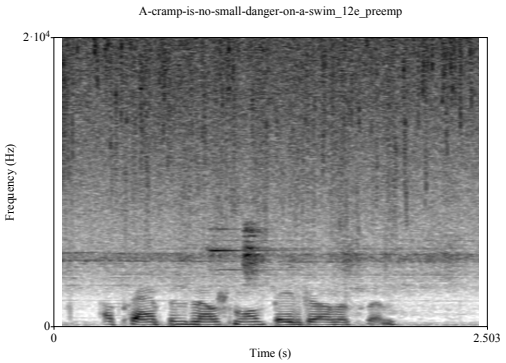
It should also be noted that the SNR in Fig. 1.14a is much higher than originally intended. For this particular example, the speech was recorded with an 80dB noise background, with the intent of obtaining a -10dB SNR. Instead, the SNR for the sentence in Fig. 1.14a is +6dB[11]. Unfortunately, this was widespread. This is attributed to a) the participant speaking louder than anticipated, resulting in a higher speech threshold, and b) the directionality of the microphone used eliminated much more background noise than anticipated.

In an attempt to see if there was recoverable information in the higher frequencies of the ear signal, the spectrogram range was increased from 5kHz to 20kHz (see Fig. 1.15a). There is certainly acoustic energy that makes it to the higher frequencies,

[11]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 noisy 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 raw noise vectors were calculated, averaged, and then used in the SNR calculation. For explicit code, see Appendix E**??**



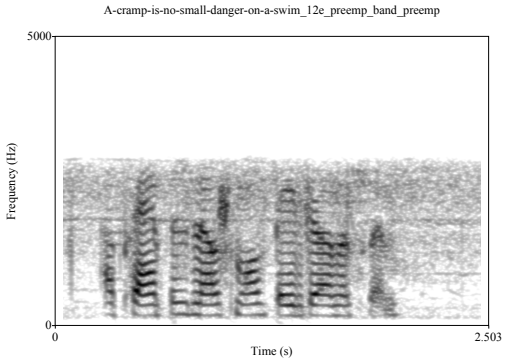
(a) Narrow band spectrogram of ear-recorded speech with 80dB 'bus' background noise.



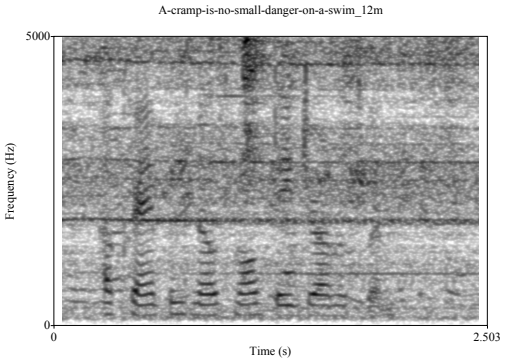
(b) Narrow band spectrogram of ear-recorded speech with 80dB 'bus' noise. The signal has been pre-emphasized.

Figure 1.15: Narrow band spectrogram 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. Note, the noise only extends to 8kHz.

but noticeable harmonic energy is not present, 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. 1.15b, which seems to confirm the lack of high-frequency speech information. It appears that while



(a) The ear-recorded speech, pre-emphasized, filtered at 2500 Hz with 500Hz slope, and pre-emphasized again.



(b) The noisy spectrogram of the mouth-recorded speech; previously seen in 1.14a, repeated here for ease of comparison.

Figure 1.16: Narrow band spectrogram of "A cramp is no small danger on a swim" recorded at the ear (1.16a) and the mouth (1.16b) and spoken by a female participant, with 80dB bus noise in the background.

Question	Mean Response	Mode Response
Did you find wearing the experimental set-up to be uncomfortable?	3.59	4
Did you find the sound of your voice to be altered, annoying, or uncomfortable?	3.85	4
Would you consider wearing such a device if in a noisy workplace or environment in order to communicate?	3.51	5
Would you be more inclined to use such a device if earmuffs were not required?	3.86	5

Table 1.1: Coded results from the post-experiment survey. Each response was given a Likert scale code of 1-5.

those fainter harmonics in the lower mid-range frequencies are more pronounced, there is no new speech information in the upper frequencies that makes 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 pre-emphasized a second time (after filtering). This can be seen in Figure 1.16a, next to the noisy mouth speech for comparison in Figure 1.16b. The average normalized maximum cross-correlation (Equation 1.1) was calculated for these transformed ear-recorded signals and the mouth-recorded signals (excluding the signals with noise), which yielded a value of 0.674 - an increase of 0.316 over that of the untransformed ear-recorded speech.

The author compiled the coded survey questions[12] and answers, which are located in Table 1.1.

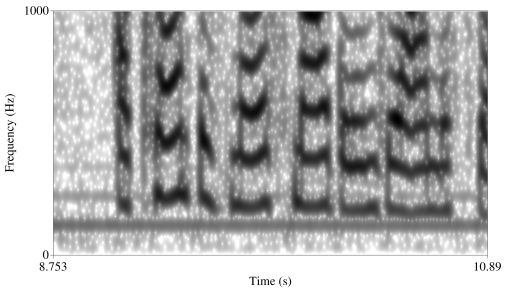
The mean response for each of the questions hovers slightly above the median response of '3'. Most participants found the recording set-up to be rather uncom-

[12]As mentioned at the end of Section 1.3.5, the actual question in the survey for Question 1 was "Can you describe your experience with wearing the experimental set-up?" Responses were coded according to their answer to Question 1 - as it appears in Table 1.1

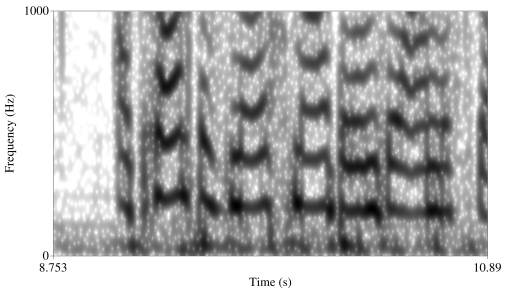
fortable; the earmuffs in particular seemed to play a role in the discomfort, as most participants would me more inclined to use the system if earmuffs weren't required. Participants seem split as to whether they would be willing to use the device in a real-life application (Q3). The mode is rather high (5) given the mean (3.51), which could be due in part to participant response bias (ie. desiring to provide an 'acceptable' response).

1.5 Limitations

During data collection, there were several issues that affected the quality of speech. The particular recorder which was used, when the gain knobs were turned up into the slightly higher range, would produce a low-frequency humming sound (cf. Figure 1.17). This was much more prominent for the mouth-recorded signals, as the gain for this channel was turned up higher. Since the headphones the researcher used were plugged into the preamplifier, this was not noticed until most participants were already recorded.



(a) The low frequency hum in the mouth-recorded signal. It can be seen at 120 Hz and subsequent harmonics with decaying amplitude.



(b) The low frequency hum in the ear-recorded signal. As in Figure 1.17a, it can be seen at 120 Hz and subsequent harmonics. It is less prominent than in Fig. 1.17a due to the gain on the recorder being lower.

Figure 1.17: Spectrograms of a low frequency hum introduced by the recorder at 120 Hz and subsequent harmonics in both mouth (Fig. 1.17a) and ear (Fig. 1.17b) recorded signals. The range of frequency is 0-1kHz.

On a physical level, the ear-plugs which were used are fairly standard audiological

silicone ear-plugs, which had a hole in the middle that was the correct size to fit the microphones that were used. The degree of noise damping these plugs offer is unknown, as they are not inherently designed for noise reduction. It is very likely that a better NRR ear-plug could be found and used in conjunction with the 30dB NRR earmuffs to achieve a greater noise reduction at even higher noise levels.

As mentioned previously, the primary limitation is that the background noise was not loud enough to be able to fit into the SNR ratios desired (+10 dB SNR, 0 dB SNR, -10 dB SNR). Instead, what was recorded was an average (across all speakers) of 31, 23, and 12 dB SNR for the 60, 70, and 80 dB condition, respectively; The 60 dB noise condition reached as high as +40 dB SNR, and the lowest 80 dB noise condition SNR was +5 dB. The three noise levels can be seen in the spectrograms in Figure 1.18 for the bus background noise. Note that the noise level can barely be observed in the lowest noise condition in Fig. 1.18a.

The three options to remedy this would be to a) ask the participants to speak more quietly, b) increase the volume of the loudspeaker, or c) use omnidirectonal microphones. Options (a) and (c) are more appealing, as it continues to keep the ambient noise outside the range of amplitude which could possibly damage hearing. However, in the case of (a), it is difficult for a participant to consciously modify their loudness and keep that consistent for the duration of the experiment; the need to repeat sentences spoken above a certain amplitude would drastically increase the length of time of the experiment. Additionally, it would be difficult for the researcher monitoring the speech to determine whether or not the speech was an appropriate loudness, particularly with the additive background noise playing at the same time.

Option (b) - increasing the noise level itself - is less appealing in that it results in a higher risk for hearing damage, despite being the most 'authentic' scenario, testing the directional microphones' capabilities at the mouth and ear to eliminate noise. Given the SNR of +6dB for the 80dB noise condition (with some 80dB noise conditions having upwards of +10dB SNR), this would mean increasing the ambient noise from 80dB to 96-100dB in order to achieve the -10dB SNR that was originally desired. Even a 0dB SNR would require a +10 dB increase to 90dB ambient noise

in order to turn the observed +10dB SNR into 0dB SNR. The risk to participants is substantially mitigated by the use of 30dB NRR earmuffs, however at noises of this magnitude, it would be difficult to find a location which insulated the sound from affecting a significant radius outside the sound booth. It is also very likely that a different loudspeaker would be required to reach the needed amplitudes without clipping.

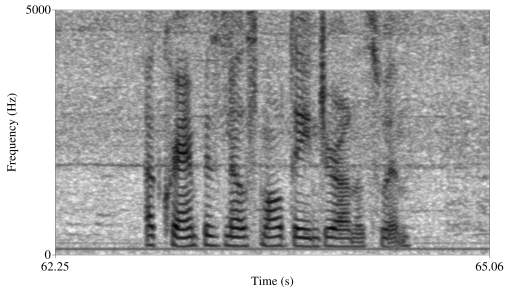
Option (c) - using an omnidirectional microphone - does not rely on speakers' ability to modulate their voice, nor does it introduce additional risk by increasing the ambient noise. The only drawback, as mentioned above, is that the use of an omnidirectional microphone results in an unfair comparison between the mouth-recorded speech and ear-recorded speech. Any real-life application will use a directional microphone at the mouth with the intention of eliminating as much ambient noise as possible. Nevertheless, this is likely the best option for future research in this area, given the restrictions described above.

1.6 Summary

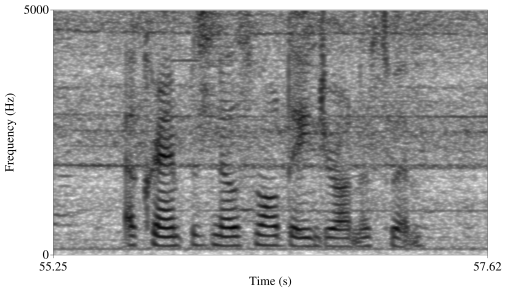
Despite some speaker variation, the speech recorded from inside the ear canal contains information up to approximately 2700Hz. This upper cut-off frequency was in the same range described by the aforementioned literature. Two, very basic acoustic transformations (pre-emphasis and bandpass filtering) were used in an attempt to create a more intelligible signal from the speech collected at the ear.

Limited benefits might be seen from a sound or sound-category specific alteration, particularly among fricative sounds and those with the majority of speech information located in frequencies above 2700Hz. It seems unlikely though, that spectral subtraction (eg. using the filters proposed by Hansen (1998) and Reinfeldt et al. (2010)) or a similar method would offer much additional benefit, as much of the upper frequencies were dampened beyond the existing noise level (c.f. Fig. 1.15a). However, it is hypothesized, given the transformed data collected at the ear, that enough information is present for the speech to be recognizable. This hypothesis

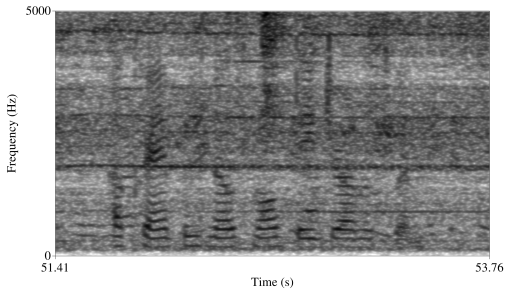
will be tested in a human perception experiment, described in Chapter **??**, and an ASR experiment, described in Chapter **??**, below.



(a) The sentence “A cramp is no small danger on a swim” with bus background noise at 60 dB. SNR for this utterance is +28 dB.



(b) The sentence “A cramp is no small danger on a swim” with bus background noise at 70 dB. SNR for this utterance is +19 dB.



(c) The sentence “A cramp is no small danger on a swim” with bus background noise at 80 dB. SNR for this utterance is +6 dB.

Figure 1.18: Spectrograms of the sentence “A cramp is no small danger on a swim” with bus background noise at 60 dB (1.18a), 70 dB(1.18a), and 80 dB (1.18a). The directional microphone filtered out most of the noise in the lower two noise level conditions.

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.

Braun, S., Neil, D., and Liu, S.-C. (2016). A Curriculum Learning Method for Improved Noise Robustness in Automatic Speech Recognition. *arXiv preprint arXiv:1606.06864*.

Brungart, D. S., Cord, M. T., Solomon, N. P., Dietrich-Burns, K., and Block, K. (2012). Evaluating the effects of hearing protection on speech production in noisy environments. In *The Listening Talker*.

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.

Gilbert, J. L., Tamati, T. N., and Pisoni, D. B. (2013). Development, Reliability, and Validity of PRESTO: A New High-Variability Sentence Recognition Test. *Journal of the American Academy of Audiology*, 24(1):26–36.

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.

Hu, Y. and Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, 49(7-8):588–601.

Huizing, E. H. (1960). Bone conduction-The influence of the middle ear. *Acta Oto-Laryngology*, (Suppl. 155):1–99.

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

Kabal, P. (2002). TSP speech database. Technical report.

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.

Lane, H. and Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research*, 14(4):677–709.

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.

Lombard, E. Le signe de l'élévation de la voix. *Annales des Maladies de L'Oreille et du Larynx. XXXVII*, (2):101–109.

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.

Stenfelt, S., Wild, T., Hato, N., and Goode, R. L. (2003). Factors contributing to bone conduction: The outer ear. *The Journal of the Acoustical Society of America*, 113(2):902–913.

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., Greenfield, E. C., and Kaufman, R. S. (1966). The occlusion of the external ear canal: Its effect upon bone conduction in cats. *Acta Oto-Laryngologica*, 61(sup213):80–104.

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.

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E.-D., and Schuller, B. (2017). Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments. *arXiv preprint arXiv:1705.10874*.