Introduction

It is self-evident that background noise makes it more difficult to hear what someone is saying. However, it has also been established that it is even more difficult for individuals to understand what someone is saying in an environment where there are multiple overlapping speakers, such as in a cafeteria or a pub - this is known as the 'Cocktail Party Problem'(Cherry, 1953). Interestingly, it has been found that it is easier to identify who is speaking and to understand what they are saying if you are familiar with the speaker, like an old friend, a spouse, or a parent (Johnsrude et.al 2013). There have been previous studies using already familiar voices, as well as lab-based artificial training on completely new voices, both of which yielded higher intelligibility than an unknown voice, and one study found that the most trained lab-based voice had the same magnitude of effect as the already familiar voice (Holmes et al 2021). Nevertheless, these voice training studies have occurred over the span of days, and there are less studies on how shorter periods of voice training might affect speech intelligibility, and how little training is necessary to yield a tangible result on listener performance. One previous study similarly looked at how effective different brief periods of voice training are, and their effect on intelligibility and familiarity, and found that ten minutes of training is enough to produce a difference in voice intelligibility (Holmes et al 2021).This study is similarly looking at brief periods of vocal training, but has an additional novelty in the inclusion of listening effort. Instead of solely focusing on the performance of the listener, this aims to determine if voice training alleviates some of the effort required in identifying and understanding voices when there is speech-based background noise. This would be useful for those with hearing loss, such as older adults, who report higher effort than those with better hearing in the same auditory situation (Stewart & Wingfield 2009). They report more fatigue afterwards, if we can reduce effort, we can reduce fatigue, and positively impact those with hearing loss.

Methodology

Listening effort may sound like a rather nebulous term, and has competing definitions, but for the purpose of this study, the following will be used: "The mental exertion required to attend to, and understand, an auditory message" (McGarrigle et al, 2014). In a nutshell, it is how hard the brain/mind must work to focus on and understand audible stimuli. In this study, listening effort is being measured in two ways: physiologically through pupillometry, but also through self-reported effort from participants. Pupillometry has several benefits: it is a time series measurement, and can show changes in effort over the duration of the task; it is an involuntary biological response from the participant; and pupil dilation is associated with effort, attention, and engagement (Winn et al 2018). This is joined by a simple question posed after each trial in the recognition and intelligibility sections, where participants must answer how much effort they exerted to answer the previous question, on a scale from 1-7 (1 being the least effort, 7 being the most effort).

The pupil dilation response to auditory cues can be affected by age, hearing loss, caffeine, pharmacological effects (ie. medication), neurological problems, and eye diseases (Winn et al 2018). Therefore, participants had to be from ages 18-35, have normal or corrected to normal vision, normal hearing, and no history of neurological, eye, or hearing issues. Finally, they were asked to not have consumed caffeine before the experiment, or wear makeup, since the makeup could interfere with the eye tracking device.. Since the voice samples were in British English, and it has been found that it is more difficult to understand and recognize voices in foreign accents and languages, participants were also required to be native British English speakers, or have moved to the UK before the age of 6. Participants were given informed consent, and gathered through either word of mouth, or online participant platforms, in this case SONA.

First, participants' hearing was tested using an audiogram. Participants were exposed to audial stimuli at different frequencies ranging from 250 to 8000 hertz. If the average volume across all frequencies was under 20 decibels, then it was considered normal hearing and the experiment proceeded as normal.

Then, the ‘Training’ section commenced, lasting roughly an hour. The participant was exposed to 10 trials of audio samples from three different voices, for a total of 30 trials. They were informed that they would be tested on whether they can match the names to the voices later, and were asked to pay attention to which name corresponded to which voice. Following this, they were trained on the three voices through 702 trials of audio samples. However, one voice had 468 trials, another 156, and the final one 78 - this created the three levels of familiarity: the “most” familiar Fam1, the medium familiar Fam2, and the least familiar Fam3. In this section, the participant was artificially trained on voices, which sets up the groundwork for later trials.

Next was the Recognition section. Here, the task was for the participant to decide whether the voice they were exposed to was familiar or not – ie. if it was one of the three voices they were trained on. If they said it was familiar, they then were asked to identify which of the three it was. For the recognition section, two new unfamiliar voices were introduced - Unfam1 and Unfam2. Throughout this section, there was background babble noise, which is essentially unintelligible overlap of speakers. Babble noise is used to replicate a noisy environment where there are many speakers, like a busy pub or restaurant. The level of babble noise in relation to the level of target voice, referred to as the Signal to Noise Ratio (SNR, units are Decibels - dB) was randomly either -3 (more noise) or 6 (less noise). For this section, pupillometry was used, as participants' eye movements were being tracked during the speech sample. After each question, the participant was asked to rate the effort they exerted during the trial from a scale of 1-7.

Note - There was a pool of the same five voices for all participants: EPS01, EPS02, EPS04, EPS05, and EPS06. However, the conditions each voice was assigned differed for each participant. In other words, for one participant EPS01 could be designated Fam1, while for another EPS01 is Unfam1 and EPS04 is Fam1, and so on.

Following this was the Adaptability section. The participant would be exposed to two overlapping sentences, one which always started with ‘Pat’ and another starting with ‘Bob’. The participant was told at the beginning of the Adaptability section to focus on the sentence with one of the two names for the entire section. They would then be asked to click the words from the sentence with the target name. The voices speaking these sentences are from the same pool of five from before. The purpose of this section was to try to even out performance between potential inputs. In other words, there are multiple possible combinations of voices, with the added element of one speaking the target sentence while the other speaks a masking sentence. The performance between all the potential combinations needed to be evened out, so that the following section could give data on effort, having accuracy more or less controlled. The evening out was done by increasing the target to masker ratio (TMR - the loudness of the target voice) if the participant made a mistake in identifying the words used in the target sentence, and decreasing TMR if they answered correctly twice in a row. The target accuracy for all conditions was 70.7%.

Finally, there was an Intelligibility Section. In terms of the task for the participant, it was nearly identical. The difference is that the voice inputs were already calibrated (no TMR changes), and the participant was asked to fill out the effort questions after each trial. This section also tracked the eye while the voice sample was being heard. Halfway through, the target name changed to the other option. The purpose of this section was to determine the listening effort required for different voices and levels of familiarity, yet at similar accuracy levels. This reaches the crux of the study: looking for effects on effort over effects on accuracy. To this end, the participants had to be briefed on what listening effort means: Not whether they think they did well, but rather how hard they felt they worked to come to their answer.

Conclusions

It was found that the accuracy of matching the voice to the name in the training section correlated with the voice's familiarity - ie. exposure to the voice. This supports the notion that short amounts of training are enough to see tangible differences in participants. It is also interesting to note that the variability of accuracy also seems to correlate with exposure - most participants were equally accurate with Fam1, less were as accurate with Fam2, and some were largely inaccurate for Fam3. This may be due to individuals finding the task easier or harder - differences in innate abilities, which become apparent when there is less exposure to a voice. Some were completely accurate in all three, while others were not. An analysis was also done to see whether the actual voice they were exposed to, ie EPS01, affected the accuracy - this was statistically insignificant, meaning that the aforementioned conclusions were reliably unaffected by which voice was familiarized.

Figure 1

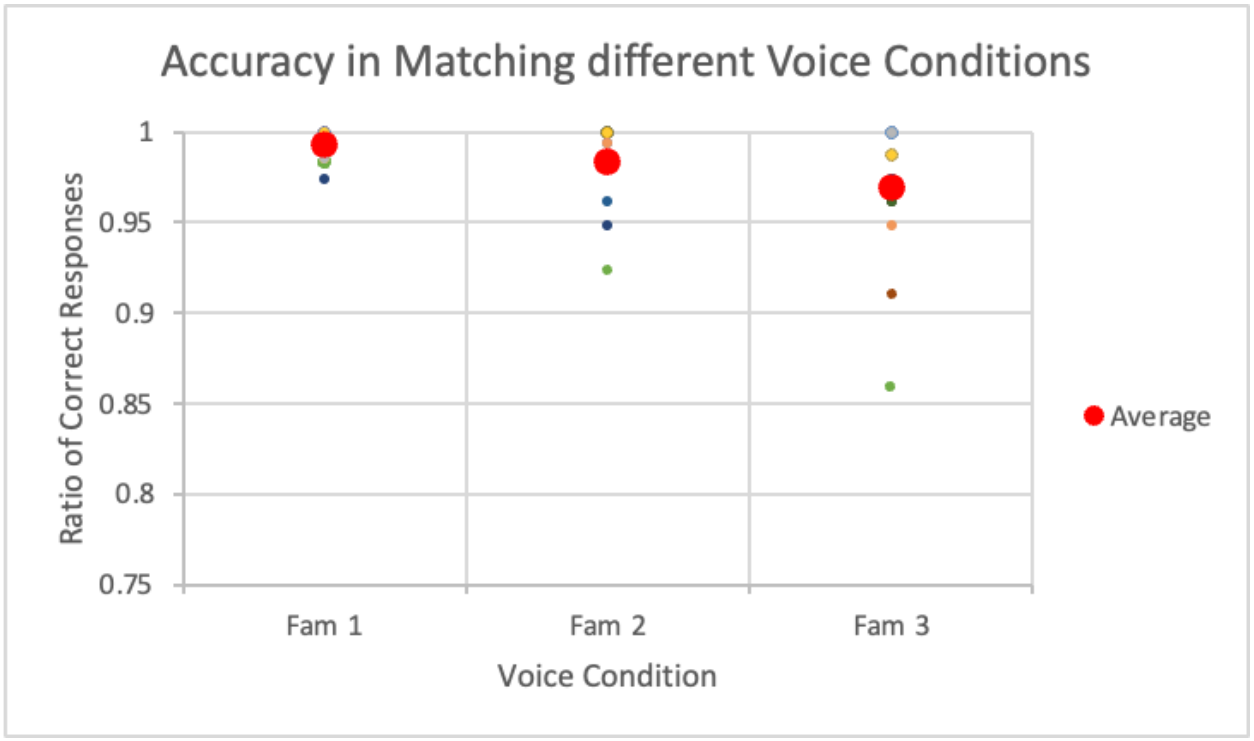


Figure 2

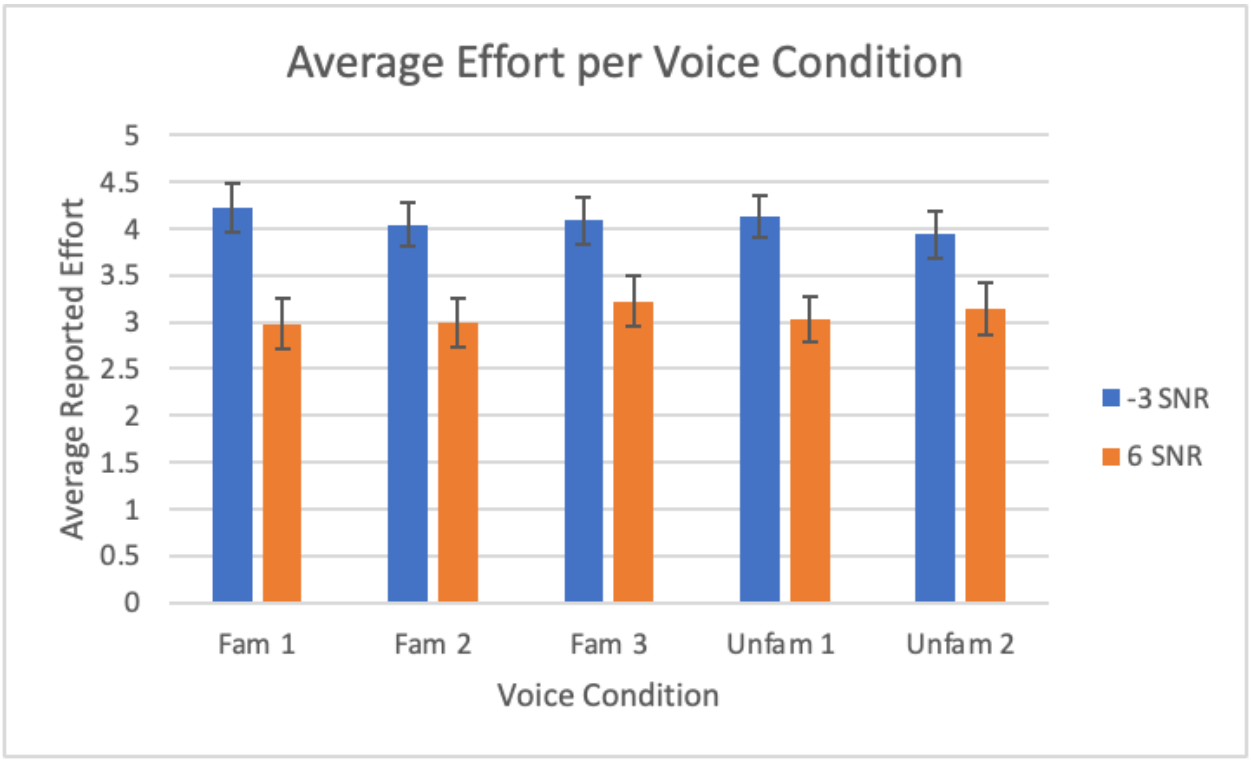
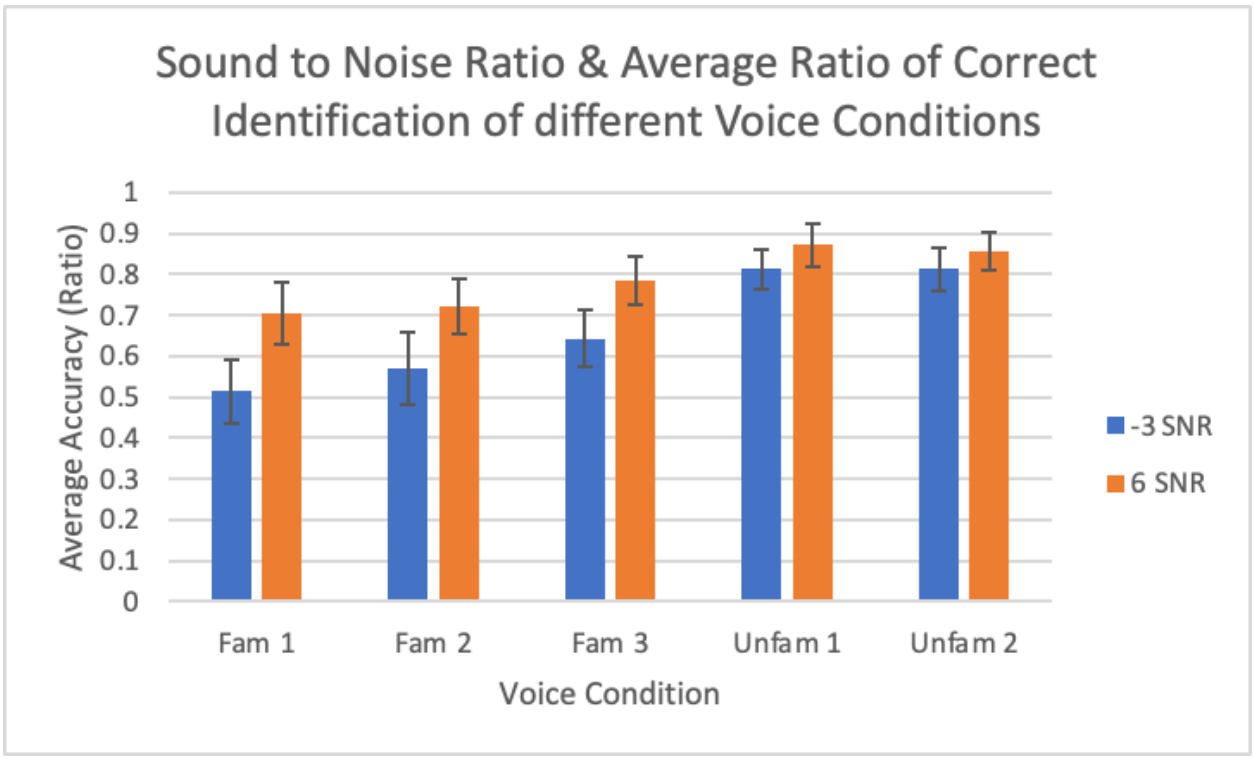


Figure 3

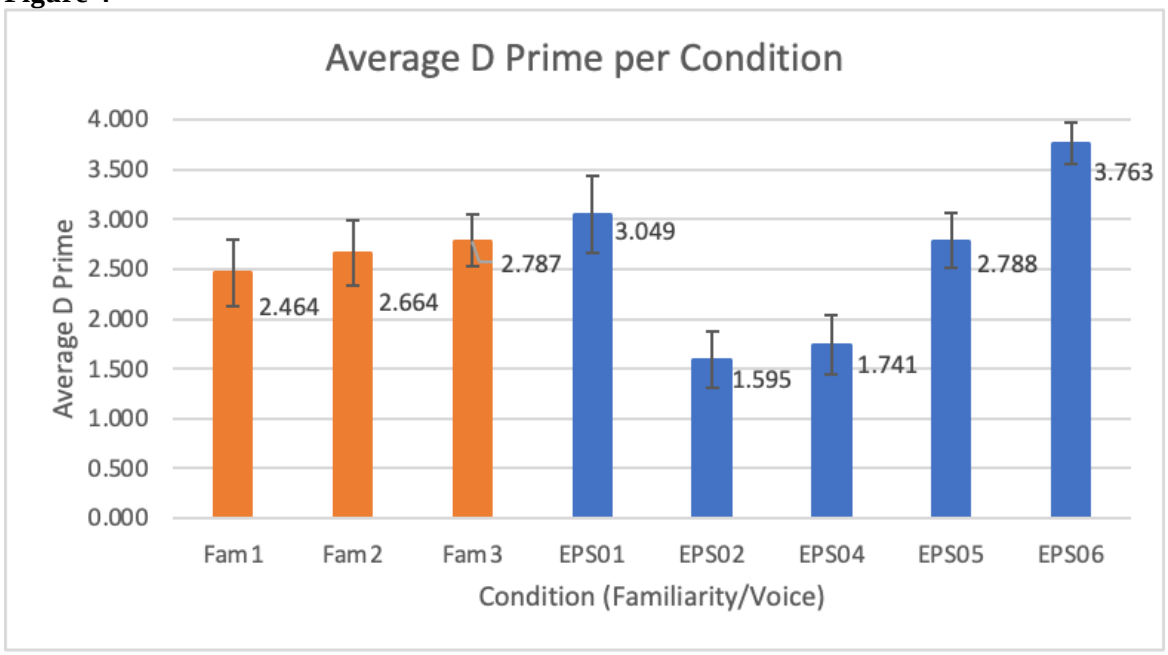


In line with previous studies, when the target voice was louder, recognition accuracy was higher. And when babble was higher, accuracy was lower. This can be seen in Figure 3 where there is a statistically significant difference in average accuracy in the first three familiarity conditions. One possible explanation why this trend does not continue for the unfamiliar voices is that to get a 'hit' for familiar voices, the exact name had to be chosen, while for unfamiliar voices, the participant only had to say that it was unfamiliar.

The new insight is the observation regarding effort. Greater effort is correlated with lower accuracy, which is unsurprising. However, for every condition there was a statistically significant difference in effort levels between SNR levels. There is a clearer relationship between SNR volume and effort than there is between accuracy and volume. The familiarity of the voice does not seem to matter for effort levels in the recognition task, which may point to a decoupling between accuracy and effort, or perhaps that there is an explanation in the conduct of the experiment, such as the aforementioned added step in counting accuracy for familiar as opposed to unfamiliar voices. Regardless, the main takeaway is that the level of the target sound, although not necessarily as effective for improving accuracy, can help alleviate the work required to do a task, and leave listeners less fatigued.

A potential conclusion from the similar results from all three familiarity conditions is that exposure to a voice does not noticeably impact accuracy or effort after a certain degree of exposure is already achieved.

Figure 4



The above graph illustrates the participants' overall ability to discriminate between voices by finding the D prime for each condition. This was done both for familiarity conditions, and each actual voice, to see which is ultimately more important in participants' ability to discriminate. It seems that familiarity was not a factor in a participant's ability to discriminate between voices - the least familiar voice was in a similar range as the most familiar. However, when it comes to the actual voices themselves, there is a significant difference. EPS02 and EPS04 had much lower D' values than the rest - meaning that participants' ability to discriminate between these two voices from the rest was significantly worse.

This may be that the voices were similar, and participants confused them with each other. In the other extreme, EPS06 had a significantly higher D' than the rest, for some reason being easier to discriminate from the other voices than the rest. This may be due to accents or intonations, but finding the reason for this could be very useful to know which voices are easier or harder to discriminate, allowing people to take this into account when speaking to individuals with hearing loss, or when attempting vocal training.

A similar trend could be seen in accuracy, where participants were significantly less accurate for EPS02 and EPS04 - but this trend was absent in the effort reporting. The fact that there was indeed lack of recognition did not mean that participants found it difficult to respond to the question, further supporting the idea that listening effort can be very distinct from accuracy.

Following up on the voice analysis in the recognition section, the same was done for training and intelligibility, but there was no statistically significant difference in accuracy rates for either training or intelligibility. This leads to the conclusion that although listeners may confuse certain voices for other voices, they are nevertheless able to understand any of the confusing voices. An analysis of which actual voice participants were exposed to, instead of familiarity was done, to check whether the voice themselves had an impact on recognition or intelligibility, rather than the familiarity. However, the intelligibility difference was not statistically significant. This may be due to the voices being similar and participants confusing them with each other, or a difficulty in understanding the voices themselves potentially due to accents or other traits in their speech.

Reflection

Regarding research, I learned that it can be boring, especially at first. When you have insufficient data, there's not much to be done but run experiments, and keep up with all the housekeeping. It can become a lot of waiting until the experiment section is done, patience is necessary when undertaking research. Having mentioned housekeeping, organization is essential to well-run research. Files must be organized, data labeled and properly put away. This becomes invaluable when something goes wrong, and you need to find the root of the issue. Moreover, it is crucial for data privacy, which I did a crash course on prior to working as an assistant researcher.

Organization can be annoying, consistent small extra work - but in the long term saves heaps of energy and time. I hadn't yet been in a position where I was required to be so organized, but it was refreshing to know where everything was at all times, having things well put together. I generally tend to lean towards chaos rather than organization, so this experience helped me gain newfound appreciation for consistent housekeeping, and made me understand why it's such a core skill to have.

I also learned about the necessity of communication and clarity. With my supervisor, and with participants. I was less consistent and less timely with my responses at first, and quickly learned the importance of good communication with my supervisor for logistical reasons but also professional ones, out of respect and politeness. As for participants, I made a mistake early on where I didn't specify that participants must not have caffeine - from then on I would inform them of that specifically, since they usually only skim through the information sheets. Things need to be double checked, and ensured they're running well. When running the experiment, I would explain to the participants, and clarify any questions, and ensure they knew what was expected and desired of them for the purpose of the experiment, while making sure they were also okay.

A big learning curve for me over the placement was data processing and analysis. I never considered numbers and data my strong suit, but I had to engage with both at length during data analysis. One interesting thing I noticed about data analysis is how you have a trove of data that you become familiar with - like becoming familiar with a map. Also, I learned how it can be manipulated in many different ways, and one needs to be careful and double check, to make sure calculations are done well and in that the presentation of data is made clearly in good faith, and that the analysis is aware of its own shortcomings.

On the skills side of things, I grappled with excel - it was finicky and frustrating and difficult for me, especially at first. Then I got the hang of things, learned how to streamline, use macros, use shortcuts, and better understood how excel works with data and formulae and graphs. I can't say I enjoyed the excel experience, but I did learn a lot from it, and am proud of how I pushed through it despite it being far from a traditional strength of mine.

Lastly, while doing the actual experiments, I realized that the most enjoyable part for me was contacting people, and telling them about the experiment itself, answering any questions and so on. I rather disliked the parts where I had to sit in an empty room, processing numbers and doing housekeeping. It was essential, and I believe I did it well, but it was a breath of fresh air when another undergraduate assistant researcher came in, as we could then chat about the experiment or data or otherwise. In fact, my favourite parts of the placement were my discussions with my supervisor and the other undergraduate about the data, the methodology, the drawbacks, and implications of our results so far. In sum, I realized just how much I enjoy, but also need, interaction with others to work at my best.

References

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America, 25,* 975–979. https://doi.org/10.1121/1.1907229

Holmes, E., To, G., & Johnsrude, I. S. (2021). How Long Does It Take for a Voice to Become Familiar? Speech Intelligibility and Voice Recognition Are Differentially Sensitive to Voice Training. Psychological Science, 32(6), 903–915. https://doi.org/10.1177/0956797621991137

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a Cocktail Party: Voice Familiarity Aids Speech Perception in the Presence of a Competing Voice. Psychological Science, 24(10), 1995–2004. https://doi.org/10.1177/0956797613482467

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: what exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International journal of audiology*, *53*(7), 433–440. https://doi.org/10.3109/14992027.2014.890296

Stewart, R., & Wingfield, A. (2009). Hearing loss and cognitive effort in older adults' report accuracy for verbal materials. *Journal of the American Academy of Audiology*, *20*(2), 147–154. https://doi.org/10.3766/jaaa.20.2.7

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends in hearing*, *22*, 2331216518800869. https://doi.org/10.1177/2331216518800869