Perception-Production Relationships in Spoken English: Comparing the Performance of Mandarin English Bilinguals and WhisperAI

Erik Glesne

Department of Cognitive Science, UCSD

PI: Dr. Sarah Creel

June 10, 2025

Abstract

In language studies, perception production relationships refer to how spoken language relates to language comprehension, and vice versa. As someone becomes more fluent in a language, they tend to become better at both understanding and speaking that language. However, the order of this effect is the subject of ongoing research. Furthermore, whether this effect is generalizable to other accents or is specific to one's own accent is the subject of ongoing research. We sought to better understand this relationship by looking at how native Mandarin bilinguals use vowel differences to distinguish between similar sounding words. Specifically, we looked at this vowel difference within final consonant voicing contrasted minimal pairs. We also compared the usefulness of this cue to Whisper, a speech to text AI model. This helped us to understand whether humans and AI are picking up on similar cues when listening to ambiguous speech. We found that humans and Whisper did use vowel duration in different ways, with Whisper showing an especially dramatic effect of vowel duration on its accuracy in correctly identifying voiced final consonant contrasts. This finding helps us better understand the differences between human and AI processing of speech data, and identifying its limitations in dealing with accented speech.

Introduction

Accent perception has social, psychological, and educational implications. Improving how well other people understand you is a common goal of those who want to become more fluent in a language, and the difficulty of changing one's native accent is a common source of frustration. In language research, we often hear the idea of a "critical period" of development, after which certain sounds from outside the speaker's native language are no longer discernible(Werker & Tees,1984). While this theory remains debated, it nonetheless informs theories about our understanding of how perception affects production and vice versa.

However, when examining accentedness through a more objective lens, it is important to keep in mind that there is not a "gold standard" English accent, even within the United States. The goal of "sounding like a native speaker" is therefore imprecise and perhaps misguided. This is especially true given that some research has suggested that native mandarin speakers show an increased ability to understand other native Mandarin speakers compared to native English speakers. The same study also suggested native Mandarin speakers in Beijing were more able to understand English produced by other native Mandarin speakers than native English speakers. (Xie, Fowler, 2013). This research suggests that accent effects show some specificity, and that being a native English speaker doesn't increase accent perception across the board. This also suggests that native Mandarin speakers are picking up on cues that native English speakers aren't, and perhaps native English speakers can also learn these cues.

The idea that by speaking with a certain accent, you become better at understanding that accent is often referred to as the Self-Specialization hypothesis. This hypothesis posits that

speaking with an accent creates a sort of positive feedback loop, where hearing one's own accent causes them to teach themselves to recognize it. This has important implications for language learning and accent perception. It suggests that exposure to a certain accent is necessary for understanding that accent better, which could inform how language is taught in an educational setting.

Minimal pairs are words that differ from each other in only one sound. Languages differ in terms of what sounds constitute a meaningful distinction, and this leads to difficulty for language learners, both in producing and perceiving these differences. For example, English uses final consonant voicing as a meaningful cue for what word something is. Words like "back" and "bag", an example of one of these minimal pairs, only differ in whether their final consonant is voiced. Certain other languages, such as Mandarin, do not have a final voicing distinction, which makes it difficult for native Mandarin speakers to reliably produce, and possibly to perceive, this difference. Because minimal pairs can exhibit differences in pronunciation that simply don't exist as distinctions in other languages, they're helpful in understanding how perception of another language develops over time. Previous studies, such as the Xie and Fowler study mentioned above, have also studied final consonant voicing minimal pairs, and their findings suggest that speakers may be more accurate on voiceless versus voiced versions of each pair. This provides a prediction we can compare our results to.

Another increasingly important issue in the environment of speech comprehension is the increasing prevalence of speech to text models. As these models become more and more widespread, increasing their performance on speakers from different accent backgrounds becomes an issue for technology accessibility. Furthermore, examining the limitations of these text to speech models can help inform our understanding of how these models learn language. Studies such as those by Georgiou(2023) have also tried examining how these models perform on accented speech.

Whisper AI is a speech to text software created by Open AI trained on an extremely large corpus of data. The majority of this data is from native English speakers. This likely leads to differences in the model's ability to interpret speech by non native English speakers. In our study, we sought to compare the performance of Whisper to Native Mandarin bilinguals in order to better understand what acoustic features are informative for differentiating similar sounding words.

Our study focuses on acoustic analyses of final consonant voicing contrast produced by Native Mandarin speakers. We sought to determine how meaningful of a cue vowel length is in identifying whether a word has a voiced or voiceless final consonant, for both native Mandarin Bilinguals and Whisper. In order to examine this cue in a naturalistic setting, we decided to use the natural variability we see in speaker's production of this cue as a predictor of accuracy, instead of experimentally controlling it

Results of our data analysis indicate that Humans and Whisper do in fact use vowel duration cues in different ways, with vowel duration being particularly significant in Whisper's ability to accurately identify voiced final consonant words.

Methods

Data collection procedure

Participants.

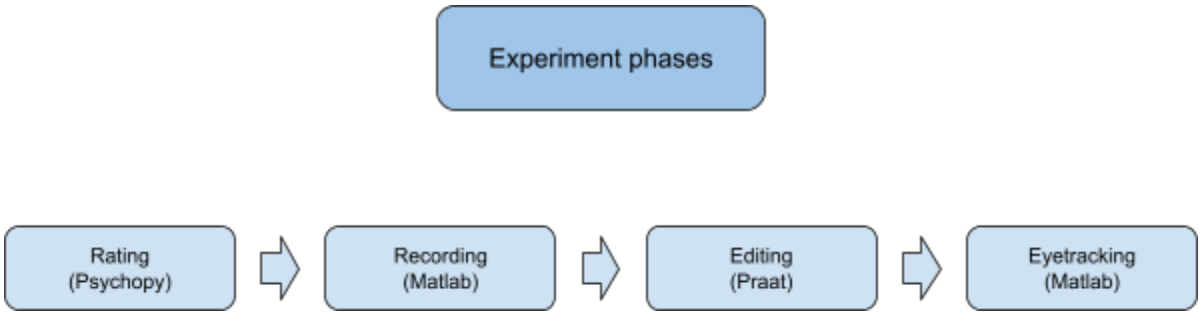
Data collection for our study is on UCSD undergraduate students. Since the research is on human subjects, the study required IRB approval, and all members of the lab completed the CITI training course for research on human subjects. All subjects signed a consent form to participate in the study. Participation is incentivised either with credit through UCSD's SONA system, or through payment with an amazon gift card. The study was advertised with flyers, which were also reviewed and approved by the IRB. Research was conducted on bilingual speakers of Mandarin and English. Participants were screened for eligibility, and excluded if they spoke another language, including Cantonese. Two participants that met eligibility criteria came into the lab at the same time to perform the experiment.

Procedure. The data collection for the initial experiment consisted of 4 phases: rating, recording, editing, and eyetracking. For the rating phase of the experiment, participants were shown 48 pictures paired with 48 words. Participants were asked to rate on a scale of 1-5 their familiarity with using a given word to describe a displayed picture. This portion of the experiment was done in Psychopy. Next, during the recording phase, participants were shown the pictures from the first phase of the experiment and recordings were taken of each participant saying all 48 words. During this part of the experiment, both members of the pair are present in the recording room in order to counterbalance the effect of hearing oneself produce a word, so for every time a participant says a word, they also hear the other participant say the same word. Next, during the editing phase, participants are asked to return to their original, separate rooms, and they are given non-linguistic tasks such as coloring sheets to pass the time. The audio recordings are then trimmed so that only the word is present in the audio file.

For the eyetracking phase, each participant is brought into a room separately that has a monitor and an eyetracking computer. For each trial, four pictures appear on the screen, and a recording of one of the words is played into their headphones. Two of the pictures represent minimal pairs, which are words that differ in just one sound. The other two pictures on the screen are distractors. The participant is asked to click on the word they heard, and their accuracy, as

Figure 1

Experiment phases



well as their eye movements throughout the trial, are both recorded. The participant hears both themselves and the other participant say all of the words. Whether they heard themselves or the other speaker first is counterbalanced across pairs. Both the recording and eyetracking phases of the experiment are done using a matlab script. The editing is done using a Praat script.

Data collection according to the above protocol was done on 24 pairs of native Mandarin speakers and was finished in 2024. During this time, we also collected data on a second "solo" version of the study, in which the number of words and overall procedure were the same except that the participants(n=48) only heard their own voices and not the voice of another participant. The following acoustic analyses were performed on both sets of these data, giving us a larger corpus of audio and accuracy data to work with. The 48 words we tested consisted of 24 minimal pairs. Of these pairs, 12 differed by final consonant voicing, meaning the word differed only in whether the final consonant was voiced. For brevity's sake, I'll use "voicing pairs" to refer to minimal pairs of words differing by final consonant voicing, and "voiced" and "voiceless" to refer to the voiced and unvoiced final consonant versions of these words, respectively.

Acoustic analyses.

To analyze the data, we initially fed all of our audio data with transcriptions to two forced aligners, Montreal Forced Aligner(MFA)(McAuliffe et al, 2017), and Charsiu(Zhu et al, 2022), in order to automatically create boundaries around the segments of each word. These forced aligners output TextGrid files in Praat, a software commonly used in linguistics research. These forced aligners are not perfect, so although they speed up the process, the boundaries still require manual editing. This editing was done by myself and another RA, Samantha Roxas. To assist us in the editing, I modified a script originally written by Will Styler, who is my secondary advisor on this project. This script pulls up the audio files and corresponding Praat textgrid objects in a folder, and allows us to speed up the process as we don't need to manually open and save each pair of audio and textgrid files. The script also allows us to edit one word at a time, so we are able to be more consistent with how we segment each word between different

participants. A subset of words was annotated twice, once by each of us, which allowed us to test our interrater reliability. Once this was done, we used a python script to analyze the results of our editing.

Interrater reliability measurement

After data was collected, we used a jupyter notebook to run the wav files and text files through Charsiu. A script made text files to label each audio file automatically, and adjustments had to be done to ensure the forced aligners were given the right words. We also needed to make sure there was an error exception so an error on a single single file wouldn't derail the entire script. We also ran the wav files and corresponding lab files(MFA uses lab files as transcriptions) through MFA using the command line. Once we had the files, we did some preliminary analysis to determine which script seemed to be performing better. Neither one appeared to make errors at a rate vastly larger than the other, but we decided the Charsiu ones were a bit closer to our desired outputs. We made a folder of the files and textgrids Charsiu created, and used a script to create a list of the wav files that Charsiu was unable to create an output for. Most of these were silence or recordings where the participant said a word other than the target. Then, Sam and I divided the list of words into two lists. The lists were divided so that each person was working on both members of the minimal pair for each pair, for example, I edited the boundaries for voicing pairs, like "bag-back," while Sam edited boundaries for vowel pairs, like "pen-pan." I then used the modified Praat script so that we could go through all of the files one word at a time. We were careful to not overwrite the original textgrid files. It was also helpful to move the files from the textgrid output folder to a saved textgrid folder, as running the script with the same word again would overwrite files that happened to be left in the folder.

The editing of the Charsiu forced alignments mostly involved slight adjustments to boundaries to ensure they lined up with the spectral indicators of different sounds. For example, sometimes we had to move the boundary in order to ensure it didn't cut off the pitch pulses and continuous formants that indicate a vowel. In addition to these more minor adjustments, sometimes Charsiu would "cut up" a word, for example showing the opening consonant repeating, followed by silence, followed by the entire word again. Editing these files involved removing extra intervals. Finally, because the Charsiu forced alignments didn't separate the release burst of final consonants from the closures, we added "h" to the final consonants to indicate final bursts after the closure. This was only relevant for words ending in plosives.

In order to check for interrater reliability, I made a script that divided all of the files into separate folders depending on whether they were coded by me or my research colleague, Sam. I then copied these files into an excel sheet and created a random value column. We could then recode the first 10% of these files, which were a random subset of each of our original encodings. Once these transcriptions were done, I used a Jupyter notebook to move the files into

separate folders to make them easier to work with. I made a folder for my original transcripts, Sam's original transcripts, my IRR output, and Sam's IRR output. Once this was finished, I ran another script by Will Styler on each of the four folders, which outputs a tsv with information about the labeled start and end times of each sound in a word. This information was read from our adjustments of the forced alignments. I then read the text output of the tsv of these into python using pandas. There were some randomly distributed parsing errors where it seems some tabs were automatically added in, so I used the line information given in the error to fix these by hand. This issue turned out to be caused by trailing spaces.

These tables contained data about the textgrid, including the start and end times of each tier. I added information about who coded which one. I also used a different script that included the word transcription for each file. On words where the speaker said something other than the target word, for example saying "horse" instead of "ride," we included "wrong word" in the transcription of the word. This allowed me to drop words containing "wrong", which would be words we would want to exclude from analysis. After I dropped these words from the list. I combined the dfs using the merge function in pandas, doing a left merge based on the filename and label, allowing me to compare how each person coded the values differently.

In order to get accurate results, some of the data had to be recoded. For example, words with two duplicate labels, such as cake (transcribed in ARPABET as K EI K) had to be edited so that each label was unique. This was so that one rater's first label wouldn't be compared to the other rater's second label, causing an inaccurate increase in discrepancy.

In order to measure interrater reliability, the subsets of recoded files were merged with the original Charsiu transcriptions, and additional start and end offset columns were added that were the difference in seconds between the two coder's labels. The proxy we used for interrater reliability was the proportion of labels that were within 20 milliseconds of each other. These measures were also done between this same subset of files and Charsiu as well as MFA, to ensure a fair comparison. Since Charsiu didn't include details about the burst following a closure, we treated the burst and closure as one interval during comparison of both raters to each other and to Charsiu in order to not artificially inflate the accuracy of the two raters compared to each other. Silences at the beginning and ends of words were also removed as they are not relevant to our current analysis.

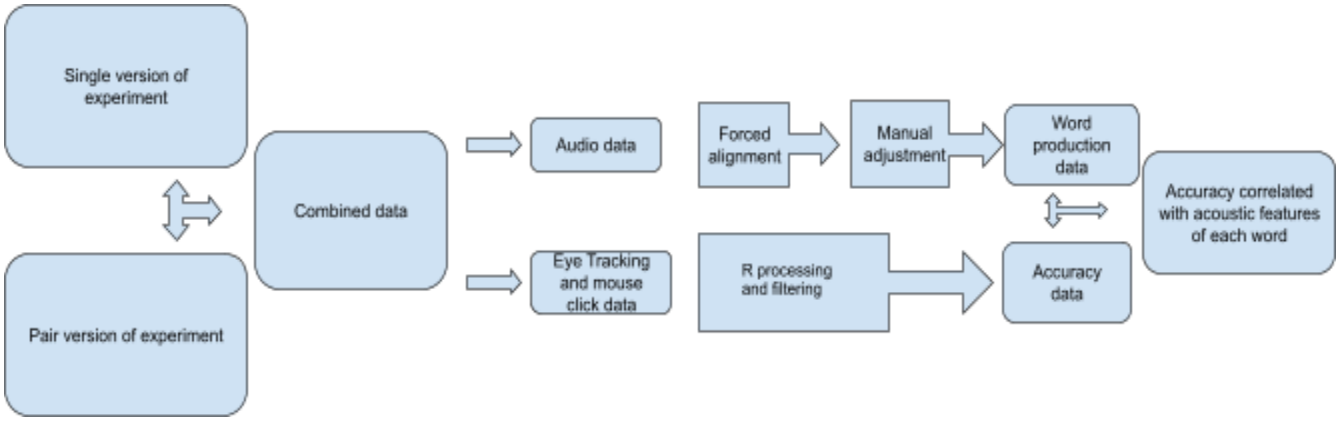
Using Accuracy Data

While accuracy data was originally pulled straight from the eyetracking file in the experiment, I later updated my script to use the preexisting R pipeline made by my advisor, Dr. Sarah Creel. Using the accuracy data output by this script in CSV format allowed me to ensure that the participants we were analyzing met our criteria for participating in the study, and that they met the cutoff criteria in terms of the number of usable audio files. This was because occasionally the participants could not remember enough of the words and therefore accuracy on identifying their own and other participants' productions of words could not be reliably measured. The cut off criteria for the number of usable word pairs was determined before results

were explored. There were also occasional issues with the audio recording or the audio editing process that resulted in some unusable files. Once I had the accuracy data, I merged it with the outputs of our manual forced alignment edits in python. This allowed for analysis of how

Figure 2

Data Analysis Pipeline



acoustic features correlated with accuracy. In order to get more information about the acoustic features, I used the start and end time of each boundary to get the duration of each label. One of the features we focused on was how vowel length correlated with response accuracy, specifically for minimal pairs that differed only by the voicing of the final consonant.

In order to examine how vowel length affects perception, accuracy data and acoustic features were used from two versions of the experiment. Once data from both versions of the experiment were combined, basic statistical analysis measures were used to ensure that the vowel length did in fact differ between the voiced and voiceless versions of each word pair. For the subset of the words that differed by final consonant voicing, we wanted to see how informative this cue was for both humans and the Whisper AI model.

Whisper AI

Whisper is a speech-to-text model created by OpenAI. One of our research goals was to better understand how well speech-to-text models largely trained on Native English speech perform on English speech produced by Native Mandarin speakers. Furthermore, we wanted to see if Whisper uses similar or different cues to humans when trying to differentiate between similar sounding words. In order to investigate this relationship, we fed the audio stimuli from the experiments into the Whisper tiny, small, base, medium, and large models.

The first step in analysis was preprocessing Whisper data, which involved cleaning up minor inconsistencies in how each word was coded. For example, words could sometimes include exclamation marks or other characters, or Whisper sometimes guessed a homophone of the word. Once the Whisper data was cleaned, we were able to analyze how acoustic features of the audio correlated with Whisper's performance on it. This was also done in python, by merging the Whisper accuracy data with the acoustic features data by filename. In order to compare

human accuracy, Whisper accuracy, and acoustic features of each word, the dataframe including the acoustic features and human accuracy data was also merged with the Whisper data by filename.

While human accuracy could simply be operationalized as whether they picked the target word, Whisper's accuracy had to account for the fact that it was not given alternatives to choose from. For example, since homoforms are by definition acoustically the same as each other, we decided to treat Whisper picking a homoform as Whisper getting that trial correct.

Data analysis process

In order to analyze the data, irrelevant information such as silences had to be removed. Furthermore, features such as vowels, closures, and bursts weren't prelabeled, so I made a script to identify which labels were vowels based on a dictionary, and also identified closures and burst durations. In order to identify what features were useful for discriminating between similar sounding words, we did EDA using features like vowel length and final burst duration. Because of our EDA and the known phenomena of vowel duration systematically varying depending on final consonant voicing, we decided to further investigate the relationship between voicing, vowel duration, and accuracy.

In order to correct for the significant amount of variation in vowel length between minimal pairs, vowel durations were z-scored within word pairs. This allowed for analysis of vowel duration differences between minimal pairs, without letting the individual differences in duration between word pairs add noise to the data. We also corrected vowel duration by using vowel nucleus duration, which included the duration of the /ɪ/ phoneme as part of the vowel length. Once data was cleaned and additional information was processed in python, the files including vowel duration for the voicing pairs of words were exported via csv into R for further plotting and analysis. In R, logistic regression models were run on the data. Initially, some of the models used adjusted vowel nucleus duration, but these were later updated to use the aforementioned z-scored within word pairs vowel duration as one of the independent variables. For both Whisper and Human data, separate logistic regression models were run to see the relationship between vowel duration and accuracy for both voiced and voiceless words. These models also accounted for the random effects of different speakers and different word pairs.

Initial tests sought to determine whether vowel length was a predictor of which word within a pair was picked, regardless of what the target word was. Vowel length was a significant predictor of whether the voiced or voiceless version of a word was picked, but this effect was potentially misleading given that words with longer vowels are naturally more likely to be the voiced version of the word. For this reason, we decided it made more sense to look at accuracy in identifying the correct word within a pair.

More complicated linear regression models were also run in r, which the interaction between z scored vowel duration and voicingness (coded as a binary variable for -0.5 as voiceless and +0.5 as voiced) as the independent variable, while also accounting for the random variation of this interaction between word pairs and speakers. These linear regression models were run on both the human accuracy and Whisper accuracy data.

Results

Interrater reliability measurements

Interrater reliability measurements conducted on the random 10% of files we used for IRR found that 78.66% of start boundaries were within 20 ms of each other. The same measurements comparing the same subset of data to the Charsiu forced alignments found that 67.59% of the start boundaries were within 20 ms of each other. Using the same criteria to measure end boundaries found that 74.74% of end boundaries for our two raters and 70.26% of end boundaries for Charsiu met this criteria. This suggests that our raters were more similar to each other than either of them was to Charsiu, with the effect showing up stronger for start boundaries than for end boundaries.

Comparison of Charsiu and MFA

Exploratory measures were also done to compare the performance of Charsiu with MFA. For these analyses, it is important to keep in mind that both transcribers started with forced alignment data output by Charsiu, so it should be thought of as exploratory analysis, rather than a true evaluation of their relative performance.

We found that, using the same 20 ms threshold cutoff as above, the start labels for Charsiu and MFA had a 69.73% concordance with each other. Humans had a 67.42% concordance with MFA.

Whisper text outputs

In general, Whisper seemed to prefer the voiceless version of the words, with its accuracy being much higher for the voiceless versions than for the voiced. For the voiceless words, its accuracy was around 83%, while it was around 43% for the voiced words. However, its accuracy to the voiceless words was very sensitive to vowel duration, which will be discussed more later.

Vowel duration overall effect

Initial analyses validated that vowel length was reliably longer preceding a voiced consonant than a voiceless one. This was consistent with our expectations. Statistical tests in R run using vowel duration as an independent variable found that for humans and the large Whisper model, vowel length was positively correlated with the correct identification of a word ending in a voiced consonant (human: $\beta = 0.15292$, $p = 0.0317$, Whisper:$\beta = 0.8732$, $p = 4.89e\text{-}12$). For human listeners, vowel length was negatively correlated with accuracy for words ending in a voiceless consonant ($\beta = -0.46004$, $p = 3.30e\text{-}08$), but this effect was not statistically significant for the Whisper model ($\beta = -0.1836$, $p = 0.19$).

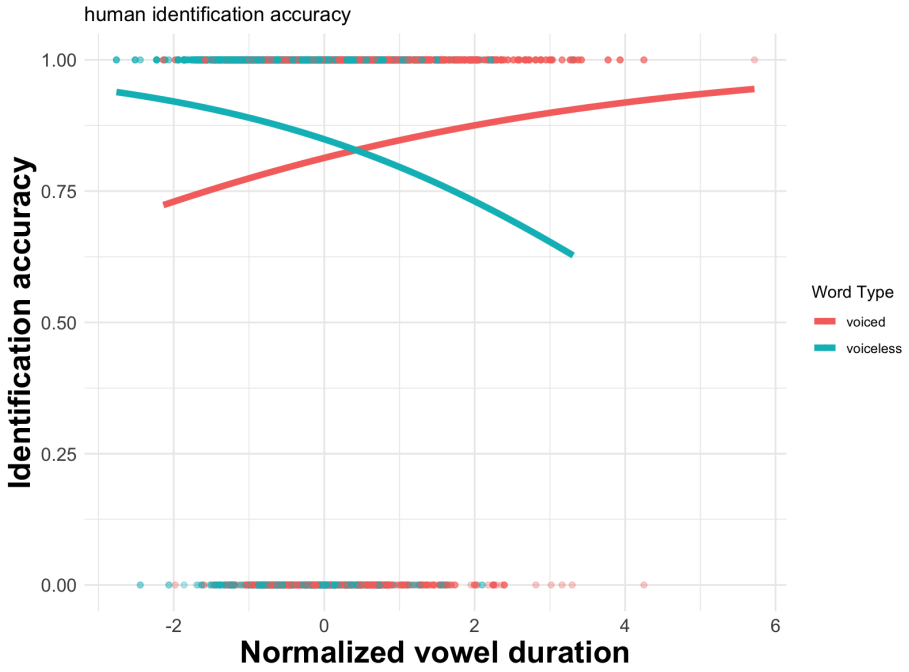
Different vowel effects / different models

Initially, my accuracy comparisons used the tiny Whisper model, which showed much lower accuracy. However, since accuracy monotonically increased with model size, I decided to use the large Whisper model to give it the fairest comparison. As the Whisper model size increased, Whisper's biases became less severe, but even the Large Whisper model's biases were quite pronounced.

Interaction effects

Figure 3a

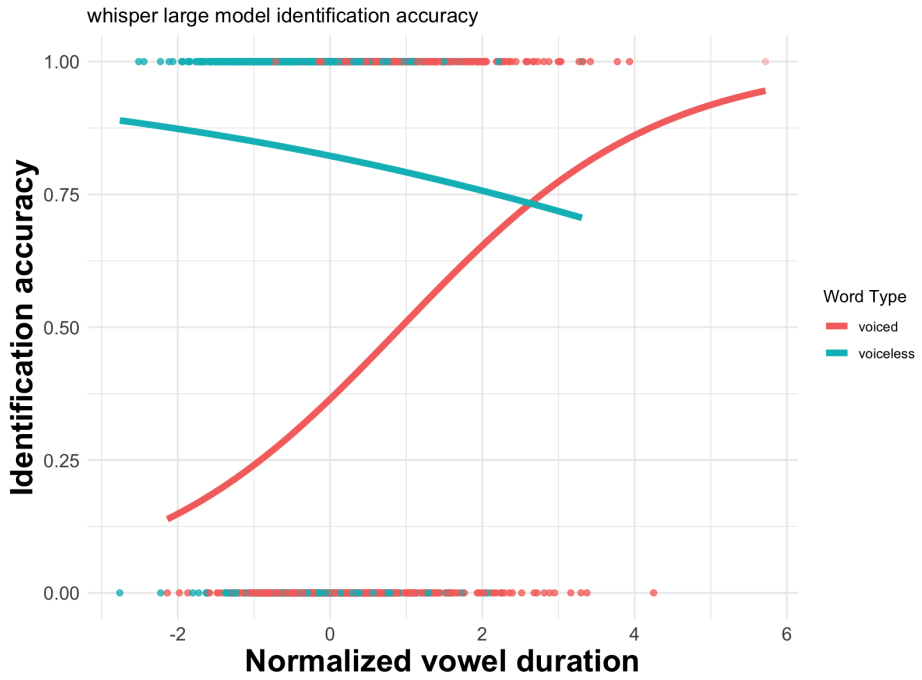
Human Accuracy Interaction Plot



Interaction models were also run on the human and Whisper large model data, as described above. These models found a significant interaction effect for humans (β= 0.76562, p =0.0012). However, for the Whisper Large model, all three effects were significant (normalized vowel duration: β=0.426, p= 0.000756, final consonant voicing: β= -3.3416, p = 6.11e-05 , interaction: β = 1.0044, p = 0.000498).

Figure 3b

Whisper Large Model Interaction Plot



Discussion

When comparing Whisper and human accuracy, it is important to keep in mind that while humans were presented with the target word and three alternatives, Whisper was not given information as to the possible words it was picking between, meaning their tasks were slightly different. However, the most common error for whisper was picking the minimal pair, so while the tasks are not exactly equivalent, they are similar enough to warrant comparison.

Charsiu/ MFA performance comparison

When performing analysis on the Charsiu data, I had to redo my data pipeline a few times to avoid biasing the results by not treating the data the exact same way every time. For example, I found out that one of the dataframes I was merging had information about another Praat tier, which made the accuracy look a lot lower than it was. This was because it included extra rows in the dataframe, effectively increasing the denominator of my measurement, which was the proportion of rows where the start offset were within 20 milliseconds of each other.

Overall, our exploratory analysis of MFA and Charsiu's performance on the data showed that they were quite similar. Remarkably, even without accounting for the fact that Charsiu was used as the starting point for our manual adjustments, the concordance of our final outputs with MFA and Charsiu were very similar, with 67.59% of Charsiu vs 67.42% of MFA start boundaries

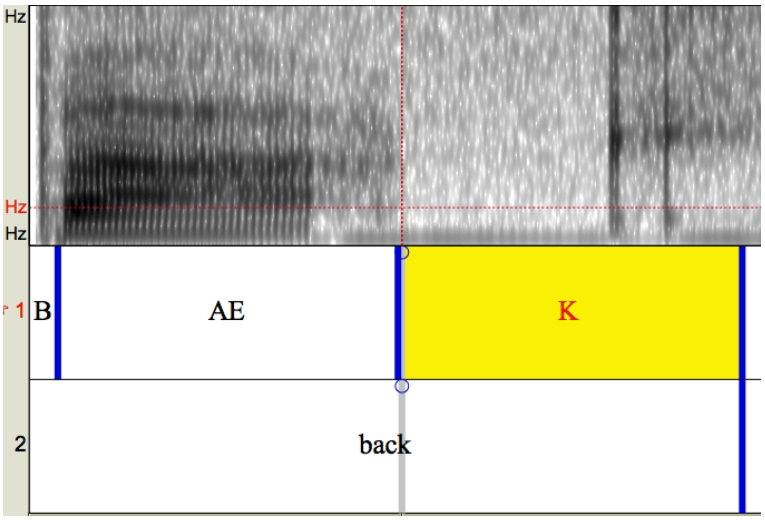
meeting our criteria. However, our raters showed a higher concordance with each other than either of them showed with either model. This suggests that the manual adjustments we did were still worthwhile. However, we were slightly surprised that the difference between our interrater reliability and our concordance with either forced alignment outputs was not greater.

There's many possible reasons for this. For example, some measures, such as burst duration, are hard to get a lot of concordance on. What some people might consider part of the release burst, another might consider aspiration after the burst. Certain qualities, such as formant cues, can lead someone to make one decision over another, but people may differ in terms of how they interpret these cues.

Our sample also included several strange, idiosyncratic differences in terms of vowel performance (see figure 4). Usually, a vowel cutoff before a closure is more defined, with both pitch pulses and formants indicating a movement into a closure. The below file almost sounded like a Whisper after the vowel, and my secondary advisor, Will Styler, suggested it might be a feature of the room, such as an echo but also thought it looked strange. We seemed to have a few of these files with this same problem, and features such as these could have added noise to the concordance between our two raters.

Figure 4

Example of strange vowel in "back"



Interaction effect models

Statistical analysis revealed that Native Mandarin speakers are more accurate at correctly identifying a member of a voicing minimal pair when the vowel preceding a voiced consonant is

longer, and when a vowel preceding a voiceless consonant is shorter. This is consistent with the vowel difference usually present in these words, and implies that these cues provide hints as to what the target word was. This is a significant finding in that it suggests that not only are second language speakers of English able to produce these secondary cues, they also may be using them to distinguish between similar sounding words. This supports the idea that as someone learns a language, they're learning to both produce and pick up on cues that aren't being explicitly taught to them.

When examining the interaction between voicingness and vowel duration, humans showed an interaction effect, but the effects of voicingness and vowel duration overall disappeared. Since our simpler models showed that vowel duration does have an effect depending on whether the word was the voiced or unvoiced version, this implies a crossover interaction where the slopes of the voiced and voiceless versions of words cancel each other out. This means that average accuracy didn't increase with duration or voicingness, but that duration had a different effect depending on the voicingness of the final consonant. These results are consistent with the idea that our sample of listeners are not significantly better at simply voiced or unvoiced words. It also suggests they do not simply perform better on this subset of words when the vowel is long or short, but instead that their performance on either the voiced or voiceless subset on each word depends on vowel duration. This interaction effect is also consistent with the idea that vowel duration is a meaningful cue for which member of a pair each word belonged to.

Another important consideration about the human data is that the average of both voicing and voiceless words is performing well above chance regardless of duration, which suggests that cues besides vowel duration are being used to distinguish between these two words. This suggests that even though final voicing minimal pairs are not present in mandarin, Mandarin speakers are still able to use a variety of cues to help disambiguate voicing minimal pairs. Other potential cues they could be using include glottalization, other vowel qualities such as formant space, or level of vocalization preceding the burst.

Whisper, on the other hand, did show a significant effect of both vowel duration and voicingness, as well as an interaction effect. This result illustrates that Whisper was simply more accurate on the voiceless version of the words. Certain word pairs, like "bag-back" showed that Whisper had extreme preferences, in this case for "back." Whisper also tended to be overall more accurate when the vowel duration was larger. However, there was still an interaction effect, which suggests that Whisper may also be using vowel length cues to differentiate between voicing pairs.

The large β coefficient for Whisper's interaction model suggests that the voiced version of words was especially sensitive to vowel duration for Whisper. A possible explanation of this effect is that due to Whisper's bias toward the voiceless version of each word pair, a more ambiguous word, such as one containing a short vowel, is likely to be counted as the voiceless version by default, and a larger vowel difference is needed to make up for this inherent bias.

While some of Whisper's biases can be explained by frequency of usage in colloquial language, there are some other possible explanations. Whisper's tiny model, especially, hallucinated quite often on our data, and some of the most common hallucinations were things like "bye", "we'll be back," and bizarrely, in some cases, "thanks for watching." Open AI says that Whisper is trained on a huge corpus of labeled audio data, but hasn't officially disclosed what that data is. Because of the nature of its hallucinations, it seems like its training data might involve a lot of youtube content.

Another limitation we noticed of using Whisper for this task is that we're using one word utterances, which are probably quite limited in its training data. In another experiment, this led to a lot of single word utterances being transcribed as profanity. This issue highlights the importance of context to these models, and enhanced ability to understand single word phrases could be a potential area of improvement for the future of these models.While having a massive corpus of prelabeled audio data makes YouTube an obvious choice for training a Text to Speech model like Whisper, it also may present problems with its generalizability to other context, and its frequent hallucinations on single word content is an example of this issue.

Future research could apply to more word pairs to get a better understanding of whether the large difference we found for Whisper is due to an inherent weakness in processing the voicing, (eg. Whisper is more likely to pick the voiceless version for all words because of something in its processing), or whether the subset of word pairs we used just happened to include a lot of Whispers 'favorite' words as the voiceless version. This latter explanation seems more likely, as two examples of words where Whisper strongly preferred the voiceless version were bag-back and peas-peace. In both cases, the voiceless version of this word is far more common in colloquial language, as well as in the sort of content we theorize Whisper was trained on, which likely contributed to Whisper's bias.

More analyses on the same data could help us better understand the nature of the perception/ production relationships we're looking at. For example, I'm currently working on using formant data on some of the other minimal pairs we worked with and seeing how that relates to accuracy for both humans and Whisper. This will be helpful in seeing if the differences in how humans and Whisper use acoustic cues are more widespread than just vowel duration.

Currently, I'm working on a similar data analysis process on data from an experiment with pairs of Mandarin-English bilinguals and Native English speakers. Comparing how acoustic measures such as vowel duration impact task performance for native English speakers will allow us to compare and contrast Whisper, native English speakers, and Mandarin bilinguals as perceivers of English. Research that examines whether English listeners perform more like Mandarin bilinguals or Whisper will help examine whether Whisper's biases seem to be due more to its training data bias, or whether it has more to do with the different ways AI versus Humans perceive language.

Limitations of using vowel duration as a predictor of accuracy

While using accuracy as a dependent variable turned out to be more fair than using whether the voiced or voiceless version of the word was picked, it doesn't entirely solve the problem of potential covariates. It is possible that, for example, people who produce voiced words with a longer vowel also produce other helpful cues more clearly, and people who produce voiced words with a short consonant don't produce other meaningful cues as reliably. If this were the case, it could cause our model to appear as though vowel duration is a meaningful predictor of accuracy for these people, when in reality people are picking up on the other cues that they're producing. This is another reason more acoustic analyses on this data could prove helpful.

Our ongoing research comparing Whisper's performance on native English speech will also help us to better understand the nature of Whisper's biases. Whisper's performance on a sample of people who presumably produce more cues that are reliably similar to Whisper's training data could perhaps show a more balanced effect of voicingness. Maybe Whisper's strong biases showed up in our case due to the absence of the cues it was used to.

The results of our analysis differ slightly from some earlier studies studying accent intelligibility benefits, including the Xie & Fowler (2013) paper mentioned in the introduction. Namely, we did not find that Native mandarin speakers are overall more accurate on voiceless consonants rather than voiced final consonants. In our case, the main effect of voicing was not significant. Interestingly, Whisper did show this effect, with its mean accuracy being much higher for voiceless final consonants. This provides a potential alternative explanation to our hypothesis that its relatively higher classification accuracy for the voiceless version of words was simply due to our sample of word pairs aligning with its training bias. Again, future research into how Whisper perceives native English speech could prove informative here. If Whisper performs more like a native English speaker did in their experiment, we would expect it to be far more accurate on voiced final consonants produced by native English speakers than for native Mandarin speakers. However, given the fact that our native Mandarin listeners did not perform significantly differently on voiced or voiceless final contrasts, this remains speculative.

Comparing and contrasting how humans and AI perceive speech also helps us understand the way in which our own processing of language differs from a pattern seeking machine, and may even have implications in the debate of Language domain specificity. In other words, if even the largest pattern recognition machines are using different cues than humans, it could provide more evidence for the domain specific understanding of language learning, which suggests that human brains are innately wired to understand language(Margolis & Laurence, 2023).

Finally, improving our understanding of how L2 speakers of English and AI differentially interpret language cues will help us to better understand our own perceptual processes. It also has implications for language education, as language teachers can emphasize focusing on word features that are known to help differentiate words. Furthermore, given the black box nature of AI models, comparisons to human interpretations of speech help give us more insight into how these models may actually be functioning. This is especially important for our sample of L2

English speakers, where the ability of models to understand them can become an accessibility issue. Knowledge of the cues technology like Whisper finds useful could prove helpful in troubleshooting issues, and more importantly, in making these technologies more robust and accessible.

Acknowledgements:

Thank you to my primary advisor, Professor Sarah Creel, for the opportunity to work in her lab, for all the support and feedback along the way, and for always helping steer me in the right direction.
Thank you to my secondary advisor, Professor Will Styler, for providing invaluable feedback on my data analysis process and for his help in decoding ambiguous spectrograms.
Thank you to my grad student advisor, Emma Miller, for her help and guidance, and her willingness to provide support throughout the course of this project.
Thank you to Professor Bradley Voytek, who provided valuable insights and for teaching me how to successfully deliver a presentation.
Finally, thank you to the rest of the Language Acquisition and Sound Recognition lab, especially Samantha Roxas, who worked closely with me in both data collection and analysis phases of the experiment.

References

Georgiou, G. P. (2023). Comparison of the prediction accuracy of machine learning algorithms in crosslinguistic vowel classification. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-42818-3

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced Aligner: Trainable Text-Speech alignment using Kaldi. *Interspeech 2022*. https://doi.org/10.21437/interspeech.2017-1386

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63. https://doi.org/10.1016/s0163-6383(84)80022-3

Xie, X., & Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics*, *41*(5), 369–378. https://doi.org/10.1016/j.wocn.2013.06.003

Zhu, J., Zhang, C., & Jurgens, D. (2022). Phone-to-Audio Alignment without Text: A Semi-Supervised Approach. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8167–8171. https://doi.org/10.1109/icassp43922.2022.9746112