# Native Mandarin Speakers Learning English Exhibit Enhanced Comprehension of Their Own Vocal Productions

**David Wu, Sarah Creel**

**Department of Cognitive Science, Language Acquisition and Sound Recognition Lab, University of California, San Diego, La Jolla, CA, USA 92093.**

***e-mail:** David Wu **d7wu@ucsd.edu**, Sarah Creel **screel@ucsd.edu**

**Data availability**

All data collected in this experiment was done at the Language Acquisition and Sound Recognition social science laboratory at the University of California, San Diego

**Acknowledgements**

Thanks: Reina Mizrahi, Sean Huang, Victor Ferreira, Bradley Voytek

**Author contributions**

Sarah Creel conceived of the experiment and developed and ran the analyses. David Wu conducted the experiment and ran the data analysis

**Competing interests** the authors declare no competing interests.

## Abstract

Second language acquisition literature has demonstrated that bilingual speaker's native language influences both their comprehension and their production accuracy of words in their target language. Dominant features of each person's native language such as phonetic and prosodic characteristics will carry over into their target language. Two bilingual language processing hypotheses have been proposed to explore native language influence in target language speech perception. The Interlanguage Speech Intelligibility Benefit (ISIB) account states that people with the same language background tend to understand each other more when they speak another language. On the other hand, our hypothesis, the Self Specialization hypothesis, states that because one's own speech is heard a plurality of the time, people will recognize their own speech idiosyncrasies and have enhanced understanding of their own vocal utterances. To further explore language perception in bilingual speakers and to reconcile between the ISIB and Self Specialization hypothesis, we recruited bilinguals with Mandarin-Chinese backgrounds learning English. They were tested on their ability to produce and comprehend speech in the target language. The participants listened to their own voice and another speaker of a similar native language background. We tested their accuracy in a visual world paradigm presented with their own speech and from another speaker, and tracked the eye movements of the participants. Our results indicated that regardless of a shared native Mandarin background, participants were able to understand themselves better than their peers, particularly more with vowel sounds than voicing sounds. The results show implications in language learning where listening to your own productions influences your comprehension of every language you speak. In this light, it would be optimal to gain more authentic input during language learning to minimize the influence of self-generated speech production biases.

## Introduction

Some of the oldest concepts of language learning such as the Wernicke-Geschwind model connect Broca's and Wernicke's area and detail the relationship between speech and comprehension (Geschwind). As the field improves and finer research is conducted, more is understood about the relationship between language perception and production. The inputs of both our own and other people's speech have been known to contribute to the production of our language.

Multilingualism has been a common point of interest for linguists and cognitive scientists. From developmental research (Cooper, Fecher, and Johnson 2018), to later acquisition (Hayes-Harb 2010) and accent perception (Dokova 2022). However, only specific researchers have been tackling the specific distinction of listener benefits. When learning a language, who do people understand best? Previous studies have suggested differentiated accounts towards language perception in bilingual speakers. Bent and Bradlow (2003) proposed the Interlanguage Speech Intelligibility Benefit (ISIB) account, which suggested that when speakers of the same native language background spoke a foreign language, they were able to understand each other better than other accents. This account was supported by Hayes and Harb (2008) where they found native Mandarin speakers also showed higher levels of comprehension of their peers and native speakers of English. Native Mandarin speakers particularly excelled in comprehending other native Mandarin speakers compared to native English speakers when they listened to Mandarin accented speech. Eger and Reinisch (2018) also found that German speakers showed better levels of comprehension of their peers when speaking a different language, irrespective of their skill levels in production. They also concluded that people tend to adapt to their native language (L1) accents when they speak their target language (L2).

**The Current Study**

As shown, previous works in the field of speech recognition had participants tested on other people's productions. They have not shown native Mandarin listeners raw vocal productions of *themselves*. Some other articles have explored this phenomenon but with more distorted listening setups. Cheung and Babel disguised speakers' voices when testing Cantonese speakers' perception of themselves compared to other people with the same native language (2022). The audio manipulation of Cheung and Babel does not fully suggest that people are better or worse at comprehending themselves (2022). Schuerman et al. tested Dutch speakers with noise vocoded speech and compared people's vocal productions to those of the "average" speaker (2014). Noise vocoded speech is meant to mimic speech of cochlear implants. This causes distortion and can lead to more variability in the data. Schuerman found that the "average" speaker was more comprehensible compared to the speaker themselves when hearing productions through noise vocoded speech. We hypothesize that second-language (L2) listeners learn their own idiosyncratic production patterns as perceptual representations–listening to their own speech is an input for their language comprehension. However, it is not necessarily just the pattern of speech, but also the timbre and the tone of the speaker's voice that is necessary to pinpoint. Many previous studies fail to show the speakers' exact voices without any alteration. Therefore, an L2-accented speaker, in this case a native Mandarin speaker learning English as an L2, should have a sharpened understanding of their own unaltered speech compared to any other L2 speaker with the same background. We refer to this as the **self-specialization hypothesis**.

We aim to test 24 pairs of native Mandarin speakers who are learning English and analyze their comprehension of themselves and someone else with the same language background. We use 24 minimal pairs, using voicing (cart/card) and vowel (pen/pan) words that have been historically shown to be challenging for native Mandarin speakers to pronounce (Flege et al., 1992; Jia et al., 2006; Li & Mok, 2012; Wang, 2007). Although in English, two words can sound completely different when pronounced by a native speaker. However, when native Mandarin and Cantonese speakers pronounce these particular words, made up of sounds not found in their native language, the lines begin to blur. Certain words that can be distinguished by a trained native ear, such as "pen" and "pan" can be interpreted and produced like homophones. This ensures that participants are not solely looking at two different words and instead are having to mentally compute the differences in the words they hear.

Before the study begins, the participants are primed to remember the particular minimal pairs. After the study, there is the Multilingual Naming Test (MINT) that gauges the participants' fluency in Mandarin and English. We want to understand how far along they are in their journey learning English and if they are a native Mandarin speaker.

Our study uses two methods of determining the participants' comprehension. The participants will hear either their own voice or the other person participating in the study with them and be shown pictures associated with the words. We calculate their accuracy by looking at the percentage of how many times they click on the correct word. We also use eye tracking to determine how fast people look at the stimuli spoken by themselves or the other participant. Participants in the study are shown to have higher word recognition percentages when they hear their own words compared to the other speaker. They are also quicker to look at the correct answer when they hear themselves compared to the other speaker as well. Where speakers show the biggest difference in favor of the self specialization hypothesis is with vowel paired words.

## Methods

Data will come from human subjects. We will recruit 24 pairs (48 total) of participants for our study where they are expected to speak, have their audio recorded, and listen to their own and other participant's recordings. We will continue to recruit participants until we meet 24 pairs that qualify for our study. The participants will not know the treatment group to which they have been assigned. All participants gave their informed consent in accordance with the protocols approved by the University of California, San Diego Human Research Protections Program.

The design of this study is within subjects. Participants are tested in pairs. There are three phases with one additional phase: the familiarization, recording, and clicking phase. In the familiarization phase, each participant sees each target word paired with a picture and are asked to rate how familiar on a scale from 1-6 they are with the word. This helps prime the participants to use these particular words in the naming phase, and also helps the researchers verify identification of the words. The data here should be consistent with the exclusion criteria. If people do not know what words they are seeing or saying, they will be excluded. In the recording phase, each participant is shown *only* the picture corresponding to the word they are supposed to recall and they speak the word. Both participants are in the room while they hear their own vocal productions and the other participant's an equal number of times. This repeats 48 times until all stimuli are shown. If a participant does not know or forgets a word, a simple "skip" or "don't know" response is permitted and will be included in the live experiment however removed from the analysis of the data. Showing only the picture allows for a more natural production of the word and also assures the participants know the words they are saying. When recording is finished, the participants are asked to go into another room to color or draw (in order to avoid exposure to other languages) while the researchers crop the sound files on the application Praat to reduce noise and focus on the word. The decibel levels are normalized for all productions. Finally, in the clicking phase, one participant is called back one at a time and they view four pictures at a time on a computer screen while hearing recordings of both their own voice or the other participant's voice. They are asked to use the mouse to click on the named picture, while eye movements are tracked. The order of recording self versus other in the naming phase and order of hearing self versus the other participant in the comprehension test are both counterbalanced across participants. In the recording phase, the participants hear their own voice twice and the other's twice per word. In total, with 24 pairs of words and two subjects, each participant should listen to spoken words 196 times for a total of 392 trials between the two. The final part of the experiment is the Multilingual Naming Test (MINT) where participants see 80 pictures of increasingly difficult objects and are asked to name them all first in English, then in Mandarin. The MINT test is a gauge of fluency in each respective language. After the experiment is completed, the participants fill out a final survey with questions that collect race and language demographics.

Prior to completion of the data collection, we will not look at any of the data directly associated with the accuracy of the experiment. The familiarization task and the MINT scores have both been viewed prior to finishing data collection. We *have* also looked at the accuracy with which they named words, as this is an exclusion criterion: if fewer than eight word pairs remain out of twenty-four after eliminating cases where one or both participants misnamed a word in the pair, then the participant pair is excluded. We enlisted the help of Whisper, a Language Learning Model (LLM) to expedite the process of detecting the correct word. If the LLM returned a word completely different than the expected target word, then there was a manual listening of the stimuli. If the LLM showed a synonym or a similar word, it counted as a pass and would be allowed to be used for the data analysis. The accuracy of Whisper compared to the real word and manual transcription is shown below (Figure 1).

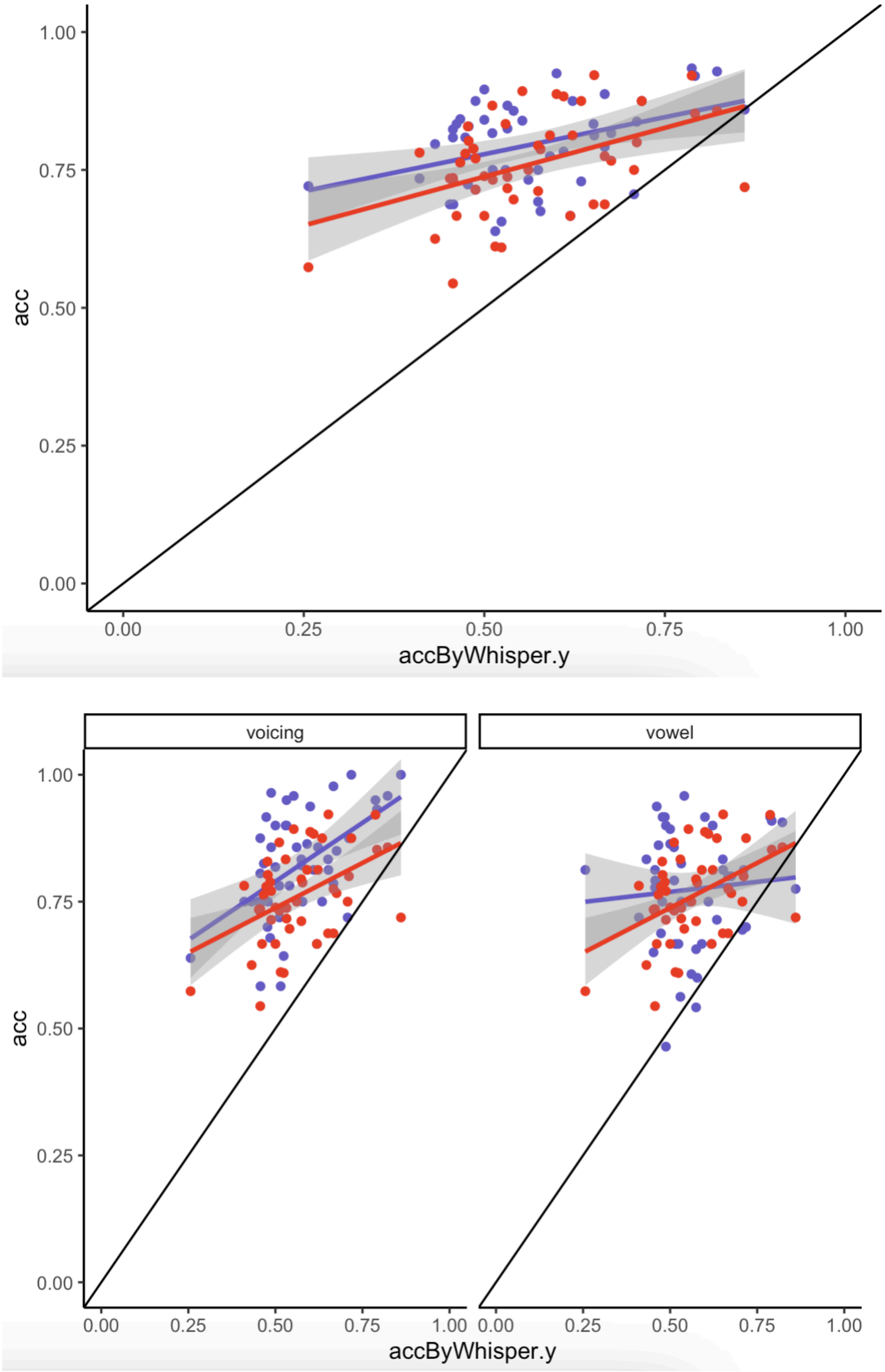

Figure 1: Whisper LLM model accuracy compared to real accuracy. Also split by vowel and voicing pairs

We have also looked at the final survey with detailed responses on language background, as this is also an exclusion criterion. Some participants report extensive knowledge of additional languages, native expertise in English, or hearing impairments, which they did not disclose prior to participation. Importantly, these exclusions were made *a priori,* NOT based on their accuracy or eye tracking data. The subjects were recruited from the UCSD SONA database for course credit or via flyers around the UCSD campus for pay. Some word of mouth recruiting in classes was conducted.  We tested 24 pairs of native Mandarin speakers (total $N = 48$). All participants must be 18 years or older.  Analyses in G*Power (Faul et al., 2009) or R software suggest that a moderate effect size of $\eta^2_P = .10$ can be detected with power > .99 at the target sample size of 48 participants. We plan to terminate data collection when we have reached 48 participants or 24 pairs who meet prespecified criteria.

For mouse-clicking accuracy: logistic mixed-effects regression (similar to an ANOVA) with subjects and items random intercepts and slopes for mouse clicking accuracy
Accuracy ~ HearingSelf * SpeakerAccuracy + (HearingSelf|Subject)
Speaker Accuracy, a continuous variable, will be assessed in two ways and z-scored:

- The speaker's English MINT score
- The relative transcription accuracy of that speaker by an ASR model (Whisper-small)

ANOVAs by participants and by items for looking proportions. Note that SpeakerAccuracy is a continuous variable.
Looks ~ HearingSelf * SpeakerAccuracy + Error(HearingSelf|Subject)
Looks ~ HearingSelf + Error(HearingSelf|Item)

**Results**

For the familiarization task, we found relative consistency when the participants rated the words. However, a scale is not completely determinant of people's familiarity with certain words. Most people can conduct a cursory response and finish soon without paying attention. However, even if most participants were familiar with most of the 48 words, the words “tan” and “peace” were the most commonly reported as less familiar. Also, certain plural versions of words needed to be recorded and participants often said the singular version i.e. “egg” instead of “eggs” for the eggs/x pair and “pea” instead of “peas” for the peas/peace pair.

### Accuracy and Eye Tracking

Across 48 different words, or 24 minimal word pairs, and 196 trials of listening to themselves and 196 trials of listening to the other participant for a total of 392 trials, people were able to understand their own vocal utterances better than the other participant. The native Mandarin speakers were able to understand themselves speaking English 81% of the time compared to 77% when they listened to the other participant (Figure 2). After systematically organizing the data and running a t-test we found a p-value of .015 which is less than .05 rejecting the null hypothesis of no difference. There is a significant difference between speaker/listener relationship comprehension. This phenomenon is most caused by the major difference between accuracy with vowel pairs compared to voicing pairs. When separating the different word pairs, voicing pair accuracy differences are minimal while vowel pair accuracy is much more stark. The difference in comprehension of voicing words was only about 1-2% with 3% error and 6% with 3% error for vowel productions (Figure 4). There was also a significant difference overall in how much vowel pairs had more comprehensibility overall compared to the voicing pairs (Figure 5). The data showed an overall enhanced level of comprehension when the participant listened to themselves speak (Figure 3).

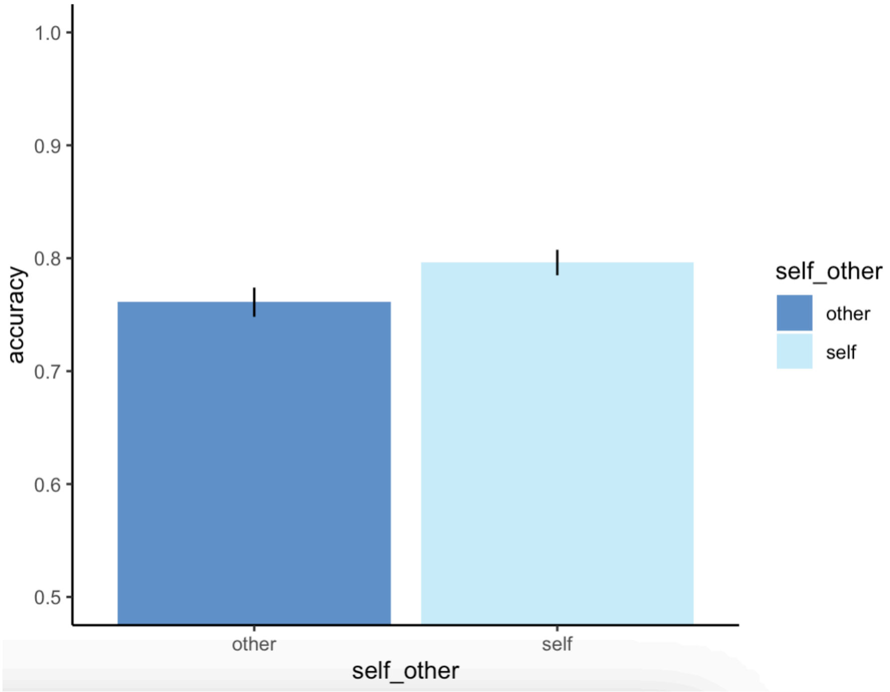

Figure 2: Accuracy differences between self-other

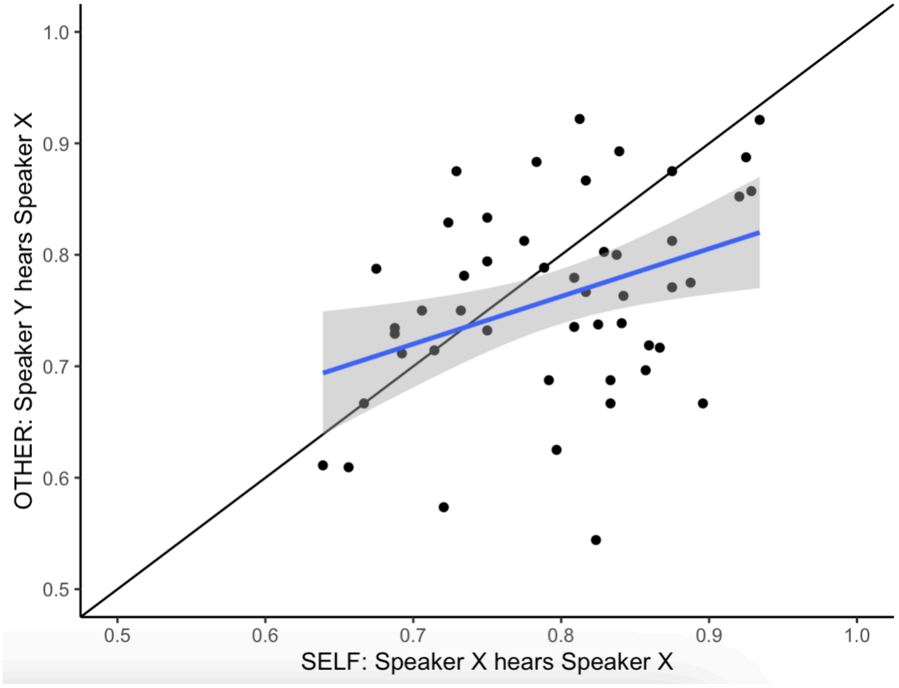

Figure 3: Self comprehension is observed to be significantly higher than no difference

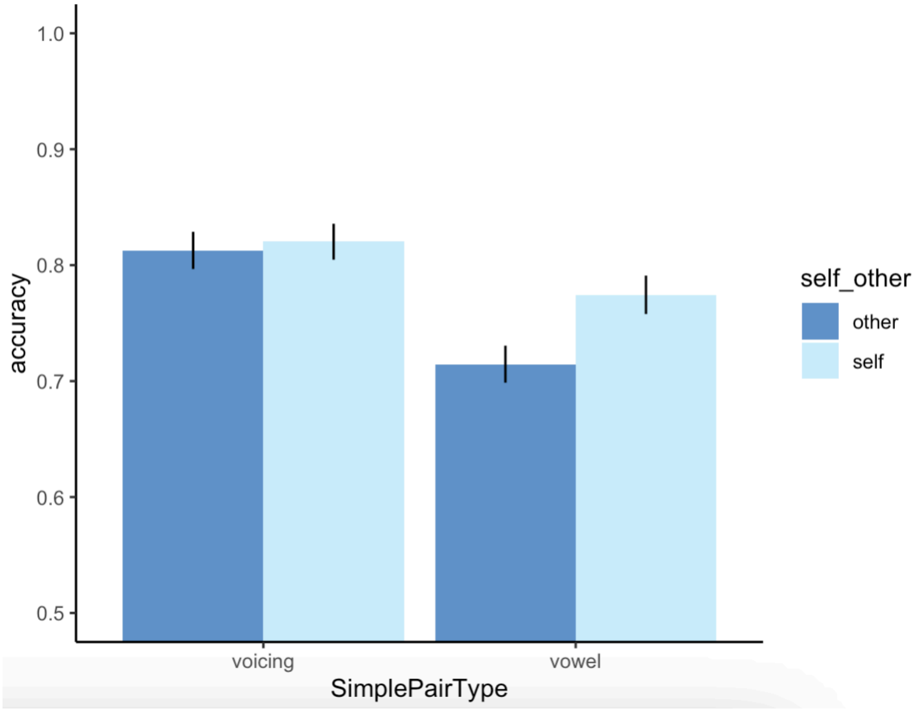

Figure 4: Accuracy differences between self-other separated by the simple pair type (voicing and vowel pair)

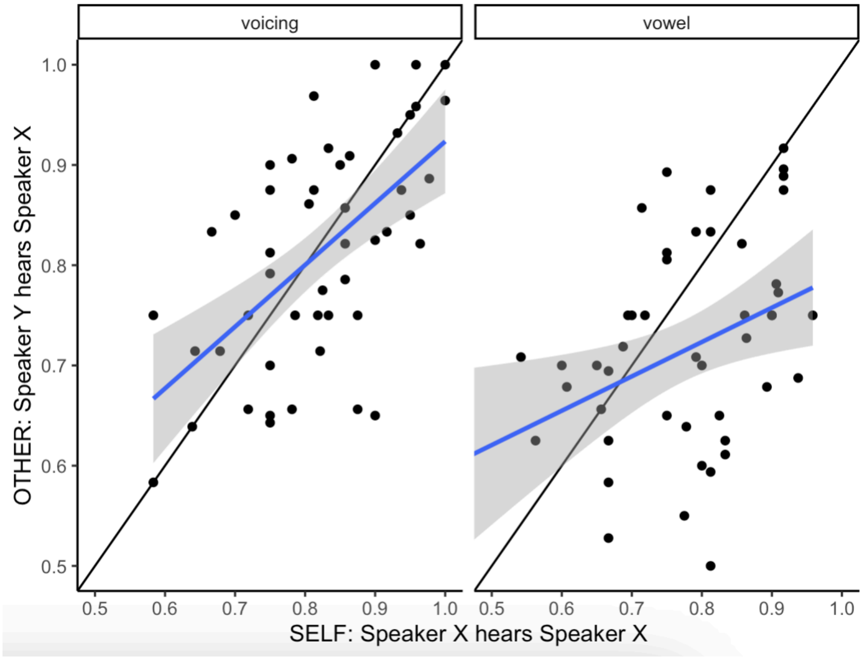

Figure 5: The tendency for speakers to understand themselves better when they are listening to vowel pairs rather than voicing pairs.

The same phenomenon can be found for the eye tracking data. Measuring the percentage of looks/fixations to the target over the word onset time showed there was a significant difference in how the participants viewed their own produced stimuli versus their experimental partner. Looks to the proper target were faster and at a higher percentage when people heard their own vocal productions. On average about 350ms faster over the course of one second (Figure 6). The word onset time, although conducted from 0ms to 2000ms, the most useful timeframe to observe is between 250 ms and 1250 ms because that is when the word starts and when fixations end. Any wider range will add more variance to the observed data due to natural saccades and changes of attention. Because of this, we determine that people are not only understanding themselves better, but are more confident at deciding their answer as well.

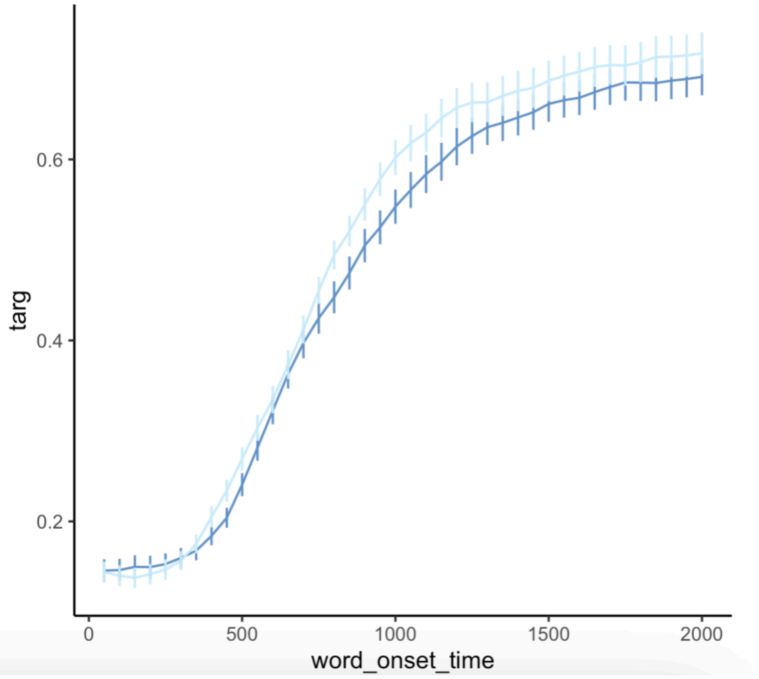

Figure 6: Eye tracking data. Percentage of looks to the target word self versus other over time.

For the MINT, we found that the participants we recruited were consistently high performing in Mandarin, however very scattered among their English vocabulary. This is the exact demographic we were looking for. To answer the self specialization hypothesis, having randomly assigned participants of all different skill levels and still finding an enhanced level of comprehension shows that it is not necessarily a factor of fluency or where people are in their journey of learning another language. Native Mandarin speakers understand their own speech better.

**Discussion**

The speech recognition aspect of this project is the most applicable to learning languages in general. Eger and Reinisch (2018) have shown that people with the same native language background will have a better time understanding what their peers speak compared to those of another language background. However, this does not specifically apply to how people can learn languages as a whole. With our data now, we have significant data backing up how we understand our own vocal productions better than someone else, even if they have the same language background. Going back to our hypothesis, our speech is strongly influential to our language development. Now we understand the extent to which our own inputs can affect our speech. As we continue to learn a new language, it would be beneficial to replace some of our own improper inputs with more authentic pronunciations—however this is not to discount the importance of the role of speaking in the journey of language learning.

The eye tracking data, although supplementary to the accuracy/clicking data can also be interpreted in more ways. Because looks to the proper target are faster, the data can also be interpreted as confidence influenced by familiarity. People look to the proper target word more confidently when they hear themselves speak compared to the other person’s voice.

One factor we did not fully address was the difference in gender pairing when conducting the experiment. If we paired the participants by gender, their vocal frequencies could be more similar and it could add either more noise or consistency into the data. Eger and Reinisch (2018) studied only a group of female speakers, which also has its limitations, but can be considered more consistent as the variable of vocal timbral quality is removed.

We also found that many participants were consistently missing certain words in the recording phase that led to the elimination of up to a majority of the data in some participant pairs. This can add more variability into the data and results as the less stimuli we have, the more one stimulus can affect the data. We need more procedures in place to increase the participant scores. However, according to our MINT scores and what we expect, people's Mandarin scores on average are higher and more consistent compared to their English scores.  These results do seem incredibly robust and consistent, but there may still be other accents and languages that can add more noise into the data. Maybe European or African accented speech will have a different effect on speech recognition.

The next steps we are taking revolve around similar concepts. The next experiment is a version where the participants are paired by matching a native English speaker with a native Mandarin speaker to test their ability to comprehend each other and themselves as well. This further tests the production lag hypothesis and pits the supposed most intelligible accents (native target, and self) against each other. The final step after that will be getting a baseline understanding of just how well participants understand themselves. This will encapsulate the entire project of speakers listening to their own unaltered speech compared to other groups and a baseline.

The self specialization hypothesis for native Mandarin speakers is significantly supported by the evidence of our experiment. Not only do they understand themselves better overall though more favorability to vowel minimal pairs compared to voicing, they are also more confident and certain about their decisions when they hear their own voice. Language strongly revolves around what the inputs are. Changing the way language is learned by introducing more authentic listening into the mix can be a strong benefit to the journey.

## References

Cheung, S., & Babel, M. (2022). The own-voice benefit for word recognition in early bilinguals. *Frontiers in Psychology, 13*, Article 901326. https://doi.org/10.3389/fpsyg.2022.901326

Creel S., Mizrahi R., Yu M. (2019). Perception precedes production in native Mandarin speakers of English. The Journal of The Acoustical Society in America https://pubs.aip.org/asa/jasa/article/146/4_Supplement/2792/704443

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America, 114*(3), 1600-1610. https://doi.org/10.1121/1.1603234

Dokovova, M., Scobbie, J. M., & Lickley, R. (Year). Matched-accent processing: Bulgarian-English bilinguals do not have a processing advantage with Bulgarian-accented English over native English speech. *Speech and Language Therapy, University of Strathclyde*, Glasgow, Scotland, UK; *Clinical Audiology Speech and Language (CASL) Research Centre, Queen Margaret University*, Edinburgh, Scotland, UK.

Eger, N. A., & Reinisch, E. (2018). The Impact of One's Own Voice and Production Skills on Word Recognition in a Second Language. Journal of Experimental Psychology: Human Perception and Performance, 44(4), 560–573. https://doi.org/10.1037/xhp0000487

Eger, N. A., & Reinisch, E. (2019). The impact of one's own voice and production skills on word recognition in a second language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(3), 552–571. https://doi.org/10.1037/xlm0000599

Flege, J. E. (1989). Chinese subjects' perception of the word-final English /t/-/d/ contrast: Performance before and after training. *Journal of the Acoustical Society of America, 86*(5), 1684-1697. https://doi.org/10.1121/1.398599

Geschwind, N. (1972). Language and the brain. *Scientific American, 226*(4), 76-83. https://www.jstor.org/stable/24927318

Hayes-Harb R., Smith B., Bent A., Bradlow T., (2008). Perception of English vowels by Mandarin learners: The interlanguage speech intelligibility benefit. *Journal of Phonetics, 36*(4), 603-615. https://doi.org/10.1016/j.wocn.2008.02.004

Hayes-Harb, R., & Smith, B. L. (2010). Individual differences in the perception of final consonant voicing among native and non-native speakers of English. *Journal of Phonetics, 38*(2), 272-280. https://doi.org/10.1016/j.wocn.2010.11.005

Hickok, G., & Poeppel, D. (2008). The cortical organization of speech processing. Journal of Phonetics. Nature Reviews Neuroscience, 8(5), 393–402. https://doi.org/10.1038/nrn2113

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *Journal of the Acoustical Society of America, 119*(2), 1117-1130. doi:10.1121/1.2151806

Li, A., & Mok, P. (2012). Cross-language perception and production of vowels by Cantonese and Mandarin learners of English. *Journal of the Acoustical Society of America, 131*(2), EL102-EL108. doi:10.1121/1.3674998

Schuerman, W. L., Meyer, A., & McQueen, J. M. (n.d.). Do we perceive others better than ourselves? A perceptual benefit for noise-vocoded speech produced by an average speaker. *PLOS one*

Wang, X. (2007). Perception of English vowels by Mandarin learners: The interlanguage speech intelligibility benefit. *Speech Communication, 51*(12), 875-888. https://doi.org/10.1016/j.specom.2007.05.004