Babble Noise: Application in Assessment

Xiaoya Wang *

August 23, 2023

Abstract

Language assessment is a crucial qualifying exam for people seeking study and working opportunities abroad. Research shows that test-takers develop non-language strategies to cope with language proficiency exams. (Cohen et al. 2023) The intelligibility of non-native English listeners can be largely influenced by babble noise compared to native-English listeners at a certain level of noise. (Florentine 1985, Jin & Liu 2012, Kozou et al. 2005) The listening performance among non-native listeners under babble noise environment has not been addressed and remains unclear. This study investigates second language assessment from a phonetic perspective, calling for a ceasefire of the traditional split between linguistics and applied linguistics.

*Department of Linguistics, University of California, Berkeley. E-mail: xiaoyawang@berkeley.edu

Contents

1 Introduction

Narrowly speaking, language assessment is a crucial qualifying exam for non-native language speakers seeking study and working opportunities abroad (Cohen et al. 2023). According to the official statement by a language assessment agency Educational Testing Services (2021), "TOEFL scores are accepted by more than 11,500 colleges, universities, and licensing agencies in more than 160 countries. The test is also used by governments, and scholarship and exchange programs worldwide." Research shows that test-takers develop test-taking strategies to cope with language proficiency exams, instead of using language strategies (Cohen 2013). Applied linguists have worked with test centers for decades to investigate test-taker behavioral patterns and adjust task designs, in order to reflecting the real language competence of test-takers and proving the validity of tests (Cohen 2006).

Previous research has indicated that the perceptual ability of non-native English listeners can be largely influenced by babble noise compared to native-English listeners at a certain level of noise. (Florentine 1985, Jin & Liu 2012, Kozou et al. 2005) And cross-linguistic study has shown that first language (L1) cues can influence listeners' perception of second language (L2) cues. Native-Korean speakers have a better performance in the English sentence recognition task than native-Chinese speakers (Jin & Liu 2012).

The listening performance among non-native listeners within a language community under babble noise environment has not been addressed and remains unclear. The relationship between intelligibility and test scores is underresearched as well.

Considering the previous study, My research questions would be: Are there differences of perceptual ability within a language community (i.e. Chinese)? Does the test results (i.e. TOEFL) align with the ability of recognition? Is it doable to embed babble noise into the task design of language assessment?

2 Methodology

The study conducted an empirical experiment. The experiment was an English sentence recognition task (Appendix 1), which has been adopted by mainstream babble noise studies in the field of linguistics. Subjects were asked to did a dictation in a blank A4 paper after hearing each English sentence under babble noise. There were no time limitation. They could go over the sentences again and again until they were confident about the final answer. And they could adjust the volume of audio at a comfortable level.

2.1 Subjects

The participants were 6 Chinese students at University of California, Berkeley. Biological gender balance was satisfied. They were native-Chinese speakers with English as the second language and most of them spent less than a year abroad in total. The latest TOEFL score of each participant was provided after the dictation task. And participants were categorized into three groups by range of scores (90-100, 100-110, 110-120). Two participants fell into the range of 100-110. One participant was below 100. One participant was above 110. Another participant lived in the United States for couple years, so the subject was considered as the control group. No speech and hearing impairment was reported.

2.2 Stimuli

The list of English sentences for recognition was selected from an irrelevant paper (Bent & Bradlow 2003), but had been used in another related and published paper (Van Engen & Bradlow 2007). The length of each sentence is adequate and the lexical item is not too advanced for L2 speakers' memory load (Chen et al. 2023). And the sentences were recorded by a white male English monolingual under a quiet environment by iPhone 14. The background noise was a file of multitalker babble on Github (Johnson 2023). The babble noise was extended by repeating the clip in Praat (Boersma & Weenink 1992–2022).

And the merging of audios was conducted in Audacity (Audacity Team 2023). The stimuli was taken for a pilot test in a small group of volunteers who did not participate in the experiment.

3 Predictions

To answer my previous research questions, my assumptions are:

1. There are differences of intelligibility within a language community. The gap between the groups of 100-110 & 110-120 is smaller than between 90-100 & 100-110.

2. The test results are aligned with the ability of recognition under babble noise. Scores of the language proficiency exam are positively correlated.

3. And it's promising to embed babble noise into the task design of listening section in the future.

4 Data Analysis

4.1 General

The words of dictation were divided into two categories, labelled as correct and incorrect respectively. And the correct words were made into a word cloud after removing a high frequency word "the" (Figure 2). And word frequency was calculated by a word cloud generator on Google Doc (Figure 1).

And something magical was observed. The most recognized words have burst onsets, which is different from the distribution in the original texts. More differently, words with consonant clusters as the onset were well-recognized while they have low occurrence. My assumption is that native-Mandarin speakers are more sensitive to consonant clusters because the type of onsets may share more similar cues in Mandarin. It is a situation where L1 gives advantage to L2 perception.

Word	Frequency
back	4
fast	4
girl	4
train	4
ball	3
found	3
kitchen	3
stood	3
bag	2
bicycles	2

Figure 1: Word List for correct word frequency.

Figure 2: Word Cloud for correct word frequency.

	Group 1 (90-100)	Group 2 (100-110)	Group 3 (110-120)
correct/incorrect (mean)	0.235	2.950	9.667

Figure 3: The Mean value of Correct words divided by incorrect words.

4.2 Individual

The positive correlates of sentence recognition results and TOEFL scores are significant, by calculating correct words divided by incorrected word (Figure 3).

5 Discussion

As the data presented, perceptual ability can be reflected on language proficiency test scores. And the gap between 100-110 & 110-120 is larger than 90-100 & 100-110, which was not expected, but indeed it is what it is. However, several subjects have reported that the background noise was not comfortable. It may also decrease the expectation and patience of participants, and influence the final result.

In the official annual report, the average TOEFL score in China was 87 in 2021. Considering the tendency of score increasing year by year and most participants took the test at least a year ago, all of the subjects were still assumed to be beyond average. And they were all well-educated students at top universities, two of them were PhD candidates. Maybe the result would be more representational and significant for the entire population if the scores of most samples were below 90.

From the perspective of experiment design, it will be more precise and scientific if only taking scores of the listening section into consideration. And it's more ideal to take the sentence recognition test just before participants take TOEFL exam.

6 Future Research

The balance between signal and noise should be taken more consideration. At the beginning of the study, the author was trying to control the signal to noise ratio (SNR). However, it was not possible to keep SNR at a single level throughout the speech without giving up the noise type babble noise. And that's the reason a Python script for adding noise was found and debugged successfully, but has not been used in the present project.

After debugging to almost midnight and realized it was not fitted to my research, Jackson and I decided to make our own Python script this summer. Our goal is to develop a script that users are able to upload signal and noise file, merging them into a single file with a fixed SNR. After assessing the feasibility, we came to the conclusion that it could be done by the current library.

7 Acknowledgement

I would like to thank my supervisor & instructor Professor Alexandra Pfiffner for her brilliant personality and introducing me to the wonderland of phonetics, Professor Keith Johnson for his expertise in the field of phonetics, Professor Zehlia Babaci-Wilhite for her unconditional love and support, Professor Andrew D. Cohen for his expertise in the field of applied linguistics, my peers in Ling 113 Experimental Phonetics & Ling 210 Advanced Phonetics for the great fellowship this semester, all my friends in the church.

Thank Julian Vargo for reading sentences with the Californian Vowel Shift. Thank Zilin Dong for volunteering in the pilot experiment. Thank Aquila Xu, Zikai Xu, Amy Tao and Yisheng Gao for participation. Special thank goes to Jackson Gao for his great contribution in debugging and data collection.

8 Appendix

Appendix 1.

The children dropped the bag.

The dog came back.

The floor looked clean.

She found her purse.

The fruit is on the ground.

Mother got a saucepan.

They washed in cold water.

The young people are dancing.

The bus left early.

The ball is bouncing very high.

Father forgot the bread.

The girl has a picture book.

The boy forgot his book.

A friend came for lunch.

The match boxes are empty.

He climbed his ladder.

The family bought a house.

The jug is on the shelf.

The ball broke the window.

They are shopping for cheese.

The pond water is dirty.

They heard a funny noise.

The police are clearing the road.

The bus stopped suddenly.

The book tells a story.

The young boy left home.

They are climbing the tree.

She stood near her window.

The table has three legs.

A letter fell on the floor.

The five men are working.

The shoes were very dirty.

They went on a vacation.

The baby broke his cup.

The lady packed her bag.

The dinner plate is hot.

A dish towel is by the sink.

She looked in her mirror.

The good boy is helping.

They followed the path.

The kitchen clock was wrong.

Someone is crossing the road.

The mailman brought a letter.

They are riding their bicycles.

He broke his leg.

The milk was by the front door.

The shirts are hanging in the closet.

The chicken laid some eggs.

The orange was very sweet.

He is holding his nose.

The new road is on the map.

She writes to her brother.

The football player lost a shoe.

The three girls are listening.

The coat is on a chair.

The train is moving fast.

The child drank some milk.

The janitor used a broom.

The ground was very hard.

The buckets hold water.

References

Audacity Team (2023), 'Audacity(r): Free audio editor and recorder', Computer application. Retrieved May 7th, 2023 from https://audacityteam.org/.

Bent, T. & Bradlow, A. (2003), 'The interlanguage speech intelligibility benefit', *The Journal of the Acoustical Society of America* **114**(3), 1600–1610.

Boersma, P. & Weenink, D. (1992–2022), 'Praat: doing phonetics by computer', Computer program. Retrieved 23 January 2022 from https://www.praat.org.

Chen, J., Antoniou, M. & Best, C. (2023), 'Phonological and phonetic contributions to perception of non-native lexical tones by tone language listeners: Effects of memory load and stimulus variability', *Journal of Phonetics* **96**, 101199.

Cohen, A. (2006), 'The coming of age of research on test-taking strategies', *Language Assessment Quarterly* **3**(4), 307–331.

Cohen, A. (2013), 'Test-taking strategies and task design', pp. 276–292.

Cohen, A., Rahmati, T. & Sadeghi, K. (2023), 'Test-taking strategies in technology-assisted language assessment', pp. 235–254.

Florentine, M. (1985), 'Speech perception in noise by fluent, non-native listeners', *The Journal of the Acoustical Society of America* **77**(S1), S106–S106.

Jin, S. & Liu, C. (2012), 'English sentence recognition in speech-shaped noise and multi-talker babble for english-, chinese-, and korean-native listeners', *The Journal of the Acoustical Society of America* **132**(5), EL391–EL397.

Johnson, K. (2023), 'Keithjohnson-berkeley - overview'. Accessed: May 8, 2023.

Kozou, H., Kujala, T., Shtyrov, Y., Toppila, E., Starck, J., Alku, P. & Näätänen, R. (2005), 'The effect of different noise types on the speech and non-speech elicited mismatch negativity', *Hearing Research* **199**(1-2), 31–39.

Van Engen, K. & Bradlow, A. (2007), 'Sentence recognition in native- and foreign-language multi-talker background noise', *The Journal of the Acoustical Society of America* **121**(1), 519–526.