Using virtual sound reproduction for studying L1/L2 speech perception in varying acoustic environments

Yusuke Hioka[1], C.T. Justine Hui[1], Hinako Masuda[2] and Catherine I. Watson[3]
[1]Acoustics Research Centre, Department of Mechanical and Mechatronics Engineering, University of Auckland, New Zealand
[2]Faculty of Science and Technology, Seikei University, Japan
[3]Department of Electrical, Computer and Software Engineering, University of Auckland, New Zealand

1aPP26

1. Background and Scope

- ✓ It is logistically challenging to study speech perception when:
  - studying the effect of varying acoustic environments; and/or
  - comparing first (L1) and second (L2) language listeners

  because participants have to travel to the venues where experiment is conducted
- ✓ Virtual sound reproduction technology could address the challenge thanks to its ability to reproduce the acoustics of arbitrary environments at any geographic locations in a controllable manner
- ✓ However, it has NOT been studied well if the results collected using virtual sound reproduction would replicate the results collected in the original real spaces
- ✓ This study investigates the difference of speech perception in varying acoustic environments between L1 and L2 language New Zealand English listeners using virtual sound reproduction technology
- ✓ The study particularly focuses on how the results collected under virtual acoustic environments assimilate to that collected in the original real acoustic environments between L1 and L2 listeners

2. Virtual sound reproduction

System overview

- ✓ Implemented 3rd order Ambisonics based system (Fig.1 top)
- ✓ Room impulse responses (RIR) in target environment were measured by Eigenmike (32-ch spherical microphone array) *https://leomccormack.github.io/sparta-site/
- ✓ The RIRs were encoded and decoded by the SPARTA toolboxes*
- ✓ Decoded RIRs were convolved with arbitrary sound sources to generate stimuli, then were rendered by 16-ch loudspeaker array (Fig.1 bottom) installed in the anechoic chamber at the University of Auckland

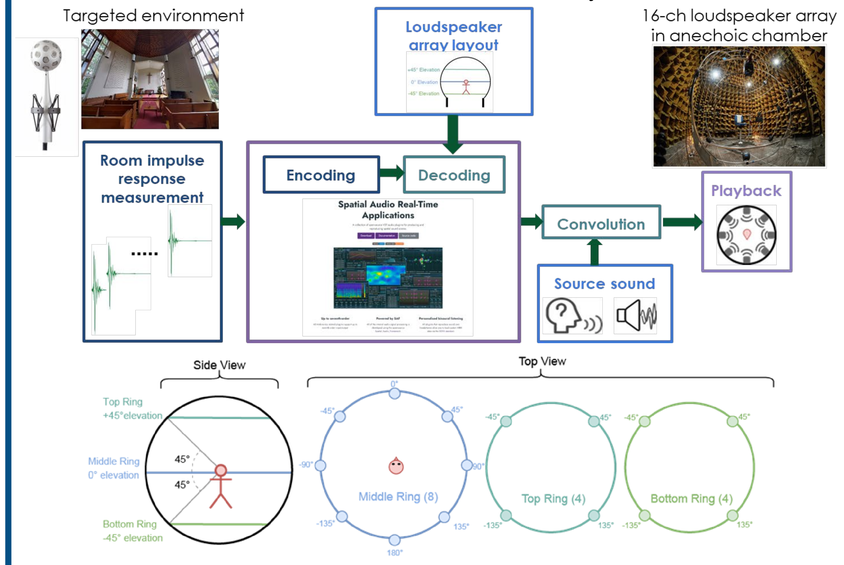


Figure 1: Virtual sound reproduction system used in the study; Top: Implementation flowchart; Bottom: Loudspeaker array configuration

3. Subjective listening test

Methodology

- ✓ Speech intelligibility test – participants transcribed spoken sentences in noise via GUI (Fig. 2). The number of correctly transcribed words was counted and scaled to 0 to 1 as *proportion correct.*
- ✓ Speech and noise were played simultaneously from various separation angles (*Speech-noise separation*; Fig. 3).
- ✓ Room acoustics of the environment was varied by testing in 2 rooms and 2 source distances (*Room acoustics*).
- ✓ Recruited participants with different language background quantified by their first exposure to English-speaking environment (*Immersion age*)
- ✓ Conducted the test both in the real rooms and under the virtual sound reproduction (*Test venue*)

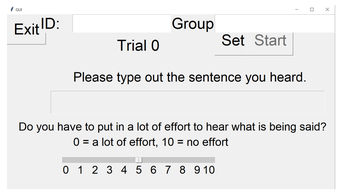


Figure 2: Graphical user interface used in the test

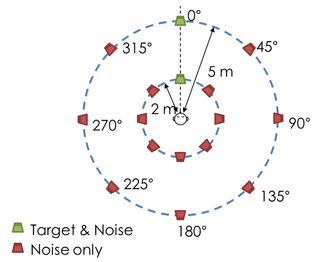


Figure 3: Position of the sound sources

Sound sources

- ✓ Target speech: Bamford-Kowal-Bench (BKB) Sentence; recordings of New Zealand (NZ) accented female voice
- ✓ Noise: Babble noise
- ✓ Target-to-noise ratio: -3 dB

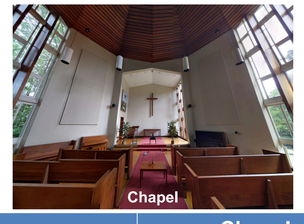
Participants

- ✓ 57 participants with normal hearing (self-reported); 18 – 49 years old
  - *Early* immersed (n = 20): born in NZ or moved to NZ before 12/13 years old
  - *Late* immersed (n = 37): moved to NZ after 14 years old

Marking rubric

- ✓ Marked by the root of words e.g. “running”, “ran” correct for “run”
- ✓ No penalties for homonyms e.g. “meat”/“meet”, “sun”/“son”
- ✓ No penalties for vowel merger in NZ English e.g. “ear”/“air”

Room acoustics



Metric	Distance	Seminar room	Chapel
Reverberation Time (T20)		0.7 s	1.7 s
Speech Clarity (C50)	2 m	15.9 dB	10.3 dB
	5 m	6.8 dB	3.3 dB

4. Results and Findings

Results

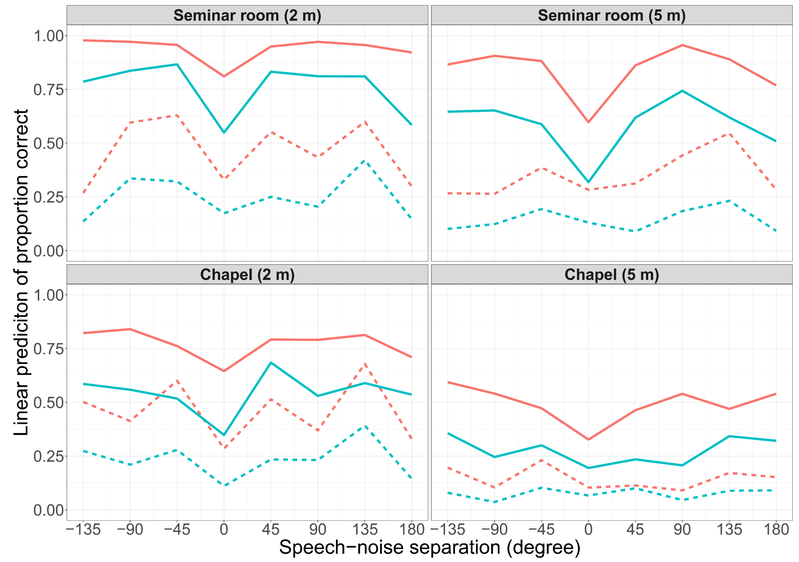


Figure 4: Results of the subjective listening test separated by *Room acoustics*

Key findings

- ✓ Four-way interaction between the *Room acoustics*, *Immersion age*, *Test venue* and *Speech-noise separation*
- ✓ *Immersion age* consistently affected speech intelligibility (Early > Late) mostly regardless of *Room acoustics*, *Speech-noise separation*, or *Test venue* (excl. Chapel (5m) under virtual sound reproduction)
- ✓ Speech intelligibility differed by *Test venue* (Real > Virtual) regardless of *immersion age,* also by *Room acoustics* (Seminar > Chapel; 2m > 5m) regardless of *Test venue* but *Immersion age* affected it differently
- ✓ Spatial release from masking (the “dip” at 0°) was observed regardless of *Immersion age*, but was affected by *Room acoustics* and *Test venue*
  - For *Room acoustics*, mainly when the source distance was shorter
  - For *Test venue*, more benefit for *Early* in virtual sound reproduction whereas more benefit for *Late* in real rooms

5. Conclusion

- Difference in speech intelligibility between L1 and L2 listeners in both real room and virtual sound reproduction was evaluated and compared
- Virtual sound reproduction closely replicated the **relative** difference in speech intelligibility between L1 and L2 listeners that was observed in the real acoustic environments
- Speech intelligibility was degraded under virtual sound reproduction compared to the real room counterparts for both L1 and L2 listeners
- The effect of spatial release from masking was observed both in L1 and L2 listeners but the trend differed by both acoustics of the environment and test venue (real room or virtual sound reproduction)