Impact of Perceptual Training on Comprehensibility in Spontaneous Speech

Efthimia Mia Moutis

PhD

University of York
Education

May 2025

Abstract

The ultimate goal of learning to speak a foreign language is to be understood that is, to achieve comprehensibility (Derwing & Munro, 2015). There is strong evidence that training focused on the pronunciation of individual sounds, such as High Variability Phonetic Training (HVPT), improves learners' ability to identify individual phonemes (Logan et al., 1991; Bradlow et al., 1997) and produce more intelligible speech at the word and sentence level (Thomson, 2011; Iino, 2019). Recent research also suggests potential gains in comprehensibility, though primarily in semi-controlled tasks, leaving it unclear whether such training benefits extend to spontaneous speech, issues that have rarely been examined through delayed testing.

This study investigated the impact of HVPT on perception, intelligibility, and comprehensibility in spontaneous oral production among Chinese learners of English with a focus on high functional load phonemic contrasts. It explored whether improved perception would transfer to word-level intelligibility and, ultimately, to greater comprehensibility in free oral production. A secondary aim was to assess the durability of these gains through delayed testing. HVPT was delivered entirely online, reflecting real-world learning contexts.

Fifty-one adult Mandarin speakers of English completed 15 HVPT lessons over three weeks. A pre-, post-, and delayed post-test design assessed perception via identification tasks, intelligibility through a read-aloud task, and comprehensibility via both semi-controlled and spontaneous speech. Results showed significant and sustained improvements in phonemic perception. Intelligibility gains were modest and varied by phoneme, while comprehensibility improved notably in the spontaneous narrative task, but only at delayed post-test. These patterns suggest a time-dependent transfer from perception to production.

Overall, the findings highlight HVPT's potential to support long-term L2 speech development. Gains in perception can cascade into improvements in intelligibility and comprehensibility, particularly when diverse speaking tasks and delayed assessments are used. The study underscores the value of integrating spontaneous speech and longitudinal evaluation into pronunciation research and pedagogy.

Table of Contents

Abstract 2

Table of Contents 3

List of tables 6

List of figures 7

Dedication 8

Acknowledgements 9

Declaration 10

Chapter 1 Introduction 11

1.1 Introduction *11*

1.2 Rationale *13*

1.3 Aims and Research Questions *16*

1.4 Context *17*

1.5 Structure of thesis *19*

Chapter 2 Literature 21

2.1 Introduction *21*

2.2 Comprehensibility *21*
2.2.1 Definition and Conceptualization 21
2.2.2 From Broad Instructional Goals to Specific Objectives 24
2.2.3 Linguistic Features Influencing Comprehensibility 26

2.3 Intelligibility *28*
2.3.1 Definition and Conceptualization 28
2.3.2 From Broad Goals to Specific Objectives in Pronunciation Intelligibility 30
2.3.3 Linguistic Features Influencing Intelligibility 31
2.3.4 Intelligibility from a Theoretical Perspective 33

2.4 Pronunciation Training *36*
2.4.1 Types of Pronunciation Training 36
2.4.2 Perceptual training 38
2.4.3 Speech Perception Theories 41

2.5 Functional Load *43*
2.5.1 Theoretical Foundations of Functional Load 44
2.5.2 Operationalizing Functional Load 45
2.5.3 Empirical Support for Functional Load 50

2.6 High Variability Phonetic Training *54*
2.6.1 Principles and Methodology of HVPT 54
2.6.2 Types of HVPT Activities 59
2.6.3 Impact of HVPT on Perception 61
2.6.4 HVPT and intelligibility in isolated word production 65
2.6.5 HVPT and comprehensibility in spontaneous speech 69

2.7 Chapter Summary *73*

Chapter 3 Methodology **76**

3.1 Introduction *76*

3.2 Pilot study *80*
3.2.1 Purpose 80
3.2.2 Participants 82
3.2.3 Piloting the training material 82
3.2.4 Piloting the perception test 84
3.2.5 Piloting the comprehensibility and intelligibility tests 87
3.2.6 Implications for the main study 89

3.3 Main Study *89*

3.3 Participants *90*
3.3.1 L2 Learners 90
3.3.2 Raters 91

3.4 Design *91*

3.5 Materials *95*
3.5.1 The online phonetic training design 95
3.5.2 Training material 96
3.5.3 Pre, post and delayed post-tests 100
3.5.4 Unstructured interview 118

3.6 Ethics *119*

Chapter 4 Results **121**

4.1 Introduction *121*

4.2 Perception of phonemic contrasts *121*
4.2.1 Introduction 121
4.2.2 Results 122

4.3 Intelligibility of pronounced words *126*
4.3.1 Introduction 126
4.3.2 Results 126

4.4 Comprehensibility of spontaneous oral productions *132*
4.4.1 Introduction 132
4.4.2 Comprehensibility of spontaneous oral productions of Task 1 133
4.4.3 Comprehensibility of spontaneous oral productions of Task 2 134

4.5 Unstructured interview feedback *137*

Chapter 5 Discussion **139**

5.1 Introduction *139*

5.2 Aims of the Study *139*

5.3 Key findings *140*

5.4 Perception *143*
5.4.1 Overall gains in perception 143
5.4.2 Retention of Perception 144
5.4.3 Perception and high functional load phonemes 145

5.5 Intelligibility *147*
5.5.1 Overall gains in intelligibility 147
5.5.2 Retention of intelligibility 148
5.5.3 Intelligibility and high functional load phonemes 149

5.6 Comprehensibility in spontaneous oral production *150*

5.6.1 Overall gains in comprehensibility 150
5.6.2 Retention in comprehensibility 151
5.6.3 Comprehensibility and High Functional Load Phonemes 153

5.7 Chapter summary *155*

Chapter 6 Conclusion 156

6.1 Introduction *156*

6.2 Aims and research questions *156*

6.3 Summary of the study *157*

6.4 Summary of key findings *158*

6.5 Significance *160*

6.6 Limitations *160*

6.6 Implications *162*
6.6.1 Theoretical implications 162
6.6.2 Pedagogical implications 164
6.6.3 Methodological implications 165

6.7 Chapter summary *166*

Appendix A *Pilot study consent and questionnaire form for learner participants* *167*

Appendix B *Pilot study consent and questionnaire form for raters* *170*

Appendix C *Main study consent and questionnaire form for learner participants* *173*

Appendix D *Main study consent and questionnaire form for raters* *176*

Appendix E *Stimuli of the Individual lessons for the high variability phonetic training* *179*

Appendix F *Learner practice sample: Comprehensibility Task 1 timed picture story* *181*

Appendix H *Learner official test: Comprehensibility Task 1 timed picture story* *183*

Appendix I *Learner official test: Comprehensibility Task 2 long turn narrative* *184*

Appendix J *Comprehensibility rating scale for Task 1 timed picture story instructions for raters* *185*

Appendix K *Comprehensibility rating scale for Task 2 long-turn narrative instructions for raters* *186*

Appendix L *Histograms and Q-Q plots of the perception test at pre, post and delayed tests* *187*

Appendix M *ANOVA results for differences in perception test at pre-, post-, and delayed post-tests on raw scores* *188*

Appendix N *Wilcoxon Signed Rank test of the intelligibility test within group on all available L2 learners at pre, post and delayed post-tests.* *189*

Appendix O *Histograms and Q-Q plots of the comprehensibility test Task 1 at pre, post and delayed tests* *190*

Appendix P *Friedman test results for comprehensibility Task 1 at pre-, post-, and delayed post-tests* *191*

Appendix Q *Histograms and Q-Q plots of the comprehensibility test Task 2 at pre, post and delayed tests* *192*

Appendix R *Friedman test results for comprehensibility Task 2 at pre-, post-, and delayed post-tests* *193*

Appendix S *Summary of Lesson Duration Data* *194*

References 195

List of tables

TABLE 1 MEAN SCORES AND STANDARD DEVIATIONS OF THE PILOT PERCEPTION TEST AT PRE, POST, DELAYED POST-TESTS...............85
TABLE 2 SHAPIRO-WILK TEST RESULTS FOR THE PERCEPTION TEST AT PRE, POST AND DELAYED POST-TESTS. 86
TABLE 3 DEPENDENT VARIABLES AND ASSESSMENT TYPES IN THE STUDY 94
TABLE 4 THE TEN VOWELS AND CONSONANT PHONEMIC CONTRASTS WITH THE HIGHEST FUNCTIONAL LOAD AS RANKED BY BROWN (1988) USED IN THE STUDY 97
TABLE 5 THE ORDER AND PURPOSE OF PRE-POST-AND DELAYED-POST TESTS..101
TABLE 6 ALLOCATION OF RATERS TO COMPREHENSIBILITY TESTS, EACH TEST RATED AT PRE-, POST-, AND DELAYED POST-TESTS STAGE BY SAME RATER FOR THE L2 LEARNERS 104
TABLE 7 NUMBER OF L2 LEARNERS INCLUDED IN THE ANALYSIS OF THE COMPREHENSIBILITY TESTS (TASK 1 & TASK 2)................107
TABLE 8 THE 80-WORD STIMULI USED IN THE INTELLIGIBILITY TEST INCLUDING TRAINED AND UNTRAINED WORDS AT PRE-, POST-, AND DELAYED POST-TESTS 108
TABLE 9 THE FINAL PHONEMIC CONTRASTS SELECTED FOR RATING AND THE TOTAL NUMBER OF WORDS RATED PER L2 LEARNER ON THE INTELLIGIBILITY TEST 106
TABLE 10 ALLOCATION OF RATERS TO INTELLIGIBILITY TESTS, EACH RATED AT PRE-, POST-, AND DELAYED POST-TEST STAGE BE SAME RATER FOR THE L2 LEARNERS INDICATED 110
TABLE 11 THE 18 L2 LEARNERS AND THE WORDS THAT WERE NOT RATED IN THE INTELLIGIBILITY TEST AT PRE-, POST-, AND DELAYED POST-TESTS..113
TABLE 12 CRONBACH ALPHA OF THE INTELLIGIBILITY TEST BASED ON THE 12 WORDS WITH THE SAME 33 L2 LEARNERS AT PRE-, POST-, AND DELAYED POST-TESTS 113
TABLE 13 THE 80-WORD STIMULI OF THE PERCEPTION TEST AT PRE-, POST-, AND DELAYED POST-TESTS 115
TABLE 14 THE CRONBACH'S ALPHA VALUE OF THE 160 ITEMS IN THE PERCEPTION TEST AT PRE-, POST-, AND DELAYED POST-TESTS117
TABLE 15 THE CRONBACH' ALPHA TEST FOR THE PERCEPTION TEST AT PRE-, POST, AND DELAYED POST-TESTS 122
TABLE 16 NORMALITY TEST OF THE PERCEPTION TEST AT PRE-, POST-, AND DELAYED POST-TESTS 123
TABLE 17 MEAN SCORES AND STANDARD DEVIATIONS OF THE PERCEPTION TEST AT PRE-, POST-, AND DELAYED POST-TESTS..........123
TABLE 18 THE EFFECT SIZE OF THE PERCEPTION TEST AT PRE-, POST-, AND DELAYED POST-TESTS...124
TABLE 19 POST HOC PAIRWISE COMPARISONS OF PERCEPTION TESTS SCORES ACROSS THE THREE TIME POINTS...........................125
TABLE 20 CRONBACH 'S ALPHA TEST ON THE DATASET OF THE 12-WORD ITEMS IN THE INTELLIGIBILITY TEST FOR ALL AVAILABLE L2 LEARNERS AT PRE-, POST-, AND DELAYED POST-TESTS...127
TABLE 21 MEAN, SD, SE, CONFIDENCE LEVEL OF MEAN AND SIGNIFICANCE LEVEL OF THE INTELLIGIBILITY TEST OF L2 LEARNERS' WORD PRODUCTIONS AT PRE-, POST-, AND DELAYED POST-TESTS FOR ALL AVAILABLE L2 LEARNERS WITHIN GROUP.....................128
TABLE 22 RESULTS OF THE RATED INTELLIGIBILITY TEST ON THE 12 UNTRAINED WORDS REPRESENTED BY PERCENTAGE AT PRE-, POST-, AND DELAYED POST-TEST FOR ALL AVAILABLE L2 LEARNERS 129
TABLE 23 POST HOC WILCOXON SIGNED RANK TEST OF INTELLIGIBILITY TEST SCORES ON THE SUBSET OF THE UNTRAINED 12-WORD ITEMS AT PRE-TEST TO POST-TEST..131
TABLE 24 POST HOC WILCOXON SIGNED RANK TEST OF INTELLIGIBILITY TEST SCORES ON THE SUBSET OF THE UNTRAINED 12-WORD ITEMS AT PRE-TEST TO DELAYED POST-TEST 131
TABLE 25 NORMALITY TEST OF THE COMPREHENSIBILITY TASK 1 AT PRE-, POST-, AND DELAYED POST-TESTS...............................133
TABLE 26 MEAN, SD, SE, AND CONFIDENCE LEVEL OF RATINGS OF THE COMPREHENSIBILITY OF LEARNERS' SPONTANEOUS ORAL PRODUCTIONS AT PRE-, POST-, AND DELAYED POST-TEST OF THE TIMED PICTURE STORY..134
TABLE 27 NORMALITY TEST OF COMPREHENSIBILITY TASK 2 AT PRE-, POST-, AND DELAYED POST-TESTS 135
TABLE 28 MEAN, SD, SE, AND CONFIDENCE LEVEL OF RATINGS OF THE COMPREHENSIBILITY OF LEARNERS' SPONTANEOUS ORAL PRODUCTIONS AT PRE-, POST-, AND DELAYED POST-TEST OF THE LONG TURN NARRATIVE..136
TABLE 29 SUMMARY OF THE KEY FINDINGS IN THE MAIN STUDY 138
TABLE 30 SUMMARY OF PRE-TO-POST GAINS AND PRE-TO-DELAYED POST GAINS OF THE PERCEPTION, INTELLIGIBILITY AND COMPREHENSIBILITY TESTS...142
TABLE 31 PHONEMIC PAIRS OF HIGH FUNCTIONAL LOAD AND ATTESTED DIFFICULTIES USED IN THE TRAINING AND THEIR RESULTS AT PRE-, POST-, AND PERCEPTION TESTS. 146

List of figures

FIGURE 1 HISTOGRAMS OF THE PERCEPTION SCORES OF THE (A) PRE-TEST, (B) POST-TEST, (C) DELAYED POST-TESTS 85

FIGURE 2 GENERAL DESIGN OF THE MAIN STUDY 93

FIGURE 3 AN INDIVIDUAL TRAINING LESSON SHOWING THE DIFFERENT TYPES OF ACTIVITIES AND THE NUMBER OF TRIALS IN THE LESSON 98

FIGURE 4 TRAINING LESSON INTERFACE PART 1 TWO-WORD FORCED-CHOICE IDENTIFICATION ACTIVITY 99

FIGURE 5 TRAINING LESSON INTERFACE PART 2 BINARY IDENTIFICATION ACTIVITY YES OR NO 100

FIGURE 6 TRAINING LESSON INTERFACE PART 3 DISCRIMINATION ACTIVITY ODD WORD OUT.....................100

FIGURE 7 THE INTERFACE USED BY THE RATERS TO RATE THE COMPREHENSIBILITY TEST TASK 1 AND TASK 2 106

FIGURE 8 THE INTERFACE USED BY THE RATERS TO RATE THE INTELLIGIBILITY TEST.....................111

FIGURE 9 THE INSTRUCTIONS OF THE PERCEPTION TEST FOR THE L2 LEARNERS.....................116

FIGURE 10 SAMPLE EXERCISE OF PART I TWO-WORD FORCED-CHOICE IDENTIFICATION ACTIVITY OF THE PERCEPTION TEST FOR L2 LEARNERS.....................116

FIGURE 11 SAMPLE EXERCISE OF PART II BINARY IDENTIFICATION YES OR NO ACTIVITY OF THE PERCEPTION TEST FOR L2 LEARNERS 117

FIGURE 12 BAR GRAPH SHOWS POINT ESTIMATES (MEAN SCORES) WITH 95% CONFIDENCE INTERVAL OF THE INTELLIGIBILITY TEST AT PRE-, POST- AND DELAYED POST-TEST FOR ALL AVAILABLE L2 LEARNERS WITHIN GROUP 128

FIGURE 13 BAR GRAPH SHOWS INDIVIDUAL WORD-LEVEL INTELLIGIBILITY IMPROVEMENTS OVERTIME ACROSS PRE-, POST- AND DELAYED POST-TESTS.....................130

Dedication

This thesis is dedicated to my family for their endless support and encouragement.

Acknowledgements

I would like to express my sincerest gratitude to my supervisor, Dr. Zoe Handley. Your guidance and exceptional support have been invaluable. I am also thankful to my TAP supervisor, David O’Reilly for your encouragement and insightful feedback.

A special thanks to all the learner participants and raters who generously took part in the study. To colleagues and friends for supporting my PhD journey. Finally, to my family for their patience and unconditional support.

Declaration

I declare that this thesis is a presentation of original work, and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Chapter 1 Introduction

1.1 Introduction

Oral production skills are fundamentally important while learning a second language. Most people mainly utilize speech as a tool to support the process of second language acquisition (Handley, 2005). The development of oral production skills in globalization enables people to interactively participate in different communication situations across multiple linguistic environments (Dong et al., 2019). Moreover, spoken language helps people strengthen their grammar and lexical skills, and these skills are especially important in multilanguage speaking societies for communication (Darcy, 2018). Accordingly, international language bodies such as the Common European Framework of Reference for Languages (CEFR) place importance to include in pronunciation training a component of communication for enhancing communicative competence (Grim & Sturm, 2016). Thus, the central aim of language teaching is to support L2 learners to achieve their communicative goals that is effective in real-world communication, understood here as authentic, spontaneous and contextually meaningful language use beyond controlled classroom or laboratory settings.

Clear speech is significant within academic contexts for successful participation and achievement. Research findings emphasize that communication success may not be dependent on teaching grammar and lexis only, as other factors such as pronunciation difficulties and lack of fluency may interfere in understanding and reducing communication effectiveness (Tavakoli & Cooke, 2024). Even more so important in high-stakes situations including group projects, presentations, and interviews require oral communication for exchanging of information and targeting goals (Rogerson-Revell, 2011, 2021). Therefore, language instruction should include in its scope explicit training in pronunciation accuracy and fluency development to further successful communication. Such instruction equips learners to reach communication in spontaneous oral productions and respond to communicative demands (Derwing et al., 1998; De Jong et al., 2012). Given these demands, it is important to understand what makes spoken language more accessible to listeners, particularly in intelligibility and comprehensibility, which are considered important thresholds in the assessment of pronunciation instruction.

Building on this, effective spoken communication depends on producing speech that is readily understandable to the listener, a goal captured by the constructs of intelligibility and comprehensibility. While these related but distinct dimensions have been extensively investigated in pronunciation research (Munro & Derwing, 1995; Trofimovich & Isaacs, 2012), recent work demonstrates that instructional designs may influence intelligibility and comprehensibility in divergent ways. For example, Derwing et al. (1998) demonstrated that prosody-focused instruction improved comprehensibility and fluency in spontaneous speech, whereas segmental-focused interventions had a more limited effect in that particular context, resulting in questioning when and how different pronunciation features should be taught. Despite interest in prosodic training, recognizing that accurate perception and production of segmental features provide the necessary foundation for intelligibility is crucially important as this supports prosodic competence to be built on effectively (Thomson & Derwing, 2015). Accordingly, future instructional designs are recommended to develop training interventions on segmental accuracy before introducing suprasegmental elements, particularly for L2 learners whose phonemic contrasts are not found in their first language inventory.

Adult learners frequently have challenges in acquiring the phonological system of a second language, especially with phonemic contrasts that are absent from their native language. For instance, Chinese learners of English may be challenged in discriminating or orally producing contrasts such as /n/-/l/ in words like "net"-"let" or /r/-/l/ in "red"-"led", as these distinctions do not exist in their L1 phonological inventory (MacKain, Best & Strange, 1981; Pruitt et al., 2006; Mora & Fullana, 2007; Wong, 2014). These perceptual and productive difficulties can have an influential role in impeding the output of communication, highlighting for further targeted interventions which include specific phonological challenges faced by L2 learners across different L1 backgrounds.

Unintelligible communication caused by pronunciation difficulties can have far-reaching social, academic, and professional consequences for L2 speakers, often resulting in reduced participation, lower self-esteem, and reduced opportunities (Crowther, 2023). Such communicative barriers reinforce the need of developing pronunciation curriculum that focus on the perceptual and productive challenges most relevant to learners' communicative goals. Also, is the need to tailor instructional approaches to learners' linguistic background needs, technological access, and long-term communicative objectives (Jenkins, 2000; Derwing & Munro, 2005; Grantham O'Brien et al., 2018). Thus, focusing these types of training interventions is important for learners to achieve communicative success and support them to overcome communicative barriers.

These considerations highlight the need for pedagogical and research rationale in systematically investigating the link between perceptual phonetic training and communicative success in L2 learners. Recent calls in the literature advocate that studies go beyond controlled laboratory conditions and examine the real-world applicability of phonological training interventions, particularly in digitally mediated environments (Saito, 2021). Therefore, linking research and pedagogical practices remains a key obstacle in furthering effective pronunciation teaching.

Through attention to these pedagogical aspects, the following chapter introduces the conceptual foundation for investigating the role of perceptual training, particularly High Variability Phonetic Training (HVPT), which can be designed to improve the intelligibility and comprehensibility of L2 speech. This study is especially timely given the uptake in dependency on technology-enhanced instruction and the demand for pedagogical practices that are modifiable, learner-centered, and empirically grounded (Barriuso & Hayes-Harb, 2018). As such, a pedagogical approach in research inclusive of technology for pronunciation training provides a possible solution for L2 learners.

1.2 Rationale

While recent advances in pronunciation pedagogy have begun to demonstrate the contribution of perception-based interventions, the overall effect on the success of communication remains inadequately understood. Most notably, High Variability Phonetic Training (HVPT) has gained empirical support in improving perception of challenging phonemic contrasts for L2 learners, especially ones that do not exist in the learner's L1 or are difficult to distinguish (Thomson, 2018). However, despite numerous studies documenting perceptual gains, questions remain as to whether these improvements extend beyond controlled environments and directly influence intelligibility and comprehensibility, constructs more reflective of real-world communication.

A key motivation for the present study arises from the current lack of research on whether HVPT can meaningfully strengthen the broader communicative goals of intelligibility, comprehensibility, and real-world interactional competence. More personally, my motivation for this study is developed from several years of teaching EAP (English for Academic Purposes) and observing that even after completing intensive academic English programmes many L2 learners continued to be challenged with intelligibility, that is pronunciation of phonemic accuracy and overall oral delivery. Speaking components were frequently limited in classroom time, and rarely was pronunciation practice integrated into EAP curricula or connected to authentic communicative use. These recurring challenges pointed out the need for evidence-based approaches that can bridge targeted phonetic training with broader aspects of comprehensibility in speech production. In parallel, recent developments in educational technology, such as artificial intelligence (AI) and speech processing, have opened new possibilities for delivering adaptive, high-variability pronunciation training on technology-mediated environments. Such tools now make it feasible to recreate the acoustic variability central to HVPT through multi-speaker input, feedback, and learner-specific sequencing. Against this backdrop, the present study provides new empirical evidence on whether phonetic training, using functional load principles, can improve comprehensibility in spontaneous speech, an area that remains largely unexplored despite its clear pedagogical relevance.

Although a strong link between perception and production has been assumed (Baker & Trofimovich, 2006), present studies have not consistently shown that perception training on its own can induce sustainable changes in production, especially in spontaneous speech. As pointed out in recent works of meta-analyses, the effect of HVPT on L2 production is typically modest, with average gains in post-training for trained items around 10% and for untrained items even lower, and little evidence for robust long-term retention or generalization to truly novel communicative contexts (Uchihara et al., 2024); moreover, these production gains are commonly assessed in semi-controlled tests (such as word reading or repetition), not in free or spontaneous speech tests. Much of the existing work has focused on isolated phoneme discrimination or production in scripted tasks, which do not fully simulate the pressures and variability of conversation in authentic contexts (Barriuso & Hayes-Harb, 2018).

Furthermore, there is a scarcity of studies that examine HVPT outcomes longitudinally. Few studies include delayed post-tests, leaving open the question of whether perceptual and productive gains are retained over time. This is a significant limitation, as lasting sustainability is essential for durable communicative development. While some studies have demonstrated retention of perceptual gains for periods ranging from two weeks to several months (Lively et al., 1994; Bradlow et al., 1999; Uchihara et al., 2024), these findings often pertain to perception rather than to productive or communicative abilities. Additionally, most HVPT studies have involved a small number of phonemic contrasts and have not accounted for the communicative weight or functional load of those contrasts. Phonemes with high functional load, those that distinguish many word pairs are more likely to impact intelligibility and overall communication (Thomson, 2018). However, targeted research examining HVPT's impact on such contrasts within ecologically valid communicative contexts is still lacking.

To resolve these issues, a more ecologically valid and communicatively grounded method to assess HVPT is required. A growing body of research recommends that perceptual training should integrate tasks that more closely mirror naturalistic language use, including spontaneous and sentence-level production, as well as intelligibility and comprehensibility ratings by proficient speakers (Barriuso & Hayes-Harb, 2018). Only by linking perception training outcomes to these higher-level communicative constructs can we accurately assess HVPT's real-world efficacy.

Therefore, the rationale for the current study is twofold: first, to explore the transferability of perceptual learning to intelligibility and comprehensibility in spontaneous speech; second, to evaluate the durability of HVPT-induced gains through delayed testing. By examining both interrelated goals, the research entails in enhancing a more in-depth understanding of perceptual training on L2 learners and its influence in fostering authentic, intelligible, and comprehensible L2 speech.

This study also responds to recent pedagogical and technological advances in the language landscape. As learning contexts are becoming more prevalent in online and hybrid forms, there is a demand for scalable pronunciation interventions capable of delivering outside traditional classroom environments. HVPT provides a possible solution for self-directed learning and is adaptable to computer-assisted and web-based platforms (Iino & Thomson, 2018; Qian et al., 2018). However, its practical effectiveness within these settings, especially in terms of communicative gains, has yet to be fully established. By implementing an online HVPT program focused on high functional load contrasts, this study is designed to close the gap between laboratory evidence and real-world practice, investigating how digitally mediated, perception-based pronunciation training can support learners' communicative competence in dynamic language learning environments.

1.3 Aims and Research Questions

This study investigates the effects of High Variability Phonetic Training (HVPT) on second language learners' development of perceptual, intelligibility, and comprehensibility skills. Insights were drawn from two key domains: L2 perceptual training and the principle of functional load. It seeks to examine whether perceptual gains acquired through HVPT can lead to improvements in intelligibility (i.e. the accurate pronunciation of words in read-aloud tasks) and, subsequently, to gains in comprehensibility during spontaneous oral production (i.e. picture story and long-turn narrative tasks).

While HVPT forms the core instructional approach of this study, the design of the training stimuli was guided by functional load theory. Rather than treating functional load as a standalone framework, this study uses it as a principled criterion for selecting phonemic contrasts with the highest communicative value. By targeting phonemes that distinguish many words or frequently occur in communicative contexts, the training aims to maximize the likelihood that perceptual improvements will bring meaningful, listener-oriented outcomes in spoken interaction.

Moreover, the study further explores whether such improvements are retained over time by incorporating a delayed post-test. This longitudinal design provides a more thorough depiction of the trajectory from perceptual learning to communicative performance and aims to support the development of pronunciation instruction that emphasizes lasting, real-world communicative benefits.

This study is guided by two core aims:

1. To determine the extent to which HVPT training on high functional load phonemic contrasts improves learners' ability to perceive English phonemes, and whether these perceptual gains transfer to the intelligibility of word production and the comprehensibility of spontaneous oral productions.

2. To examine the retention of these gains in the long-term with a delayed post-test, contributing new data on the long-term efficacy of HVPT.

The study is guided by these questions:

RQ1: What is the impact of high variability phonetic training (HVPT) on high functional load phonemes on:
a) L2 learners' perception of phonemic contrasts?
b) the intelligibility of words read aloud by L2 learners?
c) the comprehensibility of L2 learners' spontaneous oral productions?

RQ2: To what extent are any gains in the above dimensions retained long-term?

This study extends the previous research by supporting integrated approaches to pronunciation instruction that go beyond isolated phoneme training. It addresses a significant knowledge gap in the present literature by bringing forth evidence that closely reflects authentic communication in real-world settings (Qian, 2018). The present project builds on the existing evidence that learner-centered HVPT, especially when combined with functional load analysis, can produce improvements in perception, segmental intelligibility as well as in global comprehensibility in spontaneous speech. It also puts forth the notion of a *Comprehensibility Transfer Pathway*, a framework capturing the cascading development from perception to intelligibility and ultimately to listener-rated comprehensibility in spontaneous oral production.

1.4 Context

This study was conducted in the post-COVID-19 era, a period that saw considerable shifts in L2 instruction toward alternative formats of online and hybrid use. Initially, the study had hoped to recruit Chinese learners of English enrolled in UK-based English language programs. However, due to pandemic-related restrictions, there were significant reductions in international student enrollment which led to difficulties in participant recruitment. As a result, the study was adapted to recruit learners from an English language program in China.

While all perceptual training along with testing were conducted online and thus unaffected in terms of delivery or access, this shift in context may have influenced participants' overall English exposure. L2 learner participants based in the UK would likely have had greater access to English-language environments outside the classroom, whereas participants residing in China remained immersed in a primarily L1 context. This contextual factor should be considered when interpreting the study's results, particularly the possibility of language exposure on phonological development beyond the experimental intervention (Jia et al., 2006; Handley & Wang, 2018).

The recent shift to training environments in online and remote forms, necessitated in part due to the global Covid-19 pandemic, introduces challenges and new opportunities for L2 pronunciation instruction. Reduced immersion in English-speaking environments may limit incidental exposure to target language input, making technology-mediated interventions like HVPT crucial for compensating for insufficient English language exposure and offering learners intensive, targeted practice regardless of location (Qian, 2018). Moreover, HVPT is particularly well suited to digital delivery because it can be automated and individualized. The present research thus offers both pedagogical and methodological insights into how perceptual-focused training can be implemented and assessed in online distance learning contexts that simulate real-world, autonomous language acquisition conditions.

Research on perceptual training has been predominantly addressing on a single phonemic contrast (Lively et al., 1993; Hattori & Iverson, 2009; Logan et al., 1991; Shinohara & Iverson, 2018) with minimal research training on more than one phonemic contrast. The present study addresses this limitation by incorporating multiple high functional load phonemic contrasts, some of which present attested difficulties (McAndrews & Thomson, 2017) for Mandarin speakers of English because of differences in L1 and L2 phonological inventories. This approach investigates beyond single-segmental contrast studies and validates the ecological validity and pedagogical importance of the observed effects of a wider range of pronunciation difficulties observed by this L2 learner population.

In brief, this study further explores how perceptual training may enrich L2 speech by examining the transfer from perception to intelligibility and comprehensibility across different communicative contexts. By further investigating these interrelated dimensions in Mandarin learners of English and incorporating considerations of functional load, the research intends to yield a deeper understanding into L2 phonological development and in oral production targeting improved communicative effectiveness.

1.5 Structure of thesis

The structure of this thesis is organized to build a coherent narrative from theoretical foundations to empirical supported findings and pedagogical implications. The following are the details of each chapter:

Chapter 1 introduces the background, rationale, and aims of the study, situating the research in relation to current issues in L2 pronunciation instruction and establishing the study's relevance.

Chapter 2 provides a literature review, covering the constructs of intelligibility and comprehensibility, theoretical models of speech perception and production, the principles of functional load, and empirical supported findings related to High Variability Phonetic Training (HVPT). This chapter also identifies key limitations in the recent literature and introduces the coined term Comprehensibility Transfer Pathway as the concept through which the study is designed.

Chapter 3 establishes the methodological foundation, including participant selection, training design, evaluation procedures, and rating methodologies, while justifying the study’s experimental choices with reference to both empirical rigor and ecological validity.

Chapter 4 introduces the results of the perception, intelligibility, and comprehensibility tests, using both descriptive statistics and inferential analysis to present the developmental changes across time.

Chapter 5 discusses the findings in relation to the defining research questions.

Chapter 6 concludes the thesis with a summary of the study’s significance, limitations, and finally its implications.

Chapter 2 Literature

2.1 Introduction

The chapter presents a detailed evaluation of theoretical frameworks and experimental evidence which supports the research conducted in this study. The main research objective investigates how high variability phonetic training affects the perception and production of high functional load phonemes and examines whether gains in intelligibility at the level of individual word transfer to comprehensibility in spontaneous oral production. The review methodically develops the research basis through an evaluation of essential theoretical concepts and experimental evidence related to L2 speech development.

The chapter starts by defining comprehensibility (Section 2.2) and intelligibility (Section 2.3) before moving on to pronunciation training approaches (Section 2.4) and functional load theory (Section 2.5) to determine its effects on phonemic contrast selection in pronunciation training. The chapter explores high variability phonetic training (Section 2.6) before presenting the chapter summary in Section 2.7. The review combines existing and contemporary research about speech perception and phonetic training and speech production to justify the investigation of perceptual discrimination transfer to word-level intelligibility and to comprehensibility in spontaneous oral production which serves as the main research question for this study.

2.2 Comprehensibility

2.2.1 Definition and Conceptualization

Comprehensibility refers to the measure of how well a listener can decode spoken content from L2 speakers in a "communication context" (Barefoot et al., 1993). In a more defined manner, the construct of comprehensibility expresses the degree to which a listener finds it easy or hard to understand the message of a speaker (Derwing & Munro, 2015). This differs from the speaker's pronunciation accuracy (intelligibility) or accentedness (Munro & Derwing, 1995). Research on second language acquisition has extensively focused on the three dimensions: accentedness, intelligibility, comprehensibility, but Munro and Derwing's research (1995) introduced comprehensibility as an independent speech dimension in second language acquisition. Overtime comprehensibility has come to be understood as a complex construct rather than being associated with only phonological terms. This development points to a new direction where emphasis in applied linguistics is on communicative effectiveness over formal accuracy (Levis, 2005). It is presently viewed as a construct that is determined by different linguistic components including accurate production of sounds; prosodic features such as stress; fluency; lexical diversity; accuracy and complexity in grammar and organization of ideas (Isaacs & Trofimovich, 2012; Crowther et al., 2015).

Some researchers have argued that comprehensibility is closely related with communicative adequacy (Isaacs & Thomson, 2013; Revesz et al., 2016). The notion of *communicative adequacy* is that language learners are successful to the extent that they achieve their communicative goals in context, including not just ease of understanding but more broadly the appropriateness, completeness, and effectiveness of the message for the situation (Isaacs & Trofimovich, 2012). The primary aim for comprehensibility and communicative adequacy is to succeed in communication. However, each construct approaches the main goal differently. Comprehensibility specifically emphasizes the listener's perspective, functioning as a proxy for the cognitive load involved in decoding speech whereas communicative adequacy includes broader pragmatic and interactional dimensions for successful real-world communication (Handley & Wang, 2023). From these perspectives, comprehensibility and communicative adequacy are viewed as two different constructs where some common features are shared as part of communication but features of comprehensibility measured at the utterance level (e.g. phrase or sentence) (Isaacs & Trofimovich, 2012) and communicative adequacy measured at a broader discourse level are unique to each construct. These features help define the constructs as two distinct constructs.

The most effective way to understand comprehensibility is that it exists within the context of a complete communicative system. This can be seen in the manner that comprehensibility is assessed. Commonly the assessment of comprehensibility uses Likert scales for subjective listener ratings (Munro & Derwing, 1995a) whereas intelligibility is commonly assessed through objective methods such as transcription tasks (Kang et al., 2018). The analytic rubrics developed by Isaacs and Trofimovich (2012) enable assessors to connect comprehensibility ratings with specific linguistic features which improves the diagnostic quality of such assessments. The assessment of comprehensibility has developed through mixed-methods research approaches (Saito et al., 2015) which integrate qualitative feedback to provide further information about listener perceptions. These assessments have gained ground as a methodological approach of language assessment for language learners and have been embedded into standardized speaking assessments as a category of speaking proficiency levels (Isaacs & Trofimovich, 2012)

The global acceptance of comprehensibility as a reachable target for language learners has also incorporated it into second language teaching methods. The educational approach now focuses on communicative effectiveness instead of native-like pronunciation achievement (Hincks, 2005; Derwing & Munro, 2005, 2015). The change in approach demonstrates increasing research evidence that specific instruction methods can enhance understanding even when accent remains present thus making it a realistic objective for adult learners in various communication situations (Derwing et al., 1998)

The construct of comprehensibility exists as a listener-based multidimensional construct which represents the perceived effort required for processing. The concept of comprehensibility has evolved from its initial phonological definition into a comprehensive measure of communicative success which differs from both intelligibility and accentedness. The growing importance of comprehensibility in educational settings and evaluation processes demonstrates the changing definition of successful L2 communication.

For the purposes of the present study, *comprehensibility* is defined as the degree of ease or effort experienced by a listener when interpreting a speaker’s message. Within this research, comprehensibility in spontaneous oral production is examined in relation to improvements in phonetic perception following High-Variability Phonetic Training (HVPT). This working definition also considers that while the study focuses on segmental-level effects, comprehensibility represents a broader communicative construct that underlies successful oral production. Thus, in this study, comprehensibility is treated as a multi-dimensional construct linking targeted phonetic gains to real-world speaking production.

2.2.2 From Broad Instructional Goals to Specific Objectives

The main broad goal of comprehensibility in second language (L2) instruction is to develop speech that listeners can easily understand regardless of the similarity to native-like pronunciation. The focus has shifted from achieving nativeness to achieving communicative efficiency because L2 learners gain more benefit from clear functional speech than from accent elimination efforts (Flege et al., 1995a; Levis, 2005). The increasing global use of English as a lingua franca in an interconnected world makes comprehensibility an important practical and inclusive goal for L2 learners. Studies in applied linguistics show that comprehensibility should be the main focus of pronunciation instruction because adult learners develop their phonological systems based on their first language (Flege et al., 1995b; Munro & Derwing, 2008) Learners should focus on developing speech that enables successful communication while reducing listener work and promoting mutual understanding instead of pursuing native-like speech.

This broad goal of comprehensibility translates into the following specific instructional and communicative objectives: (a) prioritize understandable speech - L2 learners should aim to practice and develop skills in oral communication. The goal should be to create speech that listeners can easily understand instead of striving for an unattainable native accent for adult learners (Derwing & Munro, 2009; Crowther et al., 2015a); (b) establish practical objectives - L2 learners must achieve communication abilities that produce effective speech for academic, professional and social contexts (Saito et al., 2022a) which sets a practical goal with important real-world value; (c) Develop multiple linguistic domains - multiple linguistic domains need to be developed by L2 learners since comprehensibility depends on segmental and prosodic features together with fluency and vocabulary and grammar. The focus of L2 learners should not be only on pronunciation but they should work to better their performance in all areas (Isaacs & Trofimovich, 2012); (d) Minimize listener process effort - the cognitive workload of listeners decreases when speech becomes more understandable which correlates precisely with comprehensibility as “perceived processing effort” (Isaacs et al., 2018; Munro & Derwing, 2015); (e) address and learn from communication breakdowns - the process of understanding and learning from communication breakdowns leads to better interaction efficiency and prevents confusion (Deterding, 2013; Crowther et al., 2023); (f) engage across diverse tasks and contexts - the ability of L2 learners to connect with cultural and linguistic environments more easily helps them develop social relationships and transition better to other cultures (Derwing et al., 2008; Crowther et al., 2015a); (g) increase self-awareness - the improvement of comprehensibility enables L2 learners to develop better self-awareness about their second language usage which is an objective and beneficial outcome for the L2 learner (Derwing & Rossiter, 2002; Isaacs et al., 2018).

However, these broad goals of comprehensibility have led to the identification of more specific instructional and developmental goals that support comprehensibility. These include designing pedagogical interventions that directly target linguistic features most strongly associated with listener processing ease such as segmental accuracy (Isaacs & Trofimovich, 2012), suprasegmental control (Hahn, 2004; Kang et al., 2010) and speech rate (Saito, 2021). Another narrower objective also involves developing tools and criteria to assess and monitor L2 learners' comprehensibility in realistic communicative tasks, including the use of listener ratings and audio-recorded self-assessment (Derwing & Munro, 2015). In applied contexts, comprehensibility-focused teaching supports learners in developing adaptive speaking strategies that adjust delivery based on listener response, including paraphrasing, repetition, and pausing (Kennedy & Trofimovich, 2010). To add to this, instructional models that emphasize comprehensibility encourage metacognitive development by helping L2 learners become aware of which speech features affect understanding and how to modify them effectively (Saito, 2021).

Overall, these narrower goals reflect a pedagogically grounded and learner-centered approach to pronunciation, one that supports incremental gains in communicative competence while extending the wider goals of communication in providing practical ways to improve real-world speech clarity. Despite the clear relevance of pronunciation to spoken communication, there is limited research in enhancing comprehensibility particularly in relation to instructional approaches. Although pronunciation is widely recognized as a key element of comprehensibility, few studies have investigated how targeted pronunciation instruction especially perceptual approaches like High Variability Phonetic Training (HVPT) can support L2 learners in becoming more understandable in spontaneous speech. HVPT is founded on the principle that improved perception leads to enhancement phonemic accuracy and intelligible production, suggesting a direct link between perceptual training communicative success. As such, this study’s focus on HVPT aims to focus on how pronunciation training may enhance comprehensibility in spontaneous L2 oral production.

2.2.3 Linguistic Features Influencing Comprehensibility

Studies conducted to determine which linguistic features in L2 speech affect native speakers' listening judgments in comprehensibility (Munro & Derwing, 1995; Isaacs & Trofimovich, 2012; Suzukida & Saito, 2019) have been numerous. The research studies have been important in informing which features of speech should be emphasized in pronunciation instruction in order to enhance communicative effectiveness. Importantly, this subsection is directly relevant to one of the major objectives of the current study: to investigate whether the benefits of High Variability Phonetic Training (HVPT) on phonemic perception generalize to spontaneous speech comprehensibility. To be able to interpret possible measurable outcomes of perceptual training, it is important to know which specific linguistic features facilitate comprehensibility in such contexts.

It has been established that comprehensibility is determined by a set of linguistic features which involve different linguistic features of speech, including segmental accuracy, prosody, fluency, lexical appropriateness and grammatical accuracy (Trofimovich & Isaacs, 2012). For example, Crowther et al., 2015a) reported that appropriate grammar and lexis usage was a significant predictor of comprehensibility. The study involved trained linguistic coders to make subjective judgments to determine the relationship between several linguistic dimensions and overall L2 comprehensibility. They reported that pronunciation, fluency, vocabulary and grammar scores were significant predictors of L2 comprehensibility, particularly in the case of cognitively demanding tasks (e.g., TOEFL iBT) compared to less cognitively demanding tasks (e.g., IELTS). This finding indicates that the relative contribution of different linguistic features to L2 comprehensibility can be influenced by the level of task complexity and cognitive demands.

Comprehensibility models based on hierarchies have been developed to understand the linguistic features that affect comprehensibility judgments. Jułkowska and Cebrian (2015) suggested that particular non-native segmental features can cause significant disruption to listener comprehension. Acoustic analysis has further refined the relationship between comprehensibility and certain segmental features. Research has determined certain vowel features that affect comprehensibility ratings and found that vowel reduction in unstressed syllables and the accurate production of tense-lax contrasts (e.g., /i/-/ɪ/) were the most important for English comprehensibility (Kang & Moran, 2014). More recent research has attempted to explain how different learners' proficiency level can impact comprehensibility judgments. Saito's (2021) comprehensive meta-analysis presents segmental accuracy and fluency as "threshold features" that must achieve a certain level of accuracy before other linguistic features such as prosody, lexical appropriateness, and grammatical accuracy as they each have an influential and significant role on comprehensibility. In other words, if a speaker's segmental accuracy and fluency are insufficient, improvements in prosody, vocabulary, or grammar may have limited impact on how easy their speech is to understand. These studies provide fundamental evidence why pronunciation especially segmental accuracy is a main area of focus for this current study.

Comprehensibility is also further influenced by listener variables. Studies have shown that comprehensibility ratings depend on the following listener factors: a) foreign accent familiarity (Gass & Varonis, 1984), b) training in linguistic skill development, c) pedagogical experience (Isaacs & Trofimovich 2012), and d) a stronger metalinguistic awareness of second language learners' comprehensibility over accentedness (Saito et al., 2019). Gass and Varonis (1984) demonstrated that listeners who are more familiar with foreign accents tend to rate L2 speech as more comprehensible, indicating that familiarity with the speaker's L1 accent can lessen perceived communication barriers. Listener training and experience are also influential listener variables in comprehensibility. Studies showed that listeners with training in linguistic skills or pedagogical backgrounds are more attuned to a broader range of linguistic cues, leading to more awareness of segmental differences and sometimes more generous comprehensibility ratings (Isaacs & Trofimovich, 2012). The research demonstrates why listener background matters in comprehensibility studies. This study involved untrained listeners to produce authentic real-world judgments while preventing pedagogical or metalinguistic bias.

Other research areas have also been concerned with speech rate as another linguistic factor that affects comprehensibility. Munro and Derwing (2001) indicated that comprehensibility ratings formed a curvilinear pattern in relation to speech rate, that is, both very slow and very fast speech were rated as less comprehensible, suggesting an optimal mid-range rate for maximizing comprehensibility. Similar results were obtained in other research studies (Wright & Tavakoli, 2016; Thomson, 2018). Understanding how speech rate interacts with other features such as pronunciation accuracy helps contextualize listener judgments across different task types used in the current study, each reflecting varying degrees of cognitive and communicative demand. The present study does not measure speech rate directly but its documented influence across task types shows its importance for understanding how a speaking context may affect comprehensibility results, although the study 's investigation is on listener ratings of spontaneous oral productions.

Considering the influence that linguistic features may have on listeners, this study aims to establish whether perception improvements resulting from HVPT would produce detectable enhancements in L2 comprehensibility.

2.3 Intelligibility

2.3.1 Definition and Conceptualization

Research into second language pronunciation centers on *intelligibility* as a key concept to describe the degree at which listeners correctly receive “a speaker's intended message” (Munro & Derwing, 1995; Derwing & Munro, 1997). Word or an utterance recognition constitutes a part of intelligibility definition because it describes listeners' skill to correctly detect linguistic forms spoken by speakers (Smith & Nelson, 1985; Kenworthy, 1987; Yorkston et al., 1996). These definitions of intelligibility place the measurement of actual message reception as an objective of communication success while comprehensibility relates to the perceived “effort required by the listener to understand the speaker” (Derwing & Munro, 2005). This difference controls the assessment of L2 speech and determines the appropriate methods for pronunciation education (Kang et al., 2020) where comprehension is assessed with comprehensibility scores and intelligibility with transcription accuracy (Barrass et al., 2020), and word recognition tests (Sheppard et al., 2017).

Intelligibility Principle proposed by Levis (2005), prioritizes teaching of pronunciation in making learners being understood by listeners as opposed to focusing on accent elimination. The principle marks a fundamental change from the *Nativeness Principle* which strives at native-like pronunciation and aims at communicative success over phonological similarity (Jenkins, 2000; Thomson & Derwing, 2015). Intelligibility involves recurring involvement between speaker-listener and various influential context factors. Of particular importance though is that the degree of intelligibility not only depends on influential context factors and on accurate phonological production of the learner but also on the listener's familiarity with accents; as well as a range of perceptual processes (Field, 2005). Further to this, Jenkins (2000) shows through their thorough investigation of communication among non-native speakers that significant phonological elements are needed to preserve intelligibility in lingua franca communication.

Intelligibility has developed as a concept throughout the years. The field of speech pathology first defined intelligibility as the number of words listeners correctly identify (Yorkston et al., 1996), later intelligibility expanded to include further language units such as syllables and connected speech among others (Gooskens et al., 2010). Early research centered on segmental precision, yet subsequent investigations showed that suprasegmental elements like lexical stress (Benrabah, 1997; Field, 2005), sentence stress (Hahn, 2004), rhythm (Tajima et al., 1997; Low, 2006; Kang et al., 2010), and intonation (Pickering, 2001) are also fundamental features in achieving intelligibility. Research studies on intelligibility have shown its methodological approaches by investigating approaches beyond basic transcription activities. Researchers developed multiple evaluation tools including multiple-choice comprehension questions (Hahn, 2004) together with cloze tests (Kang et al., 2018) forced-choice identification tasks (Thomson, 2018) and sentence verification tasks (Gass & Varonis, 1984). The analytical scope of intelligibility assessment has grown from word and sentence to discourse level through methods including passage comprehension (Zielinski, 2008) and in some studies oral retelling or summary tasks (Kang et al., 2010). Research has also explored intelligibility and its relation to other speech dimensions. Munro and Derwing (1995) study showed that intelligibility stands separately from accentedness and comprehensibility although the three are interconnected; speech can remain heavily accented yet very intelligible. The discovery holds important teaching implications because it shows that all accent characteristics do not have the same influence on speakers' communication (Levis, 2005; Crowther et al., 2015b).

Intelligibility consists of multiple dimensions which combine both perceptual accuracy and communication. Importantly, its transition from native-like standards to intelligibility-based instruction demonstrates the increasing priority of practical communication in L2 learning. In this study**,** *intelligibility* is defined as the extent to which a listener is able to accurately recognize a speaker's intended words. It is examined in relation to improvements in phonetic perception following High-Variability Phonetic Training (HVPT). While this definition places emphasis on word-level accuracy, it also considers that intelligibility represents a wider communicative construct that enables successful message transmission in oral production.

Next, the objectives for intelligibility in second language education follows.

2.3.2 From Broad Goals to Specific Objectives in Pronunciation Intelligibility

The overarching goal of intelligibility research in L2 pronunciation teaching is to achieve correct perception of L2 learners' words and sounds by listeners. This goal places importance on identification of phonological segments, phonemes, and individual lexical items rather than overall message comprehension. This represents a new direction from the importance to gain native pronunciation and places importance on listeners identifying speech elements at the level of individual phonemes and words. This goal can be divided into two main strands which are further discussed in detail.

These specific learning objectives describe what learners should *achieve* in phonological intelligibility: (a) identifying critical phonemic features - this objective is to identify the specific segment, phonemic contrasts, and sound segments most significantly impact word recognition. With this objective in mind, systematic research led to the creation of frameworks such as Lingua Franca Core (see section 2.3.4) and the functional load approach (see section 2.5.1) to signify the most important consonants, vowels and phonemic distinctions for maintaining basic word intelligibility (Jenkins, 2000; Munro and Derwing, 2006); (b) realistic phonological targets – this is to determine achievable sound targets for L2 learners that focus on sound intelligibility instead of native-like pronunciation. The key point is that even if speakers maintain their accent, but are able to develop phonemic distinctions, listeners can accurately identify words even when the pronunciation is not native. (Levis, 2005; Saito, & Plonsky, 2019); (c) ensuring context-specific sound intelligibility –the objective is for learners to improve accurate production of phonemes for their specific communicative contexts. Producing accurate phonemes is necessary for word intelligibility and are dependent on factors such as the acoustic environment and the phonological repertoire of typical interlocutors (Pennington, 1999; Tavakoli & Cooke, 2024).

These teaching and assessment objectives describe the learning goals that are supported through instructional practices and assessment tools: (a) assessment approaches for word-level –this refers to the development of valid assessment methodologies that specifically measure whether listeners can correctly identify each lexical item and sounds produced by L2 speakers. This includes controlled word-recognition tests, minimal pair discrimination tasks, and other focused measures of segmental accuracy in pronunciation (Isaacs, & Trofimovich, 2012; Harding, 2017) (b) segmental teaching practices – a practical objective to identify instructional approaches that most effectively enhance the intelligibility of specific problematic sounds. Instructional and practical methods for improving the production of crucial phonemes that affect word recognition are identified through methods such as minimal pair drills, phonetic training, and segmental feature feedback (Derwing et al., 1998; Kissling, 2013).

These specific objectives support the overall broader goal of developing L2 pronunciation that enables accurate word recognition, the foundation upon which higher-level comprehensibility can be built.

2.3.3 Linguistic Features Influencing Intelligibility

Research shows that specific phonological elements in the speech of L2 speakers play a major role in determining how easily listeners can understand them. This research indicates that particular phonological features should be emphasized in pronunciation training to enhance communication.

Intelligibility is greatly influenced by the prosodic features of stress, rhythm and intonation. Field (2005) found that lexical stress errors can cause word recognition failure even when segmental accuracy is achieved. The study demonstrates that prosody plays a crucial role in marking word boundaries and focusing on essential information in the speech signal. Further studies Hahn (2004) reported that the proper placement of sentence stress (nuclear stress) significantly improved L2 speech intelligibility for native speakers indicating that prosodic elements may be as vital as segmental accuracy for intelligibility. However, several studies and review articles argue that segmental pronunciation (the correct production of individual consonant and vowel sounds) is fundamentally important, sometimes even positing it as the "foundation" for other linguistic features such as suprasegmentals (stress, intonation, rhythm), and that intelligibility and communication break down without segmental intelligibility even if suprasegmental features are well-produced (Jenkins, 2000; Bent et al., 2007a). Studies on Mandarin speakers of English found that 76% of the variance in their intelligibility was explained by seven segmental features (Rogers & Dalby, 2005) which infers that suprasegmental features cannot compensate for segmental errors that render a word unrecognizable. Thus, intelligible segmentals are a prerequisite for other pronunciation features to matter (Saito, 2011).

The placement of segmental errors in words and sentences might affect their contribution to intelligibility. Jenkins (2000) reported that consonant errors placed at the beginning of the word negatively impacted intelligibility than consonant errors placed at the end of the word in English. Zielinski (2008) also found that errors on stressed syllables affected word recognition more than errors on unstressed syllables suggesting a connection between segmental accuracy and prosodic features.

Manner of articulation consonant contrasts seem to have the most significant effect on the intelligibility of the speech. Deterding (2013) investigated communication breakdowns in English as a lingua franca interaction; and found that confusions between stops, fricatives, and approximants (e.g. /b/ vs. /v/, /r/ vs. /l/) were common sources of misunderstandings. This result is consistent with acoustic-phonetic studies that demonstrate that manner features are reliable indices of phoneme identity in difficult listening conditions (Wright & Tavakoli, 2016) and thus are critical for maintaining intelligibility.

The effects of vowel errors on intelligibility are generally less than those of consonant errors (Bent et al., 2007; Qian et al., 2018), and similar findings were found on comprehensibility (Munro & Derwing, 2006; Isaacs & Trofimovich, 2012), but certain vowel contrasts play an essential role in understanding. The following studies (Catford, 1987; Jia et al., 2006; Bent et al., 2007b; Munro & Derwing, 2006) demonstrated that tense-lax and front-back confusions (e.g. /i/ vs. /ɪ/) were the most critical for intelligibility. The results support the acoustic-phonetic study of vowel perception (Hillenbrand et al., 1995) which indicates that confusions in primary acoustic dimensions for vowel identification are most detrimental to intelligibility. As a results of these findings, it is recommended that pronunciation teaching should focus on phonological features that most affect intelligibility for specific learners in their target communicative contexts.

2.3.4 Intelligibility from a Theoretical Perspective

Theoretically, intelligibility is now more commonly understood as a context-dependent construct influenced from the communicative exchange between speaker, listener, and context. Contemporary research increasingly positions intelligibility within broader communicative theoretical models concerned with the evolving interaction between the interlocutors and the communicative context. Rather than assuming that intelligibility is determined solely by the accuracy of pronunciation, these models argue that it is influence by multiple "linguistic, cognitive, and social variables" (Smith & Nelson, 1985; Field, 2005; Zielinski, 2008; Lindblom, 1990; Trofimovich & Isaacs, 2012; Kang et al., 2018). Three theoretical perspectives are directly aligned with aims of this research and are discussed below.

The development of the *Lingua Franca Core* (LFC) arose from continuous discussions about the goals of pronunciation teaching for international users of English. Jenkins (2000, 2007) proposed the LFC to enhance communication among global users of English where spoken interaction often occurs between non-native speakers. She argued that native-speaker norms were neither necessary nor attainable for most second language learners of English. Her work redirected attention toward promoting effective spoken interaction among L2 users as a realistic and pedagogical goal. Moving away from native-like articulation, the LFC framework identifies specific phonological features that play a critical role in intelligibility, including consonant distinctions (excluding interdental fricatives), vowel duration, and appropriate word stress (Jenkins, 2000; Walker, 2010; Kang, 2013). Features such as native-like rhythm, variation in pitch, or exact vowel quality are considered less essential. This theoretical shift has relevance to pedagogical practices as it encourages teachers and directs learners' awareness to phonemic contrasts that are effective in supporting intelligibility over native-like articulation (Munro & Derwing, 2006; Sakai & Moorman, 2018; Saito & Plonsky, 2019). While a number of researchers argue that the LFC requires further empirical testing using targeted research methods to validate which phonological features are essential for intelligibility across a broader range of ELF speakers, rather than assuming a one-size-fits-all core (Pickering, 2006; Haslam & Zetterholm, 2016); the LFC model remains influential for its focus on intelligibility-based pedagogy and its specification of important features for enhanced communication.

In the framework of the present study, functional load principles (Brown, 1988) directly guided the selection of high-impact phonemic contrasts, while the LFC provided a complementary theoretical perspective that supported prioritising phonemic features most relevant to intelligibility in global communication. Certain of the contrasts selected for the present study such as /l–r/ and /m-n/ also correspond to consonant distinctions identified within the Lingua Franca Core as essential for word meaning and intelligibility in international English communication. These phonemes carry a high functional load in English, meaning that confusion between them tends to lead to lexical misunderstandings (Catford, 1987; Sewell, 2017). By focusing perceptual training on contrasts central to intelligibility, the study extends the underlying principles of the LFC into an empirical context, testing whether perception-based improvements transfer to enhanced comprehensibility in spontaneous speech. In this way, the LFC complements the study's theoretical orientation towards intelligibility-focused instruction, while functional load principles specifically informed the choice of contrasts. The LFC thus serves as a supportive framework that places the study within current discussions on intelligibility and communicative success in spontaneous, listener-dependent speech beyond controlled task settings.

In parallel, usage-based theories of language learning offer a developmental and cognitive explanation for how intelligibility evolves. These theories suggest that language competence emerges through repeated and contextualized language (Bybee (2001), Ellis (2002, 2012), and Tomasello (2003). According to this view, speech sound representations become more longstanding and retrievable when learners are repeatedly exposed to these forms and produce them in communicatively relevant situations. From this perspective, intelligibility is an outcome of correct phoneme articulation and the result of accumulated experience with the sound system in use (Munro & Derwing, 1995; Derwing & Munro, 2015; Saito & Akiyama, 2017). Perceptual training, especially in the form of High Variability Phonetic Training (HVPT), is closely consistent with this model as it offers learners repeated exposure to critical contrasts across different phonetic environments and speaker conditions. This consistent exposure with target contrasts may strengthen the perceptual encoding of difficult contrasts and lead to more intelligible output over time.

Of comparable significance are attention-based models of speech perception that draw attention to cognitive processing functions underlying listener comprehension. The *Ease of Language Understanding* (ELU) model (Rönnberg et al., 2019) demonstrates a compelling account of how processing effort varies depending on the match between incoming speech and the listener's stored phonological representations. When this match is high, as in the case of familiar accents or well-pronounced speech, processing is fast and automatic. However, when speech input differs from what the listeners anticipate (e.g unfamiliar accents or mispronunciations) they must recruit additional cognitive resources to interpret the message. This added mental effort can negatively impact listener's understanding, even if the speaker's articulation is broadly accurate (Derwing & Munro, 2005; Adank et al., 2009). In this light, perceptual training may reduce listeners' processing load by helping L2 speakers produce speech that is easier to decode, and in turn support more efficient communication.

Taken together, these three theoretical perspectives provide a multidimensional framework for understanding intelligibility and offer important contributions for instructional design. These frameworks not only justify the necessity of perceptual training, but they also guide the design of interventions such as high variability phonetic training with the purpose to improve intelligibility and support broader communicative success in L2 development.

The next section introduces various types of pronunciation training before turning to High Variability Phonetic Training (HVPT) as a specific form of perceptual instruction.

2.4 Pronunciation Training

Pronunciation training involves a variety of instructional methods designed to improve L2 learners' accuracy in production of individual segments and communicative effectiveness. These methods differ in their conceptual foundations, delivery strategies, and the particular speech features they address. The following section reviews the main types of pronunciation instruction commonly applied in second language learning, examining the development of learners' spoken performance. Special emphasis is placed on perceptual training as it is the one of the main aims the present study.

2.4.1 Types of Pronunciation Training

The development of L2 pronunciation pedagogy has been influenced by many different instructional methods which provide distinct perspectives about how learners learn to produce more accurate and intelligible speech. The various methods used in L2 pronunciation instruction have particular theories and pedagogical techniques which help frame the role of perceptual training via High Variability Phonetic Training (HVPT). The subsections that follow discuss the main types of pronunciation training and explains the study's training intervention selection of the present study.

The main focus of Articulatory Training is to help learners gain control over their speech production mechanisms through speech-motor coordination. Learners gain better understanding of the production of speech sounds through explicit instruction about the placement and movement of articulators such as the tongue, lips and jaw (Aliaga-Garcia & Mora, 2009; Saito & Plonsky, 2019). The use of sagittal diagrams, ultrasound imaging and real-time visual feedback (Li, 2015) helps learners to be more aware of placement and movement. Pronunciation training targeting articulation is well-suited for segmental production errors, but it tends to neglect the perceptual aspect of phonological development. L2 learners who have difficulty in distinguishing non-native phonemes due to L1 interference, articulatory training may not have adequately target the perceptual challenges (Hazan et al., 2005).

Articulatory training can support production accuracy however the training may be insufficient in supporting perceptual sensitivity which is central to the aims of the current study.

Production-Based Training relies on repeated oral practice and communicative tasks to reinforce accurate pronunciation. Typical activities include minimal pair drills, repetition, shadowing, and scripted dialogues, often informed by skill acquisition theory (DeKeyser, 2007). These types of training aim to develop procedural knowledge and automatization of target forms (Lyster, 2007; Saito et al., 2020). However, they often presuppose that learners have already perceived phonemic contrasts accurately. When learners lack perceptual clarity, such as in cross-linguistic phonological conflict, production-based methods may be insufficient (Bradlow et al., 1997; Nagle, 2017). Since the aim of the current study is to address phonemic perception directly, this production-oriented model is not employed.

Technology-Enhanced Training incorporates digital medium to deliver multimodal pronunciation instruction. Computer-Assisted Pronunciation Training (CAPT) platforms provide learners with acoustic visualizations, speech recognition, and individualized feedback (Hardison, 2004; Thomson, 2018). Undeniably, training on web-based programs and using mobile applications programs make training more convenient and adaptable, especially relevant for online research environments. Although both types of online programs integrate both articulatory and perceptual elements, their innovation primarily lies in delivery format. The delivery format in the present study is through a digital platform but its purpose of using technology is different as it used as a digital medium to facilitate perceptual training specifically High Variability Phonetic Training (HVPT). The central aim is not to evaluate the technological medium itself, but to examine whether perceptual learning delivered online can facilitate transfer to improved comprehensibility in spontaneous oral production.

The unified instructional frameworks of Integrated Approaches combine perception-oriented and production-focused strategies. These approaches use the close connection between perception and production to help learners move through instruction stages starting with phonological analysis followed by scaffolded exercises and ending with communicative implementation (Celce-Murcia et al., 2010). The complete framework provides substantial educational benefits yet creates challenges for researchers who want to identify specific perception-focused training effects. Given that the present study investigates whether exposure to perceptual input alone through HVPT can produce measurable improvements in intelligibility and comprehensibility, models that are integrated may not be ideally for distinguishing perceptual effects from other instructional variables.

Each of these approaches contributes to L2 pronunciation instruction, yet each also presents limitations when the present study's aim is to isolate and assess perceptual learning outcomes. Articulatory and production-based models prioritize speech output but these models may not provide an emphasis or further support to resolve the perceptual foundations such as detecting phonetic differences (Flege, 1995a; Best & Tyler, 2007) mapping unfamiliar sounds onto existing or new phonological categories (Saito, 2020, Li, 2015) and establishing perceptual representations of L2 phonemes that underlie accurate pronunciation (Flege, 1995a; Bradlow et al., 1997; Huensch & Tremblay, 2015). Technology-enhanced training offers accessible and scalable platforms, which was applied in this study to deliver perceptual training in an online environment. However, the technological tools were not the object of investigation; instead, they functioned as a means to facilitate HVPT. Integrated approaches merge perception and production elements which makes it difficult to establish the effects of perceptual training. Therefore, the present study adopts an HVPT framework which provides online perceptual training with an aim to evaluate whether perception-focused input can independently support gains in intelligibility and comprehensibility.

Taking the above into account, the current study turns its attention to perceptual training as a primary intervention. Specifically, it examines the efficacy of High Variability Phonetic Training (HVPT), a method designed to enhance phonemic perception and test whether these perceptual gains transfer to intelligibility and comprehensibility particularly in spontaneous speech. The following subsection (2.4.2) explores the rationale and mechanisms of perceptual training in greater detail.

2.4.2 Perceptual training

Perceptual training refers to structured instructional interventions that aim to enhance second language (L2) learners' ability to accurately perceive phonetic distinctions in the target language. The central premise underlying this approach that perception often precedes and facilitates accurate production has become a foundational concept in contemporary L2 phonological development research (Flege, 1995a; Best & Tyler, 2007; Iverson et al., 2012). The present study focuses on High Variability Phonetic Training (HVPT), a perceptually based training paradigm, and investigates whether perceptual learning can transfer to intelligibility and comprehensibility in spontaneous speech. This section thus builds a focused argument for the theoretical and pedagogical importance of perceptual training, setting the groundwork for the implementation of HVPT in the current study.

A growing body of evidence supports the notion that persistent pronunciation difficulties in L2 learners often stem from perceptual, rather than articulatory, limitations. Adult learners in particular are constrained by entrenched L1 phonological categories, which shape how they perceive and categorize non-native sounds (Flege, 1995a; Iverson et al., 2005). Because of this, interventions that target perception directly, especially those using varied, naturalistic input, have been shown to facilitate more stable and accurate L2 phonemic representations (Bradlow et al., 1997; Thomson, 2011; Huensch & Tremblay, 2015).

While some studies have questioned whether perception development is a necessary prerequisite for production (e.g., Gick et al., 2008; Olson, 2014), an increasingly accepted view is that both perception and production interact in the phonological development (Bradlow et al., 1997; Baese-Berk, 2019). For the present study, which seeks to determine whether perceptual leads to improvements in comprehensibility during spontaneous speech, this bidirectional model offers critical justification for focusing on perceptual training as an independent variable.

One of the key advantages of perceptual training is the potential to generalize beyond the training context a phenomenon otherwise referred to as *transfer*. Pruitt et al. (2006) identify seven levels of transfer that provide a framework for evaluating the efficacy of perceptual training. There is transfer to:

1. new tasks – such as moving from discrimination to identification tasks
2. new instances from the same speaker – applying learned distinctions to novel stimuli
3. new talkers – applying perceptual skills across a range of speakers and accents

4. new phonetic contexts – distinguishing contrasts in different diverse segmental environments
5. new syllabic environments – recognizing sounds in different word positions
6. new contrasts of the same feature – applying acquired phonological structures to similar phonological distinctions
7. connected speech – applying skills learned in isolated word contexts to fluent, real-time communication

These transfer conditions are especially relevant to the current study, which aims to determine whether perceptual learning achieved through HVPT can extend to intelligibility and comprehensibility during spontaneous speech production. Such transfer would indicate that learners have not simply memorized stimuli but have internalized phonological contrasts robustly enough to apply them flexibly in authentic communicative contexts.

Another rationale for including perceptual training in L2 pronunciation instruction is that it aids in developing intelligible, comprehensible speech which is also the primary concern of listener-oriented models of communication (Bradlow et al., 1997; Derwing & Munro, 2005; Levis, 2005; Thomson, 2011). Traditional pronunciation training tends to emphasize native-like articulation as the main objective. However, recent frameworks such as Levis's (2005) Intelligibility Principle prioritize listener comprehension over accent reduction. According to this view, training that promotes perceptual sensitivity to phonemic contrasts is more beneficial than training that aims at imitating native speaker norms (Thomson & Derwing, 2015; Qian, 2018). This orientation is particularly important in multilingual spoken contexts where intelligibility, not accent, determines communicative success.

Furthermore, perceptual training offers a promising way to address age-related limitations in learning new phonological contrasts. Although adults commonly possess less neural flexibility than children, targeted approaches like High Variability Phonetic Training (HVPT) can still help them improve by retraining their attention and auditory processing systems, enabling meaningful gains in second language speech perception (Lively et al., 1993;Kuhl et al., 2008). In the context of the current study, this makes HVPT a particularly promising method for adult L2 learners who are challenged with specific phonemic contrasts due to L1 interference. Neuroimaging studies support these findings by showing that changes in "phonological processing areas of the brain are related to the neural activity associated with training speech perception skills" (Callan et al., 2003; Sakai & Moorman, 2018). These findings provide neurological support for the idea that perceptual training can alter the way second language sounds are represented in the brain which result in refined phoneme production and improved listener comprehension.

Although perceptual training may take different forms, High Variability Phonetic Training is one of the most rigorously studied and widely adopted models for perceptual training (Alliaga-Garcia et al., 2009; Li, 2015). By systematically having L2 learners exposed to target contrasts produced by different speakers and situated in diverse phonetic contexts, HVPT promotes the abstraction of invariant acoustic cues (Logan et al., 1991; Lively et al., 1993). This encourages the formation of speaker-independent phonemic categories a process that is believed to be essential for generalization and long-term retention.

However, despite the strong theoretical and empirical foundations for perceptual training, further research is needed to determine HVPT long-term effects and real-world applicability. While many studies report immediate post-training improvements in perceptual phoneme discrimination and, in some cases, production, fewer have examined the extent to which these improvements persist over time or transfer to spontaneous, unscripted speech (Uchihara et al., 2024). The current study builds on prior work by extending analysis to a delayed post-test and by evaluating listener-rated comprehensibility in free speech tasks, thereby offering empirical support for HVPT's effects beyond controlled and semi-controlled tasks.

2.4.3 Speech Perception Theories

The rationale supporting perceptual training within L2 pronunciation instruction, particularly High Variability Phonetic Training (HVPT), is well supported by theoretical models that explain how learners discri40minate and classify unfamiliar speech sounds. Specifically, they explain why L2 learners are challenged with certain phonemic contrasts, how perceptual categories are formed or blocked, and how structured input, as seen in HVPT, may assist in the reorganization of perceptual systems in ways that improve intelligibility and comprehensibility.

Two of the most influential frameworks are the *Perceptual Assimilation Model* (PAM) and the *Speech Learning Model* (SLM). These models inform the design of perceptual training interventions by providing mechanisms for understanding how L1–L2 phonetic relationships affect acquisition, and how targeted training can reduce L1-L2 perceptual challenges.

The Perceptual Assimilation Model (Best, 1995; Best & Tyler, 2007) proposes that learners interpret non-native phonemes to similar phonemes found in their native language. Depending on how L2 sounds are influenced by nearby sounds to L1 categories whether to two different categories (Two-Category), to the same category with different goodness-of-fit ratings (Category-Goodness), or to a single category with similar fit (Single-Category) learners will experience varying levels of perceptual difficulty. For instance, Japanese learners that are challenged with the English /r/-/l/ contrast have been interpreted using PAM as resulting from Single-Category assimilation to the Japanese /r/ (Best & Strange, 1992; Takagi, 1993; Aoyama et al., 2004). In this case, PAM-L2 predicts that, with increased exposure and experience, L2 learners are able to change their assimilation patterns and form new, more accurate phonological categories. HVPT is designed to foster this change by exposing learners to L2 contrasts across multiple speakers and phonetic environments, thereby encouraging attention to the acoustic dimensions that define category boundaries even when those dimensions differ from those in their L1 (Best & Tyler, 2007).

The Speech Learning Model (Flege, 1995a; Flege & Bohn, 2021) offers complementary insights. SLM asserts that the new phonetic category formation relies on the perceived phonetic distance between the L2 sound and the closest L1 equivalent. If the L2 sound is perceived as sufficiently distinct, a new category is likely to form; if not, it will often be "merged" with an existing L1 category. Crucially, SLM maintains that adults retain the capacity to form new categories, supporting the inclusion of adult participants in perceptual training studies as demonstrated in this study. SLM suggests that training should highlight subtle phonetic differences that L2 learners might otherwise overlook due to L1 interference. HVPT confronts this issue by exposing L2 learners to variation in input such as different speakers and contexts. This helps learners to not rely on speaker-specific features and instead focus on the core phonetic cues that distinguish each sound across voices. (Lively et al., 1993;

Bradlow et al., 1997; Logan et al., 1991; Iverson et al., 2005), thereby facilitating robust category formation (Thomson, 2018; Barriuso & Hayes-Harb, 2018; Sakai & Moorman, 2018).

The Automatic Selective Perception (ASP) model (Strange, 2011) adds a focus on attentional processes. ASP argues that native listeners develop automatic routines to filter speech input, privileging only the acoustic cues relevant for L1 distinctions. When processing L2 input, these routines can result in neglecting cues critical for distinguishing L2 phonemes. As such, perceptual training is not only as exposure, but a cognitive retraining that helps learners redirect attention to new, linguistically relevant dimensions. HVPT is an appropriate method for this purpose, as it reshapes the processing of L2 sounds by exposing learners to a wide range of unfamiliar tokens that disrupt L1-based expectations (Strange, 2011).

Taken together, these theoretical models provide a foundation for the present study's focus on HVPT and its role in facilitating L2 learners' perception, intelligibility, and comprehensibility, particularly in spontaneous speech. They also help explain why a perceptual training approach incorporating both identification and discrimination tasks, as in this study, may be especially effective: it engages with both categorization of new sounds (PAM, SLM) and the attentional realignment necessary for recognizing unfamiliar acoustic cues (ASP).

The previous section presented a range of pronunciation training approaches including articulatory and perceptual methods and discussed key theories of speech perception, the effectiveness of such training interventions often depends on the selection of phonemic contrasts they target. The following section introduces functional load theory as a principled approach to identifying contrasts that are most likely to influence intelligibility and comprehensibility in L2 oral production.

2.5 Functional Load

Phonology defines *functional load* as the degree of importance that phonemic contrasts hold for maintaining meaning distinctions in spoken language. Jakobson (1931) and Trubetzkoy (1939) first introduced the concept which Martinet (1952, 1955) expanded; followed by Hockett (1955, 1967); and then King (1967) developed methods to measure phonemic opposition for functional load (Brown, 1988; Sewell, 2017; Lin, 2019). The work of Catford (1987) and Brown (1988, 1991) brought functional load back into educational focus because they demonstrated that teaching pronunciation should target the most important communicative contrasts which subsequent research (i.e. Munro & Derwing, 2006) has confirmed. The concept has received modern research attention which shows the value of functional load for instructional guidance especially in perceptual training methods such as High Variability Phonetic Training (HVPT) that use high-functional-load contrasts to achieve maximum intelligibility and communicative effectiveness (Munro & Derwing, 2006; Sewell, 2021).

Section 2.5 - 2.5.3 builds on foundational work of functional load and the role it has within the context of perceptual training, particularly High Variability Phonetic Training (HVPT), and its function in guiding the training intervention of this study.

2.5.1 Theoretical Foundations of Functional Load

The notion of functional load emerges from structuralist and functionalist views of language as an efficient system, where not all phonemic distinctions are equal in weight in communicating their meaning. Martinet's (1955) principle of economy suggested that "phonological systems strive to balance" maximal communicative efficiency with minimal articulatory effort (Lin, 2019). In this view, phonemic contrasts that distinguish many words or occur in high-frequency words, carry greater functional load and are more resistant to loss or merger over time (Jakobson, 1931; Martinet, 1952; Sewell, 2017).

From an information-theoretic standpoint, functional load can be quantified as the "informational entropy lost" when two phonemes are neutralized (Hockett, 1955; Surendran & Niyogi, 2003). Phonemes that appear in more frequent contexts contribute more to a language's communicative efficiency. This changes the perspective of the structural concern of phonemes to assessments of communicative use (Sewell, 2017).

In pronunciation pedagogy, the notion of functional load was notably operationalized by Catford (1987) and Brown (1988), who proposed that errors in high functional load contrasts such as /p/-/b/ or /i/-/ɪ/ are more detrimental to intelligibility than errors in contrasts like /θ/-/ð/, which distinguish relatively few minimal pairs. Their claims were the beginning for what Levis (2005) termed the Intelligibility Principle, emphasizing that intelligibility, rather than native likeness, should be the goal of pronunciation instruction. This view has crucial significance with the aims of the current study, which seeks to measure intelligibility and comprehensibility outcomes from training that explicitly targets high functional load contrasts.

Although functional load identifies which phonemic contrasts carry greater communicative importance within a language, it does not directly predict which contrasts will be difficult for second language learners. In this respect, contrastive analysis offers a complementary viewpoint to functional load, approaching the issue from another angle. It compares the phonological systems of a learner's L1 and the target L2 to identify possible difficulties arising from the absence or difference of specific sounds (Lado, 1957; Wardhaugh, 1970; Eckman, 2008). In the present study, contrastive analysis is not used as a predictive framework but as a linguistic reference, acknowledging that some of the high functional load contrasts selected (e.g., /l–n/, /æ–ʌ/) have also been consistently reported in the literature as challenging for L1 Mandarin speakers (Jia et al., 2006; Qian et al., 2018). This distinction shows that while contrastive analysis highlights where difficulties may occur, functional load determines how much these contrasts matter for intelligibility.

While the theoretical background of functional load is widely accepted, fewer studies have explicitly examined the pedagogical consequences of training multiple high functional load contrasts simultaneously, especially using HVPT. Much of the literature has focused on one or two contrasts (e.g., /r/-/l/, /i/-/ɪ/) in isolation (Hwang & Lee, 2016). However, the communicative value of training multiple contrasts that individually carry high functional load remains an under-explored but an important area of this study. This study addresses that gap.

2.5.2 Operationalizing Functional Load

The concept of functional load has had an important role in theoretical and applied linguistics, by providing a method for prioritizing phonological contrasts in both research and language teaching (Jakobson, 1931; Martinet, 1952; Hockett, 1955; Catford, 1987; Surendran & Niyogi, 2003; Sewell, 2017). While the concept of functional load is widely acknowledged, that is that some sound distinctions have a more important role for distinguishing meaning than others, the challenge of operationalizing and quantifying this concept has led to a variety of methods, each with its own implications (Brown, 1988; Lin, 2019; Isaacs & Trofimovich, 2017; Sewell, 2021).

Minimal Pair Method

Historically, the most accessible and widespread method for quantifying functional load has been the counting of minimal pairs (King, 1967; Catford, 1987; Brown, 1988; Jenkins, 2003; Surendran & Niyogi, 2006; Munro & Derwing, 2006). This approach defines "functional load as a measure of the number of minimal pairs which can be found for a given opposition" (King, 1967; Catford, 1987; Sewell, 2017) and is a preferred method for precise and direct pedagogical applicability. For example, contrasts such as /p/ - /b/ and /t/ - /d/ in English are recognized as carrying a high functional load, as they distinguish large numbers of frequent, familiar words (Brown, 1988; Munro & Derwing, 2006; Isaacs & Trofimovich, 2017). A list ranking the order of minimal pairs was suggested by Brown (1988). They provide a rank ordering of Received Pronunciation (RP) phoneme pairs by functional load that is, it shows which pairs of sounds, when confused, are more likely to affect communication, according to variables such as frequency, number of minimal pairs, and occurrence in native accents. The minimal pairs method has led to the recommendation that pronunciation instruction should prioritize these high functional load contrasts (Henderson, 2008; McAndrews & Thomson, 2017).

However, the minimal pair method has limitations. Not all minimal pairs are equally communicatively valuable; some involve rare or contextually obvious words, and the method does not account for the frequency of occurrence of the contrasts in everyday speech or consider positional or contextual effects or sociolinguistic variation (Levis et al., 2016; Henderson, 2008; Sewell, 2017). One such example as noted by Catford (1987) is that the dental fricatives /θ/ and /ð/ carry low functional weight despite being phonetically marked and frequent in some language contexts, because they distinguish relatively few minimal pairs (Henderson, 2008; Isaacs & Trofimovich, 2017).

Information-Theoretic (Entropy-Based) Approaches

Information-theoretic (entropy-based) approaches to functional load were first developed in the mid-20th century (Hockett, 1955) and have since been elaborated by a number of scholars (Surendran & Niyogi, 2003, 2006; Lin, 2019). These methods define the functional load of a phonemic contrast as the "change in lexical entropy, the reduction in uncertainty induced by neutralizing the contrast within a corpus" (Surendran & Niyogi, 2003). This framework weights contrasts according to the frequency of the words involved, resulting in more precision in the communicative result of possible sound mergers (Surendran & Niyogi, 2003; Sewell,2017). For example, Surendran and Niyogi (2003) demonstrated that neutralizing a contrast between two highly frequent words results in a greater increase in lexical uncertainty (i.e. a higher functional load) than neutralizing a contrast between rare words, even if the number of minimal pairs is the same. This means that entropy-based measures are particularly sensitive to the practical risk of miscommunication in real-world usage (Isaacs & Trofimovich, 2017). Empirical studies confirm that while entropy-based and minimal pair measures are often correlated, they can diverge in how they rank particular segmental contrasts, especially where word frequency plays a key role (Lin, 2019). This divergence demonstrates the conceptual and practical differences between the methods. However, entropy-based approaches are computationally intensive and require large, frequency-annotated corpora as well as statistical expertise, which can limit their accessibility for pedagogical purposes or for use in smaller, under-resourced languages (Brown, 1988).

Corpus-Based Computational Modeling

Advances in corpus linguistics and computational phonology have led to increasingly sophisticated methods for estimating functional load. These approaches produce datasets to incorporate not only minimal pairs and entropy, but also other variables such as phonological neighborhood density, part-of-speech information, and even syntactic context (Lin, 2019). For example, Wedel et al. (2013) demonstrated that phoneme pairs which typically merge have lower functional load, and that variables like syntactic category and word frequency further predict the likelihood of contrast loss (see section 2.5.3 for a detailed discussion of Wedel et al., 2013 study). Similarly, Oh et al. (2015) used a corpus-based approach to compare the phonological system and lexicon in several languages, showing that functional load values calculated over entire corpora are insightful into how phoneme contrasts contribute to lexical distinction and how values are different across many languages. These computational methods provide detail, can compare among different multiple languages and show the context nature of phonological contrasts in real language use (Sewell, 2017). However, the success of these methods relies on how good and representative the data is, which can become difficult for instructors or non-specialists without technical expertise to use or understand them (Isaacs & Trofimovich, 2017; Brown, 1988).

Despite approaches to estimating functional load in English, they generally agree in identifying the same contrasts as high or low in functional importance. Across multiple studies, the dental fricatives /θ/ and /ð/ are consistently ranked as low functional load, while contrasts such as /n/–/l/ are recognized as highly significant for distinguishing words in English (Henderson, 2008). This was demonstrated in a corpus and entropy-based analyses that the loss of the /n/–/l/ (ranked 10 as high functional load in Brown's list, 1988) distinction would result in a high degree of homophony and communicative confusion, whereas the neutralization of /θ/ and /ð/ (ranked 5 as low functional load in Brown's list, (1988) would affect relatively few lexical items (Lin, 2019).

This agreement is echoed in experimental studies. Munro and Derwing (2006) empirically demonstrated that errors involving high functional load contrasts (i.e. /n/-/l/) have a significantly greater negative impact on comprehensibility and intelligibility than errors on low functional load contrasts (i.e. /ð/-/d/). Such findings are supported by analyses of segmental errors in learner speech and their effect on listeners (Henderson, 2008).

The agreement between theoretical, computational, and empirical approaches provides credibility to functional load as a guide for instructional design and for pronunciation teaching (McAndrews & Thomson, 2017). By prioritizing high functional load contrasts, teachers and curriculum designers can maximize communicative value for learners, helping them avoid errors most likely to interfere with listener comprehension (Munro & Derwing, 2006). Moreover, historical phonological research shows that high functional load contrasts are less likely to be lost through sound change, as language users tend to preserve distinctions that serve to differentiate many words (Wedel et al., 2013). This long-lasting stability also supports the pedagogical emphasis on teaching high functional load contrasts since supports the current goals of communication and extensive language growth needed for development (Lin, 2019).

However, it is important to note, that functional load may not accommodate every learner. Some phonological errors, though low in functional load, may carry substantial social stigma or be relevant for learners seeking native-like varieties (2008 Henderson). Moreover, functional load can vary by position, register, or discourse context, and also interact with other factors which may include articulation, L1 transfer, and sociolinguistic goals (Brown, 1988; Sewell, 2017; Isaacs & Trofimovich, 2017). These complexities highlight the need for a context-aware application of the functional load principle in instructional settings, acknowledging that the communicative goals may intersect with goals set by the leaner and their social contexts.

Overall, while each method for calculating functional load may be beneficial, the findings provide strong validation for the principle that functional load should guide instructional focus, especially in interventions that intend to focus on developing intelligibility and comprehensibility. Despite this growing consensus, it is still uncommon to see HVPT studies that integrate multiple high-load contrasts in a single training design. Most prior research isolates one contrast, which may not be a true representation of real-world communicative demands. This gap suggests that, although the theoretical rationale for targeting high functional load contrasts is sound, its practical applicability in instructional contexts remains underexplored.

The current study directly addresses this limitation by targeting ten high functional load contrasts in its HVPT training intervention. These contrasts were selected based on Brown's (1988) list that had the highest functional load ranking and known difficulty for Mandarin speakers. This decision is solely intended to better simulate real-world communicative challenges and examine whether training multiple functionally weighted contrasts can promote wider transfer effects.

In summary, functional load can be operationalized through minimal pairs (Brown, 1988; Catford, 1987), entropy-based measures (Surendran & Niyogi, 2003), large-scale corpus modeling (Wedel et al., 2013), or experimental validation (Jenkins, 2003; Munro & Derwing, 2006; Suzukida & Saito, 2019). These methods collectively further support the principle that instructional priorities should be determined by the communicative value of phonological contrasts (Brown,1988; Sewell, 2021). The present study's design synthesizes these insights, aiming for targeted, high-impact improvements in intelligibility by foregrounding high functional load features and addressing the practical demands of authentic communication.

2.5.3 Empirical Support for Functional Load

A significant volume of research has established the empirical validity of the functional load principle. Foundational figures in this area include Wang (1967), King (1967), Catford (1987) and Brown (1991) who developed early theoretical and quantitative frameworks for functional load and its implications for phonological systems (Brown, 1988). More recent decades have seen a shift toward direct experimental and corpus-based investigations, notably by Munro and Derwing (2006), Wedel et al., (2013), Bent et al. (2007b), Surendran and Niyogi (2006), and contemporary applied linguists such as Kang and Moran (2014), and Suzukida and Saito (2019). These studies collectively demonstrate that high functional load contrasts those that distinguish the most minimal pairs and carry the greatest communicative weight are most critical for intelligibility and comprehensibility in both L1 and L2 contexts (Sewell, 2017)).

Particularly influential research comes from Wedel et al., (2013) who approached the role of functional load from a historical and cross-linguistic perspective. In their large-scale corpus analyses, they examined phoneme mergers across eight languages including English, Korean, French, and Dutch and found that phoneme pairs that eventually merged during language change had significantly lower functional load than those that remained distinct. Using both minimal pair counts and information-theoretic entropy measures, they demonstrated that high functional load contrasts were robustly preserved over time, while low-load contrasts were more prone to merger and loss (Lin, 2019; Sewell, 2021). Their findings provide evidence that the preservation of high functional load contrasts is not random but is instead driven by communicative necessity; maintaining distinctions with greater lexical impact helps to avoid homophony and ensures clearer communication. This resistance to merger supports the core rationale for prioritizing high functional load contrasts in both research and pedagogy.

Kang and Moran (2014) investigated how the distribution of segmental errors, analyzed through functional load, relates to proficiency and comprehensibility in L2 English speech. Using speech samples from candidates at four proficiency levels on the Cambridge ESOL General English Examinations (B1–C2), the researchers coded vowel and consonant substitution errors as either high or low functional load according to established criteria (i.e. Brown, 1991; Catford, 1987). Their analysis revealed that the number of high functional load segmental errors (both vowels and consonants) decreased significantly as proficiency increased, whereas low functional load errors did not show a consistent pattern across levels. Importantly, high functional load errors were much more prevalent at lower English language levels and were closely associated with lower comprehensibility ratings. In contrast, as learners advanced, segmental errors were reduced.

In the domain of L2 perception and intelligibility, Bent et al., (2007b) offered critical experimental evidence for the perceptual consequences of functional load. Their study systematically manipulated segmental errors in spoken English and measured listener word recognition and message comprehension. The results showed high functional load consonant contrasts such as those distinguishing a large number of minimal pairs led to further declines in word recognition rates and greater overall degradation of message comprehension than comparable errors in low functional load contrasts. This effect was seen across different listener groups and tasks which demonstrates that communication was a disruptive pattern for learners. This is caused by misperception that is associated to the functional load of the segment (McAndrews & Thomson, 2017). The study also reinforced to target high functional load contrasts in pronunciation instruction, as these distinctions are disproportionately influential for successful spoken communication.

Cross-linguistic differences in functional load rankings add an important additional degree of difficulty to pronunciation teaching. As Brown (1991) demonstrated, what constitutes a high-load contrast in English may not be the same in German or Spanish. Therefore, pedagogical applications of functional load must be targeted to the specific language being learned. Importantly, a particular consideration is the relevancy for Mandarin L1 speakers learning English, since many targeted contrasts (i.e. /æ/-/ʌ/, /l/-/r/) do not exist in Mandarin and thus require greater instructional attention. Recent research by Lin (2019) further investigated the role of functional load in a learner’s L1 and its influence on perception, acquisition, and learnability of new contrasts in the L2. Through a combination of simulations, corpus studies, and perceptual experiments, they found that learners are more likely to successfully acquire and perceive high functional load contrasts in the target language even when these contrasts are absent from their L1 because such contrasts are more perceptually prominent and carry greater communicative consequences. These findings support the importance of ranking high functional load contrasts in pronunciation instruction for Mandarin speakers learning English.

Research has also demonstrated that learners of all language levels and ages benefit from high variability phonetic training (HVPT). For example, Hwang and Lee (2016) showed that HVPT targeting high functional load (FL) contrasts led to significant perceptual gains among young Korean learners. This confirms that FL-informed instruction is a highly productive method for adults and also for children and low-level learners. The study included 40 Korean elementary school students primarily 6th graders that participated in 18 sessions of HVPT covering all English vowel and consonant sounds, rather than focusing on a limited set of contrasts. Phoneme identification abilities were assessed before and after training, with comparisons made between the two groups: experimental and control. The results provide evidence that the functional load principle is empirically effective in pronunciation training for learners of varying ages and proficiency levels. Notably, the greatest improvements were observed for contrasts with high functional load, which are most critical for distinguishing meaning in English, while smaller gains were seen for low FL contrasts. Further to this, it was found that perceptual gains were especially substantial for sounds that were poorly identified prior to training, and a learnability gradient was established based on identification accuracy.

By demonstrating that both young learners and those at lower proficiency levels have the potential to achieve perceptual gains, especially on high FL contrasts, through HVPT, the study extends beyond FL-based instructional design for adult learners to include children and beginners. These findings suggest that prioritizing high FL contrasts in instructional materials can increase the impact of pronunciation training regardless of age or initial proficiency. The results support the view that targeting high FL contrasts in training benefits learners across the entire learning spectrum, as these phonemic distinctions are both more learnable and have a significant role for intelligibility and communicative success. This provides a strong empirical foundation for FL-based instructional priorities in L2 pronunciation pedagogy. Thus, Hwang and Lee (2016) reinforce the use of functional load as an important criterion in L2 pronunciation teaching for learners at different stages of learning . This finding, complements and extends earlier work in the field.

Extensive empirical and theoretical evidence supports the functional load principle is an approach which can be applied as a reference model for L2 pronunciation instruction. Integrating functional load principle into HVPT, the present study aims to promote perceptual gains and improvements in intelligibility and comprehensibility. Unlike most prior studies, which have relied on small phonemic inventories, the present design is novel in tracking the effects of HVPT on multiple high functional load contrasts across perception, intelligibility, and comprehensibility including their transfer into spontaneous speech and retention over time. While prior studies have sometimes trained or tested multiple contrasts, few, if any, have followed the full trajectory from perceptual learning through to real-world communicative outcomes and delayed effects using a multi-tiered assessment approach (Lin, 2019; Hwang & Lee, 2016). This approach is relevant for spontaneous speech, where listeners depend on a range of phonemic cues and overall speech clarity. To evaluate whether perceptual training gains transfer meaningfully into real-time, unstructured communication, the present study applies a multi-tiered methodology, including spontaneous oral production, naïve listener ratings, and delayed post-testing. In sum, the principled selection of high functional load contrasts ensures that training is aligned with real-world communicative needs and provides a theoretically grounded basis for interpreting outcomes across perception, intelligibility, and comprehensibility dimensions.

Building on the principles of functional load theory, which inform the selection of phonemic contrasts most likely to affect intelligibility and comprehensibility, the following section (2.6) turns to High Variability Phonetic Training (HVPT). HVPT represents one of the most prominent and empirically supported models of perceptual training and serves as the core instructional framework investigated in the present study.

2.6 High Variability Phonetic Training

High Variability Phonetic Training (HVPT) is a specific form of auditory perceptual training designed to improve learners' ability to perceive difficult non-native phonemic contrasts. The training method exposes learners through repetition to minimal pairs containing target contrasts, presented in varying phonetic contexts by multiple talkers. This variability is intended to promote the development of robust phonological categories which function independently from specific voices and contexts and potentially transfer to new situations. Research demonstrates that HVPT effectively enhances both the perception and, to some extent, the production of L2 sounds. However, most HVPT research has focused on perception, often using only a single task format typically forced-choice identification with relatively limited attention to whether perceptual gains transfer to intelligibility or comprehensibility, particularly in spontaneous speech. This study addresses that gap by employing both identification and discrimination tasks in HVPT and evaluate their impact on transfer to listener comprehension of spontaneous L2 oral production.

2.6.1 Principles and Methodology of HVPT

High Variability Phonetic Training (HVPT) is based on the premise that perceptual learning tends to be more successful when learners experience a range of phonetic variation. By presenting minimal pairs produced by multiple speakers in diverse phonetic environments, The specific training supports learners to identify stable phonetic sounds that reliably distinguish phonemes across contexts (Logan et al., 1991; Lively et al., 1993). This approach draws upon models of speech perception (Pierrehumbert, 2001), which state that phonological categories are formed and further developed through repeated engagement and listening to diverse speech samples. As learners internalize multiple exemplars, they begin to identify common phonetic features that underlie phonemic categories, independent of speaker-specific or contextual noise. This theoretical grounding makes HVPT especially effective in supporting generalization to unfamiliar voices and real-world communicative settings.

A central methodological feature of HVPT is its application of multiple talkers during training. Speaker variability is a core pedagogical mechanism that strengthens perceptual learning.

Early evidence from Lively et al. (1993) showed that Japanese learners that trained on the English /r/-/l/ contrast using multiple speakers achieved greater generalization to untrained voices than those trained with a single speaker. The basic concept is that learners need to rely on sound patterns that remain stable across speakers, such as vowel quality shifting, instead of relying on speaker-specific differences (i.e. accent). Later studies built on this finding. Brosseau-Lapré et al. (2013), for instance, found that training with ten speakers resulted in significantly better generalization than training with three speakers, further supporting that speaker diversity enhances perceptual robustness. Similarly, Wong (2014) reported that learners trained with variable speaker input performed better on novel-talker tests and demonstrated more flexible application of their perceptual gains. Taking together the former findings support the inclusion of talker variability as a consistent outcome in HVPT design (Thomson, 2018).

Of similar significance is the structure of HVPT tasks (see section 2.5.2 for further details on tasks). The traditional format involves a forced-choice identification task, where learners select from among two or more options after hearing a stimulus. Immediate feedback follows each response, signaling whether it was correct or incorrect. This immediate corrective feedback helps learners reposition their perceptual boundaries and learn to focus more precisely on the sound characteristics that determine phonemic contrasts (Handley et al., 2009). The current study incorporates real-time feedback within both identification and discrimination tasks, an approach consistent with findings suggesting this design supports more sustained learning (Qian, 2018). Some newer designs have expanded to include ABX or oddity tasks to adjust difficulty and maintain learners' interest throughout the training (Iverson et al., 2005; Thomson, 2018). However, most implementations still rely on identification tasks, which suggests the needs to examine whether training formats might be optimized further to support progress among learners in diverse language ability levels.

The effectiveness of HVPT is closely linked to the nature of the stimuli used. Unlike training approaches that rely on synthesized or isolated sounds, HVPT emphasizes natural speech recordings. Natural stimuli preserve the acoustic variability and coarticulatory subtle differences present in authentic speech, thereby equipping learners for real-world listening conditions. For example, Logan, Lively, and Pisoni (1991) demonstrated that Japanese learners trained with natural, multi-talker recordings of English /r/ and /l/ that improved in perception tasks and also transferred this learning to unfamiliar speakers and untrained words, an outcome not observed with synthetic or single-talker stimuli (see section 2.5.3 for details of the study). Similarly, Iverson et al. (2005) found that training with naturally produced speech was at least as effective as, and less labor-intensive than, training with digitally manipulated or synthesized stimuli. In line with this evidence, the present study employs naturally produced stimuli from multiple speakers in a variety of phonetic contexts, thus in line with concepts of HVPT to maximize transferability and ecological validity in L2 speech perception training.

Moreover, HVPT frequently uses minimal pairs such as “road” vs. “load” or “rake vs. “lake” (Bradlow et al., 1997; Shinohara & Iverson, 2018) to sharpen learners’ sensitivity to phonemic contrasts that differ by only one sound. These pairs help reduce lexical guessing and highlight subtle phonetic distinctions, particularly for contrasts that may not exist in the learner’s first language. In the context of the present research, the use of minimal pairs is an important feature, because it places targeted attention on specific contrasts as well as natural voices and varied input for learners. This approach agrees with previous findings that HVPT using minimal pairs fosters more robust and generalizable L2 phoneme categories (Li, 2015). By selecting minimal pairs that challenge learners’ L1 perceptual boundaries, the present study examines perceptual improvements but further extends to examine for potential transfer to novel lexical and phonetic environments.

Furthermore, phonetic context variability plays also an important role in promoting effectiveness of HVPT even further. In their investigation, Wang and Munro (2004) and Hardison (2003) demonstrated that exposing learners to target phonemes in a range of phonetic environments, such as different word positions and varying consonantal and vocalic surroundings, enhanced their ability to transfer newly acquired perceptual skills to untrained items. This variability helps learners notice the key sound features of each contrast even when other sounds around them change. Such adaptability is important for learners so they can still understand phonemes correctly even when those sounds are blended or altered in fast, connected speech. This can also be seen in recent experimental work by Zhang et al. (2021) which provides strong evidence that phonetic context variability is a central influence of generalization in HVPT. In their study, adult Mandarin speakers were trained to perceive difficult English contrasts using HVPT protocols that manipulated not only talker variability but also the range of phonetic environments. The results revealed that participants that interacted with greater phonetic context variability across different consonantal and vocalic frames showed robust gains in both perception and production that generalized to new lexical items and unfamiliar speakers, even in the absence of extensive talker variability. Importantly, Zhang et al., (2021) found that while talker variability has traditionally been emphasized, it was the inclusion of enhanced acoustic and contextual variability that most reliably induced transfer and retention of learning, suggesting that phonetic context variability is essential for long-term, real-world perceptual gains. Although the present study includes a range of phonetic contexts, it does not isolate these variables in investigated the data, as the central focus is on the transfer of HVPT to comprehensibility in spontaneous speech. Nonetheless, the inclusion of phonetic context variability is supported by a substantial body of evidence as a best practice for promoting perceptual learning that endures outside the limited environment of laboratory tasks.

Although HVPT has demonstrated effectiveness, certain methodological limitations have been noted in the literature. A recurring methodological critique in the HVPT literature is that many studies have placed emphasis on immediate post-test assessments but lack testing long-term retention. In fact, research shows that only 30% of surveyed HVPT studies include a delayed post-test, pointing towards longitudinal research to determine whether learning is retained in the long term and if it gradually develops in L2 speech production over time (Thomson, 2018). This limitation is echoed in a recent meta-analysis (Uchihara et al. 2024) which reports insufficient evidence for the long-term retention of production improvements. Another significant gap is the dominant use of controlled practice, such as reading words and sentences, rather than on spontaneous or interactive speech. The meta-analysis notes that HVPT is more effective at improving explicit, form-focused production skills but leaves open the question of whether develops the automatic processing skills for fluent, spontaneous speech. A further limitation shows that the overuse of reading tasks and calling for more ecologically valid measures, such as elicited imitation or picture naming, to better capture learners' real-world pronunciation abilities (Thomson, 2018). Supporting this, it has been found that while HVPT led to improvements in delayed imitation tasks, these gains were not further tested in tasks or contextual situations for spontaneous speech productions from L2 learners (Thomson & Derwing, 2016). Together, the former studies emphasize that further contributions in HVPT are needed to include both long-term retention and the transferability of gains to authentic, communicative settings.

Beyond these concerns, it has also been noted that the repetitive and decontextualized nature of HVPT, typically involving numerous cycles of minimal-pair identification and discrimination tasks with little variety, can reduce learner motivation as training continues over time (Thomson, 2018). On the other hand, some researchers argue that a repetition is precisely what reinforces robust phonemic categories. Especially, including tasks that are scaffolded and refreshed (e.g. via stimulus rotation, adaptive difficulty or gamified variations) can lessen the negative motivational effects. For example, Huensch (2016) highlighted the use of custom computer videogames that embedded target sounds into the gameplay. This design allowed learning to occur without explicit labelling; that is, learners dd not need to consciously identify which phonemes they heard. It avoided overt categorization and direct corrective feedback while producing stronger perceptual improvements than longer, traditional training. An approach such as this, supports learner motivation by offering a more stimulating alternative to conventional HVPT tasks while retaining the effectiveness of perceptual training. Similarly, Saito et al. (2022b) emphasized integrating multimodal and communicative tasks to improve learners; motivation without compromising HVPT training effectiveness. Their research illustrates how incidental and multimodal HVPT can enhance phonetic category restructuring even when learners' attention is divided. These approaches show how adaptive, engaging training methods have the potential to balance perceptual precision with learner engagement.

Overall, HVPT is an evidence-based perceptual training paradigm designed to promote robust phonological category formation through exposure to naturalistic, variable speech input. Its principles and features (speaker and phonetic variability, minimal pairs and real-time feedback) aim at training learners to perceive L2 contrasts accurately. While HVPT has traditionally focused on perception, its broader instructional value is found in supporting potential gains in intelligibility and comprehensibility. In this study, the above principles are operationalized in an online training context, combining forced-choice identification and discrimination tasks to evaluate perceptual training's potential to impact spontaneous speech performance in L2 learners.

2.6.2 Types of HVPT Activities

High Variability Phonetic Training (HVPT) typically includes of an identification type and a discrimination type task in the perceptual training, in which, each has a different role but also a complimentary one in shaping learners' phonological representations and improving the perception of non-native contrasts. The present study includes both types of tasks in the HVPT design in order to understand the possible maximal benefits of the training intervention and to address the limitations found in past studies especially in the area of transferability in comprehensibility to spontaneous oral production

A typical identification task involves categorizing a single auditory stimulus into a phonemic category, usually through a forced-choice response format. These tasks are important in assisting learners form abstract phonological categories by encouraging them to associate variable tokens with a consistent linguistic label through listening activities (Strange & Dittman, 1984; Lively et al., 1993). In doing so, identification tasks facilitate categorical perception, and learners begin to perceive speech not as a stream of sounds but as discrete linguistic units. Logan and Pruitt (1995) reported that such tasks result in lasting improvements and promote transfer to untrained items. The effectiveness of identification tasks has demonstrated that these types of tasks can assist learners to overcome L1-based perceptual biases. For instance, Japanese learners, after undergoing high-variability phonetic training using identification tasks, significantly improved their ability to distinguish the English /r/-/l/ contrast. The repeated exposure to difficult contrasts and corrective feedback enabled learners to reform their perceptual categories, and as a result improve their phoneme categorization and their confidence in perceptual judgments (Hazan et al., 2005).

In contrast, a typical discrimination task requires learners to judge whether two or more stimuli are the same or different. It is also implemented in same–different, AX, or oddity paradigms and are especially useful in the early stages of training. These task types enhance learners' sensitivity to subtle acoustic distinctions, particularly within phonemic categories, and may help learners tune into features that are otherwise perceptually assimilated due to

L1 interference (Logan & Pruitt, 1995). For example, Werker and Tees (1984) used AX discrimination tasks to examine English listeners' sensitivity to non-native place of articulation contrasts in Hindi. They found that performance improved when the interval between stimuli was shortened, indicating reliance on auditory memory and acoustic salience rather than established phonological categories. Furthermore, sustained exposure to discrimination tasks enhanced within-category perceptual acuity, suggesting that such tasks can refine auditory sensitivity even without explicit categorization (Logan et al., 1991)

Importantly, evidence suggests that combining identification and discrimination tasks may produce more comprehensive perceptual learning, especially when learners face strong L1 interference. Carlet and Cebrian (2015) observed that while identification training was more effective for improving perception of trained contrasts, discrimination tasks contributed to generalization to untrained sounds, enhancing learners' overall perceptual flexibility. Handley et al. (2009) directly compared oddity discrimination and identification training in Mandarin speakers learning the English /r/-/l/ contrast. They found both tasks to be effective in furthering learning, with no significant difference in outcomes. This suggests that including both tasks can offer variety and enhance perceptual learning.

Further research has also combined both types of tasks in HVPT, reinforcing the claim that their integration yields complementary perceptual benefits. For example, Carlet (2017) compared identification and categorical discrimination training for English vowels and found that both methods led to significant perceptual gains, but that identification training produced larger improvements. However, the discrimination group showed more improvement on untrained contrasts, suggesting that a mixed-task protocol can optimize both within-category sensitivity and between-category robustness. Learners exposed to such dual-task training protocols may have gains in categorization accuracy while also sharpening their sensitivity to subtle, within-category distinctions.

Research from the field of cognitive psychology also supports the idea that different types of perceptual tasks tap into distinct cognitive processes. Identification relies more on category learning and long-term memory, while discrimination may engage working memory and attention more directly (Flege, 1995b; Shinohara & Iverson, 2018). Therefore, alternating between these tasks in training may provide more activation of the neural systems involved in L2 phonetic learning. Neuroimaging studies have shown that perceptual learning through HVPT engages distributed networks across auditory, motor, and attentional brain regions. Specifically, training-induced changes in mismatch negativity (MMN) responses and categorical perception have been observed, indicating enhanced pre-attentive neural discrimination and more native-like categorical processing after HVP (Zhang et al., 2009). Having both task types in this study's training design reflects this multidimensionality, aiming to activate complementary pathways that contribute to long-term phonological change.

The pedagogical value of task variation should not be underestimated. Learners differ in their aptitude, attention span, and cognitive style, and task variety may help sustain motivation and cognitive interaction during training. In an instructional setting, task diversification can also serve as a diagnostic tool to reveal which contrasts learners perceive reliably and which require further attention. Saito et al., (2022b) reviewed multimodal and variable HVPT approaches and noted that including different task types sustains engagement and may also better accommodate individual differences in learning strategy and aptitude, resulting in long-term gains. This suggests that learners with strong analytic skills may perform better on identification tasks, while those with heightened auditory sensitivity may excel at discrimination. Designing training that accommodates these individual differences aligns with learner-centered pedagogy and may facilitate wider access of HVPT.

The present study incorporates both identification and discrimination tasks within its HVPT design. It seeks to extend existing research by investigating whether combining both types of tasks can improve learners' perceptual precision of segments and support transfer to intelligibility and comprehensibility in spontaneous speech, a dimension of L2 speech development that remains underexplored in HVPT literature.

2.6.3 Impact of HVPT on Perception

Literature demonstrates that High Variability Phonetic Training (HVPT) improves second language (L2) speech perception abilities across different language combinations and phonetic distinction sets. Early research showed that learners who experience multiple talkers and phonetic contexts during training develop strong flexible phonological categories which enables perceptual learning and leads to stimulus generalization. This section examines the perception effects of HVPT.

Lively, Logan and Pisoni (1991) developed HVPT research through their study of English /r/-/l/ difficulties experienced by Japanese learners. The researchers enrolled Japanese adults who had spent various amount of time in the United States for their study which required forced-choice identification testing. They listened to naturally spoken minimal pair words (such as "rock" vs. "lock") that five native English speakers delivered through different phonetic contexts including syllable-initial, medial and final positions. The participants obtained instant feedback on their responses during 15 training sessions that each lasted 40 minutes across three weeks. Participants achieved significant improvement in their /r/-/l/ contrast identification through the training; gained scores from 65% to 85% for trained stimuli. The training materials led participants to demonstrate better identification of the /r/-/l/ contrast when hearing new words or unknown speakers. Participants demonstrated higher accuracy when identifying the /r/-/l/ contrast in both unfamiliar words and voices of different speakers which indicated their learning generalization abilities. However, the transfer to untrained contexts was limited. The identification of /r/-/l/ by learners depended on phonetic context but they performed better when the syllable-final position occurred compared to the initial position. The study showed that training materials which incorporate multiple talkers and diverse phonetic contexts enable adult L2 learners to create meaningful perceptual restructuring. The training achieved this by presenting participants with multiple natural speech examples which directed their attention toward essential acoustic features for /r/ and /l/ differentiation instead of speaker-specific traits. The training guided future research HVPT investigations.

Lively et al. (1994) investigated the same group of Japanese adult learners who studied English /l/-/r/ to evaluate how perceptual improvements persist over time. The experimental group of 19 participants completed high variability phonetic training (HVPT) which consisted of identification tasks with naturally spoken words from multiple native English speakers across different phonetic positions. The control group consisted of 23 Japanese speakers who only took the tests without any training. The researchers ran multiple assessments which began with pre-tests followed by post-tests and crucial delayed post-tests conducted three and six months after training to assess long-term retention. The delayed tests presented new words to participants which they had not encountered during training along with audio from known and unknown speakers to assess generalization. The experimental participants showed marked improvement in their ability to identify the /l/-/r/ contrast between the pre-test and post-test but the control participants maintained their initial scores. The experimental group maintained their perceptual improvements at 98% accuracy during the three-month post-training assessment and their results remained 4.5% better than pre-test scores during the six-month evaluation. The improvement showed evidence of generalization across new words and unfamiliar speakers which suggested that learning occurred through category formation instead of memorization of specific training items. The research established HVPT as an effective method for producing immediate perception benefits which lead to long-term learning retention with broad applicability when training incorporates multiple speakers and different phonetic contexts. The study establishes HVPT as a valuable educational tool for establishing long-lasting changes in L2 speech perception abilities.

HVPT transferability received further investigation through Pruitt et al. (2006) who trained both native English and Japanese speakers to perceive Hindi stop consonant contrasts. The experiment used naturally produced minimal pairs presented by multiple native Hindi speakers to obtain forced-choice identification results with immediate feedback. Post-training assessments evaluated both trained and untrained items, as well as unfamiliar speakers. The two learning groups achieved significant improvements in perception that extended past the trained materials. The results demonstrated that the speed of improvement together with the extent of improvement relied on learner L1 language backgrounds which affected HVPT outcome results. The findings showed that HVPT effectively enhances L2 learning targets as well as basic phonetic abilities across multiple languages. Research shows HVPT works effectively for teaching L2 vowel perception in addition to the English /r/-/l/ sound contrast.

The study by Iverson and Evans (2009) trained English language learners from Spain and Germany who were based in the UK to identify English vowels through discrimination and identification tasks. Both learner groups achieved perceptible improvements in vowel recognition though the degree of enhancement differed according to their native language.

Lambacher et al. (2005). trained Japanese learners on five American English sounds (/iː/, /ɪ/, /æ/, /ʌ/, /ɑː/) during a six-week program through identification tasks with high-variability minimal pairs. The research demonstrated both effective perception improvement and production benefits. Advancing these studies, Nishi and Kewley-Port (2007, 2008) demonstrated that High Variability Phonetic Training (HVPT) with naturally produced vowel stimuli from multiple speakers across various phonetic contexts produces better and more generalizable English vowel perception improvements for Japanese and Korean learners when the training includes all nine vowel categories. The participants who received training with the complete vowel set achieved better identification accuracy and improved word recognition and speaker generalization and maintained their gains longer than participants who received training with only three challenging vowels. The studies demonstrate that HVPT learners need to experience the complete set of target language vowel categories to achieve better and longer-lasting phonetic learning outcomes.

Aliaga-García and Mora (2009) provided Spanish and Catalan learners with training on vowel (/æ/-/ʌ/) and consonant (/b/-/p/, /t/-/d/, /l/-/r/) contrasts through high-variability minimal pairs in a laboratory experiment. Learners performed identification tasks across six two-hour training sessions which included stimuli from multiple talkers in different phonetic contexts. The research showed perceptual improvement occurred for every sound contrast while results persisted between different tasks and testing environments. The study demonstrated that perception-production links exist through its findings of improved production abilities. The research demonstrated that HVPT delivers flexible training methods for multiple segmental features which work with different first language speakers. The training of suprasegmental and segmental contrasts through HVPT has received limited research attention. Wang et al. (1999; 2003) used high-variability identification tasks with multiple talkers in variable contexts to train Mandarin-speaking learners of English for both vowel and tonal distinctions. Participants showed measurable improvements in their perception abilities which extended to new items and demonstrated some retention of skills throughout the training period. The research indicates HVPT principles can be used to train other challenging perceptual features including tone and intonation which play essential roles in achieving both speech intelligibility and natural speech quality.

Research studies on HVPT and perception have built a strong foundation to prove its effectiveness in enhancing students' ability to recognize L2 phonemes across various languages and learner types and different segmental targets. HVPT delivers reliable perceptual changes which extend past training environments when it combines variable speakers and phonetic elements. The degree of transfer resistance varies across different contrasts while individual learner factors such as first language background together with perceptual skills and training methods influence the final results. Perception task-based studies form the core of existing research although few investigations have connected perception improvement to actual communication performance. The essential question persists regarding whether HVPT-induced perception improvements lead to production improvements that subsequently result in better intelligibility and comprehensibility in spontaneous L2 oral production. The following sections address these questions starting with HVPT effects on intelligibility.

2.6.4 HVPT and intelligibility in isolated word production

HVPT was initially developed to enhance learners' speech perception, , it has further extended in investigating its potential in enhancing L2 speech production intelligibility at the word level. Research investigating whether perceptual training transfers to more intelligible production of isolated words and controlled sentences reveals a complex, yet positive connection between perception and production.

The research by Bradlow et al. (1997) examined Japanese learners' English /r/-/l/ sound production intelligibility after HVPT training. Japanese adult participants received 45 sessions of HVPT training, each for 20 - 30 minutes, during a 3–4-week period while focusing on the /r/-/l/ distinction. The research involved 11 Japanese adult participants who learned to identify words from five native English speakers in different phonetic contexts and received instant feedback during the training sessions. The participants were required to repeat the target words both before and after the training period. The native English listeners completed identification tests using forced-choice methods. Native English listeners successfully identified participant targeted phonemic sounds at improved rates following the training period. Extending on the study, Bradlow et al. (1999) conducted delayed post-tests at three months. The results demonstrated that HVPT produced enduring long-term production improvements which remained stable at three months even without direct articulatory training. The research findings demonstrate that perceptual representation development through HVPT produces enhanced speech intelligibility during production as well as sustained communication advantages.

The initial research findings have led multiple investigations to extend HVPT across different L1-L2 language pairs and phonetic targets. The investigation by Kangatharan et al. (2020) evaluated the impact of HVPT training on English vowel intelligibility in Greek-speaking L2 learners. Twenty participants took part in five online HVPT sessions which lasted for one hour each. The training consisted of natural spoken tokens from different English proficient speakers in various phonetic contexts together with instant feedback delivery for learners. The evaluation of intelligibility relied on twenty native English listeners who transcribed learner word productions before and after the study. Results showed that native listeners achieved better transcription accuracy after the study which proved that HVPT resulted in significant gains in word intelligibility. Iino (2019) conducted research to determine if HVPT training enhances Japanese learners' production of English fricatives including /f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/. The HVPT training system utilized computer-assisted web-based software to deliver natural speech tokens from various native English speakers with instant feedback to participants. The training took place at a university CALL lab during multiple weeks. Native listeners performed transcription tasks to measure speech intelligibility. The training results showed improved speech clarity on average although the effectiveness varied between different fricatives and phonetic contexts. Concluding, HVPT achieves successful results for multiple phoneme categories while enabling perceptual learning to further production accuracy of challenging segmental differences.

Findings by Uchihara et al. (2024) indicate trained items show a 10.5% improvement in production intelligibility yet untrained items only show a 4.5% improvement while retention and generalization present restricted progress. Several studies present various limitations together with conflicting viewpoints in their research outcomes. The word-level intelligibility gains from HVPT training fail to show how gains will transfer to speech forms which include connected sentences and spontaneous conversation. The study by Wong (2014) found that Hong Kong Cantonese speakers improved English /ɛ/-/æ/ words with both HVPT or LVPT.

However, HVPT produced superior results which transferred better to unfamiliar words and voices. The results failed to show statistical significance for sentence-level intelligibility which indicates HVPT enhances isolated word production only when extra training is provided.

Studies investigating the duration of HVPT production improvements have focused on the sustainability of these gains. Carlet (2017), as well as Rato and Rauber (2015) discovered that participants sustained most of their advancements in L2 vowel production and intelligibility during the two-week to three-month follow-up period but with differences based on phoneme categories and learner profiles. The intelligibility assessments were based on word reading and sentence reading tasks as well as native speaker ratings and acoustic analyses. The research indicates that HVPT creates long-lasting speech production changes mainly after learners achieve sufficient perceptual gains that become stable. Duration studies on HVPT was also conducted throughout an entire year by Nagle (2017) who studied Spanish speakers as they developed their English vowel perception skills and production abilities. Intelligibility tests included identification tasks for perception and elicited imitation and word reading for production. Better skills in vowel perception among listeners developed before their production abilities improved according to the study findings. The time gap between perception and intelligibility improvement supports the fact that students require time to link new phonetic knowledge to their articulatory abilities. The research findings presented in Lee et al. (2015) and Sakai and Moorman (2018) supports this interpretation. The studies tested intelligibility through delayed imitation tasks, word reading and identification tasks. The production gains showed a slight decrease between the pretest and delayed post-test ($g = 0.26$) yet remained above the baseline performance level. The meta-analysis demonstrated that training effects failed to persistently generalize to untrained items and spontaneous speech and thus suggested more research with innovative methods.

The study by Saito (2013) examined how HVPT perceptual training would enable the transfer of acquired skills to sentence-level intelligibility. The study involved 49 Japanese university students who acted as both experimental and control subjects. Experimental participants received HVPT training for English /ɹ/ segmental contrasts from multiple native English speakers in several 20–30-minute sessions which occurred across two to three weeks. The training session contained no explicit production instruction for students. The evaluation of production post-training included word and sentence reading assessments that native listeners scored for accuracy and comprehensibility. The results indicated that target sound intelligibility improved more than the control group in word-level and novel word contexts. The gains at the sentence level remained modest while showing less consistency which suggests perceptual training must be combined with production-focused tasks to achieve better complex spoken output.

Huensch and Tremblay (2015) investigated the effects of High Variability Phonetic Training (HVPT) on both the perception and production of English palatal codas (/ʃ/, /ʧ/, /ʤ/) among 24 Korean EFL learners over 8 daily sessions of 20 minutes each. By employing training materials produced by multiple talkers and eliciting responses in both isolated words and carrier sentence contexts, their study robustly targeted the development of robust phonological categories a key theoretical benefit of HVPT. Their findings indicated that participants achieved significant gains in perceptual discrimination of the target contrasts. However, improvements in production, while present, were relatively small. Most importantly, there was no direct, statistically significant correlation between the degree of perceptual improvement and the extent of production gains at the individual level. As noted by the authors, better sound discrimination abilities among learners do not automatically lead to improved speech output that is intelligible or accurate. The researchers found that perceptual gains do not necessarily lead to better production skills because learners differ in motivation and phonetic experience and articulation habits or because perceptual improvement may need extended practice to impact production abilities.

A key limitation of the study consistent with broader HVPT research is that the training is relatively short in duration and in intensity. The training periods used in Huensch and Tremblay (2015) along with other studies (Wong, 2014; Thomson, 2018) may be sufficient to modify perceptual categories but fail to provide enough time for new perceptual skills to integrate into stable speech production. The study's production assessment methods focused on specific tasks (word reading and sentence reading), but these tasks might not demonstrate complete gains from more natural spontaneous or communicative speech settings (Thomson & Derwing, 2016).The study's small participant number of N=24 restricts both the ability to detect minor perception-production relationships and limits generalization to wider populations. The study by Huensch and Tremblay (2015) shows how perception-based learning drives L2 phonological development, however, their research demonstrates the typical HVPT challenge of perceptual gains that fail to produce measurable production improvements. This finding indicates the necessity of developing extended training methods that combine learning components and using genuine speech assessment tools for proper understanding of L2 speech perception-production relationship.

The research shows that HVPT establishes a complex but nevertheless beneficial relationship which enhances production clarity. The direct teaching of articulation is not part of HVPT, but its perceptual accuracy leads to better intelligible speech mainly in structured word and sentence tasks. The investigations support that word-level intelligibility improves across different L1-L2 combinations and phonetic distinctions while showing some evidence of item expansion and time-based stability. The study results also show that the outcomes depend on training stimuli as well as assessment tasks and learner-specific variables. Research additionally shows a strong relationship between HVPT and the potential to advance intelligibility in controlled testing environments but lacks understanding about HVPT's impact in advancing spontaneous speech. The ability of HVPT to support comprehensibility in unstructured listener-rated speech remains an area that needs further investigation because real-world communication demands segmental clarity alongside holistic understanding from listeners. The following section evaluates findings from past to more recent research to determine HVPT's effects on comprehensibility in spontaneous L2 oral production.

2.6.5 HVPT and comprehensibility in spontaneous speech

The goal of pronunciation instruction is to enhance learners' communicative effectiveness during real-world, spontaneous interaction. Multiple studies show that High Variability Phonetic Training (HVPT) successfully enhances segmental contrast perception for learners' intelligibility at the word level but its effects on comprehensibility in spontaneous oral production remain uncertain. "Comprehensibility, defined as the listener's perceived ease or difficulty in understanding a speaker", is influenced by linguistic elements (i.e. lexical and grammatical choices) and suprasegmental factors (i.e. fluency and prosody (Hahn, 2004; Isaacs et al., 2018; Suzuki & Kormos, 2020). This complex nature of comprehensibility raises important theoretical and pedagogical questions about whether training methods focused primarily on segmental perception such as HVPT can produce measurable gains in speech that is spontaneously produced and listener-rated in naturalistic conditions.

Among the relatively limited body of research addressing this question, Thomson and Derwing (2016) conducted a foundational HVPT study involving 34 adult ESL learners in Canada. The study was specifically designed to examine whether perceptual training using HVPT could facilitate improvements in L2 speech production, emphasizing comprehensibility. The research involved one-month HVPT training with 40 sessions each lasting 20-30 minutes. The HVPT training presented ten Canadian English vowel stimuli (/i/, /ɪ/, /e/, /ɛ/, /æ/, /ɒ/, /ʌ/, /o/, /ʊ/, and /u/) from multiple talkers through different phonetic conditions to maintain high variability. Production was assessed at pre- and post-training through two separate production tests: a) delayed imitation task where learners repeated modeled sentences after a brief pause, and b) a spontaneous sentence production task where learners created their own sentences including the trained vowels. While the output results demonstrated gains in the delayed imitation task, no improvement was observed in the spontaneous speech production task. The findings demonstrate that HVPT alone may not be sufficient to transfer to unstructured and communicatively complex speech which limits its practical application for comprehensibility in spontaneous oral production. Further limitations show that the study's assessment of comprehensibility was based solely on a semi-controlled production task rather than truly free, spontaneous speech. Also, it did not include a delayed post-test so no conclusions could be made if gains were retained in the long term from the training intervention. The study's limitations prevent the findings from applying to authentic communicative environments.

A related but distinct contribution to this line of research comes from Lengeris (2018), who investigated whether HVPT could support perceptual and production gains among Greek learners of English. This study utilized a controlled laboratory environment to deliver auditory training over ten sessions, each lasting approximately 15–20 minutes across two weeks. The training focused on English vowel contrasts (/iː/, /ɪ/, /ʊ/, /uː/, /ʌ/, /ɑː/, /ɒ/, /ɔː/, /e/, /æ/) and incorporated identification tasks with instant feedback. High variability was ensured by including stimuli from multiple native speakers in diverse phonetic contexts. Production was evaluated using both read-aloud sentences and a semi-spontaneous picture-based task. The results revealed improvement in vowel production across both conditions, as confirmed by acoustic analyses (F1/F2 formant measures) and listener-based comprehensibility ratings provided by native English speakers. These findings suggest that HVPT has possibilities to impact speech beyond highly structured tasks. However, the speech samples produced in the picture-description task, while more natural than reading, were still elicited in a semi-controlled context. As such, the speech samples may not be a full representation of freely produced, interactional language, thus, limiting the applicability to which findings can be generalized to comprehensibility in spontaneous oral production.

A further study by Saito et al. (2022b), examined the effects of HVPT on Japanese EFL learners' ability to produce [r]–[l] and [æ]–[ʌ] contrasts. The training was conducted under high-variability conditions using multiple speakers and phonetic contexts. Post-training performance was evaluated using a timed picture-description task, chosen for spontaneous production. Results showed that improvements in spontaneous production were more evident for the easier contrast ([æ]–[ʌ]) than the more challenging [r]–[l] contrast. Interestingly, gains were limited to learners who already had some explicit awareness of the phonemic distinctions. Although the study demonstrated that perceptual training could influence production in quasi-spontaneous contexts, it did not assess comprehensibility using naïve listener ratings. Instead, assessments focused on expert analyses of segmental accuracy, which leaves open the question of whether such gains are perceptible or helpful in communicative terms.

A pivotal reference in the literature on pronunciation instruction and comprehensibility is Zhang and Yuan (2020), which is frequently cited as a methodological reference for evaluating L2 comprehensibility through explicit pronunciation instruction. Zhang and Yuan (2020) recruited 90 second-year undergraduate students from a mainland Chinese university, all of whom were Mandarin native speakers majoring in English. Participants were divided into three classroom groups: a segmental instruction group, a suprasegmental instruction group, and a no-specific-pronunciation control group. Each group received instruction from the same teacher, and lessons were delivered twice a week for 18 weeks. Segmental instruction focused on problematic contrasts such as /ɪ/ vs. /æ/, /θ/ vs. /s/, and /v/ vs. /w/, following a PPP model that included explicit instruction, minimal pair practice, and contextualized production; suprasegmental instruction, in contrast, targeted intonation, linking and word stress. There was no explicit pronunciation training for the control group. Speech was assessed at pretest, post-test, and delayed post-test (20 days later) using sentence reading (controlled task) and picture-based narrative description (semi-spontaneous task). Native speaker raters scored comprehensibility on a 9-point Likert scale. Findings showed that both experimental groups improved in the controlled task, but only the suprasegmental group showed significant, lasting improvement in the spontaneous task. However, the study did not use HVPT and lacked tasks that included interactive or fully spontaneous speech. It also did not assess whether gains in perception transferred to production, nor did it examine intelligibility or perception in depth. Furthermore, the delayed post-test occurred only 20 days after the intervention, indicating a potential insufficient time frame to test long-term retention or integration. Its findings, while useful, are limited in transferability because of classroom setting, lack of online delivery, and semi-controlled nature of most activities, which raise questions about transfer to real-world communication.

These limitations are echoed in recent syntheses and meta-analyses (Tavakoli & Cooke, 2024; Crowther & Isbell, 2023), which note the scarcity of studies that assess true, real-time communicative comprehensibility following pronunciation training. This limitation has become a broader methodological concern in HVPT research, that most studies claiming to assess spontaneous oral production rely on semi-controlled elicitation methods such as picture-description tasks or guided dialogues. While the tasks enable more speech variability and creativity than word or sentence reading, they do not measure the complete cognitive and communicative requirements of language spontaneity. As a result, the current evidence of HVPT's efficacy in facilitating improvements in comprehensibility in spontaneous speech remains limited or indirect. Moreover, research studies rarely use truly open-ended speaking tasks or interactional settings reflecting real communicative situations, which places a limit on how much can be interpreted from the recent findings. This concern was also reflected by Thomson (2018) and Huensch and Tremblay (2015) that recommend further investigations in experimental designs in HVPT research, particularly with respect to comprehensibility outcomes. They point out that future research needs to add delayed post-tests together with open-ended production tasks and listener-based assessments from naive raters as the current research design prevents us from determining if perceptual training interventions result in meaningful improvements for real-life language usage.

The research findings show that HVPT has proven useful for segmental perception improvement and controlled speech production but its effects on comprehensibility in spontaneous speech need further investigation. The study's methodological constraints including semi-controlled tasks and delayed post-tests have made it difficult to determine if perceptual gains can be maintained and applied to real-life communication.

2.7 Chapter Summary

The literature chapter examined the theoretical founding and related studies along with their findings on high variability phonetic training (HVPT), high functional load phonemes, and the transfer of perceptual learning to intelligibility and comprehensibility in L2 speech.

The chapter started with the definitions of comprehensibility and intelligibility and then explained how they are related to L2 speech but are distinct from each other. “Comprehensibility refers to the listener's perception of how easy or difficult it is to understand a speaker's message” (Derwing & Munro, 2015), and is influenced by segmental accuracy, prosody, fluency, lexical appropriateness, along with grammatical precision. Research has shown that comprehensibility can improve significantly through instruction as well as experience, providing a practical role in pronunciation teaching. While, intelligibility, is referred to “as the extent to which a listener can actually understand a speaker's intended message” (Munro & Derwing, 1995) which is an objective measure of communicative success. Studies have narrowed to specific phonological features that have shown to impact to a significant degree intelligibility, with high functional load contrasts playing a particularly important role.

The chapter then investigated a number of possible approaches to pronunciation training reflecting on their merits and also discussing their constraints. These included articulatory, production-based, technology-enhanced training as well as perceptual training and integrated approaches. The discussion moved on the Speech perception theories, including the Perceptual Assimilation Model (PAM) and the Speech Learning Model (SLM), explaining why certain L2 sound contrasts are particularly challenging for L2 learners and how perceptual training may potentially inform the challenging contrasts . The discussed theories suggest that most pronunciation difficulties are due to perceptual confusions rather than articulatory limitations, supporting the application perceptual training in L2 pronunciation instruction.

Within the chapter, the concept of functional load was examined, and the possible influential role it has on pronunciation instruction. The notion of functional load puts forth that phonemic contrasts that distinguish many phonemic contrasting pairs in a language carry greater communicative weight, thus having a greater impact on intelligibility when they are not accurately produced or perceived. Research has supported this principle, showing that errors involving high functional load contrasts have a more detrimental effect on intelligibility than errors involving low functional load contrasts. This finding suggests that prioritizing high functional load contrasts in perceptual training may yield greater communicative benefits than targeting low functional load contrasts.

Lastly, High Variability Phonetic Training (HVPT) was examined in detail, including its principles, methodology, and empirical support. HVPT involves exposing learners to phonetic contrasts produced by multiple speakers in various phonetic contexts, helping them develop robust phonological categories that can generalize to potential new contexts and speakers. Research has discussed HVPT's effectiveness in perception along with the training gains of L2 phonemic contrasts, with evidence of transfer to production in some contexts. However, questions remain about the extent to which these benefits transfer to intelligibility in word production and in turn to comprehensibility in spontaneous speech.

The following have been identified as gaps from the literature:

Although research has demonstrated HVPT's effectiveness for enhancing perception of L2 phonemic contrasts, fewer studies have examined its impact on comprehensibility in spontaneous oral production.

1. The relationship between perceptual training, intelligibility, and comprehensibility requires further investigation, in particular the transferability of learning across these dimensions.

2. Most existing studies rely on semi-controlled or scripted production tasks (i.e. picture descriptions), which do not fully reflect spontaneous, communicatively authentic language use.
3. Few studies include delayed post-tests, making it unclear whether HVPT-related gains are retained over time.
4. Listener ratings are often provided by trained experts rather than naïve listeners, which limits the ecological validity of comprehensibility outcomes.
5. Studies that measure perception, intelligibility, and comprehensibility together within a single, ecologically valid framework remain rare.
6. Sample sizes in HVPT comprehensibility research tend to be small, limiting the transferability and statistical power of findings.
7. HVPT studies are typically conducted in classroom or laboratory settings, which reduces external validity. The present study delivers HVPT entirely online, thereby increasing ecological validity by simulating real-world, self-directed learning conditions.

In responding to the above limitations, the current study aims in providing a further understanding of how perceptual training can improve L2 speech development and propose further effective approaches to pronunciation instruction that prioritize communicative success over native-like accuracy. The findings offer implications for enhancing practice in L2 pronunciation teaching, potentially suggesting insights of how perceptual training can best enhance comprehensibility in real-world communication.

To frame this contribution more precisely, this study develops the term Comprehensibility Transfer Pathway (CTP) to describe the proposed developmental trajectory in which HVPT leads to improvements in phonemic perception, which in turn support intelligibility in word-level production and lead to gains in comprehensibility in spontaneous oral production. This pathway captures the integrated nature of perception and production outcomes and also provides a conceptual method which the study’s design, analysis, and interpretations are grounded.

The following chapter outlines the design of the present study, which was developed in direct response to these challenges.

Chapter 3 Methodology

3.1 Introduction

As established in Chapter 2, perception and speech production of second language learners of English has mainly been researched on single phonemic training (Shinohara & Iverson, 2018) with limited studies on intelligibility of speech production (Lambacher et al., 2005). Past studies on intelligibility have suggested that gains in intelligibility may lead to gains in comprehensibility (Bradlow et al., 1997; Sakai & Moorman, 2018; Thomson, 2018). Research has not yet examined whether High Variability Phonetic Training (HVPT) leads to perception improvements that transfer into production (intelligibility) and subsequently in comprehensibility in spontaneous oral production.

Most perceptual training studies use controlled or semi-controlled tasks to assess outcomes, such as word reading or picture narration (Derwing et al., 1998; Thomson & Derwing, 2016). Although these tasks are useful for measuring speech, they may not fully represent learners' abilities in authentic communication. In fact, free, spontaneous oral production tasks are insufficiently investigated in HVPT research, despite their relevance for evaluating real-world communicative effectiveness (Saito & Plonsky, 2019).

Most prior research (including Zhang and Yuan, 2020) is restricted to classroom-based interventions and instruction is delivered by a teacher in a group format, and assessments are embedded in the curriculum. This reduces external validity and poses challenges in applying findings to autonomous, self-directed learners or to online learning environments. However, the present study simulates the conditions of real-world language learning, where individuals engage with online training lessons and independently work in their own time.

Whereas previous HVPT studies have often assessed perception or controlled settings in oral production, this study tests transfer at all levels: from perception to intelligibility, and in turn to comprehensibility in spontaneous speech context. This provides a holistic, functionally relevant picture of the impact of perceptual training and addresses the critical gap found in the literature (Saito & Plonsky, 2019).

The inclusion of a four-week delayed post-test is an advance in this area of research, as this allows to assess the retention of long-term gains resulting from HVPT. Most prior studies, including Lengeris (2018) and Zhang and Yuan (2020), measured only immediate or short-term effects, leaving open the question of whether gains persist beyond the training period.

Furthermore, by targeting multiple contrasts with high functional load this study ensures that any measured improvements will have practical value in real-world communication. As Zielinski (2008) and Munro & Derwing (2006) emphasize, targeting high functional load contrasts increase the possibility that gains in perception and production may transfer into measurable improvements in intelligibility and comprehensibility.

In sum, this study aims to make several contributions. It directly explores what can be described as a Comprehensibility Transfer Pathway a trajectory through which HVPT- driven perceptual learning impacts intelligibility and ultimately listener-rated comprehensibility in spontaneous oral productions. It provides the first ecologically valid, online HVPT intervention with high functional load phonemes and 15 training lessons with assessment at all levels of the perception–production–comprehensibility chain. It overcomes the classroom or laboratory bias and includes technology enhanced language learning realities. It introduces a more refined methodological framework by including a delayed post-test, free narrative speech, naïve listener ratings, and multiple high functional load contrasts. It builds directly on the empirical gaps outlined in prior literature, particularly those highlighted in meta-analyses such as Saito and Plonsky (2019) and Saito (2021), which have called for research assessing pronunciation training outcomes in spontaneous and ecologically valid communicative contexts.

An additional strength of the present study is the number of participants. Most HVPT studies in the literature have small participant samples, often ranging between 10 and 20 per group, with a mean of approximately 17 participants (Sakai & Moorman, 2018; Thomson, 2018). Studies focusing on speech production and comprehensibility typically employ limited numbers due to the considerable burden placed on human raters. Even studies that include both intelligibility and comprehensibility ratings rarely exceed 20–30 participants (i.e. Derwing & Munro, 1997; Saito et al., 2018). In contrast, the present study includes 51 participants, a sample size exceeding the average and ranks among one the most extensive in this area of research. Furthermore, each participant's speech was evaluated across multiple tasks two measuring comprehensibility (i.e. timed picture story and long turn narrative) and a third assessing intelligibility each rated by human listeners. The multi-layer rating procedures is founded in the multiple assessment tasks, the use of naïve listener judgments, the measurement of intelligibility and comprehensibility across controlled and free spontaneous tasks, and the delayed post-test evaluations. This robust design allows for a more in-depth investigation and understanding of the transfer effects of HVPT on real-world communication. This multi-task rating design, although methodologically demanding, enhances the strength of the research and further extends consistency through its findings. The inclusion of naïve listener raters and complex rating framework strengthens the empirical rigor of this research but additionally responds directly to recent calls in the literature for broader applicability and methodological transparency in HVPT and L2 speech production research (Uchihara et al., 2024).

Expanding on the empirical gaps discussed in earlier sections, recent meta-analyses (Saito & Plonsky, 2019; Saito, 2021) reinforce the finding that most pronunciation instruction studies including HVPT demonstrate strong effects in controlled tasks but significantly weaker or statistically unstable effects in spontaneous speech. These studies put forth the importance to undertake research that examines both controlled and open-ended production tasks, rated by naïve listeners and to reliably represent communicative effectiveness. The present study follows these recommendations and aims to provide new data in this underrepresented area.

With the emergence of blended and fully online modes, it becomes important to test the efficacy of pronunciation (and HVPT in particular) outside the laboratory or classroom setting. Recent studies on computer-assisted pronunciation training (CAPT) and mobile-assisted language learning (MALL) signify the value for digital, individualized, and scalable interventions (Thomson & Derwing, 2015; Thomson, 2018). The present study leverages these technologies and offers evidence for their communicative value.

Ultimately, pronunciation instruction should be judged according to the impact on communicative effectiveness, not just on technical accuracy or laboratory test scores (Derwing & Munro, 2005; Thomson & Derwing, 2015; Tavakoli & Cooke, 2024). By assessing transfer to spontaneous, listener-rated comprehensibility, the present study moves the focus away from linguistic form to communicative function.

Taken together, these findings demonstrate that while HVPT has demonstrated value in enhancing segmental perception and, in some cases, controlled production, its impact on comprehensibility in spontaneous, listener-rated speech remains underexplored. Key methodological limitations of reliance on semi-controlled tasks and absence of delayed post-tests have prevented a clear understanding of whether perceptual gains can be retained and generalized to authentic communication. Furthermore, few studies have assessed the entire progression from perception to intelligibility to comprehensibility within a single, ecologically valid framework. These gaps point to the importance to further an in-depth investigation that examines the transferability of HVPT's impact across all three dimensions using free speech tasks that reflect real-world language use.

Therefore, the main aim of this study is to investigate to what extent improvements in the perception of English phonemic contrasts in isolated words transfer to improvements in intelligibility of word production and in turn to improvements in comprehensibility in spontaneous oral productions among Chinese learners of English.

The participants in this study were L2 learners who spoke Mandarin Chinese as their first language. Studies have shown that Chinese students of English experience speech comprehensibility difficulties because of segmental speech errors (Crowther et al., 2015). It was therefore expected that segmental phonemic contrast training would enhance both perception and intelligibility for the L2 learners and in turn improve comprehensibility in spontaneous oral productions.

This study puts forth the concept of "Comprehensibility Transfer Pathway" (CTP) a coined term to capture the hypothesized trajectory from perception improvement through HVPT to intelligibility gains and ultimately to enhanced comprehensibility in spontaneous speech. This framework underpins the study design and provides an integrated perspective used to construct the research questions.

To achieve this aim, the following research questions guided the study:

Research Question 1: What is the impact of high variability phonetic training (HVPT) on high functional load phonemes on:

a) L2 learners’ perception of phonemic contrasts?

b) the intelligibility of words read aloud by the L2 learners?

c) the comprehensibility of L2 learners' spontaneous oral productions?

Research Question 2: To what extent are any gains in the above dimensions retained long term?

This chapter begins with the pilot study. Then the main study is presented in the following order: participants, research design, materials and the design of the pre, post and delayed-post-tests followed by the ethics.

3.2 Pilot study

This section provides the design outline and outcomes of a single pilot study conducted to inform the main investigation. This pilot study aimed to evaluate the feasibility, effectiveness, how easily understandable and well-structured the training materials were for learners, testing instruments, and procedures. Further to this, the pilot study helped determine the practicality of the research design in real-world conditions and informed critical adjustments made to the main study.

Pilot studies, including this one, are a standard step in experimental research (Van Teijlingen & Hundley, 2001) and provide many functions, including testing instruments for reliability, identifying operational obstacles, and enhancing the methodology. Testing these functions are relevant in pronunciation research, where training design, stimulus quality, and listener reliability are potential factors in impacting the results (Derwing & Munro, 2015).

3.2.1 Purpose

The primary aim of the present study is to examine whether improvements in the perception of English phonemic contrasts with high variability phonetic training (HVPT), transfer to

a) increased intelligibility in word-level production, and b) increased comprehensibility in spontaneous oral production.

The pilot study aimed to test the training intervention and the testing procedures for perception, intelligibility, and comprehensibility. The pilot study also served to assess the feasibility of the training format and the instruments collecting the data. Further to this, testing features that are typical in experimental studies involving technology-based High Variability Phonetic Training (HVPT), as outlined by Thomson (2018), and Qian (2018) is important. These features include:

a. Evaluate the practical use and technical functionality of the online HVPT platform.

b. Assess the suitability of the test items and procedures for perception, intelligibility, and comprehensibility.

c. Confirm whether the high functional load phonemic contrasts selected for training were valid targets.

d. Explore logistical and instructional factors, such as lesson order, audio playback features, and test participant fatigue.

e. Trial the process of consent, the gathered data, and manage the data to assure ethical compliance and participant clarity.

Duration of study

The pilot study was conducted between January 30 and April 18, 2023, with 16 participants, in which 10 completed the full procedure. The reasons given by the participants for not continuing with the study was mainly attributed to academic workload, particularly near the end of term, highlighting the importance of suitable planning time in applied linguistic research (Thomson & Derwing, 2015).

3.2.2 Participants

The 10 learner participants were enrolled in an online English language course offered by a British university and were based in the UK during the study. All participants signed a consent form and completed a short background questionnaire (see Appendix A). The sample included 3 males and 7 females, all over 18 years of age, whose first language was Chinese Mandarin. They reported between 6 to 15 years of English language study. One participant disclosed hearing issues but chose to continue in the study; the remaining nine reported no hearing difficulties.

The two raters (one male and one female) also signed a consent form and completed a short background questionnaire (see Appendix B). Both raters were proficient speakers of English. They had never received training in English language teaching, had not participated in English language learning programs. Both raters were over 18 years old and confirmed that they had no hearing issues. In the main study, an additional question was added to the rater questionnaire related to their familiarity with the Chinese Mandarin language which was not considered in the pilot version of the questionnaire.

All ten learner participants and both raters who participated in the pilot study were compensated with a £5 Amazon voucher upon completing all tests and the training intervention.

3.2.3 Piloting the training material

Initially, two experimental groups were designed for the study: one would train on phonemes with high functional load (FL), and another on phonemes identified as attested difficulties for Mandarin speakers based on previous literature (i.e., Li, 2016; Qian, 2018). However, to increase the ecological validity of the study and align with current best evidence regarding the communicative utility of phoneme contrasts (Wedel et al., 2013) the second condition was removed. It is worth noting that six of the phonemic contrasts categorized under high functional load also appeared in the list of attested difficulties for Mandarin speakers, resulting in substantial overlap between the material stimuli for the two groups. While training materials for the attested difficulties group were developed, this condition was never implemented.

Each learner participant completed 15 HVPT lessons, each with 120 trials (totaling 1,800 trials). Each lesson consisted of three parts: (1) a two-word forced-choice identification activity, where participants selected the correct word from two minimal pair options; (2) a binary identification activity, in which participants answered 'yes' or 'no' to whether the word matched the one shown; and (3) a discrimination task, where participants chose the odd word out from a group of three. These varied formats aimed to engage different perceptual processes and enhance phonemic distinction. The selection of these activity types was supported in established models of L2 phonological development (Logan, Lively, & Pisoni, 1991). Forced-choice identification tasks encourage categorical phoneme perception and have been widely used in HVPT research to train learners in segmental contrast recognition (Bradlow et al. 1997). Binary identification tasks facilitate the development of accuracy in word recognition under variability, while discrimination tasks, such as the odd-one-out format, foster learners' sensitivity to fine-grained acoustic differences (Strange & Dittmann, 1984). Theoretical justification for each activity type is discussed more extensively in the main study design section (see Section 3.5.2.4).

Lessons targeted five consonant and five vowel pair contrasts that were selected based on published research on functional load and theoretical models of communicative efficiency (Wedel et al., 2013; Qian, 2018a).

In response to learner participant feedback during the pilot, the review lessons originally interspersed between training lessons, that is after every second training lesson a review lesson followed; however, these were reorganized. In the main study, all review lessons were moved to follow the 10 core lessons, avoiding confusion and improving lesson flow.

A further technical enhancement was made to allow learner participants to replay audio stimuli and correct their responses before advancing to the next trial. In the pilot study, learner participants were only able to listen to the stimulus once and had to select a response; regardless of whether their response was correct or incorrect, the system moved them on to the next trial. Immediate feedback was provided, but there was no opportunity to replay the audio or retry the item.

In contrast, the main study incorporated two major changes: (1) an audio replay button was added, allowing learner participants to listen to the word as many times as needed, and (2) the system required participants to select the correct response before progressing to the next item. This enhancement was intended to support deeper perceptual learning through repetition and error correction and aligns with recommendations for HVPT training design that emphasize immediate feedback and flexible interaction (Iverson et al., 2005). Providing learners with these affordances also promotes engagement and autonomy, particularly in online settings (Thomson, 2018).

Participants were able to complete the training without technical difficulties. Feedback indicated that the user interface was intuitive, and the trial feedback was clear. As the training was hosted on the Gorilla platform, user activity was automatically logged, allowing the researcher to verify engagement and completion (Anwyl-Irvine et al., 2020).

3.2.4 Piloting the perception test

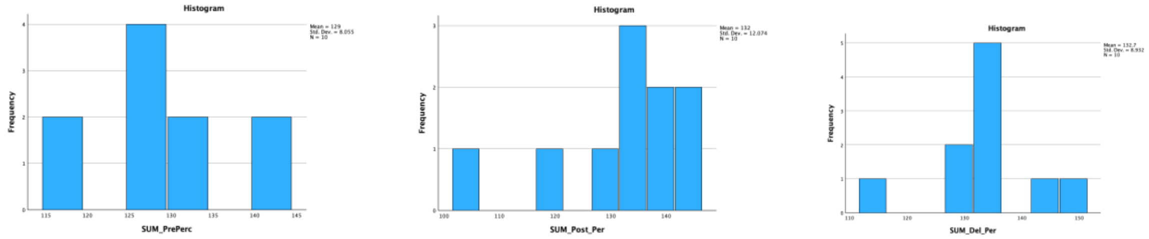
Tests were administered in the following order; first the comprehensibility tests (timed picture story Task 1 followed by long turn narrative Task 2), then the intelligibility test and lastly the perception test.

In the pilot phase, the perception test consisted of 200 trials, equally split between forced-choice identification and binary (yes/no) identification tasks. The trials were reduced to 160 trials, that is 80 forced choice and 80 binary tasks, for the full-scale study to mitigate learner participant fatigue and enhance the learners' rate in completing the test without impacting the internal reliability of the test items. This adjustment was made in line with research that stresses importance in having a full balanced data with participant manageability (Gilakjani & Ahmadi, 2011). Ten participants completed the perception test at three time points (pre-test, post-test, and delayed post-test). Table 1 displays the mean scores and standard deviations of the pilot perception test at pre, post and delayed post-tests.

***Table 1** Mean scores and standard deviations of the pilot perception test at pre, post, delayed post-tests.*

					95% Confidence Interval	
Perception Tests	*N*	*M*	*SD*	*SE*	Lower Bound	Upper Bound
Pre-test	10	129.00	8.06	2.55	123.24	134.76
Post-test	10	132.00	12.07	3.82	123.36	140.64
Delayed post-test	10	132.70	8.93	2.83	126.31	139.09

The results showed a modest upward trend in perception scores over time, with mean scores increasing from 129.00 (SD = 8.06) at pre-test, to 132.00 (SD = 12.07) at post-test, and 132.70 (SD = 8.93) at delayed post-test. Although these changes were not statistically significant, they indicated a potential training-related improvement in perceptual ability following high variability phonetic training (HVPT). To assess data distribution, histograms were generated and the Shapiro–Wilk test for normality was performed. Figures 1–3 illustrate the score distributions at each time point for the perception tests.



(a) perception pre-test (b) perception post-test (c) perception delayed post-test

***Figure 1** Histograms of the pilot perception scores of the (a) pre-test, (b) post-test, (c) delayed post-tests*

The test showed no violation of normality for the pre-test (W = 0.950, p = .664) and delayed post-test (W = 0.925, p = .402). The post-test scores approached significance (W = 0.856, p = .069), suggesting a mild deviation from normality.

***Table 2** Shapiro-Wilk test results for the pilot perception test at pre, post and delayed post-tests.*

Perception Test	W	df	p-value
Pre-test	0.950	10	0.664
Post-test	0.856	10	0.069
Delayed post-test	0.925	10	0.402

Note: p > .05 indicates normality.

Given the small sample size (N = 10), non-parametric statistical methods were used in subsequent analyses to ensure robustness. Effect sizes were computed using Plonsky and Oswald's (2014) adaptation of Cohen's d for within-subject designs. The effect size for the comparison between pre- and post-test scores was d = 0.29, interpreted as a small effect. The effect size between pre- and delayed post-test scores was d = 0.43, indicating a medium effect. These values reflect a moderate upward trend in phonemic perception performance, even if not statistically significant, and provide support for the selected phonemic contrasts to be included in the full-scale study.

A Friedman test was conducted to examine differences in perception scores across the three time points. The results revealed no statistically significant difference, $\chi^2(2) = 4.00$, p = .135. The mean ranks nonetheless suggested a consistent upward trend from pre-test (1.50) to post-test (2.20) and delayed post-test (2.30). This progression demonstrates a potential training effect that may not have reached statistical significance because of limited statistical power from the small, collected sample size.

To further explore differences between specific time points, Wilcoxon Signed-Rank Tests were conducted. No significant difference was found between pre-test and post-test scores (Z = –1.023, p = .306), nor between pre-test and delayed post-test scores (Z = –1.636, p = .102). Despite no evidence of statistical significance, the results showed a consistent upward trend in scores and small-to-medium effect sizes suggest that participants may have begun to internalize phonemic distinctions targeted through HVPT.

This pilot study though limited in size provided initial evidence of the potential benefits of HVPT for enhancing L2 phonemic perception. The positive trajectory of mean scores and modest effect sizes reinforced the decision to proceed with these contrasts in the full-scale study, while also helping refine the test design to support greater feasibility, reproducibility and scalability.

Following the administration of the perception test, it became important to consider further the participant selection criteria. Notably, no initial screening was conducted based on participants' levels of English. This decision was deliberate for the pilot stage; however, the results warranted a closer investigation of whether such a criterion should be incorporated in the full-scale study. The perception test results did not show a consistent pattern between IELTS speaking scores and perception gains. For instance, participants with IELTS 5.0 sometimes showed no gain, while those with 5.5 or 6.0 showed improvement. However, variability in perception tests scores existed among all learners although having different English language levels. Due to incomplete reporting of IELTS scores by all participants and the small sample size, it was concluded that language proficiency scores would not be a reliable criterion for inclusion. Moreover, language levels are not always a predictor of segmental perception gains in HVPT studies (Qian et al., 2018)., This decision also supports more inclusive educational environments that reflect the diverse profiles of L2 learners and the range of L2 learner profiles.

3.2.5 Piloting the comprehensibility and intelligibility tests

The speaking production component included two tests:

Comprehensibility Test: Task 1: A six-frame timed picture story (a family picnic sequence). Task 2: A long-turn narrative task where learners described shop near where they live (at pre- and delayed post-test) and described their favourite shop (at post-test). Learners were given one minute of preparation time. During piloting, it was observed that in Task 1 picture story, most learner participants were ready before the full minute preparation time elapsed. In the full-scale study, this prep time was reduced to simplify the procedure in collecting empirical data without impacting learners' performance.

The two raters rated the comprehensibility test, Task 1 and Task 2, using a 0 - 5 rating scale adapted from De Jong et al. (2012). Each rater randomly rated the same five participants across all three test phases (pre, post, delayed), listening to 30-second speech extracts with audio files accessed via the Gorilla platform. The researcher was present during the online rating sessions to support consistency and answer any questions, helping to ensure standardization across evaluators. The reliability of this rating method was evident in the consistency of the scores, although formal inter-rater reliability statistics were not calculated due to the small sample size.

Intelligibility Test: Each learner participant read 80 words aloud (40 trained + 40 untrained). Recordings were captured via Zoom and later segmented into individual word audio files.

Each of the two raters was assigned 5 participants across three time points, full scoring (5 learners x 80 words x 3 phases = 1,200 words) which was not feasible. To reduce the rating load, only 4 words were selected per participant per test phase. This resulted in a manageable total of 60 words per rater. Raters listened to the recordings of each word and had to choose between two visually presented word options using the Gorilla platform. The same four words were rated for each L2 learner participant across all three time points (pre, post, delayed) to preserve measurement consistency and enable clearer comparisons of intelligibility gains. These four words were carefully selected in advance and kept constant to control for lexical variation across phases. They represented two minimal pairs involving phonemic contrasts that were both high in functional load and attested as difficult for Mandarin speakers of English (Li, 2016; Qian, 2018), ensuring both pedagogical relevance and cross-linguistic sensitivity. This rating task was designed to emulate standard two-word forced-choice identification tasks commonly used in intelligibility research (Munro & Derwing, 1995).

Statistical analysis was not conducted for either the comprehensibility or intelligibility pilot data. The primary goal was to assess feasibility, rating clarity, and data collection logistics. With only five learner participants rated per rater and a reduced set of speech samples, the statistical power would be too low for meaningful inference. Pilot studies of this nature often rely on procedural feedback and descriptive observations rather than inferential statistics (van Teijlingen & Hundley, 2002).

3.2.6 Implications for the main study

Implications of the full-scale study is first the reorganization of the training sequence (10 core lessons followed by 5 review lessons) enhanced learner orientation and minimized confusion. Second, the addition of an audio replay feature, combined with a revised trial format requiring correct responses before progression, was positively received and parallels with theoretical and empirical recommendations for perceptual learning in L2 phonology (Iverson et al., 2005; Thomson, 2018). Third, the pilot demonstrated that language proficiency scores did not systematically predict training outcomes; thus, these were not used as inclusion criteria in the full-scale study (Qian, 2018). Fourth, the HVPT training on high functional load contrasts led to perceptual gains and demonstrated small but consistent improvements, warranting their retention. Fifth, while no inferential statistics were calculated for the speaking tasks, scoring procedures were designed to be consistent and manageable, with two raters following common assessment guidelines. Lastly, in response to pilot observations, the full-scale study will include post-training interviews with the intention to collect richer data on learner experience and training usability (Saito, 2011).

In summary, the pilot study confirmed the feasibility of the training protocol, testing instruments, and procedures of online data collection. They revealed no major design flaws and provided critical information to support and enhance test delivery, training design, and rater management. For the full-scale study, improvements include expanded rater pools, streamlined lesson review structure, adjusted task timing, and added learner feedback mechanisms. These refinements aim to increase the ecological validity, reliability, and overall robustness of the study design.

3.3 Main Study

The following sections of the main study are presented in the order of 3.3 Participants, 3.4 design including the duration and the procedure, 3.5materials, with subsections of 3.5.1

online phonetic training design, 3.5.2 training materials including the 3.5.2.1 audio stimuli, 3.5.2.2 training stimuli, 3.5.2.3 training lesson trials, and 3.5.2.4 training lesson activities. Followed by 3.5.3 pre, post, and delayed-post-tests and lastly by 3.6 ethics.

3.3 Participants

3.3.1 L2 Learners

Online L2 learner participants were recruited from an English language program at a university in Northeast England and from a university in eastern China. The total number of participants was 51. All 51 participants completed the pre-tests and post-tests but only 47 completed the delayed post-tests.

Among the 51 L2 learners, 14 attended an online English language course at a British university in Northeast England, in which 11 were based in China, 2 in northeast England and 1 in Taiwan. The remaining 37 L2 learners attended an English language course on campus at a Chinese public university based in eastern China.

All 51 L2 learner participants shared the same first language, Chinese Mandarin. The majority of them were in the age group of eighteen to twenty years old and most had six to ten years of studying English. Only two of them had lived abroad. Thirty-five were females, thirteen were males, and 3 preferred not to say. Regarding hearing issues, forty-five of the L2 learners indicated having no hearing issues while six reported they had hearing issues. However, all six of them chose to take part in the study.

Further to this, the study included all learner participants irrespective of their English level or the results of their pre-test scores (Iverson et., 2012; Barriuso & Hayes-Harb, 2018) as long as they fulfilled the demographic questionnaire's main criteria. Although there was variability in the learners' pre-test perception scores (64% to 92.5%), the present study did not investigate these different levels. It compared within-group changes across pre-, post-, and delayed post-tests as this approach parallels with typical language program instructional settings that commonly include mixed-ability groupings (Perrachione et al., 2011; Thomson, 2018). This inclusive approach was further supported by prior research showing that HVPT benefits learners at all proficiency levels, from beginners to more advanced users (Hwang & Lee, 2016).

3.3.2 Raters

Five proficient speakers of English (4 females, 1 male) based in the UK were recruited to assess learners' oral productions. That is, the intelligibility and comprehensibility tasks. The research chose non-language teachers as raters. The selection is based on ecological validity which defines comprehensibility as the ease with which average listeners can understand messages during real-world communication (Derwing & Munro, 2005). Non-language teacher raters can evaluate the reception of L2 speech by everyday listeners because they lack training in linguistically informed assessment while these listeners focus on overall communicative effectiveness rather than conscious analysis of pronunciation features. The listener-based nature of comprehensibility requires evaluation from people who match the intended audience of L2 speakers in their natural communication settings (Isaacs & Trofimovich, 2012). The study recruited raters without language instruction training because this approach produced judgments about L2 speech comprehensibility that would generalize to the wider English-speaking population.

All five raters reported none or relatively low familiarity with Chinese accented English (3 with no familiarity, 2 with minimum familiarity). The majority of them spoke an additional language (1 French, 1 Spanish, 1 Spanish and Hebrew, 2 no other language). Their age ranged from 18 to over 42 years old; 2 (18-24 years old), 1 (25-30), 2 (over 42). None of the raters reported hearing issues. All raters were compensated with a 5-pound Amazon voucher upon completing the rating of the comprehensibility and intelligibility tests.

3.4 Design

This study examined the impact of a high variability pronunciation training intervention (Bradlow et al., 1997) on L2 learners' perception, on the intelligibility of word production and on the comprehensibility of L2 learners' spontaneous oral productions through controlled, semi-controlled and free speech assessments.

The study used multiple assessment methods to determine whether High Variability Phonetic Training (HVPT) perceptual gains transfer to oral productive skills. The general design of the study included a perception test that consisted of two parts: a forced-choice identification and a binary yes/no identification task to evaluate phonemic perception; a read-aloud word task served to evaluate the intelligibility of word pronunciation; and semi-controlled and free speech tasks to assess comprehensibility in spontaneous oral productions. The multiple assessment approach shows how L2 speech development involves many dimensions because perception, intelligibility and comprehensibility represent related yet distinct constructs (Munro & Derwing, 1995; Levis, 2005). The assessment of speech follows established models which recommend using both controlled and free spontaneous speaking tasks to evaluate intelligibility and comprehensibility (Derwing & Munro, 2015). The assessment of intelligibility uses a read-aloud word task to provide consistent results while picture description and monologue tasks provide more naturalistic language use to measure comprehensibility in real-world communication (Isaacs & Trofimovich, 2012). The study combined different tasks to monitor how perceptual learning develops the ability to communicate more clearly in formal and spontaneous contexts.

This study followed a quantitative, within-participant design using a single-arm structure which involved fifty-one Chinese learners of English. All learners completed an intervention, pre-, post-, and four-week delayed post-tests (Lively et al., 1994; Yeon, 2008) to measure changes over time, allowing each individual to serve as their own control. An unstructured interview was also administered immediately after the intelligibility post-test on learners' experiences and perceptions of the training intervention. Figure 2 summarizes the general design of the study.

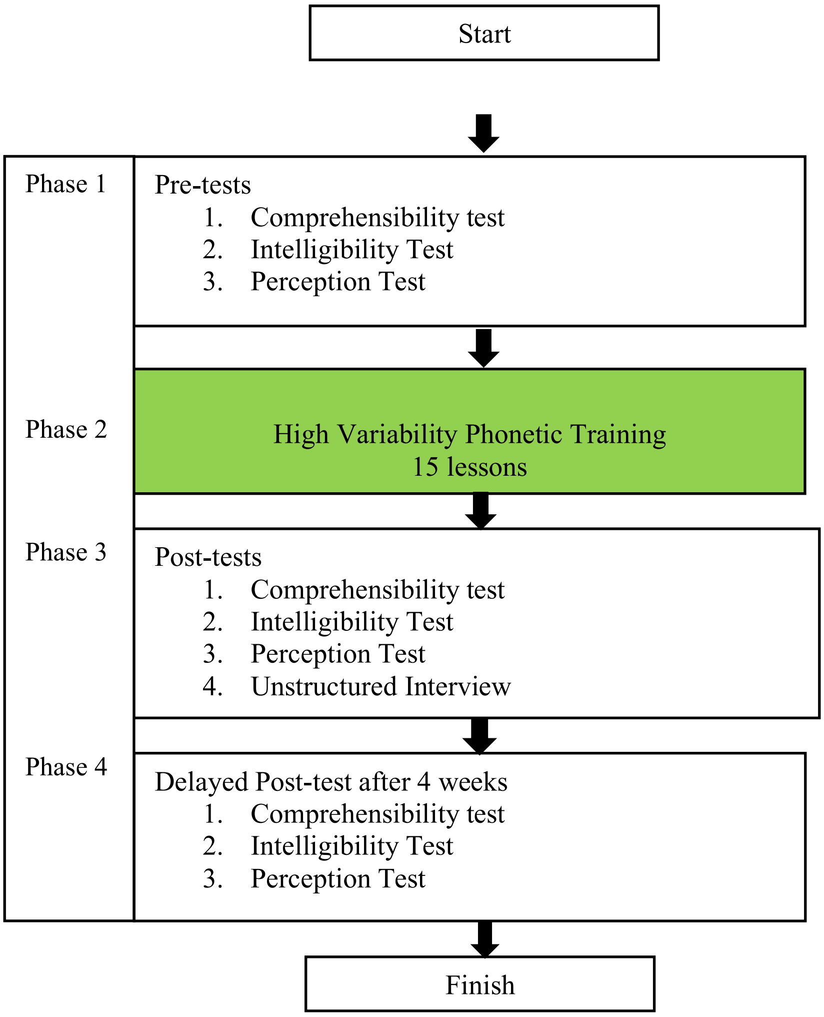


Figure 2 *General design of the main study*

The main dependent variables were (1) L2 learners' perception of phonemic contrasts (2) the intelligibility of word production, and (3) the comprehensibility of spontaneous oral production (see Table 3).

Table 3 *Dependent variables and assessment types in the study*

Variables	**Type of Assessment**
L2 learners' comprehensibility in spontaneous oral production	Task 1: timed-picture story Task 2: long turn narrative speaking
L2 learners' intelligibility of word production	Individual words read aloud
L2 learners' perception of phonemic contrasts	Phonemic contrasts (two-word forced choice identification and binary [yes/no] identification activities)

Duration of the main study

The study was conducted over a period of four months which started on May 30, 2023, and finished on October 1, 2023. Not all L2 learners began at the same time. One cohort of L2 learners began the online training between May 30th – June 8th while the second cohort of L2 learners began the training between August 1st – August 7th. Each L2 learner began and completed the online training at their own pace and in their own time to allow for flexibility to the learner to adapt the online training to their individual daily schedules. The learners were encouraged to complete a single training lesson everyday day or every two days.

Due to the nature of the online intervention, it was not fully feasible to control L2 learners' training time. However, the flexibility of completing the online training at their own pace parallels real world learning conditions, allowing L2 learners' autonomy in how often they engaged with the training and when they decided to take breaks. An approach that caters to L2 learners' needs and personalized learning experiences (Lai et al., 2022). Each learner completed the training asynchronously at their own pace through the Gorilla Experiment Builder platform. On average, participants completed one lesson every 1.4 days (N = 43) or every 3.5 days (N = 8), depending on their individual schedules. As the system allowed users to leave lessons open without logging off, exact completion times could not be systematically recorded. However, based on log data from a subset of learners, each training lesson typically took between 8.42 and 21.45 minutes to complete (see Appendix S for summary of lesson duration data, including total recorded time and number of learners per lesson), although some sessions appeared longer when learners left lessons open while inactive or had no recorded time data because the training platform remained logged in even when learners were not actively engaged.

Following the training intervention, immediate post-tests were administered the day following their last training lesson and up to 1 week later, except for one L2 learner that completed the post-tests 3 weeks after the intervention. Delayed post-tests were administered between 4 and 4.5 weeks after the post-tests. Overall, learners experienced a complete cycle of training and testing approximately 16.5 weeks per participant, from pre-test to delayed post-test.

Procedure of the main study

The L2 learners began with the comprehensibility tests (task 1 the timed-picture story, and task 2 the long turn narrative) then the intelligibility test (read aloud words), and finally the perception test. The perception test was completed last to avoid possible training effects from the two previous tests (Fouz-González, 2020) such as explicitly drawing L2 learners' attention to specific features.

3.5 Materials

3.5.1 The online phonetic training design

The high variability phonetic training was designed and presented online on the *Gorilla* Experiment Builder platform (https://gorilla.sc/). The online platform allowed the researcher access to real time data including current stage in training, time commitment, reaction time, incorrect and correct responses. The L2 learners were provided with immediate feedback for their responses. All participants, L2 learners and raters, required a computer/laptop or tablet and access to the internet.

3.5.2 Training material

3.5.2.1 Audio stimuli

Audio recordings of natural stimuli were extracted from the Oxford English Dictionary (OED) online database. OED uses multiple human speakers including both female and male voices for the audio recording of their dictionary words. Including both natural stimuli and multiple speakers in high variability phonetic training is important as both are effective contributors to phonetic learning. Natural stimuli maintain the accurate acoustic distinctions in forming phonetic categories (Deng et al., 2018; Iverson et al., 2005), and the variability of the speakers allows the L2 learners to generalize the learnt category across different speakers and communicative contexts (Wong, 2014). Together these elements, help promote deeper and more lasting improvements in perception and production of non-native speech sounds.

A total of 480 audio stimuli were downloaded to the researcher's laptop for use in the online phonetic training lessons, the perception test, and the intelligibility test. Although speaker variability was present, the exact number of individual speakers featured in the stimuli is unknown, as the OED does not provide specific information about the speaker identities or numbers.

3.5.2.2 Training stimuli

As mentioned, the online phonetic training used audio recordings of words from the OED for the training stimuli. The training stimuli were phonemic contrasts of vowels and consonants with a high functional load of 10 (1 = low, 10 = high) from Brown's (1988) list. Phonemic contrasts with high functional load are contrasts for which there are largest number of minimal pairs and as such play a significant role in determining the comprehensibility of speech (Kewley-Port, 2007). The minimal pairs in the study were contrasted in different phonetic contexts; initial-middle-final position (**p**ump-**b**ump; di**pp**er-di**bb**er; ta**p**-ta**b**). Variability in phonetic contexts can assist learners to further adapt their perceptual strategies to identify and produce sounds but can also lead to better generalization of learning for more different real-life contexts (Bradlow et al., 1997; Iverson et al., 2005; Wong, 2014). The design follows the study's rationale that training phonemes with higher communicative weight, those most likely to disrupt meaning when misperceived, are capable of bringing more functionally significant gains in intelligibility and comprehensibility. The concept of functional load served in this study as a systematic selection tool to identify the contrasts which best supported communicative effectiveness. Table 4 presents the ten vowel and consonant phonemic contrasts with the highest functional load, as ranked by Brown (1988), that were used in the study with examples of phonemic contrasts for each pair sound.

Table 4 *The ten vowels and consonant phonemic contrasts with the highest functional load as ranked by Brown (1988) used in the study*

Vowels (Functional Load ranked at 10)		Consonants (Functional Load ranked at 10)	
	example		example
/ɛ - æ /	beg - bag	/p - b/	tap -tab
/æ - ʌ/	lack - luck	/m - n/	mode - node
/æ - ɒ/	cap - cop	/l - r/	collect - correct
/ɔ: - əʊ/	corn - cone	/p - f/	wipe - wife
/ʌ - ɒ/	bus - boss	/n - l/	nine - line

3.5.2.3 Training lesson trials

The experimental intervention included 15 training lessons with a total of 1800 trials (see Appendix E for stimuli used at each individual lesson). L2 learners received training on ten lessons targeting high functional load (FL) phonemic contrasts, followed by five review lessons. Each of the ten initial lessons focused on a single phonemic contrast (e.g., /ɛ - æ/), using a set of 40 different words. These 40 words were presented across 120 trials, with random repetition across the three sections of each lesson.

The five review lessons combined two previously trained phonemic contrasts per session. For example, Review Lesson 1 covered /ɛ - æ/ and /p – b/, incorporating 40 words from each original lesson for a total of 80 words. These words were also randomly repeated across 120 trials. The duration of each training lesson depended on each learners’ performance.

3.5.2.4 Training lesson activities

The training lessons comprised of three different perceptual activities with immediate feedback on all responses. The aim of these activities was to improve English phonemic contrast among learners through a two-word forced-choice identification activity, a binary (yes/no) identification activity and an odd-one-out discrimination activity. The present study adopted these activities because they replicated previous HVPT procedures (e.g., Bradlow et al., 1997; Iverson et. al, 2005; Thomson & Derwing, 2016: Qian, 2018) and both types of activities have shown to benefit learners (Carlet & Cebrian, 2015; Flege, 1995b). The identification activities (forced-choice and binary formats) helped learners identify and sort target phonemes, whereas the discrimination activity helped them detect small acoustic differences between similar sounds.

The training lesson activities needed to support phonemic accuracy and perceptual flexibility because comprehensibility requires real-time listener judgments in less structured speech. By training learners in different types of perceptual activities, the training intervention aimed to improve learners' phonemic contrasts thus leading to improvement in comprehensibility in spontaneous speech. Figure 3 shows an individual training lesson with the different types of activities and the number of trials in the lesson.

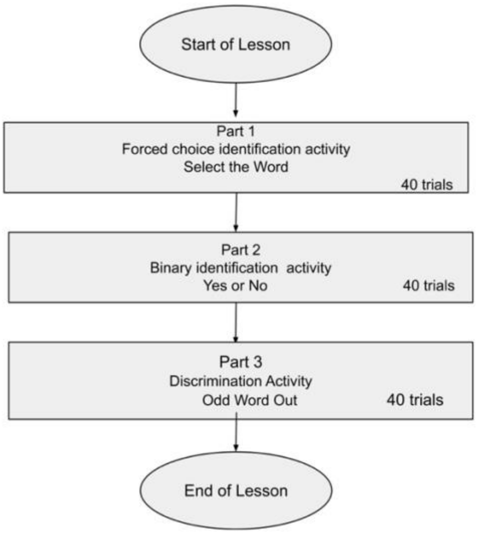


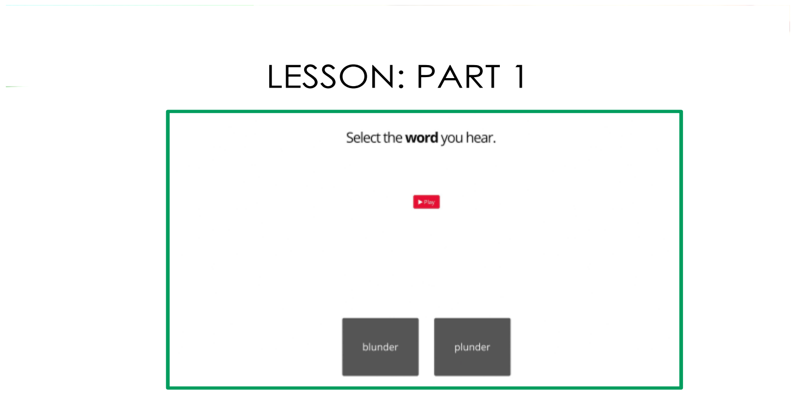
Figure 3 *An individual training lesson showing the different types of activities and the number of trials in the lesson*

The HVPT training lessons consisted of three parts, each corresponding to a different perceptual activity, as shown in the interface screenshots.

Part 1: Two-word forced-choice identification activity (Figure 4). The L2 learners listened once to the audio by pressing the play button and selected one of the two words displayed on the screen that they believed matched the word they heard.

Part 2: Binary identification activity (Figure 5). A single word was displayed on the screen, and the L2 learners listened once to the audio by pressing the play button. They then selected either the YES or NO button on the screen, depending on whether they believed the audio matched the written word.

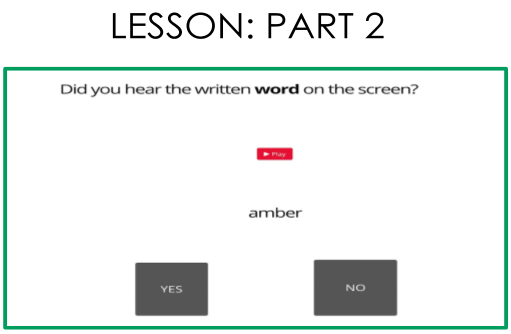
Part 3: Discrimination activity – odd-one-out (Figure 6). The L2 learners listened once to three audio clips by pressing each of the play buttons and selected one of the three-word buttons (Word 1**,** Word 2**,** or Word 3) that they believed was the odd one out.



LESSON: PART 1

Select the **word** you hear.

Figure 4 *Training lesson interface Part 1 Two-word forced-choice identification activity*

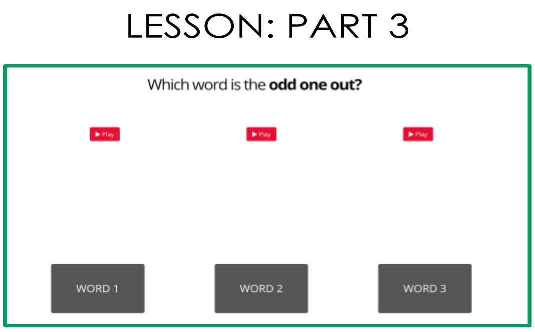


LESSON: PART 2

Did you hear the written **word** on the screen?

amber

Figure 5 *Training lesson interface Part 2 Binary identification activity Yes or No*



LESSON: PART 3

Which word is the **odd one out?**

Figure 6 *Training lesson interface Part 3 Discrimination activity odd word out*

3.5.3 Pre, post and delayed post-tests

A variety of research methods using high variability phonetic training have been used to assess L2 learners' oral speech productions. These methods included controlled oral speech assessments such as reading sentences (DeJong 2012; Derwing et al., 2014) and reading words aloud (Uchihara et al., 2024). Other employed methods were semi-controlled oral speech assessments such as timed picture description (Isaacs & Trofimovich, 2012; Suzukida & Saito, 2019).

This section describes the three types of tests that were administered at pre, post and delayed post-test. The administered tests were a comprehensibility test which included two tasks, an intelligibility test and a perception test.

The comprehensibility and intelligibility tests were video recorded on Zoom and then downloaded on the researcher's MacBook Air laptop and converted to audio format recordings. All tests were the same and repeated at pre, post and delayed-post-test with one exception, the comprehensibility test Task 2 from pre to post test. The comprehensibility test Task 2 post-test had a similar topic but not exactly the same as the pre-test (see section 3.5.3.1). Table 5 shows the order and purpose of each test in the study.

Table 5 *The order and purpose of pre-post-and delayed-post tests*

The Order of Administered Tests	Purpose
Comprehensibility Test	
Task 1: Timed picture story speaking task	To elicit samples of the learners' spoken English in spontaneous speech
Task 2: Long-turn narrative speaking task	
Intelligibility Test	To assess the learner's pronunciation of read aloud words
Read aloud 80 words	
Perception Test	To measure whether learners can identify the target phonemes.
160 trials	

3.5.3.1 Comprehensibility test

The purpose of the comprehensibility test was to elicit samples of the L2 learners' spoken English for the analysis of the learners' spontaneous oral productions. The test comprised of two tasks which took place via Zoom. Each session was video recorded and immediately converted to audio format for storage and analysis. The comprehensibility test involved a timed-picture story task and an IELTS- style long turn narrative speaking task. Each task was rated on a 0 - 5-point scale by five highly proficient speakers of English.

Stimuli comprehensibility tests

The material for the comprehensibility test included two tasks: Task 1 a timed picture story speaking task and Task 2 an IELTS style long-turn narrative speaking task. The reason for using a picture story task and an IELTS long-turn style narrative task is that the former has limitations in eliciting spontaneous speech as it is a semi-controlled task supported by visual prompts. On the other hand, the long-turn narrative is less constrained and less structured, allowing the learner more freedom to create content, generate ideas and produce extended speech (Foster & Skehan, 1996). This differentiation is important given that the present study aims to evaluate the impact of improvements in perception and intelligibility transfer to comprehensibility in spontaneous oral production.

The two tasks provided a full assessment of learners' comprehensibility in spontaneous oral production at two different levels of spontaneity. The timed-picture story task included the same visual prompt across all learners, whereas the long turn narrative task reflected real speech production by requiring speakers to develop their own ideas. Furthermore, a long-turn narrative task, elicits more stretches of spontaneous speech from learners (Suzukida & Saito, 2019) and different learner responses. The goal in designing these two activities was to provide a more holistic assessment of the impact of spontaneous oral production of HPVT beyond controlled test conditions.

Comprehensibility Task 1 involved a timed picture story speaking task (see Appendix H). of a family going on a picnic day in a six-picture sequence. This task was the same across pre, post and post-delayed tests. Utilizing the same tasks across the three time points allowed for a more reliable assessment to measure L2 learners' oral performance as there was a direct comparison across the three time points (Isaacs & Trofimovich 2012)

Comprehensibility Task 2 comprised of an IELTS style long-turn narrative speaking task (see Appendix I). For the pre-test, learners were asked to talk about a shop near where they lived now that they sometimes visited. There were also a few supporting questions to guide them. The same question was asked at delayed post-test. However, for the post-test, learners had a similar topic but a different topic. At post-test, they were asked to talk about their favourite shop they visited with a few supporting questions to guide them (see Appendix I). It was important to not have the same topics as this was a free speech task, and the aim was to assess spontaneous oral productions in a more natural context. This approach was better suited to real-life communication, where the L2 learners came across similar but not identical topics and their language had to be adapted accordingly (Saito, 2020; Zhang & Yuan, 2020).

Procedure for L2 learner comprehensibility tests

The comprehensibility test Task 1 and Task 2 were conducted and video recorded on Zoom. The Zoom video recordings were converted to audio files, coded and saved. First task 1 was conducted immediately followed by Task 2.

For Task 1 (timed picture story speaking task), the L2 learners were told that they would be looking at a set of pictures for 10 seconds and to describe the story they see as best they can. They would be speaking for up to 2 minutes and they would be recorded. When the L2 learners confirmed they understood the task, they practised on a set of four pictures to become familiarized with the task (see Appendix F). Immediately afterwards, they moved on to the real test. Once again, they were instructed that they had 10 seconds to look at the set of six pictures before speaking for 2 minutes (see Appendix H). The decision to allow 10 seconds preparation time is that past studies using visual prompts have allowed from 5 seconds (Suzukida & Saito, 2019; Ellis, 2005) to a few minutes (Derwing et al., 2008) preparation time but with no consistency found throughout the studies. Whereas 10 seconds has been found to be sufficient planning time for learners to organize their ideas minimally and “minimize conscious speech monitoring” before their oral production (Saito & Hanzawa, 2016; Zhang & Yuan, 2020). Therefore, taking into consideration the past studies 10 seconds was deemed as appropriate preparation time for the task.

Following Task 1, L2 learners immediately moved on to Task 2 (long-turn narrative speaking task). For Task 2 the L2 learners were told that they would be speaking on a topic for up to 2 minutes and that they would be recorded. The topic would be shown to them before speaking and they would have 1 minute to think about what they plan to say The 1-minute preparation time allowed the learners to organize their thoughts without becoming overly dependent on rehearsed structures (De Jong et al., 2012; Saito et al., 2020) but also represented real-life speaking contexts which require quick thinking before oral speech production. The learners were told that notetaking was not allowed. Once the L2 learners confirmed they understood the task, they practised using a sample topic (see Appendix G). Immediately afterwards, they began the real test. Once again, they were instructed that they had up to 1 minute (Crowther et al., 2015) preparation time before beginning to speak for 2 minutes (see Appendix I). The researcher was not engaged in the conversation of Task 1 nor Task 2.

Procedure for rater comprehensibility tasks

Five raters rated the comprehensibility of the L2 learner's spontaneous oral productions. For each task, each rater rated the productions of ten L2 learners. They rated the pre, post and delayed post-tests of those ten L2 learners presented. To control for rater effect, these raters rated the pre, post and delayed tests for the same L2 learners. To control for order effects, the order in which the L2 learners in the pre, post and delayed post-tests were presented to each rater was randomized. Table 6 shows the allocation of raters to comprehensibility tests, and that each test was rated at pre-, post-, and delayed post-test stage by the same rater for the L2 learners indicated in Table 6.

Table 6 *Allocation of raters to comprehensibility tests, each test rated at pre-, post-, and delayed post-tests stage by same rater for the L2 learners*

	Rater Allocation	
	Comprehensibility test	
Participants[a]	**Task 1: Timed Picture Story**	**Task 2: Long Turn Narrative Speaking**
1 - 10	Rater 1	Rater 5
11-20	Rater 2	Rater 3
21-30	Rater 3	Rater 4
31-40	Rater 4	Rater 2
41-51	Rater 5	Rater 1

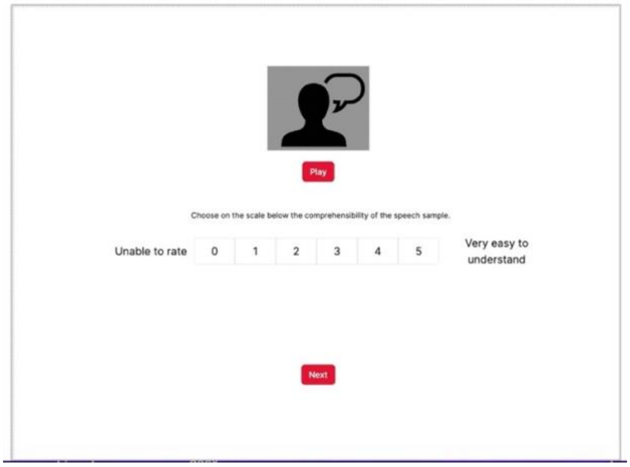
[a]Participants 41-47 completed pre-,post- and delayed tests, participants 48-51 completed pre- and post-tests only.

Raters were asked to make comprehensibility judgments on L2 learners' speech samples to indicate how difficult or easy it was to understand the speech samples. Similar instructions were provided in past studies and were noted as reliable methods (Munro et al. 2006; Derwing et al., 2008).

To rate comprehensibility, 30-second speech samples from L2 learners were selected at the beginning of their audio recordings from Task 1 and Task 2 (Derwing et al., 2008; Crowther et al., 2015a). They were rated based on the 0 - 5-point (from 0= being unable to rate, 5 = being very easy to understand) based on DeJong et al., (2012) rating scale. The rating procedure took place synchronously online over Zoom using the *Gorilla* Experiment Builder platform for the rating task.

Raters had a short training session just before starting the real rating procedure. For the training, the raters were told that they would be listening and rating Chinese learners describing a picture story of a family going on a picnic (see Appendix J). The picture was shown to the raters to familiarize themselves with it. The raters were told they would be rating how easy it is to understand what the L2 learners are trying to convey using a scale from 0 -5 (see Appendix J). They were informed they would be rating 30 audios which were 30 seconds each. The raters were introduced to the rating scale and a copy was immediately sent to them. A link from the Gorilla website was sent to the raters to begin practice rating three 30-second audio samples on the timed picture story. The raters logged in to the Gorilla website with their anonymous assigned identification numbers and followed the instructions on the screen. The practice audio samples of the timed picture story were from the pilot study's L2 learners. Immediately afterwards, they moved on to the real test. Once again, the instructions were repeated for the raters before starting the rating of the real audio.

For the long turn narrative Task 2, the raters were informed that speaking task followed exactly the same rating scale and interface and therefore, no practice audio samples deemed necessary. The long turn narrative topic was introduced before starting the rating procedure and they were informed again that they would be rating 30 audios which were 30 seconds each. The raters were told that they would be listening and rating to Chinese learners talking about a shop near them or about their favourite shop (see Appendix K). The raters were told they would be rating how easy it is to understand what the L2 learners are trying to convey using the same scale from 0 -5 which they used for Task 1 (timed picture story). A new link was sent to the raters to log in again to the Gorilla website to begin Task 2. Figure 7 shows the interface that was used by the raters to rate Task 1 and Task 2 of the comprehensibility tests.



Choose on the scale below the comprehensibility of the speech sample.

Unable to rate 0 1 2 3 4 5 Very easy to understand

***Figure** 7 The interface used by the raters to rate the comprehensibility test Task 1 and Task 2*

Data analysis comprehensibility tests

For the comprehensibility test, the raters listened to the 30-second audio responses of the L2 learners and rated the comprehensibility of each response on a scale from 0 -5. The raters' listening scores were compared across the comprehensibility tests at pre, post and delayed post-tests. This was done for Task 1 (timed picture story) and the same procedure was followed for Task 2 (long turn narrative speaking).

As mentioned in section 3.4.1, assessment data was collected from 51 Chinese Mandarin native speakers who were second language learners of English (L2). However, four L2 learners did not complete the delayed post-tests. Thus, data was reported for 47 L2 learners at pre, post and delayed post-tests and 51 L2 learners at pre and post-tests. The number of L2 learners are presented in Table 7.

***Table** 7 Number of L2 learners included in the analysis of the comprehensibility tests (Task 1 & Task 2)*

Comprehensibility Tests (Task 1 & Task 2)	
Type of test	*N*
Pre	51
Post	51
Delayed post	47

3.5.3.2 Intelligibility test

The intelligibility test was presented on Zoom, video recorded and then converted to audio recordings which were saved in files. The purpose of the intelligibility test was to assess learner's pronunciation of words read aloud. Learners' pronunciation was assessed by listening raters' total number of correct responses in a two-word forced-choice discrimination task.

Stimuli for L2 learner intelligibility test

The intelligibility test involved 80 stimulus words. These included 4 trained and 4 untrained words from each of the 10 high FL phonemic contrasts. The intelligibility test was the same across all pre, post and post-delayed tests. Each of the 80 stimuli was shown on an individual power point slide over Zoom. The table shows the phonemic contrasts and number of words that each L2 learner was tested in the intelligibility test. This included trained and untrained words at pre, post, and delayed-posts tests. Table 8 presents the word stimuli including trained and untrained words that the L2 learners read aloud in the intelligibility test at pre, post and delayed post-tests.

Table 8 *The 80-word stimuli used in the intelligibility test including trained and untrained words at pre-, post-, and delayed post-tests*

Intelligibility Test		
Phonemic Pair	Trained words	Untrained Words
/ɛ -æ/	and -end land - lend	pat – pet expend - expand
/p - b/	rip – rib harper - harbour	pail – bail Pete - beat
/æ-ʌ/	chump – champ umber – amber	cat – cut app - up
/m-n/	sunning – summing mime - mine	homing – honing nail - mail
/æ-ɒ/	ox – axe bland - blond	tag – tog mac - mock
/l-r/	alive – arrive finery - finally	rack – lack jelly - Jerry
/ɔ: - əʊ/	portion – potion mow - more	board – bode torn - tone
/p - f/	snipping – sniffing top - toff	perry – ferry strife - stripe
/ʌ - ɒ/	buddy – body rock - ruck	snob – snub rubbery - robbery
/n-l/	collect – connect lead - need	light – night mealtime - meantime

Procedure for L2 learner intelligibility test

The intelligibility test was the second test administered to the L2 learners immediately after the comprehensibility test. The L2 learners were told that they would be reading aloud 80 individual words and would be recorded on Zoom. First, they would listen to the pronunciation of all 80 words and then they would read the words aloud while being recorded. The presentation of each word was presented on an individual power point slide. Once they understood this, they proceeded to the test.

Stimuli used in the intelligibility rater task

A high number of stimuli (11,920 words) was produced from the 80 words read aloud by the 51 L2 learners. The 80 words included four trained and four untrained words from each of the 10 phonemic contrasts. However, as rating such a large number of stimuli was not feasible due to rater fatigue and time constraints, each rater assessed 10 L2 speakers at the pre-, post-

and delayed post-test stages on 12 untrained words drawn from six phonemic contrasts. It was decided to remove the following phonemic contrasts from the rating procedure: /ɔ: - əʊ/, /p - b/, /p - f/, and /m - n/, as these contrasts are not considered attested difficulties for Chinese learners (Rogers, 1997; Rogers & Dalby, 2005; Jia et al., 2006), despite being ranked high in functional load. The final six high functional load phonemic contrasts (/æ - ɛ/, /æ - ʌ/, /æ - ɒ/, /ɒ - ʌ/, /l - n/, and /l – r/) were rated based on a subset of 12 untrained words per L2 learner at each testing stage (pre-, post-, and delayed post-test). Table 9 represents the final phonemic contrasts selected for rating and the total number of words rated per L2 learner.

Table 9 *The final phonemic contrasts selected for rating and the total number of words rated per L2 learner on the intelligibility test*

Intelligibility Test				
Phonemic Contrasts Tested on Pre, Post, Delayed Tests		**Final Phonemic Contrasts used in Rating**		
Phonemic Contrast Pairs	No. of Words	Phonemic Contrast Pairs	No. of Words	Words Used Pre/Post/Delayed test
/æ - ɛ/	8	/æ - ɛ/	2	pat- pet
/æ - ʌ/	8	/æ - ʌ/	2	app-up
/æ - ɒ/	8	/æ - ɒ/	2	tag-tog
/ɒ - ʌ/	8	/ɒ - ʌ/	2	robbery-rubbery
/ɔ: - əʊ/	8			
/l-n/	8	/l - n/	2	light-night
/p - b/	8			
/p - f/	8			
/l – r/	8	/l – r/	2	jelly-Jerry
/m -n/	8			
Total trained & untrained words	80	Total untrained words	12	
Total trained & untrained words per pre-post-delayed tests	240 words	Total untrained words per pre-post-delayed tests	36	

Procedure for rater intelligibility task

For the intelligibility test, the same five raters were allocated in the same procedure as in the comprehensibility tests. That is, each rater rated the intelligibility of ten L2 learners. They rated the pre, post and delayed post-tests of those ten L2 learners presented. To control for rater effect, these raters rated the pre, post and delayed tests for the same L2 learners. To control for order effects, the order in which the L2 learners in the pre, post and delayed post-tests were presented to each rater was randomized. Table 10 represents the allocation of raters to the intelligibility test, and that each test was rated at pre-, post-, and delayed post-test stage by the same rater for the L2 learners.

Table 10 *Allocation of raters to intelligibility tests, each rated at pre-, post-, and delayed post-test stage be same rater for the L2 learners indicated*

	Rater Allocation
Participants[a]	**Intelligibility Test**
1 - 10	Rater 3
11-20	Rater 4
21-30	Rater 5
31-40	Rater 1
41-51	Rater 2

[a]Participants 41-47 completed pre-,post- and delayed tests, participants 48-51 completed pre- and post-tests only.

Raters were asked to make intelligibility judgments on L2 learners' pronunciation of individual words. To rate intelligibility, raters were asked to listen to the audio of the L2 learners' pronunciation of each read aloud word and rate the intelligibility of each word by choosing between a two-word forced choice task. In the task, two words appeared on the screen and raters chose one of the words they believed best represented the word they heard. The particular type of task was chosen as it assisted raters to focus on the distinction of the phonemic contrasting sound (Hazan et al., 2005). The rating procedure took place synchronously online over Zoom and used the *Gorilla* Experiment Builder platform (https://gorilla.sc/) for the rating task.

The procedure was as follows. Upon completing the rating of the comprehensibility tests, the raters moved on to the intelligibility test. Raters were told that they would be listening and rating the pronunciation of each word read aloud by the learners. The interface of the task was shown to the raters. All five raters assessed 36 words per learner (12 - pre, 12 - post and 12 - delayed post-test), with the exception of 4 learners who did not complete the delayed post-test. For these learners, the rater assessed 24 words each (12 - pre and 12 - post). The raters were not administered any practice samples because of the simplicity of the task. Once the raters understood the instructions, a link from the Gorilla website was sent to them to begin rating. They logged in to the Gorilla website with their anonymous assigned identification numbers and followed the instructions on the screen. Figure 8 shows the interface that was used by the raters to rate the intelligibility test.

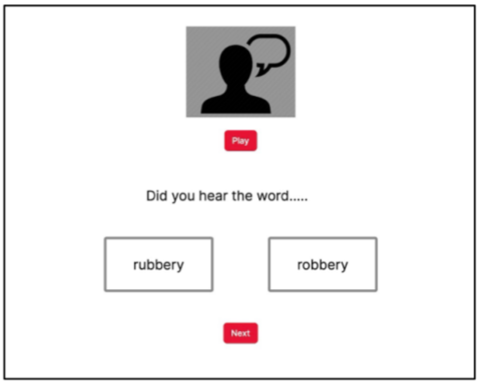

Did you hear the word.....

Figure 8 *The interface used by the raters to rate the intelligibility test*

Data Analysis intelligibility test

Rater data was collected through the raters' responses. Each rater listened to the L2 learners' pronunciation of the 12 untrained words and rated the intelligibility of each item using a two-forced-choice identification task. The raters chose between two words displayed on the monitor. The chosen word was coded as 1 (correct word) or 0 (incorrect word). Raters' listening scores were summed across the 12 untrained items for each L2 learner and then aggregated across all learners. This procedure was applied at all three time points: pre-test, post-test, and delayed post-test.

However, among the 51 L2 learners, 18 had missing ratings on word items at one or more of the three time points (pre-test, post-test, or delayed post-test). A power analysis was performed to confirm whether the remaining 33 participants had sufficient statistical power and if it was a reliable sample size to run the data analysis on the intelligibility tests using a repeated measures ANOVA to evaluate the changes across the three-time points: pre-test, post-test, delayed post-test (Field, 2024). The G*Power 3.1 (Faul et al., 2007) showed that a sample size of 28 was needed for a medium effect size (f = .25), based on Cohen's (1988) conventions for repeated measures ANOVA, with .80 power (α = .05) across the three-time points. While Cohen's framework was used for the power analysis, subsequent effect size interpretations in this thesis (e.g. perception results) draw on Plonsky and Oswald's (2014) L2 specific benchmarks, which provide field appropriate thresholds for interpreting practical significance in second language research. The sample size 33 used in the study was higher than the threshold and resulted in a power of .81 which was sufficient to run the data analysis. Based on the power analysis results, it was decided to exclude these 18 learners from the analysis to maintain reliable longitudinal comparisons.

The final intelligibility analysis was conducted on a subset of twelve words that had complete ratings across all participants and testing times, allowing for balanced group-level comparisons of mean performance at the three measurement points. Missing data were handled through listwise deletion, which ensured that only participants with complete ratings were retained for analysis. While repeated-measures ANOVA was an appropriate method for comparing mean performance across the three testing times at the group level, it does not account for random variation across individual participants or items. Alternative approaches such as multiple imputation or mixed-effects regression modelling, as discussed in Plonsky and Oswald (2017) and Mifka-Profozic et al. (2020), could provide a more flexible approach for future research by modelling this variability directly. However, the key focus of the present study was on group-level trends rather than individual trajectories, and the dataset was balanced after the exclusion of incomplete cases, making the ANOVA approach suitable for the current research aims. Table 11 shows the 18 L2 learners and the words that were not rated on the subset of untrained 12-word items at any of the testing stages.

Table 11 *The 18 L2 learners and the words that were not rated in the intelligibility test at pre-, post-, and delayed post-tests*

	Intelligibility Test	
	Participants	Words not rated
Pre-test	3	light
	1	robbery
Post-test	1	up
	1	jelly
	1	rubbery
	1	tag
Delayed post-test	2	app
	1	jelly
	1	light
	2	robbery
	2	rubbery
	2	tag
Total Participants	18	

A Cronbach's Alpha was calculated to validate the internal consistency within the subset of the 12-word items with the same 33 L2 learners who had complete data across the three time points. The Cronbach's Alpha values of the 12-word items were not acceptable at pre-test (a = -.590), post-test (a = .118) and at delayed post-test (a = .094) as they were below the acceptable value (a = .70) (Field, 2024). The Cronbach's Alpha values results are shown in Table 12.

Table 12 *Cronbach Alpha of the intelligibility test based on the 12 words with the same 33 L2 learners at pre-, post-, and delayed post-tests*

Intelligibility Test	*N* of L2 Learners	*N* of Items	Alpha Value
Pre-test	33	12	-.590
Post-test	33	11*	.118
Delayed post-test	30	11**	.094

** The internal consistency was calculated on 11 items. One item (the word Jerry) was removed as there was no variance across the participants' scores.*

*** One item (the word Tog) was removed as there was no variance across the participants' scores.*

Based on these results, a recalculation of Cronbach's Alpha was applied to all available L2 learners (47 at pre-test, 47 at post-test, 36 at delayed post-test) to expand the dataset. The Cronbach's Alpha results also revealed low internal consistency within the subset of the 12-word items (see Table 19, Chapter 4 section 4.3.2). As such, the results should be interpreted with caution.

3.5.3.3 Perception test

The perception test was designed and delivered via the Gorilla software platform and was administered as a 15-minute timed test. The purpose of the test was to measure whether learners could identify the target phonemes. It involved phonemic contrasts presented through identification activities: forced-choice identification and binary (yes/no) identification trials. Learners' performance was assessed based on the total number of correct responses.

Stimuli perception test

The perception test included 160 trials and had two parts. Part I comprised of 80 two-word forced choice identification trials and Part II consisted of 80 binary (yes/no) identification trials. The identification-type tasks were chosen because they are frequently used to test learners' ability to categorize phonemic contrasts in L2 speech perception (Iverson et al., 2005).

The test comprised of four trained and four untrained stimulus words per phonemic contrast from the 10 high FL phonemic contrasts. The total stimulus words were 80 and were the same for Part I and Part II of the perception test. The perception test was the same across pre, post and post-delayed tests. Table 13 presents the 80-word stimuli of the perception test at pre, post and delayed post-tests.

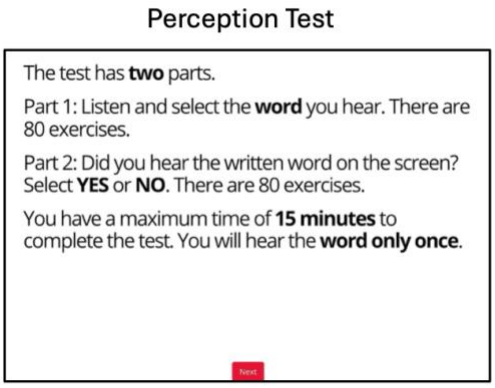
Table 13 *The 80-word stimuli of the perception test at pre-, post-, and delayed post-tests*

Perception Test		
Phonemic Pair	Trained words	Untrained Words
/ɛ -æ/	cattle - kettle access – excess (stress on 1st syllable)	dab - deb arrant - errant
/p - b/	pumpkin - bumpkin swap - swab	pillow – billow dipper - dibber
/æ-ʌ/	match - much ankle - uncle	snag – snug hatch - hutch
/m-n/	lining – liming dime - dine	pan - pam smear - sneer
/æ-ɒ/	accident - occident jobber - jabber	stamp - stomp job - jab
/l-r/	towelling - towering royal - loyal	bland - brand long - wrong
/ɔː - əʊ/	porker -poker war - woe	cork - coke potable - portable
/p - f/	coffer – copper cheep - chief	puddle - fuddle hoop - hoof
/ʌ - ɒ/	model - muddle bunny - bonny	pop – pup logging - lugging
/n-l/	line – nine teller - tenor	yearling - yearning hault - haunt

Procedure for L2 learner perception test

The perception test was the last test immediately administered after the L2 learners completed the intelligibility test. It was accessed via the Gorilla platform with an anonymous assigned public identification number.

For the perception test, L2 learners were informed after the completion of the intelligibility test that a 15-minute perception test would follow. The L2 learners remained in the Zoom platform while completing the perception test online. Once they understood the instructions, a link from the Gorilla website was sent to them to begin the test. The L2 learners logged in to the Gorilla website with their anonymous assigned public identification numbers and followed the instructions on the screen. The instructions informed them that the perception test consisted of two parts, the type of activities they would be completing and the duration of the test. Before the start of each part of the test, a sample exercise was provided. When they completed the test, the researcher confirmed that the testing was complete and ended the Zoom session. Figure 9 represents the instructions of the perception test for the L2 learners. Figure 10 and figure 11 show sample exercises before the start of Part I and Part II of the perception test for the L2 learners.



Perception Test

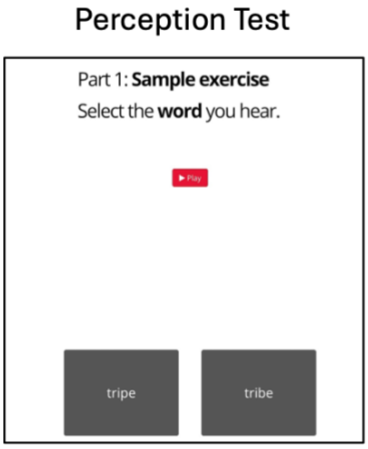
The test has **two** parts.

Part 1: Listen and select the **word** you hear. There are 80 exercises.

Part 2: Did you hear the written word on the screen? Select **YES** or **NO**. There are 80 exercises.

You have a maximum time of **15 minutes** to complete the test. You will hear the **word only once**.

Figure 9 *The instructions of the perception test for the L2 learners*

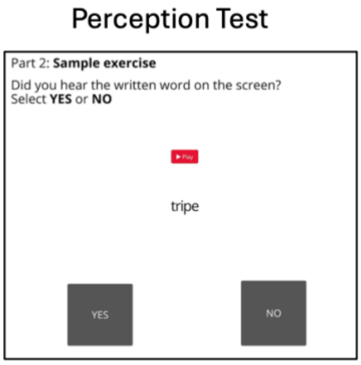


Perception Test

Part 1: **Sample exercise**

Select the **word** you hear.

Figure 10 *Sample exercise of Part I two-word forced-choice identification activity of the perception test for L2 learners*



Perception Test

Part 2: **Sample exercise**

Did you hear the written word on the screen?
Select **YES** or **NO**

tripe

Figure 11 *Sample exercise of Part II binary identification Yes or No activity of the perception test for L2 learners*

Data analysis perception test

For the perception test, the L2 learners' responses of the 160 trials were summed across items for each participant and then averaged across participants. This was done for pre, post and delayed-post perception tests. The final dataset consisted of L2 learners' overall scores across pre, post and-delayed post-tests.

Cronbach's Alpha test was used to measure the internal reliability of the 160 items in the perception test. There were 51 L2 learners at pre and post-tests and 47 L2 learners at delayed post-test for the analysis. Table 14 represents the Cronbach's Alpha results of the perception test across the three time points.

Table 14 *The Cronbach's Alpha value of the 160 items in the perception test at pre-, post-, and delayed post-tests*

Perception Test	N of L2 learners	N of items	Alpha value
Pre-test	51	160	.80
Post-test	51	160	.81
Delayed post-test	47	160	.80

The perception test showed high internal consistency across all three conditions at pre, post and delayed post-tests as they were above the acceptable Cronbach's alpha value of a=.70 (Field, 2024).

3.5.4 Unstructured interview

An unstructured interview was employed in the study immediately after the intelligibility post-test to gather learners' reflections on the online phonetic training. This component represents a novel addition to HVPT research, as few studies have combined perceptual training outcomes with learner feedback on their experience of the training process. It served as a valuable supplement to the quantitative measures used in the study as it provides further insights into the L2 learners' experiences (Isaacs & Trofimovich, 2017). Although exploratory in nature, this interview offered useful context for interpreting the training outcomes and identifying areas for improvement in future implementations of online HVPT. The interviews were conducted and video recorded on Zoom. Each video recording was downloaded to the researcher's MacBook Air laptop and converted to audio format recordings.

Procedure for L2 learners

Immediately after the intelligibility post-test was completed, the L2 learners remained on Zoom to continue with the unstructured interview. They were informed that their responses would be recorded and used to better understand their learning experience. All 51 L2 learners agreed to participate in the unstructured interview.

Data analysis unstructured interview

Once all the data were collected, the corresponding L2 learner's responses were individually transcribed. For each case at hand, thematic units which pertained to the central issues were established and subsequently analyzed. A preliminary thematic analysis was conducted following general qualitative procedures in past studies. The recommendations included the following steps:

(1) “becoming acquainted with the data by means of transcription which incorporates reading and re-reading the material; (2) generating preliminary codes throughout the data set; (3) searching for possible themes by gathering codes; (4) testing themes to verify that they are logical in regard to the codes as well as the data set; (5) ascribing meanings to themes; (6) creating the report” (Braun & Clarke , 2006: Harding, 2017),

However, as the feedback was brief and limited in scope, the data were used descriptively to contextualize learner experience rather than as a core analytical component. The interview responses thus served primarily as supplementary reflections to support the interpretation of the quantitative results.

The next section below presents the ethical guidelines that the study adhered to in conducting this research.

3.6 Ethics

This section describes the ethical procedures followed throughout the study, covering participant recruitment, consent, risk minimization, data protection, and institutional approval. The ethical approval for this study was granted by the Department of Education at the University of York. The study provided all participants including L2 learner participants and raters’ complete details regarding the study’s objectives and procedures, information regarding their participation and an informed consent was obtained in writing (see Appendix C for learner consent form; see Appendix D for rater consent form).

All learner and rater participants were over the age of 18 and were recruited via email. Those who took part in the pilot study, learners and raters were based in the UK. In the main study, learners were based in China, but raters were based in the UK. The pilot study learners and all raters (both pilot and main) received a £5 Amazon voucher as a token of appreciation. Learners in the main study, based outside the UK, were not offered monetary compensation due to feasibility limitations. Instead, the leaner participants received free access to an online phonetic training program which offered benefits for L2 learners who had no experience with such training.

Participation was entirely voluntary for all learners and raters. The recruitment materials explicitly indicated that the research study was independent of learner participants' academic performance and that withdrawal from any stage of the study without any prior reasons could be done by notifying the researcher via email. All learner participants completed the online phonetic training in private, independently, through the secure online *Gorilla* experiment platform in their own time. However, the researcher was present via Zoom throughout all assessment sessions, including those with learners and raters, and provided guidance and supported participants before they began the assessment process. This ensured consistent understanding of the procedures to be followed and limited any future lack of clarity. Although the study was considered low risk, clear communication was an important factor in supporting participant confidence.

All participants were assigned anonymous identification codes, and no personally identifying information was linked to the research data. Raters could only access anonymized audio files, and no names, images, or contextual learner data were shared. The data, which included assessments and recordings of speech samples, were stored on the researcher's password-protected computer. The data were completely anonymized and coded prior to analysis. All research data will be retained indefinitely as part of the University's Research Data Service and may be used within academic publications, with all identities remaining anonymous and protected.

The study overall made every effort to uphold best ethical principles regarding all participants, privacy, measures in securing the data, and cultural requirements of both UK-based and international participants.

Chapter 4 Results

4.1 Introduction

The aim of this study was to investigate whether improvements in the production of English phonemic contrasts in isolated words transfer to improvements in intelligibility of word production (Pruitt, et al., 2006) and to improvements in comprehensibility in spontaneous oral production among Chinese Mandarin learners of English (L2 speakers). The study aims to test whether there were significant gains from pre to post-test on measures of perception, intelligibility and comprehensibility on L2 learners' performance after the training intervention; that is, the online high variability phonetic training which is based on the high functional load theory. This chapter discusses the results of the study in relation to the following questions:

Research Question 1: What is the impact of high variability phonetic training (HVPT) on high functional load phonemes on:

a) L2 learners' perception of phonemic contrasts?

b) the intelligibility of words read aloud by the L2 learners?

c) the comprehensibility of L2 learners' spontaneous oral productions?

Research Question 2: To what extent are any gains in the above dimensions retained long term?

This chapter is divided in the following sections: perception tests results in section 4.2; the results of the intelligibility tests in section 4.3; and the results of the comprehensibility test including Task 1 and Task 2 in section 4.4.

4.2 Perception of phonemic contrasts

4.2.1 Introduction

The rationale for assessing L2 learners' perception on the perception test was to determine whether the high variability phonetic training was effective for L2 learners to better discriminate phonemic contrasts.

This section 4.2, presents results of the Cronbach's Alpha test of the perception test then discusses assumptions and descriptives, calculates the effect size, and compares mean across three time points by discussing the results of the one-way repeated measures ANOVA test. It also looks at the results of the post hoc pairwise comparisons of the tests scores across the three time points (pre, post, delayed post-test).

4.2.2 Results

Cronbach's Alpha test was used to measure the internal consistency reliability of the 160 items in the perception test. The results are presented below in Table 15.

Table 15 *The Cronbach' Alpha test for the perception test at pre-, post, and delayed post-tests*

Perception test	N of L2 Learners	N of items	Alpha Value
Pre-test	51	160	.80
Post-test	51	160	.81
Delayed post-test	51	160	.80

The results of the Cronbach test revealed that the perception test had high internal consistency of the items across all three time points as they were above the acceptable Cronbach's alpha value of α=.70 (Field, 2024).

Histograms of the L2 learners' perception test scores were plotted (see Appendix L). These revealed that all data fitted the normal distribution for all three time points (pre, post and delayed post-tests). To measure the significance of normality a one-tailed Shapiro-Wilk test was applied. This showed a significant normal distribution for all pre, post, delayed post-tests with a *p*-value (*sig*) > 0.05. The results are shown in Table 16.

Table 16 *Normality test of the perception test at pre-, post-, and delayed post-tests*

	Shapiro-Wilk test (*one-tailed*)		
Perception Test	W	df	*p* *
Pre-test	.989	51	.900
Post-test	.973	51	.294
Delayed post-test	.959	47	.101

* Shapiro-Wilk test Sig. *p* >.05 level

Since the results of the histograms revealed normal data distribution and did not violate the assumptions and the Shapiro-Wilk tests had significant values of p>0.05, it was possible to use parametric tests for the data analysis of the intelligibility test (Field, 2024).

The means scores were calculated across participants for the within group design of the perception test at pre, post and delayed post-tests. Descriptives were run separately on each time point (pre, post, delayed post-test) and compared. The results of the L2 learners' mean scores, standard deviation, standard error and confidence level are presented in Table 17.

Table 17 *Mean scores and standard deviations of the perception test at pre-, post-, and delayed post-tests*

					95% Confidence Interval	
Perception Tests	*N*	*M*	*SD*	*SE*	Lower Bound	Upper Bound
Pre-test	51	125.67	9.60	1.34	122.97	128.37
Post-test	51	130.86	9.40	1.32	128.22	133.51
Delayed post-test	47	132.02	9.13	1.33	129.34	134.70

The results show that the online phonetic training was effective for L2 learners to discriminate phonemic contrasts as there was an increase across means scores from pre to post-test and pre to delayed post-test. The small increase from post to delayed post-test in mean scores suggests that L2 learners retained the knowledge but may also have gained improvements over time.

The effect size of the perception tests was calculated with respect to pre to post-test, post to delayed post-test and pre to delayed post-test. The results are shown in Table 18.

Table 18 *The effect size of the perception test at pre-, post-, and delayed post-tests*

Perception Test	Cohen's *d*
Pre to post	.62
Post to delayed post*	.09
Pre to delayed post*	.49

**Hedge's correction used because of different sample sizes, post to delayed post-test and pre to delayed post-test calculated at 47 participants*

The study follows Plonsky and Oswald's (2014) L2 specific benchmarks because they provide more realistic thresholds for applied linguistics data than Cohen's (1988) general psychology standards. This recognises that effect sizes in L2 research tend to be smaller due to the complexity and variability of linguistic and learner factors. The effect size of the perception test (d = .62) from pre- to post-test is near a medium effect (Plonsky benchmark d = .70). This suggests that there is a noticeable difference of L2 learners' improvement between pre to post-tests, but it is a moderate one. From post to delayed post-test there is a very small effect size *d*=.09 which seems to be below the small effect size benchmark *d*=.40. This suggests there were minimal changes between the post and delayed-post-tests, therefore the L2 learners' gains had mainly retained overtime. From pre to delayed post-test the effect size is *d*=.49 indicating slightly more than a small effect size (Plonsky *d*=40). This represents an improvement in the L2 learners' performance in the perception test from pre to delayed post-test. This suggests that L2 learners had learning gains from the online phonetic training which were sustained over time.

A one-way repeated measures ANOVA test was applied to determine whether the gained differences between the perception tests at pre, post and delayed post-tests were statistically significant. The ANOVA requires homogeneity of participants across all time points (pre, post, delayed post-tests), however in this study there were 51 participants for pre and post-test but 47 for the delayed post-test. Thus, when running the ANOVA, it automatically excluded the 4 participants across all three time points (pre,post, delayed post-tests) who had missing data. To confirm whether 47 participants across the three time points was a sufficient sample to run the ANOVA or to continue with a mixed effects regression model, a G*Power 3.1 was used to calculate the power for 47 participants. The results showed that with 47 participants the study had a power of .80 which is commonly accepted as the threshold in studies (Saito, 2011; Qian et al., 2018). It was decided to continue with the ANOVA test (see Appendix M) as the power was sufficient, and it included participants with complete data which furthered the integrity of the analysis. While mixed-effects regression modelling could also have been applied to capture by-participant and by-item variability, the present analysis focused on mean differences at the group level, with scores averaged per participant across trials. ANOVA was therefore suitable for addressing the study's primary research question concerning overall changes in perception accuracy over time. Using the Greenhouse-Geisser correction, results indicated a significant overall difference between the L2 learners' performance across the three time points (pre, post, and delayed post-test) F(1.62,74.36)=7.51, *p*=.002; an overall effect size of .140 (partial η^2) showed that time accounted for 14 per cent of the variance in test scores suggesting that L2 learners' performance improved over time.

Post hoc pairwise comparisons (Bonferroni corrected) were applied to determine which specific point of time (pre, post, delayed post-test) significantly differed. The results are shown below in Table 19.

Table 19 *Post hoc pairwise comparisons of perception tests scores across the three time points*

			Mean Difference			95% Confidence Interval	
Time (1)	Time (J)	*N*	(I -J)	SE	*p**	Lower Bound	Upper Bound
Pre-test	Post-test	51	5.13	1.27	<.001	8.27	1.99
Pre-test	Delayed post-test	47	6.40	1.88	.004	11.07	1.74
Post-test	Delayed post-test	47	1.28	2.01	1.00	6.28	3.72

** mean difference significant at the .05 level*

*** p values were adjusted using the Bonferroni correction across the three pairwise comparisons (Pre-Post, Pre-Delayed, Post-Delayed), which can cause some adjusted values to round up to 1.00*

The results of the post hoc pairwise comparisons showed a significant improvement between the pre- and post-test scores (mean difference = 5.13, *p*<.001). Similarly, the pre-test to delayed post-test scores also had a significant improvement (mean difference = 6.40, *p*=.004). However, there was no significant difference between post-test to delayed post-test (mean difference = 1.28, *p*=1.00). The *p*-value of 1.00 is the Bonferroni-adjusted value: the uncorrected *p*-value was also non-significant, indicating the learners' performance remained effectively unchanged between post- and delayed post-test.

4.3 Intelligibility of pronounced words

4.3.1 Introduction

The rationale for assessing L2 learners' intelligibility in words read aloud was to test for transfer of the high variability phonetic training from perception to production in the pronunciation of read aloud words. The intelligibility test was measured at three points of data collection (pre, post, delayed post-tests). The intelligibility of L2 learners' word productions were rated by highly proficient speakers of English in a forced-choice discrimination task. In other words, the raters listened to an audio and were presented with two words on the screen, and chose which word best matched what they heard.

Section 4.3.2 presents results of the Cronbach's Alpha test for examining internal consistency of the untrained subset of the 12-word items of the intelligibility test and then presents its assumptions and descriptives. Followed by further analysis on individual word level.

4.3.2 Results

The collected data for the intelligibility test had 18 out of 51 L2 learners who had missing ratings on words either at pre-test, post-test or delayed post-test. A Cronbach Alpha test was first calculated on the same 33 L2 learners who had completed data across all three time points (see Chapter 3 section 3.5.3.2 Table 10). This revealed a low internal consistency (Cronbach's Alpha value of α=.70) (Field, 2024) within the subset of the untrained 12-word items. A second Cronbach Alpha test was recalculated on the dataset of all available L2 learners (see Table 19). The findings also indicated a lack of internal consistency within the subset of the 12-word items across all three time points. The results are presented in Table 20.

Table 20 *Cronbach 's Alpha test on the dataset of the 12-word items in the intelligibility test for all available L2 learners at pre-, post-, and delayed post-tests*

Intelligibility Test	*N* of L2 Learners	*N* of Items	Alpha Value
Pre-test	47	12	-.157
Post-test	47	12	.021
Delayed post-test	36	12	-.012

**Listwise deletion based on all variables in the procedure.*
***Intelligibility pre and post-test excluded 4 participants; intelligibility delayed post-test excluded 11 participants*

Due to the lack of internal consistency within the subset of the 12-words, the intelligibility test results should be interpreted with caution. Further to this, the time points were independently analyzed by comparing two time points (pre-to-post-tests, pre-to-delayed post-tests, and post-to-delayed post-test) using a non-parametric test, the Wilcoxon Signed-Rank test and then analyzed on an individual word-level using all available L2 learners to maximize data.

The means scores were totalled across the 12 items for each L2 learner and then aggregated across all L2 learners for the intelligibility test at pre, post and delayed post-tests. The results of the rater scores for the intelligibility tests within group showing the mean scores, standard deviation, standard error, confidence level of mean and significance level for all available L2 learners are presented in Table 21.

Table 21 *Mean, SD, SE, confidence level of mean and significance level of the intelligibility test of L2 learners' word productions at pre-, post-, and delayed post-tests for all available L2 learners within group*

					95% Confidence Interval	
Intelligibility Tests	*N*	*M*	*SD*	*SE*	Lower Bound	Upper Bound
Pre-test	47	9.77	1.146	.167	9.43	10.10
Post-test	47	9.94	1.169	.170	9.59	10.28
Delayed post-test	36	9.08	1.296	.216	8.64	9.52

**Intelligibility pre and post-test excluded 4 participants; intelligibility delayed post-test excluded 11 participants*

Whereas figure 12 represents a visual image via a bar graph the point estimates (mean scores) with 95% confidence interval of the intelligibility test at pre-, post- and delayed posts-test for all available L2 learners within group.

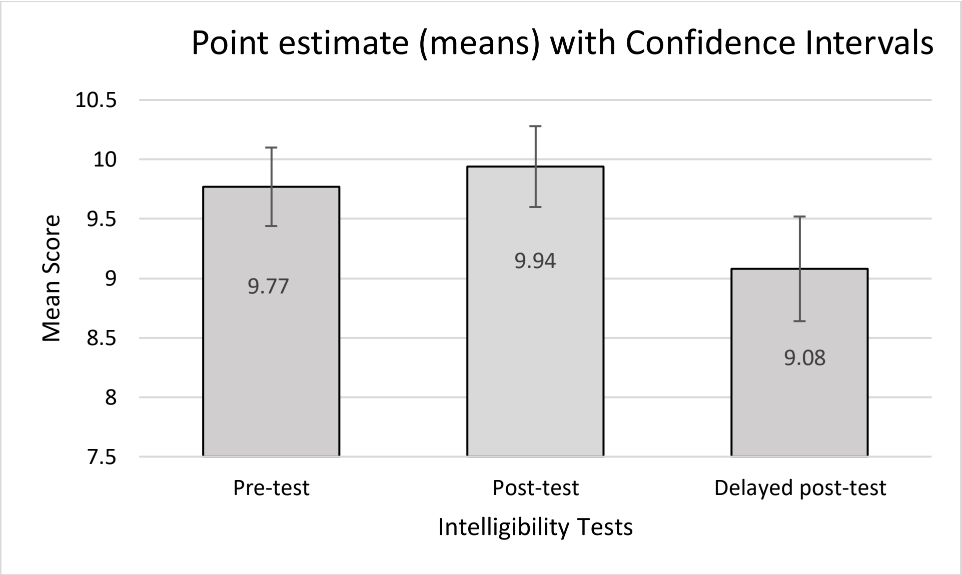


Figure 12 *Bar graph shows point estimates (mean scores) with 95% confidence interval of the intelligibility test at pre-, post- and delayed post-test for all available L2 learners within group*

A Wilcoxon Signed-Ranked test was conducted for statistical differences between time points within group. The results showed an improvement immediately after the phonetic training in group-level from pre-test to post-test but was not statistically significant (Z = -.346, *p* =. 730) (see Appendix N). However, this improvement was not retained and showed a decline in the delayed post-test (pre-test to delayed post-test Z = -2.675, p = .007).

Further statistical analysis was conducted on an individual word-level. The scores of each individual word were summed at pre, post and delayed post-tests. These scores represent the number of L2 learners that read the word aloud and scored 1 (correct) as intelligible by the listening raters. Table 22 shows the results of the rated intelligibility test on the 12 untrained words represented by percentage at pre-test, post-test and delayed post-test for all available L2 learners.

Table 22 *Results of the rated intelligibility test on the 12 untrained words represented by percentage at pre-, post-, and delayed post-test for all available L2 learners*

	Pretest		Post-test		Delayed post-test	
Untrained Word	N	Percentage	N	Percentage	N	Percentage
App /æ/	51	92	51	90	45	80
Up /ʌ/	51	90	50	86	47	66
Jelly /l/	51	78	50	86	46	80
Jerry /r/	51	96	51	100	47	96
Light /l/	48	98	51	94	46	96
Night /n/	51	92	51	94	47	85
Pat /æ/	51	49	51	47	47	45
Pet /ɛ/	51	73	51	80	47	57
robbery /ɒ/	50	82	51	86	44	82
rubbery /ʌ/	51	41	50	44	45	33
Tag /æ/	51	86	50	92	44	86
Tog /ɒ/	51	96	51	98	47	98

The following bar graph (Figure 13) demonstrates on an individual word-level the improvements of the intelligibility tests overtime across pre-test, post-tests and delayed posts-tests.

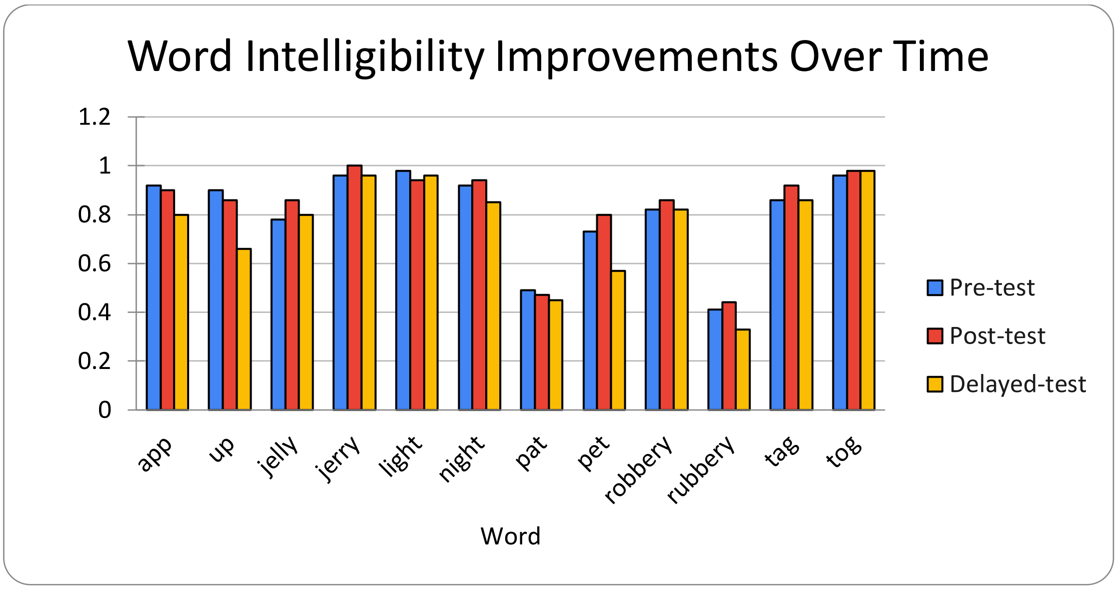


Figure 13 Bar graph shows individual word-level intelligibility improvements overtime across pre-, post- and delayed post-tests

The results showed an improvement immediately after the phonetic training on individual word-level for 8 out of the 12 words (jelly +.08, Jerry +.04, night +.02, pet +.07, robbery +.04, rubbery +.04, tag +.06, tog +.02) from pre-test to post-test. Words that retained improvement from pre-test to delayed post-test were the words jelly (+.02) and tog (+.02). While the words up (-.24), app (-.12), light (-.02) and pat (-.04) saw a decline from pre-test to delayed post-test.

The Shapiro-Wilk tests revealed that all 12-word items did not meet assumptions of normality across pre-test, post-test, and delayed post-test (p < .001), except for the word *Jerry* in the post-test which had no results due to a lack of variance. Further to this, a series of non-parametric Wilcoxon Signed Rank tests were applied to investigate differences in word-level intelligibility between the time points pre-test to post-test and pre-test to delayed post-test. The results are shown below in Table 23 and Table 24.

Table 23 *Post hoc Wilcoxon Signed Rank Test of intelligibility test scores on the subset of the untrained 12-word items at pre-test to post-test*

Word Comparison	N	Negatives	Positives	Ties	Z Value	*p* Value
app_post - app_pre	51	3	2	46	-0.447	0.655
up_post - up_pre	50	4	2	44	-0.816	0.414
jelly_post - jelly_pre	50	3	7	40	-1.265	0.206
jerry_post - jerry_pre	51	0	2	49	-1.414	0.157
light_post - light_pre	48	2	0	46	-1.414	0.157
night_post - night_pre	51	3	4	44	-0.378	0.705
pat_post - pat_pre	51	10	9	32	-0.229	0.819
pet_post - pet_pre	51	2	6	43	-1.414	0.157
robbery_post - robbery_pre	50	4	6	40	-0.632	0.527
rubbery_post - rubbery_pre	50	9	10	31	-0.229	0.819
tag_post - tag_pre	50	2	5	43	-1.134	0.257
tog_post - tog_pre	51	1	2	48	-0.577	0.564

* *Significant at the 0.05 level*

Table 24 *Post hoc Wilcoxon Signed Rank Test of intelligibility test scores on the subset of the untrained 12-word items at pre-test to delayed post-test*

Word Comparison	N	Negatives	Positives	Ties	Z Value	*p* Value
app_del - app_pre	45	7	1	37	-2.121	0.034
up_del - up_pre	47	13	2	32	-2.84	0.005
jelly_del - jelly_pre	46	5	5	36	0.0	1.0
jerry_del - jerry_pre	47	2	2	43	0.0	1.0
light_del - light_pre	44	2	0	42	-1.414	0.157
night_del - night_pre	47	6	3	38	-1.0	0.317
pat_del - pat_pre	47	6	6	35	0.0	1.0
pet_del - pet_pre	47	11	4	32	-1.807	0.071
robbery_del - robbery_pre	43	7	5	31	-0.577	0.564
rubbery_del - rubbery_pre	45	16	11	18	-0.962	0.336
tag_del - tag_pre	44	2	3	39	-0.447	0.655
tog_del - tog_pre	47	1	1	45	0.0	1.0

* *Significant at the 0.05 level*

The Wilcoxon Signed Rank test results for the intelligibility test on the 12 untrained words indicated a stable performance with no significant differences between time points for 10 out of the 12 words including 'jelly', 'Jerry', 'light', 'night', 'pat', 'pet', 'robbery', 'rubbery', 'tag', 'tog'. For the word 'app' and 'up' the results revealed a significant decline in performance scores from pre-test to delayed post-test (Z = -2.121, p = 0.034; Z = 2.84, p = 0.005).

Summary of intelligibility test results

Overall, the intelligibility test showed no significant differences from pre to post-test within group and no significant differences at the individual word level except for the word 'app' and 'up' revealing a significant decline in performance from pre to delayed post-test.

4.4 Comprehensibility of spontaneous oral productions

4.4.1 Introduction

The rationale for assessing L2 learners' spontaneous oral productions on the comprehensibility tests was to ensure that the high variability phonetic training can contribute to the enhancement of the L2 Learners' comprehensibility in spontaneous oral productions. The comprehensibility tests involved two different tasks to measure spontaneous oral productions. Task 1 a timed picture story which was a semi-controlled spontaneous speech task and Task 2 a long-turn narrative speaking which was a freer spontaneous speech task. The comprehensibility tests were measured over three points of data collection (pre, post, delayed post-tests). The L2 learners' oral productions in terms of being comprehensible were rated by highly proficient speakers of English on a scale of 0 - 5.

Before exploring and analysing data with respect to our research questions, assumptions and descriptives were discussed of the comprehensibility test. Then tests of statistical differences were applied. Task 1 timed picture story is first explored followed by Task 2 long-turn narrative speaking.

4.4.2 Comprehensibility of spontaneous oral productions of Task 1

To determine if the collected data from the comprehensibility in spontaneous oral productions of Task 1 timed picture story were normally distributed or not, histograms and the Shapiro-Wilk normality tests were applied. The results of the Shapiro-Wilk test are shown in Table 25.

Table 25 *Normality test of the comprehensibility Task 1 at pre-, post-, and delayed post-tests*

	Shapiro-Wilk test		
Comprehensibility Test	W	df	*p* *
Task 1 Timed picture story			
Pre-test	.898	51	<.001
Post-test	.897	51	<.001
Delayed post-test	.905	47	.001

* Shapiro-Wilk test Sig. >.05 level

The histograms (see Appendix O) and z-scores for skewness and kurtosis confirmed normal distribution of raters' listening scores for L2 learners' comprehensibility in spontaneous oral productions during the timed picture story task: pre-tests z = .234, post-tests z = –.807, and delayed post-tests z = –.170 (all < ±1.96). However, the Shapiro-Wilk test of normality showed that they did not meet the assumptions of normality, pre-test (W = .898, *p* < .001), post-test (W = .897, *p* <.001) delayed post-test (W = .905, *p* = .001). Therefore, it was not possible to use parametric tests to analyse the data from the comprehensibility tests and non-parametric tests were applied to analyse the data.

The means scores were calculated across participants for the comprehensibility Task 1 timed picture story at pre, post and delayed post-tests. Means, standard deviation, standard error and confidence level are reported in Table 26.

***Table 26** Mean, SD, SE, and confidence level of ratings of the comprehensibility of learners' spontaneous oral productions at pre-, post-, and delayed post-test of the timed picture story*

					95% Confidence Interval	
Comprehensibility Tests	*N*	*M*	*SD*	*SE*	Lower Bound	Upper Bound
Task 1 Timed picture story						
Pre-test	51	3.20	.917	.128	2.98	3.45
Post-test	51	3.41	.983	.138	3.14	3.69
Delayed post-test	47	3.30	.976	.142	3.01	3.58

The means and standard deviations for the comprehensibility Task 1 (timed picture story) presented in Table 14, show that L2 learners' comprehensibility ratings in spontaneous oral productions revealed an improvement after the online phonetic training, post-test *M* = 3.41 (SD = .983) compared to pre-test *M* = 3.20 (SD = .917). The delayed post-test *M* = 3.30 (SD = .976) had a higher mean score than the pre-test, which appears to suggest that L2 learners still maintained their improvements after four weeks.

A non-parametric equivalent to a one-way repeated measures design, Friedman test, was used to check for statistical differences across multiple time points (pre-test, post-test and delayed post-test) of the raters scores of the L2 learners' comprehensibility in spontaneous oral productions in the timed picture story. The scores reflected the degree to which raters understood the L2 learners' production, the higher the score, the more comprehensible the oral production was. The Friedman test indicated that there was no statistically significant difference in L2 learners' comprehensibility in spontaneous oral production across the three time points, x^2(2, N=47) = .964, *p* = .623 in the timed picture story task (see Appendix P). These inferential results support the observed trends found in the descriptive analyses.

4.4.3 Comprehensibility of spontaneous oral productions of Task 2

To determine if the collected data from the comprehensibility tests Task 2 long turn narrative speaking were normally distributed or not, histograms and the Shapiro-Wilk normality tests were applied similarly to Task 1. The results of the Shapiro-Wilk test are shown in Table 27.

Table 27 *Normality test of comprehensibility Task 2 at pre-, post-, and delayed post-tests*

	Shapiro-Wilk test		
Comprehensibility Tests	W	df	*p* *
Task 2 Long turn narrative			
Pre-test	.867	51	<.001
Post-test	.893	51	<.001
Delayed post-test	.903	47	<.001

* Shapiro-Wilk test Sig. >.05 level

The histograms (see Appendix Q) indicated a normal distribution of raters' listening scores for L2 learners' comprehensibility in spontaneous oral productions during the long-turn narrative speaking task. This was supported by z-scores within the acceptable range (±1.96), pre-test z = .492, post-test z = –.240, and delayed post-test z = .202. However, the Shapiro-Wilk test showed that they did not meet the assumptions of normality, long turn pre-test (W = .867, *p* < .001), long turn narrative post-test (w = .893 *p* <.001) long turn narrative delayed post-test (w = .903 *p* = .001). Therefore, non-parametric tests were used to analyse the data from Task 2.

The means scores were calculated across participants for the comprehensibility Task 2 long turn narrative speaking at pre, post and delayed post-tests. Descriptives were run separately on each time point (pre, post, delayed post-test) and compared as was done for Task 1 timed picture story (see section 4.4.2). Means, standard deviation, standard error and confidence level are reported are shown in Table 28.

Table 28 *Mean, SD, SE, and confidence level of ratings of the comprehensibility of learners' spontaneous oral productions at pre-, post-, and delayed post-test of the long turn narrative*

					95% Confidence Interval	
Comprehensibility Tests	*N*	*M*	*SD*	*SE*	Lower Bound	Upper Bound
Task 2 Long turn narrative						
Pre-test	51.	3.20	.872	.122	2.95	3.44
Post-test.	51	3.04	.937	.131	2.78	3.30
Delayed post-test	47	3.32	1.024	.149	3.02	3.62

The means and standard deviations for comprehensibility Task 2 (long turn narrative) shown in Table 28, show that L2 learners' comprehensibility ratings in spontaneous oral productions had a decrease in performance after the training, pre-test *M* = 3.20 (*SD* = 0.872) to post-test *M* = 3.04 (*SD* = .937). However, there was an improvement over time in the delayed post-test *M* = 3.32 (*SD* = 1.024).

The Friedman test was used to check for statistical differences across multiple time points (pre-test, post-test and delayed post-test) of the raters scores of the L2 learners' comprehensibility in spontaneous oral productions in Task 2. The higher the scores, the more comprehensible the speech production was. The results of the Friedman test revealed that there was no statistically significant difference in L2 learners' comprehensibility in spontaneous oral production across the three time points, x^2(2, N=47) = 3.036, *p* = .219 in the long turn narrative speaking task (see Appendix R) These inferential results support the observed trends found in the descriptive analyses.

Summary of comprehensibility test results

The former findings suggest that the online high variability phonetic training had a positive impact on L2 learners' comprehensibility in spontaneous oral productions in the timed picture story task (semi-controlled) with some loss of gains in the delayed post-test. Whereas the long turn narrative task (freer speech) showed gains over time but not immediately after the phonetic training. However, L2 learners' comprehensibility in speech production remained relatively constant across the three time points with no significant changes in speaking performance. Table 29 represents the key results of the present study.

4.5 Unstructured interview feedback

Although participant feedback was collected through post-test interviews, a preliminary thematic review was conducted to identify general impressions of the online phonetic training. The feedback primarily reflected surface-level observations related to learner engagement, and perceived learning gains. However, most responses were brief and limited in depth, providing little additional insight into the study’s main focus on perception, intelligibility, and comprehensibility. Consequently, the interview data were not incorporated into the main statistical findings but were considered descriptively to contextualize learner experience and inform reflections on the design and delivery of the online HVPT training.

Table 29 represents a summary of the key findings in relation to the research questions of the main study.

Table 29 *Summary of the key findings in the main study*

Research Question	Focus Area	Key Finding	Significance
1. What is the impact of HVPT on high FL phonemes on L2 learners' perception of phonemic contrasts and to what extent are any gains retained in the long term?	Phonemic identification	Learners' perception of identifying phonemic contrasts showed gains at the post and delayed post gains.	Significant
2. What is the impact of HVPT on high FL phonemes on L2 learners' the intelligibility of words read aloud and to what extent are any gains retained in the long term?	Word-level production	Learners showed immediate gains in intelligibility of word production at the post-test. No gains in delayed post-test.	Not significant
3. What is the impact of HVPT on high FL phonemes on the comprehensibility of L2 learners' spontaneous oral productions and to what extent are any gains retained in the long term?	Spontaneous speech (picture story; long-turn narrative)	Picture story (semi-controlled): Learners showed small gains in comprehensibility at the post-test. Gains partially retained at the delayed post-test. Long-turn narrative (free spontaneous): No immediate gains at post-test. Improvement retained in the long-term.	Not significant

Chapter 5 Discussion

5.1 Introduction

Chapter 5 discusses the main results of this study and links them to previous studies within the field. Section 5.2 revisits the main aims of the study, followed by section 5.3, which outlines the main results. Sections 5.4, present the results of the perception test, Section 5.5 the results of the intelligibility test, and section 5.6 the results of the comprehensibility assessment of spontaneous oral production. Finally, section 5.7 provides a summary of the discussion chapter.

Taken together, the findings suggest a possible trajectory through which high variability phonetic training (HVPT) may support improvements across different dimensions of L2 speech development. In this study, perceptual gains preceded improvements in word-level intelligibility and were followed by modest, delayed gains in listener-rated comprehensibility in spontaneous speech. While these observed patterns offer preliminary support for what this study has termed a potential *Comprehensibility Transfer Pathway*, the findings are interpreted carefully, given that multiple interacting factors may also contribute to the development of comprehensibility in spontaneous speech over time. This interconnected framework informs the interpretation of the results and connects the dimensions of perception, production, and real-world communicative effectiveness assessed in this study.

5.2 Aims of the Study

The research question that guided this study was to determine whether gains in the perception of English phonemic contrasts can be extended to gains in (a) word intelligibility (Pruitt et al., 2006) and (b) in turn to gains in the comprehensibility of spontaneous oral production. This question of transferability from perceptual training to intelligibility in isolated word production and ultimately to comprehensibility in spontaneous oral production addresses a fundamental gap in understanding how phonological skills develop across different domains of L2 speech. Although there is a number of studies that have confirmed the effectiveness of HVPT for perception and intelligibility, its application to spontaneous oral production remains poorly understood. Several studies have demonstrated that perception training improves phonemic categorization (Logan, Lively & Pisoni, 1991) and word-level intelligibility (Thomson, 2018) when minimal pairs with high functional load are applied (Munro & Derwing, 2006). Nonetheless, there has been limited empirical evidence of how these gains can be transferred to comprehensibility in spontaneous oral production, especially in freely generated oral production. The present study is one of the first to address this gap in a systematic way and will make several new contributions.

5.3 Key findings

The key findings of the study indicate that a short, intense HVPT program with multiple high functional load phonemic pairs may be useful for L2 learners. Based on research to date, this study is one of the first to experimentally investigate the transfer effects of HVPT to comprehensibility in spontaneous oral production, thus filling a significant gap in the pronunciation training research. It is also the first study to investigate the effectiveness of HVPT on a long-turn narrative task. Unlike other studies that have mostly employed controlled or semi-controlled tasks, such as picture description or spot-the-difference tasks, the current study employed a less structured and more free speech task, which provided new insights into the challenges of generalizing HVPT benefits to real-world communication.

Another key finding are the positive delayed effects observed in comprehensibility in spontaneous oral productions in the long turn narrative task which suggest that HVPT may indirectly improve the clarity of speech especially when high functional load phonemes are targeted. This builds on previous studies (Saito & Plonsky, 2019) to show that comprehensibility improvements can be achieved through perception training only, without any explicit instruction in prosody or articulation. These findings provide more precision on Saito's (2022a) position that comprehensibility is a developmental construct that is shaped by exposure and interaction. Although this study did not involve interactive practice, the results suggest that perceptual input alone can lead to internal reorganization that will later enhance comprehensibility. This is consistent with longitudinal evidence which shows that L2 perceptual sensitivity is a predictor of later production accuracy in naturalistic contexts (Flege, 1995a; Nagle, 2017).

Importantly, this study contests the idea that comprehensibility enhancement has to be achieved through output practice or explicit articulatory feedback (Derwing & Munro, 2015). Rather, it proposes that the strong perceptual representations that can be gained through HVPT can be a basis for gradual production improvements even in challenging tasks such as spontaneous oral productions. This interpretation is in line with theoretical models that consider perception as a precursor to production development (Qian et al., 2018; Baese-Berk, 2019). Furthermore, past research suggests that phonemes with high functional load impact word intelligibility (Kang & Moran, 2014; Sewell, 2017). The present study extends previous findings by illustrating that these high FL phonemes also enhance comprehensibility in spontaneous oral productions. Including these phonemes in HVPT seem to help with perception and intelligibility and clarity of spontaneous oral production over time.

More importantly, this study offers support for the *Intelligibility Principle* proposed by Levis (2005) and Foote et al. (2016) that states that the focus of pronunciation teaching should be on features that are functional in communication rather than on features that are typical of accent. In the present study, this distinction is reflected in the focus on high functional load contrasts referring to phonemic distinctions that, when confused, change word meaning and therefore disrupt communication. These phonemic contrasts can be separated from accent features (e.g. vowel quality differences or prosodic patterns) that may influence perceived nativeness but do not commonly impede understanding. By demonstrating gains in perception and modest improvements in comprehensibility in spontaneous speech for contrasts with clear communicative value, the study provides empirical support for prioritising functionally important phonemes over accent-related detail.

In addition, this study uses semi-controlled and free speech tasks to assess the impact of HVPT in a more natural way than the conventional read-aloud or sentence-repetition tasks. The findings suggest that phonetic improvement can be observed at different levels of task difficulty. While the former (semi-controlled task) may exhibit rapid development, the latter (free speech task) may require more time for noticeable change to manifest. This is important for understanding the actual communicative benefits of pronunciation training.

The present study findings are in agreement with other studies on task effects in L2 phonology (Crowther et al., 2018), which show that speech production is quite different based on task difficulty and planning time. This only goes to support the theoretical models that stress the link between cognitive processing demands and phonological accuracy (Kormos, 2006), which means that HVPT benefits may first manifest in less cognitively demanding tasks before generalizing to more complex spontaneous speech. Table 30 summarizes the gains and significance of pre-to-post and pre-to-delayed post-tests of the perception, intelligibility and comprehensibility tests.

Table 30 *Summary of pre-to-post gains and pre-to-delayed post gains of the perception, intelligibility and comprehensibility tests*

Test	**Pre to Post**		**Pre to Delayed**	
Perception	gains	Sig	gains	Sig
Intelligibility				
Within group	gains	Not Sig	No gains	Not Sig
Word level				
app	No gains	Not Sig	No gains	Sig
up	No gains	Not Sig	No gains	Sig
jelly	gains	Not Sig	gains	Not Sig
Jerry	gains	Not Sig	No gains	Not Sig
light	No gains	Not Sig	No gains	Not Sig
night	gains	Not Sig	No gains	Not Sig
pat	No gains	Not Sig	No gains	Not Sig
pet	gains	Not Sig	No gains	Not Sig
robbery	gains	Not Sig	No gains	Not Sig
rubbery	gains	Not Sig	No gains	Not Sig
tag	gains	Not Sig	gains	Not Sig
tog	gains	Not Sig	No gains	Not Sig
Comprehensibility				
Task 1 (picture story)	gains	Not Sig	gains	Not Sig
Task 2 (long turn narrative)	No gains	Not Sig	gains	Not Sig

* Sig = a statistically significant difference at 0.05

Considering the evaluation of the entire dataset collected in this study and with previous research findings, there are a number of possible explanations for these results.

5.4 Perception

This section addresses the research question (RQ 1a): What is the impact of high variability phonetic training (HVPT) on high functional load phonemes on L2 learners' perception of phonemic contrasts and (RQ 2) to what extent are any gains retained long term?

5.4.1 Overall gains in perception

The findings showed gains across all three time points (pre, post, delayed post-test) in perception. Significant gains were from pre-test to post and from pre-test to delayed post-test. The results support past findings that training with a wide range of phonetic variations enhances phonemic categorization (Bradlow et al., 1997; Lively et al., 1993; Thomson, 2018). Moreover, this observation indicates that learners enhanced phonemic awareness skills and also that these improvements were retained over time. Thus, reinforcing the idea that perceptual learning has long-term stability. Furthermore, perceptual gains were more pronounced for those contrasts that are particularly difficult for Chinese speakers of English speakers, including the phonemic pairs /æ - ʌ/, /ʌ - ɒ/, and /l - r/. However, the phonemic pair /ɛ - æ/ pair showed a slight decrease in performance at the post-test, which could be due to similar acoustic sounds found in learner's L1 phonemic categories (Qian, 2018).

Post to delayed post-test revealed a consistency of gains, although not significant, in perception scores which is in line with the idea that when a learner improves their perception, these improvements tend to be rather stable, especially when they are exposed to HVPT (Earle & Myers, 2014; Qian et al., 2018). Moreover, this insinuates that even an HVPT program of short duration can result in long-term progress of phonemic representations, which are a key precursor to a positive outcome of L2 speech production (Strange & Shafer, 2008). A possible further explanation of the gains in perception is that higher variability training produces stronger formation in phonemic categories than low variability training, thus supporting the theoretical rationale of HVPT and suggests that HVPT results in changes in perception at a basic processing level (Uchihara et al., 2024).

The findings of the present study parallel with an increasing number of studies that indicate that phonetic training, especially when given in short, focused sessions, can produce positive results. In fact, studies (Shinohara & Iverson, 2018) have shown that improvements in perception can be achieved with training duration that are comparable to the design of the present study. This challenges the notion that prolonged training periods are required for perceptual restructuring, instead stressing the importance of the quality and variability of the input rather than the length of exposure.

Overall, the enhancements in perception across pre, post and delayed post-tests, support the use of high functional load phonemic pairs in HVPT as an effective training method. This study demonstrates the effectiveness of this training approach through both significant statistical findings and meaningful effect sizes and retention of gains in L2 phonemic perception.

5.4.2 Retention of Perception

The perception test found long-term gains between post to delayed post-test which support the notion that perceptual gains remain strong especially with high-variability input (Earle & Myers, 2014; Qian et al., 2018). Although there was an increase at the delayed post-test, though not significant, it possibly indicates that perceptual learning continues to improve well beyond the training period. This observation is consistent with memory consolidation theories which state perceptual learning becomes stabilized and potentially enhanced while the learner is offline (Davis & Gaskell, 2009; Tyler et al., 2022). Moreover, these long-term perceptual gains which lasted four-weeks after the training had completed, confirm that high variability phonetic training enables learners to learn acoustic subtle differences. As a result, leading to stronger and more generalizable phonemic categories (Gabay et al., 2015). The study's HVPT training method inclusive of natural high variability stimuli helped create perceptual representations which proved long term stability because learners maintained consistent performance throughout the study.

The learners consistently maintained their learning gains between phonemic contrasts at different levels of perceptual difficulty (e.g., between the easier /l-r/ contrast and the more difficult /ɛ-æ/ distinction) which suggests that HVPT establishes fixed category representations regardless of what the original perceptual distance is between L1 and L2. The results of this study support recent modifications to the *Perceptual Assimilation Model* (PAM-L2) that perceptual training can restructure even the most difficult "single-category" assimilations by including an adequate amount of acoustic variability and attentional focus in the training (Logan et, 1991; Sakai & Moorman, 2018).

The delayed post-test took place four weeks after training without any phonetic instruction or focused practice. The sustained gains after this four-week period indicate HVPT creates enduring changes in perceptual processing which continue to exist without ongoing reinforcement. These findings support past longitudinal studies (Bradlow et al., 1999; Nishi & Kewley-Port, 2007) that have demonstrated that perceptual training effects maintain stability throughout long periods of time which indicates permanent changes to L2 phonological representations.

5.4.3 Perception and high functional load phonemes

The present study demonstrated that focusing on HVPT training with high FL phonemic pairs produced significant improvements in perceiving speech sounds. This aligns with the theoretical assumption that phonemes which are ranked higher in communicative importance tend to receive greater attentional focus. The perception gains measured for phonemic pairs with high functional load [i.e., /m-n/, /ɔ:-əʊ/, and /p-b/] with the exception of /p - f/ (see Table 44) but not attested difficulties for Chinese learners of English, confirms this. Previous research supports this finding by demonstrating that high FL contrasts receive greater attentional focus during processing while producing more consistent outcomes in phonetic learning (Munro & Derwing, 2006).

The results also demonstrate the importance of phonemic pairs that are both high FL and attested difficulties for Chinese learners of English (e.g., /æ - ʌ/, /æ - ɒ/, /ʌ - ɒ/, /l - r/, /n - l/) can be an effective method in HVPT when teaching sounds that are crucial for communication and hard to distinguish. This result is supported by the notion that to achieve optimal phonetic learning, one should focus on phonemic contrasts that are important in communication (Bundgaard-Nielsen et al., 2016).

The patterns of gains between phonemic pairs help explain the relationship between functional load and L1 interference in perceptual learning processes. The phonetic training resulted in minimal changes for the /ɛ-æ/ contrast because this phonemic distinction showed a small decrease from 53% to 52% despite receiving training while most high FL phonemic contrasts made significant progress (e.g. /ʌ-ɒ/ improved from 74% to 77%). Studies on Chinese learners of English learning English vowel contrasts provides evidence for this interpretation by demonstrating that the /ɛ-æ/ distinction persists as a difficult contrast for Chinese learners of English because it maps to one vowel category in Mandarin (Jia et al., 2006). This discovery demonstrates the effect of FL on L1-L2 phonological connections by indicating that high FL phonemes typically improve perceptual learning yet severe L1-L2 phonological similarities need specialized training methods beyond basic HVPT. Table 31 represents the phonemic pairs of high functional load and attested difficulties used in the training and their results at pre- and post-perception tests.

Table 31 *Phonemic pairs of high functional load and attested difficulties used in the training and their results at pre-, post-, and perception tests.*

High functional load phonemic pairs used in training	*Attested difficulties for Chinese learners of English**	*Pre-test perception test*	*Post-test perception test*
/m-n/		72%	*79%*
/æ-ʌ/	/æ-ʌ/	*79%*	*82%*
/æ-ɒ/	/æ-ɒ/	93%	*96%*
/ɛ -æ/	/ɛ -æ/	*53%*	*52%*
/ɔ: - əʊ/		*62%*	*69%*
/ʌ - ɒ/	/ʌ - ɒ/	*74%*	*77%*
/l-r/	/l-r/	*90%*	*94%*
/n-l/	/n-l/	*87%*	*91%*
/p-b/		*78%*	*83%*
/p-f/		*96%*	*96%*

* Swan and Smith, 2001; Jia et al., 2006; Nilsen and Nilsen, 2010; Barriuso & Haves-Harb, 2018.

Overall, the findings show that using high FL phonemic pairs in HVPT resulted in significant perceptual gains (RQ 1a), with improvements largely maintained over time (RQ 2). These findings provide further evidence in understanding the relationship between functional load and perceptual learning, which may suggest that phonetic training should aim in prioritizing phonemic pairs that have communicative value but to also consider the L1 phonological systems of learners.

5.5 Intelligibility

This section addresses the research question (RQ 1b): What is the impact of high variability phonetic training (HVPT) on high functional load phonemes on the intelligibility of words read aloud by the L2 learners and (RQ 2) to what extent are any gains retained long term?

5.5.1 Overall gains in intelligibility

The post-test results showed small improvements in words such as pet, robbery and pen. The results support the idea that HVPT can increase the intelligibility of word items by making certain phonemic contrasts easier to perceive (Thomson, 2018; Uchihara et al., 2024).

The training had the most impact on phonemes which presented challenges for FL and L1 transfer such as /æ - ʌ/, /ʌ - ɒ/. A minimal pair example is ‘robbery – rubbery’ where ‘robbery’ showed an increase (pre- 82%, post- 85%, delayed post- 82%) whereas ‘rubbery’ showed a decrease (pre- 41%, post- 44%, delayed post- 33%) suggesting that some contrasts may be less amenable to change due to vowel similarity or difficulties with the articulation of the vowels (Aliaga-Garcia & Mora, 2009; Thomson et al, 2009).

The different rates of improvement between word items indicate complex processes of perception and production in L2 phonological acquisition. The improved production of words with phonemic pairs such as /ɛ - æ/ (pet – pat) and /ɒ - ʌ/ (robbery - rubbery) shows that perceptual training can improve production accuracy by improving phonological representations. The results also support the idea that well developed perceptual categories can help with hearing and distinguishing speech sounds and thus support articulatory movements by linking sensory input to motor output (Jia et al., 2006; Fouz-Gonzalez, 2020). Nonetheless, the different improvements across words reflect the complex dynamics of the perception production relationship. The research suggests that articulatory and lexical frequency represent factors which affect the transition between perception and production (Thomson & Derwing, 2016; Nagle & Baese-Berk, 2022). The results of the current study are in line with this perspective, suggesting that while perceptual training is likely to improve intelligibility, the extent of improvement will be dependent on certain phonological and lexical factors.

The HVPT post gains is consistent with studies which found that improvement in discrimination of L2 phonemic contrasts is associated with improved articulatory movements in production (Lambacher et al., 2005). This means that HVPT may have an impact on speech production systems by improving the target sounds for articulation. Interestingly, the present study challenges the notion that enhancements in speech production require articulatory practice as the present study did not include any explicit instruction in pronunciation. The study's result is supported with recent findings which state that enhanced perceptual representations on their own can increase production accuracy through the automatic adjustment of the continuous feedback cycle: perception – production (Franken et al., 2017).

Moreover, the intelligibility test which was rated using a two-word forced-choice identification task made it possible to use objective scoring of learners' intelligibility production. Though, limitations were put forth due to the high volume of collected data from L2 learners' audio words, and the final rated words were reduced in items (from 80 to 12 words for each learner at pre, post and delayed post-tests) due to concerns of time management, feasibility and rater fatigue. Although this made the task easier to handle, there was no internal reliability among the word items.

5.5.2 Retention of intelligibility

The delayed post-test results for intelligibility presented inconsistent findings which suggested problems with long-term retention. The results suggest that speech production does not persist in the long-term unless additional input is provided (Iverson & Evans, 2009; Uchihara et al., 2024). One such example are the words 'up', 'app', and 'rubbery' which showed decreased intelligibility during the delayed-post-test. The varying intelligibility scores at delayed post-test is further examined in how specific sounds within each word impacted the HVPT long-term gains.

The research results showed consonant contrasts maintained better long-term stability than vowel contrasts. The vowel phonemic contrasts /æ - ʌ/ and /ʌ - ɒ/ showed performance decline in the delayed post-test. The results demonstrate that Chinese English learners possibly face difficulties with vowel contrasts because their L1 lacks corresponding phonemic distinctions (Thomson et al., 2009; Qian, 2018).

The retention results support the *usage-based* theoretical framework of second language phonemic development (Bybee, 2008; Tyler, 2019) which states that word content and frequency determine phoneme accuracy in production. The results may suggest that particular words which appeared frequently in learners' everyday communication maintained their phonemic accuracy better than words that received less usage. The results can also be explained by the “overgeneralization effect” (Bradlow et al., 1999: Baese-Ber & Samuel, 2016) where at the beginning of learning, learners apply phonemic contrasts without discrimination before learning to use them correctly in specific contexts. The initial gains observed at post-test may have been overgeneralized and learners had not yet incorporated the distinctions into their speech.

5.5.3 Intelligibility and high functional load phonemes

The results show not all high functional load phonemic pairs produced improvement, suggesting training only on high functional load may not be enough for effective training, and potentially, other elements other than phonemic contrasts may have an influential role on learner attention and on training results.

These results suggest that high functional load phonemic contrasts were perceived as challenges for the learners. These results indicate that functional load may become insufficient to handle interference problems in perception and articulation when there are significant differences in the phonological system between learners’ first language and second language. The results also showed that high variability phonetic training on high FL phonemic contrasts may also be influenced by other variables such as individual differences for e.g. perceptual sensitivity (Thomson, 2018); lexical properties for e.g. familiarity (Wong, 2014; Cheng et al., 2019) and the phonetic context for e.g. sound position (Li, 2015) that may affect the outcome of the training. This means that the training intervention results go beyond the high functional load of the targeted contrasts and are based on many variables (Cheng et al., 2019).

The analysis shows that high variability phonetic training may be most effective when it focuses on high functional load phonemes but should also consider L2 learners' language background including first language transfer effects. The results demonstrate that HVPT training on high functional load phonemic contrasts can support improvements in individual word-level intelligibility (RQ 1b). However, the extent and consistency of these gains may vary across L2 learners and items (RQ 2), which may relate to first language effects, lexical-specific characteristics and phonetic contexts.

5.6 Comprehensibility in spontaneous oral production

This section addresses the research question (RQ 1c): What is the impact of high variability phonetic training (HVPT) on high functional load phonemes on the comprehensibility of L2 learners' spontaneous oral productions and (RQ 2) to what extent are any gains retained long term?

5.6.1 Overall gains in comprehensibility

The immediate post-test results did not reveal any statistically significant improvements in comprehensibility performance for both assessment tasks. Interestingly, Task 2 in the long-turn narrative task showed slight improvements at the delayed post-test.

Task 1 (timed picture story) was a semi-controlled speech task with a set of six pictures. The mean score in the timed picture story task (Task 1) increased between pre-test and post-test but with no significant gains. However, these increases demonstrate that phonemic knowledge can be applied towards planned oral production. The second task (Task 2 long turn narrative) demonstrated different results because it demanded unstructured and extensive oral production with no visual support. The results showed a decrease immediately after the training intervention before showing an improvement at delayed post-test. These results match the *Cognitive Load Theory* (Sweller, 1988; Kormos, 2006) which predicts that performance might decrease briefly while learners become more automatic before reaching stability. The observed delay is consistent with the proceduralization model of L2 speech development (DeKeyser, 2007; DeKeyser & Criado, 2013) that shows perceptual gains will affect spontaneous oral production after multiple instances of use and implicit repetition. However, it should be noted that the variations in learners' performance may also indicate statistical regression to the mean which shows natural score fluctuations during retesting instead of actual progress in learning.

The results from Task 1 (picture story) and Task 2 (long turn narrative) show different transfers of HVPT to different communicative contexts. The observation of post-test gains in Task 1 shows that pronunciation accuracy is influenced by cognitive workload and pre-task preparation (Crowther et al., 2015). Whereas the gains in the delayed effect in Task 2 reveal that learners need more mental effort during speech production which makes them pay less attention to phonological forms, thus, learners need more time in freer speech to show improvements in comprehensibility in spontaneous oral production (Guion & Pederson, 2007). The research by Lin (2001, 2003) on Chinese L2 speakers of English shows that learners' pronunciation improves in spontaneous oral production when they shift their attention from task preparation to phoneme accuracy.

A final observation is that learners may experience a period of "destabilization" (when learners restructure their sound categories to accept new perceived sounds that leads to performance disruption before stable integration happens (Gass & Selinker, 2008). This may interpret the results of Task 2 where there was a decline in the post-test and an increase in the delayed post-test.

5.6.2 Retention in comprehensibility

The comprehensibility retention patterns demonstrated distinct variations between the two assessment tasks. The results from Task 1 (timed picture story) indicated minimal post-test gains which partially declined at the delayed post-test but stayed higher than baseline measurements while Task 2 (long-turn narrative) presented delayed improvement at the four-week post-training assessment. The various results demonstrate how skill retention in L2 phonology behaves differently between semi-controlled and freer spontaneous speech contexts.

The partial decline in Task 1 comprehensibility scores from post-test to delayed post-test matches results from Saito and Hanzawa (2018) and Pattamadilok et al. (2022) which indicate that controlled speech improvements depend on explicit monitoring processes which fade away when students do not receive reinforcement. The delayed improvement in Task 2 demonstrates a distinct process of phonological development for spontaneous speech because it depends less on explicit monitoring and more on automatic integration of perceptual knowledge into speech processing (Segalowitz, 2010).

The current research by Saito et al. (2020) and Nagle (2017) also shows delayed effects in L2 pronunciation development especially for features that need major changes in existing phonological categories. Research indicates that spontaneous speech pronunciation improvements develop through a non-linear pattern which requires internal reorganization before noticeable changes appear. The research findings show that HVPT specifically causes delayed development patterns because it sets off perceptual reorganization processes that affect production outcomes.

The quick gains in intelligibility differ from the delayed effect of comprehensibility which matches the developmental timelines described by Tyler (2019) and Darcy and Mora (2014) that phonemic contrasts at word-level reveal quicker gains than applying phonemic contrasts into connected speech which reveal gradual gains but are more dependent on automatic processing skills and the need to communicate effectively.

Moreover, this delayed improvement agrees with the idea that learners need some time to incorporate the perceived gains in their spontaneous oral production. This finding is in line with the Skill Acquisition Theory (DeKeyser, 2007) which states that the incorporation of perceptual knowledge into automatic use takes time and practice. Unlike explicit pronunciation instruction, which may produce immediate but shallow effects, HVPT seems to foster deeper phonological restructuring that supports long-term development

5.6.3 Comprehensibility and High Functional Load Phonemes

The patterns of comprehensibility development especially the delayed improvements in spontaneous oral productions of the long turn narrative (Task 2) need to be considered in relation to the specific design of the HVPT training which targeted ten high functional load (HFL) phonemic pairs: /m-n/, /æ-ʌ/, /æ-ɒ/, /ɛ-æ/, /ɔː-əʊ/, /ʌ-ɒ/, /l-r/, /n-l/, /p-b/, and /p-f/. These contrasts were selected based on their high lexical contrastiveness in English, and six of them (/æ-ʌ/, /æ-ɒ/, /ɛ-æ/, /ʌ-ɒ/, /l-r/, and /n-l/) have been consistently documented as attested difficulties for Chinese learners of English. The inclusion of a limited yet high impact set of contrasts allows for a closer examination of how these phonemic pairs influence comprehensibility in spontaneous oral productions.

The results from the long turn narrative task 2 show that even a short intensive HVPT intervention can be beneficial to spontaneous oral production development when paired with perceptual training. One interpretation of the results is that HVPT enhanced learners' phonological awareness which enabled them to refine their output production in the long term. Although, task 2 had no immediate improvements in production accuracy, becoming more aware of key contrasts gradually may have refined learners' speech, particularly in tasks where attentional demands are high, and feedback is implicit (Lengeris, 2018).

The observed improvements could stem from cross-task transfer because learners generalized phonological features. The training focused on ten targeted high functional load phonemic contrasts, yet learners began to apply their enhanced perceptual awareness to untrained lexical items and related phonemic categories (Carlet & Cebrian, 2015). One such instance is the improved vowel distinction perception between /ʌ-ɒ/ that may have helped learners articulate more English words correctly. The generalization effect is important particularly when the trained phonemic contrasts are both high functional load and attested difficulties for learners, since these contrasts most possibly impede learners' comprehensibility in spontaneous oral productions (Qian, 2018).

The results show that HVPT training focusing on specific high functional load phonemic contrasts, including attested difficult ones, may possibly lead to noticeable gains in comprehensibility in spontaneous oral production for L2 learners (RQ 1c), with these gains observed at the delayed post-test for the free speech task rather than immediately (RQ 2). The findings also indicate that these effects developed gradually, suggesting the need for perceptual training that supports phonological restructuring and provides cognitive reinforcement across different time points (pre, post and delayed post-test) and across different communicative task types. While these results are consistent with the idea of a *Comprehensibility Transfer Pathway* (the gradual influence of perceptual gains on spontaneous speech) they should be interpreted with caution. The observed improvements were modest and appeared only in the delayed post-test, suggesting that HVPT may be one of several interacting influences rather than the sole cause of change. Other contributing factors could include learners' ongoing exposure to English, instructional context, motivation, or individual variation in phonological awareness and cognitive processing. It is also plausible that broader speech variables such as rate, pausing, or prosodic phrasing affected listeners' judgments of comprehensibility. Future work could examine these dimensions in more detail to clarify the mechanisms underlying delayed improvement patterns.

The present study should acknowledge every potential reason that could explain the delayed improvements in comprehensibility in spontaneous oral productions. All 51 L2 learners received the same HVPT intervention through online delivery yet their learning environments remained distinct. The study included 37 Chinese participants who studied English with a British-instructor at a Chinese university while fourteen participants studied online English with a British-instructor at a UK institution with eleven students in China and one in Taiwan and two in the UK. The observed gains in the delayed effect in the long turn narrative task 2 might have received support from additional English input exposure that learners experienced through their educational environments. The present study's single-arm design without a control group prevents the researcher from establishing HVPT as the sole cause of comprehensibility improvements.

5.7 Chapter summary

The research in this chapter investigates how high variability phonetic training (HVPT) affects high functional load phonemes in three areas: perception, intelligibility and comprehensibility in spontaneous oral production. The results showed that phonemic perception improved significantly, and these gains were maintained at the delayed post-test while word-level intelligibility received selective improvements. The results also showed that comprehensibility in spontaneous speech improved later on especially in the long-turn narrative task. The results demonstrate the intricate time patterns of how perceptual training affects L2 speech development by showing that HVPT starts a cascading effect which leads to the progressive transfer of improved phonemic representations to production.

Chapter 6 Conclusion

6.1 Introduction

This chapter brings together the aims and research questions of the study (section 6.2), followed by a summary of the study (section 6.3) and of the key findings (section 6.4). The significance of the study (section 6.5), after which the main limitations are outlined (section 6.6). The chapter concludes with the theoretical, pedagogical, and methodological implications arising from the research, and a final summary of the chapter (section 6.7).

Building on the results presented earlier, this chapter also reflects on the broader pattern that emerged across perception, word-level intelligibility, and comprehensibility in spontaneous speech. The findings suggest a possible developmental pathway referred to here as the *Comprehensibility Transfer Pathway* through which gains in phonemic perception may support later improvements in comprehensibility in spontaneous production. The conceptual model helps frame the interpretation of the study as a whole and offers a way of understanding how perceptual training may relate to changes in real-world communicative performance.

6.2 Aims and research questions

This study investigated the impact of high variability phonetic training (HVPT) focused on high functional load phonemes on perception, intelligibility, and comprehensibility in spontaneous oral productions of L2 speech The overarching research question explored whether improvements in the perception of English phonemic contrasts would transfer to gains in the intelligibility of word production and, in turn, to greater comprehensibility in spontaneous oral production. A secondary aim was to investigate the extent to which these gains were still present in a delayed post-test administered four weeks after the training.

To achieve the aims, the following questions were proposed:

Research Question 1: What is the impact of high variability phonetic training (HVPT) on high functional load phonemes on:

a) L2 learners’ perception of phonemic contrasts?

b) the intelligibility of words read aloud by the L2 learners?

c) the comprehensibility of L2 learners' spontaneous oral productions?

Research Question 2: To what extent are any gains in the above dimensions retained long term?

These questions were investigated through a longitudinal design that assessed outcomes immediately after training and again at a four-week delayed post-test, allowing for examination of post effects and retention patterns.

6.3 Summary of the study

The research design used pre-, post-and delayed post-test measures with 51 Chinese learners of English. The study involved 15 lessons of HVPT instruction which focused on 10 high functional load phonemic contrasts that included six phonemic contrasts of attested difficulties known for Chinese learners of English. The online training program allowed participants to learn at their own pace in which L2 learners completed the training on average in three weeks.

The study evaluated perception through two identification tasks that included multiple minimal pairs which targeted the phonemic contrasts from the training. The evaluation of intelligibility used a word read aloud assessment which required proficient English speakers to identify words produced by L2 learners in minimal pair forced-choice tasks. The assessment of comprehensibility used two spontaneous speech tasks including a semi-controlled timed picture story task (Task 1) and a freer, long-turn narrative task (Task 2).

The comprehensibility of spontaneous oral production was rated by proficient English speakers who used a 0 – 5-point rating scale. The study used multiple assessment methods to measure perception and intelligibility and comprehensibility across three time points which gave a complete understanding of HVPT effects on L2 speech development from immediate post-training to long-term retention. The research included both controlled and unstructured speech evaluations to analyze training outcomes under various task requirements and processing limitations.

6.4 Summary of key findings

The following is a summary of the key findings in relation to the research questions.

RQ1a: What is the impact of high variability phonetic training on high functional load phonemes on L2 learners' perception of phonemic contrasts and (RQ 2) to what extent are any gains retained in the long term?

The study revealed statistically significant gains in learners' perception of high functional load phonemic contrasts from pre-test to post-test and further gains at the delayed post-test. This indicated that L2 learners retained in the long-term their improved phonemic discrimination skills over the four-week period after training. Furthermore, the phonemic contrasts revealed that most high functional load pairs showed improvement, including gains in phonemes that are considered attested difficulties for Chinese Mandarin speakers, such as /æ-ʌ/ , /ʌ-ɒ/ , and /l-r/ . Only the /ɛ-æ/ contrast showed a slight decline (53% to 52%), possibly due to phonetic similarity and the absence of an equivalent sound in Chinese Mandarin vowels.

Overall, these findings demonstrate that HVPT effectively enhances L2 learners' ability to discriminate between phonemic contrasts, particularly those with high functional load. The stability of these perceptual gains in the long-term propose that the training induced robust changes in learners' phonological representations. Thus, developing a fundamental base for prospective enhancements in oral production.

RQ1b: What is the impact of high variability phonetic training (HVPT) on high functional load phonemes on the intelligibility of words read aloud by the L2 learners and (RQ 2) to what extent are any gains retained in the long term?

The post-test intelligibility results demonstrated small but quantifiable improvements in certain words, yet these changes did not reach statistical significance when analyzing within-group data. The delayed post-test results showed variability where some words maintained their gains while other words demonstrated decreased performance.

The variability in intelligibility gains indicate that perceptual training affects production accuracy, but this transfer depends on phonological complexity and articulatory demands and L1 interference. The phonemic vowel contrasts that Chinese Mandarin speakers find difficult to distinguish (e.g., /ʌ/ in up and rubbery) demonstrated reduced improvement and instability compared to easier contrasts.

The intelligibility results demonstrate that the connection between perception and production in L2 phonological development remains complex. The results show that HVPT strengthens perceptual representations but the improvement in production accuracy depends on the particular phonological features being targeted.

RQ1c: What is the impact of high variability phonetic training (HVPT) on high functional load phonemes on the comprehensibility of L2 learners' spontaneous oral productions and (RQ 2) to what extent are any gains retained in the long term?

The results of the comprehensibility assessment demonstrated complex patterns of development between the two speech tasks. The results from Task 1 (timed picture story) indicated a minimal gain between pre-test and post-test, but this change was not statistically significant. Task 2 (long-turn narrative) demonstrated a small decrease from pre-test to post-test, before showing improvement at delayed post-test.

The pattern of Task 2 shows a decline in performance after post-test followed by improvement at delayed post-test which indicates that perceptual learning needs time to transfer to spontaneous oral production. The initial decrease in performance might stem from increased cognitive demand that L2 learners needed to monitor their pronunciation quality during the challenging narrative task. The results from delayed post-test showed that perceptual gains successfully transferred to spontaneous oral production after learners integrated and proceduralized their new phonological representations.

The different developmental patterns between tasks demonstrate how task requirements affect L2 oral production. The semi-controlled Task 1 provided L2 learners with more opportunities to monitor their phonology which led to faster improvements. The free speech nature of Task 2 demanded more cognitive resources for content planning which delayed the appearance of phonological improvements until these enhanced perceptual representations became automatically accessible.

Overall transfer effects

The study shows that transfer occurs between perception – intelligibility - comprehensibility but it happens through a complex time-dependent process instead of a linear trajectory. The results show that perceptual improvements appeared immediately and retained long term but intelligibility in word production showed variable selective improvements which did not always persist; and comprehensibility in spontaneous oral production needed time to develop especially in less structured situations.

The temporal sequence indicates that HVPT starts a developmental process which allows improved phonemic representations to influence production in progressively complex communicative contexts. The time needed for internal reorganization and integration of enhanced phonological knowledge into the L2 learner's production system explains the observed delay in comprehensibility improvements in spontaneous oral production particularly in the long turn narrative.

6.5 Significance

The research results deliver important insights for second language pronunciation research. The research reveals that high variability phonetic training with high functional load phonemic pairs leads to perceptual advantages which then impact both intelligibility and comprehensibility in spontaneous oral production. The research also shows that the training gains are retained in the long term and proves the effectiveness of delayed post-testing as a method to evaluate training durability. Further to this, independent online learning with HVPT has real world applicability of perception-based instruction.

6.6 Limitations

Upon completing the study, the following limitations are acknowledged. A primary limitation was that the initial intelligibility test contained 80 words but only 12 words were chosen for rating due to the large number of collected stimuli (11,920 words). As a result, the reduction in the number of words might have reduced the scale's reliability. The selection of phonemic contrasts for testing was based on attested difficulties and high functional load, but the limited set of 12 words was not able to completely cover the learners' phonetic range. A secondary limitation was that these selected 12-word items showed low internal consistency according to Cronbach's alpha which restricted longitudinal comparisons. The low internal consistency indicate that rater perception, word difficulty, and the binary rating system may have contributed to the observed variability. A future recommendation is to apply a Likert-style scale which may provide better measurement of small changes in intelligibility. A third limitation relates to the intelligibility test and the statistical analysis used. The study addressed missing data for the intelligibility test through listwise deletion to maintain a balanced dataset; however, future work could explore alternative treatments such as imputation or mixed-effects regression modelling to retain a larger sample. It also employed repeated-measures ANOVA to examine changes across the three testing times. While this approach was appropriate for analysing group-level trends and balanced data, it does not account for potential random variability associated with individual learners or items. Mixed-effects regression modelling could offer a more flexible framework for future research by incorporating such variability and capturing more fine-grained patterns across participants and test items. Similar considerations apply to the perception test, where mean scores were analysed at the group level using ANOVA. Although this approach was appropriate for the study’s aims, future studies could adopt mixed-effects regression to model by-participant and by-item variability in greater detail.

A fourth limitation is the lack of prosody instruction which may explain why comprehensibility transfer remained limited. The training focused on segmental features, yet comprehensibility can also possibly be affected by suprasegmental elements (Derwing et al., 1998; Saito, 2021). A fourth limitation is that the sample consisted of Chinese speakers of English, and findings may not generalize to learners with other L1 backgrounds. Further cross-linguistic studies are needed to explore whether HVPT effects vary across L1 groups with different phonological profiles. A fifth limitation is that the study employed a longitudinal design with a four-week delayed post-test, but even longer follow-up periods might reveal different retention patterns in tracking the long-term stability of HVPT effects across perception, intelligibility, and comprehensibility dimensions (Thomson & Derwing, 2015)

A sixth limitation is that the pre-test perception scores revealed a wide range of perceptual abilities (64% to 92.5%), but the analysis focused on overall within-group gains. As a result, the findings do not account for how learner-internal variables may mediate responsiveness to HVPT. A following limitation is that the assessment of intelligibility and comprehensibility in spontaneous oral productions in the current study depended on human raters. This dependency may contribute to variability in listening judgments. The comprehensibility test contained two tasks where raters evaluated different sets of 10 learners for each task. The observed differences between Task 1 and Task 2 results may have been due to rater differences instead of actual performance variations. A possible suggestion to have eliminated rater variability and have yielded stronger findings would have been to have used the same rater to evaluate the same learners for both tasks. A final limitation is that the study did not include a control group that received the HVPT intervention. Only an experimental group was used, which limits the ability to determine whether the observed gains were solely attributable to the training itself or potentially influenced by external factors, such as the instructional context in which the learners were situated.

Despite the former limitations, the study delivers important findings about how HVPT could enhance L2 phonological development.

6.6 Implications

This study offers a range of implications across theoretical, pedagogical, and methodological domains. The following subsections outlined below are discussed in relation to the how the findings contribute to second language pronunciation training: section 6.6.1 theoretical implications, section 6.6.2 pedagogical implications, and section 6.6.3 methodological implications.

6.6.1 Theoretical implications

The research at the theoretical level advances current models of second language phonological development by questioning basic perception-production models and providing fresh insights into developmental processes.

The cascading pattern of improvement across perception, word-level intelligibility, and comprehensibility in spontaneous oral production challenges traditional linear models of the perception–production relationship. This progression supports what this study puts forth as the *Comprehensibility Transfer Pathway*, a developmental trajectory in which perceptual training leads to intelligibility gains, which then lead to improvements in comprehensibility in spontaneous oral production. The concept provides a refined method to understand how perceptual learning affects real-world L2 communication success.

The results show that the process is more complex and time-dependent than direct and immediate transfer because it is influenced by task demands, cognitive processing and the structural complexity of speech. These findings support multidimensional models of L2 phonological development (Flege, 1995a; Huensch & Tremblay, 2015; Flege & Bohn, 2021; Melnik-Leroy et al., 2022) which recognize that perception and production follow distinct but interrelated developmental trajectories.

The result of the delayed effect of the comprehensibility test in the long turn task provides empirical support for what has been referred to as the "incubation period" (Saito & Plonsky, 2019), in phonological learning. It proposes that learners may be processing and assimilating new phonemic distinctions internally even after the training is complete and are gradually only seen in the production of speech.

Phonemic contrasts with different high functional loads show selective improvements which support theoretical models that explain phonological development through sound communicative value (Catford, 1987; Jenkins, 2003). The continued challenges with /ɛ-æ/ contrasts demonstrate that L1-specific obstacles significantly affect the learning process. The research demonstrates the necessity for L2 phonological acquisition models to consider both functional load and cross-linguistic transfer.

The results from semi-controlled and free speech tasks demonstrate how task characteristics affect speech performance. Research supports models which demonstrate that planning time (Foster & Skehan, 1996; Yuan & Ellis, 2003) and cognitive load (Levelt, 1989, 1999; Derwing et al., 1998) and communicative pressure (Schmidt, 1992, 2001) determine the accessibility of phonological knowledge (Segalowitz, 2010). A task-sensitive approach is necessary to understand L2 speech development when analyzing pronunciation results.

The delayed post-test results show a U-shaped pattern because learners first experienced a decline before achieving improvement. The observed pattern contradicts traditional models which predict continuous progress while supporting dynamic systems approaches (Gass & Selinker, 2008; Larsen-Freeman & Cameron, 2008; Qian, 2018) that recognize regression and destabilization and reorganization as typical learning processes.

6.6.2 Pedagogical implications

The research results provide multiple practical applications for pronunciation instruction in L2 learning environments.

The substantial perceptual gains from a short training period show that HVPT can be used effectively within standard classroom time limits. Language programs should use short phonetic training modules as efficient supplements to their regular communication-focused instruction.

The observed improvements in high functional load contrasts validate educational methods which first teach phonemes that are most important for communication. Educational professionals should direct their efforts toward phonemic distinctions which create multiple phonemic contrasts and are also attested difficulties for students from certain L1 backgrounds.

The successful online implementation of HVPT demonstrates that technology-based pronunciation training methods can be effective. The approach provides flexible learning opportunities for students and enables phonetic training to be part of a hybrid learning approach which expands educational accessibility.

The results showing a delayed effect of transferability to spontaneous oral production, demonstrate the need to establish realistic expectations about pronunciation development. Teachers along with learners need to recognize that perceptual enhancements in speech do not automatically appear in spontaneous speech but develop through time as phonological representations become automatic speech processes.

The moderate comprehensibility gains from HVPT indicate that this method works best when used alongside other teaching methods. The combination of perceptual training with explicit articulation guidance and prosodic instruction and communicative practice will improve spontaneous speech outcomes.

The late development of comprehensibility gains demonstrates why longitudinal assessment plays an essential role in pronunciation teaching methods. Educators should use delayed assessments instead of post-tests to evaluate training effectiveness because these tests reveal the complete path of phonological development.

Beyond these classroom applications, HVPT can be applied in a range of teaching contexts and complements communicative language teaching by strengthening the perceptual foundation learners need before engaging in spoken interaction. Although the training in this study focused on perception rather than production, it can support communication-based lessons by helping learners recognise key phonemic contrasts that often lead to misunderstandings in real-world speech. Short HVPT modules could be used before oral tasks in a communicative syllabus or combined with pronunciation and speaking activities in hybrid or online courses. HVPT can also be integrated into digital pronunciation tools or apps that use artificial intelligence for speech analysis and feedback. Embedding HVPT within AI-assisted learning platforms would allow perception training to work alongside automatic speech recognition, creating a research-informed model of pronunciation learning that connects perception, awareness, and production through gradual, task-based progression.

6.6.3 Methodological implications

The research demonstrates how high functional load principles can be used to enhance HVPT design through methodological approaches.

Phonemic pairs with high functional load enable HVPT to adopt a more meaningful communicative approach. Research studies prior to this work focused on difficult contrasts without evaluating their communicative importance (Bradlow et al., 1997; Logan et al., 1991) but this study demonstrates that selecting contrasts based on functional load leads to better perceptual and intelligibility results. A key design strength of the present study is its integration of functional load as a principled basis for training design. The study used functional load as a key variable to select phonemic contrasts which ensured that HVPT focused on features with the most communicative value. Further to this, the current study differs from previous research by training 10 phonemic contrasts with high functional load instead of only 1–3. The expanded design enhances ecological validity because it mirrors the various phonological obstacles that students encounter during actual communication.

Moreover, the training model connects theoretical principles to learner-specific needs through its alignment of 10 high functional load contrasts in which six of the ten are attested difficulties for Chinese learners of English. The innovative methodological approach enables researchers to create HVPT programs which combine linguistic principles with learner-specific needs. A last methodological implication is that the online self-paced training system advances digital language learning tools while showing that effective phonetic instruction can be delivered outside laboratory settings. The method enables flexible pronunciation training which can be adapted to different learning environments.

6.7 Chapter summary

The present study yielded research-based findings that phonemic perception development leads to better intelligibility in word production before leading to comprehensibility in spontaneous oral production. The study demonstrates that transfer occurs through complex time-dependent mechanisms which impact differently speech dimensions and task categories. It further illustrates the cascading effect of HVPT through significant perception and selective word-level intelligibility gains which eventually result in delayed gains of comprehensibility in spontaneous oral production.

Moreover, the short intensive self-paced online training program achieved success in delivering an effective phonetic intervention which suggests it has the potential to be accessible by other learners with different backgrounds. Overall, the research shows that HVPT may guide students toward better phonemic perception and more understandable spontaneous oral speech. Learners who receive training on fundamental communicative phonemic contrasts may gradually develop more understandable L2 speech.

Appendix A Pilot study consent and questionnaire form for learner participants

The Impact of Perceptual Training on Comprehensibility in Spontaneous Speech

Dear participant,

Thank you very much for your time. I would like to invite you to take part in this research project.

Before agreeing to take part, please read this information sheet carefully and let me know if anything is unclear or you would like further information.

Please also read the information about General Data Protection Regulation (GDPR). For information about General Data Protection Regulation (GDPR) please follow the link: https://www.york.ac.uk/education/research/gdpr_information/

Then, if you wish to continue with the study, select the 'Consent' button at the end of the Consent Form. A short questionnaire will follow.

Sincerely,

Mia Moutis , PhD student

Department of Education, University of York

Information Sheet

Purpose of the study

The study explores what type of pronunciation training is most effective for Chinese Mandarin learners of English. In particular, it explores the impact of pronunciation training on unplanned speech, with a focus on how easy it is for listeners to understand your speech following the training.

What would this mean for you?

The study takes place online over a training period of up to three weeks. You will participate in an online training. If you are allocated to the control group you will receive no training but you will be given the opportunity to do the training at a later date. Specifically:

♣ Before the training, you will be asked to complete a listening and speaking test.

♣ The training will involve one training session per day. Each session will last up to 15 minutes and involve a series of listening exercises.

♣ After the training, you will be asked to complete the listening and speaking tests again.

♣ One month after the training, you will be asked to complete the listening and speaking tests as final.

If you want to find out more about the study, please email me at mia.moutis@york.ac.uk .

Participation is voluntary

Participation is optional. If you do decide to take part, please complete the online consent form below and maintain a copy of this information by downloading a copy of the completed consent form. If you change your mind at any point during the study, including during data collection, you will be able to withdraw your participation without having to provide a reason by emailing me at mia.moutis@york.ac.uk . All participants will be reimbursed for their participation in the study with a five-pound Amazon voucher for the whole intervention upon completing the study.

Anonymity and confidentiality

The data that you provide such as audio recordings and test results will be stored by code number. Any information that identifies you will be stored separately from the data. You are free to withdraw from the study by emailing me at any time during data collection and up to seven days after the last test.

Storing and using your data

Data will be stored on a password protected computer. Data will be fully anonymized seven days after the last date of data collection is completed.

I am practising Open Science and anonymized data will be managed professionally and stored indefinitely with the University's Research Data York service.

The data that I collect such as audio recordings and test responses may be used in anonymous format in different ways including publications, presentations and online. Please indicate on the consent form enclosed with a tick if you are happy for this anonymized data to be used in the ways listed.

Questions or concerns If you have any questions about this participant information sheet or concerns about how your data is being processed, please feel free to contact me, Mia Moutis by email mia.moutis@york.ac.uk, or the Chair of Ethics Committee via email education-research-admin@york.ac.uk. If you are still dissatisfied, please contact the University's Data Protection Officer at dataprotection@york.ac.uk

I hope that you will agree to take part. If you are happy to participate, please tick "I consent" to all of the four boxes below. Please keep this information sheet for your own records.

Thank you for taking the time to read this information.

Yours sincerely,

Mia Moutis

Consent Form

Please tick each box if you are happy to take part in this research.

Statement of consent	Tick each box
I confirm that I have read and understood the information given to me about the above-named research project and I understand that this will involve me taking part as described above.	
I understand that participation in this study is voluntary and that if I wish to withdraw, I can do so at any time during data collection and up to seven days after the last test.	
I understand that my data will be initially identifiable and then anonymized one week after the last test. The anonymous data may be used in publications, presentations and online.	
I confirm that I have read the information about GDPR.	

Write First and Last name:
Date:
I consent to take part in the above study.

A short background questionnaire follows.

Questionnaire

1. What is your first?

2. What is your last name?

3. What do you identify with?

4. What is your age group?

5. What is your first language?

6. Have you lived in an English-speaking country?

7. At what age did you begin learning English?

8. Do you have any known hearing difficulties?

Thank you for agreeing to take part in the study.

Appendix B Pilot study consent and questionnaire form for raters

The Impact of Perceptual Training on Comprehensibility in Spontaneous Speech

Dear participant,

My name is Mia Moutis, and I am currently carrying out a research project The Impact of Perceptual Training on Comprehensibility in Spontaneous Speech. I would like to invite you to take part in this research project.

Before agreeing to take part, please read this information sheet carefully and let me know if anything is unclear or you would like further information.

Please also read the information about General Data Protection Regulation (GDPR). For information about General Data Protection Regulation (GDPR) please follow the link: https://www.york.ac.uk/education/research/gdpr_information/

Then, if you wish to continue with the study, select the 'Consent' button at the end of the Consent Form. A short questionnaire will follow.

Sincerely,

Mia Moutis , PhD student

Department of Education, University of York

Information Sheet

Purpose of the study

The study is designed to investigate different types of pronunciation training to see which one is more optimal for Chinese Mandarin learners of English in being more comprehensible in spontaneous speech. The study will do this by focusing on an online short training program that will use specific vowels and consonants that are considered challenging for Chinese Mandarin learners of English.

What would this mean for you?

The study will be conducted online. You will be involved in the evaluation process, and you will be asked to rate learners' speech before and after the learner's training. The evaluation process will take up to an hour in total. If you want to find out more about the study, please email me at mia.moutis@york.ac.uk . You will be reimbursed with a 5-pound Amazon voucher after completing the whole evaluation process.

Participation is voluntary

Participation is optional. If you do decide to take part, please complete the online consent form below and maintain a copy of this information by downloading a copy of the completed consent form. If you change your mind at any point during the study, you will be able to withdraw your participation without having to provide a reason by emailing me at mia.moutis@york.ac.uk .

Anonymity and confidentiality

The data that you provide such as ratings of learner productions and data collection, will be stored by code number. Any information that identifies you will be stored separately from the data. You are free to withdraw from the study by emailing me at any time during data collection and up to seven days after completing the rating.

Storing and using your data

Data will be stored on a password protected computer. Data will be fully anonymized seven days after the last date of data collection is completed.

I am practising Open Science and anonymized data will be managed professionally and stored indefinitely with the University's Research Data York service.

The data that I collect such as ratings of learner productions may be used in anonymous format in different ways including publications, presentations and online. Please indicate on the consent form enclosed with a tick ⌋ if you are happy for this anonymized data to be used in the ways listed.

Questions or concerns If you have any questions about this participant information sheet or concerns about how your data is being processed, please feel free to contact me, Mia Moutis by email mia.moutis@york.ac.uk, or the Chair of Ethics Committee via email education-research-admin@york.ac.uk. If you are still dissatisfied, please contact the University's Data Protection Officer at dataprotection@york.ac.uk

I hope that you will agree to take part. If you are happy to participate, please tick "I consent" to all of the four boxes below. Please keep this information sheet for your own records.

Thank you for taking the time to read this information.

Yours sincerely,

Mia Moutis

Consent Form

Please tick each box if you are happy to take part in this research.

Statement of consent	Tick each box
I confirm that I have read and understood the information given to me about the above-named research project and I understand that this will involve me taking part as described above.	
I understand that participation in this study is voluntary and that if I wish to withdraw, I can do so at any time during data collection and up to seven days after completing the rating.	

I understand that my data will be initially identifiable and then anonymized seven days after the last test. The anonymous data may be used in publications, presentations and online.	
I confirm that I have read the information about GDPR	

Write First and Last name:
Date:
I consent to take part in the above study.

A short background questionnaire follows.

Questionnaire

1. What is your first name?
2. What is your last name?
3. What do you identify with?
4. What is your age group?
5. What is your first language?
6. What is your level of English?
7. Do you speak another language? If yes, please state which language.
8. Have you ever attended an English language or linguistics course?
9. Have you ever taught an English language or linguistics course?
10. Do you have any known hearing difficulties?

Thank you for agreeing to take part in the study.

Appendix C Main study consent and questionnaire form for learner participants

UNIVERSITY of York

The Impact of Perceptual Training on Comprehensibility in Spontaneous Speech

Dear Student,
My name is Mia Moutis and I am currently carrying out a research project *The Impact of Perceptual Training on Comprehensibility in Spontaneous Speech*. I would like to invite you to take part in this research project.
Before agreeing to take part, please read this information sheet carefully and let me know if anything is unclear or you would like further information.

Please also read the information about General Data Protection Regulation (GDPR). For information about General Data Protection Regulation (GDPR) please follow the link:
https://www.york.ac.uk/education/research/gdpr_information/

Purpose of the study
The study explores what type of pronunciation training is most effective for Chinese Mandarin learners of English. In particular, it explores the impact of pronunciation training on unplanned speech, with a focus on how easy it is for listeners to understand your speech following the training.

What would this mean for you?
The study takes place online over a training period of up to three weeks. You will participate in one of the types of training (all forms of the training have been shown to be beneficial). If you are allocated to the control group you will receive no training but you will be given the opportunity to do the training at a later date. Specifically:

- Before the training, you will be asked to complete a listening and speaking test.
- The training will involve one training session per day. Each session will last up to 15 minutes and involve a series of listening exercises.
- After the training, you will be asked to complete the listening and speaking tests again and a short audio recorded open interview describing what you thought of the pronunciation training.
- One month after the training, you will be asked to complete the listening and speaking tests as final.

If you want to find out more about the study, please email me at mia.moutis@york.ac.uk .

Participation is voluntary
You will be given the opportunity to comment on a written record of your interview. If you wish to do so, please email the researcher (mia.moutis@york.ac.uk) within seven days of the interview taking place.
Participation is optional. If you do decide to take part, please complete the online consent form below and maintain a copy of this information by downloading a copy of the completed consent form. If you change your mind at any point during the study, including during data

collection, you will be able to withdraw your participation without having to provide a reason by emailing me at mia.moutis@york.ac.uk .

Anonymity and confidentiality

The data that you provide such as audio recordings and test results will be stored by code number. Any information that identifies you will be stored separately from the data. You are free to withdraw from the study by emailing me at any time during data collection and up to seven days after the last test.

Storing and using your data

Data will be stored on a password protected computer. Data will be fully anonymized seven days after the last date of data collection is completed.

I am practising Open Science and anonymized data will be managed professionally and stored indefinitely with the University's Research Data York service.

The data that I collect such as audio recordings and test responses may be used in anonymous format in different ways including publications, presentations and online. Please indicate on the consent form enclosed with a tick ☑ if you are happy for this anonymized data to be used in the ways listed.

Questions or concerns

If you have any questions about this participant information sheet or concerns about how your data is being processed, please feel free to contact me, **Mia Moutis** by email **mia.moutis@york.ac.uk**, or the Chair of Ethics Committee via email education-research-admin@york.ac.uk. If you are still dissatisfied, please contact the University's Data Protection Officer at dataprotection@york.ac.uk

I hope that you will agree to take part. If you are happy to participate, please tick "I agree" to all of the four boxes below. Please keep this information sheet for your own records.

Thank you for taking the time to read this information.

Yours sincerely
Mia Moutis

Please tick each box if you are happy to take part in this research.

Statement of consent	Tick each box
I confirm that I have read and understood the information given to me about the above-named research project and I understand that this will involve me taking part as described above.	
I understand that participation in this study is voluntary and that if I wish to withdraw, I can do so at any time during data collection and up to seven days after the last test.	
I understand that my data will be initially identifiable and then anonymised one week after the last test. The anonymous data may be used in publications, presentations and online.	
I confirm that I have read the information about GDPR.	

Write first and last name:
Name:
Signature:
Date:
I consent to take part in the above study.

A short background questionnaire follows.

Questionnaire

1. What is your first name?
2. What is your last name?
3. What do you identify with?
4. What is your age group?
5. What is your first language?
6. Have you lived in an English-speaking country?
7. At what age did you begin learning English?
8. Do you have any known hearing difficulties?

Thank you for agreeing to take part in the study.

Appendix D Main study consent and questionnaire form for raters

UNIVERSITY of York

The Impact of Perceptual Training on Comprehensibility in Spontaneous Speech

Dear Participant,
My name is Mia Moutis, and I am currently carrying out a research project *The Impact of Perceptual Training on Comprehensibility in Spontaneous Speech*. I would like to invite you to take part in this research project.
Before agreeing to take part, please read this information sheet carefully and let me know if anything is unclear or you would like further information.

Please also read the information about General Data Protection Regulation (GDPR). For information about General Data Protection Regulation (GDPR) please follow the link**:**
https://www.york.ac.uk/education/research/gdpr_information/

Purpose of the study
The study is designed to investigate different types of pronunciation training to see which one is more optimal for Chinese Mandarin learners of English in being more comprehensible in spontaneous speech. The study will do this by focusing on an online short training program that will use specific vowels and consonants that are considered challenging for Chinese Mandarin learners of English.

What would this mean for you?
The study will be conducted online. You will be involved in the evaluation process, and you will be asked to rate learners' speech before and after the learner's training. The evaluation process will take up to two hours in total. If you want to find out more about the study, please email me at mia.moutis@york.ac.uk . You will be reimbursed with a 5-pound Amazon voucher after completing the whole evaluation process.

Participation is voluntary
Participation is optional. If you do decide to take part, please complete the online consent form below and maintain a copy of this information by downloading a copy of the completed consent form. If you change your mind at any point during the study, you will be able to withdraw your participation without having to provide a reason by emailing me at mia.moutis@york.ac.uk

Anonymity and confidentiality
The data that you provide such as ratings of learner productions and data collection, will be stored by code number. Any information that identifies you will be stored separately from the data. You are free to withdraw from the study by emailing me at any time during data collection and up to seven days after completing the rating.
Storing and using your data

Data will be stored on a password protected computer. Data will be fully anonymized seven days after the last date of data collection is completed.

I am practising Open Science and anonymized data will be managed professionally and stored indefinitely with the University's Research Data York service.

The data that I collect such as ratings of learner productions may be used in anonymous format in different ways including publications, presentations and online. Please indicate on the consent form enclosed with a tick ☑ if you are happy for this anonymized data to be used in the ways listed.

Questions or concerns

If you have any questions about this participant information sheet or concerns about how your data is being processed, please feel free to contact me, **Mia Moutis** by email **mia.moutis@york.ac.uk**, or the Chair of Ethics Committee via email education-research-admin@york.ac.uk. If you are still dissatisfied, please contact the University's Data Protection Officer at dataprotection@york.ac.uk

I hope that you will agree to take part. If you are happy to participate, please tick "I agree" to all of the four boxes below. Please keep this information sheet for your own records.

Thank you for taking the time to read this information.

Yours sincerely
Mia Moutis

Please tick each box if you are happy to take part in this research.

Statement of consent	Tick each box
I confirm that I have read and understood the information given to me about the above-named research project and I understand that this will involve me taking part as described above.	
I understand that participation in this study is voluntary and that if I wish to withdraw, I can do so at any time during data collection and up to seven days after completing the rating.	
I understand that my data will be initially identifiable and then anonymized seven days after the last test. The anonymous data may be used in publications, presentations and online.	
I confirm that I have read the information about GDPR	

Write First and Last name:
Name:
Date:

I consent to take part in the above study.

A short background questionnaire follows.

Questionnaire

1. What is your first name?
2. What is your last name?
3. What do you identify with?
4. What is your age group?
5. What is your first language?
6. What is your level of English?
7. Do you speak another language? If yes, please state which language.
8. Have you ever attended an English language or linguistics course?
9. Have you ever taught an English language or linguistics course?
10. How familiar are you with Chinese accented English?
11. Do you have any known hearing difficulties?

Thank you for agreeing to take part in the study.

Appendix E Stimuli of the Individual lessons for the high variability phonetic training

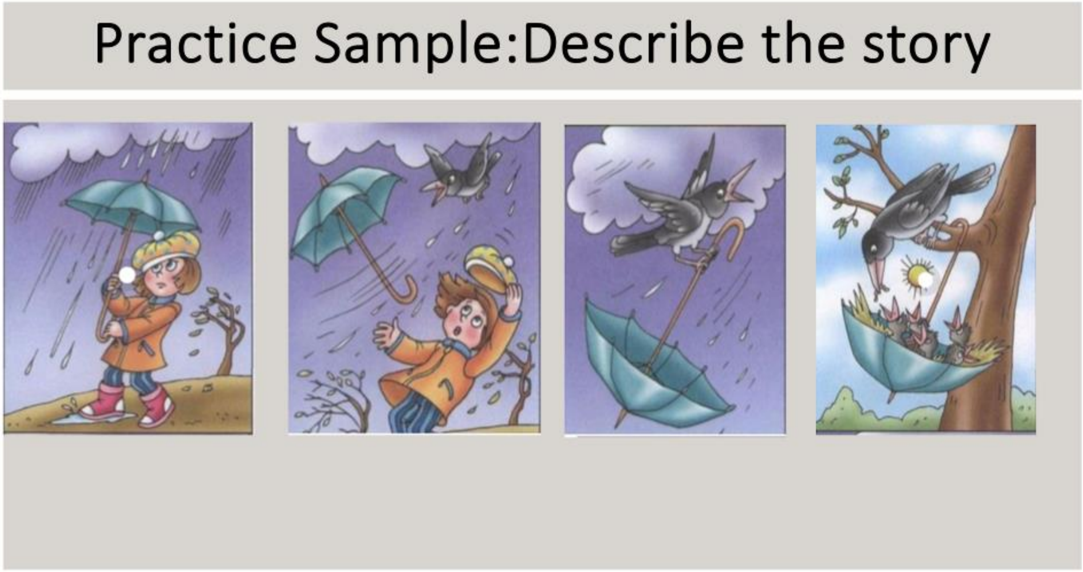
Lesson 1 /ɛ - æ/	Lesson 2 /p - b/	Lesson 3 /æ - ʌ/	Lesson 4 /m-n/	Lesson 5 /æ-ɒ/	Lesson 6 /l-r/	Lesson 7 /ɔ: - əʊ/
affluent effluent	tripe - tribe	bag-bug	mode - node	baggy - boggy	flee - free	corn - -cone
latter - letter	stable - staple	crash-crush	rummer runner	occident accident	collect correct	yoke - York
bag - beg	blunder plunder	match-much	teen - team	topping - tapping	alive arrive	sorely solely
and - end	oppressed abreast	lack-luck	beam - bean	hallow - hollow	climb crime	portion potion
tan - ten	gab - gap	ankle-uncle	nope - mope	add - odd	glass grass	toe - tore
past - past	pumpkin bumpkin	bubble-babble	lining - liming	lorry - Larry	late - rate	load - lord
excess access	rip - rib	grab-grub	tum -tun	valley - volley	towelling towering	store - stow
amber ember	unpacked unbacked	natty - nutty	night - might	bonding banding	rush - lush	morn moan
fellow - fallow	crumble crumple	attar - utter	warming warning	packet - pocket	mallow marrow	porker poker
bally - belly	swap - swab	slang - slung	then - them	opposition apposition	write light	norm gnome
land - lend	rabid - rapid	damp dump	smack - snack	spot - spat	splinter sprinter	court - coat
rant - rent	pump - bump	plank - plunk	chrone chrome	adapt - adopt	present pleasant	bone borne
vessel - vassal	tab - tap	chump champ	sunning summing	chop - chap	rune loon	corm comb
track - trek	simple symbol	gammy gummy	timed - tined	ox - axe	finery finally	hone - horn
areca - Erica	robe - rope	umber amber	snuggle smuggle	bland - blond	blacken bracken	shore - show
mention mansion	harper harbour	jug - jag	dime - dine	cap -cop	royal loyal	war - woe
flesh - flash	prude - brood	shatter shutter	span - spam	flop - flap	play - pray	mow - more
cattle - kettle	slap - slab	lamp - lump	nuzzle muzzle	apt - opt	balling boring	tote - tort
malady melody	mopping mobbing	campus compass	gleaming gleaning	jobber - jabber	flame frame	porch poach
jam - gem	tabor - taper	tussle tassel	mime - mine	shack - shock	corrie collie	folk - fork

Appendix E continued

Lesson 8 /p - f/	Lesson 9 /ʌ - ɒ/	Lesson 10 /n-l/	Review Lesson 1	Review Lesson 2	Review Lesson 3	Review Lesson 4	Review Lesson 5
append offend	shot - shut	line - nine	All words from Lessons 1 & 2	All words from Lessons 3 & 4	All words from Lessons 5 & 6	All words from Lessons 7 & 8	All words from Lessons 9 & 10
faint - paint	boss - bus	slow - snow					
reaping reefing	cop - cup	life - knife					
wife - wipe	buddy body	collect connect					
snipping sniffing	lock -luck	let - net					
coffer copper	hubby hobby	lame - name					
puppy puffy	fund - fond	shill - shin					
leap - leaf	smug smog	belly Benny					
differ dipper	model muddle	lead - need					
gulf - gulp	dock - duck	teller tenor					
plash -flash	long - lung	belt - bent					
fashion passion	rock - ruck	sleek sneak					
refutation reputation	stock - stuck	meal - mean					
beep - beef	bunny bonny	rail - rain					
fat - pat	podgy pudgy	lock - knock					
cheep chief	toff - tuff	lap - nap					
suffer supper	wrong rung	nag - nag					
peel - feel	lug - log	correlation coronation					
top - toff	jog - jug	mulching munching					
defend depend	muss - moss	trail - train					

Appendix F Learner practice sample: Comprehensibility Task 1 timed picture story

Your task is to speak about the picture story for 2 minutes in English. You have 10 seconds to think about what you plan to say. Please begin speaking when the researcher begins the recording.



Practice Sample:Describe the story

Appendix G Learner practice sample: Comprehensibility Task 2 long turn narrative

Your task is to speak on the topic below for 2 minutes in English. You have 1minute to think about what you plan to say. Notetaking is not allowed. Please begin speaking when the researcher begins the recording.

A Typical Day

Talk about a typical day at work, school or college.

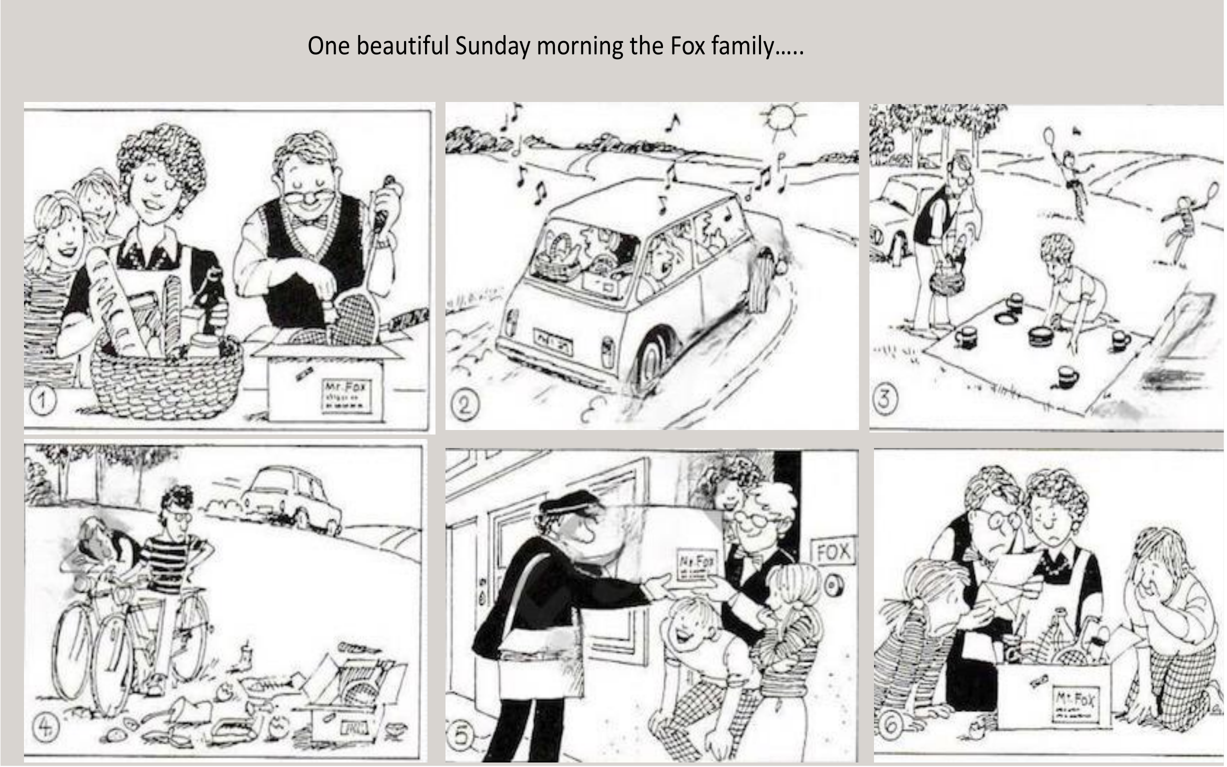
Supporting Questions:

- What you do
- When you do it
- How long you have had this routine
- Explain what you would like to change in your work or study routine

Appendix H Learner official test: Comprehensibility Task 1 timed picture story

Timed picture story (same version across all pre-, post- and delayed post-tests)

Your task is to speak about the picture story for 2 minutes in English. You have 10 seconds to think about what you plan to say. Please begin speaking when the researcher begins the recording.



One beautiful Sunday morning the Fox family.....

Appendix I Learner official test: Comprehensibility Task 2 long turn narrative

Task 2 Long turn narrative speaking (**Pre-test and delayed post-test version**)

Your task is to speak on the topic below for 2 minutes in English. You have 1minute to think about what you plan to say. Notetaking is not allowed. Please begin speaking when the researcher begins the recording.

A Shop Near Me
Talk about a shop near where you live now that you sometimes visit.

Supporting Questions:

- What sort of products or services it sells
- What the shop looks like
- Where it is located

And explain why you visit this shop.

Task 2 Long turn narrative speaking **(Post-test version)**

Your task is to speak on the topic below for 2 minutes in English. You have 1minute to think about what you plan to say. Notetaking is not allowed. Please begin speaking when the researcher begins the recording.

A Favourite Shop
Talk about your favourite shop that you visit.

Supporting Questions:

- what sort of products or services it sells
- what the shop looks like
- where it is located

And explain why you visit this shop.

Appendix J Comprehensibility rating scale for Task 1 timed picture story instructions for raters

In this task you will listen to Chinese learners describing a picture story. The rating scale refers to how much effort it takes to understand what someone is trying to convey. Your task is to rate how easy it is to understand what someone is trying to convey. If you can understand with ease, then the speaker is highly comprehensible. However, if you struggle or cannot understand what is being said at all, then a speaker has low comprehensibility.

Please read and apply the following scale:

0	Unable to rate	Response is unrelated to the topic and has no relevant information.
1	**Very difficult** to understand	The response lacks coherence or is not effective. It is **very difficult** to understand the response.
2	**Difficult** to understand	The response lacks coherence, and it is **difficult** to understand.
3	**Some effort** to understand	With **some effort**, it is possible to understand the response. The speaker's message comes across and is somewhat coherent.
4	**Easy** to understand	It is **easy** to understand the response. The speaker's message comes across clearly and is coherent. The speaker takes account of the communicative situation.
5	**Very easy** to understand	It is **very easy** to understand the response. The speaker's message comes across very clearly and is very coherent. The speaker carefully takes account of the communicative situation.

Source: de Jong et al. (2012)

Step 2: Practice Audio Samples

- Please rate the practice audio samples using the scales provided on–screen.
- When completed, you will be assigned the real audio samples.

Step 3: Real Audio Samples

- Please rate the real audio samples using the scales provided on-screen.

Thank you for taking part in the study.

Appendix K Comprehensibility rating scale for Task 2 long-turn narrative instructions for raters

In this task you will listen to Chinese learners talking about a shop near them or about their favourite shop. The rating scale refers to how much effort it takes to understand what someone is trying to convey. Your task is to rate how easy it is to understand what someone is trying to convey. If you can understand with ease, then the speaker is highly comprehensible. However, if you struggle or cannot understand what is being said at all, then a speaker has low comprehensibility.

Please read and apply the following scale:

0	Unable to rate	Response is unrelated to the topic and has no relevant information.
1	**Very difficult** to understand	The response lacks coherence or is not effective. It is **very difficult** to understand the response.
2	**Difficult** to understand	The response lacks coherence, and it is **difficult** to understand.
3	**Some effort** to understand	With **some effort**, it is possible to understand the response. The speaker's message comes across and is somewhat coherent.
4	**Easy** to understand	It is **easy** to understand the response. The speaker's message comes across clearly and is coherent. The speaker takes account of the communicative situation.
5	**Very easy** to understand	It is **very easy** to understand the response. The speaker's message comes across very clearly and is very coherent. The speaker carefully takes account of the communicative situation.

Source: de Jong et al. (2012)

Step 2: Practice Audio Samples

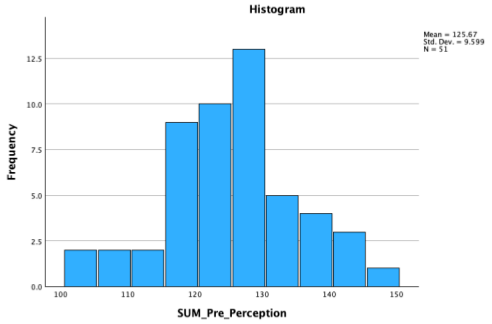
- Please rate the practice audio samples using the scales provided on–screen.
- When completed, you will be assigned the real audio samples.

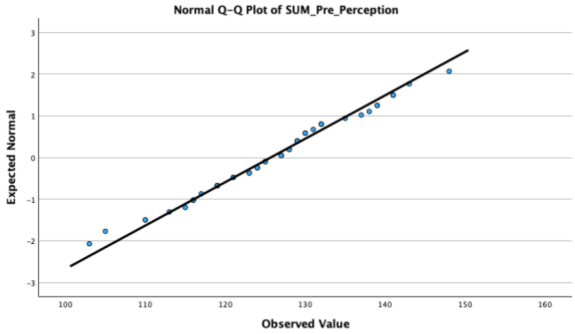
Step 3: Real Audio Samples

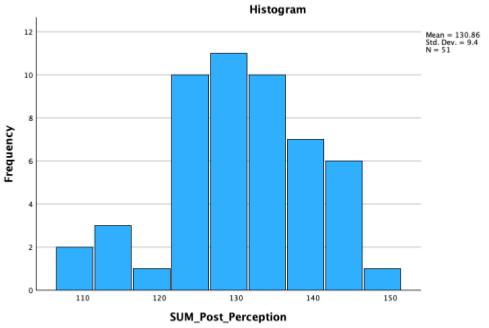
- Please rate the real audio samples using the scales provided on-screen.

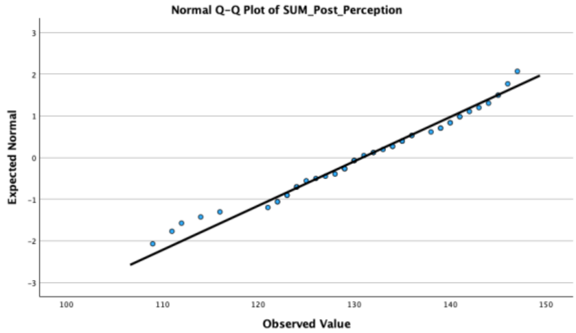
Thank you for taking part in the study.

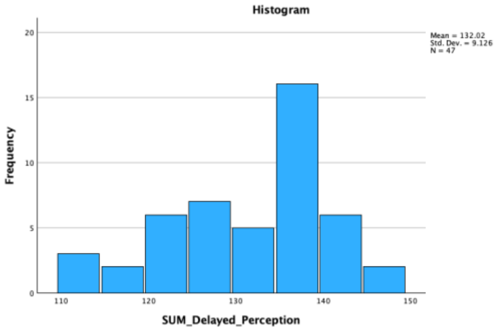
Appendix L Histograms and Q-Q plots of the perception test at pre, post and delayed tests

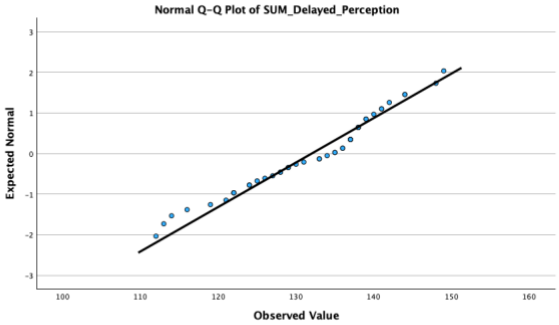












Appendix M ANOVA results for differences in perception test at pre-, post-, and delayed post-tests on raw scores

This appendix presents the ANOVA results for differences in the perception test at pre, post and delayed tests on raw scores. It includes the F-statistic and p-value to indicate statistical significance.

Effect	**SS**	**df**	**MS**	**F**	***p***	**η^2**
Within subjects	1080.01	1.617	668.10	7.51	0.002	.14

** Due to violation of sphericity the Greenhouse Geisser correction was applied*

Appendix N Wilcoxon Signed Rank test of the intelligibility test within group on all available L2 learners at pre, post and delayed post-tests.

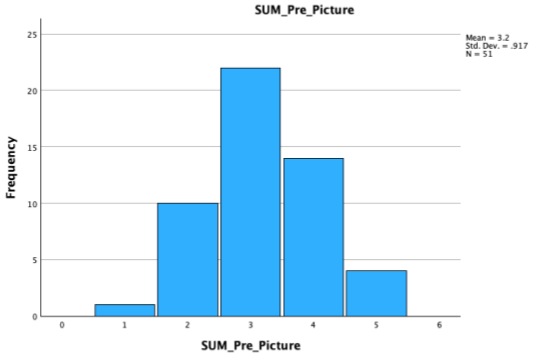
This appendix presents the Wilcoxon Signed Rank test of the intelligibility test within group on a dataset of all available L2 learners at pre, post and delayed post-tests.

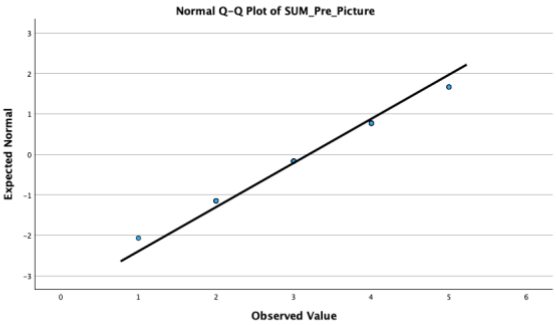
		Mean Rank				
Time	*N*	Negative	Positive	Ties	*Z*	*p* value
Pre-test to Post-test	43	14	19	10	-.346	.730
Pre-test to Delayed post-test	33	20	5	8	-2.675	.007
Post-test to Delayed post-test	33	22	7	4	-2.627	.009

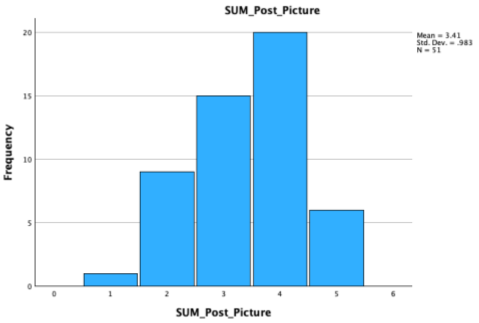
**Intelligibility pre-test excluded 8 participants; intelligibility post-test excluded 18; and delayed post-test excluded 11 participants*

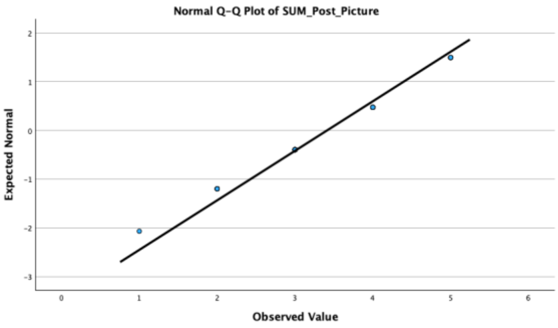
Appendix O Histograms and Q-Q plots of the comprehensibility test Task 1 at pre, post and delayed tests

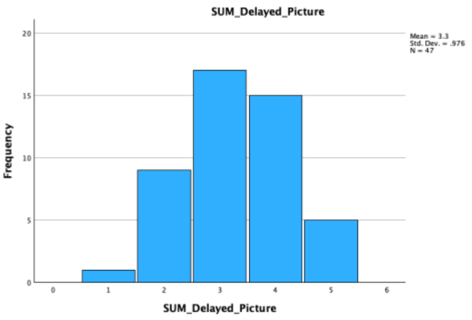
This appendix presents the histograms and Q-Q plots of raters’ listening scores for L2 learners’ comprehensibility in spontaneous productions during the timed picture story task 1 across pre, post and delayed post-tests.

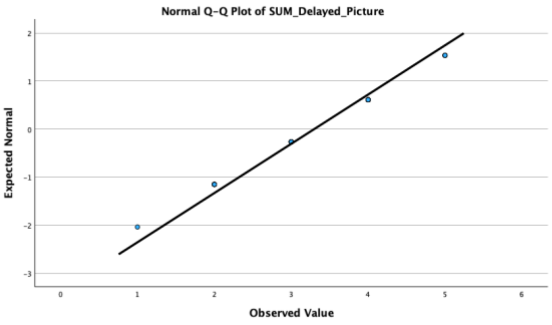










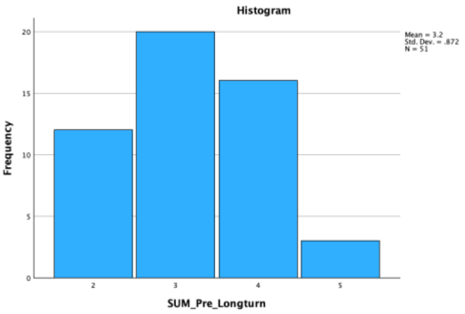


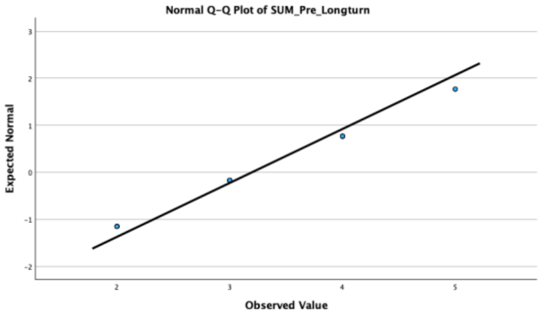
Appendix P Friedman test results for comprehensibility Task 1 at pre-, post-, and delayed post-tests

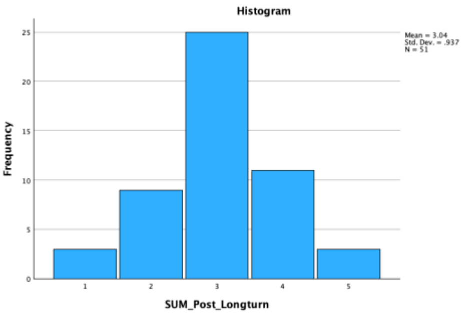
Task	N	Chi-Square ($x^{2)}$	(*df*)	Asymp. Sig.(*p*)	Mean Rank Pre-test	Mean Rank Post-test	Mean Rank Delayed post-test
Timed picture story	47	0.964	2	0.618	1.90	2.07	2.02

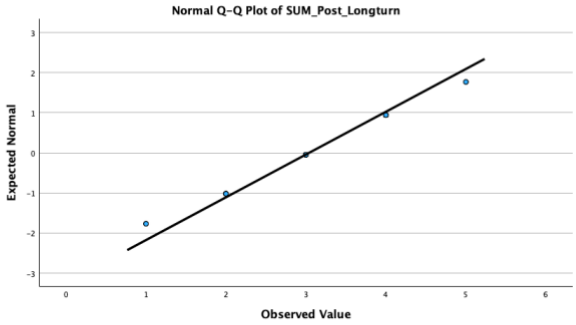
Appendix Q Histograms and Q-Q plots of the comprehensibility test Task 2 at pre, post and delayed tests

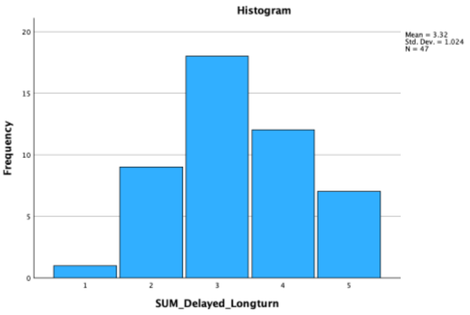
This appendix presents the histograms and Q-Q plots of raters' listening scores for L2 learners' comprehensibility in spontaneous productions during the long-turn narrative Task 2 across pre, post and delayed post-tests

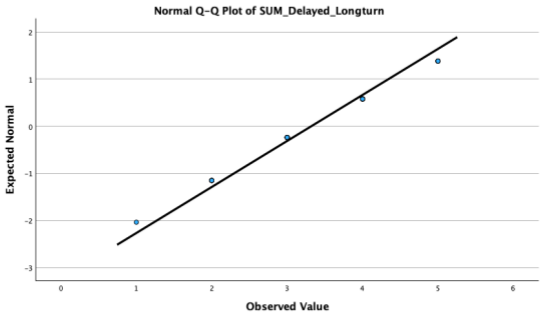












Appendix R Friedman test results for comprehensibility Task 2 at pre-, post-, and delayed post-tests

Task	*N*	Chi-Square (x$^{2)}$	(*df*)	Asymp. Sig.(*p*)	Mean Rank Pre-test	Mean Rank Post-test	Mean Rank Delayed post-test
Long- Turn Narrative	47	3.036	2	0.219	2.01	1.84	2.15

Appendix S Summary of Lesson Duration Data

This appendix presents the summary of lesson duration data based on available log data from a subset of learners. The table includes the total recorded time (in minutes), the number of learners who completed each lesson, and the average lesson duration as recorded by the Gorilla Experiment Builder platform.

Lesson	**Total Learners (N)**	**Total Recorded Time (min)**	**Average Time of Lesson Duration (min)**
Lesson 1	35	548.46	15.67
Lesson 2	12	171.54	14.30
Lesson 3	8	87.66	17.53
Lesson 4	9	98.18	10.90
Lesson 5	11	124.58	11.32
Lesson 6	9	124.99	13.89
Lesson 7	10	147.74	14.77
Lesson 8	27	266.47	9.89
Lesson 9	11	137.28	12.48
Lesson 10	5	107.26	21.45
Review 1	13	221.46	17.03
Review 2	7	84.05	12.01
Review 3	7	115.90	16.56
Review 4	6	50.49	8.42
Review 5	13	144.09	11.08

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance, 35*(2), 520–529. https://doi.org/10.1037/a0013552

Aliaga-García, C., & Mora, J. C. (2009). Assessing the effects of phonetic training on L2 sound perception and production. In Watkins MA, Rauber A.S and Baptista B.O. editors. *Recent Research in Second Language Phonetics/phonology: Perception and Production, 2*(31). Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics, 32*, 233–250. https://doi.org/10.1016/S0095-4470(03)00036-6

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, *81*, 981–1005. https://doi.org/10.3758/s13414-019-01725-4

Baker, W., & Trofimovich, P. (2006). Perception and production of second language speech sounds: The relationship and its implications for pronunciation instruction. *Studies in Second Language Acquisition, 28*(2), 197-221.

Barefoot, S. M., Bochner, J. H., Johnson, B. A., & vom Eigen, B. A. (1993). Rating deaf speakers' comprehensibility: An exploratory investigation. *American Journal of Speech-Language Pathology, 2*(3), 31-35. https://doi.org/10.1044/1058-0360.0203.31

Barrass, D., Baffoe-Djan, B. J., Rose, H., & Boggs, A. J. (2020). Intelligibility and comprehensibility of Korean English Speakers' Phonological Features in Lingua Franca Listening Contexts. *The Journal of Asis TEFL, 17*(1), 1-318. https://doi.org/10.18823/asiatefl.2020.17.1.8.124

Barriuso, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *The CATESOL Journal, 30*(1), 177–202.31. https://doi.org/10.5070/B5.35970

Benrabah, M. (1997). Word-stress: A source of unintelligibility in English. *International Review of Applied Linguistics in Language Teaching*, *35*(3), 157–165. https://doi.org/10.1515/iral.1997.35.3.157

Bent, T., Bradlow, A. R., & Smith, B. L. (2007a). Production and perception of temporal patterns in native and non-native speech. *Phonetica, 64*(2-3), 131–147. https://doi.org/10.1159/000144077

Bent, T., Bradlow, A. R., & Smith, B. L. (2007b). Phonemic errors in different word positions and their effects on intelligibility of non-native speech. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 331–347). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.17.28ben

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics, 20*(3), 305–330. https://doi.org/10.1016/S0095-4470(19)30637-0

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171-204). Baltimore, MD: York Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). https://doi.org/10.1075/lllt.17.07bes

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299–2310. https://doi.org/10.1121/1.418276

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*(5), 977–985. https://doi.org/10.3758/BF03206911

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brosseau-Lapré, F., Rvachew, S., & Brosseau-Lapré, F. (2013). The effects of high-variability phonetic training on the perception of /r/ and /l/ by Japanese learners of English. *Journal of Phonetics, 41*(5), 363–378.

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly, 22*(4), 593-606. https://doi.org/10.2307/3587258

Brown, A. (1991). *Pronunciation models*. Singapore: Singapore University Press.

Bybee, J. (2001). *Phonology and language use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511612886

Bybee, J. (2008). Usage-based grammar and second language acquisition. In P. Robinson & N. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 216–236). New York, NY: Routledge.

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., & Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *NeuroImage, 19*(1), 113–124. https://doi.org/10.1016/S1053-8119(03)00020-X

Carlet, A. F. (2017). *L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study* (Doctoral dissertation). Universitat Autònoma de Barcelona.

Carlet, A. F., & Cebrian, J. (2015). Identification vs. discrimination training: Learning effects for trained and untrained sounds. *In Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)* (pp. 1-5). International Phonetic Association. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0752.pdf

Catford, J. C. (1987). Phonetics and the teaching of pronunciation: A systemic description of English phonology. *In J. Morley (Ed.). Current Perspectives on pronunciation: Practices anchored in theory*, (pp. 87-100). Tesol Press.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation: A course book and reference guide (2nd ed.).* Cambridge University Press

Cheng, B., Zhang, X., Fan, S., & Zhang, Y. (2019). The Role of Temporal Acoustic Exaggeration in High Variability Phonetic Training: A Behavioral and ERP Study. *Frontiers in Psychology, 10*, 1178. https://doi.org/10.3389/fpsyg.2019.01178

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Crowther, D., & Isbell, D. (2023). Second language speech comprehensibility: a research agenda. *Language Teaching*, *56* (4), 1-17. Cambridge University Press.

Crowther, D., Saito, K., Isaacs, T., & Trofimovich, P. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, *49*(4), 814–837. https://doi.org/10.1002/tesq.203

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015a). Does a speaking task affect second language comprehensibility? *Modern Language Journal, 99*(1), 80-95. https://doi.org/10.1111/modl.12185

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition, 40*(2), 443–457. https://doi.org/10.1017/S027226311700016X

Darcy, I. (2018). Powerful and Effective Pronunciation Instruction: How Can We Achieve It? *The CATESOL Journal, 30*(1), 13–36. https://doi.org/10.5070/B5.35963

Darcy, I., & Mora, J. C. (2014). Attention control and inhibition influence phonological development in a second language. *In Proceedings of the International Symposium on the Acquisition of Second Language Speech* (Vol. 5, pp. 115-129). Montreal: Concordia Working Papers in Applied Linguistics (COPAL).

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1536), 3773–3800. https://doi.org/10.1098/rstb.2009.0111

De Jong, N.H., Steinel, M.P., Florijn, A.F., Schoonen, R., & Hulstijn, J. (2012) Facets of speaking proficiency. *Studies in Second Language Acquisition 34*(1) 5–34. https://doi.org/10.1017/S0272263111000489

DeKeyser, R. (Ed.). (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667275

DeKeyser, R., & Criado, R. (2013). Automatization, skill acquisition and practice in second language acquisition. In M. García Mayo, M. G. Mangado, & M. M. Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 94–111). John Benjamins. https://doi.org/10.1002/9781405198431.wbeal0067

Deng, Z., Chandrasekaran, B., Wang, S., & Wong, P. C. M. (2018). Training-induced brain activation and functional connectivity differentiate multi-talker and single-talker speech training. *Neurobiology of Learning & Memory, 151*, 1–9. https://doi.org/10.1016/j.nlm.2018.03.009

Derwing, T. M., Fraser, H., Kang, O., & Thomson, R. I. (2014). L2 accent and ethics: Issues that merit attention. In A. Mahboob & L. Barratt (Eds.), *Englishes in multilingual contexts* (pp. 63–80). Springer. https://doi.org/10.1007/978-94-017-8869-4_5

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42,* 476–490. https://doi.org/10.1017/S026144480800551X

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.42

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*(3), 359–380. https://doi.org/10.1093/applin/amm041

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*(3), 393–410. https://doi.org/10.1111/0023-8333.00047

Derwing, T. M., & Rossiter, M. J. (2002). ESL learners' perceptions of their pronunciation needs and strategies. *System*, *30*(2), 155–166. https://doi.org/10.1016/S0346-251X(02)00012-X

Deterding, D. (2013). *Misunderstandings in English as a Lingua Franca: An Analysis of ELF Interactions in South-East Asia*. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110288599

Dong, H., Clayards, M., Brown, H., & Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ*, 7, e7191. https://doi.org/10.7717/peerj.7191.

Earle, F. S., & Myers, E. B. (2014). Building phonetic categories: An argument for the role of sleep. *Frontiers in Psychology, 5*, 1192. https://doi.org/10.3389/fpsyg.2014.01192

Eckman, F. R. (2008). Markedness and the contrastive analysis hypothesis. In M. H. Long & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 285-297). Wiley-Blackwell.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 122–139). Cambridge University Press. https://doi.org/10.1017/S0272263102002024

Ellis, R. (Ed.). (2005). *Planning and task performance in a second language*. John Benjamins Publishing Company. https://doi.org/10.1075/lllt.11

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17–44. https://doi.org/10.1017/S0267190512000025

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399–423. https://doi.org/10.2307/3588487

Field, A. (2024). *Discovering statistics using IBM SPSS statistics* (6th ed.). SAGE Publications.

Flege, J. E. (1995a). Second language speech theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-377). York Press.

Flege, J. E. (1995b). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics, 16*, 425-442. https://doi.org/10.1017/S0142716400066029

Flege, J. E., & Bohn, O.-S. (2021). The revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995a). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America, 97*(5), 3125–3134. https://doi.org/10.1121/1.413041

Flege, J. E., Takagi, N., & Mann, V. (1995b). Japanese adults can learn to produce English /ɹ/ and /l/ accurately. *Language and Speech, 38*, 25–55. https://doi.org/10.1177/002383099503800102

Franken, M. K., Acheson, D. J., McQueen, J. M., & Hagoort, P. (2017). Individual differences in speech perception: The effect of auditory acuity and vowel space density on vowel production. *The Journal of the Acoustical Society of America*, *142*(4), 2007–2018. https://doi.org/10.1121/1.5006899

Foote, J. A., Trofimovich, P., Collins, L., & Urzúa, F. S. (2016). Pronunciation teaching practices in communicative second language classes. *The Language Learning Journal, 44*(2), 181–196. https://doi.org/10.1080/09571736.2013.784345

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18, 299-323*. https://doi.org/10.1017/S0272263100015047

Fouz-González, J. (2020). Using apps for pronunciation training: An empirical evaluation of the English File Pronunciation App. *Language Learning & Technology, 24*(1), 62–85. https://doi.org/10.64152/10125/44709

Gabay, Y., Dorfman, R., & Holt, L. L. (2015). Auditory learning and generalization: The role of variability in training. *Journal of Experimental Psychology: Human Perception and Performance, 41*(1), 1–13. https://doi.org/10.1037/xhp0000073

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3rd ed.). New York, NY: Routledge. https://doi.org/10.4324/9780203932841

Gass, S., & Varonis, E. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 65–87. https://doi.org/10.4324/9780203932841

Gick, B., Bernhardt, B., Bacsfalvi, P., & Wilson, I. (2008). Ultrasound imaging applications in second language acquisition. *Phonology and Second Language Acquisition, 36*, 315-328. https://doi.org/10.1075/sibil.36.15gic

Gilakjani, A. P., & Ahmadi, M. R. (2011). A Study on Factors Affecting EFL Learners' English Pronunciation Learning and the Strategies for Improvement. *Journal of Language Teaching and Research, 2(*6), 1247–1254. https://doi.org/10.4304/jltr.2.5.977-988

Gooskens, C., van Heuven, V. J., van Bezooijen, R., & Pacilly, J. J. (2010). Is spoken Danish less intelligible than Swedish? , *Speech Communication, 52*, 1022–1037. https://www.let.rug.nl/gooskens/pdf/publ_speechcom_2010.pdf

Grantham O'Brien, M., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., Strik, H., Levis, J. M., Munro, M. J., Foote, J. A., & Muller Levis, G. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation, 4*(2), 182–207. https://doi.org/10.1075/jslp.17001.obr

Grim, F., & Sturm, J. (2016). *Where does pronunciation stand in the 21st century foreign language classroom? Educators' and learners' views* [In *Proceedings of Pronunciation in Second Language Learning & Teaching 7*]. Iowa State University. https://apling.engl.iastate.edu/wp-content/uploads/sites/221/2016/08/PSLLT7_July29_2016_B.pdf

Guion, S. G., & Pederson, E. (2007). Investigating the role of attention in phonetic learning. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 57–77). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.17.09gui

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*, 201–223. https://doi.org/10.2307/3588378

Handley, Z. L. (2005). *Evaluating Text-to-Speech (TTS) synthesis for use in Computer-Assisted Language Learning (CALL)* (Doctoral dissertation, University of Manchester).

Handley, Z., Sharples, M. & Moore, D. (2009). Training novel phonemic contrasts: a comparison of identification and oddity discrimination training. *In Speech and Language Technology in Education (SLaTE) 2009 Conference*, (pp. 3-5), Wroxall, England. https://doi.org/10.21437/SLaTE.2009-31

Handley, Z., & Wang, H. (2018). What is the impact of study abroad on oral fluency development? *A comparison of study abroad and study at home (ELT Research Papers 17.05).* British Council. https://www.teachingenglish.org.uk/sites/teacheng/files/pub_h136_final_web.pdf

Handley, Z. L., & Wang, H. (2023). What do the measures of utterance fluency employed in Automatic Speech Evaluation (ASE) tell us about oral proficiency? *Language Assessment Quarterly*, *21*(1), 3-32. https://doi.org/10.1080/15434303.2023.2283839

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24*(4), 495-522. https://doi.org/10.1017/S0142716403000250

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology, 8*(1), 34–52. https://doi.org/10.64152/10125/25228

Harding, L. (2017). What do raters need in a pronunciation scale? The user's view. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 12–36). Multilingual Matters.

Haslam, M., & Zetterholm, E. (2016). The importance of aspirated initial stops in English as a lingua franca. In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds.), Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference (pp. 66-75). Ames, IA: Iowa State University.

Hattori, K., & Iverson, P. (2009). English /r/–/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America, 125*(1), 469–479. https://doi.org/10.1121/1.3021295

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360-378. https://doi.org/10.1016/j.specom.2005.04.007

Henderson, A. (2008). Towards intelligibility: Designing short pronunciation courses for advanced field experts. *Asp: La revue du Geras*, *53-54*, 89-110. https://doi.org/10.4000/asp.369

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Hincks, R. (2005). *Computer support for learners of spoken English* (Doctoral thesis, KTH Royal Institute of Technology, Stockholm, Sweden).

Hockett, C. F. (1955). *A manual of phonology*. Baltimore: Waverly Press.

Hockett, C. F. (1967*).* The quantification of functional load. *Word, 23,* 320–339. https://doi.org/10.1080/00437956.1967.11435484

Huensch, Amanda. (2016). Perceptual phonetic training improves production in larger discourse contexts. Journal of Second Language Pronunciation, 2,183-207. https://doi.org/10.1075/jslp.2.2.03hue

Huensch, A., & Tremblay, A. (2015**).** Effects of perceptual phonetic training on the perception and production of second language syllable structure. *Journal of Phonetics, 52*, 105–120. https://doi.org/10.1016/j.wocn.2015.06.007

Iino, A. (2019). Effects of HVPT on perception and production of English fricatives by Japanese learners of English. In F. Meunier, J. Van de Vyver, L. Bradley & S. Thouësny (Eds), *CALL and complexity – short papers from EUROCALL 2019,*186-192. https://doi.org/10.14705/rpnet.2019.38.1007

Iino, A., & Thomson, R. I. (2018). Effects of web-based HVPT on EFL learners' recognition and production of L2 sounds. In P. Taalas, J. Jalkanen, L. Bradley, & S. Thouësny (Eds.), *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018* (pp. 106–111). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.26.821

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*, 135-159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition,* 34(3), 475–505. https://doi.org/10.1017/S0272263112000150

Isaacs, T., & Trofimovich, P. (Eds.). (2017). *Second language pronunciation assessment: Interdisciplinary perspectives.* Bristol: Multilingual Matters. https://doi.org/10.21832/ISAACS6848

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018).Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*(2),193-216. https://doi.org/10.1177/0265532217703433

Iverson, P., & Evans, B. (2009). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *Journal of the Acoustical Society of America, 126*(2), 866–881. https://doi.org/10.1121/1.3148196

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America, 118*(5), 3267–3278. https://doi.org/10.1121/1.2062307

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics, 33*(1), 145–160. https://doi.org/10.1017/S0142716411000300

Jakobson, R. (1931). Prinzipien der historischen Phonologie. *Travaux du cercle linguistique de Prague*, *4*, 246–267.

Jenkins, J. (2000). *The phonology of English as an international language: New models, new norms, new goals.* Oxford, UK: Oxford University Press.

Jenkins, J. (2003). *World Englishes: A resource book for students*. London: Routledge.

Jenkins, J. (2007). *English as a Lingua Franca: Attitude and identity*. Oxford: Oxford University Press.

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. *The Journal of the Acoustical Society of America*, *119*(2), 1118–1130. https://doi.org/10.1121/1.2151806

Jułkowska, I. A., & Cebrian, J. (2015). Effects of listener factors and stimulus properties on the intelligibility, comprehensibility and accentedness of L2 speech. *Journal of Second Language Pronunciation, 1*, 211-237. https://doi.org/10.1075/jslp.1.2.04jul

Kang, O. (2013). Impact of rater characteristics on ratings of international teaching assistant's oral performance. *Language Assessment Quarterly, 9*(3), 249–269. https://doi.org/10.1080/15434303.2011.642631

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly, 48*(1), 176–206. https://doi.org/10.1002/tesq.152

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English. *The Modern Language Journal, 94*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning, 68*(1), *115–146.* https://doi.org/10.1111/lang.12270

Kang, O., Thomson, R. I., & Moran, M. (2020). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics, 41*(4), 453-480. https://doi.org/10.1093/applin/amy053

Kangatharan, J., Giannakopoulou, A., & Uther, M. (2020). *Intelligibility of speech improves after perceptual vowel training in L2 learners of English* [Poster presentation]. *In Proceedings of the 2nd Workshop on Speech Perception and Production across the Lifespan* (UCL, London, UK).

Kennedy, S., & Trofimovich, P. (2010). Language awareness and second language pronunciation: A classroom study. *Language Awareness, 19*(3), 171–185. https://doi.org/10.1080/09658416.2010.486439

Kenworthy, J. (1987). *Teaching English pronunciation*. London: Longman.

Kewley-Port, D. (2007). Vowels and speech intelligibility. *Journal of The Acoustical Society of America, 122*(5). https://doi.org/10.1121/1.2942774

King, R. (1967). Functional load and sound change. *Language, 43*(4), 831–852. https://doi.org/10.2307/411969

Kissling, E. M. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *The Modern Language Journal*, 97(3), 720–744. https://doi.org/10.1111/j.1540-4781.2013.12029.x

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). Philosophical Transactions of the Royal Society B: Biological Sciences, 363, 979–1000. https://doi.org/10.1098/rstb.2007.2154

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers.* University of Michigan Press.

Lai, Y., Saab, N. & Admiraal, W. (2022). Learning Strategies in Self-directed Language Learning Using Mobile Technology in Higher Education: A Systematic Scoping Review. *Educ Inf Technol* **27**, 7749–7780. https://doi.org/10.1007/s10639-022-10945-5

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(2), 227-247. https://doi.org/10.1017/S0142716405050150

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford, UK: Oxford University Press.

Hwang, H., & Lee, H.-Y. (2016). Gradient of learnability in teaching English pronunciation to Korean learners. The Journal of the Acoustical Society of America, 139(4), 1859–1872. https://doi.org/10.1121/1.4945716

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, 36(3), 345–366. https://doi.org/10.1093/applin/amu040

Lengeris, A. (2018). Computer-based auditory training improves second-language vowel production in spontaneous speech. *The Journal of the Acoustical Society of America*, *144*(3). https://doi.org/10.1121/1.5052201

Levelt, W. J. M. (1989). Speaking: From intention to articulation. *Cambridge*, MA: MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

Levelt, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences*, 3(6), 223–232. https://doi.org/10.1016/S1364-6613(99)01319-4

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369–377. https://doi.org/10.2307/3588485

Levis, J., Le, H., Lucic, I., Simpson, E., & Vo, S. (Eds.). (2016). *Proceedings of the 7th Annual Pronunciation in Second Language Learning and Teaching Conference* (ISSN 2380-9566). Iowa State University.

Li, Feng. (2016). Contrastive Study between Pronunciation Chinese L1 and English L2 from the Perspective of Interference Based on Observations in Genuine Teaching Contexts. *English Language Teaching*, *9*(10), 90-100. https://doi.org/10.5539/elt.v9n10p90

Li, Y. (2015). *The effect of audiovisual perception training on L2 speech perception and production by L1 Mandarin learners of English* (Doctoral dissertation, University of York). University of York Research Database. https://doi.org/10.5430/ijelt.v3n2p14

Lin, I. (2019). Functional load and L2 perception: An investigation of cross-linguistic effects with English and French learners. [*Doctoral dissertation, University of California, Los Angeles].*

Lin, Y. (2001). Syllable simplification strategies: A stylistic perspective. Language Learning, 51, 681-718. https://doi.org/10.1111/0023-8333.00171

Lin, Y. (2003). Interphonology variability: Sociolinguistic factors affecting L2 simplification strategies. Applied Linguistics, 24, 439–464. https://doi.org/10.1093/applin/24.4.439

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Kluwer. https://doi.org/10.1007/978-94-009-2037-8_16

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*(3), 1242-1255. https://doi.org/10.1121/1.408177

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y. I., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America, 94*(3), 2076-2087. https://doi.org/10.1121/1.410149

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/and/l/: A first report. *The Journal of the Acoustical Society of America*, 89(2), 874–886. https://doi.org/10.1121/1.1894649

Logan, J. S., & Pruitt, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross Language Research* (pp. 351-378). Timonium, MD: York Press

Low, E. L. (2006). Rhythm in the Singapore English speech of undergraduates. In D. Deterding, A. Brown, & E. L. Low (Eds.), *English in Singapore: Phonetic Research on a Corpus* (pp. 59–66). Singapore: McGraw Hill.

Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach.* John Benjamins. https://doi.org/10.1075/lllt.18

MacKain, K., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied PsychoLinguistics*, 2, 369–390. https://doi.org/10.1017/S0142716400009796

Martinet, A. (1952). Function, structure, and sound change. *Word*, 8(1), 1–32. https://doi.org/10.1080/00437956.1952.11659416

Martinet, A. (1955*). Économie des changements phonétiques*. Berne: Francke.

McAndrews, M. M., & Thomson, R. I. (2017). Establishing an empirical basis for priorities in pronunciation teaching. *Journal of Second Language Pronunciation, 3*(2), 267–287. https://doi.org/10.1075/jslp.3.2.05mca

Melnik-Leroy, K., Saito, K., & Baese-Berk, M. (2022). The perception-production link in second language speech: A systematic review and meta-analysis. *Studies in Second Language Acquisition*. Advance online publication.

Mifka-Profozic, N., O'Reilly, D., & Guo, J. (2020). Sensitivity to syntactic violation and semantic ambiguity in English modal verbs: A self-paced reading study. Applied Psycholinguistics, 41, 1017-1043. https://doi.org/10.1017/S0142716420000338

Mora, J. C., & Fullana, N. (2007). Production and perception of English /h9/-/H/ and /æ/-/⍰/ in a formal setting: Investigating the effects of experience and starting age. *Proceedings of the 16th International Congress of Phonetic Sciences,* 1613-1616. Saarbrücken, Germany.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M., & Derwing, T. (1995a). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. Language and Speech, 38, 289–306. https://doi.org/10.1177/002383099503800305

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. Studies in Second Language Acquisition, 23(2), 451–468. https://doi.org/10.1017/S0272263101004016

Munro, J. M., & Derwing, M. T. (2006). The functional load principal in ESL pronunciation instruction: An exploratory study. *Science Direct*, 34, 520-531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. Language Learning, 58(3), 479–502. https://doi.org/10.1111/j.1467-9922.2008.00448.x

Nagle, C. L. (2017). Examining the temporal structure of the perception–production link in second language acquisition: A longitudinal study. *Language Learning, 68*(1), 234–270. https://doi.org/10.1111/lang.12275

Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *Modern Language Journal,* 102(1), 199–217. https://doi.org/10.1111/modl.12461

Nilsen, D. L., & Nilsen, A. P. (2010). Pronunciation contrasts in English (2nd ed.). Waveland Press.

Nishi, K., & Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research, 50(6),* 1496–1509. https://doi.org/10.1044/1092-4388(2007/103)

Nishi, K., & Kewley-Port, D. (2008). Nonnative speech perception training using vowel subsets: Effects of vowels in sets and order of training. Journal of Speech, Language, and Hearing Research, 51(6), 1480–1493.

Oh, Y. M., Coupé, C., Marsico, E., and Pellegrino, F. (2015). Bridging Phonological System and Lexicon: Insights from a Corpus Study of Functional Load. J. *Phonetics* 53, 153–176. https://doi.org/10.1016/j.wocn.2015.08.003

Olson, D. J. (2014). Benefits of visual feedback and perceptual training for L2 vowel acquisition. *Language Learning & Technology*, 18(3), 173–192. https://doi.org/10.64152/10125/44389

Pattamadilok, C., Welby, P., & Tyler, M. D. (2022). The contribution of visual articulatory gestures and orthography to speech processing: Evidence from novel word learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 48(10), 1542–1558. https://doi.org/10.1037/xlm0001036

Pennington, M. C. (1999). Computer-aided pronunciation pedagogy: Promise, limitations, directions. Computer Assisted Language Learning, 12(5), 427–440. https://doi.org/10.1076/call.12.5.427.5693

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America, 130*(1), 461–472. https://doi.org/10.1121/1.3593366

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly, 35*(2), 233–255. https://doi.org/10.2307/3587647

Pickering, Lucy. (2006). Current research on intelligibility in English as a lingua franca. *Annual Review of Applied Linguistics. 26.* 219 - 233. https://doi.org/10.1017/S0267190506000110

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*, 137–157. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.45.08pie

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. Studies in Second Language Acquisition, 39(3), 579-592. https://doi.org/10.1017/S0272263116000231

Pruitt, J. S., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *Journal of the Acoustical Society of America, 119*(3), 1684-1696. https://doi.org/10.1121/1.2161427

Qian, M., Chukharev-Hudilainen, E., & Levis, J. (2018). A system for adaptive high- variability segmental perceptual training: Implementation, effectiveness, transfer. *Language Learning & Technology*, *22*(1), 69–96. https://doi.org/10.64152/10125/44582

Qian, M. (2018). *An adaptive computational system for automated, learner-customized segmental perception training in words and sentences: Design, implementation, assessment* (Doctoral dissertation, Iowa State University). Iowa State University Digital Repository. https://lib.dr.iastate.edu/etd/17293

Rato, A., & Rauber, A. (2015). The effects of perceptual training on the production of English vowel contrasts by Portuguese learners. In the Scottish Consortium for ICPhS 2015

(Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences. Paper number 656*. Glasgow, UK: Glasgow University.

Revesz, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. Applied Linguistics, 37(6), 828-848. https://doi.org/10.1093/applin/amu069

Rogers, C. L. (1997). Intelligibility of Chinese-accented English (Unpublished doctoral dissertation). Indiana University, Bloomington.

Rogers, C. L., & Dalby, J. (2005). Forced choice analysis of segmental production by Chinese-accented English speakers. *Journal of Speech, Language, and Hearing Research, 48*(2), 306-322. https://doi.org/10.1044/1092-4388(2005/021)

Rogerson-Revell, P. (2011). *English phonology and pronunciation teaching*. London: Bloomsbury. https://doi.org/10.5040/9781350934177

Rogerson-Revell, P. (2021). Computer-assisted pronunciation training (CAPT): Current issues and future directions. *RELC Journal*, 52(1), 189–211. https://doi.org/10.1177/0033688220977406

Rönnberg, J., Holmer, E., & Rudner, M. (2019). Cognitive hearing science and ease of language understanding. International Journal of Audiology, 58(5), 247–261. https://doi.org/10.1080/14992027.2018.1551631

Saito, K. (2011). Examining the role of explicit phonetic instruction in native-like and comprehensible pronunciation development: An instructed SLA approach to L2 phonology. *Language Awareness*, 20(1), 45–59. https://doi.org/10.1080/09658416.2010.540326

Saito, K. (2013). Effects of high variability phonetic training on L2 pronunciation: A study of Japanese learners' acquisition of English /ɹ/. Language Learning, 63(3), 499-531. https://doi.org/10.1111/lang.12015

Saito, K. (2015). The role of age of acquisition in late second language oral proficiency attainment. *Studies in Second Language Acquisition, 37*(4), 713–743. https://doi.org/10.1017/S0272263115000248

Saito, K. (2020). Multi- or Single-Word Units? The Role of Collocation Use in Comprehensible and Contextually Appropriate Second Language Speech. Language Learning. *A Journal of Research in Language Studies*, 70(2), 548-588. https://doi.org/10.1111/lang.12387

Saito, K. (2021). What characterizes comprehensible and nativelike pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, 55(1), 1-30. https://doi.org/10.1002/tesq.3027

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language Japanese speech. Journal of Second Language Pronunciation, 3(2), 199–217. https://doi.org/10.1075/jslp.3.2.02sai

Saito, K., & Hanzawa, K. (2016). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. Applied Psycholinguistics, 37(5), 813-840. https://doi.org/10.1017/S0142716415000259

Saito, K., & Hanzawa, K. (2018). The role of input in second language oral ability development in foreign language classrooms: A longitudinal study. Language Teaching Research, 22(3), 398-417. https://doi.org/10.1177/1362168816679030

Saito, K., Hanzawa, K., Petrova, K., Kachlicka, M., Suzukida, Y., & Tierney, A. (2022b). Incidental and multimodal high variability phonetic training: Potential, limits, and future directions. Language Learning, 72(4), 1049–1091. https://doi.org/10.1111/lang.12503

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T., & Minematsu, N. (2022a). Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, 1–30. https://doi.org/10.1017/S0272263122000080

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning, 69*(3), 652–708. https://doi.org/10.1111/lang.12345

Saito, K., Sun, H., Magne, V., & Ilkan, M. (2020). Perception- vs. production-based pronunciation instruction: *An investigation into the relative effects for inexperienced Japanese EFL learners.* Unpublished manuscript. https://doi.org/10.1177/2041669520958430

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. Studies in Second Language Acquisition, 41(5), 1133–1149. https://doi.org/10.1017/S0272263119000226

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2018). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. Applied Psycholinguistics, 37(2), 217–240. https://doi.org/10.1017/S0142716414000502

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics,* 39(1), 187–224. https://doi.org/10.1017/S0142716417000418

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition,* 14(4), 357–385. https://doi.org/10.1017/S0272263100011189

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), Cognition and second language instruction (pp. 3–32). Cambridge: *Cambridge University Press.* https://doi.org/10.1017/CBO9781139524780.003

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge. https://doi.org/10.4324/9780203851357

Sewell, A. (2017). Functional load revisited: Reinterpreting the findings of 'lingua franca' intelligibility studies. *Journal of Second Language Pronunciation, 3*(1), 57–79. https://doi.org/10.1075/jslp.3.1.03sew

Sewell, A. (2021). Functional load and the prioritization of pronunciation targets. *Frontiers in Communication, 6,* Article 627378. https://doi.org/10.3389/fcomm.2021.627378

Sheppard, C., Elliott, N., & Baese-Berk, M. (2017). The intelligibility of Chinese-accented English: A multi-listener, multi-talker investigation. *Journal of Second Language Pronunciation, 3*(1), 1–26. https://doi.org/10.1016/j.jeap.2017.01.006

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/*. Journal of Phonetics, 66*, 242–251. https://doi.org/10.1016/j.wocn.2017.11.002

Smith, L. E., & Nelson, C. L. (1985). International intelligibility of English: Directions and resources. *World Englishes*, 4(3), 333–342. https://doi.org/10.1111/j.1467-971X.1985.tb00423.x

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model*. Journal of Phonetics*, 39(4), 456–466. https://doi.org/10.1016/j.wocn.2010.09.001

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics, 36*(2), 131-145. https://doi.org/10.3758/BF03202673

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners: The re-education of selective perception*. In J. G. Hansen Edwards & M. L. Zampini (Eds.), Phonology and second language acquisition* (pp. 153-192). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/sibil.36.09str

Surendran, D., & Niyogi, P. (2003). Measuring the functional load of phonological contrasts. *In Proceedings of the 15th International Congress of Phonetic Sciences* (ICPhS 2003) (pp. 2117–2120). Barcelona.

Surendran, D., and Niyogi, P. (2006). “Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals,” in Competing Models of Language Change: Evolution and Beyond, edited by O. Nedergaard Thomsen (John Benjamins, Amsterdam, the Netherlands), pp. 43–58. https://doi.org/10.1075/cilt.279.05sur

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, *25*(3), 431–450. https://doi.org/10.1177/1362168819858246

Swan, M., & Smith, B. (2001). Learner English: A teacher’s guide to interference. Cambridge University Press. https://doi.org/10.1017/CBO9780511667121

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Takagi, N. (1993). *Perception of American English /r/ and /l/ by adult Japanese learners of English: A unified view.* (Doctoral dissertation). University of California.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics*, 25, 1–24. https://doi.org/10.1006/jpho.1996.0031

Tavakoli, P., & Cooke, S. (2024). Comprehensibility in language assessment: A broader perspective. Equinox Publishing Ltd. https://doi.org/10.3138/9781800504349

Thomson, R. I. (2011). Computer assisted pronunciation Training: Targeting second language vowels: Perception improves pronunciation, *CALICO Journal 28*(3): 744–65. https://doi.org/10.11139/cj.28.3.744-765

Thomson, R. I. (2018). High Variability [Pronunciation] Training (HVPT): A proven technique about which every language teacher and learner ought to know, Journal *of Second Language Pronunciation,* 4(2), pp. 208–23. https://doi.org/10.1075/jslp.17038.tho

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, 1–20. https://doi.org/10.1093/applin/amu076

Thomson, R. I., & Derwing, T. M. (2016). Is phonemic training using nonsense or real words more effective? In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds.), Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference (pp. 88-89), Ames, IA: Iowa State University.

Thomson, R. I., Nearey, T. M., & Derwing, T. M. (2009). A modified statistical pattern recognition approach to measuring the crosslinguistic similarity of Mandarin and English vowels. *The Journal of the Acoustical Society of America, 126*(3), 1447–1460. https://doi.org/10.1121/1.3177260

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Harvard University Press.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. Bilingualism: Language and Cognition, 15, 905–916. https://doi.org/10.1017/S1366728912000168

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck & Ruprecht.

Tyler, M. D. (2019). PAM-L2 and the Acquisition of L2 Speech Perception. In O.-S. Bohn & M. J. Munro (Eds.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge: Cambridge University Press.

Tyler, M., Pattamadilok, C., & Welby, P. (2022). Visual information and speech processing: Effects of orthographic and articulatory gestures on perceptual learning and consolidation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 48(9), 1543–1555. https://doi.org/10.1037/xlm0001036

Uchihara, T., Karas, M., & Thomson, R. I. (2024). Does perceptual high variability phonetic training improve L2 speech production? A meta-analysis of perception-production connection. *Applied Psycholinguistics*, *45*(4), 591-623. https://doi.org/10.1017/S0142716424000195

van Teijlingen, E., & Hundley, V. (2002). The importance of pilot studies. Social Research Update, 35. https://doi.org/10.7748/ns2002.06.16.40.33.c3214

Walker, R. (2010). Teaching the Pronunciation of English as a Lingua Franca. Oxford University Press.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, 113(2), 1033–1043. https://doi.org/10.1121/1.1531176

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System, 32*(4), 539–552. https://doi.org/10.1016/j.system.2004.09.011

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, *106*, 3649-3658. https://doi.org/10.1121/1.428217

Wardhaugh, R. (1970). The contrastive analysis hypothesis. *TESOL Quarterly, 4*(2), 123-130. https://doi.org/10.2307/3586182

Wedel, A., Jackson, S., & Kaplan, A. (2013). Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech*, 56(3), 395–417. https://doi.org/10.1177/0023830913489096

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

Wong, J. W. S. (2012). The effects of high and low variability phonetic training on the perception and production of English vowels /e/-/æ/ by Cantonese ESL learners with high and low L2 proficiency levels. In *Proceedings of the 14th Australasian International Conference on Speech Science and Technology* (pp. 524-528). Singapore: ISCA. https://doi.org/10.21437/Interspeech.2014-129

Wright, C. E. M., & Tavakoli, P. (2016). New directions and developments in defining, analyzing and measuring L2 speech fluency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 73–77. https://doi.org/10.1515/iral-2016-9990

Yeon, S. H. (2008). Training English word-final palatals to Korean speakers of English. Applied Language Learning, 18 (1-2), 51-61. https://eric.ed.gov/?id=EJ811066

Yorkston, K. M., Strand, E. A., & Kennedy, M. R. T. (1996). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. American Journal of Speech Language Pathology, 5, 55-66. https://doi.org/10.1044/1058-0360.0501.55

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics,* 24(1), 1–27. https://doi.org/10.1093/applin/24.1.1

Zhang, X., Cheng, B., Qin, D., & Zhang, Y. (2021). Is talker variability a critical component of effective phonetic training for nonnative speech? *Journal of Phonetics*, 87, 101071. https://doi.org/10.1016/j.wocn.2021.101071

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Stevens, E. B., ... & Nemoto, I. (2009). Neural signatures of phonetic learning in adulthood: A magnetoencephalography study. *NeuroImage, 46*(1), 226–240. https://doi.org/10.1016/j.neuroimage.2009.01.028

Zhang, R., & Yuan, Z. (2020). Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition*, 1-14. https://doi.org/10.1017/S0272263120000121

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System, 36*, 69–84. https://doi.org/10.1016/j.system.2007.11.004