Natural versus concatenative speech in high-variability phonetic training: A pilot

Abstract

In the quest to develop effective speech perception training, research has increasingly reported positive results with high-variability phonetic training (HVPT). Presenting utterances in different phonetic environments and by multiple speakers, HVPT provides a more representative input to map onto developing L2 sound systems. While there is much success surrounding HVPT, we still do not yet know what components of HVPT help second language learners. One characteristic that plays a major role in HVPT is the speech input, such as using “natural,” human voices versus synthetic voices. This study aimed to explore possible effects on segmental perception through high-variability training with either human voices or concatenative synthetic voices. Six non-native English-speaking participants trained with the /ɪ/-/ɨ/ contrast in one of the conditions. The study followed a pretest-training-posttest design and included one training session in between. The learning management system Canvas was used to display all materials. While this study highlighted the need to examine HVPT, it did not find conclusive results. Further studies will be needed to explore HVPT and how to maximize its gains.

Introduction

• High Variability Phonetic Training (HVPT) = Perceptual training that uses stimuli in different phonetic environments and by multiple speakers (high-variability) and provides feedback (Logan et al., 1991; Lively et al., 1993, 1994)
• High variability provides a more representative input to map onto developing L2 sound systems
• HVPT stimuli have used both human and Text-to-Speech (TTS) voices—both of which have shown perceptual improvement (Logan et al., 1991; Qian et al., 2018)
• However, little has been done to examine the voices used
• Particularly, whether human and TTS inputs are equal in effectiveness is unknown

RESEARCH QUESTION

After completing high-variability perceptual training for the /ɪ/-/ɨ/ contrast, does a natural speech input provide more gain in non-native learners of English’s perception performance than a concatenative synthetic input?

Methodology

Participants
• 6 non-native English speakers in the Applied Linguistics Graduate Program at a large midwestern university

Stimuli
• Human voice condition:
  • Trained with recordings of two native American English speakers, 1 male and 1 female (Praat)
• Concatenative TTS condition:
  • Trained with two TTS-generated voices, 1 male and 1 female (Amazon Polly)
  • Minimal pair contrast: /ɪ/ and /ɨ/
  • In /b/-/t/, /h/-/t/, and /n/-/t/ environments

Procedure
• Pretest-Training-Posttest design
• Identification task
• Immediate feedback with training

Statistical Analysis
• Measured “gain” through differences between posttest and pretest scores
• Between-subjects 1-t-test

Results
• TTS-generated voice condition produced greater mean identification scores than the natural voice condition
• Both groups either improved or maintained their initial identification performances of the /ɪ/-/ɨ/ contrast on the posttest

Discussion and Conclusions

• TTS voices > Human voices?
  • Unlikely – In first language acquisition, babies need human input
  • Results have low reliability
  • Extremely small sample size
  • No control group - although many similar studies justify not needing one
  • Participants were not initially tested on whether they have difficulty with /ɪ/-/ɨ/ contrast
  • Only one training session
  • /ɪ/-/ɨ/ contrast “trouble-spots” = systematic
  • Phonetic environment played a bigger role for the contrast rather than contrast itself

Implications
• Learning management systems lend themselves as an affordance to computer-assisted phonetic training
• Convenient, easy set-up, easy grading, platform familiarity
• Integrating pronunciation research into pedagogical practice
• Raising linguistic awareness