Incremental cue training: a study of lexical tone learning by non-tonal listeners

Yanyu Li, Laurence White & Ghada Khattab

Newcastle University
y.li218@newcastle.ac.uk, laurence.white@newcastle.ac.uk, ghada.khattab@newcastle.ac.uk

ABSTRACT

Speech perception requires sensitivity to acoustic dimensions relevant to meaningful distinctions. Training with exaggerated cues has been used to direct second language (L2) learners' attention to dimensions relevant for segmental contrasts; here we applied similar incremental cue training to lexical tone contrasts. Two groups of native English listeners were trained, either on a fixed tonal contrast (*high-rising* vs *high-falling*) or with incremental stimuli, where pitch movements relevant to the rising/falling contrast were exaggerated at training outset and reduced, stepwise, to the fixed contrast level by training completion. An ABX task tested discrimination, with feedback during training. Final tonal discrimination accuracy was similar between groups, but prior learning trajectories contrasted: the Incremental group had higher initial performance, i.e., with the exaggerated contrasts, whilst the Fixed group improved gradually over training. Finally, despite contrasting training, the groups had comparable pre-test to post-test improvement on a separate untrained tonal contrast (*high-level* vs *mid-rising*).

Keywords: lexical tone, perception, incremental cue training, second language acquisition

1. INTRODUCTION

Speech perception is shaped by early exposure to first languages (L1s). Attention-to-dimension models, such as the Native Language Magnet Model (NLM) [1, 2], describe the perceptual system as a distorted space where speech sound discrimination is based on perceived acoustic distances. Critical dimensions for sound discrimination are "stretched", leading to increased perceptual sensitivity, while the less relevant dimensions are "compressed", resulting in reduced sensitivity [3–8]. Perceptual distance along critical dimensions is additionally warped for efficient categorisation: distance within category prototypes is compressed, while distance at category boundaries is stretched.

Such adaptation to L1 exposure leads to efficient L1 perception, but potentially interferes with non-native perception (what we call here "L2 perceptual bias"). Thus, non-native listeners behave differently from L1 listeners in phonetic discrimination. For instance, Japanese L1 listeners tend to confuse English /l/ and /r/ [9, 10]. Iverson et al. [3] found that in perceiving the contrast, Japanese listeners relied more on F2, which was a less important dimension for English L1 listeners. Similarly, Spanish L1 listeners' perception of English /i/-/ɪ/ is influenced more by vowel duration than spectral features [11, 12]. L2 perceptual bias is also seen in perception of suprasegmental features, such as lexical tone. For instance, in lexical tone perception, English L1 listeners rely more on pitch height, while Thai and Mandarin L1 listeners are differentially sensitive to pitch contour [13–15].

Despite L2 perceptual bias, the adult listener's perceptual space is not invariable. Francis et al. [16] trained Mandarin and English L1 listeners on Cantonese tone perception. Multi-dimensional scaling (MDS) showed that L1 listeners of both languages increased sensitivity to previously less-attended dimensions. Mandarin listeners improved discrimination on two Cantonese *rising* tones 23 and 25 (as represented in Chao digits [17]), with common pitch trajectory, but differing pitch height. English listeners improved discrimination between Cantonese 23 and 33, which contrast in pitch trajectory but are similar in pitch height. Likewise, Chandrasekaran et al. [13] trained English L1 listeners on Mandarin tones which contrasted mainly on pitch contour. MDS indicated that listeners improved discrimination based on pitch trajectory, but showed no training benefit for pitch height-based discrimination.

As phonetic learning seems to be strongly influenced by perceptual attention [13, 16], training techniques based on attention-to-dimension theories – including adaptive cue training and inhibition training – have been tested for segmental learning [11, 19, 20]. There is evidence of listener discrimination improvement with these techniques, whilst their relative benefits are still debated [11, 19]. Adaptive or incremental cue training, for example, aims to redirect listeners' attention to previously less attended dimensions to improve phonetic perception. Training expands critical acoustic dimensions to boost phonetic discrimination at the beginning of training. This enhancement gradually reduces to a natural range by the end of training.

Whilst there is evidence that the tonal perceptual space can be modified by adult experience [13, 16], the efficacy of incremental cue training for L2 lexical tone learning has not been tested. Therefore, the present study examines whether the initial use of exaggerated pitch cues to lexical tone contrasts improves inexperienced English listeners' tonal discrimination, compared to a training condition using non-exaggerated cues.

2. METHOD

2.1. Participants

We tested 90 adult native English-speaking listeners (F=74, M=16). Only the 86 participants without tonal language experience were included in the analyses. These 86 participants were randomly assigned to the Incremental (N=40) and Fixed groups (N=46).

2.2. Stimuli

The stimuli were 36 stylised tonal syllables, combining 4 pseudo-tones and 9 syllables. The four synthesised tones shared the same onset pitch height, to encourage listeners' attention to pitch contour. The four tones were *high-rising* 45, *high-falling* 41, *level* 44, and *falling-rising* 425 (Figure 1).

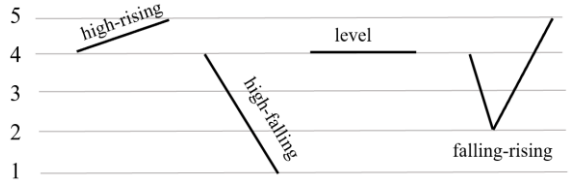

Figure 1: The pseudo-tone system used as the basis for the experimental stimuli.

These tones were presented in nine consonant-vowel (CV) syllables consisting of three voiceless aspirated plosives /p/, /t/, /k/ and three open long vowels /a/, /i/, /u/. All syllable combinations obeyed Mandarin and English phonotactic constraints.

2.2.1. Recording

Two female Mandarin L1 speakers recorded the nine syllables with four Mandarin tones (*high-level* 55, *mid-rising* 35, *falling-rising* 214, *high-falling* 51) in a sound-attenuated booth. Each syllable was embedded in a Mandarin sentence, e.g., (English translation) "I read '/paa/ 35' three times". All sentences were presented three times on a screen. Speakers were asked to read sentences aloud in a normal, even pace.

2.2.2. Pitch manipulation

Extraction. For each speaker, one example of each token was selected from the three recorded repetitions. The target tonal syllables were then extracted from the carrier utterances using Praat [21].
Manipulation. The recorded syllables were overlaid with the four pseudo-tones via Pitch-Synchronous Overlap and Add (PSOLA) [22, 23] in Praat. The PSOLA manipulation is intended to change F0 trajectories without affecting other acoustic features.

The endpoints of the pseudo-tones, as represented in Chao digits, were converted into Hertz values according to each speaker's fundamental frequency (F0) range, based on the formula "T = [(lg X - lg L) / lg H - lg L] * 5" [24], where H and L are each speaker's highest and lowest F0 points and X is the F0 point to be converted. The converted F0 values for the onset and offset of each tone are shown Table 1, also the turn point for the falling-rising tone.

Pseudo-tone (Chao digits)	Speaker 1 (Hz)	Speaker 2 (Hz)
High-rising **45**	297-375	294-335
High-falling **41**	297-147	294-201
Level **44**	297-297	294-294
Falling-rising **425**	297-240-375	294-260-335

Table 1: F0 values (Hz) of the four pseudo-tones.

The extracted Mandarin syllables were manipulated to the pseudo-tone pitch values (Table 1). Mandarin tone *mid-rising* 35 was converted to pseudo-tones *high-rising* 45 and *falling-rising* 425, Mandarin *high-falling* 51 to pseudo-tone *high-falling* 41, and Mandarin *high-level* 55 to pseudo-tone *level* 44. The natural *falling-rising* tone 214 was not used in the conversion, due to phrase-medial realisation as 21, as well as creakiness and other phonetic confounds.
Test stimuli. The contrasting tonal pair *level* 44 and *falling-rising* 425 (Table 1) were used in pre- and post-training tests.
Training stimuli. The contrasting tonal pair *high-rising* 45 and *high-falling* 41 were used in training blocks. For the incremental group, the F0 endpoints of the tones were manipulated to enhance the difference between *rising* and *falling*. F0 offset values for *high-rising* 45 were raised by two 30Hz steps. Likewise, F0 offsets for *high-falling* 41 was lowered by two 30Hz steps. Table 2 shows the F0 values for each level of enhancement for each tone.

2.3. Procedure

The experiment consisted of 5 blocks, taking about 30 minutes per participant. Blocks 1 and 5 were the pre-test and post-test, using the pseudo-tone contrast *level* 44 *vs falling-rising* 425. Blocks 2, 3, and 4 were training blocks, using the contrast *high-rising* 45 *vs high-falling* 41. Each testing or training block contained 36 ABX trials: 9 syllables x 2 tonal orders (AB/BA) x 2 test stimuli for each order (ABA/ABB).

In each ABX trial, participants heard two segmentally-identical syllables (spoken by the same speaker), one for each pseudo-tone, and then the same syllable (spoken by the other speaker), with one of the contrasting pseudo-tones (i.e., A or B). All trials were randomised within each block.

Pseudo-tone (Chao digits) by block	Speaker 1 (Hz)	Speaker 2 (Hz)
High-rising **45**		
Block 2	297-435	294-395
Block 3	297-405	294-365
Block 4 (baseline)	297-375	294-335
High-falling **41**		
Block 2	297-87	294-141
Block 3	297-117	294-171
Block 4 (baseline)	297-147	294-201

Table 2: Enhanced and baseline F0 values (Hz) for pseudo-tones 45 and 41 for the Incremental group. Block 2 has the most exaggerated contrasts; Block 4 is identical to the Fixed group baseline.

In Blocks 1 and 5, participants had up to 5 seconds to respond in each trial, to indicate whether the third sound heard was the same as the first or the second. No feedback was provided regarding the response.

In Blocks 2 to 4 (training), the task was the same, but the timeout was 10 seconds and visual feedback was given ("correct"/ "incorrect"). For the Fixed group, the stimuli were the unaltered (baseline) syllables for all blocks. For the Incremental group, the most exaggerated versions of the pseudo-tones were used in Block 2, intermediate versions in Block 3, and the baseline contrasts in Block 4 (Table 2).

3. RESULTS

Separate accuracy analyses were conducted for test and training blocks. In test blocks, as shown in Figure 2, there was an overall accuracy improvement from pre-test (Block 1) to post-test (Block 5) for both Incremental (67% to 76%) and Fixed groups (70% to 76%.). In training blocks, the Fixed group's accuracy improved from 69% to 75% from Block 2 to Block 4, whilst the Incremental group's accuracy dropped from 81% (Block 2) to 76% (Block 4). Thus, whilst the Incremental group performed numerically better at the start of training, with exaggerated stimuli, the Fixed group performed comparably at the end of training, when the two groups were exposed to the same (baseline) stimuli.

We analysed these accuracy patterns via logistic mixed-effect regression models, with Group (Incremental vs Fixed) and Block (coded as a numeric variable) as fixed effects. There was a random intercept for participants and a random slope for block-by-participant. A random intercept for syllable type was excluded because it did not improve model fit. Significance of predictors was determined by likelihood ratio tests.

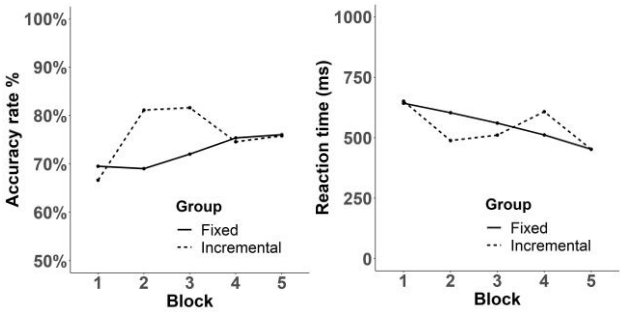

Figure 2: Tone discrimination accuracy and reaction time across the five blocks (Blocks 1 & 5: test-only; Blocks 2, 3 & 4: training) by group (Incremental vs Fixed)

3.1. Discrimination accuracy analyses

3.1.1. Test block accuracy

There was an effect of Block in the test-only blocks, $\chi^2(1) = 31.96$, $p < .001$: thus, there was an overall improvement in accuracy from pre-test to post-test (Figure 2 left: Block 1 *vs* Block 5) in distinguishing the pseudo-tone contrasts *level* and *falling-rising*. There was no effect of Group, $\chi^2(1) = 0.53$, $p = .47$, and no Group x Block interaction, $\chi^2(1) = 0.37$, $p = .54$.

3.1.2. Training block accuracy

In training blocks, Group had an effect, $\chi^2(1) = 20.44$, $p < .001$, indicating that the Incremental group performed better than the Fixed group overall. There was also an effect of Block, $\chi^2(2) = 8.69$, $p = .003$, and a Block x Group interaction, $\chi^2(2) = 20.10$, $p < .001$. Post-hoc pairwise analyses were conducted between every two blocks within each group and between groups for each block. Comparing Blocks 2 and 4, the Incremental group dropped accuracy, $\chi^2(1) = 10.07$, $p = .002$, whereas the Fixed group improved accuracy, $\chi^2(1) = 13.31$, $p < .001$. Looking at adjacent blocks, the Incremental group decreased accuracy between Blocks 3 and 4, $\chi^2(1) = 19.09$, $p < .001$, whilst the Fixed group improved accuracy between Blocks 2 and 3, $\chi^2(1) = 4.76$, $p = .03$. Comparing within blocks, the Incremental group were more accurate than the Fixed group in Block 2, $\chi^2(1) = 9.42$, p = .002, and Block 3, $\chi^2(1) = 5.80$, p = .02. Despite their contrasting learning trajectories, the two groups had comparable Block 4 discrimination accuracy, $\chi^2(1) = 0.10$, p = .75 (Figure 2 left: Blocks 2 – 4).

3.2. Reaction time analyses

Reaction times (RTs) on correct-response trials were analysed via linear mixed-effect regression models.

The fixed effects and random effects structure was as for the accuracy data. We eliminated RT outliers, via lower and upper thresholds for each participant based on individual means and standard deviations (SDs). Thus, RTs below or above 2.5SD from the mean were removed from further analyses.

3.2.1. Test block reaction times

RTs for both groups reduced from pre-test to post-test (Figure 2 right: Block 1 *vs* Block 5), $\chi^2(1) = 46.57$, $p < .001$. There was no effect of Group, $\chi^2(1) = 0.15$, $p = .70$, nor an interaction between Group and Block, $\chi^2(1) = 0.01$, $p = .92$.

3.2.2. Training reaction times

For training block RTs, there was no effect of Block, $\chi^2(2) = 0.54$, $p = .50$, or Group, $\chi^2(1) = 0.08$, $p = .77$, nor an interaction, $\chi^2(2) = 1.33$, $p = .25$.

4. DISCUSSION AND CONCLUSION

Overall, listeners' tonal discrimination accuracy improved with experience. Firstly, performance was better and reaction times were quicker on the untrained tonal contrast (*level* 44 vs *falling-rising* 425) in the final block (after training) than in the initial block. This improvement did not differ between Incremental and Fixed training groups.

For the trained tonal contrast (*high-rising* 45 *vs high-falling* 41), learning trajectories for the two training groups were in opposite directions. The Fixed group started with lower accuracy which increased over training. The Incremental group started with higher accuracy which decreased over training. The two groups had comparable accuracy at the end of training, i.e., when both were exposed to the same baseline tonal contrast (Block 4).

For the trained tonal contrast (Blocks 2-4), RTs did not show statistically-robust (overall or pairwise) variation. There was a non-significant numerical tendency for faster RTs in more accurate blocks, which at least runs counter to any possible speed/accuracy trade-off.

The overall training improvement, on both trained and untrained stimuli, is consistent with previous findings on lexical tone learning [13, 16, 25]. However, the higher accuracy for the Incremental group at training outset was not seen in other studies which also used enhanced stimuli at training outset [11, 19]. The initial performance benefit in the Incremental group appears to be due to the stimulus enhancement, where pitch trajectory differences between the pseudo-tonal contrasts were maximised and required less effort to distinguish. That this difference was observed here, but not in other similar studies, may relate to training design contingencies.

First, the study on training Spanish listeners on English /i/-/ɪ/ [11] used adaptive stimuli and found them to be more effective than fixed stimuli. One difference with our findings is in the training implementation. In Kondaurova and Francis [11], participants had to reach an accuracy threshold in consecutive sessions before stimulus enhancement was reduced. Potentially this adaptive manipulation is more sensitive to individuals' learning trajectories than our training regime, where the enhancement reduced automatically in successive training blocks.

Second, in our study, lack of Block 2 to 3 accuracy fall-off for the Incremental group suggests that we used relatively salient levels of cue enhancement in both blocks, clearly above the just-noticeable-difference for tonal contours [26]. Starting with a less exaggerated contrast tailored to provide a reliable, but minimal discrimination boost – along with more, but smaller steps of decreasing enhancement – may help to maintain higher performance back to baseline.

Third, studies of L2 segmental learning using enhanced stimuli [11] have typically implemented training for 4 to 10 days [11, 19]; where found, the boost to learning manifested as improvement over multiple sessions [11]. Insofar as accuracy is high in early training blocks, our findings support an overall beneficial effect of enhanced stimuli, which may be sustained with extended training, although we note that Iverson et al. [19] found no ultimate benefit for initially-enhanced stimuli over multiple training sessions. The nature of the learned L2 contrast, the degree of enhancement and the length/frequency of training are all likely to influence relative outcomes.

Finally, there were no robust response latency differences between the Incremental and Fixed groups (notwithstanding the numerically contrasting patterns across training blocks, see Figure 2). These null results were comparable to [11], which used an adaptive training method over a longer period. This reinforces our accuracy data conclusions regarding the similar final impact, over one session, of Incremental and Fixed stimuli. Further work will examine how stimulus enhancement affects reaction time over extended training.

In summary, F0 exaggeration of training stimuli initially boosted discrimination of lexical tones. The improvement in performance for this Incremental group, relative to the Fixed group (who were presented with non-exaggerated stimuli throughout), was, however, not sustained when stimulus exaggeration was eliminated in the final training block. Discrimination of untrained tonal contrasts improved equally for the Fixed and Incremental training groups.

5. ACKNOWLEDGEMENTS

This study is supported by the Chinese Scholarship Council (CSC).

6. REFERENCES

[1] P. K. Kuhl, B. T. Conboy, S. Coffey-Corina, D. Padden, M. Rivera-Gaxiola, and T. Nelson, 'Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e)', *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 363, no. 1493, pp. 979–1000, Mar. 2008, doi: 10.1098/rstb.2007.2154.

[2] P. K. Kuhl and P. Iverson, 'Linguistic experience and the "perceptual magnet effect"', *Speech Percept. Linguist. Exp. Issues Cross-Lang. Res.*, pp. 121–154, 1995.

[3] P. Iverson *et al.*, 'A perceptual interference account of acquisition difficulties for non-native phonemes', *Cognition*, vol. 87, no. 1, pp. B47–B57, Feb. 2003, doi: 10.1016/S0010-0277(02)00198-1.

[4] A. L. Francis, N. Kaganovich, and C. Driscoll-Huber, 'Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English', *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1234–1251, Aug. 2008, doi: 10.1121/1.2945161.

[5] P. W. Jusczyk, 'From general to language-specific capacities: the WRAPSA Model of how speech perception develops', *J. Phon.*, vol. 21, no. 1, pp. 3–28, Jan. 1993, doi: 10.1016/S0095-4470(19)31319-1.

[6] R. M. Nosofsky, 'Attention, Similarity, and the Identification-Categorization Relationship', p. 19, 1986.

[7] R. L. Goldstone, 'Feature distribution and biased estimation of visual displays.', *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 19, no. 3, p. 564, 1993.

[8] R. L. Goldstone, 'Influences of categorization on perceptual discrimination.', *J. Exp. Psychol. Gen.*, vol. 123, no. 2, p. 178, 1994.

[9] K. Aoyama, J. E. Flege, S. G. Guion, R. Akahane-Yamada, and T. Yamada, 'Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/', *J. Phon.*, vol. 32, no. 2, pp. 233–250, Apr. 2004, doi: 10.1016/S0095-4470(03)00036-6.

[10] H. Goto, 'Auditory perception by normal Japanese adults of the sounds "l" and "r."', *Neuropsychologia*, vol. 9, pp. 317–323, 1971, doi: 10.1016/0028-3932(71)90027-3.

[11] M. V. Kondaurova and A. L. Francis, 'The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods', *J. Phon.*, vol. 38, no. 4, pp. 569–587, Oct. 2010, doi: 10.1016/j.wocn.2010.08.003.

[12] P. Escudero and P. Boersma, 'Bridging the gap between L2 speech perception research and phonological theory', *Stud. Second Lang. Acquis.*, vol. 26, no. 4, pp. 551–585, 2004.

[13] B. Chandrasekaran, P. D. Sampath, and P. C. M. Wong, 'Individual variability in cue-weighting and lexical tone learning', *J. Acoust. Soc. Am.*, vol. 128, no. 1, pp. 456–465, Jul. 2010, doi: 10.1121/1.3445785.

[14] J. Gandour, 'Tone perception in Far Eastern languages', *J. Phon.*, vol. 11, no. 2, pp. 149–175, Apr. 1983, doi: 10.1016/S0095-4470(19)30813-7.

[15] J. Gandour, 'TONE DISSIMILARITY JUDGMENTS BY CHINESE LISTENERS/声调异同辨别的测试', *J. Chin. Linguist.*, pp. 235–261, 1984.

[16] A. L. Francis, V. Ciocca, L. Ma, and K. Fenn, 'Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers', *J. Phon.*, vol. 36, no. 2, pp. 268–294, 2008.

[17] Y.-R. Chao, 'A system of tone letters', *Maître Phon.*, 1930.

[18] P. C. Wong and T. K. Perrachione, 'Learning pitch patterns in lexical identification by native English-speaking adults', *Appl. Psycholinguist.*, vol. 28, no. 4, pp. 565–585, 2007.

[19] P. Iverson, V. Hazan, and K. Bannister, 'Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults', *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 3267–3278, 2005.

[20] D. G. Jamieson and D. E. Morosan, 'Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones', *Percept. Psychophys.*, vol. 40, no. 4, pp. 205–215, Jul. 1986, doi: 10.3758/BF03211500.

[21] P. Boersma and D. Weenink, 'Praat: doing Phonetics by Computer', 2005. https://www.fon.hum.uva.nl/praat/ (accessed Dec. 16, 2021).

[22] P. Boersma, 'Praat: doing phonetics by computer', *Httpwww Praat Org*, 2006.

[23] E. Moulines and J. Laroche, 'Non-parametric techniques for pitch-scale and time-scale modification of speech', *Speech Commun.*, vol. 16, no. 2, pp. 175–205, Feb. 1995, doi: 10.1016/0167-6393(94)00054-E.

[24] Y. Wang, A. Jongman, and J. A. Sereno, 'Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training', *J. Acoust. Soc. Am.*, vol. 113, no. 2, pp. 1033–1043, Feb. 2003, doi: 10.1121/1.1531176.

[25] Y. Wang, M. M. Spence, A. Jongman, and J. A. Sereno, 'Training American listeners to perceive Mandarin tones', *J. Acoust. Soc. Am.*, vol. 106, no. 6, pp. 3649–3658, Dec. 1999, doi: 10.1121/1.428217.

[26] C. Liu, 'Just noticeable difference of tone pitch contour change for English- and Chinese-native listeners', *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3011–3020, Oct. 2013, doi: 10.1121/1.4820887.