Phonological acceptability is not isomorphic with phonological grammaticality of stimulus

Acceptability judgement studies have become a staple of experimental research on phonology (Albright, 2009; Hayes & Wilson, 2008; Scholes, 1966). An important assumption such studies make is that acceptability ratings are a direct indicator of the grammaticality of a particular stimulus. However, research in perception has shown convincingly that native listeners perceive auditory illusions when presented with sound sequences that do not respect the phonotactic constraints of their language (Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999; Hallé, Segui, Frauenfelder, & Meunier, 1998; amongst others). These two lines of research present a paradox: if indeed participants are hearing auditory illusions in illicit phonotactic contexts, then how can listeners be expected to rate stimuli containing such sequences poorly? One way out of this paradox is to posit that the acceptability ratings are based on representations that are prior to the "repair" of illicit phonotactic sequences, i.e., acceptability judgements are based on the original veridical percept before any perceptual repairs are made to the input. Therefore, a participant can give low ratings to an illicit sequence and still perceive an auditory illusion. Here, we show that this position is untenable. Therefore, a simplistic assumption that acceptability judgements are a direct reflection of grammaticality of the presented stimulus is untenable. Furthermore, we show that acceptability judgements appear to be based on the perceived input (which might include illusory vowels), and not on the actual input.

Experiment 1: Crucially, Korean does not allow [cm, bm] sequences within words, but does allow [cim, cɯm, bim, bɯm] sequences. Based on prior work, we expect Korean listeners to hear illusory vowels in stimuli with [cm, bm] sequences. We designed an experiment to see if we could correlate the presence/absence of illusory vowels in stimuli with illicit phonotactic sequences with the corresponding acceptability judgement. *For each stimulus presented*, the participants identified the medial vowel in the stimulus task (choices: [i, ɯ, nothing]) immediately followed by an acceptability judgement task (scale: 1-5, 1="impossible", 5="possible"). The experiment was conducted on 29 native Korean speakers (18 women and 11 men, age: 19-25 years) in Daegu, South Korea, and the instructions were given to them in Korean. We presented participants with nonce words of the form $V_1C_1V_2$ma, where, V_1 = [a, i, u]; C_1 = [c, b]; V_2 = [i, ɯ, Ø (Null)]; C_1 fillers (unreported here) = [m, ng, n]. All the tokens had stress on the first vowel, and were natural recordings by a trained phonetician.

The identification results for the test items are presented in Figure 1. As has been observed in previous research (Durvasula & Kahng, 2015), for illicit [bm] sequences (green, 1st col.), only /ɯ/ is a potential illusory vowel; however, for illicit [cm] sequences (green, 2nd col.), both /i/ and /ɯ/ are potential illusory vowels. The results of the acceptability rating task for the crucial stimuli with no medial vowel (V_1C_1ma stimuli) are presented in Figure *2*. The best logistic mixed effects model fitted to the acceptability judgments revealed, crucially, that there was a clear difference between the goodness ratings of *the same stimuli with no medial vowel*, when the participants identified an illusory vowel compared to when they did not ($\hat{\beta}$=1.48, SE= 0.12, p<0.0001). The results suggest that the acceptability rating of a stimulus with no medial vowel is contingent on whether the participant perceived an illusory vowel in the stimulus or not: when they do hear an illusory vowel, the goodness ratings are much higher than when they do not. Furthermore, hierarchical clustering revealed that the goodness ratings of all the illusory vowels consistently patterned with those of the stimuli with full medial vowels.

Experiment 2: In Experiment 1, it is possible that the acceptability judgements were contingent on the responses in the identification task only because participants were explicitly asked for the

identification of the medial vowel before the acceptability judgement (so, the observed relationship could have been a task effect). Therefore, in Experiment 2, we gave another set of 20 Korean participants (10 women and 10 men, age: 19-25 years) in Daegu, South Korea, just an acceptability judgement task with exactly the same stimuli as in Experiment 1. If the acceptability judgments for the illicit sequences [V_1bma, V_1cma] in Experiment 1 were contingent on the illusory vowel perception *only* because of a task effect, then the average goodness rating of the illicit stimuli would have been higher than otherwise as the acceptability judgments for the illusory vowel cases would have raised the overall average rating for the illicit stimulus. Crucially, then in Experiment 2, the average goodness rating for the same stimuli with no medial vowels should be (much) lower than Experiment 1. In Figure 3 below, we show the mean goodness ratings for different stimuli in Experiments 1 and 2. As can be seen, the patterns of acceptability in the two experiments were nearly identical. The best logistic mixed effects model fitted to the acceptability judgments did not include Experiment as a fixed effect (it had just a fixed effect of C_1); this suggests that the patterns of acceptability in the two experiments are not observably different, and that the results in Experiment 1 are not an artifact of the specific task.

Discussion: We have shown using two experiments that participants' acceptability judgements are contingent on the *perceived* representation and not the *actual* input. Therefore, just as it has been argued for syntactic knowledge (Chomsky, 1965; Schütze, 1991), acceptability judgements in phonology cannot be seen as direct reflections of the grammaticality of the actual input. As a consequence, researchers need to make a better effort at understanding the perceived representation in order to better understand/model acceptability judgements.

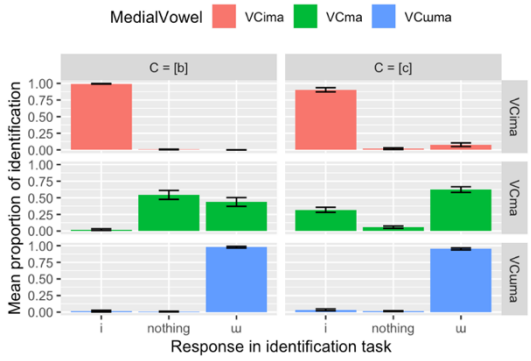

Figure 1: Identification responses for stimuli with different C_1 = [c, b], and V_2 = [i, ɯ, Ø]

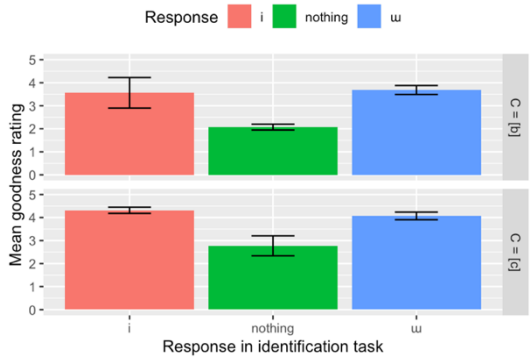

Figure 2: Mean goodness rating for V_1C_1ma stimuli with C_1 = [c, b], for different responses chosen

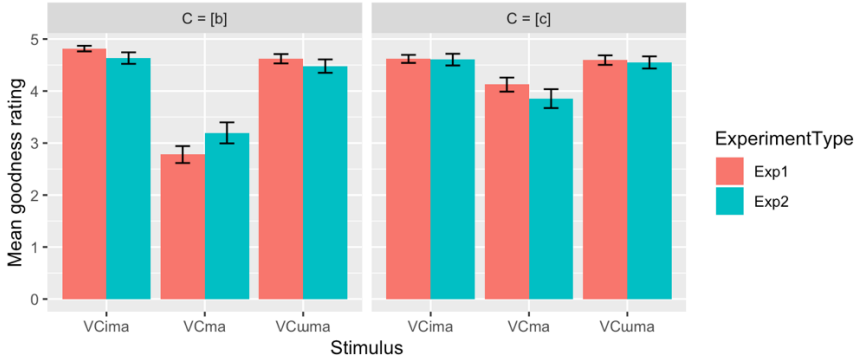

Figure 3: Comparison of mean goodness ratings for test stimuli in Experiments 1 & 2