L1 EFFECTS ON NAIVE PERCEPTION AND PRODUCTION OF [ʉ] BY MANDARIN AND ITALIAN SPEAKERS

Xuanda Chen[1,3] Meghan Clayards[1,2,3] Heather Goad[1,3] Donghyun Kim[4]

[1]Department of Linguistics, [2]School of Communication Sciences and Disorders,
[3]CRBLM, McGill University [4]Kumoh National Institute of Technology, South Korea
{xuanda.chen, meghan.clayards, heather.goad}@mcgill.ca, heydonghyun@gmail.com

ABSTRACT

Accurately predicting the difficulty of learning novel sounds based on linguistic experience is a challenge. This paper explores the effectiveness of two approaches, the spatial and dimension approaches, in predicting naive perception and production of novel sounds. The spatial approach measures the gradient acoustic-phonetic distance, while the dimension approach leverages discrete sound constructs. We investigate Mandarin and Italian speakers' performance on discrimination and imitation tasks involving the novel sound [ʉ]. The results show that Mandarin speakers discriminate [ʉ] better than Italian speakers, but both groups have similar performance in imitation. Our findings suggest that the dimension approach is more effective in capturing early learning difficulty of novel sounds, and that listeners perform better when the dimensions used to construct the novel sounds are contrastive in their native language.

Keywords: perception; imitation; similarity

1. INTRODUCTION

Cross-linguistic studies revealed that linguistic experience strongly conditions listeners' perception and production of novel sounds [1]. Novel sounds have been observed to range in difficulty for speakers of different languages. For example, it is well-known that Japanese speakers have trouble discriminating between English /l/ and /r/ [2], but Chinese speakers do not [3]. Thus, given the sound structure of one's native language, what predictions can we make about which novel speech sounds will be hard to perceive and produce? A large number of studies have appealed to the notion of 'similarity' to answer this question. A novel sound should cause difficulties when it is similar to existing L1 sounds [4]. However, how 'similarity' between novel sounds and L1 sounds should be defined is still a matter of debate. This paper compares two different approaches to quantifying sound similarity: the spatial approach and the dimension approach.

The spatial approach, exemplified by the Speech Learning Model (SLM), defines similarity between sounds based on their acoustic-phonetic distance. SLM predicts that a greater distance between an L2 and the nearest L1 sound results in easier acquisition of the L2 sound [4]. The SLM makes the correct predictions for Japanese listeners' acquisition of the English /l/-/r/ distinction [5]. While the SLM makes no prediction for naive learners, if the SLM's prediction holds at the onset of learning, sounds that are closer in acoustic-phonetic distance to L1 sounds will be harder to perceive and produce than those that are further away.

The dimension-based approach quantifies sound similarity in cross-language perception by using dimensions to classify sounds. These dimensions, such as vowel backness or roundedness, describe how a language contrasts pairs of phonemes. Distinctive feature theory [6] shares similar concepts, but unlike features, dimensions do not aim to capture a language's phonological processes.

These dimensions can be expressed acoustically, such as using F3 (e.g. [7]), or articulatorily, like lip rounding (e.g. [8]). We assume that if a dimension is contrastive, it can be dissociated from other dimensions and repurposed to build new sounds (cf. e.g. [3, 9]). Thus, speakers of that language should be better able to perceive and produce novel combinations using that dimension than speakers of a language that does not use that dimension contrastively.

The two approaches predict naive perception and production based on different aspects of linguistic experience. For example, Mandarin has three high vowels, /i/, /u/ and /y/, while Italian only has two, /i/ and /u/. The phonetic realization of the two peripheral vowels is similar in the two languages, with the only difference being the presence or absence of front rounded /y/. Suppose that Mandarin and Italian speakers are presented with the novel high central rounded vowel [ʉ]. Which language group will have better naive perception and production?

Under the spatial approach, the crucial distance metric for Mandarin is [y]-[ʉ], while for Italian, it is [u]-[ʉ]. The acoustic distance in Mandarin is smaller than in Italian, and thus, Italian speakers, due to the greater distance between [ʉ] and its nearest L1 sound, should have better naive perception and production than Mandarin speakers. The dimension approach makes the opposite prediction. Italian does not have contrasts in roundedness (rounding is bound to the backness dimension). By contrast, Mandarin speakers have a /y/-/i/ contrast, meaning that the roundedness dimension is contrastive in this language. Thus, Mandarin speakers should have better naive perception and production of [ʉ] because contrastive roundedness is available to capture [ʉ].

Our research question is which approach will be borne out? We use an AX discrimination task to assess perception and an imitation task to assess production. We use an ABX identification task to confirm that the crucial distance metric is different between Mandarin and Italian, and it can be used to test the spatial approach's assumptions.

2. METHODS

2.1. Participants

We recruited 27 Mandarin speakers (mean age = 22.5, standard deviation = 3.14) and 27 Italian speakers (mean age = 26.2, standard deviation = 6.60). All participants were monolingual speakers of either Mandarin or Italian. None reported any speech or hearing impairments.

2.2. Stimuli

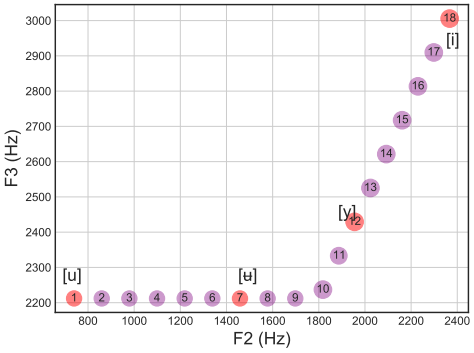

Figure 1: The high vowel stimuli.

The stimuli were 60 msec isolated high vowels on the [i]-[y]-[u] continuum following [10], synthesized in Praat [11]. We first selected four sound steps from the continuum to represent the categories of [i], [y], [ʉ], and [u] (see Figure 1). F1 was 300 Hz for all stimuli. F2 values were manually determined, and F3 values were automatically calculated using the equation $F3 = 1.4 * (F2 - 220)$ [12]. Two endpoints, representing [i] (step 18) and [u] (step 1) categories, were selected based on the estimation of mean values for Mandarin and Italian prototypical [i] and [u] [13, 14]. We chose [y] (step 12) based on the judgment of the first author, a native speaker of Mandarin, and [ʉ] (step 7) based on the first recording of [ʉ] on the course webpage [8], available at https://linguistics.berkeley.edu/

acip/course/ipachart/vowels.html. Equidistant steps were created to fill in the gaps between categories for a total of 18 stimuli. The stimuli and related files are available at the project page https://osf.io/7k2j8/?view_only=a26c2961e5cf4d6a99ccf50c23bceee3.

2.3. Procedure

The experiment was performed using an online experiment service implemented in jsPsych [15]. Participants accessed the experiment in their own homes and used their own recording devices and headphones (they had to pass a headphone test before the main experiment [16]). They received written instructions about the experiment in their native language, and were required to give responses by pressing buttons on the keyboard. Each participant completed three tasks, and each task contained a practice session.

ABX Task Participants had to decide whether the last sound they heard (sound steps between 1 and 12) was the same as the first ([y] step 12) or the second ([u] step 1) [17]. Participants were tested on 40 trials using stimuli from the [y]-[ʉ]-[u] range, with an inter-stimulus interval (ISI) of 1500 msec [18].

AX Task Participants had to determine whether two stimuli in a trial were the same or different, with an ISI of 1500 msec. We tested all possible different sound pairs within 9 steps (e.g., step 2-10 but not step 2-11), and 36 same pairs (e.g., step 1-1) for a total of 144 trials in three blocks.

Imitation Task Participants were instructed to produce a sound as close as possible to the presented stimulus. 18 stimuli were presented 4 times in random order. In each trial, the stimuli were played twice, with an ISI of 1500 msec. We used the default sampling rate of 22.05 kHz during recording.

2.4. Measurements

For the ABX task, we used a logistic mixed-effect regression to model the response pattern, i.e., to determine if [ʉ] was categorized as an instance of [y] or [u] [19]. The normalized predictors included continuous VOWELS and categorical LANGUAGE. The model included random intercepts for participants with p-values calculated by *lmerTest* [20]. For the AX task, goodness of sound discrimination was measured as the average accuracy of response to a sound pair for each stimulus across languages. We performed ANOVA tests to examine the difference in [ʉ] accuracy across the two languages. For the imitation task, each imitation sample was manually measured at the temporal midpoint to obtain formant values (Hz) in Praat. Goodness of imitation was measured as the Euclidean distance between production and target. Imitation distance was normalized across participants with the Lobanov method [21]. We also performed ANOVAs to examine the difference in imitation in [ʉ] across the two languages.

3. RESULTS

Due to limited space, we will focus on results for [ʉ] only. Figure 2 shows GLM fits to the response data from the ABX task. Step

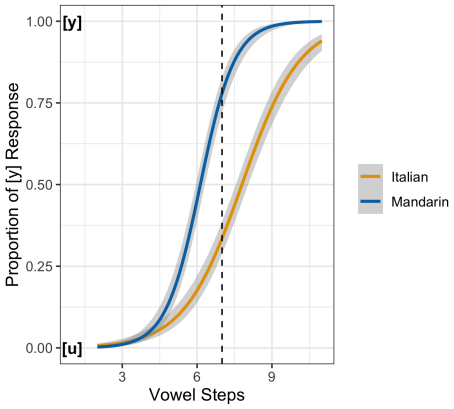

Figure 2: Vowel categorization

7, [ʉ], is marked with a dashed line. The model found that participants categorized the stimuli more often as [y] as the stimuli shifted from [u] to [y] [$\hat{\beta}$=7.154, σ=0.377, p<0.005] and that Mandarin speakers were more likely to categorize the stimuli as [y]

than Italian speakers [$\hat{\beta}$=2.058, σ=0.344, p<0.005]. From the plot we see that [ʉ] is categorized as [y] more than 50% of the time for Mandarin speakers and as [u] more than 50% of the time for Italian speakers. The results confirm that the crucial distance metric is different for Mandarin and Italian participants, and that [ʉ] is closer to a native sound for Mandarin than for Italian listeners.

Figure 3 provides boxplots for the discrimination data from the AX task and imitation accuracy for the two languages. The ANOVA test showed that Mandarin participants have higher discrimination accuracy for [ʉ] than Italian participants [F(1, 52)=8.812, p=0.004]. The ANOVA also found that the two groups of speakers do not significantly differ from each other [F(1, 192)=0.002, p=0.965] in imitation.

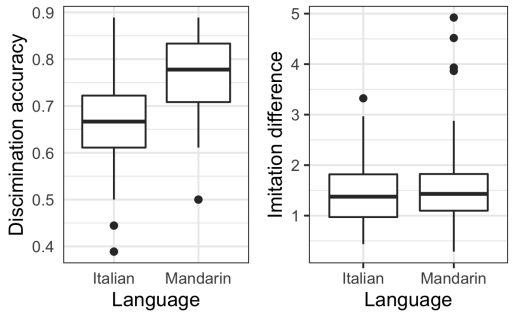

Figure 3: Discrimination and imitation of [ʉ]

4. DISCUSSION AND CONCLUSION

The goal of the present study was to test which of the two different approaches to calculating similarity makes correct predictions for perception and production on first exposure to a non-native sound, [ʉ]. The categorization results show that Mandarin participants tend to categorize [ʉ] as /y/, while Italian participants tend to categorize it as /u/. This confirms the assumption that the crucial distance metric is [u]-[ʉ] for Italian and [y]-[ʉ] for Mandarin, allowing us to test the spatial approach (gradient acoustic-phonetic distance). Mandarin speakers have higher discrimination accuracy than Italian speakers, which is not predicted by the spatial approach but supports the dimension approach. The imitation results do not support one approach over the other.

Our results suggest that acoustic-phonetic distance may not be sufficient to explain all sound similarity problems. On first exposure to a novel sound, speakers may benefit from a contrastive dimension in their language to perceive a novel sound that involves that dimension, though this is not guaranteed to facilitate production. This suggests that speakers learn representations for sounds with discrete parts. Similar ideas can be found in the dimension-based perceptual interference account [7] and in the Redeployment Hypothesis (RH) [9]. The RH states that a phonological feature [F] can be redeployed from the L1 grammar to represent a non-native contrast by combining with another feature [G] to form [F, G]. Both the dimension approach and the RH suggest that success in naive perception depends on speakers' knowledge of sound constructs and their ability to reuse these components of sounds, either dimensions or phonological features.

In conclusion, we have argued that the impact of linguistic experience on perceiving and producing a novel sound can be characterized as experience in using particular dimensions in L1. If the novel sound includes a contrastive dimension in L1, perceiving a novel sound combination using that dimension on first exposure may be less difficult. Conversely, perceiving it may be more challenging if the dimension is absent. The imitation results are particularly intriguing as they indicate that neither approach provides accurate predictions. It is possible that additional factors, such as retrieval of incorrect motor plans, influence the imitation of novel sounds. Future research should investigate the underlying reasons for these findings, as well as the connection between perception and production.

5. REFERENCES

[1] K. Miyawaki, J. J. Jenkins, W. Strange, A. M. Liberman, R. Verbrugge, and O. Fujimura, "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English," *Perception & Psychophysics*, vol. 18, no. 5, pp. 331–340, 1975.

[2] H. Goto, "Auditory perception by normal Japanese adults of the sounds 'L' and 'R'," *Neuropsychologia*, 1971.

[3] C. Brown, "The interrelation between speech perception and phonological acquisition from infant to adult," *Second language acquisition and linguistic theory*, vol. 1, pp. 4–64, 2000.

[4] J. E. Flege, "Assessing constraints on second-language segmental production and perception," *Phonetics and phonology in language comprehension and production*, vol. 6, pp. 319–355, 2003.

[5] 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/," *Journal of Phonetics*, vol. 32, no. 2, pp. 233–250, 2004.

[6] N. Chomsky and M. Halle, *The sound pattern of English*. New York: Harper and Row, 1968.

[7] P. Iverson, P. K. Kuhl, R. Akahane-Yamada, E. Diesch, A. Kettermann, C. Siebert *et al.*, "A perceptual interference account of acquisition difficulties for non-native phonemes," *Cognition*, vol. 87, no. 1, pp. B47–B57, 2003.

[8] P. Ladefoged and K. Johnson, *A course in phonetics [6th ed.]*. Boston, Massachusetts: Wadsworth Publishing, 2010.

[9] J. Archibald, "Second language phonology as redeployment of phonological features," *Canadian Journal of Linguistics*, vol. 50, no. 1-4, pp. 1000–1030, 2005.

[10] B. L. Rochet, "Perception and production of second-language speech sounds by adults," *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research*, pp. 379–410, 1995.

[11] P. Boersma, *Praat: doing phonetics by computer*, 2018. [Online]. Available: http://www.praat.org/

[12] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *The Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 2088–2113, 1989.

[13] P. M. Bertinetto and M. Loporcaro, "The sound pattern of standard Italian, as compared with the varieties spoken in florence, milan and rome," *Journal of the International Phonetic Association*, vol. 35, no. 2, pp. 131–151, 2005.

[14] J. Yang, *Acoustic properties of vowel production in Mandarin-English bilingual and corresponding monolingual children*. The Ohio State University, 2014.

[15] J. R. De Leeuw, "jspsych: A javascript library for creating behavioral experiments in a web browser," *Behavior Research Methods*, vol. 47, no. 1, pp. 1–12, 2015.

[16] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig, and M. Chait, "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, vol. 53, no. 4, pp. 1551–1562, 2021.

[17] W. Strange, V. L. Shafer *et al.*, "Speech perception in second language learners: The re-education of selective perception," *Phonology and Second Language Acquisition*, vol. 36, pp. 153–192, 2008.

[18] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Perception & Psychophysics*, vol. 37, no. 1, pp. 35–44, 1985.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[20] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[21] B. M. Lobanov, "Classification of Russian vowels spoken by different speakers," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 606–608, 1971.