INTRODUCTION

Speech signals vary substantially in a number of their key properties, with the variability deriving from, among other things, talkers' age and gender. Speech processing requires resolution of this variation, necessitating interpretation of age and gender information in the signal. In some signals, the age and gender are not clear from acoustic information alone. In these cases there may be substantial individual variation in judgments of age and gender. In this study, we examine the interplay between the interpretation of age and gender across language communities.

VLAM CORNER VOWEL STIMULI

The *Variable Linear Articulatory Model* (VLAM, Boë and Maeda, 1997) is a computational model of the articulatory system and its speech production capacities. Midsagittal representations, such as those depicted in Figure 1 (bottom), are wrought by configuring "articulatory blocks" (Maeda, 1990, 1991) corresponding to jaw height, tongue body position, tongue dorsum position, tongue apex position, lip protrusion, lip height, and larynx height. The VLAM is age-varying and capable of representing vocal tract lengths ranging from those of infants to young adults, calibrated to age based on Beck (1996). Given an age in years, the set of all articulatory configurations of the VLAM at that age that do not result in occlusion of the oral cavity yield a corresponding *maximal vowel space* (Boë *et al.*, 1989) for that age. Corner vowel stimuli ([i], [u], [a]) were generated by the VLAM set at seven different ages: 6 months, 2, 4, 5, 10, 16, and 21 years, and are indexed numerically as 1, 6, and 36, respectively, in each of the maximal vowel spaces pictured in Figure 1 (top). For each age, the corner vowel stimuli were part of a set of 38 "prototype" vowel stimuli that were presented for a vowel categorization task (Ménard *et al.*, 2009; Munson *et al.*, 2010).

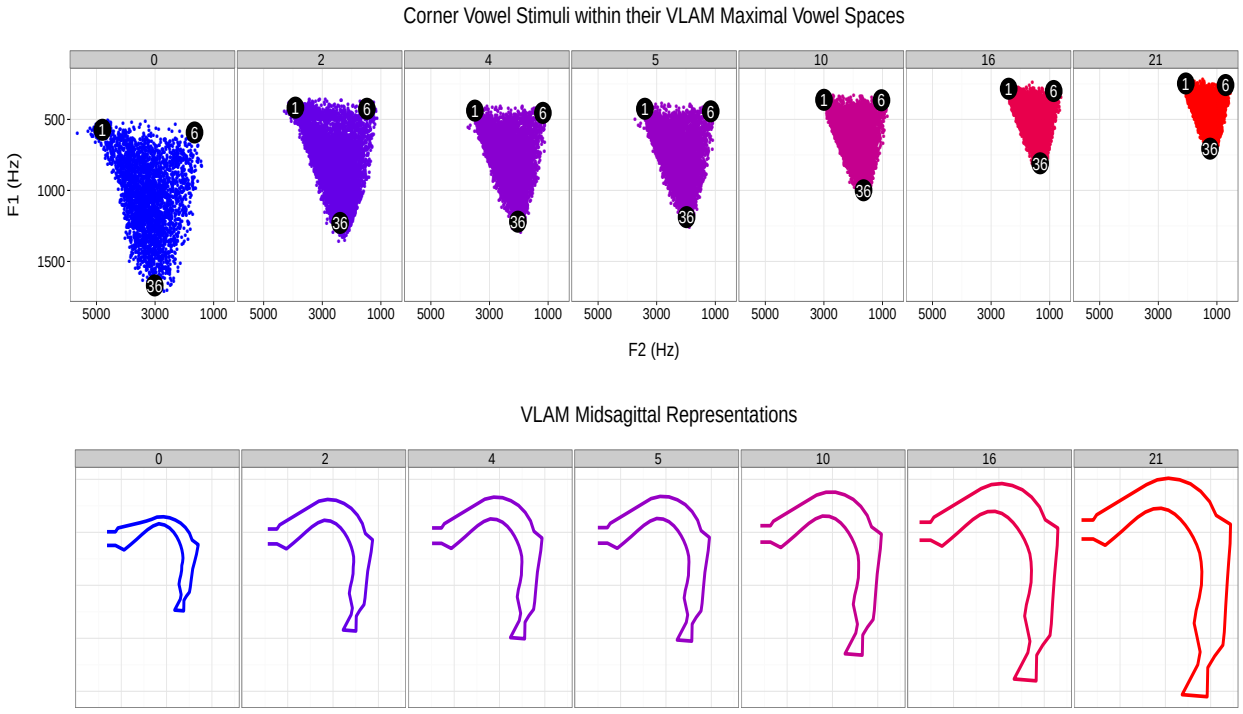

FIGURE 1: (top) Corner vowel stimuli ([i], [u], [a]) generated by the VLAM set at seven different ages: 6 months, 2, 4, 5, 10, 16, and 21 years, and indexed numerically as 1, 6, and 36, respectively. (bottom) Midsagittal representations wrought by the VLAM at each vocal tract age.

Experimental procedure

Subjects included 15 native speakers of Cantonese recruited in Hong Kong by Benjamin Munson, 21 native speakers of American English recruited in Columbus by Chanelle Mays, and 21 native speakers of Japanese recruited in Tokyo by Kiyoko Yoneyama. Subjects were played each stimulus and listeners assigned to each stimulus an age in years, and a gender along a visual analog scale ranging from "definitely male" to "definitely female" (or the equivalent in Japanese or Cantonese) similar to that depicted in Figure 2 below. Responses along the gender scale were converted to numbers ranging from "definitely male" (0 on the scale) to "definitely female" (650 on the scale).

Figure 2: Visual analogue scale used by subjects to provide gender ratings in response to corner vowel stimuli.

The age and gender ratings were a final task block in an experiment in which the listeners from each language group first identified each of the larger set of 38 stimuli, blocked by vocal tract age, as one of a set of vowel phoneme categories for that language. Prior to each block, listeners were told the vocal tract age of the child whose vowels they would listen to in that block. The order of vocal tract ages was randomized across listeners. Listeners responded by clicking on a keyword (for English and Cantonese), or a hiragana symbol unambiguously representing the vowel in isolation (Japanese) (see Munson *et al.*, 2010).

Age and gender responses

Initial analysis of the response patterns to the stimuli yielded the results shown in Figures 3-5.

Figure 3 shows boxplots (median, interquartile range, and full range of values) for the age judgments (left) and the gender judgments (right) from all 57 listeners to all 6 trials for each vocal tract age. The age judgements generally increase as the vocal tract age of the VLAM increases, with more variable responses for the three oldest vocal tracts, and especially variable responses to vocal tract age 10. The gender judgements demonstrate ambiguity in interpretation up to age 10, which is resolved by sexual dimorphism at the simulated age 16.

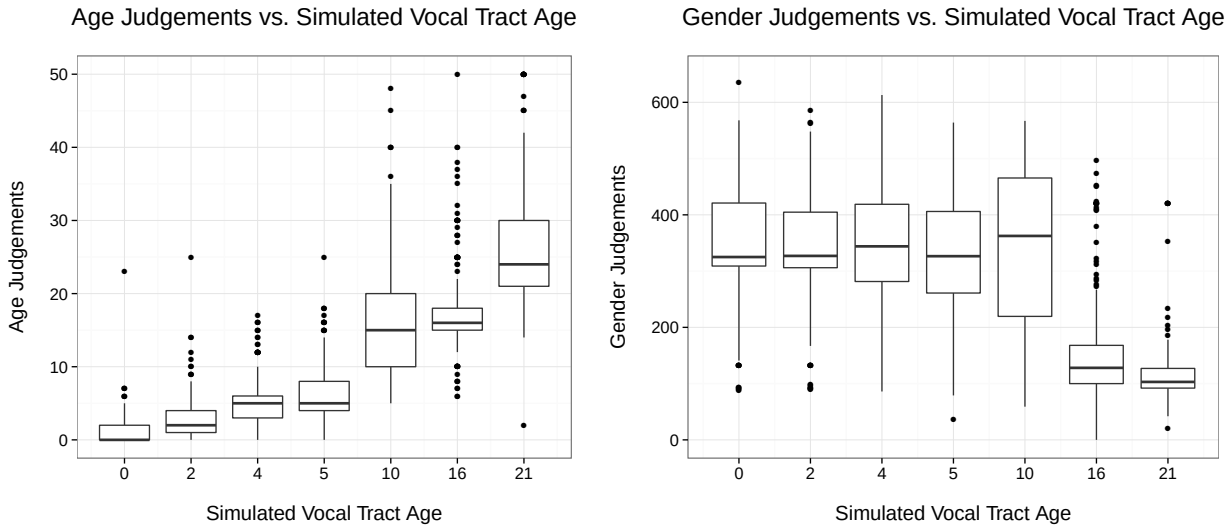

Figure 3: Box plots for age (left) and gender (right) judgements from all 57 subjects in response to corner vowel stimuli (2 presentations each of tokens 1, 6, and 36) produced by the VLAM set at ages 6 months, 2, 4, 5, 10, 16, and 21 years.

Figure 4 shows a column scatter plot grouped by simulated vocal tract age for the age judgments from all 57 listeners to all 6 trials for each vocal tract age, with the color indicating the assigned gender rating. A bifurcation is evident in the interpretation of age and gender for the age 10 stimuli, which subjects rated as either a younger male or older female, suggesting a nonuniformity in the resolution of variability during processing.

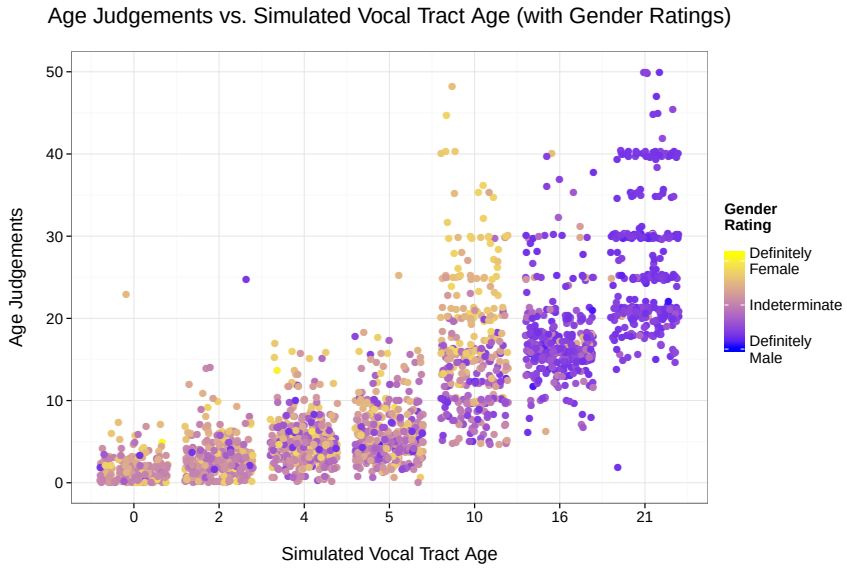

FIGURE 4: A column scatter plot grouped by simulated vocal tract age for the age judgments from all 57 listeners to all 6 trials for each vocal tract age, with the color indicating the assigned gender rating.

Figure 5 shows a set of column scatter plots arranged by vocal tract age and language, with data points grouped by vowel stimuli, for the age judgments from all 57 listeners to all 6 trials for vocal tract ages 10, 16, and 21 years, with the color indicating the assigned gender rating. The Japanese subjects are assigning an older rating to 21 year old vowels, and this is especially true for stimulus 6.

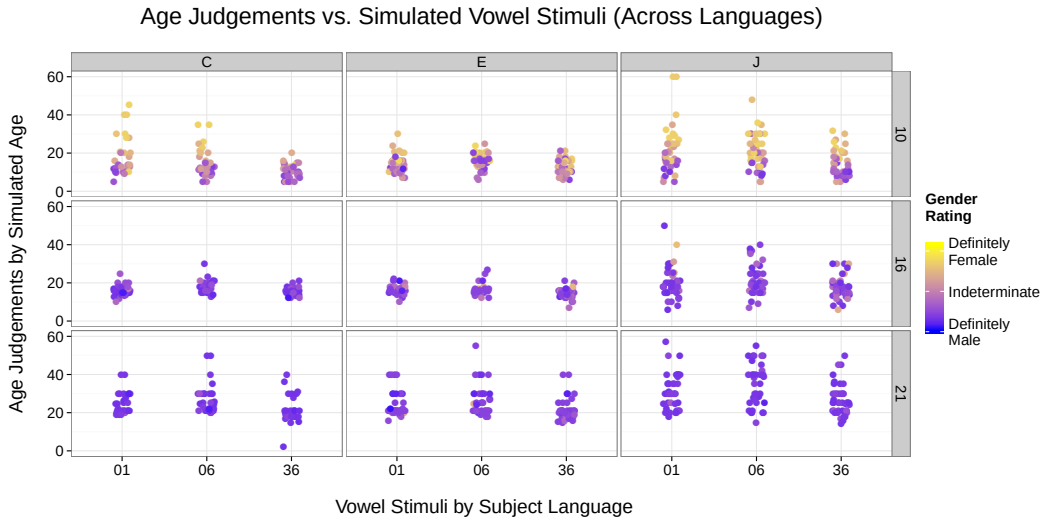

FIGURE 5: Column scatter plots arranged by vocal tract age and language, with data points grouped by vowel stimuli, for the age judgments from all 57 listeners to all 6 trials for vocal tract ages 10, 16, and 21 years, with the color indicating the assigned gender rating.

DISCUSSION

Age responses do increase as the simulated vocal tract age increases, but the interpretation of age is variable across listeners and interacts substantially with the interpretation of gender for the simulated 10-year-old. Gender judgments suggest a strong ambiguity up until age 10, at which point gender judgements become bimodal, and contingent on age judgement. Of course, this does not mean that natural language-specific stimuli would be as ambiguous, since there could well be language- (and culture-)specific cues to age and gender that are not available in these synthesized stimuli. For example, the much older responses to stimulus 6 in the Japanese speakers' judgements could be an artifact of the quality of this stimulus, which is [u] rather than the unrounded [ɯ] that is the prototypical value for the high back vowel of Japanese. The rounded quality of the [u] might make it be heard as "careful speech" by Japanese listeners (see Okada, 1999), which could suggest an older, more deliberate talker. Exploring these language effects further may reveal the depth of influence of culture on aspects of perception, especially vowel normalization. While further experiments will be necessary to understand the full extent of language effects, the current preliminary results strongly support models of vowel normalization across different talker types as a learned process rather than an innate hard-wired one.

ACKNOWLEDGMENTS

Work supported by NSF grants BCS 0729277 (to Munson) and BCS 0729306 (to Beckman).

REFERENCES

Beck, J. M. (**1996**). "Organic variation of the vocal apparatus", in *Handbook of Phonetic Sciences*, 256–297 (Blackwell, Cambridge, England).

Boë, L.-J. and Maeda, S. (**1997**). "Modélization de la croissance du conduit vocal. Éspace vocalique des nouveaux-nés et des adultes. Conséquences pour l'ontegenèse et la phylogenèse", in *Journée d'Études Linguistiques: "La Voyelle dans Tous ces États"*, 98–105 (Nantes, France).

Boë, L.-J., Perrier, P., Guérin, B., and Schwartz, J.-L. (**1989**). "Maximal vowel space", in *EUROSPEECH 09*, 281–284 (Paris, France).

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model", in *Speech Production and Speech Modeling*, edited by W. Hardcastle and A. Marchal, 131–149 (The Netherlands: Kluwer Academic Publishers).

Maeda, S. (**1991**). "On articulatory and acoustic variabilities", Journal of Phonetics **19**, 321–331.

Ménard, L., Davis, B., Boë, L.-J., and Roy, J.-P. (**2009**). "Producing American-English vowels during vocal tract growth: A perceptual categorization study of synthesized vowels", Journal of Speech, Language and Hearing Research .

Munson, B., Ménard, L., Beckman, M. E., Edwards, J., and Chung, H. (**2010**). "Sensorimotor maps and vowel development in English, Greek, and Korean: A cross-linguistic perceptual categorizaton study (A)", Journal of the Acoustical Society of America **127**, 2018.

Okada, H. (**1999**). "Japanese", in *Handbook of the International phonetic Association: a guide to the use of the International Phonetic Alphabet*, 117–119 (Cambridge University Press, Cambridge, U.K.; New York, New York).