Vowel perception at formant-harmonic crossovers

May Pik Yu Chan[1], Jianjing Kuang[1]
[1]*University of Pennsylvania*

The source-filter theory [1] which underpins much of modern phonetics research, often assumes the source and filter to be independent. This predicts that source-related attributes (namely, F0) cannot be influenced by the filter-related attributes (e.g. vowels). This long-held assumption has resulted in successful descriptions, syntheses and processing of speech production (e.g. [2]) and continues to be upheld in most current studies of speech, despite Klatt and Klatt's [3] warning from decades ago that such assumption is better made for low-pitched male speech than for high-pitched females and children's speech. Indeed, some studies in speech have suggested that vowels and pitch do interact. For example, it is well-known that high vowels tend to have higher F0 than low vowel counterparts, also known as intrinsic pitch (e.g. [4]), though the mechanism behind it remains unclear. Many working theories have attempted to account for this effect from both production (e.g. aerodynamic coupling between the larynx and the vocal tract leads to F0-F1 correlation (e.g. [5]); high vowels pulls on the larynx and increases vocal fold tension (e.g. [6])) and perception perspectives (e.g. F0-F1 distance determines vowel height perception (e.g. [7]); talker normalization (e.g. [8])). The singing literature hints at more complicated interactions between vowel and pitch, where a broader pitch range is involved. Bozeman [9] observed that when H2 crosses over F1, a timbral shift occurs, resulting in more closed sounding vowels. Formant-harmonic crossovers pose an interesting case for understanding the nature of vowel perception. To provide better insights into the interaction between vowel and pitch, we systematically investigate the perceptual effects of formant-harmonic crossovers with synthesized stimuli.

A set of vowels /i, u, a, ɔ/ in nine F0 steps were Klatt synthesized for a modified XAB discrimination task. The reference tokens (A/B) were versions of the 4 vowels that varied in F1 (+/-50 Hz) or F2 (+/-100Hz). The target token (X) varied in F0 along a continuum of 9 steps, each 1 semitone apart, and sharing identical filter functions as the reference tokens. At step 5, the F1 of the target vowel aligned to H2. Reference tokens had the average F1-F3 values of the male speakers from the Peterson & Barney (1952) dataset. For example, the 'X' token of /i/ has an F1 of 267 Hz, and an F0 value of 133.5 Hz at step 5. Its A/B tokens had F1s of either 317 Hz or 217 Hz. The F0 range of the X tokens depend on the vowels' F1; /i, u, a, ɔ/ had F0 ranges of 105.9-168.2 Hz, 121.8-193.4 Hz, 284.9-452.3 Hz and 225.4-357.8 Hz respectively. Fig. 1 illustrates the F0 step manipulation. 63 listeners judged whether the vowel quality of 'X' was more similar to tokens 'A' or 'B'. The order of presentation was fully randomized.

Perceptual results (Fig. 2) show a general V-shaped response pattern. Vowels tend to be perceived as higher (lower F1) when H2 crosses over F1 (step 5), with the vowel height being perceived as lower both before and after H2 crossing over F1. No systematic differences across vowels were found. A logistic regression model fitted to participants' responses with vowel formant, F0 step, vowel category and their interactions confirm these observations. Predicted marginal effects from the model are summarized in Fig. 3. *What is the mechanism for this V-shaped pattern?* It cannot be explained by any existing working theories of intrinsic pitch, as they would all suggest a linear relationship between vowel and pitch perception. We predict that listeners could be sensitive to energy changes under spectral peaks, and thus conducted a linear predictive coding (LPC) analysis on our stimuli, which successfully replicates the non-linearities of the V-shaped pattern that is consistent with perceptual results as Fig. 4 illustrates. Our results suggest that human perception similarly tracks spectral peaks in vowel perception.

We conclude that the relationship between F0 and vowel perception bears multifaceted psychoacoustic grounding, and that source-filter information could interact in perception, as in the case of harmonic-formant crossovers. The mechanism for vowel and pitch interaction in perception is comparable to formant tracking with LPC, suggesting that the auditory system searches for spectral peaks in the signal in a similar way.

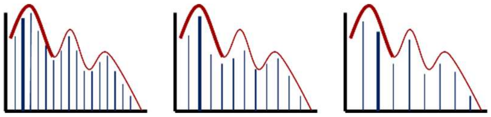

Fig. 1. Illustration of synthesized stimuli (left = before Step 5, middle = Step 5, right = after Step 5).

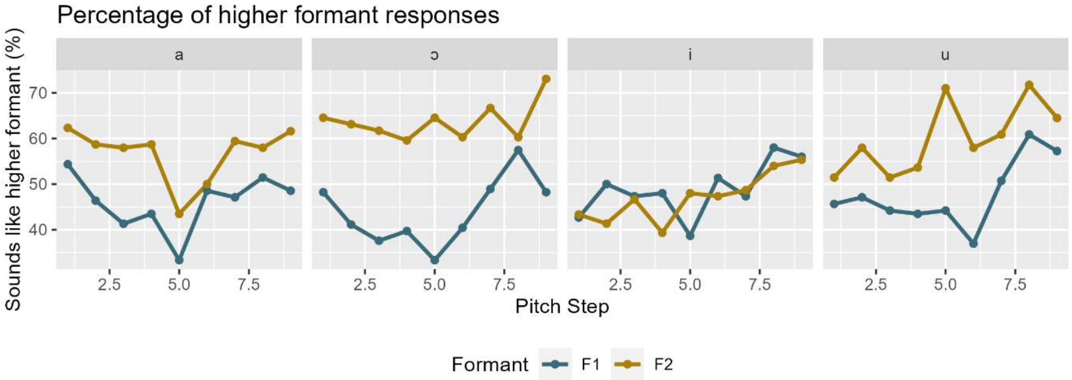

Fig. 2. Percentage of responses of token X sounding like it has a higher formant (F1 / F2).

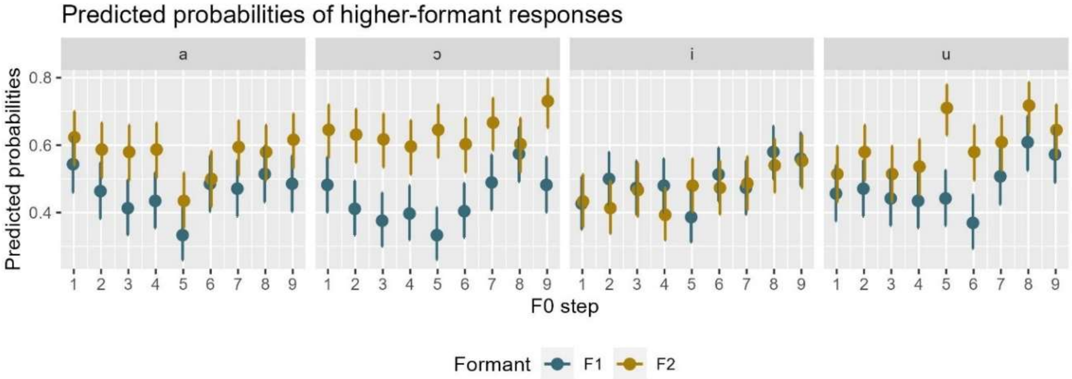

Fig. 3. Predicted marginal effects of logistic model with 95% Confidence Interval error bars.

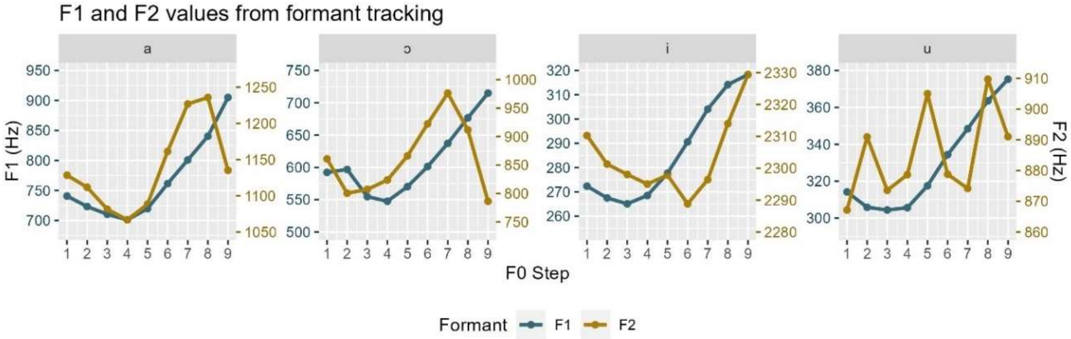

Fig. 4. Formant tracked F1 (Left y-axis, blue) and F2 (Right y-axis, yellow) values.

References

[1] Fant, G. (1980). The relations between area functions and the acoustic signal. *Phonetica*.

[2] Markel, J. D., & Gray, A. J. (1976). *Linear prediction of speech.*

[3] Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*

[4] Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *J. Phon.*

[5] Ewan, W. G., & Ohala, J. J. (1979). Can intrinsic vowel F be explained by source/tract coupling?. *J. Acoust. Soc. Am.*

[6] Chen, W. R., Whalen, D. H., & Tiede, M. K. (2021). A dual mechanism for intrinsic f0. *J. Phon.*

[7] Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *J. Acoust. Soc. Am.*

[8] Johnson, K. (1990). Contrast and normalization in vowel perception. *J. Phon.*

[9] Bozeman, K. W. (2010). The Role of the First Formant in Training the Male Singing Voice. *Journal of Singing*.