THE UNIVERSITY OF ARIZONA

Identification of consonants produced by acoustically driven modulations of male and female vocal tracts

Brad H. Story & Kate Bunton

Speech, Language, and Hearing Sciences, University of Arizona

Acoustic-based control of vocal tract modulation

- In the current version of the TubeTalker model of speech production, an utterance is specified as a sequence of relative acoustic events along a time axis (Story & Bunton, 2017;2019;2021).
- These events specify directional changes of the vocal tract resonance frequencies relative to a neutral tract shape, and are called resonance deflection patterns (RDPs). When associated with a temporal event function, the RDPs are transformed via acoustic sensitivity functions, into time-varying modulations of the vocal tract shape that, in turn, affect the temporal variation of the formants.

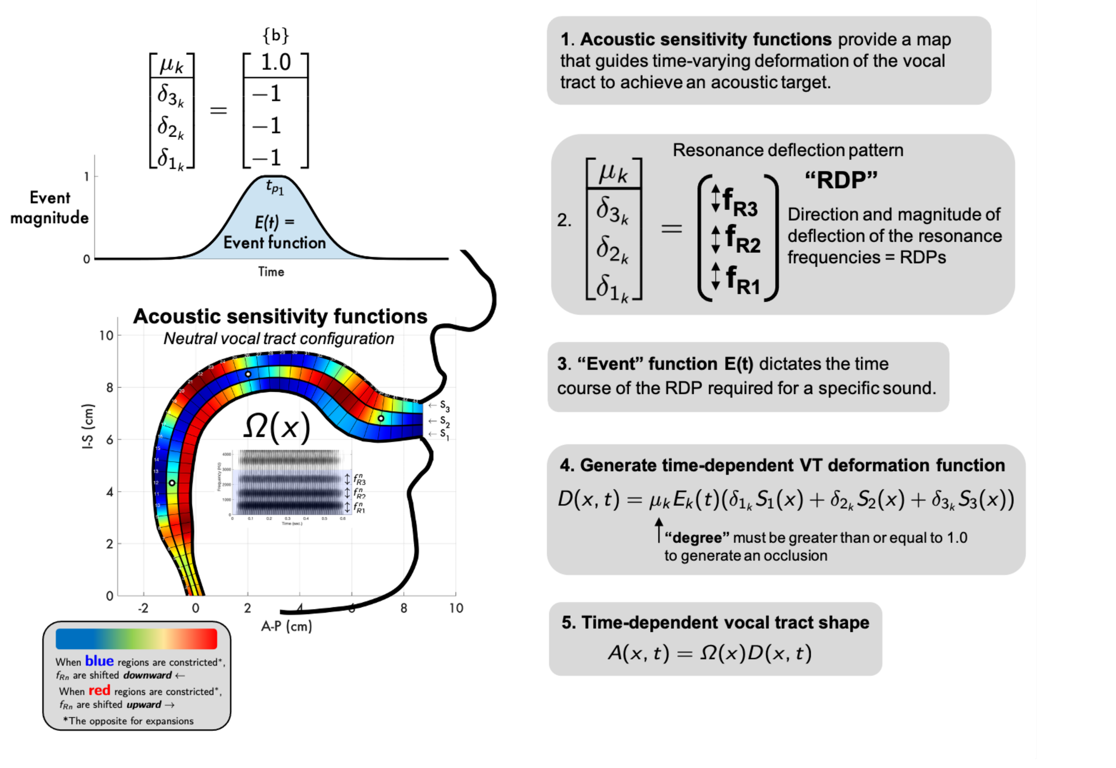

Figure 1: Transformation of a discrete representation of a phonetic segment (RDP) into time-varying vocal tract modulations.

Example: RDPs transformed to speech

- RDPs intended to represent phonetic targets /t/, /ɪ/, /k/ are transformed via overlapping event functions into a time-varying vocal tract area function, and subsequently synthetic speech via the TubeTalker system (Story, 2013).
- This modeling approach does not require any explicit specification of vocal tract shaping parameters (e.g., constriction location); modulation of the vocal tract is based on achieving the acoustic targets specified by the RDPs.
- The δs in an RDP are ordered from bottom to top to emulate the vertical frequency axis in a spectrogram, and they have the effect of shifting the formant frequencies upward or downward in frequency to encode a message.

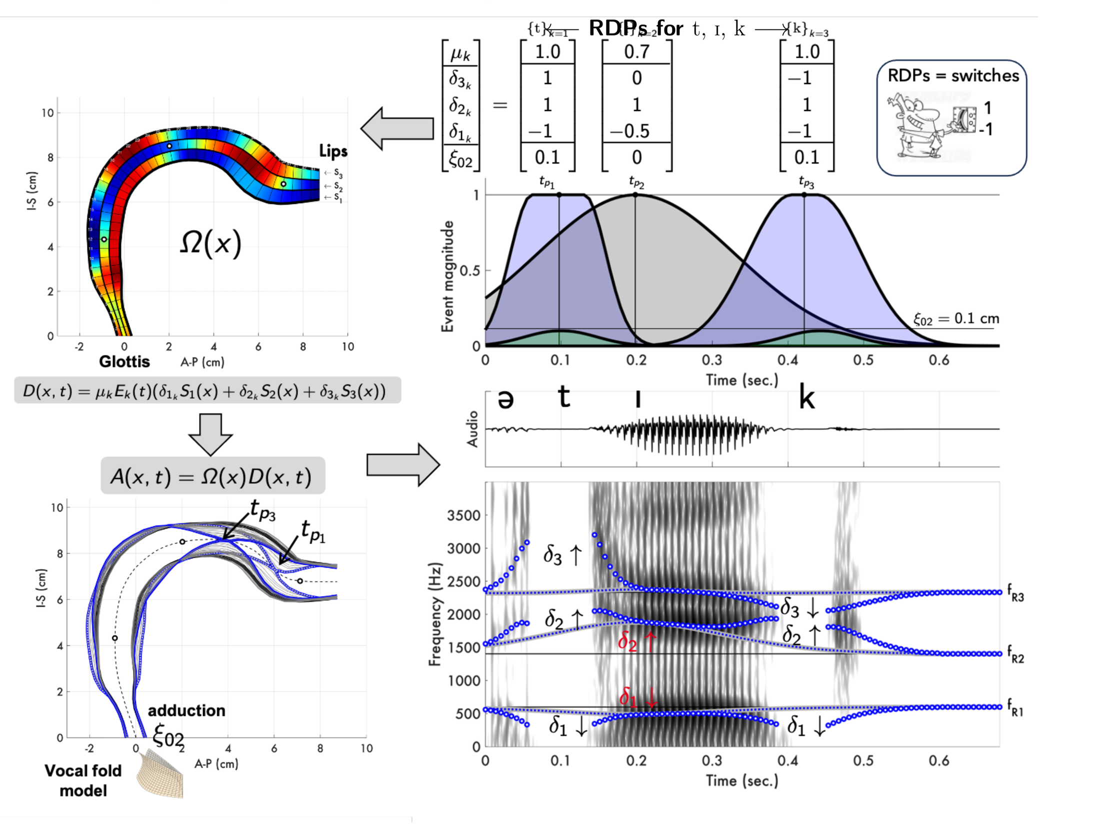

Figure 2: Transformation of three sequential RDPs into the synthesized utterance "a tick". The δs with arrows are shown in the spectrogram to indicate their effect on the resonance frequencies (gray and blue lines) and ultimately the formant frequencies in the output speech.

- RDPs can be conceived as a bank of switches where production of contrastive phonetic segments can be generated by changes in the pattern of switch settings.
- The consonants in the example above were generated with binary settings (1 or -1) of the three elements of the RDP (i.e., the δs).
- The motivation of this study was to better understand how incremental variation, rather than binary change, of the individual elements of the RDPs (specifically δ_2 and δ_3) affect the vocal tract shape and the identification of the consonants.

Specific aim and construction of VCV continua

- **Specfic aim:** To determine the effect of variations of magnitude and polarity of RDPs on consonant identification while all other parameters are held constant.

  Hypothesis 1: Consonant ID will shift from /d,t/ to /g,k/ when the polarity of δ_3 switches from positive to negative.

  Hypothesis 2: Consonant ID will shift from /b,p/ to /g,k/ when the polarity of δ_2 switches from negative to positive.

- Demonstrations of constructing the VCV continua for male and female speech production systems, and for the voiced and unvoiced cases are shown below.

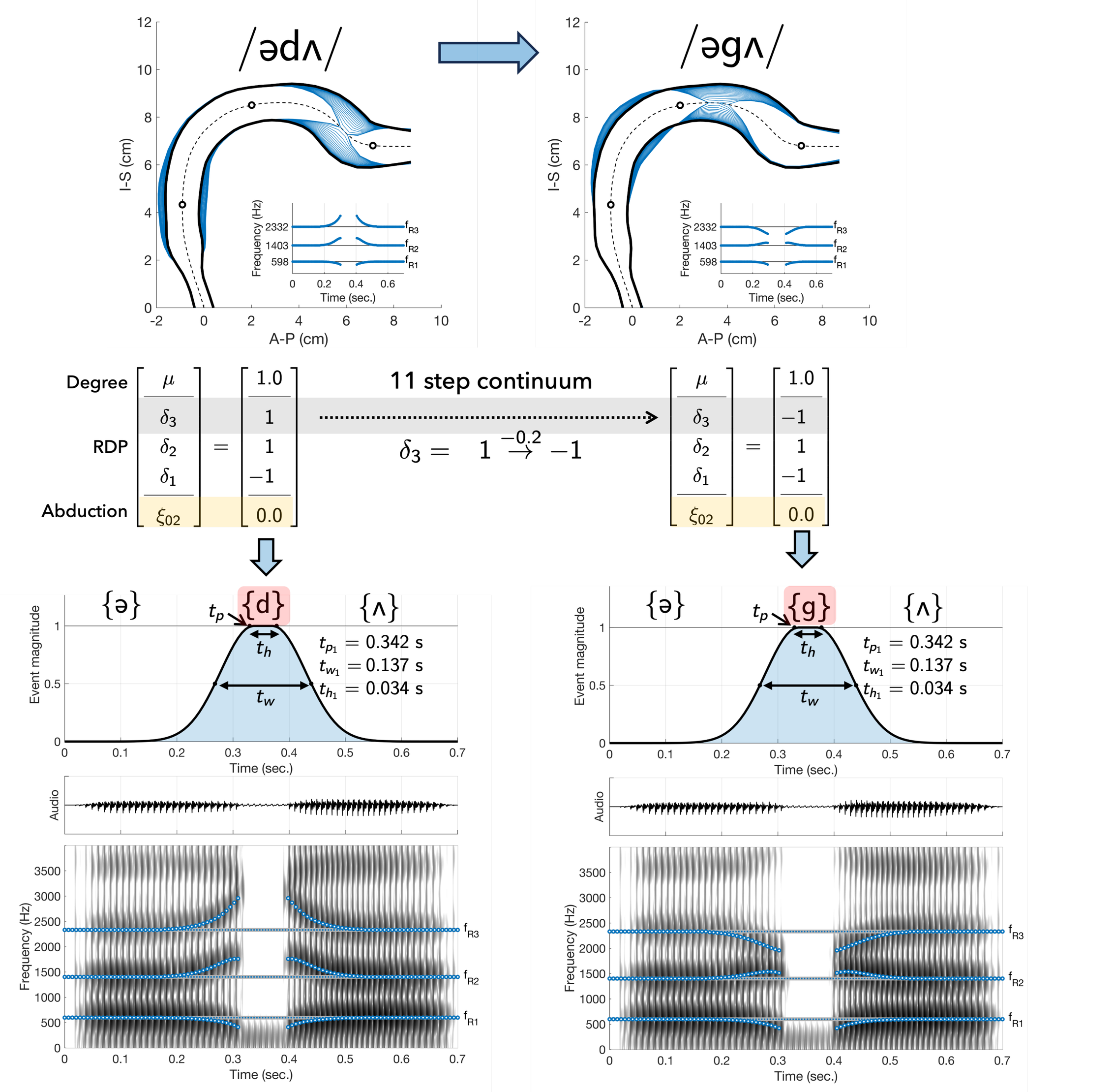

Figure 3: Continuum: {d} → {g}. Adult male vocal tract versions of /ədʌ/ and /əgʌ/ are shown in the upper panel; the inset plots show calculated resonance frequencies as a function of time. Lower panels indicate RDPs, event functions, temporal parameters, synthesized waveforms, and spectrograms.

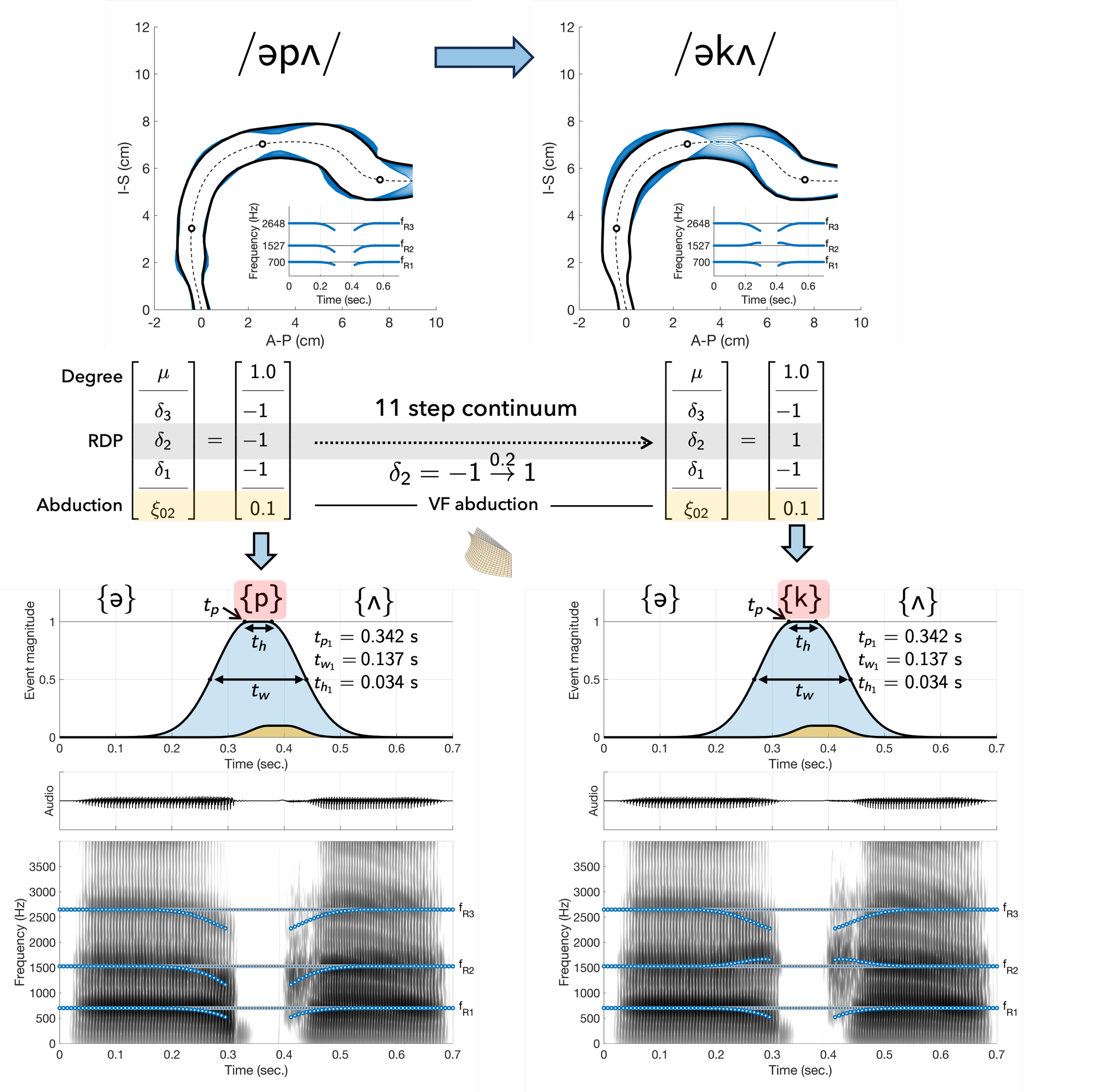

Figure 4: Continuum: {p} → {k}. Adult female vocal tract versions of /əpʌ/ and /əkʌ/ in upper panel. Lower panels show RDPs, event functions, temporal parameters, synthesized waveforms, and spectrograms.

Consonant identification experiment

- Eight VCV continua were generated and presented to listeners: Male/female and voiced/unvoiced versions of δ_3 variation (d→g & t→k) and δ_2 variation (b→g & p→k).
- The ALVIN interface (Hillenbrand and Gayvert, 2005) was used for presentation of the VCVs and collection of listener responses.
- VCVs were presented to 14 naive listeners (12F, 2M, mean age = 22.3) over a loudspeaker in a sound booth. Listeners were asked to identify the consonant using a forced-choice paradigm where they chose from b, d, g, p, t or k.
- Presentation of VCVs was blocked by speech system scaling (male, female) and by δ_3 and δ_2. Within each block the VCV order was randomized and presented five times.

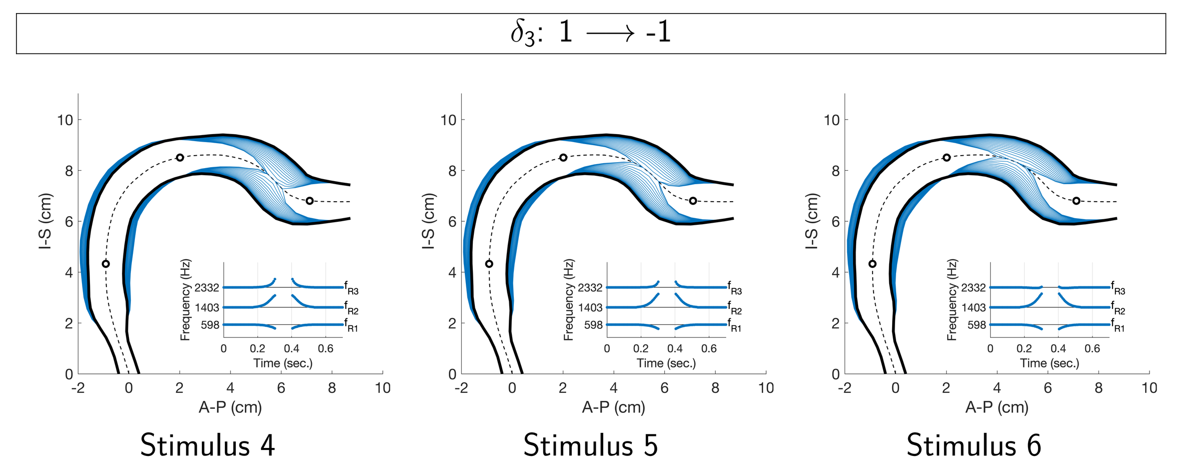

Figure 5: Male vocal tract configurations sampled in the middle of the 11 point VCV continua based on incrementing δ_3 from 1 to -1 (i.e., {d} → {g} and {t} → {k}).

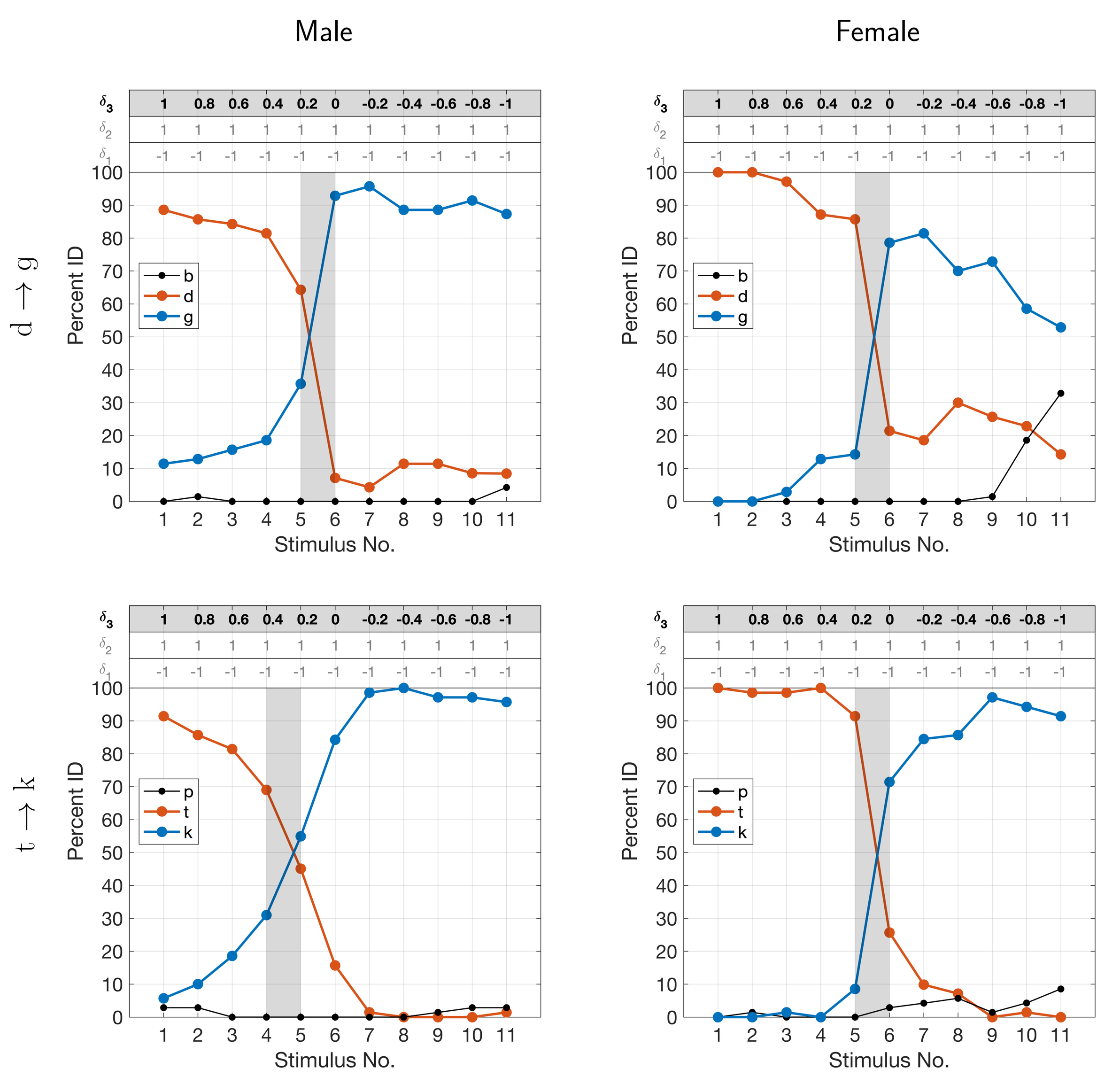

Figure 6: Consonant identification results for 11 point VCV continua based on incrementing δ_3 from 1 to -1 (i.e., {d} → {g} and {t} → {k}).

References

Hillenbrand J., and Gayvert R. T. (2005). Open source software for experiment design and control. J. Spch. Hear. Res., 48, 45-60.

Story, B.H., (2013). Phrase-level speech simulation with an airway modulation model of speech production, Computer Speech and Language. 27(4), 989-1010.

Story, B. H., and Bunton, K., (2017). An acoustically-driven vocal tract model for stop consonant production, Speech Comm., 87, 1-17.

Story, B. H., and Bunton, K., (2019). A model of speech production based on the acoustic relativity of the vocal tract, J. Acoust. Soc. Am., 146(4), 2522-2528.

Story, B. H., and Bunton, K. (2021). Identification of voiced stop consonants produced by acoustically driven vocal tract modulations. JASA Express Letters, 1(8), 085203:1-6.

Consonant identification experiment continued

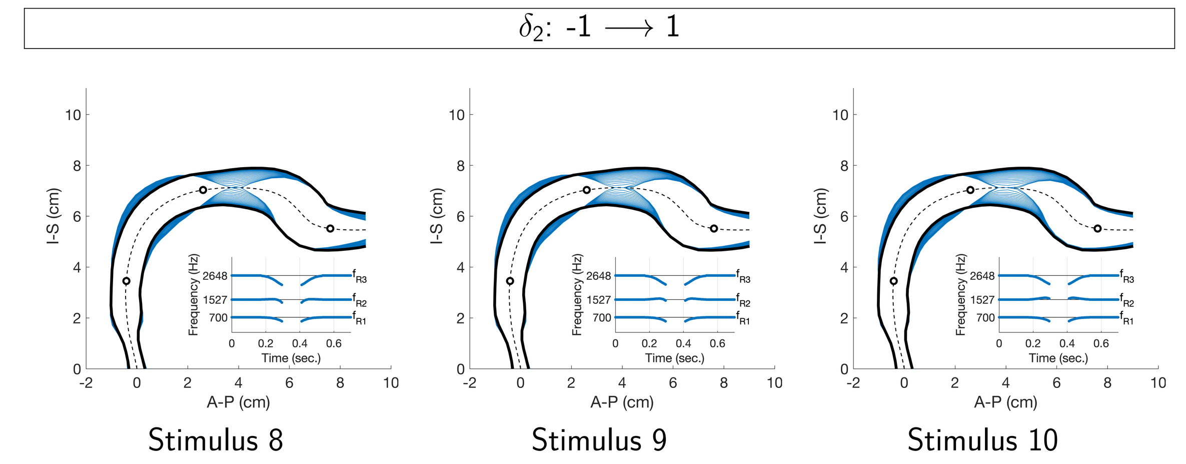

Figure 7: Female vocal tract configurations sampled near the end of the 11 point VCV continua based on incrementing δ_2 from -1 to 1 (i.e., {b} → {g} and {p} → {k}).

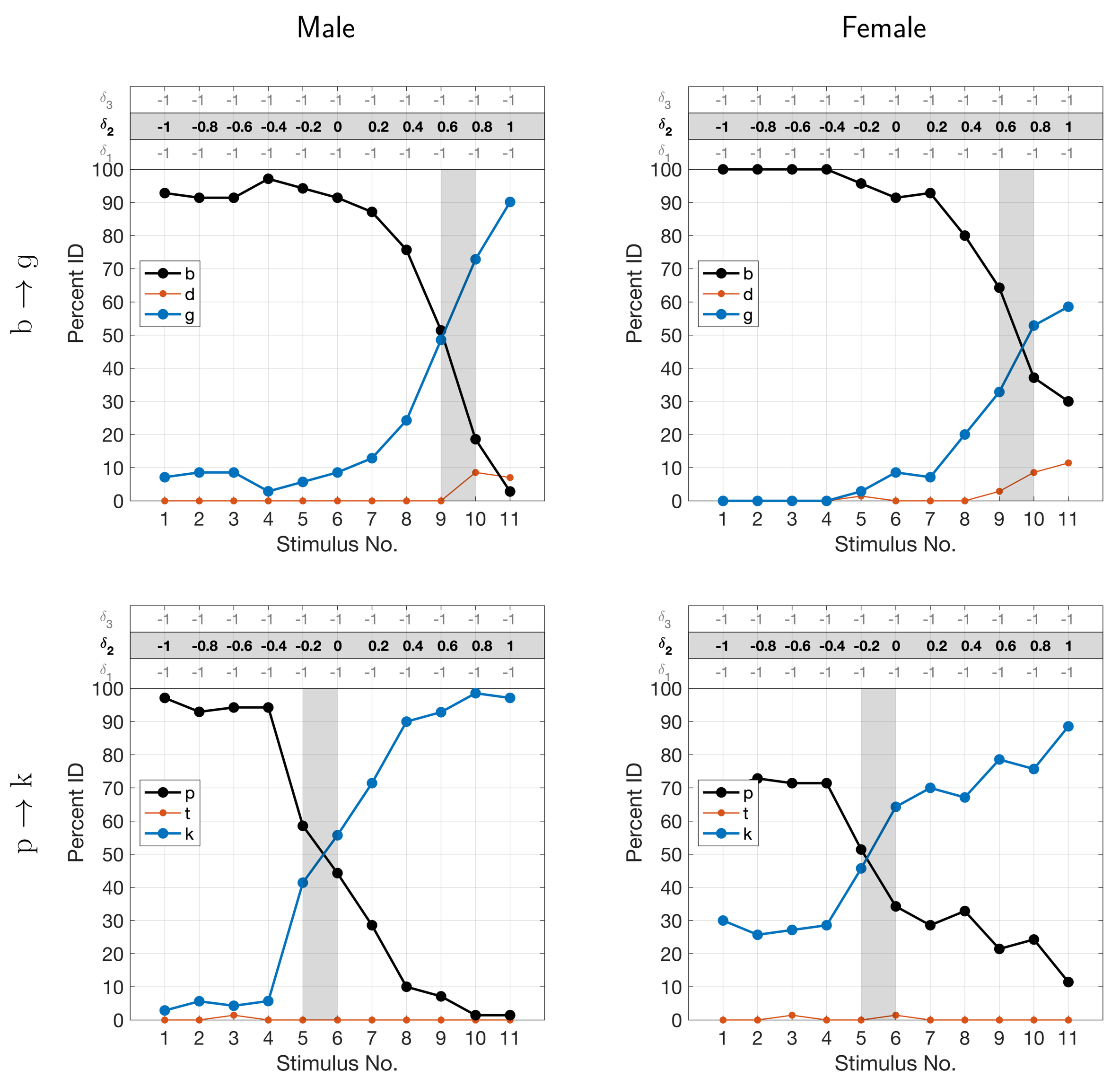

Figure 8: Consonant identification results for 11 point VCV continua based on incrementing δ_2 from -1 to 1 (i.e., {b} → {g} and {p} → {k}).

Conclusions

- For the continua generated by incrementing δ_3 from 1 to -1 (Fig. 6), listener identification switched from an alveolar to a velar consonant between stimuli 5 and 6, except for the male unvoiced condition; in the latter, the switch occurred between stimuli 4 and 5.
- For continua generated by incrementing δ_2 from -1 to 1 (Fig. 8), listener identification switched from a bilabial to a velar consonant between stimuli 9 and 10 for the voiced male and female conditions, but between stimuli 5 and 6 in the unvoiced male and female conditions. This result is surprising considering that the vocal tract modulations were identical for the voiced and unvoiced stimuli along each continuum. As can be seen in Fig. 7, the second resonance is not deflected upward until stimulus 9, even though δ_2 shifted to a positive value at stimulus 7.
- In attempt to understand the previous result, the δ_2 continuum (from 1 to -1) was simulated again, but with $\delta_3 = 0.25$. With this smaller value, the upward deflection of the second resonance was larger when δ_2 was positive. This may shift the perceptual boundary. Future work will include determining whether listeners identify consonants in these stimuli differently.
- The results suggest that RDPs are an effective discrete representation of phonetic segments that can be transformed into intelligible speech by modulation of the vocal tract shape guided by acoustic sensitivity functions, but there is much to understand about how particular combinations of RDPs affect listener responses.

Acknowledgements

Research supported by a Galileo Circle grant. Thanks to Ann Ross for assistance in collection of listener responses.