Simulation of sentence-level speech with an acoustically-driven model of speech production

Brad H. Story & Kate Bunton

Speech, Language, and Hearing Sciences, University of Arizona

QR code to access audio demonstrations of simulated speech

Model of speech production

- In this model of speech production, speech segments are encoded by specifying relative acoustic events along a time axis that consist of directional changes of the vocal tract resonance frequencies called resonance deflection patterns (RDPs) (Story & Bunton, 2019).

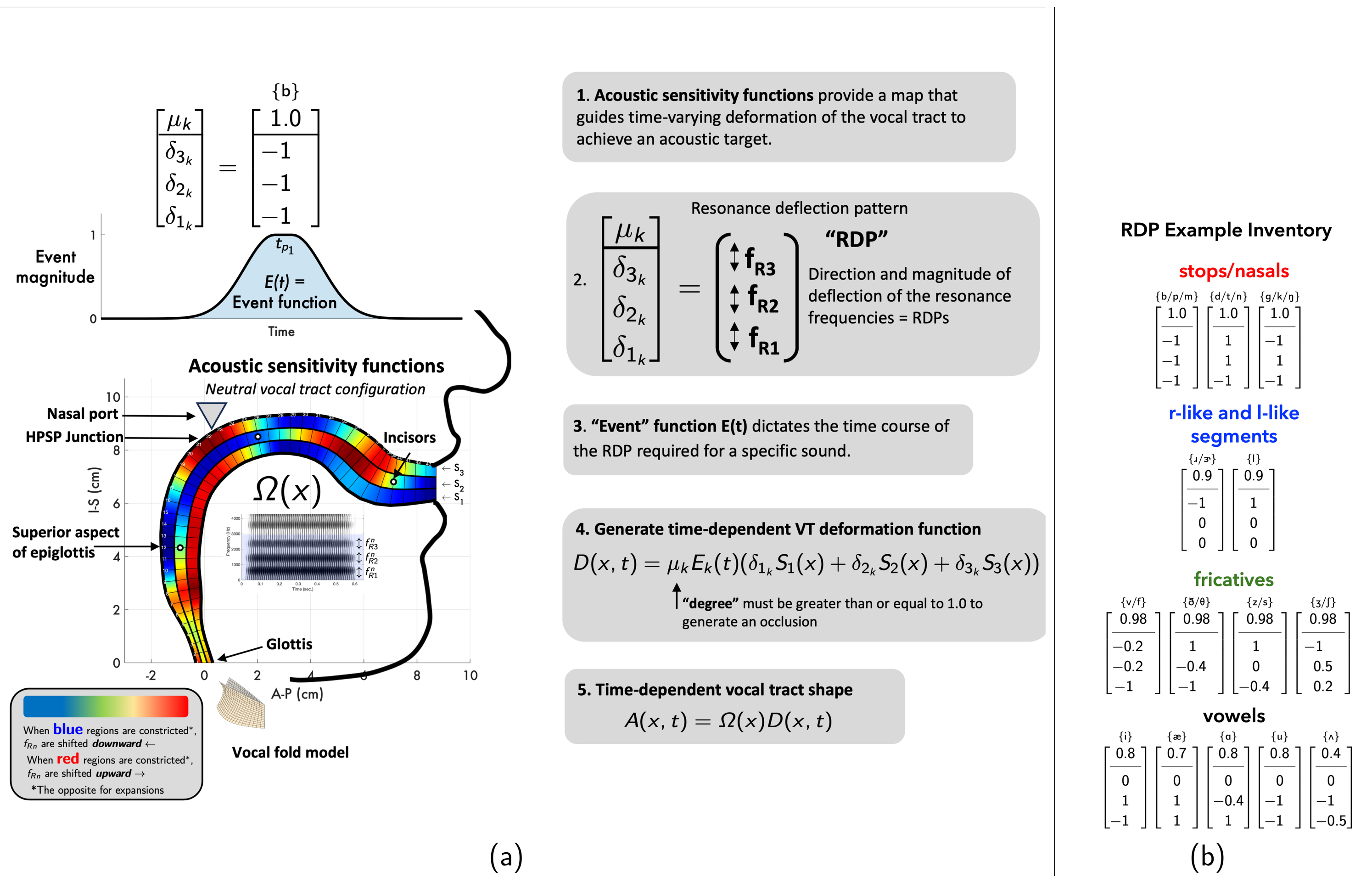

Figure 1: (a) Transformation of a discrete representation of a phonetic segment (RDP) into time-varying vocal tract modulations. (b) RDP inventory for specfication of phonetic segments.

- The **purpose of this study** was to demonstrate the use of this model to generate word-level and sentence-level speech. The aim is to continue to develop the model as a tool for understanding aspects of speech production, speech perception, and speech intelligibility.

Control tiers for speech events

- Simulation of an utterance is controlled by a "bookkeeping" system in which any given event is represented as an encapsulation of its attributes. The first row shows the events for "abracadabra" scheduled along a time axis.
- Other than the initial neutral configuration, an explicit specification of vocal tract shaping parameters (e.g., constriction location) is not required; modulation of the vocal tract is based on achieving the acoustic targets specified by the RDPs.
- Events that modulate glottal adduction, nasal coupling, fundamental frequency, and tracheal pressure are also specified along the time axis as indicated in rows 2-5.

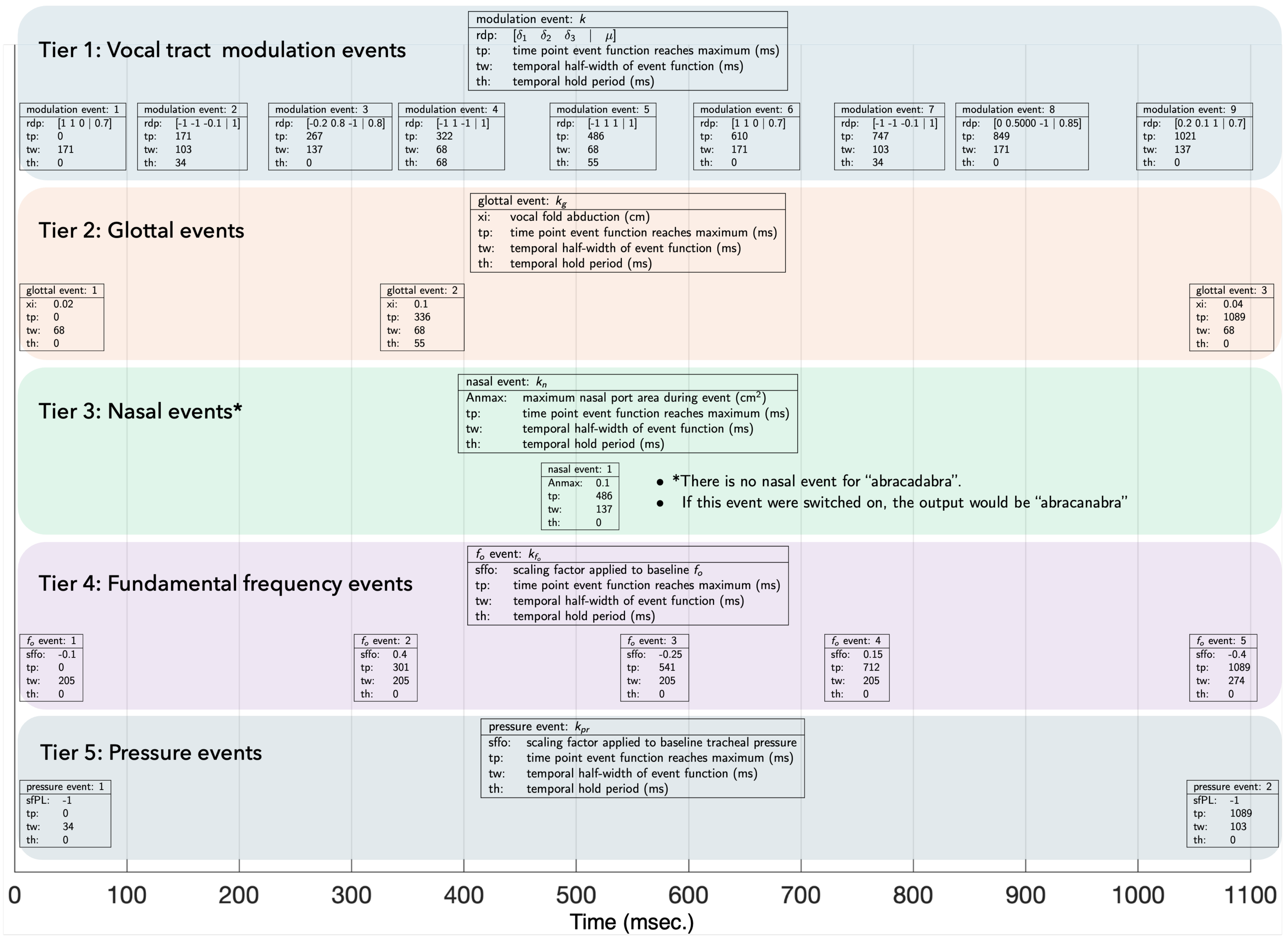

Figure 3: Structure for controlling modulation events in the speech production system.

Example simulations

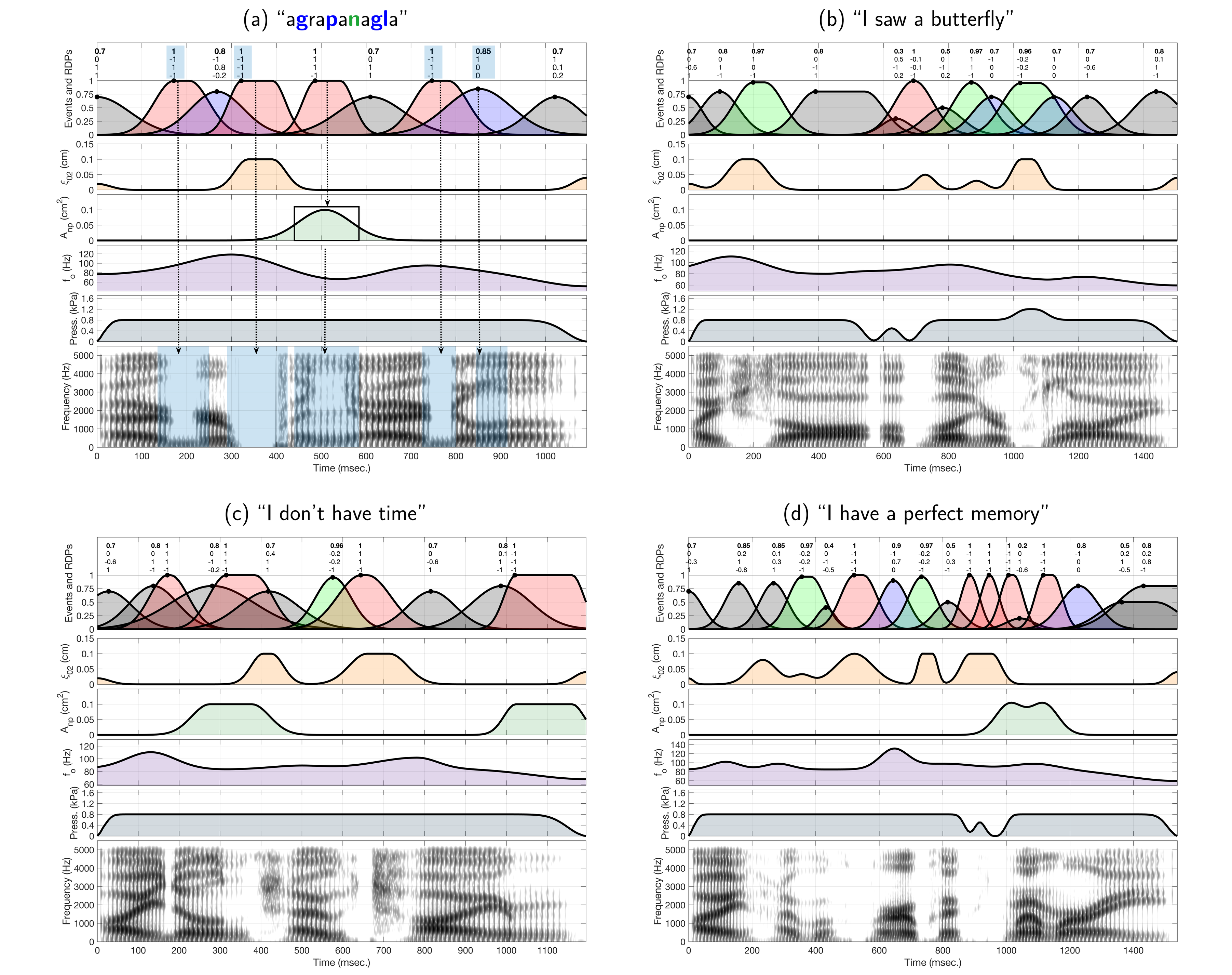

Figure 5: (a) Switch-flipping demonstration that transforms "abracadabra" into "agrapanagla." (b-d) Demonstrations of simulating sentence-level speech. RDP color code: gray = vowel, red = stop, blue = liquid, green = fricative.

Timing of simulated phonetic events

- Temporal parameters of the RDP (vocal tract) events are based on visual inspection of the waveform and spectrogram of natural speech, along with the time derivative of the Hilbert transform envelope. The example in Fig. 2 is based on a recording of **"abracadabra"** produced by an adult male talker.

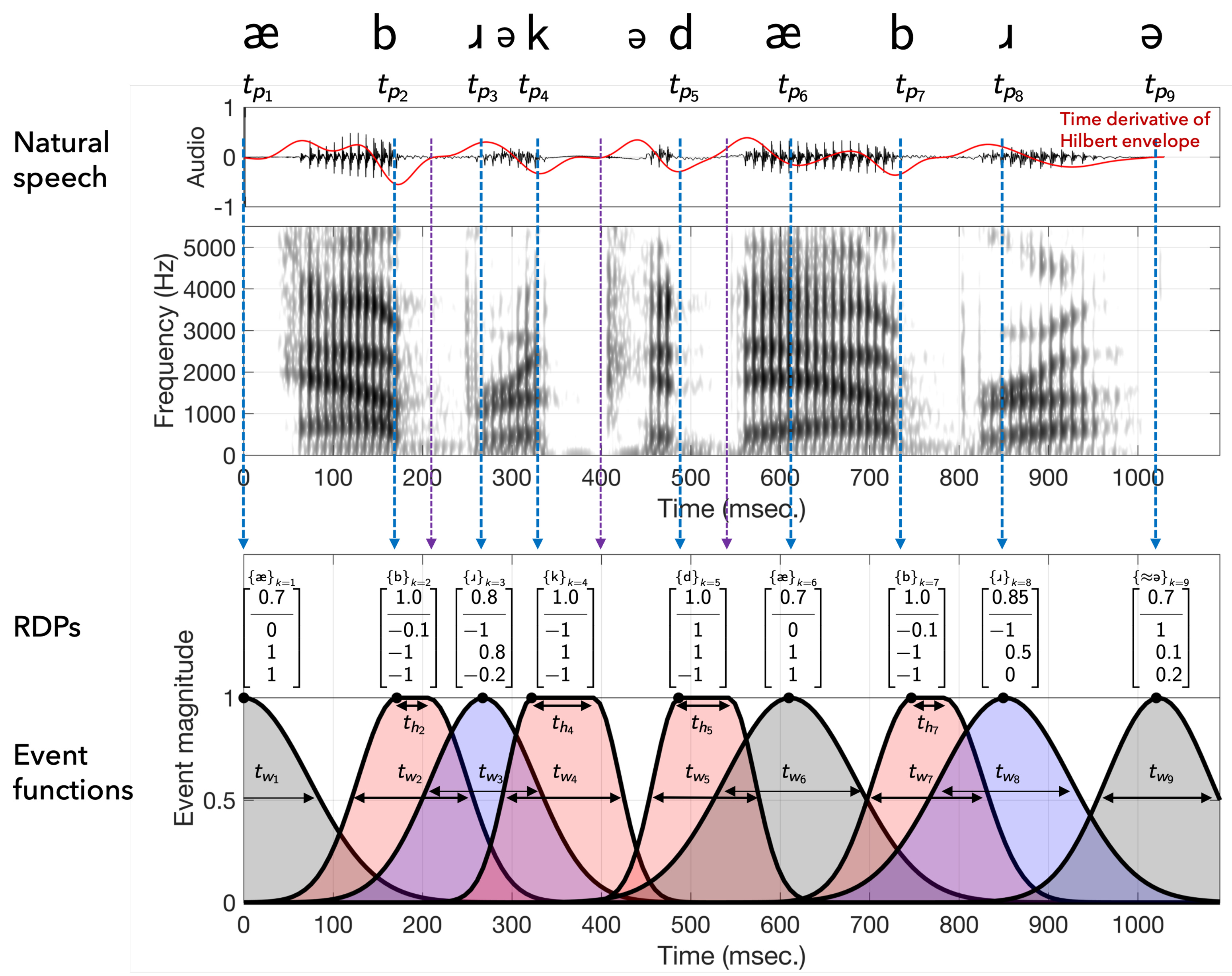

Figure 2: Determining temporal parameters of event functions from natural speech. Color code: gray = vowel, red = stop, blue = r-like or l-like.

Simulation of speech

- The RDP events for "abracadabra," along with the temporal variation of vocal fold abduction, nasal port area, fundamental frequency, and tracheal pressure are shown in Fig. 4a. To indicate their prominence, the event functions have been scaled by their respective degrees (μ). The bottom panel displays a WB spectrogram of the resulting simulated speech.
- A 3D version of the time-varying area function resulting from the overlapping RDP events is shown in the upper part of Fig. 4b. Vocal tract shape variations are also shown in the lower part but in the form of a pseudo-midsagittal plot and 2D area function plot; also shown are the time varying vocal tract resonances.
- The RDPs can be conceived as a bank of switches in which a change from one switch pattern to another generally evokes a predictable perceptual response.

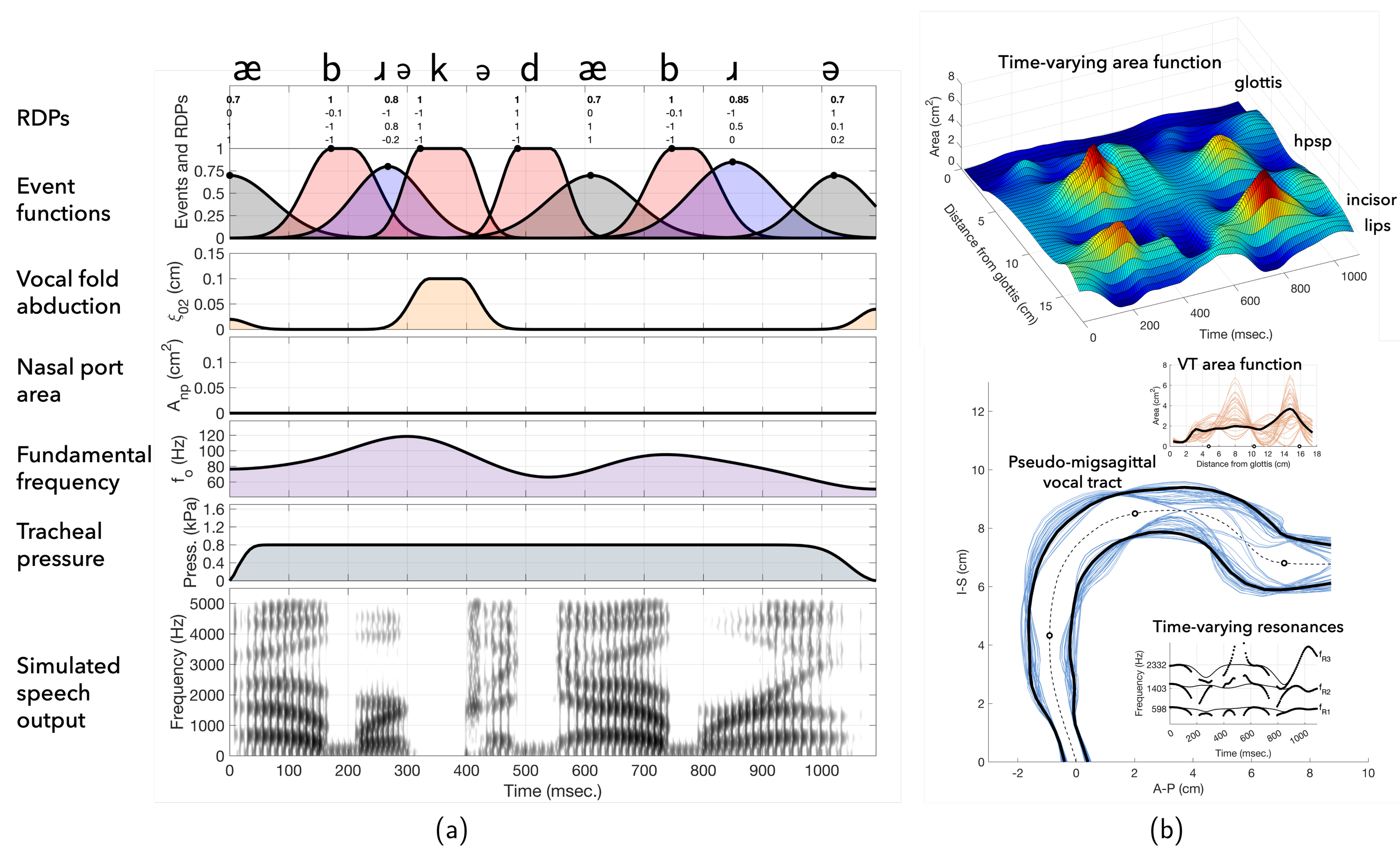

Figure 4: (a) Event plot for "abracadabra," (b) Time-varying area function in 3D and 2D forms, along with calculated vocal tract resonances.

Long sentence simulation and conclusion

- As a final demonstration, the simulated sentence shown in Fig. 6 contains 28 RDP events over about 2.2 seconds that generates 11 syllables (5 syl/sec). These events are augmented by glottal, nasal, fundamental frequency and pressure events.
- The simulations 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.
- The QR code in the upper right corner of this poster provides a link to a website where all simulations can be played.

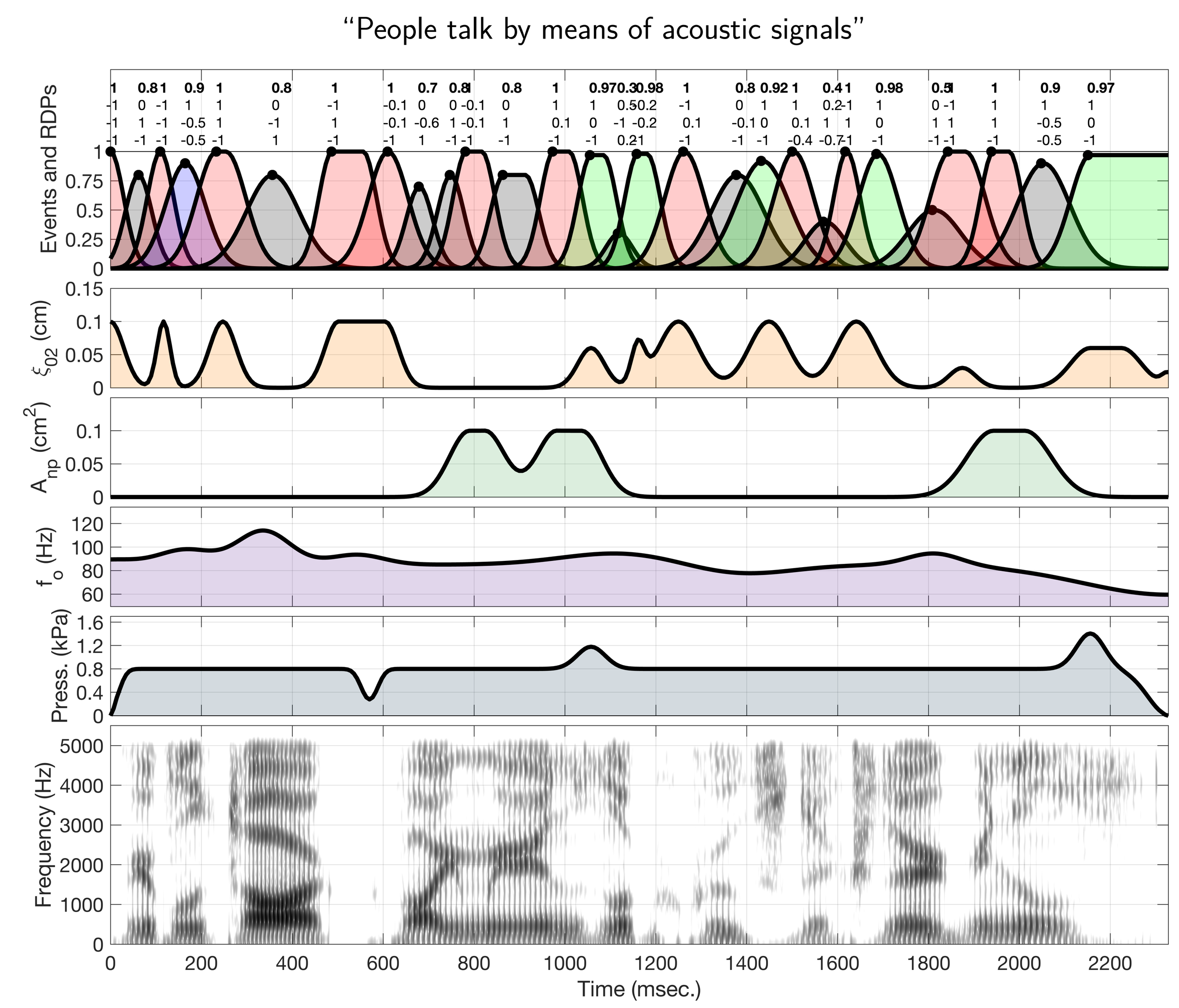

Figure 6: Simulated sentence. RDP color code: gray = vowel, red = stop, blue = liquid, green = fricative.

References

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.

Acknowledgements

Research supported in part by NIH 5R01DC017998 and Galileo Circle Fellows grant from the University of Arizona.