Intelligibility of synthetic words generated by transformation of a sequence of discrete acoustic events into modulation of the vocal tract shape

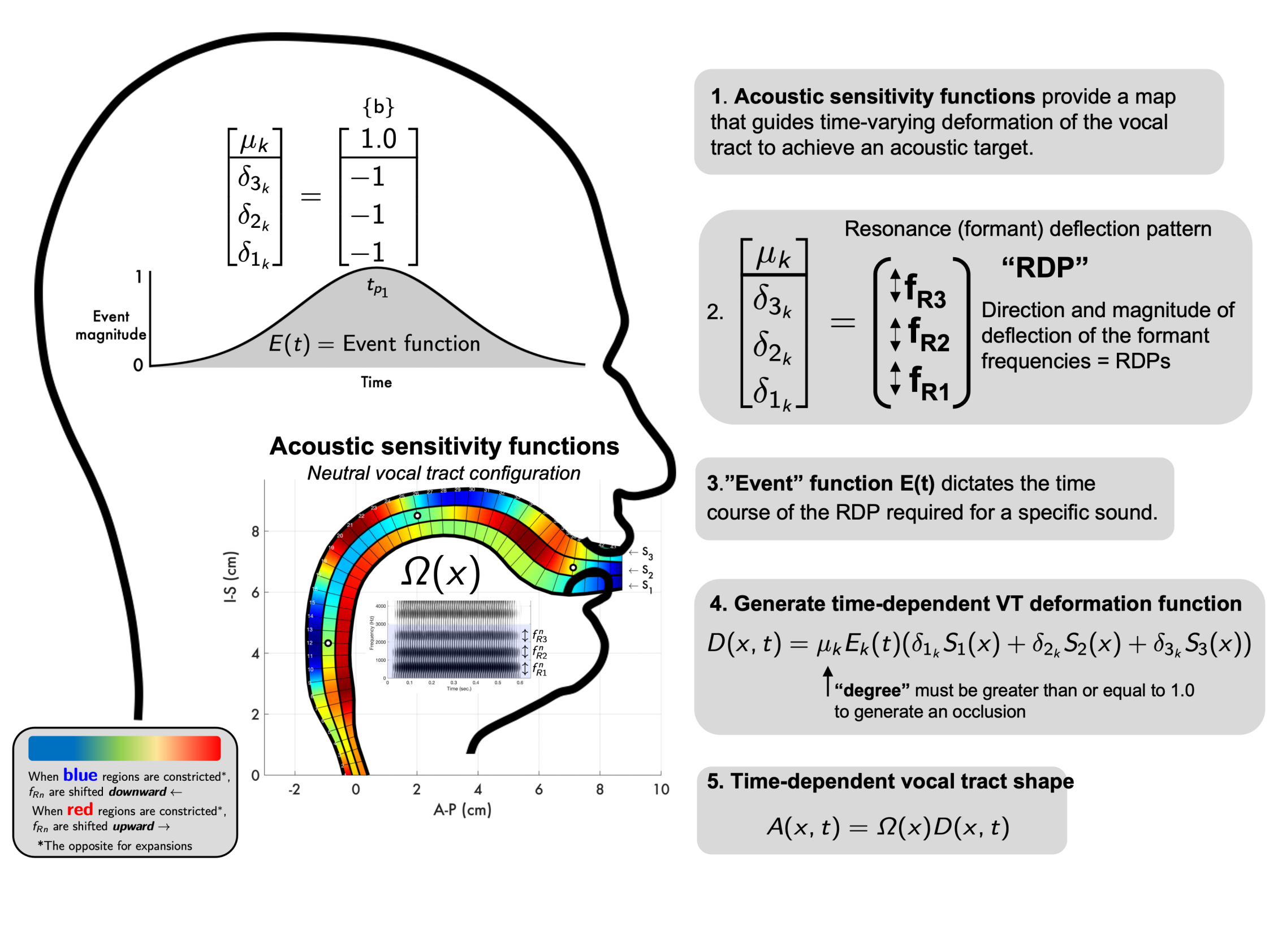
Brad H. Story & Kate Bunton

Speech, Language, and Hearing Sciences, University of Arizona

Story, B. H., and Bunton, K. (2022). Intelligibility of synthetic words generated by transformation of a sequence of discrete acoustic events into modulation of the vocal tract shape, Presented at the 183rd Meeting of the Acoustical Society of America, JASA, 152, A173; https://doi.org/10.1121/10.0015931

1. Background

- 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 consist of directional changes of the vocal tract resonance frequencies called resonance deflection patterns (RDPs) that, when associated with a temporal event function, are transformed via acoustic sensitivity functions, into time-varying modulations of the vocal tract shape.


{b}
$\begin{bmatrix}\mu_k\\ \delta_{3_k}\\ \delta_{2_k}\\ \delta_{1_k}\end{bmatrix} = \begin{bmatrix}1.0\\ -1\\ -1\\ -1\end{bmatrix}$
1
Event magnitude
0
t_{p_1}
$E(t)$ = Event function
Time
Acoustic sensitivity functions
Neutral vocal tract configuration
$\Omega(x)$
I-S (cm)
A-P (cm)
S_3
S_2
S_1
When blue regions are constricted*, f_{Rn} are shifted downward ←
When red regions are constricted*, f_{Rn} are shifted upward →
*The opposite for expansions

1. **Acoustic sensitivity functions** provide a map that guides time-varying deformation of the vocal tract to achieve an acoustic target.

2. Resonance (formant) deflection pattern

$$\begin{bmatrix}\mu_k\\ \delta_{3_k}\\ \delta_{2_k}\\ \delta_{1_k}\end{bmatrix} = \begin{pmatrix}\updownarrow f_{R3}\\ \updownarrow f_{R2}\\ \updownarrow f_{R1}\end{pmatrix}$$

"RDP" Direction and magnitude of deflection of the formant frequencies = RDPs

3. **"Event" function E(t)** dictates the time course of the RDP required for a specific sound.

4. **Generate time-dependent VT deformation function**

$$D(x,t) = \mu_k E_k(t)(\delta_{1_k}S_1(x) + \delta_{2_k}S_2(x) + \delta_{3_k}S_3(x))$$

"degree" must be greater than or equal to 1.0 to generate an occlusion

5. **Time-dependent vocal tract shape**

$$A(x,t) = \Omega(x)D(x,t)$$

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

1.1. Example 1

- In this example, RDPs intended to represent phonetic targets /b/, /ɪ/, /d/ are transformed via overlapping event functions into a time-varying vocal tract area function, and subsequently synthetic speech via the TubeTalker system (Story, 2013).
- Other than the initial neutral configuration, 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.

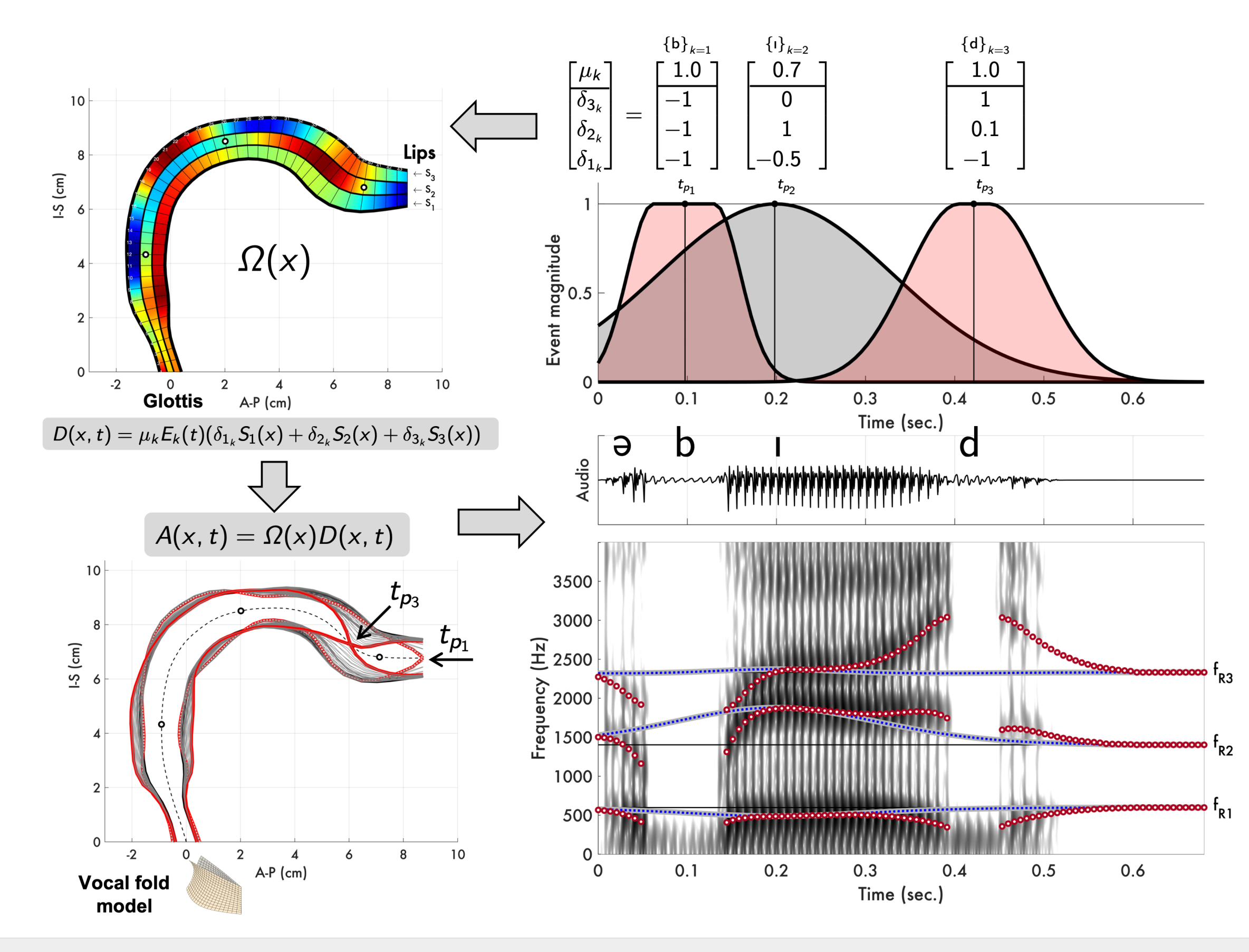


Figure 2: Transformation of three sequentially-arranged RDPs into the synthesized utterance "a bid".

1.2. Example 2

- In this example, RDPs are specified to represent phonetic targets /t/, /ɪ/, /k/.
- Unlike Example 1, there are additional timing functions associated with ξ_{02}, the degree of vocal fold adduction, that generate an abductory maneuver during the initial and final consonants so that they are unvoiced. These are shown in green in the upper right panel of Figure 3.

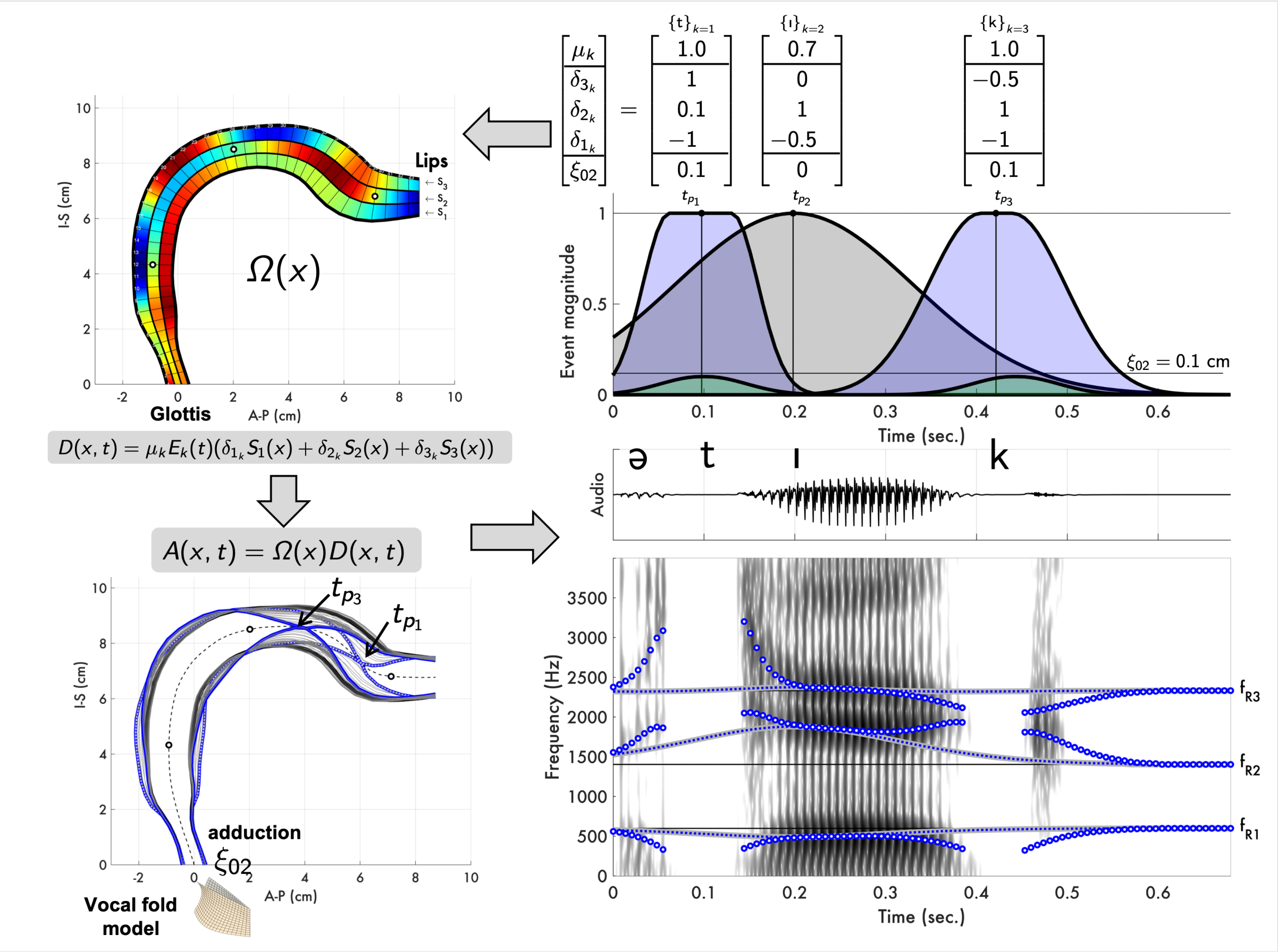


Figure 3: Transformation of three sequentially-arranged RDPs into the synthesized utterance "a tick".

2. Specific Aim

- The **specific aim of this study** was to determine if listeners' recognition of synthetic words was aligned with the RDP settings. That is, if RDPs are conceived as a bank of switches, can word recognition be guided by changes in the pattern of switch settings.

3. Synthesis of C_1VC_2 words

- C_1VC_2 words were generated by combining RDPs from the first row below for C_1 and C_2 with an RDP from the second row for the vowel V. The RDPs were sequenced in exactly the same temporal pattern and associated with the same event functions as demonstrated in Examples 1 & 2.
- With three consonantal RDPs that can be voiced or unvoiced, along with four possible vowels, 144 C_1VC_2 synthetic words were generated.

RDP inventory: Stop consonants

$$\begin{bmatrix}\mu_k\\ \delta_{3_k}\\ \delta_{2_k}\\ \delta_{1_k}\\ \xi_{02}\end{bmatrix} = \overset{\{b/p\}}{\begin{bmatrix}1.0\\ -1\\ -1\\ -1\\ 0/0.1\end{bmatrix}} \text{ and/or } \overset{\{d/t\}}{\begin{bmatrix}1.0\\ 1\\ 0.1\\ -1\\ 0/0.1\end{bmatrix}} \text{ and/or } \overset{\{g/k\}}{\begin{bmatrix}1.0\\ -0.5\\ 1\\ -1\\ 0/0.1\end{bmatrix}}$$

RDP inventory: Vowels

$$\begin{bmatrix}\mu_k\\ \delta_{3_k}\\ \delta_{2_k}\\ \delta_{1_k}\end{bmatrix} = \overset{\{ɪ\}}{\begin{bmatrix}0.7\\ -0\\ -1\\ -0.5\end{bmatrix}} \text{ or } \overset{\{ɛ\}}{\begin{bmatrix}0.4\\ 0\\ 1\\ 0.3\end{bmatrix}} \text{ or } \overset{\{æ\}}{\begin{bmatrix}0.8\\ 0\\ 1\\ 1\end{bmatrix}} \text{ or } \overset{\{ʌ\}}{\begin{bmatrix}0.5\\ -0.5\\ -1\\ 0.1\end{bmatrix}}$$

4. Word intelligibility experiment

- Based on a survey of 10 undergraduate students, nonsense, unusual, or inappropriate American English words were eliminated from the collection of 144 C_1VC_2 synthetic words, leaving the 66 words shown in the table below as stimuli for the listening experiment.

Vowel	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
/ɪ/	bɪb	bɪd	bɪg	bɪt	dɪp	dɪd	dɪg	gɪg	pɪg	pɪt	pɪk	tɪp	tɪk	kɪd	kɪt	kɪk			
/ɛ/	bɛd	bɛg	bɛt	dɛd	dɛt	dɛk	gɛt	pɛg	pɛt	pɛk	tɛk	kɛg							
/æ/	bæd	bæg	bæt	bæk	dæb	dæd	gæp	gæg	pæd	pæt	pæk	tæb	tæd	tæg	tæp	tæk	cæb	cæp	cæt
/ʌ/	bʌd	bʌg	bʌt	bʌk	dʌd	dʌg	dʌk	gʌt	pʌb	pʌg	pʌp	pʌt	pʌk	tʌb	tʌg	tʌk	cʌb	cʌp	cʌt

- Word intelligibility was tested by presenting target C_1VC_2 words to 11 naive listeners (7F, 4M, mean age = 21) over a loudspeaker in a sound booth. Listeners were asked to indicate what they heard by choosing a word from a matrix that included the target and near-neighbor words. Word presentation was blocked by vowel and randomized. Each word was presented five times.
- ALVIN (Hillenbrand et al., 2015) was used for presentation of the words and collection of listener responses. Example matrices are shown in Figure 4.

Vowel /æ/

Vowel /ʌ/

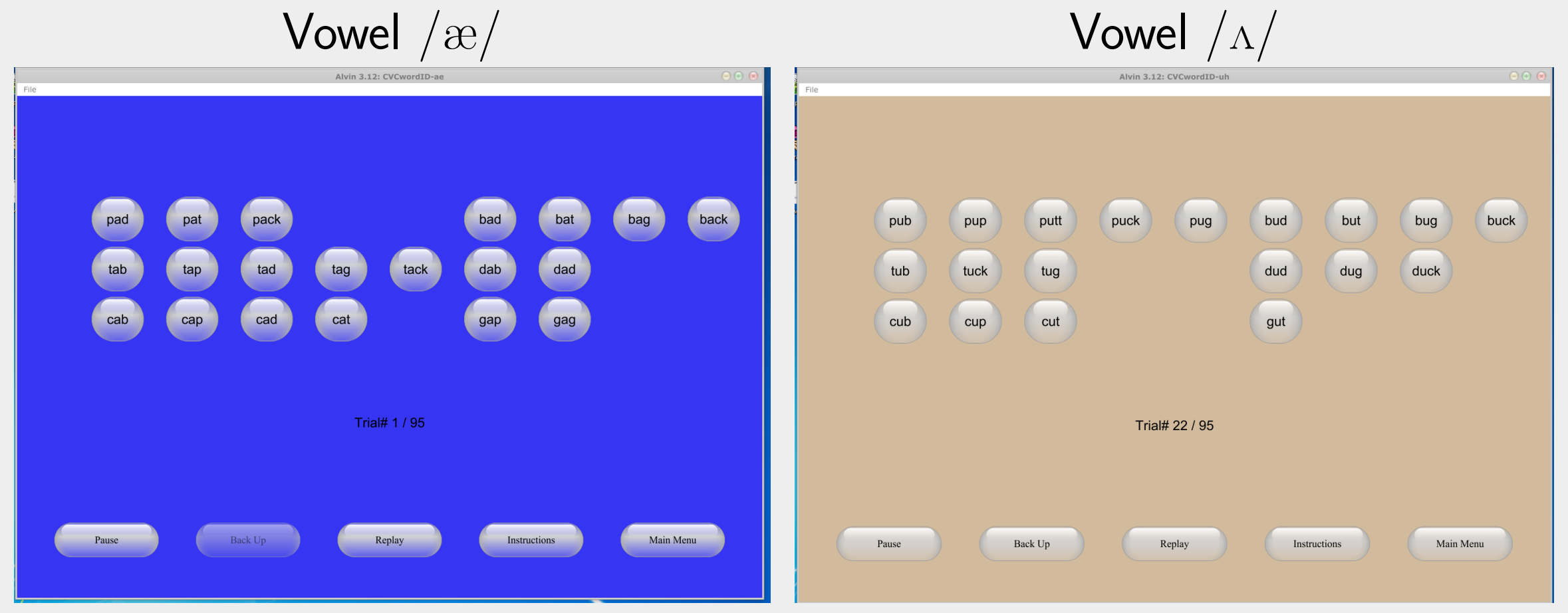


Figure 4: Examples of the user interface in ALVIN.

5. Results

- The listener responses for words in each vowel context are shown in the form of confusion matrices. All data is presented as percentage of the total number of responses to each word. The target words are indicated on the vertical axis of a matrix and the listener responses are shown on the horizontal axis.
- The values along the diagonal indicate the percentage of trials in which the listener heard the intended target word. The off-diagonal values are the percentage of trials heard as something other than the intended target. The green and yellow highlight indicate place and voicing errors, respectively.
- A confusion matrix for words in the /ʌ/ vowel context is not shown because the listener responses were 99.9% aligned with the target words

Vowel context: /ɪ/

Listener Response (97.4% aligned with target)

Target word

	bɪb	bɪd	bɪg	bɪt	dɪp	dɪd	dɪg	gɪg	pɪg	pɪt	pɪk	tɪp	tɪk	kɪd	kɪt	kɪk
bɪb	100															
bɪd		100														
bɪg			100													
bɪt				100												
dɪp					95			4							1	
dɪd		2				98										
dɪg							89	11								
gɪg								100								
pɪg									100							
pɪt										96		2			2	
pɪk									2		96					2
tɪp												96			4	
tɪk													89			11
kɪd														100		
kɪt															100	
kɪk																100

5. Results cont'd

Vowel context: /ɛ/

Listener Response (99.7% aligned with target)

Target word

	bɛd	bɛg	bɛt	dɛd	dɛt	dɛk	gɛt	pɛg	pɛt	pɛk	tɛk	kɛg
bɛd	100											
bɛg		100										
bɛt	2		98									
dɛd				100								
dɛt					98							2
dɛk						100						
gɛt							100					
pɛg								100				
pɛt									100			
pɛk										100		
tɛk											100	
kɛg												100

Vowel context: /æ/

Listener Response (97.8% aligned with target)

Target word

	bæd	bæg	bæt	bæk	dæb	dæd	gæp	gæg	pæd	pæt	pæk	tæb	tæd	tæg	tæp	tæk	cæb	cæp	cæt
bæd	98								2										
bæg		98						2											
bæt		2	96							2									
bæk				98					2										
dæb					100														
dæd					2	96			2										
gæp							98	2											
gæg							2	98											
pæd									100										
pæt								2		98									
pæk											100								
tæb												93			2		5		
tæd						2							98						
tæg														100					
tæp												2			93			5	
tæk																100			
cæb																	98	2	
cæp									2								2	96	
cæt																			100

6. Conclusions

- In general, the intelligibility of CVC words generated by specification of RDPs was greater than 97% across the four vowel contexts. Many of the cases where listener response were not aligned with the intended target word could be explained by single voicing or place errors.
- The results suggest that RDPs are an effective discrete representation of phonetic segments that can be transformed into speech by modulation of the vocal tract shape guided by acoustic sensitivity functions. The RDPs can be conceived as a bank of switches in which a change from one switch pattern to another evokes a predictable perceptual response. In theory, any given switch pattern specified by an RDP is relative to the acoustic characteristics of any phonetically-neutral, idiosyncratic configuration of the vocal tract (e.g., male, female, child).
- Future work will include repeating the experiment for differently-scaled speech production systems.

Acknowledgements

Research supported by a College of Science Galileo Circle grant. Thanks to Shelby Kramer for assistance in collection of listener responses.

References

Hillenbrand, J. M., Gayvert, R. T., & Clark, M. J. (2015). Phonetics exercises using the Alvin experiment-control software. Journal of Speech, Language, and Hearing Research, 58(2), 171-184.

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.