Automatic analysis of phonemic context-dependent cue productions in acoustic cue-labeled speech

Jeung-Yoon Choi, Sofie Chung, and Stefanie Shattuck-Hufnagel

Massachusetts Institute of Technology

This work describes a framework for automatic analysis of context-dependent cue productions in acoustic cue-labeled speech, enabling detailed investigation of context-governed patterns of subphonemic variation. In the process described here, context refers to the underlying phonological environment provided by the word sequence, while cue production includes the implemented acoustic cues present in a specific production (as opposed to the cues that might be predicted from the lexical entries, or that might occur in a different utterance). As an illustrative application of the analysis method, cue labels for a subset of the TIMIT database (Garofalo et al. 1990) were analyzed for context-dependent cue production patterns. This subset of the database was labeled with a set of 40 basic acoustic cues (Huilgol et al. 2019), which include four categories of cues: 8 types of Landmark acoustic cues (Stevens 2002), which are related to manner features; 9 types of vowel/glide place cues; 15 types of consonant place cues; and 8 types of nasalization and glottal configuration cues.

In this approach, in order to analyze the cue production pattern for a cue-labeled sentence, a baseline cue production pattern is first generated from the underlying phoneme sequence (Soh et al. 2019). This baseline pattern is then aligned/matched with the entire set of cues that appears in the cue-labeled file for the current production; this step is required because some predicted cues are missing and non-predicted cues may be added. In the analysis described here, context was constrained to the immediate phonemic environment, defined as a phoneme triple [previous phoneme, target phoneme, following phoneme]. In order to study the patterns of cue variation for a particular phonemic segment, e.g. /t/, in a particular context of this type, we identify all of the tokens of [‘*’, ‘t’, ‘*’], where ‘*’ indicates any phoneme, and extract the aligned acoustic cues that were produced for /t/ in all of these contexts (Torres 2023).

In the current work, the focus is on examining just one of the four categories of cues, i.e. the Landmark cue productions; other acoustic cues are left for future analysis. An example of a possible Landmark cue production for /t/ in this context is the default production of a Stop closure followed by a Stop release, noted as the context-production pair [‘*’, ‘t’, ‘*’]:[‘Sc’, ‘Sr’]. Another possible cue production pattern, traditionally transcribed as a flap, would be annotated in this framework as the presence of a Glide Landmark for an intervocalic /t/, because Glide Landmarks occur when the vocal tract is narrowed but not enough to stop airflow or generate noise. Such a token is denoted as a [‘V’, ‘t’, ‘V’]:[‘Sc*’, ‘G-+’, ‘Sr*’] context-production pair. The symbol ‘Sc*’ indicates either the realization or the absence of a Stop closure for the /t/, and similarly for ‘Sr*’. The ‘G-+’ symbol indicates the presence of a Glide-type acoustic cue for a flapped /t/. The option to include all combinations of Stop closure – Glide – Stop release is critical for capturing cases in which a Stop closure and/or Stop release can be observed along with a Glide-like minimum (Yun et al. 2020).

Using this automatic analysis method, it is possible to identify all context-production pairs for a given phonological segment in a given cue-labeled speech sample, and group them according to the cue patterns that were actually produced. In this representative study of /t/ production, it was possible to quantitatively describe the patterns of systematic subphonemic variation in the production of Landmark cues and the contexts in which they occur, including the default production (stop closure and release), flapping, stop closure absence, stop release absence (unreleased /t/), and fricativization.

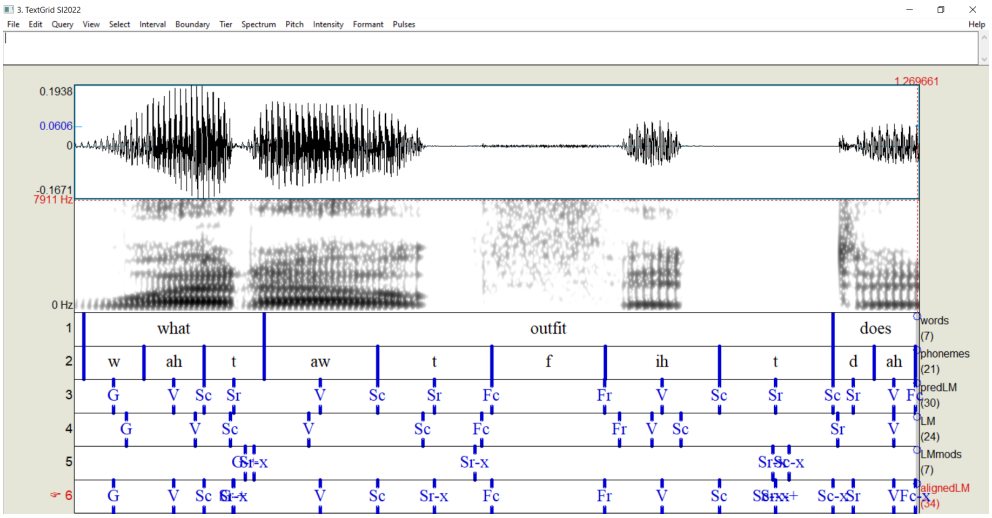

Fig. 1 Praat TextGrid showing tiers for words and associated phonemes, as well as predicted (baseline), realized, modified, and aligned Landmark acoustic cues from an example TIMIT sentence

context	production	wordpos	type	count	speakers
('V', 't', 'V')	**('Sc', 'G-+', 'Sr-x')**	**word-final**	['flapping', 'Sr absence']	1	['DR2/FAEM0/SI2022']
('V', 't', 'F')	**('Sc', 'Sr-x')**	**word-medial**	['Sr absence']	1	['DR2/FAEM0/SI2022']
('V', 't', 'S')	**('Sc', 'Sr-x')**	**word-final**	['Sr absence']	1	['DR2/FAEM0/SI2022']

Fig. 2 Results of automatic analysis of all /t/ productions for the TIMIT sentence in Fig. 1. Out of 3 underlying /t/ phonemes, the algorithm extracts and tabulates 1 production of a flapped /t/ with a stop closure but without a stop release, and 2 productions of /t/ without a stop release

References

Garofalo, J. S., Lamel, L. F., & Fisher, W. M. (1990). The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NIST.

Huilgol, S., Baik, J., & Shattuck-Hufnagel, S. (2019). A framework for labeling speech with acoustic cues to linguistic distinctive features. *The Journal of the Acoustical Society of America*, *146*(2), EL184-EL190.

Soh, C., Talkar, T., Choi, J. Y., & Shattuck-Hufnagel, S. (2019, December). Toward a feature-cue-based analysis of modification patterns in speech: Alignment of canonical and realized acoustic cue labels. In *Proceedings of Meetings on Acoustics* (Vol. 39, No. 1). AIP Publishing.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*(4), 1872-1891.

Torres, D. C. (2023). *An algorithm for characterizing context-governed speech production patterns*, MEng thesis, Massachusetts Institute of Technology.

Yun, S., Choi, J. Y., & Shattuck-Hufnagel, S. (2020). A landmark-cue-based approach to analyzing the acoustic realizations of American English intervocalic flaps. *The Journal of the Acoustical Society of America*, *147*(6), EL471-EL477.