3aSC5. Spectro-temporal cues for perceptual recovery of reduced syllables from continuous, casual speech

Laura Dilley, Meisam K. Arjmandi, Zachary Ireland, Matt Lehet

Michigan State University, Department of Communicative Sciences and Disorders, East Lansing, MI

ABSTRACT

Function words may be highly reduced, with little to no discontinuity marking their onsets to cue their segmentation from continuous speech. We investigated whether reduced function words lacking onset discontinuities have residual timing cues that could be used for word segmentation. Participants (n = 51) briefly viewed sentences and spoke them from memory to elicit casual speech. They were randomly assigned to either a "function-word present" condition (n = 29) in which experimental items contained a critical function word expected to frequently blend spectrally with context, or a "function-word absent" set (n = 22) with phonetically matched items lacking the critical word. Acoustic analyses confirmed that in "function-word present" sentences, critical words lacked detectable onset discontinuities 60% of the time. Critically, in the "function-word present" condition, portions of speech containing critical function words were longer, both in terms of absolute duration and normalized for context speech rate, compared with matched portions in the "function-word absent" condition, even when the former were highly reduced and lacked onset discontinuities. These findings suggest that relative duration cues provide substantial information which may be used by listeners for segmentation of highly reduced syllables from continuous speech.

BACKGROUND

- Function words (*a*, *are*, *our*, *or*, or *her*) are very critical in comprehending the content of speech
- **Segmentation** of words such as function words in continuous speech is a perceptual challenge (Klatt, 1979; Dilley et al., 1996).
  - Function words are frequently **co-articulated** in sentences such as:
    Deena didn't have any leisure or time.
  - Even when producing sentences containing function words, speakers produced many instances of no-discontinuity and relatively few instances of clear discontinuity (Fig. 1) (Dilley et al., 2016).
  - In instances of co-articulation many **acoustic cues** linked to **boundary detection** and **word segmentation** are not present (Stevens, 2000; Drullman, 1994).
  - Cues such as spectral information surrounding the function word and the distal speech rate (timing of non-adjacent context speech) affect whether listeners hear function words (e.g., Dilley & Pitt, 2010).

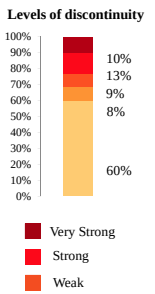
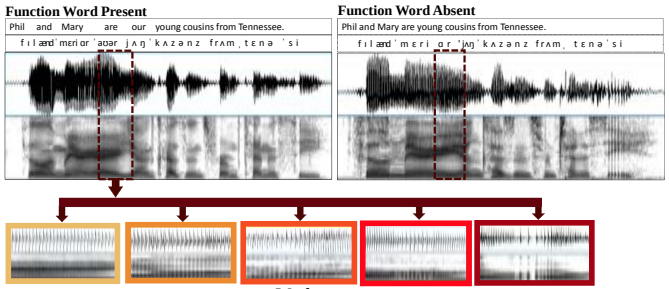


Figure 1. Percentage of words in " function word present" condition by degree of discontinuity.

MATERIALS AND METHODS

- Two corpuses of spoken sentences with either:
  - **Function words present** (*100 sentences, 29 speakers*), or
  - **Function words absent** (*100 lexically matched sentences, 22 speakers)*
- Phonetic analysts classified function-word-present sentences according to the degree of discontinuity before the function word. Function-word-absent sentences had the corresponding region identified (marked with a dotted line in Fig. 2):

Figure 2. An illustration of spectro-temporal cues in two conditions of "function word present" and "function word absent". Four typical items reflecting different levels of discontinuity and the applied criteria for classification are presented.



RESEARCH QUESTIONS:

- **What spectral cues proximal to the function word might help listeners identify a function word onset? How frequently are spectral discontinuities absent at function word onsets?**
- **What temporal cues before the function word might help a listener determine if there was a heavily co-articulated function word present in the speech stream?**

ACOUSTIC ANALYSIS:

Proximal spectral cues characterizing function word boundaries

Characterization of the word boundary:
Three regions were identified (Figure 3):
- The **region of discontinuity** (**rod**)
- A region that included 4 pitch cycles **before** the rod (**r1**)
- A region that included 4 pitch cycles **after** the rod (**r2**)

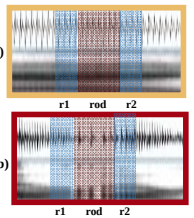


Figure 3. Demonstration of r1, rod, and r2 regions for two speech excerpt with (a) "no discontinuity" and (b) "strong discontinuity".

Analysis 1: Hand-coded spectral slope measures

- **H1*-H2*:** Indexes spectral tilt and correlates with voice quality (breathy, modal, glottalized) (e.g., Gordon & Ladefoged, 2001)
- **H1*-A1:** Indexes spectral tilt and first formant bandwidth (Hanson, 1997) as a measure of voice quality
- Corrections (denoted by asterisks) were applied by subtracting the quantity of $20*\log10[F1^2/(F1^2-f^2)]$ from H1 and H2, where f was the frequency corresponding to the harmonic (Hanson, 1997).
- **Results**: These spectral measures were effective in distinguishing variation across regions of analysis and levels of discontinuity for /h/-initial words but not vowel-initial words (Figure 4).

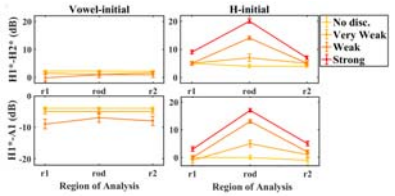


Figure 4. Variation of hand-coded measures of spectral slope (H1-H2, H1-A1) across three regions of r1, rod, and r2 and among different levels of discontinuity. These measures are shown for two sets of vowel-initial and h-initial function words.

Analysis 2: Automated characterization of spectro-temporal speech quality

- **Spectral Rolloff:** Captures noisy excerpts such as pauses with low-level energy (Lerch, 2012)

$$v_{SR} = i\Big|_{\sum_{k=0}^{i}|X(k)| \approx k.\sum_{k=0}^{k-1}|X(k)|}$$

- **Spectral Skewness:** Characterizes the distribution of spectral information across three regions of *r1, rod, and r2* (Lerch, 2012)

$$v_{SSK} = \frac{2.\sum_{k=0}^{k-1}(|X(k)| - \mu_{|X|})^3}{K.\sigma_{|X|}^3}$$

- **Root Mean Square:** Parametrizes the energy distribution (intensity) in temporal domain across three regions under study (Lerch, 2012)

$$v_{RMS} = 20.log_{10}(\sum_{i=0}^{N}(x(n))^2$$

- These three spectral measures are predictive of the level of discontinuity in coarticulation-prone contexts preceding function word (Figure 5).
- There is a clear decrease/increase in the values of these spectral shape measures among regions of analysis which is suggestive of a boundary with a certain level that depends on the degree of variation of these measures over time (Figure 5).
- These are proximal spectral cues signaling a potential boundary. Our next analysis aimed to characterize distal temporal cues and their interaction with proximal cues.

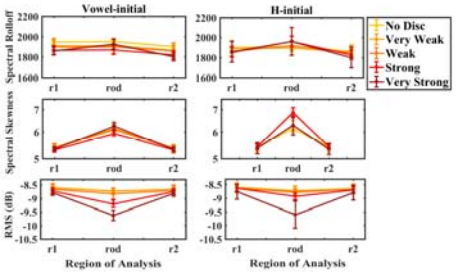


Figure 5. Variation of three automatically calculated measures of "spectral rolloff", "spectral skewness", and "root-mean-square" across three regions of r1, rod, and r2 and among different levels of discontinuity. These measures capture the variation of function word boundary discontinuities over time and across different levels of discontinuities. The pattern of spectral rolloff does not perfectly model the variation among different levels of discontinuity.

ACOUSTIC ANALYSIS:

Distal Temporal cues: Characterizing function word present (with no discontinuity) versus function word absent conditions

- **Proximal cues**
  - **Proximal duration:** The duration of the vocalic region corresponding to the preceding syllable up through the end of a (possible) function word in both corpora (Figure 1, dotted lines).
- **Relative distal cues**
  - **Relative Phoneme-based cue:** The ratio of the duration of vocalic region to the phoneme-based distal speech rate starting from the beginning of the utterance to the start of the vocalic region.
  - **Relative syllable-based cue:** The ratio of rate in vocalic region to the syllable-based distal speech rate starting from the beginning of the utterance to the start of vocalic region.
  - **Relative Fourier-based cue** (Tilsen & Johnson, 2008)**:** The ratio of implied rate of the vocalic region (1/duration) to the Fourier-based speech rate estimated from speech amplitude envelope in context speech.

Predicting function word presence and absence with mixed models (subject and sentence as random effects):

Acoustic Cue	Coef.	*t*-statistic	*p*-value	95% Conf. Interval	
Proximal duration	-7.20	-0.68	0.49818	-28.04	13.64
Relative Phoneme-based Rate	-13.624	-2.5516	**0.010769***	-24.094	-3.1552
Relative Syllable-based Rate	1.501	0.08261	0.93751	-33.855	36.855
Relative Fourier-based Rate	-1.389	-1.6882	**0.091481***	-3.0023	0.22426
Proximal duration*Relative Phoneme-based Rate	70.326	1.872	0.061301	-3.3335	143.99
Proximal duration*Relative Syllable-based Rate	18.663	1.4936	0.13539	-5.8375	43.164
Proximal duration*Relative Fourier-based Rate	16.347	2.867	**0.004171***	5.1673	27.526

Table 1. The results for fitting a logistic regression model to predict the existence/absence of a function word based on using different proximal, proximal-to-distal cues and their interaction as input variables.

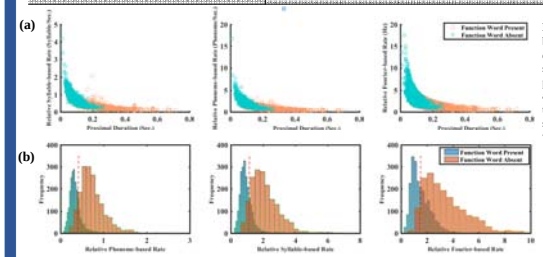


Figure 6. (a) Interaction between proximal and relative cues (proximal-to-distal) in separation of function word present from function word absent condition. (b) Relative timing cues are very strong in signaling whether function word is present or absent.

CONCLUSIONS

- A striking majority (~77%) of function words showed little to no evidence of discontinuity. Therefore, speakers frequently blend boundaries for these words, presenting a common perceptual problem that listeners must overcome.
- The present results characterize the spectro-temporal statistics available in casual speech that might allow a listener to determine a function word boundary, or determine if there was a function word at all (cf. Dilley & Pitt, 2010).
  - We show proximal differences in multiple characterizations of the spectral information available at the function word boundary *across levels of discontinuity*, demonstrating that the degree of co-articulation can be characterized based on local contextual information.
  - Further, we show distinct statistical distributions of proximal and distal temporal information in speech when *a function word is present but blended* (i.e., lacking discontinuity) versus speech in which *a function word is absent.*
- These results support prior experimental work showing that listeners use *distal context speech rate* to perceptually recover blended words (Dilley & Pitt, 2010; Heffner et al., 2013).
- The present research thus contributes to understanding neurocognitive processes underlying speech perception. These findings shed light on what cues are available for the perceptual recovery of function words when discontinuities to which neurons might entrain are lacking.

References

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *J. Acoust. Soc. Am.* 113, 1001–1024.
Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21(11), 1664-1670.
Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24(4), 423-444.
Dilley, L., Arjmandi, M. K., Ireland, Z., Heffner, C., & Pitt, M. (2016). Glottalization, reduction, and acoustic variability in function words in American English. The Journal of the Acoustical Society of America, 140(4), 3114-3114.
Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* 95, 1053–1064.
Ghitza, O., Giraud, A., and Poeppel, D. (2013). Neuronal oscillations and speech perception: critical band temporal envelopes are the essence. Frontiers in Human Neuroscience,
Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. Journal of Phonetics, 29(4), 383-406.
Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. The Journal of the Acoustical Society of America, 101(1), 466-481.
Heffner, C. C., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2013). When cues combine: how distal and proximal acoustic cues are integrated in word segmentation. *Language and Cognitive Processes*, 28(9), 1275-1302.
Klatt, D. H. (1979). "Synthesis by rule of segmental durations in English sentences," in Frontiers of Speech Communication Research, edited by B. Lindblom and S. Öhmann, Academic, New York, pp. 287–299.
Lerch, A. (2012). An introduction to audio content analysis: Applications in signal processing and music informatics. John Wiley & Sons.
Shockey, L. (2003). Sound Patterns of Spoken English Blackwell, Malden, MA.
Stevens, K. (1998). Acoustic Phonetics, MIT Press, Cambridge, MA.
Tilsen, S., & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. The Journal of the Acoustical Society of America, 124(2), EL34-EL39.

Acknowledgements

We gratefully acknowledge support of NSF grant BCS 1431063. We thank members of the MSU Speech Perception-Production Lab, especially Chase Smitterberg and Rachel Jansen.