Predicting echo thresholds from speech onset characteristics

Scott D. Miller
Department of Psychology and Waisman Center, University of Wisconsin, Madison, Wisconsin 53706
sdmiller1@wisc.edu

Ruth Y. Litovsky
Department of Communicative Disorders and Waisman Center, University of Wisconsin, Madison, Wisconsin 53705
litovsky@waisman.wisc.edu

Keith R. Kluender
Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706
krkluend@wisc.edu

Abstract: Echo threshold variability has previously been examined using stimuli that are carefully controlled and artificial (e.g., clicks and noise bursts), while studies using speech stimuli have only reported average thresholds. To begin to understand how echo thresholds might vary among speech sounds, four syllables were selected in pairs that contrasted abruptness vs gradualness of onset envelopes. *Fusion* and *discrimination suppression thresholds*, two echo thresholds commonly used to study the *precedence effect*, differed among syllables. Results were used to evaluate two predictive heuristics adapted from *perceptual center (p-center)* models.

PACS numbers: 43.66.Pn, 43.71.Rt, 43.72.Dv, 43.66.Qp [QJF]
Date Received: November 23, 2008 **Date Accepted:** January 16, 2009

1. Introduction

Everyday listening conditions are characterized by multiple sound sources and numerous reflections. As a first step toward understanding auditory perception in reverberant conditions, studies have employed a simple experimental paradigm with only a single first-arriving or *lead* sound and a single later-arriving or *lag* sound. The lag is usually an exact replica of the lead, simulating the first reflection to reach the ear in a reverberant environment. At short lead-lag delays, the lead dominates or suppresses perception of lag information along a number of dimensions. This is known as the *precedence effect*. Examples of dimensions over which it operates include perception of the lag as a discrete event (usually assessed subjectively via a *fusion* task) and extraction of lag directional cues (assessed objectively via tasks such as *discrimination suppression*). One common goal has been to quantify echo thresholds, i.e., the lead-lag delay at which the lead ceases to dominate the lag.[1]

While some of the very first work in this area used continuous speech as one kind of stimulus,[2–4] most progress since then has been made using artificial stimuli, which allow greater experimental control but lack acoustic characteristics intrinsic to speech and other naturally produced sounds. The majority of precedence effect studies to date have used clicks or noise bursts that are only a few milliseconds in duration. Such stimuli do not result in temporal overlap between the lead and lag at inter-stimulus delays corresponding to psychophysical echo thresholds.[1] In contrast, a single syllable in speech is usually 150 ms or longer, much longer than echo thresholds reported in a small number of studies for complex sounds such as continuous speech and music.[3,5] Clicks and noise bursts have been generated in laboratories to eliminate periodicities and recognizable transients,[1] while pure tones have been used to eliminate spectral complexity.[6] Stimulus onset characteristics, which play a critical role in the precedence

effect,[6–8] are often held constant and have only been manipulated in a simplified linear fashion.[6] Information in speech is carried by just the kinds of spectral and temporal characteristics that are eliminated from most artificial stimuli.

Systematic variability among echo thresholds for various speech sounds has not previously been reported, and it is not clear how current auditory models might be best adapted to make accurate predictions about such differences. Precedence effect studies that have used speech stimuli[2,3,5] have done so with the implicit simplifying assumption of equivalent thresholds across various speech sounds, reporting only the average results across all speech stimuli used. This assumption has been explicitly built into the few precedence effect models that have attempted to account for experimental results with speech stimuli.[9,10] The current literature does not provide any greater precision in the characterization of speech echo thresholds than the following general observation from the first precedence effect report,[4] p. 335: "The interval over which fusion takes place is not the same for all kinds of sounds. The upper limit of the interval was found to be about 5 ms for single clicks and is apparently much longer, perhaps as much as 40 ms, for sounds of a complex character."

Existing literature points to some acoustic characteristics likely to be most relevant to variability among speech echo thresholds, should there prove to be any. In particular, binaural detection and localization results with non-speech stimuli have demonstrated the distinct importance of onset characteristics for both periodic (i.e., pure tone[6]) and aperiodic (i.e., noise[7,8]) stimuli. The other consistent finding has been the dominance of low-frequency information ($<$1.5 kHz) over high-frequency information in precedence effect phenomena.[11,12]

One can also look to the speech perception literature, where waveform onsets and low-frequency information have emerged as critical predictors of so-called *perceptual centers* or *p-centers*.[13] The p-center is the beat or perceived moment of occurrence of a sound event as distinguished from the absolute acoustic onset.[14] From the p-center literature, we adopted Scott's[13] *frequency-dependent amplitude increase model* (FAIM) for the prediction of speech echo thresholds. Specifically, we replicated Scott's[13] procedure for the calculation of *rise time* values, the metric determined to most accurately predict p-centers. The extension of rise times to the prediction of echo thresholds was consistent with Rakerd and Hartmann's[6] identification of *onset rate* (increase in sound pressure per unit time) as the best predictor of binaural localization ability, but a frequency weighted application of this principle seemed more promising for use with naturally produced sounds. In addition, FAIM rise times have the advantages of being clearly specified and relatively easy to implement with speech stimuli.

The current study addressed two separate but related questions: (1) do echo thresholds for speech sounds vary systematically, and (2) can we predict these thresholds from acoustic characteristics? The first was addressed by applying a repeated-measures analysis of variance (ANOVA) to echo thresholds for different speech sounds; the second required a predictive heuristic, in this case FAIM rise times, and was addressed via correlational analysis.

2. Method

2.1 Participants

Nine young adults, six females and three males, initially volunteered to participate. To continue in the study, participants were required to pass a 20 dB hearing screening at 500, 1000, 2000, 4000, and 8000 Hz for each ear and have a symmetrical hearing profile, defined as no more than a 10 dB discrepancy between right and left ear thresholds at a particular frequency. None had any history of hearing difficulties, but one male was deemed ineligible for participation due to an asymmetrical hearing profile. One female did not return for testing after the initial practice session, leaving a final group of seven participants. All participants spoke English as their first and primary language.

2.2 Stimuli

Four syllables were recorded from an adult female: [bʌ], [dʌ], [wʌ], and [yʌ] (transliterated: "buh," "duh," "wuh," and "yuh"). These stimuli were selected as pairs, [bʌ] vs [wʌ] and [dʌ] vs

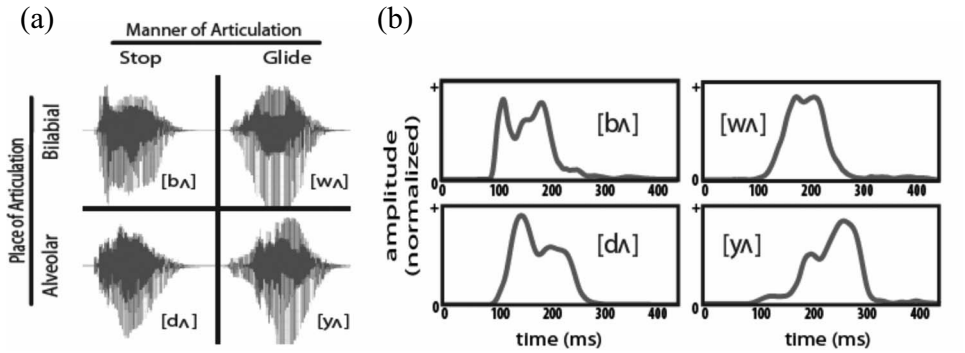

Fig. 1. (a) Raw waveforms of four syllables used as stimuli. Stimuli were digitally sampled at 44.1 kHz, and amplitude levels were root-mean-square equalized (see Multimedia link in text). (b) Outputs from filtering procedure used to calculate onset rise times for each stimulus.

[yʌ], which would differ significantly in terms of rise times but would otherwise have similar spectral and temporal characteristics. The voiced stop consonants in [bʌ] and [dʌ] have abrupt waveform onsets compared to the glides in [wʌ] and [yʌ], which have gradual onsets [see Fig. 1(a)].

[Mm. 1. [bʌ], [wʌ], [dʌ], and [yʌ] stimulus files concatenated in wav format (134 KB).]

Rise times were calculated following the procedure of Scott.[13] Each syllable was passed through a gammatone filter[15] with a center frequency of 578 Hz and an equivalent rectangular bandwidth of 4.0. These outputs, which contained significant information several hundred hertz above and below the center frequency, were full-wave rectified and smoothed with a 25 Hz Butterworth filter [see Fig. 1(b) for results]. The onset time (in milliseconds) between 10% and 90% of the maximum amplitude of each smoothed waveform was calculated, producing a frequency-dependent rise time for each recorded syllable. Rise times (in parentheses following each syllable) predicted that echo thresholds would be ordered as follows, from lowest to highest: [bʌ] (14.83), [dʌ] (33.22), [wʌ] (43.65), and [yʌ] (91.54).

2.3 Equipment

A customized MATLAB program was developed for randomization and presentation of stimuli, which was accomplished via a Tucker-Davis Technologies System 3 multiple input-output processor. Testing took place in a standard IAC sound booth (reverberation time, T_{60}=250 ms), where an array of loudspeakers (Cambridge SoundWorks Henry Kloss Center/Surround IV) was positioned in the horizontal plane. Each speaker in the semi-circular array was at ear level, 1.4 m from the approximate center of the seated participant's head. A computer monitor was placed under the front loudspeaker, and participants entered responses via a mouse by clicking on appropriate icons as instructed.

2.4 Design and procedure

Two tasks were used to measure echo thresholds: fusion and discrimination suppression. All seven participants completed both tasks. Before data collection began, each participant completed a 2-h practice session to establish a consistent response criterion on each task. The procedure for each task and thresholds (50% two sounds perceived on the fusion task, and 70.9% correct, i.e., $d'=1.0$, on the discrimination task) was similar to those described by Freyman *et al.*[16] The following delays, which always produced temporal overlap between lead and lag, were used for both tasks: 1, 5, 10, 15, 20, 25, 30, 35, 40, and 45 ms. For both tasks the following conditions applied: (a) The same stimulus was always presented as both lead and lag at 60 dBA, (b) participants were instructed to face the computer monitor in front (0° azimuth), (c) the lead was always from −45° azimuth, and (d) participants initiated each trial at their own pace using

Table 1. Both raw and centered group-mean thresholds are reported for each syllable on each task. Centered values are relative to the task mean, facilitating cross-task comparisons. Syllable error terms are *within subjects standard deviations* (relative to centered thresholds), consistent with the design of the study. Task mean error terms are *between subjects standard deviations* (relative to raw thresholds).

	Task									
	Fusion					Discrimination				
Group mean	[bʌ]	[dʌ]	[wʌ]	[yʌ]	Task mean	[bʌ]	[dʌ]	[wʌ]	[yʌ]	Task mean
Raw	17.00	14.79	19.23	21.73	18.19	10.91	6.85	13.26	13.83	11.14
Centered	−1.19	−3.40	1.04	3.54	18.19	−0.23	−4.29	2.12	2.69	11.14
Std. dev.	2.36	3.30	2.35	2.15	6.79	3.14	3.33	4.53	2.53	5.72

onscreen controls. During the fusion task, the lag was always presented from +45° azimuth. Participants, who were informed that the task involved an auditory "illusion" and that two sounds were always presented, made a subjective decision as to whether they perceived there to be one or two sounds. No feedback could be given, but participants were instructed to only select two if they were certain they had heard a second sound coming from a location to the right of center. During the discrimination task, the lag was presented from either +35° or +55°. Participants were instructed to decide whether the lag originated from the left or right of a visual reference point at +45° (a two-alternative forced choice), and feedback was given after each response.

The method of constant stimuli was used, and blocks alternated between tasks in a counterbalanced manner. Within each block, all stimuli and delays were used an equal number of times, and the presentation order was completely randomized across stimuli and delays; lag locations were included in this block randomization on the discrimination task. Across all blocks of the fusion task, 20 repetitions were completed for each stimulus at each delay, resulting in 800 total trials. For the discrimination task, 40 repetitions (20 from each lag location) resulted in a total of 1600 trials. Participants were allowed to take breaks as often and for as long as they wanted during testing. They were encouraged to take a minimum of one break every 200 trials. Trials were typically completed across three 2-h testing sessions spaced over several days.

3. Results and discussion

3.1 Echo thresholds

Group mean thresholds are reported in Table 1. Task means, 18.19 ms for fusion and 11.14 ms for discrimination suppression, were slightly higher than those reported in the literature for click stimuli (e.g., 5–9 ms for thresholds on both tasks[16] using identical threshold criteria and free field procedures) but slightly lower than fusion thresholds typically reported for connected speech (e.g., 30–50 ms[3]). Litovsky *et al.*[1] provided a review of thresholds found on both tasks with different types of stimuli. In the present study, the use of single syllables as stimuli, rather than sentences or concatenated word lists, probably explains why thresholds fell somewhere midway between previous results with artificial stimuli, on the one hand, and continuous speech, on the other.

3.2 Analysis of variance

For all statistical tests, a level of $p < 0.05$ was used to determine significance. When task (fusion vs discrimination), manner of articulation (stop vs glide), and place of articulation (bilabial vs alveolar) were entered as factors in a three-way repeated-measures ANOVA, there was a significant main effect of manner, $F(1,6) = 11.45$, $p = 0.02$, and a significant manner by place interac-

tion, $F(1,6)=15.67$, $p=0.01$. No other main effects or interactions were significant. This supported our primary hypothesis that echo thresholds for speech sounds would differ across syllables.

The main effect of manner was in the predicted direction, with stops as a class having lower group mean thresholds than glides, while the interaction of place with manner affected the strength but not the direction of this effect. Therefore, this analysis of onset effects, at the admittedly gross level of syllable category, also provided support for our second hypothesis that the abrupt onsets of stops would tend to result in lower thresholds. Paired-sample *t*-tests reached significance for [dʌ] vs [yʌ], [bʌ] vs [yʌ], and [wʌ] vs [yʌ] on the fusion task and [dʌ] vs [bʌ] and [dʌ] vs [yʌ] on the discrimination task.

3.3 Correlational analyses

Correlations between rise times and group mean echo thresholds were in the predicted direction, but neither slope was statistically significant: $r=0.82$, $r^2=0.67$, and $p=0.18$ for the fusion task, and $r=0.58$, $r^2=0.33$, and $p=0.42$ for the discrimination task. Although the small number of data points meant statistical power was low for this analysis, rise times have two characteristics that may account for their failure to predict observed differences among echo thresholds more accurately. They falsely assume linearity in waveform onsets, and they do not relate onset characteristics to the waveform as a whole.[17] An alternative p-center metric, Howell's[18] *center of gravity* can be adapted to address both of these potential shortcomings. It had never been applied in a frequency-specific manner, but the limited success of frequency-dependent rise times in the current study seemed to warrant the test of a hybrid *frequency-dependent center of gravity* (FCoG) model. In deriving FCoGs, Scott's[13] filtering procedure was again used—and consistent with this procedure, the syllable was defined as beginning and ending at the points where the waveform amplitude was equal to 10% of the maximum amplitude—but the remaining steps were suggested by Howell.[19] The duration from the beginning of the syllable to the point at which the integral (area under the curve) equaled one-half of the integral for the entire syllable was expressed as a ratio over the duration of the entire syllable, i.e., as a unitless proportion. From a signal detection theoretical perspective, this would seem to capture the detectability of the lag onset as well as the masking potential of the continuous and offset portions of the lead in a single metric.

The FCoGs (in parentheses following each syllable) predicted that echo thresholds would be ordered as follows, from lowest to highest: [dʌ] (0.22), [bʌ] (0.41), [wʌ] (0.49), and [yʌ] (0.59). Correlations for the fusion task, $r=0.976$, $r^2=0.952$, and $p=0.02$, and the discrimination task, $r=0.982$, $r^2=0.970$, and $p=0.02$, indicated that FCoGs accounted for almost all of the variance in echo thresholds. Both slopes were significant even after a conservative Bonferroni correction was applied to this *post hoc* analysis.

4. Conclusions

In this investigation, echo thresholds differed as a function of syllable-initial phonetic features and correlated significantly with FCoGs, shedding light on properties of speech stimuli worth exploring further in relation to the precedence effect and speech perception in reverberant situations. Predictive heuristics tested in this study fit most naturally with other signal detection theoretical approaches in the binaural literature (see Saberi and Petrosyan[20] for review) and in the broader perceptual sciences in that they focus on the information-bearing properties of the stimulus. This approach is less popular than physiologically oriented modeling in the binaural literature (see Stern and Trahiotis[21] for review), which focuses on the transformational processes accomplished by the auditory system, but it has shown merit here as an entrée into the characterization of variability among echo thresholds for speech stimuli. In the long run, this complementary approach should facilitate improvements in physiological models as well.

Without experimental data from a much wider variety of stimuli (e.g., different phoneme combinations and talkers), it is difficult to know whether FCoGs or FAIM rise times can be used to predict all speech echo thresholds. Perhaps the most critical test for any predictor will come with initial voiceless consonants, in particular voiceless fricatives. These onsets present a

challenge because they lack both the low-frequency spectral peaks and periodicities characteristic of voiced consonant and vowel sounds. There are no explicit models for how high-frequency-aperiodic and low-frequency-periodic cues are combined to produce precedence effect phenomena when both are sequentially present in a single stimulus, as they are, for example, in syllables that begin with voiceless fricatives. These challenges may require a more complex model, such as the p-center model of Pampino–Marschall,[22] which determines *partial onset events* as well as *syllabic centers of gravity* from frequency-specific loudness dynamics within multiple critical bands. If multivariate predictors prove to be necessary, their application to continuous stimuli may benefit from a learning algorithm and training procedure (e.g., the neural network model of Wilson and Darrell[23]).

Acknowledgments

This work was supported by grants from the NIH-NIDCD (Grant No. R01 DC030083 to R.Y.L. and T32 DC005359 to Susan Ellis Weismer, University of Wisconsin, Madison).

References and links

[1]R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," J. Acoust. Soc. Am. **106**, 1633–1654 (1999).
[2]H. Haas, "The influence of a single echo on the audibility of speech," J. Audio Eng. Soc. **20**, 145–159 (1972).
[3]J. P. A. Lochner and J. F. Burger, "The subjective masking of short time delayed echoes by their primary sounds and their contribution to the intelligibility of speech," Acustica **8**, 1–10 (1958).
[4]H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," Am. J. Psychol. **62**, 315–336 (1949).
[5]B. Rakerd, W. M. Hartmann, and J. Hsu, "Echo suppression in the horizontal and median sagittal planes," J. Acoust. Soc. Am. **107**, 1061–1064 (2000).
[6]B. Rakerd and W. M. Hartmann, "Localization of sound in rooms, III: Onset and duration effects," J. Acoust. Soc. Am. **80**, 1695–1706 (1986).
[7]T. Houtgast and R. Plomp, "Lateralization threshold of a signal in noise," J. Acoust. Soc. Am. **44**, 807–812 (1968).
[8]T. Houtgast and S. Aoki, "Stimulus-onset dominance in the perception of binaural information," Hear. Res. **72**, 29–36 (1994).
[9]O. Schwartz, J. G. Harris, and J. C. Principe, "Modeling the precedence effect for speech using the gamma filter," Neural Networks **12**, 409–417 (1999).
[10]B. Supper, T. Brookes, and F. Rumsey, "A new approach to detecting auditory onsets within a binaural stream," Paper presented at the 114th Convention of the Audio Eng. Soc., Amsterdam, Netherlands (March 2003). Retrieved from http://epubs.surrey.ac.uk/recording/23 (Last viewed June 25, 2008).
[11]B. G. Shinn-Cunningham, P. M. Zurek, N. I. Durlach, and R. K. Clifton, "Cross frequency interactions in the precedence effect," J. Acoust. Soc. Am. **98**, 164–171 (1995).
[12]D. J. Tollin and G. B. Henning, "Some aspects of the lateralization of echoed sound in man. II. The role of the stimulus spectrum," J. Acoust. Soc. Am. **105**, 838–849 (1999).
[13]S. K. Scott, "P-centers in speech: An acoustic analysis," Ph.D. dissertation, University College London, UK, 1993.
[14]J. Morton, S. Marcus, and C. Frankish, "Perceptual centers (p-centers)," Psychol. Rev. **83**, 405–408 (1976).
[15]M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Apple Computer Technical Report No. 35; Retrieved from http://citeseer.ist.psu.edu/8863.html (Last viewed June 25, 2008).
[16]R. L. Freyman, R. K. Clifton, and R. Y. Litovsky, "Dynamic processes in the precedence effect," J. Acoust. Soc. Am. **90**, 874–884 (1991).
[17]While rise times accurately predict p-centers of natural stimuli (Ref. 13), perception of onsets is complicated by the presence of background sounds (Ref. 24), and it remains an open question whether absolute onsets are better than relative onsets for predicting echo thresholds of natural stimuli in reverberant conditions.
[18]P. Howell, "Prediction of the p-center from the distribution of energy in the amplitude envelope: I," Percept. Psychophys. **43**, 90–93 (1988).
[19]Howell himself (Ref. 18) used a procedure to calculate centers of gravity that assumed linearity in the amplitude envelope by modeling the syllable onset as a triangle and the remainder of the syllable as a rectangle, but he suggested an alternative procedure that takes integrals directly from rectified complex waveforms and does not assume linearity. This alternative is more straightforward theoretically, but the computations are more complex.
[20]K. Saberi and A. Petrosyan, "A detection-theoretic model of echo inhibition," Psychol. Rev. **111**, 52–66 (2004).
[21]R. M. Stern and C. Trahiotis, "Models of binaural interaction," in *Handbook of Perception and Cognition: Vol. 4. Hearing*, 2nd ed., edited by B. C. J. Moore (Academic, New York, 1995), pp. 347–386.

[22]B. Pompino-Marschall, "On the psychoacoustic nature of the P-center phenomenon," J. Phonetics **17**, 175–192 (1989).
[23]K. W. Wilson and T. Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," IEEE Trans. Audio, Speech, Lang. Process. **14**, 2156–2164 (2006).
[24]J. Vos and R. Rasch, "The perceptual onset of musical tones," Percept. Psychophys. **29**, 323–335 (1981).