The influence of amplitude envelope information on resolving lexically ambiguous spoken words

Christine M. Szostak[a] **and Mark A. Pitt**
Department of Psychology, Ohio State University, Columbus, Ohio 43210
cszostak@shorter.edu, pitt.2@osu.edu

Abstract: Abstract: Prior studies exploring the contribution of amplitude envelope information to spoken word recognition are mixed with regard to the question of whether amplitude envelope alone, without spectral detail, can aid isolated word recognition. Three experiments show that the amplitude envelope will aid word identification only if two conditions are met: (1) It is not the only information available to the listener and (2) lexical ambiguity is not present. Implications for lexical processing are discussed.

PACS numbers: 43.71.-k, 43.71.Gv, 43.71.Rt, 43.71.Sy [AC]
Date Received: May 11, 2014 **Date Accepted:** July 30, 2014

1. Introduction

When listening to speech, the listener must resolve ambiguities to ensure comprehension succeeds. One common source of ambiguity is masking by environmental noise. An understanding of the robustness of speech cues to noise can provide insight into the mechanisms involved in word recognition. One cue that continues to receive considerable attention is amplitude (temporal) envelope (AE) with studies demonstrating its usefulness in recognizing phrases and isolated words (Doelling *et al.*, 2014; Shannon *et al.*, 1995).

The current study examines a lower bound on AE processing, exploring under what conditions the AE of a single phoneme aids recognition. Specifically, if noise masks a phoneme that is critical for identifying the word produced by a talker (e.g., *=ake,* where = represents noise), the listener must somehow recognize that the intended word was *lake* rather than *rake* or *bake*. If the AE of the phoneme can uniquely specify the phoneme, there is no lexical ambiguity, and processing should be unimpaired. If AE is ambiguous, the presence of lexical competitors should impede recognition. Prior work on this question has yielded mixed results, so our goal was to provide clarity.

Samuel (1987) was interested in exploring the degree to which lexical competitors (for present purposes, rhyme competitors, e.g., *=ake*) influenced spoken word recognition. Using the phonemic restoration paradigm, he compared the restoration of words that had at least one rhyme competitor (e.g., *r/locket*) to words that had no competitors (e.g., *lengthen*). Noise was either added to the initial phoneme or replaced the initial phoneme. He found that the presence of a rhyme competitor yielded less restoration (i.e., better discrimination between added and replaced stimuli) than when there were no competitors, indicating that lexical influences on phoneme restoration are influenced by the amount of lexical competition.

Samuel (1987) used signal-correlated noise (SCN: Schroeder, 1968) to create the noise altered stimuli so as to retain the syllabic quality of the masked portion of the speech. To ensure that participants could not use AE to identify what word was spoken, he chose word pairs the onsets of which differed only in place of articulation

[a]Present address: Department of Social Sciences, Shorter University, 315 Shorter Ave., 5E Alumni Hall, Rome, Georgia, 30165.

(e.g., *r/locket*), a feature found to be highly difficult to identify by AE alone (cf., Rosen, 1992).

Bashford, Warren, and Brown (1996) found that the AE of SCN contains cues to phoneme identity, raising questions about the use of SCN by Samuel (1987). Of most relevance for the current study is their second experiment in which listeners heard monosyllabic words in isolation. In the manipulations of interest, approximately 50% of each word was replaced by either stochastic noise (SN) or by SCN. Bashford *et al.* (1996) reasoned that if AE aids word recognition, participants should be more accurate in the SCN condition than in the SN condition. This is exactly what was found. Participants were approximately 7% better in identifying words in the SCN than the SN condition, prompting Bashford *et al.* (1996) to question whether the use of Samuel (1987) of AE influenced word identification.

Three differences between the two studies suggest that there could be other reasons for the discrepant outcomes: (1) Bashford *et al.* (1996) did not control the number of rhyme competitors. Thus there is no way to know whether the AE of the noise served as a discriminating cue for a single lexical item or multiple words. That is, in the study by Samuel (1987), did AE aid listeners more when no competitors existed? (2) Samuel (1987) chose stimuli such that the target word (e.g., *locket*) and the primary rhyme competitor (e.g., *rocket*) were minimal pairs differing only in place of articulation, while Bashford *et al.* (1996) were not concerned with this property of their stimuli. Given that work by Shannon *et al.* (1995) demonstrated that place of articulation is a highly confusable acoustic phonetic feature when recognition is based largely on AE, it is possible that AE information was not sufficient to aid performance in the study by Samuel (1987). (3) In the study by Samuel (1987), the entire target phoneme and any residual acoustic features were noise altered. In contrast, Bashford *et al.* (1996) presented stimuli wherein alternating 200 ms sections of a train of words were replaced by noise. These replaced portions did not always correspond to an entire phoneme. Because of this, their noise masker may have covered only the final portion of one phoneme and the onset of the next, and so on. As a consequence, cues such as spectral detail may have strengthened the influences of AE. Alone, as in the case of the study by Samuel (1987), AE may provide no useful information to aid word recognition.

The purpose of the present study was to explore these differences across studies and determine whether AE alone can disambiguate lexically ambiguous spoken words (e.g., *r/lake*).

2. Experiment 1

The purpose of experiment 1 was to replicate Bashford *et al.* (1996) to ensure that our stimuli yielded the recognition advantage for SCN that they reported. Although our interest was in comparing words that differ in the number of rhyme competitors, use of the noise interruption methodology of Bashford *et al.* (1996) precluded us from making such comparisons in this first experiment. Because the noise was not confined to the disambiguating phoneme (e.g., *=ake*), lexical ambiguity was not controlled.

2.1 Method

2.1.1 Participants

Thirty-two students received course credit in an introductory psychology course in exchange for participation. All were native speakers of American English with no reported hearing deficits.

2.1.2 Stimuli

The experimental stimuli were 12 monosyllabic word quads. Three of the words in each quad were considered ambiguous as they shared the same rhyme (e.g., *lake*, *rake*, and *bake*) and one was unique, having no rhyme competitors (e.g., *length*). The first ambiguous word (target condition) shared the same onset as the unique condition word (e.g., *lake* vs *length*). Word onsets in these two conditions were only liquids, nasals, and glides.

The second ambiguous word (subtle condition) had an onset that differed from the target condition only in place of articulation (e.g., *rake*). The third ambiguous word (distinct condition, e.g., *bake*) differed from the target word in at least one feature that was not place of articulation, yielding an AE that was different from the other conditions. Words across the four conditions did not differ in word frequency, F (3,44) = 0.7, $p < 0.6$. Ninety-six monosyllabic filler words were also included to hide the rhyme competitor manipulation and to help ensure there was greater phonetic variety in word onsets across the stimuli. Digital recordings of the stimuli were made.

In Bashford *et al.* (1996), a list of spoken words was interrupted every 200 ms with noise that was 200 ms in duration. The starting point of the first noise interruption relative to list onset was randomized across participants. We simplified this methodology by fixing noise onset to one of two positions in each word. In the original condition, 200 ms of noise was pseudo-randomly placed somewhere within or immediately following the first 200 ms of word onset, and re-occurred every 200 ms until the end of the word was reached. In the mirrored condition, a mirror image of each stimulus was constructed such that the noise occurred in the opposite 200 ms portions of the word. Two types of noise were inserted: SN, created by calculating the root mean square amplitude of the 200 ms section of speech, and then for each digital sample, randomly sampling a value within the full dynamic range. SCN was created using the method described in Schroeder (1968). Following Bashford *et al.,* both noise portions matched the RMSA level of the speech they replaced, resulting in a signal-to-noise ratio of 0 dB. This matched with the stimulus construction reported by Samuel (1987) and the specific conditions of interest from Bashford *et al.* (1996, experiment 2).

2.1.3 Procedure

Noise type and noise location were manipulated between participants to avoid stimulus repetition. Stimuli were split such that half of the participants heard the SN stimulus versions ($n = 8$ original, $n = 8$ mirrored) and the other half heard the SCN stimulus versions ($n = 8$ original, $n = 8$ mirrored). Stimuli were presented across four blocks such that only one member of each of the 12 word quads occurred in a block and at least one filler item separated target words. Participants were told that they would hear words that would be disrupted by noise over headphones and were instructed to type the word that was spoken.

2.2 Results and discussion

Participant responses were scored as correct only if the intended word was reported. Listeners more accurately identified words when the speech signal was disrupted by SCN (50.5%) than by SN (45.1%),[1] for an effect size of 5.4%, an effect that was similar to the approximately 7% effect reported by Bashford *et al.* (1996). There was also an effect of noise position with overall accuracy being higher in the mirrored (53.6%) than original (42.0%) condition. Inspection of the data by items suggests that this effect is due to greater voicing confusion in word final phonemes for participants in the original than in the mirrored condition.

A two-factor analysis of variance (ANOVA) with noise type and noise position (original vs mirrored) as factors yielded a reliable main effect of noise type, F (1,28) = 9.3, $p = 0.005$ and a main effect of noise position F (1,28) = 43.1, $p < 0.001$. The interaction was not reliable ($F < 2.0$). The results provide a clear replication of Bashford *et al.* (1996; experiment 2) in showing that speech intelligibility is better in SCN than SN.

3. Experiment 2

The technique in experiment 1 of replacing every other 200 ms of speech with noise contains insufficient control to conclude that resolution of a lexically ambiguous word (=*ake*) is possible with AE alone because the words were not always made ambiguous using this noise insertion method. The purpose of experiment 2 was to correct this.

Experiment 1 was run again, but noise covered fully and only the disambiguating (onset) phoneme in each word in line with Samuel (1987).

If the AE present in the SCN condition can serve as a bottom-up cue to ambiguity resolution (e.g., *lake* vs *rake*), accuracy in word identification should be higher in the SCN than in the SN condition. If this occurs, differences should emerge across the four word conditions. In the SN condition, rhyme differences (unique vs the three ambiguous conditions) should emerge such that accuracy will be higher in the unique condition (because of the absence of competitors) while the lexical ambiguity in the remaining three conditions should yield uniformly poor accuracy. In contrast, in the SCN condition, differences between the three ambiguous conditions should be found due to the presence of AE cues that convey information about the identity of the onset phoneme.

3.1 Method

3.1.1 Participants

Twenty new participants from the same pool meeting the same criteria as those of experiment 1 participated.

3.1.2 Stimuli

The same 144 stimuli used in experiment 1 were again used in this experiment. The noise modification (e.g., SCN vs SN) was performed in an identical manner to that of experiment 1 except that only the onset phoneme was noise altered. Care was taken to ensure that no evidence of the phoneme was visible or audible within the clear portion of the word.

3.1.3 Procedure

The procedure was identical to experiment 1.

3.2 Results and discussion

The data were scored as described in experiment 1. Figure 1 shows mean accuracy broken down by word type (unique, target, subtle, and distinct) and noise type (SCN, SN). The results were nearly identical across noise type, with mean accuracy in the SCN and SN conditions being 11.9% and 10.3%, respectively. A two-factor ANOVA was performed with noise type as a between-subjects variable and word type as a within-subjects variable. The main effect of noise type ($F < 0.4$) and the interaction of these variables ($F < 0.1$) proved unreliable. The main effect of word type was reliable, $F(3,54) = 19.1$, $p < 0.001$. Planned comparison t-tests with a Bonferroni correction compared the unique condition with each of the three ambiguous conditions (the data were collapsed over noise type for this analysis because there was no main effect of noise type nor did noise type interact with word type). The unique condition was reliably

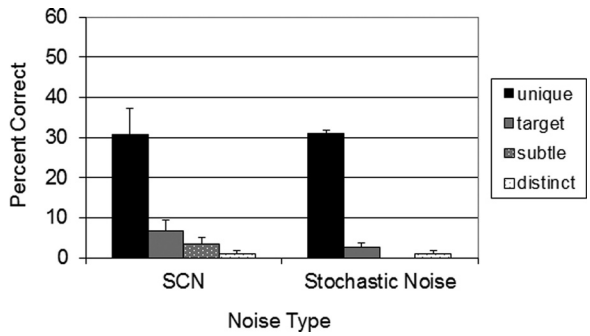

Fig. 1. Mean percent correct as a function of stimulus condition and noise type in experiment 2. Error bars represent standard error of estimate. SCN refers to signal correlated noise.

different from all three of the ambiguous conditions [target condition: t (19) = 4.8, $p < 0.001$; subtle condition: t (19) = 5.1, $p < 0.001$; distinct condition: t (19) = 3.3, $p < 0.005$], indicating that listeners were reliably more accurate in identifying a word with no lexical competitors than one having lexical competitors.

When the placement of noise was more finely controlled, obliterating only the disambiguating phoneme, no differences in SCN and SN were found. These data provide no evidence that AE serves as a valuable bottom-up cue for resolving lexical ambiguities when the clear portion of the speech alone cannot disambiguate the word. The lack of a noise effect was found across all of the ambiguous conditions, even in the distinct condition where the phoneme's AE had a much more distinctive shape (e.g., *lake* vs *bake*). The preceding outcomes are in line with Samuel (1987), who found that listeners were more likely to report a phoneme as having been present when only one competitor was available than when multiple competitors were available.

4. Experiment 3

The purpose of experiment 3 was to test whether a floor effect was responsible for the null effect of noise in experiment 2. Because the majority of words were CVCs, they could have been particularly difficult to recognize when such a large portion (1/3) of each word (initial consonant) was replaced by noise. By providing a brief glimpse of the target phoneme, intelligibility should improve and a noise effect might emerge. We shifted the noise forward into the word by 30% of the phoneme's length providing a glimpse of the start of the phoneme. Accuracy was expected to be higher than in experiment 2 because the onset phoneme should be more identifiable. A main effect of noise type should be found if SCN contributes to the identity of the initial phoneme.

4.1 Method

4.1.1 Participants

Participants were 20 new individuals from the same pool who met the same criteria as those in experiment 1.

4.1.2 Stimuli

The design was simplified to just the target (e.g., *lake*) and subtle (e.g., *rake*) conditions because these were the conditions of most interest. There were 12 rhyme competitor word pairs (24 stimuli) and 24 filler words. The start and end points of the noise were shifted later in time by 30% of the target phoneme's length.

4.1.3 Procedure

The procedure was identical to experiment 2 except that the experiment was completely within subjects. Half of the participants heard the SN condition stimuli first, and the other half heard the SCN condition stimuli first. After listeners heard both blocks of stimuli, the two blocks were repeated in the same order to increase the number of observations per cell.

4.2 Results and discussion

The data were scored as in experiment 2. Participants were nearly identical in their accuracy across the two noise types, regardless of word type (target condition SCN = 22.7%, SN = 25.1%; subtle condition SCN = 16.4%, SN = 16.3%). An ANOVA indicated that only the main effect of word type (target vs subtle) was reliable [$F(1,19) = 8.3$, $p = 0.01$]. The main effect of noise type and the interaction between these two variables proved unreliable (both Fs < 1.7). One reason for the lack of a noise effect may be learning across blocks. We therefore analyzed only the first block of data. This ensured that each item was heard only once and participants heard only one type of noise. Only word type approached reliability, F (1,18) = 3.5, $p < 0.08$. Both the noise type and the interaction between the two variables proved unreliable (Fs < 0.4).

Although SCN did not improve word identification, it could convey gross information about phoneme identity, such as phone class (e.g., stop, nasal). Analyses in which the data were rescored using these broader categories did not yield a performance advantage for SCN. The results of experiment 3 reinforce those of experiment 2: When lexical identity is ambiguous, listeners are not aided by the AE of the noise-replaced phoneme.

5. Conclusion

Three experiments explored the question of whether AE alone can resolve a lexical ambiguity. Experiment 1 served as a successful replication of Bashford *et al.* (1996), showing that when intermittent noise bursts disrupt a portion of the signal, word identification is better when the bursts preserve the AE of the obliterated speech. Experiments 2 and 3 showed that when noise was confined to obscuring only the disambiguating (initial) phoneme, the SCN advantage disappeared.

Combined, the findings suggest that AE information alone cannot aid resolution of a lexical ambiguity. The fact that accuracy in the unique condition was comparable across the two types of noises reinforces this conclusion. If the information conveyed by the AE of the initial phoneme had been sufficient, or even partially helpful for identification, then listeners should have identified the words in the SCN condition more accurately than in the SN condition. That they did not do so shows how uninformative AE can be in this situation. Note that by providing listeners with just a very short (30 ms) glimpse of the spectral content of the initial phoneme (experiment 3), accuracy increased noticeably, but importantly, there was no greater advantage when the AE was retained (SCN condition) than when it was not retained (SN condition).

The current results can be interpreted as providing a lower bound on the usefulness of AE for spoken word identification. They suggest that alone, when the processor must encode words in which more than one lexical option is available (e.g., = *ake*), AE alone provides little useful information. What this suggests for lexical processing is that, although AE alone might activate a set of competitors that is a good general fit to the intended word (e.g., activating *lake*, *rake*, and *bake*), it is insufficient to eliminate competitors. In other words, AE can contribute to activation of a set of lexical competitors, but it is minimally useful in distinguishing between them. This idea fits with the findings reported by Samuel (1987). Recall that he found that listeners were more likely to report a phoneme as having been present when a word had no competitors versus when competitors existed. Otherwise, he should have found no difference across conditions.

We are not suggesting that AE is never useful in spoken word recognition. Studies using noise-vocoded speech have demonstrated that the time-varying modulation of amplitude can be quite informative (e.g., Obleser *et al.*, 2012). In such studies, stimuli tend to be short phrases or sentences. AE likely conveys suprasegmental information (e.g., syllable duration, speech rate, possibly even rhythm) that can aid comprehension. In these richer communicative contexts, including when mouth movements are visible, it may be possible to find evidence of AE disambiguating word perception, such as constraining the interpretation of a noise-covered syllable.

Acknowledgments

We thank Hartman Brawley, Chloe Meyer, and Jason Miao for assistance with stimulus construction and data collection and Eric Healy and Carla Youngdahl for comments on the manuscript.

[1]Item analyses proved identical to subject analyses. Therefore only subject analyses were reported.

References and links

Bashford, J. A., Warren, R. M., and Brown, C. A. (**1996**). "Use of speech-modulated noise adds strong 'bottom-up' cues for phonemic restoration," Percept. Psychophys. **58**, 342–350.

Doelling, K. B., Arnal, L. H., Ghitza, O., and Poeppel, D. (**2014**). "Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing," Neuroimage **85**, 761–768.
Obleser, J., Herrmann, B., and Henry, M. J. (**2012**). "Neural oscillations in speech: Don't be enslaved by the envelope," Front. Hum. Neurosci. **6**, 1–4.
Rosen, S. M. (**1992**). "Temporal information in speech: Acoustic, auditory, and linguistic aspects," Phil. Trans. Royal Soc. London B **336**, 367–373.
Samuel, A. G. (**1987**). "Lexical uniqueness effects on phonemic restoration," J. Mem. Lang. **26**, 36–56.
Schroeder, M. R. (**1968**). "Reference signal for signal quality studies," J. Acoust. Soc. Am. **44**, 1735–1736.
Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.