Temporal effects in priming of masked and degraded speech

Richard L. Freyman,^{a)} Charlotte Morse-Fortier, and Amanda M. Griffin

Department of Communication Disorders, University of Massachusetts, 358 North Pleasant Street, Amherst, Massachusetts 01003, USA

(Received 24 March 2015; revised 13 July 2015; accepted 14 July 2015; published online 10 September 2015)

When listeners know the content of the message they are about to hear, the clarity of distorted or partially masked speech increases dramatically. The current experiments investigated this priming phenomenon quantitatively using a same-different task where a typed caption and auditory message either matched exactly or differed by one key word. Four conditions were tested with groups of normal-hearing listeners: (a) natural speech presented in two-talker babble in a non-spatial configuration, (b) same as (a) but with the masker time reversed, (c) same as (a) but with target-masker spatial separation, and (d) vocoded sentences presented in speech-spectrum noise. The primary manipulation was the timing of the caption relative to the auditory message, which varied in 20 steps with a resolution of 200 ms. Across all four conditions, optimal performance was achieved when the initiation of the text preceded the acoustic speech signal by at least 400 ms, driven mostly by a low number of “different” responses to Same stimuli. Performance was slightly poorer with simultaneous delivery and much poorer when the auditory signal preceded the caption. Because priming may be used to facilitate perceptual learning, identifying optimal temporal conditions for priming could help determine the best conditions for auditory training.

© 2015 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4927490>]

[EB]

Pages: 1418–1427

I. INTRODUCTION

The effect of prior knowledge of message content on the perception of degraded speech is easy to demonstrate in listeners with normal hearing. One of the most striking demonstrations is with speech degraded by noise-excited vocoding, which removes periodic temporal fine structure and reduces spectral detail. If the number of spectral channels is limited sufficiently, the acoustic signal is barely recognized as speech and is mostly unintelligible. However, if the listener is given a preview of the message content beforehand, either auditorily or through a written caption, the same vocoded auditory signal is not only clearly recognized as speech, but listeners are often impressed that the message is quite intelligible. This perceptual priming or cueing effect has been called “pop out” (Davis *et al.*, 2005; Loebach *et al.*, 2010).

Among the various reasons why it would be useful to explore this phenomenon further is its potential utility in promoting perceptual learning of degraded speech. For example, Davis *et al.* (2005), Hervais-Adelman *et al.* (2008), and Hervais-Adelman *et al.* (2011) demonstrated that subjects’ recognition performance for words and sentences processed with noise vocoding (e.g., Shannon *et al.*, 1995) improved rapidly during a single session, and reached the highest level of performance when subjects received feedback after each trial that should have elicited the pop-out effect. Specifically, after they were asked to identify a vocoded utterance in an open-set recognition task, subjects were

given a clear unprocessed presentation of the signal before a second presentation of the vocoded utterance. It was hypothesized that this type of feedback, which made the vocoded signal sound clearer, would lead to improved performance on open-set recognition with new utterances on subsequent trials. The advantage of this form of feedback (clear presentation before distorted presentation) as opposed to the reverse (distorted presentation before clear presentation) was observed within the first ten trials (Davis *et al.*, 2005), although growth in performance was not steeper for the clear feedback-first condition thereafter. Improved performance was also reported when the feedback was presented orthographically just before and concurrently with the vocoded utterance. Earlier studies (Schwab *et al.*, 1985; Greenspan *et al.*, 1988) also trained subjects on recognizing synthetic speech using feedback with captions that began simultaneously with the auditory signal.

If perceptual learning is assisted by the experience of priming or pop out, one could hopefully maximize perceptual learning by creating conditions that lead to optimal priming. However, measuring the effect of priming itself is not straightforward because once the message is known, a listener could simply remember it and repeat it whether or not the degraded auditory signal was perceived correctly. Therefore, it is difficult to measure what listeners actually hear during a degraded presentation following a clear presentation. A few different approaches have been used to try to get around this problem. For example, Freyman *et al.* (2004) investigated the priming of sentences that were presented in a highly confusable competing speech background. Only the first two key words of a three-key-word nonsense sentence were primed. The listener’s task was to identify a

^{a)}Electronic mail: rlf@comdis.umass.edu

third key word that was not primed. To the extent that priming helped listeners perceptually extract the target message from the background, the improved perception was predicted to continue as the sentence continued via processes related to auditory streaming, improving recognition of the third key word. Using this partial priming technique, both Freyman *et al.* (2004) and Ezzatian *et al.* (2011) found substantial benefits of priming for competing English speech maskers, and not as much for a control condition in which the masker was noise. Yang *et al.* (2007) found similar results using this priming technique with Chinese speech stimuli.

A second technique used a same-different task in which subjects were asked to determine whether a written caption and a masked utterance were identical or instead contained one changed key word (Jones and Freyman, 2012). The effectiveness of priming was measured by comparing listeners' performance for two orders of presentation: caption before auditory stimulus or auditory stimulus before caption. The idea behind this approach was that if prior knowledge of an auditory signal causes it to be better organized perceptually, it will be easier to detect a change in the auditory signal. The analogy used by Jones and Freyman (2012) was with the drawing by R. C. James, often shown in introductory psychology textbooks, of a seemingly random set of lines and dots that becomes perceptually organized into the image of a Dalmatian canine once the viewer is cued. Jones and Freyman (2012) reasoned that changes to the drawing of lines and dots would be easier to detect once it was perceived as a Dalmatian. They proposed, analogously, that a change in a degraded speech signal would also be easier to detect if it was more perceptually organized. The results of their study demonstrated large benefits of priming in the presence of interfering speech, steady noise, and speech-envelope-modulated noise. In all these conditions, listeners were better able to identify whether the text and degraded speech were the same or different when the text preceded the speech (priming condition) than when it followed it.

Several studies have used subjective intelligibility judgments or subjective ratings of clarity to estimate the effect of priming (Rankovic and Levy, 1997; Sohoglu *et al.*, 2012, 2014). Jacoby *et al.* (1988) found that subjects perceived background noise to be reduced in loudness following priming. Freyman *et al.* (2013) found that listeners were biased toward the judgment that a primed message was processed with less-severe filtering than an unprimed message, whether or not it actually was.

Few studies have focused on the potential effects of fine differences in timing of the delivery of the prime in relation to the auditory message. It could be hypothesized that the immediacy and reduced memory load of simultaneous delivery would lead to the most significant pop-out effects, but it could also be predicted that simultaneous delivery would lead to a dividing of attention that would reduce priming effects. Since priming effects have been linked to auditory perceptual learning, the question of which temporal relationships between primes and target signals lead to optimal priming effects could be important. Only one study that we are aware of (Sohoglu *et al.*, 2014) contained an experiment that specifically manipulated in fine steps the timing of captions

relative to the auditory targets, which were isolated words in this case. Listeners were asked to judge the clarity of the words as the onset of a visual caption was varied in 11 steps between leading the auditory presentation by 1.6 s to trailing the auditory stimulus by 1.6 s. Sohoglu *et al.* (2014) found that clarity was judged to be better when captions were simultaneous with or preceded the auditory signal than when the caption was delayed. No difference was observed between clarity judgments for simultaneous presentation and judgments for the conditions where the caption led the auditory signal over the tested range up to 1.6 s.

The current study adapted the objective same-different task from Jones and Freyman (2012) to measure the importance of the timing of the delivery of a written prime or caption in relation to the auditory signal, and whether there were any major variations in the patterns of results across different masking and distortion conditions. Printed captions were delivered with 20 different temporal offsets that ranged from 2 s before to ~2 s after the initiation of an acoustic sentence. Four different combinations of target and masker, including one distortion condition, were used to determine the generality of temporal effects in priming. Two of the variations also included a measure of listeners' confidence in their judgments as a function of the temporal relationship of the caption and auditory message.

II. METHODS

A. Subjects

Eighty normal-hearing listeners, 20 per experiment (77 females, 3 males), with audiometric thresholds ≤ 20 dB hearing level (HL) at octave frequencies between 250 and 8000 Hz participated in this experiment. The mean age was 21.2 yr, ranging between 19 and 34 years old. Subjects were undergraduate students at the University of Massachusetts Amherst and received extra course credit for their participation in the study.

B. Stimuli

1. Target stimuli

The target stimuli were taken from a corpus of 320 non-sense sentences developed by Helfer (1997) and spoken by a college-aged female native speaker of standard American English. The use of sentence-length material rather than isolated words creates some imprecision in the specification of exactly when subjects were reading and hearing the same words as the sentences progressed. However, our preliminary informal listening prior to conducting the study suggested that the pop-out effect was not quite as compelling for isolated words, leading to our decision to use sentences. This issue is considered again in Sec. IV.

The sentences were syntactically but not semantically correct and contained three key words, for example, "A shop can frame a dog." These stimuli have been used in previous experiments (Helfer, 1997; Freyman *et al.*, 1999; Freyman *et al.*, 2001, 2004; Jones and Freyman, 2012). The three key words in each sentence cannot be determined from the semantic context of the sentence. On average, the total

duration of the auditory presentation of the sentence was 1.6 s (range 1.12–2.11 s).

Replacement words (foils) were developed by Jones and Freyman (2012) for all key words in the sentences. In foil trials, one key word selected from the three possible sentence positions was replaced by a foil word before presentation on the computer monitor. Foil words were chosen according to the following rules: (1) foils could not rhyme with the target word, (2) foils had the same number of syllables as the target word, (3) foils had the same stress pattern as the target word, (4) foils were not pronounced with more than one stress pattern, (5) no foil could be used more than once, (6) foils did not make sense in the context of the sentence, and (7) foils were the same part of speech as the target word (e.g., verbs were replaced with verbs).

2. Vocoding

In the vocoded condition (condition D), the target sentences were processed using four-channel vocoding with a noise carrier using the algorithm of Qin and Oxenham (2003). The frequency range of 80 to 6000 Hz was divided into four channels of equal bandwidth according to the equivalent rectangular bandwidth scale (Glasberg and Moore, 1990), using digital sixth-order Butterworth band-pass filters. Envelopes were extracted from the filter outputs by digitally low-pass filtering rectified signals with a cutoff frequency of the smaller of 300 Hz or half the bandwidth, using a second-order Butterworth filter. White noise filtered to have the same bandwidth as the filtered signals was multiplied by the appropriate envelope in the time domain to create noises that matched the temporal envelopes in each channel. The four modulated noises were summed to create a broadband four-channel speech-envelope-modulated noise (noise-excited vocoding) for each of the 320 sentences.

3. Maskers

Three maskers were employed in the present study: (1) steady state speech-spectrum noise (SSN) that was spectrally shaped to match the long-term spectrum of the target speech, (2) two-talker babble (TTB), and (3) time-reversed TTB. The TTB was composed of two different female talkers speaking a series of nonsense sentences similar to the target sentences. The specific nonsense sentences used in the masker were different from those used for the target stimuli. Pauses between sentences were removed using audio editing software to create a 35-s long continuous stream from each masking talker, which were then matched in root-mean-square amplitude and mixed together. The time-reversed TTB masker was created by taking the previously described TTB masker and reversing it in the time domain, producing a speech-like, but unintelligible masker.

C. Apparatus and procedure

The experiment was conducted in a double-walled sound-treated booth (IAC model #1604) (Industrial Acoustics Company, Inc., Bronx, NY) measuring 2.76 m × 2.55 m. Listeners were seated on a chair placed against

one wall of the booth and were instructed to face the front loudspeaker, but were not physically restrained. Two Realistic Minimus 7 loudspeakers (Radio Shack) were positioned at a distance of 1.3 m from the approximate center of the subject's head (1.2 m high). One loudspeaker was positioned directly in front (0° azimuth) and one was positioned to the right (60° azimuth) of the listener. For each auditory stimulus presentation, the target and masker were digitally mixed on a computer at the required signal-to-noise (S-N) ratio and output through a two-channel sound card (Creative Sound Blaster) at 22.05 kHz, lowpass filtered at 8.5 kHz (TDT FLT 5) (Tucker Davis Technologies, Alachua, FL), attenuated (TDT PA4), amplified (TDT HBUF5), power amplified (TOA P75D, TOA Canada Co., Mississauga, ON), and delivered to the loudspeakers. On each trial, a section of masker waveform of the same duration as the target waveform was selected randomly from the 35-s stream and presented simultaneously with the target. No rise or fall times were imposed. Text was presented on a computer monitor positioned in the sound-treated room directly in front of the listener, but well below ear level so as to produce minimal interference with the direct wave from either sound source. Daily calibration was conducted using a handheld sound level meter. The presentation level of the target sentences was 44 dBA and always from the 0-degree loudspeaker. Masker levels depended on the S-N ratio used in the different conditions.

D. Conditions

Four different conditions were tested, each employing 20 unique subjects. The S-N ratios were chosen based on pilot listening for each condition with simultaneous onset of caption and auditory signal (0-s offset) with the purpose of avoiding floor or ceiling effects. In condition A, unprocessed nonsense sentences were presented in the two-talker (TTB) masker at –3 dB S-N ratio from the speaker positioned at 0°. This condition with a target female talker presented in a background of female maskers without target-masker spatial separation has been shown to produce a great deal of informational masking (e.g., Freyman *et al.*, 2001). In condition B, the same target sentences were delivered in the presence of the TTB masker that had been time-reversed and presented from the 0° loudspeaker at an S-N ratio of –7 dB. This condition may also produce informational masking (Freyman *et al.*, 2001), but there is no confusion about which intelligible words to pay attention to. In condition C, the same unprocessed target stimuli were presented in the presence of the two-talker interference at –10 dB S-N ratio (in each of two masker loudspeakers) in a spatially separated configuration. The spatial separation between target and masker was created by presenting the masker from two loudspeakers (front and right) and imposing a 4-ms delay on the masker arriving from the front loudspeaker. Due to the precedence effect (Wallach *et al.*, 1949; Yost and Soderquist, 1984; Litovsky *et al.*, 1999), this masker was heard to the right, well separated from the target presented from the front loudspeaker. Thus, this condition is similar to condition A except that, as a result of the target-masker spatial separation, the speech masker is not likely to produce very much

informational masking (Freyman *et al.*, 2001). In condition D, the target stimuli were vocoded (see Sec. II B 2 for details) and presented with the SSN masker at 0 dB S-N ratio with both target and masker from the loudspeaker positioned at 0° azimuth. The vocoded condition was included to complement the work done by Davis *et al.* (2005) on priming with vocoded speech. The noise masking was necessary because the same-different task is otherwise too easy with the non-rhyming foils, even with only four channels of vocoding. Across all four conditions, the goal was to cover a range of degrees of informational masking, spatial and non-spatial masking, and both unprocessed and spectrally degraded speech, in order to acquire a comprehensive picture of temporal effects in priming.

E. Listener task

The task for listeners was to judge whether the sentence displayed on screen matched the auditory target. There was no time limit placed on their response. Subjects were instructed that the presentation order of the visual and auditory stimuli would vary throughout the experiment and they should maintain visual focus on the computer screen for the entirety of the experiment. Subjects were given four listening clues and instructed as follows: “When the sentences are different: (1) only one word will change, which could be at the beginning, middle, or end of the sentence, (2) the replaced word will not rhyme with the target word, (3) half of the trials are indeed different (foils) and (4) if you feel it improves your performance, please rehearse the sentence in your head or out loud.”

For two of the conditions (A and B), the task for the subject included an extra component beyond the basic same-different decision. A rating scale was employed in order to gain insight regarding subjects’ confidence in their responses. The question, “Were the sentences the same?” appeared on the computer screen, and subjects chose between six confidence indicators: “Sure Same”/“Sure Different,” “Probably Same”/“Probably Different,” or “Maybe Same”/“Maybe Different.” The six choices were displayed on the computer monitor and subjects used a computer mouse to select their response. When the sentences were the same, correct responses were “Sure Same,” “Probably Same,” and “Maybe Same.” When the sentences were different, correct responses were “Sure Different,” “Probably Different,” and “Maybe Different.” For conditions C and D, the task was similar, but confidence judgments were not requested. At the end of each trial, the question “Were the sentences the same?” appeared on the computer screen. Subjects responded by using the mouse to select one of two printed response options: “yes” for same or “no” for different.

F. Specification of temporal offsets

The start time of the auditory stimulus in relation to the initiation of the display of the caption is referred to as the temporal offset. Captions were presented on the computer monitor for 2.0 s. Relative to the beginning of this 2-s period, 20 different offsets for the start of the auditory message were

investigated, ranging from -2.0 to $+1.8$ s in 200-ms steps. Negative values of offset indicate that the auditory stimulus began before the visual stimulus, while positive values indicate that the visual prime appeared before the auditory stimulus. For example, in the 1.8-s offset condition, the auditory stimulus began 1.8 s after the visual display was initiated. Because the total duration of the caption was 2 s, there was only a 0.2-s overlap of the visual display and auditory signal in this condition. An offset value of 0-s indicated simultaneous delivery of both auditory and visual stimuli. The simultaneity was verified by inspecting video and audio recordings of zero offset trials, and was expected to be accurate to within one cycle of the refresh rate of the monitor (70 Hz).

Offset times were divided into four blocks each containing five consecutive offsets: (A) -2.0 to -1.2 , (B) -1.0 to -0.2 , (C) 0.0 – 0.8 , and (D) 1.0 – 1.8 s. The reason for this partial blocking of offset times was that a totally random assignment across an ~ 4 -s range seemed difficult for pilot subjects to adjust to. Each of the 4 offset blocks contained 60 trials for a total of 240 trials per subject. Within a 60-trial block, each of the 5 offset times was presented 12 times (6 Same trials and 6 Different trials). Within the six Different trials, each foil position (key word 1, 2, or 3) in the sentence was used twice. All of these variables were randomized within the block of 60 trials, and the order of the 240 sentences was completely randomized for each individual subject. The order of blocks was counterbalanced across subjects such that five subjects listened to block A (see above) first, five listened to block A second, five listened to block A third, and five listened to block A fourth. The same was true for blocks B, C, and D. These criteria were fulfilled for each of the four conditions.

III. RESULTS

The main results of the same-different experiment conducted in four different conditions are displayed in Fig. 1. Each panel shows performance in d' as a function of the offset of the auditory relative to the visual stimuli. The value of -2.0 s on the left-hand side of the abscissa indicates that the auditory stimulus began 2 s before the display of the text on the computer screen. Zero offset indicates simultaneous presentation of auditory and visual stimuli. The offset of $+1.8$ s indicates that the printed sentence preceded the auditory sentence by 1.8 s. Other values are intermediate.

Values of d' were calculated by treating the experiment as a simple yes/no task. Estimation of sensitivity for same-different tasks often is more complex (see Macmillan and Creelman, 2005, p. 215), but the treatment of the current data as yes/no arises from the assumption that the reading of the caption itself introduces no sensory variability. A hit was defined as a response of “Different” for trials in which the written and auditory messages were different. A false alarm was defined as a response of “Different” when written and auditory messages were the same. This choice of what stimulus-response combination was labeled a hit is consistent with examples given in Macmillan and Creelman (2005, p. 215).

Each data point was based on 240 experimental trials (12 trials per listener $\times$ 20 listeners) computed from the total

FIG. 1. Average performance in d' for all four processing and masking conditions as a function of the temporal difference between the initiation of the auditory stimulus and the text presentation. The averages are based on a different set of 20 listeners for each condition. Condition A used natural speech and a two-talker masker (both female) co-located with the target (also female). S-N ratio was -3 dB. Condition B was the same as condition A, but the masker was time reversed. S-N ratio was -7 dB. Condition C was the same as condition A, but with an additional spatially separated masker source. S-N ratio was -10 dB. Condition D used the same target stimuli but processed with four-channel noise-excited vocoding presented at a 0-dB S-N ratio in speech-spectrum unmodulated noise. The light lines for conditions A and B display d_a , a measure of sensitivity obtained from the fitted ROC (see the text).

percentage of hits and false alarms across all subjects. This analysis of the data (as opposed to averaging the d' values obtained from individuals) was desired because of the undefined d' values that occurred frequently when hits or false alarms were either 100% or 0% for individual subjects. It also, unfortunately, prevents a straightforward assessment of the variability of the data points using the standard error of the mean d' . To estimate the 95% confidence intervals of each of the data points in Fig. 1, the results were resampled 2000 times using the bootstrap method described by Efron and Tibshirani (1986). The size of these confidence intervals was relatively stable across the 80 total conditions, on average showing a confidence range from $\sim 0.45 d'$ units above the mean to $0.40 d'$ units below the mean data points shown in Fig. 1.

The lighter lines in Figs. 1(A) and 1(B) display d_a (see Macmillan and Creelman, 2005, p. 60), a measure of sensitivity based on the full receiver operating characteristic (ROC) that had been determined from the confidence ratings. The values of d_a were computed from the fits to the ROCs determined through maximum-likelihood estimation (see Macmillan and Creelman, 2005, p. 70). Because of the generally good correspondence between d_a and d' , only the d' measure will be considered further in this paper. However, the confidence judgments themselves will be discussed later.

The purpose of the pilot testing for each of the four conditions was to achieve S-N ratios that produced a similar and middle range of performance at the 0-s offset to avoid both floor and ceiling effects. The results for this simultaneous

condition show that the desired result was mostly achieved with d' performance, shown in Fig. 1, ranging from ~ 1.4 to 1.9 across the four target and masker conditions. The four different functions are grossly similar across the range of delays. Although there are some non-monotonic regions of the functions, the overall trend was for performance to rise as the offset of the auditory stimulus relative to the text was advanced from -2 s to 0 s (approaching the center dashed line from the left). For positive offsets (caption leading auditory presentation) shown by the data to the right of the dashed lines, overall performance continued to improve somewhat, although there was clearly a great deal of fluctuation within some of the functions and variation across the conditions in the extent of this fluctuation.

Inspection of the hit and false alarm data suggested that these fluctuations in d' , as well as some of the differences between functions, occurred with very small fluctuations in hit and false alarm rates and may not be of further interest. Because the overall pattern of results was relatively similar across the four conditions, the hit and false alarm rates were averaged and plotted in Fig. 2 to characterize these overall trends. Each hit and false alarm value was based on 480 trials (6 trials per listener $\times$ 80 listeners).

The data in Fig. 2 reveal that changes in false alarm rates (when Same trials were reported as “different”) were most responsible for the changes in performance as a function of auditory temporal offset. When the acoustic signal began well before the caption (approaching the left edge of the figures), the false alarm rate was $\sim 30\%$ across the four

target/masker conditions. When the caption was presented before the auditory message (right side of the figures), the false alarm rates were quite low, averaging $\sim 10\%$ or lower. These trials, where the caption was the same as the auditory message and was presented before the auditory message, were expected to produce the pop-out effect described in the Introduction, and it was rare in these cases for subjects to incorrectly label these Same trials as “different.”

In contrast, hit rates were not as high as false alarm rates were low (there was a bias toward reporting “same”), and hit rates changed less dramatically as the timing of the auditory message was manipulated relative to the presentation of the text. The hit rates reflect correct reporting on trials in which the text and auditory message were different, so these were not fully priming trials, even when the caption preceded the text. That is, the presentation of the foil rather than the target word during the text presentation meant that one of the three key words was never primed at any onset asynchrony, including those where the caption was presented first. No semantic cues were available in the nonsense sentences to help predict these words. Therefore, it is perhaps understandable that hit rates were less affected by temporal offsets than were false alarm rates.

To characterize the overall trends in the data, the hit and false alarm data were each fitted with three linear equations, the limits of which were chosen based on visual inspection of data. A variety of fits were possible for each of the functions separately, but we were able to achieve a reasonably good fit despite constraining the breakpoints to be the same for both hits and false alarms. The segments were: from -2.0 s to -1.2 s offset, from -1 to $+0.4$ s, and from $+0.6$ to $+1.8$ s. The first segment shows a slight downward trend for both hits and false alarms. In the second segment, hit rates increased, although not dramatically, while false alarm rates fell relatively steeply. In the third segment, both hits and false alarm rates showed essentially flat functions. The steady change in false alarm rates in the middle segment

FIG. 2. Average (and ± 1 standard error) hit and false alarm rates across the four conditions. Each hit and false alarm value was based on 480 trials (6 trials per listener $\times$ 80 listeners). Lines are three-segment linear least-squares fits to the data, with breakpoints determined by visual inspection.

presumably reflects changes in the amount of overlap between auditory and text presentations. Perhaps even small amounts of temporal overlap led to pop-out effects that resulted in reduced false alarm rates.

The hit and false alarm rates taken from the equation fits were used to compute an aggregate d' at each of the 20 offsets. This function is displayed in Fig. 3, which summarizes the effect of temporal asynchrony of auditory stimuli and text on performance on the same-different task across the different conditions of distortion and masking. The figure shows little change in performance over the range of -2 s to -1.2 s and steady growth in performance from -1.2 to $+0.4$ s, with very little additional growth for still more positive offsets. As mentioned in Sec. II, the mean duration of the acoustic sentences was 1.6 s. This was also the duration of the growth segment in d' performance and may reflect increasing and then full overlap of hearing the auditory stimulus and reading and processing of the caption.

To verify that these trends seen in Fig. 3 were supported statistically by the raw data, t -tests were conducted to test for differences at 5 key temporal offsets along these functions, using all 80 d' values (1 for each subject) at each offset. As mentioned earlier, within individual subjects, there were sometimes occurrences of either 100% or 0% hits or false alarms where d' could not be calculated. In such cases, for the purposes of this statistical analysis, it was necessary to apply a correction factor before calculating d' from the hit and false alarm results. A proportion of 1.0 was adjusted to 0.917 ($1 - 1/2N$, where $N = 6$ trials), and a proportion of 0.0 was adjusted to 0.083 ($1/2N$), as recommended by Macmillan and Creelman (2005, p. 8). Along the lower left flat portion of the function, the difference between -2.0 s and -1.2 s was not significant ($p = 0.250$). Within the rising portion of the function the difference between -1.2 s and 0 s was highly significant ($p < 0.0001$). The difference between 0 s and $+0.4$ s was also significant ($p < 0.001$). At the upper right portion of the function, the difference between $+0.4$ s

FIG. 3. Performance in d' across all 4 conditions and 80 listeners using hit and false alarms calculated from the fitted functions shown in Fig. 2.

and +1.8 s was not significant ($p = 0.645$). It should be noted that these t -tests were conducted *post hoc*, after observation of the functions, and thus must be interpreted with appropriate caution. However, the trends seen in Fig. 3 derived from the fitted hit and false alarm rates in Fig. 2 do appear to be well supported by these statistical analyses of the raw d' results. The results indicate that it was better to present the auditory stimulus simultaneously with the caption than before the caption, but that it was still better to delay the acoustic stimulus by ~ 0.5 s or more.

The confidence judgments obtained in the presence of the natural and time-reversed two-talker maskers allowed for the measurement of ROCs in those conditions. Figure 4 displays example ROCs for the data at three selected offsets, -2 s, 0 s, and 1.4 s (see Macmillan and Creelman, 2005, p. 55, for the computational procedure). Slopes of the ROCs fitted with a maximum-likelihood estimation procedure (Macmillan and Creelman, 2005, p. 354) averaged 1.09 across both types of maskers and all 20 delays. No obvious pattern was observed across temporal offsets. The fact that the slopes were not very different from 1.0, and the d' and d_a metrics were shown to be reasonably consistent in Fig. 1, suggests that d' was an appropriate metric for describing the results.

Figure 5 displays the number of correct “sure,” “probably,” and “maybe” responses as a function of temporal offset for the natural and time-reversed masker conditions in the top and bottom pairs of panels, respectively. The left panels show the three different types of correct responses to Same stimuli and the right panels to Different stimuli. Incorrect responses cluster toward the bottom of each panel and only averages of the three incorrect response classes are shown (by dashed lines) to avoid excessive clutter. The total number of possible responses was 120 (6 Same or Different stimuli per condition $\times$ 20 listeners). The plots show that the number of “sure” responses of either type increased as the temporal offset of the auditory stimulus was delayed, while the number of “probably” or “maybe” responses varied less. As the temporal offset approached simultaneous delivery and increased further to conditions where the caption preceded the acoustic speech signal, correct “sure same” responses exceeded those of correct “sure different” responses substantially. That is, highest confidence was obtained when the caption preceded the acoustic presentation and the text and auditory message were identical, revealing the effect of priming. For such conditions, the subjects responded correctly with “sure same” on roughly two-thirds of the trials ($\sim 80/120$).

IV. DISCUSSION

The purpose of this study was to attempt to quantify pop-out or priming effects and discover the conditions in which these effects are optimal. We determined that the timing of a written caption relative to the initiation of an auditory message was important. Specifically, this timing influenced listeners’ ability to judge whether the auditory and text messages were the same. The perception of increased salience and clarity of the auditory message

following the delivery of a caption was inferred from objective performance in the same-different task.

The current experiment investigated the effects of priming in three natural speech conditions in which the auditory message was masked by a two-talker speech stream that was either co-located or spatially separated (known to cause a large difference in the amount of informational masking), or was presented in a condition of masker time-reversal so that the masker could not be understood. The study also included a four-channel noise vocoding condition. The timing of the delivery of the written caption affected performance in all four conditions. However, despite having 20 subjects per condition, the detailed fluctuations in most of the functions discouraged an attempt to highlight or test for any differences that may exist among them. Instead (see Figs. 2 and 3), we focus here on the commonality in the general shape across four different conditions of masking and distortion and use the increased power of combining across 80 subjects to describe these trends better.

Across these functions, changes in performance were driven largely by changes in false alarm rates. At delays where the captions led the auditory signal by 400 ms or more, it was rare for listeners to report that the visual and auditory messages were different when they were actually the same. This is evidence of the positive effects of priming in these conditions. Confidence ratings were measured for the nonsense sentences presented in natural or time-reversed two-talker masking co-located with the target (see Figs. 4 and 5). Analysis of these ratings suggests that priming led to highly confident “same” judgments when captions and auditory messages were indeed the same. Lower confidence was observed for correct “different” responses. This asymmetry, as well as asymmetries in the rate of change of hit and false alarm rates as a function of temporal offset, are most easily explained by assuming that the trends observed reveal an effect of temporal offset on the perception of pop-out due to priming, as opposed to a more general effect of which stimulus, easy or hard, is presented first.

The results also indicate that for unprimed conditions (where the initiation of the caption followed the onset of the auditory message), the degree of delay was important. For example, in the current study, performance for a delay of -1.2 s was worse than that for a delay of -0.4 s with respect to false alarm rates and overall d' (see Figs. 1–3). This finding was generally consistent with results from measures of subjective clarity of isolated distorted words presented in varying temporal proximity to printed text (Sohoglu *et al.*, 2014). At least for our sentence-length materials, it appears that some overlap of auditory message and caption is better than no overlap. In these cases, it could be the case that the caption was read quickly while the auditory nonsense sentence (averaging 1.6 s in duration) was still ongoing. That is, it is possible that for some sentences the last words of the captions were read ahead of or at least simultaneously with the last words in the auditory message, even when the caption was slightly delayed. To investigate this possibility, it may be helpful to collect eye-tracking data simultaneously with the collection of the auditory perceptual data.

FIG. 4. ROC functions computed from confidence judgment data obtained for the TTB and TTB reversed conditions for three sample values of offset.

Given these uncertainties, it may appear that it would have been better to conduct this study on the timing of primes with simpler stimuli like single words rather than sentences. As mentioned in Sec. II, our preliminary informal listening suggested a stronger pop-out effect from priming in sentences versus isolated words. Also, [Hervais-Adelman et al. \(2008\)](#) noted that absolute performance growth with priming training was not as strong for the isolated words used in that study in comparison to the longer sentence-length material used by [Davis et al. \(2005\)](#). The effects of prior captions on subjective judgments of clarity ([Sohoglu et al., 2014](#)) for isolated words seem relatively modest compared to what one might expect based on the dramatic increase in clarity that priming of sentences appears to create on informal listening, although a definitive conclusion about this must await a direct comparison between word and sentence length material using the same task. Finally, sentences may have more validity and interest for real-world listening. So, although using sentences presents some complexities and difficulties in interpretation, there may be some advantages as well.

As described in the Introduction, research on the effects of priming on intelligibility judgments and on training have presented written captions both ahead of and simultaneously with the initiation of the auditory speech signal, without a

great deal of attention paid to this variable in most of these studies. At least for the stimuli and task used here, it was found that although simultaneous and prior presentations are both effective, simultaneous presentation was not the optimal condition. Better discrimination performance was achieved when captions were delivered a few hundred milliseconds to a few seconds ahead of the auditory message. It was probably important for the subject to have the opportunity to read some or all of the words before they heard them in order for the full pop-out effect to occur, which was facilitated by the short delay. It is not yet clear how well the details of the forms of these functions can be extended to other types of target speech materials, or whether the specific duration over which the caption is available to read is important (2 s, in this case). Clarification of these issues will have to await further investigation.

To the extent that they can be generalized, the current results and future studies on this topic may give direction for the design of training protocols where listeners attempt to improve their speech recognition for words and sentences through feedback that includes some form of priming in short-term (e.g., [Davis et al., 2005](#); [Hervais-Adelman et al., 2008](#); [Loebach et al., 2010](#); [Hervais-Adelman et al., 2011](#)) and longer-term training protocols (e.g., [Burk and Humes, 2007](#); [Kuchinsky et al., 2014](#)). Also, priming in the form of textual captions has been used in computerized aural rehabilitation programs designed mostly for home use (e.g., LACE, [Sweetow and Henderson Sabes, 2004, 2006](#), and “The Listening Room,” produced by [Advanced Bionics](#)). In the LACE paradigm, optional feedback is provided via a written caption delivered ~ 1 s before the masked or degraded stimulus is replayed. This offset is well within the region shown to produce optimal priming effects in the current study and in the data of [Sohoglu et al. \(2014\)](#).

The current data demonstrate the effectiveness of priming in overcoming masking of various different types, as well as distortion. However, the existence of the pop-out effect for a particular masking condition does not lead directly to an assumption that it can be useful for training in each of these conditions, particularly, in the generalization to stimuli that subjects do not experience during training. Some conditions, such as spectral and temporal fine structure degradation through vocoding, may be more amenable to perceptual learning than other conditions, e.g., understanding speech masked by steady background noise. For those conditions that show both priming and learning effects, it will be interesting to determine whether conditions that optimize priming also optimize perceptual learning.

Other means of delivering primes could potentially be effective in the same manner as textual captions. For example, [Freyman et al. \(2004\)](#) showed that visual delivery of the prime by caption, auditory delivery of the prime by the target talker, and auditory delivery of the prime by a different talker, all produced almost the identical priming effect, albeit in a different task from the current one. [Davis et al. \(2005\)](#) used visual captions and clear auditory presentations interchangeably in their perceptual learning experiments. These results suggest that it is the message itself that is most important, not the means by which the message is presented.

FIG. 5. The number of correct “sure,” “probably,” and “maybe” responses as a function of offset for the TTB and TTB reversed conditions in the top and bottom pairs of panels, respectively. The left panels show the correct responses to Same stimuli and the right panels show the correct responses to Different stimuli. The total number of possible responses was 120 (6 Same or Different stimuli per condition $\times$ 20 listeners). Dashed lines show the average of the “sure,” “probably,” and “maybe” incorrect responses.

If that is the case, then other means of delivering primes might be effective as well. One example we have considered is the case of simultaneous delivery of sign and speech (“Simultaneous Communication,” e.g., [Marmor and Pettito, 1979](#); [Whitehead et al., 1997](#)), which is the language of instruction at some schools for the Deaf. For hearing-impaired listeners who are competent signers, signs presented to them simultaneously with speech could possibly produce priming effects, enhancing the internal representation of the auditory signal. Although there is much research to be done first, if priming effects in simultaneous communication can also be demonstrated in listeners with hearing impairment, then the precise timing of the delivery of the signs in relation to the speech may also be important (see [Whitehead et al., 1997](#)). Thus, the current study’s results may be helpful in guiding future studies on the question of timing of signs and speech in simultaneous communication.

Finally, current and future work on the effects of the timing of primes in relation to an auditory message could ultimately advance our understanding of the mechanisms that underlie increased perceptual clarity following priming. For example, [Sohoglu et al. \(2012\)](#) discovered that cortical responses in the superior temporal gyrus are different when speech is enhanced by prior knowledge than when it is

enhanced by reducing the level of distortion. Thus, although speech may sound clearer as a result of prior knowledge, it seems not to be processed in the same way as less distorted acoustic signals. As research continues to attempt to understand the pop-out effect mechanistically, knowledge about timing effects in priming should be helpful in informing theories of the basis of the phenomenon.

ACKNOWLEDGMENTS

The authors wish to thank Derina Boothroyd and Decia DeMaio for their contributions to data collection and analysis, and Neil Macmillan and Caren Rotello for guidance with signal detection theory computations and interpretation. Associate Editor Emily Buss and two anonymous reviewers made valuable constructive criticisms of an earlier version of the paper. The authors are grateful for the financial support of National Institute on Deafness and Other Communication Disorders Grant No. DC01625.

- Burk, M. H., and Humes, L. E. (2007). “Effects of training on speech recognition performance in noise using lexically hard words,” *J. Speech Lang. Hear. Res.* **50**, 25–40.
- Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (2005). “Lexical information drives perceptual learning of

- distorted speech: Evidence from the comprehension of noise-vocoded sentences,” *J. Exp. Psych. Gen.* **134**, 222–241.
- Efron, B., and Tibshirani, R. (1986). “Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy,” *Stat. Sci.* **1**, 54–75.
- Ezzatian, P., Li, L., Pichora-Fuller, K., and Schneider, B. (2011). “The effect of priming on release from informational masking is equivalent for younger and older adults,” *Ear Hear.* **32**, 84–96.
- Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2001). “Spatial release from informational masking in speech recognition,” *J. Acoust. Soc. Am.* **109**, 2112–2122.
- Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2004). “Effect of number of masking talkers and auditory priming on informational masking in speech recognition,” *J. Acoust. Soc. Am.* **115**, 2246–2256.
- Freyman, R. L., Griffin, A. M., and Macmillan, N. A. (2013). “Priming of lowpass-filtered speech affects response bias, not sensitivity, in a bandwidth discrimination task,” *J. Acoust. Soc. Am.* **134**, 1183–1192.
- Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (1999). “The role of perceived spatial separation in the unmasking of speech,” *J. Acoust. Soc. Am.* **106**, 3578–3588.
- Glasberg, B. R., and Moore, B. C. J. (1990). “Derivation of auditory filter shapes from notched-noise data,” *Hear. Res.* **47**, 103–138.
- Greenspan, S. L., Nusbaum, H. C., and Pisoni, D. B. (1988). “Perceptual learning of synthetic speech produced by rule,” *J. Exp. Psych.: Learn., Mem., Cognit.* **14**, 421–433.
- Helfer, K. S. (1997). “Auditory and auditory-visual perception of clear and conversational speech,” *J. Speech Lang. Hear. Res.* **40**, 432–443.
- Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., and Carlyon, R. P. (2008). “Perceptual learning of noise vocoded words: Effects of feedback and lexicality,” *J. Exp. Psych.: Hum. Percept. Perf.* **34**, 460–474.
- Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., and Carlyon, R. P. (2011). “Generalization of perceptual learning of vocoded speech,” *J. Exp. Psych.* **37**, 283–295.
- Jacoby, L. L., Allan, L. G., Collins, J. C., and Larwill, L. K. (1988). “Memory influences subjective experience: Noise judgments,” *J. Exp. Psych.: Learn., Mem., Cognit.* **14**, 240–247.
- Jones, J. A., and Freyman, R. L. (2012). “Effect of priming on energetic and informational masking in a same-different task,” *Ear Hear.* **33**, 124–133.
- Kuchinsky, S. E., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., and Eckert, M. A. (2014). “Speech perception training for older adults with hearing loss impacts word recognition and effort,” *Psychophysiology* **51**, 1046–1057.
- Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (1999). “The precedence effect,” *J. Acoust. Soc. Am.* **106**, 1633–1654.
- Loebach, J. L., Pisoni, D. B., and Svirsky, M. A. (2010). “Effects of semantic context and feedback on perceptual learning of speech processed through an acoustic simulation of a cochlear implant,” *J. Exp. Psych.: Hum. Percept. Perf.* **36**, 224–234.
- Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User’s Guide*, 2nd ed. (Erlbaum, Mahwah, NJ).
- Marmor, G., and Pettito, L. (1979). “Simultaneous communication in the classroom: How well is English grammar represented?,” *Sign Lang. Stud.* **1023**, 99–136.
- Qin, M. K., and Oxenham, A. J. (2003). “Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers,” *J. Acoust. Soc. Am.* **114**, 446–454.
- Rankovic, C. M., and Levy, R. M. (1997). “Estimating articulation scores,” *J. Acoust. Soc. Am.* **102**, 3754–3761.
- Schwab, E. C., Nusbaum, H. C., and Pisoni, D. B. (1985). “Some effects of training on the perception of synthetic speech,” *Hum. Factors* **27**, 395–408.
- Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). “Speech recognition with primarily temporal cues,” *Science* **270**, 303–304.
- Sohoglu, E., Peelle, J. E., Carlyon, R. P., and Davis, M. H. (2012). “Predictive top-down integration of prior knowledge during speech perception,” *J. Neurosci.* **32**, 8443–8454.
- Sohoglu, E., Peelle, J. E., Carlyon, R. P., and Davis, M. H. (2014). “Top-down influences of written text on perceived clarity of degraded speech,” *J. Exp. Psych.: Hum. Percept. Perf.* **40**, 186–199.
- Sweetow, R. W., and Henerson Sabes, J. H. (2004). “The case for LACE: Listening and auditory communication enhancement training,” *Hear. J.* **57**, 32–38.
- Sweetow, R. W., and Henerson Sabes, J. H. (2006). “The need for development of an adaptive Listening and Communication Enhancement (LACE) Program,” *J. Am. Acad. Audio.* **17**, 538–58.
- The Listening Room™ Online Auditory Rehabilitative Website. Advanced Bionics. Web (Last viewed February 2, 2015).
- Wallach, H., Newman, E. B., and Rosenzweig, M. R. (1949). “The precedence effect in sound localization,” *Am. J. Psychol.* **62**, 315–336.
- Whitehead, R. L., Schiavetti, N., Whitehead, B. H., and Metz, D. E. (1997). “Effect of sign task on speech timing in simultaneous communication,” *J. Commun. Disord.* **30**, 439–455.
- Yang, Z., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B. A., and Li, L. (2007). “The effect of voice cuing on releasing Chinese speech from informational masking,” *Speech Commun.* **49**, 892–904.
- Yost, W. A., and Soderquist, D. R. (1984). “The precedence effect: Revisited,” *J. Acoust. Soc. Am.* **76**, 1377–1383.