Evidence for independent time-unit processing of speech using noise promoting or suppressing masking release (L)

Eric W. Healy, Carla L. Youngdahl, and Frédéric Apoux

Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210

(Received 4 March 2013; revised 31 July 2013; accepted 23 December 2013)

The relative independence of time-unit processing during speech reception was examined. It was found that temporally interpolated noise, even at very high levels, had little effect on sentence recognition using masking-release conditions similar to those of Kwon *et al.* [(2012). *J. Acoust. Soc. Am.* **131**, 3111–3119]. The current data confirm the earlier conclusions of Kwon *et al.* involving masking release based on the relative timing of speech and noise. These data also indicate substantial levels of independence in the time domain, which has implications for current theories of speech perception in noise. © 2014 Acoustical Society of America.
[<http://dx.doi.org/10.1121/1.4861363>]

PACS number(s): 43.66.Mk, 43.66.Dc, 43.71.An [PBN]

Pages: 581–584

I. INTRODUCTION

The relative independence with which various time-frequency (T-F) units of speech may be processed is relevant to the glimpsing model of speech perception in noise (Cooke, 2006; Apoux and Healy, 2013) and to binary-masking techniques used to separate speech from noise (e.g., Wang, 2005). In the glimpsing model, the auditory system bases the speech percept on T-F units containing relatively clean speech and essentially disregards T-F units dominated by noise. In ideal binary masking, the signal-to-noise ratio (SNR) of each T-F unit is analyzed to determine if it is dominated by speech or by noise. Speech is then represented by only the speech-dominant units, while noise-dominant units are discarded. It has been found that binary-masking can substantially raise intelligibility, despite large numbers of discarded (silent) T-F units in the output stimuli (Anzalone *et al.*, 2006; Brungart *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2009).

For the auditory system to perform a glimpsing-like analysis to extract speech from noise, it must be able to process the T-F units containing speech independently from those containing noise, even when those units are interleaved in both frequency and time. Information concerning the relative independence of spectral frequencies or time windows can be gained from psychoacoustic studies involving simple non-speech stimuli. Indeed, the concept of the critical band (e.g., Fletcher, 1940) and the techniques used to measure it (e.g., Patterson, 1976) are based on the assumption that various spectral frequencies are analyzed independently. Other data exist to indicate that robust integration of information across frequencies is also possible (e.g., Hall *et al.*, 1984; Yost and Sheft, 1989). Work involving non-simultaneous masking (e.g., Miller, 1947a; Luscher and Zwislocki, 1949) and temporal integration (e.g., Munson, 1947) provide information concerning independence of events in time. However, the application of psychoacoustic findings to speech perception is likely not straightforward. Speech involves highly structured acoustic patterns in which trajectories of change over both time and frequency convey

considerable amounts of information. Further, speech perception involves not only the reception of these complex acoustic patterns, but also top-down cognitive processes including linguistic pattern matching and memory that can both potentially affect the way in which the signals are processed. As a result, there are numerous examples of discrepancies between psychoacoustic and speech-perception results, and the relative independence of various speech T-F units is therefore best studied using speech stimuli.

Apoux and Healy (2009, 2010) demonstrated high levels of independence in the spectral-frequency domain by showing that interleaved bands of noise had to be considerably elevated in level to affect speech recognition. The evaluation of T-F unit independence in the time domain is more difficult to assess. Studies involving gated speech indicate that speech can be integrated across numerous silent gaps and remain intelligible (e.g., Miller and Licklider, 1950). But it is seemingly more relevant to assess independence in the time domain when there is a need to ignore intervening time periods containing noise. One option would be to examine gated speech having temporally interpolated noise. However, in this case, the noise may not be processed independently and may instead be used as a substrate to reconstruct the missing portions of speech (Warren *et al.*, 1994).

Studies involving masking release (MR) may offer the best opportunity to assess independence of speech units in the time domain. These studies indicate that intelligibility is increased when fluctuations are introduced to a masker (e.g., Miller, 1947b; Carhart *et al.*, 1966), thus providing glimpses of speech in the masker dips. Kwon *et al.* (2012) examined MR under conditions specifically designed to promote or suppress it. In one condition (–MR), masker fluctuations occurred simultaneously with speech-energy fluctuations, thus maximizing the masking of speech by noise and minimizing MR. In another condition (+MR), masker fluctuations occurred in inverse proportion to speech-energy fluctuations, thus minimizing the masking of speech by noise and maximizing MR. The speech reception threshold (SRT) for normal-hearing (NH) listeners was found by Kwon *et al.* to be –22 dB in the +MR condition, indicating that noise had

to be 22 dB *above* the level of the speech for 50% sentence recognition. In a second experiment involving gated noise, NH listeners produced essentially 100% sentence intelligibility at an SNR of -25 dB. If these large negative SNR values are accurate, then it may be concluded that very high levels of independence exist in the time domain during speech processing, because even extremely high levels of temporally interpolated noise have little effect on intelligibility.

However, before this conclusion can be made, an alternative interpretation for the large negative SNRs observed in Kwon *et al.* must be ruled out. As Fig. 1 shows, silence present at the beginning and end of the speech files caused a considerable amount of noise energy to fall outside the speech duration in the +MR condition. As a result, the actual SNR in the +MR condition was likely more favorable than reported when only the duration of the speech is considered. Further, this elevated SNR, and not the temporal relationship between speech and noise, could have affected the conclusions made by Kwon *et al.* (2012) regarding masking release in the +MR versus -MR conditions.

In the current study, this issue was examined. Specifically, silences preceding and following each sentence were eliminated so that the overall amount of noise energy was the same across conditions when measured during the utterance. The current study therefore had two goals. One was to clarify a potential confound in Kwon *et al.* (2012) and the other broader goal was to examine the relative independence with which T-F units may be processed in the time domain.

II. METHOD

A group of ten female listeners between the ages of 19 and 38 yrs participated (mean = 24.3). All had pure-tone audiometric thresholds of 20 dB hearing level (HL) or better from 250 to 8000 Hz (ANSI, 2004), and none had prior exposure to the sentence materials used. They were recruited from undergraduate courses at The Ohio State University and received a monetary incentive.

The construction of stimuli and design of procedures followed closely that in Kwon *et al.* (2012). The same male-speaker recordings of IEEE sentences (IEEE, 1969) were employed. However, prior to processing, all silence preceding and following the sentence in each file was eliminated.

In every condition, speech was masked with noise shaped to the spectrum of each individual sentence using a bank of 1/3-octave bandpass filters centered from 120 to 7680 Hz. The duration of each noise matched that of each individual sentence, and SNR was defined by the overall average RMS level of the sentence relative to the overall average RMS level of the noise.

A steady condition was created by mixing speech-shaped noise with each sentence. A -MR condition was created by mixing each sentence with noise that matched the amplitude envelope of that sentence. To create this condition, the broadband speech envelope was first obtained using full-wave rectification and low-pass filtering at 20 Hz (elliptic filter, $N = 3$). Speech-modulated noise was then created by multiplying speech-shaped noise by the envelope of each sentence. Finally, the noise and speech were mixed. A +MR condition was created by mixing each sentence with noise that was modulated by the inverse of the sentence envelope. To create this condition, the level of speech-shaped noise in each 50-ms window was adjusted so that it was inversely proportional to the level of the speech in each window (see Fig. 1 and Kwon *et al.*, 2012, for details). The use of 20-Hz envelope smoothing in -MR and 50-ms windows in +MR allowed the temporal fluctuation rate of the noise envelopes to match across conditions.

The SRT was measured in each condition using an adaptive procedure having a 1-down/1-up decision rule (Levitt, 1971). Each trial involved presentation of one sentence, and correct trials were defined as those in which three or more of the five keywords in the sentence were reported. The SNR was decreased after each correct trial and increased after each incorrect trial. Blocks of trials began with the SNR at +5 dB. It was adjusted in 4-dB steps during the first two reversals and in 2-dB steps thereafter. Blocks were terminated after eight reversals and the SNRs at the final four reversal points were averaged to represent the SRT. Three such blocks were averaged to obtain the threshold estimate for each subject in each condition. The protocol called for additional testing if the standard deviation of the first three blocks exceeded 3 dB for any subject in any condition, but no additional testing was required. Condition order and sentence-to-condition correspondence was randomized for each listener.

FIG. 1. Waveforms of (A) a sentence, (B) noise modulated by the inverse of the speech envelope, (C) noise modulated by the speech envelope, (D) steady noise. Waveforms on the right show the three noise conditions. From Kwon *et al.* (2012).

Stimuli were delivered from a PC using Echo Gina3G digital-to-analog converters. They were delivered diotically via Sennheiser HD280 headphones and scaled so that the overall presentation level for each trial was 65 dBA when measured in a flat-plate coupler (Larson Davis AEC101). Subjects were seated in a double-walled booth with the experimenter, who controlled the delivery of stimuli and recorded the accuracy of responses.

III. RESULTS

Figure 2 shows the SRT in each of the three conditions (steady, -MR, +MR). Panel A displays individual NH data from [Kwon et al. \(2012\)](#) and panel B displays individual data from the current subjects. Panel C displays group mean performance in each condition for both studies. As can be seen, performance was highly similar across the two studies in the steady condition (group means within 0.2 dB across studies) and in the -MR condition (group means within 0.3 dB across studies). The mean SRT in +MR, the condition of main interest, was 2.9 dB higher in the current study than in [Kwon et al. \(2012\)](#). A series of planned comparisons

(uncorrected two-tailed t tests) revealed that performance was equivalent between the previous and current datasets in the steady condition [$t(18) = 0.46$, $p = 0.65$] and in the -MR condition [$t(18) = 0.91$, $p = 0.38$]. Performance differed significantly in the +MR condition [$t(18) = 4.08$, $p < 0.01$].

IV. DISCUSSION

A first conclusion involves the comparison of the current results to those of [Kwon et al. \(2012\)](#). It is clear from Fig. 2 that the elimination of silence from the beginning and end of the sentence files had no effect on performance in the steady or -MR conditions. This should be expected because these beginning and end regions do not substantially affect the overall average RMS of the steady noise, and the silences in these regions contribute only minimally to the average RMS of the -MR noise. In contrast, an increase in threshold averaging 3 dB was observed in the +MR condition. This can be understood in terms of the elimination of noise bursts preceding and following each sentence and the redistribution of this noise energy throughout the actual sentence duration. As a result of this redistribution, the SNR was effectively

FIG. 2. Panels A and B display mean speech reception thresholds (and standard deviations) for each subject in each noise condition. Panel A displays data from [Kwon et al. \(2012\)](#) and panel B displays the current data. Panel C displays group mean performance (and standard deviations) in each condition in each study.

lower currently than in the +MR condition of Kwon *et al.* (2012), despite equal nominal SNRs.

Although the current results indicate that thresholds in the +MR condition of Kwon *et al.* (2012) were affected by an inflated SNR, the question of importance involves the magnitude of this effect relative to the overall magnitude of MR observed. After the current reduction in masking release of 3 dB, a large amount of release remains in the condition designed to promote it. Further, Fig. 2(C) shows that the pattern of thresholds across conditions is the same across studies. It may therefore be concluded that the influence of SNR inflation in Kwon *et al.* (2012) was modest, and that the conclusions made there are relatively unaffected by this aspect of processing. Here, we concur that vastly different amounts of MR can be obtained by NH listeners under conditions that vary only in the timing of noise fluctuations relative to the timing of speech fluctuations. Further, the cochlear-implant users tested by Kwon *et al.* (2012) produced little MR under conditions that were even slightly more favorable than those employed here. Thus, this earlier conclusion that masking release in cochlear-implant users is severely limited, even under optimal conditions, is also confirmed.

A second conclusion involves the independence of speech time-unit processing. For the auditory system to execute a glimpsing-like analysis to extract speech from noise, it must perform the same type of analysis employed during ideal binary masking. Each T-F unit must be assessed to determine if it is dominated by speech or by noise, and those T-F units dominated by speech must be grouped to form the foundation for a speech percept. Further, those T-F units dominated by noise must be essentially disregarded, despite that they are interleaved with speech-dominated units in both time and frequency. Both the assessment of individual T-F units and the perceptual separation of interleaved T-F units require that these units be processed with considerable independence.

The current results support the conclusion that high levels of independence exist in the time domain during speech processing, by demonstrating that temporally interpolated noise must be present at very high levels (19 dB re. level of speech) to reduce intelligibility of words in sentences to 50%. This result suggests that the auditory system is able to extract and combine T-F units containing speech, while ignoring interpolated T-F units containing noise, even when the levels of noise are substantially elevated.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication Disorders (DC008594 to E.W.H. and DC009892 to F.A.). We are grateful to Bomjun Kwon and Sarah Yoho for helpful discussions of this work.

- ANSI (2004). *ANSI S3.21 (R2009). American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (American National Standards Institute, New York).
- Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (2006). "Determination of the potential benefit of time-frequency gain manipulation," *Ear. Hear.* **27**, 480–492.
- Apoux, F., and Healy, E. W. (2009). "On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence," *Hear. Res.* **255**, 99–108.
- Apoux, F., and Healy, E. W. (2010). "Relative contribution of off- and on-frequency spectral components of background noise to the masking of unprocessed and vocoded speech," *J. Acoust. Soc. Am.* **128**, 2075–2084.
- Apoux, F., and Healy, E. W. (2013). "A glimpsing account of the role of temporal fine structure information in speech recognition," in *Basic Aspects of Hearing: Physiology and Perception*, edited by B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel (Springer, New York), pp. 119–126.
- Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2006). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.* **120**, 4007–4018.
- Carhart, R., Tillman, T. W., and Johnson, K. R. (1966). "Binaural masking of speech by periodically modulated noise," *J. Acoust. Soc. Am.* **39**, 1037–1050.
- Cooke, M. P. (2006). "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.* **119**, 1562–1573.
- Fletcher, H. (1940). "Auditory patterns," *Rev. Mod. Phys.* **12**, 47–65.
- Hall, J. W., Haggard, M. P., and Fernandes, M. A. (1984). "Detection in noise by spectro-temporal pattern analysis," *J. Acoust. Soc. Am.* **76**, 50–56.
- IEEE (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 225–246.
- Kwon, B. J., Perry, T. T., Wilhelm, C. L., and Healy, E. W. (2012). "Sentence recognition in noise promoting or suppressing masking release by normal-hearing and cochlear-implant listeners," *J. Acoust. Soc. Am.* **131**, 3111–3119.
- Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.* **49**, 467–477.
- Li, N., and Loizou, P. C. (2008). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.* **123**, 1673–1682.
- Lüscher, E., and Zwislocki, J. (1949). "Adaptation of the ear to sound stimuli," *J. Acoust. Soc. Am.* **21**, 135–139.
- Miller, G. A. (1947a). "Masking effect of periodically pulsed tones as a function of time and frequency," *J. Acoust. Soc. Am.* **19**, 798–807.
- Miller, G. A. (1947b). "The masking of speech," *Psychol. Bull.* **44**, 105–129.
- Miller, G. A., and Licklider, J. C. R. (1950). "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.* **22**, 167–173.
- Munson, W. A. (1947). "The growth of auditory sensation," *J. Acoust. Soc. Am.* **19**, 584–591.
- Patterson, R. D. (1976). "Auditory filter shapes derived with noise stimuli," *J. Acoust. Soc. Am.* **59**, 640–654.
- Wang, D. L. (2005). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.
- Wang, D., Kjems, U., Pedersen, M., Boldt, J., and Tunner, T. (2009). "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.* **125**, 2336–2347.
- Warren, R. M., Bashford, J. A., Jr., Healy, E. W., and Brubaker, B. S. (1994). "Auditory induction: Reciprocal changes in alternating sounds," *Percept. Psychophys.* **55**, 313–322.
- Yost, W. A., and Sheft, S. (1989). "Across-critical-band processing of amplitude-modulated tones," *J. Acoust. Soc. Am.* **85**, 848–857.